

RESOURCE ARTICLE

Mutant non-coding RNA resource in mouse embryonic stem cells

Jens Hansen^{1,*‡}, Harald von Melchner² and Wolfgang Wurst^{1,3,4,5,‡}

ABSTRACT

Gene trapping is a high-throughput approach that has been used to introduce insertional mutations into the genome of mouse embryonic stem (ES) cells. It is performed with generic gene trap vectors that simultaneously mutate and report the expression of the endogenous gene at the site of insertion and provide a DNA sequence tag for the rapid identification of the disrupted gene. Large-scale international efforts assembled a gene trap library of 566,554 ES cell lines with single gene trap integrations distributed throughout the genome. Here, we re-investigated this unique library and identified mutations in 2202 non-coding RNA (ncRNA) genes, in addition to mutations in 12,078 distinct protein-coding genes. Moreover, we found certain types of gene trap vectors preferentially integrating into genes expressing specific long non-coding RNA (lncRNA) biotypes. Together with all other gene-trapped ES cell lines, lncRNA gene-trapped ES cell lines are readily available for functional *in vitro* and *in vivo* studies.

KEY WORDS: Long non-coding RNA, Mutation, Gene trap, Mouse, *Mus musculus*, Mutagenesis

INTRODUCTION

The comprehensive annotation of the mouse genome has identified over 21,000 protein-coding genes (PCGs), along with more than 15,000 non-coding RNA (ncRNA) genes. To address their function, platforms for large-scale mutagenesis in embryonic stem (ES) cells have been implemented, with the ultimate goal to convert all mutant ES cell lines into mice for subsequent phenotyping. Using high-throughput gene trapping and targeting, the International Knockout Mouse (IKMC) and International Mouse Phenotyping (IMPC) consortia have created an unprecedented resource comprising mutant ES cell lines harboring mutations in ~18,500 unique PCGs. Of these, over 5000 have been converted into mice and subjected to high-throughput phenotyping

(www.mousephenotype.org) (Bradley et al., 2012; Collins et al., 2007; Kaloff et al., 2016; Lloyd et al., 2020; Rosen et al., 2015; Skarnes et al., 2011, 2004). Moreover, genes thus far inaccessible by targeting or trapping are now being addressed individually using CRISPR/Cas9 technology (Brandl et al., 2015; Wefers et al., 2017).

Unlike gene targeting, gene trap strategies rely on generic vectors capable of simultaneously mutating and reporting gene expression at the insertion site as well as providing a sequence tag for seamless gene identification (Friedel and Soriano, 2010). Multiple gene trap vectors have been developed and used in high-throughput screens to generate large libraries of mutant ES cell lines. The vast majority of the ES cell lines assembled by the international consortia were produced with promoter trap vectors, most of which comprise a promoterless reporter and/or selectable marker gene flanked by a 5' splice acceptor (SA) and a 3' polyadenylation (pA) sequence (Table S1). Their integrations into an intron of an expressed gene elicits splicing of upstream exons to the reporter gene, resulting in a fusion transcript terminating at the gene trap's pA site and thus truncating the endogenous transcript (Friedrich and Soriano, 1991; Gossler et al., 1989; Skarnes et al., 1992; Wiles et al., 2000; Wurst et al., 1995; Zambrowicz et al., 2003, 1998). Variants thereof either contain type II transmembrane domains fused to the reporter for trapping secretory pathway genes (De-Zolt et al., 2006) or lack a splice acceptor for trapping exons, in which case the reporter is translated from in-frame read-through fusion transcripts (Hicks et al., 1997; von Melchner et al., 1992). Although in theory the latter vector (also referred to as 'exon traps') should be activated exclusively from in-frame integrations into exons, in practice a large proportion of these vectors are activated from integrations into introns by adjacent cryptic splice sites (Osipovich et al., 2004). A significantly lower number of ES cell lines were produced with vectors referred to as 'polyA traps', in which the reporter genes are flanked by a 5' constitutive promoter and a 3' splice donor site, enabling downstream splicing. PolyA trap integrations into introns are expressed from their exogenous promoter and, therefore, unlike most other gene trap vectors, are activated independently of target gene expression (Ishida and Leder, 1999; Niwa et al., 1993; Salminen et al., 1998; Stanford et al., 2006; Yoshida et al., 1995). In a further application, ES cell lines were also generated with gene trap vectors containing both promoter and polyA trap modules, although selection overwhelmingly relied on the promoter trap cassettes (Zambrowicz et al., 2003). Finally, to enable conditional mutagenesis, a significant proportion of ES cell lines were produced with promoter traps equipped with site-specific recombination systems (Schnütgen, 2006; Schnütgen et al., 2005). Overall 566,554 gene-trapped ES cell lines have been produced by the IKMC and can be accessed via the Mouse Genome Informatics (MGI) website (www.informatics.jax.org) (Ringwald et al., 2011). The database covers gene trap integrations into protein-coding and non-coding genes, including long and small non-coding RNA genes.

Long non-coding RNAs (lncRNAs) are defined by a gene length greater than 200 nucleotides, of which 9072 have been annotated in the Ensembl 83 (genome build GRCm38) database. Based on their

¹Institute of Developmental Genetics, Helmholtz Zentrum München GmbH, German Research Center for Environmental Health, Ingolstädter Landstr. 1, D-85764 Neuherberg, Germany. ²Department of Molecular Hematology, University Hospital Frankfurt, Goethe University, D-60590 Frankfurt am Main, Germany.

³Technische Universität München-Weihenstephan, c/o Helmholtz Zentrum München, Ingolstädter Landstr. 1, D-85764 Neuherberg, Germany. ⁴German Center for Neurodegenerative Diseases (DZNE), Site Munich, Feodor-Lynen-Str. 17, D-81377 Munich, Germany. ⁵Munich Cluster for Systems Neurology (SyNergy), Adolf-Butenandt-Institut, Ludwig-Maximilians-Universität München, Feodor-Lynen-Str. 17, D-81377 München, Germany.

*Present address: Institute of Metabolism and Cell Death, Helmholtz Zentrum München GmbH, German Research Center for Environmental Health, Ingolstädter Landstr. 1, D-85764 Neuherberg, Germany.

‡Authors for correspondence (hansen@helmholtz-muenchen.de; wurst@helmholtz-muenchen.de)

ORCID iD: J.H., 0000-0001-7520-1577; W.W., 0000-0003-4422-7410

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0>), which permits unrestricted use, distribution and reproduction in any medium provided that the original work is properly attributed.

Handling Editor: Monica J. Justice
Received 12 October 2020; Accepted 14 December 2020

position relative to PCGs, lncRNA genes were subdivided by the GENCODE consortium into five major classes: (1) long intergenic non-coding RNAs (lincRNAs) located between two protein-coding genes ($n=3579$); (2) antisense lncRNAs transcribed from the opposite strand of coding genes ($n=2189$); (3) ‘sense overlapping’ lncRNAs transcribed from the same strand of protein coding genes ($n=23$ genes); (4) ‘sense intronic’ lncRNAs transcribed from the introns of coding genes ($n=253$); and (5) ‘bidirectional promoter’ lncRNAs transcribed from the opposite strand within the promoter region of a protein-coding gene ($n=12$) (Frankish et al., 2019; Harrow et al., 2012). In addition, several lncRNA genes of numerically minor significance are distributed between the following biotypes (1) ‘processed transcript’ biotype, defined by noncoding transcripts without an open reading frame, (2) ‘3’ overlapping’, defined as short non-coding transcripts transcribed from the 3’UTR, (3) ‘macro’, defined by unspliced lncRNA of several kb in size; and (4) ‘to-be-experimentally-confirmed’ (TEC), defined by non-spliced polyadenylated transcripts with an open reading frame, which, pending further experimental validation, presumably encode novel proteins (Frankish et al., 2019; Harrow et al., 2012).

As key regulators of global gene expression, lncRNAs are involved in the regulation of nearly all fundamental biological processes, including development, cell cycle, differentiation, pluripotency, apoptosis, autophagy and cell migration (Fritah et al., 2014). Hence, it is not surprising that deregulation of lncRNA expression can lead to a wide spectrum of diseases (Rinn and Chang, 2012). However, only a minority of lncRNAs have been functionally validated thus far in tissue culture experiments and knockout mice (Bond et al., 2009; Gomez et al., 2013; Grote and Herrmann, 2015; Li et al., 2013; Liu et al., 2014; Nakagawa et al., 2014; Oliver et al., 2015; Sauvageau et al., 2013; Zhang et al., 2013). Given their biological significance, a large-scale analysis of individual lncRNA function(s) seems highly desirable. To facilitate this endeavor, we re-analyzed the existing gene trap libraries and identified 31,069 ES cell lines with gene trap insertions in 2202 unique ncRNA genes (Tables S4 and S5). This freely available resource should significantly support the functional lncRNA annotation effort.

RESULTS

The international gene trap resource

The MGI web portal provides the largest data set of gene trap sequence tags (GTSTs) from mutant murine ES cells generated worldwide by the consortia, institutions and corporations listed with their respective contributions in Table 1. MGI periodically updates vector integration sites by mapping existing GTSTs to the latest mouse genome sequence build (Ringwald et al., 2011). Presently, the database contains 854,155 GTSTs, of which 566,554 are unique. Systematic in-depth analysis of this database revealed 339,779 GTSTs (60%) corresponding to annotated genes and 226,773 (40%) to intergenic regions. For easy accessibility for the user, gene trap clones for a specific gene can be found in the MGI web portal by gene symbol or identifier. All trapped alleles are listed together with information about the vector, the insertion point, the sequence tags and the available mouse lines. Alternatively, a user can search a specified genomic region for gene trap integrations by using the MGI genome browser displaying the gene trap tracks (see tab ‘Search’ and follow the link ‘Mouse Genome Browsers’).

Distribution of gene trap integrations between major gene biotypes

According to their predicted function, the GENCODE consortium (www.genecodegenes.org) subdivides genes into PCGs, lncRNA genes, short non-coding RNA (sncRNA) genes and pseudogenes. Based on this classification, we identified 12,078 (82.1%) of the gene trap integrations in unique PCGs, 2060 (14.0%) in lncRNA genes, 142 (1.0%) in sncRNA genes and 426 (2.9%) in pseudogenes (Table 2). Overall, this corresponds to 55.1% of annotated PCGs and 22.7% of annotated lncRNA genes (Table 2; Tables S4 and S5). Gene trap integrations were significantly enriched in multiple-exon PCGs and processed lncRNA genes consistent with the vast majority of gene trap vectors, for which activation is based on upstream splicing (Tables S2 and S3). Regarding the position of insertion sites relative to transcription start sites, the majority of vectors with SA sites preferred the 5’ ends of both PCGs and lncRNA genes because the larger the 5’ sequence appended to the reporter the less likely the latter will maintain its function. By contrast, polyA trap vectors

Table 1. The international gene trap resource

Project	Sequence tags	ES cell clones	Trapped genes	Unique genes trapped	URL
OmniBankII	465,736	287,701	133,093	10,176	www.tigm.org
OmniBank	198,901	166,637	112,003	10,693	www.lexpharma.com
GGTC	50,854	31,791	31,984	5131	www.genetrap.de
EUCOMM	68,592	26,917	22,893	4718	www.mousephenotype.org
CMHD	24,321	19,859	9801	3856	www.cmhd.ca/genetrap
BayGenomics	14,179	11,547	11,293	3838	www.mmrrc.org
ESDB	13,839	9349	7554	1907	www.cmmr.ca
SIGTR	11,885	8019	7656	3130	www.mmrrc.org
FHCRC	1627	1394	1013	776	labs.icahn.mssm.edu/sorianolab
HER	1661	1389	640	432	www.vanderbilt.edu
EGTC	993	818	784	588	egtc.jp
NAIST	758	625	570	451	bsw3.naist.jp/NAISTrap
TIGEM	809	508	495	347	genetrap.tigem.it
Total	854,155	566,554	339,779	14,712*	

Shown are the number of gene trap sequence tags, embryonic stem (ES) cell clones, trapped mouse genes and unique genes trapped in the individual projects. OmniBankII, Texas A & M Institute for Genomic Medicine (TIGM), USA (Hansen et al., 2008); OmniBank, Lexicon Pharmaceuticals, USA (Zambrowicz et al., 2003); German Gene Trap Consortium (GGTC), Germany (Hansen et al., 2003); European Conditional Mouse Mutagenesis (EUCOMM), EU (Friedel et al., 2007); Centre for Modelling Human Disease (CMHD), Canada (To et al., 2004); BayGenomics, USA (Stryke et al., 2003); Embryonic Stem Cell Database (ESDB), Canada (Hicks et al., 1997); Sanger Institute Gene Trap Resource (SIGTR), UK (Lloyd, 2003); Fred Hutchinson Cancer Research Center (FHCRC), Soriano Lab Gene Trap Resource, USA (Friedrich and Soriano, 1991); H. E. Ruley (HER), Vanderbilt University, USA (Osipovich et al., 2005); Exchangeable Gene Trap Clones (EGTC), Japan (Niwa et al., 1993); Nara Institute of Science and Technology (NAIST), Japan (Shigeoka et al., 2005); Telethon Institute of Genetics and Medicine (TIGEM), Italy (Cobellis et al., 2005). *Total number of unique genes trapped on the basis of the entire resource.

Table 2. Distribution of mouse genes and gene trap integrations among gene biotypes

Gene biotype	Annotated mouse genes	Genes trapped (%)	% of biotype trapped
Protein-coding	21,917	12,078 (82.1)	55.1
lncRNA	9072	2060 (14.0)	22.7
Pseudogene	9174	426 (2.9)	4.6
sncRNA	6084	142 (1.0)	2.2
Others	699	6 (0.04)	0.9
Total	46,946	14,712 (100)	–

lncRNA, long non-coding RNA; sncRNA, short non-coding RNA.

overwhelmingly select for integration into the 3' ends of both PCGs and lncRNA genes, as more upstream integrations are generally lost due to nonsense-mediated decay (NMD) (Shigeoka et al., 2005; Stanford et al., 2006) (Fig. 1).

Distribution of gene trap insertions between specific ncRNA biotypes

Seventy one percent of the trapped lncRNAs (1455 of 2060) belonged to the lincRNA (806) and antisense RNA (649) biotypes, which together are the most prevalent lncRNAs in the mouse genome (Table 3). Consistent with the general preference of gene traps to mutate larger, multiple-exon genes, only between 1% and 4% of sncRNAs were trapped primarily by vectors lacking SA sites (Table 3). Although in PCGs only ~0.1% to 0.8% of gene trap insertions occurred in non-spliced genes, insertions into non-spliced lncRNA genes occurred up to 100 times more frequently (1-13%), reflecting the much higher proportion of non-spliced genes among lncRNAs (Table 4; Table S2).

Regarding gene trap integrations into specific lncRNA biotypes, we found a significant relationship between vector type and lncRNA biotype. Fig. 2 shows that the retroviral promoter trap vectors VICTR74 and VICTR76 used for creating the OmniBankII library integrated with much higher frequency into lincRNA and TEC genes than any other similarly structured vectors. Although the reasons for this preference remain unknown, it is likely that the somewhat more sensitive ES cell culture and selection protocols employed for OmniBankII (Hansen et al., 2008) enabled a more efficient isolation of these rather weakly expressed genes (Derrien et al., 2012; Djebali et al., 2012). As TEC genes represent genomic regions presumably encoding novel proteins, the gene trap libraries provide a useful resource for characterizing novel PCGs. Unlike promoter traps, polyA trap vectors, which are activated

independently of gene expression, captured lncRNA genes at a much higher rate than any other vectors. For example, the polyA trap vectors GepNMDi3, Gen-SD5, pGTNMDf, pGTR1.3 and Gep-SD5 (To et al., 2004) were all found with high frequency in antisense and lincRNA genes, most of which are either weakly expressed or not expressed at all in ES cells (Ghosal et al., 2013; Jia et al., 2013; Loewer et al., 2010) (Fig. 2).

Gene trap activation mechanisms in ncRNA genes

Depending on the trapped lncRNA biotype, gene trap integrations were activated by different mechanisms. For example, intron integrations in multiple-exon lincRNAs such as growth arrest-specific transcript 5 (*Gas5*) were activated from the sense strand similar to the activations seen in PCGs (Fig. 3A). By contrast, integrations into the first intron of the *1110002L01Rik* antisense lncRNA, which overlaps the 3' end of the kinesin family member 3C (*Kif3c*) and the 5'UTR of the additional sex combs-like 2 (*Asxl2*) PCGs, was transcribed in antisense direction to *Asxl2* and *Kif3c* (Fig. 3B). Neither of the PCGs was physically affected by the integration, although mutation of the antisense *1110002L01Rik* transcript could, in principle, interfere with the expression of either gene. A promoter trap integration into the *D0830050J10Rik* bidirectional promoter lncRNA encoded from the opposite strand of the v-raf-leukemia viral oncogene 1 (*Raf1*) PCG was transcribed from the same bidirectional promoter (Fig. 3C), and an integration into the *Gm12971* sense intronic lncRNA was transcribed from its own promoter located in the 14th intron of the *Pum1* PCG (Fig. 3D). Fig. 3E shows an integration into a sense overlapping lncRNA exemplified by Sox1 overlapping transcript (*Sox1ot*), which hosts the SRY (sex determining region Y)-box 1 (*Sox1*) PCG in the first intron. In this arrangement, the fusion transcript initiating at the *Sox1ot* promoter terminates at the gene trap pA site residing in the seventh *Sox1ot* exon (Fig. 3E). Finally, Fig. 3F shows a polyA trap activation from an integration into the last intron of the *4932443L11Rik* processed transcript lncRNA gene by including the gene trap as a portable exon.

DISCUSSION

In this study, we re-analyzed a library of 566,554 mutant mouse ES cell lines produced in multiple large-scale gene trap mutagenesis projects. Although the library of mutant ES cell lines was originally produced to study the function of PCGs, the present analysis revealed that the library contains 31,069 ES cell lines with mutations in 2202 unique ncRNA genes, in addition to the ES cell lines with mutations

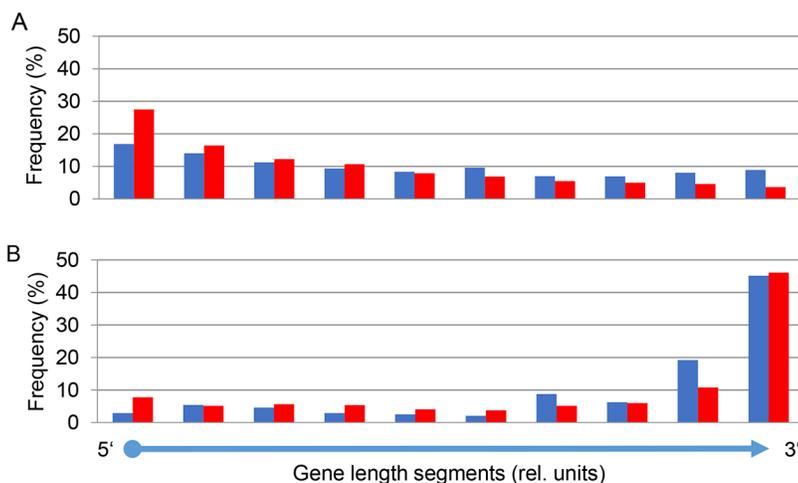


Fig. 1. Integrations of gene trap vectors into gene length segments of long non-coding RNA (lncRNA) genes and protein-coding genes (PCGs), measured as distance from the transcription start site. Each individual gene has been subdivided into ten equal-length segments. lncRNA genes are shown in blue, PCGs in red. (A) Promoter trap vectors with splice acceptor ($n=196,565$). (B) PolyA-trap vectors ($n=4220$). All integrations were analyzed with genomic PCR technologies (see Table S1 for composition of the vector classes). The significance of the integration pattern into gene length segments was studied with a G-test of goodness-of-fit (all P -values $<10^{-16}$). rel. units, relative units.

Table 3. Distribution of gene trap integrations in ncRNA biotypes

	Gene biotype	Mouse ncRNA genes (%)	Trapped loci (% of biotype)
lincRNA	lincRNA	3579 (39.4)	806 (22.5)
	Antisense	2189 (24.1)	649 (29.6)
	Processed transcript	749 (8.3)	338 (45.1)
	TEC	2264 (24.9)	218 (9.6)
	Sense intronic	253 (2.8)	35 (13.8)
	Sense overlapping	23 (0.3)	8 (34.8)
	Bidirectional promoter	12 (0.1)	6 (50.0)
	Other	3 (<0.1)	0 (0)
	Total	9072 (100)	2060 (22.7)
sncRNA	miRNA	2202 (36.1)	45 (2.0)
	snoRNA	1508 (24.8)	56 (3.7)
	snRNA	1382 (22.7)	19 (1.4)
	rRNA	354 (5.8)	5 (1.4)
	scaRNA	51 (0.8)	2 (3.9)
	Other	587 (9.8)	15 (2.6)
	Total	6084 (100)	142 (2.3)

lincRNA, long intergenic non-coding RNA; miRNA, microRNA; ncRNA, non-coding RNA; snRNA, small nuclear RNA; snoRNA, small nucleolar RNA; rRNA, ribosomal RNA; scaRNA, small Cajal body-specific RNA; TEC, to-be-experimentally-confirmed.

in 12,078 unique PCGs, and provides a useful resource for the functional characterization of many ncRNAs. The cell lines can be used *in vitro* to explore the role of ncRNAs in controlling ES cell pluripotency and differentiation (Chakraborty et al., 2012; Dinger et al., 2008; Fisher et al., 2017; Guttman et al., 2011; Sheik Mohamed et al., 2010) and can be readily converted into mutant mice for functional studies at organismal level. It is also worthwhile noting that all traps contain a LacZ reporter, easily enabling the *in vivo* analysis of lincRNA activity at cellular level, which is particularly useful in mutant mouse embryo phenotyping (Dickinson et al., 2016).

The GENCODE reference human and mouse genome annotation database contains three major functional categories of genes: PCGs, non-coding genes and pseudogenes (Harrow et al., 2012). Although gene trap insertions have been found in all these gene classes, a significant proportion involved intergenic regions (Table 2). Considering that 75% of the human genome is covered by primary transcripts and 62% by processed transcripts (Djebali et al., 2012), it is not surprising that 40% of all gene trap integrations were activated from non-annotated genomic regions, thus reflecting the high untapped potential of the gene trap approach for novel gene discovery. In line with this, the existing gene trap resource provides a unique means for resolving the biological significance of not yet annotated genes (Chi, 2016).

Table 4. Distribution of gene trap vector integrations into spliced and non-spliced (1-exon) genes

Classes of vector types	Gene biotype	Number of gene traps	Non-spliced genes (%)	Spliced genes (%)
Retroviral promoter trap vectors without splice acceptor	protein-coding	7973	0.1	99.9
	lincRNA	546	5.9	94.1
Retroviral promoter trap vectors with splice acceptor	protein-coding	164,145	0.2	99.8
	lincRNA	16,060	11.5	88.5
Plasmid promoter trap vectors	protein-coding	20,756	0.1	99.9
	lincRNA	952	13.4	86.6
Combined retroviral promoter and polyA trap vector	protein-coding	104,173	0.1	99.9
	lincRNA	11,648	2.3	97.7
Retroviral polyA trap vectors	protein-coding	2296	0.8	99.2
	lincRNA	837	1.2	98.8
Plasmid polyA trap vectors	protein-coding	5782	0.2	99.8
	lincRNA	557	5.0	95.0

Vectors are merged in classes based on their activation mechanism (see Table S1 for composition of the vector classes).

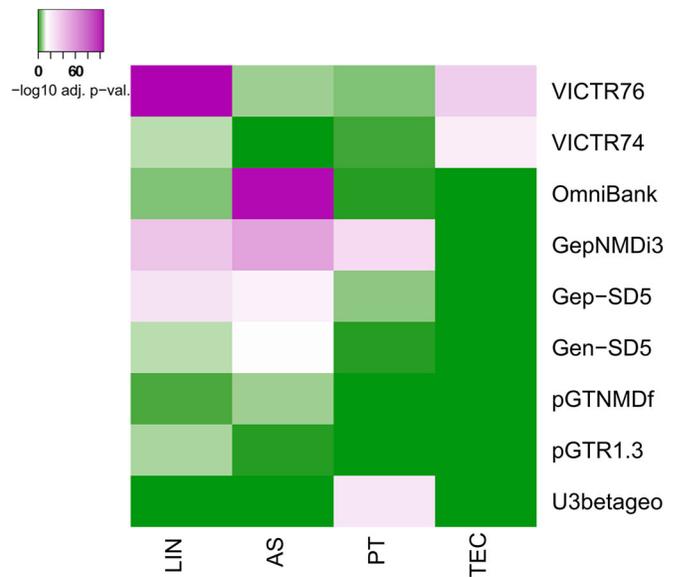


Fig. 2. Heatmap showing the distribution of individual gene trap vector integrations into different gene biotypes. Colors represent adjusted *P*-values ($-\log_{10}$) of Fisher's exact test for the significant enrichment of gene trap events in at least one gene biotype. *P*-values are corrected for multiple hypothesis testing using the procedure of Benjamini and Hochberg (1995) for false discovery rate estimation. Vectors belong to functional classes as listed in Table S1. AS, antisense; LIN, long intergenic non-coding RNA (lincRNA); PT, processed transcript; TEC, to-be-experimentally-confirmed.

Comparison of the integration targets of the different types of gene trap vectors revealed that, owing to fusion transcript size constrictions, promoter trap vectors preferentially integrated near the 5' ends of both PCGs and multiple-exon lincRNAs (Fig. 1A). However, polyA trap vectors overwhelmingly inserted near the 3' ends of PCGs and lincRNA genes to produce relatively short fusion transcripts unsusceptible to NMD (Fig. 1B) (Shigeoka et al., 2005; Stanford et al., 2006). In confirmation of previous observations suggesting that gene expression is an important trappability-defining factor for both promoterless and polyA trap vectors (Nord et al., 2007), we found that 90% of the lincRNA genes trapped with promoterless or polyA trap vectors are expressed in ES cells (data not shown).

As ~900 lincRNAs harbored multiple gene trap integrations at different locations, the ES cell library also provides allelic series for a multitude of lincRNA genes that are extremely useful for specifying distinct functional domains. For example, trapping different regions of the *Gas5* gene resulted in a series of *Gas5* truncation alleles affecting

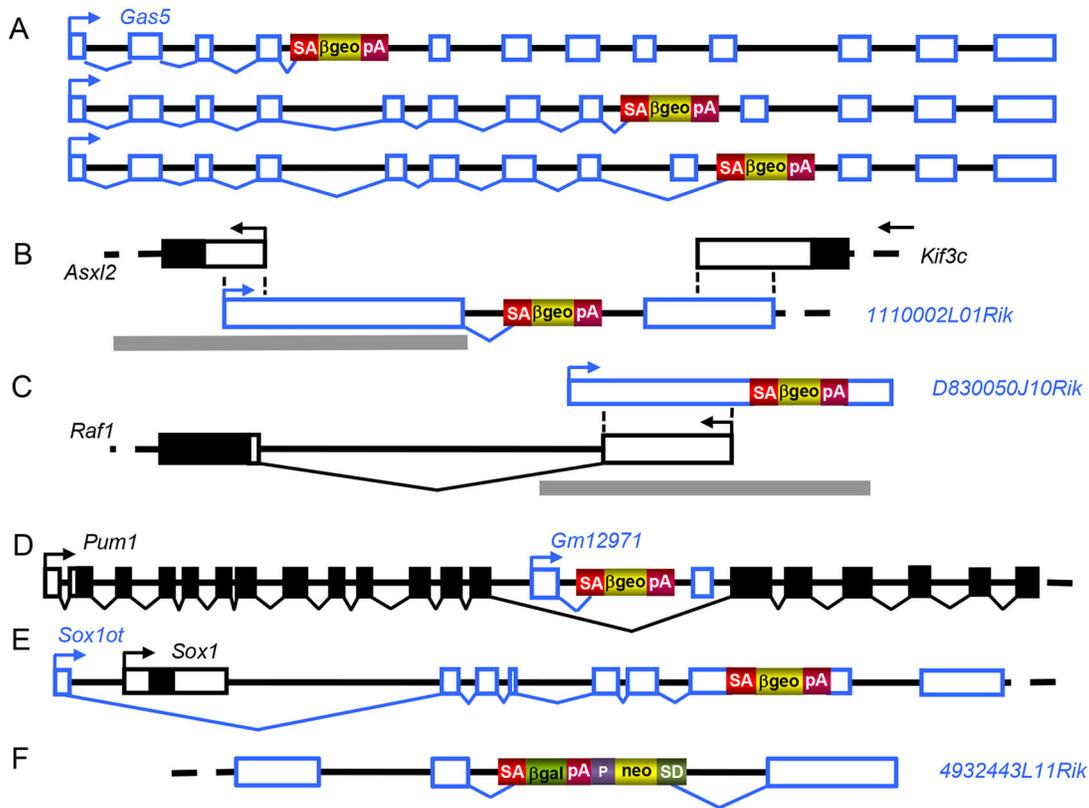


Fig. 3. Gene trap activation from integrations into various lncRNA biotypes. Exons of lncRNA genes are shown in blue, exons of PCGs in black. Filled bars represent coding sequence, open bars represent non-coding sequence. Transcription start sites and transcriptional orientation are indicated by arrows. Introns are shown as solid black lines, incomplete introns as dashed black lines. Promoters are indicated by thick gray lines. The elements of gene trap selection cassettes are shown in color. βgeo, β-galactosidase-neomycin phosphotransferase fusion gene; βgal, β-galactosidase gene; neo, neomycin phosphotransferase gene; P, promoter; pA, polyA; SA, splice acceptor; SD, splice donor. (A-F) Gene trap vector integrations are shown in the *Gas5* lincRNA gene (A), the antisense *1110002L01Rik* lncRNA gene (B), the bidirectional promoter *D830050J10Rik* lncRNA gene (C), the sense intronic *Gm12971* lncRNA gene (D), the sense overlapping *Sox1ot* lncRNA gene (E) and the processed transcript *4932443L11Rik* lncRNA gene (F). For further explanations, see text.

different protein functions (Fig. 3). *Gas5* is a tumor suppressor gene involved in several types of cancer and encodes several molecular functions over its length (Ma et al., 2016), including (1) a glucocorticoid response element (GRE) that competes with DNA for binding to the glucocorticoid receptor DNA-binding domain encoded by a stem-loop structure within the *Gas5* exon 12 (Kino et al., 2010); (2) a mir-21-binding function in exon 4 acting as a miRNA sponge regulating mir-21 levels, which are important in development, cancer, cardiovascular disease and inflammation (Zhang et al., 2013); and (3) an eIF4E-binding function, a key factor of the translation initiation complex (Hu et al., 2014). As shown in Fig. 3A, all these specific functions can be addressed by simply selecting the appropriate gene trap clones for *in vitro* and *in vivo* studies. In support of the *in vivo* value of the lncRNA gene trap lines, Miard et al. (2017) recently published a *Malat1* lncRNA knockout mouse produced with a VICTR74-expressing OmniBankII gene trap clone (IST14461G11). The *Malat1* lncRNA is overexpressed in many types of cancers, including hepatocellular carcinoma, and induces cell proliferation in several cell lines *in vitro*. Although its inactivation had no effect on liver carcinogenesis in mice treated with the genotoxic agent diethylnitrosamine (DEN), DEN-treated knockout mice developed a robust hypercholesterolemia, implicating *Malat1* in the regulation of cholesterol metabolism (Miard et al., 2017).

Finally, mutant alleles of lncRNAs containing a reporter gene can nowadays be established *de novo* using CRISPR/Cas9 knock-in strategies in mouse ES cells or mouse zygotes (Wefers et al., 2017;

Yao et al., 2018). However, notwithstanding the simplicity of the technology, the generation of allelic series including proper quality controls is still quite time consuming, requiring rigorous genotyping to exclude frequently occurring on-target mutations such as large deletions, insertions, inversions and translocation (Boroviak et al., 2017; Kosicki et al., 2018).

Although the functional characterization of all PCGs is well underway, currently comprising ~5000 already phenotyped mouse mutants, the next big challenge will be the functional dissection of all non-coding genes for which the existing mutant lncRNA ES cell library provides an unprecedented resource.

MATERIAL AND METHODS

Gene trap data

Gene trap sequence tags and their mouse genome coordinates were downloaded from the MGI web portal (www.informatics.jax.org; download on 19 January 2016). We filtered the data set with the objective to finally have one representative sequence tag with a high-quality alignment per vector integration, which was unequivocally mapped to the genome. First, we discarded sequence tags that did not result in a unique high-quality alignment. Insertions that resulted in multiple high-quality alignments and non-successful mappings were also discarded. In a final step all high-quality alignments with the mouse genome, which were indicated as ‘non-representative’, were filtered out.

Genome data

Software to identify the genomic locus for each gene trap vector insertion site was written in Perl 5.8.8 programming language and uses BioPerl

libraries. Genome features at each locus mutated by a gene trap vector integration event were retrieved from the Ensembl database (Yates et al., 2016) using the Ensembl application programming interface (Release 83; www.ensembl.org; genome build GRCh38). Gene models were categorized into biotypes according to the reference gene sets for the mouse published by the GENCODE consortium (version M8 August 2015) (Harrow et al., 2012).

Statistical testing

To study the significance of gene trap vector integration frequencies over gene length we used a G-test of goodness-of-fit. To determine whether gene trap insertions with a specific vector are over-represented in a given gene biotype, i.e. more integrations are present in genes of a specific gene biotype than expected by chance, a two-by-two contingency table was constructed and Fisher's exact test was performed. The procedure was repeated for each gene trap vector, and adjusted *P*-values were computed to control the false discovery rate (Benjamini and Hochberg, 1995). Categories with a *P*-value not greater than the corresponding adjusted *P*-value were considered significant. The false discovery rate constraint was set to 0.01. All statistical analyses were performed with R statistical software (R v3.3.1; www.r-project.org), using packages stats, RVAideMemoire, gplots and graphics.

Acknowledgements

We thank all our colleagues for generating these large-scale gene trap resources. The excellent technical assistance and system administration of Bernd Lentjes is gratefully acknowledged.

Competing interests

The authors declare no competing or financial interests.

Author contributions

Conceptualization: J.H., H.v.M., W.W.; Methodology: J.H.; Software: J.H.; Validation: J.H., H.v.M., W.W.; Formal analysis: J.H.; Investigation: J.H., H.v.M., W.W.; Resources: J.H., W.W.; Data curation: J.H.; Writing - original draft: J.H., W.W.; Writing - review & editing: J.H., H.v.M., W.W.; Visualization: J.H., H.v.M., W.W.; Supervision: W.W.; Project administration: W.W.; Funding acquisition: W.W.

Funding

This work was supported by National Genome Research Network Plus Project 'From Disease Genes to Protein Pathways' [FKZ 01GS0858] by Bundesministerium für Bildung, Wissenschaft und Forschung, and the projects 'EUCOMM' [FP6 grant number LSHM-CT-2005-01893] and 'I-DCC: The International Data Coordination Centre' [FP7-HEALTH-2007-2.1.2-6-223592], by the European Commission.

Supplementary information

Supplementary information available online at <https://dmm.biologists.org/lookup/doi/10.1242/dmm.047803.supplemental>

References

- Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B (Methodological)*, **57**, 289-300. doi:10.1111/j.2517-6161.1995.tb02031.x
- Bond, A. M., Vangompel, M. J. W., Sametsky, E. A., Clark, M. F., Savage, J. C., Disterhoft, J. F. and Kohtz, J. D. (2009). Balanced gene regulation by an embryonic brain ncRNA is critical for adult hippocampal GABA circuitry. *Nat. Neurosci.* **12**, 1020-1027. doi:10.1038/nn.2371
- Boroviak, K., Fu, B., Yang, F., Doe, B. and Bradley, A. (2017). Revealing hidden complexities of genomic rearrangements generated with Cas9. *Sci. Rep.* **7**, 12867. doi:10.1038/s41598-017-12740-6
- Bradley, A., Anastassiadis, K., Ayadi, A., Battey, J. F., Bell, C., Birling, M.-C., Bottomley, J., Brown, S. D., Bürger, A., Bult, C. J. et al. (2012). The mammalian gene function resource: the international knockout mouse consortium. *Mamm. Genome* **23**, 580-586. doi:10.1007/s00335-012-9422-2
- Brandl, C., Ortiz, O., Röttig, B., Wefers, B., Wurst, W. and Kühn, R. (2015). Creation of targeted genomic deletions using TALEN or CRISPR/Cas nuclease pairs in one-cell mouse embryos. *FEBS Open Bio* **5**, 26-35. doi:10.1016/j.fob.2014.11.009
- Chakraborty, D., Kappei, D., Theis, M., Nitzsche, A., Ding, L., Paszkowski-Rogacz, M., Surendranath, V., Berger, N., Schulz, H., Saar, K. et al. (2012). Combined RNAi and localization for functionally dissecting long noncoding RNAs. *Nat. Methods* **9**, 360-362. doi:10.1038/nmeth.1894
- Chi, K. R. (2016). The dark side of the human genome. *Nature* **538**, 275-277. doi:10.1038/538275a
- Cobellis, G., Nicolaus, G., Iovino, M., Romito, A., Marra, E., Barbarisi, M., Sardiello, M., Di Giorgio, F. P., Iovino, N., Zollo, M. et al. (2005). Tagging genes with cassette-exchange sites. *Nucleic Acids Res.* **33**, e44. doi:10.1093/nar/gni045
- Collins, F. S., Rossant, J. and Wurst, W. (2007). A mouse for all reasons. *Cell* **128**, 9-13. doi:10.1016/j.cell.2006.12.018
- De-Zolt, S., Schnütgen, F., Seisenberger, C., Hansen, J., Hollatz, M., Floss, T., Ruiz, P., Wurst, W. and von Melchner, H. (2006). High-throughput trapping of secretory pathway genes in mouse embryonic stem cells. *Nucleic Acids Res.* **34**, e25. doi:10.1093/nar/gnj026
- Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D. G. et al. (2012). The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res.* **22**, 1775-1789. doi:10.1101/gr.132159.111
- Dickinson, M. E., Flenniken, A. M., Ji, X., Teboul, L., Wong, M. D., White, J. K., Meehan, T. F., Weninger, W. J., Westerberg, H., Adissu, H. et al. (2016). High-throughput discovery of novel developmental phenotypes. *Nature* **537**, 508-514. doi:10.1038/nature19356
- Dinger, M. E., Amaral, P. P., Mercer, T. R., Pang, K. C., Bruce, S. J., Gardiner, B. B., Askarian-Amiri, M. E., Ru, K., Solda, G., Simons, C. et al. (2008). Long noncoding RNAs in mouse embryonic stem cell pluripotency and differentiation. *Genome Res.* **18**, 1433-1445. doi:10.1101/gr.078378.108
- Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F. et al. (2012). Landscape of transcription in human cells. *Nature* **489**, 101-108. doi:10.1038/nature11233
- Fisher, C. L., Marks, H., Cho, L. T.-Y., Andrews, R., Wormald, S., Carroll, T., Iyer, V., Tate, P., Rosen, B., Stunnenberg, H. G. et al. (2017). An efficient method for generation of bi-allelic null mutant mouse embryonic stem cells and its application for investigating epigenetic modifiers. *Nucleic Acids Res.* **45**, e174. doi:10.1093/nar/gkx811
- Frankish, A., Diekhans, M., Ferreira, A.-M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J. M., Sisu, C., Wright, J., Armstrong, J. et al. (2019). GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47**, D766-D773. doi:10.1093/nar/gky955
- Friedel, R. H. and Soriano, P. (2010). Gene trap mutagenesis in the mouse. *Methods Enzymol.* **477**, 243-269. doi:10.1016/S0076-6879(10)77013-0
- Friedel, R. H., Seisenberger, C., Kaloff, C. and Wurst, W. (2007). EUCOMM the European conditional mouse mutagenesis program. *Brief Funct. Genomic. Proteomic.* **6**, 180-185. doi:10.1093/bfpg/elm022
- 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. doi:10.1101/gad.5.9.1513
- Fritah, S., Niclou, S. P. and Azuaje, F. (2014). Databases for lncRNAs: a comparative evaluation of emerging tools. *RNA* **20**, 1655-1665. doi:10.1261/rna.044040.113
- Ghosal, S., Das, S. and Chakrabarti, J. (2013). Long noncoding RNAs: new players in the molecular mechanism for maintenance and differentiation of pluripotent stem cells. *Stem Cells Dev.* **22**, 2240-2253. doi:10.1089/scd.2013.0014
- Gomez, J. A., Wapinski, O. L., Yang, Y. W., Bureau, J.-F., Gopinath, S., Monack, D. M., Chang, H. Y., Brahic, M. and Kirkegaard, K. (2013). The NeST long ncRNA controls microbial susceptibility and epigenetic activation of the interferon- γ locus. *Cell* **152**, 743-754. doi:10.1016/j.cell.2013.01.015
- Gossler, A., Joyner, A. L., Rossant, J. and Skarnes, W. C. (1989). Mouse embryonic stem cells and reporter constructs to detect developmentally regulated genes. *Science* **244**, 463-465. doi:10.1126/science.2497519
- Grote, P. and Herrmann, B. G. (2015). Long noncoding RNAs in organogenesis: Making the difference. *Trends Genet.* **31**, 329-335. doi:10.1016/j.tig.2015.02.002
- Guttman, M., Donaghey, J., Carey, B. W., Garber, M., Grenier, J. K., Munson, G., Young, G., Lucas, A. B., Ach, R., Bruhn, L. et al. (2011). lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature* **477**, 295-300. doi:10.1038/nature10398
- Hansen, J., Floss, T., Van Sloun, P., Füchtbauer, E.-M., Vauti, F., Arnold, H.-H., Schnütgen, F., Wurst, W., von Melchner, H. and Ruiz, P. (2003). A large-scale, gene-driven mutagenesis approach for the functional analysis of the mouse genome. *Proc. Natl. Acad. Sci. USA* **100**, 9918-9922. doi:10.1073/pnas.1633296100
- Hansen, G. M., Markesich, D. C., Burnett, M. B., Zhu, Q., Dionne, K. M., Richter, L. J., Finnell, R. H., Sands, A. T., Zambrowicz, B. P. and Abuin, A. (2008). Large-scale gene trapping in C57BL/6N mouse embryonic stem cells. *Genome Res.* **18**, 1670-1679. doi:10.1101/gr.078352.108
- Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B. L., Barrell, D., Zadissa, A., Searle, S. et al. (2012). GENCODE: the reference human genome annotation for the ENCODE project. *Genome Res.* **22**, 1760-1774. doi:10.1101/gr.135350.111
- 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. doi:10.1038/ng0897-338

- Hu, G., Lou, Z. and Gupta, M. (2014). The long non-coding RNA GAS5 cooperates with the Eukaryotic translation initiation factor 4E to regulate c-Myc translation. *PLoS ONE* **9**, e107016. doi:10.1371/journal.pone.0107016
- Ishida, Y. and Leder, P. (1999). RET: a poly A-trap retrovirus vector for reversible disruption and expression monitoring of genes in living cells. *Nucleic Acids Res.* **27**, e35-e42. doi:10.1093/nar/27.24.e35
- Jia, W., Chen, W. and Kang, J. (2013). The functions of microRNAs and long non-coding RNAs in embryonic and induced pluripotent stem cells. *Genomics Proteomics Bioinformatics* **11**, 275-283. doi:10.1016/j.gpb.2013.09.004
- Kaloff, C., Anastassiadis, K., Ayadi, A., Baldock, R., Beig, J., Birling, M.-C., Bradley, A., Brown, S. D. M., Bürger, A., Bushell, W. et al. (2016). Genome wide conditional mouse knockout resources. *Drug Discov. Today: Dis. Models* **20**, 3-12. doi:10.1016/j.ddmod.2017.08.002
- Kino, T., Hurt, D. E., Ichijo, T., Nader, N. and Chrousos, G. P. (2010). Noncoding RNA gas5 is a growth arrest- and starvation-associated repressor of the glucocorticoid receptor. *Sci. Signal.* **3**, ra8. doi:10.1126/scisignal.2000568
- Kosicki, M., Tomberg, K. and Bradley, A. (2018). Repair of double-strand breaks induced by CRISPR-Cas9 leads to large deletions and complex rearrangements. *Nat. Biotechnol.* **36**, 765-771. doi:10.1038/nbt.4192
- Li, L., Liu, B., Wapinski, O. L., Tsai, M.-C., Qu, K., Zhang, J., Carlson, J. C., Lin, M., Fang, F., Gupta, R. A. et al. (2013). Targeted disruption of Hotaïr leads to homeotic transformation and gene derepression. *Cell Rep.* **5**, 3-12. doi:10.1016/j.celrep.2013.09.003
- Liu, F.-L., Liu, T.-Y. and Kung, F.-L. (2014). FKBP12 regulates the localization and processing of amyloid precursor protein in human cell lines. *J. Biosci.* **39**, 85-95. doi:10.1007/s12038-013-9400-1
- Lloyd, K. C. K. (2003). The mutant mouse regional resource center program. *Breast Cancer Res.* **5**, 7. doi:10.1186/bcr666
- Lloyd, K. C. K., Adams, D. J., Baynam, G., Beaudet, A. L., Bosch, F., Boycott, K. M., Braun, R. E., Caulfield, M., Cohn, R., Dickinson, M. E. et al. (2020). The deep genome project. *Genome Biol.* **21**, 18. doi:10.1186/s13059-020-1931-9
- Loewer, S., Cabili, M. N., Guttman, M., Loh, Y.-H., Thomas, K., Park, I. H., Garber, M., Curran, M., Onder, T., Agarwal, S. et al. (2010). Large intergenic non-coding RNA-RoR modulates reprogramming of human induced pluripotent stem cells. *Nat. Genet.* **42**, 1113-1117. doi:10.1038/ng.710
- Ma, C., Shi, X., Zhu, Q., Li, Q., Liu, Y., Yao, Y. and Song, Y. (2016). The growth arrest-specific transcript 5 (GAS5) is a pivotal tumor suppressor long noncoding RNA in human cancers. *Tumor Biol.* **37**, 1437-1444. doi:10.1007/s13277-015-4521-9
- Miard, S., Girard, M.-J., Joubert, P., Carter, S., Gonzales, A., Guo, H., Morpurgo, B., Boivin, L., Golovko, A. and Picard, F. (2017). Absence of Malat1 does not prevent DEN-induced hepatocarcinoma in mice. *Oncol. Rep.* **37**, 2153-2160. doi:10.3892/or.2017.5468
- Nakagawa, S., Shimada, M., Yanaka, K., Mito, M., Arai, T., Takahashi, E., Fujita, Y., Fujimori, T., Standaert, L., Marine, J.-C. et al. (2014). The lncRNA Neat1 is required for corpus luteum formation and the establishment of pregnancy in a subpopulation of mice. *Development* **141**, 4618-4627. doi:10.1242/dev.110544
- 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. doi:10.1093/oxfordjournals.jbchem.a124049
- Nord, A. S., Vranizan, K., Tingley, W., Zambon, A. C., Hanspers, K., Fong, L. G., Hu, Y., Bacchetti, P., Ferrin, T. E., Babbitt, P. C. et al. (2007). Modeling insertional mutagenesis using gene length and expression in murine embryonic stem cells. *PLoS ONE* **2**, e617. doi:10.1371/journal.pone.0000617
- Oliver, P. L., Chodroff, R. A., Gosal, A., Edwards, B., Cheung, A. F. P., Gomez-Rodriguez, J., Elliot, G., Garrett, L. J., Lickiss, T., Szele, F. et al. (2015). Disruption of Visc-2, a brain-expressed conserved long noncoding RNA, does not elicit an overt anatomical or behavioral phenotype. *Cereb. Cortex* **25**, 3572-3585. doi:10.1093/cercor/bhu196
- Osipovich, A. B., White-Grindley, E. K., Hicks, G. G., Roshon, M. J., Shaffer, C., Moore, J. H. and Ruley, H. E. (2004). Activation of cryptic 3' splice sites within introns of cellular genes following gene entrapment. *Nucleic Acids Res.* **32**, 2912-2924. doi:10.1093/nar/gkh604
- Osipovich, A. B., Singh, A. and Ruley, H. E. (2005). Post-entrapment genome engineering: first exon size does not affect the expression of fusion transcripts generated by gene entrapment. *Genome Res.* **15**, 428-435. doi:10.1101/gr.3258105
- Ringwald, M., Iyer, V., Mason, J. C., Stone, K. R., Tadepally, H. D., Kadin, J. A., Bult, C. J., Eppig, J. T., Oakley, D. J., Briois, S. et al. (2011). The IKMC web portal: a central point of entry to data and resources from the international knockout mouse consortium. *Nucleic Acids Res.* **39**, D849-D855. doi:10.1093/nar/gkq879
- Rinn, J. L. and Chang, H. Y. (2012). Genome regulation by long noncoding RNAs. *Annu. Rev. Biochem.* **81**, 145-166. doi:10.1146/annurev-biochem-051410-092902
- Rosen, B., Schick, J. and Wurst, W. (2015). Beyond knockouts: the International knockout mouse consortium delivers modular and evolving tools for investigating mammalian genes. *Mamm. Genome* **26**, 456-466. doi:10.1007/s00335-015-9598-3
- Salminen, M., Meyer, B. I. and Gruss, P. (1998). Efficient poly A trap approach allows the capture of genes specifically active in differentiated embryonic stem cells and in mouse embryos. *Dev. Dyn.* **212**, 326-333. doi:10.1002/(SICI)1097-0177(199806)212:2<326::AID-AJA17>3.0.CO;2-1
- Sauvageau, M., Goff, L. A., Lodato, S., Bonev, B., Groff, A. F., Gerhardinger, C., Sanchez-Gomez, D. B., Hacısuleyman, E., Li, E., Spence, M. et al. (2013). Multiple knockout mouse models reveal lincRNAs are required for life and brain development. *eLife* **2**, e01749. doi:10.7554/eLife.01749
- Schnütgen, F. (2006). Generation of multipurpose alleles for the functional analysis of the mouse genome. *Brief Funct. Genomic. Proteomic.* **5**, 15-18. doi:10.1093/bfpg/ell009
- Schnütgen, F., De-Zolt, S., van Sloun, P., Hollatz, M., Floss, T., Hansen, J., Altschmied, J., Seisenberger, C., Ghyselink, N. B., Ruiz, P. et al. (2005). Genomewide production of multipurpose alleles for the functional analysis of the mouse genome. *Proc. Natl. Acad. Sci. USA* **102**, 7221-7226. doi:10.1073/pnas.0502273102
- Sheik Mohamed, J., Gaughwin, P. M., Lim, B., Robson, P. and Lipovich, L. (2010). Conserved long noncoding RNAs transcriptionally regulated by Oct4 and Nanog modulate pluripotency in mouse embryonic stem cells. *RNA* **16**, 324-337. doi:10.1261/ma.1441510
- Shigeoka, T., Kawaichi, M. and Ishida, Y. (2005). Suppression of nonsense-mediated mRNA decay permits unbiased gene trapping in mouse embryonic stem cells. *Nucleic Acids Res.* **33**, e20. doi:10.1093/nar/gni022
- Skarnes, W. C., Auerbach, B. A. and Joyner, A. L. (1992). A gene trap approach in mouse embryonic stem cells: The lacZ reported is activated by splicing, reflects endogenous gene expression, and is mutagenic in mice. *Genes Dev.* **6**, 903-918. doi:10.1101/gad.6.6.903
- Skarnes, W. C., von Melchner, H., Wurst, W., Hicks, G., Nord, A. S., Cox, T., Young, S. G., Ruiz, P., Soriano, P., Tessier-Lavigne, M. et al. (2004). A public gene trap resource for mouse functional genomics. *Nat. Genet.* **36**, 543-544. doi:10.1038/ng0604-543
- Skarnes, W. C., Rosen, B., West, A. P., Koutsourakis, M., Bushell, W., Iyer, V., Mujica, A. O., Thomas, M., Harrow, J., Cox, T. et al. (2011). A conditional knockout resource for the genome-wide study of mouse gene function. *Nature* **474**, 337-342. doi:10.1038/nature10163
- Stanford, W. L., Epp, T., Reid, T. and Rossant, J. (2006). Gene trapping in embryonic stem cells. *Methods Enzymol.* **420**, 136-162. doi:10.1016/S0076-6879(06)20008-0
- Stryke, D., Kawamoto, M., Huang, C. C., Johns, S. J., King, L. A., Harper, C. A., Meng, E. C., Lee, R. E., Yee, A., L'Italien, L. et al. (2003). BayGenomics: a resource of insertional mutations in mouse embryonic stem cells. *Nucleic Acids Res.* **31**, 278-281. doi:10.1093/nar/gkg064
- To, C., Epp, T., Reid, T., Lan, Q., Yu, M., Li, C. Y. J., Ohishi, M., Hant, P., Tsao, N., Casallo, G. et al. (2004). The centre for modeling human disease gene trap resource. *Nucleic Acids Res.* **32**, D557-D559. doi:10.1093/nar/gkh106
- von Melchner, H., DeGregori, J. V., Rayburn, H., Reddy, S., Friedel, C. and Ruley, H. E. (1992). Selective disruption of genes expressed in totipotent embryonic stem cells. *Genes Dev.* **6**, 919-927. doi:10.1101/gad.6.6.919
- Wefers, B., Bashir, S., Rossius, J., Wurst, W. and Kühn, R. (2017). Gene editing in mouse zygotes using the CRISPR/Cas9 system. *Methods* **121-122**, 55-67. doi:10.1016/j.ymeth.2017.02.008
- Wiles, M. V., Vauti, F., Otte, J., Füchtbauer, E.-M., Ruiz, P., Füchtbauer, A., Arnold, H.-H., Lehrach, H., Metz, T., von Melchner, H. et al. (2000). Establishment of a gene-trap sequence tag library to generate mutant mice from embryonic stem cells. *Nat. Genet.* **24**, 13-14. doi:10.1038/71622
- Wurst, W., Rossant, J., Prideaux, V., Kownacka, M., Joyner, A., Hill, D. P., Guillemot, F., Gasca, S., Cado, D., Auerbach, A. et al. (1995). A large-scale gene-trap screen for insertional mutations in developmentally regulated genes in mice. *Genetics* **139**, 889-899.
- Yao, X., Zhang, M., Wang, X., Ying, W., Hu, X., Dai, P., Meng, F., Shi, L., Sun, Y., Yao, N. et al. (2018). Tild-CRISPR allows for efficient and precise gene Knockin in mouse and human cells. *Dev. Cell* **45**, 526-536.e5. doi:10.1016/j.devcel.2018.04.021
- Yates, A., Akanni, W., Amode, M. R., Barrell, D., Billis, K., Carvalho-Silva, D., Cummins, C., Clapham, P., Fitzgerald, S., Gil, L. et al. (2016). Ensembl 2016. *Nucleic Acids Res.* **44**, D710-D716. doi:10.1093/nar/gkv1157
- Yoshida, M., Yagi, T., Furuta, Y., Takayanagi, K., Kominami, R., Takeda, N., 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. doi:10.1007/BF01969122
- 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. doi:10.1038/33423
- Zambrowicz, B. P., Abuin, A., Ramirez-Solis, R., Richter, L. J., Piggott, J., BeltrandelRio, H., Buxton, E. C., Edwards, J., Finch, R. A., Friddle, C. J. 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. doi:10.1073/pnas.2336103100
- Zhang, Z., Zhu, Z., Watabe, K., Zhang, X., Bai, C., Xu, M., Wu, F. and Mo, Y.-Y. (2013). Negative regulation of lincRNA GAS5 by miR-21. *Cell Death and Differ.* **20**, 1558-1568. doi:10.1038/cdd.2013.110