PCP
PLANT & CELL PHYSIOLOGY

Special Issue – Regular Paper

# PosMed-plus: An Intelligent Search Engine that Inferentially Integrates Cross-Species Information Resources for Molecular Breeding of Plants

Yuko Makita[1,3], Norio Kobayashi[1,3], Yoshiki Mochizuki[1], Yuko Yoshida[1], Satomi Asano[1], Naohiko Heida[1], Mrinalini Deshpande[1], Rinki Bhatia[1], Akihiro Matsushima[1], Manabu Ishii[1], Shuji Kawaguchi[1], Kei Iida[1], Kosuke Hanada[2], Takashi Kuromori[2], Motoaki Seki[2], Kazuo Shinozaki[2] and Tetsuro Toyoda[1,*]

[1]Bioinformatics And Systems Engineering (BASE) division, RIKEN, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa, 230-0045 Japan

[2]Plant Science Center (PSC), RIKEN, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa, 230-0045 Japan

**Molecular breeding of crops is an efficient way to upgrade plant functions useful to mankind. A key step is forward genetics or positional cloning to identify the genes that confer useful functions. In order to accelerate the whole research process, we have developed an integrated database system powered by an intelligent data-retrieval engine termed PosMed-plus (Positional Medline for plant upgrading science), allowing us to prioritize highly promising candidate genes in a given chromosomal interval(s) of *Arabidopsis thaliana* and rice, *Oryza sativa*. By inferentially integrating cross-species information resources including genomes, transcriptomes, proteomes, localizomes, phenomes and literature, the system compares a user's query, such as phenotypic or functional keywords, with the literature associated with the relevant genes located within the interval. By utilizing orthologous and paralogous correspondences, PosMed-plus efficiently integrates cross-species information to facilitate the ranking of rice candidate genes based on evidence from other model species such as Arabidopsis. PosMed-plus is a plant science version of the PosMed system widely used by mammalian researchers, and provides both a powerful integrative search function and a rich integrative display of the integrated databases. PosMed-plus is the first cross-species integrated database that inferentially prioritizes candidate genes for forward genetics approaches in plant science, and will be expanded for wider use in plant upgrading in many species.**

**Keywords:** Omics • Omic space • Superbrain • Web application.

**Abbreviations:** NER, named entity recognition; QTL, quantitative trait locus; RFLP, restriction fragment length polymorphism; SSR, simple sequence repeat.

## Introduction

Molecular breeding is an efficient way to upgrade plant functions that are pertinent to solving a variety of problems facing mankind, including changes in the environment and shortages of food, energy and bio-based materials. The efficiency of molecular breeding depends on our ability to access and utilize the available molecular information from the viewpoint of these valuable phenotypes. The recent accumulation of plant genome sequences, and increasing knowledge of gene functions in the published literature and various omics databases, is expected to accelerate the efficiency of molecular breeding. Plant-upgrading science, or molecular-based science for upgrading plant functions, requires a coherent information platform that integrates the available molecular knowledge and plant functions that have been investigated with forward genetics approaches.

To construct such a coherent integrated information tool for plants, we focused on *Arabidopsis thaliana* and rice, *Oryza sativa*, both well studied land plants. In particular, the genomic sequence for Arabidopsis has been determined

[3]These authors contributed equally to this work.

*Corresponding author: E-mail, toyoda@base.riken.jp; Fax: +81-45-503-9553.

*Plant Cell Physiol.* 50(7): 1249–1259 (2009) doi:10.1093/pcp/pcp086, available online at www.pcp.oxfordjournals.org

(Arabidopsis Genome Initiative 2000) and various levels of genome-wide information have been accumulated, such as a collection of full-length cDNAs (Seki et al. 2002, Yamada et al. 2003), expression profiles (Goda et al. 2008, Matsui et al. 2008), proteomes (Baerenfaller et al. 2008) and interactomes (Cui et al. 2006). Reverse genetics approaches, whereby each gene is systematically knocked-out or knocked-in to observe the resulting phenotypic changes, have elucidated the relationships between phenotype and the responsible genes genome-wide (Kuromori et al. 2004, Kondou et al. 2009). These reverse genetics experiments are particularly efficient for model plants, such as Arabidopsis and rice, and the resulting data are accumulating in both public databases (Hirochika et al. 2004, Kuromori et al. 2006, Swarbreck et al. 2008) and the literature (Coletti et al. 2001). The results are useful not only for understanding gene functions, but also for upgrading the valuable phenotypes of crops by altering their genomes: molecular breeding. On the other hand, forward genetics approaches have elucidated many quantitative trait loci (QTLs) of agriculturally important traits, such as increasing grain number (Ashikari et al. 2005), grain width and weight (Song et al. 2007, Shomura et al. 2008), salt tolerance (Ren et al. 2005) and reducing heading date (Takahashi et al. 2001, Doi et al. 2004).

Thus, the coherent information system needed for molecular breeding must be capable of realizing the intelligent inferential association of phenotypes and genes, through various types of input information including experimental data described in the literature and molecular networks inferred by bioinformatics. This system is necessary to fill the gap that exists between the knowledge accumulated by reverse genetics experiments in model plants and the QTL knowledge narrowed down by forward genetics approaches for useful crops (**Fig. 1**). Our proposed data integration model is analogous to an artificial intelligence-oriented artificial neural network approach, categorized as connectionist–symbolic hybrid integration (Sun et al. 1997). Here we introduce PosMed-plus (Positional Medline for plant-upgrading science), which is an application of the GRASE search engine (Kobayashi et al. 2008), to the linked data generated from various plant science databases, in order to accelerate the prioritization of candidate genes for positional cloning in crops.

PosMed-plus is a plant science version of PosMed, which was initially established to assist in candidate selection for positional cloning work in mice, humans and rats (Yoshida et al. 2009), and has been widely used to prioritize candidates to follow forward genetics approaches restricting the chromosomal intervals responsible for diseases (Moritani et al. 2006, Kato et al. 2008). Among several software tools available to prioritize positional candidate genes (van Driel et al. 2005, Adie et al. 2006, Aerts et al. 2006, Seelow et al. 2008), PosMed was evaluated as highly effective in comparison with two other similar tools designed for the candidate selection in forward genetics (Thornblad et al. 2007).

PosMed-plus is the first information tool to prioritize candidate genes for forward genetics approaches in plant science, and will contribute to a wider use of such systems in plant-upgrading sciences in many plant species. PosMed-plus is also integrated seamlessly with other data-browsing systems to supply both a powerful integrative search function and a rich integrative display of the integrated databases. For each candidate gene, the accumulated omics information from genomes to phenomes is displayed by OmicBrowse (Toyoda et al. 2007) and can be downloaded by OmicDownload which joins the table of candidate genes selected by PosMed-plus with tables of other annotations such as full-length cDNAs, microarray and whole-genome tiling array data, genome annotations, genetic markers, polymorphisms and gene ontology (Matsushima et al. 2009). PosMed-plus will be continuously maintained by RIKEN in order to make a significant contribution to a wide range of plant sciences, and expanded to utilize other plant information resources such as data for wheat and poplar and a manual association of literature references with genes of other plant species. PosMed-plus is available at http://omicspace.riken.jp/.

## Results

### A neural network representation of the statistical algorithm for searching complex literature and omics data

PosMed-plus prioritizes candidate genes for positional cloning by employing our original database search engine GRASE (Kobayashi et al. 2008). GRASE uses an inferential process similar to an artificial neural network comprising documental neurons (or 'documentrons') that represent each document contained in databases such as MEDLINE (**Fig. 2a**). Given a user-specified query, PosMed-plus initially performs a full-text search of each documentron in the first layer artificial neurons, and then calculates the statistical significance of the connections between the hit documents and the second layer artificial neurons representing each Arabidopsis gene. When a chromosomal interval(s) in Arabidopsis is specified, PosMed-plus explores the second layer and third layer artificial neurons representing genes within the chromosomal interval, by evaluating the combined significance of the connections from the hit documentrons to the genes. When a chromosomal interval(s) in rice is specified, PosMed-plus further explores the fourth layer artificial neurons representing rice genes within the chromosomal interval by utilizing orthologous and paralogous correspondence between Arabidopsis genes and rice genes. For the output display, PosMed-plus shows the ranked genes with evidence documents highlighted with the users' keyword.

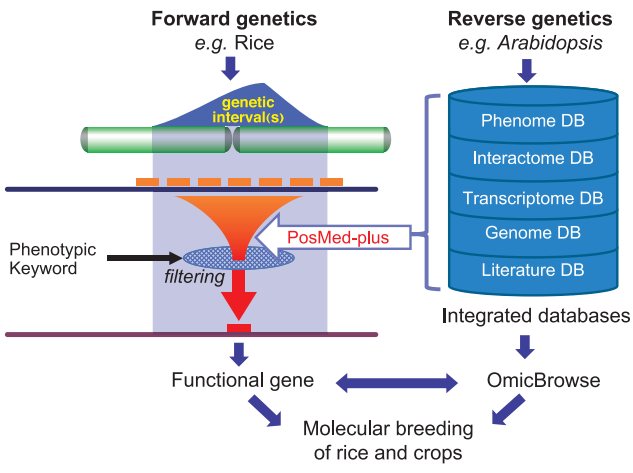


**Fig. 1** PosMed-plus accelerates forward genetics gene discoveries (left half of the chart) by integrating the omics knowledge collected from reverse genetics (right half of the chart). PosMed-plus assists users to narrow down the candidate responsible genes from those existing within the chromosomal intervals. OmicBrowse assists users to look into every piece of detailed information of each candidate gene. The entire system is designed to support coherently molecular breeding research and plant-upgrading sciences.

PosMed-plus is, therefore, a powerful tool that immediately ranks the candidate genes by connecting phenotypic keywords to the genes, with connections representing both gene–gene interactions and other biological interactions such as metabolite–gene, phenotype–gene, subcellular localization–gene, co-expression, protein–protein interactions (PPIs), and ortholog and paralog data. By utilizing orthologous and paralogous connections, PosMed-plus can facilitate the ranking of rice genes based on evidence found in other plant species. The system is an artificial superbrain (Yoshida et al. 2009) that has already learned a vast amount of biological knowledge ranging from genomes to phenomes (or 'omic space'), and supports the prioritization of positional candidate genes in both rice and *A. thaliana*.

## Manual curation work connecting Arabidopsis genes to the literature

The accuracy of PosMed-plus is strongly correlated with its ability to make correct associations between each gene and documents. This is because GRASE utilizes these associations to execute direct searches and inference searches that are supported by co-citations. To increase Posed-plus's accuracy, we employed manual curation to make connections between Arabidopsis genes and the literature. Our original curation method is based on named entity recognitions (NERs; see Materials and Method for details). Rather than connecting every literature reference to genes, specialized curators create search rules to retrieve all the correct references from titles, abstracts and MeSH terms. In order to validate the effectiveness of our method, we compared our curation results with TAIR annotation (**Table 1**).

The number of total MEDLINE hits for PosMed-plus was 28% larger than the TAIR data set. This is because our method has an advantage over TAIR in terms of updating new data. On the other hand, TAIR has 2.4 times more genes with associated MEDLINE records. One reason for this is that TAIR annotators extract gene–reference relationships not only based on abstracts but also using the whole article text. As a result, TAIR sometimes connects many genes to each article describing omic research. In contrast, PosMed-plus only focuses on literature with gene and/or synonym names in the abstract. Generally, non-omic research that addresses the functions of a small number of genes will mention the gene names in the abstract. Our curation results suggest that only around 15% of the total genes have been functionally analyzed in Arabidopsis. We also compared the number of gene–reference pairs. Approximately 65% of PosMed-plus data matched with TAIR data. Although PosMed-plus has more MEDLINE references, TAIR has a greater number of gene–reference pairs. This is because TAIR tends to extract many genes from a single source of literature.

## General search paths of PosMed-plus

Using the search functionalities of GRASE, PosMed-plus supports the following four types of searches:

(i) Direct search: GRASE searches genes located in the user's chromosomal interval by performing a full-text search against the set of databases with the user's keyword, i.e. the following search path is realized: keyword→document (e.g. literature)→gene→chromosomal interval (**Fig. 2b**, top).

(ii) Inference search: by applying gene–gene relationships over the genes extracted by a direct search that is not located in the user's chromosomal interval, GRASE discovers further genes that are indirectly related to the keyword via gene–gene relationships, i.e. the following search path is realized: keyword→document (e.g. literature)→gene1→gene2→chromosomal interval. The link between gene1 and gene2 is supported by omics data (**Fig. 2b**, bottom).

(iii) Cross-species search: this is an extension of the direct search (i) to the rice genome. The connections from Arabidopsis genes to rice genes are supported by ortholog and paralog data (**Fig. 2c**, top).

(iv) Cross-species inference search: this is an extension of the inference search (ii) to the rice genome. As for (iii) above, ortholog and paralog data connect Arabidopsis genes to rice genes (**Fig. 2c**, bottom).

In the final stage, these types of search results are integrated into a ranked gene list by species. In the following section we describe examples to illustrate the powerful applications of PosMed-plus.

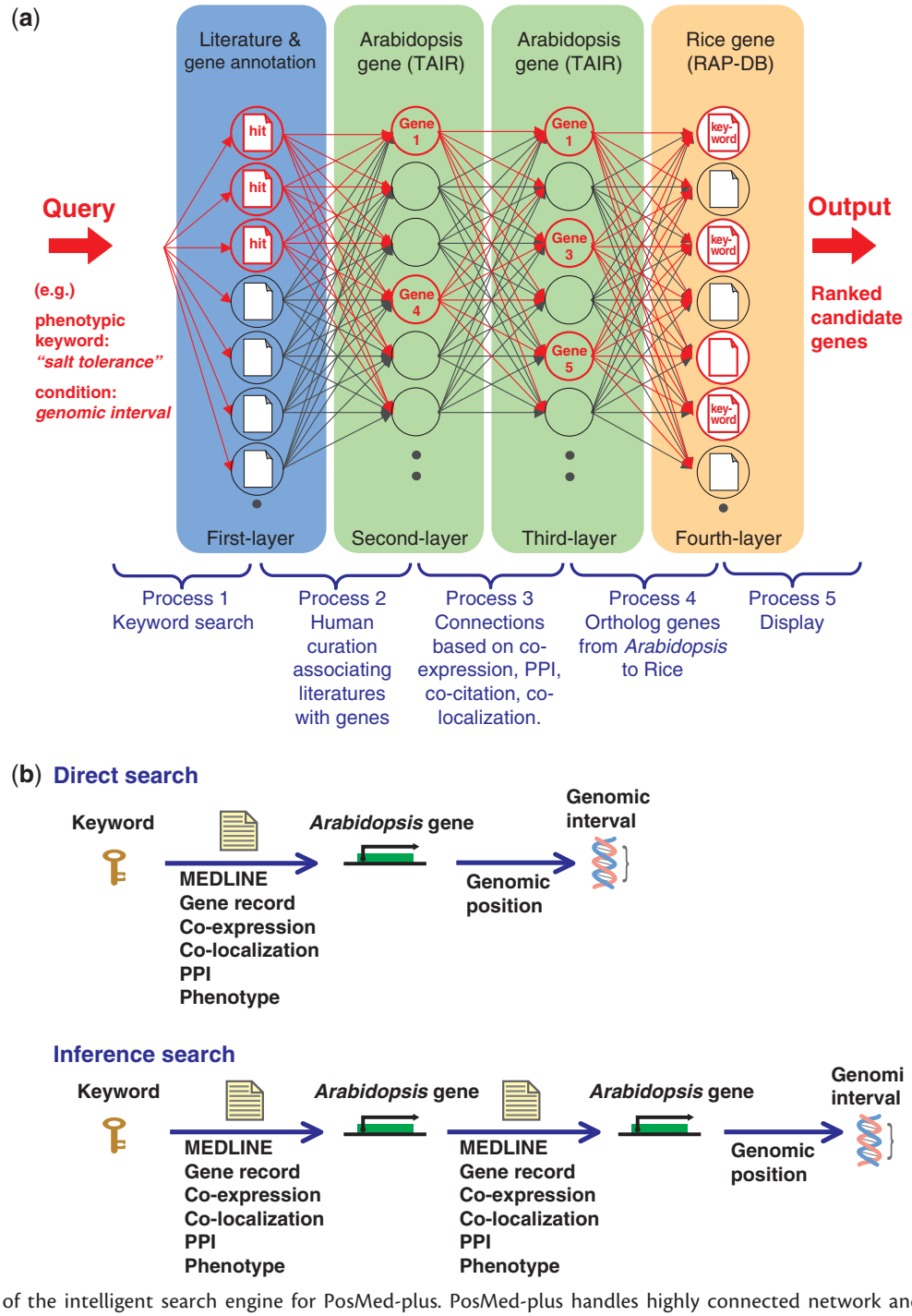


**Fig. 2** (a) Model of the intelligent search engine for PosMed-plus. PosMed-plus handles highly connected network and reply candidate genes depending on the correlation with the phenotypic keyword. Further details are given in Yoshida et al. (2009) and Kobayashi et al. (2008). (b) Data flow of PosMed-plus search and comparison with the direct search and the inference search in Arabidopsis. (c) Data flow of PosMed-plus cross-species search.

*1. Search with user-specified keywords and chromosomal intervals.* A typical application of PosMed-plus is searching with phenotypic keywords and chromosomal intervals suggested by linkage analysis. As an example, we retrieved 'drought tolerance'-related genes in the chromosomal interval from 0 to 10 Mbp on chromosome 6 in the rice genome (**Fig. 3A**). In this example, PosMed-plus retrieved 24 candidate genes ranked by statistical significance between the

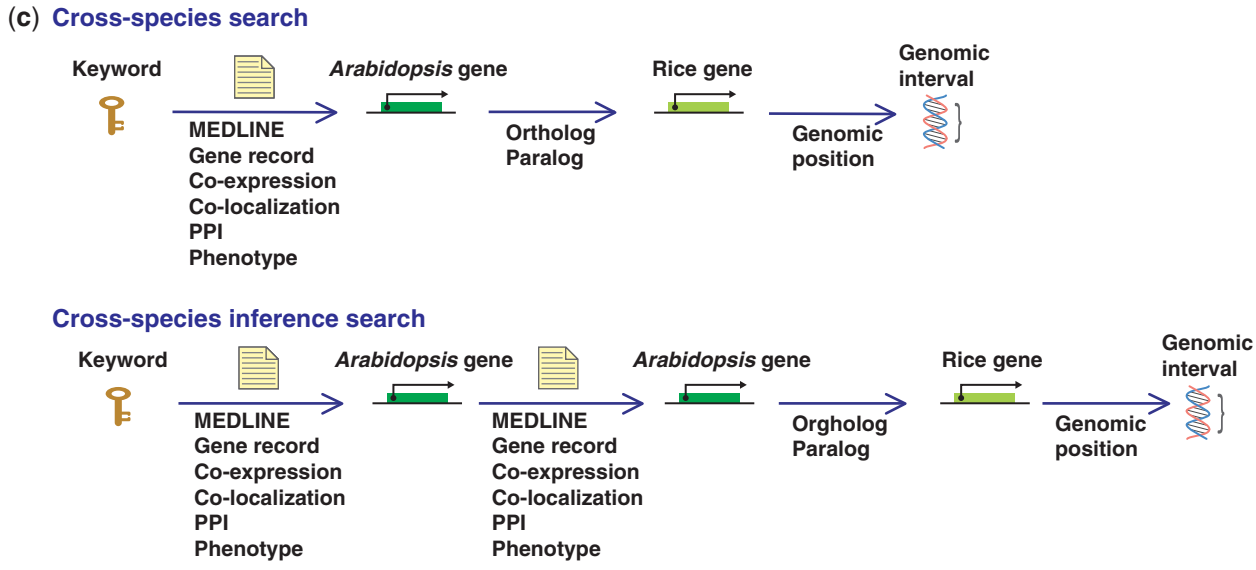


**Fig. 2** Continued

**Table 1** Comparison of curation results between PosMed and TAIR

| | No. of MEDLINE data | No. of genes | No. of gene–reference pairs |
|---|---|---|---|
| PosMed | 15,165 | 4,773 | 30,155 |
| TAIR | 11,894 | 11,503 | 38,905 |

user's keyword and each gene. Although PosMed-plus found >200,000 documents, it returned results within 1 s. Users can download all the candidate genes together with the associated gene annotations, using the 'download rank list' button in the left blue box (**Fig. 3D**). PosMed-plus also supports an 'expert mode' that allows users to select possible search paths and confirm the number of resulting genes for each search path. By clicking on a gene name listed in the gene search result page shown in **Fig. 3B**, PosMed-plus shows the supporting evidence for each candidate gene. To confirm the expression pattern of candidate genes with a genome browser, we provide a link to OmicBrowse from the gene location (**Fig. 3C**). OmicBrowse covers four genome versions for Arabidopsis and two for rice, and each genome is mapped to omic-type databases and a total of 78 data sources (**Table 2**).

For chromosomal intervals, users can select according to location (e.g. 10 Mbp) or in relation to restriction fragment length polymorphism (RFLP) and simple sequenc repeat (SSR) markers.

2. *Search with phenotypic keywords*. PosMed-plus also allows users to find genes related to phenotypic keywords. For example, searching with the keyword 'rumpled leaves', PosMed-plus shows four known cases via the direct search and one new candidate gene via the inference search. For the four known cases, PosMed-plus shows the link to RAPID (RIKEN Arabidopsis Phenome Information Database) and users can confirm the phenotypes with pictures. PosMed-plus also shows the evidence documents in the inference path to the AT1G51500 candidate gene. In this case, AT1G51500 is retrieved via the AT1G17840 gene that is one of the four known genes found in the direct search. They are highly connected with co-expression, PPI and co-citation data.

3. *Reference search with gene IDs*. It is difficult to retrieve all the appropriate references based on gene names, because of the wide variation of synonyms. Moreover, sometimes the same abbreviated names are used for functionally different genes, causing false-positive hits. In PosMed-plus, we carefully extracted these gene–reference relationships manually. Therefore, users can retrieve the curated results with the gene ID (e.g. AGI code) even if abstracts do not contain the gene ID itself.

4. *Search for omics data*. As shown in **Fig. 4**, PosMed-plus integrates various data such as gene annotations, co-expressions, subcellular localizations, phenotypes and PPIs. Users can select any document set (the default setting is to search everything) and retrieve the required data, all with the same interface. PosMed-plus links not only to the original databases but also to our genome browser, OmicBrowse. OmicBrowse also assists users in accessing various omics data and in downloading the data (Matsushima et al. 2009).

## In silico positional cloning after QTL analysis in rice

To validate the efficacy of PosMed-plus, we checked whether PosMed-plus could successfully retrieve correct genes that


**Fig. 3** An example search result for *Oryza sativa* genes against the query keyword 'drought tolerance' and the genomic interval between 0 and 10 Mbp on chromosome 6 in the IRGSP build 4 genome. Users can apply their queries at the top of the output display (A). To select genomic interval visually, PosMed-plus cooperates with the Flash-based genomic browser OmicBrowse (Matsushima et al. 2009). The tab labeled 'All Hits' (B) shows a list of selectable document sets to be included in the search. Changing the 'Simple Mode' to the 'Expert Mode' allows fine retrievals. PosMed search results are ranked in (C). Clicking the ranked gene names shows all the documents (see **Fig. 4**). The 'Position' links guide users to see the gene on our original genome browser, OmicBrowse. Users can download at most 300 candidate genes and their annotations from (D).

**Table 2** Data description for PosMed-plus

| | No. of documents | Data sources | Data contents | Reference |
|---|---|---|---|---|
| MEDLINE | 69,060 | MEDLINE | MEDLINE title, abstract and MeSH term | Coletti et al. (2001) |
| Arabidopsis gene record | 33,003 | TAIR, UniProt | Gene descriptions (annotations) | Swarbreck et al. (2008), UniProt Consortium (2009) |
| Rice gene record | 29,389 | RAP-DB | Gene descriptions (annotations) | Tanaka et al. (2008) |
| At co-expression | 44,082 | ATTED-II | Microarray based co-expression prediction | Obayashi et al. (2009) |
| At localization | 8,404 | SUBA2 | Experimentally validated subcellular localizatioin | Heazlewood et al. (2007) |
| AtPPI | 24,418 | AtPID | Protein–protein interaction | Cui et al. (2008) |
| | 214 | RAPID | RIKEN Arabidopsis Phenome Information DB | Kuromori et al. (2006) |
| At phenotype | 1,697 | TAIR | Phenotype informations from TAIR | Swarbreck et al. (2008) |
| | 1,784 | Literature | Manually collected original data | |
| Homologous genes | 1,553,922 | Original data | Homologous genes between Arabidopsis and rice | Hanada et al. (2008) |
| Rice markers | 1,712 | RAP-DB | RFLP marker | Harushima et al. (1998) |
| | 15,623 | | SSR marker | McCouch et al. (2002) |

have been identified by QTL analysis. Three examples are described below.

Ren *et al.* (2005) isolated the *SKC1* gene and through QTL analysis found that it encoded an $Na^+$-selective transporter. In this example, we need to prioritize candidate genes without the functionally related keyword 'transporter'. Instead of the functional keyword, we retrieved genes with the phenotypic keyword 'salt tolerance' and selected the genomic interval between the markers C955 and E50811 on chromosome 1. PosMed-plus returned the Os01g0307500 (cation transporter family protein) gene with a high ranking. This is because the keyword 'salt tolerance' was mapped to the sodium ion transmembrane transporter gene AT4G10310, and Os01g0307500 was suggested as a homolog of AT4G10310.

Using a no-pollen type of male-sterile mutant (*xs1*), Zuo *et al.* (2008) revealed that mutant microspores are abnormally condensed and agglomerated to form a deeply stained cluster at the late microspore stage. This results in cessation of the microspore vacuolation process, and, therefore, the mutant forms lack functional pollen. This mutation is controlled by a single recessive gene, termed *VR1* (vacuolation retardation 1), which is located between the molecular markers RM17411 and RM5030 on chromosome 4. We searched candidate genes with a phenotypic keyword 'sterility' in the suggested chromosome region. PosMed-plus suggested the Os04g0605500 gene (similar to calcium-transporting ATPase) as the homolog of the Arabidopsis calcium-transporting ATPase, AT3G21180. Since Schiøtt et al. (2004) found that mutation of AT3G21180 results in partial male sterility, we conclude that PosMed-plus found an appropriate candidate.

Lastly, Zhang et al. (2008) found a male sterility mutant of anther dehiscence in advance, *add(t)*, between the markers R02004 and RM300 on chromosome 2. In this search, PosMed-plus returned RNA-binding region RNP-1, Os02g0319100 and Disease resistance protein family protein, Os02g0301800 with strong homology with Arabidopsis genes. PosMed-plus retrieved the Os02g0319100 gene as a homolog of Arabidopsis mei2-like protein 5, AT1G29400. As supportive evidence, Kaur et al. (2006) showed that multiple mutants of all the Arabidopsis mei2-like (AML) genes displayed a sterility phenotype. The other candidate gene, Os02g0301800, was derived via an inference search. First, PosMed-plus retrieved the keyword 'sterility' in a document describing the AT2G26330 gene. Then, AT2G26330 was linked to AT5G43470 supported by three co-citations. Finally, Os02g0301800 was returned as a homolog of AT5G43470. PosMed-plus originally suggested the Os02g0301800 gene because AT2G26330 is linked to the keyword 'sterility' in a document. However, this document states that AT2G26330 causes aberrant ovule development and female-specific sterility. Since Zhang et al. (2008) focused on male sterility, we conclude that Os02g0319100 is the appropriate candidate.

## Discussion

PosMed has been widely used to prioritize candidate genes after QTL analysis in mice and successfully identify responsible genes, as reported previously. In this paper, we aimed to create a supportive tool for molecular breeding in plants, and describe an extension of PosMed to the model plants *A. thaliana* and *O. sativa*. PosMed-plus is a useful tool to

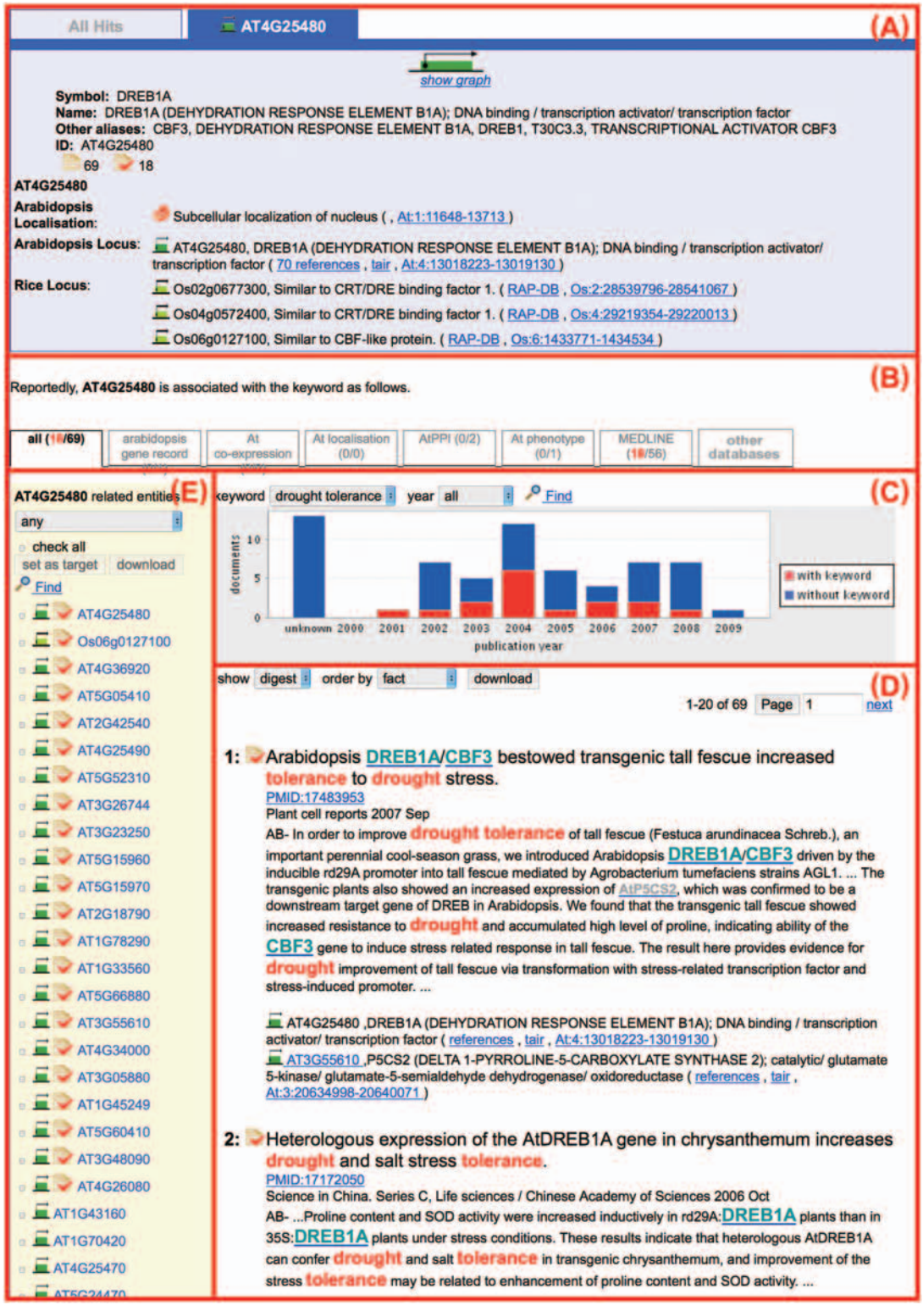


**Fig. 4** Detailed documents screen on PosMed-plus. This page shows document sets for the AT4G2580 gene, linked from the best hit (Os06g0127100) in **Fig. 3** (C). Gene descriptions are shown at (A). Users can select the type of documents from the Arabidopsis gene record, At co-expression, At localization, AtPPI, At phenotype or MEDLINE (B). Additionally, several types of transcriptome data such as tilling array and full-length cDNA data, genome annotations and markers are stored at the 'other databases' tab (B). The bar chart represents the number of related documents per year. Red shows the number of documents with a user-specified keyword and blue shows the number of documents without a user-specified keyword. All documents are shown at (D). The AT4G2580-related genes are listed in (E).

assist positional cloning in silico. At the same time, PosMed-plus integrates various kinds of omics data and assists users by allowing them to access several omics databases at a time.

In order to expand our positional cloning support system to other important crop plants, we have been preparing ortholog and paralog information (Hanada et al. 2008). We hope and expect that PosMed-plus will contribute towards solving many of the world's environmental and food problems by supporting QTL analysis for useful plants.

## Materials and Methods

### Data source

In order to construct PosMed-plus in Arabidopsis, we combined the following four kinds of omics data (**Table 2**). First, genome and related functional annotations were obtained from TAIR (Swarbreck et al. 2008) and UniProt (UniProt Consortium 2009). For transcriptome data, we used co-expression information derived from ATTED-II (Obayashi et al. 2009). Using 1,388 samples of GeneChip data, ATTED-II calculated the geometric mean of the correlation rank of gene1 to gene2 and of gene2 to gene1. For interactome data, we used PPI data from AtPID (Cui et al. 2008) and subcellular localization from SUBA (Heazlewood et al. 2007). AtPID (*Arabidopsis thaliana* Protein Interactome Database) integrated their data from several bioinformatics prediction methods and manually collected information from the literature. SUBA provide their subcellular localization data from multiple data sources, such as mass spectrometry, green fluorescent protein (GFP) and manually collected data. Currently, SUBA provide subcellular localization information for 8,068 non-redundant proteins which represents >25% of all Arabidopsis genes. Additionally, we collected phenome data from the following three data sources: RAPID, TAIR and our own manual collection from the literature. After combining these three databases, we held phenotype information for 2,500 non-redundant proteins.

Since OmicBrowse is designed as a scalable system for maintaining numerous genome annotation data sets, it already combined 74 databases and their different versions (**Table 3**). In addition to genome and transcriptome data, OmicBrowse contains marker, ontology, polymorphism and other data (Toyoda et al. 2007, Matsushima et al. 2009).

### Manual high-accuracy curation for mapping from Arabidopsis gene to MEDLINE abstract

In order to develop a set of document databases for our original search engine for PosMed, we developed a method for mapping between Arabidopsis genes and MEDLINE abstracts, based on an NER (Leser et al. 2005) technique that extracts named entities such as genes from a document. Since false-positive relationships can arise from a primitive NER method that simply checks for the appearance of a name in a document, we instead employ a full-text search engine for NER, with logical queries defined as a list of names or words related to a gene concatenated with logical operators such as 'AND', 'OR' and 'NOT'. More specifically, as a base query we computationally collected all the synonym names for each gene from TAIR and UniProt, connected these synonyms with the logical 'OR' operation, and added 'Arabidopsis' with the 'AND' operation. Using these base queries, we performed a full-text search against a set of documents including MEDLINE title, abstract and MeSH terms (Coletti et al. 2001). In order to reduce false-positive hits and true-negative hits, we carefully edited these queries manually through trial and error by performing a full-text search for each trial against the document set. For example, in order to detect all MEDLINE documents for the AT1G03880 (cruciferin B, CRB) gene, yet eliminating false-positive hits with the synonym 'CRB' which represents 'chloroplast RNA binding', we defined the following logical operation: ('AT1G03880' OR 'CRU2' OR 'CRB' OR 'CRUCIFERIN 2' OR 'CRUCIFERIN B') AND ('Arabidopsis') NOT ('chloroplast RNA binding'). This curation method is effective in updating with the latest publications. Once we curate a query, the query can be reused to extract gene–document relationships by performing a full-text search against those new document sets.

### PosMed-plus RANKING

In order to prioritize the positional candidate genes, PosMed-plus first calculates the statistical significance between the user's keyword and each gene. Then, a $2\times2$ contingency table $\begin{pmatrix} \text{i} & \text{ii} \\ \text{iii} & \text{iv} \end{pmatrix}$ is generated that consists of the following:

(i) the number of documents that match both the keyword and the gene
(ii) the number of documents that match the keyword but not the gene
(iii) the number of documents that match the gene but not the keyword
(iv) the number of documents that match neither the keyword nor the gene.

The *P*-value is then computed using Fisher's exact test.

For an inference search, we statistically evaluate the relevance between gene1 and gene2 using the Fisher's exact test. Thereafter, we compute the total *P*-value as $P = 1 - (1 - Ps)(1 - Pr)$, where *Ps* is the *P*-value of the first association search between the user's keyword and each gene, and *Pr* is the *P*-value of the gene–gene relationship applied in the second association search.

To treat biological data such as PPIs using this method, all biological data are described as sentences (e.g. protein A interacts with protein B) and they are stored as document sets in PosMed-plus.

**Table 3** Data description for OmicBrowse in *Arabidopsis* and rice

| Species | Genome version | Genome[a] | Transcriptome[b] | Proteome[c] | Marker | Ontology | Others[d] | Total |
|---|---|---|---|---|---|---|---|---|
| Arabidopsis | TAIR6 | 2 | 24 | 1 | 1 | 2 | 4 | 34 |
| | TAIR7 | 2 | 1 | 1 | 1 | 2 | 6 | 13 |
| | TAIR8 | 2 | 8 | | 1 | | 1 | 12 |
| | MIPSv20037 | 2 | 10 | 1 | | 2 | | 15 |
| Rice | IRGSP build3 | 1 | 1 | | | | | 2 |
| | IRGSP build4 | 1 | 1 | | | | | 2 |

[a]In addition to the various genome versions, Omic Browse contains the Entrez gene.
[b]Transcriptome data consist mainly of tiling array data and additionally full-length cDNA and expressed sequence tag data.
[c]As proteome data, we have the Genomes TO Protein Structures and Functions (GTOP) database (http://spock.genes.nig.ac.jp/~genome/gtop.html).
[d]The category 'others' contains polymorphism, T-DNA and other data. OmicBrowse is available at http://omicspace.riken.jp/gps/full.jsp?hHead=At.

## Implementation

PosMed-plus was developed as a web-oriented tool using Java and Java Servlet. Detailed information is provided in Kobayashi et al. (2008) and Yoshida et al. (2009). Users can freely access PosMed-plus with a conventional web browser, and no plug-ins need to be installed. However, for Windows we recommend the use of Microsoft Internet Explorer 7 or later, or Firefox 2 or later, and for Macintosh we recommend Safari 2 or later, or Firefox 2 or later. PosMed-plus is freely available at http://omicspace.riken.jp/PosMed-plus/.

## Funding

The Japanese Ministry of Education, Culture, Sports, Science and Technology Special Coordination Funds.

## Acknowledgements

We thank Koji Doi and Michiel J. L. de Hoon for critically reading the manuscript.

## References

Adie, E., Adams, R., Evans, K., Porteous, D. and Pickard, B. (2006) SUSPECTS: enabling fast and effective prioritization of positional candidates. *Bioinformatics* 22: 773–774.

Aerts, S., Lambrechts, D., Maity, S., Van Loo, P., Coessens, B., et al. (2006) Gene prioritization through genomic data fusion. *Nat. Biotechnol.* 24: 537–544.

*Arabidopsis* Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408: 796–815.

Ashikari, M., Sakakibara, H., Lin, S., Yamamoto, T., Takashi, T., Nishimura, A., et al. (2005) Cytokinin oxidase regulates rice grain production. *Science* 309: 741–745.

Baerenfaller, K., Grossmann, J., Grobei, M., Hull, R., Hirsch-Hoffmann, M., Yalovsky, S., et al. (2008) Genome-scale proteomics reveals *Arabidopsis thaliana* gene models and proteome dynamics. *Science* 320: 938–941.

Coletti, M. and Bleich, H. Medical subject headings used to search the biomedical literature. *J. Amer. Med. Inform. Assoc.* 8: 317–323.

Cui, J., Li, P., Li, G., Xu, F., Zhao, C., Li, Y., et al. (2008) AtPID: *Arabidopsis thaliana* protein interactome database—an integrative platform for plant systems biology. *Nucleic Acids Res* 36: D999–D1008.

Doi, K., Izawa, T., Fuse, T., Yamanouchi, U., Kubo, T., Shimatani, Z., et al. (2004) Ehd1, a B-type response regulator in rice, confers short-day promotion of flowering and controls FT-like gene expression independently of Hd1. *Genes Dev.* 18: 926–936.

Goda, H., Sasaki, E., Akiyama, K., Maruyama-Nakashita, A., Nakabayashi, K., Li, W., et al. (2008) The AtGenExpress hormone and chemical treatment data set: experimental design, data evaluation, model data analysis and data access. *Plant J.* 55: 526–542.

Hanada, K., Zou, C., Lehti-Shiu, M., Shinozaki, K. and Shiu, S. (2008) Importance of lineage-specific expansion of plant tandem duplicates in the adaptive response to environmental stimuli. *Plant Physiol.* 148: 993–1003.

Harushima, Y., Yano, M., Shomura, A., Sato, M., Shimano, T., Kuboki, Y., et al. (1998) A high-density rice genetic linkage map with 2275 markers using a single F2 population. *Genetics* 148: 479–494.

Heazlewood, J., Verboom, R., Tonti-Filippini, J., Small, I. and Millar, A. (2007) SUBA: the *Arabidopsis* subcellular database. *Nucleic Acids Res.* 35: D213–D218.

Hirochika, H., Guiderdoni, E., An, G., Hsing, Y., Eun, M., Han, C.D., et al. (2004) Rice mutant resources for gene discovery. *Plant Mol. Biol.* 54: 325–334.

Kato, N., Watanabe, Y., Ohno, Y., Inoue, T., Kanno, Y., Suzuki, H., et al. (2008) Mapping quantitative trait loci for proteinuria-induced renal collagen deposition. *Kidney Int.* 73: 1017–1023.

Kaur, J., Sebastian, J. and Siddiqi, I. (2006) The Arabidopsis-mei2-like genes play a role in meiosis and vegetative growth in Arabidopsis. *Plant Cell* 18: 545–559.

Kobayashi, N. and Toyoda, T. (2008) Statistical search on the Semantic Web. *Bioinformatics* 24: 1002–1010.

Kondou, Y., Higuchi, M., Takahashi, S., Sakurai, T., Ichikawa, T., Kuroda, H., et al. (2009) Systematic approaches to using the FOX hunting system to identify useful rice genes. *Plant J.* 57: 883–894.

Kuromori, T., Hirayama, T., Kiyosue, Y., Takabe, H., Mizukado, S., Sakurai, T., et al. (2004) A collection of 11,800 single-copy Ds transposon insertion lines in *Arabidopsis*. *Plant J.* 37: 897–905.

Kuromori, T., Wada, T., Kamiya, A., Yuguchi, M., Yokouchi, T., Imura, Y., et al. (2006) A trial of phenome analysis using 4,000 Ds-insertional mutants in gene-coding regions of *Arabidopsis*. *Plant J.* 47: 640–651.

Leser, U. and Hakenberg, J. (2005) What makes a gene name? Named entity recognition in the biomedical literature. *Brief Bioinform.* 6: 357–369.

Matsui, A., Ishida, J., Morosawa, T., Mochizuki, Y., Kaminuma, E., Endo, T.A., et al. (2008) *Arabidopsis* transcriptome analysis under drought, cold, high-salinity and ABA treatment conditions using a tiling array. *Plant Cell Physiol.* 49: 1135–1149.

Matsushima, A., Kobayashi, N., Mochizuki, Y, Ishii, M., Kawaguchi, S., Ishii, M., et al. (2009) OmicBrowse: a Flash-based high-performance graphics interface for genomic resources. *Nucleic Acids Res.* (in press)

McCouch, S., Teytelman, L., Xu, Y., Lobos, K., Clare, K., Waltam, M., et al. (2002) Development and mapping of 2240 new SSR markers for rice (*Oryza sativa* L.). *DNA Res.* 9: 199–207.

Moritani, M., Togawa, K., Yaguchi, H., Fujita, Y., Yamaguchi, Y., Inoue, H., et al. (2006) Identification of diabetes susceptibility loci in db mice by combined quantitative trait loci analysis and haplotype mapping. *Genomics* 88: 719–730.

Obayashi, T., Hayashi, S., Saeki, M., Ohta, H. and Kinoshita, K. (2009) ATTED-II provides coexpressed gene networks for *Arabidopsis*. *Nucleic Acids Res.* 37: D987–991.

Ren, Z., Gao, J., Li, L., Cai, X., Huang, W., Chao, D., et al. (2005) A rice quantitative trait locus for salt tolerance encodes a sodium transporter. *Nat. Genet.* 37: 1141–1146.

Schiøtt, M., Romanowsky, S., Baekgaard, L., Jakobsen, M., Palmgren, M. and Harper, J. (2004) A plant plasma membrane $Ca^{2+}$ pump is required for normal pollen tube growth and fertilization. *Proc. Natl Acad. Sci. USA* 101: 9502–9507.

Seelow, D., Schwarz, J. and Schuelke, M. (2008) GeneDistiller—distilling candidate genes from linkage intervals. *PLoS ONE* 3: e3874.

Seki, M., Narusaka, M., Kamiya, A., Ishida, J., Satou, M., Sakurai, T., et al. (2002) Functional annotation of a full-length *Arabidopsis* cDNA collection. *Science* 296: 141–145.

Shomura, A., Izawa, T., Ebana, K., Ebitani, T., Kanegae, H., Konishi, S., et al. (2008) Deletion in a gene associated with grain size increased yields during rice domestication. *Nat. Genet.* 40: 1023–1028.

Song, X., Huang, W., Shi, M., Zhu, M. and Lin, H. (2007) A QTL for rice grain width and weight encodes a previously unknown RING-type E3 ubiquitin ligase. *Nat. Genet.* 39: 623–630.

Sun, R., and Alexandre F. Connectionist–Symbolic Integration From Unified to Hybrid Approaches. Lawrence Erlbaum Associates Inc., London.

Swarbreck, D., Wilks, C., Lamesch, P., Berardini, T., Garcia-Hernandez, M., Foerster, H., et al. (2008) The *Arabidopsis* Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Res.* 36: D1009–D1014.

Takahashi, Y., Shomura, A., Sasaki, T. and Yano, M. (2001) Hd6, a rice quantitative trait locus involved in photoperiod sensitivity, encodes the alpha subunit of protein kinase CK2. *Proc. Natl Acad. Sci. USA* 98: 7922–7927.

Tanaka, T., Antonio, B., Kikuchi, S., Matsumoto, T., Nagamura, Y., Numa, H., et al. (2008) The Rice Annotation Project Database (RAP-DB): 2008 update. *Nucleic Acids Res.* 36: D1028–D1033.

Thornblad, T., Elliott, K., Jowett, J. and Visscher, P. (2007) Prioritization of positional candidate genes using multiple web-based software tools. *Twin Res. Hum. Genet.* 10: 861–870.

Toyoda, T., Mochizuki, Y., Player, K., Heida, N., Kobayashi, N. and Sakaki, Y. (2007) OmicBrowse: a browser of multidimensional omics annotations. *Bioinformatics* 23: 524–526.

UniProt Consortium (2009) The Universal Protein Resource (UniProt) 2009. *Nucleic Acids Res.* 37: D169–D174.

van Driel, M., Cuelenaere, K., Kemmeren, P., Leunissen, J., Brunner, H. and Vriend, G. (2005) GeneSeeker: extraction and integration of human disease-related information from web-based genetic databases. *Nucleic Acids Res.* 33: W758–W761.

Yamada, K., Lim, J., Dale, J., Chen, H., Shinn, P., Palm, C.J., et al. (2003) Empirical analysis of transcriptional activity in the *Arabidopsis* genome. *Science* 302: 842–846.

Yoshida, Y., Makita, Y., Heida, N., Asano, S., Matsushima, A., Ishii, M., et al. (2009) PosMed (Positional Medline): prioritizing genes with an artificial neural network comprising medical documents to accelerate positional cloning. *Nucleic Acids Res.* (in press)

Zhang, Y., Li, Y., Zhang, J., Shen, F., Huang, Y. and Wu, Z. (2008) Characterization and mapping of a new male sterility mutant of anther advanced dehiscence (t) in rice. *J. Genet. Genomics* 35: 177–182.

Zuo, L., Li, S., Chu, M., Wang, S., Deng, Q., Ding, L., et al. (2008) Phenotypic characterization, genetic analysis, and molecular mapping of a new mutant gene for male sterility in rice. *Genome* 51: 303–308.

(Received March 31, 2009; Accepted June 10, 2009)