TransCistor reveals the landscape of cis-regulatory long noncoding RNAs

Authors

Panagiotis Chouvardas123, Marc Zimmerli1#12, Daniel Hanhart23, Mario Moser#12, Hugo Guillen-Ramirez1245, Sanat Mishra6, Roberta Esposito12, Taisia Polidori12, Maro Widmer12, Raquel García-Pérez7, Marianna Kruithof-de Julio23, Dmitri Pervouchine8, Marta Melé7, Rory Johnson*1245

1. Department of Medical Oncology, Inselspital, Bern University Hospital, University of Bern, 3010 Bern, Switzerland.
2. Department for BioMedical Research, University of Bern, 3008 Bern, Switzerland.
3. Department of Urology, Inselspital, Bern University Hospital, Bern, Switzerland.
4. School of Biology and Environmental Science, University College Dublin, Dublin D04 V1W8, Ireland.
5. Conway Institute for Biomolecular and Biomedical Research, University College Dublin, Dublin D04 V1W8, Ireland.
6. Indian Institute of Science Education and Research, Mohali, India.
7. Life Sciences Department, Barcelona Supercomputing Centre, 08034, Barcelona, Spain.
8. Center of Life Sciences, Skolkovo Institute of Science and Technology, Moscow, Russia.

# Equal contribution.

* Correspondence to rory.johnson@dbmr.unibe.ch
Abstract

Long noncoding RNAs (lncRNAs) can positively and negatively regulate expression of target genes encoded in cis. However, the extent, characteristics and mechanisms of such cis-regulatory lncRNAs (cis-lncRNAs) remain obscure. Until now, they have been defined using inconsistent, ad hoc criteria that can result in false-positive predictions. Here, we introduce TransCistor, a framework for defining and identifying cis-lncRNAs based on enrichment of targets amongst proximal genes. Using transcriptome-wide perturbation experiments for 190 human and 133 mouse lncRNAs, we provide the first large-scale view of cis-lncRNAs. Our results ascribe cis-regulatory activity to only a small fraction (~10%) of lncRNAs, with a prevalence of activators over repressors. cis-lncRNAs are detected at similar rates by RNA interference (RNAi) and antisense oligonucleotide (ASO) perturbations. We leverage this cis-lncRNA catalogue to evaluate mechanistic models for cis-lncRNAs involving enhancers and chromatin folding. Thus, TransCistor places cis-regulatory lncRNAs on a quantitative foundation for the first time.

Main

The first characterised long noncoding RNAs (lncRNAs), H19 and XIST, were both found to have cis-regulatory activity: their perturbation by loss-of-function (LOF) led to increased expression of protein-coding genes encoded “in cis” - i.e. within a relatively short linear distance on the same chromosome1,2. Genes whose expression responds to lncRNA LOF are considered “targets” of that lncRNA, while the direction of this change (up or down) defines the lncRNA as a “repressor” or “activator”, respectively. Since then, numerous more cis-regulatory lncRNAs have been reported 3,4. Conversely, other lncRNAs have no apparent positional preference for their targets, and are termed trans-lncRNAs 5. This cis/trans duality provides a fundamental framework for understanding regulatory lncRNAs6, yet the global prevalence of cis- and trans-regulatory lncRNAs remains poorly defined.

Within reported cis-lncRNAs there appears to be great diversity, in terms of regulatory activity (activators and repressors), distance of the target (ranging from one hundred basepairs4 to hundreds of kilobases5) and number of targets (one4 to many8). Two overarching molecular mechanisms have been proposed for cis-lncRNAs: enhancer elements and chromatin folding6. Some cis-activating lncRNAs, termed “enhancer lncRNAs” (e-lncRNAs), have been found to overlap DNA-encoded enhancer elements10–12, similar to lncRNAs more generally13. The expression and splicing of the e-lncRNA transcripts correlate with enhancer activity, implying that RNA processing somehow promotes target gene activation. Similarly, it has been proposed that cis-lncRNAs find their targets via spatial proximity, determined by chromatin looping or within the confines of local topologically-associating domains (TADs)6. An attractive corollary of these models is that cis-regulatory lncRNAs may act via non-sequence dependent mechanisms, perhaps involving phase separation14 and local concentration gradients15. It has recently been posited that lncRNAs proceed through an evolutionary trajectory commencing with fortuitous cis-regulatory activity before acquiring targeting capabilities and graduating to trans-regulation16. Nonetheless, these conclusions are drawn from piecemeal studies of individual lncRNAs, and a holistic view of cis and trans lncRNAs, the features that distinguish them, and resulting clues to their molecular mechanisms and biological significance, await a comprehensive catalogue of cis-lncRNAs.
Despite their importance, we lack a rigorous and agreed definition for cis-IncRNAs. Until now, they have been defined by the existence of ≥1 proximal targets. Targets are defined as those whose expression changes in response to IncRNA LOF, as measured using single-gene (RT-PCR) or whole-transcriptome (RNA-seq, CAGE, microarray) techniques. “Proximity” is defined on a case-by-case basis, using a wide range of windows spanning $10^2$ to $10^5$ bp. A single proximal target is usually considered sufficient. The problem with this approach is that, as the total number of targets and/or cis-window size increase, so will the chance of observing ≥1 cis-target genes by random chance.

In this study, we consider cis-IncRNAs from a quantitative perspective. We show that conventional definitions are prone to high false positive rates. We introduce statistical methods for definition of cis-IncRNAs at controlled false discovery rates, and use them to classify regulatory IncRNAs across hundreds of perturbation datasets. Our results enable us to estimate, for the first time, the prevalence of cis-acting IncRNAs and evaluate hypotheses regarding their molecular mechanisms of action.

## Results

### A quantitative definition of cis-IncRNAs

We began by investigating the usefulness of the present “naïve” definition of cis-IncRNAs, based on the existence of ≥1 target gene within a local genomic window. To create a dataset of IncRNA-target relationships, we collected 382 IncRNA LOF experiments targeting 188 human IncRNAs from a mixture of sources (including the recently published dataset of ASO knockdowns in human dermal fibroblasts from the FANTOM consortium) and 137 experiments for 131 IncRNAs in mouse (Figure 1A). To this we added 6 hand-curated LOF experiments targeting previously reported cis-acting IncRNAs (UMLILO, XIST x2, Chaserr, Paupar and Dalii). We employed a functional definition of “targets”, as genes whose steady-state levels significantly change in response to a given IncRNA’s LOF (Figure 1B). We further define targets as activated or repressed, where they decrease / increase in response to IncRNA LOF, respectively. Finally, we define genes to be proximal / distal, if their annotated transcription start site (TSS) lies inside / outside a defined distance window of the IncRNA’s TSS, respectively. In the dataset, 126 IncRNAs were represented by ≥1 independent experiments, and the median number of target genes identified per experiment was 65 (Supplementary Figure S1).

Using a range of cis-window sizes from 50 kb to 1 Mb centred on the IncRNAs’ TSSs, we evaluated the fraction of IncRNAs that would be defined as cis-IncRNAs under the naïve definition. This approach defines ~2 to 12% of IncRNAs as cis-regulators (Figure 1C, line). To test whether this rate is greater than random chance, we shuffled the target / non-target labels of all protein-coding genes and repeated this analysis. The rate of cis-IncRNA predictions in this random data overlapped the true rates in all windows (Figure 1C, boxplots), suggesting that the conventional definition of cis-IncRNAs yields high rates of false-positive predictions.

To overcome this issue, we adopted the following definition for cis-IncRNAs: cis-IncRNAs are those whose targets are significantly enriched amongst proximal genes. This definition has the advantage of being quantitative and statistically testable. Below we incorporate this definition into two alternative methods for identifying cis-IncRNAs.
TransCistor: digital and analogue identification of cis-IncRNAs

The first method employs a simple definition of proximal genes, being those whose TSS falls within a defined window centred on the lncRNA TSS. We developed a pipeline, TransCistor-digital, which takes as input a processed whole-transcriptome list of target genes (“regulation file”), and tests for statistical enrichment in proximal genes using the hypergeometric distribution (Figure 1D). Although in principle any sized window may be used, we reasoned that the most biologically-meaningful would be the local TAD, in line with previous studies19. Chromatin folding varies to an extent between cell types20. Therefore, TransCistor-digital calculates enrichment across a set of experimentally-defined cell-type-specific TADs (40 human, 3 mouse)21 and aggregates the resulting p-values by their harmonic mean.

The above fixed-window approach is intuitive yet has drawbacks. Several reported cis-IncRNAs have individual targets that are not immediately adjacent7, and might be overlooked by the fixed-window approach. Furthermore, many IncRNAs may have no neighbouring genes in their local TAD, or no identified local TAD. Therefore, we developed an alternative method that dispenses with fixed windows, while still examining proximity biases in targets. This method, TransCistor-analogue, defines a distance statistic as the mean TSS-to-TSS distance of all same-chromosome targets of a given lncRNA (Figure 1E). To estimate statistical significance, a null distribution is calculated by randomisation of target labels (Figure 1E). Now, cis-IncRNAs are defined as those having a distance statistic that is lower than a majority of simulations.

We sought to test the performance of TransCistor-digital and evaluate the global landscape of cis-IncRNAs. After filtering out unusable datasets (having no cis-targets or no overlapping TAD), 168 datasets remained. The majority of p-values produced by this analysis follow the null distribution, underlining the conservative statistical behaviour of TransCistor (Figure 2A,B). We discovered 19 cis-acting IncRNAs (12 activators, 7 repressors), with a relatively relaxed false discovery rate (FDR) threshold of 0.5, while no cis-IncRNAs are simultaneously classified as activator and repressor (Figure 2C). Amongst the top-ranked cis-IncRNAs is UMLILO, previously described to activate multiple genes in its local genomic neighbourhood8. UMLILO exhibits a significant enrichment of activated targets amongst proximal genes, which is not observed for repressed targets (Figure 2D,E). Analysis of the entire perturbation dataset by TransCistor-analogue identified 20 cis-IncRNAs (15 activators, 5 repressors, FDR≤0.5). Statistical behaviour is good (Figure 2F,G), while cis-IncRNAs once again are cleanly split between activators and repressors (Figure 2H).

The usefulness of these methods is supported by their internal and external consistency. Together, the TransCistor approaches correctly identify previously-described cis-activators H1922, JPX23, Evx1os24 and DA12594225 amongst the top ranked cis-activators, while XIST is amongst the top repressors26 (Figure 2J). Both human and mouse orthologues of CHASERR (ENSG00000272788) are identified as cis-repressors4. Of the hits, two are concordantly classified by ≥1 independent perturbation experiments (XIST & DNAAF3-AS1 classified as cis-repressors based on two separate experiments each) (Figure 2I,J). We observed agreement between the two TransCistor methods, with 6 cis-IncRNAs in common (DA125942, linc1427, RAD51-AS1, H19, Xist, DNAAF3-AS1) (p-value < 0.05, hypergeometric test) (Figure 3A).

TransCistor predicted cis-regulatory activity for a number of known IncRNAs that have never been described as such in prior literature. These include BANCR (cis-activator), and SBF2-AS1,
LASTR, NORAD, DANCER (cis-repressors). However, the latter two are only identified in one out of multiple independent perturbation experiments (6/5 for NORAD/DANCER respectively). In the case of DANCER, the cis definition arises from the repression of two same-strand small RNAs (has-mir-4449, SNORA26). It is not yet clear if these results reflect false-positive or false-negative predictions. To investigate this, we merged all hits across experiments and repeated the analysis, but here we found no cis signal, suggesting that they are false-positive predictions. On the other hand, analysis of an independent dataset for SBF2-AS1 from different cells (A549 lung adenocarcinoma) and perturbation (siRNA) yielded concordant cis-repressor prediction from TransCistor-analogue (Supplementary Figure S2). This is strong evidence that SBF2-AS1 is a novel cis-repressive lncRNA.

Surprisingly, TransCistor failed to find evidence supporting two previously reported cis-lncRNAs, Paupar27 and Dai28. Inspection of the originating microarray data revealed that, for neither case, do the claimed cis-target genes pass cutoffs of differential expression (Supplementary Figure S3).

Overall, if we consider lncRNAs where at least one method in one dataset is called as cis-acting, then our data implicates 10% (33/323) of lncRNAs as cis-regulators (Figure 3B). When broken down by direction of regulation, we find that 7% (23) of these are activators and 3% (10) are repressors, of which none overlap. We henceforth define the remaining 290 tested lncRNAs as trans-lncRNAs. Together, these findings indicate that TransCistor is capable of identifying known and novel cis-lncRNAs, and a relatively small minority of lncRNAs display significant cis-activity.

TransCistor identifies cis-lncRNA independently of perturbation technology

The perturbation experiments contained a mixture of RNA interference (RNAi) and antisense oligonucleotide (ASO) LOF perturbations. While early experiments were performed using the two RNAi approaches of siRNA and shRNA, it is widely thought that these principally degrade targets in the cytoplasm29,30 or ribosome31. In contrast, ASOs are becoming the method of choice to knock down lncRNAs, since they are thought to act on nascent RNA in chromatin32. If correct, then one would expect ASO perturbations to have greater power to discover cis-lncRNAs. To test this, we compared predictions from each perturbation technology (Figure 3C). Surprisingly, we observed broadly similar rates of cis-lncRNA identification between perturbation methods. However, ASO experiments discover similar rates of activators and repressors, while RNAi perturbations yield an apparent excess of activators over repressors.

We conclude that TransCistor is capable of discovering cis-lncRNAs across perturbation types. While the small numbers preclude statistical confidence, these findings broadly support the use of RNAi in targeting nuclear lncRNAs and identifying cis-lncRNAs, although the possibility for perturbation-specific biases should be further investigated.

Association of cis-lncRNAs with enhancer elements

It has been widely speculated that cis-lncRNAs, particularly activators (ie e-lncRNAs), act in concert with DNA enhancer elements to upregulate target gene expression3,9,12. Our catalogue of cis-lncRNAs represents an opportunity to independently test this. We calculated the rate of overlap of lncRNAs with enhancers using epigenomics data across human tissues (Figure 4A, Supplementary Figure S4). Analyses were performed at a variety of epigenome thresholds (the minimum number of samples required to define a given epigenomic state) and window sizes (the distance from the lncRNA TSS to the nearest epigenome element).
This analysis revealed several intriguing relationships between cis-lncRNAs and enhancer elements. First, we noted an enrichment of super-enhancers at the TSS of cis-activator lncRNAs (boxed, Figure 4B). Inspection of overlaps at other thresholds and window sizes revealed a similar effect (Supplementary Figure S4). Second, we observed an enrichment of Enhancer(1) elements around the TSS of cis-repressor lncRNAs (boxed, Figure 4B), which similarly was corroborated by analyses with a variety of thresholds (Supplementary Figure S4). More broadly, we observed a generalised enrichment of various enhancer element annotations with cis-lncRNAs (Figure 4B, left column). However, we do not observe a preference for such enrichment in cis-activator over cis-repressor lncRNAs (Figure 4B, right column). Overall, within the limits of statistical power given our relatively small sample size, these findings are consistent with a relationship between cis-lncRNAs and enhancer elements.

Some cis-lncRNAs are brought into spatial proximity to their targets by chromatin looping

A second key mechanistic model posits that regulatory interactions between cis-lncRNAs and target genes are effected by close spatial proximity, brought about by chromatin looping (Figure 5A). To measure proximity, we utilised published Hi-C interactions from a range of human cell lines. We evaluated the importance of proximity for regulatory targeting, by combining an asymptotic regression model to predict an “expected interaction” at a given linear genomic distance, with a logistic regression model to evaluate whether strong deviations from this expectation were indicative of targeting (Figure 4C). This approach revealed a significant (p-value ≤ 0.1) contribution of spatial proximity to targeting for two cis-activator lncRNAs: UMLILO (8 cell lines) and DA125942 (1 cell line) (Figure 4D). In both cases, previous studies have implicated chromatin looping in target identification. An excellent example is represented by HUVEC cells, where UMLILO target genes tend to be located in higher proximity (Interaction, y-axis), compared to other non-targets at similar distances in linear DNA (x-axis) (Figure 4E). An alternative inverse square model yielded the same two lncRNAs (Supplementary Figure S5). Together, this indicates that for a subset of cis-lncRNAs, spatial proximity may determine identity of target genes.

Discussion

We have described TransCistor, a modular quantitative method for identification of cis-regulatory lncRNAs. We applied it to a corpus of perturbation datasets to create the first large-scale survey of cis-regulatory RNAs. We evaluated the performance of TransCistor in light of the present state-of-the-art and used the resulting catalogue of cis-lncRNAs to address fundamental questions regarding their prevalence and molecular mechanisms.

TransCistor-digital and -analogue represent practical tools for cis-lncRNA discovery. Previous studies used a “naïve” criterion of ≥1 cis-target gene within an arbitrarily-sized window; however, we show that this method is prone to predominantly false-positive predictions at ≥50 kb windows. TransCistor improves on this situation by making predictions at a defined false discovery rate (FDR). The two distinct statistical methods are designed to capture a range of cis-activity, from lncRNAs regulating the most proximal neighbour gene’s expression within the local TAD, such as Chaserr, to those regulating a more distal target amongst other non-target genes, such as CCAT1-L. The value of resulting predictions is supported by good statistical behaviour as judged by quantile-quantile (QQ) analysis, consistency between methods and datasets, and recall of
numerous known cis-IncRNAs, including founding members H19 and XIST. TransCistor is made available both as a webserver and standalone software. It is compatible with a wide range of input data, since “regulation” files can be readily generated from any experimental dataset comprising IncRNA perturbation and global readout of gene expression changes, including two decades of experiments from microarrays to RNA-sequencing and future parallelised CRISPR LOF methods such as Perturb-Seq.

This work builds on important previous attempts to comprehensively discover cis-regulatory IncRNAs. Basu and Larsson utilised gene expression correlation as a means for inferring candidate cis-regulatory relationships. Very recently, de Hoon and colleagues employed genome-wide RNA-chromatin and chromatin folding to train a predictive model for cis-regulatory IncRNAs. While these methods are valuable, they infer target genes based on indirect correlates of cis-regulation, which may not reflect causation. Furthermore, we failed to find evidence that chromatin folding links cis-IncRNAs to their target genes in all cases. What distinguishes TransCistor from these approaches, is its use of LOF perturbations to directly identify gene targets. We argue that, due to its direct and functional nature, this approach should be considered the gold standard evidence for defining cis-regulatory relationships.

Our results afford important insights into the regulatory IncRNA landscape. Notwithstanding the caveats discussed above, we provide the first global estimate of cis-IncRNA prevalence, suggesting they represent a modest fraction (10%) of the total, with a slight prevalence of activators over repressors. These values are certainly impacted by a variety of errors discussed above, which we hope will be corrected by future, larger-scale studies. The preponderance of cis-activators may be an artefact of RNAi perturbations, which appear to yield an excess of activators over repressors, with no apparent explanation yet. Our results shed light on cis-IncRNAs’ molecular mechanisms, finding evidence supporting their relationship with enhancer elements and, in some cases, a preference to loop into spatial proximity to targets. Surprisingly, we observed evidence that enhancers are associated with both activator and repressor IncRNAs.

Finally, it is worth revisiting the assumptions we make when interpreting IncRNA perturbation experiments. These involve a small oligonucleotide with perfect sequence complementarity to a IncRNA target in RNA and DNA, and assess the outcome in terms of steady state RNA levels. Two key assumptions are made. Firstly, any change in downstream gene expression is assumed to occur through changes in the targeted IncRNA transcript. It is well known that small oligos are not only capable of hybridising to genomic DNA, but also to affect local chromatin modifications, raising the possibility of chromatin/DNA-mediated cis-regulatory mechanisms. The second assumption is more fundamental: that, when local gene changes are observed to occur, such changes reflect the biological function of the IncRNA. The alternative explanation is that perturbations of a IncRNA lead to changes to local gene expression, but that this is a by-product of altering IncRNA expression (e.g. by disrupting local transcription factories), and that the evolutionarily-selected function of the IncRNA is something quite different. In other words, is observed cis-activity a reflection of genuine, adaptive biological regulatory pathway, or is it merely a technical artefact without biological relevance? Testing these alternative explanations will be an interesting challenge for the future, facilitated by the tools provided here.
Methods

TransCistor

TransCistor was developed under the R statistical software (v4.0). Gene locations were extracted from GENCODE annotation file in GTF format (v38 for human, v25 for mouse)\textsuperscript{32} and were converted into a matrix. The TransCistor input consists of a "regulation file", containing all genes and a flag indicating their regulation status: 1 (upregulated after perturbation; repressed by the lncRNA), -1 (downregulated after perturbation; activated by the lncRNA) or 0 (not target). Regulation status can be defined by the user, and here is based on differential expression after IncRNA perturbation. The perturbed IncRNA itself is removed from the regulation file to avoid false positive predictions. Results are visualized with ggplot2 (v3.3.5), ggpubr (v0.4), pheatmap (v1.0.12) packages and custom in-house generated scripts.

TransCistor includes two modules; Digital and Analogue. TransCistor-digital defines cis-IncRNAs based on statistical overrepresentation of proximal targets, defined as targets in the same topologically associated domain (TAD) as the lncRNA. Membership of a TAD is defined based on a gene’s TSS. Digital TransCistor utilizes a collection of TADs for human and mouse cell types accessed via the 3D-Genome Browser\textsuperscript{21}. For each cell type, TransCistor identifies the IncRNA TAD, estimates the number of proximal (within TAD) and distal (outside TAD) targets / non-targets (separately for activated and repressed). Then, it tests for overrepresentation of proximal targets over distal targets by the hypergeometric test. The p-values for all the cell types are then integrated by their harmonic mean. P-values are corrected for multiple hypothesis testing using the False Discovery Rate (FDR) method and taking into account the experiments which show at least one proximal target. TransCistor-analogue evaluates whether the mean distance of targets from the same chromosome are closer than random chance. Distance is defined by TSS to TSS. Analysis is performed separately for activated and repressed targets. Then, the random distribution is calculated, by randomly shuffling the regulation flags on genes within the same chromosome, and recalculating the test statistic each time. By default, 1000 simulations are performed. Finally, the empirical p-value is calculated from the proportion of simulations with a statistic less than the true value.

Both modules of TransCistor are available as a standalone R package and along with all regulation files (https://github.com/pchouvardas/TransCistor) and Rshiny webserver (https://transcistor.unibe.ch/). The input comprises metadata about the lncRNA, and a regulation file containing target gene information that can be readily derived from any transcriptome-wide data including RNA-sequencing, CAGE and microarray experiments.

Collecting and processing perturbation datasets

The FANTOM perturbation datasets were downloaded from the Core FANTOM6 repository\textsuperscript{18}. The differential expression results were transformed into regulation files by applying an adjusted p-value threshold of 0.05 and using custom bash scripts. The respective metadata were also downloaded from FANTOM6 and were integrated to the GENCODE annotation matrix. Perturbation experiments were removed because they target protein coding genes, and an additional 19 were removed because target IncRNAs had no ENSEMBL identifier. The LncRNA2Target datasets were downloaded from the webserver (Version 2.0)\textsuperscript{43} and targets were defined by using an adjusted p-value cutoff of 0.05. The IncRNA locations were manually obtained...
from the website or original publications, when necessary. The rest of the datasets were accessed through the original publications and post-processed to generate the regulation files. All regulation files are available from the project Github repository, linked above.

Analysis of chromatin states

Chromatin states annotations were retrieved from three sources: EpiMap\textsuperscript{44}, genoSTAN\textsuperscript{45}, and dbSUPER\textsuperscript{46}. EpiMap consists of 18 chromatin states across 833 samples, genoSTAN identifies promoter and enhancer regions genome-wide across 127 samples, and dbSUPER aggregates 82234 human superenhancers from 102 cell types/tissues. The annotations were relabelled as follows: Superenhancer – dbSUPER’s superenhancers; Enhancer(1) – genoSTAN’s enhancers; Enhancer(2.1) – EpiMap’s Genic enhancer 1; Enhancer(2.2) – EpiMap’s Active enhancer 1; Enhancer(2.3) – EpiMap’s Weak enhancer; Promoter(1) – genoSTAN’s promoters; and Promoter(2) – EpiMap’s Active TSS.

The human TSS annotations were intersected with chromatin states at several genomic windows (1 bp, 100 bp, 1000 bp, 10000 bp) and a given state-TSS intersection was counted only if it was present in more samples than a given threshold (0, 1, 5, or 10 samples). For each pair of genomic window and filter, a contingency matrix was computed for each pair of predicted labels (\textit{cis}-activator vs. \textit{cis}-repressor, \textit{cis}-activator vs. \textit{trans}, and \textit{cis}-repressor vs. \textit{trans}) or the grouped label (\textit{cis} vs. \textit{trans}), counting the number of TSSs falling into each category. Fisher’s exact test was used to compute the p-value of each contingency matrix.

Chromatin folding analysis

HiC interaction data was obtained using the python package “hic-straw” (v1.2.1) (https://github.com/aidenlab/straw), using human HiC datasets from Aiden laboratory\textsuperscript{33}. The gene coordinates reported by TransCistor were converted to hg19 to match the HiC data. The binning resolution was set to 25 kb, and interaction scores were normalized by Knight-Ruiz matrix balancing method. Due to gaps in the HiC matrices, ~7% of IncRNA: (non-)target interactions were approximated by using a “next best” pair of bins, for which an interaction score was available, instead of the correct binning. In 6.8% of cases this only required replacing either one of the ideal bins by a direct neighbour and for the remaining 0.2% either shifting both genes by one bin or one of the genes by two bins. An estimate for the expected interaction at a given distance was then calculated by fitting a regression model to the HiC data with the interaction score as the response and the TSS distance between the two genes as the explanatory variable. An asymptotic regression model was chosen for this step (‘SSasymp’ and ‘nls’ of the R base package ‘stats’ v4.0.3). Due to model limitations only \textit{cis}-IncRNAs identified by TransCistor digital were included in this analysis. For 2/12 IncRNAs from this subgroup (RAD51-AS1, NARF-AS2), model generation failed for one or more of the cell types. Modelling the interaction as a function of the inverse square distance was also considered (‘glm’ also from ‘stats’). This model had the advantage of not failing for either combination of \textit{cis}-IncRNA and cell type, but fit the data less well and it had a clear bias to underestimate the interaction in close 2D proximity and overestimate interaction further away (Supplementary Figure S5). The significance of interaction on the targeting status was then assessed by fitting a logistic regression model to predict whether a gene is a target of a given IncRNA based on the difference between observed and expected interaction (again using the ‘glm’ function).
**Figures**

**Figure 1**

A. Datasets overview

<table>
<thead>
<tr>
<th></th>
<th>Experiments</th>
<th>LncRNAs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>(385)</td>
<td>72</td>
</tr>
<tr>
<td>Mouse</td>
<td>(140)</td>
<td>130</td>
</tr>
<tr>
<td>Human</td>
<td>(190)</td>
<td>130</td>
</tr>
</tbody>
</table>

B. Target identification

RNA degradation

siRNA/shRNA/ASO

RISC / RnaaseH

C. Naive cisDefinitions

Regulation

Activated

Repressed

Data

Real data

Simulations

Window [50kb to 1mb]

1 target within window [%]

D. TransCistor Digital

Proximal

Non-targets (3x)

IncrNA targets (5x)

Genome-wide (excluding TAD)

Distal

2427 Targets

11842 Non-targets

E. TransCistor Analogue

Actual Data

Simulations

Calculate mean distance

Distances

Calculate mean distance

Shuffle gene labels within chromosome

Distances

Mean Distance $\bar{x}$

$p = \frac{\#(x_{\text{simulated}} \leq x_{\text{real}})}{\# \text{simulations total}}$
Figure 1: TransCistor is a quantitative framework for classifying cis- and trans-regulatory lncRNAs.

A) Origin of the perturbation data: The y-axis displays the number of perturbation experiments (left panel) or individual lncRNA genes (right panel) per model organism (x-axis). Note that the difference between number of perturbation experiments & lncRNAs arises from the fact that many lncRNA genes are represented by > 1 experiment. The bars are split/color-coded according to the origin dataset. The 6 hand curated perturbation experiments are the ones targeting UMLILO, XIST (2), Chaserr, Paupar and Dali. B) Definition of target genes: A target gene is defined as one whose expression significantly changes after loss-of-function perturbation of a given lncRNA (pink). The direction of that change (down/up) defines the target as activated/repressed (green, orange), respectively. C) Evaluating accuracy of naïve cis-lncRNA definition: The plot displays the number of lncRNAs classified as “cis-regulatory” using a definition of ≥1 proximal target genes (y-axis), while varying the size of the genomic window (centred on the lncRNA TSS) within which a target is defined as “proximal” (x-axis). Line: real data from Panel A; Boxplot: Simulations created by 50 random shuffles of the target labels across all annotated genes. D) TransCistor-digital method: TransCistor-digital evaluates the enrichment of targets (green) in proximal regions, defined as those residing within the same topologically associating domain (TAD) as the lncRNA TSS (pink) (left panel), compared to the background target rate in the rest of the genome (“Distal”) (centre panel). Cis-lncRNAs are defined as those having a significantly higher proximal target rate, defined using hypergeometric test (right panel). E) TransCistor-analogue method: A distance statistic is defined as the mean genomic distance (bp) of all targets (green) on the same chromosome as the lncRNA (pink) (left panel). 1000 simulations are performed where target labels are shuffled across genes within the same chromosome (centre panel). Cis-lncRNAs are defined as those whose real statistic (dashed line) falls below the majority of simulations (right panel).
Figure 2: Large scale classification of cis-lncRNAs in human and mouse.

A) Quantile-quantile plot displays the random expected (x-axis) and observed (y-axis) p-values for lncRNAs (points) tested for activated targets by TransCistor-digital. The grey diagonal y=x line indicates the expectation if no hits were present. B) As for (A), for TransCistor-digital and repressed targets. C) Comparison of activator and repressor activity detected by TransCistor-digital. For each lncRNA (points), their false-discovery rate (FDR)-adjusted significance is plotted on the x-axis (activator) and y-axis (repressor). Note the absence of lncRNAs that are both activators and repressors. D) UMLILO, an example cis-activator: The plot shows the number of
genes, divided by targets / non-targets (colour / grey), location (distal/proximal) and regulation direction (activated/repressed). *UMLILO* is classified as a *cis*-activator, due to the significant excess (8) of proximal activated targets. Statistical significance (uncorrected) is displayed above.

**E)** *UMLILO* genomic locus: Vertical bars denote gene TSS. Grey: non-targets; green: activated targets; pink: *UMLILO*. **F)** As for (A), for TransCistor-analogue and activated targets. **G)** As for (B), for TransCistor-analogue and repressed targets. **H)** As for (C), for TransCistor-analogue. **I)** *DNAAF3-AS1*, an example *cis*-repressor identified by TransCistor-analogue. Shown is the target distance statistic (x-axis) for real data (vertical bar) and simulations (boxes). The number of simulations in each distance bin is displayed on the y-axis. **J)** As for (I), for a second perturbation experiment.
Figure 3: Rate of cis-lncRNA across perturbations, datasets and species.

A) Summary of TransCistor results: The values represent numbers of experiments classified in the bins indicated on the two axes, at a cutoff of FDR ≤ 0.5. The names of lncRNA genes are displayed. Previously-described cis-lncRNAs are red. B) The rate of lncRNA genes defined to be cis-regulatory based on our analysis. Note that one single experiment is sufficient to label a lncRNA gene as cis-regulatory. C) The rate of experiments defined as cis-regulatory, broken down by perturbation method.
Figure 4: Intersection of cis-lncRNAs with enhancer elements and generation of a chromatin conformation based model.

A) Method of calculating overlap by enhancer annotations (horizontal purple bars) of IncRNA TSS (pink bar). Overlaps are considered while varying two important thresholds: numbers of individual enhancer annotations are considered the minimum necessary (epigenome threshold) and various sized windows around the TSS of overlap calculation between extended TSS region (span). Only the TSS spans with overlaps in more samples than a given epigenome threshold are considered.

B) Enrichment results epigenome threshold=1 and span=100 bp. Rows show enrichment for super-enhancer, enhancer, and promoter states while comparing the TSS according to their mechanism of action (see Methods). P-values were computed using Fisher’s Exact Test.

C) A model for proximity-driven target selection: (Left panel) Chromatin folding brings IncRNA (pink) into spatial proximity with proximal genes, which are subsequently targeted (green). (Right panel) Chromatin proximity maps, such as provided by HiC methodology, enable one to evaluate the spatial proximity (y-axis) of targets, while normalising for confounder of linear 2D DNA distance (x-axis). These parameters were modelled using an Asymptotic regression model (right panel, inset).

D) Evaluating the contribution of proximity to target selection in human cells: The model significance of cis-lncRNAs (identified by TransCistor-digital) (x-axis) was evaluated across HiC interaction data from a panel of human cell lines (y-axis). Colour scale shows uncorrected p-values. Green cells indicate cases where target genes tend to be significantly more proximal than non-targets. No cases of the inverse were observed.

E) Example data for UMLILO in HUVEC cells. Note that target genes (green) tend to be more spatially proximal (y-axis) than non-target genes (grey) at a similar TSS-to-TSS genomic distance (x-axis).
Supplementary Figures

Supplementary Figure 1

A) Histogram displaying the numbers of separate perturbation experiments (x-axis) available for each lncRNA gene (y-axis). B) Histograms displaying the number of significantly changing genes (targets) (x-axis) for each perturbation experiment (y-axis). Regulated genes represent the union of activated and repressed genes.

Supplementary Figure S1: Summary statistics of perturbation datasets.

A) Histogram displaying the numbers of separate perturbation experiments (x-axis) available for each lncRNA gene (y-axis). B) Histograms displaying the number of significantly changing genes (targets) (x-axis) for each perturbation experiment (y-axis). Regulated genes represent the union of activated and repressed genes.
Supplementary Figure 2: Analysis of cis-regulation by SBF2-AS1 in independent datasets.

(A) TransCistor-analogue results for FANTOM ASO knockdown targeting SBF2-AS1 in human dermal fibroblasts. B) As for (A), but for independent data from A549 cells treated with siRNA. C) Numbers indicate the genes in each category, classified by their regulation in the two distinct datasets in (A) and (B).
Supplementary Figure S3: Analysis of Dali and Paupar target genes using public microarray data.

Supplementary Figure 4
Supplementary Figure S4: Enrichment of enhancers, super-enhancers and promoters in cis-
IncRNAs. Each row represents a different enhancer annotation (see Methods). Columns
represent comparisons between indicated pairs of IncRNA classes. Heatmaps display the
enrichment of overlap at different genomic windows around the IncRNA TSSs (span, x-axis) and
the minimum number of observed samples required to define an enhancer (epigenome threshold,
y-axis). Statistical significance is indicated where p-value ≤ 0.05 by Fisher’s exact test (1-sided),
and not corrected for multiple hypothesis testing.
Supplementary Figure S5: Alternative inverse square model for target gene interaction.

A) and B) are equivalent to main Figure 4 panels D and E respectively but using an inverse square model instead of an asymptotic regression. C) Comparison of asymptotic (solid line) and inverse square (dashed line) regression models.
Supplementary Data Files

Data File 1: Information for all IncRNAs studied in this work.
Acknowledgements

We thank the other members of GOLD Lab for insightful feedback and discussions. Hyeonjoo Lee and Jin-Wu Nam (Hanyang University, Republic of Korea) kindly shared Xist knockdown RNA-sequencing data. All computation was carried out on the University of Bern Interfaculty Bioinformatics Unit cluster maintained by Rémy Bruggman and Pierre Berthier. We acknowledge administrative and logistical support from Basak Ginsbourger, Rahel Tschudi, Marla Rittiner, Beatrice Stalder, Willy Hofstetter and Patrick Furer (DBMR, University of Bern). HG-R is supported by a Marie Skłodowska-Curie Actions postdoctoral fellowship (R21950). Work in the GOLD Lab is funded by the Swiss National Science Foundation through the National Centre of Competence in Research (NCCR) “RNA & Disease” (51NF40-182880), Project funding “The elements of long noncoding RNA function” (31003A_182337), Sinergia project “Regenerative strategies for heart disease via targeting the long noncoding transcriptome” (173738), by the Medical Faculty of the University and University Hospital of Bern, by the Helmut Horten Stiftung, Swiss Cancer Research Foundation (4534-08-2018), and Science Foundation Ireland through Future Research Leaders award 18/FRL/6194.
References


