SUPPLEMENTAL METHODS

Chromatin Map Data

Chromatin data for H3K4me3 and H3K36me3, for mouse Embryonic Stem Cells (mES), mouse Embryonic Fibroblasts (MEF), and mouse neural precursor cells (NPC) were taken from Mikkelsen et al. 2007 and were downloaded from (ftp://ftp.broad.mit.edu/pub/papers/chipseq/). Chromatin data in mouse lung fibroblasts and human lung fibroblasts were generated as previously described [1].

Identifying K4-K36 Enriched Domains

To identify regions of enriched chromatin marks we employ a sliding window approach: we slide windows, score each window based on the number of ChIP fragments, compute a threshold for significance, and use the significant windows to define intervals. Specifically: (**i**) We fix a window size *w* and slide it across each position of the genome. For each position, we compute a score, S_w, as the number of reads aligned within the window. (**ii**) To identify windows that have significantly more reads than would be expected by chance, we define a null model based on the randomization of read locations across the genome. This null model is estimated as a Poisson distribution where λ is defined as the number of reads in the library divided by the number of possible non-overlapping windows of size *w*. (**iii**) Given the null model, we choose a threshold *T* on the score such that the genome-wide probability that the Score S_w exceeds the threshold T by chance is less than 0.01 ($Prob(S_w>T)<0.01$). We therefore cannot compute this probability exactly, since the scores S_w occur in overlapping windows they are not independent values or multiple testing corrected values. We therefore estimate it genome-wide across overlapping windows using the scan statistic procedure [Naus and Wallenstein]. Therefore, windows that

pass this threshold *T* are significantly enriched after multiple testing correction. (**iv**) We retain only windows that pass this threshold *T*, and merge overlapping significant windows into a single contiguous interval. We refine the boundaries of this interval by taking the maximum contiguous subsequence. (**vi**) To generalize for multiple window sizes, we compute a threshold for each window size separately and repeat the above procedure, merging windows of different sizes. (**v**) Finally, we score each interval and test if it is significantly enriched using the same scan statistic approach introduced above. The result is a set of intervals and their *p*-values.

To identify the intervals that encode intergenic K4-K36 domains we applied this approach to independently find K4 and K36 regions. We filtered all K4 and K36 regions that overlapped with known annotations (as described below). We identified all K4 and K36 intervals that were adjacent. To define a K4-K36 domain we required that the interval from the K4 region through the end of the K36 region was significantly enriched for K36 using the same scan statistic approach. We then filter the list by regions that are at least 5Kb in length.

All results were produced in the March 2006 (MM8) freeze of the Mouse genome. Code to perform this analysis is available from the authors (MG).

Filtering Gene Lists

We filtered the list of K4-K36 domains to eliminate all regions annotated as containing a protein coding gene in mouse or orthologous protein coding genes in human, rat, or dog.

We obtained the list of all human protein coding genes as determined by Clamp et al. 2007 in the Human genome (Hg17) from (http://www.broad.mit.edu/mammals/alpheus/data/) and used the

liftOver (http://genome.ucsc.edu/cgi-bin/hgLiftOver) tool to identify their orthologous location in the mouse genome (MM8). We also used the list of allRefSeq protein coding genes (MM8) along with all RefSeq genes annotated in the Human (Hg18), Rat (Rn4), and Dog (canFam2) genomes. All refSeq gene lists were obtained from the UCSC Genome Browser (http://hgdownload.cse.ucsc.edu/downloads.html). The liftOver tool was similarly used to place genes from other species in the mouse genome (MM8). In our analysis, we eliminated all regions that overlapped any portion of a protein coding locus, including introns, exons, and UTRs. We also excluded all regions that overlap a known miRNA gene obtained from the MIRBASE database.

RNA Preparation and Sources

We purchased total RNA for mouse lung, brain, testes, and ovary (Ambion). We isolated RNA from Mouse whole embryo, forelimb, and hindlimb from developmental time points E9.5, 10.5 and 13.5. These mice were generated using timed mating embryo isolation and dissection. We obtained mES, mEF, and NPC RNA extracted from cell lines using the Qiagen RNAEasy Kit. Bone Marrow dendritic cells were extracted as previously described[2], and stimulated with various ligands (see below). We extracted RNA after 6 hours using the Qiagen RNAEasy Kit.

Tilling array design, hybridization and analysis.

High resolution DNA tiling arrays containing 2.1 million features were designed on the Nimblegen platform (HD2) to represent a random sampling of ~400 intergenic K4-K36 domains identified in the mouse genome. Total RNA from mES, mLF, NPC and mEF was amplified using poly-dT and labeled as described[3]. Arrays were hybridized and washed according to the

Nimblegen protocols and kits (Roche/Nimblegen). Array image files were processed using Nimblescan (Roche/Nimblegen) and arrays were normalized by mean centering the data.

A second array was designed on the Nimblegen platform (HD1) arrays containing 300,000 and representing ~150 K4-K36 domains. We hybridized mES, mEF, mLF, NPC, BMDC, TLR2, TLR4, TLR9, lung (Ambion), brain (Ambion), testis (Ambion), ovary (Ambion), whole embryo, forelimb, and hindlimb to this array from developmental time points E9.5, 10.5 and 13.5. Total RNA was amplified and labeled for array as described[3]. For both arrays we tiled across all Hox genes as well as handful of other genes as controls.

A third array was designed on the Nimblegen platform (HD1) to tile all lincRNA exons and control regions. Total RNA from $p53^{+/+}$ and $p53^{-/-}$ mEFs across the doxirubicin time course (see below) was amplified, labelled, and hybridized to the array as described[3].

All hybridization data is publicly available from GEO under accession number GSEXXX.

Determining Transcribed Segments From Tiling Arrays

To identify transcribed regions of K4-K36 domains, we hybridized poly-A RNA to a tiling microarray. We developed a statistical algorithm to identify peaks in hybridization, representing likely exons in a mature transcript.

We normalized the data by dividing each probe value by the average probe intensity across the array. We scanned the K4-K36 domains using sliding windows of width *w*. We computed a score defined as the sum of the normalized probe intensities. To determine the significance of this score we permuted the intensity values assigned to each probe and recalculated the statistic. We took the value for each permutation as the maximum score obtained for any random region. We performed 1000 permutations and assigned a multiple testing corrected p-value to each region based on its rank within this distribution. All regions with a p-value less than 0.05 were retained. After determining the transcribed segments from each sample on the array, we defined exons as the union of all bases covered by a transcribed segment.

RNA blot analysis.

RNA blot analysis was performed on Ambion first choice RNA blots (Ambion). The blots contained RNA from various mouse tissues including heart, brain, liver, spleen, kidney, whole embryo, lung, thymus, testes, and ovary. Probes were designed to selected lincRNA exons, as determined by tiling arrays, and hybridized to the RNA blot. Probes were prepared by PCR of genomic regions followed by biotin incorporation using the North2South ® Biotin Labeling Kit (Pierce). Probes were hybridized to the RNA blot for 14-15 hrs using the North2South ® Hybridization Kit (Pierce). The resulting chemiluminescence was detected using a CCD camera. The probes were made by RT-PCR, the primers and corresponding genomic locus is detailed in supplemental table 6.

RT-PCR

RT-PCR analysis was performed on cDNA libraries made from total RNA from mouse embryo (13.5), lung, brain, MEF, NPC, and ES cells reverse transcribed using Superscript II (Invitrogen) and a poly-dT /random hexamer primer mix.

To validate the presence of individual lincRNA exons and their connectivity within a locus we designed primers within and across exon boundaries using the *Primer3* computer program. PCR was performed as previously described[11] on reverse transcribed cDNAs. We performed a negative control using a no RT reaction and a positive control using the mouse GAPDH gene. The PCR products were analyzed by gel electrophoresis. To confirm splicing across exons, the PCR products were purified with QIAquick PCR Clean-up kits (Qiagen) and then sequenced, using the forward primer. To characterize apparent alternative splicing, the products were run on 2% NuSieve agarose (Lonza) gels and the multiple bands purified with a QIAquick Gel Extraction kit (Qiagen) and sequenced. The primers used are detailed in supplemental table 6.

Multiple Species Alignments

All multiple species alignments were the MULTIZ alignments obtained from the UCSC genome browser (build MM9, http://hgdownload.cse.ucsc.edu/goldenPath/mm9/multiz30way/).

Coding Potential

We tested for protein coding potential of K4-K36 domains by determining the maximum CSF[4,5] score observed across the entire genomic locus. We downloaded the alignments from UCSC and computed the CSF scores across sliding windows of 90 base pairs. We then scanned all 6

possible reading frames in each window. After computing a score for each window, we defined the 'max CSF score' for a K4-K36 domain to be the maximum observed score across the region.

We also computed a 'normalized CSF score' for each individual exon. The 'normalized CSF score' for each exon was defined to be the CSF score for each exon divided by the nucleotide length of the exon.

PhastCons Enrichment Within K4-K36 Domains and Promoter Regions

We downloaded the phastConsElements30way from the UCSC Genome Browser (MM9). We counted the number of phastCons elements within each K4-K36 domain as well as the number of these elements within random, size matched, genomic regions. We constructed a distribution based on the random genomic regions. A p-value was computed based on the rank of the K4-K36 domain's rank within the random genomic distribution. This statistic was similarly applied to the promoter regions of lincRNAs.

Computing Pi Constraint

To detect sequence constraint within large ncRNAs, we chose to use a method that explicitly models the rate of mutation as well as the level of constraint. This is especially relevant for detecting constrained sequences in noncoding regions of the genome since many of these sites are known to be degenerate and can tolerate mutations between certain nucleotides.

Briefly the method we used to identify purifying selection uses a probabilistic neutral model of evolution. Given a phylogenetic tree *T* and a substitution rate matrix *Q*, constrained regions will be evident because they are poor fits to the neutral model. In this framework, selection can be apparent in 2 ways either through contraction of the tree length that depends on the intensity of selection (ω) or through a mutation pattern (π) that is not concordant with the rate matrix.

We compute a log-odds score, Pi LOD score, which is the estimate of the sequence evolution compared to neutrally evolving sequences. Sitewise LOD score estimation provides low sensitivity to determine conservation, we therefore integrated across multiple bases. We chose 12mers based on empirically testing the tradeoff between sensitivity and specificity for various kmers. Since the estimation of functional constraint is site specific, we can determine the log-odds score for a region by adding the log-odds scores for each base contained in the region. (Garber et al. 2008, *in preparation*). When the total tree length is less than 1 a Pi LOD score is not computed.

Exon Conservation and K4-K36 Pi LOD Enrichment

To identify functional constraint within exons of large ncRNAs, we analyzed each exon separately. We computed the Pi LOD score for each 12-mer contained within the exon. We took the maximum 12-mer value for each exon. In order to normalize for the size differences between different exons we computed a size matched random score. To do this we randomly generated size matched regions of the genome and divided the observed LOD score by the average LOD score from the random regions. This normalization procedure produces a score for each exon in the genome that reflects a size-independent level of constraint on each exon. The

Observed/Expected score can be interpreted as an enrichment level of the LOD score compared with the genomic average. The distributions of this normalized score were then compared among multiple different classes of genomic units, specifically protein coding introns, exons, and untranslated regions (UTRs), as well as known large non-coding RNAs and non-coding cDNA sequences. This statistic is robust to detecting regions of the genome that, while highly constrained in sequence, are not necessarily highly conserved over the entirety of the region. We performed the same analysis for the K4-K36 domain, using 75nt windows.

CAGE and RNA Pol2 Enrichment

For each promoter region, we computed the number of CAGE tags or ChIP-Seq reads for PolII. We compared the number of aligned reads in the promoters to the number of aligned reads in random regions of similar size (excluding repetitive regions of the genome). We computed enrichment with a wilcoxon rank sum statistic between the promoters and random genomic DNA.

CAGE data were downloaded from http://fantom31p.gsc.riken.jp/cage/download/mm5/ and the regions were mapped to the MM8 build using the liftOver tool (http://genome.ucsc.edu/) CAGE scores were computed by summing the number of reads in each tag cluster (Carninci et al. 2006).

RNA Polymerase II ChIP-Seq data was generated as previously described[1] in mES cells.

Oct4/Nanog Enrichments in ES-specific lincRNAs

We used data generated by Loh et al 2006[6]. Briefly, Chromatin Immunoprecipitation (ChIP) was performed using antibodies against Oct4 and Nanog in mES cells. The resulting library was sequenced using 454 sequencing and the 'paired end reads' were mapped to the genome. We downloaded the read clusters mapped on the mouse genome (build MM5) from http://www.nature.com/ng/journal/v38/n4/suppinfo/ng1760_S1.html. We used the liftOver tool (http://genome.ucsc.edu/) to place the reads on the MM8 build of the mouse genome. We defined binding events as clusters with at least 3 independent ChIP sequencing reads, as described in Loh et al. 2006.

In order to determine the enrichment of intergenic Oct4/Nanog binding sites we counted the number of intergenic Oct4/Nanog binding sites that overlapped with a K4me3 peak in the four cell types. Next we counted how many of these regions coincided with the promoter of a lincRNA in the four cell types. We then counted the number of these lincRNA promoter binding events in ES cells and the number that had strong enrichment levels specifically in ES cells. A hypergeometric statistic was applied to determine if the intergenic binding of Sox2 and Oct4 was enriched at lincRNA promoter regions (K4) compared to other intergenic non-lincRNA K4 regions.

Luciferase Reporter Assay

We amplified individual regions of the lincRNA-Sox2 promoter using AccuPrime *Pfx* polymerase (Invitrogen) and cloned the products into the pCR 2.1TOPO vector (Invitrogen). Each region was subsequently cloned into pGL3 firefly Luciferase Reporter Vector (Promega).

293T cells were transiently transfected in triplicate using FuGENE 6 transfection reagent (Roche) and analyzed 24 hours post-transfection by Promega Dual-Luciferase Reporter Assay kit. Analysis was performed using the Veritas Microplate Luminometer system. Expression of the promoter regions was detected by firefly luciferase activity and was determined by obtaining the relative value compared to the transfection control plasmid (CMV *Renilla* luciferase).

Comparison with Previous Transcript Maps

We downloaded the cDNAs sequenced by the FANTOM consortium from (ftp://fantom.gsc.riken.jp/FANTOM3/). We defined two sets of FANTOM transcripts: the first was the ncRNA conservative set, as provided on their site, and the second was all FANTOM cDNA transcripts. We computed significant overlap between the genomic locus of a lincRNA and a FANTOM unit by asking how much of a K4-K36 domain was covered by a FANTOM unit and how much of a FANTOM unit was covered by a K4-K36 unit. We identified all cases in which a transcript overlapped at least 25% of a K4-K36 domain or vice versa. We performed a similar analysis between exons determined by our tiling arrays and FANTOM exons.

Protein Coding Gene Expression Profiles

We obtained Affymetrix 430 2.0 mouse gene expression data for all RNA samples profiled on our lincRNA array. For ES, MEF, NPC (GSE8024) and brain, lung, testis, and ovary (GSE9954) arrays were already available in the Gene Expression Omnibus (GEO) and in these cases we downloaded the data. For Forelimb, Hindlimb, and Whole Embryo for days 9.5, 10.5, and 13.5, we generated our own data using Affymetrrix 430 2.0 arrays. For dendridic cells we generated data for, unstimualted, TLR2 stimulated, TLR4 stimulated, and TLR9 stimulatedcells using

Affymetrrix 430A arrays (RNA isolated as mentioned above). All data were deposited in GEO (GSE XXX) and are publicly available.

Correlation Matrix Clustering

We generated a correlation matrix between lincRNAs by computing the Pearson correlation coefficient between all pairs of lincRNAs. A matrix was constructed whoe entries are the correlation coefficients. This matrix was clustered and visualized using the Gene Pattern platform for integrative genomics (http://genepattern.broad.mit.edu/) using a Euclidian distance metric and complete linkage clustering[7]. The same procedures were used to produce, cluster, and visualize the lincRNA-Protein coding gene matrix and the lincRNA-Functional Term matrix.

Gene Set Enrichment Analysis and Functional Term Clustering

Gene Set Enrichment Analysis was performed as previously described[8]. Briefly, we used each lincRNA as a profile, computedthe Pearson correlation for each protein coding gene and then ranked the protein coding genes by their correlation coefficient. The rank of these genes was used to identify significant gene sets, using the weighted Kolmogorov–Smirnov (KS) test[8]. Gene sets were permuted 1000 times to obtain FDR corrected p-values. We constructed an association matrix between lincRNAs and terms. We then performed biclustering on this matrix to identify significant lincRNAs associated with functional terms. Biclusters were obtained using the Samba algorithm implemented in Expander software package (http://www.cs.tau.ac.il/~rshamir/expander/expander.html).

Identifying Differentially Expressed Genes in DNA Damage Stimulated Cells

Tp53$^{LSL/+}$ heterozygous mice were intercrossed and fibroblasts were derived from p53$^{LSL/LSL}$ and p53$^{+/+}$ embryos as decribed previously[9]. Sub-confluent cultures were infected on two consecutive days with adenoviruses expressing green fluorescent protein (AdGFP) or Cre recombinase (AdCre) (University of Iowa Genetics Core Facility). Cells were then seeded overnight into 10 cm dishes and treated with 500 nM doxorubicin (Sigma) for the indicated time course. Cells were harvested into Trizol reagent (Invitrogen) and total RNA was extracted for subsequent analysis as described[10].

In parallel, cells were harvested for analysis of p53 protein expression. A monoclonal antibody to mouse p53 (Gift from Kristian Helin) was used for protein blotting and detected by enhanced chemiluminescence (GE Healthcare) per manufacturer's instructions. Hsp90 monoclonal antibody served as a loading control (BD Biosciences).

We identified differentially expressed genes, protein coding and lincRNA, using the Patterns from Gene Expression (http://www.cbil.upenn.edu/PaGE/) program[11]. Briefly, we determined differential expression between p53$^{+/+}$ MEFs compared to p53$^{-/-}$ MEFs at paired times (paired t-test). We filtered the list by genes that were specifically induced across the time points.

Motif Enrichments

Motifs were represented by Position Weight Matrix (PWM) downloaded from the TRANSFAC matrix database v8.3 (http://www.gene-regulation.com/pub/databases.html)[12]. Specifically, for the P53 motif we used the Transfac P53.01 matrix. Given a PWM, for each nucleotide position in a promoter, we calculated an affinity score defined as the log likelihood (LOD score) for observing the sequence given the PWM versus a given random genomic background. We then found the best conserved motif instance over the entire promoter region for each PWM. An

instance was considered conserved if its conservation score was in the top 5% of the genome distribution.

We computed this score for each lincRNA promoter and computed enrichment of the motif for our experimentally determined set compared with all lincRNA promoters. To ensure that enrichment was not due to nucleotide bias within the promoter, we shuffled the PWM and computed enrichment for the true PWM compared to the shuffled PWMs. Enrichment was computed using a two-sided Wilcoxon rank-sum test between the set and the background. We then computed an FDR to correct for testing of multiple PWMs.

Functional Enrichment of lincRNA's neighbor genes

To test for positional bias in lincRNA locations we tested for enrichment of lincRNA neighbors with specific functional annotations. We identified all immediate 3' and 5' neighbors of the 1248 (conservative) mouse lincRNA which resulted in 1703 unique genes. We used Mouse genome build MM8 and the RefSeq gene coordinates to identify nearest neighboring genes. We used a hypergeometric test to calculate an enrichment p-value for each Gene Ontology (GO) term annotated in the mouse. GO annotatiuons were taken from the Gene Ontology consortium (http://www.geneontology.org/)[13]. We filtered GO terms containing at least 5 genes and no more than 500. We identified 56 significant terms (FDR < 0.05) accounting for ~515 neighboring genes (Figure 1c; Supplementary Dataset 2). Specifically, we observe a significant over representation of a set of manually curated ($p< 5 \times 10^{-21}$) and known transcription factors ($p<10^{-24}$). To ensure that the transcription factor enrichments identified were not due to positional biases of TFs in the genome, we generated random sets of "lincRNAs" controlling for intergenic spacing and identified their neighboring genes. This method provides a genome-wide randomization model accounting for variable gene density in the genome. We applied GO analysis to these random set and calculated the proportion of each GO annotation enriched. We

generated 100 random permutation sets of which 90% had no significant GO terms enriched. Of the remaining 10% fewer than 3 terms were observed in any set and transcription factor activity was never observed. We performed similar analyses with the DAVID tool (http://david.abcc.ncifcrf.gov/) and confirm all of these results[14].

Bone marrow dendritic cell (BMDC) cultures

Bone marrow was harvested from 6-8 week old female mice and cultured for 6 days in GM-CSF[2] supplemented medium. Non-adherent cells were sorted using anti-CD11c-beads (Miltenyi Biotech) according to manufacturers guidelines. CD11c positive cells where replated $1.5*10^6$ cells/plate on day 7. BMDCs were left untreated or stimulated with 100 ng/ml LPS for 6 hours or stimulated with 250 ng/ml Pam3CSK4 for 6 hours (TLR2 stimulation) or with CpG oligonucleotide 1uM for 6 hours (TLR9 stimulation) or with poly-inosine:cytosine (polyI:C) 2ug/ml for 6 hours (TLR3 stimulation) . Cells were then collected by scraping and RNA was purified using the miRNAEasy RNA isolation kit (Qiagen). RNA integrity was verified using bioanalyzer (Agilent).

Real-time quantitative PCR.

cDNA was generated by the use of High-Capacity cDNA Archive Kit (Applied Biosystems). Real-time PCR assays were performed using SYBR Green I as a fluorescent dye on a lightCycler 480 (Roche), according to the manufacturer's guidelines. Experiments were carried out in triplicate, and relative gene expression was normalized to glyceraldehyde-3-phosphate dehydrogenase (GAPDH) RNA levels. Real-time PCR primer pairs for protein coding genes were designed using ProbeLibrary (https://www.roche-applied-science.com/sis/rtpcr/upl/index.jsp), primer pairs for lincRNA were designed using primer3

(http://frodo.wi.mit.edu/) with similar settings. Primer sequences are available in Supplementary Table 6.

1. Mikkelsen, T.S. *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448,** 553-560 (2007).
2. Palliser, D. *et al.* A role for Toll-like receptor 4 in dendritic cell activation and cytolytic CD8+ T cell differentiation in response to a recombinant heat shock fusion protein. *J Immunol* **172,** 2885-2893 (2004).
3. Rinn, J.L. *et al.* Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* **129,** 1311-1323 (2007).
4. Lin, M.F. *et al.* Revisiting the protein-coding gene catalog of Drosophila melanogaster using 12 fly genomes. *Genome Res* **17,** 1823-1836 (2007).
5. Lin, M.F., Deoras, A.N., Rasmussen, M.D. & Kellis, M. Performance and scalability of discriminative metrics for comparative gene identification in 12 Drosophila genomes. *PLoS computational biology* **4,** e1000067 (2008).
6. Loh, Y.H. *et al.* The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nature genetics* **38,** 431-440 (2006).
7. Reich, M. *et al.* GenePattern 2.0. *Nature genetics* **38,** 500-501 (2006).
8. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* **102,** 15545-15550 (2005).
9. Ventura, A. *et al.* Cre-lox-regulated conditional RNA interference from transgenes. *Proceedings of the National Academy of Sciences of the United States of America* **101,** 10380-10385 (2004).
10. Rinn, J.L., Bondre, C., Gladstone, H.B., Brown, P.O. & Chang, H.Y. Anatomic demarcation by positional variation in fibroblast gene expression programs. *PLoS genetics* **2,** e119 (2006).
11. Grant, G.R., Liu, J. & Stoeckert, C.J., Jr. A practical false discovery rate approach to identifying patterns of differential expression in microarray data. *Bioinformatics (Oxford, England)* **21,** 2684-2690 (2005).
12. Matys, V. *et al.* TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic acids research* **34,** D108-110 (2006).
13. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics* **25,** 25-29 (2000).
14. Dennis, G., Jr. *et al.* DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol* **4,** P3 (2003).