TEXAS A&M INSTITUTE FOR GENOMIC MEDICINE

Genomics and Bioinformatics Services

Providing Genomics and Bioinformatics Services to the Texas A&M System, Texas, and the World

A Gene Trap Resource to Characterize Cancer-related lncRNAs in C57BL/6 ES Cells & Mice

Michael P. McLeod[1,2], Andrei Golovko[1], Noushin Ghaffari [2], Esmaeil Atashpaz Gargari[2], Huiping Guo[1], Charles D. Johnson[2] Stephen H. Safe[3], Indira Jutooru[3] and Benjamin Morpurgo[1]

[1]Texas A&M Institute for Genomic Medicine; [2]Texas A&M AgriLife Genomics and Bioinformatics Services, [3]Texas A&M College of Veterinary Medicine, Dept. of Physiology & Pharmacology. College Station, TX, USA.

Abstract

Texas A&M Institute for Genomic Medicine (TIGM) houses the world's largest library of knockout C57BL/6N ES cells and provides transgenic mice for researchers around the world. Long non-coding RNAs (lncRNA) are non-protein coding transcripts longer than 200 nucleotides. A handful of studies have implicated lncRNAs in a variety of disease states and demonstrated their involvement in oncogenesis. The 500,000 sequence tags, which were produced in the course of generating the 350,000 gene trapped clones, were compared to the latest build (38) of the mouse genome using BLAST. Our screening of more than 18,000 clones has identified over 1,000 inactivated ncRNAs with including a number of lncRNAs. One such clone, IST14461G11, was used to establish a colony of homozygous mutant Malat1 mice in pure C57BL/6N genetic background. Real time PCR-based expression analysis confirmed absence of the targeted transcript in the mutant line.

Introduction

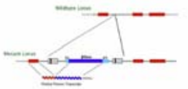
The past years have seen an explosion in the number of detected RNA transcripts with no apparent protein-coding potential. This has led to speculation that non-protein-coding RNAs (ncRNAs) might be as important as proteins in the regulation of vital cellular functions. However, there has been significantly less progress in actually demonstrating the functions of these transcripts. Although some ncRNAs act as molecular switches that regulate gene expression, the function of many ncRNAs is unknown.

Genome-wide profiling revealed that many transcribed non-coding ultraconserved regions exhibit distinct profiles in various human cancer states (Calin 2007). For example, Malat1 was originally identified as an abundantly expressed ncRNA that is up-regulated during metastasis of early-stage non-small cell lung cancer and its overexpression is an early prognostic marker for poor patient survival rates (Fu 2006). The highly conserved mouse homologue of Malat1 was found to be highly expressed in hepatocellular carcinoma (Lin 2007) and we have identified Malat1 expression and function in pancreatic cancer. We are proposing to analyze our library of the gene-trapped embryonic stem (ES) cell clones in the C57BL/6 background to identify clones with inactivated lncRNAs and use them to produce animal models. These lines will then be made available to the entire research community for studies on the function on lncRNAs in cellular homeostasis and cancer.

Technology

The basic gene trap vectors we have used include a reporter gene downstream of a splice acceptor sequence (Fig.1).

Figure 1. Gene trap vectors used in C57BL/6 library (β-geo version shown). LTR, long terminal repeat; SA, splice acceptor sequence; β-geo, galactosidase/neomycin phosphotransferase fusion gene; pA, polyadenylation sequence .



They are designed to function when inserted in an intron, to produce incorrect splicing of the target gene such that all exons downstream of the insertion site are not expressed. The gene trap cassette is inserted in a retroviral vector. Retroviruses insert as a single copy per locus, with no rearrangement of flanking sequences. They have a preference for insertions at the 5' end of genes, often upstream of the initiator ATG, and the splice acceptor sequence we use does not appear to be bypassed by the RNA-splicing machinery. As a result, the majority of the mutations generated using our gene trap vectors are predicted to lead to null alleles.

References

1. Calin GA, Liu CG, Ferracin M, et al. (September 2007). "Ultraconserved regions encoding ncRNAs are altered in human leukemias and carcinomas". Cancer Cell 12 (3): 215–29.

2. Fu X, Ravindranath L, Tran N, Petrovics G, Srivastava S (March 2006). "Regulation of apoptosis by a prostate-specific and prostate cancer-associated noncoding gene, PCGEM1". DNA and Cell Biology 25 (3): 135–41

3. Lin R, Maeda S, Liu C, Karin M, Edgington TS (February 2007). "A large noncoding RNA is a marker for murine hepatocellular carcinomas and a spectrum of human carcinomas". Oncogene 26 (6): 851–8.

Results

Analyzing and annotating TIGM C57BL/6N ES clones (pilot study)

BLAST was used to compare the 498,995 insertion sequence tags from the 335,900 TIGM clones to build 38 of the mouse genome. Default settings were used and information regarding the top hit was stored in an Oracle database. This included the name of the IST, the start and stop of the alignment on the query IST, the start and stop of the alignment on the chromosome, and if the alignment was in the same or opposing orientation. The last field of information is needed to determine if the gene trap insertion is in the appropriate direction when compared to the direction of the gene in the genome.

In a similar manner, mouse and human ncRNA data from RFAM (http://rfam.sanger.ac.uk/) and fRNAdb (http://www.ncrna.org/frnadb/) were also mapped to build 38 of the mouse genome and that data. This data was converted to a GFF v.3 format file and loaded into the same Oracle database. Also, the GFF file for build 38 containing more than 680,000 annotated regions was downloaded from NCBI, and loaded into the Oracle database (ftp://ftp.ncbi.nlm.nih.gov/genomes/M_musculus/GFF/).

SQL queries were then made to select clones which have a gene trap insertion within the same region as a ncRNA were selected which were found within the coordinates of the ncRNA and in the appropriate direction to effectively interrupt transcription.

The bioinformatics efforts to date have revealed more than 18,000 clones have insertions which appear to inactivate 1,000 unique long ncRNAs (those greater than 200bp). Included in these are lncRNAs such as Malat1, Tsix and Air. Table 1 shows an example of the output produced in the course of remapping the clones and ncRNAs:

Clone mapping results									ncRNA mapping results		
ist	chr_start	chr_end	strand	ist_start	ist_end	expect	HSP	chromosome	Gene ID, name	chr_start	chr_end
IST14461G11	5802420	5802475	+	2	56	7.00E-18	94	Chr 19	gi|372099091|FR0393287|Hepcarcin **MALAT1**	5795690	5802671
IST13169D9	103481277	103481042	-	1	236	2.00E-120	436	Chr X	gi|372099090|NC_000086.7|FR0353021| **Tsix**	103481273	103484957
IST13831H9	12750221	12750105	-	1	117	2.00E-52	209	Chr 17	gi|372099093|NC_000083.6|FR0378282| **Air**	12741949	12805881

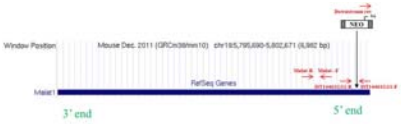
Table 1. Sample output produced as a result of mapping TIGM clones and lncRNAs to the latest mouse genome build. Clones were found to carry mutations in lncRNAs of high scientific value: Malat1, Tsix and Air.

Generation of Malat1 knockout

TIGM C57BL/6N ES cell clone IST14461G11 was found to carry mutation in lncRNA Malat1 (Figure 2) and was chosen for further work. The clone was expanded and the genomic sequence surrounding the gene trap insertion site was determined as follows (the insertion site is denoted with an asterisk *):

```
CAGGCATTCAGGCAGCGAGAGCAGAGCAGCGTAGAGCAGCACAGCTGAGCTCGTGAGGCAGGAGACTCAGCCCGAGGAAA
TCGCAGATAAGTTTTTAATTAAAAAGATTGAGCAGTAAAAAGAATTAGAACTCTAAACTTAAGCTAATAGAGTAGCTTATCGAA
ATATTACTTAGTCTTAATAATCTAAGAAGATCTTAAGAGATAACATGAAGGCTTATTTAAACAGTTTGAAAAAGGAAATGAGGA
GA*AAAGTATTTGTACTGTATAATGGAGGCTGACCAGAGCAGTTTAGGAGATTGTAAAGGGAGGTTTTGTGAAGTTCTAAAAG
GTTCTAGTTTGAAGGTCGGCCTTGTAGATTAAAACGAAGGTTACCTAAATAGAATCTAAGTGGCATTTAAAACAGTAAAGTTGT
AGAGAATAGTTTGAAAATGAGGTGTAGTTTTAAAAGATTGAGAAAAGTAGGTTAAGTTGACGGCCGTTATAAAAATCCTTCGA
CTG
```

Figure 2 Schematic representation of the mutated locus Malat1. IST14461G11-F, IST14461G11-R and Downstream rev are the primers used in genotyping. Malat1-F and Malat1-R are primers used to confirm knockout via RT-PCR.


NEO
Window Position
Mouse Dec. 2011 (GRCm38/mm10) chr19:5,795,690-5,802,671 (6,982 bp)
RefSeq Genes
Malat1
3' end
5' end

The IST14461G11 clone was used to establish a colony of homozygous mutant Malat1 mice in pure C57BL/6N genetic background. Mutant mice were generated using standard procedures. In short, a mutant ES cell clone was expanded and microinjected into albino B6 host blastocysts to generate germline chimeras. Those were bred to C57BL/6 females for germline transmission of the mutant Malat1 allele. The correct mutation was confirmed using PCR-based genotyping protocol (Table 2) using primers specific for genomic insertion site and for the vector (Figure 2).

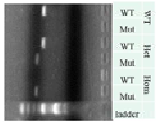
Table 2 Schematic representation of the mutated locus Malat1. IST14461G11-F, IST14461G11-R and Downstream rev are the primers used in genotyping. Malat1-F and Malat1-R are primers used to confirm knockout via RT-PCR.

Primer Sequences (5' to 3'):	
Mutant Oligos: IST14461G11-F + Downstream Rev (337bp)	
Wt Rxn Oligos: IST14461G11-F + IST14461G11-R (468bp)	
IST14461G11-F	AGAGCAGAGCAGCGTAGAGC
IST14461G11-R	TAACGGCCGTCAACTTAACC
Downstream rev	CCAATAAACCCTCTTGCAGTTGC

Generation of Malat1 knockout (Continued)

The heterozygous mice were subsequently bred to obtain homozygous mutants. As shown in Figure 3, the wild type Malat1 amplicon (468 bp) was only detected in Wt and Het animals, whereas the mutant product (270 bp) was amplified in both Null and Het animals.

Figure 3. Genotyping results showing wild type (WT), heterozygous (Het) and homozygous (Hom) mutants using primer sets for wild type (WT) and mutant (Mut) alleles in TIGM clone IST14461G11 for Malat1.



Malat1 knockout confirmation

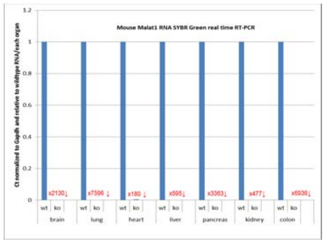
To confirm that gene trap-based inactivation resulted in the anticipated reduction of the gene expression, Malat1 mRNA expression levels in mouse brain, lung, heart, liver, pancreas, kidney and colon were determined by real-time PCR on the Step one plus (Applied Biosystems) instrument. SYBR Green one step Real-time RT-PCR was performed with 20ng total RNA using iScript one-step RT-PCR and SYBR Green mix (BIO-RAD). The assay was performed using a primer pair amplifying a 170 nt fragment of the transcript downstream of the gene trap insertion (Figure 3). Expression levels were normalized to Gapdh expression using primers specific to this gene.

Primer sequences used in this study were as follows:

Malat1-F	ggcagaatgcctttgaagag
Malat1-R	ggtcagctgccaatgctagt
Gapdh-F	ggcattgctctcaatgacaac
Gapdh-R	gccatgtaggccatgaggt

As shown in Figure 4, Malat1 expression was completely eliminated in all tissues in homozygous mice compared with wild type siblings. These findings indicate that gene trap insertions in long non-coding RNAs can effectively knock out genes and create true mutant alleles.

Figure 4 Fold change in Malat1 gene expression in knockout mice as compared to the wild type.



Overall, our preliminary data indicates that genomic localization of the gene trap insertions from TIGM library along with mapping of the non-coding transcriptome can serve as a very effective tool to segment out a collection of functional mutations in thousands of murine lncRNAs that, in turn, can be used to produce a repository of novel mutant mouse models for various research areas.

Conclusions and Future Directions

TIGM maintains the world's largest library of stable mouse knockout embryonic stem (ES) cells in the C57BL/6 background, with a total of over 350,000 clones representing more than 10,000 unique protein-coding genes. The library has been successfully mapped to the most recent genome build and distilled down to a collection of non-coding RNA. Simple, rapid and inexpensive access to a collection of already created mutant lines will greatly facilitate the availability for academic, government and private research laboratories. As a proof of concept, we have successfully established a mutant mouse line carrying a homozygous mutation in Malat1, an lncRNA involved in several different types of cancer, including pancreatic. Future efforts will include continuing analysis of the TIGM collection to discover and verify additional inactivated lncRNAs. We will continue annotating our clone collection against the most recent mouse genome builds as they become available, and extrapolating mutations to identify novel genes that have been trapped. As a result, we will be adding more clone-gene associations to the database with the focus on long non-coding RNA. These additional clones will be made easily accessible to the scientific community and can be found on the TIGM website using text or sequence searching tools. TIGM is also planning to produce more mutant mouse lines with disrupted lncRNAs of the highest scientific value, most of which will include potential cancer targets, and make them available to the scientific community. Finally, the existing Malat1 knockout line will be further analyzed for its ability to influence the onset and progression of different types of cancer.

For more information please visit www.tigm.org or contact us at info@tigm.org.