Int. J. Dev. Biol. 42: 1025-1036 (1998)

Comprehensive mammalian genetics: history and future prospects of gene trapping in the mouse

BRIAN P. ZAMBROWICZ* and GLENN A. FRIEDRICH

Lexicon Genetics, The Woodlands, Texas, USA

ABSTRACT Gene trapping has matured into a tool with tremendous potential for mammalian biology. It both mutates and helps identify genes and can be streamlined so that many thousands of insertions can be characterized. In only a few years most of the genome of the mouse will be tagged and mutated using the latest gene trap designs. By creating such a resource, costly and time consuming alternative methods of mutagenesis and gene identification can be avoided allowing biologists to concentrate on determining gene function *in vivo*. This will mean a major shift in how the genome will be mined for new drug targets. Notably, gene discovery via gene traps does not suffer from the limitations of other methods as it is not biased by expression level. Mouse strains with specific gene mutations can be easily derived from a gene trap library constructed using embryonic stem cells. These strains will help determine the role of the gene product in mammalian physiology and hence the relevance of the gene product to human disease.

KEY WORDS: *mutagenesis, mouse embryonic stem cells, gene trapping, genomics*

Introduction

Genomics is biology's superconducting supercollider –our first big science project. And now, before the sequence of the human genome is even at the halfway mark, we have moved on to the next big projects: functional genomics and pharmacogenomics. The goal of this stepped-up assault on the genome is drug development. How can biologists find those proteins in the body that could be targets of new drugs? The answers to this question only start with gene sequence. What is required are methods to quickly find genes that code for proteins involved in relevant disease physiology and to then move those proteins on to high throughput assays to find molecules that modulate their function. This is a tall order and one that is being attacked on many fronts. Our perception is that you must go to the organism and study the gene's function in its natural biological context –the first step being to remove the gene and study the consequences. Methods that rely on expression patterns or protein structure are useful only as aids to the actual *in vivo* biology. The proposal that quick and high-throughput biology can aid drug development necessitates a relevant and versatile model system. We present here the combination of gene trapping technologies with the techniques developed in the past decades which make the mouse the best model system for studying how genes work in mammals. These techniques are well illustrated in the pages of this issue and are part of the legacy of techniques and science built by Dr. Brinster and others over the past decades.

Gene trapping is a classic technique, used in bacteria, plants and animals as long as we've known about the nature of genes (and before if you count Barbara McClintock's maize jumping genes as traps). The greatest single advantage of the technique, as we will describe below, is that a gene trap both causes a mutation and helps clone the gene –a combination of classical and reverse genetics. We will describe the technique as it has been implemented in mammalian cells and try to demonstrate that it is one of the premier tools for functional genomics. Our goal is to use traps to mutate every gene in the mouse genome and to provide a sequence tag for each of these catalogued mutations. Below we describe the types of traps that could be used to this end and the issues involved in creating a comprehensive library of mutations. One particular technique is described in more detail and illustrated with some preliminary data. This technique provides a method to isolate a mutation of choice from an arrayed library of gene trap cell lines. Finally, we will briefly consider some issues that will arise once a comprehensive gene trap library is available, such as data storage and analysis and moving on to the actual biology in the form of phenotypic analysis.

Background: genomics and gene trapping

The accumulation of sequence data has allowed the identification of many thousands of genes within the last few years. While some functional information is known about a few of these genes,

***Address for reprints:** Lexicon Genetics, 4000 Research Forest Drive, The Woodlands, Texas 77381, USA. FAX: 281.364.0155. e-mail: brian@lexgen.com, glenn@lexgen.com

0214-6282/98/$10.00

Printed in Spain

the function remains unknown for the vast majority. A major challenge for biology for the foreseeable future will be to understand the function of every gene. Currently a number of approaches are being developed to move more rapidly from gene sequence information to function, a process that has been termed “functional genomics”. As a first step, sophisticated bioinformatics programs are being used to organize the data and group it based on similarities at the nucleotide and amino acid level. High throughput methods are also being developed and used for studying expression patterns of large numbers of genes at the RNA and protein levels. These techniques include array technologies (Schena, 1996), SAGE (Velculescu, *et al.*, 1995), differential display (Liang and Pardee, 1995), 2-dimensional gels (Shevenko *et al.*, 1996; Wise *et al.*, 1997) and mass spectroscopy. Other technologies such as yeast two-hybrid screens (Fields and Sternglanz, 1994; Fromont-Racine *et al.*, 1997) are being used to develop protein interaction maps. These methods will provide important information about related amino acid and DNA sequences, define expression at the RNA and protein level, indicate factors that modify expression and identify potential interacting proteins, but this information may only provide hints to the actual function of the gene.

The genetic study of gene function in model organisms will be an important component of functional genomics. A variety of organisms will be valuable for understanding how gene products work and the genetic pathways in which they reside. Efforts are already underway to mutate all yeast genes to define their function (Bassett *et al.*, 1996; Goffeau *et al.*, 1996; Shoemaker *et al.*, 1996) and the powerful genetic screens possible in organisms such as *Drosophila melanogaster* and *Caenorhabditis elegans* are being used to define genetic pathways. However, if one is interested in gene function within a mammalian system then the mouse will provide an important model system. Embryonic stem (ES) cell technology allows the manipulation of genetic material in cell culture and selection for rare genetic events. These mutations can be studied in mice by production of chimeras, germline transmission and breeding to homozygosity. The creation of loss of function mutants in the mouse has already provided valuable resources for investigators studying gene function at an organismal, cellular and biochemical level and will continue to play an important role in understanding the function of the many thousands of genes.

A variety of methods can be used to create mutations in mice including chemicals (Brown and Peters, 1996), x-rays (You *et al.*, 1997), gene targeting by homologous recombination (Bradley, 1993; Gossler and Zachgo, 1993; Ramirez-Solis *et al.*, 1993), and gene trapping (Friedrich and Soriano, 1991; Evans *et al.*, 1997). While all of these techniques provide useful means to create mutations in mice, gene trapping has certain advantages over the other methods. Chemical or x-ray induced mutations require a large mouse colony and extensive breeding if one wishes to look at recessive phenotypes. Also, once these mutations are created, identification of the mutated gene may require a significant amount of time even with the latest mapping techniques and the possibility of-linked mutations must be ruled out. Gene targeting by homologous recombination is an excellent method for creating specific mutations in known genes. Two disadvantages of gene targeting are the time required to produce targeting vectors for each gene one wishes to mutate and the requirement for prior information about gene sequence and structure. While it should be possible to mutate all genes by homologous recombination, limitations in the speed of the process mean that many years will be required to accomplish the task. Gene trapping is a rapid method of producing and tagging mutations in the mouse genome with no requirement for prior knowledge of the gene. Although one cannot create a designed mutation using gene trapping, it provides a protocol to randomly mutate the genome and identify the gene mutated in a particular ES cell line. We will describe how the gene trapping approach can be a valuable tool both for rapidly creating loss of function mutants in ES cells as well as for identifying genes involved in specific functional pathways.

Methods of trapping genes

There are four basic methods that have been used to trap genes in mammalian cells: the enhancer trap, promoter trap, gene trap, and polyA trap. Enhancer trap vectors contain a reporter gene with a minimal promoter that is insufficient to activate transcription (O’Kane and Gehring, 1987; Bellen *et al.*, 1989; Bier *et al.*, 1989; Wilson *et al.*, 1989). Upon integration into the genome, the reporter gene is expressed only if it lands in a position that allows a cis-acting regulatory element to activate the promoter (Fig. 1B). Thus it is possible to identify transcriptionally active regions of the genome. Enhancer traps were first used in *Escherichia coli* and subsequently in *Drosophila*. Many screens have been carried out in *Drosophila* allowing the identification of multiple lines with restricted patterns of expression during development (Carlson, 1993; Sentry *et al.*, 1994). These lines have provided marker expression patterns for the study of development and in some cases nearby genes have been identified that are regulated by the enhancer. Enhancer traps are based on the knowledge that enhancers can act at a distance and because of this it can be difficult to identify the gene whose enhancer has been trapped. Cloning the gene requires isolating genomic DNA flanking the insertion site of the enhancer trap vector. Additionally the effect of the enhancer trap insertion can vary from case to case often resulting in little effect on the expression and function of the endogenous gene. While a powerful approach to identify enhancers, these vectors have not been used extensively for mutagenesis in ES cells. In one study, 5 enhancer trap lines were produced and only one resulted in a phenotype (Allen *et al.*, 1988; Kothary *et al.*, 1988). Enhancer trapping, due to the difficulty of identifying the trapped gene and uncertainty of the mutagenic effect, is not a preferred method for large scale mutagenesis in ES cells.

Promoter trap vectors are designed to identify transcriptionally active cellular promoters (Hicks *et al.*, 1997). They are Moloney murine leukemia virus based and take advantage of the fact that sequences can be inserted into the U3 region of the retrovirus LTR without compromising the retrovirus life cycle (Fig. 1C). Promoter trap vectors contain a promoterless reporter or selectable marker gene placed in the U3 region of the LTR positioned so the 5' end of the reporter gene is about 40bp from the point of integration. This places the reporter gene in a position where it can be acted upon by sequences surrounding the point of integration. These vectors are not expressed unless they integrate into an exon or near the promoter of an expressed gene. When passaged through the germline, the frequency of phenotypes observed (6 out of 16 examined) indicates that such insertions often result in loss of gene function (DeGregori *et al.*, 1994). To identify the trapped gene, DNA flanking the retrovirus integration site is subcloned and sequenced using genomic libraries, PCR methods or a shuttle vector. Hicks *et al.* (1997) have reported the use of a shuttle vector to obtain sequence tags from 400 promoter trap events in ES cells. One drawback of promoter traps is that the sequence tag may or may not contain exonic sequence thus compromising the ability to identify ESTs or cDNAs that match the trapped gene. Promoter traps have been used extensively in ES cells and have proven to be a rapid method for making and identifying mutations with sequence tags.

Gene trap vectors were developed as an efficient means of producing mutations in ES cells (Gossler *et al.*, 1989; Friedrich and Soriano, 1991; Skarnes *et al.*, 1992). They contain a reporter or selectable marker gene that is preceded by a strong splice acceptor sequence but no promoter (Fig. 1D). These vectors are not expressed unless they integrate into an intron of an expressed gene. Upon integration into an expressed gene, a fusion transcript is obtained containing 5' exon sequences of the trapped gene fused to the reporter gene sequences. This fusion transcript allows the trapped gene to be identified by either 5' RACE or cDNA cloning methods (Skarnes *et al.*, 1992; Chen *et al.*, 1994; Friedrich *et al.*, 1997; Townley *et al.*, 1997; Zambrowicz *et al.*, 1997). Although some advancements have been made in 5' RACE technology, it remains a difficult procedure to undertake for large numbers of insertion events. Gene trap vectors have proved highly efficient in mutating genes. When 60 random gene trap lines were produced using SAβgeo (a splice acceptor and the β*galactosidase/neomycin phosphotransferase* fusion gene) in a retrovirus, and bred to homozygosity, half demonstrated an overt phenotype indicating that these gene trap vectors are mutagenic in many positions and are more mutagenic than enhancer and promoter traps (Friedrich and Soriano, 1991). A number of these retrovirus based gene trap lines have been examined in more detail to monitor the effect of the insertion on expression of the endogenous gene. Northern blot and sensitive PCR assays have demonstrated loss of the endogenous transcript and have indicated no splicing around the SAβgeo insertion (Chen *et al.*, 1994; Deng and Behringer, 1995; Zambrowicz *et al.*, 1997; Ross *et al.*, 1998). When electroporated, SAβgeo has also proven to be mutagenic producing severe hypomorphs or null alleles (Serafini *et al.*, 1996; Friedrich *et al.*, 1997). These data give a reasonable assurance that gene trap vectors will dramatically reduce or abolish expression of the trapped gene and indicate that gene trapping is an excellent protocol for mutating genes, especially if it can be combined with a method of rapid sequence acquisition.

Promoter and gene trap vectors produce loss of function mutations in ES cells and tag the trapped gene for identification. However, they have some important limitations. One critical limitation is that they can only trap genes expressed in the experimental cell type. This limits the number of genes that can be mutated by these methods. If one wishes to trap the maximal number of genes in ES cells, it is essential to have methods that trap genes that are not expressed in ES cells. Methods are also required to improve the efficiency of obtaining a sequence tag from the trapped gene and preferably to obtain coding sequences. The methods required for obtaining sequence tags from gene and promoter traps include several steps such as dialysis and bacterial electroporations that are not easily automated (Hicks *et al.*, 1997; Townley *et al.*, 1997). These gene and promoter trap sequence tags often contain 5' untranslated regions of cDNAs or non-transcribed genomic DNA which are not optimal for database searches as they are currently underrepresented. It would be preferable to have methods with a higher probability of obtaining sequence tags that contain coding sequence.

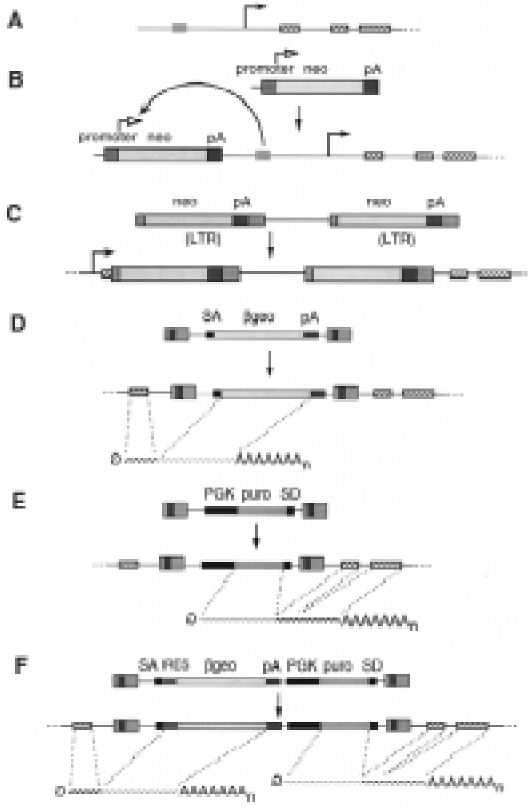


Fig. 1.**Vectors for gene trapping. (A)** *Schematic representation of a hypothetical gene consisting of 3 exons (striped rectangles), a promoter (arrow) and an enhancer (shaded box).* **(B)** *The enhancer trap vector contains a minimal promoter (open arrow), selectable marker gene (neo for neomycin phosphotransferase) and a polyadenylation sequence (pA). Upon integration near the 3 exon gene depicted in A, the enhancer can activate expression of the selectable marker from the minimal promoter. This promoter trap vector contains a selectable marker (neopA) in both long terminal repeats (LTR) of a retrovirus. When it integrates into the promoter or an exon of the 3 exon gene, transcription of the selectable marker gene is activated.* **(C)** *This gene trap vector contains a splice acceptor sequence (SA) fused to a selectable marker gene (βgeo for βgalactosidase/neomycin phosphotransferase fusion gene) and is depicted in retrovirus form between two LTRs. When it integrates into an intron of the 3 exon gene, a fusion transcript is created between 5′ exons of the trapped gene and βgeo sequences.* **(E)** *This polyA trap vector contains a promoter (PGK) fused to a selectable marker gene (puro for puromycin N-acetyltransferase) and followed by a splice donor sequence (SD) and is depicted as a retrovirus. When it integrates into the 3 exon gene, it produces a fusion transcript between puro sequences and 3′ exons from the trapped gene.* **(F)** *This combination gene/polyA trap vector contains components of both the gene trap (D) and polyA trap (E) vector and upon integration into an intron of the 3 exon gene, produces two fusion transcripts.*

PolyA addition traps are a more recent addition to the gene trapping methods and provide a means to both trap non-expressed genes and obtain coding sequence tags (Niwa *et al.*, 1993; Yoshida *et al.*, 1995). PolyA trap vectors contain a promoter active in ES cells directing the expression of a selectable marker gene (Fig. 1E). The selectable marker gene does not contain a polyA

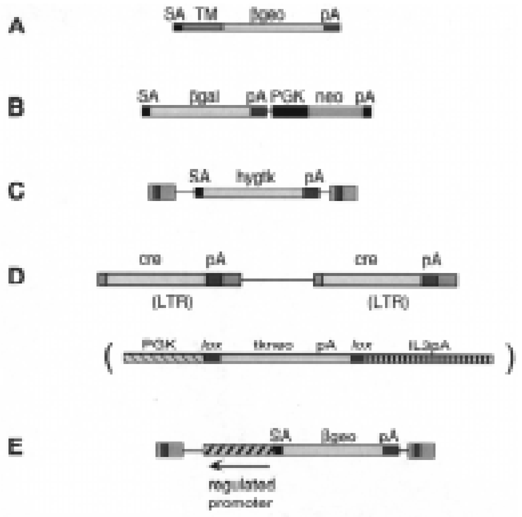


Fig. 2. Gene trap vectors for selection of subgroups of gene trap events. (A) *This secretion trap vector contains a SA sequence fused to a transmembrane (TM) encoding sequence fused to the 5' end of a selectable marker gene (βgeo).* **(B)** *This vector contains a selectable marker sequence (neopA) expressed from a promoter (PGK). All insertion events may be selected with G418 and the gene trap portion of the vector (SAβgalpA) is used to examine regulated expression of trapped genes.* **(C)** *This is a promoter trap vector containing cre-recombinase (cre) in the LTRs. In parentheses is the switch reporter gene consisting of a promoter (PGK), a thymidine kinase/neo fusion gene (tkneo) flanked by recombination sites recognized by cre (lox) and the interleukin-3 gene. When a gene is trapped that is expressed, cre is expressed and causes removal of tkneo resulting in expression of IL3 from the PGK promoter.* **(E)** *This is a combination gene trap antisense expression vector containing a regulated promoter in reverse orientation to express antisense message from the trapped gene as well as the typical components of a gene trap vector (Fig. 1C).*

addition signal and is therefore not expressed unless it lands within an intron of a gene and traps downstream exons that include a polyadenylation signal. This design should allow the trapping of non-expressed genes as the trapping vector carries its own promoter. The gene trap fusion construct now contains selectable marker sequence fused with exons from the trapped gene on the 3' end. This type of fusion allows the efficient process of 3' RACE to be used to identify sequence tags and these 3' sequences are more likely to contain coding sequences. When first attempted, these methods were shown to trap genes and in one case examined the insertion abolished the expression of the trapped gene (Yoshida *et al.*, 1995). In that case, the vector was a combination of a gene trap and a polyA trap vector suggesting that this combination may be optimal for maintaining all the advantages of gene trap vectors while adding the ability to trap non-expressed genes and more rapidly obtain sequence tags. However in these first reports, there was no demonstration that non-expressed genes could be trapped and the sequence acquisition was inefficient. We have modified the use of polyA trapping vectors (Fig. 1F) resulting in an efficient process for obtaining sequence tags and have demonstrated the ability to trap non-expressed genes (Zambrowicz *et al.*, 1998).

The DNA vectors used for trapping genes can be introduced into ES cells by a variety of methods but the two most common methods have been electrophoration and retroviral infection (Friedrich and Soriano, 1991; Skarnes *et al.*, 1992, 1995). Electroporation can often produce multiple integrants of the vector and the insertion can cause deletions and rearrangements of DNA that could affect genes other than the one trapped (Niwa, *et al.*, 1993). It is not known what percentage of the genome can be targeted by electroporation. In spite of these caveats, electroporation has been used successfully to trap and study the function of a number of genes (Skarnes *et al.*, 1995; Rijkers and Ruther, 1996; Serafini *et al.*, 1996; Friedrich *et al.*, 1997; Torres *et al.*, 1997). Retrovirus infection has also been used extensively to introduce trapping vectors into ES cells (Friedrich and Soriano, 1991; Hicks *et al.*, 1997). Retrovirus integration into DNA is characterized by no loss or rearrangement of DNA. A single copy of the retrovirus is integrated when multiplicity of infection is controlled and a 4 to 6bp duplication of DNA is seen at either end of the integration. One limitation to the use of retroviruses may be non-random integration into the genome. Several reports have suggested that retroviruses have some hot-spots for integration but the majority of integration events are essentially random (Chang *et al.*, 1993; Withers-Ward *et al.*, 1994). With the recent development of transposable elements such as *mariner* (Guerios-Filho and Beverley, 1997), *Sleeping beauty* (Ivics *et al.*, 1997) and LINE-1 (Moran *et al.*, 1996; Sassaman *et al.*, 1997), and the ability to obtain integration events in the mammalian genome, one can envision these being developed as delivery methods for integration of gene trap vectors. These methods will need to be developed further for efficient integration in ES cells and to control for single integration events. It remains unknown once again what percentage of the genome will be open to integration by these elements and what integration preferences they may have.

Methods of screening gene trap mutations

Gene trapping has proven to be an efficient method of producing novel mutations in mouse genes. The challenge has not been in making the mutations but rather in deciding which mutations to choose for further study and how to proceed rapidly in the identification and analysis of the mutations. Early gene trap work in the mouse proceeded directly from trapping a gene in ES cells to producing mice from those ES cell lines (Friedrich and Soriano, 1991; Skarnes *et al.*, 1992). These studies proved the concepts worked and identified a variety of interesting genes but it became apparent that screening through mutations via phenotype was a time consuming and expensive endeavor. Although a tedious approach, the phenotypic screen remains valuable for obtaining phenotypes that might not be predicted. One example is the *Gtl2lacZ* gene trap line that produces a parental origin-dependent phenotype, characterized by dwarfism in animals inheriting the mutation from the father but showing a reduction in penetrance and expressivity when inheriting the mutation from the mother (Schuster-Gossler *et al.*, 1996). This potential imprinted gene may not have been identified unless mice were made and examined. To avoid relying entirely on phenotypic screens, pre-screening methods have been developed with the intention of identifying which gene trap cell lines might be of most interest for further analysis. These methods also indicate the potential of the gene trap method for dissecting cellular pathways and suggest the power of the technique is limited only by the imagination of investigators and the screens they can devise.

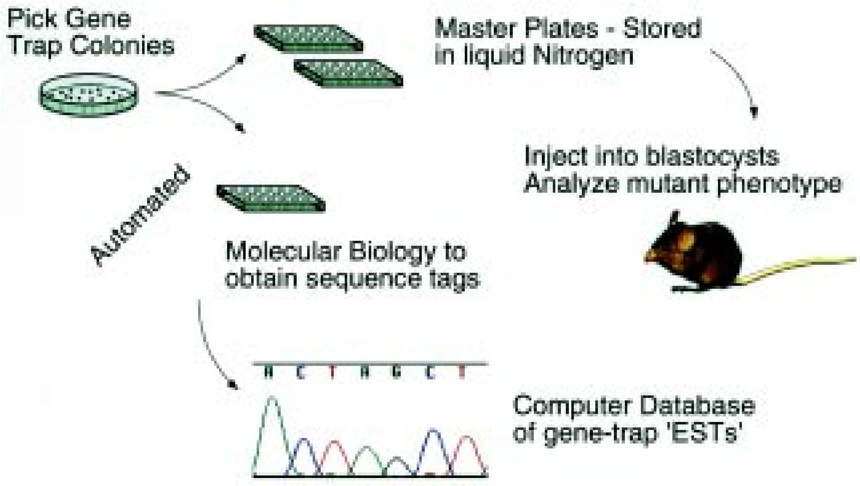

Pick Gene Trap Colonies
Master Plates - Stored in liquid Nitrogen
Inject into blastocysts Analyze mutant phenotype
Automated
Molecular Biology to obtain sequence tags
A C T A G C T
Computer Database of gene-trap 'ESTs'

Fig. 3. A flowchart for the large-scale acquisition, storage and cataloguing of gene traps in mouse ES cell lines. *The system combines simple tissue culture, automated molecular biological methods and a relational sequence database and is designed around robust gene trapping vectors delivered via retroviral infection. With such a system is has become possible to mutate most genes in the genome of the mouse, the most widely studied model mammalian organism.*

Expression screens

One of the important concepts that was demonstrated in early gene trap work was that the βgalactosidase reporter gene could be used to examine the expression pattern of the endogenous trapped gene. Skarnes *et al.* (1992) compared the expression of the βgal reporter gene as determined by X-gal staining with the expression of the endogenous trapped gene as determined by *in situ* hybridization. They found that when examined at three developmental timepoints, there was a good match between βgal expression and expression of the endogenous gene. This demonstrated that X-gal staining could be used as a method to identify interesting patterns of expression in gene trap lines providing one method for selecting lines for further analysis. Friedrich and Soriano (1991) examined the expression pattern of X-gal in 28 lines of gene trap mice and found both widespread and restricted patterns. Wurst *et al.* (1995) expanded the use of X-gal screening of gene trap lines by producing chimeras from 279 separate gene trap lines and examining expression during development. They found that at embryonic day 8.5, 13% had restricted expression patterns, 32% had widespread expression patterns and 55% had no detectable expression pattern. The lines with no detectable expression at E8.5, were examined again at embryonic day 12.5 and one third were found to be expressed. While these results demonstrate the feasibility of the approach, the production of chimeras necessitates a considerable time and effort. If one is interested only in genes with restricted patterns of expression in a subset of tissues it could require the screening of large numbers of clones before any clone of interest is identified. These results suggested that the development of methods to pre-screen gene trap clones before production of mice would be extremely valuable.

Secretion trap screens

Investigators examining gene traps in ES cells realized that some lines had different subcellular expression patterns and did not exhibit the standard cytoplasmic staining (Skarnes *et al.*, 1995). These results suggested that the creation of fusion proteins resulting from the attachment of amino acids encoded by 5' portions of the trapped gene to βgal could result in targeting of βgal to specific subcellular locations. These clones could identify genes containing sequences that deliver them to locations critical for their function. Skarnes *et al.* (1995) decided to test this hypothesis by fusing signal sequences, that target transmembrane and secreted proteins to the endoplasmic reticulum (ER) to the 5' end of βgal. They found that if the βgal had a signal sequence, it was targeted to the lumen of the ER and inactivated within that environment. However, by following the signal sequence with a transmembrane domain from CD4, they found that βgal activity was restored and localized to the (ER). This led to the development of "secretion trap" vectors containing a splice acceptor sequence followed by a transmembrane sequence and then βgal (Fig. 2A). It was predicted that traps in genes containing a signal sequence could be identified by X-gal staining of the ER. This was demonstrated when the authors obtained 5' RACE products from a number of secretion trap clones and found that they were in fact transmembrane or secreted proteins. This approach allowed a pre-screening for a subset of proteins known to be critical for many developmental processes. These experiments indicated the power of *in vitro* staining of cells and others have gone on to differentiate ES cells *in vitro* to identify genes expressed in specific cell lineages (Rijkers and Ruther, 1996; Baker *et al.*, 1997). The ability to differentiate ES cells into a variety of cell types including neurons, glia, chondrocytes, cardiomyocytes and hematopoietic cells can be used to identify trapped genes expressed in lineages of particular interest.

Gene induction screens

Reporter gene expression has been further exploited for the identification of trap events in genes that are regulated by specific induction events. Forrester *et al.* (1996) used a gene trap vector containing SAβgal and an expressed neomycin selectable marker gene (Fig. 2B). Using this vector they could select for vector insertion events using G418 independent of trapping an expressed gene. Colonies selected for the insertion event were divided among replica plates and subjected to retinoic acid induction or no induction followed by X-gal staining. In this way they could identify trapping events in genes whose expression was either induced (increased βgal staining) or repressed (decreased βgal staining) upon retinoic acid treatment. These results indicate the potential to screen for gene traps responsive to a variety of induction events including growth factors, transcription factors, responses to apoptosis etc. These results also point to gene trapping as a method not only to trap and mutate genes in ES cells but also as a strategy to identify genes involved in a variety of cellular responses and developmental processes.

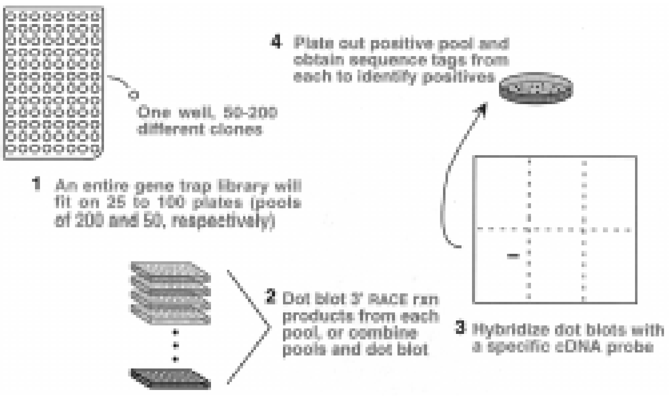

4 Plate out positive pool and obtain sequence tags from each to identify positives
One well, 50-200 different clones
1 An entire gene trap library will fit on 25 to 100 plates (pools of 200 and 50, respectively)
2 Dot blot 3' RACE rxn products from each pool, or combine pools and dot blot
3 Hybridize dot blots with a specific cDNA probe

Fig. 4. Isolating specific gene trap mutations from an arrayed library. *One of several possible methods is diagrammed that will allow a specifically mutated ES cell line to be isolated from a large library of cell lines. This particular method relies on the hybridization of cDNA sequences from the gene of interest to pooled sets of 3' RACE products derived from splice-donor/polyA traps.*

Gene trapping combined with methods to monitor induction of expression of the trapped gene have now been used in a variety of cell types. Gogos *et al.* (1996) trapped genes using a retrovirus containing SAβgeo in C2C12 myoblasts that can be induced to differentiate into myotubes. They looked for genes in which the βgal expression was up-regulated during this differentiation process. One of the genes they identified was cathepsin B. Kerr *et al.* (1991) used transfection of a gene trap vector containing SAβgeo followed by FACs sorting to identify gene trap events that were responsive to LPS treatment in a B-lineage cell line. LPS can induce the differentiation of some B-cell lineages. They made replica plates of the gene trap clones, treated them with LPS and looked for gene traps in which βgal expression was modulated. They identified 3 repressed and 2 induced gene trap lines and the 5' RACE products of all 5 genes were novel. Others have developed sorting and selection schemes to identify induced or suppressed gene trap events. Gogos *et al.* (1997) used gene trap vectors containing SAhygtk (a hygromycin/thymidine kinase fusion gene, Fig. 2C) or SAβgeo. They trapped genes in NIH 3T3 cells and either FACs sorted (for SAβgeo gene traps) or gancyclovir treated (for SAhygtk gene traps) to remove constitutively expressed gene traps. They then induced myoD expression to examine genes regulated by myoD using either FACs sorting (for SAβgeo traps) or hygromycin selection (for SAhygtk traps) to identify up-regulated genes. These protocols proved useful for identifying gene traps induced by myoD expression and eliminated the steps of clone picking and replica plating thus reducing the effort required to identify regulated genes. Similarly, others have trapped and identified genes induced during differentiation of myeloid precursor cells into appropriate lineages or P19 cells into neurons (Imai *et al.*, 1995; Jonsson *et al.*, 1996). Russ *et al.* (1996) developed a cre-lox based switch to allow selection for genes up-regulated during programmed cell death in a hematopoietic precursor cell line (Fig. 2D). The switch allowed them to select for genes that are induced by factor deprivation in a cell line that responds to such deprivation by undergoing apoptosis. Once again the need for clone picking and replica plating was eliminated. In addition, the use of cre-recombinase promoter trap vectors in combination with a separate switch that allowed selection after recombination theoretically provides a method to identify induction events even if they are transitory since once the gene is induced and the switch is flipped, the cells are marked even after the gene is turned back off. These advances indicate that gene trapping is not restricted to ES cells and can be used in a variety of mammalian cell lines chosen based upon their ability to undergo cellular processes of interest.

Functional screens

It has also been demonstrated that gene trapping can be used to identify genes involved in specific cellular processes using screens based on phenotypes. Hubbard *et al.* (1994) used promoter trapping in CHO cells to mutate genes involved in glycosylation. They selected for mutants with reduced cell surface expression of Neu5Ac by selecting for resistance to wheat germ agglutinin and identified four separate integration events falling within a 796bp region of the genome. The ability to identify mutations in this way may be dependent on haploinsufficiency or may result from selection events in cell lines that have already lost or are susceptible to loss of portions of the genome containing the second gene copy. Li and Cohen (1996) also took advantage of a phenotypic screen in cell culture by devising a combination gene trap antisense expression vector to identify trap events in tumor suppressor genes. The vector contained SAβgeo in one orientation to trap and mutate one copy of potential tumor suppressor genes and a promoter in the opposite orientation to direct the expression of antisense transcripts from the gene trap locus (Fig.

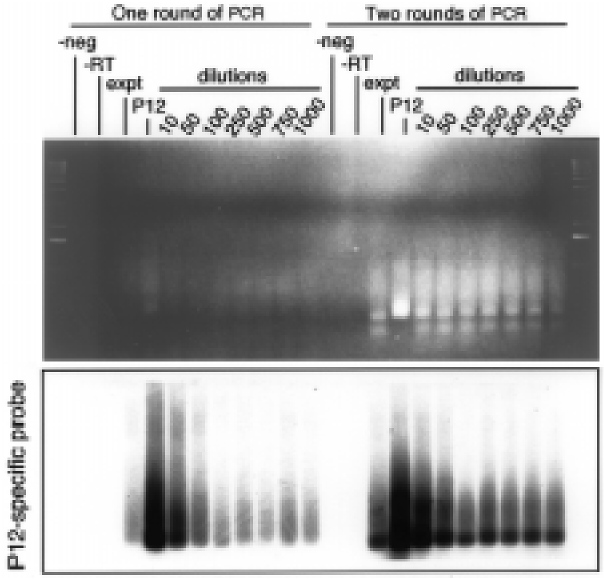


Fig. 5. Detection of a gene-specific sequence by hybridization after dilution into pools of heterologous gene trap cell lines. *Cells from the 'P12' gene trap clone were mixed with cells from other uncharacterized gene trap lines in the ratios indicated. RNA from each set of cells was isolated for 3' RACE reactions designed to amplify the gene-trap specific transcript (Zambrowicz et al., 1998). Products of one or two rounds of amplification were run out on an agarose gel (top), blotted and probed with P12-specific sequences (bottom). (-neg: water negative control; -RT: no reverse transcriptase negative control; expt.: experimental dilution of 1 in 200; P12: no dilution, P12 cells only).*

2E). They hoped that the antisense expression would be sufficient to reduce or eliminate the function of the second wild type allele of the trapped gene. They were able to identify NIH 3T3 clones that could grow in soft agar after a gene trap event and identified the *Tsg101* gene as a potential tumor suppressor gene. These two experiments indicate that gene trapping can be used in combination with protocols that allow for the identification or selection of phenotypes of interest. The power of gene trapping to identify genes involved in specific pathways may only be limited by the ingenuity of the investigator in devising phenotypic identification or selection protocols. The combination of inductive and phenotypic screens, that are possible using gene traps as mutagens, along with the new methods of rapid sequence identification of trapped genes, provide powerful methods of genetic screening in mammalian cells.

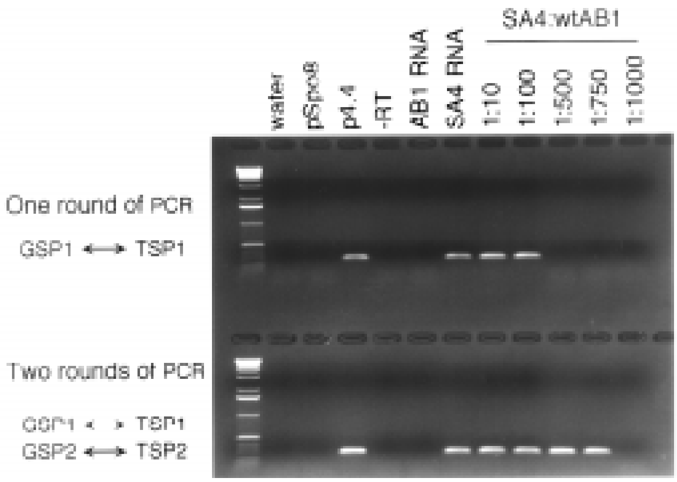


Fig. 6. Detection of a gene-specific sequence by RT-PCR after dilution into pools of heterologous gene trap cell lines. *Cells containing the gene trap 'SA4' were diluted at the ratios indicated with wild-type AB1 ES cells. RNA from each pool was used for RT-PCR amplification using primers specific for the SA4 gene (GSP1 and GSP2, for* ***g****ene* ***s****pecific* ***p****rimer) and for the gene trap sequences (TSP1 and TSP2, for* ***tr****ap* ***s****pecific* ***p****rimer). Products of the reactions are shown after one round of PCR and after a further, nested round of PCR. (water: negative control; pSpo8: plasmid control containing only gene cDNA sequence; p4.4: plasmid control containing the gene trap fusion cDNA sequences; -RT: no reverse transcriptase negative control).*

Large scale gene trapping

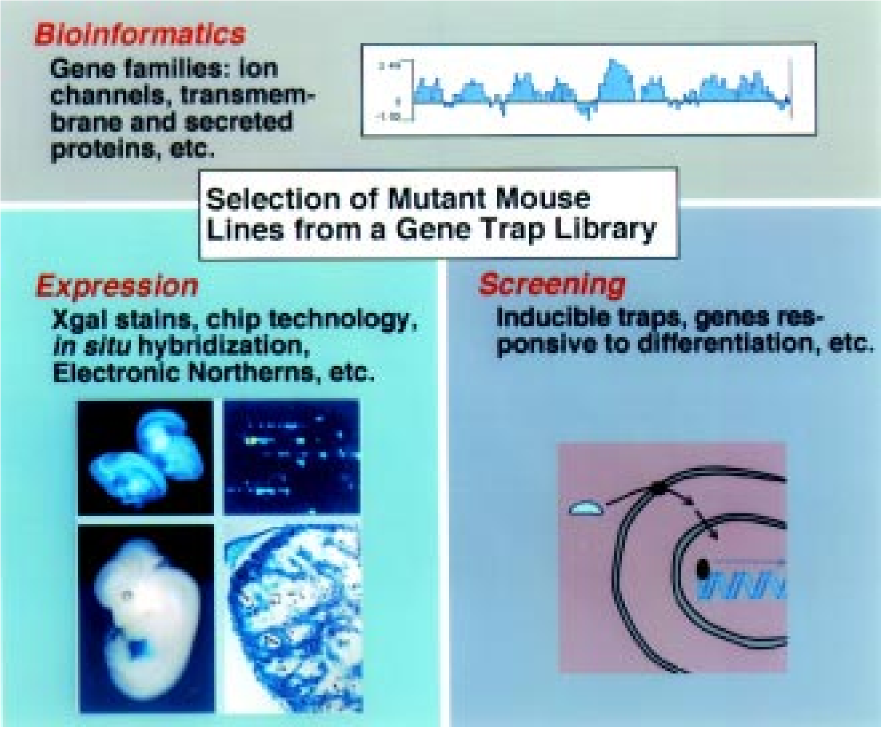
Creating libraries of gene trap mutations

With the increasing amount of gene sequence information, it has become apparent that one excellent way to pre-select gene traps is to obtain sequence tags of trapped genes. This is a considerable challenge as obtaining sequence tags from each gene trap has been a tedious process often requiring 5' RACE or production of genomic or cDNA libraries. More recently a number of approaches have been developed that are increasing the speed of obtaining gene trap sequence tags. Townley *et al.* (1997) and Chowdhury *et al.* (1997) developed more efficient methods of 5' RACE and were able to obtain sequence tags from 40% of 150 gene trap lines and 55 gene trap events, respectively. Hicks *et al.* (1997) reported the use of shuttle vectors to identify flanking sequence from 400 promoter trap events. Both these methods are significant improvements in the speed of sequence acquisition but both methods require a number of steps that are difficult to automate. We have modified polyA trap methods and automated them in the 96-well format allowing us to obtain sequence tags from over 20,000 gene trap events (Zambrowicz *et al.*, 1998). These sequences are placed in a relational database and through bioinformatics and searching protocols it is possible to identify trap events in genes of interest for further study (Fig. 3) This library of ES cells with traps in thousands of genes will play an important role in the study of gene function. With such a library it will become possible to screen through large numbers of mutant mice for specific phenotypes of interest. Candidate genes to be analyzed may be identified through bioinformatics, expression, positional cloning or a variety of other methods.

We believe the most efficient method of pre-screening trap events is by sequence. In combination with the identification of full length sequence for all genes, it should become possible to match any sequence tag with a known gene. Once a gene trap has been identified with a sequence tag it can be linked with all other information pertaining to that gene which may include expression pattern and interacting proteins. The sequence tag can also be used to obtain expression information by using it as a probe for methods such as Northern blots, RNase protection assays, *in situ* hybridization or more sophisticated array techniques. Mouse chimeras or heterozygotes can be used to examine X-gal staining patterns. The gene trapped ES cells can also be subjected to any other pre-screening method such as *in vitro* differentiation or induction screening. In summary the ability to identify each gene trap with a sequence tag is likely to be the most rapid method to prescreen and catalogue mouse mutations providing a pre-made functional genomics resource ready to be mined for phenotypic analysis.

One example of the power of identifying gene traps by sequence tags is the case of the *netrin-1* gene trap line (Serafini *et al.*, 1996). This line was produced by secretion trapping and identified by 5' RACE. The *netrin-1* protein is secreted and can attract or repel neurons in the developing neural tube. The availability of the pre-identified *netrin-1* gene trap clone allowed investigators to rapidly examine the phenotype of the resulting mice. The gene trap produced a hypomorph that confirmed the function of the gene and matched the phenotype found in mice with a disruption of the DCC (deleted in colon cancer) gene, the *netrin* receptor (Fazeli *et al.*, 1997). As libraries of ES cell mutants grow this ability to move rapidly to the study of mouse mutants will dramatically increase.

The gene trapping method is also a powerful gene identification method. When polyA trap vectors are used they identify genes regardless of expression and about 60% of our sequence tags do not have matches in GenBank (Zambrowicz *et al.*, 1998). Unlike approaches used to obtain ESTs, gene traps are not biased against trapping genes expressed at low levels or only transiently. A library of gene trap sequences contains genes that may only be detected using multiple normalized cDNA libraries from many tissues. The high frequency of novel sequences obtained by our gene trap methods indicates the power of the methods to identify novel transcribed sequences. Even after lengthening these novel sequences in either direction along the cDNA, many remain unknown and are likely newly discovered genes.



Bioinformatics

Gene families: ion channels, transmembrane and secreted proteins, etc.

Selection of Mutant Mouse Lines from a Gene Trap Library

Expression

Xgal stains, chip technology, *in situ* hybridization, Electronic Northerns, etc.

Screening

Inducible traps, genes responsive to differentiation, etc.

Fig. 7. Integrated functional genomics is possible with a well characterized and comprehensive gene trap library constructed in mouse ES cell lines. *Starting from anonymous coding sequence, bioinformatics will be able to classify protein products into known families or will be able to provide some hints about function due to certain structural features. The temporal and spatial distribution of the mRNA can be determined en mass by chip array technologies or more directly via in situ hybridization (whole-mount or sections). Genes can be further classified based on their presence in certain in vitro functional assays. Ultimately, all these data are the foundation to the physiological information that will be derived once a mutant mouse is generated quickly from an ES cell line containing a gene trap mutant allele.*

Rapid access to individual mutant lines within a gene trap library

We have undertaken the construction of a gene trap library with mutations in every gene of the mouse genome. As discussed in the previous section, the acquisition of sequence information from each of thousands of gene trap clones is an automated process and we have so far accumulated over 20,000 such sequences. The current rate of accumulation is 500-1000 gene trap lines per week. The obvious and easiest way to access a mutation of interest is by sequence, and the highest likelihood of success is when the library is completed. A method to circumvent the requirement for sequence information from the entire library would be to adapt pooling and screening strategies to quickly obtain a clone of interest that is mixed in with many other clones. Such a method would obviate the need to pick 100,000 to 500,000 individual gene trap clones. Similar strategies have been described in other organisms, for example *C. elegans* (Zwaal *et al.*, 1993).

One can take advantage of retroviruses to generate a large number of integrants for a complete gene trap library in one experiment. The resulting colonies could be pooled and screened for particular clones with insertions into any gene of interest. To generate such a library in one experiment only requires a packaging cell line that gives an adequate titer and a means to mass infect ES cells. The former is usually not a consideration since a packaging cell line with a relatively low titer of only 1,000 cfu/mL would require only half a liter of virus-containing media to provide infectious particles for 500,000 integrations, or about 5 integrations for every gene in the mouse genome.

A proposed method (represented schematically in Fig. 4) is as follows. Five genome equivalents of gene trap clones (500,000) are generated by mass infection. The resulting infected cells are plated onto 96-well plates in pools of 100 to 200 colonies (i.e., 50 or 25 plates, respectively). Each of these clones has a distinct gene trap event, and there is no need to pick them individually. The cells are grown to confluence and split 1:3. Two plates are for freezing and storage and the remaining plate is processed for screening. This processing follows a standard automated 3’ RACE procedure as described above for polyA-type traps (Fig. 1F). This results in PCR reaction products representing cDNA sequence from all 100-200 clones in a given pool. These PCR products are gridded onto a nylon membrane and used for hybridization (for example, using a 96-well vacuum manifold, or ‘dot blotting’ apparatus). To locate a clone of interest, a full length or partial cDNA is labeled and hybridized against the gridded PCR products. Positives indicate that a gene trap retrovirus had integrated into the gene of interest.

This hybridization strategy relies on the type of message generated by a polyA-trap insertion (Zambrowicz *et al.*, 1998). As depicted in Figure 1E and F, the splicing that is concomitant with proper termination and poly-adenylation of the message generates a fusion transcript. It is this fusion transcript that is driven by the trap promoter in ES cells and that is amplified specifically in the 3’ RACE procedure. Since the trap will integrate into an intron, the PCR products are unique to the trapped gene and begin at discrete exon boundaries. Full length cDNA probes or partial probes (preferably near the 3’ end, such as those generated from ESTs) would be specific for insertions into their corresponding genes.

Once a positive pool is identified in this manner, the corresponding pool of cells is thawed and plated at a low density to give individual colonies. A set of 200-300 clones are isolated and processed by automation to provide the gene trap sequence tag from each. The particular cell line in the pool with the mutation of interest is identified by sequence.

Experiments were undertaken to prove that the various aspects of this method were feasible. First, it was important to determine that a positive clone in a background of negative clones could be detected by hybridization after 3’ RACE and second, it was crucial to show that positives could be recovered from pools of up to 200 clones.

A gene trap cell line with a random insertion into a novel, uncharacterized gene was used as a known positive to test the level of detection possible with hybridization. A subfragment of the 3' portion of the cDNA from this gene (labeled 'P12') was subcloned to provide a hybridization probe. Cells with a gene trap insertion into P12 were diluted with cells from other clones with gene trap insertions. Figure 5 shows the results from a hybridization experiment using the P12 specific probe on the PCR products generated from these pools. The primers used to amplify the fusion cDNAs in the pools of cells were specific to the gene trap on the 5' side and for the tailed poly-dT RT primer on the 3' side (Zambrowicz *et al*., 1998). The results from a series of such experiments indicate that this method is quite sensitive and that the positive cell line could be detected by hybridization down to a dilution of 1 in 1000 background clones using two rounds of PCR for 3' RACE. Note that in this experiment, the PCR products were separated on an agarose gel before transferring to a hybridization membrane whereas in an experiment using the entire pooled library of gene trap lines, each set of PCR products would be dotted onto the membrane directly.

An alternative to finding a positive pool by hybridization would be to detect it by a specific RT-PCR. Within a hypothetical fusion transcript generated by a polyA trap are known sequences juxtaposed in a unique configuration. Sequences of the selectable marker are now attached 5' of sequences from a gene of interest only in clones where the gene trap has landed in, and mutated that particular gene. It should be possible to detect such events in a background of negative clones by PCR specific for such a fusion transcript. A preliminary experiment, the results of which are shown in Figure 6, indicates that a sensitivity of at least 1 in 500 is possible with such a method. Sensitivity could obviously be increased if the RT-PCR products are blotted and probed for sequences upstream of the gene-specific primers. This method is not as robust as simple hybridization after 3' RACE. The biggest disadvantage is that unique RT-PCR sets would have to be performed for each gene an investigator desires to extract from the library, whereas with hybridization of 3' RACE products, the membranes can be used with multiple different probes. When the membranes start to suffer from use, a new set can be made from frozen 3' RACE products.

The second experimental hurdle noted above is the ability to recover a positive clone from a pool of up to 200 clones. This is a concern given that subpopulations of cells will have different growth rates in culture. It is important to note, however, that in the final experimental design there are relatively few cell divisions between the time the infections are made and when positive pools are plated at low density. There is little time then for any particular set of subclones to outgrow and dominate the culture before the pool is frozen into storage. After the positive pools are identified and thawed, the first step in culture is to separate the clones into colonies by plating at a low density. A simple experiment in which a set of independent gene trap lines was grown together through a half dozen cell divisions, plated to form single colonies and then processed for sequence tags demonstrated that the distribution of different clones within the pool was sufficiently broad to recover all the component clones in a modest sized set of subclones.

The pooling techniques described here do not obviate the need to make a complete library of gene trap lines, each clone with a sequence tag. What the techniques could provide is stop gap access to desired mutations if they have yet to appear in the library. And once the library is completed, the pooling technique could still be useful if an investigator desires insertions of different types of vectors into a gene of interest –for example vectors that cause conditional or inducible mutations.

Integrated databases

The age of genomics has forced biologists to become sophisticated computer users. The size of genomics databases and the need to continuously analyze and cross reference sequence data is a significant computing problem (Doolittle, 1994). And the problem is growing now that many biologists and biotechnology firms are moving beyond sequence toward designing and using large scale techniques to determine gene function and expression. The types of data generated from all these approaches are diverse –but it is essential that they can be seemlessly accessed and that relations between them can be established (Fields, 1992).

The prospects are slim right now that this Tower of Babel will be tamed. To accommodate major pharmaceutical and large biotechnology companies, there are new companies devoted solely to writing robust bioinformatics software. Unfortunately, the nature of the market and competition between these start-ups means that standards of data representation cannot exist. Even under the large umbrella of U.S. federal funding there is more than one competing standard.

The consequences of this free-for-all are not yet apparent. After all, the one basic data type –DNA or protein sequence– is the day to day working material of the molecular biologist and we have no problems in extracting maximal information from limited numbers of sequences. However, when expression data, structural data, sequence polymorphisms, mapping data, species comparisons, predicted protein structures and so on are included into the same database structures, the problem of management is eclipsed only by the problem of interpretation. The computer needs to become a tool that is not only a data manager, but also has a hand in interpretation – it is impossible for humans to make correlations in such vast arrays of data.

Indeed, we think overcoming this computing challenge will be a difficult feat to accomplish in the coming decade. The reason is simple, and is basically a consequence of the limits of reductionism. Genomic data today, of whichever type, is in isolation from the organism from which it is derived. Gene and protein sequence describe the coding potential of just one entity that plays a part in the function of an organism. So does all the associated data: expression patterns of the gene, the structure of the gene, etc. And the current state of that one gene is just a snapshot in geological time over which evolution is acting. The problem is even more stark when one considers how evolution dabbles with what's on hand. Ponder for a moment the fact that the human genome probably has over 2000 tyrosine kinase-like genes. To solve the problem of transferring signals to the nucleus, why not use what is there and does the job already?

One way to take a more holistic approach is to go right to the organism, take away the gene's function and examine the consequences. That is why we would define 'functional genomics' as simply doing the biology and genetic experiments on the organism. The computer can really help little once this process is undertaken. And the only way that this enterprise differs from what we biologists have been doing for decades is the matter of scale. How many genes can we study and learn about their functions and how quickly can we do so? The new databases are a tool to this end and are a sort of 'ground zero' or provide a minimum amount of required data needed before proceeding to make mutations and studying the organism (Fig. 7). There are no biological answers '*in silico*', there is just data and methods to organize data.

Phenotypic analysis in drug target discovery

Functional genomics was formulated by drug industries (biotechnology and pharmaceutical) and academic circles which believe that there are real possibilities that the next generation of drugs will come directly from analysis of the human genome (Fields, 1997). Obviously drugs work in general by binding and modulating certain proteins in the body, and those proteins are encoded by genes. So logic suggests that if a protein whose function is in some aspect of physiology that is perturbed by disease, then finding a molecule that targets that protein is the best way to discover new drugs. This logic is so compelling that billions have been spent on genomics before one drug discovered by this process –starting with a protein product of a newly sequenced gene and moving all the way through candidate molecules to clinical trials– has been brought to market.

But how might this logical process be implemented in the lab to provide new drugs for clinical trials? How can the various components of genomics and functional genomics be integrated to provide drug targets and small molecules? While no one can provide precise answers, we are certain that the genetic analysis of model organisms will be crucial. Simply making a mutation and observing the phenotypic consequence can provide invaluable information regarding the physiological role of a gene. And it is this information that directly impinges on determining whether a particular gene product can be a valid drug target.

An illustrative example is the mutation in the prostacyclin receptor generated in a line of mice by Murata *et al.* (1997). Prostacyclin is a member of a family of small lipids called the prostanoids which mediate localized responses of various types. Prostacyclin is involved in vasodilation, inflammation and pain sensation; its receptor is one of the large serpentine, or 7-transmembrane spanning (7 TM) family. One of the oldest and most common drugs in existence, aspirin (and aspirin-like drugs such as indomethacin) is known to affect prostanoid synthesis. Genetically perturbing a protein on this aspirin-targeted pathway results in mice whose phenotype resembles wild-type animals treated with the drug. Homozygotes have reduced responses to pain and inflammatory stimuli –essentially at levels of wild-type animals treated with indomethacin. The phenotype recapitulates the effect of a small molecule antagonist.

These experiments clearly operated in the reverse of the proposed functional genomics process. The drug is in hand and a mutation of a protein on the affected pathway would be expected to result in a disease-relevant phenotype. The lesson is valuable, however; it becomes important to ask which and how many proteins are present in the body that, when bound by a small molecule will, for example, reduce inflammation in general or in specific tissues? To answer this, it should be possible to use genetics to screen for an expected and defined phenotype in order to quickly define new drug target proteins. Clearly the prostacyclin receptor is such a target. It is a member of a large family of proteins, many of which are known drug targets. These G-protein coupled receptors (GPCRs, or 7 TMs) are the targets of about 20% of the top 100 pharmaceutical drugs (Campbell, 1996). Also, the prostacyclin receptor is involved in two important physiological processes –both of which can be usefully modulated to treat a variety of maladies.

Another indirect example may help illustrate how the functional genomics process of anonymous gene sequence to phenotype to drug might work. The recently discovered orexin neuropeptides regulate feeding behavior by acting on the hypothalamus (Sakurai *et al.*, 1998). They were discovered by their ability to bind to, and activate in a cell assay, orphan GPCRs. The neuropeptides stimulate feeding behavior in rats when administered directly to the brain through catheters. The orexin receptors are prime targets for which a small molecule development screen could find drugs that modulate appetite in humans. The means to this end in which research moved from an interesting gene sequence through a cell-based screen and an animal model phenotype will be generally applicable. (Even though the phenotype induced in the rat was not from a mutation, it is likely that perturbation of feeding behavior will result when the orexin receptors in mouse are mutated.) How this process might operate in practice remains to be seen; one possibility is presented below (see also Friedrich, 1996).

The starting line is coding sequence from transcripts. Bioinformatics, as discussed above, is most useful in categorizing anonymous gene sequences into families. For the purposes of drug discovery, the relevant gene families are few but large (for example, there are more than 1000 GPCRs). These sets of potential drug targets are flagged due to precedence: we know that proteins of similar structure and cellular location are bound by drugs and thereby have a therapeutic effect. Expression data provides more data to narrow the search. If our interest is diseases related to the immune system, candidates must at least be found in the thymus, spleen or bone marrow. If they are expressed in one of these tissues in addition to a variety of others, this is not a criterion for eliminating that gene from further consideration. Few genes are expressed only in the tissues where they function – there are no evolutionary constraints that make this the rule. Therefore, while exceptionally valuable, interpretation of expression data needs to be approached with caution. It may turn out that such data is useful as often for eliminating candidates that are not expressed in a relevant tissue as when choosing candidates based on where they are expressed.

With information from bioinformatics and mRNA expression, and perhaps from *in vitro* functional assays, one will be left with a large set of potentially valuable genes, including new drug targets. The best means to get at function to determine which will become the focus of small molecule screens is genetics (Fig. 7). As the examples above illustrate, the phenotype resulting from mutation of certain genes will be relevant to disease. It is these proteins that are identified by phenotypic analysis that are going to be the most valuable. It is easy to imagine how the prostacyclin receptor or the orexin receptor, where they not yet discovered, could be identified by genetic analysis. There are many such proteins waiting to be discovered. And the mouse, we have been arguing, will become the preeminent model organism given a sufficiently powerful genetic system. With gene traps throughout the genome creating a frozen library of pluripotent ES cells, we will have such a resource. An added advantage is that each trap possesses a sequence tag. We can then undertake this "sequence to drugs" discovery process with an organism that is a workable model of human disease, physiology and genetics.

Ultimately, it is not technological advances such as chip expression analysis, better bioinformatics algorithms and databases, or mouse gene trap mutation libraries that will deliver drugs. It is the intelligent mustering of these tools to do the same biology that has been the rule for decades, only doing it more efficiently, more quickly and more directed toward specific ends: discovering new drug targets.

Challenges

Investigators using gene trapping face a number of challenges. Trapping all genes will be a major obstacle and will likely require the development of a variety of vectors, but our current data suggests that gene trapping has a tremendous number of potential target genes. Maintaining germline transmission for such a large resource of ES cells will require strict quality control. It will also be essential to optimize vectors for the ability to abolish gene expression and it will be useful to develop vectors that allow tissue-specific or temporal gene mutation. While the challenges in creating such a gene trap library are tremendous, realizing the potential of this resource will provide even more challenges. The large numbers of mutations represented in a resource containing thousands of gene trap events will challenge investigators to wisely choose what mice to study and to devise the proper phenotypic screens to identify the consequences of the mutations. This will be fundamentally important for discovering new drug targets. The utilization of this gene trap library to study gene function along with devising and carrying out novel genetic screens to be used in combination with gene trapping should keep investigators busy for years.

References

ALLEN, N.D., CRAN, D.G., BARTON, S.C., HETTLE, S., REIK, W. and SURANI, M.A. (1988). Transgenes as probes for active chromosomal domains in mouse development. *Nature. 333:* 852-855.

BAKER, R.K., HAENDEL, M.A., SWANSON, B.J., SHAMBAUGH, J.C., MICALES, B.K. and LYONS, G.E. (1997). *In vitro* preselection of gene-trapped embryonic stem cell clones for characterizing novel developmentally regulated genes in the mouse. *Dev. Biol. 185:* 201-214.

BASSETT, D.E., JR., BASRAI, M.A., CONNELLY, C., HYLAND, K.M., KITAGAWA, K., MAYER, M.L., MORROW, D.M., PAGE, A.M., RESTO, V.A., SKIBBENS, R.V. and HIETER, P. (1996). Exploiting the complete yeast genome sequence. *Curr. Opin. Genet. Dev. 6:* 763-766.

BELLEN, H.J., O'KANE, C.J., WILSON, C., GROSSNIKLAUS, U., PEARSON, R.K. and GEHRING, W.J. (1989). P-element-mediated enhancer detection: a versatile method to study development in *Drosophila. Genes Dev. 3:* 1288-1300.

BIER, E., VAESSIN, H., SHEPHERD, S., LEE, K., MCCALL, K., BARBEL, S., ACKERMAN, L., CARRETTO, R., UEMURA, T., GRELL, E., JAN, L.Y. and JAN, Y.N. (1989). Searching for pattern and mutation in the *Drosophila* genome with a *P-lacZ* vector. *Genes Dev. 3:* 1273-1287.

BRADLEY, A. (1993). Site-directed mutagenesis in the mouse. *Recent Prog. Horm. Res. 48:* 237-251.

BROWN, S.D.M. and PETERS, J. (1996). Combining mutagenesis and genomics in the mouse-closing the phenotype gap. *Trends Genet. 12:* 433-435.

CAMPBELL, P. (1996). Intelligent drug design. *Nature 384 (Suppl.):* 1-5.

CARLSON, J. (1993). Molecular genetics of *Drosophila* olfaction. *Ciba Found. Symp. 179:* 162-166.

CHANG, W., HUBBARD, C., FRIEDEL, C. and RULEY, E. (1993). Enrichment of insertional mutants following retrovirus gene trap selection. *Virology 193:* 737-747.

CHEN, Z., FRIEDRICH, G.A. and SORIANO, P. (1994). Transcriptional enhancer factor 1 disruption by a retroviral gene trap leads to heart defects and embryonic lethality in mice. *Genes Dev.* 2293-2301.

CHOWDHURY, K., BONALDO, P., TORRES, M., STOYKOVA, A. and GRUSS, P. (1997). Evidence for the stochastic integration of gene trap vectors into the mouse germline. *Nucleic Acids Res. 25:* 1531-1536.

DEGREGORI, J., RUSS, A., VON MELCHNER, H., RAYBURN, H., PRIYARANJAN, P., JENKINS, N.A., COPELAND, N.G. and RULEY, H.E. (1994). A murine homolog of the yeast RNA1 gene is required for postimplantation development. *Genes Dev. 8:* 265-276.

DENG, J.M. and BEHRINGER, R.R. (1995). An insertional mutation in the BTF3 transcription factor gene leads to an early postimplantation lethality in mice. *Transgenic Res. 4:* 264-269.

DOOLITTLE, R.F. (1994). Protein sequence comparisons: searching databases and aligning sequences. *Curr. Biol. 5:* 24-28.

EVANS, M.J., CARLTON, M.B.L. and RUSS, A.P. (1997). Gene trapping and functional genomics. *Trends Genet. 13:* 370-374.

FAZELI, A., DICKINSON, S.L., HERMISTON, M.L., TIGHE, R.V., STEEN, R.G., SMALL, C.G., STOECKLI, E.T., KEINO-MASU, K., MASU, M., RAYBURN, H., SIMONS, J., BRONSON, R.T., GORDON, J.I., TESSIER-LAVIGNE, M. and WEINBERG, R.A. (1997). Phenotype of mice lacking functional deleted in colorectal cancer *(DCC)* gene. *Nature 386:* 796-804.

FIELDS, C. (1992). Data exchange and inter-database communication in genome projects. *Trends Biotechnol. 10:* 58-61.

FIELDS, S. (1997). The future is function. *Nature Genet. 15:* 325-327.

FIELDS, S. and STERNGLANZ, R. (1994). The two-hybrid system: an assay for protein-protein interactions. *Trends Genet. 10:* 286-292.

FORRESTER, L.M., NAGY, A., SAM, M., WATT, A., STEVENSON, L., BERNSTEIN, A., JOYNER, A.L. and WURST, W. (1996). An induction gene trap screen in embryonic stem cells: identification of genes that respond to retinoic acid *in vitro. Proc. Natl. Acad. Sci. USA 93:* 1677-1682.

FRIEDRICH, G.A. (1996). Moving beyond the genome projects. *Nature Biotechnol. 14:* 1234-1237.

FRIEDRICH, G. and SORIANO, P. (1991). Promoter traps in embryonic stem cells: a genetic screen to identify and mutate developmental genes in mice. *Genes Dev. 5:* 1513-1523.

FRIEDRICH, G., HILDEBRAND, J.D. and SORIANO, P. (1997). The secretory protein Sec8 is required for paraxial mesoderm formation in the mouse. *Dev. Biol. 192:* 364-374.

FROMONT-RACINE, M., RAIN, J.-C. and LEGRAIN, P. (1997). Toward a functional analysis of the yeast genome through exhaustive two-hybrid screens. *Nature Genet. 16:* 277-282.

GOFFEAU, A., BARRELL, B.G., BUSSEY, H., DAVIS, R.W., DUJON, B., FELDMANN, H., GALIBERT, F., HOHEISEL, J.D., JACQ, C., JOHNSTON, M., LOUIS, E.J., MEWES, H.W., MURAKAMI, Y., PHILIPPSEN, P., TETTELIN, H. and OLIVER, S.G. (1996). Life with 6000 genes. *Science 274:* 563-567.

GOGOS, J.A., LOWRY, W. and KARAYIORGOU, M. (1997). Selection for retroviral insertions into regulated genes. *J. Virol. 71:* 1644-1650.

GOGOS, J.A., THOMPSON, R., LOWRY, W., SLOANE, B.F., WEINTRAUB, H. and HORWITZ, M. (1996). Gene trapping in differentiating cell lines: regulation of the lysosomal protease cathepsin-B in skeletal myoblast growth and fusion. *J. Cell Biol. 134:* 837-847.

GOSSLER, A. and ZACHGO, J. (1993). Gene targeting: A practical approach. In *Gene trapping strategies: screens for developmentally regulated genes and insertional mutagenesis* (Ed. A.L. Joyner). Vol. Oxford Univ. Press, New York, pp. 181-227.

GOSSLER, A., JOYNER, A., ROSSANT, J. and SKARNES, W.C. (1989). Mouse embryonic stem cells and reporter constructs to detect developmentally regulated genes. *Science 244:* 463-465.

GUERIOS-FILHO, F.J. and BEVERLEY, S.M. (1997). Trans-kingdom transposition of the *Drosophila* element mariner within the protozoan *Leishmania. Science 276:* 1716-1719.

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.

HUBBARD, S.C., WALLS, L., RULEY, H.E. and MUCHMORE, E.A. (1994). Generation of Chinese hamster ovary cell glycosylation mutants by retroviral insertional mutagenesis. *J. Biol. Chem. 269:* 3717-3724.

IMAI, Y., SUZUKI, Y., MATSUI, T., TOHYAMA, M., WANAKA, A. and TAKAGI, T. (1995). Cloning of a retinoic acid-induced gene, GT1, in the embryonal carcinoma cell line P19: neuron-specific expression in the mouse brain. *Mol. Brain Res. 31:* 1-9.

IVICS, Z., HACKETT, P.B., PLASTERK, R.H. and IZSVAK, Z. (1997). Molecular reconstruction of *Sleeping Beauty*, a *Tc1*-like transposon from fish, and its transposition in human cells. *Cell 91:* 501-510.

JONSSON, J.-I., WU, Q., NILSSON, K. and PHILLIPS, R.A. (1996). Use of a promoter-trap retrovirus to identify and isolate genes involved in differentiation of a myeloid progenitor cell line in vitro. *Blood 87:* 1771-1779.

KERR, W. and HERZENBERG, L.A. (1991). Gene-search viruses and FACS-Gal permit the detection, isolation, and characterization of mammalian cells with *in situ* fusions between cellular genes and *Escherichia coli* lacZ. In *Methods: A Companion to Methods in Enzymology* (Ed. Vol. 2.). pp. 261-271.

KOTHARY, R., CLAPOFF, S., BROWN, A., CAMPBELL, R., PETERSON, A. and ROSSANT, J. (1988). A transgene containing *lacZ* inserted into the dystonia locus is expressed in neural tube. *Nature 335:* 435-437.

LI, L. and COHEN, S.N. (1996). *tsg101*: A novel tumor susceptibility gene isolated by controlled homozygous functional knockout of allelic loci in mammalian cells. *Cell 85:* 319-329.

LIANG, P. and PARDEE, A.B. (1995). Recent advances in differential display. *Curr. Opin. Immunol. 7:* 274-280.

MORAN, J.V., HOLMES, S.E., NAAS, T.P., DEBERARDINIS, R.J., BOEKE, J.D. and KAZAZIAN, H.H., Jr. (1996). High frequency retrotransposition in cultured mammalian cells. *Cell 87:* 917-927.

MURATA, T., USHIKUBI, F., MATSUOKA, T., HIRATA, M., YAMASAKI, A., SUGIOMOTO, Y., ICHIKAWA, A., AZE, Y., TANAKA, T., YOSHIDA, N., UENO, A., OH-ISHI, S. and NARUMIYA, S. (1997). Altered pain perception and inflammatory response in mice lacking prostacyclin receptor. *Nature 388:* 678-682.

NIWA, H., ARAKI, K., KIMURA, S., TANIGUCHI, S., WAKASUGI, S. and YAMAMURA, K. (1993). An Efficient Gene-Trap Method Using Poly a Trap Vectors and Characterization of Gene-Trap Events. *J. Biochem. 113:* 343 - 349.

O'KANE, C.J. and GEHRING, W.J. (1987). Detection *in situ* of genomic regulatory elements in *Drosophila. Proc. Natl. Acad. Sci. USA 84:* 9123-9127.

RAMIREZ-SOLIS, R., DAVIS, A.C. and BRADLEY, A. (1993). Gene targeting in embryonic stem cells. In *Guide to techniques in mouse development* (Ed. P.M. Wassarman and M.L. DePamphilis). Methods in Enzymology, Vol. 225. Academic Press, Inc., New York, pp. 855-878.

RIJKERS, T. and RUTHER, U. (1996). Sequence and expression pattern of an evolutionarily conserved transcript identified by gene trapping. *Biochim. Biophys. Acta 1307:* 294-300.

ROSS, A.J., WAYMIRE, K.G., MOSS, J.E., PARLOW, A.F., SKINNER, M.K., RUSSELL, L.D. and MACGREGOR, G.R. (1998). Testicular degeneration in *Bclw*-deficient mice. *Nature Genet. 18:* 251-256.

RUSS, A.P., FRIEDEL, C., BALLAS, K., KALINA, U., ZAHN, D., STREBHARDT, K. and VON MELCHNER, H. (1996). Identification of genes induced by factor deprivation in hematopoietic cells undergoing apoptosis using gene-trap mutagenesis and site-specific recombination. *Proc. Natl. Acad. Sci. USA 93:* 15279-15284.

SAKURAI, T., AMEMIYA, A., ISHII, M., MATSUZAKI, I., CHEMELLI, R.M., TANAKA, H., WILLIAMS, S.C., RICHARDSON, J.A., KOZLOWSKI, G.P., WILSON, S., ARCH, J.R.S., BUCKINGHAM, R.E., HAYNES, A.C., CARR, S.A., ANNAN, R.S., MCNULTY, D.E., LIU, W.-S., TERRETT, J.A., ELSHOURBAGY, N.A., BERGSMA, D.J. and YANAGISAWA, M. (1998). Orexins and Orexin receptors: a family of hypothalmic neuropeptides and G protein-coupled receptors that regulate feeding behavior. *Cell 92:* 573-585.

SASSAMAN, D.M., DOMBROSKI, B.A., MORAN, J.V., KIMBERLAND, M.L., NAAS, T.P., DEBERARDINIS, R.J., GABRIEL, A., SWERGOLD, G.D. and KAZAZIAN, H.H., Jr. (1997). Many humans L1 elements are capable of retrotransposition. *Nature Genet. 16:* 37-43.

SCHENA, M. (1996). Genome analysis with gene expression microarrays. *Bioessays 18:* 427-431.

SCHUSTER-GOSSLER, K., SIMON-CHAZOTTES, D., GUENET, J.-L., ZACHGO, J. and GOSSLER, A. (1996). *$Gtl2^{lacz}$*, an insertional mutation on mouse Chromosome 12 with parental origin-dependent phenotype. *Mammal. Genome 7:* 20-24.

SENTRY, J.W., GOODWIN, S.F., MILLIGAN, C.D., DUNCANSON, A., YANG, M. and KAISER, K. (1994). Reverse genetics of *Drosophila* brain structure and function. *Prog. Neurobiol. 42:* 299-308.

SERAFINI, T., COLAMARINO, S.A., LEONARDO, E.D., WANG, H., BEDDINGTON, R., SKARNES, W.C. and TESSIER-LAVIGNE, M. (1996). Netrin-1 is required for commissural axon guidance in the developing vertebrate nervous system. *Cell 87:* 1001-1014.

SHEVENKO, A., JENSEN, O.N., PODTELEJNIKOV, A.V., SAGLIOCCO, F., WILM, M., VORM, O., MORTENSEN, P., SHEVCHENKO, A., BOUCHERIE, H. and MANN, M. (1996). Linking genome and proteome by mass spectrometry: large-scale identification of yeast proteins from two dimensional gels. *Proc. Natl. Acad. Sci. USA 93:* 14440-14445.

SHOEMAKER, D.D., LASHKARI, D.A., MORRIS, D., MITTMANN, M. and DAVIS, R.W. (1996). Quantitative phenotypic analysis of yeast deletion mutants using a highly parallel molecular bar-coding strategy. *Nature Genet. 14:* 450-456.

SKARNES, E., AUERBACK, B.A. and JOYNER, A.L. (1992). A gene trap approach in mouse embryonic stem cells: the *lacZ* reporter is activated by splicing, reflects endogenous gene expression, and is mutagenic in mice. *Genes Dev. 6:* 903-918.

SKARNES, W., MOSS, J., HURTLEY, S. and BEDDINGTON, R. (1995). Capturing genes encoding membrane and secreted proteins important for mouse development. *Proc. Natl. Acad. Sci. USA 92:* 6592 - 6596.

TORRES, M., STOYKOVA, A., HUBER, O., CHOWDHURY, K., BONALDO, P., MANSOURI, A., BUTZ, R. and GRUSS, P. (1997). An *alpha-E-catenin* gene trap mutation defines its function in preimplantation development. *Proc. Natl. Acad. Sci. USA 94:* 901-906.

TOWNLEY, D.J., AVERY, B.J., ROSEN, B. and SKARNES, W.C. (1997). Rapid sequence analysis of gene trap integrations to generate a resource of insertional mutations in mice. *Genet. Res. 7:* 293-298.

VELCULESCU, V.E., ZHANG, L., VOGELSTEIN, B. and KINZLER, K.W. (1995). Serial analysis of gene expression. *Science 270:* 484-487.

WILSON, C., PEARSON, R.K., BELLEN, H.J., O'KANE, C.J., GROSSNIKLAUS, U. and GEHRING, W.J. (1989). P-element-mediated enhancer detection: an efficient method for isolating and characterizing developmentally regulated genes in *Drosophila. Genes Dev. 3:* 1301-1313.

WISE, M.J., LITTLEJOHN, T.G. and HUMPHERY-SMITH, I. (1997). Peptide-mass fingerprinting and the ideal covering set for protein characterisation. *Electrophoresis 18:* 1399-1409.

WITHERS-WARD, E.S., KITAMURA, Y., BARNES, J.P. and COFFIN, J.M. (1994). Distribution of targets for avian retrovirus DNA integration *in vivo. Genes Dev.* 1473-1487.

WURST, W., ROSSANT, J., PRIDEAUX, V., KOWNACKA, M., JOYNER, A.L., HILL, D. and GUILLEMOT, F. (1995). A large-scale gene-trap screen for insertional mutations in developmentally regulated genes in mice. *Genetics 139:* 889-899.

YOSHIDA, M., YAGI, T., FURUTA, Y., TAKAYANAGI, K., KOMINAMI, R., TAKEDA, Y., TOKUNAGA, T., CHIBA, J., IKAWA, Y. and AIZAWA, S. (1995). A new strategy of gene trapping in ES cells using 3'RACE. *Transgenic Res. 4:* 277-287.

YOU, Y., BERGSTROM, R., KLEMM, M., LEDERMAN, B., NELSON, H., TICKNOR, C., JAENISCH, R. and SCHIMENTI, J. (1997). Chromosomal deletion complexes in mice by radiation of embryonic stem cells. *Nature Genet.* 285-288.

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.

ZAMBROWICZ, B.P., IMAMOTO, A., FIERING, S., HERZENBERG, L.A., KERR, W.G. and SORIANO, P. (1997). Disruption of overlapping transcripts in the ROSA b-geo 26 gene trap strain leads to widespread expression of β-galactosidase in mouse embryos and hematopioetic cells. *Proc. Natl. Acad. Sci. USA 94:* 3789-3794.

ZWAAL, R.R., BROEKS, A., VAN MEURS, J., GROENEN, J.T.M. and PLASTERK, R.H.A. (1993). Target-selected gene inactivation in Caenorhabditis elegans by using a frozen transposen insertion mutant bank. *Proc. Natl. Acad. Sci. USA 90:* 7431-7435.