Nucleic Acids Research, 2013, 1–6
doi:10.1093/nar/gkt474

PosMed: ranking genes and bioresources based on Semantic Web Association Study

Yuko Makita[1,2], Norio Kobayashi[1,2], Yuko Yoshida[1,2], Koji Doi[1,2], Yoshiki Mochizuki[1,2], Koro Nishikata[1,2], Akihiro Matsushima[1,2], Satoshi Takahashi[1,2], Manabu Ishii[1,2], Terue Takatsuki[3], Rinki Bhatia[1], Zolzaya Khadbaatar[1], Hajime Watabe[1], Hiroshi Masuya[3] and Tetsuro Toyoda[1,2,*]

[1]Bioinformatics and Systems Engineering Division (BASE), RIKEN, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan, [2]Integrated Database Unit, Advanced Center for Computing and Communication (ACCC), RIKEN, 2-1, Hirosawa, Wako, Saitama 351-0198, Japan and [3]Technology and Development Unit for Knowledge Base of Mouse Phenotype, BioResource Center (BRC), RIKEN, 3-1-1 Koyadai, Tsukuba, Ibaraki 305-0074, Japan

Received February 26, 2013; Revised May 7, 2013; Accepted May 8, 2013

ABSTRACT

Positional MEDLINE (PosMed; http://biolod.org/PosMed) is a powerful Semantic Web Association Study engine that ranks biomedical resources such as genes, metabolites, diseases and drugs, based on the statistical significance of associations between user-specified phenotypic keywords and resources connected directly or inferentially through a Semantic Web of biological databases such as MEDLINE, OMIM, pathways, co-expressions, molecular interactions and ontology terms. Since 2005, PosMed has long been used for *in silico* positional cloning studies to infer candidate disease-responsible genes existing within chromosomal intervals. PosMed is redesigned as a workbench to discover possible functional interpretations for numerous genetic variants found from exome sequencing of human disease samples. We also show that the association search engine enhances the value of mouse bioresources because most knockout mouse resources have no phenotypic annotation, but can be associated inferentially to phenotypes via genes and biomedical documents. For this purpose, we established text-mining rules to the biomedical documents by careful human curation work, and created a huge amount of correct linking between genes and documents. PosMed associates any phenotypic keyword to mouse resources with 20 public databases and four original data sets as of May 2013.

INTRODUCTION

Mouse bioresources contribute to the study of human genes and diseases (1,2). To elucidate the function of all mouse genes, the International Knockout Mouse Consortium systematically generates mutant embryonic stem cells for every protein-coding gene (3), and the International Mouse Phenotype Consortium produces knockout mice and carries out high-throughput phenotyping of each line (4). Including other mouse resources, >24 000 mouse strains are registered in the International Mouse Strain Resource (IMSR) (5). To enhance the value of bioresources, we applied our original statistical search engine called the General and Rapid Association Study Engine (GRASE) and provided this as a web-oriented service called Positional MEDLINE (PosMed) (6–9). PosMed not only allows users to retrieve mouse bioresources directly with phenotypic keywords described in bioresource annotations, but also inferentially through corresponding documents for genes, diseases, drugs, ontologies, pathways, metabolites, molecular interactions and MEDLINE abstracts. With this inferential association search function, PosMed discovers wider resources than simple keyword search and accelerates the utilization of bioresources, especially those having fewer phenotypic annotations. In particular, knockout strains are not fully used when the targeted gene has an unknown function and no observed phenotype. PosMed connects these functionally unknown genes to known genes using molecular interactions, pathway information and/or co-citations and enables the suggestion of unobserved phenotypic bioresources as a search result.

PosMed is also applicable to the functional interpretation of genetic variants detected by exome sequencing

*To whom correspondence should be addressed. Tel: +81 48 467 9267; Fax: +81 48 462 1365; Email: tetsuro.toyoda@riken.jp

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

studies. When users submit a list of genes and a phenotypic keyword, PosMed ranks the genes by statistical relevance between the keyword and each gene (6–9). These search functions are implemented using Semantic Web Association Study (SWAS) technology. The Semantic Web and linked data originally aim to provide a common framework that allows data to be shared and reused across application, enterprise and community boundaries (10,11). For most biological research purposes, however, association studies of linked data provide more analytical insights into biological systems than simple pattern-matching queries of the data (6). To take advantage of the Semantic Web of biological linked data, we propose to extend the methodology of association studies to the methodology called a 'Semantic Web Association Study (SWAS)' (Figure 1). A typical example of SWAS is the 'Genome-wide Association Study (GWAS)', which focuses only on the association between allelic variants and phenotypes in different individuals. In expanding the methodology of GWAS, SWAS explores more distant correlations among genes, functions, publications, alleles, lines, phenotypes and any subset specified by a user's keywords. Because the conventional Semantic Web Resource Description Framework (RDF) (http://www.w3.org/RDF/) and query language SPARQL (http://www.w3.org/TR/rdf-sparql-query/) do not adequately support statistical evaluation of semantic links, we developed the GRASE for the implementation of PosMed (6,8).

General usage of PosMed for bioresources

PosMed prioritizes genes, bioresources, diseases, metabolites or drugs depending on the statistical relevance between a user's keyword and biological documents. The algorithm computing *P*-value is described in our previous publications (6,7). PosMed provides paths connecting the user's keyword to the targeted resources. Figure 2 shows an example of two-step inferential or indirect search result associating a mouse resource with the keyword 'diabetes'. Although the mouse strain 'B6.129S6-Gcg<tm1Yhys>' was not directly annotated with 'diabetes', PosMed suggested it via mouse gene 'gcg, glucagon', which has thousands of documents annotated with 'diabetes'. PosMed provides up to three steps of inferential search function. For more examples such as specified genomic interval queries, please see the tutorial provided on our PosMed Web site.

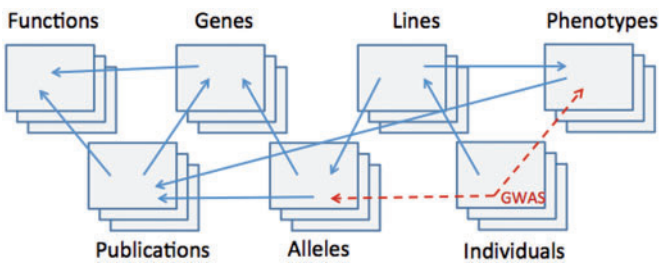

Figure 1. Concept of SWAS, which calculates the statistical significance of associations between any sets of resources connected through a web of semantic links (solid arrows), while GWAS associates only between alleles and phenotypes (dashed arrows).

Advanced search with selection of biological documents and search paths

PosMed provides several options for users to select the search paths, the documents used for the search, and the search scoring method from 'expert mode' of the advanced search setting page (Figure 3). With the expert mode menu, users can also select whether or not to use statistical significance association or Boolean association methods to associate biological items such as genes, chemicals and bioresources to a user's query directly or indirectly through user-selected search paths. The statistical significance associates biological items based on the *P*-values using Fisher's exact test of co-occurrence of the linked items in documents of OMIM, pathways, protein–protein interaction, gene ontology, phenotype ontology and other annotations, while the Boolean method associates the linked items co-occurring in the documents equally by ignoring the degree of significance (6).

PosMed assists functional interpretation after exome sequencing

Exome sequencing studies usually find several hundred to several thousand genetic variants by comparing samples and controls. To help prioritize the thousands of candidate genes for which PosMed calculates the ranking, PosMed accepts a list of gene IDs with the user's descriptions of the gene variants. The descriptions in the uploaded file are displayed together with the ranked gene, allowing users to interpret the functionality of the gene variations (Figure 4). Detailed pages for each gene assist functional interpretation by showing biological documents such as MEDLINE, gene annotations, OMIM, bioresources, pathway information, molecular interactions, ontologies and links to related databases (Table 1).

Extension of data coverage

Since previous publication, we updated 17 databases to include ~10 million biological documents (Table 1; 7,9). To enhance the inferential search function for bioresources retrieval, we newly installed the following three biological documents: mammalian phenotype ontology (MP), human disease ontology (DO) and the International Classification of Diseases (ICD-10) and updated semantic links between each biomedical document. For example, we re-annotated mouse gene to MEDLINE by defining named entity recognition (NER) rules to retrieve correct publications (6–9). For human genes, we connected to publications via mouse homologs and newly defined NER rules for 2249 human non-homolog genes against mouse. Users can download these NER rules from our Web site.

In our previous publications, we used molecular interactions and co-expression to make links from fewer annotated biological resources to well annotated resources. These relationships are important to show more candidates. On the other hand, PosMed accuracy is

Figure 2. Example inferential search result followed by direct search results for retrieving a mouse bioresource associated with the keyword 'diabetes'. PosMed shows the path connecting from a user's keyword to the resource, a resource description and linked biological documents (B). To download all candidate mouse strains, click 'check all' at the top of 'Hit resources' and download them as a text file (C).

strongly affected by low-quality data. Because Omics data are accumulated with various experimental methods, we selected high-quality data and removed low-trust data such as the classical yeast two-hybrid of protein–protein interaction (28).

For Semantic Web compliant data preparation, we used RIKENBASE or the RIKEN Scientists' Networking System (SciNetS) (29), and public data are downloadable though Biophenome Linked Open Databases (BioLOD) (http://biolod.org). At least once a month we update

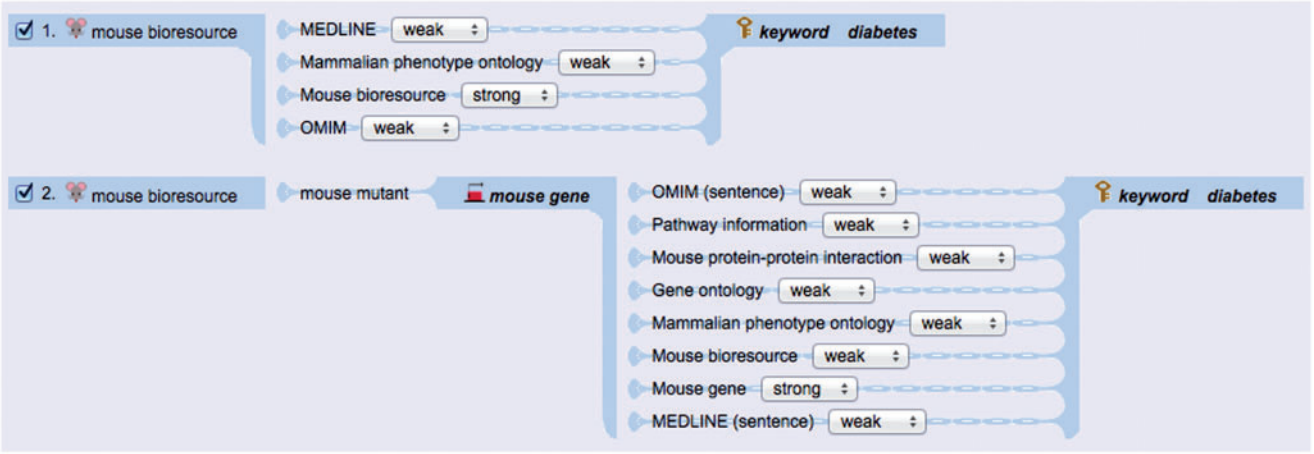

Figure 3. A partial example snapshot for 'expert mode'. The upper path (1.) shows direct search with MEDLINE, mammalian phenotype ontology, mouse bioresources and OMIM documents. The lower path (2.) shows an example inferential path via gene. Users can select the scoring method of each document from 'strong', 'weak' or 'none' in the menu. The 'strong' scoring method uses a Boolean function and the *P*-value becomes 0 when the document has at least one keyword. The 'weak' method computes *P*-value using Fisher's exact test. If a user selects 'none', the biological document is not used (6,7). In this mode users can confirm all PosMed search paths for biological documents.

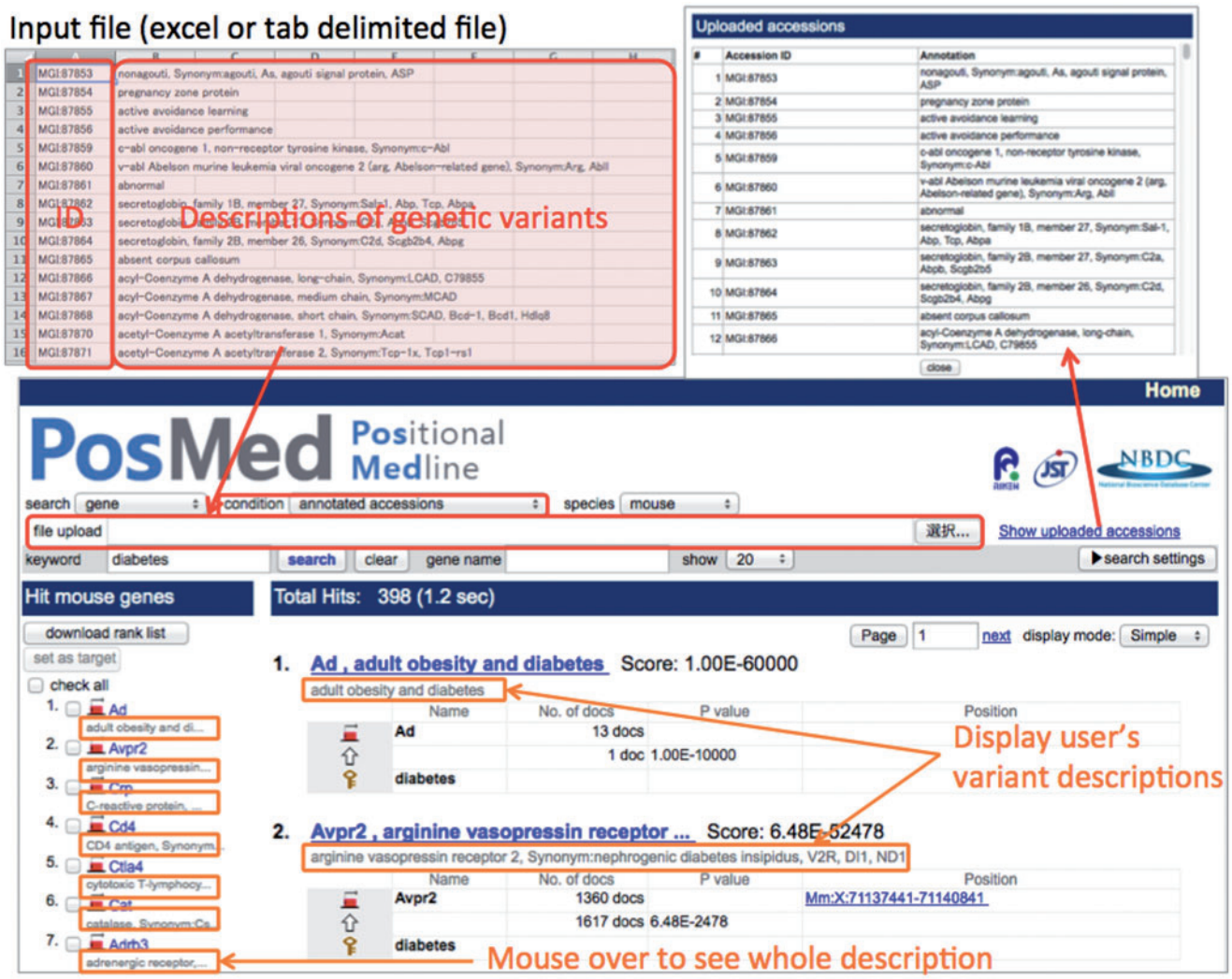

Figure 4. File upload function and display of users' descriptions. Users can upload an excel file with gene IDs and descriptions by the user. PosMed ranks the genes listed within the files by statistical relevance between the user's keyword and each gene, and displays the ranked genes together with the descriptions uploaded by the user.

Table 1. Updated biological documents for PosMed 2013

Document set	No. of documents	Data contents	References
Mouse bioresource	19 280	Mouse strain information registered at IMSR.	(5)
	5115	Mouse strain information from RIKEN BioResource center.	(12)
Human gene	37 287	Gene annotation of HGNC	(13)
Mouse gene	85 726	Gene annotation of MGI	(14)
Rat gene	36 634	Gene annotation of RGD	(15)
Arabidopsis gene	32 041	Gene annotation of TAIR	(16)
Rice gene	29 389	Gene annotation of RAP-DB	(17)
Disease	20 054	Online Mendelian Inheritance in Man	(18)
	2037	Manually collected our original data	(7)
	12 131	ICD-10, International Statistical Classification of Diseases and Related Health Problems	(19)
Metabolite	49 983	A comprehensive species-metabolite relationship database (KNApSAcK)	(20)
MEDLINE	9 378 134	MEDLINE titles, abstracts and MeSH terms	(21)
Pathway information	3809	Pathway information from REACTOME	(22)
Protein–protein interaction	73 645	Protein–Protein Interactions in Human and Mouse from rom IntAct and *Arabidopsis* from AtPID	(23,24)
Gene ontology	12 787	Gene ontology data	(25)
Human disease ontology	2282	Human disease ontology data	(26)
Mammalian phenotype ontology	7440	Mammalian phenotype ontology data	(27)

All data sources and links to the original DBs are described at http://omicspace.riken.jp/Data/.

PosMed data and its search index over the 10 million biological documents.

Implementation

PosMed is implemented as a web application that users can access freely via their web browsers without log in. Although users can use a conventional web browser and a web browser plug-in is not needed, for Windows we recommend Microsoft Internet Explorer 9 or later, Firefox 18 or later and Google Chrome 24 or later. For Macintosh we recommend Safari 5 or later and Firefox 18 or later.

The web server is developed in Java and contains 11 Linux servers, including 10 distributed servers using GRASE engines (6) that perform direct search and inferential search in parallel, and one head server performing as both the Java Servlet user interface and the coordinator that evokes parallel search requests to the distributed servers and composes their results to rank the resultant data items. This architecture realizes scalability, so the search process can still be done in a few seconds even though our data sets are extended since our previous manuscript (7,9).

Although since the first launch of the PosMed service we have often been requested by users to implement a system to support batch queries, we do not support this yet because of machine resource limitations (PosMed consumes ~1 to several seconds per query). Batch queries will be supported by securing additional machine resources in the future.

DISCUSSION

Since 2005, PosMed has been widely used to prioritize candidate genes after Quantitative Trait Locus (QTL) analysis in mice and successfully identify responsible genes (30). This time, we added a file upload function (described in Figure 4) to modify the application of QTL analysis to exome sequencing studies. For data content, we added three new databases and ontologies to expand PosMed inferential search to bioresources. These data sets allow PosMed to discover bioresources with phenotypes, while most other databases only support genetic information. We expect our work to assist active use of bioresources.

Several eminent databases have released RDF files, but not so many scientists use Semantic Web technology actively. This may be partially because bioinformaticians like to calculate the statistical significance of associations of the RDF connections rather than a simple Boolean retrieval of the connections. To solve this problem, we propose our original methodology SWAS for statistical searching of the biological Semantic Web data. PosMed executes SWAS to rank significantly enriched groups of biological resource data items that can be associated with a user-specified query through the big data of Medline, omics data sets, other semantic data and so on. Our results confirm that such enrichment analysis using our SWAS methodology is effective (31) and provides many practical usage cases of enrichment studies including biological resource ranking problems. In the near future of big-data-driven science, the SWAS methodology needs to be added to the SPARQL end point services worldwide for any user to execute enrichment study over linked open data distributed around the world.

ACKNOWLEDGEMENTS

We thank Hiromi Toyoshima for data installation to RIKEN SciNetS, Sayoko Shimoyama and Nobuhiko Tanaka for their advice on the PosMed GUI, as well as Venkata Jyothi, Ranganath Gudimella, Jeevan Kumar Suroju, Nageswara Rao Reddy Neelapu and Suresh

Avvari for their contribution to PosMed gene and MEDLINE data connection. We also thank David Gifford for critical reading of the manuscript.

FUNDING

National Bioscience Database Center (NBDC) of the Japan Science and Technology Agency (JST). Funding for open access charge: Grant from the NBDC of the JST.

Conflict of interest statement. None declared.

REFERENCES

1. Rosenthal,N. and Brown,S. (2007) The mouse ascending: perspectives for human-disease models. *Nat. Cell Biol.*, **9**, 993–999.
2. Schofield,P.N., Gkoutos,G.V., Gruenberger,M., Sundberg,J.P. and Hancock,J.M. (2010) Phenotype ontologies for mouse and man: bridging the semantic gap. *Dis. Model Mech.*, **3**, 281–289.
3. Skarnes,W.C., Rosen,B., West,A.P., Koutsourakis,M., Bushell,W., Iyer,V., Mujica,A.O., Thomas,M., Harrow,J., Cox,T. *et al.* (2011) A conditional knockout resource for the genome-wide study of mouse gene function. *Nature*, **474**, 337–342.
4. Collins,F.S., Finnell,R.H., Rossant,J. and Wurst,W. (2007) A new partner for the international knockout mouse consortium. *Cell*, **129**, 235.
5. Strivens,M. and Eppig,J.T. (2004) Visualizing the laboratory mouse: capturing phenotype information. *Genetica*, **122**, 89–97.
6. Kobayashi,N. and Toyoda,T. (2008) Statistical search on the semantic web. *Bioinformatics*, **24**, 1002–1010.
7. Yoshida,Y., Makita,Y., Heida,N., Asano,S., Matsushima,A., Ishii,M., Mochizuki,Y., Masuya,H., Wakana,S., Kobayashi,N. *et al.* (2009) PosMed (Positional Medline): prioritizing genes with an artificial neural network comprising medical documents to accelerate positional cloning. *Nucleic Acids Res.*, **37**, W147–W152.
8. Kobayashi,N. and Toyoda,T. (2011) Prioritizing genes with an artificial neural network comprising medical documents to accelerate positional cloning in biological research. Artificial Neural Networks. In: Kenji,S. (ed.) *Artificial Neural Networks - Methodological Advances and Biomedical Applications*, Chapter 9, InTech, New York, USA, pp. 173–196.
9. Makita,Y., Kobayashi,N., Mochizuki,Y., Yoshida,Y., Asano,S., Heida,N., Deshpande,M., Bhatia,R., Matsushima,A., Ishii,M. *et al.* (2009) PosMed-plus: an intelligent search engine that inferentially integrates cross-species information resources for molecular breeding of plants. *Plant Cell Physiol.*, **50**, 1249–1259.
10. Berners-Lee,T., Hendler,J. and Lassila,O. (2001) The Semantic Web. *Sci. Am. Mag.*, **284**, 34–43.
11. World Wide Web Consortium (W3C). (2011) W3CSemantic Web activity. Available at: http://www.w3.org/2001/sw/ (22 May 2013, date last accessed).
12. Yoshiki,A., Ike,F., Mekada,K., Kitaura,Y., Nakata,H., Hiraiwa,N., Mochida,K., Ijuin,M., Kadota,M., Murakami,A. *et al.* (2009) The mouse resources at the RIKEN BioResource center. *Exp. Anim.*, **58**, 85–96.
13. Gray,K.A., Daugherty,L.C., Gordon,S.M., Seal,R.L., Wright,M.W. and Bruford,E.A. (2013) Genenames.org: the HGNC resources in 2013. *Nucleic Acids Res.*, **41**, D545–D552.
14. Blake,J.A., Bult,C.J., Kadin,J.A., Richardson,J.E., Eppig,J.T. and Group,M.G.D. (2011) The Mouse Genome Database (MGD): premier model organism resource for mammalian genomics and genetics. *Nucleic Acids Res.*, **39**, D842–D848.
15. Dwinell,M.R., Worthey,E.A., Shimoyama,M., Bakir-Gungor,B., DePons,J., Laulederkind,S., Lowry,T., Nigram,R., Petri,V., Smith,J. *et al.* (2009) The Rat Genome Database 2009: variation, ontologies and pathways. *Nucleic Acids Res*, **37**, D744–749.
16. Lamesch,P., Berardini,T.Z., Li,D., Swarbreck,D., Wilks,C., Sasidharan,R., Muller,R., Dreher,K., Alexander,D.L., Garcia-Hernandez,M. *et al.* (2012) The *Arabidopsis* Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.*, **40**, D1202–D1210.
17. Sakai,H., Lee,S.S., Tanaka,T., Numa,H., Kim,J., Kawahara,Y., Wakimoto,H., Yang,C.C., Iwamoto,M., Abe,T. *et al.* (2013) Rice Annotation Project Database (RAP-DB): an integrative and interactive database for rice genomics. *Plant Cell Physiol.*, **54**, e6.
18. Amberger,J., Bocchini,C.A., Scott,A.F. and Hamosh,A. (2009) McKusick's Online Mendelian Inheritance in Man (OMIM). *Nucleic Acids Res.*, **37**, D793–D796.
19. World Health Organization. (2004) *International Statistical Classification of Diseases and Health Related Problems (The) ICD-10*, Vol. 1. World Health Organization, Geneva, Switzerland.
20. Afendi,F.M., Okada,T., Yamazaki,M., Hirai-Morita,A., Nakamura,Y., Nakamura,K., Ikeda,S., Takahashi,H., Altaf-Ul-Amin,M., Darusman,L.K. *et al.* (2012) KNApSAcK family databases: integrated metabolite-plant species databases for multifaceted plant research. *Plant Cell Physiol.*, **53**, e1.
21. Coletti,M. and Bleich,H. (2001) Medical subject headings used to search the biomedical literature. *J. Am. Med. Inform. Assoc.*, **8**, 317–323.
22. Croft,D., O'Kelly,G., Wu,G., Haw,R., Gillespie,M., Matthews,L., Caudy,M., Garapati,P., Gopinath,G., Jassal,B. *et al.* (2011) Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.*, **39**, D691–D697.
23. Kerrien,S., Aranda,B., Breuza,L., Bridge,A., Broackes-Carter,F., Chen,C., Duesbury,M., Dumousseau,M., Feuermann,M., Hinz,U. *et al.* (2012) The IntAct molecular interaction database in 2012. *Nucleic Acids Res.*, **40**, D841–D846.
24. Li,P., Zang,W., Li,Y., Xu,F., Wang,J. and Shi,T. (2011) AtPID: the overall hierarchical functional protein interaction network interface and analytic platform for Arabidopsis. *Nucleic Acids Res.*, **39**, D1130–D1133.
25. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
26. Osborne,J.D., Flatow,J., Holko,M., Lin,S.M., Kibbe,W.A., Zhu,L.J., Danila,M.I., Feng,G. and Chisholm,R.L. (2009) Annotating the human genome with Disease Ontology. *BMC Genomics*, **10(Suppl. 1)**, S6.
27. Smith,C.L. and Eppig,J.T. (2012) The Mammalian Phenotype Ontology as a unifying standard for experimental and high-throughput phenotyping data. *Mamm. Genome*, **23**, 653–668.
28. Caufield,J.H., Sakhawalkar,N. and Uetz,P. (2012) A comparison and optimization of yeast two-hybrid systems. *Methods*, **58**, 317–324.
29. Masuya,H., Makita,Y., Kobayashi,N., Nishikata,K., Yoshida,Y., Mochizuki,Y., Doi,K., Takatsuki,T., Waki,K., Tanaka,N. *et al.* (2011) The RIKEN integrated database of mammals. *Nucleic Acids Res.*, **39**, D861–D870.
30. Masuya,H., Yoshikawa,S., Heida,N., Toyoda,T., Wakana,S. and Shiroishi,T. (2007) Phenosite: a web database integrating the mouse phenotyping platform and the experimental procedures in mice. *J. Bioinformatics Comput. Biol.*, **5**, 1173–1191.
31. 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.