Functional identification of cis-regulatory long noncoding RNAs at controlled false-discovery rates

AUTHORS

Panagiotis Chouvardas[1,2], Bhavya Dhaka[3,4]. Marc Zimmerli[1,2], Daniel Hanhart[1,2], Mario Moser[1,2], Hugo Guillen-Ramirez[1,2,3,4], Sanat Mishra[5], Roberta Esposito[1,2], Taisia Polidori[1,2], Maro Widmer[1,2], Raquel García-Pérez[6], Marianna Kruithof-de Julio[2,7], Dmitri Pervouchine[8], Marta Melé[6], Rory Johnson[*1,2,3,4]

[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]School of Biology and Environmental Science, University College Dublin, Dublin D04 V1W8, Ireland.

[4]Conway Institute for Biomolecular and Biomedical Research, University College Dublin, Dublin D04 V1W8, Ireland.

[5]Indian Institute of Science Education and Research, Mohali, India.

[6]Life Sciences Department, Barcelona Supercomputing Centre, 08034, Barcelona, Spain.

[7]Department of Urology, Inselspital, Bern University Hospital, Bern, Switzerland.

[8]Center of Life Sciences, Skolkovo Institute of Science and Technology, Moscow, Russia.

* To whom correspondence should be addressed. Email: rory.johnson@ucd.ie

ABSTRACT

Long noncoding RNAs (lncRNAs) are believed to regulate expression of neighbouring target genes. However, such 'cis-lncRNAs' are presently defined using ad hoc criteria that, we show, are prone to false-positive predictions. The resulting lack of confident cis-lncRNA catalogues hinders our understanding of their extent, characteristics and mechanisms. Here, we introduce TransCistor, a framework for defining and identifying cis-lncRNAs based on enrichment of targets amongst proximal genes. TransCistor's simple and conservative statistical models are compatible with functionally-defined target gene maps generated by existing and future technologies, and generate predictions at controlled false discovery rates. Using transcriptome-wide perturbation experiments for 268 human and 134 mouse lncRNAs, we provide the first large-scale survey of cis-lncRNAs. Our results suggest that cis-activity is confined to a minority of lncRNAs, with a prevalence of activators over repressors. Cis-lncRNAs are detected by both RNA interference (RNAi) and antisense oligonucleotide (ASO) perturbations. Mechanistically, we find weak associations of cis-lncRNAs with enhancer elements. Moreover, they are not distinguished from other lncRNAs by evolutionary conservation nor nuclear enrichment, nor are they frequently linked to target genes by chromatin looping. In summary, our study will enable researchers to evaluate the regulatory mode of their lncRNA of interest at controlled false discovery rates and places cis-lncRNAs on a quantitative foundation for the first time.

INTRODUCTION

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 chromosome (Brown et al. 1991; Thorvaldsen, Duran, and Bartolomei 1998). Protein-coding 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 (U. A. Ørom et al. 2010; Rom et al. 2019). Conversely, other lncRNAs have no apparent positional preference for their targets and are termed trans-lncRNAs (Leucci et al. 2016). This cis/trans duality provides a fundamental framework for understanding regulatory lncRNAs (Gil and Ulitsky 2020), 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 base pairs (Rom et al. 2019) to hundreds of kilobases (Xiang et al. 2014)) and a number of targets (one(Rom et al. 2019) to many (Fanucchi et al. 2019)). Two interrelated molecular mechanisms have been proposed: enhancer elements and chromatin folding (Gil and Ulitsky 2018). Some cis-activating lncRNAs, termed "enhancer lncRNAs" (e-lncRNAs), have been found to overlap DNA-encoded enhancer elements (Gil and Ulitsky 2018; Engreitz et al. 2016; Ulf Andersson Ørom and Shiekhattar 2011; Jennifer Y. Tan et al. 2020), similar to lncRNAs more generally (Marques et al. 2013). 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) (Gil and Ulitsky 2020). In contrast, trans-acting lncRNAs are thought to diffuse through the nucleus or cytoplasm and find their targets via molecular recognition, for example by hybridisation (Lee et al. 2017). An attractive corollary of these models is that cis-regulatory lncRNAs may act via non-sequence-dependent mechanisms, perhaps involving phase separation (Hnisz et al. 2017) and local concentration gradients (Wu, Yang, and Chen 2021). 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-regulation (Palazzo and Koonin 2020). 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 lncRNA regulatory modes.

Regulatory lncRNA catalogues will require a rigorous and agreed definition for cis-lncRNAs, which is presently lacking. Until now, they have been defined simply by the existence of ≥1 proximal target. Targets are defined as those whose expression changes (even weakly) in response to lncRNA LOF, as measured using single-gene (RT-PCR) or whole-transcriptome (RNA-seq, CAGE, microarray) techniques (Leucci et al. 2016; U. A. Ørom et al. 2010; Ramilowski et al. 2020). "Proximity" is defined on a case-by-case basis, using a wide range of windows spanning 10^2 to 10^5 bp (Xiang et al. 2014). 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 gene by random chance.

For example, consider a lncRNA having 10 proximally-encoded genes, and 2000 targets genome-wide (10% of all protein-coding genes); one would expect to observe 1 proximal target by random chance alone (10% of 100). Therefore, the conventional "naïve" definition of cis-lncRNAs, where key parameters of global target number, window size and target definition remain unconsidered or undefined, suffers from an inherent risk of false-positive predictions.

In this study, we consider cis-lncRNAs from a quantitative perspective. We show that conventional definitions are prone to high false positive rates. We introduce statistical methods for the definition of cis-lncRNAs at controlled false discovery rates and use them to classify regulatory lncRNAs across hundreds of perturbation datasets. The resulting catalogue of cis-lncRNAs enables us to evaluate hypotheses regarding their molecular mechanisms of action.

MATERIAL AND 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 humans, v25 for mouse) (Frankish et al. 2020) 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 lncRNA perturbation. The perturbed lncRNA 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-lncRNAs based on the 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 (Wang et al. 2018). For each cell type, TransCistor identifies the lncRNA TAD and estimates the number of proximal (within TAD) and distal (outside TAD) targets / non-targets, separately for activated and repressed genes. Then, it tests for the overrepresentation of proximal targets over distal targets using the twoby2Calibrate R package. Statistical significance is estimated based on the mid-p-value calibrated Fisher's test, for each TAD dataset/cell type. The p-values for all the cell types are then integrated by their harmonic mean. The 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 is 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, 10,000 simulations are performed.

Finally, the empirical p-value is calculated from the proportion of simulations with a test statistic less than the true value.

Both modules of TransCistor are available as a standalone R package along with all regulation files (https://github.com/gold-lab/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 (Ramilowski et al. 2020; Yip et al. 2022). 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 into the GENCODE annotation matrix. 31 perturbation experiments were removed because they target protein-coding genes, and an additional 19 were removed because target lncRNAs had no ENSEMBL identifier. The LncRNA2Target datasets were downloaded from the webserver (Version 2.0) (Cheng et al. 2019) and targets were defined by using an adjusted p-value cutoff of 0.05. The lncRNA 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 at the project Github repository, linked above.

TransCistor concordance score

To evaluate the consistency of lncRNA classification for TransCistor, we calculated a concordance score based on lncRNAs with more than one perturbation experiment for both modes separately. We then randomly shuffled the classification labels (cis-activator, cis-repressor, not significant, or no target/TAD found.) 1000 times to create a null distribution and recalculated the score. The actual scores for TransCistor-digital and analogue were then compared with the null distribution to assess whether they provided consistent classifications.

Analysis of subcellular localisation and expression

The data used in the initial subcellular localization analysis (Supplementary Figure S6) was downloaded from lncATLAS (Mas-Ponte et al. 2017). The list of lncRNAs considered cis- or trans-acting respectively is based on the predicted activity as reported in Supplementary File 1. Excluding genes with no associated ENSEMBL ID, mouse genes, and genes for which either the cytosolic or nuclear expression level was missing in the lncATLAS data resulted in a reduction from 33 to 20 cis- and 290 to 133 trans-acting lncRNAs for the purpose of this comparison. The Wilcoxon rank sum test was then used to check for significant differences between cis- and trans-acting lncRNAs (1-sided, alternative: cis > trans) for a) log-transformed cytosolic/nuclear ratios, and b) total log2 FPKM cell expression.

LncRNA evolutionary conservation analysis

Data were obtained from the LnCompare (Carlevaro-Fita et al. 2019) data tables. PhastCons scores were utilized for the hg38 human genome assembly to obtain conservation scores of both promoter and exon regions. Three models were employed namely 7, 20, and 100 species.

For each category, we first classified the lncRNAs into two groups: cis and trans (same groups as for subcellular localisation). Then, we compared the PhastCons scores between these two groups using a Welch test, with a one-sided hypothesis that cis lncRNAs have greater conservation compared to trans lncRNAs. The comparisons were performed separately for promoter and exon regions of the lncRNAs in each of the three models.

Analysis of chromatin states

Chromatin states annotations were retrieved from three sources: EpiMap (Hoon, Rahimzadeh, and Bustos 2021), genoSTAN (Zacher et al. 2017), and dbSUPER (Khan and Zhang 2016) EpiMap consist 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 windows and filter, a contingency matrix was computed for each pair of predicted labels (*cis*-activator vs. cis-repressor, cis-activator vs. trans, and cis-repressor vs. *trans*) or the grouped label (*cis* vs. *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 were obtained using the Python package "hic-straw" (v1.2.1) (https://github.com/aidenlab/straw), using human HiC datasets from Aiden laboratory (Durand et al. 2016). 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 lncRNA: (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 with 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. After visual inspection of the data, 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 as well as unclear comparability of TAD-based and TAD-independent cis-regulation, only cis-lncRNAs identified by TransCistor-digital were included in this analysis. For 2/12 lncRNAs from this subgroup (RAD51-AS1, NARF-AS2), model generation failed for one or more of the cell types considered. 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 cis-lncRNA and cell type, but fit the data visually 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 lncRNA based on the difference between observed and expected interaction (glm function also from 'stats').

RESULTS

A quantitative, functional definition of cis-lncRNAs

To understand the lncRNA-target relationship we explored transcriptional profile changes from multiple perturbation studies. We employed a functional definition of "targets", as genes whose steady-state levels significantly change in response to a given lncRNA's LOF (Figure 1A). We further define targets as activated or repressed, where they decrease / increase in response to lncRNA loss-of-function (LOF), respectively. Overall we collected 488 lncRNA LOF experiments targeting 268 human lncRNAs from a mixture of sources, including the recently published datasets of ASO knockdowns in human dermal fibroblasts and induced pluripotent stem cells (iPSCs) from the FANTOM consortium (Ramilowski et al. 2020; Yip et al. 2022). To this we added 140 experiments for 134 lncRNAs in mouse (Figure 1B). Among these we included 6 hand-curated previously-reported cis-acting lncRNAs (*UMLILO*, *XIST x2*, *Chaserr*, *Paupar* and *Dali*). Of the total lncRNAs, 130 were represented by at least one independent experiment, and the median number of target genes identified per experiment was 55 (Supplementary Figure S1 B).

We first evaluated the performance of the widely-used naïve definition for cis-lncRNAs, defined as ≥1 target within an arbitrary distance window. Using a range of window sizes from 50 kb to 1 Mb centred on the lncRNAs' TSSs, we evaluated the fraction of lncRNAs that would be defined as cis-acting under this definition. This approach defines ~2 to 12% of lncRNAs 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 genes and repeated this analysis. The rate of cis-lncRNA predictions in this random data overlapped the true rates in all windows (Figure 1C, boxplots), suggesting that the naïve definition of cis-lncRNAs yields high rates of false-positive predictions.

To overcome this issue, we adopted a new definition of cis-lncRNAs: *cis-lncRNAs 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-lncRNAs.

TransCistor: digital and analogue identification of cis-lncRNAs

The first cis-lncRNA identification 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 (Materials and Methods) in proximal genes (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 studies (Jennifer Yihong Tan et al. 2017). Chromatin folding varies to an extent between cell types. Therefore, TransCistor-digital calculates enrichment across a set of experimentally-defined cell-type-specific TADs (45 human, 3 mouse) (Wang et al. 2018) and aggregates the resulting P-values by their harmonic mean.

The above TAD-window approach is effective, yet has drawbacks. Several reported cis-lncRNAs have individual targets that are not immediately adjacent (Xiang et al. 2014), and might be

overlooked by the digital approach. Furthermore, many lncRNAs 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 testing for proximal enrichment of 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. Statistical significance can be estimated empirically, by generating a null distribution based on randomisation of target labels (Figure 1E). Now, cis-lncRNAs are defined as those having a distance statistic that is lower than the majority of randomisations.

We sought to test the performance of TransCistor-digital and evaluate the global landscape of cis-lncRNAs. After filtering out unusable datasets (having no targets, or no overlapping TAD), 195 datasets remained. We discovered 23 cis-acting lncRNAs (14 activators, 9 repressors), at a false discovery rate (FDR) threshold of 0.25 (Figure 2A-C). The majority of p-values produced by this analysis follow the null distribution, underlining the conservative statistical behaviour of TransCistor (Figure 2A,B). All cis-lncRNAs have a uniquely activator or repressor assignment, with the exception of *Evx1os*, which is classified as having both activating and repressing characteristics (Figure 2C). Amongst the top-ranked cis-lncRNAs is *UMLILO*, previously reported to activate multiple genes in its local genomic neighbourhood (Fanucchi et al. 2019). *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, on the other hand, identified 15 cis-lncRNAs (9 activators, 6 repressors, FDR≤0.25). Statistical behaviour is good (Figure 2F,G), while cis-lncRNAs are cleanly split between activators and repressors (Figure 2H). *LINC00240* was identified as the most significant activator cis-lncRNA by both TransCistor-digital and analogue (Figure 2I,J). Similar to previous findings we observed a strong cis-regulation of nearby histone genes (Figure 2I, Supplementary Figure S2 A). (Ramilowski et al. 2020; Yip et al. 2022).

The usefulness of these methods is supported by their internal and external consistency. Together, the TransCistor approaches correctly identify previously-described cis-activators *H19 (Forné et al. 1997)*, RP11-398K22.12 (Ramilowski et al. 2020; Yip et al. 2022), *JPX (Carmona et al. 2018), Evx1os (Luo et al. 2016)*, LINC01615 (Hu et al. 2022) and *DA125942 (Maass et al. 2012)* amongst the top-ranked cis-activators, while both independent experiments for *XIST* are amongst the top repressors (Engreitz et al. 2013).

TransCistor predicted cis-regulatory activity for several known lncRNAs that have never been described as such in prior literature. These include *CAT2, NARF-AS2, BANCR, CD27-AS1* (*cis*-activators), and *SBF-AS1*, *LASTR, NORAD*, and *DANCR* (*cis*-repressors). However, the latter two are only identified in one out of multiple independent perturbation experiments (6 and 5 for *NORAD* and *DANCR* respectively). In the case of *DANCR*, 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 signal, suggesting that they are false-positive predictions. On the other hand, analysis of an independent dataset for *SBF-AS1* from different cells (A549 lung adenocarcinoma) and perturbation (siRNA), which was not included in our original dataset, yielded concordant cis-repressor prediction from TransCistor-analogue (Supplementary Figure S2 B-D). Both human and mouse orthologues of *CHASERR* (ENSG00000272888) are identified as cis-*repressors (Rom et al. 2019)* (Figue2B). Of

the hits, two are concordantly classified by ≥2 independent experiments (*XIST* & *DNAAF3-AS1* classified as cis-repressors based on two separate experiments each). (Supplementary Figure S2 E-F). Together, these data support the accuracy of TransCistor in predicting cis-lncRNAs.

Conversely, TransCistor failed to find evidence supporting several previously reported cis-lncRNAs, namely *Paupar (Vance et al. 2014)* and *Dali* (Chalei et al. 2014). Inspection of the originating microarray data revealed that, for neither case, the reported cis-target genes pass cutoffs of differential expression (Supplementary Figure S3), suggesting that these lncRNAs are not cis-regulatory.

A summary of the entire set of TransCistor predictions is found in Figure 3A. We observed a significant overlap between the two TransCistor methods for classified activators (P=0.0001) and repressors (P=0.001), with 5 cis-lncRNAs in common (LINC00240, linc1427, DA125942 DNAAF3-AS1and XIST). Overall, if we consider lncRNAs where at least one method in one dataset is called cis-acting, then our data implicate 7.46% (30/402) of lncRNAs as cis-regulators. When broken down by direction of regulation, we find that 4.97% (20/402) are activators and 2.48% (10/402) are repressors, with one being identified as both (Figure 3B). Henceforth, we defined the remaining 372 tested lncRNAs provisionally as 'trans-lncRNAs'.

To further test the value of TransCistor predictions, we evaluated the consistency of cis predictions between independent perturbation experiments of the same lncRNA. Using simulations to evaluate significance, we observed a significant (P<0.001) concordance between experiments for TransCistor-digital and analogue methods (Figure 3C). Together, multiple lines of evidence support the ability of TransCistor to identify known and novel cis-lncRNAs.

TransCistor cis-lncRNA identification across perturbation technologies

Our transcriptomic dataset 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 located in the cytoplasm (Stalder et al. 2013; Zeng and Cullen 2002) or ribosome(Biasini et al. 2021). In contrast, ASOs are becoming the method of choice to knock down lncRNAs, since they are thought to act on nascent RNA in chromatin (Lai et al. 2020). Given that cis-lncRNAs presumably act locally in chromatin, then one would expect ASO perturbations to have greater power to discover cis-lncRNAs. To test this, we evaluated predictions from each perturbation technology separately (Figure 3D). Surprisingly, we observed broadly similar rates of cis-lncRNA identification between perturbation methods, supporting the notion that RNAi is readily active in the nucleus (Johnson and Corey 2023). However, ASO experiments appear to discover similar rates of activators and repressors, while RNAi perturbations yield an apparent excess of activators over repressors, together suggesting that differences do exist between RNAi and ASO perturbations.

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.

Clues to cis-lncRNA mechanisms from localisation, expression and evolutionary conservation

We next sought insights into cis-lncRNA mechanisms by examining a range of features related to functionality and subcellular localisation. Although it has been previously postulated that

cis-acting lncRNAs are more localized in the nucleus than the cytosol and have an overall lower RNA expression level (Gil and Ulitsky 2020), we surprisingly found no evidence of any difference in nuclear/cytoplasmic localization between cis-lncRNAs and other lncRNAs (Figure 4A, Supplementary Figure S4 A). Thus, cis-activity does not impact a lncRNA's rate of nuclear export. Similarly, we evaluated the whole-cell expression level of cis-lncRNAs and observed a trend for average expression levels to exceed those of trans-lncRNAs in a number of cases, although these differences did not reach statistical significance (Figure 4B, Supplementary Figure S4 B).

Evolutionary conservation and gene expression patterns are considered to yield important clues to lncRNA functionality (Necsulea et al. 2014). Promoter conservation and high expression are taken as general evidence for functionality (Derrien et al. 2012), whereas conservation specifically in exons is reflective of the functionality of mature RNA transcripts (Haerty and Ponting 2015). Some mechanistic models for cis-lncRNAs posit that they act through non-sequence dependent features (Gil and Ulitsky 2020), predicting that cis-lncRNAs' exons display background levels of evolutionary conservation, whereas their promoter regions (controlling expression) are more conserved. To test this, we evaluated the evolutionary conservation for promoter and exons for three different vertebrate phylogenies (Figure 4C-D). This revealed no discernible difference in exonic or promoter conservation for cis-lncRNAs over trans-lncRNAs. A lack of statistical power due to a low sample size and high variance should be noted.

Together, these findings suggest that cis-lncRNAs have similar characteristics to other lncRNAs in terms of expression, subcellular localisation and overall functionality.

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 expression (Gil and Ulitsky 2018; U. A. Ørom et al. 2010; Jennifer Y. Tan et al. 2020). Our catalogue of cis-lncRNAs represents an opportunity to independently test this. To do so, we calculated the rate of overlap of lncRNAs with enhancers using epigenomics data across human tissues (Figure 5A-B, Supplementary Figure S5). 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 trends of association between cis-lncRNAs and enhancer elements, although none reached statistical significance at the given sample size (Figure 5B). Broadly, we observed a generalised enrichment of various enhancer element annotations with cis-lncRNAs, but not preferably for activators. Inspection of overlaps at other thresholds and window sizes revealed a similar effect (Supplementary Figure S5). Within the limits of statistical power given our relatively small sample size, these findings suggest only a weak relationship between cis-lncRNAs and enhancer elements.

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

A second mechanistic model posits that regulatory interactions between cis-lncRNAs and target genes are defined by spatial proximity, brought about by chromatin looping (Figure 5C). Once again, our cis-lncRNA catalogue makes it possible to test this. To measure proximity, we

utilised published Hi-C interactions from a range of human cell lines (Durand et al. 2016). 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 5C). This approach revealed a significant (-log10(P)≥ 1.3) contribution of spatial proximity to targeting for cis-activator lncRNAs UMLILO (8 cell lines) (Figure 5D). An alternative inverse square model also yielded DA125942 (8 cell lines) in addition with UMLILO (Supplementary Figure S6C). In both cases, previous studies have implicated chromatin looping in target identification (Fanucchi et al. 2019; Maass et al. 2012). 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 5E). Together, this indicates that for a subset of cis-lncRNAs, spatial proximity may determine the identity of target genes.

DISCUSSION

We have described TransCistor, a pair of quantitative methods for the 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. Their 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 (Rom et al. 2019), to those regulating a more distal target amongst other non-target genes, such as CCAT1-L (Xiang et al. 2014). 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-lncRNAs, 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 lncRNA perturbation and global readout of gene expression changes, including two decades of experiments from microarrays to RNA-sequencing and cap analysis of gene expression (CAGE). Importantly, TransCistor is ready to deploy with recently-developed and future parallelised CRISPR LOF methods such as Perturb-Seq (Dixit et al. 2016), raising the possibility of comprehensive mapping of cis- and trans-regulatory lncRNAs in the near future. It is important to note, however, that identification of cis-target genes requires whole transcriptome datasets, meaning that signature methods based on subsets of genes, such as LINCS (Evangelista et al. 2022), will not be suitable.

This work builds on important previous attempts to comprehensively discover cis-regulatory lncRNAs. Basu and Larsson utilised gene expression correlation as a means for inferring candidate cis-regulatory relationships (Basu and Larsson 2018). Very recently, de Hoon and colleagues employed genome-wide RNA-chromatin and chromatin folding to train a predictive model for cis-regulatory lncRNAs (Agrawal et al. 2022). While these methods are valuable, they infer target genes based on indirect correlates of cis-regulation, which often do not reflect causation (Goyal et al. 2017). Furthermore, we failed to find evidence that chromatin

folding links cis-lncRNAs 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.

A key insight from this work is the low statistical power available to identify cis-lncRNAs and the consequent high rate of false-positive predictions. 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). Nevertheless, the statistics of cis-lncRNA prediction depend on the distribution of target and non-target genes around the lncRNA in question. This means that statistical power to identify cis-lncRNAs is inherently constrained by biology. Several likely examples of false positive predictions are well-studied lncRNAs NORAD and DANCR, which are represented by numerous perturbation experiments in our dataset, where only one was called a hit. Conversely, several factors likely contribute to false negatives, including low statistical sensitivity (i.e. presence of few cis-targets relative to trans-targets) and correction for multiple testing. Furthermore, it is likely that direct target mRNAs in turn regulate numerous, indirect downstream genes, creating a high background that further reduces detection sensitivity. This latter issue might be addressed in future by performing transcriptomic analysis at short timepoints immediately after lncRNA LOF to observe direct target genes. Finally, the majority of lncRNAs were tested in only a single cell type, and it is entirely possible that lncRNAs display cis-regulatory activity in a cell-type specific manner. These considerations should prove useful in the design of future experiments to identify cis-lncRNAs. To minimise false-positive predictions, we recommend that colleagues perform multiple independent perturbations (e.g. two or more distinct ASO sequences), perform transcriptomic analysis at earliest possible timepoints, and only consider cis-lncRNAs on the basis of ≥2 consistent results.

Our results afford important insights into the regulatory lncRNA landscape. Notwithstanding the caveats discussed above, we provide the first global estimate of cis-lncRNA prevalence, suggesting they represent a modest fraction (8%) 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. Our results shed light on cis-lncRNAs' molecular mechanisms, finding evidence challenging the notion that spatial proximity defines lncRNA targets. Surprisingly, cis-lncRNAs are not preferably localized to the nucleus, nor are more evolutionarily conserved or more expressed than trans-acting lncRNAs. On the other hand, we find non-significant associations of cis-lncRNAs with enhancer elements, which may point to a mechanistic relationship that awaits further investigation with greater sample sizes. Overall, the data presented here do not resoundingly support present views of cis-lncRNAs' molecular mechanism.

Finally, it is worth revisiting the assumptions we make when interpreting lncRNA perturbation experiments. These involve a small oligonucleotide with perfect sequence complementarity to a lncRNA target in both 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 lncRNA transcript. It is well known that small oligos are not only capable of hybridising to genomic DNA (Scharner et al. 2019) but

also affecting local chromatin modifications (Marasco et al. 2022), raising the possibility of chromatin/DNA-mediated cis-regulatory mechanisms that could be misinterpreted as lncRNA-mediated effects. The second assumption is more fundamental:, when local gene changes are observed to occur, such changes reflect the biological function of the lncRNA (Doolittle 2018; Graur et al. 2013). The alternative explanation is that perturbations of a lncRNA lead to changes to local gene expression, but that this is a by-product of altering lncRNA expression (e.g. by disrupting local transcription factories), and that the evolutionarily-selected function of the lncRNA is something quite different. In other words, is observed cis-activity a reflection of a 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.

DATA AVAILABILITY

Most of the data underlying this article were derived from sources in the public domain: FANTOM (https://fantom6-collaboration.gsc.riken.jp//webdav/home/), lncRNA2Target (http://bio-annotation.cn/lncrna2target/). XIST knockdown RNA-sequencing data were kindly provided by Hyeonjoo Lee and Jin-Wu Nam (Hanyang University, Republic of Korea). TransCistor source code, all regulation files and related information can be found on GitHub (https://github.com/gold-lab/TransCistor). The webserver can be accessed at https://transcistor.unibe.ch/.

SUPPLEMENTARY DATA

Supplementary Data (Supplementary File 1) are available online.

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).

FUNDING

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

1. Agrawal, Saumya, Ivan V. Kulakovskiy, Jessica Severin, Masaru Koido, Tanvir Alam, Imad Abugessaisa, Andrey Buyan, et al. 2022. "Systematic Identification of Cis-Interacting lncRNAs and Their Targets." bioRxiv. https://doi.org/10.1101/2021.01.13.426305.
2. Basu, S., and E. Larsson. 2018. "A Catalogue of Putative Cis-Regulatory Interactions Between Long Non-Coding RNAs and Proximal Coding Genes Based on Correlative Analysis Across Diverse Human Tumors." G3 8 (6). https://doi.org/10.1534/g3.118.200296.
3. Biasini, A., B. Abdulkarim, S. de Pretis, J. Y. Tan, R. Arora, H. Wischnewski, R. Dreos, M. Pelizzola, C. Ciaudo, and A. C. Marques. 2021. "Translation Is Required for miRNA-Dependent Decay of Endogenous Transcripts." The EMBO Journal 40 (3). https://doi.org/10.15252/embj.2020104569.
4. Brown, C. J., A. Ballabio, J. L. Rupert, R. G. Lafreniere, M. Grompe, R. Tonlorenzi, and H. F. Willard. 1991. "A Gene from the Region of the Human X Inactivation Centre Is Expressed Exclusively from the Inactive X Chromosome." Nature 349 (6304): 38–44.
5. Carlevaro-Fita, Joana, Taisia Polidori, Monalisa Das, Carmen Navarro, Tatjana I. Zoller, and Rory Johnson. 2019. "Ancient Exapted Transposable Elements Promote Nuclear Enrichment of Human Long Noncoding RNAs." Genome Research 29 (2): 208–22.
6. Carmona, Sarah, Benjamin Lin, Tristan Chou, Katti Arroyo, and Sha Sun. 2018. "LncRNA Jpx Induces Xist Expression in Mice Using Both Trans and Cis Mechanisms." PLoS Genetics 14 (5): e1007378.
7. Chalei, V., S. N. Sansom, L. Kong, S. Lee, J. F. Montiel, K. W. Vance, and C. P. Ponting. 2014. "The Long Non-Coding RNA Dali Is an Epigenetic Regulator of Neural Differentiation." eLife 3 (November). https://doi.org/10.7554/eLife.04530.
8. Cheng, Liang, Pingping Wang, Rui Tian, Song Wang, Qinghua Guo, Meng Luo, Wenyang Zhou, Guiyou Liu, Huijie Jiang, and Qinghua Jiang. 2019. "LncRNA2Target v2.0: A Comprehensive Database for Target Genes of lncRNAs in Human and Mouse." Nucleic Acids Research 47 (Database issue): D140.
9. Derrien, Thomas, Rory Johnson, Giovanni Bussotti, Andrea Tanzer, Sarah Djebali, Hagen Tilgner, Gregory Guernec, et al. 2012. "The GENCODE v7 Catalog of Human Long Noncoding RNAs: Analysis of Their Gene Structure, Evolution, and Expression." Genome Research 22 (9): 1775–89.
10. Dixit, A., O. Parnas, B. Li, J. Chen, C. P. Fulco, L. Jerby-Arnon, N. D. Marjanovic, et al. 2016. "Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens." Cell 167 (7). https://doi.org/10.1016/j.cell.2016.11.038.
11. Doolittle, W. Ford. 2018. "We Simply Cannot Go on Being so Vague about 'function.'" Genome Biology 19 (1): 1–3.
12. Durand, Neva C., Muhammad S. Shamim, Ido Machol, Suhas S. P. Rao, Miriam H. Huntley, Eric S. Lander, and Erez Lieberman Aiden. 2016. "Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments." Cell Systems 3 (1): 95–98.
13. Engreitz, Jesse M., Jenna E. Haines, Elizabeth M. Perez, Glen Munson, Jenny Chen, Michael Kane, Patrick E. McDonel, Mitchell Guttman, and Eric S. Lander. 2016. "Local

Regulation of Gene Expression by lncRNA Promoters, Transcription and Splicing." Nature 539 (7629): 452–55.
14. Engreitz, Jesse M., Amy Pandya-Jones, Patrick McDonel, Alexander Shishkin, Klara Sirokman, Christine Surka, Sabah Kadri, et al. 2013. "The Xist lncRNA Exploits Three-Dimensional Genome Architecture to Spread across the X Chromosome." Science 341 (6147): 1237973.
15. Evangelista, John Erol, Daniel J. B. Clarke, Zhuorui Xie, Alexander Lachmann, Minji Jeon, Kerwin Chen, Kathleen M. Jagodnik, et al. 2022. "SigCom LINCS: Data and Metadata Search Engine for a Million Gene Expression Signatures." Nucleic Acids Research 50 (W1): W697–709.
16. Fanucchi, Stephanie, Ezio T. Fok, Emiliano Dalla, Youtaro Shibayama, Kathleen Börner, Erin Y. Chang, Stoyan Stoychev, et al. 2019. "Immune Genes Are Primed for Robust Transcription by Proximal Long Noncoding RNAs Located in Nuclear Compartments." Nature Genetics 51 (1): 138–50.
17. Forné, T., J. Oswald, W. Dean, J. R. Saam, B. Bailleul, L. Dandolo, S. M. Tilghman, J. Walter, and W. Reik. 1997. "Loss of the Maternal H19 Gene Induces Changes in Igf2 Methylation in Both Cis and Trans." Proceedings of the National Academy of Sciences of the United States of America 94 (19): 10243–48.
18. Frankish, Adam, Mark Diekhans, Irwin Jungreis, Julien Lagarde, Jane E. Loveland, Jonathan M. Mudge, Cristina Sisu, et al. 2020. "GENCODE 2021." Nucleic Acids Research 49 (D1): D916–23.
19. Gil, Noa, and Igor Ulitsky. 2018. "Production of Spliced Long Noncoding RNAs Specifies Regions with Increased Enhancer Activity." Cell Systems 7 (5): 537–47.e3.
20. ———. 2020. "Regulation of Gene Expression by Cis-Acting Long Non-Coding RNAs." Nature Reviews. Genetics 21 (2): 102–17.
21. Goyal, Ashish, Evgenij Fiškin, Tony Gutschner, Maria Polycarpou-Schwarz, Matthias Groß, Julia Neugebauer, Minakshi Gandhi, Maiwen Caudron-Herger, Vladimir Benes, and Sven Diederichs. 2017. "A Cautionary Tale of Sense-Antisense Gene Pairs: Independent Regulation despite Inverse Correlation of Expression." Nucleic Acids Research 45 (21): 12496–508.
22. Graur, Dan, Yichen Zheng, Nicholas Price, Ricardo B. R. Azevedo, Rebecca A. Zufall, and Eran Elhaik. 2013. "On the Immortality of Television Sets: 'Function' in the Human Genome According to the Evolution-Free Gospel of ENCODE." Genome Biology and Evolution 5 (3): 578–90.
23. Haerty, W., and C. P. Ponting. 2015. "Unexpected Selection to Retain High GC Content and Splicing Enhancers within Exons of Multiexonic lncRNA Loci." RNA 21 (3). https://doi.org/10.1261/rna.047324.114.
24. Henninger, J. E., O. Oksuz, K. Shrinivas, I. Sagi, G. LeRoy, M. M. Zheng, J. O. Andrews, et al. 2021. "RNA-Mediated Feedback Control of Transcriptional Condensates." Cell 184 (1). https://doi.org/10.1016/j.cell.2020.11.030.
25. Hnisz, Denes, Krishna Shrinivas, Richard A. Young, Arup K. Chakraborty, and Phillip A. Sharp. 2017. "A Phase Separation Model for Transcriptional Control." Cell 169 (1): 13–23.
26. Hoon, Dave S. B., Negin Rahimzadeh, and Matias A. Bustos. 2021. "EpiMap: Fine-Tuning Integrative Epigenomics Maps to Understand Complex Human Regulatory Genomic Circuitry." Signal Transduction and Targeted Therapy 6 (1): 1–3.

27. Hu, Z., C. Yang, S. Guo, Y. Li, and Y. Li. 2022. "LINC01615 Activates ZEB2 through Competitively Binding with miR-3653-3p to Promote the Carcinogenesis of Colon Cancer Cells." Cell Cycle 21 (3). https://doi.org/10.1080/15384101.2021.2015670.
28. Johnson, K. C., and D. R. Corey. 2023. "RNAi in Cell Nuclei: Potential for a New Layer of Biological Regulation and a New Strategy for Therapeutic Discovery." RNA 29 (4). https://doi.org/10.1261/rna.079500.122.
29. Khan, A., and X. Zhang. 2016. "dbSUPER: A Database of Super-Enhancers in Mouse and Human Genome." Nucleic Acids Research 44 (D1). https://doi.org/10.1093/nar/gkv1002.
30. Lai, F., S. S. Damle, K. K. Ling, and F. Rigo. 2020. "Directed RNase H Cleavage of Nascent Transcripts Causes Transcription Termination." Molecular Cell 77 (5). https://doi.org/10.1016/j.molcel.2019.12.029.
31. Lee, B., A. Sahoo, J. Marchica, E. Holzhauser, X. Chen, J. L. Li, T. Seki, et al. 2017. "The Long Noncoding RNA SPRIGHTLY Acts as an Intranuclear Organizing Hub for Pre-mRNA Molecules." Science Advances 3 (5). https://doi.org/10.1126/sciadv.1602505.
32. Leucci, E., R. Vendramin, M. Spinazzi, P. Laurette, M. Fiers, J. Wouters, E. Radaelli, et al. 2016. "Melanoma Addiction to the Long Non-Coding RNA SAMMSON." Nature 531 (7595). https://doi.org/10.1038/nature17161.
33. Luo, Sai, J. Yuyang Lu, Lichao Liu, Yafei Yin, Chunyan Chen, Xue Han, Bohou Wu, et al. 2016. "Divergent lncRNAs Regulate Gene Expression and Lineage Differentiation in Pluripotent Cells." Cell Stem Cell 18 (5): 637–52.
34. Maass, Philipp G., Andreas Rump, Herbert Schulz, Sigmar Stricker, Lisanne Schulze, Konrad Platzer, Atakan Aydin, et al. 2012. "A Misplaced lncRNA Causes Brachydactyly in Humans." The Journal of Clinical Investigation 122 (11): 3990–4002.
35. Marasco, L. E., G. Dujardin, R. Sousa-Luís, Y. H. Liu, J. N. Stigliano, T. Nomakuchi, N. J. Proudfoot, A. R. Krainer, and A. R. Kornblihtt. 2022. "Counteracting Chromatin Effects of a Splicing-Correcting Antisense Oligonucleotide Improves Its Therapeutic Efficacy in Spinal Muscular Atrophy." Cell 185 (12). https://doi.org/10.1016/j.cell.2022.04.031.
36. Marques, Ana C., Jim Hughes, Bryony Graham, Monika S. Kowalczyk, Doug R. Higgs, and Chris P. Ponting. 2013. "Chromatin Signatures at Transcriptional Start Sites Separate Two Equally Populated yet Distinct Classes of Intergenic Long Noncoding RNAs." Genome Biology 14 (11): R131.
37. Mas-Ponte, David, Joana Carlevaro-Fita, Emilio Palumbo, Toni Hermoso Pulido, Roderic Guigo, and Rory Johnson. 2017. "LncATLAS Database for Subcellular Localization of Long Noncoding RNAs." RNA 23 (7): 1080.
38. Necsulea, A., M. Soumillon, M. Warnefors, A. Liechti, T. Daish, U. Zeller, J. C. Baker, F. Grützner, and H. Kaessmann. 2014. "The Evolution of lncRNA Repertoires and Expression Patterns in Tetrapods." Nature 505 (7485). https://doi.org/10.1038/nature12943.
39. Ørom, U. A., T. Derrien, M. Beringer, K. Gumireddy, A. Gardini, G. Bussotti, F. Lai, et al. 2010. "Long Noncoding RNAs with Enhancer-like Function in Human Cells." Cell 143 (1). https://doi.org/10.1016/j.cell.2010.09.001.
40. Ørom, Ulf Andersson, and Ramin Shiekhattar. 2011. "Long Non-Coding RNAs and Enhancers." Current Opinion in Genetics & Development 21 (2): 194–98.
41. Palazzo, Alexander F., and Eugene V. Koonin. 2020. "Functional Long Non-Coding RNAs Evolve from Junk Transcripts." Cell 183 (5): 1151–61.

42. Ramilowski, J. A., C. W. Yip, S. Agrawal, J. C. Chang, Y. Ciani, I. V. Kulakovskiy, M. Mendez, et al. 2020. "Functional Annotation of Human Long Noncoding RNAs via Molecular Phenotyping." Genome Research 30 (7). https://doi.org/10.1101/gr.254219.119.
43. Rom, Aviv, Liliya Melamed, Noa Gil, Micah Jonathan Goldrich, Rotem Kadir, Matan Golan, Inbal Biton, Rotem Ben-Tov Perry, and Igor Ulitsky. 2019. "Regulation of CHD2 Expression by the Chaserr Long Noncoding RNA Gene Is Essential for Viability." Nature Communications 10 (1): 5092.
44. Scharner, Juergen, Wai Kit Ma, Qian Zhang, Kuan-Ting Lin, Frank Rigo, C. Frank Bennett, and Adrian R. Krainer. 2019. "Hybridization-Mediated off-Target Effects of Splice-Switching Antisense Oligonucleotides." Nucleic Acids Research 48 (2): 802–16.
45. Stalder, L., W. Heusermann, L. Sokol, D. Trojer, J. Wirz, J. Hean, A. Fritzsche, et al. 2013. "The Rough Endoplasmatic Reticulum Is a Central Nucleation Site of siRNA-Mediated RNA Silencing." The EMBO Journal 32 (8). https://doi.org/10.1038/emboj.2013.52.
46. Tan, Jennifer Y., Adriano Biasini, Robert S. Young, and Ana C. Marques. 2020. "Splicing of Enhancer-Associated lincRNAs Contributes to Enhancer Activity." Life Science Alliance 3 (4). https://doi.org/10.26508/lsa.202000663.
47. Tan, Jennifer Yihong, Adam Alexander Thil Smith, Maria Ferreira da Silva, Cyril Matthey-Doret, Rico Rueedi, Reyhan Sönmez, David Ding, Zoltán Kutalik, Sven Bergmann, and Ana Claudia Marques. 2017. "Cis-Acting Complex-Trait-Associated lincRNA Expression Correlates with Modulation of Chromosomal Architecture." Cell Reports 18 (9): 2280–88.
48. Thorvaldsen, J. L., K. L. Duran, and M. S. Bartolomei. 1998. "Deletion of the H19 Differentially Methylated Domain Results in Loss of Imprinted Expression of H19 and Igf2." Genes & Development 12 (23): 3693–3702.
49. Vance, K. W., S. N. Sansom, S. Lee, V. Chalei, L. Kong, S. E. Cooper, P. L. Oliver, and C. P. Ponting. 2014. "The Long Non-Coding RNA Paupar Regulates the Expression of Both Local and Distal Genes." The EMBO Journal 33 (4). https://doi.org/10.1002/embj.201386225.
50. Vancura, Adrienne, Alejandro H. Gutierrez, Thorben Hennig, Carlos Pulido-Quetglas, Frank J. Slack, Rory Johnson, and Simon Haefliger. 2022. "Is Evolutionary Conservation a Useful Predictor for Cancer Long Noncoding RNAs? Insights from the Cancer LncRNA Census 3." Non-Coding RNA 8 (6). https://doi.org/10.3390/ncrna8060082.
51. Wang, Yanli, Fan Song, Bo Zhang, Lijun Zhang, Jie Xu, Da Kuang, Daofeng Li, et al. 2018. "The 3D Genome Browser: A Web-Based Browser for Visualizing 3D Genome Organization and Long-Range Chromatin Interactions." Genome Biology 19 (1): 151.
52. Wu, Man, Liang-Zhong Yang, and Ling-Ling Chen. 2021. "Long Noncoding RNA and Protein Abundance in lncRNPs." RNA 27 (12): 1427–40.
53. Xiang, Jian-Feng, Qing-Fei Yin, Tian Chen, Yang Zhang, Xiao-Ou Zhang, Zheng Wu, Shaofeng Zhang, et al. 2014. "Human Colorectal Cancer-Specific CCAT1-L lncRNA Regulates Long-Range Chromatin Interactions at the MYC Locus." Cell Research 24 (5): 513–31.
54. Yip, C. W., C. C. Hon, K. Yasuzawa, D. M. Sivaraman, J. A. Ramilowski, Y. Shibayama, S. Agrawal, et al. 2022. "Antisense-Oligonucleotide-Mediated Perturbation of Long Non-Coding RNA Reveals Functional Features in Stem Cells and across Cell Types." Cell Reports 41 (13). https://doi.org/10.1016/j.celrep.2022.111893.

55. Zacher, Benedikt, Margaux Michel, Björn Schwalb, Patrick Cramer, Achim Tresch, and Julien Gagneur. 2017. "Accurate Promoter and Enhancer Identification in 127 ENCODE and Roadmap Epigenomics Cell Types and Tissues by GenoSTAN." PloS One 12 (1). https://doi.org/10.1371/journal.pone.0169249.
56. Zeng, Y., and B. R. Cullen. 2002. "RNA Interference in Human Cells Is Restricted to the Cytoplasm." RNA 8 (7). https://doi.org/10.1017/s1355838202020071.

TABLES AND FIGURES

Figure 1

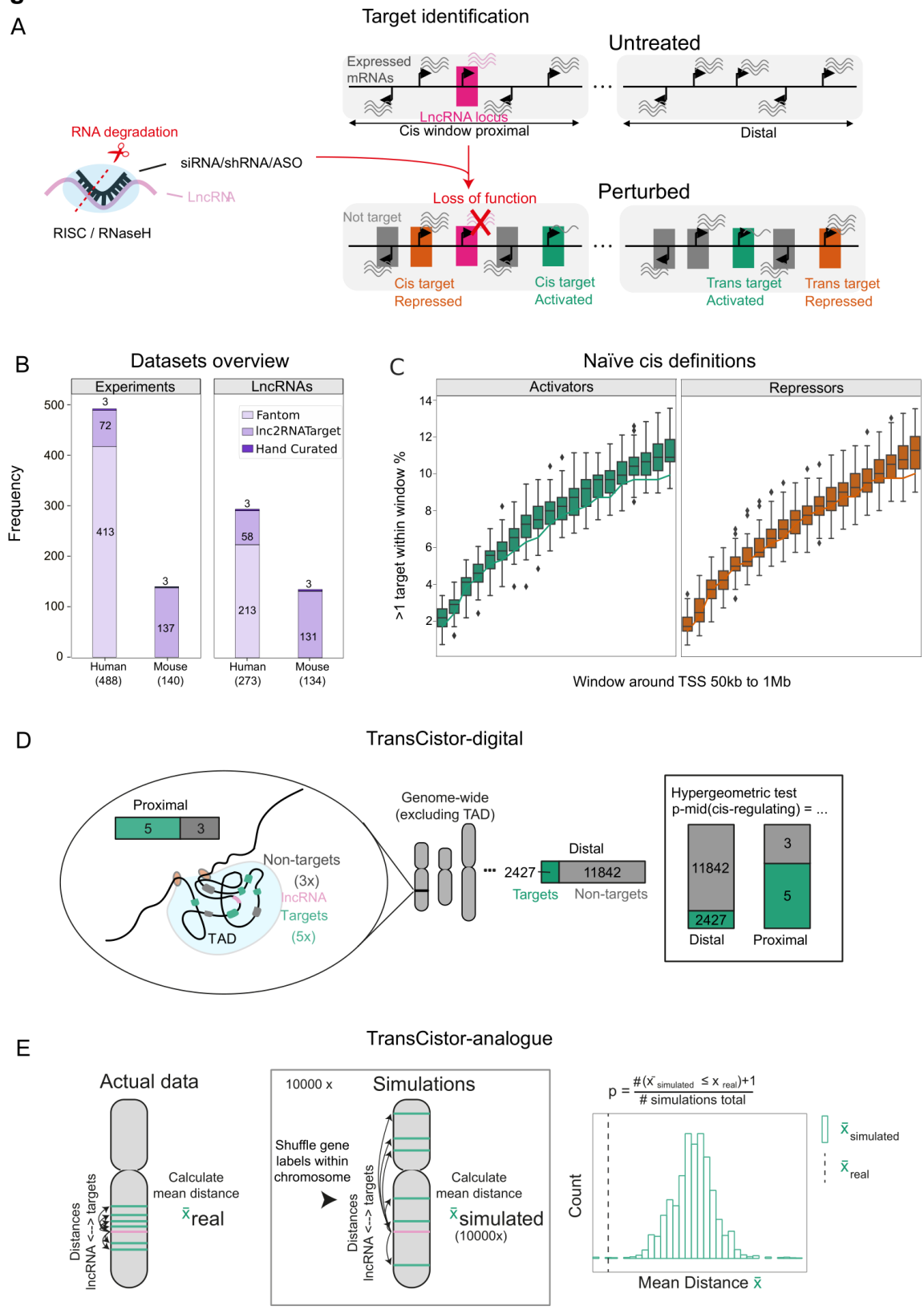


Figure 1: TransCistor is a quantitative framework for classifying cis- and trans-regulatory lncRNAs.

A) 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. B) The perturbation datasets used here: Data were mainly obtained from two sources, FANTOM[47] and Lnc2Target[48] (x-axis). The y-axis displays the number of individual experiments (left panel) or individual lncRNA genes (right panel) (note that the difference arises from the fact that many lncRNA genes are represented by >1 experiment). Some lncRNAs are present in both datasets. 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 p-mid adjusted 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). 10000 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

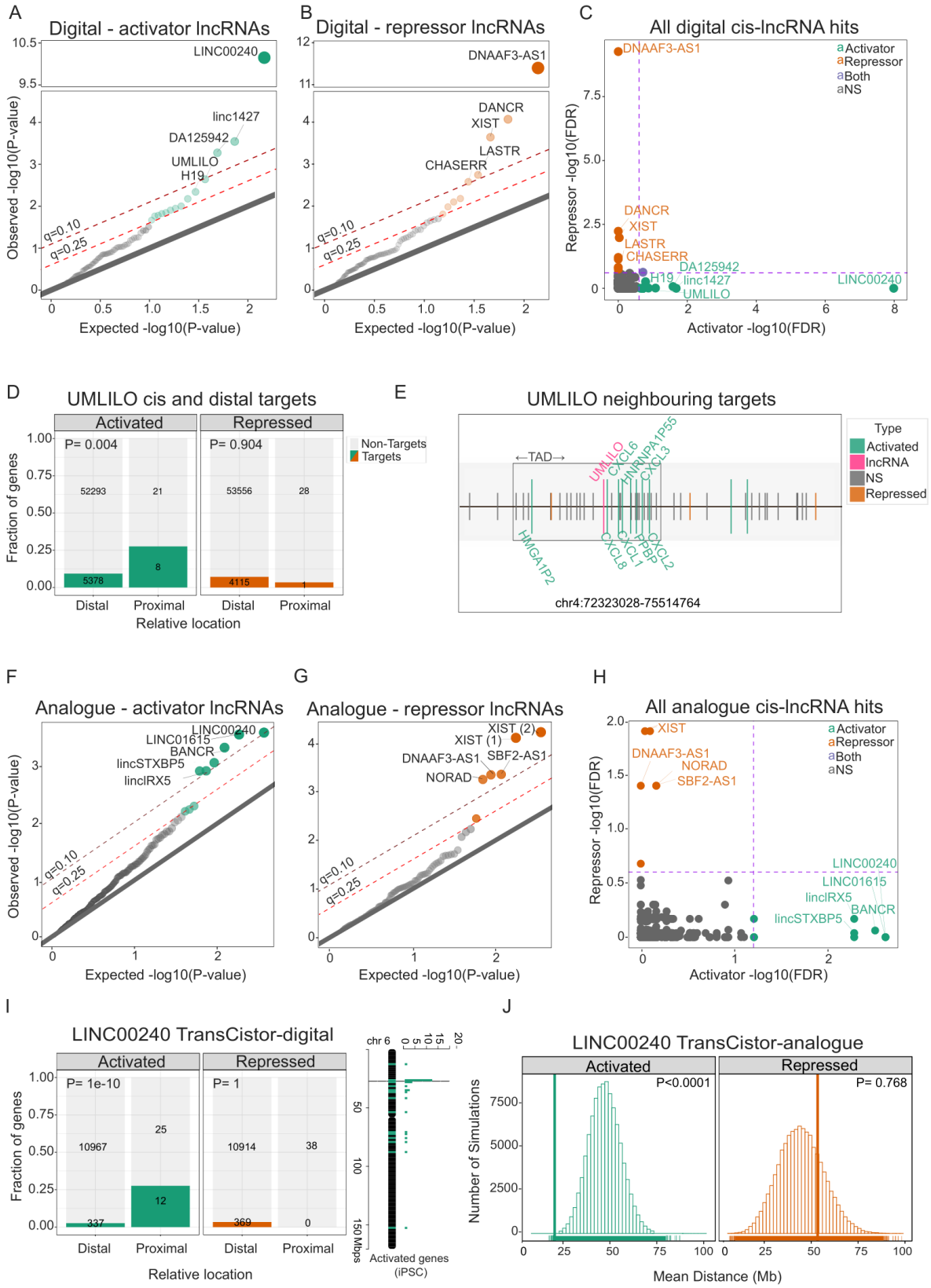

A
Digital - activator lncRNAs
LINC00240
linc1427
DA125942
UMLILO
H19
q=0.10
q=0.25
Observed -log10(P-value)
Expected -log10(P-value)
B
Digital - repressor lncRNAs
DNAAF3-AS1
DANCR
XIST
LASTR
CHASERR
C
All digital cis-lncRNA hits
Activator
Repressor
Both
NS
Repressor -log10(FDR)
Activator -log10(FDR)
D
UMLILO cis and distal targets
Activated
Repressed
P= 0.004
P= 0.904
Non-Targets
Targets
Fraction of genes
Distal
Proximal
Relative location
E
UMLILO neighbouring targets
←TAD→
UMLILO
CXCL6
HNRNPA1P55
CXCL3
HMGA1P2
CXCL8
CXCL1
PPBP
CXCL2
chr4:72323028-75514764
Type
Activated
lncRNA
NS
Repressed
F
Analogue - activator lncRNAs
LINC00240
LINC01615
BANCR
lincSTXBP5
lincIRX5
G
Analogue - repressor lncRNAs
XIST (2)
XIST (1)
DNAAF3-AS1
SBF2-AS1
NORAD
H
All analogue cis-lncRNA hits
XIST
DNAAF3-AS1
NORAD
SBF2-AS1
LINC00240
LINC01615
lincIRX5
BANCR
lincSTXBP5
I
LINC00240 TransCistor-digital
P= 1e-10
P= 1
chr 6
Activated genes (iPSC)
J
LINC00240 TransCistor-analogue
P<0.0001
P= 0.768
Number of Simulations
Mean Distance (Mb)

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. Light and dark red dotted lines indicate an FDR cutoff of 0.25 and 0.10 respectively. 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. Black box: Topologically associated domain F) As for (A), for TransCistor-analogue and activated targets. G) As for (B), for TransCistor-analogue and repressed targets. Two experiments supporting the XIST lncRNA appear in the top significant classifications represented by (1) & (2) H) As for (C), for TransCistor-analogue. I) and J) LINC00240, an example cis-activator identified by both TransCistor-diital and TransCistor-analogue. I) as for D). J) 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.

Figure 3

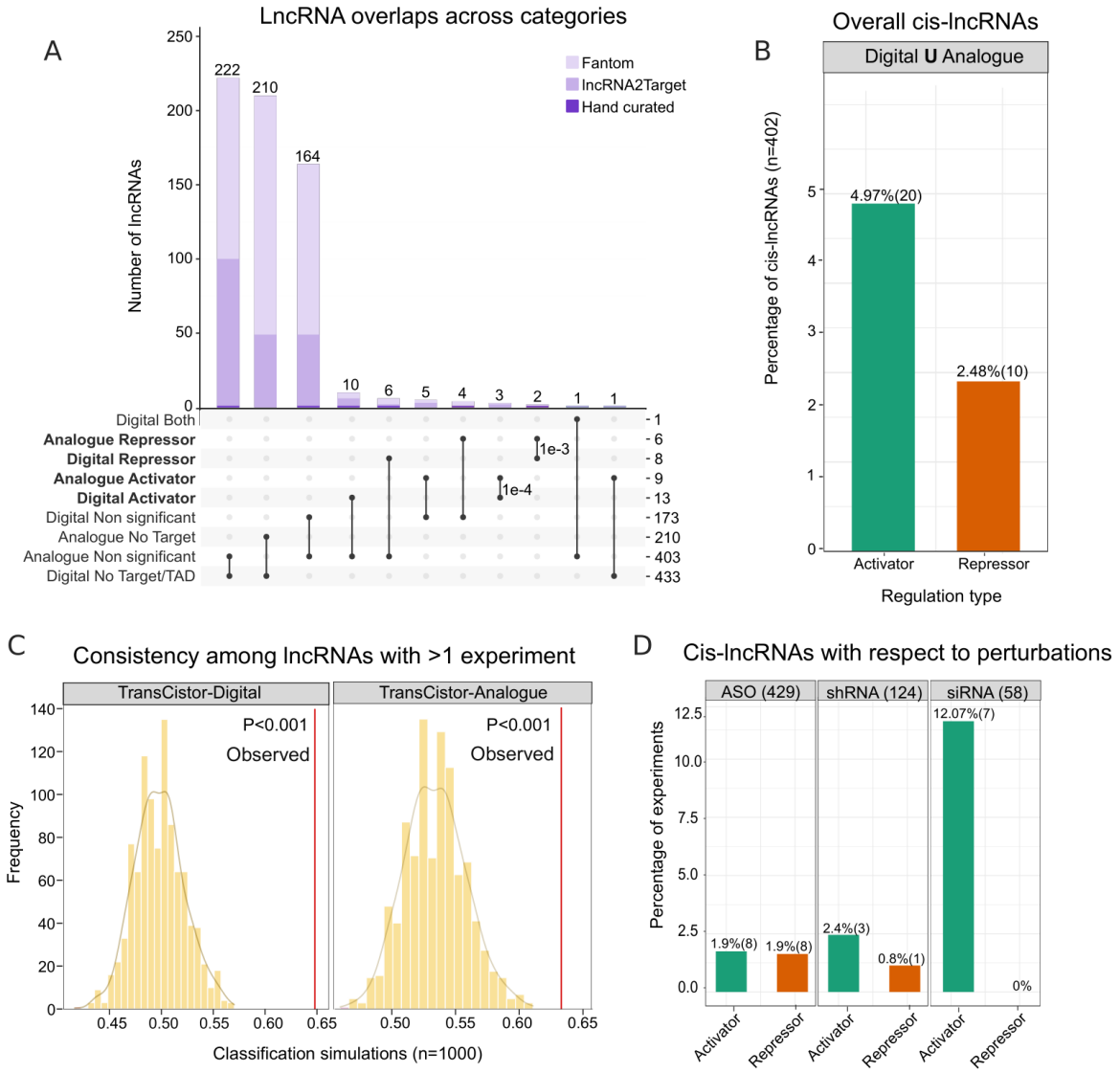


Figure 3: Rate and consistency of cis-lncRNA inference.

A) Summary of TransCistor results across datasets. Significant overlaps are noted with the respective p-value. Reported p-values are based on a hypergeometric distribution. B) The rate of lncRNA genes defined to be cis-regulatory based on our analysis (union of digital and analogue). Note that one single experiment is sufficient to label a lncRNA gene as cis-regulatory. C) The concordance score is calculated for TransCistor-Digital and Analogue, based on the consistency of classification for lncRNAs with multiple supporting experiments (>1). To assess the significance, the score is recalculated for 1000 simulations with shuffled classification labels. The real concordance score, indicated by the red line, exceeds the distribution obtained from the simulated data. D) The rate of experiments defined as cis-regulatory, broken down by perturbation method.

Figure 4

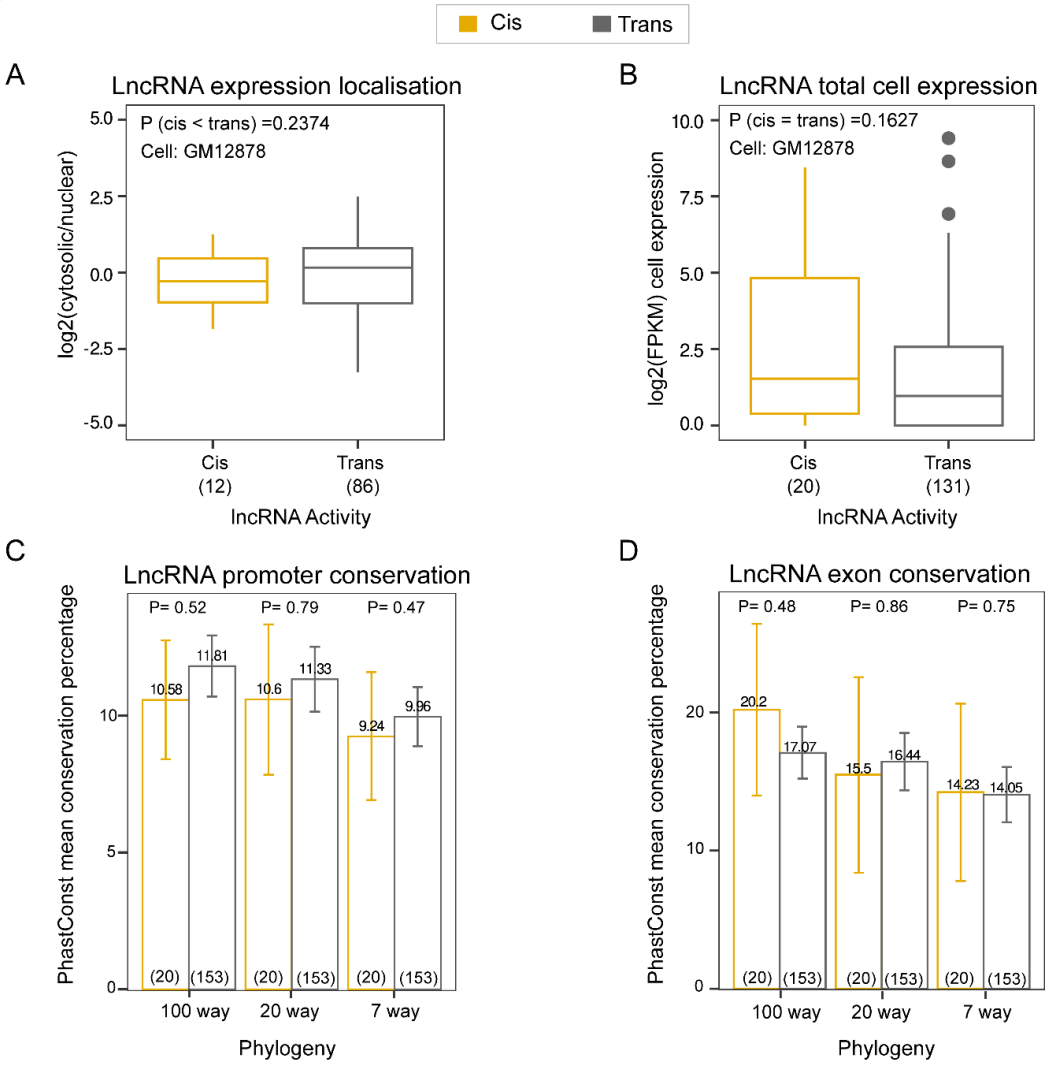


Figure 4: Localisation, expression and conservation comparison among cis- and trans-lncRNAs.

A) The distribution of the (log2-scaled) ratio of cytosolic to nuclear expression levels for cis- and trans-lncRNAs in GM12878 cells. Number of genes in each group are displayed below, and represent the subset of lncRNA genes for which localisation data was available. Reported p-values for significance of between group differences are based on one-tailed Wilcoxon-rank-sum test. B) As for A) exploring the difference in total levels of cell expression. Reported p-values for significance are based on two-tailed Wilcoxon-rank-sum test. C) D) Barplots display the mean evolutionary conservation for the indicated features (x-axis). Number of genes in each group are displayed in brackets, and represent the subset of lncRNA genes for which PhastCons score was available for its promoter and exonic regions. Error bars represent standard deviation. Evolutionary conservation is calculated using PhastCons scores, where promoter conservation represents the percentage of promoter nucleotides overlapped by PhastCons conserved elements, and exon conservation represents the percentage of merged exonic nucleotides overlapped by PhastCons conserved elements. The reported p-values indicate the significance difference between cis- and trans-lncRNAs for each feature, determined using a one-sided Welch test with the alternative hypothesis of 'greater' (cis > trans).

Figure 5

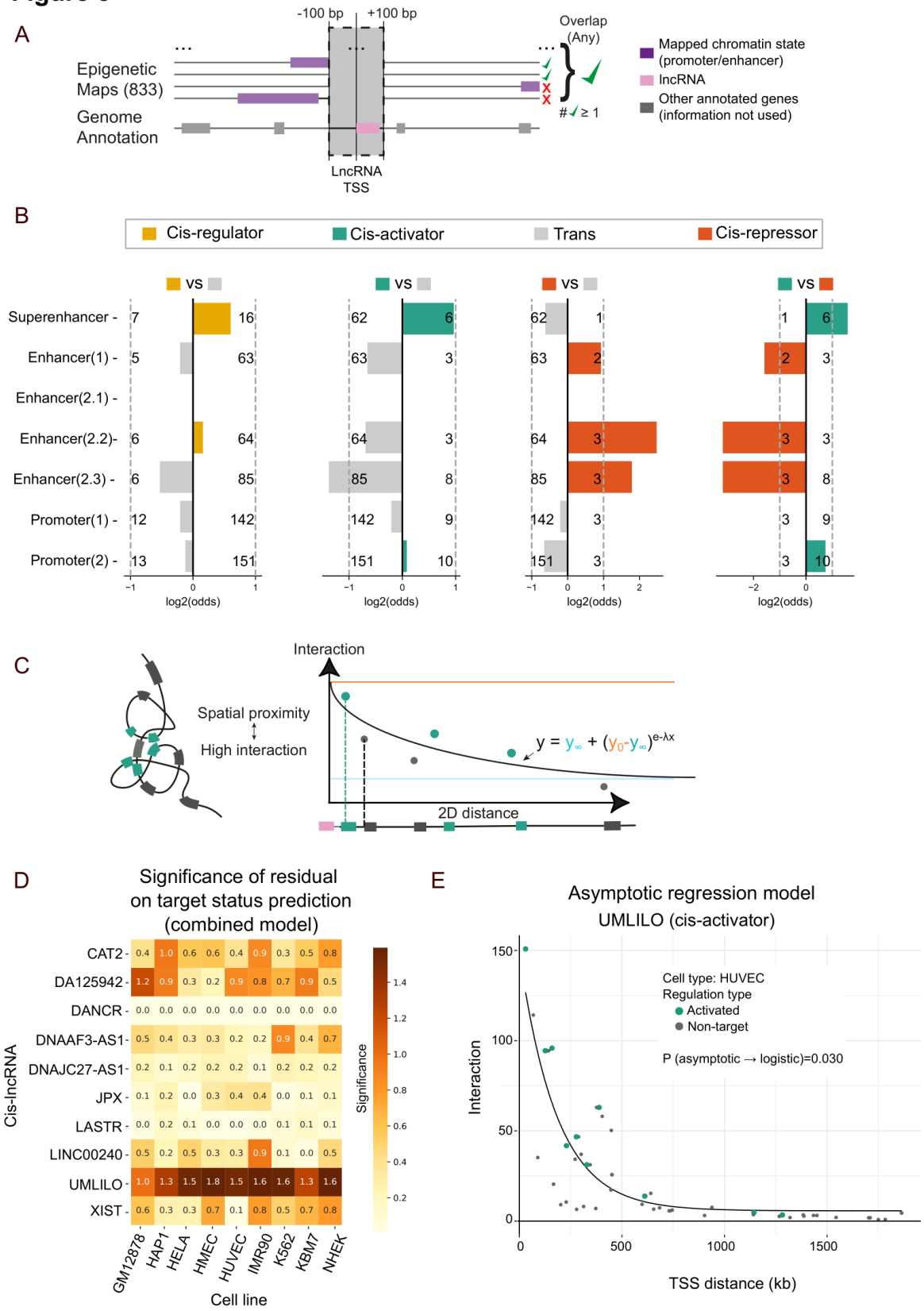
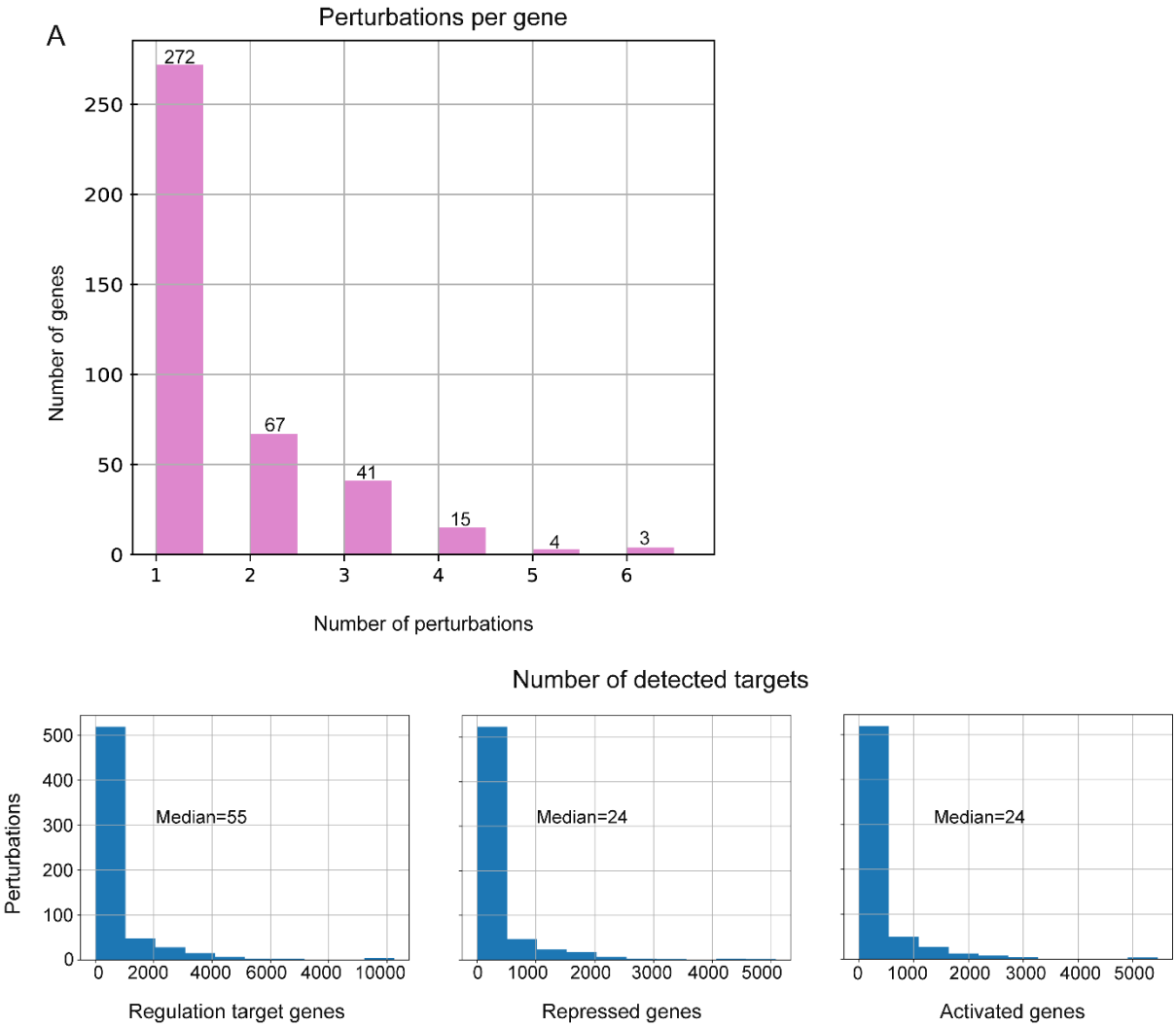

A
-100 bp
+100 bp
Overlap (Any)
Epigenetic Maps (833)
Genome Annotation
LncRNA TSS
≥ 1
Mapped chromatin state (promoter/enhancer)
lncRNA
Other annotated genes (information not used)
B
Cis-regulator
Cis-activator
Trans
Cis-repressor
vs
Superenhancer
Enhancer(1)
Enhancer(2.1)
Enhancer(2.2)
Enhancer(2.3)
Promoter(1)
Promoter(2)
log2(odds)
C
Spatial proximity
High interaction
Interaction
y = y∞ + (y0-y∞)e-λx
2D distance
D
Significance of residual on target status prediction (combined model)
Cis-lncRNA
CAT2
DA125942
DANCR
DNAAF3-AS1
DNAJC27-AS1
JPX
LASTR
LINC00240
UMLILO
XIST
GM12878
HAP1
HELA
HMEC
HUVEC
IMR90
K562
KBM7
NHEK
Cell line
Significance
E
Asymptotic regression model
UMLILO (cis-activator)
Cell type: HUVEC
Regulation type
Activated
Non-target
P (asymptotic → logistic)=0.030
Interaction
TSS distance (kb)

Figure 5: 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 lncRNA TSS (pink bar). Overlaps are considered while varying two thresholds for defining a lncRNA-enhancer overlap: minimum numbers of individual enhancer annotations (epigenome threshold) and window size. Only the TSS spans with overlaps in more samples than a given epigenome threshold are considered. B) Enrichment results for 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). C) A model for proximity-driven target selection: (Left panel) Chromatin folding brings lncRNA (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. Darker colors 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).

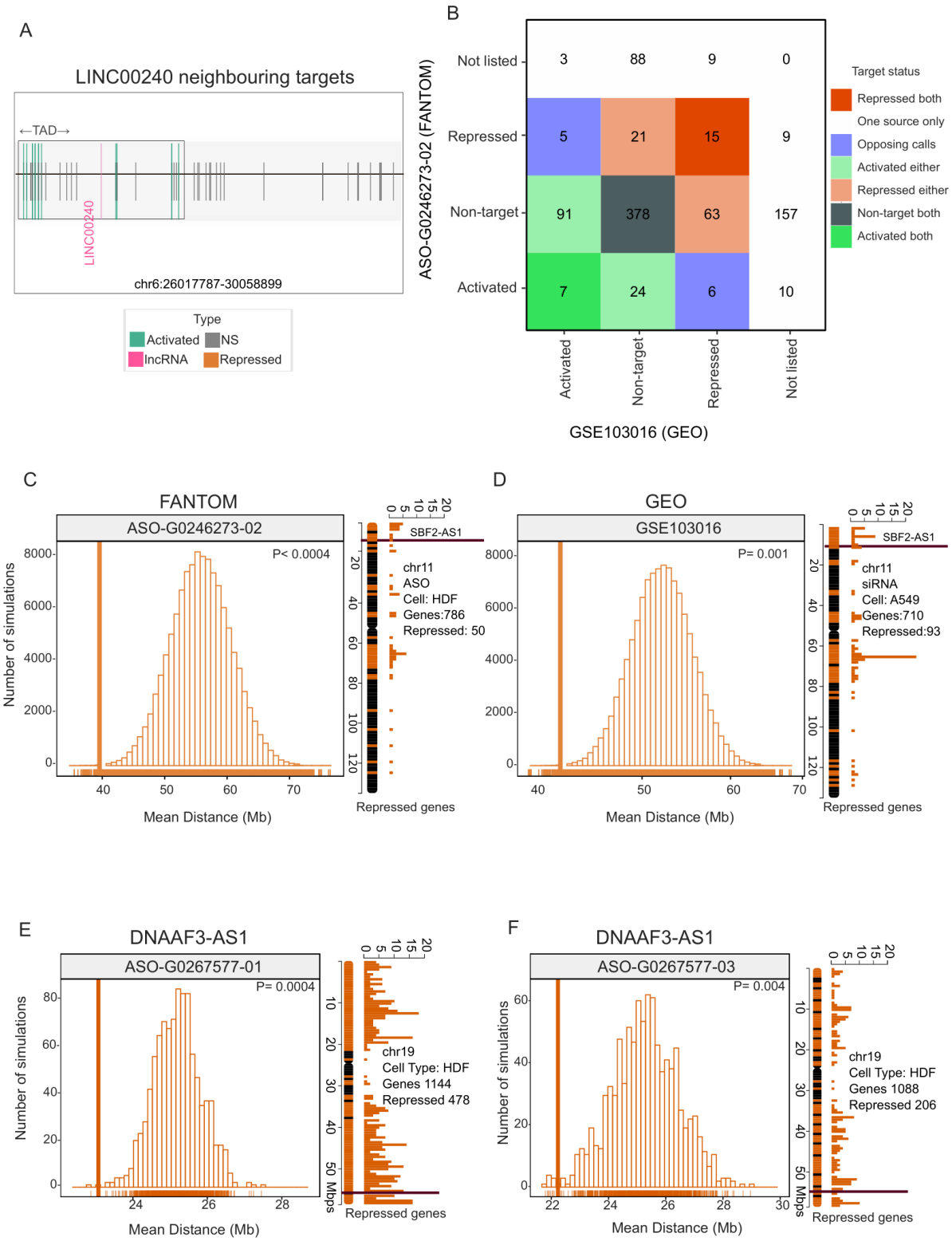
Supplementary Figure 1



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.

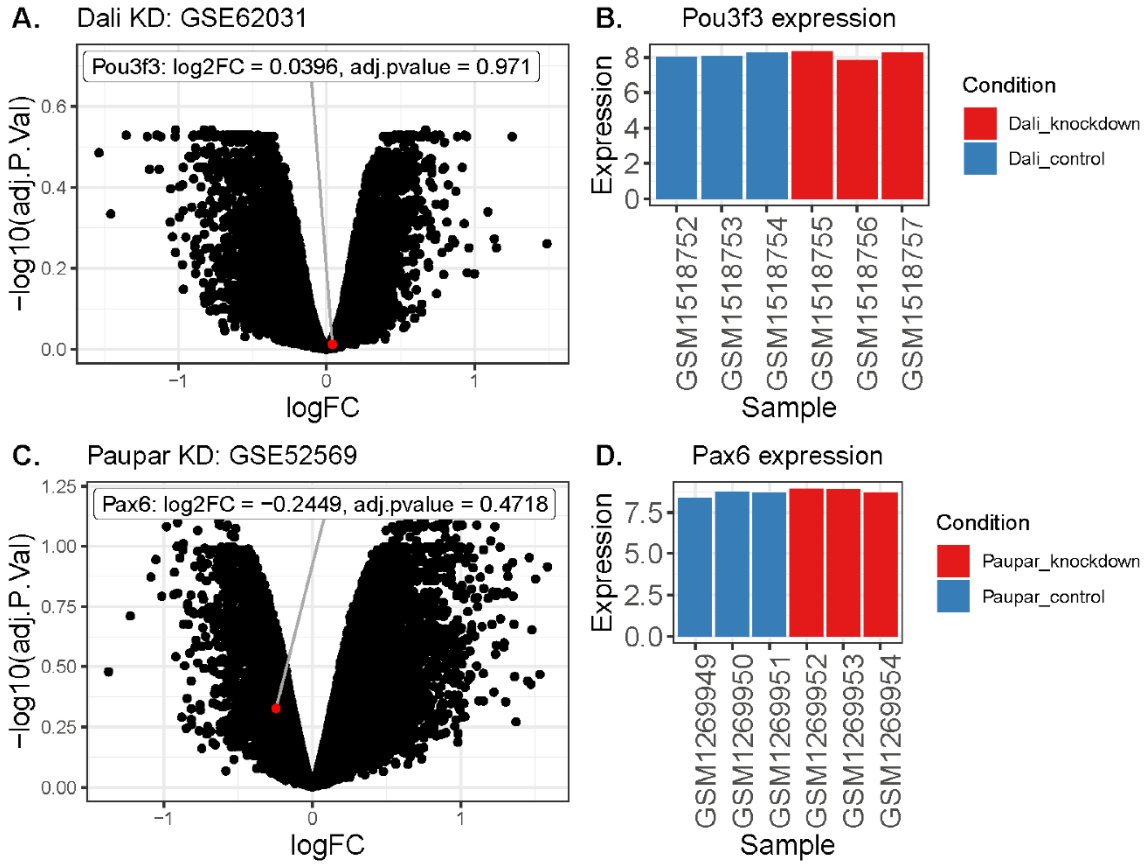
Supplementary Figure 2



Supplementary Figure S2: Exploring independent and supporting datasets for significant cis lncRNAs.

A) LINC00240 genomic locus: Vertical bars denote gene TSS. Grey: non-targets; green: activated targets; pink: LINC00240. Black box: Topologically associated domain. B) SBF2-AS1 target regulation overlap between FANTOM ASO knockdown and independent siRNA GEO dataset. Numbers indicate the genes in each category, classified by their regulation in the two distinct datasets. C) TransCistor-analogue results for FANTOM ASO knockdown targeting SBF2-AS1 in human dermal fibroblasts. D) As for (C), but for independent data from A549 cells treated with siRNA. E) DNAAF3-AS1, an example cis-repressor identified by TransCistor-analogue. Shown are 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. F) As for (J), for a second perturbation experiment.

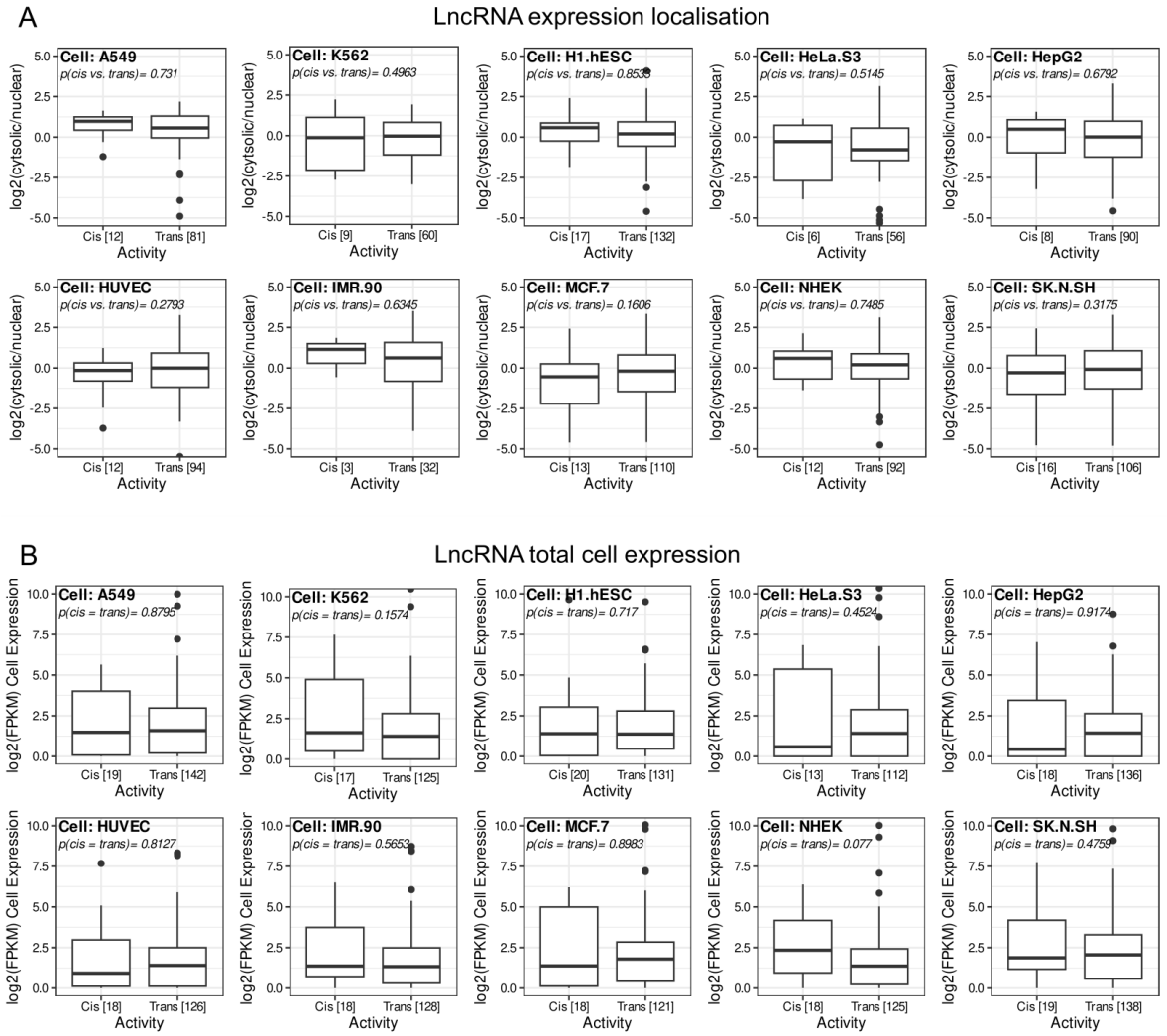
Supplementary Figure 3



Supplementary Figure S3: Analysis of Dali and Paupar target genes using public microarray data.

A) Global transcriptome changes upon Dali knockdown were obtained from Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE62031). B) Neighbour gene and putative target Pou3f3 mRNA expression in control and Dali knockdown samples. C) Global transcriptome changes upon Paupar knockdown were obtained from Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE52569). D) Neighbour gene and putative target Pax6 mRNA expression in control and Paupar knockdown samples.

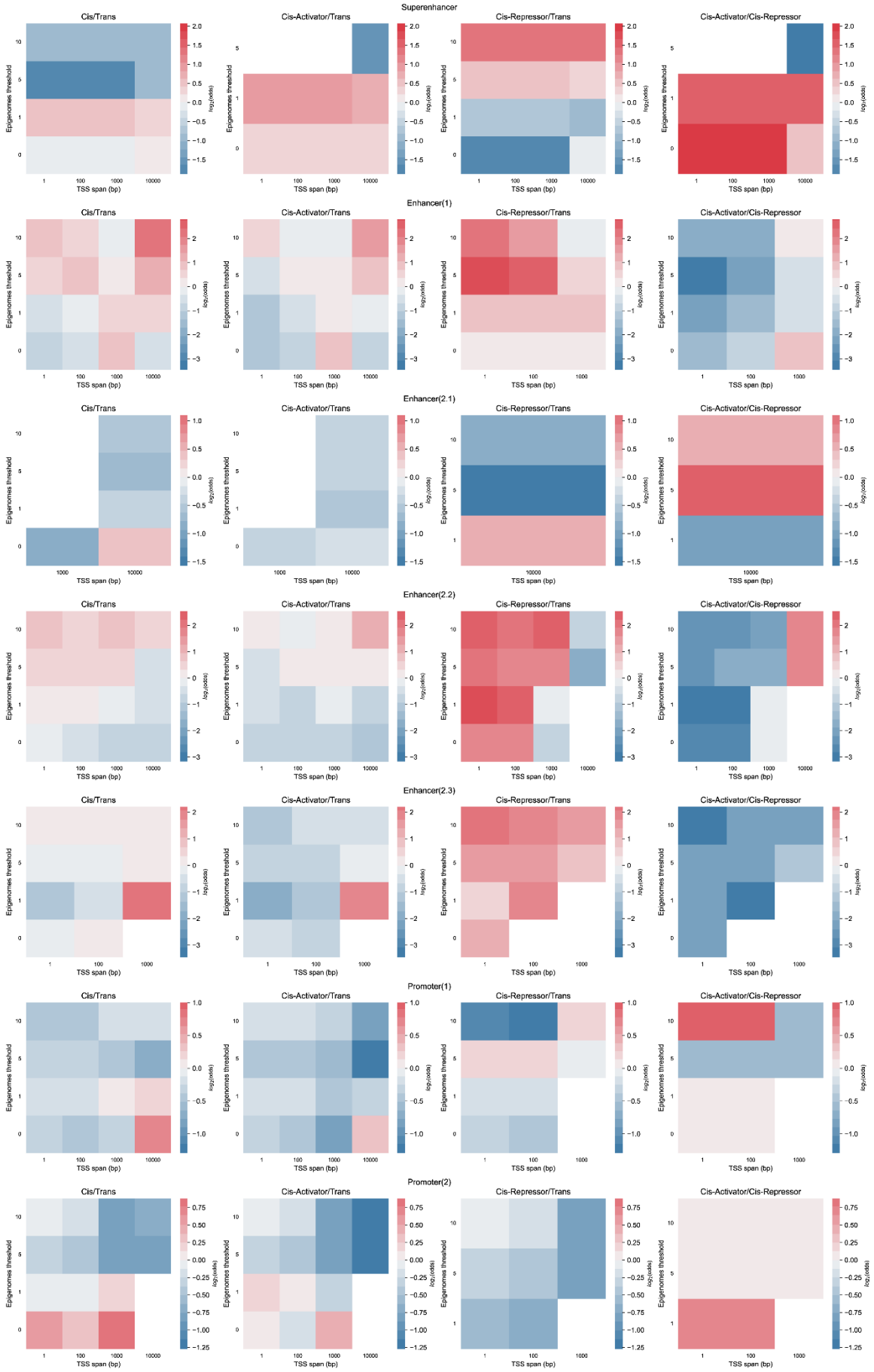
Supplementary Figure 4



Supplementary Figure S4: Cytosolic/nuclear localisation of cis/trans lncRNAs.

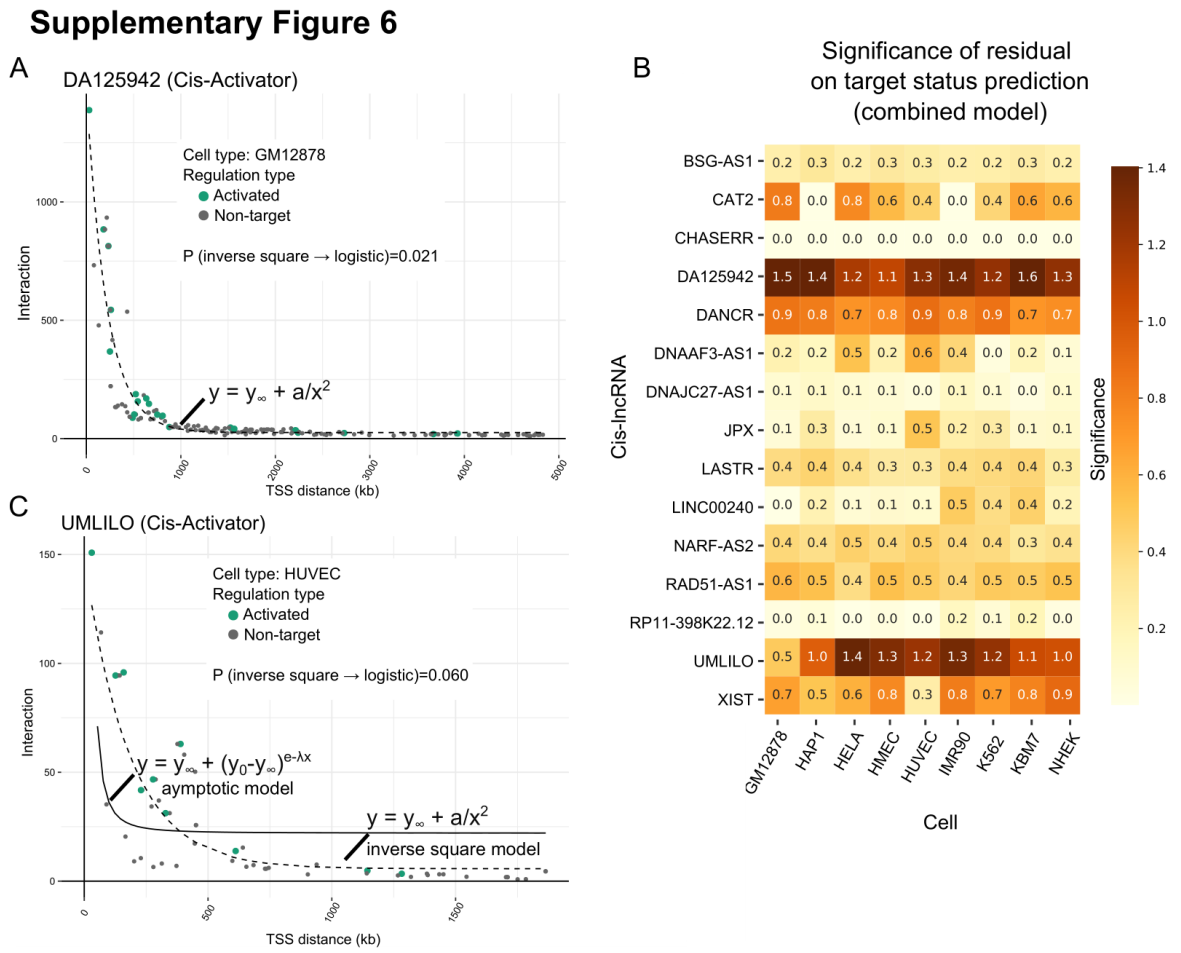
Boxplots showing the distribution of the (log2-scaled) ratio of cytosolic to nuclear expression levels for cis-/ vs. trans-regulatory lncRNAs in different cell types. The number of genes in each group is displayed below and represents the subset of lncRNA genes for which localisation data was available. Reported p-values for the significance of between-group differences are based on a one-tailed Wilcoxon-rank-sum test (alternative: cis > trans). B) As for A) exploring the difference in total levels of cell expression. Reported p-values for significance are based on a two-tailed Wilcoxon-rank-sum test.

Supplementary Figure 5



Supplementary Figure S5: Enrichment of enhancers, super-enhancers and promoters in *cis*-lncRNAs.

Each row represents a different enhancer annotation (see Methods). Columns represent comparisons between indicated pairs of lncRNA classes. Heatmaps display the enrichment of overlap at different genomic windows around the lncRNA 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 S6: Alternative inverse square model for target gene interaction.

A) and B) are equivalent to main Figure 5 panels E and D 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.