Integrative Annotation of Human Large Intergenic Non-Coding RNAs

Reveals Global Properties and Specific Subclasses

Moran N. Cabili, Cole Trapnell, Loyal Goff, Magdalena Koziol, Barbara Tazon-Vega, Aviv Regev*, John L. Rinn*

* These authors contributed equally to this work.

Supplementary Material

Library preparation protocol 2
RNA-Seq read mapping 3
RNA-Seq transcriptome assembly 3
lincRNA classification pipeline 4
Stringent set classification 7
Classification of Transcripts of Uncertain Coding Potential (TUCP) 8
Assembly performance estimation on protein coding genes 8
Chromatin state of actively transcribed genes 9
Chromatin state of enhancers 9
Random permutation model of intergenic transcripts 10
Functional enrichment analysis of coding gene sets 11
Estimating expression abundance: note 11
Normalization of expression vectors for tissue specificity calculation 12
K- means clustering of tissue expression patterns 12
Expression correlation of lincRNAs and their neighboring genes 13
Refined alignment of human lincRNAs and their mouse orthologs 14
Sequence conservation level estimation in novel transcripts with potential coding capacity 14
lincRNAs in disease associated regions 15
Supplementary tables 16
Supplementary Methods

Library preparation protocol

RNA was extracted from in-house cultured cell lines (lung and foreskin fibroblasts) using TRizol (Invitrogen) following the manufacturer instructions. After extraction, a DNAse treatment was applied using Turbo DNAse (Ambion) and a second RNA extraction with TRizol was performed. Human tissue total RNA samples were obtained from Ambion.

All samples were sequenced using a standard Illumina mRNA-Seq Protocol. The standard mRNA-seq library preparations for both the 16 individual tissues and the 8 additional samples, was done using the “Illumina mRNA-Seq Sample Prep Kit” (Part #RS-100-0801). Briefly, poly A+ RNA was purified from 100ng of total RNA with oligo-dT beads. Purified mRNA was then fragmented with divalent cations under elevated temperature. First strand cDNA synthesis was performed with random hexamers and reverse transcriptase. Second strand cDNA synthesis was performed using RNAseH and DNA PolI. Following cDNA synthesis, the double stranded products were end repaired, a single “A” base was added, and Illumina PE adaptors were ligated onto the cDNA products. The ligation products were purified using gel electrophoresis. The target size range on the gel for these libraries was ~300 bp for the Human Body Map samples and ~350-500 for the 8 additional samples, such that the final library for sequencing would have cDNA inserts with sizes of ~200 bp or 160-380 bp long, respectively. Following gel purification, the adapter ligated cDNA was then amplified with 15 cycles of PCR for the Human Body Map samples and 14 cycles for the 8 additional samples.
**RNA-Seq read mapping**

All sequenced reads were aligned to the human genome (NCBI 37, Hg19) using the spliced read aligner TopHat version V1.1.4 (Trapnell et al. 2009). Briefly, using a two step mapping processes, TopHat first uses Bowtie (Langmead et al. 2009) to align reads that are directly mapped to the genome (with no gaps). It then determines the possible location of gaps in the alignment based on canonical and non-canonical splice sites flanking the aligned reads. Finally, it uses gapped alignments to align the reads that were not aligned by Bowtie in the first step. For this analysis, we used 2 iterations of TopHat alignments to maximize the use of splice site information derived across all samples. To this end, the reads from each sample were first aligned using the paired-end alignment option when possible (default parameters and ‘min-anchor=5’, ‘min-isoform-fraction =0’). Next, we generated a pooled splice-sites (or “junctions”) file by combining all predicted splice sites across all alignments. We then re-aligned each sample using the pooled splice-sites file (using ‘raw-juncs’ and ‘no-novel-juncs’ parameters).

**RNA-Seq transcriptome assembly**

The transcriptome of each sample was assembled from the mapped reads separately by both Scripture (Guttman et al. 2010) and Cufflinks (Trapnell et al. 2010). Briefly, Cufflinks and Scripture are ab-initio transcriptome assemblers that reconstruct the transcriptome based on RNA-Seq reads that were aligned to the genome using a spliced read aligner (e.g. TopHat) (Haas and Zody 2010). Both methods use spliced read information to determine exons connectivity, but with two different approaches. Scripture uses a statistical segmentation model to distinguish expressed loci from experimental noise and uses spliced reads to assemble expressed segments. It reports all statistically significantly expressed isoforms in a given locus. Cufflinks uses a
probabilistic model to simultaneously assemble and quantify the expression level of a minimal set of isoforms that provides a maximum likelihood explanation of the expression data in a given locus.

The main difference between the output of the two methods results from the different approach each assembler uses when reconstructing low abundant transcripts (Garber et al. 2011). Whereas Scripture employs a significance threshold on expression levels to filter out transcripts and reports all significantly expressed isoforms (Guttman et al. 2010), Cufflinks will report a minimal set of isoforms that explains the expression of a given locus without employing such a threshold (Trapnell et al. 2010). The observed discrepancy across lincRNA loci is comparable to that previously observed for low expressed protein-coding genes reconstructed by these assemblers (Garber et al. 2011).

To obtain transcriptome assemblies from the read alignments, Cufflinks version V1.0.0 was run using default parameters (and ‘min-frags-per-transfrag=0’) and Scripture version 1.0 was run with default parameters, but without using paired-end information (to avoid conflicts that occurred while running Cufflinks abundance estimation mode in later steps). In case several lanes were available per sample the corresponding aligned reads were pooled to a single source file prior to running the assemblers.

**lincRNA classification pipeline**

After obtaining a unique set of assembled isoforms from all processed tissue assemblies and known annotations we ran the set through the following filters:
(1) **Size selection.** We excluded single exon transcripts and ones smaller than 200 bases.

(2) **Minimal read coverage threshold.** We ran Cufflinks with its transcript abundance calculation mode to estimate the read coverage of each transcript across the 24 tissues and cell types. We eliminated transcripts with a maximal coverage below 3 reads per base. This coverage threshold was set by optimizing the sensitivity and specificity of identifying full length vs. partial length transcripts of protein coding genes annotated in RefSeq or non-coding genes annotated in UCSC. To this end, we calculated the number of full length and partial length transcripts identified at each coverage threshold (considering the maximal coverage threshold in which a transcript was identified across all tissues). We used area under the curve (AUC) calculations to determine the optimal threshold for the coding and non-coding sets and took their average as the final threshold.

(3) **Filter of known non-lincRNAs annotations.** We eliminated all transcripts that had an exon overlapping a transcript from any of the following sets: (a) coding genes annotated in RefSeq, UCSC or GENCODE 4, (b) microRNA, tRNAs, snoRNAs, rRNAs annotated in Ensembl, (c) pseudogenes. See **Supplementary Table 5** for specific details on each annotation set.

(4) **Positive coding potential threshold.** We estimated for each transcripts the degree of evolutionary pressure on sequence substitutions acting to preserve an open reading frame. To this end, we scored the coding potential of all remaining transcripts using PhyloCSF (phylogenetic codon substitution frequency) (Lin et al. 2011). Briefly, PhyloCSF determines whether a multi-
species nucleotide sequence alignment in a specific locus is more likely to represent a protein coding than a non-coding transcript. To do so it applies a probabilistic model that examines the over-representation of evolutionary signatures characteristic of alignments of conserved coding regions, such as the high frequencies of synonymous codon substitutions and conservative amino acid substitutions. We ran PhyloCSF using a multiple sequence alignment of 29 mammalian genomes (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/multiz46way/) to obtain the best scoring ORF greater than 29 amino acids across all three reading frames. We excluded from our catalog all transcripts that contained such an open reading frame with a PhyloCSF score greater than 100. The PhyloCSF score threshold was determined to optimize the specificity and sensitivity in correctly classifying coding and non-coding transcripts annotated in RefSeq (RefSeq coding and RefSeq lincRNAs; see below and in Supplementary Table 5). PhyloCSF =100 corresponds to a false negative rate of 6% for coding genes (6% of coding genes are classified as non-coding) and a false positive rate of ~10% (9.5% of noncoding transcripts are classified as coding).

(5) Known protein domain filter. We evaluated which of the remaining transcripts contained a known protein coding domain. To this end, we translated each transcripts in all three possible frames and used HMMER-3 (Finn et al. 2011) to identify occurrence of any of the 31,912 known protein family domains documented in the Pfam database (release 24; used both PfamA and PfamB ; (Finn et al. 2008)). Briefly, HMMER-3 uses hidden Markov models (HMMs) to scan each amino acid sequence and classify whether it resembles any of the known domains in the database. We have excluded transcripts with a Pfam hit that was declared significant.
(6) Intergenic classification. We classified all the remaining transcripts that did not overlap the genomic region of a known non-lincRNA annotation as potential lincRNAs. All transcripts that passed all the filters above are available in Supplementary Dataset 1.

To derive a unique set of lincRNAs that includes previous annotations we used Cuffcompare to integrate the RNA-Seq derived lincRNAs with the predetermined set of lincRNAs previously annotated by RefSeq, UCSC or GENCODE 4. The publically available lincRNA sets were derived by running specific steps of our lincRNA classification pipeline on the transcripts annotated in the public datasets (Figure 1a, See Supplementary Table 5 for specific details).

Notably, there were 364 loci (4%) for which the number of isoforms identified by Scripture was over 10 (and often over few dozens or hundreds). Such complexity of isoform makes it impossible to estimate abundance in our downstream analysis. We therefore considered only the isoforms identified by Cufflinks and public annotations for these loci and provide the additional possible isoforms at http://www.broadinstitute.org/genome_bio/human_lincrnas.

Stringent set classification

We defined a stringent lincRNA set that includes transcripts loci for which at least one lincRNA isoform: (i) was reconstructed in at least two tissues, or (ii) was reconstructed by two assemblers in the same tissue. An isoform was considered as reconstructed by two sources in case an exact sub-sequence of exon-intron boundaries was identified in the isoforms from the two sources. That is, at least two neighboring exons were independently assembled by two sources. This
criterion can still be sensitive to errors in the alignments as both assembly approaches use the same input alignments.

**Classification of Transcripts of Uncertain Coding Potential (TUCP)**

We defined the *Transcripts of Uncertain Coding Potential* (TUCP) set as the set of transcripts that passed the size and abundance filters of our pipeline but were either: (i) classified as pseudogenes by Vega (Wilming et al. 2008), (ii) had a PhyloCSF score that exceeded our determined threshold, or (iii) had a Pfam domain hit. We included only the maximum likelihood isoforms reconstructed by cufflinks for this set to decrease the isoform complexity in each locus.

While our abundance estimation for pseudogenes may be biased by incorrect read mapping to corresponding paralogous loci, our results estimating global characteristics of expression levels and tissue specificity were similar regardless of whether pseudogene transcripts were included or excluded from the TUCP set (data not shown).

**Assembly performance estimation on protein coding genes**

We estimated the number of protein coding transcripts annotated in RefSeq that were correctly assembled by Cufflinks, Scripture or both by counting how many of the known transcripts had a partially compatible or fully compatible transcript (*Supplementary Figure 1a*) assembled from any of our 24 tissues and cell-lines. A transcript was considered as having a fully compatible assembled isoform if the exact exon-intron chain was recovered in the assembled transcript, and partially compatible if at least 2 exact exons and their connecting intron were recovered in the assembled transcript (*Supplementary Figure 1a*).
Chromatin state of actively transcribed genes

We screened for evidence of a chromatin signature of histone 3 lysine 4 tri methylation (H3K4me3) across the promoter followed by a H3K36me3 along the transcribed region (K4-K36 domains) across lincRNAs loci. To this end, we used publically available segmentations of K4 and K36 chromatin domains detected from ChIP-Seq (chromatin immunoprecipitation followed by sequencing) data across 9 ENCODE cell lines (Ernst et al. 2011). These cell lines (with the one exception of lung fibroblast), are not included in our 24 tissue/cell line compendium. We identified all lincRNAs that had an enrichment of H3K4me3 within +/- 2 KB of the transcript start site and enrichment of H3K36me3 somewhere along the transcripts genomic region in the same cell type. We defined lincRNAs with such overlap as K4-K36 lincRNAs whenever neither the K4 nor the K36 chromatin domains were also overlapping a coding gene. We also included in the K4K36 lincRNA set those lincRNAs that overlapped genomic regions previously classified as transcribing human lincRNAs based on similar chromatin state maps (Khalil et al. 2009). Details of the random model generation are specified below.

Chromatin state of enhancers

To evaluate the number of lincRNAs that coincide with enhancers we examined genomic regions previously classified as containing a chromatin domain characterizing enhancers (Ernst et al. 2011). In brief, Ernest et al. applied a multivariate hidden Markov model to classify the chromatin state of 200 base windows along the genome across 9 encode cell lines based on the combinatorial patterns of 9 chromatin marks. We enumerated the number of lincRNAs that had an exon overlap with a region that was classified as a strong enhancer (state #4) at any of the 9 cell types. Since the chromatin state classification were assigned to just a sub region of the
transcript and such sub region may have been classified as having different states in different cell type, we also applied a more stringent criteria to define potential enhancer regions. In this criterion, we now focused on the 200 bases windows in which an enhancer state was classified in the majority of cells in which the chromatin state classification was different than heterochromatin (state #13). This includes all regions that were classified as enhancer in only one or few tissues and not by other alternative states (such as: promoters, transcription elongation etc’).

Enhancers sets used for the comparison with mouse lincRNAs from a same/similar cell types were obtained as following. Enhancers that were defined in mouse embryonic stem cells based on regions bound by CHD7 and p300 and also enriched with H3K4me1 and H3K27ac were obtained from (Zentner et al. 2011). Intergenic enhancer elements that produce eRNA in mouse cortical neurons based on binding to p300/CBP, enrichment in H3K4me1 and RNA-Seq support (> 7 reads) were obtained from (Kim et al. 2010).

**Random permutation model of intergenic transcripts**

We used the following random permutation model to calculate the enrichment of our lincRNA set with different genomic features (e.g. K4K36 domains, enhancer regions). For each calculation we generated a 1,000 random sets that preserved the same transcript structures distribution of our original lincRNA set but were uniformly sampled from the un-annotated fraction of the genome. Thus, given our transcript set we have randomly re-positioned each transcript in the un-annotated fraction of the genome. An enrichment P-value was estimated by the number of times the number of overlaps between a random set and the genomic feature of
interest exceeded that of our lincRNA set. We defined the un-annotated fraction of the genome by removing centromer regions and all regions that were annotated with a transcript by GENCODE 4.

**Functional enrichment analysis of coding gene sets**

We estimated the enrichment of different coding gene sets with Gene Ontology (GO) (Ashburner et al. 2000) functional terms using the David Bioinformatics Tool (Huang et al. 2008) and reported the results for GO-FAT biological process terms. GO-FAT are a sub-set of the GO annotation set derived by David by eliminating broad GO-terms that are high in the GO-term tree hierarchy. This is designed to avoid redundancy of annotation sets and overshadowing of the broad terms when applying multiple testing corrections.

**Estimating expression abundance: note**

In some cases, the abundance estimation does not represent a specific isoform of interest but rather a different transcript. This is due to cases when there is read coverage across the exons but no spliced reads supporting the specific transcript of interest (as displayed using the Integrative Genome Viewer (Robinson et al. 2011) in [Supplementary Figure 14](#)). This problem will affect any transcript abundance estimation method that doesn’t explicitly constrain requirements for spliced read support of a specific isoform. To the best of our knowledge such method is not yet available. To address this issue, we flagged each transcript in the catalog that did not have spliced read support in the tissue where it is maximally expressed and excluded these transcripts from the neighbor correlation analysis (below).
Normalization of expression vectors for tissue specificity calculation

To calculate the tissue specificity scores of a transcript we needed to convert the transcript’s expression vector to an abundance density (as the JS metric is applied on discrete probability distributions). To this end, we added a pseudo-count of 1 to the raw FPKM (fragments per kilobase of exons per million fragments mapped) expression vector of each transcript and applied a log 2 normalization to obtain a non negative expression vector. We then normalized this expression vector to a density vector by dividing by the total expression counts. Formally:

\[ V' = \frac{\log_2(V + 1)}{\sum_{i=1}^{n} \log_2(v_i + 1)} \]

where \( V = (v_1, ..., v_n) \) is the original raw FPKM abundance estimation of the transcript and \( V' \) is the new normalized density vector.

K- means clustering of tissue expression patterns

We applied K-means clustering with random seeds to obtain clusters of lincRNAs and protein coding genes using the \( JS_{dist} \) (Methods) as a distance metric. Applying the JS distance metric allowed better separation of heterogeneous clusters to tissue specific clusters as compared to using a Euclidean distance or Pearson correlation as a distance metric. For the clusters presented in this paper we used K=30 after optimizing the selection of K to minimize the distances of data within clusters while maximizing the distance between clusters using a Silhouette function (Rousseeuw 1987) (Supplementary Figure 15). Briefly, we used K-means clustering with 31 values of K (k=20,25,30,…,80). For each run, we calculated the Sillhouette function on each transcript’s expression pattern \( e^t \):

\[ Si(e^t) = \frac{b(e^t) - a(e^t)}{\max(a(e^t), b(e^t))} \]

where
(3) \[ a(e^i) = E(Dist(e^i, e^j) \mid e^i \in c^x \text{ and } e^j \in c^x), \text{ where } c^x \text{ is the cluster to which } e^i \text{ was assigned.} \]

(4) \[ b(e^i) = \min_{c_{cox}} E(Dist(e^i, e^j) \mid e^i \not\in c_{cox} \text{ and } e^j \in c_{cox}) \]

That is, \( a(e^i) \) is the average distance of a sample to all other samples in its cluster and \( b(e^i) \) is the minimal average distance of a sample from all other clusters.

We then calculated two summary statistics:

(i) the mean silhouette value across all transcripts: \( A = E(S_i(e^i)), \ i \in (1 \ldots n) \)

(ii) the mean of mean cluster Silhouette scores across \( m \) clusters:

\[ B = E(S_{i_{cluster}}(c^x)), x \in (1 \ldots m) \]

\[ \text{where, } S_{i_{cluster}}(c^x) = E(S_i(e^i)| e^i \in c^x) \]

\( K=30 \) obtained a maximal Sillhouette score according to both statistics and was chosen for our analysis.

**Expression correlation of lincRNAs and their neighboring genes**

We estimated the expression correlation between the expression pattern of a lincRNA and its closest coding gene neighbor by calculating the Pearson correlation coefficient between their density-normalized expression vectors (as described above). Density normalization was applied in order to compare vectors of similar magnitudes (as the expression levels of coding genes is 10 orders of magnitude higher than that of lincRNAs).
Refined alignment of human lincRNAs and their mouse orthologs

To assess the alignment quality of the TransMap lincRNAs and their syntenic orthologs we realigned the transcript sequence of all human lincRNAs and their mouse orthologs using the Fast Statistical Alignment algorithm with default parameters (Bradley et al. 2009). We compared the fraction of aligned bases and fraction of identical bases with respect the human reference sequence across 4 sets: (a) human lincRNAs and their mouse orthologs, (b) random sequence pairs, (c) randomly selected syntenic blocks and (d) orthologous coding genes known to RefSeq (Figure 4f and Supplementary Figure 10c-d). The random sequence pairs were obtained by shuffling the human lincRNAs and mouse ortholog pairs. Randomly selected syntenic block were obtained by uniformly sampling 1 KB blocks from the un-annotated fraction of the genome that is also syntenically mapped to mouse.

Sequence conservation level estimation in novel transcripts with potential coding capacity

We used the SiPhy algorithm and software package (http://www.broadinstitute.org/genome_bio/siphy/ ;(Garber et al. 2009)) to estimate $\omega$, which quantifies how well a sequence substitution pattern across a multiple sequence alignment fits a neutral selection model, or is constrained by purifying selection. Specifically, $\omega$ represents the deviation ('contraction' or 'extension') of the phylogenetic tree’s branch length in a specific position in the genome sequence compared to the neutral tree, based on the total number of substitutions estimated from the alignment of the region of interest across 20 placental mammals (build Hg18, http://hgdownload.cse.ucsc.edu/goldenPath/hg18/multiz44way/). We estimated $\omega$ across the exons of: (1) coding genes annotated in RefSeq, (2) TUCP transcripts after excluding pseudogenes, (3) Trans-mapped lincRNAs, (4) the stringent set lincRNAs, (5) intronic segments
that were size-matched to their neighboring coding exon and were randomly selected from the intron, and (6) ancestral repeats of human and mouse. The human-mouse ancestral repeats were downloaded from the UCSC Genome Browser (Kent et al. 2002) and a subset of 3000 repeats was uniformly sampled and used for this analysis.

**lincRNAs in disease associated regions**

We downloaded a list of disease-associated SNPs \( P < 5 \times 10^{-5} \) from the National Human Genome Research Institute (NHGRI) catalog of published genome wide association studies (Hindorff et al. 2009). We then extracted regions which are in linkage disequilibrium (LD) with each of the SNPs by first finding the left-most and right-most SNPs that are in LD \( (R^2 > 0.5) \) with the aforementioned SNP and then finding the closet recombination hotspots as described in (Raychaudhuri et al. 2009). We than crossed our lincRNAs and TUCP sets with disease/trait associated LD regions. The results are reported in **Supplementary Dataset 2 and 6**.

Transcription factor binding motif were first identified using AliBaba 2.1 through TRANSFAC (Matys et al. 2003) and then scanned to determine whether the motif is conserved by running SiPhy (Garber et al. 2009) in comparison to 5000 randomly chosen coding gene promoter regions (+/- 2KB from transcription start site).
Supplementary Tables

Supplementary Table 1. Transcript reconstruction comparison between biological replicates.

<table>
<thead>
<tr>
<th>% of stringent set transcript loci from the low coverage sample that were identified by a matching multi-exon or single-exon isoform in the high coverage sample</th>
<th>Human Lung Fibroblast</th>
<th>Brain</th>
<th>Testes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>69%</td>
<td>74.5%</td>
<td>81.6%</td>
</tr>
</tbody>
</table>
Supplementary Table 2. Number of transcripts across species.

<table>
<thead>
<tr>
<th>Species</th>
<th>Species name</th>
<th>Number of transcripts trans-mapped across species</th>
<th>Number of ESTs</th>
<th>Number of RefSeq genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mouse</td>
<td><em>Mus musculus</em></td>
<td>2313560</td>
<td>4367822</td>
<td>28539</td>
</tr>
<tr>
<td>Zebra fish</td>
<td><em>Danio rerio</em></td>
<td>792572</td>
<td>1646898</td>
<td>16158</td>
</tr>
<tr>
<td>Cow</td>
<td><em>Bos Taurus</em></td>
<td>685264</td>
<td>1621863</td>
<td>13964</td>
</tr>
<tr>
<td>Rat</td>
<td><em>Rattus norvegicus</em></td>
<td>424072</td>
<td>1121086</td>
<td>17172</td>
</tr>
<tr>
<td>Madka fish</td>
<td><em>Oryzias latipes</em></td>
<td>389820</td>
<td>703843</td>
<td>584</td>
</tr>
<tr>
<td>Chicken</td>
<td><em>Gallus gallus</em></td>
<td>258332</td>
<td>618531</td>
<td>5468</td>
</tr>
<tr>
<td>Dog</td>
<td><em>Canis familiaris</em></td>
<td>121789</td>
<td>399902</td>
<td>1212</td>
</tr>
<tr>
<td>Tetraodon</td>
<td><em>Tetraodon nigroviridis</em></td>
<td>103162</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rhesus</td>
<td><em>Macaca mulatta</em></td>
<td>34270</td>
<td>68320</td>
<td>2566</td>
</tr>
<tr>
<td>Orangutan</td>
<td><em>Pongo pygmaeus abelli</em></td>
<td>29657</td>
<td>46027</td>
<td>3555</td>
</tr>
<tr>
<td>Zebra finch</td>
<td><em>Taeniopygia guttata</em></td>
<td>27017</td>
<td>92700</td>
<td>170</td>
</tr>
<tr>
<td>Horse</td>
<td><em>Equus caballus</em></td>
<td>15803</td>
<td>37539</td>
<td>996</td>
</tr>
<tr>
<td>Rabbit</td>
<td><em>Oryctolagus cuniculus</em></td>
<td>13561</td>
<td>36332</td>
<td>1334</td>
</tr>
<tr>
<td>Stickleback fish</td>
<td><em>Gasterosteus aculeatus</em></td>
<td>9042</td>
<td>301622</td>
<td></td>
</tr>
<tr>
<td>Opossum</td>
<td><em>Monodelphis domestica</em></td>
<td>1712</td>
<td>459</td>
<td>467</td>
</tr>
<tr>
<td>Platypus</td>
<td><em>Ornithorhynchus anatinus</em></td>
<td>486</td>
<td>47782</td>
<td>383</td>
</tr>
</tbody>
</table>
Supplementary Table 3: disease associated intergenic regions containing a lincRNA that is expressed in a disease related tissue.

<table>
<thead>
<tr>
<th>Disease/trait</th>
<th>SNP</th>
<th>P value; OR/ beta</th>
<th>lincRNAs</th>
<th>Tissue</th>
<th>Reference</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thyroid cancer</td>
<td>rs944289</td>
<td>2E-9 1.37</td>
<td>XLOC_010996</td>
<td>Thyroid</td>
<td>(Gudmundsson et al. 2009)</td>
<td>Strikingly high expression in the thyroid</td>
</tr>
<tr>
<td>Waist-hip ratio</td>
<td>rs1055144</td>
<td>9.9E-25 0.04</td>
<td>XLOC_006016</td>
<td>Adipose</td>
<td>(Heid et al. 2010)</td>
<td>Ortholog transcript in cow, some EST support</td>
</tr>
<tr>
<td>LDL cholesterol</td>
<td>rs2126259</td>
<td>7E-12 0.02</td>
<td>XLOC_006706</td>
<td>Liver</td>
<td>(Waterworth et al. 2010)</td>
<td>Transcript identified by GENCODE with support by RNA-Seq.</td>
</tr>
<tr>
<td>Bipolar disorder</td>
<td>rs472913</td>
<td>2E-7 1.18</td>
<td>XLOC_000856</td>
<td>Brain</td>
<td>(Scott et al. 2009)</td>
<td>Partial isoform identified by GENCODE</td>
</tr>
<tr>
<td>Prostate cancer</td>
<td>rs9284813</td>
<td>5E-9</td>
<td>XLOC_002726</td>
<td>Prostate</td>
<td>(Takata et al. 2010)</td>
<td>Partial isoform identified by GENCODE</td>
</tr>
</tbody>
</table>
Supplementary Table 4. RNA-Seq datasets.

<table>
<thead>
<tr>
<th>Tissue/ Cell type</th>
<th>Dataset</th>
<th>Sequencing platform</th>
<th>Read length</th>
<th>Number of aligned reads</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human lung fibroblasts (hLF)</td>
<td>Rinn lab</td>
<td>Illumina genome analyzer II (GAII)</td>
<td>75 base pairs (bp), paired ends (PE)</td>
<td>33849576</td>
</tr>
<tr>
<td>hLF2</td>
<td>Rinn lab</td>
<td>GAII</td>
<td>75 pb, PE (2 lanes)</td>
<td>136310398</td>
</tr>
<tr>
<td>Foreskin fibroblasts</td>
<td>Rinn lab</td>
<td>GAII</td>
<td>75 pb, PE</td>
<td>58787488</td>
</tr>
<tr>
<td>Brain</td>
<td>Rinn lab</td>
<td>GAII</td>
<td>75 pb, PE</td>
<td>50661137</td>
</tr>
<tr>
<td>HeLa</td>
<td>Rinn lab</td>
<td>GAII</td>
<td>75 pb, PE</td>
<td>30134104</td>
</tr>
<tr>
<td>Liver</td>
<td>Rinn lab</td>
<td>GAII</td>
<td>75 pb, PE</td>
<td>32132725</td>
</tr>
<tr>
<td>Placenta</td>
<td>Rinn lab</td>
<td>GAII</td>
<td>75 pb, PE</td>
<td>47384953</td>
</tr>
<tr>
<td>Testes</td>
<td>Rinn lab</td>
<td>GAII</td>
<td>75 pb, PE</td>
<td>49171324</td>
</tr>
<tr>
<td>Adipose</td>
<td>Illumina Human Body Map2</td>
<td>High-Seq</td>
<td>75 pb single end (SE), 50 bp , PE</td>
<td>221715491</td>
</tr>
<tr>
<td>Adrenal gland</td>
<td>Illumina Human Body Map2</td>
<td>High-Seq</td>
<td>75 pb SE, 50 bp PE</td>
<td>218793183</td>
</tr>
<tr>
<td>Brain</td>
<td>Illumina Human Body Map2</td>
<td>High-Seq</td>
<td>75 pb SE, 50 bp PE</td>
<td>204772783</td>
</tr>
<tr>
<td>Breast</td>
<td>Illumina Human Body Map2</td>
<td>High-Seq</td>
<td>75 pb SE, 50 bp PE</td>
<td>219398164</td>
</tr>
<tr>
<td>Colon</td>
<td>Illumina Human Body Map2</td>
<td>High-Seq</td>
<td>75 pb SE, 50 bp PE</td>
<td>244916925</td>
</tr>
<tr>
<td>Heart</td>
<td>Illumina Human Body Map2</td>
<td>High-Seq</td>
<td>75 pb SE, 50 bp PE</td>
<td>246232044</td>
</tr>
<tr>
<td>Kidney</td>
<td>Illumina Human Body Map2</td>
<td>High-Seq</td>
<td>75 pb SE, 50 bp PE</td>
<td>229699125</td>
</tr>
<tr>
<td>Liver</td>
<td>Illumina Human Body Map2</td>
<td>High-Seq</td>
<td>75 pb SE, 50 bp PE</td>
<td>239123970</td>
</tr>
<tr>
<td>Lung</td>
<td>Illumina Human Body Map2</td>
<td>High-Seq</td>
<td>75 pb SE, 50 bp PE</td>
<td>240042489</td>
</tr>
<tr>
<td>Lymph Node</td>
<td>Illumina Human Body Map2</td>
<td>High-Seq</td>
<td>75 pb SE, 50 bp PE</td>
<td>249015289</td>
</tr>
<tr>
<td>Ovary</td>
<td>Illumina Human Body Map2</td>
<td>High-Seq</td>
<td>75 pb SE, 50 bp PE</td>
<td>241958761</td>
</tr>
<tr>
<td>Prostate</td>
<td>Illumina Human Body Map2</td>
<td>High-Seq</td>
<td>75 pb SE, 50 bp PE</td>
<td>251395450</td>
</tr>
<tr>
<td>Skeletal Muscle</td>
<td>Illumina Human Body Map2</td>
<td>High-Seq</td>
<td>75 pb SE, 50 bp PE</td>
<td>242833112</td>
</tr>
<tr>
<td>White Blood Cells</td>
<td>Illumina Human Body Map2</td>
<td>High-Seq</td>
<td>75 pb SE, 50 bp PE</td>
<td>248288643</td>
</tr>
<tr>
<td>Testes</td>
<td>Illumina Human Body Map2</td>
<td>High-Seq</td>
<td>75 pb SE, 50 bp PE</td>
<td>241527185</td>
</tr>
<tr>
<td>Thyroid</td>
<td>Illumina Human Body Map2</td>
<td>High-Seq</td>
<td>75 pb SE, 50 bp PE</td>
<td>235187522</td>
</tr>
</tbody>
</table>
Supplementary Table 5. Publically available annotations used for this study.

<table>
<thead>
<tr>
<th>Dataset name</th>
<th>Downloaded from (date)</th>
<th>refs</th>
<th># of transcript s</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein coding transcripts known to RefSeq</td>
<td>The UCSC genome browser ; NCBI37/Hg19 (July 2010)</td>
<td>(Pruitt et al. 2002)</td>
<td>31911</td>
<td>Extracted only transcripts with an identifier beginning with NM_</td>
</tr>
<tr>
<td>microRNA and snoRNAs set1</td>
<td>The UCSC genome browser; NCBI37/Hg19 (November 2010)</td>
<td>(Lestrade and Weber 2006; Griffiths-Jones et al. 2008)</td>
<td>1341</td>
<td>Data was collected from miRBASE and snoRNABase databases.</td>
</tr>
<tr>
<td>miRNA, miRNA_pseudogene, rRNA, scoRNA, snorNA, tRNAs</td>
<td>Biomart Ensembl ; NCBI37/Hg19 (November 2010; Ensembl Release 60)</td>
<td>(Flicek et al. 2011)</td>
<td>6219</td>
<td>Based on Ensembl annotation</td>
</tr>
<tr>
<td>Pseudogenes set1 – Vertebrate Genome Annotation (Vega) pseudogenes</td>
<td>The UCSC genome browser NCB137/Hg19; (July 2010; Vega Release 38)</td>
<td>(Wilming et al. 2008)</td>
<td>9534</td>
<td>Classification scheme: <a href="http://vega.sanger.ac.uk/info/about/gene_and_transcript_types.html">http://vega.sanger.ac.uk/info/about/gene_and_transcript_types.html</a> Eliminated the ones that overlapped a RefSeq NR_</td>
</tr>
<tr>
<td>Pseudogenes set2 - Yale Gerstein Group Pseudogenes</td>
<td>From : <a href="http://pseudogenes.org/">http://pseudogenes.org/</a> (July 2010 ; build 37)</td>
<td>(Zhang et al. 2006)</td>
<td>16363</td>
<td></td>
</tr>
<tr>
<td>UCSC coding genes</td>
<td>The UCSC genome browser NCBI37/Hg19 (November 2010); The kgTxInfo primary table</td>
<td>(Hsu et al. 2006; Karolchik et al. 2008)</td>
<td>69201</td>
<td>Extracted transcripts annotated as coding (ref classification 2008 ; <a href="http://genome.ucsc.edu/cgi-bin/hgGene?hgg_do_kgMethod">http://genome.ucsc.edu/cgi-bin/hgGene?hgg_do_kgMethod</a> =1, <a href="http://genome.ucsc.edu/cgi-bin/hgGene?hgg_do_txInfoDescription=1">http://genome.ucsc.edu/cgi-bin/hgGene?hgg_do_txInfoDescription=1</a>)</td>
</tr>
<tr>
<td>GENCODE coding genes</td>
<td>GENCODE Version 4, from : ftp://ftp.sanger.ac.uk/pub/gencode/ (Nov 2010)</td>
<td>(Harrow et al. 2006)</td>
<td>29782</td>
<td>Downloaded file : GENCODE_v4.annotation.GR Ch37.gtf.gz Extracted genes that were manually annotated ( level 1-2) with the types : gene_type = &quot;protein_coding&quot;, transcript_status = &quot;KNOWN&quot;</td>
</tr>
<tr>
<td>Non-coding transcripts known</td>
<td>The UCSC genome browser ;</td>
<td>(Pruitt et al. 2002)</td>
<td>4883</td>
<td>Extracted only transcripts with an identifier beginning with</td>
</tr>
<tr>
<td>to RefSeq to UCSC</td>
<td>NCBI37/Hg19 (July 2010)</td>
<td>NR_</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------------------------</td>
<td>-----------------------</td>
<td>-----</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Non-coding transcripts known to UCSC</strong></td>
<td>The UCSC genome browser NCBI37/Hg19 (November 2010); The kgTxInfo primary table</td>
<td>(Hsu et al. 2006; Karolchik et al. 2008) 7673</td>
<td>Extracted transcripts annotated as non-coding (ref classification 2008)</td>
<td></td>
</tr>
<tr>
<td><strong>Non-coding transcripts known to GENCODE</strong></td>
<td>GENCODE Release 4 From : ftp://ftp.sanger.ac.uk/pub/GENCODE/ Nov 2010</td>
<td>(Harrow et al. 2006) 14598</td>
<td>Extracted genes that were annotated with the types : gene_type = &quot;processed_transcript&quot;, or &quot;lincRNA&quot;, or &quot;non_coding&quot;; transcript_status = &quot;processed_transcript&quot;, or &quot;lincRNA&quot;, or &quot;non_coding&quot;, or &quot;misc_RNA&quot;; Based on the Havana annotation guidelines : <a href="http://www.sanger.ac.uk/research/projects/vertebrategenome/havana/assets/guidelines.pdf">http://www.sanger.ac.uk/research/projects/vertebrategenome/havana/assets/guidelines.pdf</a></td>
<td></td>
</tr>
<tr>
<td><strong>RefSeq derived lincRNAs</strong></td>
<td></td>
<td></td>
<td>Run “non coding transcripts known to RefSeq” set through steps 1,3,6 of our lincRNA classification pipeline</td>
<td></td>
</tr>
<tr>
<td><strong>GENCODE derived lincRNAs</strong></td>
<td></td>
<td></td>
<td>Run “non coding transcripts known to GENCODE” set through steps 1,3,4,5,6 of our lincRNA classification pipeline</td>
<td></td>
</tr>
<tr>
<td><strong>UCSC derived lincRNAs</strong></td>
<td></td>
<td></td>
<td>Run “non coding transcripts known to UCSC” set through steps 1,3,4,5,6 of our lincRNA classification pipeline</td>
<td></td>
</tr>
</tbody>
</table>
References:


