TransCistor reveals the landscape of *cis*-regulatory long noncoding RNAs

Authors

Panagiotis Chouvardas[123], Marc Zimmerli[#12], Daniel Hanhart[#23], Mario Moser[#12], Hugo Guillen-Ramirez[1245], Sanat Mishra[6], Roberta Esposito[12], Taisia Polidori[12], Maro Widmer[12], Raquel García-Pérez[7], Marianna Kruithof-de Julio[23], Dmitri Pervouchine[8], Marta Melé[7], 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 chromosome[1,2]. 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 lncRNAs[6], 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 basepairs[4] to hundreds of kilobases[7]) and number of targets (one[4] to many[8]). Two overarching molecular mechanisms have been proposed for *cis*-lncRNAs: enhancer elements and chromatin folding[9]. Some *cis*-activating lncRNAs, termed “enhancer lncRNAs” (e-lncRNAs), have been found to overlap DNA-encoded enhancer elements[10–12], similar to lncRNAs more generally[13]. 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 separation[14] and local concentration gradients[15]. 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[16]. 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*-lncRNAs. Until now, they have been defined by the existence of ≥1 proximal targets. Targets are defined as those whose expression changes in response to lncRNA LOF, as measured using single-gene (RT-PCR) or whole-transcriptome (RNA-seq, CAGE, microarray) techniques[3,5,17]. "Proximity" is defined on a case-by-case basis, using a wide range of windows spanning 10^2 to 10^5 bp[7]. 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*-lncRNAs from a quantitative perspective. We show that conventional definitions are prone to high false positive rates. We introduce statistical methods for definition of *cis*-lncRNAs at controlled false discovery rates, and use them to classify regulatory lncRNAs across hundreds of perturbation datasets. Our results enable us to estimate, for the first time, the prevalence of *cis*-acting lncRNAs and evaluate hypotheses regarding their molecular mechanisms of action.

Results

A quantitative definition of *cis*-lncRNAs

We began by investigating the usefulness of the present "naïve" definition of *cis*-lncRNAs, based on the existence of ≥1 target gene within a local genomic window. To create a dataset of lncRNA-target relationships, we collected 382 lncRNA LOF experiments targeting 188 human lncRNAs from a mixture of sources (including the recently published dataset of ASO knockdowns in human dermal fibroblasts from the FANTOM consortium[18]) and 137 experiments for 131 lncRNAs in mouse (Figure 1A). To this we added 6 hand-curated LOF experiments targeting previously reported *cis*-acting lncRNAs (*UMLILO*, *XIST x2*, *Chaserr*, *Paupar* and *Dali*). We employed a functional definition of "targets", as genes whose steady-state levels significantly change in response to a given lncRNA's LOF (Figure 1B). We further define targets as activated or repressed, where they decrease / increase in response to lncRNA 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 lncRNA's TSS, respectively. In the dataset, 126 lncRNAs 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 lncRNAs' TSSs, we evaluated the fraction of lncRNAs that would be defined as *cis*-lncRNAs under the naïve 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 protein-coding 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 conventional definition of *cis*-lncRNAs yields high rates of false-positive predictions.

To overcome this issue, we adopted the following definition for *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.

105 **TransCistor: digital and analogue identification of *cis*-lncRNAs**

106 The first method employs a simple definition of proximal genes, being those whose TSS falls
107 within a defined window centred on the lncRNA TSS. We developed a pipeline, *TransCistor-*
108 *digital*, which takes as input a processed whole-transcriptome list of target genes ("regulation
109 file"), and tests for statistical enrichment in proximal genes using the hypergeometric distribution
110 (Figure 1D). Although in principle any sized window may be used, we reasoned that the most
111 biologically-meaningful would be the local TAD, in line with previous studies[19]. Chromatin folding
112 varies to an extent between cell types[20]. Therefore, TransCistor-digital calculates enrichment
113 across a set of experimentally-defined cell-type-specific TADs (40 human, 3 mouse)[21] and
114 aggregates the resulting p-values by their harmonic mean.

115 The above fixed-window approach is intuitive yet has drawbacks. Several reported *cis*-lncRNAs
116 have individual targets that are not immediately adjacent[7], and might be overlooked by the fixed-
117 window approach. Furthermore, many lncRNAs may have no neighbouring genes in their local
118 TAD, or no identified local TAD. Therefore, we developed an alternative method that dispenses
119 with fixed windows, while still examining proximity biases in targets. This method, *TransCistor-*
120 *analogue*, defines a distance statistic as the mean TSS-to-TSS distance of all same-chromosome
121 targets of a given lncRNA (Figure 1E). To estimate statistical significance, a null distribution is
122 calculated by randomisation of target labels (Figure 1E). Now, *cis*-lncRNAs are defined as those
123 having a distance statistic that is lower than a majority of simulations.

124 We sought to test the performance of TransCistor-digital and evaluate the global landscape of
125 *cis*-lncRNAs. After filtering out unusable datasets (having no *cis*-targets or no overlapping TAD),
126 168 datasets remained. The majority of p-values produced by this analysis follow the null
127 distribution, underlining the conservative statistical behaviour of TransCistor (Figure 2A,B). We
128 discovered 19 *cis*-acting lncRNAs (12 activators, 7 repressors), with a relatively relaxed false
129 discovery rate (FDR) threshold of 0.5, while no *cis*-lncRNAs are simultaneously classified as
130 activator and repressor (Figure 2C). Amongst the top-ranked *cis*-lncRNAs is *UMLILO*, previously
131 described to activate multiple genes in its local genomic neighbourhood[8]. *UMLILO* exhibits a
132 significant enrichment of activated targets amongst proximal genes, which is not observed for
133 repressed targets (Figure 2D,E). Analysis of the entire perturbation dataset by TransCistor-
134 analogue identified 20 *cis*-lncRNAs (15 activators, 5 repressors, FDR≤0.5). Statistical behaviour
135 is good (Figure 2F,G), while *cis*-lncRNAs once again are cleanly split between activators and
136 repressors (Figure 2H).

137 The usefulness of these methods is supported by their internal and external consistency.
138 Together, the TransCistor approaches correctly identify previously-described *cis*-activators *H19*[22],
139 *JPX*[23]*, Evx1os*[24] and *DA125942*[25] amongst the top ranked *cis*-activators, while *XIST* is amongst
140 the top repressors[26] (Figure 2J). Both human and mouse orthologues of *CHASERR*
141 (ENSG00000272888) are identified as *cis-repressors*[4]. Of the hits, two are concordantly classified
142 by ≥1 independent perturbation experiments (*XIST* & *DNAAF3-AS1* classified as *cis*-repressors
143 based on two separate experiments each) (Figure 2I,J). We observed agreement between the
144 two TransCistor methods, with 6 *cis*-lncRNAs in common (*DA125942, linc1427, RAD51-AS1,*
145 *H19, Xist, DNAAF3-AS1*) (p-value < 0.05, hypergeometric test) (Figure 3A).

146 TransCistor predicted *cis*-regulatory activity for a number of known lncRNAs that have never been
147 described as such in prior literature. These include *BANCR* (*cis*-activator), and *SBF2-AS1*,

148 *LASTR, NORAD*, *DANCR* (*cis*-repressors). However, the latter two are only identified in one out
149 of multiple independent perturbation experiments (6/5 for *NORAD/DANCR* respectively). In the
150 case of *DANCR*, the *cis* definition arises from the repression of two same-strand small RNAs
151 (has-mir-4449, SNORA26). It is not yet clear if these results reflect false-positive or false-negative
152 predictions. To investigate this, we merged all hits across experiments and repeated the analysis,
153 but here we found no *cis* signal, suggesting that they are false-positive predictions. On the other
154 hand, analysis of an independent dataset for *SBF2-AS1* from different cells (A549 lung
155 adenocarcinoma) and perturbation (siRNA) yielded concordant *cis*-repressor prediction from
156 TransCistor-analogue (Supplementary Figure S2). This is strong evidence that *SBF2-AS1* is a
157 novel *cis*-repressive lncRNA.

158 Surprisingly, TransCistor failed to find evidence supporting two previously reported *cis*-lncRNAs,
159 *Paupar*[27] and *Dali*[28]. Inspection of the originating microarray data revealed that, for neither case,
160 do the claimed *cis*-target genes pass cutoffs of differential expression (Supplementary Figure S3).

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

167 **TransCistor identifies *cis*-lncRNA independently of perturbation technology**

168 The perturbation experiments contained a mixture of RNA interference (RNAi) and antisense
169 oligonucleotide (ASO) LOF perturbations. While early experiments were performed using the two
170 RNAi approaches of siRNA and shRNA, it is widely thought that these principally degrade targets
171 in the cytoplasm[29,30] or ribosome[31]. In contrast, ASOs are becoming the method of choice to knock
172 down lncRNAs, since they are thought to act on nascent RNA in chromatin[32]. If correct, then one
173 would expect ASO perturbations to have greater power to discover *cis*-lncRNAs. To test this, we
174 compared predictions from each perturbation technology (Figure 3C). Surprisingly, we observed
175 broadly similar rates of *cis*-lncRNA identification between perturbation methods. However, ASO
176 experiments discover similar rates of activators and repressors, while RNAi perturbations yield an
177 apparent excess of activators over repressors.

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

182 **Association of *cis*-lncRNAs with enhancer elements**

183 It has been widely speculated that *cis*-lncRNAs, particularly activators (ie e-lncRNAs), act in
184 concert with DNA enhancer elements to upregulate target gene expression[3,9,12]. Our catalogue of
185 *cis*-lncRNAs represents an opportunity to independently test this. We calculated the rate of
186 overlap of lncRNAs with enhancers using epigenomics data across human tissues (Figure 4A,
187 Supplementary Figure S4). Analyses were performed at a variety of epigenome thresholds (the
188 minimum number of samples required to define a given epigenomic state) and window sizes (the
189 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[33]. 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 [8,25]. 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*[4], to those regulating a more distal target amongst other non-target genes, such as *CCAT1-L*[7]. 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 future parallelised CRISPR LOF methods such as Perturb-Seq[34].

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[35]. Very recently, de Hoon and colleagues employed genome-wide RNA-chromatin and chromatin folding to train a predictive model for *cis*-regulatory lncRNAs[36]. While these methods are valuable, they infer target genes based on indirect correlates of *cis*-regulation, which may not reflect causation[37]. 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.

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 (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*-lncRNAs' 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 lncRNAs.

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 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[38], but also to affect local chromatin modifications[39], 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 lncRNA[40,41]. 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 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)[42] 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 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[21]. For each cell type, TransCistor identifies the lncRNA 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[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. 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)[43] 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 from the project Github repository, linked above.

Analysis of chromatin states

Chromatin states annotations were retrieved from three sources: EpiMap[44], genoSTAN[45], and dbSUPER[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.

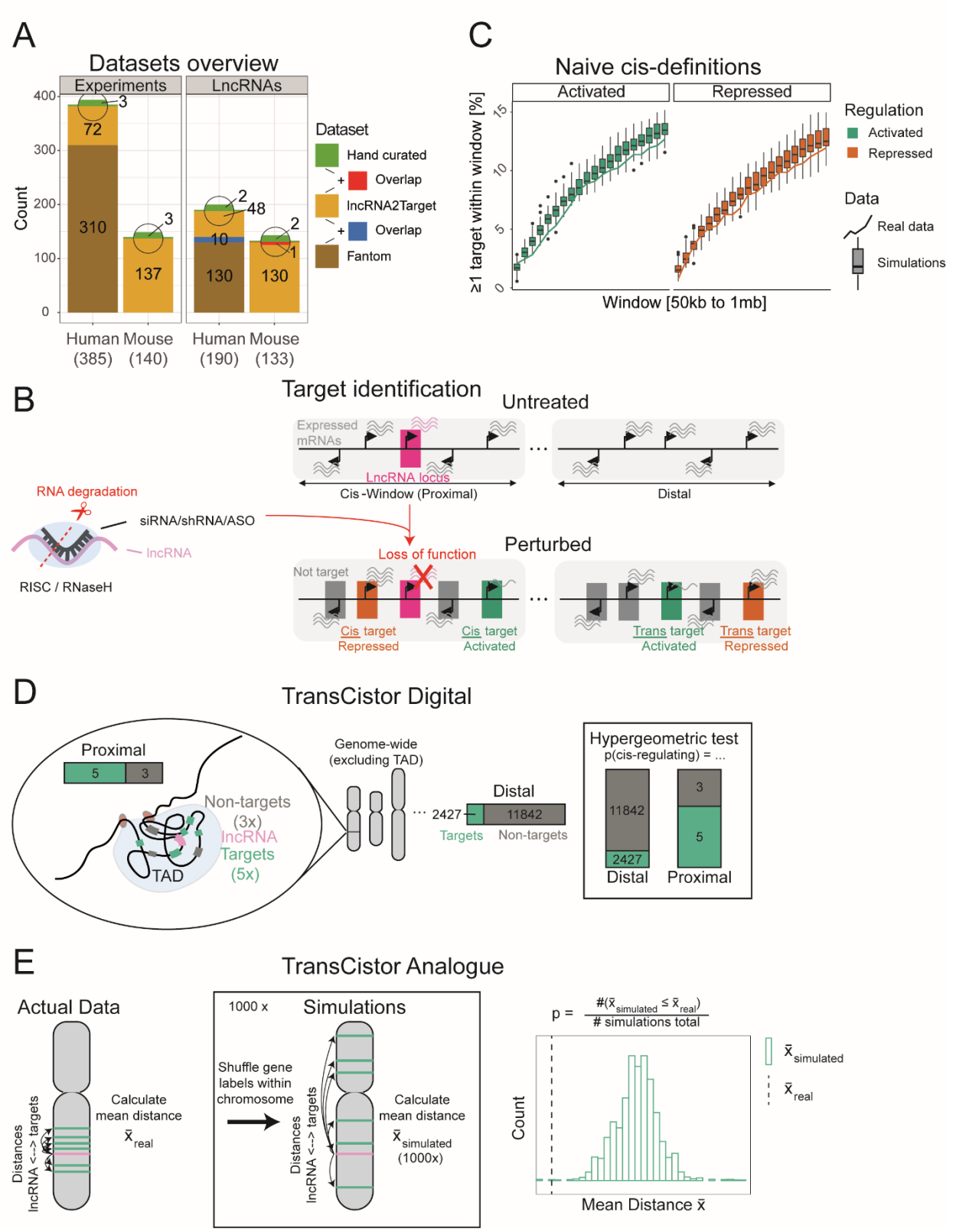
The human TSS annotations were intersected with chromatin states at several genomic windows (1 bp, 100 bp, 1000 bp, 10000 bp) and a given state-TSS intersection was counted only if it was present in more samples than a given threshold (0, 1, 5, or 10 samples). For each pair of genomic window and filter, a contingency matrix was computed for each pair of predicted labels (*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 was obtained using the python package "hic-straw" (v1.2.1) (https://github.com/aidenlab/straw), using human HiC datasets from Aiden laboratory[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 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 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 *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. 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 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 (again using the 'glm' function).

Figures

Figure 1



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

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

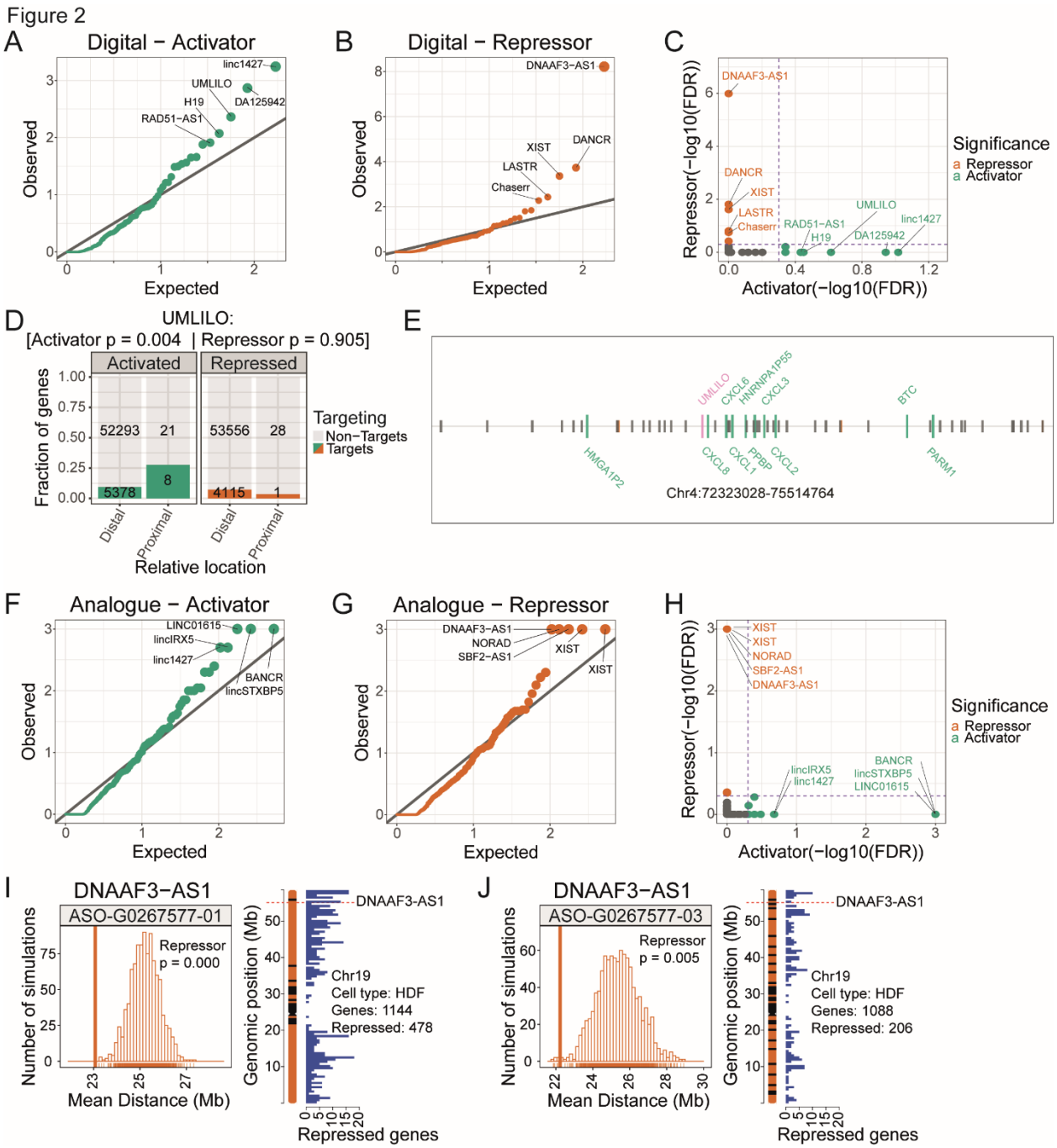


Figure 2: Large scale classification of *cis*-lncRNAs in human and mouse.

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

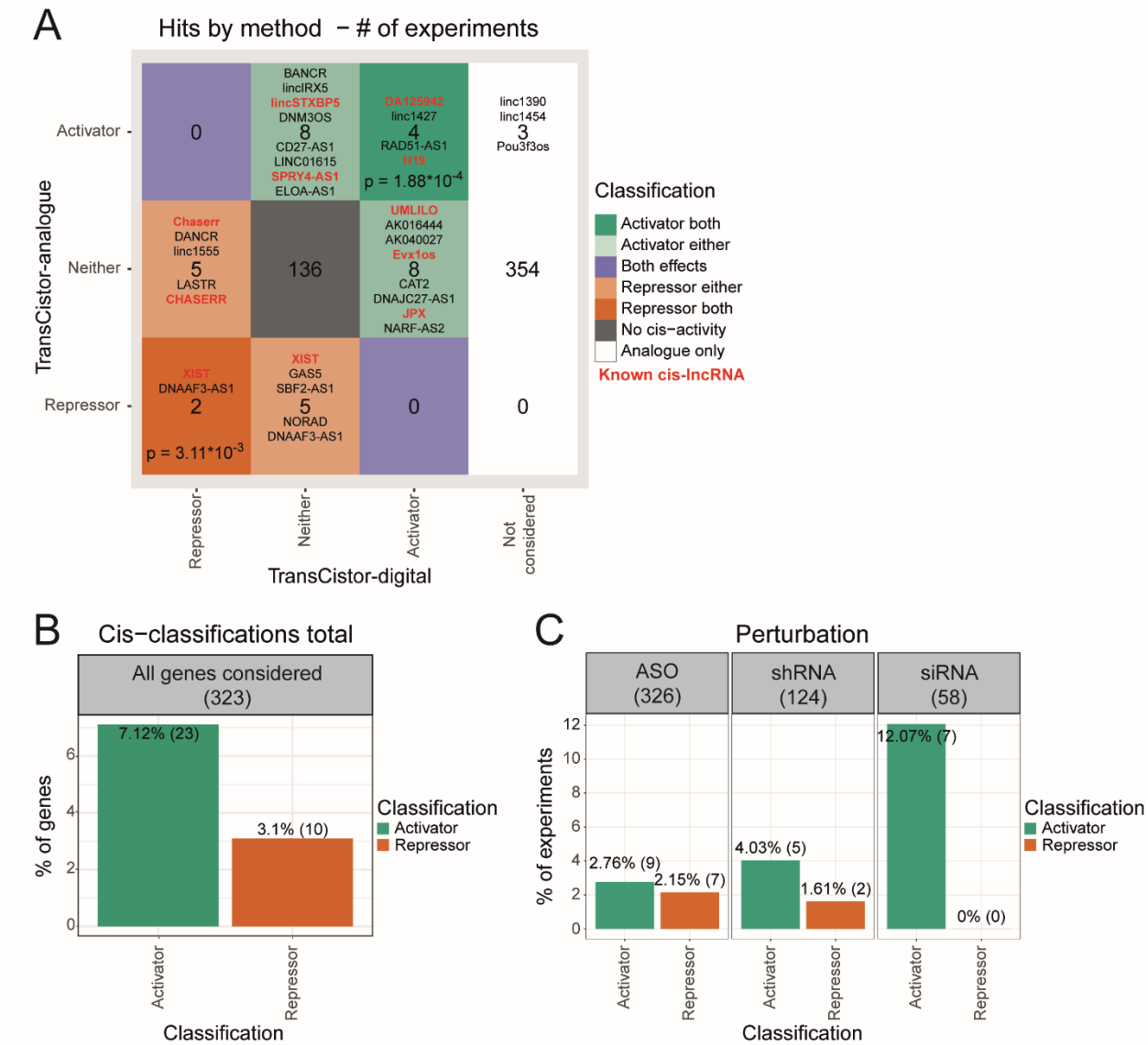
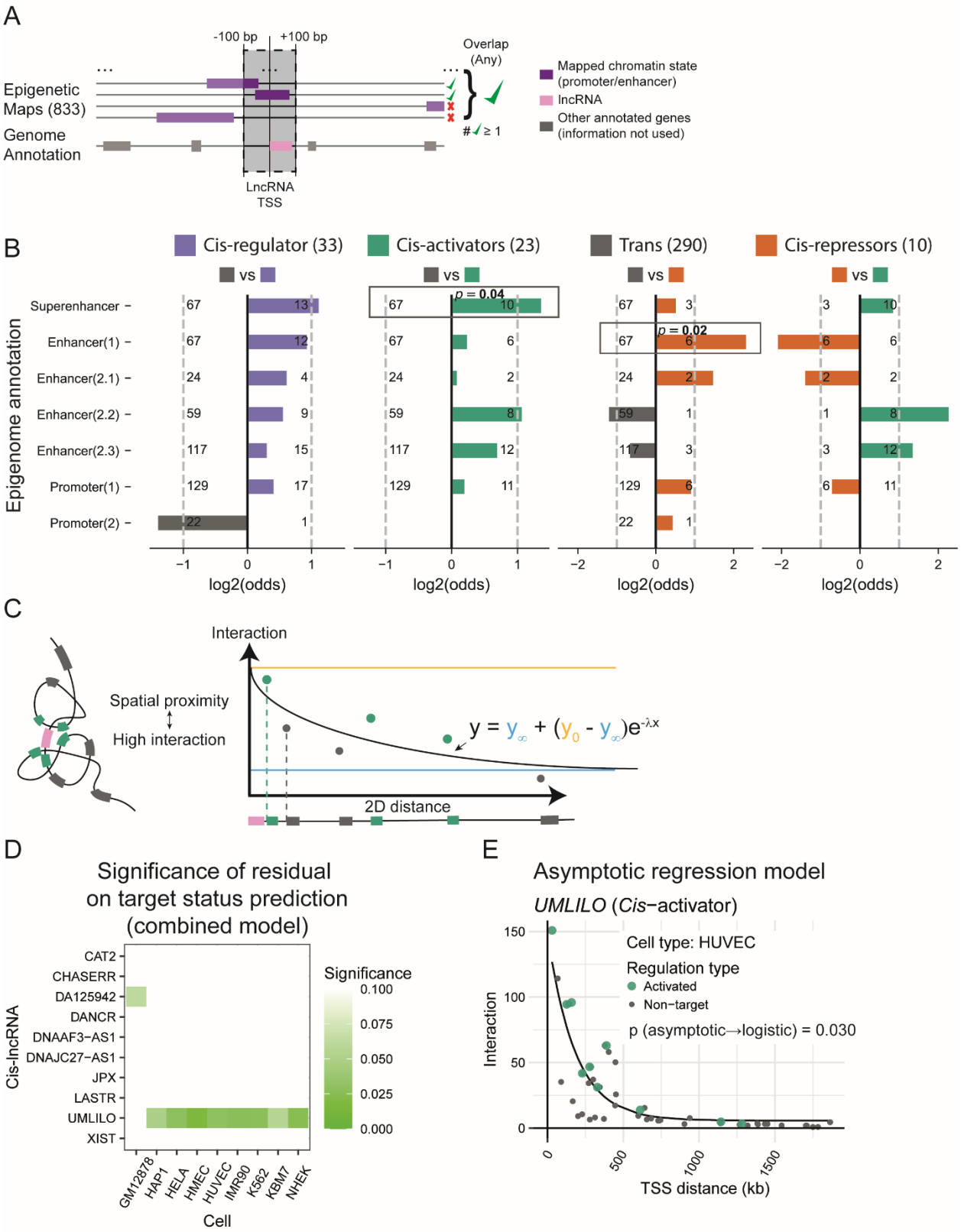


Figure 3: Rate of *cis*-lncRNA across perturbations, datasets and species.

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

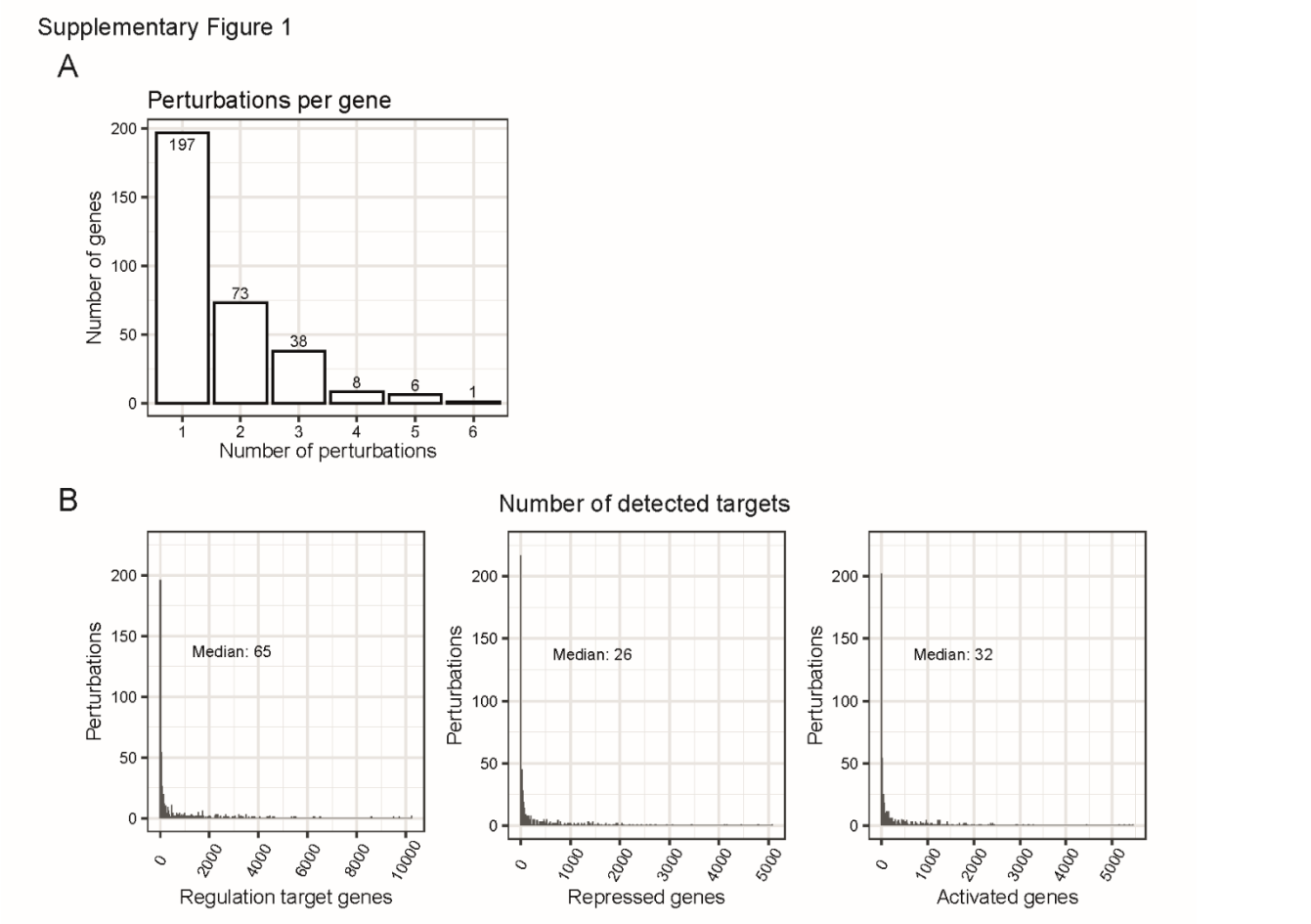

A
-100 bp
+100 bp
Overlap (Any)
Epigenetic Maps (833)
Genome Annotation
LncRNA TSS
≥ 1
Mapped chromatin state (promoter/enhancer)
lncRNA
Other annotated genes (information not used)
B
Cis-regulator (33)
Cis-activators (23)
Trans (290)
Cis-repressors (10)
vs
Epigenome annotation
Superenhancer
Enhancer(1)
Enhancer(2.1)
Enhancer(2.2)
Enhancer(2.3)
Promoter(1)
Promoter(2)
p = 0.04
p = 0.02
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
CHASERR
DA125942
DANCR
DNAAF3-AS1
DNAJC27-AS1
JPX
LASTR
UMLILO
XIST
Significance
0.100
0.075
0.050
0.025
0.000
GM12878
HAP1
HELA
HMEC
HUVEC
IMR90
K562
KBM7
NHEK
Cell
E
Asymptotic regression model
UMLILO (Cis−activator)
Cell type: HUVEC
Regulation type
Activated
Non−target
p (asymptotic→logistic) = 0.030
Interaction
TSS distance (kb)

412 Figure 4: Intersection of *cis*-lncRNAs with enhancer elements and generation of a chromatin
413 conformation based model.

414 A) Method of calculating overlap by enhancer annotations (horizontal purple bars) of lncRNA TSS
415 (pink bar). Overlaps are considered while varying two important thresholds: numbers of individual
416 enhancer annotations are considered the minimum necessary (epigenome threshold) and various
417 sized windows around the TSS of overlap calculation between extended TSS region (span). Only
418 the TSS spans with overlaps in more samples than a given epigenome threshold are considered.
419 B) Enrichment results epigenome threshold=1 and span=100 bp. Rows show enrichment for
420 super-enhancer, enhancer, and promoter states while comparing the TSS according to their
421 mechanism of action (see Methods). P-values were computed using Fisher's Exact Test. C) A
422 model for proximity-driven target selection: (Left panel) Chromatin folding brings lncRNA (pink)
423 into spatial proximity with proximal genes, which are subsequently targeted (green). (Right panel)
424 Chromatin proximity maps, such as provided by HiC methodology, enable one to evaluate the
425 spatial proximity (*y*-axis) of targets, while normalising for confounder of linear 2D DNA distance
426 (*x*-axis). These parameters were modelled using an Asymptotic regression model (right panel,
427 inset). D) Evaluating the contribution of proximity to target selection in human cells: The model
428 significance of *cis*-lncRNAs (identified by TransCistor-digital) (*x*-axis) was evaluated across HiC
429 interaction data from a panel of human cell lines (*y*-axis). Colour scale shows uncorrected p-
430 values. Green cells indicate cases where target genes tend to be significantly more proximal than
431 non-targets. No cases of the inverse were observed. E) Example data for *UMLILO* in HUVEC
432 cells. Note that target genes (green) tend to be more spatially proximal (*y*-axis) than non-target
433 genes (grey) at a similar TSS-to-TSS genomic distance (*x*-axis).

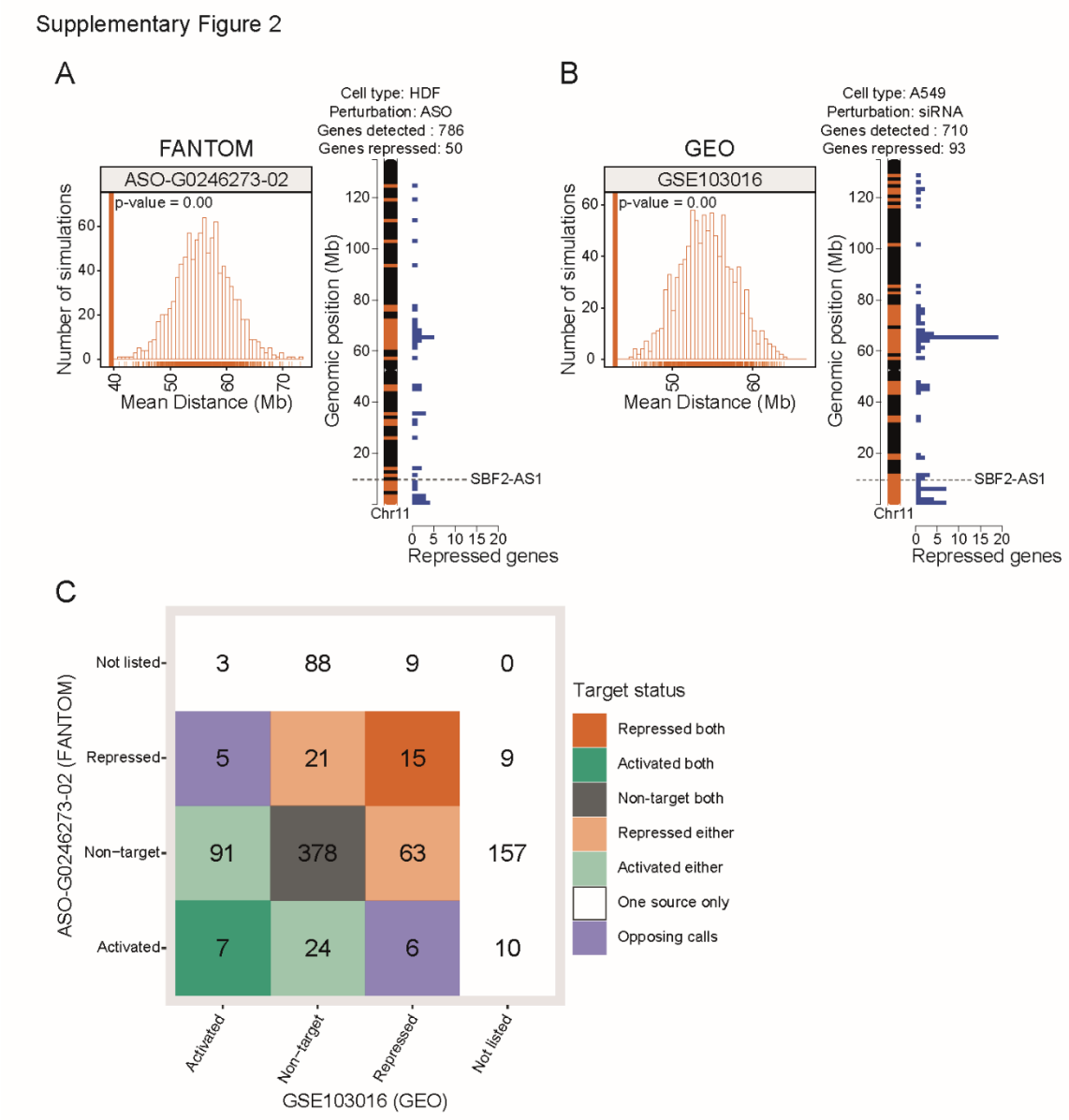
434

435 # Supplementary Figures



437 <u>Supplementary Figure S1: Summary statistics of perturbation datasets.</u>

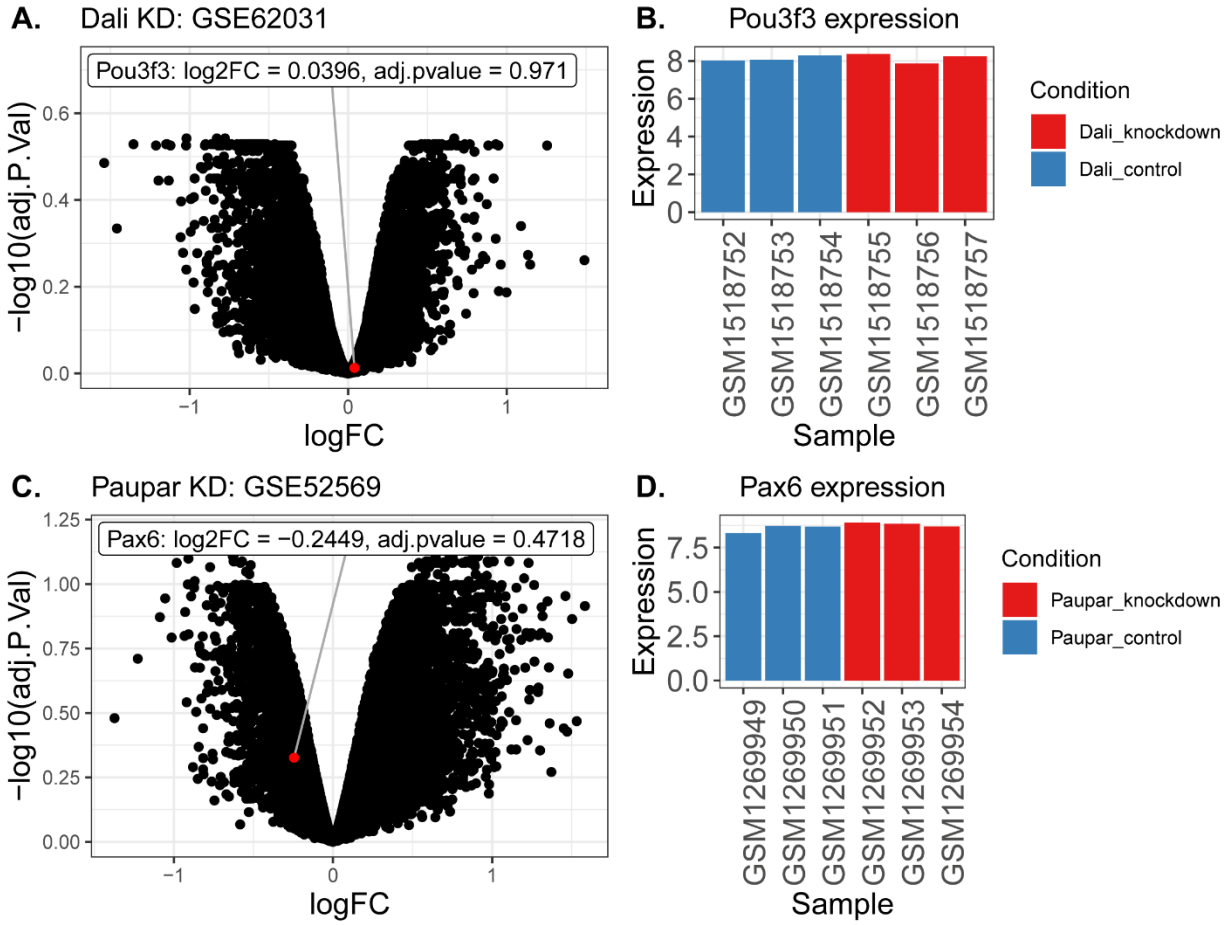
438 A) Histogram displaying the numbers of separate perturbation experiments (*x*-axis) available for
439 each lncRNA gene (*y*-axis). B) Histograms displaying the number of significantly changing genes
440 (targets) (*x*-axis) for each perturbation experiment (*y*-axis). Regulated genes represent the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



Supplementary Figure S2: Analysis of *cis*-regulation by *SBF2-AS1* in independent datasets.

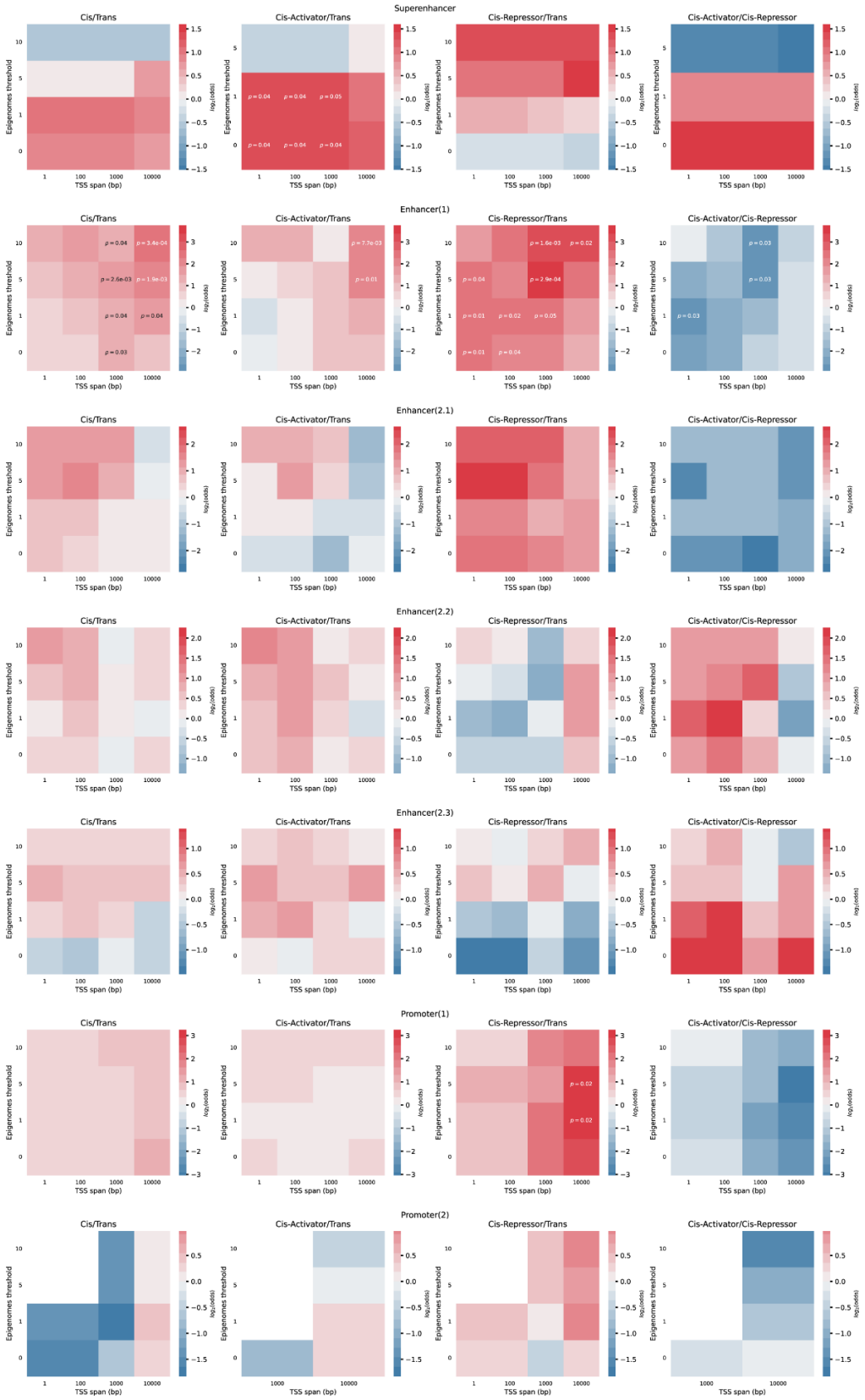
(A) TransCistor-analogue results for FANTOM ASO knockdown targeting *SBF2-AS1* in human dermal fibroblasts. B) As for (A), but for independent data from A549 cells treated with siRNA. C) Numbers indicate the genes in each category, classified by their regulation in the two distinct datasets in (A) and (B).

Supplementary Figure 3

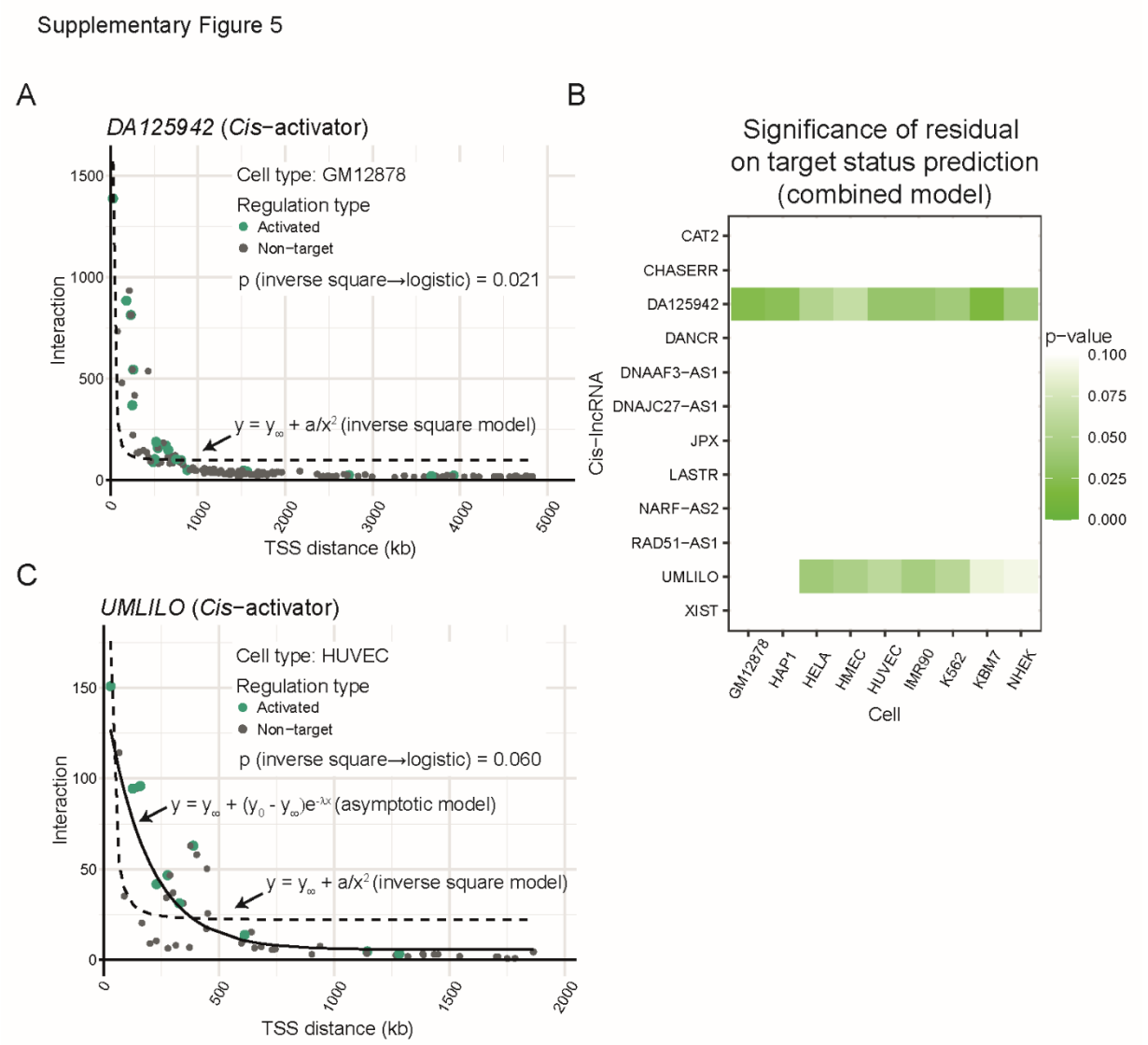


Supplementary Figure S3: Analysis of *Dali* and *Paupar* target genes using public microarray data. A) Global transcriptome changes upon *Dali* knockdown were obtained from Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE62031). B) Neighbour gene and putative target *Pou3f3* 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



457 Supplementary Figure S4: Enrichment of enhancers, super-enhancers and promoters in *cis*-
458 lncRNAs. Each row represents a different enhancer annotation (see Methods). Columns
459 represent comparisons between indicated pairs of lncRNA classes. Heatmaps display the
460 enrichment of overlap at different genomic windows around the lncRNA TSSs (span, *x*-axis) and
461 the minimum number of observed samples required to define an enhancer (epigenome threshold,
462 *y*-axis). Statistical significance is indicated where p-value ≤ 0.05 by Fisher's exact test (1-sided),
463 and not corrected for multiple hypothesis testing.



465 Supplementary Figure S5: Alternative inverse square model for target gene interaction.

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

469 ## Supplementary Data Files

470 Data File 1: Information for all lncRNAs 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

1. Thorvaldsen, J. L., Duran, K. L. & Bartolomei, M. S. Deletion of the H19 differentially methylated domain results in loss of imprinted expression of H19 and Igf2. *Genes & Development* **12**, 3693 (1998).

2. Brown, C. J. *et al.* A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome. *Nature* **349**, 38–44 (1991).

3. Ørom, U. A. *et al.* Long Noncoding RNAs with Enhancer-like Function in Human Cells. *Cell* **143**, 46–58 (2010).

4. Rom, A. *et al.* Regulation of CHD2 expression by the Chaserr long noncoding RNA gene is essential for viability. *Nature Communications* **10**, 1–15 (2019).

5. Leucci, E. *et al.* Melanoma addiction to the long non-coding RNA SAMMSON. *Nature* **531**, 518–522 (2016).

6. Gil, N. & Ulitsky, I. Regulation of gene expression by cis-acting long non-coding RNAs. *Nature Reviews Genetics* vol. 21 102–117 Preprint at https://doi.org/10.1038/s41576-019-0184-5 (2020).

7. Xiang, J. F. *et al.* Human colorectal cancer-specific CCAT1-L lncRNA regulates long-range chromatin interactions at the MYC locus. *Cell Research* **24**, 513–531 (2014).

8. Fanucchi, S. *et al.* Immune genes are primed for robust transcription by proximal long noncoding RNAs located in nuclear compartments. *Nature Genetics 2018 51:1* **51**, 138–150 (2018).

9. Gil, N. & Ulitsky, I. Production of spliced long noncoding RNAs specifies regions with increased enhancer activity. *Cell Syst* **7**, 537 (2018).

10. Engreitz, J. M. *et al.* Local regulation of gene expression by lncRNA promoters, transcription and splicing. *Nature* **539**, 452–455 (2016).

11. Ørom, U. A. & Shiekhattar, R. Long non-coding RNAs and enhancers. *Curr Opin Genet Dev* **21**, 194–198 (2011).

12. Tan, J. Y., Biasini, A., Young, R. S. & Marques, A. C. Splicing of enhancer-associated lincRNAs contributes to enhancer activity. *Life Sci Alliance* **3**, (2020).

13. Marques, A. C. *et al.* Chromatin signatures at transcriptional start sites separate two equally populated yet distinct classes of intergenic long noncoding RNAs. *Genome Biology* **14**, R131 (2013).

14. Hnisz, D., Shrinivas, K., Young, R. A., Chakraborty, A. K. & Sharp, P. A. A Phase Separation Model for Transcriptional Control. *Cell* **169**, 13–23 (2017).

15. Wu, M., Yang, L. Z. & Chen, L. L. Long noncoding RNA and protein abundance in lncRNPs. *RNA* **27**, 1427 (2021).

16. Palazzo, A. F. & Koonin, E. V. Functional Long Non-coding RNAs Evolve from Junk Transcripts. *Cell* **183**, 1151–1161 (2020).

523 17. Ramilowski, J. A. *et al.* Functional Annotation of Human Long Non-Coding RNAs
524 via Molecular Phenotyping. *bioRxiv* 700864 (2019) doi:10.1101/700864.

525 18. Ramilowski, J. A. *et al.* Functional annotation of human long noncoding RNAs via
526 molecular phenotyping. *Genome Research* **30**, 1060–1072 (2020).

527 19. Tan, J. Y. *et al.* cis -Acting Complex-Trait-Associated lincRNA Expression
528 Correlates with Modulation of Chromosomal Architecture. *Cell Reports* **18**, 2280–2288
529 (2017).

530 20. Marstrand, T. T. & Storey, J. D. Identifying and mapping cell-type-specific
531 chromatin programming of gene expression. *Proc Natl Acad Sci U S A* **111**, E645–E654
532 (2014).

533 21. Wang, Y. *et al.* The 3D Genome Browser: A web-based browser for visualizing 3D
534 genome organization and long-range chromatin interactions. *Genome Biology* **19**, 1–12
535 (2018).

536 22. Forne, T. *et al.* Loss of the maternal H19 gene induces changes in Igf2 methylation
537 in both cis and trans. *Proc Natl Acad Sci U S A* **94**, 10243 (1997).

538 23. Carmona, S., Lin, B., Chou, T., Arroyo, K. & Sun, S. LncRNA Jpx induces Xist
539 expression in mice using both trans and cis mechanisms. *PLoS Genet* **14**, (2018).

540 24. Luo, S. *et al.* Divergent lncRNAs Regulate Gene Expression and Lineage
541 Differentiation in Pluripotent Cells. *Cell Stem Cell* **18**, 637–652 (2016).

542 25. Maass, P. G. *et al.* A misplaced lncRNA causes brachydactyly in humans. *The*
543 *Journal of Clinical Investigation* **122**, 3990 (2012).

544 26. Engreitz, J. M. *et al.* The Xist lncRNA Exploits Three-Dimensional Genome
545 Architecture to Spread Across the X Chromosome. *Science (1979)* **341**, 1237973 (2013).

546 27. Vance, K. W. *et al.* The long non-coding RNA Paupar regulates the expression of
547 both local and distal genes. *EMBO J* **33**, 296–311 (2014).

548 28. Chalei, V. *et al.* The long non-coding RNA Dali is an epigenetic regulator of neural
549 differentiation. *Elife* **3**, 1–24 (2014).

550 29. Stalder, L. *et al.* The rough endoplasmatic reticulum is a central nucleation site of
551 siRNA-mediated RNA silencing. *The EMBO Journal* **32**, 1115 (2013).

552 30. Zeng, Y. & Cullen, B. R. RNA interference in human cells is restricted to the
553 cytoplasm. *RNA* **8**, 855–60 (2002).

554 31. Biasini, A. *et al.* Translation is required for miRNA-dependent decay of
555 endogenous transcripts. *The EMBO Journal* **40**, (2021).

556 32. Lai, F., Damle, S. S., Ling, K. K. & Rigo, F. Directed RNase H Cleavage of Nascent
557 Transcripts Causes Transcription Termination. *Molecular Cell* **77**, 1032-1043.e4 (2020).

558 33. Durand, N. C. *et al.* Juicer Provides a One-Click System for Analyzing Loop-
559 Resolution Hi-C Experiments. *Cell Systems* **3**, 95–98 (2016).

560 34. Dixit, A. *et al.* Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell
561 RNA Profiling of Pooled Genetic Screens. *Cell* **167**, 1853-1866.e17 (2016).

562 35. Basu, S. & Larsson, E. A Catalogue of Putative cis-Regulatory Interactions
563 Between Long Non-coding RNAs and Proximal Coding Genes Based on Correlative
564 Analysis Across Diverse Human Tumors. *G3 Genes|Genomes|Genetics* **8**, 2019–2025
565 (2018).

566 36. Agrawal, S. *et al.* Systematic identification of cis-interacting lncRNAs and their
567 targets. *bioRxiv* 2021.01.13.426305 (2022) doi:10.1101/2021.01.13.426305.

568 37. Goyal, A. *et al.* A cautionary tale of sense-antisense gene pairs: independent
569 regulation despite inverse correlation of expression. *Nucleic Acids Res* **45**, 12496–12508
570 (2017).

571 38. Scharner, J. *et al.* Hybridization-mediated off-target effects of splice-switching
572 antisense oligonucleotides. *Nucleic Acids Research* **48**, 802 (2020).

573 39. Marasco, L. E. *et al.* Counteracting chromatin effects of a splicing-correcting
574 antisense oligonucleotide improves its therapeutic efficacy in spinal muscular atrophy. *Cell*
575 **185**, 2057-2070.e15 (2022).

576 40. Doolittle, W. F. We simply cannot go on being so vague about “function.” *Genome*
577 *Biology* **19**, 1–3 (2018).

578 41. Graur, D. *et al.* On the Immortality of Television Sets: “Function” in the Human
579 Genome According to the Evolution-Free Gospel of ENCODE. *Genome Biology and*
580 *Evolution* **5**, 578–590 (2013).

581 42. Frankish, A. *et al.* GENCODE 2021. *Nucleic Acids Res* **49**, D916–D923 (2021).

582 43. Cheng, L. *et al.* LncRNA2Target v2.0: a comprehensive database for target genes
583 of lncRNAs in human and mouse. *Nucleic Acids Research* **47**, (2019).

584 44. Hoon, D. S. B., Rahimzadeh, N. & Bustos, M. A. EpiMap: Fine-tuning integrative
585 epigenomics maps to understand complex human regulatory genomic circuitry. *Signal*
586 *Transduction and Targeted Therapy 2021 6:1* **6**, 1–3 (2021).

587 45. Zacher, B. *et al.* Accurate Promoter and Enhancer Identification in 127 ENCODE
588 and Roadmap Epigenomics Cell Types and Tissues by GenoSTAN. *PLOS ONE* **12**,
589 e0169249 (2017).

590 46. Khan, A. & Zhang, X. dbSUPER: a database of super-enhancers in mouse and
591 human genome. *Nucleic Acids Research* **44**, D164–D171 (2016).

592 47. Ramilowski, J. A. *et al.* Functional annotation of human long noncoding RNAs via
593 molecular phenotyping. *Genome Research* **30**, 1060–1072 (2020).

594 48. Cheng, L. *et al.* LncRNA2Target v2.0: a comprehensive database for target genes
595 of lncRNAs in human and mouse. *Nucleic Acids Research* **47**, D140 (2019).

596