PBmice: an integrated database system of piggyBac (PB) insertional mutations and their characterizations in mice

Ling V. Sun¹,²,*, Ke Jin², Yiming Liu³, Wenwei Yang¹,², Xing Xie³, Lin Ye², Li Wang²,⁴, Lin Zhu³, Sheng Ding¹,², Yi Su³, Jie Zhou², Min Han¹,²,⁵, Yuan Zhuang¹,²,⁶, Tian Xu¹,²,⁷, Xiaohui Wu¹,², Ning Gu³,⁶ and Yang Zhong²,⁴,*

¹Institute of Developmental Biology and Molecular Medicine, ²School of Life Sciences, ³School of Information Science and Engineering, Fudan University, Shanghai 200433, ⁴Shanghai Center for Bioinformation Technology, Shanghai 200235, China, ⁵Howard Hughes Medical Institute, Department of MCDB, University of Colorado at Boulder, Boulder, CO 80309, ⁶Department of Immunology, Duke University Medical Center, Durham, NC 27710 and ⁷Howard Hughes Medical Institute, Department of Genetics, Yale University School of Medicine, Boyer Center for Molecular Medicine, New Haven, CT 06536, USA

Received August 13, 2007; Revised September 13, 2007; Accepted September 17, 2007

ABSTRACT

DNA transposon piggyBac (PB) is a newly established mutagen for large-scale mutagenesis in mice. We have designed and implemented an integrated database system called PBmice (PB Mutagenesis Information CEnter) for storing, retrieving and displaying the information derived from PB insertions (INSERTs) in the mouse genome. This system is centered on INSERTs with information including their genomic locations and flanking genomic sequences, the expression levels of the hit genes, and the expression patterns of the trapped genes if a trapping vector was used. It also archives mouse phenotyping data linked to INSERTs, and allows users to conduct quick and advanced searches for genotypic and phenotypic information relevant to a particular or a set of INSERT(s). Sequence-based information can be cross-referenced with other genomic databases such as Ensembl, BLAST and GBrowse tools used in PBmice offer enhanced search and display for additional information relevant to INSERTs. The total number and genomic distribution of PB INSERTs, as well as the availability of each PB insertional LINE can also be viewed with user-friendly interfaces. PBmice is freely available at http://www.idmshanghai.cn/PBmice or http://www.scbit.org/PBmice/.

INTRODUCTION

Large-scale mutagenesis is critical to functional characterization of the mammalian genomes. In mice, gene targeting, chemical mutagenesis and insertional mutagenesis are available to achieve this goal. The International Knockout Mouse Consortium (IKMC) has recently been formed to systematically produce null alleles for individual genes in mice (1–3). This approach focuses on annotated protein-coding regions but provides limited insight into the rest of the genome. In contrast, ENU (N-ethyl-N-nitrosoureia) based chemical mutagenesis has the capacity to introduce mutations throughout the entire genome (4–11). Several large-scale ENU mutagenesis studies have been conducted in the recent years and produced a large collection of mutant mice with interesting phenotypes. However, the majority of these mutant strains remain to be mapped to the single gene level. Insertional mutagenesis has been shown to be one of the most efficient means to generate a large number of identifiable mutants throughout the entire genome of several model organisms (12–17). With an appropriate insertion vector, insertional mutagenesis can produce a large number of mutations at low cost with a high speed. Each insertion site is tagged with the vector sequence and thus can be easily mapped.

Retroviruses have been used in mouse genetics for more than two decades (18,19). They led to discovery of a number of proto-oncogene and tumor suppressor loci (19). However, successful application of retroviruses in
large-scale insertional mutagenesis remains limited to gene trapping in ES cells (20,21). In recent years, the progress of introducing transposon tools into mammalian genetic studies has offered new opportunities for mutagenesis in mice (22–30). Several transposons, such as Sleeping Beauty (SB), Minos, Tol2, Mos1 and piggyBac (PB), were shown to be active in mammalian cells and/or in mice. Among them, the PB transposon appears to be a promising option. Appreciation of PB as an important genetic manipulation tool in various organisms including mice has been growing rapidly in recent years, as accumulating evidence has indicated that PB elements are capable of efficient transposition in animals from diverse genera of insects, flatworms, mice, as well as mammalian cell lines (28,29,31–38). PB elements, originally found in the cabbage looper moth Trichoplusia ni, are DNA transposons carrying a unique functional transposase (PBase) of 594 amino acids (39–41). It has been reported to be more active than Tol2, Mos1 and SB in mammalian cells (30), and has been shown to be effective mutagen for insertional mutagenesis in mice.

Currently, a genome-wide PB mutagenesis project is underway in the Institute of Developmental Biology and Molecular Medicine (IDM) at Fudan University in Shanghai, China. To collect and disseminate the large amount of genetic and phenotypic information from this project, we have designed and implemented an integrated database system called PBmice (PB Mutagenesis Information CEnter), which provides a user interface to query and display data obtained from PB insertions (INSERTs) in the mouse genome and their characterizations.

SYSTEM DESIGN AND IMPLEMENTATION

Data sources
Experimental data in PBmice are derived from the on-going project of PB insertional mutagenesis in mice at IDM. Each PB transposition event produces a PB insertion, which is termed as an INSERT in PBmice and carried by a mouse line (LINE). The flanking genomic sequence of an INSERT is characterized by inverse PCR and determined for its location in the mouse genome by WU-BLAST 2.0 search against the Ensembl mouse genome sequences (Ensembl release 45, based on GenBank m36 mouse assembly, http://www.ensembl.org/Mus_musculus/index.html) (42). Expression level of the gene hit by an INSERT is determined by real-time RT-PCR. Other information about the INSERT, such as the mouse strain and LINE carrying the INSERT and the PB construct used to generate the INSERT, are also collected in the database. When a trapping vector is used to generate the INSERT, the expression pattern of the reporter gene of the vector, an indication of the expression pattern of the endogenous gene affected in the INSERT, is also examined and collected in the database. Associated information of a LINE carrying one or more INSERTs includes the existence of live animals and storage status, as well as its phenotyping data. When an INSERT hits a functional genomic unit, such as an annotated protein-coding gene, a gene for a miRNA, or a regulatory element, this interruption of genomic function may be shown as phenotypes in the hosting LINE. Such characterizations are also recorded and documented in PBmice.

Data integration
A Java tool has been developed for PBmice to integrate gene data from Ensembl and GenBank are incorporated together. MGI-ID of a gene is used to integrate the information of the same gene downloaded from Ensembl and GenBank in PBmice. Gene information contains gene names with synonyms, descriptions, genomic locations, transcript information and IDs linked to existing databases, such as Ensembl, GenBank and Mouse Genome Informatics (MGI) (http://www.informatics.jax.org/), for links to its detailed information in those databases. Additionally, users can obtain protein information from ExPASy (http://us.expasy.org/) and gene function information from the Gene Ontology Database (http://www.geneontology.org/). The integrated items are converted into data files in GFF format (http://www.sanger.ac.uk/Software/formats/GFF/) for retrieving and displaying with GBrowse (http://www.gmod.org/wiki/index.php/Gbrowse), allowing for a more straightforward and user-friendly view.

User interface
INSERTs and INSERT-related data form the central core of PBmice. This system provides detailed information about an INSERT and associated information of the LINE carrying the INSERT and the gene hit by the INSERT.

To search for an INSERT of interest, either the quick search or the advanced search method can be used (Figure 1). The quick search allows a user to conduct a simple search just with the name of an INSERT, a LINE, a strain or a phenotype. A default entry results in a complete list of all INSERTs. The partial-word-match method is embedded in the query engine as well. Advanced search is provided for users with more defined interests to narrow down the search area, such as (i) to search for the existence of INSERTs in terms of their relative positions to a known INSERT or Gene; (ii) to search for INSERTs in a certain genomic area on a chromosome of interest; (iii) to search for INSERTs associated with the expression pattern characterized for a certain organ in an animal of certain developmental stage or (iv) to search for INSERTs associated with a certain phenotype. The first and second search criteria can be independently combined with the rest of search criteria but not with each other during an advanced search.

The INSERT Detail interface presents the detailed information of an INSERT after a successful search, including chromosomal location and flanking genomic sequence of the INSERT, the expression level of the gene that is hit by the INSERT, and the PB construct used to generate the INSERT (Figure 2). Of all the information displayed in the interface, the Chromosome Location Prox item refers to the genomic position immediately preceding
the INSERT, while the **Chromosome Location Dist** item refers to the position immediately following the INSERT. The **Gene hit** item points to the information of the gene with its Ensembl ID, which is clickable for users to obtain detailed information of this gene. The **Orientation** item of this INSERT is defined as following: the orientation of PB is defined as from PBL to PBR (29), while the chromosome orientation is defined by Ensembl. If the orientation of PB is the same as the orientation of the inserted chromosome, the orientation of the INSERT is denoted as ‘+’, otherwise, the orientation of this INSERT will be ‘−’.

The **DNA Sample** item shows the name of the DNA sample used for mapping the INSERT. The **Expression Level** item indicates the expression level of the gene hit by the INSERT in the mutant. The **GBrowse** item offers an annotated view of a 40 kb genomic fragment with the position of the INSERT in the middle and is linked to the GBrowse tool embedded in PBmice for a user-adjustable view of the annotation of the genomic fragment around the INSERT. The **LINE** item contains the name of the mouse LINE that harbors this INSERT. The rest of the information categories titles are self-explanatory.

Figure 1. Search interfaces. (a) The *Quick Search* interface. This is also the homepage of PBmice. (b) The *Advanced Search* interface.
In the PBmice database, the information of a LINE is associated with the INSERT(s) it carries. The LINE Detail interface reveals the properties of a LINE carrying one or more INSERTs, including the status, expression pattern and phenotyping data of the LINE (Figure 3). In this interface, the Status item indicates the storage status of the LINE (live animal, frozen sperm and/or frozen embryo) and related information for strain distribution. When a LINE carries trapping vector-related INSERT(s), the Expression pattern item will indicate any information that was characterized to some extent for the expression of the reporter protein. The Phenotype item provides clickable categories of phenotypes characterized for this LINE. The detailed phenotypic data contains not only the description of the phenotype and assays used to depict the phenotype, but also the sample information, such as the heterozygosity or homozygosity, the developmental stage, the sex of the animals or the organ examined. Users can also retrieve related INSERTs/LINEs based on similarity of phenotypes through advanced search (Figure 1b).

PBmice also provides information of a gene when hit by one or more INSERT(s). Both the Gene Information interface and the Transcript interface exhibit the identity and associated information of a gene (Figure 4a and b). The Gene Information interface presents the name and synonyms, the description and the genomic location of a gene and the links to outside databases, such as Ensembl, MGI and GenBank (Figure 4a). The Transcript interface displays links to existing databases for users to retrieve detailed information about each related transcript, protein and gene function information of the gene as shown in Figure 4a (Figure 4b). UniProt-ID(s), EMBL-ID(s) and GenBank-Protein-ID(s) are linked to information about the protein expressed by this gene stored in ExPASy containing UniProt Knowledge base (Swiss-Prot and TrEMBL) through the Multi-Protein Survey System (MPSS) developed by Bioinformation Center, Shanghai Institutes for Biological Sciences (http://www.biosino.org/MPSS/index.jsp). GO-ID(s) are linked to the associated gene function information from the Gene Ontology database. UNIGENE-ID(s) are linked to the associated transcripts information from GenBank.

**Example query**

Taking the INSERT AF0-47T6 that hits into the gene Pkd2 as an example to navigate PBmice. Pkd2 was cloned in 1996 and proved to contribute to autosomal dominant polycystic kidney disease (ADPKD) when mutated (43). Patients suffering from ADPKD grow cysts in their...
locations of all INSERTs into introns, exons, intergenic
regions or ND (undetermined). Introduction and step-by-
step search examples are provided in the Help interface of PBmice.

AVAILABILITY AND FUTURE DIRECTIONS
PBmice is freely available at http://www.idmshanghai.cn/
PBMice, or http://www.scbit.org/PBmice/. It is maintained
by the IDM at Fudan University. All data sources used by
PBmice are reviewed and updated quarterly. During the
next 3 years, development of this system will include:
(i) new INSERTs will be released from the on-going
genome-wide PB mutagenesis project that aims to create
10,000 insertions in mice; (ii) corresponding changes will
be made following any substantial change in the mouse
genome sequence information from new releases of linked
databases such as Ensembl. In particular, the position of
the existing INSERTs will be remapped with a new round
of Blast search to provide updated information for users;
(iii) a dynamic data management system will be incorpo-
rated into PBmice so that collaborators have the
opportunities to access to primary data prior to its release
and (iv) a policy and practical mechanism will be
developed to allow dissemination of mutant strains and
relevant reagents for non-profit academic research.

ACKNOWLEDGEMENTS
This research was supported by grants from Chinese
Key Projects for Basic Research (973) (Grant No. 2006CB806700, 2005CB321905 and 2002CB512801),
Hi-tech Research and Development Project (863) (Grant No.
2007AA022012), National Natural Science
Foundation of China (Grant No. 60473124), Shanghai
Pujiang Program (Grant No. 05PJ14024) and Program for
New Century Excellent Talents in University. We thank
Rick Yan from Plano, Texas (TASM) for reading the
draft version of the manuscript. Funding to pay the Open
Access publication charges for this article was provided by
Hi-tech Research and Development Project (863) (Grant No.
2007AA022012).

Conflict of Interest Statement. None declared.

REFERENCES
1. Austin, C.P., Battey, J.F., Bradley, A., Bucan, M., Capecci, M.,
Collins, F.S., Dove, W.F., Duyk, G., Dymecki, S. et al. (2004) The
knockout mouse project. Nat. Genet., 36, 921–924.
2. Auwerx, J., Avner, P., Baldock, R., Ballabio, A., Balling, R.,
Barbacid, M., Berns, A., Bradley, A., Brown, S. et al. (2004) The
European dimension for the mouse genome mutagenesis program.
Nat. Genet., 36, 925–927.
3. Collins, F.S., Rossant, J. and Wurst, W. (2007) A mouse for all
reasons. Cell, 128, 9–13.
4. Clark, A.T., Goldowitz, D., Takahashi, J.S., Vitaterna, M.H.,
Siepak, S.M., Peters, J.L., Frankel, W.N., Carlson, G.A., Rossant, J.
et al. (2004) Implementing large-scale ENU mutagenesis screens in
North America. Genetica, 122, 51–64.
5. Fahren, A.M., Bazan, J.F., Papanastasio, P., Nelms, K.A. and
Goodnow, C.C. (2001) A genomic view of immunology. Nature, 409,
836–838.
6. Hube de Angelis, M.H., Flaswinkel, H., Fuchs, H., Rathkolb, B.,
Soewarto, D., Marschall, S., Heffner, S., Pargent, W., Wenshe, K.
et al. (2000) Genome-wide, large-scale production of mutant mice by ENU mutagenesis. Nat. Genet., 25, 444–447.

7. Nolan,P.M., Peters,J., Strivens,M., Rogers,D., Hagan,J., Spurr,N., Gray,I.C., Vizor,L., Brooker,D. et al. (2000) A systematic, genome-wide, phenotype-driven mutagenesis programme for gene function studies in the mouse. Nat. Genet., 25, 440–443.

8. Nolan,P.M., Peters,J., Vizor,L., Strivens,M., Washbourne,R., Hough,T., Wells,C., Glenister,P., Thornton,C. et al. (2000) Implementation of a large-scale ENU mutagenesis program: towards increasing the mouse mutant resource. Mamm. Genome, 11, 500–506.

9. Pargenti,W., Heffner,S., Schable,K.F., Soewarto,D., Fuchs,H. and Hrabe de Angelis,M. (2000) MouseNet database: digital management of a large-scale mutagenesis project. Mamm. Genome, 11, 590–593.

10. Rathkolb,B., Decker,T., Fuchs,E., Soewarto,D., Fella,C., Heffner,S., Pargenti,W., Van Sloun,P., Hollatz,M., Floss,T., Stanford,W.L., Cohn,J.B. and Cordes,S.P. (2001) Gene-trap mutagenesis: past, present and future. Nat. Rev. Genet., 25, 1121–1128.

11. Soewarto,D., Fella,C., Teubner,A., Rathkolb,B., Pargenti,W., Hees,C., Hehl,C., Wolf,E., Balling,R. et al. (2000) The large-scale Munich ENU-mouse-mutagenesis screen. Mamm. Genome, 11, 507–510.

12. Cooley,L., Kelley,R. and Spradling,A. (1988) Insertional mutagenesis of the Drosophila genome with single P elements. Science, 239, 1233–1274.

13. Zhang,P. and Spradling,A.C. (1994) Insertional mutagenesis of Drosophila heterochromatin with single P elements. Proc. Natl Acad. Sci. USA, 91, 3539–3543.

14. Amsterdam,A., Burgess,S., Golling,G., Chen,W., Sun,Z., Townsend,K., Farrington,S., Haldimann,S. and Hopkins,N. (1999) A large-scale insertional mutagenesis screen in zebrafish. Genes Dev., 13, 2713–2724.

15. Amsterdam,A., Yoon,C., Allende,M., Becker,T., Kawakami,K., Burgess,S., Gaiano,N. and Hopkins,N. (1997) Retrovirus-mediated insertional mutagenesis in zebrafish and identification of a molecular marker for embryonic germ cells. Cold Spring Harb. Symp. Quant. Biol., 62, 437–450.

16. Koowa,H., Breuss,R.A. and Hasegawa,P.M. (2006) Identification of plant stress-responsive determinants in Arabidopsis by large-scale forward genetic screens. J. Exp. Bot., 57, 1119–1128.

17. Kuromori,T., Wada,T., Kamiya,A., Yaguchi,M., Yokouchi,T., Imura,Y., Takabe,H., Sakurai,T., Akiyama,K. et al. (2006) A trial of phenome analysis using 4000 Dos-insertional mutants in gene-encoding regions of Arabidopsis. Plant J., 47, 640–651.

18. Varma,H.E. (1983) Using retroviruses as insertional mutagens to identify cellular oncogenes. Prog. Clin. Biol. Res., 119, 23–35.

19. Uren,A.G., Kool,J., Berns,A. and van Lohuizen,M. (2005) Retroviral insertional mutagenesis: past, present and future. Oncogene, 24, 7650–7672.

20. Stanford,W.L., Cohn,J.B. and Cordes,S.P. (2001) Gene-trap mutagenesis: past, present and beyond. Nat. Rev. Genet., 2, 756–768.

21. Schnungen,F., De-Zolt,S., Van Sloun,P., Hofflitzke,F., Wolf,E. and Balling,R. et al. (2000) Implementation of a large-scale ENU mutagenesis program: towards increasing the mouse mutant resource. Mamm. Genome, 11, 500–506.

22. Araki,Y., Saito,A., Nakamura,Y., Nakamura,Y. and Araki,H. (2006) A trial of phenome analysis using 4000 Dos-insertional mutants in gene-encoding regions of Arabidopsis. Plant J., 47, 640–651.

23. Varma,H.E. (1983) Using retroviruses as insertional mutagens to identify cellular oncogenes. Prog. Clin. Biol. Res., 119, 23–35.

24. Uren,A.G., Kool,J., Berns,A. and van Lohuizen,M. (2005) Retroviral insertional mutagenesis: past, present and future. Oncogene, 24, 7650–7672.

25. Stanford,W.L., Cohn,J.B. and Cordes,S.P. (2001) Gene-trap mutagenesis: past, present and beyond. Nat. Rev. Genet., 2, 756–768.

26. Schnungen,F., De-Zolt,S., Van Sloun,P., Hofflitzke,F., Wolf,E. and Balling,R. et al. (2000) Implementation of a large-scale ENU mutagenesis program: towards increasing the mouse mutant resource. Mamm. Genome, 11, 500–506.

27. Araki,Y., Saito,A., Nakamura,Y., Nakamura,Y. and Araki,H. (2006) A trial of phenome analysis using 4000 Dos-insertional mutants in gene-encoding regions of Arabidopsis. Plant J., 47, 640–651.

28. Varma,H.E. (1983) Using retroviruses as insertional mutagens to identify cellular oncogenes. Prog. Clin. Biol. Res., 119, 23–35.

29. Uren,A.G., Kool,J., Berns,A. and van Lohuizen,M. (2005) Retroviral insertional mutagenesis: past, present and future. Oncogene, 24, 7650–7672.