Published online 16 October 2007 *Nucleic Acids Research, 2008, Vol. 36, Database issue* ***D741–D746***
doi:10.1093/nar/gkm825

The UniTrap resource: tools for the biologist enabling optimized use of gene trap clones

Guglielmo Roma[1], Marco Sardiello[1], Gilda Cobellis[1,2], Pedro Cruz[1], Giampiero Lago[1], Remo Sanges[1,3] and Elia Stupka[1,3,*]

[1]Telethon Institute of Genetics and Medicine (TIGEM), Via P. Castellino, 111, 80131, Napoli, [2]Dipartimento di Patologia Generale, Seconda Universita' di Napoli, Via De Crecchio 7, 80100 Napoli and [3]CBM S.c.r.l., c/o Area Sciente Park, Basovizza- SS14, Km 163,5 Trieste, 34012, Italy

Received August 8, 2007; Revised and Accepted September 20, 2007

ABSTRACT

We have developed a comprehensive resource devoted to biologists wanting to optimize the use of gene trap clones in their experiments. We have processed 300 602 such clones from both public and private projects to generate 28 199 'UniTraps', i.e. distinct collections of unambiguous insertions at the same subgenic region of annotated genes. The UniTrap resource contains data relative to 9583 trapped genes, which represent 42.3% of the mouse gene content. Among the trapped genes, 7 728 have a counterpart in humans, and 677 are known to be involved in the pathogenesis of human diseases. The aim of this analysis is to provide the wet lab researchers with a comprehensive database and curated tools for (i) identifying and comparing the clones carrying a trap into the genes of interest, (ii) evaluating the severity of the mutation to the protein function in each independent trapping event and (iii) supplying complete information to perform PCR, RT-PCR and restriction experiments to verify the clone and identify the exact point of vector insertion. To share this unique resource with the scientific community, we have designed and implemented a web interface that is freely accessible at http://unitrap.cbm.fvg.it/.

INTRODUCTION

One of the major challenges in the post-genomic era is to determine the role that every gene plays in the development and function of complex organisms, such as mammals. Due to its overall genetic similarity to humans and the availability of specific, advanced techniques for tailoring the genome, the mouse is currently the best model organism system to elucidate gene function and study human diseases (1,2). Mouse knockout phenotypes are also useful in drug discovery to study the pharmacological effects of drugs against the major protein targets of the pharmaceutical industry (3). Over the past few years several methodologies have been developed to carry out large-scale insertional mutagenesis in the mouse. Among these, gene trapping allows systematic, cost-effective generation of mutations in murine embryonic stem (ES) cells, which can be subsequently used to generate mutant mice (4).

Gene trapping is a high-throughput mutagenesis approach that generates sequence-tagged insertions in the genome of ES cells, many of which interrupt the coding sequence of a gene. This technique relies on the random integration in the genome of a DNA construct (the 'vector') that carries a splicing acceptor and/or a donor sequence and a reporter/selector gene. If the construct integrates into an intron, a trapped gene-selection marker fusion mRNA may be transcribed that allows both clone selection and identification of the trapped locus (5). Although gene trapping was developed as a mutagenesis approach, we have recently demonstrated its value also as a powerful tool for gene discovery (6).

Large-scale efforts to generate libraries of gene-trap insertions are under way worldwide (7–12). On one hand, several academic gene trap projects have generated trapped ES cell clones (~85 000), for which sequence tags have been deposited into the NCBI Genome Survey Sequences Database (dbGSS) (13). Recently, these projects have joined together by developing the International Gene Trap Consortium (IGTC) database that provides a centralized access to the free, publicly available ES cell gene trap libraries (14). On the other hand, the private biotechnology company Lexicon Genetics has developed OmniBank, the largest collection of mutant ES cells currently available (>270 000 clones) (15), distributed by the Texas Institute of Genomic Medicine (TIGM). Approximately 200 000 sequence tags of OmniBank clones have been deposited into the NCBI dbGSS but are not found in the IGTC database.

*To whom correspondence should be addressed. Tel: +39 040 375 7718; Fax: +39 040 375 7710; Email: elia.stupka@cbm.fvg.it

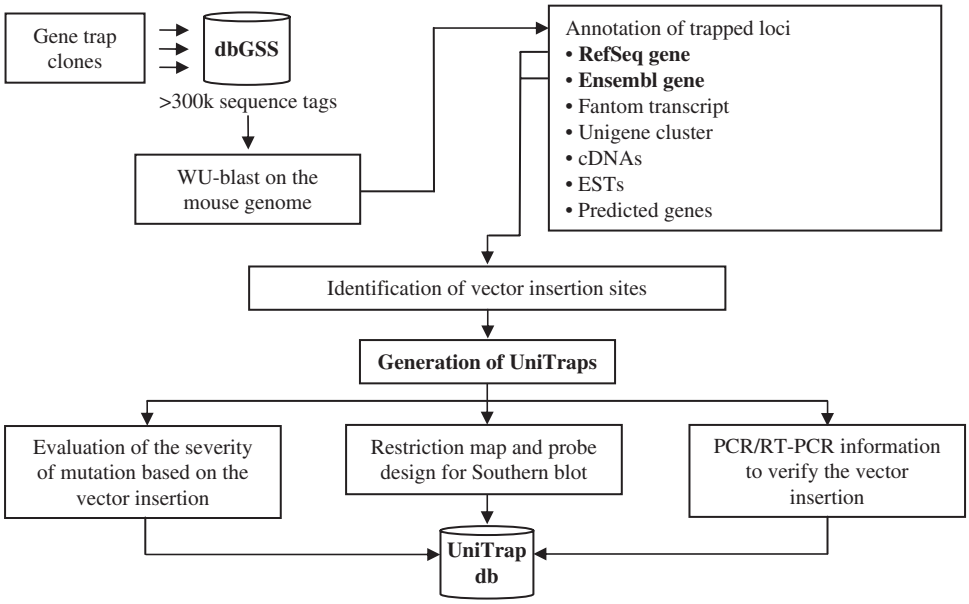

Figure 1. The UniTrap pipeline. Gene trap projects currently submit their data to dbGSS. The pipeline downloads the sequence tags and maps them to the mouse genome. Then, it checks for the annotation of the trapped region in order to identify the trapped gene and the putative vector insertion site (only RefSeq and Ensembl genes are considered). Additional annotations are checked, including Fantom transcripts, Unigene clusters, further cDNAs and ESTs, and *ab initio* predicted genes. Independent ES clones showing the same vector insertion site are grouped into 'UniTraps'. For each UniTrap, the pipeline (i) evaluates the severity of the mutation to the protein function, (ii) calculates restriction maps and probes for Southern blot analysis and (iii) designs PCR and RT-PCR experiments to verify the clone and identify the exact site of vector insertion.

The availability of such a high number of mutant ES clones, along with the strong interest toward their potential use exerted by biologists worldwide, have posed the need for a unified tool to (i) get access to exhaustive information on both public and private gene trap clones, and (ii) provide information to aid the standardized experimental design for the further characterization and validation of the vector insertions.

Here we present UniTrap, a curated collection of trapped ES cell clones generated by public and private gene trap projects. We gathered >300 000 tags from dbGSS and used an in-house developed bioinformatics pipeline to define 'UniTraps', distinct collections of ES cell clones that share vector insertions within introns of well-known genes (annotated in RefSeq and/or Ensembl). This resource aims at providing wet lab researchers with a comprehensive database and curated tools for (i) identifying and comparing the clones carrying a trap vector into the genes of interest, (ii) evaluating the effect and the severity of the mutation to the protein function in each independent trapping event and (iii) supplying complete information to perform PCR, RT-PCR and restriction experiments to verify the clone and identify the exact point of vector insertion. To share this unique resource with the scientific community, we have designed and implemented a web interface that is freely accessible at http://unitrap.cbm.fvg.it/.

UNITRAP DATA

Assignment of gene trap sequences to genes

The most common approach to identify the vector integration site is the rapid amplification of cDNA ends (RACE), which amplifies a portion of the fusion transcript between the endogenous gene and the reporter gene. Sequencing the RACE product provides a sequence tag for the identification of the trapped gene (5). Recently, however, several gene trap efforts, e.g. the German Gene Trap Consortium (10) have switched to the utilization of genomic PCR for the identification of traps, also called 'splinkerette PCR' or SPLK (16). The Unitrap project houses both types of sequences, and they can be distinguished clearly in the website.

Gene trap projects currently submit their sequence tags to the NCBI Genomic Survey Sequences Database (dbGSS), along with specific information regarding the cell lines and vectors used. In order to automate the identification and the characterization of the trapped genes, sequence tags are regularly downloaded from dbGSS and an in-house developed pipeline performs the following sequential analyses (Figure 1):

(i) Mapping of trap tags to the mouse genome. As a first step, the pipeline identifies the trapped locus of a specific ES clone by aligning its sequence tag against a repeat masked version of the mouse genome (NCBI Mouse Build 36, http://www.ncbi.nlm.nih.gov/genome/guide/mouse/) using WUBLAST (17) with an e-value cutoff of $1E{-}05$. The blast output is parsed to extract the genomic locations with BioPerl (18), using a cutoff of 96% percentage identity. For each tag, the candidate genomic sites are ranked based on the percentage of sequence identity in the alignment, the length of the aligned tag portion, and the number of aligned

exons, and the best ranking genomic site is selected. However, since some genes have multiple copies, sequence tags may align with similar scores in different genomic locations. To avoid ambiguous or erroneous mapping of trap tags, our algorithm has been optimized to distinguish the actual trapped-gene locus from recent pseudogenes (which in certain cases obtain higher alignment score due to the lack of intron gaps) or, in case of duplicated genes, to display all the possible locations of the vector insertion.

(ii) Identification of the trapped gene. Once a given trapped locus has been identified, the pipeline predicts which gene is disrupted and which exon flanks the vector insertion; we refer to this exon as the 'trapped exon'. Using a local version of the Ensembl database and the Ensembl API (19), the overlap with exons of known genes is checked for each tagged locus. For this analysis, only RefSeq genes (curated mRNAs having accession prefix NM and NR) (20) and Ensembl genes (21) are taken in consideration. In the case of overlapping genes, the trap tag is assigned based on the number of tagged exons, the length of the overlap and its ability to identify exon boundaries. Genes are subsequently annotated utilizing Ensembl for major protein features such as transmembrane domains, signal peptides, domains from PFAM, SMART and PROSITE, fingerprints from PRINTS and Superfamily classification, which are clearly displayed. The Ensembl database is used to annotate the human orthologs of trapped genes and their potential involvement in the development of monogenetic or multifactorial/polygenic diseases as reported by the On-Line Mendelian Inheritance in Man (OMIM) database (22).

(iii) Prediction of the vector insertion site. The putative vector insertion site is predicted according to the vector specifications reported in dbGSS; it may be anywhere within the intron located between the trapped exon and its 'flanking exon', which corresponds to either the upstream or the downstream exon based on the type of RACE-PCR used for the amplification of the fusion transcript. In a proportion of cases, trap tags identify novel exons within a known gene (6); in this case, if the identification of the trapped exon and its flanking exon is not predictable, the insertion site is classified as 'ambiguous'. To better manage this issue, our pipeline checks for the presence of FANTOM3 cDNAs (23), EST clusters collected in the Unigene dataset (24), further cDNAs and ESTs aligned to the mouse genome, as well as exons of *ab initio* genes predicted by Genscan (25).

Defining Unitraps: collections of gene trap clones within the same subgenic locus

Despite the fact that the number of mutated ES clones is currently one order of magnitude higher than the number of known mouse genes, it has been recently estimated that only 50% of the known genes has been trapped (6); most genes have been trapped more than once, within different introns (6).

Although different gene trap vectors have different and specific structures, the impact of their insertion on gene function depends mostly on the trapped intron. Indeed, independent ES clones carrying different vectors in the same intron will produce similarly truncated versions of the protein encoded by the trapped gene. For this reason, our pipeline groups ES cell clones into 'UniTraps'; in a given UniTrap collection, all clones share an analogous vector insertion at the sub-genic level (Figure 1).

For each UniTrap, our pipeline shows the impact of the vector insertion on the gene function by examining the number of amino acid residues deleted from the mutated protein. The severity of the mutation is clarified by showing the percentage of amino acids after which the trap insertion occurred (e.g. 'Insertion after 8% of the polypeptide chain'). We estimate that 3 983 genes (41.5% of all trapped genes) have been completely knocked out by gene trapping, because of a vector insertion upstream of all coding exons.

UniTrap currently contains 300 602 gene trap sequence tags that describe the clones generated by several public and private gene trap projects (Table 1). These sequence tags have been processed by our annotation pipeline to generate 28 199 'UniTraps' within 9 583 trapped genes (42.3% of mouse genome coverage). This represents the number of trapped genes in which at least one unambiguous vector insertion has been predicted. Among these genes, 7.728 have a counterpart in humans, including 677 genes known to be involved in the pathogenesis of human diseases.

Table 1. Gene trap projects available in UniTrap

Gene trap project	ES cell clones
Baygenomics	14 375
Centre for Modelling Human Disease (CMHD)	13 166
Embryonic Stem Cell Database	9736
Exchangeable Gene Trap Clones (EGTC)	336
Functional Genomics of Inflammation at Vanderbilt University	1665
German Gene Trap Consortium (GGTC)	35 491
Lexicon Pharmaceuticals	198 902
Nara Institute of Science and Technology (NAIST)	310
Sanger Institute Gene Trap Resource	11 886
Soriano Lab at Fred Hutchinson Cancer Research Center (FHCRC)	1627
Telethon Institute of Genetics and Medicine (TIGEM)	1343
Texas Institute of Genomic Medicine (TIGM)	11 765
Total	300 602

Taking gene trap clones from the web to the bench

Once an ES clone has been acquired to generate a mutant mouse, it is necessary to proceed to the preliminary characterization of the trapped gene (Figure 1). This is aimed at establishing the exact insertion site of the

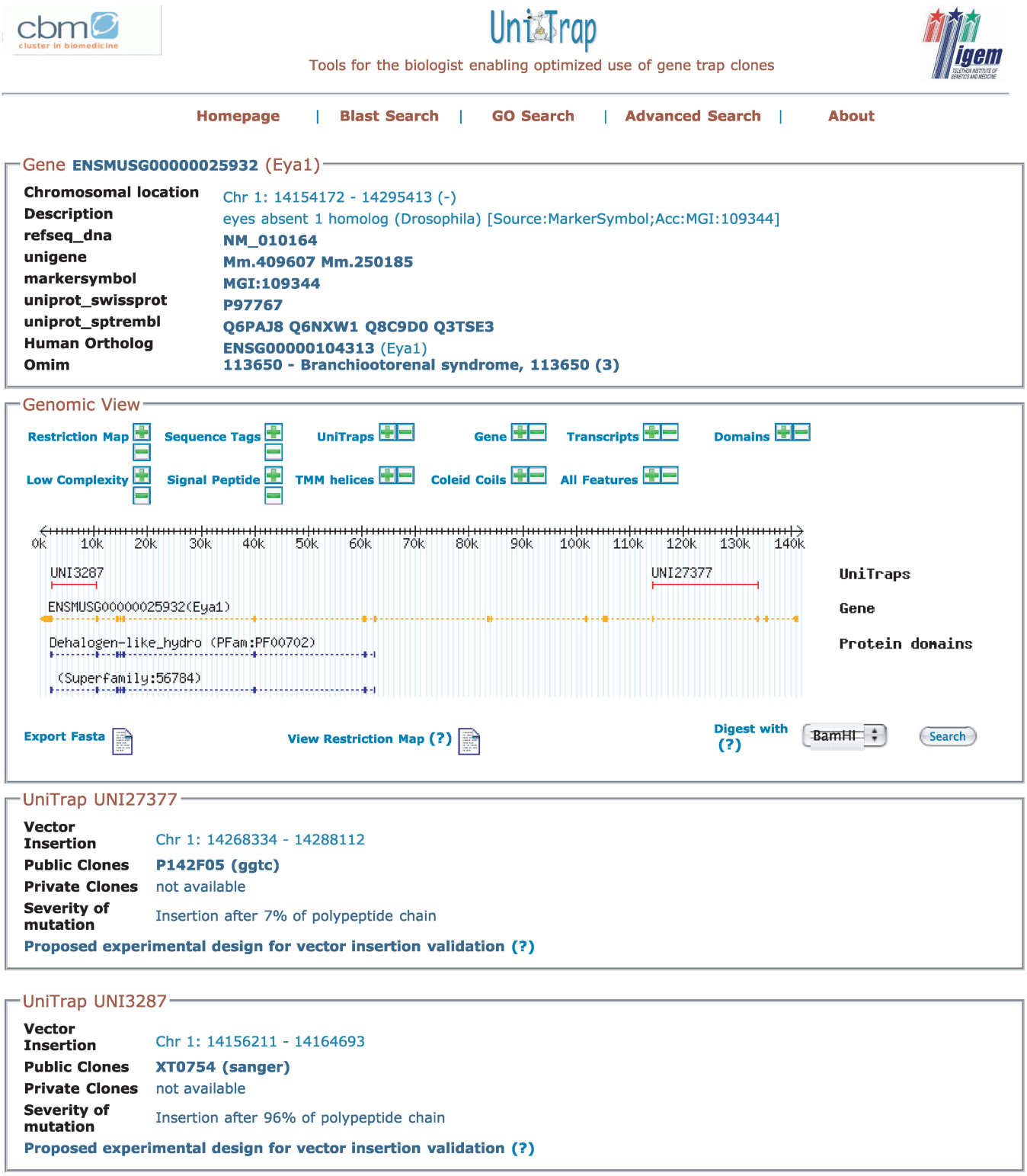

Figure 2. The UniTrap resource: data display. The image provides information regarding a trapped gene, i.e. *eya1*. Researchers can compare each distinct gene-trap insertion along with other genomic features, such as protein domains, restriction sites, etc., through a dynamic graphical representation of the genomic region of interest. A physical map of the trapped locus can be visualized to retrieve the sequences of restriction fragments or PCR primers to be used for the amplification of gene-specific probes. For each UniTrap, the page shows (i) the predicted vector insertion site, (ii) the list of public and private ES cell clones available, (iii) the severity of the mutation on the protein function and (iv) a link to retrieve the proposed experimental design for vector insertion validation.

trap vector, as well as verifying the integrity of the neighbouring DNA.

To confirm the identity of the trapped gene, RT-PCR can be performed on transcripts extracted from the mutant clone. UniTrap provides 'forward' primer sequences for each trapped gene. These primers are designed with the program EPRIMER3 (26) and match to the exon upstream of the vector insertion site. Among all possible primer sequences for a given gene, UniTrap selects the one that is predicted to work better in combination with the universal reverse primer LacZrt (TGGCGAAAGGGGGATGTG), which matches to the vector reporter gene β-*gal*. Since the wild-type allele also needs to be tested, primer pairs are designed that match the exons located upstream and downstream of the vector insertion site, respectively. The primer sequences provided can also be utilized for mouse genotyping.

To determine the exact site of the vector insertion, genomic PCR can be performed on DNA extracted from the mutant clone. A 'forward' primer is designed with EPRIMER3 on the intronic sequence predicted to host the vector; if this intron is longer than 3 kb, multiple primers (distant 3 kb from each other) are designed. These primers can be used in combination with either one of the following universal 'reverse' primers for trap vectors: GATGTGCTGCAAGGC GATTA (L232) or CCAGGGTTTTCCCAGTCACG (LacZ).

Finally, to perform a general control of the genomic locus bearing the insertion, a Southern blot analysis can be performed on properly digested genomic DNA. For each trapped gene, a restriction map of the genomic region is provided using REMAP (26) and the most common restriction endonucleases (BamHI, EcoRI, HindIII and XbaI). Restriction fragments are calculated using RESTRICT (26), and pairs of primer sequences to amplify fragment-specific probes (EPRIMER3) are also provided.

Unitrap interface

UniTrap data is stored into a MySQL database and it is freely accessible through a web interface at the address: http://unitrap.cbm.fvg.it. The UniTrap web site has been designed to provide the wet lab researchers with user-friendly tools to use gene trap clones for their gene of interest. Briefly, it allows several entry points for accessing the data:

(i) Searching by any key terms, such as gene symbols, accession numbers, gene ontology terms, human orthologs and IDs of major databases;
(ii) Searching by sequence comparison against a local database of trapped gene sequences through the BLAST algorithm; both nucleotide and amino acid sequences are allowed;
(iii) Specifying, as single or combined queries in the advanced search form, gene/clone features such as severity of mutation, presence of human orthologs, involvement in human disease based on OMIM, availability in public ES cell lines, etc.;
(iv) Choosing a genomic region by either indicating its chromosomal coordinates or clicking on the mouse karyotype image;

Each of these queries generates a list of trapped genes that meet the search criteria. Information regarding the gene of interest is visualized in a specific page where a clickable graphical representation allows a close look of all the features present in the same genomic region (Figure 2). A physical map of the trapped locus is visualized to retrieve the sequences of restriction fragments or PCR primers to be used for the amplification of gene-specific probes. Moreover, a list of all the UniTraps for the gene of interest, with information regarding public and private gene trap clones, is provided. For each UniTrap, the putative impact of the gene trap vector insertion on the protein function is shown and comprehensive information is provided to perform PCR and RT-PCR experiments in order to verify the clone and identify the exact site of vector insertion.

CONCLUSIONS

UniTrap provides a unique resource for the biologist to optimize use of both public and private gene trap clones accessible on line at: http://unitrap.cbm.fvg.it. Its ultimate goal is to aid biologists wanting to choose and utilize gene trap clones available for a gene of interest. The portal allows them to quickly find genes of interest, easily compare available clones for mutation severity, and finally it aids in the initial characterization of the chosen clone, including primers for RT-PCR, genomic PCR and Southern blot analysis.

ACKNOWLEDGEMENTS

We thank Marco De Simone and Mario Traditi for their technical support as well as Gaetano Tripoli and Antonio Romito for helpful discussion. Funding to pay the Open Access publication charges for this article was provided by was provided by the European Commission (grant n. 512003 and 513769.

Conflict of interest statement. None declared.

REFERENCES

1. Austin,C.P., Battey,J.F., Bradley,A., Bucan,M., Capecchi,M., Collins,F.S., Dove,W.F., Duyk,G., Dymecki,S. *et al.* (2004) The knockout mouse project. *Nat. Genet.*, **36**, 921–924.
2. Collins,F.S., Rossant,J. and Wurst,W. (2007) A mouse for all reasons. *Cell*, **128**, 9–13.
3. Zambrowicz,B.P. and Sands,A.T. (2003) Knockouts model the 100 best-selling drugs – will they model the next 100? *Nat. Rev. Drug Discov.*, **2**, 38–51.
4. Raymond,C.S. and Soriano,P. (2006) Engineering mutations: deconstructing the mouse gene by gene. *Dev. Dyn.*, **235**, 2424–2436.
5. Stanford,W.L., Cohn,J.B. and Cordes,S.P. (2001) Gene-trap mutagenesis: past, present and beyond. *Nat. Rev. Genet.*, **2**, 756–768.
6. Roma,G., Cobellis,G., Claudiani,P., Maione,F., Cruz,P., Tripoli,G., Sardiello,M., Peluso,I. and Stupka,E. (2007) A novel view of the transcriptome revealed from gene trapping in mouse embryonic stem cells. *Genome Res.*, **17**, 1051–1060.

7. To,C., Epp,T., Reid,T., Lan,Q., Yu,M., Li,C.Y., Ohishi,M., Hant,P., Tsao,N. *et al.* (2004) The Centre for Modeling Human Disease Gene Trap resource. *Nucleic Acids Res.*, **32**, D557–D559.
8. Cobellis,G., Nicolaus,G., Iovino,M., Romito,A., Marra,E., Barbarisi,M., Sardiello,M., Di Giorgio,F.P., Iovino,N. *et al.* (2005) Tagging genes with cassette-exchange sites. *Nucleic Acids Res.*, **33**, e44.
9. Stryke,D., Kawamoto,M., Huang,C.C., Johns,S.J., King,L.A., Harper,C.A., Meng,E.C., Lee,R.E., Yee,A. *et al.* (2003) BayGenomics: a resource of insertional mutations in mouse embryonic stem cells. *Nucleic Acids Res.*, **31**, 278–281.
10. Wiles,M.V., Vauti,F., Otte,J., Fuchtbauer,E.M., Ruiz,P., Fuchtbauer,A., Arnold,H.H., Lehrach,H., Metz,T. *et al.* (2000) Establishment of a gene-trap sequence tag library to generate mutant mice from embryonic stem cells. *Nat. Genet.*, **24**, 13–14.
11. Zambrowicz,B.P., Friedrich,G.A., Buxton,E.C., Lilleberg,S.L., Person,C. and Sands,A.T. (1998) Disruption and sequence identification of 2,000 genes in mouse embryonic stem cells. *Nature*, **392**, 608–611.
12. 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. *Nat. Genet.*, **16**, 338–344.
13. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2005) GenBank. *Nucleic Acids Res.*, **33**, D34–D38.
14. Nord,A.S., Chang,P.J., Conklin,B.R., Cox,A.V., Harper,C.A., Hicks,G.G., Huang,C.C., Johns,S.J., Kawamoto,M. *et al.* (2006) The International Gene Trap Consortium Website: a portal to all publicly available gene trap cell lines in mouse. *Nucleic Acids Res.*, **34**, D642–D648.
15. Zambrowicz,B.P., Abuin,A., Ramirez-Solis,R., Richter,L.J., Piggott,J., BeltrandelRio,H., Buxton,E.C., Edwards,J., Finch,R.A. *et al.* (2003) Wnk1 kinase deficiency lowers blood pressure in mice: a gene-trap screen to identify potential targets for therapeutic intervention. *Proc. Natl Acad. Sci. USA*, **100**, 14109–14114.
16. Horn,C., Hansen,J., Schnutgen,F., Seisenberger,C., Floss,T., Irgang,M., De-Zolt,S., Wurst,W., von Melchner,H. *et al.* (2007) Splinkerette PCR for more efficient characterization of gene trap events. *Nat. Genet.*, **39**, 933–934.
17. 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.
18. Stajich,J.E., Block,D., Boulez,K., Brenner,S.E., Chervitz,S.A., Dagdigian,C., Fuellen,G., Gilbert,J.G., Korf,I. *et al.* (2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.*, **12**, 1611–1618.
19. Hubbard,T.J., Aken,B.L., Beal,K., Ballester,B., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cunningham,F. *et al.* (2007) Ensembl 2007. *Nucleic Acids Res.*, **35**, D610–D617.
20. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **33**, D501–D504.
21. Curwen,V., Eyras,E., Andrews,T.D., Clarke,L., Mongin,E., Searle,S.M. and Clamp,M. (2004) The Ensembl automatic gene annotation system. *Genome Res.*, **14**, 942–950.
22. Hamosh,A., Scott,A.F., Amberger,J.S., Bocchini,C.A. and McKusick,V.A. (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, **33**, D514–D517.
23. Carninci,P., Kasukawa,T., Katayama,S., Gough,J., Frith,M.C., Maeda,N., Oyama,R., Ravasi,T., Lenhard,B. *et al.* (2005) The transcriptional landscape of the mammalian genome. *Science*, **309**, 1559–1563.
24. Schuler,G.D. (1997) Pieces of the puzzle: expressed sequence tags and the catalog of human genes. *J. Mol. Med.*, **75**, 694–698.
25. Burge,C. and Karlin,S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, **268**, 78–94.
26. Rice,P., Longden,I. and Bleasby,A. (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.*, **16**, 276–277.