*Published online 16 November 2006* *Nucleic Acids Research, 2007, Vol. 35, Database issue* ***D879–D882***
*doi:10.1093/nar/gkl814*

# PlantQTL-GE: a database system for identifying candidate genes in rice and *Arabidopsis* by gene expression and QTL information

**Huazong Zeng[1,2], Lijun Luo[2], Weixiong Zhang[3], Jie Zhou[1], Zuofeng Li[1], Hongyan Liu[2], Tiansheng Zhu[2], Xiangqian Feng[1] and Yang Zhong[1,4,*]**

[1]School of Life Sciences, Fudan University, Shanghai, 200433, China, [2]Shanghai Agro-Biological Gene Center, Shanghai 201106, China, [3]Department of Computer Science and Engineering, Washington University in St Louis, St Louis, MO 63130, USA and [4]Shanghai Center for Bioinformation Technology, Shanghai 200235, China

Received July 31, 2006; Revised October 3, 2006; Accepted October 4, 2006

## ABSTRACT

**We have designed and implemented a web-based database system, called *PlantQTL-GE*, to facilitate quantitatine traits locus (QTL) based candidate gene identification and gene function analysis. We collected a large number of genes, gene expression information in microarray data and expressed sequence tags (ESTs) and genetic markers from multiple sources of *Oryza sativa* and *Arabidopsis thaliana.* The system integrates these diverse data sources and has a uniform web interface for easy access. It supports QTL queries specifying QTL marker intervals or genomic loci, and displays, on rice or *Arabidopsis* genome, known genes, microarray data, ESTs and candidate genes and similar putative genes in the other plant. Candidate genes in QTL intervals are further annotated based on matching ESTs, microarray gene expression data and *cis*-elements in regulatory sequences. The system is freely available at http://www.scbit.org/qtl2gene/new/.**

## INTRODUCTION

Quantitative trait locus (QTL) analysis is an effective method for locating chromosomal regions harboring genetic variants that affect a continuously distributed, polygenic phenotype (1). However, the identification of genes affecting complex traits is one of the most difficult tasks in genetics (2). The primary challenge is to identify and clone the responsive genes underlying a QTL. Recently, several computational methods, including comparative genomics, combined cross analysis, interval-specific and genome-wide haplotype analysis, have been developed for narrowing animal QTLs (3). However, no bioinformatics tool for QTL-based candidate gene identification in plants is currently available.

Rice (*Oryza sativa*) and *Arabidopsis thaliana,* model plants for monocotyledonous and dicotyledonous species, respectively, have many quantitative traits of agronomic importance and biological significance. Most of these quantitative traits are the results of interactions of multiple genetic variations as well as interactions of genetic variations and environmental factors. In recent years, many quantitative traits in rice and *A.thaliana* have been discovered by QTL mapping. Specifically, as of the end of 2005, more than 7000 QTLs controlling various complex traits have been located on different chromosome regions in rice (http://www.gramene.org/qtl/index.html).

In practice, the candidate gene approach has been applied in plant genetics in the past for the characterization and cloning of QTLs (4). It is complementary to map-based cloning for identifying the genes placed within the QTL intervals. Using the sequenced and (partially) annotated *A.thaliana* (5) and rice genomes (6–8), it now becomes feasible to retrieve genomic sequences, select candidate genes within a QTL interval, and analyze gene expression data [i.e. expressed sequence tags (ESTs) and microarray gene expression data] for discovering novel genes in rice and *A.thaliana*. As a result, the candidate gene approach has been commonly used to select appropriate candidate genes and relevant gene expression data that are possibly associated with particular QTLs. With such refined association information among candidate genes, their expression information and QTLs, further functional analysis of genes and quantitative trait (e.g. linkage analysis) can be effectively carried out.

Therefore, it is an important task to develop a bioinformatics platform for candidate gene finding that integrates the information of gene expression and QTLs. To this end, we have designed and developed an integrated database system, called PlantQTL-GE, where GE stands for Genes and Expressions, for identifying candidate genes and searching for relevant gene expression information from microarray gene expression and EST data. This resource provides a novel tool to assist the user to focus on those candidate genes that are restricted to the QTLs of interest. In practice, the user

*To whom correspondence should be addressed. Tel: +86 21 55664436; Fax: +86 21 65642468: Email: yangzhong@fudan.edu.cn

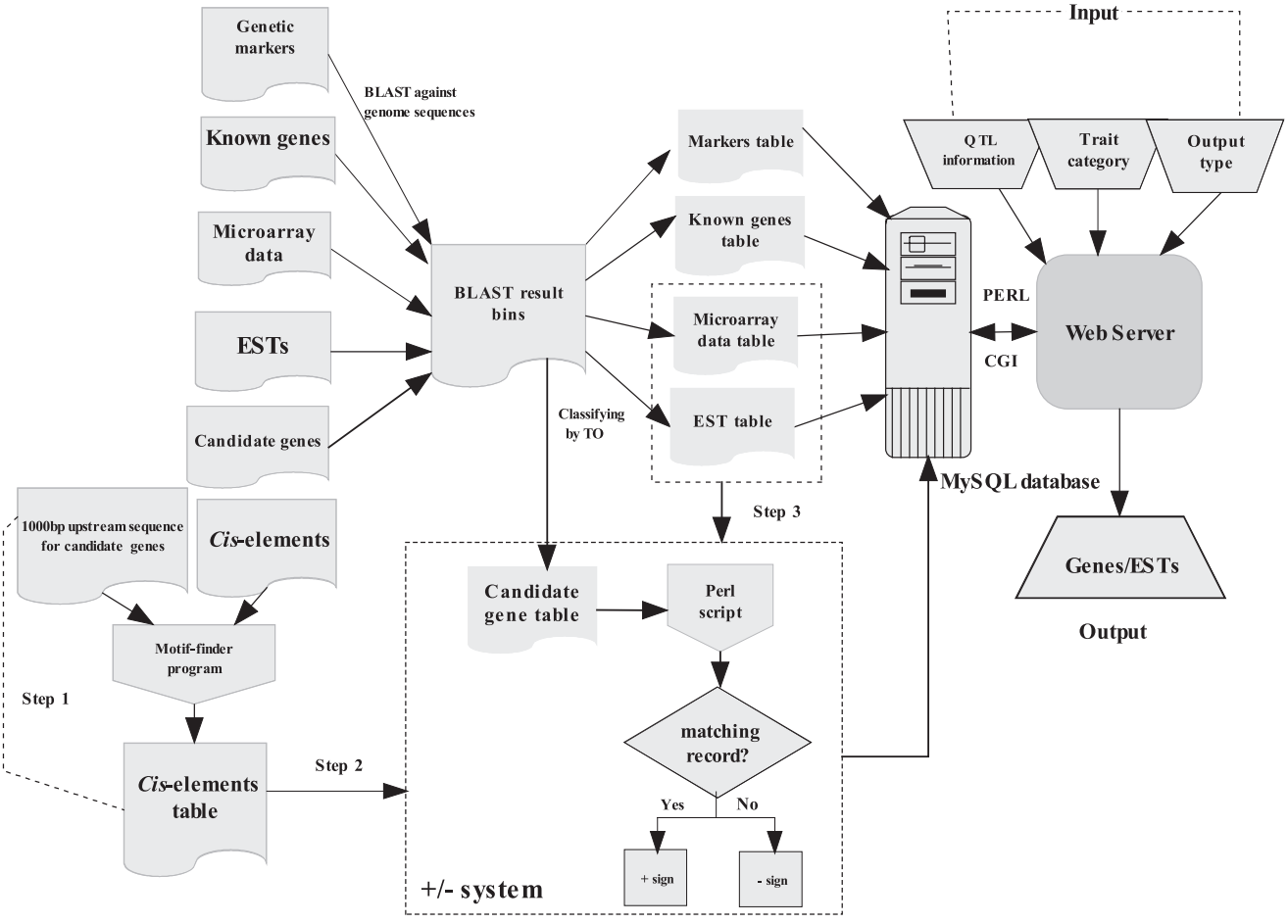


**Figure 1.** A sketch of the system architecture and data flow of PlantQTL-GE, which integrates information of genomic sequences, genes, gene expression data, ESTs, genetic markers and trait ontology.

could get a list of candidate genes located in the interested QTL region, and then isolate, map, and characterize these genes through future experiments such as differential display analysis.

## DATA SOURCES AND CONTENT

### Data sources

Various known functional genes, microarray gene expression data, ESTs, genetic markers, QTLs and candidate genes of rice and *Arabidopsis* were collected from the literature and various public databases, and organized and maintained in an integrated database in MySQL (Figure 1). Rice QTLs were further classified into seven categories and *A.thaliana* QTLs into four categories, according to the trait ontology (TO) (http://www.gramene.org/plant_ontology/index.html#to) in GRAMENE database (9). Then, information of known functional genes related to different trait categories and their corresponding EST and microarray data were extracted from literature. The microarray data included both up and down regulated genes or probes. The ESTs were retrieved from NCBI EST database (http://www.ncbi.nlm.nih.gov/dbEST/) by using relevant keywords of traits and classified based on trait ontology. Genetic markers with known genomic positions were retrieved from the marker databases in GRAMENE (release 19) (ftp://ftp.gramene.org/pub/gramene/release19/data/sequence_annotation/gff3) and TAIR (http://www.arabidopsis.org/servlets/Search?action=new_search&type=marker). Gene annotations were retrieved from TIGR (http://www.tigr.org/tdb/e2k1/osa1/data_download.shtml) for rice and TAIR (http://www.arabidopsis.org) for *A.thaliana*, respectively. The loci information of a sequence was obtained by aligning it against the genomic sequences using BLAST with an *E*-value of 1e-05. The release of rice Nipponbare genomic sequences used in the study is IRGSP Build 3.0 (http://rgp.dna.affrc.go.jp/E/IRGSP/Build3/build3.html) and the release of *A.thaliana* genomic sequences is TIGR v5 (http://www.tigr.org/tdb/e2k1/ath1/index.shtml).

To help select candidate genes, we used known *cis*-regulatory elements to annotate candidate genes. In particular, we took all 25 conserved *cis*-elements from the PLACE plant motif database (http://www.dna.affrc.go.jp/PLACE/) which are associated with quantitative traits defined in trait ontology mentioned above. These 25 motifs are listed at our website http://www.scbit.org/qtl2gene/new/Cis_element.html.

As of June, 2006, PlantQTL-GE contained 1558 and 1896 known genes, 3633 and 9270 microarray data entries, 883 598 and 162 596 ESTs, 21 523 and 2283 genetic markers,

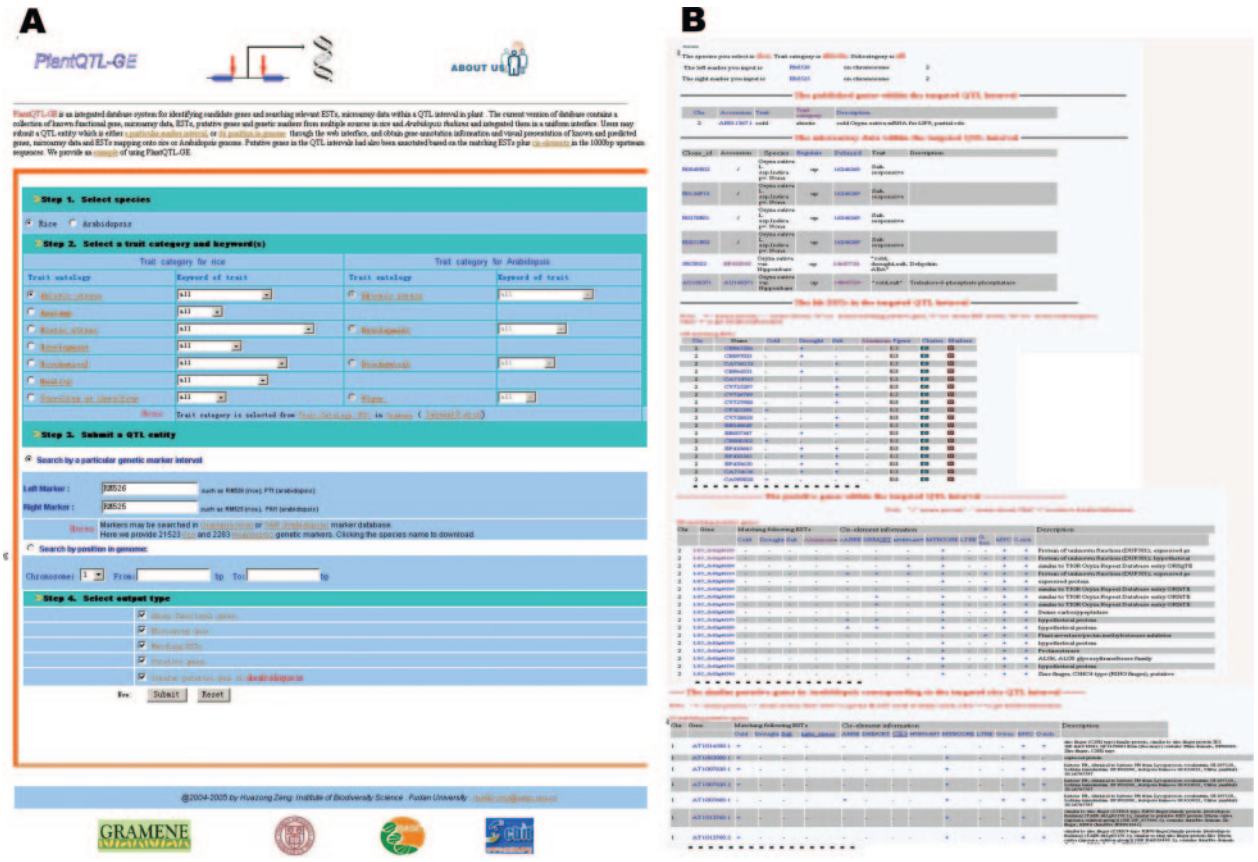


**Figure 2.** (**A**) The interface of PlantQTL-GE; (**B**) A snapshot of the result of searching for genetic elements related to a QTL between markers RM526 and RM525, which controls dry-weight of grains.

as well as 58 687 and 31 527 annotated genes for rice and *Arabidopsis*, respectively.

### QTL-based candidate gene identification

Since EST and microarray gene expression data as well as *cis*-elements have been proven to be very useful for identifying candidate genes of interest (10,11), we used them to discover and annotate candidate genes. For the genes within the two genetic markers delineating a QTL of interest, our method executes the following three-step procedure:

(i) In Step 1, sequences of length 1000 bp upstream of rice and *Arabidopsis* candidate genes within the interval of a given QTL are retrieved from GRAMENE and TAIR database, respectively. A motif-scanning method named Motif-finder (12) is then applied to locate the *cis*-elements in the upstream sequences of targeted genes. The resulting genomic loci and occurrence number of each *cis*-element in the promoter regions of the genes are stored in a *cis*-element table in the MySQL database.

(ii) The candidate genes are subsequently annotated using the identified *cis*-elements, in Step 2. Specifically, a gene is annotated as plus (+) or minus (−) with respect to a *cis*-element if its promoter contains the *cis*-element or it does not, respectively. Such *cis*-element annotations are stored in the *cis*-element fields in a gene annotation table.

(iii) Similarly, in Step 3 the genes are also annotated with respect to gene expression data (i.e. EST and microarray gene expression data) which are relevant to the quantitative trait of consideration. A sequence is annotated as plus (+) with respect to a gene expression if it matches to the corresponding EST or microarray probe, or as minus (−), otherwise. The gene expression annotation information is recorded in the gene expression fields of the annotation table.

### Similarity searches of putative genes

We compared the sequences of *Arabidopsis* genes against the genomic sequences of rice using BLAST with a stringent *E*-value of 1e-15. We carried out a reciprocal analysis by comparing the sequences of rice genes against the genomic sequences of *A.thaliana* with the same stringency level. As a result, 2445 *Arabidopsis* putative genes were mapped onto the rice genome and 4588 rice putative genes were mapped onto the *Arabidopsis* genome. The results are stored in the MySQL database.

## SYSTEM IMPLEMENTATION AND QUERY INTERFACE

We have implemented the PlantQTL–GE system, which is available online at http://www.scbit.org/qtl2gene/new/. It has a web-based query interface, supported by a MySQL database with enhanced Perl scripts.

The interface, shown in Figure 2A, allows the user to search information about candidate genes and relevant ESTs and microarray gene expression data within a QTL interval in rice or *A.thaliana*. The following steps should be followed to retrieve genetic information associated with a QTL of interest

(i) Select a plant species. Currently, rice and *A.thaliana* are available.
(ii) Select a query option, which can be *linked genetic markers* flanking the targeted QTL interval or *chromosomal positions* in the selected genome, and then input two markers or the start and stop genomic positions of the markers.
(iii) Select a trait category and one or more keywords on traits related to the targeted QTL specified by the user.
(iv) Select an output type, which can be a combination of known genes, microarray data, ESTs, candidate genes, and similar putative genes in the other species.

The result of a query includes all selected genes, microarray and/or ESTs within the targeted QTL region. The entities in the output are also hot linked in that further detailed information of the entities can be obtained by following the corresponding hyperlinks.

A snapshot of a case study of searching for genetic elements related to a QTL between markers RM526 and RM525, which controls dry-weight of grains (an abiotic related trait) (13), is shown in Figure 2B. The complete output of this example is available in the supplementary material in our website (http://www.scbit.org/qtl2gene/new/example.pdf).

## SYSTEM UPDATE AND FUTURE WORK

We will update PlantQTL-GE database bimonthly to accommodate new information of genes (from the Entrez-Nucleotide database, http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Nucleotide), microarray data (from public databases and literature) and ESTs (from the EST database, http://www.ncbi.nlm.nih.gov/dbEST). To do so, we have written and tested a Perl script based on the NCBI eUtils tools (http://eutils.ncbi.nlm.nih.gov/).

We also plan to expand PlantQTL-GE system to support more plant species, such as *Populus trichocarpa* (http://genome.jgi-psf.org/Poptr1/Poptr1.home) whose whole genome sequence is now available. In addition, we will enhance the *cis*-element annotation by including motifs identified by a motif-finding method, such as WordSpy (14), on all genes responsive to biotic and abiotic stresses.

## ACKNOWLEDGEMENTS

The research was supported by grants from The Rockefeller Foundation, Shanghai Science and Technology Committee (05DJ14008 and 03DJ14015) and National Science Foundation of China (30570109). W.Z. was supported in part by United States of America, National Science Foundation Grants ITR/EIA-0113618 and IIS-0535257. We thank Shanghai Center for Bioinformation Technology for the webserver and Dr Y. X. Li, S. W. Yu, L. Chen, and P. Hao for comments and suggestions. Funding to pay the Open Access publication charges for this article was provided by Chinese Ministry of Education key program 106068 and program for NCET.

*Conflict of interest statement*. None declared.

## REFERENCES

1. Abiola,O., Angel,J.M., Avner,P., Bachmanov,A.A., Belknap,J.K., Bennett,B., Blankenhorn,E.P., Blizard,D.A., Bolivar,V., Brockmann,G.A. *et al*. (2003) The nature and identification of quantitative trait loci: a community's view. *Nature Rev. Genet*., **11**, 911–916.
2. Risch,N.J. (2000) Searching for genetic determinants in the new millennium. *Nature*, **405**, 847–856.
3. DiPetrillo,K., Wang,X., Stylianou,I.M. and Paigen,B. (2005) Bioinformatics toolbox for narrowing rodent quantitative trait loci. *Trends Genet*., **21**, 683–692.
4. Pflieger,S., Lefebvre,V. and Causse,M. (2001) The candidate gene approach in plant genetics: a review. *Mol. Breed*., **7**, 275–291.
5. The *Arabidopsis* Genome Initiative. (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796–815.
6. Goff,S.A., Ricke,D., Lan,T.H., Presting,G., Wang,R., Dunn,M., Glazebrook,J., Sessions,A., Oeller,P., Varma,H. *et al*. (2002) A draft sequence of the rice Genome (*Oryza sativa* L. ssp. *japonica*). *Science*, **296**, 92–100.
7. Yu,J., Hu,S., Wang,J., Wong,G.K., 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.
8. Matsumoto,T., Wu,J.Z., Kanamori,H., Katayose,Y., Fujisawa,M., Namiki,N., Mizuno,H., Yamamoto,K., Antonio,B.A., Baba,T. *et al*. (2005) The map-based sequence of the rice genome. *Nature*, **436**, 793–800.
9. Jaiswal,P., Ware,D., Ni1,J., Chang,K., Zhao,W., Schmidt,S., Pan,X., Clark,K., Teytelman,L., Cartinhour,S. *et al*. (2002) Gramene: development and integration of trait and gene ontologies for rice. *Comp. Funct. Genom*., **3**, 132–136.
10. Boguski,M.S., Tolstoshev,C.M. and Bassett,D.E.,Jr (1994) Gene discovery in dbEST. *Science*, **265**, 1993–1994.
11. Zhang,W., Ruan,J., Ho,T.-H., You,Y., Yu,T. and Quatrano,R.S. (2005) *Cis*-regulatory element based targeted gene finding: genome-wide identification of abscisic acid- and abiotic stress-responsive genes in *Arabidopsis thaliana*. *Bioinformatics*, **21**, 3074–3081.
12. Shah,N.H., King,D.C., Shah,P.N. and Fedoroff,N.V. (2003) A tool-kit for cDNA microarray and promoter analysis. *Bioinformastics*, **19**, 1846–1848.
13. Liu,H., Zou,G., Liu,G., Hu,S., Li,M., Yu,X., Mei,H. and Luo,L. (2005) Correlation analysis and QTL identification for canopy temperature, leaf water potential and spikelet fertility in rice under contrasting moisture regimes. *Chin. Sci. Bull*., **50**, 317–326.
14. Wang,G. and Zhang,W. (2006) A steganalysis-based approach to comprehensive identification and characterization of functional regulatory elements. *Genome Biol*., **7**, R49.