D642–D648 Nucleic Acids Research, 2006, Vol. 34, Database issue
doi:10.1093/nar/gkj097

The International Gene Trap Consortium Website: a portal to all publicly available gene trap cell lines in mouse

Alex S. Nord[1,3], Patricia J. Chang[1], Bruce R. Conklin[2], Antony V. Cox[3], Courtney A. Harper[1], Geoffrey G. Hicks[4], Conrad C. Huang[1], Susan J. Johns[1], Michiko Kawamoto[1], Songyan Liu[4], Elaine C. Meng[1], John H. Morris[1], Janet Rossant[5], Patricia Ruiz[6], William C. Skarnes[3], Philippe Soriano[7], William L. Stanford[8], Doug Stryke[1], Harald von Melchner[9], Wolfgang Wurst[10], Ken-ichi Yamamura[11], Stephen G. Young[12], Patricia C. Babbitt[1] and Thomas E. Ferrin[1,*]

[1]University of California, San Francisco, 600 16th Street, San Francisco, CA 94143-2240, USA, [2]Gladstone Institute of Cardiovascular Disease, University of California San Francisco Department of Medicine and Pharmacology, 1650 Owens Street, San Francisco, CA 94158, USA, [3]Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, UK, [4]Manitoba Institute of Cell Biology, University of Manitoba, 675 McDermot Avenue, Winnipeg, Manitoba, Canada R3E 0V9, [5]The Hospital for Sick Children, Toronto, Ontario, Canada M5G 1X8, [6]Center for Cardiovascular Research, Charité Universitätsmedizin and Department of Vertebrate Genomics, Max Planck Institute for Molecular Genetics, 14195 Berlin, Germany, [7]Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, Seattle, WA 98109-1024, USA, [8]University of Toronto, 4 Taddle Creek Road, Toronto, Ontario, Canada M5S 3G9, [9]Department of Molecular Hematology, University of Frankfurt Medical School, 60590 Frankfurt am Main, Germany, [10]GSF Research Center for Environment and Health, Institute for Developmental Genetics, Ingolstaedter Landstrasse 1, D-85764 Neuherberg, Germany, [11]Institute of Molecular Embryology and Genetics, Kumamoto University, 2-2-1 Honjo, Kumamoto 860-0811, Japan and [12]University of California, Los Angeles, 650 Charles E. Young Dr So., Los Angeles, CA 90095, USA

Received August 16, 2005; Revised October 7, 2005; Accepted October 16, 2005

ABSTRACT

Gene trapping is a method of generating murine embryonic stem (ES) cell lines containing insertional mutations in known and novel genes. A number of international groups have used this approach to create sizeable public cell line repositories available to the scientific community for the generation of mutant mouse strains. The major gene trapping groups worldwide have recently joined together to centralize access to all publicly available gene trap lines by developing a user-oriented Website for the International Gene Trap Consortium (IGTC). This collaboration provides an impressive public informatics resource comprising ~45 000 well-characterized ES cell lines which currently represent ~40% of known mouse genes, all freely available for the creation of knockout mice on a non-collaborative basis. To standardize annotation and provide high confidence data for gene trap lines, a rigorous identification and annotation pipeline has been developed combining genomic localization and transcript alignment of gene trap sequence tags to identify trapped loci. This information is stored in a new bioinformatics database accessible through the IGTC Website interface. The IGTC Website (www.genetrap.org) allows users to browse and search the database for trapped genes, BLAST sequences against gene trap sequence tags, and view trapped genes within biological pathways. In addition, IGTC data have been integrated into major genome browsers and bioinformatics sites to provide users with outside portals for viewing this data. The development of the IGTC Website marks a major advance by providing the research community with the data and tools necessary to effectively use

*To whom correspondence should be addressed. Email: tef@cgl.ucsf.edu

public gene trap resources for the large-scale characterization of mammalian gene function.

INTRODUCTION

The large and continually growing number of genome sequencing projects provides an opportunity to greatly advance our understanding of genetics and disease. One of the keys to realizing this goal is the development of genomic resources to elucidate functional characteristics of these genes, especially in mammalian genomes. The mouse is an especially useful mammalian model system, providing an excellent subject for studies of gene function because of its short generational span, ease of handling, and the close structural and functional similarity of its genome to that of humans (1,2). Furthermore, using this organism, scientists have access to a wide range of procedures for genetic manipulation, including the use of embryonic stem (ES) cells to create mice with defined single-gene mutations using gene targeting and gene trapping techniques (3).

Gene trapping is a high-throughput method of creating mutagenized ES cells for use in generating knockout and other mutant mouse strains for research in functional genomics (4). Second generation gene trap vectors have recently enhanced the value of the method by offering the potential for creating conditional and other desired alleles using site-specific recombination (5–7). Major scientific initiatives are currently underway in North America and Europe to knock out every mouse gene in ES cells in order to characterize gene function and provide insight into systems associated with human disease (8,9).

A number of gene trap projects have already made notable progress toward this goal by generating resources of gene trap mouse ES cell lines harboring well-characterized insertional mutations (6,7,10–14), although until now the individual gene trap projects have been isolated, providing only details about their own cell lines. The International Gene Trap Consortium (IGTC) is a collaboration representing the major public gene trap resources worldwide, whose mission is to offer the scientific community access to all publicly available gene trap cell lines on a non-collaborative basis for nominal handling fees (15). The centralization of gene trap resources provides many advantages to the research community, allowing more effective utilization of the experimental opportunities offered by gene trap cell lines through standardized protocols for the identification and annotation of sequences from trapped loci, and the increased availability of experimental protocols. As reported here, the release of the IGTC Website (www.genetrap.org) marks a major advance, generating a standardized informatics pipeline and providing in one place both easy access to all publicly available gene trap cell lines and sophisticated tools for analysis of resource data. Gene trap centers currently involved in this effort are listed in Table 1.

IGTC: RESOURCE OVERVIEW

The IGTC Website centralizes access to all publicly available gene trap cell line data for the first time. This repository was created to address the needs of the international gene trap community by providing researchers with the data and informatics tools necessary to find gene trap cell lines with mutations in genes and loci of interest. IGTC member projects produce gene trap cell lines and directly submit gene trap sequence tags to the Genome Survey Sequences Database (dbGSS) division of GenBank at NCBI (16). Data from all publicly available cell lines are downloaded from dbGSS and subjected to the IGTC identification and annotation pipeline, which then automatically populates the MySQL database used to generate the annotation information presented on the Website. The IGTC Website has been designed to provide easy user access to the extensive array of assembled gene trap informatics data. Interface options include homology searches using BLAST, search and browse capabilities, and viewing trapped genes within biological pathways. The project also includes the integration of gene trap data at major genome browsers and other informatics data sites in order to offer a variety of outside portals to this data. Cell line requests from the IGTC site are forwarded to the originating gene trap resource, where the cell line is removed from cryogenic storage and sent to the user for experimental analysis. The IGTC site also provides useful documentation, on-line tutorials and scientific overviews on gene trapping and the use of gene trap cell lines.

GENE TRAP IDENTIFICATION AND ANNOTATION PIPELINE

Gene trap mutations are characterized through a process of sequencing, identification and annotation. This process involves obtaining cDNA or genomic sequence upstream or downstream of the insertion site and identifying and annotating the locus at which the insertion occurs. A full identification and annotation protocol has been developed for the IGTC that integrates genomic and transcript-based identification approaches and adds information from other major informatics

Table 1. IGTC members

IGTC members	Cell lines	Website
Baygenomics (USA)	9848	www.baygenomics.ucsf.edu/
Centre for Modelling Human Disease (Toronto, Canada)	4137	www.cmhd.ca/genetrap/
Embryonic Stem Cell Database (University of Manitoba, Canada)	8559	www.EScells.ca/
Exchangeable Gene Trap Clones (Kumamoto University, Japan)	49	egtc.jp/show/index
German Gene Trap Consortium (Germany)	13031	www.genetrap.de/
Sanger Institute Gene Trap Resource (Cambridge, UK)	7354	www.sanger.ac.uk/PostGenomics/genetrap/
Soriano Lab Gene Trap Database (FHCRC, Seattle, USA)	1627	www.fhcrc.org/science/labs/soriano/trap.html
Telethon Institute of Genetics and Medicine–TIGEM (Naples, Italy)	1435	core.tigem.it/genetrap/public/
TOTAL	44605	

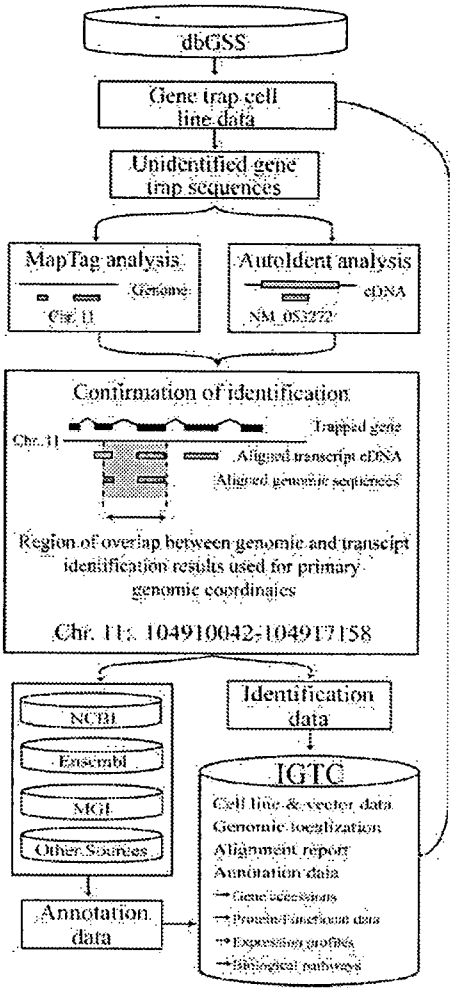


Figure 1. IGTC identification and annotation pipeline. IGTC members submit gene trap cell line data to dbGSS. The first step of the IGTC pipeline is the download of all publicly available gene trap cell line data from dbGSS. Gene trap sequences are then processed through the dual identification protocol based on genomic localization and transcript alignment using MapTag and AutoIdent, respectively. Returned homologous genomic regions and transcripts are aligned to genomic sequence to generate confirmed overlapping sequence regions, which are used as the primary identification data. Confirmed genomic coordinates are queried against major informatics databases to obtain annotation data for the genomic locus identified, and the returned data are entered into the IGTC database.

resources, such as genome browsers and specialized informatics sites to annotate the identified loci (Figure 1). Use of both genomic and transcript-based annotation methods were incorporated to increase confidence in identification and localization on the genome of trapped loci, providing a significant improvement over identification strategies previously in use. This pipeline was designed to be both robust and flexible, allowing the incorporation of multiple methods of identification and mapping and the future integration of new genomic data resources that may be developed.

Gene trap cell line sequences are submitted by each individual gene trap resource to dbGSS along with relevant ancillary information about the cell line, including its original source and the vector used in the trapping experiment. Within dbGSS, IGTC sequences are grouped using LinkOut (17) (http://www.dlib.org/dlib/march02/03inbrief.html KWAN), which links gene trap dbGSS entries to the IGTC Website. Requiring gene trap cell lines to be entered into dbGSS ensures that all IGTC lines are in the public domain, and that all data are consistent, transparent and freely accessible. Researchers can access additional information about methodology and protocols from the original trapping experiment via links to the individual gene trap project sites (Table 1).

The two complementary programs, MapTag (15) and AutoIdent (12), are used to locate the trapped gene on the mouse genome by analyzing the similarity between the unidentified gene trap sequence and genomic and transcript sequences, respectively. The minimum region of genomic overlap serves as the primary identification data, from which all subsequent annotation is derived. Finally, the identified map coordinates are annotated with gene features obtained from major genomic and informatics databases.

Genomic localization: MapTag

MapTag was developed as an automated method of identifying homologous genomic regions for gene trap sequences using the Ensembl database (18) and SSAHA algorithm (19). MapTag identifies matches in genomic sequence and assembles individual related stretches of genomic similarity using a basic splice model that takes into account exon boundaries, allowing the processing of nucleotide sequences of cDNA or genomic origin. The program filters alignment results, applying simple heuristics using the overall match length, percent identity and exon coverage, and filtering to remove pseudogenes, to determine the best match. As gene trap sequence tags are typically short and imperfect, they can be difficult to identify to a unique locus. The protocol used by MapTag greatly improves the ability to differentiate between matches that show comparable levels of similarity by correctly selecting the result that exhibits a correspondence to the insertion site. The program returns the identified genomic region and an estimate of the match confidence.

Transcript identification: AutoIdent

The pipeline also uses transcript identification to provide an independent and orthogonal identification method. AutoIdent, an automated protocol developed by BayGenomics, uses the BLAST (20) algorithm to identify the most similar sequences in the GenBank non-redundant nucleotide database at NCBI. When the program identifies many high-scoring matches to very similar (synonymous) genes, AutoIdent adds steps to filter results and condense synonymous transcripts to obtain a single result. The program applies stringent criteria for acceptance of a high-quality gene identification but allows more relaxed criteria to identify multiple matching sequences or to a homologous sequence in another species. At the end of the process, AutoIdent returns the best match transcript along with alignment data.

Identification reconciliation and confirmation

Genomic and transcript identification data are stored in the IGTC database. The pipeline then compares results from the two identification protocols, using overlap between the

genomic coordinates from MapTag and the genomic localization of transcripts returned by AutoIdent to confirm the identification. Gene trap sequences are assigned as localized when the genomic and transcript map coordinates overlap, or when only one protocol returns map coordinates. If the identified coordinates conflict or neither protocol returns map data, the gene trap sequence is classified as unlocalized and does not go through the rest of the annotation pipeline. This reconciliation step assures that each identified cell line is mapped to a genomic insertion locus with high confidence and in a manner that is fully documented.

Annotation of gene trap cell lines

Gene trap cell line annotation is based on the confirmed genomic coordinates as the primary identification data. Map coordinates are used to query the Ensembl and Entrez (21) databases to obtain gene features for the identified locus. These accessions are used as primary keys to further query Ensembl, Entrez and the Mouse Genome Informatics (MGI) resource (22) for secondary annotation data associated with the trapped genes, including major gene accession systems, Gene Ontology classifications (GO), protein domain, structure and function, PubMed and phenotype data, homology and orthology data, and microarray probesets. In addition to supporting an extensive array of informatics data, the IGTC site also includes tissue-specific expression data from the SymAtlas project (23) and biological pathway and GO hierarchy diagrams with trapped genes marked, produced using the GenMAPP program (24).

The IGTC database

The IGTC uses the open source MySQL database platform (www.mysql.com). The database is populated in an automated process using results from the MapTag and AutoIdent identification protocols and is structured to optimize information access via Web queries. New gene trap cell line entries in dbGSS are downloaded weekly and run through the IGTC pipeline. Identification results from MapTag are updated with each Ensembl build and AutoIdent is programmed to regularly BLAST sequences in the database to update accession numbers and other changes in the information provided by GenBank. Annotation data generation is synchronized with the Entrez, MGI and Ensembl databases, and the IGTC database is updated by downloading information from these sites as necessary. The information in the IGTC database is available upon request in a tab-delimited or database compatible format.

ACCESSING GENE TRAP DATA

The IGTC site provides a user-friendly approach to gene trap data, allowing researchers to access the gene trap database from a sequence, accession number or ID, expression or pathway perspective using a variety of interfaces for searching and viewing gene trap data. The site is organized around the cell line annotation page, where the user can view all annotation data for a selected cell line (Figure 2). Primary identification and annotation data appear at the top of the page with a link to detailed identification results. This is followed by expandable lists of secondary annotation data, which provide information useful in the selection and analysis of cell lines of potential interest. Below this is a section containing details about the cell line and gene trap vector, including primer sequences and a description of the vector properties. To aid in the comparison of insertion sites of different cell lines that trap the same gene, the bottom of the cell line annotation page contains a diagram showing the gene trap sequence aligned to the transcripts returned by AutoIdent. The Website provides a similar annotation page for all trapped genes, organized by gene ID. Researchers can also access gene trap data via links from other major informatics resources, including the genome browsers and primary accession pages at NCBI (25), Ensembl (26) and UCSC (27).

The IGTC Website offers users diverse ways to find gene trap cell lines, ranging from searches using protein data, microarray probesets or nucleotide sequences, to screening for traps placed in the context of biological pathways or in genes that demonstrate a particular expression profile. Figure 3 lists the ways in which gene trap data can be accessed, categorized by access type, details about available data type and data access point. Users can search the IGTC database by accession number, ID or keyword and chromosomal location. Results from database searches are displayed as a list of cell line IDs or gene symbols, which link to the individual cell line or gene annotation page. These lists can be exported as tab-delimited files for use in spreadsheet programs or custom databases. Users can also browse the database by MGI Marker Symbol, Gene Name or chromosome location. BLAST analysis can be performed using nucleotide sequence for a gene or locus of interest to search against gene trap sequences or genomic sequence of trapped genes. Trapped genes can be viewed with a pathway perspective using biological pathway diagrams and functional GO groupings, which are colored by the number of cell lines available for each gene. Users can also search for traps in genes with a designated expression profile by selecting a tissue of interest and choosing the expression level of the gene relative to the median tissue expression.

Finally, users can browse gene trap cell lines displayed at major external informatics sites, such as genome browsers and gene pages. IGTC cell lines have been mapped to genomic sequence using the NCBI map viewer, UCSC genome browser and Ensembl genome browser. IGTC gene traps are maintained at these sites either as standard map tracks or as user-configured tracks. (Users should follow site-specific directions to view gene trap data using these browsers.) In addition, gene trap cell lines are listed on some of the major gene pages and in mouse strain resource databases, and the IGTC maintains a full list of partnering sites. By integrating gene trap data into the larger context provided by these bioinformatics resources, the IGTC can reach more potential users who are interested in genomic resources and functional genetics.

SUMMARY

The establishment of the IGTC database and Website marks a major advance in making large-scale mouse knockout resources available to the scientific community. Through the collaboration of gene trap projects worldwide, a standardized

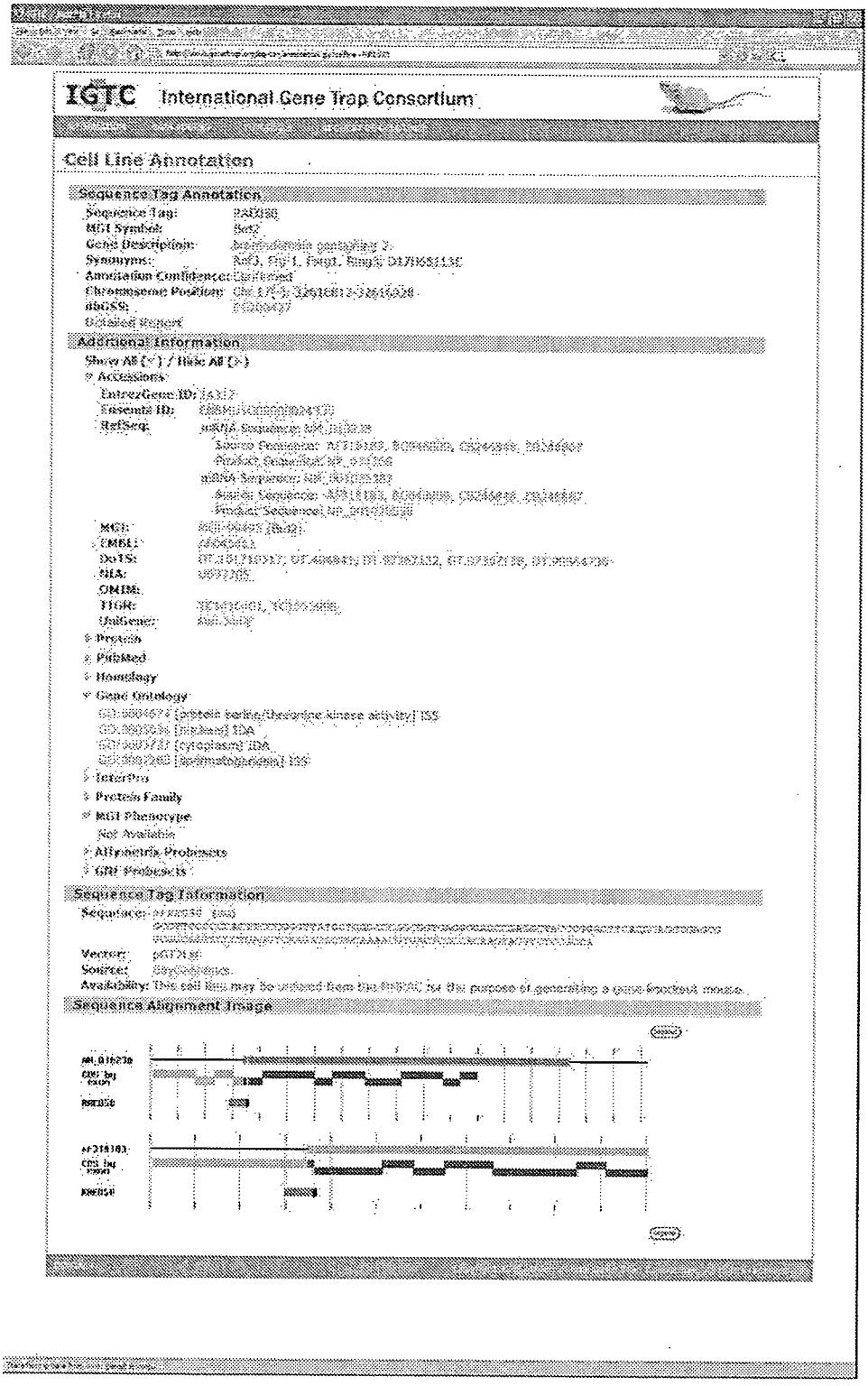


Figure 2. IGTC Website cell line annotation page. The image provides an example of the main data page for gene trap cell lines. Users can view primary identification and annotation data, with a link to detailed reports from the IGTC pipeline. An extensive amount of related annotation data is also presented to give researchers more information about the trapped locus. Finally, the page shows details about the cell line and vector and an image showing all gene trap cell line sequences aligned to the trapped gene. From this page, users are directed to the IGTC member site for cell line requests.

ICON	Access Type	Data Access Details	Access Point
?	Search	Search the IGTC database for trapped genes by keyword, accession, or cell line features	IGTC Website
	Browse	Browse trapped genes by MGI Marker Symbol, gene name, or chromosome	IGTC Website
	BLAST	Search for homologous gene trap sequences using BLAST alignment	IGTC Website
	Biological Pathways	View trapped genes visually using biological pathway diagrams and GO hierarchies	IGTC Website
	Gene Expression	Search for trapped genes that match a desired tissue-specific expression profile	IGTC Website
	Nucleotide Sequence	Search the Genome Survey Sequence database at NCBI for gene trap sequences	dbGSS
	Genome Resources	View gene trap cell line data integrated at major bioinformatics websites	Ensembl, Entrez, MGI
	Genome Browsers	Visually browse genome sequence with gene trap data mapped on to chromosomes	NCBI, UCSC, Ensembl
	Mouse Resources	Find available gene trap cell lines at mutant mouse resource websites	Mouse strain resources

Figure 3. Gene trap data access. Several methods for users to find gene trap cell lines of interest are illustrated.

identification and annotation pipeline has been developed to analyze gene trap data and offer the data access tools necessary for maximum resource utility. For the first time, researchers investigating gene function in the mouse can query all publicly available gene trap ES cell lines at a single site. The IGTC Website contains approximately 45 000 cell lines harboring mutations in nearly 40% of mouse genes (15), including many genes with gene trap cell lines representing multiple mutant alleles. Researchers can search and browse the resource based on accession numbers or IDs, keywords, sequence data, tissue expression and biological pathways, and can also access the IGTC site from other major informatics sites. Furthermore, they can easily request IGTC cell lines for functional characterization of gene function and disease models. For example, BayGenomics cell lines are available from the Mutant Mouse Regional Resource Center (MMRRC) at the University of California Davis (www.mmrrc.org). The MMRRC-UC Davis also offers to microinject ES cells to derive knockout mice for investigators. Soon, the Soriano and TIGEM collections will also be available from the MMRRC. The IGTC Website will grow and continue to develop new tools and features as the diversity of trapped genes and experimental options for using gene trap cell lines expand.

ACKNOWLEDGEMENTS

This work has been supported by NIH grants U01 HL66600 and P41 RR01081, Canadian Institutes of Health grant GOP-36055 and by The Wellcome Trust Sanger Institute. The authors thank Carol Bult at Mouse Genome Informatics, Deanna Church at the National Center for Biotechnology Information and Bob Kuhn at University of California at Santa Cruz for their helpful suggestions. Funding to pay the Open Access publication charges for this article was provided by the University of California, San Francisco.

Conflict of interest statement. None declared.

REFERENCES

1. Waterston,R.H., Chinwalla,A.T., Cook,L.L., Delehaunty,K.D., Fewell,G.A., Fulton,L.A., Fulton,R.S., Graves,T.A., Hillier,L.W., Lindblad-Toh,K. *et al.* (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.
2. Sung,Y.H., Lee,H.-W. and Song,J. (2004) Functional genomics approach using mice. *J. Biochem. Mol. Biol.*, **37**, 122–132.
3. Nolan,P.M. (2000) Generation of mouse mutants as a tool for functional genomics. *Pharmacogenomics*, **1**, 243–255.
4. Stanford,W.L., Cohn,J.B. and Cordes,S.P. (2001) Gene-trap mutagenesis: past, present and beyond. *Nature Rev. Genet.*, **2**, 756–768.
5. Branda,C.S. and Dymecki,S.M. (2004) Talking about a revolution: The impact of site-specific recombinases on genetic analyses in mice. *Dev. Cell*, **6**, 7–28.
6. Schnutgen,F., De-Zolt,S., Van Sloun,P., Hollatz,M., Floss,T., Hansen,J., Altschmied,J., Seisenberger,C., Ghyselinck,N.B., Ruiz,P. *et al.* (2005) Genomewide production of multipurpose alleles for the functional analysis of the mouse genome. *Proc. Natl Acad. Sci. USA*, **102**, 7221–7226.
7. Cobellis,G., Nicolaus,G., Iovino,M., Romito,A., Marra,E., Barbarisi,M., Sardiello,M., Di Giorgio,F.P., Iovino,N., Zollo,M. *et al.* (2005) Tagging genes with cassette-exchange sites. *Nucleic Acids Res.*, **33**, e44.
8. Austin,C.P., Battey,J.F., Bradley,A., Bucan,M., Capecchi,M., Collins,F.S., Dove,W.F., Duyk,G., Dymecki,S., Eppig,J.T. *et al.* (2004) The knockout mouse project. *Nature Genet.*, **36**, 921–924.
9. Auwerx,J., Avner,P., Baldock,R., Ballabio,A., Balling,R., Barbacid,M., Berns,A., Bradley,A., Brown,S., Carmeliet,P. *et al.* (2004) The European dimension for the mouse genome mutagenesis program. *Nature Genet.*, **36**, 925–927.
10. Hicks,G.G., Shi,E.-G., Li,X.-M., Li,C.-H., Pawlak,M. and Ruley,H.E. (1997) Functional genomics in mice by tagged sequence mutagenesis. *Nature Genet.*, **16**, 338–344.
11. Wiles,M.V., Vauti,F., Otte,J., Fuchtbauer,E.-M., Ruiz,P., Fuchtbauer,A., Arnold,H.-H., Lehrach,H., Metz,T., Von Melchner,H. *et al.* (2000) Establishment of a gene-trap sequence tag library to generate mutant mice from embryonic stem cells. *Nature Genet.*, **24**, 13–14.

12. Stryke,D., Kawamoto,M., Huang,C.C., Johns,S.J., King,L.A., Harper,C.A., Meng,E.C., Lee,R.E., Babbit,P.C., Ferrin,T.E. *et al.* (2003) BayGenomics: a resource of insertional mutations in mouse embryonic stem cells. *Nucleic Acids Res.*, **31**, 278–281.
13. Hansen,J., Floss,T., Wurst,W., Van Sloun,P., Schnütgen,F., Von Melchnert,H., Füchtbauer,E.-M., Vauti,F., Arnold,H.-H. and Ruiz,P. (2003) A large-scale, gene-driven mutagenesis approach for the functional analysis of the mouse genome. *Proc. Natl Acad. Sci. USA*, **100**, 9918–9922.
14. To,C., Epp,T., Reid,T., Lan,Q., Yu,M., Li,C.Y., Ohishi,M., Hant,P., Tsao,N., Casallo,G. *et al.* (2004) The Centre for Modeling Human Disease Gene Trap resource. *Nucleic Acids Res.*, **32**, D557–559.
15. Skarnes,W.C., Nord,A.S., Cox,T., von Melchner,H., Wurst,W., Hicks,G., Young,S.G., Conklin,B.R., Ruiz,P., Soriano,P. *et al.* (2004) A public gene trap resource for mouse functional genomics. *Nature Genet.*, **36**, 543–544.
16. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2005) GenBank. *Nucleic Acids Res.*, **33**, D34–D38.
17. Kwan,Y.K. (2002) LinkOut: Explore beyond PubMed and Entrez. *D-Lib Mag.*, **8**.
18. Hubbard,T., Andrews,D., Caccamo,M., Cameron,G., Chen,Y., Clamp,M., Clarke,L., Coates,G., Cox,T., Cunningham,F. *et al.* (2005) Ensembl 2005. *Nucleic Acids Res.*, **33**, D447–D453.
19. Ning,Z., Cox,A.J. and Mullikin,J.C. (2001) SSAHA: a fast search method for large DNA databases. *Genome Res.*, **11**, 1725–1729.
20. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
21. Maglott,D., Ostell,J., Pruitt,K.D. and Tatusova,T. (2005) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **33**, D54–D58.
22. Eppig,J.T., Bult,C.J., Kadin,J.A., Richardson,J.E., Blake,J.A., Anagnostopoulos,A., Baldarelli,R.M., Baya,M., Beal,J.S., Bello,S.M. *et al.* (2005) The Mouse Genome Database (MGD): from genes to mice—a community resource for mouse biology. *Nucleic Acids Res.*, **33**, D471–475.
23. Su,A.I., Wiltshire,T., Batalov,S., Lapp,H., Ching,K.A., Block,D., Zhang,J., Soden,R., Hayakawa,M., Kreiman,G. *et al.* (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl Acad. Sci. USA*, **101**, 6062–6067.
24. Dahlquist,K.D., Salomonis,N., Vranizan,K., Lawlor,S.C. and Conklin,B.R. (2002) GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. *Nature Genet.*, **31**, 19–20.
25. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S., Helmberg,W. *et al.* (2005) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **33**, D39–D45.
26. Stalker,J., Gibbins,B., Meidl,P., Smith,J., Spooner,W., Hotz,H.-R. and Cox,A.V. (2004) The Ensembl Web site: mechanics of a genome browser. *Genome Res.*, **14**, 951–955.
27. Karolchik,D., Baertsch,R., Diekhans,M., Furey,T.S., Hinrichs,A., Lu,Y.T., Roskin,K.M., Schwartz,M., Sugnet,C.W., Thomas,D.J. *et al.* (2003) The UCSC Genome Browser Database. *Nucleic Acids Res.*, **31**, 51–54.

J. Biochem. 140, 793–798 (2006)
doi:10.1093/jb/mvj208

Negative Selection with the *Diphtheria toxin A fragment* Gene Improves Frequency of Cre-Mediated Cassette Exchange in ES Cells

Kimi Araki[1,*], Masatake Araki[2] and Ken-ichi Yamamura[1]
[1]Department of Developmental Genetics, Institute of Molecular Embryology and Genetics, and [2]Institute of Resource Development and Analysis, Kumamoto University, Honjo 2-2-1, Kumamoto 860-0811

Received August 12, 2006; accepted October 11, 2006

The Cre-*lox* system is an important tool for genetic manipulation in embryonic stem cells. We previously reported that the cassette exchange strategy using the mutant *lox*66/71 and *lox*2272 combination showed high recombination efficiency and stability. However, the efficiency was strongly affected by the position of chromosomal target *lox* sites. To enrich successful cassette exchange events, even in clones showing lower recombination efficiency, we have improved exchange vector. The *Diphtheria toxin A fragment* gene was placed in the un-exchanged region for negative selection and the *puromycin N-acetyltransferase* gene, instead of the *neomycin phosphotransferase* gene, was used for positive selection. By reducing random integration, the frequency of successful cassette exchange increased up to 2–4 fold. Furthermore, by adding the third *lox* site to induce intrarmolecular recombination, the recombination efficiency of cassette exchange itself was improved, and the frequency increased to maximum 5 fold, in which the percentage of exchanged clones reached to 50–70%. This strategy should be useful for other recombinase-mediated cassette exchanges.

Key words: cassette exchange, Cre recombinase, *Diphtheria toxin A fragment* (*DT-A*) gene, embryonic stem (ES) cells, site-directed recombination.

Abbreviations: bsr, blasticidin S resistant; DT-A, diphtheria toxin A fragment; ES, embryonic stem; neo, neomycin phosphotransferase; NLSLacZ, *lacZ* gene fused with the nuclear localizing signal, pA, polyadenylation; pac, puromycin *N*-acetyltransferase; Pgk, phosphoglycerate kinase-1; RMCE, Recombinase Mediated Cassette Exchange; tk, thymidine kinase; X-gal, 5-bromo-4-chloro-3-indolyl β-D-galactopyranoside.

The Cre-mediated site-specific targeting system is a powerful tool for genome engineering in mammals, especially in mouse embryonic stem (ES) cells, because it allows precise and repeated knock-ins of any DNA to target *lox* sites introduced by gene targeting or gene trapping. (*1–3*) However, intermolecular recombination between wild-type *lox*P sites, *i.e.*, integrative recombination, is inefficient due to re-excision through intramolecular recombination (*4*).

In order to perform Cre-mediated insertion or replacement, two kinds of mutant *lox* sites have been developed. One is a pair of *lox* sites with a 5 bp mutation in the left or right end of the *lox* sequence, such as *lox*66/71 (*5*, *6*). Recombination between a chromosomally located *lox*71 site and a *lox*66 site on a targeting plasmid results in site-specific integration of the plasmid producing a double mutant *lox* site at both ends and a wild-type *lox*P site. Since the binding affinity of the double *lox* mutant site for Cre recombinase is reduced, the integrated plasmid is stably retained. The other mutant site is a heterospecific *lox* site that has mutation(s) in the central 8 bp spacer region (*7–9*). The recombination using heterospecific *lox* sites is termed Recombinase Mediated Cassette Exchange (RMCE) (*10*), in which the recombination does not occur between two *lox* sites differing in the spacer region whereas *lox* sites with identical spacer regions can be recombined efficiently. Until now, *lox*511 (*11*), *lox*2272 (*12*) and *lox*5171 (*13*) have been successfully used in ES cells.

Recently, we have shown that the combination of *lox* 66/71 and the heterospecific *lox*2272 site gave high recombination efficiency and stability even with Cre recombinase (*14*), and developed an exchangeable gene trap vector carrying *lox*71, *lox*P and *lox*2272 (*3*). Using the trap vector, we can initially carry out random insertional mutagenesis in ES cells, and then replace the reporter gene in the trap vector with any gene of interest to be expressed under the control of the trapped promoter through RMCE.

In RMCE, a targeting plasmid carrying an integrated floxed cassette is co-electroporated with a Cre-expression vector into cells in their circular forms to reduce random integration. However, random integration of targeting plasmid also occurs at a considerable frequency, probably due to nicks in plasmid DNA strands. In our previous study, the percentage of random integrants was over 50% even in the clone that showed the highest RMCE frequency, indicating that random integration is more efficient than RMCE. The most effective method to eliminate random integrants is negative selection using the *thymidine kinase* (*tk*) gene of the Herpes simplex virus (*15*). In this method, the *tk* gene is placed on a chromosomal target construct between two heterospecific *lox* sites, and recombined clones, where the *tk* gene should be removed by

*To whom correspondence should be addressed. E-mail: arakimi@gpo.kumamoto-u.ac.jp

replacement with other DNA on the targeting vector, are selected with ganciclovir. Several groups reported that almost all clones obtained after electroporation had targeted replacement (*12*, *16*, *17*). This *tk* negative selection is very useful because it is not necessary to have any positive marker on targeting vectors. On the other hand, the necessity of placing the *tk* gene in a chromosomal target construct is quite inconvenient for gene trapping, since it is preferable that trap vectors should consist of minimum elements to avoid unexpected effect(s) caused by introduction of vector elements. Actually, male sterility caused by ectopic expression of the *tk* gene in testis has been reported (*18*).

In this study, we aimed to increase recombination frequency by improving exchange vectors, not chromosomal target constructs. We used the *Diphtheria toxin A fragment* (*DT-A*) gene (*19*) for negative selection to reduce random integrants, changed the *neomycin phosphotransferase* (*neo*) gene into the *puromycin N-acetyltransferase* (*pac*) gene, and added a *lox*66 site to induce intra-molecular recombination within the exchange vector. With the improved exchange vector, the frequency of exchanged clones was increased to 2–5 folds.

MATERIALS AND METHODS

Plasmids—Plasmids p66-2272 and pCAGGS-Cre were described previously (*14*, *20*). The plasmid containing the MC1 promoter–DT-A gene was kindly provided by Dr. S. Aizawa (*19*). However, The original MC1-DT-A cassette had no polyadenylation (pA) signal and the MC1 promoter of the cassette carried only one copy of the enhancer sequence. The MC1-DT-A-pA cassette used in this study was constructed by replacing the original promoter by the MC1 promoter from MC1-neo-pA cassette (STRATAGENE, La Jolla, USA) and by adding the pA signal of the rabbit β-grobin gene. The pMC1-DTA plasmid was constructed by inserting the MC1-DT-A-pA cassette into p66-2272. The pPGK-DTA and pPac-DTA were constructed by replacing the MC1-neo-pA cassette into the Pgk-neo-pA and Pgk-pac-pA cassette, respectively. The pPac-DTA-66 was constructed by inserting a *lox*66 fragment into pPac-DTA. The sequences of all *lox* sites in these plasmids were confirmed by DNA sequencing.

Cell Culture and Electroporation—ES cell culture and establishment of cell lines carrying a single copy of target *lox* sites was described previously (*14*). For RMCE, the cells ($3–6 \times 10^6$ cells/0.8 ml in PBS) were electroporated at 400 V and 125 μF, and plated into two 10 cm plates. G418 selection was started after 24 h of electroporation at 200 μg/ml for 7 days. For puromycin selection, cells were fed 2 μg/ml of puromycin containing medium for 24 h × 2 times on day 1 and 4 after electroporation. On day 8, colonies were stained with 5-bromo-4-chloro-3-indolyl β-D-galactopyranoside (X-gal) or picked and expanded for DNA analysis.

A series of 5 exchange vectors with a cell line was performed on the same day, and each series of electroporation was repeated at least four times on independent days.

PCR Analysis—Genomic DNA (0.05–0.1 μg) was subjected to 32 cycles of amplification (each cycle consisted of 1 min at 94°C, 1.5 min at 58°C and 1.5 min at 72°C) using AmpliTaq polymerase (Perkin-Elmer). Primer sequences are as follows; AG2, 5′-CTGCTAACCATGTTCATGCC-3′; LZUS3, 5′-GCGCATCGTAACCGTGCAT-3′; bsr-2, 5′-GCAGAAATCGGAGGAAGAAG-3′; bsr-3, 5′-CAACTCCCTACACATACCAC-3′; FRT-S, 5′-GCTTCAAAAGCGCTCTGAAG-3′; DTA1, 5′-TACCACGGGACTAAACCTGG-3′; DTA2, 5′-CGCTTAACGCTTTCGCCTGT-3′.

Statistical Analyses—The recombination efficiencies and relative number of blue or white colonies were evaluated by non-repeated measures ANOVA. Where a significant difference ($p < 0.05$) was identified, the differences were analyzed further with SNK tests for multiple comparisons.

RESULTS

Experimental Design—The strategy to assess the RMCE frequencies used previously (*14*) is outlined in Fig. 1A. The chromosomal target is the CAG promoter-*lox*71-*Blasticidin S resistant* (*bsr*) gene-pA-*lox*511-FRT-*lox*2272. Exchange vectors contain *lox*66-the promoter-less *lacZ* gene fused with the nuclear localizing signal (*NLSlacZ*)-a selection marker gene-*lox*2272. ES cell lines carrying the chromosomal target are co-electroporated with the exchange vector and Cre-expression vector in their circular forms, and then selected with appropriate drug according to the selection marker gene. The *cre* gene on the expression vector is transiently expressed, and Cre protein mediates site-specific recombination between the chromosomal *lox*71 and *lox*66 on the targeting plasmid, and the chromosomal *lox*2272 and *lox*2272 on the targeting plasmid, resulting in cassette exchange of the *bsr* gene with the *NLSlacZ*-selection marker cassette. Since there is no negative selection marker on the chromosomal target construct, both random integrants and site-specific recombinants become drug resistant, but only the colonies where RMCE had occurred are stained blue with X-gal, since the *NLSLacZ* gene is inserted downstream of the CAG promoter. The percentage of blue colonies represents the frequency of RMCE. When only the exchange vector was electroporated, no blue colonies appeared (data not shown), indicating that gene trapping events hardly occur with electroporation using the exchange vectors in circular forms.

We used four ES cell lines, 71-5F2-7, 71-5F2-10, 71-5F2-23 and 71-5F2-26 carrying a single copy of the chromosomal target construct. The cell lines 71-5F2-10 and 71-5F2-23 were used in our previous study (*14*), and showed lower and higher recombination frequency, respectively. We added two cell lines, 71-5F2-7 and 71-5F2-26, which showed lower recombination frequency, to examine enrichment effects of exchange vectors on several different "inactive" positions. Since the original 71-5F2-7, 71-5F2-10 and 71-5F2-26 clones were contaminated with wild-type ES cells at a considerable percentage (data not shown), all four lines were re-cloned through limiting dilution and Blasticidin S selection. Table 1 shows the RMCE frequencies and the number of blue colonies in each re-cloned line electroporated with pCAGGS-Cre and p66-2272, which contains only the MC1-*neo*-pA cassette as a positive selection marker (Fig. 1B, top) and gave the best frequency in the previous study. Since the conditions of electroporation were optimized for Cre-mediated recombination, the frequency in 71-5F2-23 increased to 35.3% from the frequency obtained in the previous study, 26.3%. In 71-5F2-23, the number of blue colonies was 3–4 times higher than in the

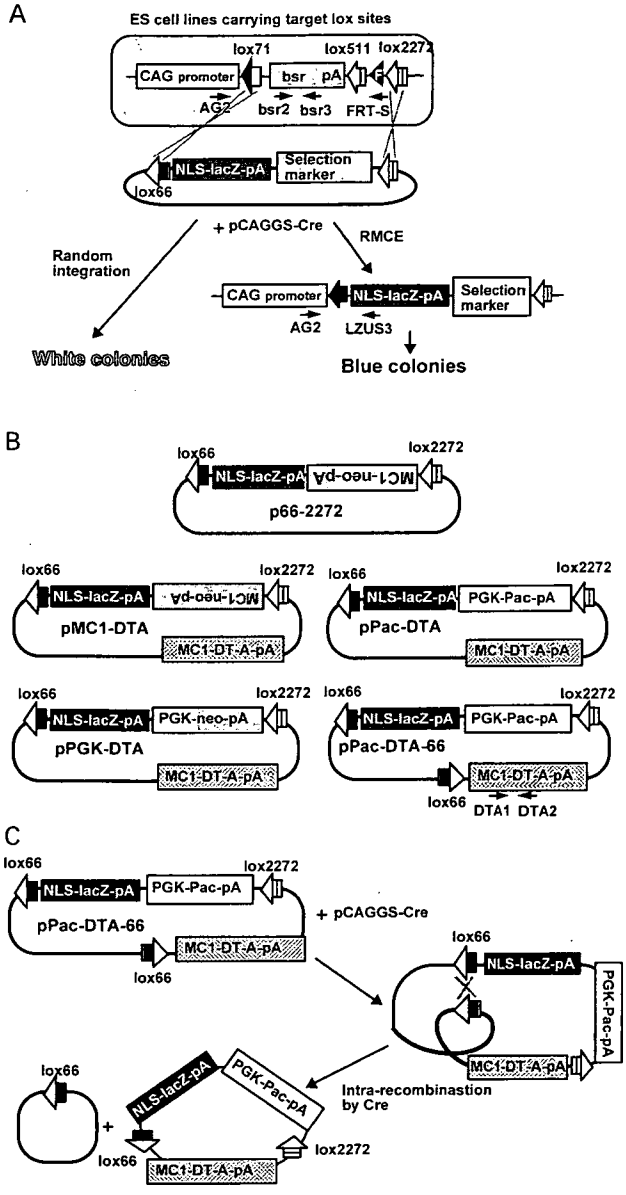


Fig. 1. **(A) Experimental strategy for comparison of frequency of RMCE.** ES cell lines carrying a single copy of the CAG-*lox*71-bsr-pA-*lox*511-FRT-*lox*2272 fragment were established. The cell lines were co-electroporated with the Cre expression vector and the targeting plasmids carrying the promoter-less *NLS-lacZ* gene. Through RMCE, the *NLSlacZ* gene is joined to the CAG promoter, resulting in positive staining with X-gal. Since the targeting plasmids contain a selection marker gene, random integrants can also appear. However, colonies with random integration are not stained because there is no promoter for the *NLSlacZ* gene. The percentage of blue colonies represents the frequency of RMCE. Positions of PCR primers used in Table 2 are shown as small arrows with the name of the primer. **(B) Targeting vectors used in this study. (C) Predicted recombination intermediate of pPac-DTA-66.** Since pPac-DTA-66 carries two *lox*66 sites, intra-molecular recombination should occur first to divide into two circular molecules. The molecule including *lox*66 and *lox*2272 becomes a substrate of RMCE.

other three clones, indicating high recombination efficiency probably due to open chromosomal configuration around the target *lox* sites. The goal of this study is to improve targeting vectors to enrich blue colonies in all cell lines, including the "lower" recombination-frequency clones.

Table 1. **Recombination frequency with p66-2272 targeting vector in clones used in this study.**

Cell line	Number of Blue colonies ± SD	Recombination Frequency (%) ± SD
71-5F2-7	121 ± 29.6	20.2 ± 5.89
71-5F2-10	92.0 ± 34.6	16.5 ± 2.67
71-5F2-23	311 ± 166	35.3 ± 7.89
71-5F2-26	67.0 ± 17.8	11.6 ± 2.44

$N = 4$ in each clone.

Exchange Vectors—Exchange vectors used in this study are shown in Fig. 1B. In pMC1-DTA, we added the MC1 promoter-*DT-A* gene-pA cassette to p66-2272 in the un-integrated region by RMCE (outside of two *lox* sites) for negative selection to reduce random integrants. The choice of selection marker gene is also important for colony formation efficiency. In pPGK-DTA, the mouse *phosphoglycerate kinase-1* (*Pgk*) promoter–*neo*-pA cassette, which shows higher colony efficiency than MC1-neo, is used, and in pPac-DTA, the *Pgk* promoter–*pac*-pA cassette, which is known to that puromycin sensitive cells are killed quite quickly and the drug selection completes within 24–48 h, is used. In addition, we constructed a pPac-DTA-66 plasmid, in which an additional *lox*66 site was placed between the DT-A cassette and plasmid vector sequence. After electroporation with Cre expression plasmid, intra-molecular recombination between *lox*66 sites should take place first, resulting in two circular molecules as shown Fig. 1C. By minimizing the size of the targeting DNA molecule, we expected to reduce random integration.

Recombination Frequencies—Percentages of blue colonies, *i.e.*, RMCE frequencies, in the 4 lines are shown in Fig. 2A. The pPac-DTA-66 plasmid (vector no. 5) gave the highest RMCE frequency in all lines with statistical significant differences. Even in the 'lower' recombination-frequency clone 71-5F2-26, the frequency exceeded 50%, and in the "high" recombination-frequency clone 71-5F2-23 line, it reached to 75%. In order to evaluate the enrichment effect, relative numbers of blue or white colonies against the number of blue colonies obtained with p66-2272 were calculated and are shown in Fig. 2B. When the DT-A cassette was added to p66-2272 (pMC1-DTA), the number of white colonies was significantly reduced to almost half (vector no. 2), indicating that negative selection of the DT-A was effective. However, since the number of blue colonies was also slightly reduced, the percentages of blue colonies did not change significantly, but were higher than in p66-2272. With the use of the PGK-neo-pA, both the number of blue and white colonies increased, but no significant changes of frequency compared to pMC1-DTA observed (vector no. 3).With the use of PGK-pac-pA, the RMCE frequency increased to 30–50% (vector no. 4). Interestingly, with pPac-DTA-66 plasmid, the numbers of blue colonies (targeted integration) were significantly increased in all lines, whereas the numbers of white colonies (random integration) were unchanged (vector no. 5). This indicates that smaller DNA molecule can access more easily to chromosomal target *lox* sites.

In order to analyze integration patterns, 71-5F2-7 and 71-5F2-10 colonies were picked from cells electroporated

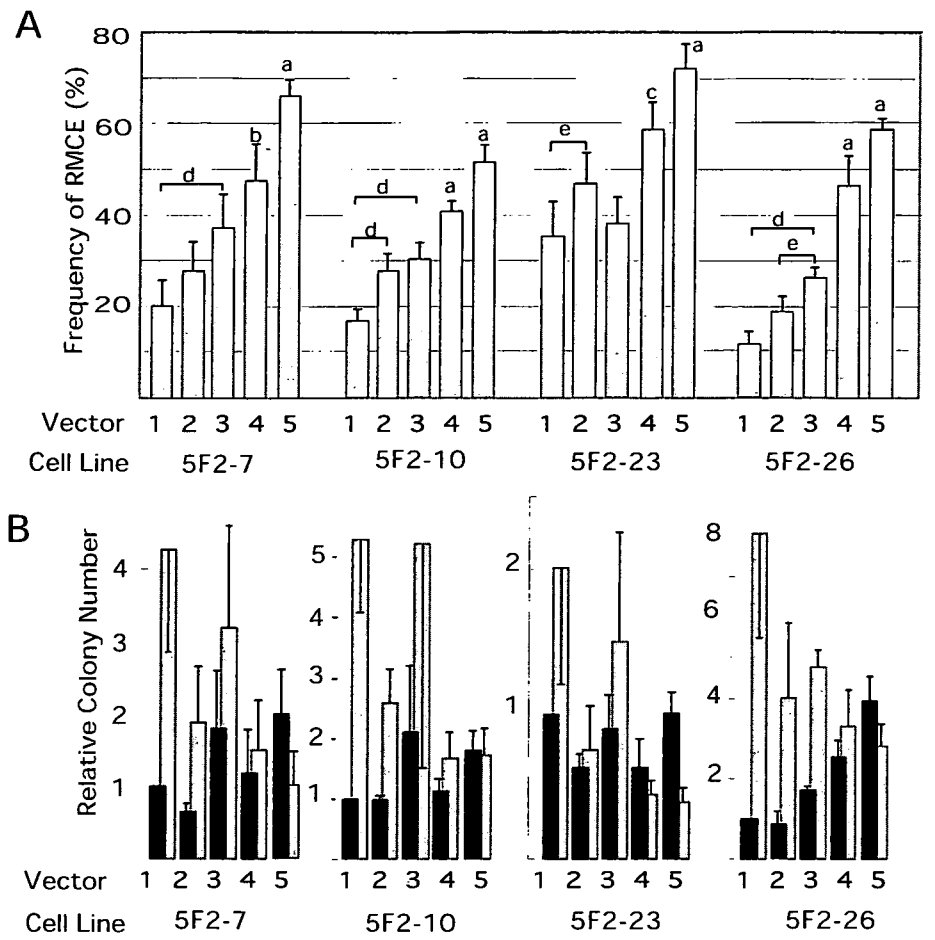


Fig. 2. **(A) Frequency of RMCE.** Twenty micrograms of each replacement plasmid and the Cre-expressing vector were co-electroporated, and after drug selection for 7 days, colonies were stained with X-gal, and the percentage of positive colonies was scored as the frequency of RMCE. Numbers under the graph indicate the targeting vector used in each electroporation. 1, p66-2272; 2, pMC1-DTA; 3, pPGK-DTA; 4, pPac-DTA, 5, pPac-DTA-66. The means ± SD of at least four independent electroporations are represented. All cell lines show significant differences at $P < 0.01$ by ANOVA analysis. Statistical significances among vectors were further analyzed by the SNK test. a, $P < 0.01$ compared to all other vectors; b, $P < 0.01$ compared to vector 1, 2 and 5 and $P < 0.05$ compared to vector 3; c, $P < 0.01$ compared to vector 1, 3 and 5 and $P < 0.05$ compared to vector 2; d, $P < 0.01$ compared to the indicated vector; e, $P < 0.05$ compared to the indicated vector. **(B) Relative Blue (solid bars) or white (gray bars) colony numbers.** The number of blue colonies obtained with p66-2272 arbitrarily set at 1.

Table 2. **Genomic DNA analysis of isolated subclones by PCR.**

Cell line	Total no. of clones analyzed	X-Gal staining (%)	PCR pimers			
			AG2/LZUS3	bsr2/bsr3	AG2/FRT-AS	DTA1/DTA2
71-5F2-7	18	Blue 8 (44)	8	0	ND	0
		White 10 (56)	0	5	5	2
71-5F2-10	23	Blue 14 (61)	14	0	ND	0
		White 9 (39)	0	6	3	0

Subclones obtained after coelectroporation with pPac-DTA-66 and pCAGGS-Cre were picked and expanded for genomic DNA preparation and PCR analysis. Part of each clone was stained with X-gal, and the number of blue or white clones is represented. PCR analyses were performed with the indicated primers (see Fig. 1), and among the each blue or white clones, the number of clones showed a band of expected size is represented. ND, not done.

with pPac-DTA-66 and pCAGGS-Cre, genomic DNAs were prepared, and PCR analysis was performed. As shown in Fig. 1A, the 5′-junction of the recombination can be amplified using the primers AG2 and LZUS3, and all blue clones gave a band of the expected size (Table 2, data of electrophoresis not shown). Then, the presence of the *bsr* gene, which should be removed through RMCE, was examined with the primers bsr-1 and bsr-2. All blue clones were negative for the *bsr* gene detecting PCR, however, only 11 clones out of 19 white clones retained the bsr gene, unexpectedly. Since it is reported that *lox*511 can be recombined with *lox*P or *lox*71 (*9*), PCR analysis with AG2 and FRT-AS primers was performed to detect recombination between *lox*71 and *lox*511. The 8 clones that did not carry the *bsr* gene exhibited a 201-bp band. We cloned the product into the T-vector and confirmed that the sequence corresponded to the recombination product *lox*71 and *lox*511. The spacer sequence of the *lox* site in the recombined product was the wild-type sequence (data not shown). The *DT-A* gene was detected in only 2 clones by PCR, indicating successful negative selection. Since the amplified region was the inside of the ORF, the promoter or pA signal might be deleted in these 2 clones.

DISCUSSION

We have shown here that the combined usage of negative selection of the *DT-A* gene, positive selection of the *Pac* gene and the third *lox* site for intra-molecular recombination efficiently enriched RMCE events. The enrichment

effect was more apparent in the 'lower' recombination-frequency clones, and we could obtain a frequency of 50% and more, which is high enough for the practical use of RMCE.

We could enrich RMCE event through two strategies, one is the use of the *DT-A* negative selection marker gene and the other is the change of positive selection marker gene. The usefulness of the *DT-A* gene which reduce random integration has been already reported in gene targeting (*21*), however, enrichment effect in gene targeting frequency by changing selection marker has not been clearly observed. In comparing two plasmids having the same promoter, pPac-DTA (Pgk-Pac) and pPGK-DTA (Pgk-neo), the number of white colonies was reduced to almost half, probably due to the difference of colony formation efficiency of the *pac* and *neo* gene. On the other hand, in comparison between pMC1-DTA (MC1-neo) and pPac-DTA (Pgk-pac), there is no statistical difference in number of white colonies or blue colonies, except of blue colonies in 71-5F2-26, nevertheless, the RMCE frequency showed statistical difference in the all lines used in this experiment. Why the use of the Pgk-*pac* cassete resulted in higher RMCE frequency? We speculate that the difference of time course in G418 and puromycin selection might be the cause of the enrichment effect. Puromycin kills almost all sensitive cells within 24hrs, whereas G418 takes 2–3 days. The quick elimination of non-transfected cells might result in reduction of spontaneously drug-resistant colonies.

In addition to reduction of random integration, we could increase recombination efficiency itself by adding the third *lox* site to induce intra-molecular recombination, which results in 33% of size reduction of targeting DNA molecule by removal of the plasmid vector sequence. Interestingly, the numbers of blue colonies with pPac-DTA-lox66 plasmid were 1.5–1.7 times higher than those with pPac-DTA plasmid, thus, the RMCE efficiencies increase approximately in inverse proportion with the size of targeting DNA molecule. This indicates that the probability of encounter of chromosomal *lox* site and targeting plasmid depends on physical mobility of DNA molecule. In the higher-recombination efficiency clone 71-5F2-23, the chromosomal configuration around *lox* sites is considered to be open, therefore, they always show high efficiency even when the molecular weight of targeting vector is relatively large. On the other hand, in the lower-recombination efficiency clone 71-5F2-26, the chromosomal configuration around *lox* sites is considered to be close, therefore, reduction of the molecular weight of targeting DNA is quite effective to increase RMCE frequency by improving accessibility to chromosomal target *lox* sites. Our results predict that RMCE using large molecular weight plasmid, *ex.* bacterial artificial chromosome, may show low frequency.

The advantage of RMCE using our strategy is the minimum requirement of the chromosomal target structure, *i.e.*, only two heterospecific *lox* sites, and wide and easy application to gene targeting or gene trapping vectors. Thus, our enrichment strategy for RMCE will be a powerful tool in genetic manipulation in ES cells.

We wish to thank Ms. K. Haruna, M. Muta, R. Minato and Y. Tsuruta for technical assistance. This work was supported in part by a Grant-in-Aid on Priority Areas from the Ministry of Education, Science, Culture and Sports of Japan, by a grant from the Osaka Foundation for Promotion of Clinical Immunology.

REFERENCES

1. Nagy, A. (2000) Cre recombinase: the universal reagent for genome tailoring. *Genesis* **26**, 99–109
2. Branda, C.S. and Dymecki, S.M. (2004) Talking about a revolution: The impact of site-specific recombinases on genetic analyses in mice. *Dev. Cell* **6**, 7–28
3. Taniwaki, T., Haruna, K., Nakamura, H., Sekimoto, T., Oike, Y., Imaizumi, T., Saito, F., Muta, M., Soejima, Y., Utoh, A., Nakagata, N., Araki, M., Yamamura, K., andAraki, K. (2005) Characterization of an exchangeable gene trap using pU-17 carrying a stop codon-β*geo* cassette. *Dev. Growth Differ.* **47**, 163–72
4. Fukushige, S. and Sauer, B. (1992) Genomic targeting with a positive-selection lox integration vector allows highly reproducible gene expression in mammalian cells. *Proc. Natl. Acad. Sci. USA* **89**, 7905–9
5. Albert, H., Dale, E.C., Lee, E., and Ow, D.W. (1995) Site-specific integration of DNA into wild-type and mutant lox sites placed in the plant genome. *Plant J.* **7**, 649–59
6. Araki, K., Araki, M., and Yamamura, K. (1997) Targeted integration of DNA using mutant lox sites in embryonic stem cells. *Nucleic Acids Res.* **25**, 868–72
7. Hoess, R.H., Wierzbicki, A., and Abremski, K. (1986) The role of the loxP spacer region in P1 site-specific recombination. *Nucleic Acids Res.* **14**, 2287–300
8. Lee, G. and Saito, I. (1998) Role of nucleotide sequences of loxP spacer region in Cre-mediated recombination. *Gene* **216**, 55–65
9. Langer, S.J., Ghafoori, A.P., Byrd, M., and Leinwand, L. (2002) A genetic screen identifies novel non-compatible loxP sites. *Nucleic Acids Res.* **30**, 3067–77
10. Bouhassira, E.E., Westerman, K., and Leboulch, P. (1997) Transcriptional behavior of LCR enhancer elements integrated at the same chromosomal locus by recombinase-mediated cassette exchange. *Blood* **90**, 3332–44
11. Bethke, B. and Sauer, B. (1997) Segmental genomic replacement by Cre-mediated recombination: genotoxic stress activation of the p53 promoter in single-copy transformants. *Nucleic Acids Res.* **25**, 2828–34
12. Kolb, A.F. (2001) Selection-marker-free modification of the murine beta-casein gene using a lox2272 [correction of lox2722] site. *Anal. Biochem.* **290**, 260–71
13. Osipovich, A.B., Singh, A., and Ruley, H.E. (2005) Post-entrapment genome engineering: first exon size does not affect the expression of fusion transcripts generated by gene entrapment. *Genome Res.* **15**, 428–35
14. Araki, K., Araki, M., and Yamamura, K. (2002) Site-directed integration of the cre gene mediated by Cre recombinase using a combination of mutant lox sites. *Nucleic Acids Res.* **30**, e103
15. Seibler, J., Schubeler, D., Fiering, S., Groudine, M. and,Bode, J. (1998) DNA cassette exchange in ES cells mediated by Flp recombinase: an efficient strategy for repeated modification of tagged loci by marker-free constructs. *Biochemistry* **37**, 6229–34
16. Soukharev, S., Miller, J.L., and Sauer, B. (1999) Segmental genomic replacement in embryonic stem cells by double lox targeting. *Nucleic Acids Res.* **27**, e21
17. Shmerling, D., Danzer, C.P., Mao, X., Boisclair, J., Haffner, M., Lemaistre, M., Schuler, V., Kaeslin, E., Korn, R., Burki, K., Ledermann, B., Kinzel, B., and Muller, M. (2005) Strong and ubiquitous expression of transgenes targeted into the beta-actin locus by Cre/lox cassette replacement. *Genesis* **42**, 229–35

18. al-Shawi, R., Burke, J., Wallace, H., Jones, C., Harrison, S., Buxton, D., Maley, S., Chandley, A. andBishop, J.O. (1991) The herpes simplex virus type 1 thymidine kinase is expressed in the testes of transgenic mice under the control of a cryptic promoter. *Mol. Cell. Biol.* **11**, 4207–16
19. Yagi, T., Nada, S., Watanabe, N., Tamemoto, H., Kohmura, N., Ikawa, Y. andAizawa, S. (1993) A novel negative selection for homologous recombinants using diphtheria toxin A fragment gene. *Anal. Biochem.* **214**, 77–86
20. Araki, K., Imaizumi, T., Okuyama, K., Oike, Y., and Yamamura, K. (1997) Efficiency of recombination by Cre transient expression in embryonic stem cells: comparison of various promoters. *J. Biochem.* **122**, 977–82
21. Yanagawa, Y., Kobayashi, T., Ohnishi, M., Tamura, S., Tsuzuki, T., Sanbo, M., Yagi, T., Tashiro, F., and Miyazaki, J. (1999) Enrichment and efficient screening of ES cells containing a targeted mutation: the use of DT-A gene with the polyadenylation signal as a negative selection maker. *Transgenic Res.* **8**, 215–21

Review Article

Gene trap mutagenesis in mice: New perspectives and tools in cancer research

Ken-ichi Yamamura[1] and Kimi Araki

Division of Developmental Genetics, Institute of Molecular Embryology and Genetics, Kumamoto University, 2-2-1 Honjo, Kumamoto 860-0811, Japan

(Received July 30, 2007/Accepted August 7, 2007/Online publication September 17, 2007)

The complete human DNA sequence of the human genome was published in 2004 and we entered the postgenomic era. However, many studies showed that gene function is much more complex than we expected, and that mutation of disease genes does not give any clue for molecular mechanisms for disease development. Since the first report on gene knockout mice in 1989, knockout mice have been shown to be a powerful tool for functional genomics and for the dissection of developmental processes in human diseases. In accordance with this successful application of knockout mice, three major mouse knockout programs are now underway worldwide, to mutate all protein-encoding genes in mouse embryonic stem cells using a combination of gene trapping and gene targeting. We developed the exchangeable gene trap method suitable for large scale mutagenesis in mice. In this method we can produce null mutation and post-insertional modification, enabling replacement of the marker gene with a gene of interest and conditional knockout. We herein discuss the effect of this gene-driven type approach for cancer research, especially for finding the genes that are related to cancer, but are paid little attention in hypothesis-driven cancer research. (*Cancer Sci* 2008; 99: 1–6)

It is a well known fact that a gene, for example the α-fetoprotein gene, expressed during development is often reactivated in cancer cells. In addition, there are some similarities between carcinogenesis and embryonic development. One important function of development involves the production and organization of all of the diverse types of cells in the body. This generation of cellular diversity is accomplished by cell proliferation, followed by cell differentiation. The processes that organize the different cells into tissues and organs are called morphogenesis and growth. Although cancer tissues seem to be disorganized, we thought that they require cell differentiation and harmonization of these differentiated cells. For example, they need new blood vessels to obtain nutrition and energy for themselves. From this point of view, we hypothesize that many of the genes involved in embryogenesis may have functions in cancer cells, not just as markers, and that we can find genes that have unexpected functions in cancer cells. Thus, we decided to use the gene trap approach to find such genes.

Enhancer trap as a tool for gene identification

A strategy to monitor transcriptionally active regions of the genome was first described in bacteria.[1] It involved the introduction of reporter constructs that require the acquisition of *cis*-acting DNA sequences in the genome to activate reporter gene expression. In this way, genes are identified based on expression information and subsequently cloned from DNA sequences flanking the site of insertion. This approach was then applied for higher organisms using modified vectors suitable for eukaryotic transcription units. Initially, the enhancer trap vector was developed based on the observation that cellular enhancers are capable of acting on heterologous promoters using cultured cell lines.[2] In drosophila, the enhancer trap approach was used successfully in a large-scale screening for genes expressed at particular developmental stages or in particular lineages.[3] The *lacZ* reporter was used to provide a sensitive way to detect expression in whole embryos. Similarly, enhancer trap vectors were found to exhibit unique temporal and spatial patterns of *lacZ* expression in transgenic mice.[4,5] Thus, this method was initially used for finding unknown genes in the genome, but not for the production of a mutation. However, an enhancer is sometimes located far from a coding region in the eukaryotic genome. Thus, it is difficult to search for the gene, because isolation of probes and screening of the cDNA library are required to identify the gene.

Exon trap and promoter trap as tools for mutagenesis

To overcome problems in the enhancer trap approach, Gossler *et al.*[6] developed a new strategy in which mouse embryonic stem (ES) cells and an exon trap vector were used to screen many integration events and rapidly clone the associated genes. Although in the published reports, the term 'gene trap' is used, the more accurate term 'exon trap' is used in this manuscript. Exon trap vectors are designed to generate spliced fusion transcripts between the reporter and the endogenous gene present at the site of integration.[7,8] The essential feature of the exon trap design is the placement of a splice acceptor in front of a promoterless *lacZ* gene. Integrations of the vector into the intron of a gene in the correct orientation were predicted to create *lacZ* fusion transcripts; and if the reading frames of the endogenous genes and *lacZ* are the same, an active β-galactosidase fusion protein should be produced. An alternative method, the promoter trap, was also developed to isolate the transcriptional control element, promoter, using cultured cells.[9] Because this vector contains only the coding sequences of the reporter gene, it is expected to require insertions into exons or the 5′ untranslated region to activate reporter gene expression. These vectors can create *lacZ* fusion products with endogenous genes and, as a result, may interfere with the normal coding capacity of the endogenous gene and thereby create a mutation. Moreover, cloning a portion of the endogenous gene directly from the *lacZ* fusion transcript eliminates the time-consuming task of searching for exons in flanking genomic DNA. Eventually, the exon trap and promoter trap approaches are applied to disrupt genes expressed in ES cells.[10,11] The use of ES cells brings several advantages. First, the exon or promoter trap technology in ES cells made it possible to carry out a large scale disruption of genes in mice.[12–15] Although the ability to carry out large scale screens has proven essential in unraveling the genetic

[1]To whom correspondence should be addressed.
E-mail: yamamura@gpo.kumamoto-u.ac.jp

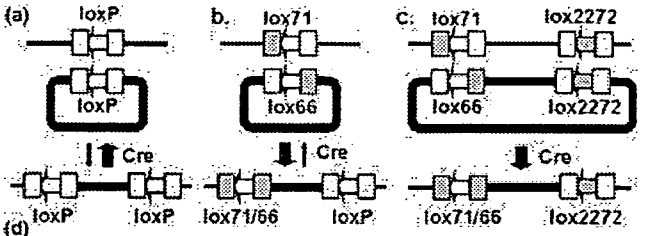
doi: 10.1111/j.1349-7006.2007.00611.x

program underlying embryogenesis in organisms such as Drosophila melanogaster, these types of screens in mammals were made difficult by the large genomic size. Furthermore, the cost and space required to house large numbers of animals and the relatively long breeding period have limited the undertaking of large scale screens. These disadvantages can be overcome by the use of ES cells. Second, it is possible to prescreen ES cells for desired insertion events and for desired expression patterns. ES cells can be induced to differentiate to many types of cells. Thus, *in vitro* preselection of gene-trapped ES clones is possible by examining the *lacZ* expression during differentiation.[16–19] Third, monitoring the *lacZ* fusion gene activity in embryos should enable one to readily visualize the pattern of endogenous gene expression during development. Thus, this method was used for a large scale screening for insertional mutation in developmentally regulated genes in mice.[12,20,21]

Development of new trap methods

We first developed a poly A trap vector.[22] As expected, the exon or promoter trap can be used for those genes expressed in ES cells. Thus, these methods can not be used for genes that are not expressed in ES cells. The constructs used in our poly A trap experiments consist of promoter-less lacZ and neomycin phosphotransferase (neo) expression unit without its own poly A addition signal sequence. We removed the poly A addition signal sequence because the neo gene was expected to be expressed only when the trap vector could use the poly A addition signal of the endogenous gene. Although we isolated six ES clones and sequenced 3′RACE (rapid amplification of cDNA end) products, there were no sequences that completely matched with the consensus sequence of the mammalian poly A addition signal. Other groups also constructed poly A trap vectors and were used to capture a broader spectrum of genes including those not expressed in ES cells.[15,23,24] However, it turned out that poly A trapping inevitably selects for the vector integration into the last introns of the trapped genes, resulting in the deletion of only a limited C-terminal portion of the protein encoded by the last exon of the trapped gene. Shigeoka *et al.*[25] demonstrated that this remarkable skewing is caused by the degradation of a selectable-marker used for poly A trapping via an mRNA-surveillance mechanism, nonsense-mediated mRNA decay (NMD). The NMD pathway is universally conserved among eukaryotes and is responsible for the degradation of mRNAs with potentially harmful nonsense mutations.[26] They also show that an internal ribosome entry site (IRES) sequence inserted downstream of the authentic termination codon (TC) of the selectable-marker mRNA prevents the molecule from undergoing NMD, and makes it possible to trap transcriptionally silent genes without a bias in the vector-integration site. Thus, this novel anti-NMD technology, termed UPATrap, could be used as one of the powerful and straightforward strategies for the unbiased inactivation of all mouse genes in the genome of ES cells.

On the other hand, there was a growing demand for production of a multipurpose allele.[27,28] As a gene has many functions, it is not enough to make one strain of null mutant mouse to examine gene function or to produce a mouse model for human disease. Actually, the gene trap vectors that have been used to generate the currently available resources induce only one type of mutation, such as the null mutation: mouse mutants generated from these libraries can show only the earliest and non-redundant developmental function of the trapped gene. Therefore, for most of the mutant strains, the significance of the trapped gene for human disease remains uncertain, because most human disorders result from late-onset gene dysfunction. In addition, between 20% and 30% of the genes targeted in ES cells are required for development and cause embryonic lethal phenotypes when transferred to the germ line, precluding functional analysis in adults.[29]



(d)

	LE	spacer	RE
loxP	ATAACTTCGTATA	GCATACAT	TATACGAAGTTAT
lox71	TACCGTTCGTATA	GCATACAT	TATACGAAGTTAT
lox66	ATAACTTCGTATA	GCATACAT	TATACGAACGGTA
lox551	ATAACTTCGTATA	GTATACAT	TATACGAAGTTAT
lox2272	ATAACTTCGTATA	GCATACTT	TATACGAAGTTAT

Fig. 1. Cre-loxP recombination system. (a) The integration reaction is inefficient with wild-type loxP sites due to the re-excision of the recombined product. (b) Integration reaction through left element/right element (LE/RE) mutant lox. (c) Integration reaction through double mutant lox. (d) Nucleotide sequences of various lox.

To solve these problems, we have developed an exchangeable gene trap vector based on a strategy for directional site-specific recombination by the Cre-*lox*P system. The Cre-*lox*P recombination system of bacteriophage P1 is currently the most powerful tool for genetic manipulation both *in vitro*[30,31] and *in vivo*.[32–36] Cre recombinase catalyzes reciprocal site-specific recombination between two *lox*P sites (Fig. 1). Consequently, Cre mediates both excisive and integrative recombination (Fig. 1a). In integrative recombination, a circular DNA carrying a *lox*P site is inserted into a *lox*P site on a chromosome. However, this integration reaction is quite inefficient, because the integrated DNA, which has *lox*P sites at both ends, is easily removed again though excisive recombination if the Cre recombinase is still present (Fig. 1a). Therefore, a special selection system in which only targeted integrants can survive is indispensable for targeted integration into *lox*P sites. Albert *et al.*[36] devised a new strategy. They identified three sets of mutant *lox* sites that favor integrative recombination over the excisive reaction. The *lox*P site is composed of an asymmetric 8 bp spacer flanked by 13 bp inverted repeats (Fig. 1d). They introduced nucleotide changes into the 13 bp left element (LE mutant *lox* site, *lox*71) or the 13 bp right element (RE mutant *lox* site, *lox*66) (Fig. 1d). Recombination between the LE mutant *lox* site and the RE mutant *lox* site produces the wild-type *lox*P site on the right and an LE + RE mutant site on the left that is poorly recognized by Cre, resulting in stable integration (Fig. 1b). Using a pair of mutant lox sites, lox71 and a lox66, we achieved site-directed DNA integration in mouse ES cells.[27] We showed that the frequency of site-specific integration via the mutant lox sites reached a maximum of 16%. In contrast, the wild-type *lox*P sites yielded very low frequencies (<0.5%) of site-specific integration events.

By applying a pair of mutant *lox* sites to gene trapping, we have developed the new gene trap vector, termed pU-Hachi.[37] The pU-Hachi consists of the SA-*lox*71-IRES-β*geo*-polyadenylation signal (pA)-*lox*P-pA-pUC. We used an IRES–βgeo unit as a selection-reporter marker, because it can trap genes whose expression levels in ES cells are very low.[13] Since the long IRES–βgeo unit interposed between the *lox*71 site and the *lox*P site can be easily removed by Cre recombinase[27], we can carry out plasmid rescue to recover the 5′-flanking genomic region as well as the 3′-flanking region. Thus, we can isolate and identify the flanking DNA sequences easily. As the *lox*71 site serves as a target for Cre-mediated insertion, we can insert a cDNA sequence joined to IRES to express it under the control of a

doi: 10.1111/j.1349-7006.2007.00611.x

trapped promoter. The poly (A) signal downstream of the *lox*P plays a role in concentrating the targeted integration event through the poly (A) trap strategy. Since the selection marker gene in a targeting vector does not have a poly (A) signal, only upon targeted integration does the selection marker gene fuse to the poly (A) signal on the trap vector, thereby making the cells drug-resistant.

Although the pU8 vector works quite well, the truncated protein may be produced when the vector is integrated downstream of the endogenous gene. This truncated protein may exert unexpected effects on phenotypes. In addition, recombination may occur between *lox*71 and *lox*P, because both carry the same spacer sequences, resulting in the deletion of a gene of interest. To overcome these problems, we improved and created the new trap vector, pU17. First, we try to produce a promoter trap vector, so that we can make a null allele in the first step. For this purpose, we inserted the three stop codons in splice acceptor, which is located upstream of the ATG of the reporter gene, β-geo, and did not use the IRES sequence. If this vector was inserted into the downstream of endogenous gene, translation would stop at the stop codon. As the β-geo gene does not have IRES sequence, it would not be translated independently. Thus, ES clones will not become drug-resistant and such clones will not be isolated. Second, we took double *lox* strategy to reduce the possibility of recombination as mentioned above. In the double *lox* strategy[38] or recombinase mediated cassette exchange[39,40] a different type of mutant lox site was used. We used heterospecific lox sites that have mutation(s) in the 8 bp spacer region. The principle is that recombination does not occur between two lox sites differing in the spacer region, whereas lox sites having the identical spacer region recombine efficiently.[41] Several groups have used lox511 (Fig. 1d), which contains a single base substitution, and demonstrated successful gene replacement in which a genomic DNA segment flanked by lox511 and loxP was replaced with another cassette flanked by lox511 and loxP located on a transfected plasmid vector.[39,42] Lee and Saito[43] developed heterospecific mutant lox sites with two base substitutions, such as lox2272 (Fig. 1d), and showed that they never recombined with the wild-type loxP site, while lox511 can recombine with loxP at low frequency using an *in vitro* system. Successful selection marker-free replacement using lox2272 and loxP was demonstrated in ES cells by Kolb[44] (Fig. 1c). So, we examined the best combination of mutant lox sites to give a high recombination efficiency and stability. We found that the frequency of site-specific integration was highest and reached a maximum of 35% when we used the combination of the LE/RE mutant and lox2272.[45] Based on these findings, we have constructed the pU17 suitable for production of a null trap allele in the first step, production of the replaced allele in the second step, and the conditionally inactivated allele in the third step by mating recombinase expressing transgenic mice[46] (Fig. 2). This is termed as the exchangeable gene trap method. The new vector, pU-17, carries the intron-*lox*71-splicing acceptor (SA)-β*geo*-*lox*P-pA-*lox*2272-pSP73-*lox*511 (Fig. 2). The SA contains three stop codons in-frame with the ATG of β-*geo* that can function as promoter trappings. We found that the trap vector was highly selective for integrations near the exon containing the start codon (Fig. 3), leading to the null mutation of trapped endogenous gene as expected. Thus, using this pU-17 trap vector, we can initially carry out random mutagenesis. Among 432 trap clones, the trapped DNAs comprise 53% (229) known coding region, 13.4% (58 clones) new or unknown genes, 12% (52 clones) known coding region but in reverse orientation, 8.8% (38 clones) no known coding region, and 12.7% (55 clones) undetermined (TransGenic Inc. Kumamoto, Japan, unpublished data, 2006) (Fig. 4). It is of interest that about 12% of clones show the insertion of a trap vector in a reverse orientation, suggesting that non-coding RNA genes may exist in these

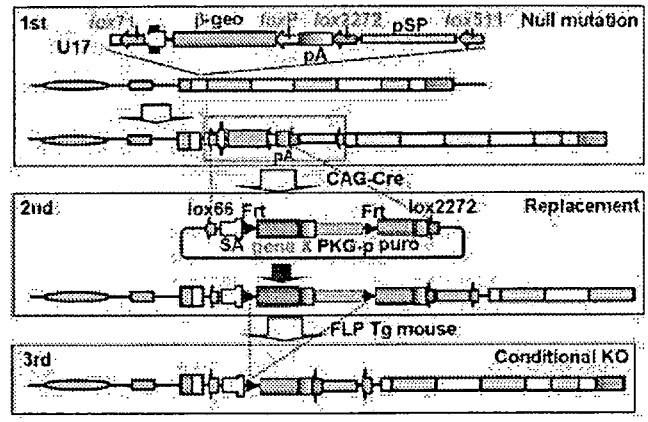


Fig. 2. Exchangeable gene trap method. Production of the null trap allele in the first step, the replaced allele in the second, and the conditionally inactivated allele in the third step is shown.

regions. Furthermore, we successfully integrated the *cre* gene into the mutant lox sites by Cre-mediated recombination. The inserted *cre* gene was stably transmitted through the germline. This cre knock-in system using the LE/RE-*lox*2272 combination should be useful for the production of various cre mice for gene targeting or gene trapping.

Exchangeable gene trap a tool for cancer research

Using the exchangeable gene trap method, we established more than 300 ES trap mouse lines and these lines are cryopreserved. The database of these lines is publicly available at the Database for the Exchangeable Gene Trap Clones (EGTC) website (http://egtc.jp/view/index). Among these, we analyzed six lines in detail, termed as Ayu3-112, Ayu8-021, Ayu8-025, Ayu8-104, Ayu8-108, and Ayu21-127, to examine whether these genes had any relation with cancer. In these lines, the integration of a trap vector resulted in the disruption of *Crebbp* (cyclic AMP response element-binding protein) gene[47,48] *Skt* (Sickle) gene[49] *Abhd2* (α/β hydrolase domain containing 2) gene[50] *c-crk* gene[51] *Impβ* (*Importin β*)/*Kpnb1* (*karyopherin β1*) gene[52] and *Lgr4* (leucine-rich repeat domain containing G protein-coupled receptor 4) gene,[53] respectively. These strains are termed as $Crebbp^{GtAyu3-112Imeg}$, $Skt^{GtAyu8-021Imeg}$, $Abhd2^{GtAyu8-025Imeg}$, $c\text{-}crk^{GtAyu8-104Imeg}$, $Impb^{GtAyu8-108Imeg}$, and $Lgr4^{GtAyu21-127Imeg}$, respectively. In the present review we will focus on three strains: $Skt^{GtAyu8-021Imeg}$, $Abhd2^{GtAyu8-025Imeg}$, and $Lgr4^{GtAyu21-127Imeg}$. These genes were initially thought to be unrelated to cancer and therefore, no-one had chosen these genes as targets for cancer research. This is because most cancer studies have been carried out using a hypothesis-driven approach. In this approach, researchers need information that suggests a relationship exists between the gene and cancer to commence work. On the contrary, the gene trap approach is a Baconian science, like natural history. This is also called an 'ignorance-driven' approach.[54] The seventeenth-century philosopher Francis Bacon suggested a system for understanding the world that begun with the accumulation of facts, based on observation. This kind of approach does not include bald theories or sudden leaps of understanding. It does not attract the same level of recognition as in hypothesis-driven types of research. Thus, many people are doubtful about this type of approach in science. However, through the gene trap approach, we can find the genes that are paid little attention in cancer research, but are actually related to cancer. Next we will briefly introduce these three lines and explain how they are related to cancer.

We established the mutant mouse line, B6; CB-$Skt^{GtAyu8021IMEG}$ (Skt^{Gt}), through gene-trap mutagenesis in embryonic stem cells.[49]

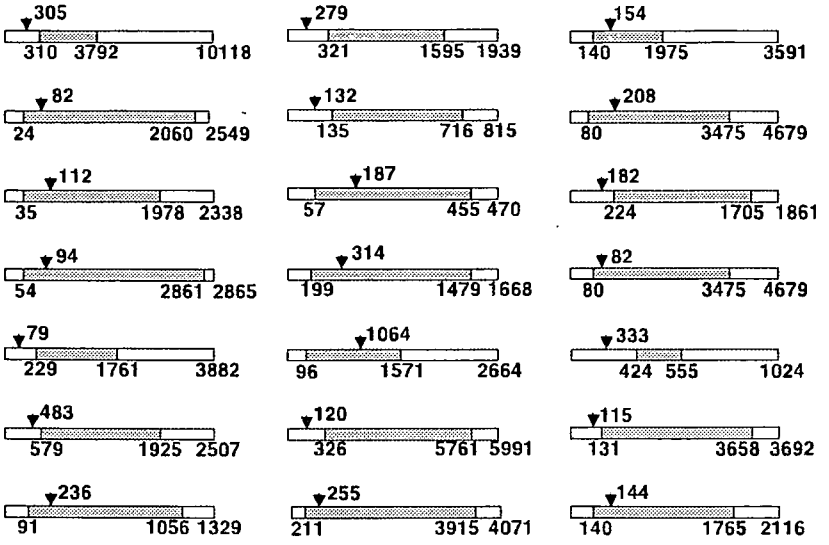


Fig. 3. Insertion sites in known genes using the pU17 vector. Black bars indicate coding region and blank boxes indicate 5′ or 3′ untranslated regions. Arrows indicate the nucleotide numbers of insertion sites. Nucleotide numbers of translation initiation site, translation termination site, and transcription termination site are shown below the genes.

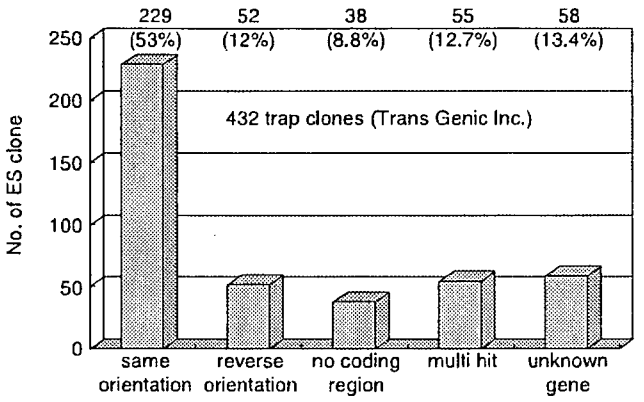


Fig. 4. Characterization of trapped DNA. In 12% of clones, the trap vector is inserted into the known coding region but in reverse orientation. In 8.8% of clones there is no known coding region around the insertion site.

The novel gene identified, called *Sickle tail* (*Skt*), is composed of 19 exons and encodes a protein of 1352 amino acids with a proline-rich region and a coiled-coil domain. Expression of a reporter gene (*β-geo*) was detected in the notochord during embryogenesis and in its derivative, the nucleus pulposus, in adult mice. Compression of some of the nuclei pulposi in the center of intervertebral discs appeared at embryonic day (E) 17.5, resulting in a kinky-tail phenotype in $Skt^{Gt(Gt}$ adult mice. Chordoma is a malignant tumor, but is characterized by slow growth, with local destruction of the bone and extension into the adjacent soft tissue. Very rarely, distant metastases are encountered. Chordomas are thought to arise from primitive notochordal remnants along the axial skeleton. The distribution of tumors matches the distribution of notochordal remnants. The Skt could be a marker for differential diagnosis of chordoma and chondrosarcoma. This differential diagnosis may be important to decide the extent of surgical dissection.

We also isolated a clone, Ayu8025, in which the trap vector was inserted into the fifth intron of α/β hydrolase domain containing 2 (Abhd2).[50] The α/β hydrolase fold family consists of hydrolytic enzymes of widely differing phylogenetic origin. Each of these family members has the same protein fold, termed as the α/β hydrolase fold.[55] The core of each enzyme is similar: an α/β-sheet of five to eight β-sheets connected by α-helices to form an α/β/α sandwich. They all have a nucleophile–histidine–acid catalytic triad, which is the best-conserved structural feature in the fold. In mice, three cDNA encoding proteins containing an α/β hydrolase fold were cloned from lung cDNA.[56] These are now termed as abhydrolase domain containing (*Abhd*) 1, 2 or 3. Reverse transcription-polymerase chain reaction (RT-PCR) analyses showed that these three genes are widely, but differentially expressed in many tissues. The expression of *Abhd1* and *Abhd3* was highest in the liver and lowest in the spleen, whereas the expression of *Abhd2* was high in the testis and the spleen. All three Abhd proteins are shown to have a single predicted amino-terminus transmembrane domain. Although proteins in this family generally act as enzymes such as acetylcholinesterases, or lipases, the specific functions of the three Abhd proteins are unknown. To obtain some clue for Abhd2 function, we first analyzed the expression pattern during embryonic development using X-gal staining. Interestingly, *Abhd2* was expressed in the vitelline vessels of yolk sacs at embryonic day (E) 12.5 and expression of Abhd2 switched from endothelial cells to vascular smooth muscle cells (SMCs) during further embryonic development. In adult, Abhd2 was expressed in many tissues including vascular and non-vascular SMCs such as intestinal SMCs. Although homozygous mutant mice were apparently normal, enhanced SMC migration in the explants SMCs culture and marked intimal hyperplasia after cuff placement were observed in homozygous mice. Our results show that *Abhd2* is involved in SMC migration and neo-intimal thickening on vascular SMCs. When we started to analyze this gene in 2002, there was no report to suggest a relationship between this gene and cancer. Using microarray analysis, Johansson *et al.*[57,58] reported that the expression of Abhd2 was enhanced in a mouse brain tumor model induced by intracerebral injection of a recombinant Moloney murine leukemia virus encoding the platelet-derived growth factor B-chain (MMLV/PDGFB) into newborn mice. In addition, using microarray analysis again, Chen *et al.*[59] reported that ABHD2 expression was increased in MCF-7 cells transfected with human Sp5, a member of the Sp transcription factor family, suggesting that ABHD2 is a downstream target of Sp5, and that ABHD2 is associated with tumorigenesis. At the moment, the role of ABHD2 in tumorigenesis is not known yet.

Another gene trap line is $Lgr4^{Gt(Ayu21\text{-}127)Imeg}$ ($Lgr4^{Gt}$) in which the trap vector was integrated into the 2nd intron of the *Lgr4* gene.[53] Northern blot analysis and quantitative RT-PCR showed

doi: 10.1111/j.1349-7006.2007.00611.x

that the *Lgr4*$^{Gt/Gt}$ mice had 10% mRNA expression of the *Lgr4* gene, suggesting that the *Lgr4*$^{Gt/Gt}$ mouse is a hypomorphic mutant. Leucine-rich repeat domain containing G-protein coupled receptor (GPCR) 4 (LGR4) family members are characterized by an extracellular domain with multiple leucine-rich repeats (LRRs).[60,61] The presence of a large extra-cellular domain is a remarkable feature that separates the LGR family members from the other GPCRs. Studies of LGRs from different species suggest that LGRs can be classified into three subtypes (A, B and C). Type A LGRs include the follicle-stimulating hormone receptor (FSHR), the lutenizing hormone receptor (LHR), and the thyroid-stimulating hormone receptor (TSHR), in which the ligands are glycoprotein hormones.[62] Type B LGR comprises three members: LGR4, also known as GPR48, LGR5, and LGR6. Type B LGR remains an orphan GPCR and its physiological functions have not yet been determined. Type C LGRs, including LGR7 and LGR8, were recently described as relaxin receptors.[63] Following the identification of LGR7 and LGR8 as relaxin receptors, the closely related relaxin3 and INSL3 have been shown to function as selective agonists for LGR7 and LGR8, respectively. The homozygous male was infertile showing morphological abnormalities in both the testes and the epididymides. In the testes, luminal swelling, loss of germinal epithelium in the seminiferous tubules, and rete testis dilation were observed. Rete testis dilation was due to a water reabsorption failure caused by a decreased expression of ESR1 and SLC9A3 in the efferent ducts. The epididymis contained short and dilated tubules and completely lacked its initial segment. Interestingly, we observed multilamination and distortion of the basement membranes with an accumulation of laminin in caput epididymidis. These results indicate that *Lgr4* has pivotal roles for the control of duct elongation through basement membrane remodeling, and the regional differentiation of the caput epididymidis. On the other hand, Gao *et al.*[64] reported that LGR4 expression is upregulated in p27-deficient HCT116 cells using microarray analysis. Forced expression of LGR4 increased both *in vitro* invasive activity and lung metastasis potency of HCT116 cells. In contrast, the depletion of endogenous LGR4 by RNA interference reduced the invasive potential of HeLa and Lewis lung carcinoma cells not only *in vitro* but also *in vivo*. Moreover, GPR48 expression was significantly associated with lymph node metastasis and inversely correlated with p27 expression in human colon carcinomas. As a reduced expression level of the cyclin-dependent kinase inhibitor p27Kip1 is associated with increased tumor malignancy and poor prognosis in individuals with various types of cancer, LGR4 may play an important role in invasiveness and metastasis of carcinoma and might therefore represent a potential prognostic marker or therapeutic target. This is consistent with our hypothesis that Lgr4 is involved in the remodeling of BM.

Conclusion

In summary, we show here that exchangeable gene trap is a powerful tool for finding genes that are related to cancer, but are paid little attention in cancer research. By combining microarray analysis and gene trap mutagenesis, we will be able to analyze function of such genes in more systematic ways. In addition, we show that we can replace the reporter gene with a gene of interest using the Cre-loxP system. We can apply this system to the vectors for homologous recombination, thus we can make the multipurpose allele at any gene locus. These methods can be applied for detailed analysis of gene function in cancer research.

Acknowledgments

This work was supported in part by a KAKENHI (Grant-in-Aid for Scientific Research) in Priority Areas 'Integrative Research Toward the Conquest of Cancer' from the Ministry of Education, Culture, Sports, Science and Technology of Japan, and a grant from the Osaka Foundation of Promotion of Clinical Immunology and a grant from Pancreas Research Foundation of Japan. We thank TransGenic Incorporation for providing unpublished data on characterization of trapped DNA and the members of our group for their support and continuous discussion.

References

1 Casadaban MJ, Cohen SN. Lactose genes fused to exogenous promoters in one step using a Mu-lac bacteriophage: *in vivo* probe for transcriptional control sequences. *Proc Natl Acad Sci USA* 1979; **76**: 4530–3.
2 Fried M, Griffiths M, Davies B, Bjursell G, La Mantia G, Lania L. Isolation of cellular DNA sequences that allow expression of adjacent genes. *Proc Natl Acad Sci USA* 1983; **80**: 2117–21.
3 O'Kane CJ, Gehring WJ. Detection in situ of genomic regulatory elements in Drosophila. *Proc Natl Acad Sci USA* 1987; **84**: 9123–7.
4 Allen ND, Cran DG, Barton SC, Hettle S, Reik W, Surani MA. Transgenes as probes for active chromosomal domains in mouse development. *Nature* 1988; **333**: 852–5.
5 Kothary R, Clapoff S, Brown A, Campbell R, Peterson A, Rossant J. A transgene containing lacZ inserted into the dystonia locus is expressed in neural tube. *Nature* 1988; **335**: 435–7.
6 Gossler A, Joyner AL, Rossant J, Skarnes WC. Mouse embryonic stem cells and reporter constructs to detect developmentally regulated genes. *Science* 1989; **244**: 463–5.
7 Brenner DG, Lin-Chao S, Cohen SN. Analysis of mammalian cell genetic regulation in situ by using retrovirus-derived 'portable exons' carrying the Escherichia coli lacZ gene. *Proc Natl Acad Sci USA* 1989; **86**: 5517–21.
8 von Melchner H, Ruley HE. Identification of cellular promoters by using a retrovirus promoter trap. *J Virol* 1989; **63**: 3227–33.
9 von Melchner H, Reddy S, Ruley HE. Isolation of cellular promoters by using a retrovirus promoter trap. *Proc Natl Acad Sci USA* 1990; **87**: 3733–7.
10 Macleod D, Lovell-Badge R, Jones S, Jackson I. A promoter trap in embryonic stem (ES) cells selects for integration of DNA into CpG islands. *Nucleic Acids Res* 1991; **19**: 17–23.
11 Friedrich G, Soriano P. Promoter traps in embryonic stem cells: a genetic screen to identify and mutate developmental genes in mice. *Genes Dev* 1991; **5**: 1513–23.
12 Wurst W, Rossant J, Prideaux V *et al.* A large-scale gene-trap screen for insertional mutations in developmentally regulated genes in mice. *Genetics* 1995; **139**: 889–99.
13 Chowdhury K, Bonaldo P, Torres M, Stoykova A, Gruss P. Evidence for the stochastic integration of gene trap vectors into the mouse germline. *Nucleic Acids Res* 1997; **25**: 1531–6.
14 Hicks GG, Shi EG, Li XM, Li CH, Pawlak M, Ruley HE. Functional genomics in mice by tagged sequence mutagenesis. *Nat Genet* 1997; **16**: 338–44.
15 Zambrowicz BP, Friedrich GA, Buxton EC, Lilleberg SL, Person C, Sands AT. Disruption and sequence identification of 2000 genes in mouse embryonic stem cells. *Nature* 1998; **392**: 608–11.
16 Forrester LM, Nagy A, Sam M *et al.* An induction gene trap screen in embryonic stem cells: Identification of genes that respond to retinoic acid *in vitro*. *Proc Natl Acad Sci USA* 1996; **93**: 1677–82.
17 Stanford WL, Caruana G, Vallis KA *et al.* Expression trapping: identification of novel genes expressed in hematopoietic and endothelial lineages by gene trapping in ES cells. *Blood* 1998; **92**: 4622–31.
18 Tate P, Lee M, Tweedie S, Skarnes WC, Bickmore WA. Capturing novel mouse genes encoding chromosomal and other nuclear proteins. *J Cell Sci* 1998; **111**: 2575–85.
19 Kluppel M, Vallis KA, Wrana JL. A high-throughput induction gene trap approach defines C4ST as a target of BMP signaling. *Mech Dev* 2002; **118**: 77–89.
20 Korn R, Schoor M, Neuhaus H *et al.* Enhancer trap integrations in mouse embryonic stem cells give rise to staining patterns in chimaeric embryos with a high frequency and detect endogenous genes. *Mech Dev* 1992; **39**: 95–109.
21 Soininen R, Schoor M, Henseling U *et al.* The mouse Enhancer trap locus 1 (Etl-1): a novel mammalian gene related to Drosophila and yeast transcriptional regulator genes. *Mech Dev* 1992; **39**: 111–23.
22 Niwa H, Araki K, Kimura S, Taniguchi S, Wakasugi S, Yamamura K. An efficient gene-trap method using poly A trap vectors and characterization of gene-trap events. *J Biochem (Tokyo)* 1993; **113**: 343–9.
23 Salminen M, Meyer BI, Gruss P. Efficient poly A trap approach allows the

capture of genes specifically active in differentiated embryonic stem cells and in mouse embryos. *Dev Dyn* 1998; **212**: 326–33.
24 Ishida Y, Leder PRET. a poly A-trap retrovirus vector for reversible disruption and expression monitoring of genes in living cells. *Nucleic Acids Res* 1999; **27**: e35.
25 Shigeoka T, Kawaichi M, Ishida Y. Suppression of nonsense-mediated mRNA decay permits unbiased gene trapping in mouse embryonic stem cells. *Nucleic Acids Res* 2005; **33**: e20.
26 Hentze MW, Kulozik AE. A perfect message: RNA surveillance and nonsense-mediated decay. *Cell* 1999; **96**: 307–10.
27 Araki K, Imaizumi T, Okuyama K, Oike Y, Yamamura K. Efficiency of recombination by Cre transient expression in embryonic stem cells: comparison of various promoters. *J Biochem (Tokyo)* 1997; **122**: 977–82.
28 Testa G, Schaft J, van der Hoeven F *et al.* A reliable lacZ expression reporter cassette for multipurpose, knockout-first alleles. *Genesis* 2004; **38**: 151–8.
29 Mitchell KJ, Pinson KI, Kelly OG *et al.* Functional analysis of secreted and transmembrane proteins critical to mouse development. *Nat Genet* 2001; **28**: 241–9.
30 Sauer B, Henderson N. Site-specific DNA recombination in mammalian cells by the Cre recombinase of bacteriophage P1. *Proc Natl Acad Sci USA* 1988; **85**: 5166–70.
31 Sauer B, Henderson N. Targeted insertion of exogenous DNA into the eukaryotic genome by the Cre recombinase. *New Biol* 1990; **2**: 441–9.
32 Lakso M, Sauer B, Mosinger B Jr *et al.* Targeted oncogene activation by site-specific recombination in transgenic mice. *Proc Natl Acad Sci USA* 1992; **89**: 6232–6.
33 Orban PC, Chui D, Marth JD. Tissue- and site-specific DNA recombination in transgenic mice. *Proc Natl Acad Sci USA* 1992; **89**: 6861–5.
34 Araki K, Araki M, Miyazaki J, Vassalli P. Site-specific recombination of a transgene in fertilized eggs by transient expression of Cre recombinase. *Proc Natl Acad Sci USA* 1995; **92**: 160–4.
35 Kuhn R, Schwenk F, Aguet M, Rajewsky K. Inducible gene targeting in mice. *Science* 1995; **269**: 1427–9.
36 Albert H, Dale EC, Lee E, Ow DW. Site-specific integration of DNA into wild-type and mutant lox sites placed in the plant genome. *Plant J* 1995; **7**: 649–59.
37 Araki K, Imaizumi T, Sekimoto T *et al.* Exchangeable gene trap using the Cre/mutated lox system. *Cell Mol Biol (Noisy-le-Grand)* 1999; **45**: 737–50.
38 Soukharev S, Miller JL, Sauer B. Segmental genomic replacement in embryonic stem cells by double lox targeting. *Nucleic Acids Res* 1999; **27**: e21.
39 Bouhassira EE, Westerman K, Leboulch P. Transcriptional behavior of LCR enhancer elements integrated at the same chromosomal locus by recombinase-mediated cassette exchange. *Blood* 1997; **90**: 3332–44.
40 Feng YQ, Seibler J, Alami R *et al.* Site-specific chromosomal integration in mammalian cells: highly efficient CRE recombinase-mediated cassette exchange. *J Mol Biol* 1999; **292**: 779–85.
41 Hoess RH, Wierzbicki A, Abremski K. The role of the loxP spacer region in P1 site-specific recombination. *Nucleic Acids Res* 1986; **14**: 2287–300.
42 Bethke B, Sauer B. Segmental genomic replacement by Cre-mediated recombination: genotoxic stress activation of the p53 promoter in single-copy transformants. *Nucleic Acids Res* 1997; **25**: 2828–34.
43 Lee G, Saito I. Role of nucleotide sequences of loxP spacer region in Cre-mediated recombination. *Gene* 1998; **216**: 55–65.
44 Kolb AF. Selection-marker-free modification of the murine beta-casein gene using a lox2272 (correction of lox2722) site. *Anal Biochem* 2001; **290**: 260–71.
45 Araki K, Araki M, Yamamura K. Site-directed integration of the cre gene mediated by Cre recombinase using a combination of mutant lox sites. *Nucleic Acids Res* 2002; **30**: e103.
46 Taniwaki T, Haruna K, Nakamura H *et al.* Characterization of an exchangeable gene trap using pU-17 carrying a stop codon-beta geo cassette. *Dev Growth Differ* 2005; **47**: 163–72.
47 Oike Y, Hata A, Mamiya T *et al.* Truncated CBP protein leads to classical Rubinstein–Taybi syndrome phenotypes in mice: implications for a dominant-negative mechanism. *Hum Mol Genet* 1999; **8**: 387–96.
48 Oike Y, Takakura N, Hata A *et al.* Mice homozygous for a truncated form of CREB-binding protein exhibit defects in hematopoiesis and vasculo-angiogenesis. *Blood* 1999; **93**: 2771–9.
49 Semba K, Araki K, Li Z *et al.* A novel murine gene, Sickle tail, linked to the Danforth's short tail locus, is required for normal development of the intervertebral disc. *Genetics* 2006; **172**: 445–56.
50 Miyata K, Oike Y, Hoshii T *et al.* Increase of smooth muscle cell migration and of intimal hyperplasia in mice lacking the alpha/beta hydrolase domain containing 2 gene. *Biochem Biophys Res Commun* 2005; **329**: 296–304.
51 Imaizumi T, Araki K, Miura K *et al.* Mutant mice lacking Crk-II caused by the gene trap insertional mutagenesis: Crk-II is not essential for embryonic development. *Biochem Biophys Res Commun* 1999; **266**: 569–74.
52 Miura K, Yoshinobu K, Imaizumi T *et al.* Impaired expression of importin/karyopherin beta1 leads to post-implantation lethality. *Biochem Biophys Res Commun* 2006; **341**: 132–8.
53 Hoshii T, Takeo T, Nakagata N, Takeya M, Araki K, Yamamura K. LGR4 Regulates the Postnatal Development and Integrity of Male Reproductive Tracts in Mice. *Biol Reprod* 2007; **76**: 303–13.
54 Sulston J, Ferry G. The common thread – a story of science, politics, ethics, and human genome. Washington, DC, USA: The Joseph Henry Press, 2003.
55 Ollis DL, Cheah E, Cygler M *et al.* The alpha/beta hydrolase fold. *Protein Eng* 1992; **5**: 197–211.
56 Edgar AJ, Polak JM. Cloning and tissue distribution of three murine alpha/beta hydrolase fold protein cDNAs. *Biochem Biophys Res Commun* 2002; **292**: 617–25.
57 Johansson FK, Brodd J, Eklof C *et al.* Identification of candidate cancer-causing genes in mouse brain tumors by retroviral tagging. *Proc Natl Acad Sci USA* 2004; **101**: 11 334–7.
58 Johansson FK, Goransson H, Westermark B. Expression analysis of genes involved in brain tumor progression driven by retroviral insertional mutagenesis in mice. *Oncogene* 2005; **24**: 3896–905.
59 Chen Y, Guo Y, Ge X *et al.* Elevated expression and potential roles of human Sp5, a member of Sp transcription factor family, in human cancers. *Biochem Biophys Res Commun* 2006; **340**: 758–66.
60 Hsu SY, Liang SG, Hsueh AJ. Characterization of two LGR genes homologous to gonadotropin and thyrotropin receptors with extracellular leucine-rich repeats and a G protein-coupled, seven-transmembrane region. *Mol Endocrinol* 1998; **12**: 1830–45.
61 Loh ED, Broussard SR, Kolakowski LF. Molecular characterization of a novel glycoprotein hormone G-protein-coupled receptor. *Biochem Biophys Res Commun* 2001; **282**: 757–64.
62 Vassart G, Pardo L, Costagliola S. A molecular dissection of the glycoprotein hormone receptors. *Trends Biochem Sci* 2004; **29**: 119–26.
63 Hsu SY, Nakabayashi K, Nishi S *et al.* Activation of orphan receptors by the hormone relaxin. *Science* 2002; **295**: 671–4.
64 Gao Y, Kitagawa K, Hiramatsu Y *et al.* Up-regulation of GPR48 induced by down-regulation of p27Kip1 enhances carcinoma cell invasiveness and metastasis. *Cancer Res* 2006; **66**: 11 623–31.

doi: 10.1111/j.1349-7006.2007.00611.x

Sakae Tanaka
Kozo Nakamura
Naoyuki Takahasi
Tatsuo Suda

Role of RANKL in physiological and pathological bone resorption and therapeutics targeting the RANKL–RANK signaling system

Authors' addresses
Sakae Tanaka[1], Kozo Nakamura[1], Naoyuki Takahasi[2], Tatsuo Suda[3]
[1]Department of Orthopaedic Surgery, Faculty of Medicine, The University of Tokyo, Tokyo, Japan.
[2]Institute for Oral Science, Matsumoto Dental University, Nagano, Japan.
[3]Saitama Medical School Research Center for Genomic Medicine, Saitama, Japan.

Correspondence to:
Sakae Tanaka
Department of Orthopaedic Surgery
Faculty of Medicine
The University of Tokyo
7-3-1 Hongo
Bunkyo-ku
Tokyo 113-0033, Japan
Tel.: +81 3 3815 5411x33376
Fax: +81 3 3818 4082
E-mail: tanakas-ort@h.u-tokyo.ac.jp

Immunological Reviews 2005
Vol. 208: 30–49

Immunological Reviews
0105-2896

Summary: Osteoclasts are primary cells for physiological and pathological bone resorption, and receptor activator of nuclear factor-κB ligand (RANKL) is critically involved in the differentiation, activation, and survival of these cells. Recently, therapeutics for pathological bone destruction targeting RANKL pathways has attracted a great deal of attention. Herein, we review the recent advances in the research on osteoclast biology and discuss the advantages and disadvantages of anti-RANKL therapies.

Introduction

The integrity of the skeletal tissues is maintained by a well-organized balance of old bone resorption by osteoclasts and new bone formation by osteoblasts (1, 2). Osteoblasts are bone-forming cells derived from multipotent mesenchymal stem cells, which also give rise to chondrocytes, muscle cells, and adipocytes, whereas osteoclasts are multinucleated giant cells differentiated for bone resorption and derived from the monocyte/macrophage lineage precursor cells (3–8). The bone remodeling cycle is a sequential process that is highly regulated. The 'activation' phase starts with the interaction between osteoblasts and osteoclast precursors, which differentiate into mature osteoclasts in the 'resorption' phase. In the 'reversal' phase, osteoclasts complete the resorption process and produce the signals that initiate bone formation directly or indirectly, and in the final 'formation' phase, mesenchymal cells differentiate into functional osteoblasts to make bone matrix. It has been estimated that about 10% of the total bone in adult humans is remodeled per year, and bone remodeling is important not only to maintain the skeletal structure and strength but also to regulate calcium homeostasis (9). In the remodeling cycle, the length of the resorption phase is very