Deducing transcriptional kinetics and molecular functions of long non-coding RNAs using allele-sensitive single-cell RNA-sequencing

Per Johnsson[1,2], Leonard Hartmanis[1], Christoph Ziegenhain[1], Gert-Jan Hendriks[1], Michael Hagemann-Jensen[1], Björn Reinius[2,3], Rickard Sandberg[1,2,*]

[1] Department of Cell and Molecular Biology, Karolinska Institutet, Stockholm, Sweden

[2] Ludwig Institute for Cancer Research, Stockholm, Sweden

[3] Department of Medical Biochemistry and Biophysics, Karolinska Institutet, Stockholm, Sweden

*Correspondence to: Rickard.Sandberg@ki.se

Abstract

An increasing number of studies have demonstrated the importance of long non-coding RNAs (lncRNAs) but it generally remains challenging to predict lncRNA functions and determine their mechanisms. Strategies are needed to generate hypotheses of lncRNA functions and to efficiently identify candidates for in-depth investigations. Here, we used allele-sensitive single-cell RNA-seq (scRNA-seq) to both characterize lncRNA transcriptional dynamics and identify lncRNA candidates for functional characterizations. Specifically, we revealed that low expression of lncRNAs is due to lowered burst frequencies and that lncRNAs show increased cell-to-cell variability compared with expression-matched mRNAs. Exploiting the cell-to-cell variability in asynchronously growing cells, we identified and experimentally validated functions of several lncRNAs involved in cell cycle progression and apoptosis. Finally, utilizing the allelic resolution, we identified pairs of lncRNAs and mRNAs with coordinated allelic expression patterns and validated their effect through knockdown experiments. Collectively, these studies demonstrated that single-cell analysis is a powerful paradigm for identifying lncRNA functions.

Introduction

Mammalian genomes encode thousands of lncRNAs[1], but identifying their molecular functions have proven difficult. Functional prediction based on primary sequence, evolutionary conservation[2], or genomic location is challenging and the functions of most lncRNAs remains unknown[3]. New experimental and computational approaches are therefore needed in order to efficiently identify lncRNAs for in-depth functional validation and characterization.

Studies of lncRNAs have shown that their average expression levels are lower than mRNAs[4–9], yet vital regulatory functions have been demonstrated for several lncRNAs. For example, the *HOTTIP* lncRNA functions through maintaining active chromatin states for *HOXA* genes, despite only having an average expression of ~0.3 RNA copies per cell[10]. It has been proposed that averaging transcriptomes over thousands of cells could mask the presence of few cells with relatively high expression of specific lncRNAs[11]. However, comprehensive analyses of transcriptional dynamics and cell-to-cell variability of lncRNAs are still missing, and most studies to date were limited to low throughput methods covering few number of genes and cells[12]. With the introduction of scRNA-seq technologies[13] and certain protocols allowing for allele-specific quantification[14], it is now feasible to characterize allele specific gene expression in individual cells for thousands of genes simultaneously. Although scRNA-seq is a powerful tool for identification of cell types[15], transient cellular states[16] and burst kinetics[17], dedicated scRNA-seq studies with allelic-resolution focusing on lncRNAs are so far limited.

Transcription of mammalian genes typically occurs in short bursts[18], where the burst frequency is the main modulator of cell-type specific differences in gene expression[17]. Burst frequencies are generally controlled by enhancers, whereas burst sizes are in part dependent

upon core promoter elements[17]. To date, analyses have mainly focused on protein-coding genes and it is unclear whether lncRNAs have similar or distinct characteristics of transcriptional kinetics.

In this study, we introduce allele-sensitive scRNA-seq to investigate transcriptional dynamics and molecular functions of lncRNAs. We first set out to study cell-to-cell variability and burst kinetics of lncRNAs and investigate if lncRNAs have altered burst kinetics compared to mRNAs. We next explore if the identification of transient cellular states with specific expressions of lncRNAs predicts their molecular functions. Using an expression-to-phenotype relationship in non-perturbed asynchronously growing cells, we identify several lncRNAs involved in cell cycle regulation as well as apoptosis. Finally, we explore patterns of allelic imbalance of lncRNAs to identify and validate *cis*-functioning lncRNAs.

Results

Detection of lncRNAs and mRNAs in individual fibroblasts.

We first investigated the expression patterns of lncRNAs and mRNAs in individual mouse fibroblasts of the cross between the distantly related CAST/EiJ and C57BL/6J strains. We profiled 533 individual adult tail fibroblasts with Smart-seq2[19,20] (**Figure S1A**) to leverage the methods high sensitivity[21] and coverage enabling allele-level RNA profiling for more than 80% of all genes[14]. For allelic expression estimates the fraction of reads covering single-nucleotide polymorphisms (SNP) were used. We first verified that non-imprinted autosomal genes (**Supplementary Table 1**) had similar overall expression from the CAST and C57 alleles (**Figure S1B**) and that we accurately detected monoallelic expression for non-escapee X-chromosomal genes[22] (**Figure S1C, Supplementary Table 2**). A total number of 24,653 genes were detected (requiring 5 or more read counts in at least 2 cells), including 15,872 mRNAs and 3,314 non-coding RNAs. On average, 9,174 protein-coding mRNAs and 408 lncRNAs were detected per cell (**Figure 1A**). The detection of thousands of lncRNAs motivated us to proceed with in-depth investigations of lncRNAs and we focused our analysis on easily separated transcriptional units since we noticed an effect on gene expression levels on closely located promoters (**Figures S1D-E**).

lncRNA expression has higher cell-to-cell variability than mRNA.

We first investigated the patterns of expression of lncRNAs and mRNAs, and as expected[4,23], we observed that lncRNAs were expressed at lower levels than mRNAs (**Figure 1B**) and detected in fewer cells (median 3% and 31% of cells, respectively) (**Figure 1C**). To further investigate if lncRNAs are expressed with higher variability between cells, we computed the squared coefficient of variation (CV^2) and observed significantly higher variability for lncRNAs

(**Figure 1D,** P=3.8e^{-293}, two-sided Wilcoxon rank-sum test). Contrasting CV^2 against the mean expression revealed that lncRNAs had higher CV^2 than mRNAs across a wide range of expression levels (**Figure 1E**). To more systematically account for possible confounding differences in mean expression of lncRNAs and mRNAs, we generated thousands of sets of randomly drawn of mRNAs with expressions matched to lncRNAs (see Methods). Consistently, lncRNAs had significantly higher expression variability than expression-matched mRNAs ($P<1e^{-4}$, permutation test) (**Figure 1F**). The increased cell-to-cell variability was also validated in human HEK293 cells (**Figure S2A**), as well as mouse embryonic stem cells (**Figure S2B**). The ability to detect the increased cell-to-cell variability was dependent on the number of lncRNAs analyzed, and when subsampling lncRNAs (and their expression-matched mRNAs) the difference declined and eventually disappeared (**Figure 1G**). The lack power when analyzing small numbers of lncRNAs might explain why a study investigating 34 lncRNAs found no difference in variability compared with mRNAs[12].

Low expression of lncRNAs is mediated by lowered burst frequency.

We next evaluated how the lowered expression of lncRNAs is achieved in terms of transcriptional bursting kinetics. To this end, we inferred burst kinetics on adult tail fibroblasts from Smart-Seq3 quantifications[24] (**Figures S3A-C**) which, in addition to Smart-Seq2, harbors unique molecular identifiers (UMIs) important for accurate burst size inference and benefits from generally improved sensitivity over Smart-seq2[17]. After quality control (**Figures S3D-E**), bursting kinetics parameters were successfully inferred for 5,623 genes (5,273 coding and 350 lncRNAs on both alleles) and burst parameters correlated well between the CAST and C57 alleles for coding and non-coding genes (**Figures 2A-B**). Focusing the analysis on separated transcriptional units (**Figure S1E**) we found that lncRNAs have significantly lowered burst

frequencies than mRNAs (P=2.2e^{-42}, two-sided Wilcoxon rank-sum test), whereas no significant difference was observed for burst sizes (P=0.09, two-sided Wilcoxon rank-sum test) (**Figures 2C-D** and **S4A-B**). Thus, the decreased expression of lncRNAs is mainly achieved by having lower frequencies of expression, likely regulated by enhancer activities[17,25–27].

Many lncRNAs are transcribed in antisense direction of protein-coding (sense) genes[28] and we next investigated if such genomic organizations could result in altered transcription kinetics. We identified loci with divergent mRNA-mRNA pairs (n=1,176), divergent mRNA-lnRNA pairs (n=336) and unidirectional mRNA (n=7,059) transcribed promoters (**Figure 2E**). In line with a previous study reporting increased expression of divergently transcribing promoters[29], we observed increased expression of divergently transcribing mRNA-mRNA as well as mRNA-lncRNA promoters, compared to unidirectional transcribing promoters (**Figures 2E and S4C**). We next explored if the increase in expression was related to alterations in burst kinetics and observed a significant increase in burst frequency for divergently transcribing promoters (**Figure 2F**). Finally, we also noted that divergent mRNA-mRNA gene pairs presented decreased burst size (P =1.8e^{-7}, two-sided Wilcoxon rank-sum test), although no such difference was observed in divergent lncRNA-mRNA gene pairs (**Figures 2G and S4D**).

Expression of lncRNAs in transient cellular phenotypes reveals function

We next explored if scRNA-seq could serve as a tool for functional annotation of lncRNAs. More specifically, we hypothesized that the dynamic expression and cellular heterogeneity carries information to their functions. To this end, we first sought out to identify lncRNAs involved in cell cycle progression. Asynchronously growing mouse fibroblast (**Figures S1A-C**) were projected into low-dimensional PCA space using the most variable[30] cell cycle genes[31]

(**Figures S5A-B, Supplementary Table 3**), clustered and the PCA coordinates were used to fit a principal curve[32]. Cells were aligned to the fit and cell cycle progression was confirmed by a subset of well-established cell cycle genes marking cells in G0, G1, G1S and G2M (**Figure 3A, Figure S5C**). To identify lncRNAs with high expression within specific phases of the cell cycle, we next applied an ANOVA test (FDR < 0.01, Benjamini-Hochberg adjusted) on gene expression over the individual cell cycle phases. The P-values were contrasted against the fold induction in the particular cell cycle phase (**Figure 3B**), revealing a total of 128 lncRNAs with cell cycle specific expression pattern (fold change > 1.5, adjusted p<0.01) (**Figure 3B**), of which we selected seven highly ranked candidates for further characterization.

To functionally evaluate the selected lncRNA candidates' potential involvement in cell cycle progression, we used the immortalized mouse embryonic fibroblast NIH3T3 cell line. The two model system expressed overlapping cell cycle genes[31] (**Supplementary Table 3**) and correlated well in expression levels (**Figure S5D**). Next, cell cycle progression of NIH3T3 cells were synchronized by serum starvation (G0/G1), thymidine block (G1S) or nocodazole treatment (G2M), and the synchronization was validated by flow cytometry (**Figure S5E**) and qRTPCR for two cell cycle marker genes (**Figure S5F**). Subsequently, all seven lncRNAs had peak expression in the predicted phases of the cell cycle, as validated by qRTPCR (**Figure S5G**). Having validated cell cycle specific expression of these lncRNA transcripts, we next selected three candidates for in-depth functional investigation. To this end, we generated individual lentiviral transduced NIH3T3 cell lines with stable shRNA-induced knockdown of each of the three lncRNA candidates (*Wincr1*, *Lockd* and *A730056A06Rik*) (**Figure 3C**). Although no clear phenotype was observed for these cells under normal growth conditions, significant effects were observed in colony formation assays (**Figure 3D**), which provide a moderate stress on

cells. Whereas the knockdown of *A730056A06Rik* (expressed in arrested cells, **Figure S5G**) resulted in the formation of more colonies, the knockdown of *Wincr1* and *Lockd* (expressed in proliferating cells) reduced the numbers of colonies formed (**Figure 3D**).

Functional investigation of the lncRNA *Lockd*

We were intrigued by a previous report[33] suggesting that *Lockd* transcription promotes the expression of the cell cycle regulator *Cdkn1b* (in *cis*) although its promoter is located 10kb upstream of the *Lockd* locus (**Figure S6A**). In our experiment, the reduction of the *Lockd* transcript had a clear functional consequence – by lowering colony formation (**Figure 3D**). To complement the stable *Lockd* knockdown, we designed two siRNAs against the *Lockd* transcript which both achieved good knockdown efficiency with less than 10% *Lockd* expression in the NIH3T3 cell line as well as in primary fibroblasts (**Figure 4A**). In agreement with the NIH3T3-shLockd stable cell line (**Figure 3D**), a consistent decrease in colony forming cells was observed upon siRNA-induced *Lockd* depletion (**Figure 4B**). In line with the previous report[33], no consistent change in RNA expression was observed for *Cdkn1b* upon knockdown of the *Lockd* transcript in NIH3T3 nor primary fibroblast cells (**Figure 4C**). However, the allele-resolved scRNA-seq data suggested co-expression of *Lockd* and *Cdkn1b* (tended to be expressed in the same cells and from the same allele) on both the CAST (Fisher's exact test, P=0.013) and C57 (Fisher's exact test, P=0.00032) alleles (**Figure S6B**).

To characterize the molecular function of *Lockd* further, we used scRNA-seq on the shControl (n=147 cells) and shLockd stable cell lines (n=144 cells). We observed that 752 genes had significantly altered expression in the *Lockd* reduced cell line (SCDE[34], adjusted p<0.05) (**Figure S6C**), likely including both direct effects and indirect effects of impaired cell cycle progression

and shRNA-induced off-target effects. To further elucidate the function of *Lockd*, we retrieved gene-gene Spearman correlations from the shControl treated stable cell line and hypothesized that the most relevant targets for *Lockd* would have a positive correlation with *Lockd* and concurrent reduced expression in the shLockd cells. Alternatively, negative correlation with increased expression upon knockdown (**Figure 4D**). This approach highly reduced the number of candidate genes (752 to 138 genes) and revealed several well-established cell cycle regulators. Particularly, three members of the kinesin superfamily (*Kif4*, *Kif11* and *Kif14*), a group of proteins known to be involved in mitosis, appeared as main candidates (**Figure 4E**). Notably, a link between these genes and Cdkn1b has been suggested. While Cdkn1b acts as a transcriptional suppressor of *Kif11* by binding to the *Kif11* promoter through a p130-E2F4 dependent mechanism[35], Kif14 regulates protein levels of Cdkn1b through a proteasome-dependent pathway[36]. Based on these previous findings, we set out to validate the effect on *Kif4*, *Kif11* and *Kif14* by measuring the expression level on qRTPCR upon siRNA induced knockdown of *Lockd* in NIH3T3 and primary fibroblast cells. The effect on *Kif11* and *Kif14* was validated in both cell lines while the effect on *Kif4* was only observed in the primary fibroblasts (**Figure 4F**). However, this is consistent with the scRNA-seq data (**Figure 4E**) where *Kif4* was more modestly affected compared to *Kif11* and *Kif14*. In summary, our data suggest *Lockd* to hold *trans* regulatory mechanisms, in addition to its previously reported *cis*-function[33]. While transcription of the *Lockd* gene functions *in cis* to promote transcription of *Cdkn1b*[33], the *Lockd* transcript appears to function in the same pathway as Cdkn1b and enhances the negative effects on cell cycle progression.

Functional investigation of the lncRNA *Wincr1*

To further investigate the molecular function of *Wincr1*, two siRNAs were designed against the *Wincr1* transcript and knockdown confirmed on qRTPCR (**Figure 5A**). In line with the NIH3T3-shWincr1 stable cell line, a decrease in colony forming cells, which scaled with the degree of knockdown (**Figure 5A**), was observed upon siRNA-induced depletion of *Wincr1* (**Figure 5B**). We next turned to our allele-resolved scRNA-seq data (**Figures S1A-C**) and observed that several genes at this locus, *Cdkn2a (*p16^{Ink4a} and p19^{ARF}p19^{Arf}), *Gm12602* and *Mtap*, presented coordinated expression with *Wincr1* (**Figures 5C and S7A**). Intriguingly, the homologous region in human has been reported to regulate the expression of CDKN2A (p16^{INK4A}) in a mechanism where the microRNA-31 host gene (MIR31HG) recruits chromatin remodeling factors to the promoter of p16^{INK4A}, [37]. However, mir31hg has a different genomic structure in mouse while *Wincr1* is absent in human cells (**Figures S7A-B**). We therefore sought to investigate if *Wincr1* had similar functions as human MIR31HG. Indeed, siRNA-induced knockdown of *Wincr1* in primary fibroblasts showed a significant increase in the expression of *Cdkn2a* (p16^{Ink4a} and p19^{Arf}) as well as *Cdkn2b* (p15^{Ink4b}) and *Mtap* (**Figure 5D**), suggesting *Wincr1* functions *in cis*. Finally, we also note that *Wincr1* has been proposed to be involved in Wnt/β-catenin signaling and to attenuate cell migration through *trans*-acting mechansisms[38]. Since *Cdkn2a* (p16^{Ink4a} and p19^{Arf})[39] as well as *Cdkn2b* (p15^{Ink4b})[39] are inactivated in NIH3T3 cells due to homozygous deletions of their chromosomal regions, our data support that *Wincr1* also maintains other *trans*-acting functions (**Figures 3D, 5B**).

Functional annotation of lncRNAs can be generalized to several phenotypes

We next evaluated if the presented approaches are applicable for other cellular states and set out to investigate if lncRNAs involved in apoptotic signaling could be identified. Since apoptotic signaling is linked to proliferation, we limited the analysis to cells in the G1 phase

(**Figure 3A**) and repeated the low-dimensional projection, now using the most variable genes related to apoptotic signaling (**Figure S8A**). We identified three clusters of cells (**Figure 6A**) and focused specifically on a subpopulation of cells that expressed several genes involved in stress signaling, i.e. *Gadd45b* (Growth arrest and DNA-damage-inducible, beta) and the p53 target gene *Cdkn1a* (**Figures 6B and S8B**). Single-cell differential expression (SCDE)[34] was applied to find lncRNAs with increased expression in this cluster of cells and a total of five highly ranked lncRNAs, to which siRNAs could be designed, were selected for further validation (**Figure 6C**). To validate these candidates, DNA damage induced apoptosis was triggered in the NIH3T3 cell line by treating cells with the chemotherapeutic and DNA crosslinking reagent Mitomycin C (MMC). DNA damage was validated on qRTCPR by measuring the expression levels of *Cdkn1a* (induced by p53) and *Gadd45b* (**Figure S8D**) and, in line with the prediction, all five candidate lncRNAs were induced at various concentrations of MMC (**Figure 6D**). To further investigate into the regulatory effects these lncRNA candidates might have on apoptosis, three of the candidates were suppressed by two siRNAs each (**Figure S8D**). The levels of apoptosis in lncRNA suppressed NIH3T3 cells was measured by AnnexinV after treatment with MMC (**Figure 6E**). Notably, the effect on apoptosis was frequently more pronounced when exposed to MMC, thus suggesting that these lncRNAs might sensitize cells to undergo apoptosis. In summary, our data support that also more subtle cellular states, such as pro-apoptotic signaling, is captured by scRNA-seq and serves as a profound tool to link lncRNAs to function.

Alllele-sensitive expression of lncRNAs reveals function

Previous bulk RNA-seq studies identified global allelic imbalance of gene expression to be pervasive across heterozygous F1 hybrid mice[40]. We next asked if allelic imbalance of lncRNAs

could reveal information about *cis* regulatory mechanisms and gene-gene interactions (**Figure 7A**). To improve power to detect gene-gene interactions, we profiled additional 218 mouse adult tail fibroblasts of the reciprocal cross (CASTxC57), resulting in a comprehensive Smart-seq2 dataset of 751 cells passing quality control (**Figures S9A-C and S1A-C**). We measured allelic distributions by counting allele sensitive read counts over both genomes (CAST / (CAST+C57) - 0.5) where a positive score reflects increased RNA expression towards the CAST genome[41]. Consistent with previous bulk RNA-seq studies[40], we confirmed >80% of mouse genes (9520 of 11377) to have RNA expression levels dependent on the genetic background (**Figure S9D**, binomial test, Benjamini and Hochberg adjusted $p<1e^{-5}$). We next applied stringent thresholds for allelic imbalance (|allelic imbalance| > 0.15) for downstream analysis where a total of 1,472 mRNAs and 252 lncRNAs (~15% of all genes with allelic information) passed the criteria (**Figures 7B and S9D**) and observed that lncRNAs generally showed greater allelic imbalance than mRNAs across a wide range of expression levels (**Figure S9E)**. Encouraged by the observation that the previously identified interaction, *Wincr1-Cdkn2b* (**Figure 5D**), passed these criteria (**Figure 9F**), we next sought to more comprehensively look for similar interactions.

We identified lncRNAs with greater/lower allelic imbalance than +/-0.15 (**Figures 7B and S9D**, 252 lncRNAs in total) and retrieved all genes with allelic information within +/- 500kb of the transcription start site (TSS). lncRNA-mRNA gene pairs were considered as candidates for downstream analysis if both presented allelic bias greater/lower than +/- 0.15, generating a total of 344 putative interactions (**Figure 7C**). We next ranked these interactions using the sum of allelic imbalance where a positive score reflects allelic imbalance towards the same allele and a negative score on alternative alleles (**Figure 7D**). Two highly ranked candidates were

experimentally tested and found to have a functional interaction, *Gsta4:C920006O11Rik* with allelic bias towards the same (C57) allele (**Figure 7E**), *and Txnrd1:1700028I16Rik* on opposite alleles (**Figure 7F**). Allele information was recovered for several other genes at these loci and no other candidate lncRNA-mRNA interactions were identified at these regions (**Figures 7E-F**). The molecular interplay was evaluated by siRNA mediated knockdown of the lncRNAs and the effects measured on qRTPCR. Primer sets that either covered splice junctions or intron-exon boundaries were applied, and a significant effect was consistently observed for *Gsta4* and *Txnrd1* on both spliced and unspliced RNA levels, thus indicating that the regulatory interplay acts on the transcriptional level (**Figures 7G-H**).

Discussion

The fraction of genomes that harbor transcriptional activity and gives rise to lncRNAs have been efficiently mapped[4], yet it is hard to determine lncRNA functions. Although examples of *cis* regulation by lncRNAs have been reported[42] they exert their effects using several modes of action. Here we sought out to map lncRNA characteristics and functions using scRNA-seq followed by functional experiments.

It has long been accepted that lncRNAs are overall expressed at lower levels than mRNAs but the underlying molecular causes have remained unclear[4]. Using allele-resolved scRNA-seq, we contrasted the transcriptional burst kinetics for several hundreds of lncRNA genes with those of mRNAs. We found that the low expression of lncRNAs is due to lowered burst frequencies (and we found no effect on burst sizes) (**Figures 2C-D**). Thus, lncRNAs are expressed more scattered in cells although at approximately the same RNA copies per burst events. We did not find evidence in primary fibroblasts or long-term cultured cells that there exist rare cells

with very high lncRNA expressions. Interestingly, loci with two genes divergently transcribed from the same promoter regions had higher burst frequencies (**Figures 2G-I**) than genes separated from nearby genes at larger distances. It is possible that the increased burst frequency stems from more promiscuous transcriptional initiation in both directions when the promoter region has been activated by loading of core transcriptional complexes.

We also revisited whether lncRNAs have increased cell-to-cell variability in expression compared to mRNAs of similar expression. Although, lncRNA expression patterns are heterogenous (**Figure 1E**), we observed in both mouse and human model systems that lncRNAs had generally higher cell-to-cell variability (quantified as CV^2) compared to expression-matched mRNAs (**Figures 1E and S2A-B**). It is possible that the larger number of lncRNAs monitored with scRNA-seq explains why earlier reports with lower gene counts did not identify any increased variability of lncRNAS[12]. In line with this argument, subsampling lncRNAs loci from the scRNA-seq data showed declined and eventually loss of power to detect the variability difference between lncRNAs and mRNAs (**Figure 1G**).

Analysis of the scRNA-seq also uncovered *cis*- and *trans*- functions of lncRNAs. We explored an expression-to-function relationship for lncRNAs involved in cell cycle progression as well as apoptosis and showed that scRNA-seq can generate hypothesis of lncRNA-functions without the need for perturbations. Furthermore, knockdown of many lncRNAs did not have a clear phenotype until exposed to a relevant stress stimulus and are therefore likely prone to be missed in large genome-wide perturbation studies. Finally, we also found it of great advantage to overlap gene-gene correlations from scRNA-seq data with differential expression upon

knockdown of lncRNAs (**Figures 4D-E**). We believe this strategy highly reduced the background and exposed the most relevant targets for our studies on *Lockd*.

In summary, our study demonstrates that allele-sensitive scRNA-seq can reveal functional signatures of lncRNAs. We here introduce computational and experimental strategies to link lncRNAs functions acting in *trans* and *cis*. We anticipate that the increasing amount of scRNA-seq data becoming available from different tissues, diseases and species will improve and facilitate functional studies of lncRNAs.

Acknowledgements

We would like to thank Belinda Pannagel, Javier Avila-Carino and the CMB flow-cytometry facility at Karolinska Institutet. This work was supported by grants to R.S. from the Swedish Research Council (2017-01062), the Knut and Alice Wallenberg Foundation (2017.0110) and the Bert L. and N. Kuggie Vallee Foundation.

Author contributions

P.J. designed experiments, sequenced transcriptomes of cells, performed computational analysis, prepared figures and wrote the manuscript. L.H. performed experiments. C.Z. performed experiments and provided support on computational analysis. G.J.H. cultured primary fibroblast cells. M.H.J. provided support on Smart-seq2 and Smart-seq3 protocols. B.R. provided support on computational analysis. R.S. designed experiments, supervised the work and wrote the manuscript.

Material and methods

Cell culture. Derivation of mouse primary fibroblasts were derived from adult CAST/EiJ x C57BL/6J or C57BL/6J x CAST/EiJ mice (approval by the Swedish Board of Agriculture, Jordbruksverket: N343/12) by skinning, mincing and culturing tail explants in DMEM high glucose (Thermo Fischer Scientific), 10% ES cell FBS (ThermoFischer), 1% Penicillin/Streptomycin (Thermo Fischer Scientific), 1% Non-essential amino acids (Thermo Fischer Scientific), 1% Sodium-Pyruvate (Thermo Fischer Scientific), 0.1mM bMercaptoethanol (Sigma) in culture dishes coated with 0.2% gelatin (Sigma). NIH3T3 cells were maintained in DMEM supplemented with 10% FBS (Thermo Fischer Scientific) and 1% Penicillin/Streptomycin (Thermo Fischer Scientific). HEK293FT cells for production of viral particles were maintained in DMEM supplemented with 10% FBS (Thermo Fischer Scientific), 1% Penicillin/Streptomycin (Thermo Fischer Scientific), 1% Non-essential amino acids (Thermo Fischer Scientific), 1% Sodium-Pyruvate (Thermo Fischer Scientific), 0.1mM NEAA (Thermo Fischer Scientific) and 2mM L-Glut (Thermo Fischer Scientific). All cells were cultured in 5% CO_2 at 37°C.

Generation of Smart-seq2 libraries. Cells were either 1) dissociated and diluted to very low cell density in medium and manually picked by mouth pipetting under 10x or 20x magnification using a thin glass capillary in picking volumes of 0.5µl medium or 2) dissociated and distributed by FACS into 96-well plates. Cells for RNA-seq were lysed in 4µl Smart-seq2 lysis buffer (1.9µl 0.4% TritonX-100; 0.1µl Recombinant RNase inhibitor (TaKaRa 40U/µl, 1µl 10µM Smarter oligo-dT, 1µl 10mM dNTPs and 0.1µl ERCC spike-in RNA (Thermo Fischer Scientific, cat no. 4456740) per reaction and snap-frozen on dry ice. Smart-seq2 cDNA libraries were prepared as described earlier[1]. Briefly, cell lysates were incubated for 3 min at 72°C and

then put on ice. Reverse transcription was performed by the addition of 6µl Smart-seq2 RT mix (0.5µl SuperScript II (Thermo Fischer Scientific); 2µl 5x SuperScriptII buffer; 0.5µl 100mM DTT; 2µl 5M betaine; 0.1µl 1mM MgCl2; 0.25µl Recombinant RNase inhibitor (TaKaRa 40U/µl); 0.1µl 100µM LNA strand switch primer; 0.55µl water) and incubation (90 min 42°C; 10 "strand-switch" cycles of (2 min 50°C; 2 min 70°C); 4°C). cDNA amplification was performed by the addition of 15µl of Smart-seq2 PCR mix (12.5µl 2x KAPA HiFi HotStart ReadyMix (KAPA Biosystems cat no. KK2602); 0.25µl 10µM ISPCR primers; 2.25µl water, per reaction) and incubation (3 min 98°C; 18 cycles of (20 s 98°C; 15s 67°C; 6 min 72°C); 5 min 70°C; 4°C). The cDNA was purified using 0.8:1 volume of AMPure XP beads (Beckman Coulter, Ref. A63882) according to the manufacturer's protocol (No. PT5163-1). The purified cDNA was inspected on an Agilent 2100 Bioanalyzer to determine cDNA concentration and size distribution, using Agilent High Sensitivity DNA chips (cat no. 5067-4626). ~1ng of cDNA libraries were tagmented using our in-house-generated Tn5 (1 µL Tn5 (12.5 µM), 2µl 10x TAPS MgCl2 Tagmentation buffer; 5µl 40% PEG8000 in final volume 20µl) and incubated 8 min at 55°C. Tn5 was inactivated and released from the DNA by the addition of 5µl 0.2% SDS and 5 min incubation at room temperature. Sequence library amplification was performed using 5µl Nextera XT Index primers (Illumina, cat no. 15032356) and 15µl PCR mix (1µl KAPA HiFi DNA polymerase (KAPA Biosystems cat no. KK202); 10µl 5x KAPA HiFi buffer; 1.5 µl 10mM dNTPs; 2.5µl water), and incubation (3 min 72°C; 30 s 95°C; 10 cycles of (10 s 95°C; 30 s 55°C; 30 s 72°C); 5 min 72°C; 4°C). DNA sequencing libraries were purified using 1:1 volume of AMPure XP beads (Beckman Coulter, Ref. A63882) according to the manufacturer's protocol (No. PT5163-1, version PR0X3693), inspected on an Agilent 2100 Bioanalyser, using Agilent High Sensitivity DNA chips, and DNA concentration was measured using a Qubit 2.0 Fluorometer (Thermo Fischer Scientific) with the Qubit dsDNA High Sensitivity Assay kit (Molecular Probes, cat no.

Q32854). Pools of samples for multiplexing, with unique Illumina barcode for each cell, were prepared according to the Nextera XT DNA Sample Preparation Guide (No. 15031942 page 46, llumina). Sequencing was carried out on an Illumina HiSeq 2000 generating 43 bp single-end reads.

Processing of RNA-seq data. Subset of primary fibroblasts analyzed in this study are part of previously published studies[2,3]. Here, the zUMIs v2.7.1b pipeline[4] was used for alignment (mouse mm10 assembly), gene quantification (Ensembl gene annotations, GRCm38.91) and allelic calling for primary fibroblasts (non siRNA treated). To pass quality control, cells were required to have **1)** >= 500'000 reads, **2)** 4000 genes expressed at >= 5 read counts **3)** distribution of allelic counts within 0.40 < allelic SNPs < 0.60 on autosomes (imprinted genes and all genes on the X-chromosome excluded, **Supplementary Table 1**) and **4)** no more than 20% of allelic counts mapped to the imprinted X-chromosome (escape genes excluded**,** **Supplementary Table 2**). Genes with at least 5 read counts in 2 cells were assigned as expressed and kept for downstream analysis.

Smart-seq3 libraries of HEK293 cells had previously been generated by *Hagemann-Jensen et al*[5]. The zUMIs v2.7.0a pipeline[4] was used for alignment (human hg38 assembly) and quantification of gene expression (Ensembl gene annotations, GRCh38.95). Cells were required to have; **1)** >= 5e5 mapped read counts / cell **2)** >= 5e5 UMI counts / cell and **3)** > = 7500 genes / cell (>= 1 read count). Genes with at least 1 read count in 3 cells were considered for downstream analysis

Smart-seq2 libraries of mES cells had previously been generated by *Ziegenhain et al*[6]. The zUMIs v2.7.2a pipeline was used for alignment (mouse mm10 assembly) and quantification of gene expression (Ensembl gene annotations, GRCm38.91). Cells were required to have; **1)** >= 4e5 mapped read counts / cell and **2)** > = 8000 genes / cell (>= 5 read counts). Genes with at least 5 read counts in 2 cells were considered for downstream analysis

Annotation of lncRNAs. The Ensembl BioMart annotations (GRCm38.p6) were used for identification of different subsets of lncRNAs. The BioMart annotations were first filtered for genes passing the quality control thresholds (see above) and lncRNAs (according to BioMart annotations) were categorized as; 1) *Divergent* (no gene-gene overlap and TSSs not separated by more than 500 bp), 2) *Convergent* (gene-gene overlap and TSSs not separated by more than 2000 bp, 3) *Intergenic* (no gene-gene overlap and at least 4'000 bp from any other expressed gene), 4) *Independent transcriptional units* (TSSs separated with at least 4'000 bp from any other expressed gene). The threshold of 4,000 bp was established by measuring the mean expression (median of sliding window, size = 51) against the distance between two TSSs.

Permutation test for CV^2. For the analysis of cell-to-cell variability only genes meeting the following criteria were considered; **1)** Not imprinted, **2)** Not encoded on the X-chromosome and **3)** separated transcriptional units (**Figure S1E**).

<u>*CV2:*</u> For each lncRNA meeting the criteria, 10 *independently transcribed* protein-coding genes having the most similar mean expression {min(|mean($RPKM_{lncRNA}$) – mean($RPKM_{mRNA}$|)} were selected. The matching allowed for the same protein-coding gene to be selected multiple times (sample replacement). For the permutation test (n=$1e^4$), one expression-matched protein-coding gene was randomly sampled for each lncRNA and the expected CV^2 (median)

was calculated for each permutation. The same protein-coding gene was not allowed more than once for each permutation (without replacement). The P-value represents the frequency of median($CV2_{sampled}$) > median($CV2_{lncRNA}$).

For estimating the number of lncRNAs needed for detection of median($CV2_{lncRNA}$) > median($CV2_{lncRNA}$), the permutation test was repeat 100 times for each size of sampling and the frequency where 95% of permutations reached median($CV2_{lncRNA}$) > median($CV2_{sampled}$) was assessed.

Transcriptional inference of bursting parameters. Expression values from Smart-seq3 libraries[5] in observed molecules were used for transcriptional inference[2]. Briefly, allele-sensitive read counts were used to distribute molecule counts (UMIs) to the *CAST* or *C57* genome. Cells having UMIs although lacking allelic read counts for individual genes were excluded from the inference while cells lacking UMIs as well as allelic information were considered as 'true' zeros and included in the analysis. For quality control filtering, genes were required to have UMIs distributed to each allele (*CAST* and *C57*) in at least 10 cells. Only non-imprinted autosomal genes were considered for downstream analysis. Only independent transcription units were considered for the analysis of lncRNAs and mRNA.

Identification of cell cycle stage of individual primary fibroblasts. The most variable genes were identified using the R package 'Seurat'[7]. Genes were first filtered for being expressed in at least 5 cells at 5 read counts. Counts were next normalized using 'LogNormalize' (setting scale.factor = 10000) and the most variable genes were identified using the 'vst' method of 'FindVariableFeatures'. We next extracted the cell cycle related genes reported by Whitfield and coworkers[8] (**Supplementary table 3**) , and used the top 50 ranked genes with the highest variability for PCA. The cell cycle phase of individual cells was identified using the first three principal components as input for the R package 'princurve' and the lambda factor used to

align individual cells along the cell cycle progression. Expression of individual genes were illustrated using a rolling mean of 15 cells (using R package 'zoo'). The assignment of cells to cell cycle phase was performed based on the expression levels of known cell cycle regulators (*Gas1*, *Ccne2*, *Ccnb1* and *Ccnd1*) using rolling mean of Seurat normalized counts.

Differential expression analysis of lncRNAs in the cell cycle. Differential expression analysis between cell cycle phases (G0, G1, G1S and G2M) was performed using one-way ANOVA test (Benjamini-Hochberg adjusted, $p<0.01$) with normalized counts (lognormalized, Seurat).

Correlation of cell cycle genes. Genes were first filtered for being expressed in at least 2 cells at 5 read counts for primary tail fibroblasts and GFP+ shControl transduced NIH3T3 cells. The R package 'Seurat' was used to log-normalize the read counts and the normalized counts were used to calculate spearman correlation of cell cycle genes[8]. For each pairwise comparison, cells lacking expression of both genes were excluded from the analysis.

Cell cycle synchronization. NIH3T3 cells (~50% confluent) were washed twice in PBS and treated either with 0.1% FBS, 2mM Thymidine or 800nM Nocodazole for 16-24 hours.

Cell cycle analysis. Cells (including supernatants) were harvested using TrypLE Express (Thermo Fischer Scientific), washed once in PBS, resuspended and fixed in 70% EtOH and stored at -20°C until analysis on flow cytometry. Prior analysis, cells were washed once in PBS followed by resuspending the samples in 500 µL staining buffer (PBS containing 40 µg/ml propidium iodide, 100 µg/ml RnaseA, 0.1% Triton X-100), incubated on ice for ~1hr and finally analyzed on flow cytometry.

qRTPCR. Total RNA was extracted using the Qiagen RNeasy mini kit (Qiagen cat no. 74106) followed by DNase treatment using the Ambion DNA-free DNA removal kit (Thermo Fisher Scientific, cat no. AM1906). Equal amounts of DNase treated RNA was used for preparing cDNA using SuperScript II (Thermo Fisher Scientific) and oligo$(dT)_{18}$ primer according to the manufacturer's recommendations. Quantification was carried out using Power SYBR Green (Thermo Fisher Scientific) on a StepOnePlus or ViiA7 Real-Time PCR system (Applied biosystems). The delta-delta Ct method was used for quantification of relative expression levels.

Cloning / Generation of lentiviral U6 expressed shRNAs. Cloning was carried out as previously described[9]. Briefly, oligos were treated with T4 PNK (New England Biolabs), annealed and ligated into the Nhe1 and Pac1 restriction sites of the pHIV7-IMPDH2-U6 construct[9].

Lentiviral production. Lentiviral particles were produced as previously described[9,10]. Shortly, HEK293FT cells were transfected with pCHGP-2, pCVM-G pCMV-rev and pHIV7-IMPDH2-U6 at 1:0.5:0.25:1.5 ratio using SuperScript II and PLUS reagent (Thermo Fisher Scientific) in serum depleted DMEM media. Media was changed ~6 post transfection to DMEM collection media (10% FBS (Thermo Fischer Scientific), 1% Penicillin/Streptomycin (Thermo Fischer Scientific), 1% Non-essential amino acids (Thermo Fischer Scientific), 1% Sodium-Pyruvate (Thermo Fischer Scientific), 0.1mM NEAA (Thermo Fischer Scientific) and 2mM L-Glut (Thermo Fischer Scientific), 0.37% Sodium Bicarbonate (Thermo Fischer Scientific) and 1x Viral Boost Reagent (Alstem Cell advancements). The viral supernatant was collected ~48 hours post transfection,

passed through a 0.45 µm filter (Sarstedt) and concentrated using PEG-it (SBI System Biosciences) according to the manufacturer's recommendations.

Generation of lentiviral transduced stable cell lines. NIH3T3 cells were transduced using low titer of lentiviral particles (< 10% of transduced cells). GFP+ cells were sorted at the CMB core facility, Karolinska Institutet.

Colony formation assay. For stable NIH3T3 cell lines cells were seeded at low density (500 cells / well) in 6 well plates. Media was changed every second to third day. After 10-14 days, cells were washed twice in PBS, fixed and stained for 20 minutes with 0.5% crystal violet, washed in H2O and finally let dry. Quantification of was carried out by manually counting the colonies and also by measuring the crystal violet absorbance at 590nm by re-solubilize the stained cells in 10% acetic acid solution.

For siRNAs, NIH3T3 cells were seeded at 1000-2,500 cells / well in 6 well plates. Transfections were carried out 24 hours after seeding and the procedure as described above repeated.

siRNA knockdown and MMC treatment. NIH3T3 and primary cells were transfected using Lipofectamine RNAiMAX Transfection Reagent (Thermo Fischer Scientific) according to the manufacturer’s protocol. A final concentration of 10nM siRNA was used. MMC treatment was initiated 24 hours post transfections.

Allelic imbalance. Allelic imbalance was calculated as ($CAST_{allelicCounts}$ /($CAST_{allelicCounts}$ + $C57_{allelicCounts}$) - 0.5). A pairwise gene comparison of lncRNA-mRNAs was applied within a

genomic region of +/- 500 kb of the lncRNA-TSS. lncRNA-mRNA gene pairs were considered as candidates if both genes presented allelic bias > 0.15 and interactions ranked based on their total sum of allelic imbalance.

References

1. Carninci, P. *et al.* Molecular biology: The transcriptional landscape of the mammalian genome. *Science (80-.).* (2005) doi:10.1126/science.1112014.
2. Johnsson, P., Lipovich, L., Grandér, D. & Morris, K. V. Evolutionary conservation of long non-coding RNAs; Sequence, structure, function. *Biochimica et Biophysica Acta - General Subjects* (2014) doi:10.1016/j.bbagen.2013.10.035.
3. de Hoon, M., Shin, J. W. & Carninci, P. Paradigm shifts in genomics through the FANTOM projects. *Mammalian Genome* (2015) doi:10.1007/s00335-015-9593-8.
4. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* (2012) doi:10.1038/nature11233.
5. Mattioli, K. *et al.* High-throughput functional analysis of lncRNA core promoters elucidates rules governing tissue specificity. *Genome Res.* (2019) doi:10.1101/gr.242222.118.
6. Cabili, M. *et al.* Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.* (2011) doi:10.1101/gad.17446611.
7. Hon, C. C. *et al.* An atlas of human long non-coding RNAs with accurate 5' ends. *Nature* (2017) doi:10.1038/nature21374.
8. Derrien, T. *et al.* The GENCODE v7 catalog of human long noncoding RNAs. *Genome Res.* (2012) doi:10.1101/gr.132159.111.

9. Ramsköld, D., Wang, E. T., Burge, C. B. & Sandberg, R. An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS Comput. Biol.* (2009) doi:10.1371/journal.pcbi.1000598.

10. Wang, K. C. *et al.* A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. *Nature* (2011) doi:10.1038/nature09819.

11. Mercer, T. R., Dinger, M. E., Sunkin, S. M., Mehler, M. F. & Mattick, J. S. Specific expression of long noncoding RNAs in the mouse brain. *Proc. Natl. Acad. Sci.* (2008) doi:10.1073/pnas.0706729105.

12. Cabili, M. N. *et al.* Localization and abundance analysis of human lncRNAs at single-cell and single-molecule resolution. *Genome Biol.* (2015) doi:10.1186/s13059-015-0586-4.

13. Sandberg, R. Entering the era of single-cell transcriptomics in biology and medicine. *Nature Methods* (2014) doi:10.1038/nmeth.2764.

14. Deng, Q., Ramsköld, D., Reinius, B. & Sandberg, R. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science (80-.).* (2014) doi:10.1126/science.1245316.

15. Zeisel, A. *et al.* Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science (80-.).* (2015) doi:10.1126/science.aaa1934.

16. La Manno, G. *et al.* RNA velocity of single cells. *Nature* (2018) doi:10.1038/s41586-018-0414-6.

17. Larsson, A. J. M. *et al.* Genomic encoding of transcriptional burst kinetics. *Nature* (2019) doi:10.1038/s41586-018-0836-1.

18. Nicolas, D., Phillips, N. E. & Naef, F. What shapes eukaryotic transcriptional bursting? *Molecular BioSystems* (2017) doi:10.1039/c7mb00154a.

19. Picelli, S. *et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.*

(2014) doi:10.1038/nprot.2014.006.

20. Picelli, S. *et al.* Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* (2013) doi:10.1038/nmeth.2639.

21. Ziegenhain, C. *et al.* Comparative Analysis of Single-Cell RNA Sequencing Methods. *Mol. Cell* (2017) doi:10.1016/j.molcel.2017.01.023.

22. Reinius, B. *et al.* Analysis of allelic expression patterns in clonal somatic cells by single-cell RNA-seq. *Nat. Genet.* (2016) doi:10.1038/ng.3678.

23. Ramsköld, D. *et al.* Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.* (2012) doi:10.1038/nbt.2282.

24. Hagemann-Jensen, M. *et al.* Single-cell RNA counting at allele- and isoform-resolution using Smart-seq3. *bioRxiv* (2019) doi:10.1101/817924.

25. Fukaya, T., Lim, B. & Levine, M. Enhancer Control of Transcriptional Bursting. *Cell* (2016) doi:10.1016/j.cell.2016.05.025.

26. Bartman, C. R., Hsu, S. C., Hsiung, C. C. S., Raj, A. & Blobel, G. A. Enhancer Regulation of Transcriptional Bursting Parameters Revealed by Forced Chromatin Looping. *Mol. Cell* (2016) doi:10.1016/j.molcel.2016.03.007.

27. Walters, M. C. *et al.* Enhancers increase the probability but not the level of gene expression. *Proc. Natl. Acad. Sci. U. S. A.* (1995) doi:10.1073/pnas.92.15.7125.

28. Katayama, S. *et al.* Antisense transcription in the mammalian transcriptome. *Science (80-.).* (2005).

29. Grinchuk, O. V., Jenjaroenpun, P., Orlov, Y. L., Zhou, J. & Kuznetsov, V. A. Integrative analysis of the human cis-antisense gene pairs, miRNAs and their transcription regulation patterns. *Nucleic Acids Res.* (2009) doi:10.1093/nar/gkp954.

30. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell

transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* (2018) doi:10.1038/nbt.4096.

31. Whitfield, M. L. Identification of Genes Periodically Expressed in the Human Cell Cycle and Their Expression in Tumors. *Mol. Biol. Cell* (2002) doi:10.1091/mbc.02-02-0030.

32. Hastie, T. & Stuetzle, W. Principal curves. *J. Am. Stat. Assoc.* (1989) doi:10.1080/01621459.1989.10478797.

33. Paralkar, V. R. *et al.* Unlinking an lncRNA from Its Associated cis Element. *Mol. Cell* (2016) doi:10.1016/j.molcel.2016.02.029.

34. Kharchenko, P. V., Silberstein, L. & Scadden, D. T. Bayesian approach to single-cell differential expression analysis. *Nat. Methods* (2014) doi:10.1038/nmeth.2967.

35. Pippa, R. *et al.* P27 Kip1 represses transcription by direct interaction with p130/E2F4 at the promoters of target genes. *Oncogene* (2012) doi:10.1038/onc.2011.582.

36. Xu, H. *et al.* Silencing of KIF14 interferes with cell cycle progression and cytokinesis by blocking the p27Kip1 ubiquitination pathway in hepatocellular carcinoma. *Exp. Mol. Med.* (2014) doi:10.1038/emm.2014.23.

37. Montes, M. *et al.* The lncRNA MIR31HG regulates p16 INK4A expression to modulate senescence. *Nat. Commun.* (2015) doi:10.1038/ncomms7967.

38. Mullin, N. K. *et al.* Wnt/β-catenin signaling pathway regulates specific lncRNAs that impact dermal fibroblasts and skin fibrosis. *Front. Genet.* (2017) doi:10.3389/fgene.2017.00183.

39. Linardopoulos, S. *et al.* Deletion and Altered Regulation of pl6INK4a and pl5INK4b in Undifferentiated Mouse Skin Tumors. *Cancer Res.* (1995).

40. Crowley, J. J. *et al.* Analyses of allele-specific gene expression in highly divergent mouse crosses identifies pervasive allelic imbalance. *Nat. Genet.* (2015)

doi:10.1038/ng.3222.

41. Eckersley-Maslin, M. A. *et al.* Random monoallelic gene expression increases upon embryonic stem cell differentiation. *Dev. Cell* (2014) doi:10.1016/j.devcel.2014.01.017.

42. Gil, N. & Ulitsky, I. Regulation of gene expression by cis-acting long non-coding RNAs. *Nature Reviews Genetics* (2020) doi:10.1038/s41576-019-0184-5.

Figure 1

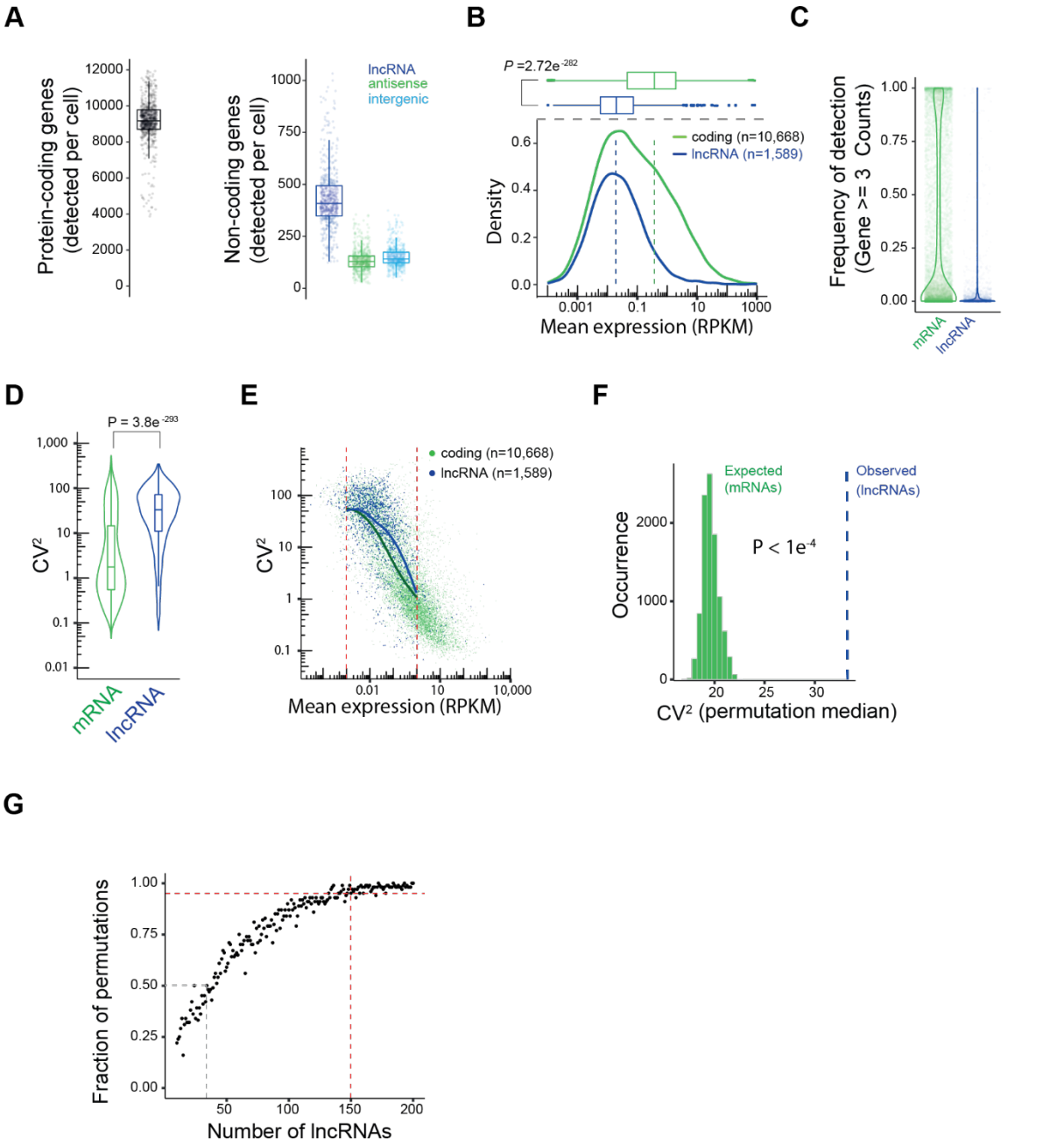

Figure 1 | Levels and variability of lncRNA and mRNA expressions

(A) Boxplots showing the detected numbers of protein-coding genes and subtypes of lncRNAs per fibroblast, based onSmart-Seq2 data (n=533 cells) and requiring 3 or more read counts for detection. **(B)** Densities and boxplots of mean lncRNAs and mRNAs expression levels across fibroblasts (n=533). Dashed lines denote the medians, p-value represents a two-sided

Wilcoxon rank-sum test. **(C)** Violin plots showing the fraction of cells that detected individual lncRNAs and mRNAs (requiring 3 or more read counts for detection). **(D)** Violin plots showing the coefficient of variation (CV^2) for lncRNAs and mRNAs expressions across fibroblast (n=533). P-value represents two-sided Wilcoxon rank-sum test**.** **(E)** Scatter plot of mean expression against the CV^2 for lncRNAs (blue) and mRNAs (green). Lines denote a smoothed fit to the rolling mean (width = 15) for lncRNAs and mRNAs, respectively. Red dotted lines denote the expression range for the smoothed. **(F)** Histogram showing median CV^2 for sampled expression-matched sets of mRNAs. Each lncRNA was matched with 10 mRNAs of similar expression followed by subsampling one expression-matched mRNA for each lncRNA, this procedure was repeated 10,000 times. The P-value represent the outcome of the permutation test, the observed CV^2 was higher than the maximum CV^2 obtained from 10,000 permutations. **(G)** Scatter plot showing the numbers of lncRNAs required to identify increased CV^2 of lncRNAs compared to expression-matched mRNAs. The X-axis represents the number of lncRNAs used for analysis. For each increment in the number of lncRNAs analysed, expression-matched mRNAs were randomly selected (this process iterated 1,000 times). The y-axis shows the fraction of sampled lncRNAs and expression-matched mRNA sets for which at least 95% of permutations reach higher CV^2 for lncRNAs than mRNAs. The grey dashed line shows the result when sampling 34 lncRNAs.

Figure 2

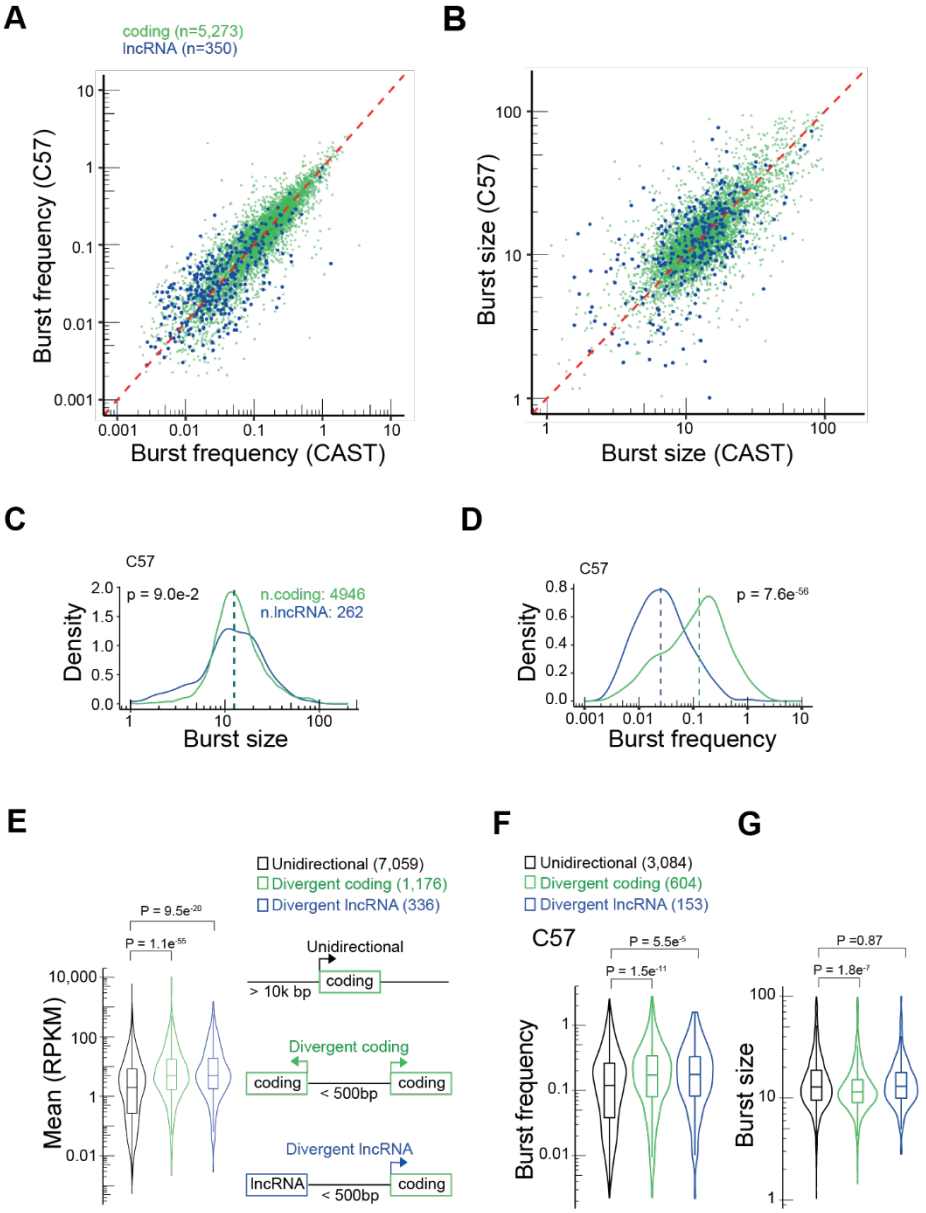

Figure 2 | Transcriptional burst kinetics of lncRNAs and divergent transcriptional loci

(A-B) Scatter plots of (A) burst frequencies and (B) burst sizes for mRNAs (green) and lncRNAs (blue) comparing the parameters inferred from the *CAST* allele against the *C57* allele**.** Red line denotes *x=y*. **(C-D)** Density plots for (C) burst sizes and (D) burst frequencies for mRNAs and lncRNAs (showing the *C57* alleles). Dashed lines represent the median burst sizes and frequencies for mRNAs and lncRNAs. The p-value represents a two-sided Wilcoxon rank-sum test. **(E)** Violin plot showing mean expression levels (in RPKM) of unidirectional mRNAs and

for mRNAs transcribed from divergent promoters (either with another mRNA or a lncRNA). The p-values computed as two-sided Wilcoxon rank-sum tests. **(F-G)** Violin plot as in (E) showing instead (F) burst frequencies or (G) burst sizes for unidirectional mRNAs or divergent gene pairs. Inference shown for C57 alleles, p-values represent a two-sided Wilcoxon rank-sum test.

Figure 3

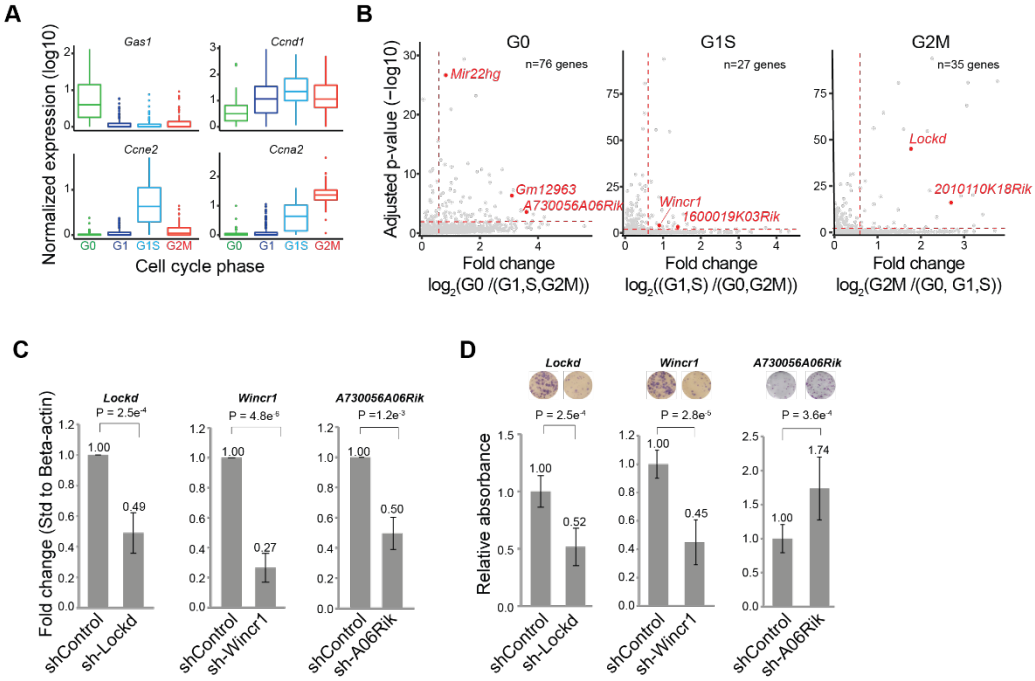

Figure 3 | Identification of cell cycle regulated lncRNAs using scRNA-seq

(A) Boxplots showing the expression levels (RPKMs) of cell cycle marker genes in cells classified to cell cycle phase. **(B)** Scatter plots showing lncRNAs with significant expression differences across cell cycle phases (y-axis, Benjamini-Hochberg adjusted ANOVA test) against the fold induction in cell cycle phase with highest expression against the other cell cycle phases. The top ranked candidates were colored red and selected for further validation. **(C)** Relative expression levels of candidate lncRNAs in lentiviral transduced NIH3T3 cells measured by qRTPCR (n =4, p-values represent a two-sided student's t-test) and in control NIH3T3 cells. **(D)** Quantification of colony forming cells in control cells and cells with stable shRNA-induced knockdown of lncRNAs, together with representative photos of staining (n=4, p-values represent a two-sided student's t-test).

Figure 4

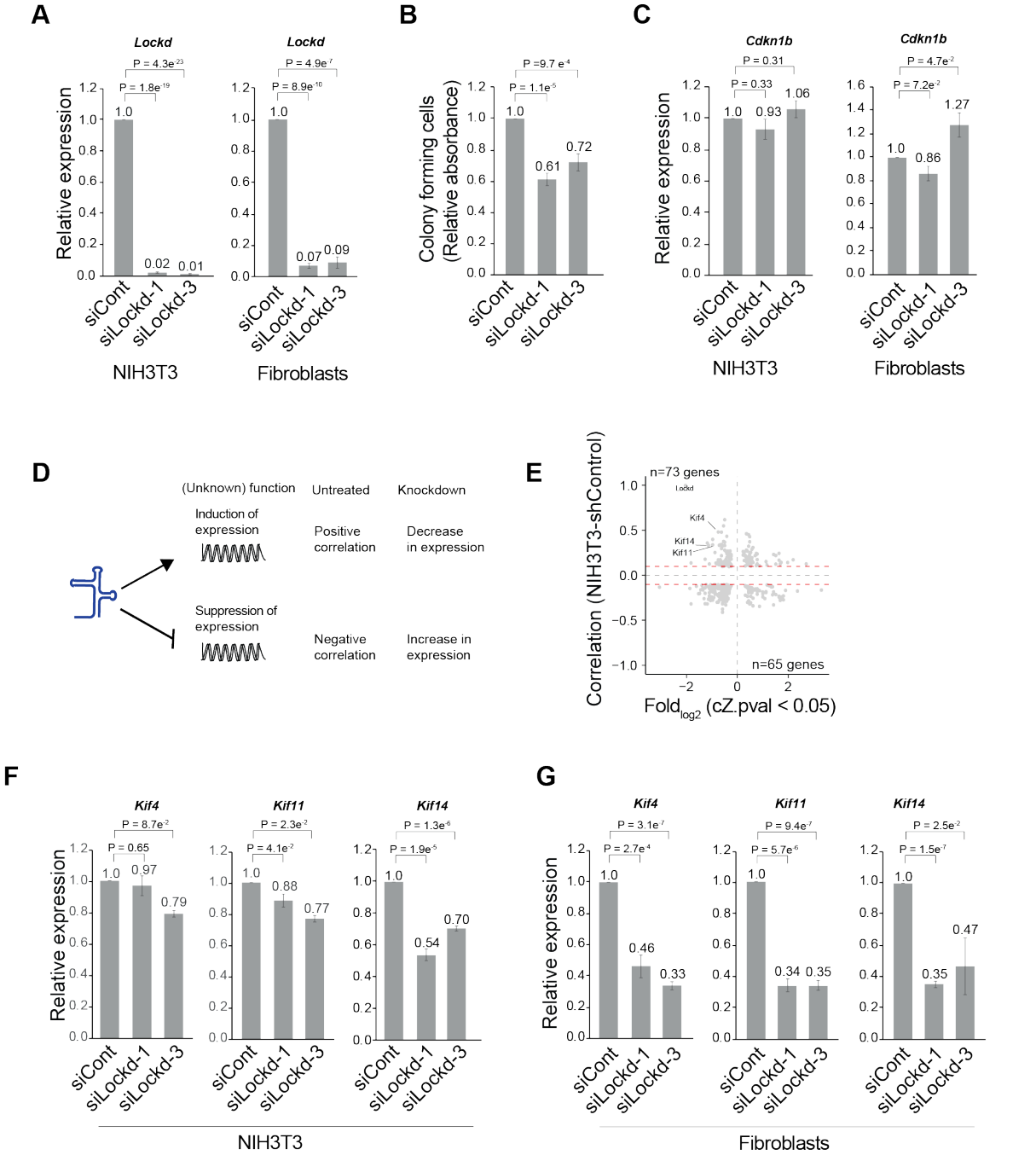

Figure 4 | Functional analysis of the proliferative phenotype after *Lockd* depletion.

A) Relative expression of *Lockd* upon siRNA induced knockdown in NIH3T3 cells (n=4) and primary fibroblasts (n=4) measured by qRTPCR, and in control cells. P-values represent a two-sided student's t-test, error bars represent the s.e.m. **(B)** Quantification of colony forming cells upon siRNA induced knockdown in NIH3T3 cells (n=5, p-values represent a two-sided student's

t-test, error bars represent the s.e.m.), and in control cells. **(C)** Relative expression of *Cdkn1b* upon siRNA induced knockdown of *Lockd* in NIH3T3 cells (n=5) and primary fibroblasts (n=4) measured by qRTPCR. P-values represent a two-sided student's t-test, error bars represent the s.e.m. **(D)** Schematic representation for the intersection of fold change (shLockd / shControl) against gene-gene correlations in shControl treated cells**.** **(E)** Scatter plot representing significant fold changes of gene expression (shLockd / shControl, adjusted p-value < 0.05) of stably transduced NIH3T3 cells (x-axis) against gene-gene correlations to *Lockd* in stably shControl-transduced NIH3T3 cells (y-axis). Genes reaching the threshold of Spearman correlations (denoted with red dashed lines) were considered for downstream analysis. **(F)** Relative expression of candidate genes upon siRNA induced knockdown of *Lockd* in NIH3T3 cells measured by qRTPCR (n=4, p-values represent a two-sided student's t-test, error bars represent the s.e.m.). **(G)** Relative expression of candidate genes upon siRNA induced knockdown of *Lockd* in primary fibroblast cells measured by qRTPCR (n=4, p-values represent a two-sided student's t-test, error bars represent the s.e.m.).

Figure 5

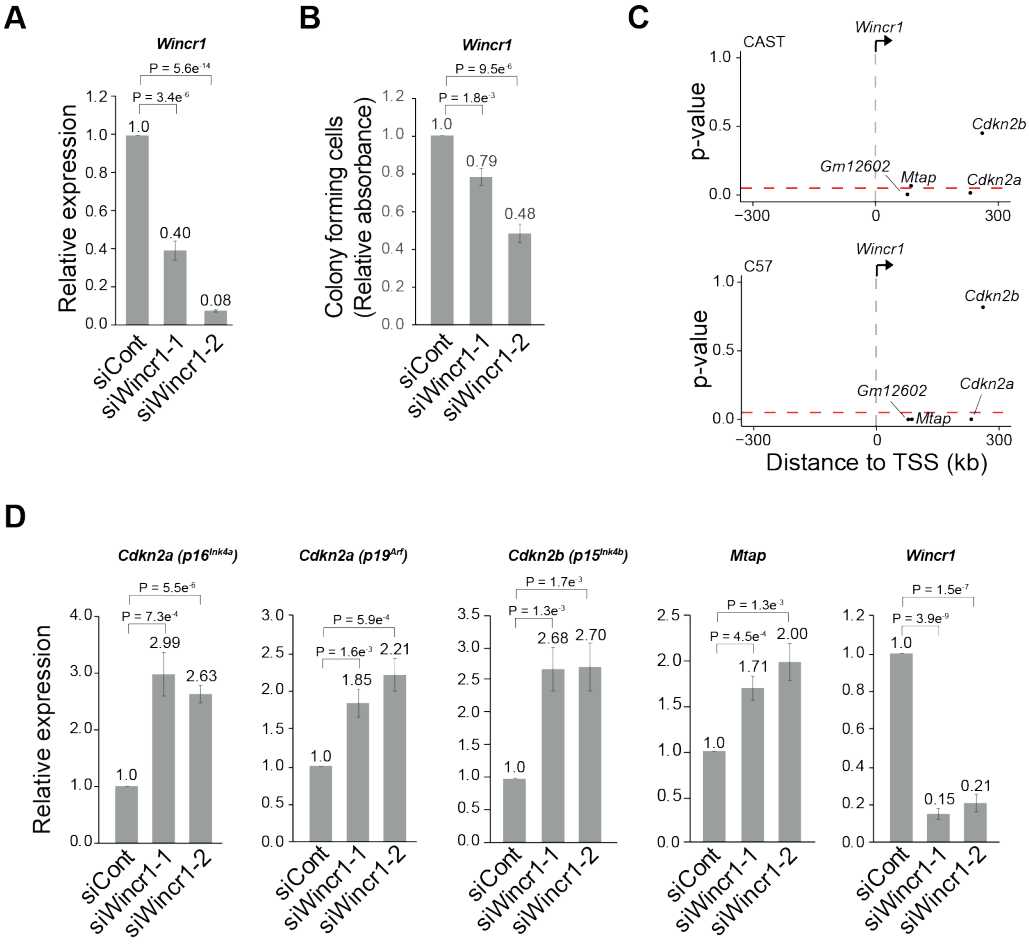

Figure 5 | Functional investigation of *Wincr1* depletion.

(A) Relative expression of *Wincr1* upon siRNA induced knockdown in NIH3T3 cells measured by qRTPCR (n=5, p-values represent a two-sided student's t-test, error bars represent the s.e.m.), and in control cells. **(B)** Quantification of colony forming cells upon siRNA induced knockdown in NIH3T3 cells (n=5, p-values represent a two-sided student's t-test, error bars represent the s.e.m.). **(C)** Scatter plots showing p-values for co-expression of *Wincr1* (Fisher's exact test, gene considered expressed if >= 3 allelic counts) against other genes in proximity (+/- 300 bp of TSS-*Wincr1*) for the CAST and C57 genomes, respectively. The red dashed lines denote threshold for significance (p<0.05). **(D)** Relative expression of candidate *cis*-interacting genes upon siRNA induced knockdown of *Wincr1* in primary fibroblast cells measured by qRTPCR (n=5, p-values represent a two-sided student's t-test, error bars represent the s.e.m.).

Figure 6

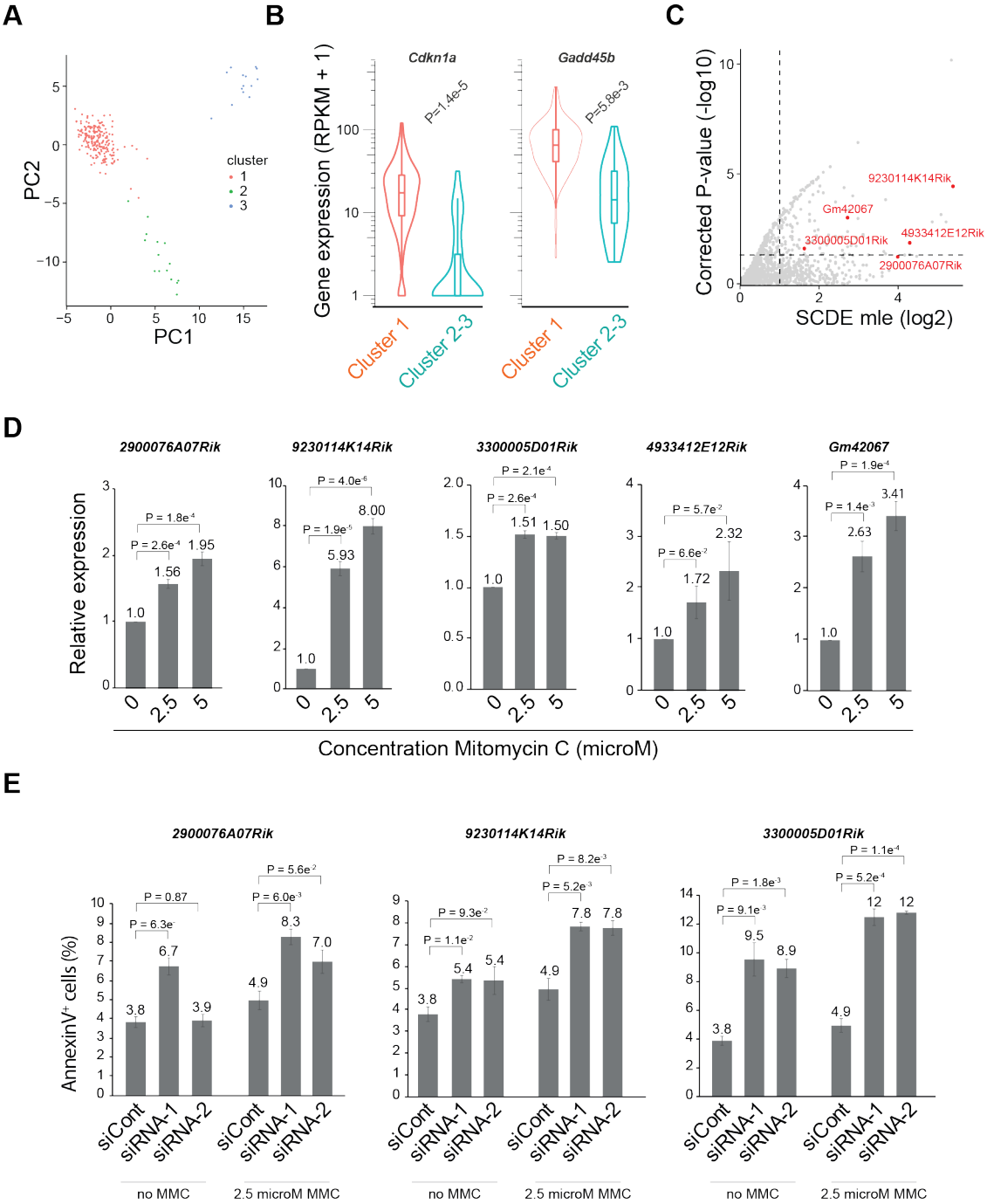

Figure 6 | Identification of lncRNAs involved in apoptosis by single-cell profiling.

(A) Low dimensional PCA projection of cells, based on the most variable genes annotated as apoptosis related. Cells are colored according to clusters (identified by the pam function of the R package 'cluster'). **(B)** Violin plots showing the expression levels of two marker genes for DNA damage (P-values represent a two-sided Wilcoxon rank-sum test). **(C)** Scatter plot

showing fold change magnitudes (x-axis) and significance levels (y-axis) for cluster 1 against clusters 2-3 identified in (A) analyzed by SCDE[1]. lncRNAs selected for validation are marked in red. **(D)** Relative expression measured by qRTPCR of candidate lncRNAs in NIH3T3 cells treated with MMC (n=4, p-values represents a two-sided student's t-test, error bars represent the s.e.m.). **(E)** Quantification of apoptosis using AnnexinV for siRNA targeted NIH3T3 cells treated with MMC (n=3, p-values represents a two-sided student's t-test, error bars represent the s.e.m.).

Figure 7

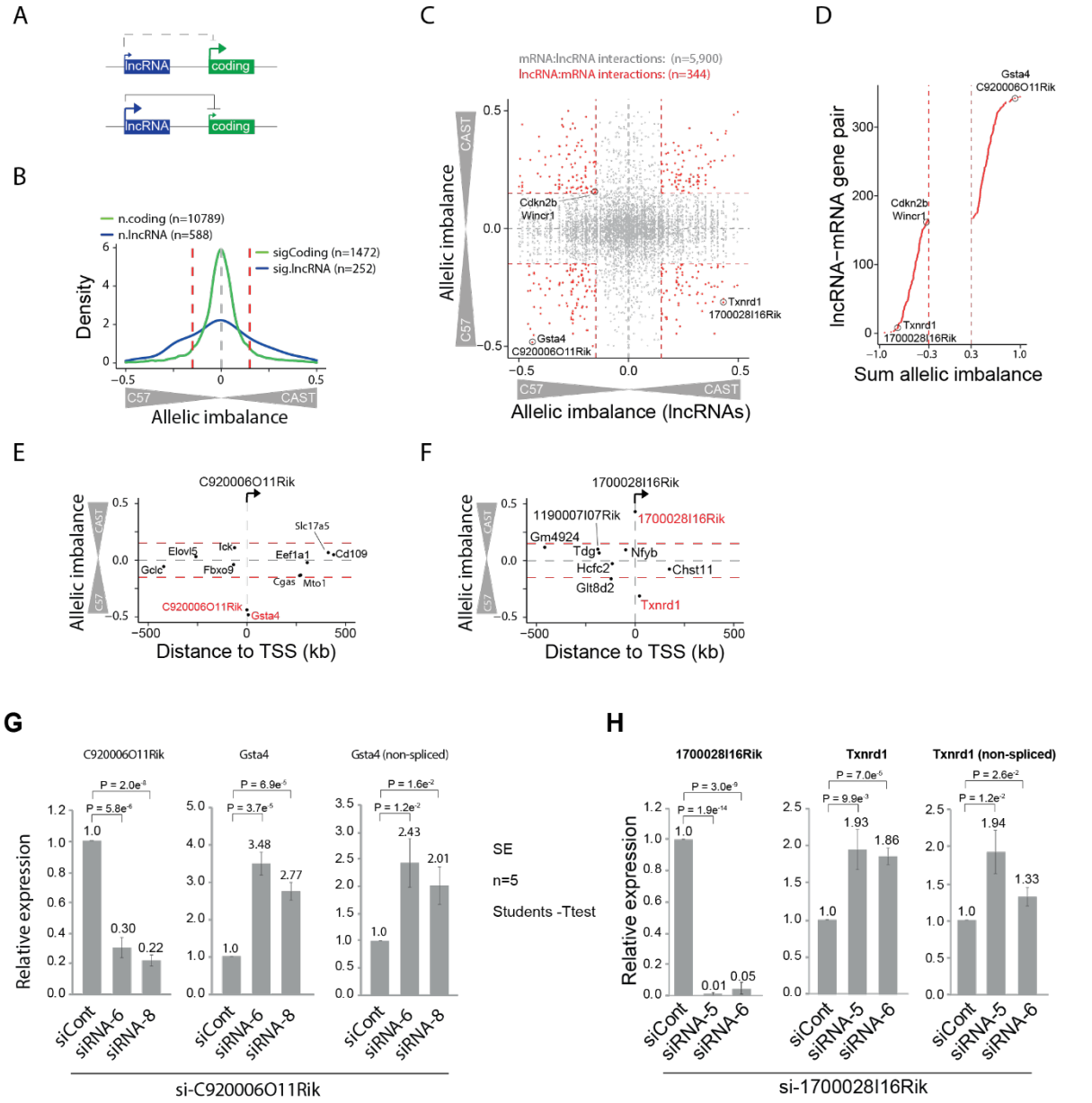

Figure 7 | lncRNAs functioning in *cis* identified by their coordinated allelic imbalances.

(A) Schematic illustration of hypothetical *cis*-effects between lncRNAs and proximal mRNAs that could manifest as coordinated allelic expression dynamics across cells. **(B)** Density plot summarizing observed allelic imbalance of lncRNAs (blue) and mRNAs (green) across fibroblasts (n=751). **(C)** Scatter plot illustrating pairwise analysis of allelic imbalance of

lncRNAs (X-axis) and mRNAs (Y-axis) within a genomic locus of 500kb of lncRNA-TSS. **(D)** Scatter plot representing candidate lncRNA-mRNA gene pairs with allelic bias > 0.15. A positive value represents allelic bias towards the same allele while a negative value represents bias on opposite alleles. **(E)** Scatter plot representing allelic bias of genes within +/-500kb of the TSS-lncRNA *C920006O11Rik*. Candidate genes are highlighted in red. **(F)** Scatter plot representing allelic bias of genes within +/-500kb of the TSS-lncRNA *1700028I16Rik*. Candidate genes are highlighted in red. **(G)** Relative expression of spliced and unspliced levels of *Gsta4* upon siRNA induced knockdown of *C920006O11Rik* measured by qRTPCR (n=5, P-values represent a two-sided student's t-test, error bars represent the s.e.m.) **(H)** Relative expression of spliced and unspliced levels of *Txnrd1* upon siRNA induced knockdown of *1700028I16Rik* measured by qRTPCR (n=5, P-values represent a two-sided student's t-test, error bars represent the s.e.m.).