Thesis for the Master of *Science*

Computational design of Genome-wide sgRNAs for long intergenic non-coding RNA knockout using Cas9

Hyeon Joo Lee

Graduate School of Hanyang University

August 2015

Thesis for the Master of *Science*

Computational design of Genome-wide sgRNAs for long intergenic non-coding RNA knockout using Cas9

Thesis Supervisor: Jin-Wu Nam

A Thesis submitted to the graduate school of Hanyang University in partial fulfillment of the requirements for the degree of *Master of Science*

Hyeon Joo Lee

August 2015

Department of Life Science
Graduate School of Hanyang University

This thesis, written by *Hyeon Joo Lee*,
has been approved as a thesis for the degree of
Master of Science.

August 2015

Committee Chairman: Hyongbum Kim

Committee member: Jin-Wu Nam

Committee member: Jiwon Shim

Graduate School of Hanyang University

Table of Contents

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Computational design of Genome-wide sgRNAs for long intergenic non-coding RNA knockout using Cas9

Hyeon Joo Lee
Department of Life Science
The Graduate School
Hanyang University

Genomic engineering with clustered, regularly interspaced, short palindromic repeats (CRISPR) system modifies predetermined genomic loci of target genes. Although the programmable nuclease, which is derived from prokaryotic cells, is widely being used to knockout genes of interest by making frame-shift mutations, yet only applicable to protein-coding genes. For functional studies of long intergenic non-coding RNAs (lincRNAs), multiple features need to be integrated for determining of target loci and to design a pair of single guide RNAs (sgRNAs) of CRISPR system to delete a target locus. Here, we computationally designed the genome-wide sgRNAs for knocking out lincRNAs using CRISPR system. The set of lincRNAs is based on the gold-standard set filtered from the GENCODE annotations, 5,882 transcripts in human and 822 transcripts in mouse. To make an archive of functional lincRNAs, ones that are expressed in at least one of cell-types examined in the ENCODE project were selected, annotating with tissue or stage specific expression pattern. The regions with RNA family motifs and embedding conserved elements are defined as candidate loci to be deleted. The

same approach is applied to annotate functional signatures of lincRNAs in other species, fly, worm and zebrafish. We also computationally designed sgRNAs with the high efficiency that can provide any pair of sgRNAs for the deletion of the candidate loci or whole lincRNA locus. To select sgRNAs with the high efficiency, we assessed them with sequence and structural features, and off-targets. We, thus, provide an online tool to guide the selection of target sites and sgRNAs for knockout any lincRNAs of interest.

Chapter 1. Introduction

Section 1. Long non-coding RNAs (lncRNAs)

Recent studies have reported the functional importance of lncRNAs. Along with protein-coding genes, several lncRNAs have regulatory roles at the multiple levels of transcription, mRNA splicing, and translation (Geisler and Coller, 2013; Guttman and Rinn, 2012). In many studies, the significance of lncRNAs have been discovered using biochemical approaches which have revealed the functions of lncRNAs by identifying the interaction between lncRNAs and other molecules like DNAs, RNAs, or proteins (Geisler and Coller, 2013; Bassett *et al.*, 2014; Esteller, 2011; Ulitsky and Bartel, 2013; Batista and Chang, 2013; Li and Chang, 2014; Gutschner and Diederichs, 2012; Du *et al.*, 2013; Lee, 2012). Recently developed protein centric biochemical purification methods, such as cross-linking immunoprecipitation followed by sequencing, have been carried out, but targets are unknown in most cases (Licatalosi *et al.*, 2008; Zhang and Darnell, 2011). RNA immunoprecipitation have been applied to identify the interaction of Xist and PRC2 (Zhao *et al.*, 2008). Also, RNA antisense purification (RAP) technique has found that Xist would be broadly localized across the X chromosome for maintenance of X chromosome inactivation (XCI) in three dimensional space (Engreitz *et al.*, 2013; Simon *et al.*, 2013). Although the functional significances of several lncRNAs have been discovered using this biochemical approaches, however, most lncRNAs are expressed in a cell specific manner and their weak expression is hardly detectable using recent technology (Engreitz *et al.*, 2014; Mili and Steitz, 2004).

Section 2. Genome editing

Since the biochemical approaches could be applied only for lncRNAs which specifically binds to other molecules, loss-of-function strategies may be more effective way for functional studies of lncRNAs. One of the loss-of-function approaches, RNA interference (RNAi)-based knockdown has been used for down-regulating endogenous lncRNAs (Guttman *et al.*, 2011; Hu, Yuan, and Lodish, 2011; Ng, Johnson, and Stanton, 2012; Klatenhoff *et al.*, 2013; Sun *et al.*, 2013). Although the RNAi approaches has been successfully applied in a handful of cases, abundant nuclear lncRNAs may not be efficiently targeted by exogenous small interfering RNAs (siRNAs) (Gutschner *et al.*, 2011) because mammalian RNAi machinery and exogenous siRNAs are in the cytoplasm and are transiently expressed. For instance, a highly abundant metastasis associated lncRNA, *MALAT1*, was not efficiently down regulated by RNAi (Gutschner *et al.*, 2011; Ji *et al.*, 2014).

Recent advances in the development of programmable nucleases have allowed loss-of-function studies for several lncRNA genes, via genome deletion of a target gene (Liu *et al.*, 2013) or termination of transcription (Gutschner *et al.*, 2011; Xiang *et al.*, 2014; Eissmann *et al.*, 2012). For example, the integration of RNA destabilizing elements to the immediate upstream region of *MALAT1* using zinc finger nucleases (ZFNs) results in permanent, efficient silencing of the lncRNA gene (Gutschner *et al.*, 2011; Eissmann *et al.*, 2012). Transcription activator-like effector nucleases (TALENs) have been used to disrupt lncRNAs (Liu *et al.*, 2013) and microRNAs (miRNAs) (Liu *et al.*, 2013; Kim *et al.*, 2013). Most recently, RNA guided engineered nucleases (RGENs), derived from the bacterial CRISPR/Cas9 system, have been used to disrupt protein-coding genes (Cho *et al.*, 2013; Cong *et al.*, 2013; Mali *et al.*, 2013; Jinek *et al.*, 2013) and non-coding RNA genes (Ho *et al.*, 2014; Han *et al.*, 2014). RGENs provide an

efficient, inexpensive, and easily programmable method for genome edition (Wang *et al.*, 2014; Zhou *et al.*, 2014). Compared to ZFNs and TALENs, they are easier to engineer for multiplex genome editing, which can achieved by using multiple guide RNAs (Kim and Kim, 2014). Given that gene deletion is one of the best ways to achieve lncRNA loss of function and that large targed deletions require multiplex genome editing to induce two double strand breaks (DSBs) that flank the deletion target, RGENs are expected to be superior to TALENs and ZFNs for lncRNA loss-of-function studies.

Chapter 2. Revised lincRNAs set

To knockout lncRNAs in a genome-wide manner, accurate annotations of lncRNA genes are necessary. Due to low expression levels, annotations of some putative lncRNAs are fragmentary and transcript boundaries are poorly defined (Derrien *et al.*, 2012; van Bakel *et al.*, 2010). To revise the boundaries of lncRNAs, genome-wide profiling of transcription start sites (TSSs) (Faulkner *et al.*, 2009) and cleavage and polyadenylation sites (CPSs) (Ulitsky *et al.*, 2011; Nam and Bartel, 2012) can be used as additional evidence. To discover unknown function of long intergenic non-coding RNAs (lincRNAs) using Cas9, we first improved the current lincRNA annotation; putative and known lincRNAs annotated from the GENCODE consortium (http://www.gencodegenes.org/) using the following computational pipeline (Figure 1A).

Section 1. Improvement of predicted gene model using CAGE-seq and polyA-seq

First, all previously annotated transcripts as lincRNAs from the GENCODE annotations (hg19: version 19, mm9: version M1) were updated with major transcription start sites (TSSs) and cleavage and polyadenylation sites (CPSs) which were detected from experimental data. We obtained deepCAGE data of 17 tissues in the FANTOM project and poly(A) position profiling by sequencing (3P-seq) data (Nam *et al.*, 2014) across many different cell types (HeLa, HEK293, Huh7, and IMR90 cells for human; mES and 3T3 cells and liver, muscle, heart, white adipose, and kidney tissues for mouse) from the NCBI gene expression omnibus (GEO) (accession number: GSE52531). The detailed method for the prediction of CPSs is described in the previous work (Nam *et al.*, 2014). The same

method was used to predict TSSs as well. Since lincRNAs were expressed with cell-type specific-manner, the deepCAGE data and 3P-seq data supported the small portion of lincRNA transcripts. Even though the percentage of updated transcripts was low, this step was significant to perform because the revised *RP11-168L7.1* gene model contains a new TSS upstream of the original TSS and includes transcription factor binding sites (TFBSs) in the region immediately upstream of the new TSS (Figure 1B). In fact, the higher number of TFBSs were located at the 2 kb upstream of the revised TSSs than either the number-matched unrevised TSSs or number-matched random sites (Figure 1C).

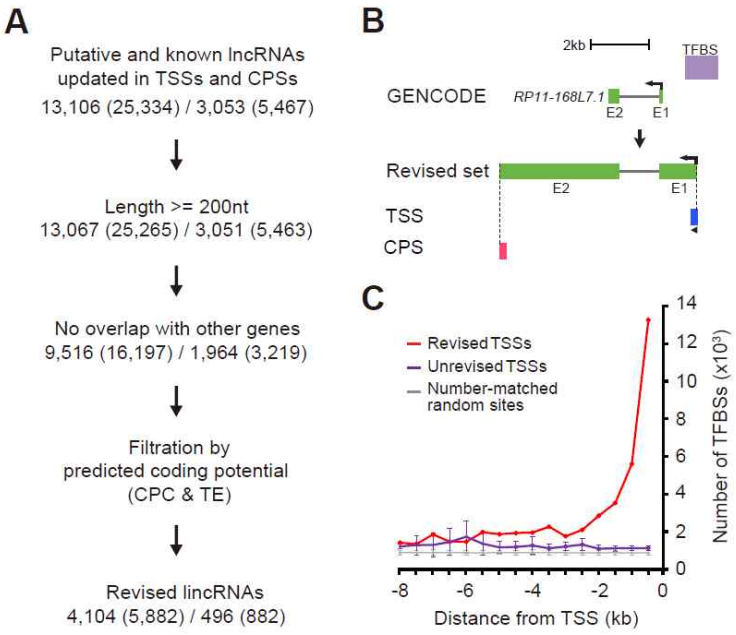


Figure 1. Revised annotations of lincRNA genes. (A) A schematic flow for revising the annotations of human and mouse lincRNA genes. (B) A revised *RP11-168L7.1* gene structures with updated TSS (blue box) and CPS (red box). The purple box indicates a cluster of transcription factor binding sites (TFBSs). (C) Frequencies of TFBSs upstream of 1,214 revised unique TSSs (red line), number-matched unrevised TSSs (purple line), and number-matched random sites (gray line). TFBSs were data profiled by chromatin

immunoprecipitation followed by sequencing for many transcription factors in the ENCODE project (Consortium, 2012). 100 cohorts of number-matched unrevised and random sites were examined, respectively, and their standard errors were indicated in each bin.

Section 2. Discrimination of lincRNAs from protein-coding genes based on coding potential

The revised lincRNAs were retained by several filtrations such as length, overlaps with other genes, and coding potential (Figure 1A). The transcripts less than 200 nt in length or that overlapped with other genes were selected to exclude from the revised set. Then, coding potential also was re-assessed for the lincRNAs to exclude potential fragments of coding transcripts. The transcripts were filtered out if its (1) coding potential calculation (CPC) score, calculated by homology-search over documented non-redundant protein sequences, that was greater than -0.3 for human (Jia *et al.*, 2010) and -0.2 for mouse (Kong *et al.*, 2007) and (2) a ribosome association rate, indicated as translation efficiency (TE) was greater than 0.03 for human and 0.08 for mouse in Ribo-seq that the actively translated RNAs were sequenced (Ingolia *et al*, 2009; Ingolia *et al*, 2011). This filtration retained 5,882 for and 822 bona fide human and mouse lincRNAs from 4,104 and 496 loci, as a revised set for knockout (Figure 1A).

Section 3. Characteristics of lincRNAs

Using the computational pipeline (Figure 1A), we enabled to improve the boundary of lincRNA transcripts, including 901 with both updated TSS and CPS and 1,594 with either TSS or CPS (Figure 2A). For mouse, 32% (262/822) of lincRNA gene models were improved with TSSs and CPSs (Figure 2B). The human lincRNAs falled within the interval from 200 bp to 1 Mb and the median length is 4,790 bp (Figure 2C). The majority (85.67%) of human lincRNAs have

multi-exonic structures (Figure 2D). For mouse, the median length of lincRNA genes is 6,316 bp and the 98.9% have multi-exonic structures (Figure 2E and F). If CRISPR-Cas9 system can deletion at least 1 Mb length, all of the lincRNA genes can be knockout (Figure 2C and E).

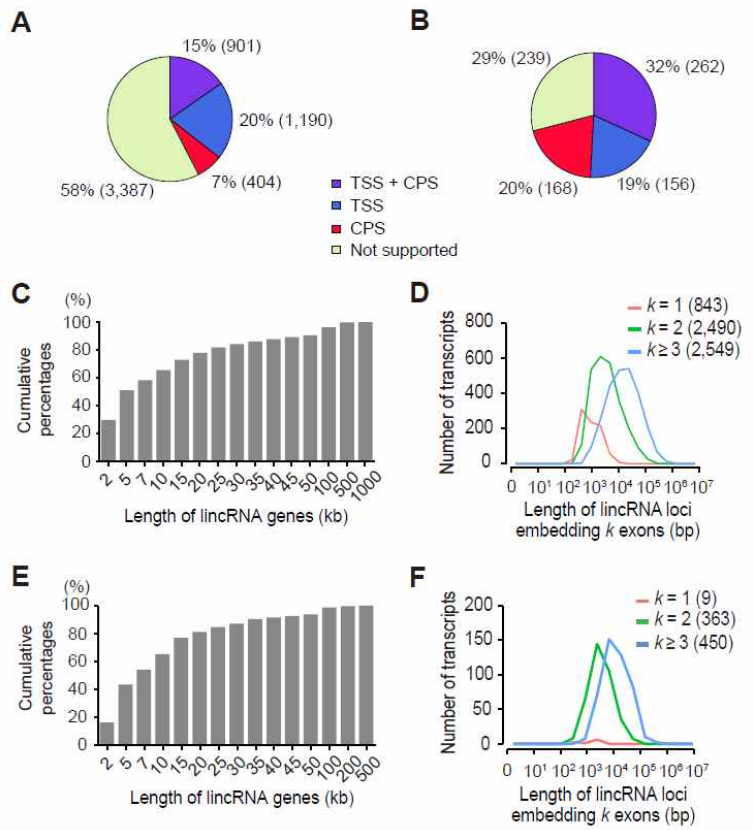


Figure 2. Characteristics of human and mouse lincRNAs. (A-B) Proportion of the revised lincRNAs with updated in TSSs and/or CPSs. (C) A cumulative distribution of the lengths of 4,104 revised human lincRNA genes. (D) Length distributions of lincRNA genes comprising k exons and k-1 introns, with k=1 indicated by a gray line, k=2 by a green line, and $k \geq 3$ by a orange line. The numbers in parentheses indicate the number of lincRNAs corresponding each group. (E) A cumulative length distribution of 496 revised mouse lincRNA genes. 99.7% of the genes were less than 500 kb. (F) Length distributions of mouse lincRNA genes comprising k exons and k-1 introns. Otherwise, the details are as in (D).

Chapter 3. Functional annotation of mammalian lincRNAs

For successful lincRNAs knockout experiment, we have to select appropriate lincRNAs with high expression in certain cell types and annotate regions where to be deleted for causing loss of their function. Here, we performed to assay functional evidence using several factors of lincRNAs.

Section 1. Cell-specific expression signatures

To annotate functional evidence about the revised lincRNAs, we assay their expression signatures in different cell types using RNA-seq data from ENCODE and NCBI GEO (Table 1A and B). For expression profiling through different cell types, we first selected transcripts that have reads per million (RPM) greater than 0.1 in at least one cell type. Then we measured their expression signatures in gene level using selected the isoforms having highest RPM. Of 2,498 lincRNAs, those with similar expression signatures were clustered using the CLICK algorithm (Sharan and Shamir, 2000) implemented in a published clustering program, Expander version 6.3.1 released on February 16, 2014 (Ulitsky *et al.*, 2010). As a result, 2,473 lincRNAs were grouped into 31 clusters with specific expression signatures in human (Table 2A). The lincRNAs in several clusters suggests that those are functional products in specific cell type (Figure 3A). Using the same approach, 452 mouse lincRNAs were grouped into 3 clusters (Figure 3B and Table 2B).

Table 1A. 33 human cell lines from four laboratories and GEO database.

Laboratory*	Cell Lines
CSHL	A549, AG04450, BJ, CD20+, GM12878, H1-hESC, HSMM, HUVEC, HeLa-S3, HepG2, IMR90, K562, MCF-7, SK-N-SH, SK-N-SH_RA
HudsonAlpha	BE2_C, ECC-1, Jurkat, PANC-1, PFSK-1, SK-N-SH, T-47D, U87
Caltech	GM12878, GM12891, GM12892, H1-hESC, HCT-116, HSMM, HUVEC, HeLa-S3, HepG2, K562, LHCN-M2, MCF-7, NHEK, NHLF
GIS	GM12878, H1-hESC, K562
GEO	H1-hESC, HEK293, HeLa, Huh7, K562

*Caltech: California Institute of Technology, CSHL: Cold Spring Harbor Laboratory, HudsonAlpha: HudsonAlpha Institute for Biotechnology, GIS: Genome Institute of Singapore, GEO: Gene Expression Omnibus

Table 1B. 54 mouse cell, tissue, or DNA samples from five laboratories.

Laboratory*	Cell or tissue samples**
CSHL	Adrenal(T), Bladder(T), CNS(T), Cerebellum(T), Colon(T), Cortex(T), Duodenum(T), FrontalLobe(T), GenitalFatPad(T), Heart(T), Kidney(T), LgIntestine(T), Limb(T), Liver(T), Lung(T), MammaryGland(T), Ovary(T), Placenta(T), SmIntestine(T), Spleen(T), Stomach(T), SubcFatPad(T), Testis(T), Thymus(T), WholeBrain(T)
Caltech	10T1/2(C), C2C12(C)
UW	416B(C), A20(C), CD19+(P), CD43-(P), Cerebrum(P), FatPad(T), HeadlessEmbryo(T), Heart(T), Kidney(T), LgIntestine(T), Liver(T), Lung(T), MEL(C), NIH-3T3(C), Patski(C), SkMuscle(T), Spleen(T), T-Naïve(P)
LICR	BoneMarrow(T), Cerebellum(T), Cortex(T), ES-Bruce4(P), Heart(T), Kidney(T), Limb(T), Liver(T), Lung(T), MEF(P), MEL(C), OlfactBulb(T), Placenta(T), SmIntestine(T), Spleen(T), Testis(T), Thymus(T), WholeBrain(T)
PSU	CH12(C), Erythrobl(P), FVLstem(P), Fvprogenitor(P), G1E(P), G1E-ER4(P), MEL(C), MEP(P)

* Caltech: California Institute of Technology, CSHL: Cold Spring Harbor Laboratory, LICR: Ludwig Institute for Cancer Research, PSU: Pennsylvania State University, UW: University of Washington

** T: Tissue, C: Cell line, P: Primary cells

Table 2A. 31 clusters using standarized RPM of 2,498 human lincRNA genes.

Name	Size	Homogeneity	Specifically expressed cell types	Laboratory*
Cluster 1	175	0.71	K562	CSHL, Caltech, GIS, GEO
Cluster 2	174	0.721	H1-hESC	CSHL, Caltech, GIS, GEO
Cluster 3	163	0.777	BE2_C	HudsonAlpha
Cluster 4	148	0.681	GM12878, GM12891, GM12892	CSHL, Caltech, GIS
Cluster 5	137	0.693	MCF-7	CSHL
Cluster 6	112	0.706	HepG2	CSHL, Caltech
Cluster 7	110	0.733	GM12878	CSHL
Cluster 8	100	0.666	HeLa-S3	CSHL, Caltech
Cluster 9	96	0.645	NHEK	CSHL, Caltech
Cluster 10	93	0.714	CD20+	CSHL
Cluster 11	83	0.726	SK-N-SH	HudsonAlpha
Cluster 12	80	0.715	T-47D	HudsonAlpha
Cluster 13	73	0.674	HSMM, LHCN-M2	CSHL, Caltech
Cluster 14	71	0.671	HUVEC	CSHL, Caltech
Cluster 15	67	0.727	ECC-1	HudsonAlpha
Cluster 16	67	0.672	HEK293	GEO
Cluster 17	67	0.729	Huh7	GEO
Cluster 18	64	0.684	SK-N-SH	CSHL, HudsonAlpha
Cluster 19	61	0.644	K562	CSHL
Cluster 20	59	0.756	Monocytes-CD14+	CSHL
Cluster 21	50	0.718	Jurkat	HudsonAlpha
Cluster 22	50	0.704	SK-N-SH	CSHL, HudsonAlpha
Cluster 23	47	0.675	HeLa	GEO
Cluster 24	46	0.662	K562	GEO
Cluster 25	45	0.765	U87	HudsonAlpha
Cluster 26	44	0.689	HeLa	GEO
Cluster 27	41	0.641	HeLa	GEO
Cluster 28	39	0.702	SK-N-SH	CSHL
Cluster 29	39	0.624	A549	CSHL
Cluster 30	38	0.684	K562	GIS
Cluster 31	34	0.737	PFSK-1	HudsonAlpha

* CSHL: Cold Spring Harbor Laboratory, Caltech: California Institute of Technology, GIS: Genome Institute of Singapore, HudsonAlpha: HudsonAlpha Institute for Biotechnology and GEO: Gene Expression Omnibus

Table 2B. 3 clusters using standarized RPM of 452 mouse lincRNA genes.

Name	Size	Homogeneity	Specifically expressed cell types	Laboratory*
Cluster 1	144	0.853	Testis	CSHL
Cluster 2	98	0.55	Cerebell, Cerebrum, Cortex, HeadlessEmbryo, Whole Brain	CSHL, UW
Cluster 3	30	0.601	Thymus	CSHL, UW

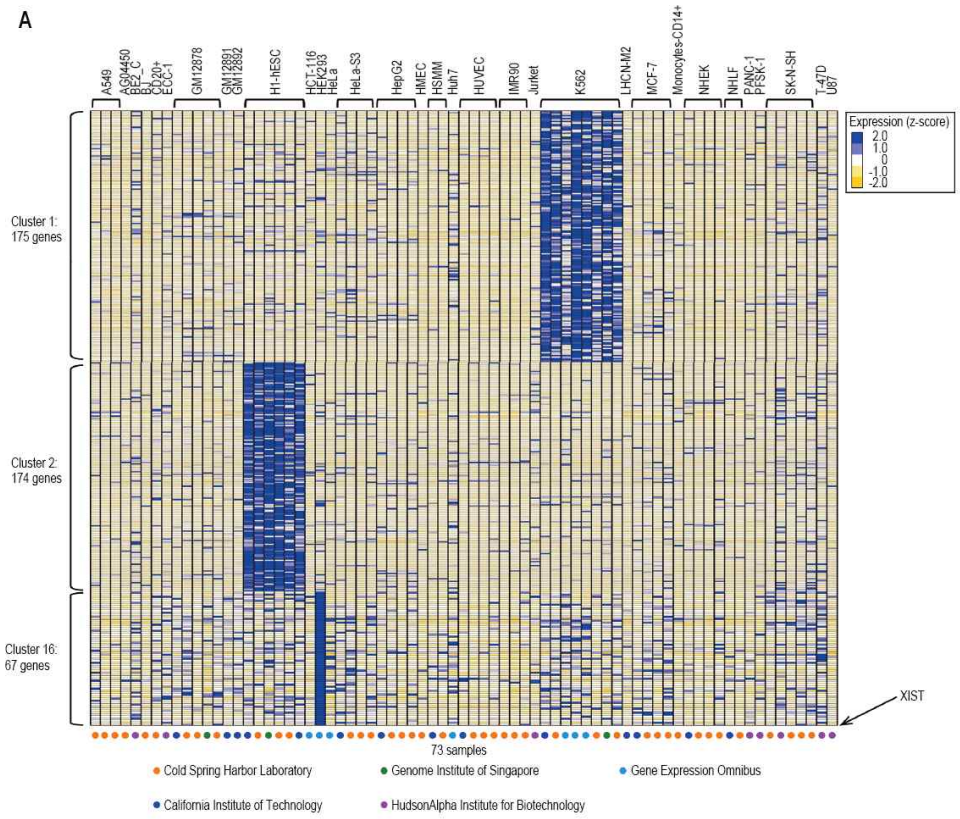
* CSHL: Cold Spring Harbor Laboratory, Caltech: California Institute of Technology, LICR: Ludwig Institute for Cancer Research, PSU: Pennsylvania State Univerisity and UW: University of Washington

Section 2. Conserved domains within lncRNAs

Beside the expression signatures, the functional evidence of lincRNAs can be explicit by existence of conserved elements within these boundary. A conservation signal for each lincRNA was obtained from the UCSC genome browser (Karolchik *et al.*, 2014). The base-wise conservation scores which were pre-calculated by the phylogenetic p-values (phyloP) over multiple genome alignments of 99 vertebrates (Margulies *et al.*, 2003) were used to annotate the conserved domains on a genome-wide scale. The conserved domains were defined as a window, in which an average conservation score is equal to or greater than 0.3, 0.5, or 0.7, respectively. The initial window size was set as 50 bp. The window is then extended by 25 bp until the average conservation score is below the predefined value (0.3, 0.5, or 0.7). As a result, we found that 40.5% (for human) and 41.5% (for mouse) of revised lincRNAs contained at least one conserved element with an average conservation score greater than 0.5; the median length of the conserved domains within the lincRNAs was approximately 184 bp in human and mouse (Figure 3C and D), which are longer than that observed in worms (Nam and Bartel, 2012).

Section 3. RNA family (Rfam) motifs

We also examined whether the lincRNAs include other Rfam motifs in exons using Rfam database which contains information of all well-known RNA families. After we scanned the lincRNAs in base-level, only few lincRNAs contained Rfam motifs. Only four lincRNAs from human and seven from mouse contained at least one such motif (Table 3A and B).



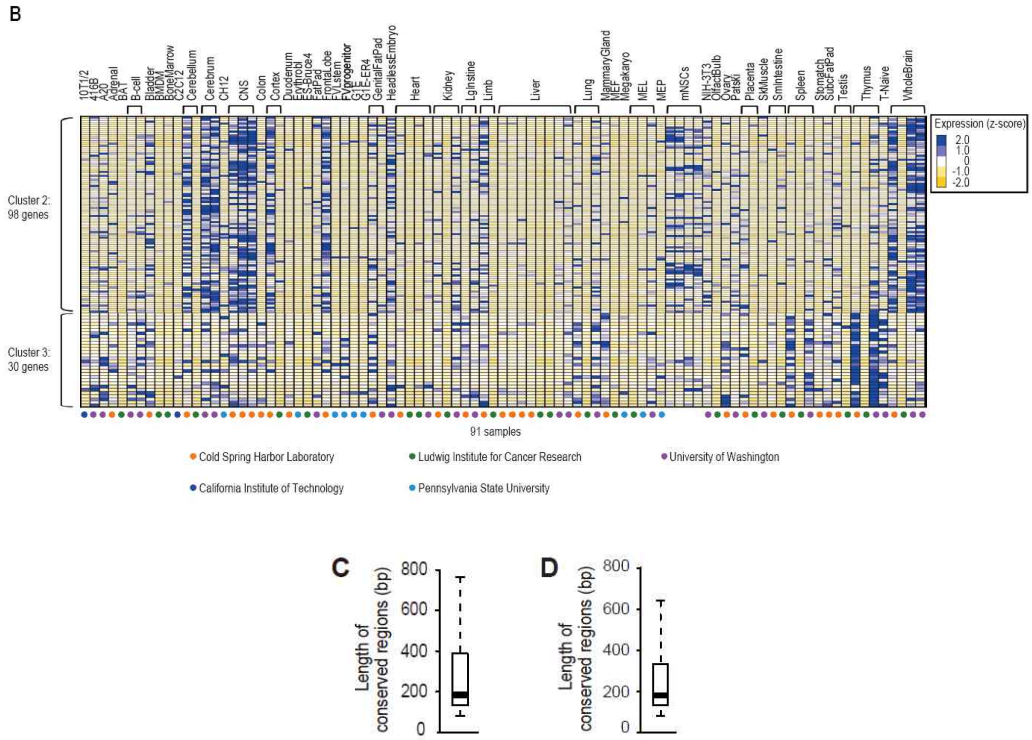


Figure 3. Functional signatures of human and mouse lincRNAs. (A) Expression heatmap of clustered human lincRNAs. (B) Expression heatmap of clustered mouse lincRNAs. (C) A box plot describing the length of conserved regions in human lincRNAs. (D) A box plot demonstrating the length of conserved regions in mouse lincRNAs.

Table 3A. List of 3 RNA families which that are embedded in human lincRNA genes.

Rfam ID	Type	Description	Frequency	Gene ID
tRNA	Gene; tRNA;	tRNA	2	XXbac-BPGBPG55C20.1, RP11-457M11.5
mir-324	Gene; miRNA;	microRNA mir-324	1	RP11-492E3.2
MIR821	Gene; miRNA;	microRNA MIR821	1	CTA-217C2.1

Table 3B. List of 4 RNA families that are embedded in mouse licnRNA genes.

Rfam ID	Type	Description	Frequency	Gene ID
tRNA	Gene; tRNA;	tRNA	4	Gm14267, Gm17549, Gm9828, AU019990
RMST	Gene; lncRNA;	Rhabdomyosarcoma 2 a	1	Gm17341
mir-1937	Gene; miRNA;	microRNA mir-1937	1	Gm14827
adapt33	Gene; lncRNA;	Adaptive response 33	1	5430416N02Rik

Chapter 4. Functional annotation of other species

The same approach for annotating functional evidence of lncRNAs in human and mouse was performed to lncRNAs in other species, like worm (Caenorhabditis elegans), fly (Drosophila melanogaster), and zebrafish (Danio rerio).

Section 1. Characteristics of lincRNAs

To exploit known functional evidence about lincRNAs in worm, fly, and zebrafish, the known lincRNAs of these species were obtained from the previous published papers (Nam and Bartel, 2012; Brown *et al.*, 2014; and Pauli *et al.*, 2012). In worm, 170 long intervening ncRNAs (lincRNAs) from 167 loci were acquired and the size of these genes ranges from 101 bp to 6 kb (Figure 4A). 3088 lincRNAs from 1875 loci in fly, and 1101 lincRNAs from 1101 loci in zebrafish were acquired from those studies (Nam and Bartel, 2012; Brown *et al.*, 2014; and Pauli *et al.*, 2012). The size of these lincRNA genes ranges from 45 bp to 100 kb and 240 bp to 400 kb (Figure 4B and C). All lincRNAs in zebrafish have multi-exonic structures (Figure 4F), since 68.82% and 45.4% of lincRNAs in worm and fly have multi-exonic structures (Figure 4D and E).

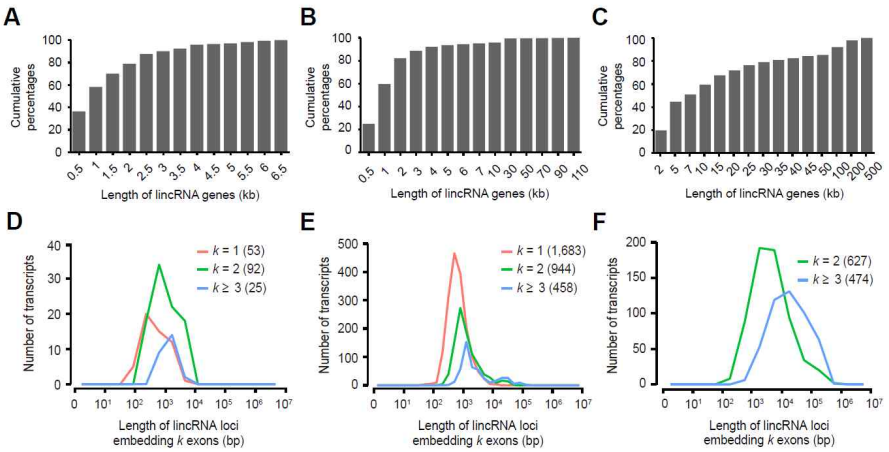
Section 2. Expression profiling

For profiling the lincRNA expression, all public RNA-seq data were obtained from NCBI GEO (GSE22410 and GSE32898) and modENCODE (https://www.modencode.org/), several developmental stages. Reads were aligned to the reference genome (ce6, May 2008 for worm and dm3, Apr. 2006 for fly)) using

Tophat and Bowtie. The in-house python script was used to calculate the expression metric, same as human and mouse. Among 170 worm lincRNAs, 167 genes had RPM greater than 0.1 in at least one stage. After we clusetered these genes using Expander, 143 lincRNAs were grouped into 5 clusters with specific expression signatures (Figure 4G). For fly, 1705 lincRNA genes with RPM greater than 0.1 in at least one stage, except 5 genes, were grouped into 6 clusters (Figure 4H).

Section 3. Conserved domains

Using the same method to annotate conserved domains in human and mouse, we found out that higher percentage of lincRNAs in worm (51.8%) and fly (87.7%) contained the conserved domains in exons, compared with human and mouse. The median length of the conserved domains was approximately 143 in worm and 133 in fly, which are shorter than human and mouse.



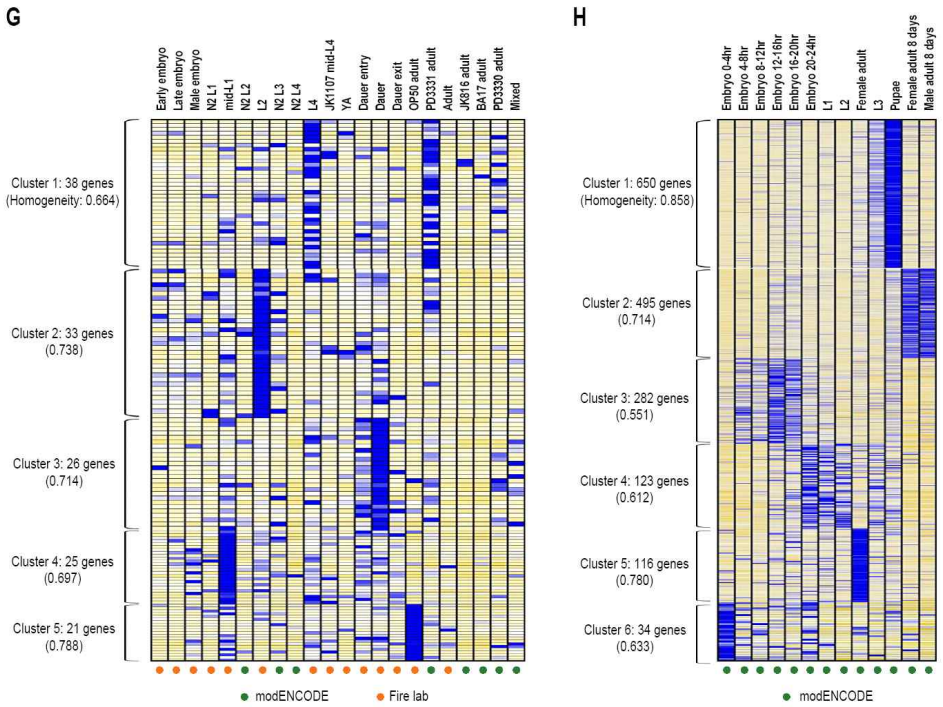


Figure 4. Characteristics of lincRNAs in worm, fly, and zebrafish. (A–C) A cumulative distribution of the lengths of 170 lincRNA genes in worm (A), 1875 lincRNA genes in fly (B), and 1101 lincRNA genes in zebrafish (C). (D–F) Length distribution of lincRNA genes in worm (D), fly (E), and zebrafish (F). Otherwise, the details are as in Figure 2D. (G) Expression heatmap of clustered worm lincRNAs. (H) Expression heatmap of clustered fly lincRNAs.

Chapter 5. Computational design of sgRNAs for CRISPR-Cas9 system

Once we have precise maps for lincRNAs, designing efficient and specific single guide RNAs (sgRNAs) for CRISPR-Cas9 system is a critical step for successful deletion of lincRNA genes. In the system, sgRNAs induce two double strand breaks (DSBs) in target loci. All candidate sgRNA target sites with a protospacer adjacent motif (PAM; NGG) and the regions flanked by these PAMs were initially identified at lincRNA loci (Figure 5A). For efficient and specific deletion, each member of an sgRNA pair requires two properties: high efficiency and specificity. The efficiency and specificity of deletion can be determined by several factors. Here, we investigated features that can be considered to predict sgRNA efficacy, such as sequence and structural features by re-analyzing public data from two independent studies (Wang *et al.*, 2014; Doench *et al.*, 2014). In addition, we also examine the off-target effects of sgRNAs.

Section 1. Sequence feature

First, the sequence features can enhance the efficacy of sgRNAs. To test these features, we measured uncertainty of sequences of 2,449 rDNA-targeting sgRNAs (Wang *et al.*, 2014) and 1,841 CDS-targeting sgRNAs (Doench *et al.*, 2014) and the sgRNA percent rank that provided by the studies considered as the efficiency of the sgRNAs. The uncertainty of guide sequences was correlated with the percent rank. The uncertainty of sequences was measured by Shannon entropy (Schneider, 1997), calculated by the sum of p_i X $\log_2$ p_i, where p_i is the probability of each i = [A or T, G or C]. The maximum entropy, 1.0, is equivalent to that when the G/C ratio is 0.5. As a result, the sgRNAs with higher Shannon

entropy tends to target loci more efficiently in both data (Figure 5B).

Section 2. Structural feature

Similarly with RNAi efficacy, secondary structures of target sites can affect the efficacy of sgRNAs. To predict the secoundary structures of target sites of sgRNAs, we examined thermodynamic stabilities of each sgRNAs in the two sets of data. The free energy values of these sgRNAs were calculated using the program RNAfold (Hofacker, 2003). As a result, the sgRNAs that are predicted to form less stable secondary structures have higher efficacy than others that are predicted to form more stable structures (Figure 5C).

Section 3. A new scoring system (INDEL score)

Based on the observation of sequence and structural features using the independent data set, a new scoring system, INDEL score, that predicted sgRNAs with a high efficacy were developed by incorporating the two features (Figure 5B and C) and the positional nucleotide preferences (Doench *et al.*, 2014) using logistic regression (Figure 5D). This regression model was modeled using 1,841 CDS-targeting sgRNAs (Doench *et al.*, 2014) by the glm function in the R package version 3.2.0. The prediction outperformed the previous scoring system (Figure 5E and F). Taken together, we developed a computational approach that selects sgRNAs with high efficiency, satisfying the following criteria: a Shannon entropy equal to or greater than 0.7, a free energy greater than −6 kcal/mol where the mean percentile is dramatically increased, and an INDEL score equal to or greater than 0.3 (Figure 5B-D).

Section 4. Off-target analysis

Last factor to be considered was off-target effect of sgRNAs. The specificity of sgRNAs can be predicted by the number of off-targets in genome. We counted the number of off-targets of genome-wide sgRNAs using multi-locus alignments. To search for putative sgRNA off-target sites, all candidate guide sequences were mapped to the reference genome using Bowtie version 1.0.0 (Langmead *et al.*, 2009), allowing mismatches. The mapped sites were considered putative off-targes if the following criteria were met: 1) they include a PAM (NGG) to the 3' end of sites and 2) they contain no more than 3 mismatches as compared with the on-target site 3) if the number of mismatches is greater than four and equal to or less than six, off-targets are allowed to have at most one mismatch in the seed region (10 bp from the 3' end of the sgRNA). After sgRNA off-target effects were examined, 63% with equal to or less than three potential off-targets were selected (Figure 5G).

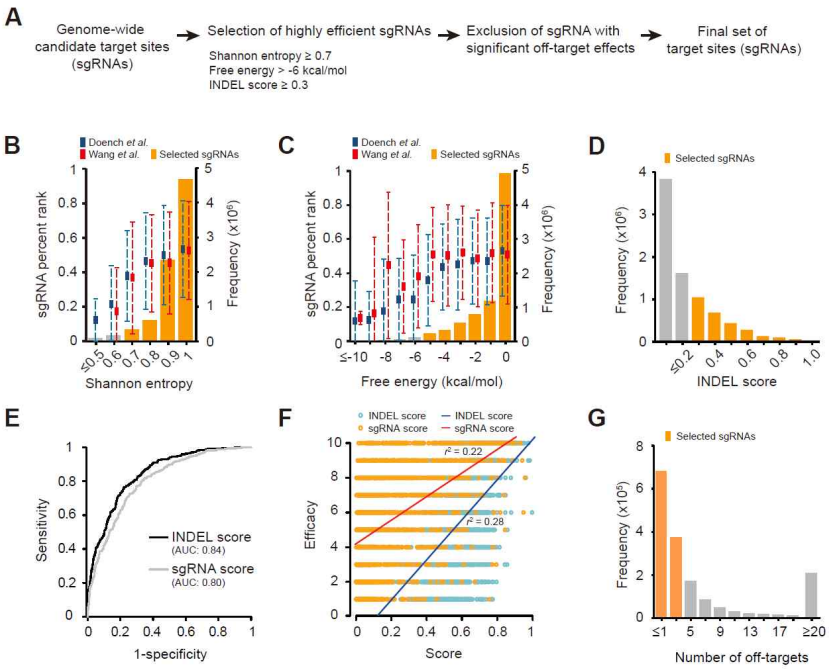


Figure 5. Genome-wide design of candidate sgRNAs for lincRNA deletion. (A) A schematic flow for designing active, specific sgRNAs. In total, 1,669,729 sgRNAs that satisfy criteria for efficiency and specificity were selected for lincRNA knockout. (B) The relationship between the median efficacy (percent rank) of sgRNAs and their Shannon entropy. Blue (Doench *et al.*, 2014) and red (Wang *et al.*, 2014) rectangles represent median values of percent ranks in each entropy bin. The dotted lines represent a standard deviation for each bin. The orange boxes indicate selected sgRNAs that satisfy INDEL's entropy criterion. (C) The relationship between efficacy of sgRNAs and their secondary structures. The orange boxes indicate selected sgRNAs that satisfy INDEL's free energy criterion. Otherwise, the details are as in (B). (D) Distribution of INDEL scores for all sgRNAs. The orange boxes indicate selected sgRNAs that satisfy the INDEL score. (E) Receiver operating characteristic (ROC) curves of INDEL scores (black line) and sgRNA scores (gray line) for 1,841 CDS-targeting sgRNAs (Doench *et al.*, 2014). The curves were drawn with mean values of 10-fold cross-validations. The area under curve (AUC) indicates the performance of the method. (F) Coefficient of determinations of predicted scores by the INDEL (blue) and by Doench *et al.* (red) and the efficacy of 1,841 CDS-targeting sgRNAs.

Chapter 6. Integration of functional annotation and computational designed sgRNA pairs for whole or segmental deletion of lincRNAs

To aid the rational design of sgRNA pairs, we developed an interactive sgRNA design system, called Integrated Design for LincRNA deletion (INDEL) (Figure 6A). The INDEL system integrates the selected sgRNAs and their associated information with lincRNA structural and functional annotations. Using this system, we can design sgRNA pairs, each member of which has the minimum number of off-targets and the best INDEL score, to delete an entire and segmental loci of the human and mouse lincRNA genes. The web system is available at http://big.hanyang.ac.kr/INDEL.

Section 1. Strategies for lincRNA knockout using sgRNA pairs

The INDEL system integrates the selected sgRNAs and their associated information with lincRNA structural and functional annotations (Figure 6A), aiding the design of sgRNA pairs that delete an entire locus (Figure 6B) or segmental regions of lincRNA genes, such as exons (Figure 6C), promoter regions (Figure 6D), conserved regions (Figure 6E) or splice sites (Figure 6F). Among these five approaches, we applied three approaches to validate for human *XIST* knockout.

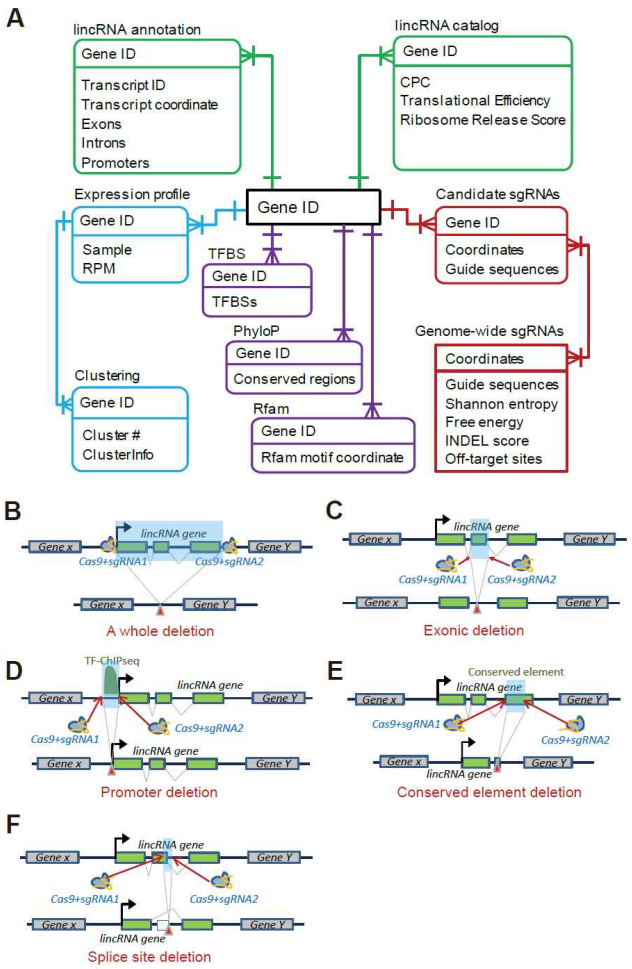


Figure 6. Integrated Design for LincRNA deletion (INDEL) system. (A) A schematic architecture of the INDEL database and its integration of structural and functional lincRNA annotations, such as coding potentials, expression signatures over multiple cell types, conserved elements, and RNA family motifs, as well as all sgRNAs designed to target lincRNA loci. A main key is the lincRNA gene ID, which connects all available tables in the INDEL database. (B-F) CRISPR/Cas9-based deletion strategies for lincRNA knockout using sgRNA pairs. Shown are strategies for deletion of a whole lincRNA locus (B) and segmental deletions of an exon (C), a promoter region (D), a conserved region (E), and a splice site (F).

Section 2. Experimental validation

The *XIST* gene encodes an approximately 19 kb spliced and polyadenylated lincRNA, which is the master regulator of XCI (Figure 7A) (Minkovsky *et al.*, 2012; Borsani *et al.*, 1991; Brockdorff *et al.*, 1992; Brown *et al.*, 1992; Lee, 2011). *XIST* constitutes both human-specific and conserved exons compared to other mammalian species (Figure 7B) (Yen *et al.*, 2007). The role of Xist in mice has been studied through its ablation, which leads to female-specific lethality in early embryonic development (Penny *et al.*, 1996; Marahrens *et al.*, 1997; Lee, 2011). However, the role of *XIST* in human cells has not been elucidated partly because it has been difficult to achieve complete loss of its function due to a lack of proper tools to ablate the gene. Here we attempted to delete both segmental regions of *XIST* and the whole *XIST* gene (including TFBSs) by making paired DSBs using the CRISPR-Cas9 system in human K562 cells (Figure 8A), which have two X chromosomes (Gribble *et al.*, 2000) and express high levels of *XIST* (Figure 8B). To evaluate the performance of sgRNAs which were selected for deletion of *XIST* based on our web tool, we used genomic DNA isolated from K562 cells transfected with plasmids encoding Cas9 (Figure 7C) and sgRNAs (Figure 7D) for the T7E1 assay. The experiments for human *XIST* knockout were performed by Ramu Gopalappa and Suresh Ramakrishna from Hyongbum Kim's Laboratory at Deparment of Pharmacology, Yonsei University College of Medicine.

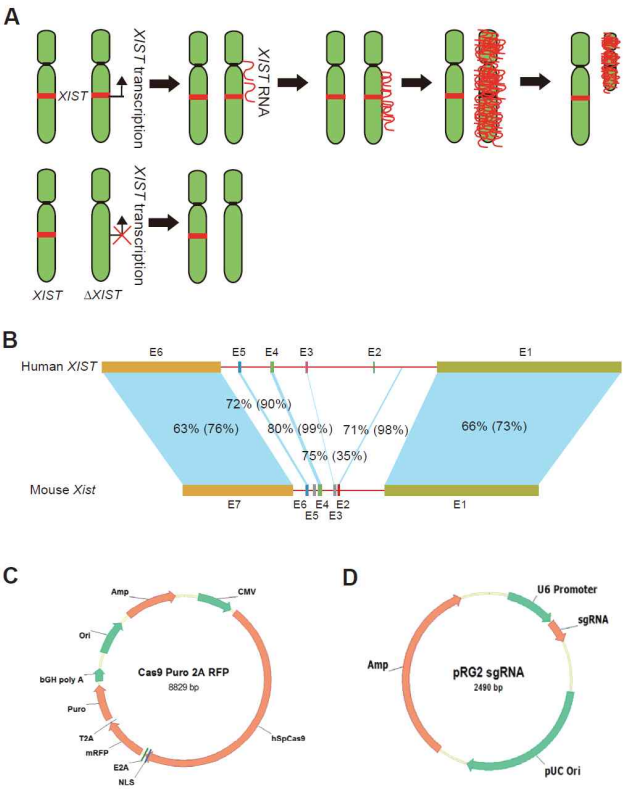


Figure 7. A conserved role of *XIST* RNA and maps of the plasmids encoding Cas9 nuclease and single guide RNA (sgRNA). (A) A schematic representation of *XIST* RNA-driven X chromosome inactivation. RNA transcribed from the *XIST* gene on the X chromosome spreads and creats an *XIST* RNA cloud, which leads to X chromosome heterochromatic formation and silencing. (B) Sequence conservation of human *XIST* and mouse *Xist* exons. Sequences from exons and introns were downloaded from the UCSC genome browser; mouse exons were aligned to human exons and introns using BLAST2 from NCBI (Johnson et al., 2008). The numbers in parentheses are sequence identities of the blocks with the longest query coverage. The vector maps of Cas9-2A-mRFP-2A-Puro (C) and pRG2-sgRNA (D) are shown. sgRNA is transcribed by the U6 promoter (pU6). hSpCas9 (human codon-optimized Cas9 nuclease derived from Streptococcus pyogenes) is expressed by the CMV promoter; E2A, equine rhinitis A virus (ERAV) 2A; mRFP, monomeric red fluorescent protein; T2A, *Thoseaasigna* virus 2A; Puro, Puromycin resistance gene. The ampicillin resistance gene (Amp) enables the selection of the transformed bacterial cells.

For the segmental and whole deletion of XIST, the total 12 sgRNAs were selected to delete partial exon 1 where is conserved, exon 2 and 3, and whole region of *XIST* were transfected in human K562 cells (Figure 8A). Before analysis, the two different sgRNA pairs were tested for each locus and the cells were treated with puromycin to enrich tranfected cells before analysis. For each locus, sgRNA pairs that led to higher mutation frequencies were selected, resulting in mutation frequencies that ranged from 21% to 56% (Figure 8C). These selected sgRNAs were used for subsequent studies. Compared with controls, we observed a polymerase chain reaction (PCR) amplicons of approximately 620, 655, or 1550 bp only in the transfected cells (Figure 8D). Then we determined the deletion frequency at fourteen days after platting the tranfected cells into 96-well plates at an average density of 0.25 cells/well. Fourteen days after plating, individual clones were isolated and analyzed. Gel electrophoresis of PCR products amplified from genomic DNA showed that 31% (5/16), 41% (7/16), and 31% (5/16) of the clones contained the targeted deletions (Figure 8E), suggesting highly efficient deletion. Among the clones containing the deletions, 25% (1/4; Figure 9A and B) for deletion of the 3' part of exon 1, 20% (1/5; Figure 10A and B) for the deletion of exon 2 and 3, and 40% (2/5; Figure 11A and B) for the whole deletion did not contain the wile-type allele, indicating that they were homozygotic knockout clones.

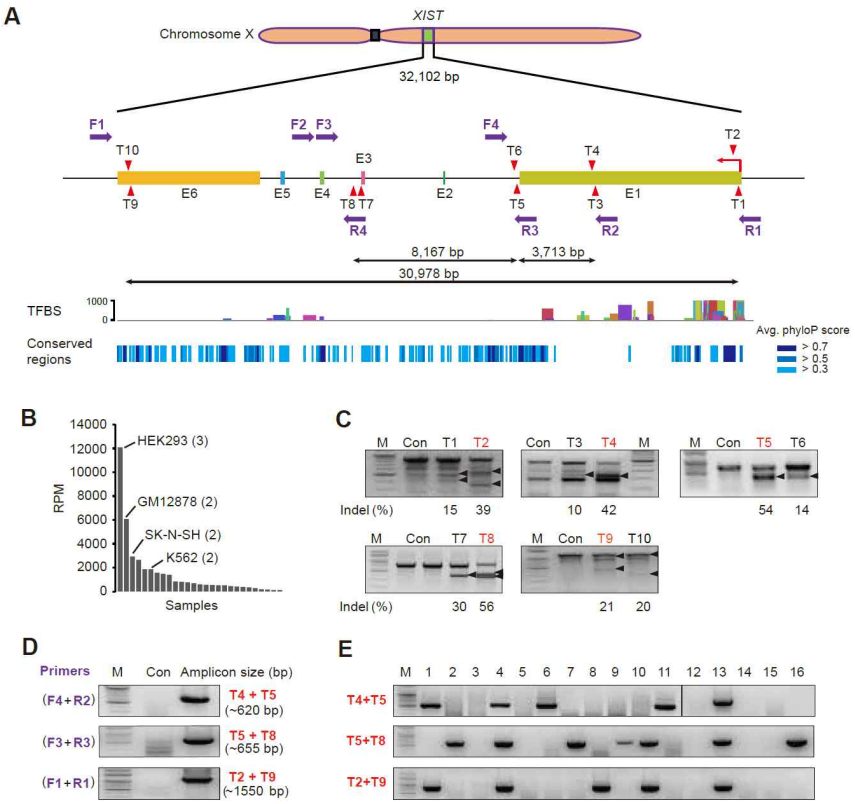


Figure 8. Paired Cas9 nuclease-mediated deletion of the human *XIST* gene. (A) A schematic depicting the human *XIST* gene structure and targeted deletion sites. The six exons are shown. Red arrowheads indicate the sgRNA binding sites (T1-T10). F1-4 and R1-3 represent PCR primers (purple arrows) used for the detection of the wild-type allele and the alleles with targeted deletions. The TFBS track demonstrates the TFBSs precompiled from the ENCODE data. The colored boxes indicate different TFBSs and the height of the box indicates the signal value. Conserved domains are marked with blue boxes. The conserved region track demonstrates the conserved domains greater than 0.3 (light blue), 0.5 (blue), and 0.7 (dark blue). (B) The expression signature (RPMs) of *XIST* RNA measured from ENCODE RNA-seqs. The numbers in parentheses indicate the X chromosome copy number. *XIST* RNA displayed the highest expression in HEK293 cells (12,085 RPM), which contain three copies of the X chromosome, as well as high expression in K562 cells (1,865 RPM), which contain two copies of the X chromosome. (C) T7E1 assay to evaluate the efficiency of each sgRNA. K562 cells were analyzed after transfection with plasmids encoding Cas9 and each sgRNA (T1-T10). Mutation frequencies (indel [%]) were

calculated from the band intensities. The more efficient sgRNAs are highlighted in red. Untransfected cells were used as controls (Con). The arrowheads indicate the expected position of DNA bands cleaved by T7E1. M: Marker lane. (D) PCR-based detection of targeted deletions. Genomic DNA isolated from K562 cells transfected with plasmids encoding Cas9 and the sgRNA pair were subjected to PCR to detect the targeted deletion. The arrowheads indicate the expected position of the amplicons from the region containing the targeted deletion. Untransfected cells were used as the control (Con). The primer and sgRNA pairs are shown on the left and right, respectively. (E) PCR-based evaluation of targeted deletions in single-cell derived clones. PCR was performed using genomic DNA isolated from individual clones derived from a population of K562 cells that were transfected with plasmids encoding Cas9 and sgRNA. The numbers on top represent the individual clone names. M: Marker lane.

We next cloned and sequenced the PCR products derived from the clonal cells with targeted deletions, which corroborated that the two targeted sites were joined and that the intervening segments of ~3,710, 8,170, and 30,980 bp were deleted from *XIST* (Figure 8E, 9E, and 10E). This sequencing also revealed that the two cleavage sites were joined without the generation of small insertions or deletions (indels) in 50% (2/4), 40% (2/5), and 0% (0/7) of the clones, whereas small indels were generated in the junction area in the remaining 50% (2/4), 60% (3/5), and 100% (7/7) of the clones, which is compatible with the error-prone property of non-homologous end joining. Among the four knockout clones, two contained heterozygous mutant alleles (clones 10 and 13 in Figure 11D) in each clone, whereas the other two clones (clone 4 in Figure 9D, clone 2 in Figure 10D) contained homozygous mutant alleles. Of these two, one (clone 2 in Figure 10D) did not contain an indel, whereas another contained homozygous indels.

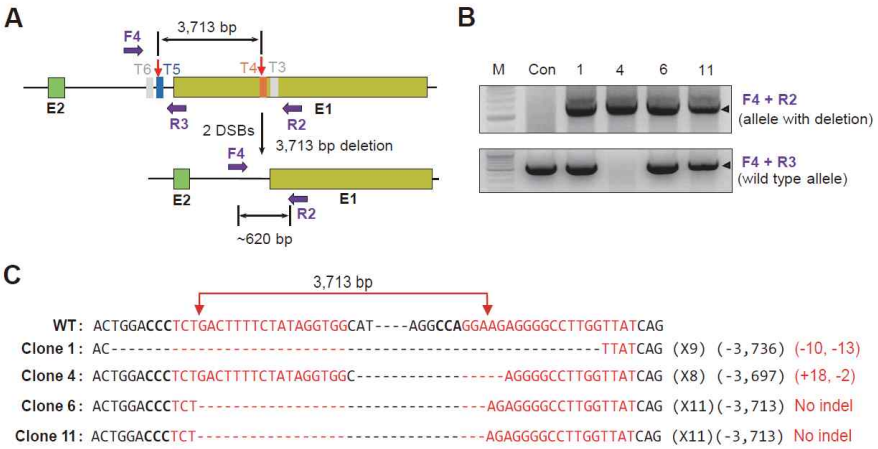


Figure 9. Cas9-induced segmental deletions of *XIST* exon 1. (A) A schematic depicting Cas9-mediated deletion of *XIST* exon 1. Targeted cleavage sites are shown with red arrows. The binding sites of sgRNAs with high efficiency (T4 and T5) are depicted using orange and sky blue squares. Black arrows labeled F4, R2, and R3 represent PCR primers used for the detection of the wild-type allele and the allele with the targeted deletion. (B) Identification of clones containing the *XIST* exon 1 deletion. The wild-type and deleted alleles were detected via PCR using different primer pairs. The arrowheads indicate the expected position of the amplicons. (C) DNA sequences of the *XIST* wild-type (WT) allele and alleles with the exon 1 deletion. Cas9 recognition sites shown in red and the protospace adjacent motif (PAM) sequence is shown in bold characters. The cleavage sites are indicated by arrowheads. Dashes indicate deleted bases. The number in the first set of parentheses represents the number of occurrences (e.g., X8, X9, and X11 indicate how many times each sequence was observed). The sequence length of the consequent large deletion is shown in the second set of parentheses. Indels are shown in the last set of parentheses.

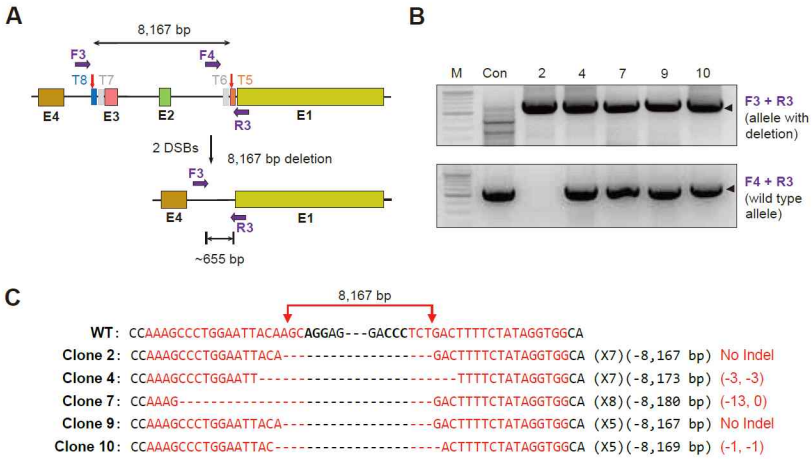


Figure 10. Cas9-induced segmental deletion of *XIST* exons 2 and 3. (A) A schematic depicting Cas9-mediated deletion of *XIST* exons 2 and 3. Targeted cleavage sites are shown with red arrows. The binding sites of sgRNAs with high efficiency (T5 and T8) are depicted using orange and sky blue squares. Purple arrows labeled F3, F4, and R3 represent PCR primers used for the detection of the wild-type allele and the allele with the targeted deletion. (B) Identification of clones containing the *XIST* exon 2 and 3 deletion. Otherwise, the details are as in Figure 9B. (C) DNA sequences of the *XIST* wild-type (WT) allele and alleles with exon 2 and 3 deletions. Otherwise, the details are as in Figure 9C.

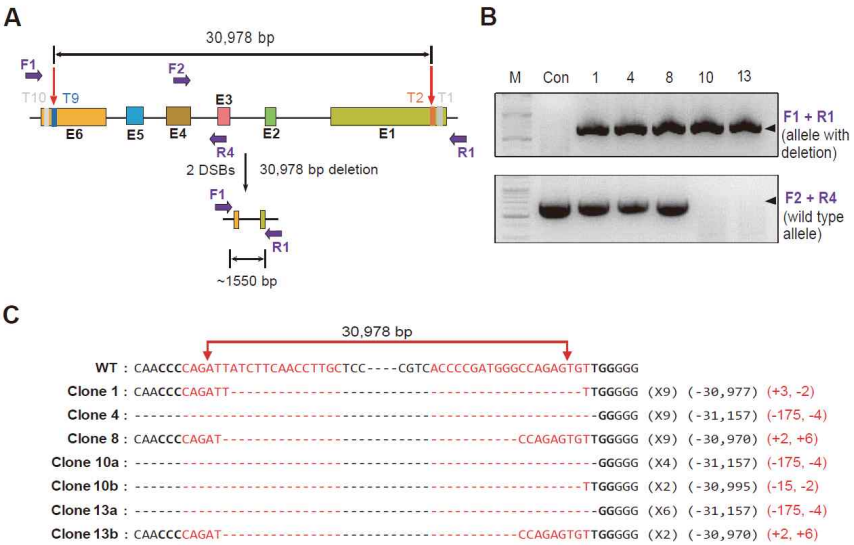


Figure 11. Cas9-induced *en bloc* deletion of *XIST*. (A) A schematic depicting Cas9-mediated deletion of the whole *XIST* gene. The binding sites of sgRNAs with high efficiency (T2 and T9) are depicted using orange and sky blue squares. Purple arrows labeled F1, F2, R1, and R4 represent PCR primers used for the detection of the wild-type allele and the allele with the targeted deletion. (B) Identification of clones containing the whole *XIST* deletion. Otherwise, the details are as in Figure 9B. (C) DNA sequences of the *XIST* wild-type (WT) allele and alleles with the whole *XIST* deletion. Otherwise, the details are as in Figure 9C.

We also explored the effect of the segmental and *en bloc XIST* deletion using fluorescent in situ hybridization (FISH) and RNA-seq. The FISH experiment were performed by Dr. Hongjae Sunwoo and Prof. Jeannie T. Lee from Lee Laboratory at Howard Hughes Medical Institute, Department of Molecular Biology, Massachusetts General Hospital, Department of Genetics, Harvard Medical School. The XIST clouds were not observed in the transfected cells with *en bloc XIST* and segmental deletion causing the loss of exon 2 and 3 (Figure 12C and D). This deletions also were conformed by the expression of *XIST* RNA in each clone. The very low expression of *XIST* RNA was observed in the clone that containing the deletion of exon 2 and 3 and deletion of whole region (Figure 12E), indicating that

XIST expression may be inhibited by deletions of exons 2 and 3 at the transcriptional or posttranscriptional level.. However, recent studies were supported this finding that the stability of the inactive X chromosome is not affected by deletion of *XIST* gene (Brown and Willard, 1994; Csankovszki *et al.*, 1999; Wutz, 2011). However, dysregulation of the expression of several other genes was observed, including X-linked genes. These changes in expression do not appear to be related to off-target effects. This finding is in agreement with recent evidence showing that the inactive X, albeit somewhat stable, is prone to reactivation once Xist is deleted (Yildirim *et al.*, 2013; Anguera *et al.*, 2012; Zhang *et al.*, 2007). In addition, the resulting transcriptome signals by the deletion of exon 2 and 3 were significantly correlated with the clones containing *en bloc XIST* deletion (Figure 12E; correlation = 0.81; $p < 2.2\text{x}10^{-16}$). Although twelve genes of top 40 most variable genes in the segmental and *en bloc XIST* deletion are significantly co-varied ($p < 2.63\text{x}10^{-10}$; hypergeometric test), the *en bloc* deletion also affected the expression of other genes, indicating that the knockout by the deletion of region lacking regulatory DNA elements is necessary to unveil the role of the target lincRNA. Although the transfected cells with deletion of partial exon 1 were observed the XIST clouds, the unusual alternative splicing events occurred in the exon 1 (Figure 12F). Taken together, segmental deletion of certain regions of *XIST* may not lead to its complete functional ablation as previously reported (Senner et al., 2011; Caparros *et al.*, 2002; Wutz *et al.*, 2002).

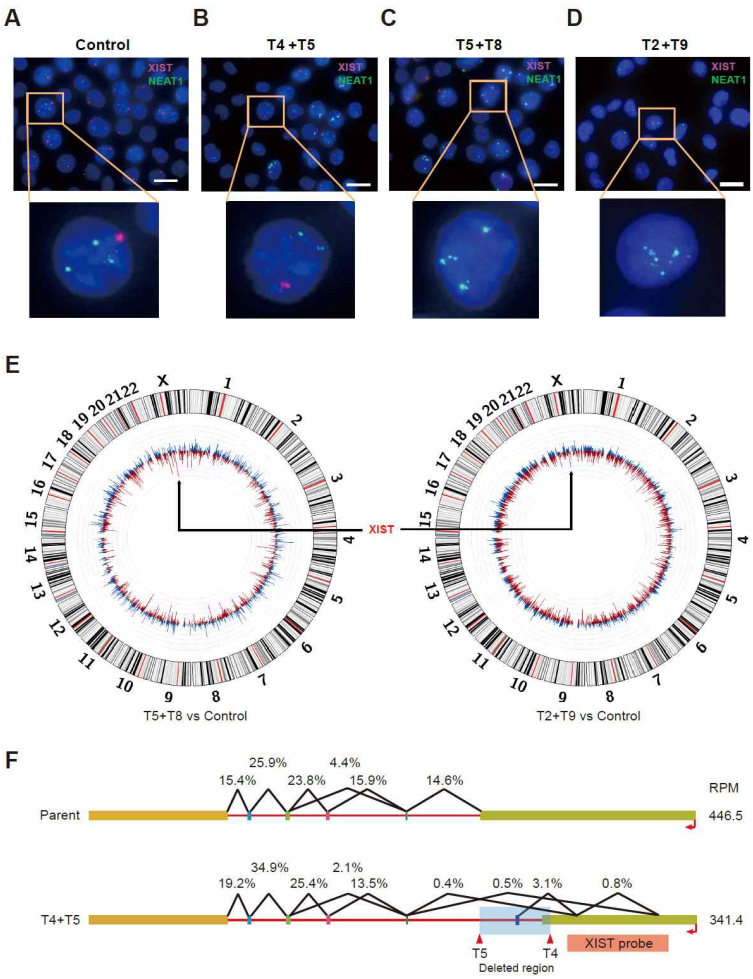


Figure 12. Cas9-mediated deletion of the *XIST* gene leads to its functional loss. (A–D) RNA FISH of *XIST* (pink) and *NEAT1* (green) in parent cells (A; control) and clones with Cas9 nuclease-induced whole (D) or segmental *XIST* deletions (B and C). (E) Transcriptome changes in a clone with a segmental *XIST* deletion (left) and a clone with a whole deletion (right). (F) Summarized diagrams of alternative splicing of *XIST* in parent cells and clones transfected with Cas9 and T4 and T5 sgRNAs (red triangles), which delete a region containing exon 1. The number over each splice junction indicates the percentage of exon junction reads at the splice junction and the rightmost number indicates the expression level (RPM) of the *XIST* gene. A light blue box indicates the deleted region.

Discussion

A large genomic deletion is a straightforward way to completely knockout long non-coding elements but generating such deletions requires a precise genomic map of the sequences encoding such elements as well as their cis-regulatory DNA elements, properly designed, specific, and efficient sgRNA pairs for the creation of paired DSBs, and a joining of the breaks without loss or gain of other functional elements. In this paper, we demonstrated a Cas9 nuclease- and guide RNA-based design for the efficient deletion of a revised set of human and mouse lincRNAs using our integrated design system, INDEL, which was implemented on an interactive web genome-browser. Our INDEL system makes possible, in an unprecidented manner, the design of sgRNAs that 1) are based on improved annotations of human and mouse lincRNA genes, ii) function as pairs for the deletion of both whole and segmental regions of these genes, again based on comprehensive functional annotations, and iii) efficiently generate mutations because of both sequence- and secondary structure-based features.

Using our system, whole and segmental regions of *XIST* were deleted in a highly efficient manner. Knockout of *XIST* did not immediately affect expression of inactivated alleles, supporting the role of *XIST* in the initiation but not maintenance of XCI. XCI maintenance is known to be dependent on DNA methylation on the X chromosome after formation of heterochromatin (Brown and Willard, 1994; Csankovszki, 1999; Wutz, 2011). Previous experiments showed that treatment with the DNA methylation inhibitor 5-AzaC leads to reactivation of X-linked genes, such as *HPRT1, FMR1-AS, and FMR1*, regardless of *XIST* expression (Jones, 1982; Ladd, 2007). In fact, these genes were not reactivated in K562 clones that lacked the *XIST* gene, indicating that knockout of *XIST* did not immediately affect the landscape of DNA methylation in somatic cells (Figure

12E).

Here we selected sgRNA with least potential for off-target cleavages using an *in silico* approach. However, recently published analyses based on high-throughput sequencing showed a large number of Cas9 nuclease-induced hard-to-predict off-target cleavages across the whole genome (Frock, 2015; Tsai, 2015; Wang, 2015). Thus, we cannot rule out the possibility of effects caused by off-target mutations that are hard to predict *in silico*. However, our *in silico* prediction is currently one of the efficient way to select sgRNA with minimal off-target effects given that it requires significant time and effort to evaluate potential off-target sites for a large number of individual sgRNAs using these *in vitro* (Kim, 2015) or *in vivo* (Frock, 2015; Tsai, 2015; Wang, 2015) approaches.

In conclusion, in this study we provide a system for the genome-wide design of sgRNA pairs for targeted *en bloc* and segmental deletion of lincRNA genes; such deletions can lead to lincRNA loss-of-function. The sgRNA design was achieved through an elaborate revision of lincRNA annotations and *in silico* predictions of sgRNA efficiency and off-target effects. These approaches have been experimentally validated as highly efficient in creating *en bloc* and segmental deletions in *XIST*. Our genome-wide sgRNA design will facilitate the application of the CRISPR/Cas9 system to elucidating lincRNA functions.

Detailed methods

lincRNA expression profiling. lincRNA expression was profiled with RNA-seq data from ENCODE (https://www.encodeproject.org/) and NCBI GEO (www.ncbi.nlm,giv/geo), both of which cover over 31 human cell types and 53 mouse cell-types and primary tissues (Tables 1A and B). The RNA-seq libraries were prepared and sequenced from multiple laboratories: California Institute of Technology (Caltech), Cold Spring Harbar Laboratory (CSHL), Genome Institute of Singapore (GIS), and HudsonAlpha Institute for Biotechnology (HudsonAlpha) for human, and Caltech, CSHL, Ludwig Institute for Cancer Research (LICR) and University of Washington (UW) for mouse. Additional RNA-seq data from HeLa, HEK293, Huh7 (GSE52530), K-562 (GSE47998 and GSE34740), and hESC-H1 (GSE51861) cells were also compiled from NCBI GEO. Reads were aligned to the reference genomes (hg19, Feb. 2009 for human and mm9, Jul. 2007 for mouse) using Tophat (version 2.0.6) with mapping parameters "--segment-length 50 -g 1 -i 61 -l 265006 --min-segment-intron 61 --max-segment-intron 265006" for human and "--segment-length 50 -g 1 -i 52 -l 240764 --min-segment-intron 52 --max-segment-intron 240764" for mouse. As expression metric, reads per million (RPM) was calculated using an in-house python script.

Processing of RNA-seq data from *XIST* knockout cells and parent cells. Strand-specific libraries of poly(A)-selected RNAs from *XIST* knockout and parent cells were constructed using a Truseq stranded mRNA prep kit (Illumina), after which RNA-seq was performed. RNA-seq reads were mapped to the human reference genome (hg19) using Bowtie version 1.0.0 with an unique-mapping parameter "-m 1" and using Tophat version 2.0.6 with mapping parameters "--segment-length 50 -g 1 -i 61 -l 265006 --min-segment-intron 61

--max-segment-intron 265006 -r 42 --mate-std-dev 15“. To examine gene expression changes (based on the GENCODE annotation version 19) between parent and knockout cells, logarithm base 2 of RPMs for each gene were calculated.

References

Anguera, M.C., Sadreyev, R., Zhang, Z., Szanto, A., Payer, B., Sheridan, S.D., Kwok, S., Haggarty, S.J., Sur, M., Alvarez, J., Gimelbrant, A., Mitalipova, M., Kirby, J.E., and Lee, J.T. (2012) Molecular signatures of human induced pluripotent stem cells highlight sex differences and cancer genes. Cell Stem Cell 11(1):75-90.

Bassett, A.R., Akhtar, A., Barlow, D.P., Bird, A.P., Brockdorff, N., Duboule, D., Ephrussi, A., Ferguson-Smith, A.C., Gingeras, T.R., Haerty, W., Higgs, D.R., Miska, E.A., and Ponting, C.P. (2014) Considerations when investigating lncRNA function in vivo. Elife 3:e03058.

Batista, P.J. and Chang, H.Y. (2013) Long noncoding RNAs: cellular address codes in development and disease. Cell 152(6):1298-1307.

Borsani, G., Tonlorenzi, R., Simmler, M.C., Dandolo, L., Arnaud, D., Capra, V., Grompe, M., Pizzuti, A., Muzny, D., Lawrence, C., Willard, H.F., Avner, P., and Ballabio, A. (1991) Characterization of a murine gene expressed from the inactive X chromosome. Nature 351(6324):325-329.

Brockdorff, N., Ashworth, A., Kay, G.F., McCabe, V.M., Norris, D.P., Cooper, P.J., Swift, S., and Rastan, S. (1992) The product of the mouse Xist gene is a 15 kb inactive Xspecific transcript containing no conserved ORF and located in the nucleus. Cell 71(3):515-526.

Brown, C.J., Hendrich, B.D., Rupert, J.L., Lafrenière, R.G., Xing, Y., Lawrence, J., and Willard, H.F. (1992) The human XIST gene: analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. Cell 71(3):527-542.

Brown, C.J. and Willard, H.F. (1994) The human X-inactivation centre is not required for 36 maintenance of X-chromosome inactivation. Nature 368(6467):154-156.

Brown, J.B., Boley, N., Eisman, R. May, G.E., Stoiber, M.H., Duff, M.O., Booth, B.W., Wen, J., Park, S., Suzuki, A.M., Wan, K.H., Yu, C., Zhang, D., Carlson, J.W., Cherbas, L., Eads, B.D., Miller, D., Mockaitis, K., Roberts, J., Davis, C.A., Frise, E., Hammonds, A.S., Olson, S., Shenker, S., Sturgill, D., Samsonova, A.A., Weiszmann, R., Robinson, G., Hernandez, J., Andrews, J.,

Bickel, P.J., Carninci, P., Cherbas, P., Gingeras, T.R., Hoskins, R.A., Kaufman, T.C., Lai, E.C., Oliver, B., Perrimon, N., Graveley, B.R., and Celniker, S.E. (2014) Diversity and dynamics of the Drosophila transcriptome. Nature http://dx.doi.org/10.1038/nature12962.

Caparros, M.L., Alexiou, M., Webster, Z., and Brockdorff, N. (2002) Functional analysis of the highly conserved exon IV of XIST RNA. Cytogenet Genome Res 99(1-4):99-105.

Cho, S.W., Kim, S., and Kim, J. (2013) Targeted genome engineering in human cells with the Cas9 RNA-guided endonuclease. Nat Biotechnol 31(3):230-232.

Cong, L., Ran, F., Cox, D. Lin, S., Barretto, R. Habib, N., Hsu, P.D., Wu, X., Jiang, W., Marraffini, L.A., and Zhang, F. (2013) Multiplex genome engineering using CRISPR/Cas systems. Science 339(6121):819-823.

Consortium EP (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489(7414):57-74.

Csankovszki, G., Panning, B., Bates, B., Pehrson, J.R., and Jaenisch, R. (1999) Conditional deletion of Xist disrupts histone macroH2A localization but not maintenance of X inactivation. Nat Genet 22(4):323-324.

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D.G., Lagarde, J., Veeravalli, L., Ruan, X., Ruan, Y., Lassmann, T., Carninci, P., Brown, J.B., Lipovich, L., Gonzalez, J.M., Thomas, M., Davis, C.A., Shiekhattar, R., Gingeras, T.R., Hubbard, T.J., Notredame, C., Harrow, J., and Guigó, R. (2012) The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res 22(9):1775-1789.

Doench, J.G., Hartenian, E., Graham, D.B., Tothova, Z., Hegde, M., Smith, I., Sullender, M., Ebert, B.L., Xavier, R.J., and Root, D.E. (2014) Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. Nat Biotechnol.

Du, Z., Fei, T., Verhaak, R.G., Su, Z., Zhang, Y., Brown, M., Chen, Y., and Liu, X.S. (2013) Integrative genomic analyses reveal clinically relevant long noncoding RNAs in human cancer. Nat Struct Mol Biol 20(7):908-913.

Eissmann, M., Gutschner, T., Hämmerle, M., Günther, S., Caudron-Herger, M., Groß, M., Schirmacher, P., Rippe, K., Braun, T., Zörnig, M., and Diederichs, S. (2012) Loss of the abundant nuclear non-coding RNA MALAT1 is

compatible with life and development. RNA Biol 9(8):1076-1087.

Engreitz, J.M., Pandya-Jones, A., McDonel, P., Shishkin, A., Sirokman, K., Surka, C., Kadri, S., Xing, J., Goren, A., Lander, E.S., Plath, K., and Guttman, M. (2013) The Xist lncRNA exploits three-dimensional genome architecture to spread across the X chromosome. Science 341(6147):1237973.

Engreitz, J.M., Sirokman, K., McDonel, P., Shishkin, A.A., Surka, C., Russell, P., Grossman, S.R., Chow, A.Y., Guttman, M., and Lander, E.S. (2014) RNA-RNA Interactions Enable Specific Targeting of Noncoding RNAs to Nascent Pre-mRNAs and Chromatin Sites. Cell 159(1):188-199.

Esteller, M. (2011) Non-coding RNAs in human disease. Nat Rev Genet 12(12):861-874.

Faulkner, G.J., Kimura, Y., Daub, C.O., Wani, S., Plessy, C., Irvine, K.M., Schroder, K., Cloonan, N., Steptoe, A.L., Lassmann, T., Waki, K., Hornig, N., Arakawa, T., Takahashi, H., Kawai, J., Forrest, A.R., Suzuki, H., Hayashizaki, Y., Hume, D.A., Orlando, V., Grimmond, S.M., and Carninci, P. (2009) The regulated retrotransposon transcriptome of mammalian cells. Nat Genet 41(5):563-571.

Frock, R.L., Hu, J., Meyers, R.M., Ho, Y.J., Kii, E., and Alt, F.W. (2015) Genome-wide detection of DNA double-stranded breaks induced by engineered nucleases. Nat Biotechnol 33(2):179-186.

Geisler, S. and Coller, J. (2013) RNA in unexpected places: long non-coding RNA functions in diverse cellular contexts. Nat Rev Mol Cell Biol 14(11):699-712.

Gribble, S.M., Roberts, I., Grace, C., Andrews, K.M., Green, A.R., and Nacheva, E.P. (2000) Cytogenetics of the chronic myeloid leukemia-derived cell line K562: karyotype clarification by multicolor fluorescence in situ hybridization, comparative genomic hybridization, and locus-specific fluorescence in situ hybridization. Cancer Genet Cytogenet 118(1):1-8.

Gutschner, T., Baas, M., and Diederichs, S. (2011) Noncoding RNA gene silencing through genomic integration of RNA destabilizing elements using zinc finger nucleases. Genome Res 21(11):1944-1954.

Gutschner, T. and Diederichs, S. (2012) The hallmarks of cancer: a long non-coding RNA point of view. RNA Biol 9(6):703-719.

Guttman, M., Donaghey, J., Carey, B.W., Garber, M., Grenier, J.K., Munson, G., Young, G., Lucas, A.B., Ach, R., Bruhn, L., Yang, X., Amit, I., Meissner, A.,

Regev, A., Rinn, J.L., Root, D.E., and Lander, E.S. (2011) lincRNAs act in the circuitry controlling pluripotency and differentiation. Nature 477(7364):295-300.

Guttman, M. and Rinn, J.L. (2012) Modular regulatory principles of large non-coding RNAs. Nature 482(7385):339-346.

Han, J., Zhang, J., Chen, L., Shen, B., Zhou, J., Hu, B., Du, Y., Tate, P.H., Huang, X., and Zhang, W. (2014) Efficient in vivo deletion of a large imprinted lncRNA by CRISPR/Cas9. RNA Biol 11(7):829-835.

Ho, T.T., Zhou, N., Huang, J., Koirala, P., Xu, M., Fung, R., Wu, F., and Mo, Y.Y. (2014) Targeting non-coding RNAs with the CRISPR/Cas9 system in human cell lines. Nucleic Acids Res.

Hofacker, I.L. (2003) Vienna RNA secondary structure server. Nucleic Acids Res 31(13):3429-3431.

Hu, W., Yuan, B., Flygare, J., and Lodish, H.F. (2011) Long noncoding RNA-mediated antiapoptotic activity in murine erythroid terminal differentiation. Genes Dev 25(24):2573-2578.

Ingolia, N.T., Ghaemmaghami, S., Newman, J.R., and Weissman, J.S. (2009) Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science 324(5924):218-223.

Ingolia, N.T., Lareau, L.F., and Weissman, J.S. (2011) Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. Cell 147(4):789-802.

Ji, Q., Zhang, L., Liu, X., Zhou, L., Wang, W., Han, Z., Sui, H., Tang, Y., Wang, Y., Liu, N., Ren, J., Hou, F., and Li, Q. (2014) Long non-coding RNA MALAT1 promotes tumour growth and metastasis in colorectal cancer through binding to SFPQ and releasing oncogene PTBP2 from SFPQ/PTBP2 complex. Br J Cancer 111(4):736-748.

Jia, H., Osak, M., Bogu, G.K., Stanton, L.W., Johnson, R., and Lipovich, L. (2010) Genome-wide computational identification and manual annotation of human long noncoding RNA genes. RNA 16(8):1478-1487.

Jinek, M., East, A., Cheng, A., Lin, S., Ma, E., and Doudna, J. (2013) RNA-programmed genome editing in human cells. Elife 2:e00471.

Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S., Madden, T.L. (2008) NCBI BLAST: a better web interface. Nucleic Acids Res 36(Web Server issue):W5-9.

Jones PA, Taylor SM, Mohandas T, Shapiro LJ (1982) Cell cycle-specific reactivation of an inactive X-chromosome locus by 5-azadeoxycytidine. Proc Natl Acad Sci U S A 79(4):1215-1219.

Kapusta, A., Kronenberg, Z., Lynch, V.J., Zhuo, X., Ramsay, L., Bourque, G., Yandell, M., and Feschotte, C. (2013) Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. PLoS Genet 9(4):e1003470.

Karolchik, D., Armstrong, J., Barber, G.P., Casper, J., Clawson, H., Diekhans, M., Dreszer, T.R., Fujita, P.A., Guruvadoo, L., Haeussler, M., Harte, R.A., Heitner, S., Hickey, G., Hinrichs, A.S., Hubley, R., Karolchik, D., Learned, K., Lee, B.T., Li, C.H., Miga, K.H., Nguyen, N., Paten, B., Raney, B.J., Smit, A.F., Speir, M.L., Zweig, A.S., Haussler, D., Kuhn, R.M., and Kent, W.J. (2014) The UCSC Genome Browser database: 2014 update. Nucleic Acids Res 42(Database issue):D764-770.

Kelley, D. and Rinn, J. (2012) Transposable elements reveal a stem cell-specific class of long noncoding RNAs. Genome Biol 13(11):R107.

Kim, D., Bae, S., Park, J., Kim, E., Kim, S., Yu, H.R., Hwang, J., Kim, J.I., and Kim, J.S. (2015) Digenome-seq: genome-wide profiling of CRISPR-Cas9 offtarget effects in human cells. Nat Methods 12(3):237-243, 231 p following 243.

Kim, H. and Kim, J.S. (2014) A guide to genome engineering with programmable nucleases. Nat Rev Genet 15(5):321-334.

Kim, Y.K., Wee, G., Park, J., Kim, J., Baek, D., Kim, J.S., and Kim, V.N. (2013) TALEN-based knockout library for human microRNAs. Nat Struct Mol Biol 20(12):1458-1464.

Klattenhoff, C.A., Scheuermann, J.C., Surface, L.E., Bradley, R.K., Fields, P.A., Steinhauser, M.L., Ding, H., Butty, V.L., Torrey, L., Haas, S., Abo, R., Tabebordbar, M., Lee, R.T., Burge, C.B., and Boyer, L.A. (2013) Braveheart, a long noncoding RNA required for cardiovascular lineage commitment. Cell 152(3):570-583.

Kong, L., Zhang, Y., Ye, Z.Q., Liu, X.Q., Zhao, S.Q., Wei, L., and Gao, G. (2007) CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. Nucleic Acids Res 35(Web Server issue):W345-349.

Ladd, P.D., Smith, L.E., Rabaia, N.A., Moore, J.M., Georges, S.A., Hansen, R.S., Hagerman, R.J., Tassone, F., Tapscott, S.J., and Filippova, G.N. (2007) An antisense transcript spanning the CGG repeat region of FMR1 is upregulated in premutation carriers but silenced in full mutation individuals. Hum Mol Genet 16(24):3174-3187.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10(3):R25.

Lee, J.T. (2011) Gracefully ageing at 50, X-chromosome inactivation becomes a paradigm for RNA and chromatin control. Nat Rev Mol Cell Biol 12(12):815-826.

Lee, J.T. (2012) Epigenetic regulation by long noncoding RNAs. Science 338(6113):1435-1439.

Li, L. and Chang, H.Y. (2014) Physiological roles of long noncoding RNAs: insight from knockout mice. Trends Cell Biol 24(10):594-602.

Licatalosi, D.D., Mele, A., Fak, J.J., Ule, J., Kayikci, M., Chi, S.W., Clark, T.A., Schweitzer, A.C., Blume, J.E., Wang, X., Darnell, J.C., and Darnell, R.B. (2008) HITS-CLIP yields genome-wide insights into brain alternative RNA processing. Nature 456(7221):464-469.

Liu, Y., Luo, D., Zhao, H., Zhu, Z., Hu, W., and Cheng, C.H. (2013) Inheritable and precise large genomic deletions of non-coding RNA genes in zebrafish using TALENs. PLoS One 8(10):e76387.

Mali, P., Yang, L., Esvelt, K.M., Aach, J., Guell, M., DiCarlo, J.E., Norville, J.E., and Church, G.M. (2013) RNA-guided human genome engineering via Cas9. Science 339(6121):823-826.

Marahrens, Y., Panning, B., Dausman, J., Strauss, W., and Jaenisch, R. (1997) Xist-deficient mice are defective in dosage compensation but not spermatogenesis. Genes Dev 11(2):156-166.

Margulies EH, Blanchette, M; NISC Comparative Sequencing Program, Haussler, D., and Green, E.D. (2003) Identification and characterization of multi-species conserved sequences. Genome 37 Res 13(12):2507-2518.

Mili, S. and Steitz, J.A. (2004) Evidence for reassociation of RNA-binding proteins after cell lysis: implications for the interpretation of immunoprecipitation analyses. RNA 10(11):1692-1694.

Minkovsky, A., Patel, S., and Plath, K. (2012) Concise review: Pluripotency and the transcriptional inactivation of the female Mammalian X chromosome. Stem Cells 30(1):48-54.

Nam, J.W. and Bartel, D.P. (2012) Long noncoding RNAs in C. elegans. Genome Res 22(12):2529-2540.

Nam, J.W., Rissland, O.S., Koppstein, D., Abreu-Goodger, C., Jan, C.H., Agarwal, V., Yildirim, M.A., Rodriguez, A., and Bartel, D.P. (2014) Global analyses of the effect of different cellular contexts on microRNA targeting. Mol Cell 53(6):1031-1043.

Ng, S.Y., Johnson, R., and Stanton, L.W. (2012) Human long non-coding RNAs promote pluripotency and neuronal differentiation by association with chromatin modifiers and transcription factors. EMBO J 31(3):522-533.

Pauli, A., Valen, E., Lin, M.F., Garber, M., Vastenhouw, N.L., Levin, J.Z., Fan, L., Sandelin, A., Rinn, J.L., Regev, A., and Schier, A.F. Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. Genome Res 22(3):577-591.

Penny, G.D., Kay, G.F., Sheardown, S.A., Rastan, S., and Brockdorff, N. (1996) Requirement for Xist in X chromosome inactivation. Nature 379(6561):131-137.

Schneider, T.D. (1997) Information content of individual genetic sequences. J Theor Biol 189(4):427-441.

Senner, C.E., Nesterova, T.B., Norton, S., Dewchand, H., Godwin, J., Mak, W., and Brockdorff, N. (2011) Disruption of a conserved region of Xist exon 1 impairs Xist RNA localisation and X-linked gene silencing during random and imprinted X chromosome inactivation. Development 138(8):1541-1550.

Sharan, R. and Shamir, R. (2000) CLICK: a clustering algorithm with applications to gene expression analysis. Proc Int Conf Intell Syst Mol Biol 8:307-316.

Simon, M.D., Pinter, S.F., Fang, R., Sarma, K., Rutenberg-Schoenberg, M., Bowman, S.K., Kesner, B.A., Maier, V.K., Kingston, R.E., and Lee, J.T. (2013) High-resolution Xist binding maps reveal two-step spreading during X-chromosome inactivation. Nature 504(7480):465-469.

Sun, L., Goff, L.A., Trapnell, C., Alexander, R., Lo, K.A., Hacisuleyman, E., Sauvageau, M., Tazon-Vega, B., Kelley, D.R., Hendrickson, D.G., Yuan, B., Kellis, M., Lodish, H.F., and Rinn, J.L. (2013) Long noncoding RNAs regulate adipogenesis. Proc Natl Acad Sci U S A 110(9):3387-3392.

Tsai, S.Q., Zheng, Z., Nguyen, N.T., Liebers, M., Topkar, V.V., Thapar, V., Wyvekens, N., Khayter, C., Iafrate, A.J., Le, L.P., Aryee, M.J., and Joung, J.K. (2015) GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. Nat Biotechnol 33(2):187-197.

Ulitsky, I., Maron-Katz, A., Shavit, S., Sagir, D., Linhart, C., Elkon, R., Tanay, A., Sharan, R., Shiloh, Y., and Shamir, R. (2010) Expander: from expression microarrays to networks and functions. Nat Protoc 5(2):303-322.

Ulitsky, I., Shkumatava, A., Jan, C.H., Sive, H., and Bartel, D.P. (2011) Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. Cell 147(7):1537-1550.

Ulitsky, I. and Bartel, D.P. (2013) lincRNAs: genomics, evolution, and mechanisms. Cell 154(1):26-46.

van Bakel, H., Nislow, C., Blencowe, B.J., and Hughes, T.R. (2010) Most "dark matter" transcripts are associated with known genes. PLoS Biol 8(5):e1000371.

Wang, T., Wei, J.J, Sabatini, D.M, and Lander, E.S. (2014) Genetic screens in human cells using the CRISPR-Cas9 system. Science 343(6166):80-84.

Wang, X., Wang, Y., Wu, X., Wang, J., Wang, Y., Qiu, Z., Chang, T., Huang, H., Lin, R.J., and Yee, J.K. (2015) Unbiased detection of off-target cleavage by CRISPR-Cas9 and TALENs using integrase-defective lentiviral vectors. Nat Biotechnol 33(2):175-178.

Wutz, A., Rasmussen, T.P, and Jaenisch, R. (2002) Chromosomal silencing and localization are mediated by different domains of Xist RNA. Nat Genet 30(2):167-174.

Wutz, A. (2011) Gene silencing in X-chromosome inactivation: advances in understanding facultative heterochromatin formation. Nat Rev Genet 12(8):542-553.

Xiang, J.F., Yin, Q.F., Chen, T., Zhang, Y., Zhang, X.O., Wu, Z., Zhang, S., Wang, H.B., Ge, J., Lu, X., Yang, L., and Chen, L.L. (2014) Human colorectal cancer-specific CCAT1-L lncRNA regulates long-range chromatin interactions at the MYC locus. Cell Res 24(5):513-531.

Yen, Z.C., Deakin, J.E., Gilbert, C., Robinson, T.J., Graves, J.A., and Waters, P.D. (2007) A cross-species comparison of X chromosome inactivation in Eutheria. Genomics 90(4):453-463.

Yildirim, E., Kirby, J.E., Brown, D.E., Mercier, F.E., Sadreyev, R.I., Scadden, D.T., and Lee, J.T. (2013) Xist RNA is a potent suppressor of hematologic cancer in mice. Cell 152(4):727-742.

Zhao, J., Sun, B.K., Erwin, J.A., Song, J.J., and Lee, J.T. (2008) Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. Science 322(5902):750-756.

Zhang, L.F., Huynh, K.D., and Lee, J.T. (2007) Perinucleolar targeting of the inactive X during S phase: evidence for a role in the maintenance of silencing. Cell 129(4):693-706.

Zhang, C. and Darnell, R.B. (2011) Mapping in vivo protein-RNA interactions at singlenucleotide resolution from HITS-CLIP data. Nat Biotechnol 29(7):607-614.

Zhou, Y., Zhu, S., Cai, C., Yuan, P., Li, C., Huang, Y., and Wei, W. (2014) High-throughput screening of a CRISPR/Cas9 library for functional genomics in human cells. Nature 509(7501):487-491.

국문요지

최근에 개발된 크리스퍼 유전자 가위는 특정 유전자가 위치한 염기서열을 절단시키므로 유전자의 기능을 없앤다. 유전자의 기능 연구를 위해 원액세포에서 유래된 크리스퍼 유전자 가위 기술이 도입되고 있다. 돌연변이에 의해 격자이동이 유발이 되면서 유전자 발현을 저해시켜 특정 유전자의 기능을 없어지게 하는데 이러한 방법은 단백질 코딩 유전자들 대상으로만 적용되어 왔다. Long intergenic non-coding RNAs (lincRNAs)의 기능 연구에 크리스퍼 유전자 가위 기술을 적용하기 위해서는 여러 다양한 특징들을 통합시켜야 한다. 본 논문에서는 크리스퍼 유전자 가위를 이용하여 lincRNA 녹아웃 라이브러리를 전산 설계하였다.

공개된 데이터베이스인 GENCODE에서 동정한 lincRNAs을 CAGE-seq과 polyA-seq 데이터를 이용하여 보다 정확하게 유전자 구조를 동정하였다. 그 다음 여러 조건을 적용한 후, 인간 lincRNAs는 총 5,882개 그리고 마우스에서는 822개의 lincRNAs를 얻었다. 동정된 lincRNAs의 기능 주석을 위해 ENCODE project 에서 얻은 실험 데이터를 이용하여 조직 또는 세포 특이적으로 발현되는 lincRNAs를 조사하였다. 또한, 기능적으로 주요할 것이라고 예측되는 conserved elements를 포함하고 있는 부분을 동정하고 알려진 RNA family motifs 존재여부는 확인하였다. 동일한 접근법을 사용하여 인간과 마우스뿐만 아니라, 예쁜 꼬마선충, 초파리, 그리고 제프라 피쉬를 대상으로도 lincRNAs 기능 주석을 하였다. 또한, 기능 주석된 부분이나 전체 유전자를 절단하기 위한 효율성 높은 single guide RNAs (sgRNAs)를 설계하였다. 전체 유전체에서 존재하는 sgRNAs 중 효율성 높은 sgRNAs만을 선정하기 위해서는 sequence 그리고 구조적 특징들을 고려해야한다. 최종적으로 특정 lincRNA을 절단하기 위해 필요한 모든 정보를 얻을 수 있는 녹아웃 라이브러리를 구축하였다.

Acknowledgement

First of all, I thank Prof. Nam who provided insight and scientific guidance. I also thank members of the Kim and Dr. Sunwoo and Dr. Jeannie T. Lee from Lee lab in Harvard Medical School who performed and allowed me to include the results of the experiment of *XIST* knockout in this thesis.

I would also like to thank all members of BIG lab in Hanyang University for helpful comments and discussions.

연구 윤리 서약서

본인은 한양대학교 대학원생으로서 이 학위논문 작성 과정에서 다음과 같이 연구 윤리의 기본 원칙을 준수하였음을 서약합니다.

첫째, 지도교수의 지도를 받아 정직하고 엄정한 연구를 수행하여 학위논문을 작성한다.

둘째, 논문 작성시 위조, 변조, 표절 등 학문적 진실성을 훼손하는 어떤 연구 부정행위도 하지 않는다.

셋째, 논문 작성시 논문유사도 검증시스템 "카피킬러"등을 거쳐야 한다.

2015년06월30일

학위명 : 석사

학과 : 생명과학과

지도교수 : 남진우

성명 : 이현주 

한 양 대 학 교 대 학 원 장 귀 하

Declaration of Ethical Conduct in Research

I, as a graduate student of Hanyang University, hereby declare that I have abided by the following Code of Research Ethics while writing this dissertation thesis, during my degree program.

"First, I have strived to be honest in my conduct, to produce valid and reliable research conforming with the guidance of my thesis supervisor, and I affirm that my thesis contains honest, fair and reasonable conclusions based on my own careful research under the guidance of my thesis supervisor.

Second, I have not committed any acts that may discredit or damage the credibility of my research. These include, but are not limited to : falsification, distortion of research findings or plagiarism.

Third, I need to go through with Coppykiller Program(Internet-based Plagiarism-prevention service) before submitting a thesis."

JUNE 30, 2015

Degree : Master

Department : DEPARTMENT OF LIFE SCIENCE

Thesis Supervisor : Nam, Jin-Wu

Name : LEE HYEON JOO (signature)