iScience, Volume *23*

Supplemental Information

The Contribution of lincRNAs at the Interface between Cell Cycle Regulation and Cell State Maintenance

Adriano Biasini, Adam Alexander Thil Smith, Baroj Abdulkarim, Maria Ferreira da Silva, Jennifer Yihong Tan, and Ana Claudia Marques

SUPPLEMENTARY INFORMATION

TRANSPARENT METHODS

Processing of cell cycle staged single cell RNA sequencing data

Mouse genome sequence and annotation files were downloaded from ENSEMBL (GRCm38.82 i.e. mm10). LincRNAs expressed in mESCs (Tan et al., 2014) were added to the transcript annotations using custom Python scripts in order to appropriately deal with novel transcripts overlapping existing loci, leading to a total of 55,596 loci. The FASTA sequences of the 96 ERCC spike-ins used in the experiment were downloaded from UCSC GoldenPath. We used CGAT ((Heger et al., 2013), v 0.2.3) to extract transcript sequences longer than 200 nucleotides, and Kallisto ((Bray et al., 2016), v 042.4, default parameters) to build a transcriptome index that included 114,842 transcripts across 49,285 loci and ERCCs. The FASTQ files for staged mESC single-cell RNA-sequencing were downloaded from ArrayExpress (accession number E-MTAB-2805). For each cell, we estimated transcript expression using Kallisto (v. 0.42.4) (Bray et al., 2016) without bootstraps. Hit counts were imported into R and summarized to the gene level using tximport (Soneson et al., 2015). After removing genes without any hits in any cells, only 38,016 genes remained.

Basic cell-level quality control was performed as previously described (Buettner et al., 2015). Briefly, we excluded cells: 1) with a cell-wise gene detection rate below 20%; 2) cells with <1.6M total hits; 3) cells with proportions of hits to ERCCs lower than 15% or higher than 65%; 4) cells with a proportion of hits to mitochondrial genes lower than 1% or higher than 10%; and 5) cells with an estimated cell size higher or lower than 2 median absolute deviations from the stage-median cell size. In total, after these filtering steps, 94/95 G1 cells, 76/88 S cells, and 74/96 G2/M cells were retained for further analysis. We then performed gene-level quality control. We considered only genes that had at least one hit count in at least 4% of samples. Hit counts were normalized using the scran package (v1.0.4, (Lun et al., 2016)). Size factors were calculated using pools of samples of sizes 10 to 35 cells from each stage. We used scLVM (Brennecke et al., 2013, Buettner et al., 2015) to filter out genes whose total variance is not significantly higher than that expected from pure technical variability (fitTechnicalNoise

parameters: mincv2=0.01, quan=0.10 ; getVariableGenes parameters: threshold = 0.10, minBiolDisp = 0.30), and subsequently also removed ERCCs & mitochondrial genes. After this final filtering step, 10,487 genes were kept.
Except for differential expression analysis, where scran-normalised counts were used, all subsequent analyses were performed on scran-normalised shifted log10 counts (sl10 = log10(normalised counts + 1)).

Processing of bulk RNA sequencing for a time-course of mESC to neurons differentiation

We used STAR (Dobin et al., 2013) (v 2.5.3a) to map bulk RNA sequencing data for a 216-hour time course of mESC differentiation into neurons (Sun et al., 2017) to the Mouse transcriptome. Transcript quantification was performed using RSEM (Li and Dewey, 2011) (v 1.3.0) and for each time point we considered the average of expression across time points. We estimated the Spearman correlation between genes of interest and pluripotency associated genes (Xu et al., 2013) using R.

Differential expression analysis

Kallisto pseudo-count data was imported into R (v 3.3.1) and prepared for DESeq2 (v 1.12.4) (Love et al., 2014) using the *tximport* (Soneson et al., 2015) package (v 1.0.3). Principal Components Analysis (PCA) was applied using R package FactoMineR (Le et al., 2008) to an expression matrix consisting of all cells passing QC in rows, and all genes passing QC in columns. The cell cycle signal typically appeared most strongly across the first two principal components, with the third component correlating strongly with the gene dropout rate (not shown). Differentially expressed genes (FDR< 5%, log2 fold change > 0.1) were called between pairs of cell cycle stages using DESeq2 (Love et al., 2014, Jaakkola et al., 2017) using appropriate contrasts.
In order to assess whether the overall lower expression of lincRNAs affected our ability to call them as DE, we performed a down-sampling experiment to bring mRNA expression down to levels comparable to lincRNAs. We first estimated the average scaling factor between mRNA and lincRNA expression in the original data. We estimated the difference between the median Transcripts Per Million reads (TPM) in quality-controlled cells between lincRNAs and mRNAs to be 14.196, i.e. on average

mRNAs are 14x more highly expressed than lincRNAs. We randomly sub-sampled the original FASTQ files to 1/14th of their original size (i.e. using a factor of 0.0714) using seqtk (v 2015.10.15). We quantified transcript expression using Kallisto and filtered cells and genes as described above prior to differential gene expression analysis. For the sub-sampled data, the number of tested genes was similar to the full data (9 923 vs 10 391). We then counted the number of mRNA genes called as DE in the full analysis that passed the quality control filters for the sub-sampled data (388), and the number of those called as DE in the sub-sampled analysis (233, i.e. 60.1%).

Co-expression analysis.

We considered a gene to be a cell cycle regulator if: 1) was annotated as a cell cycle gene (Buettner et al., 2015); 2) it was annotated with a GO term containing the string "cell cycle" or 3) had a Human ortholog annotated in CycleBase (Santos et al., 2015). Pluripotency-associated genes were extracted from the ESCAPE database (Xu et al., 2013).

Co-expression analysis with cell cycle genes was performed using the single cell RNA sequencing data with staged mESCs (Buettner et al., 2015). We considered all loci expressed (normalized shifted log10>0.14) in at least 50 cells (out of a possible 244). The expression cut-off was defined based on inspection of the shifted log10 expression values. We estimated the Spearman correlation between all these loci and considered a gene pair to be significantly correlated if the absolute Spearman correlation between their expression was > 0.10 and the correlation test p-value was < 0.10. The correlation between any given pair was only calculated in cells where both genes were expressed. Correlations for gene pairs where there were less than 20 such value pairs were set to 0.

Co-expression analysis with pluripotency genes was performed using the time course of mESC to neurons differentiation data, as described above.

Enrichment analyses

Enrichment analyses for Gene Ontology Biological Process, KEGG and REAC terms in *Mus musculus* were carried out using R package "gProfileR" (v. 0.6.1) (Reimand et al., 2016). Terms were considered to be significantly enriched if the associated

corrected p-value was below 10% relative to the specified background of mESC expressed genes.

Enrichment in binding of mESC transcription factors (TFs) (Chen et al., 2008) at CC-lincRNA promoters (defined as +/-1 kb from their annotated TSS) was estimated using the Genome Association Tester (GAT) (Heger et al., 2013). GAT tests whether the enrichment of TF binding at CC-lincRNA promoters is different from what would be expected based on 10,000 random samplings (with replacement) of intergenic segments with the same length and GC content as the CC-lincRNA promoters. The gat-compare tool was used to test the significance of the enrichment of TF binding at CC-lincRNA promoters relative to that at non CC-lincRNA promoters. Core TF binding sites were obtained from (Chen et al., 2008).

Loci tissue specificity metrics in Mouse

We manually selected 150 Mouse paired-end total RNA bulk RNA-sequencing datasets from the ENCODE project (Consortium, 2012), covering a range of tissues, sexes and ages, and added 3 publicly-available Mouse embryonic stem cell datasets (Tan et al., 2015). We defined a “tissue.simple” grouping of samples, based on a high-level description of the tissue of origin (e.g. central nervous system, or heart) and of the developmental stage (adult or embryonic). Gene expression was estimated using Kallisto and imported into R using tximport as described above. Library size normalization was performed with DESeq2, and counts were transformed into shifted log10 normalised values. Values below a background level of 0.1 were set to 0. This cut-off was obtained based on the distribution of shifted log10 expression values. Genes with no expression values above 0 in any sample were discarded. Expression values were averaged across technical and biological replicates using the median. Tau was calculated as described in (Kryuchkova-Mostacci and Robinson-Rechavi, 2017), briefly as Tau = sum(1-(v/max(v)))/(length(v)-1), where v is a vector of average expression values for different tissues.

Cell Culture

ES-E14-Tg2a (E14) cells were grown on 0.1% gelatin-coated tissue culture dishes, in DMEM (Thermo Fischer, 41965-039) supplemented with 1x Non-Essential Amino

Acids (Thermo Fischer 11140-035), 50 uM β-mercaptoethanol (Thermo Fischer 31350-10), 15% Fetal Bovine Serum (Thermo Fischer 10499-044), 500 Units/ml of Pennicilin/Streptomycin (Thermo Fischer, 15140122), and 100 Units/ml of Recombinant mouse LIF Protein (Merck ESG1107). Culture were seeded at an average density of ~ $3.8*10^4$ cells/cm^2 and passaged every 48 h.

Analysis of cell cycle stage and subcellular gene expression

We resuspended 10^6 cells in 1 ml of PBS, containing 1 uL/mL of fluorescent reactive dye (LIVE/DEAD) and incubated in the dark for 30 minutes at room temperature. Cells were then washed with PBS, spun, and resuspended in fresh medium. We add 2 uL of DNA Hoechst 33324 dye (20 mM) and cells were incubated at 37ºC for 20 minutes. Following incubation, cells were centrifuged at 4ºC for 4 minutes at 400 g and the resulting cell pellet was resuspended in 500 ul of 3% FBS in PBS with 0.1% EDTA and kept on ice. Unstained and single dye controls prepared in parallel were used for FACS calibration.

Subcellular fractionation and RNA extraction of mESCs was carried out using the PARIS kit (Thermo Fisher, AM1921) according to the manufacturer's instructions. Genomic DNA removal from the cytosolic and nuclear RNA fractions was performed through RNeasy (Qiagen, 74104) on-column DNAse digestion (Qiagen, 79254) was performed according to manufacturers instructions.

Cells were sorted in an AstriosEQ (Beckmann Coulter) cell sorter and collected in 500 uL of RTL buffer from the RNeasy extraction kit (QIAGEN). Forward and side scatter (FSC & SSC, 488nm DPSS laser) were used as is common for size and doublet exclusion. We excluded dead cells based on LIVE/DEAD fluorescence (488nm DPSS laser) and used DNA content (355nm DPSS laser) to define three gates: G1, S and G2/M (Supplementary Figure S5).

RNA was extracted using RNeasy mini kit (Qiagen) according to manufacturer's instructions. RNA was reverse transcribed using Quatitect Reverse Transcription Kit (Qiagen 205313). Quantitative PCR analysis was performed using SYBR Green Quantitative PCR Master Mix (Roche 0692404001) in a Light Cycler 96 Real Time PCR system (Roche). Briefly, the RTqPCR reaction was assembled in 10uL with

forward and reverse primers at a final concentration of .5 uM each, and SYBR Green Quantitative PCR Master Mix at a final concentration of 1X.

Candidate lincRNA overexpression and cell cycle analysis.

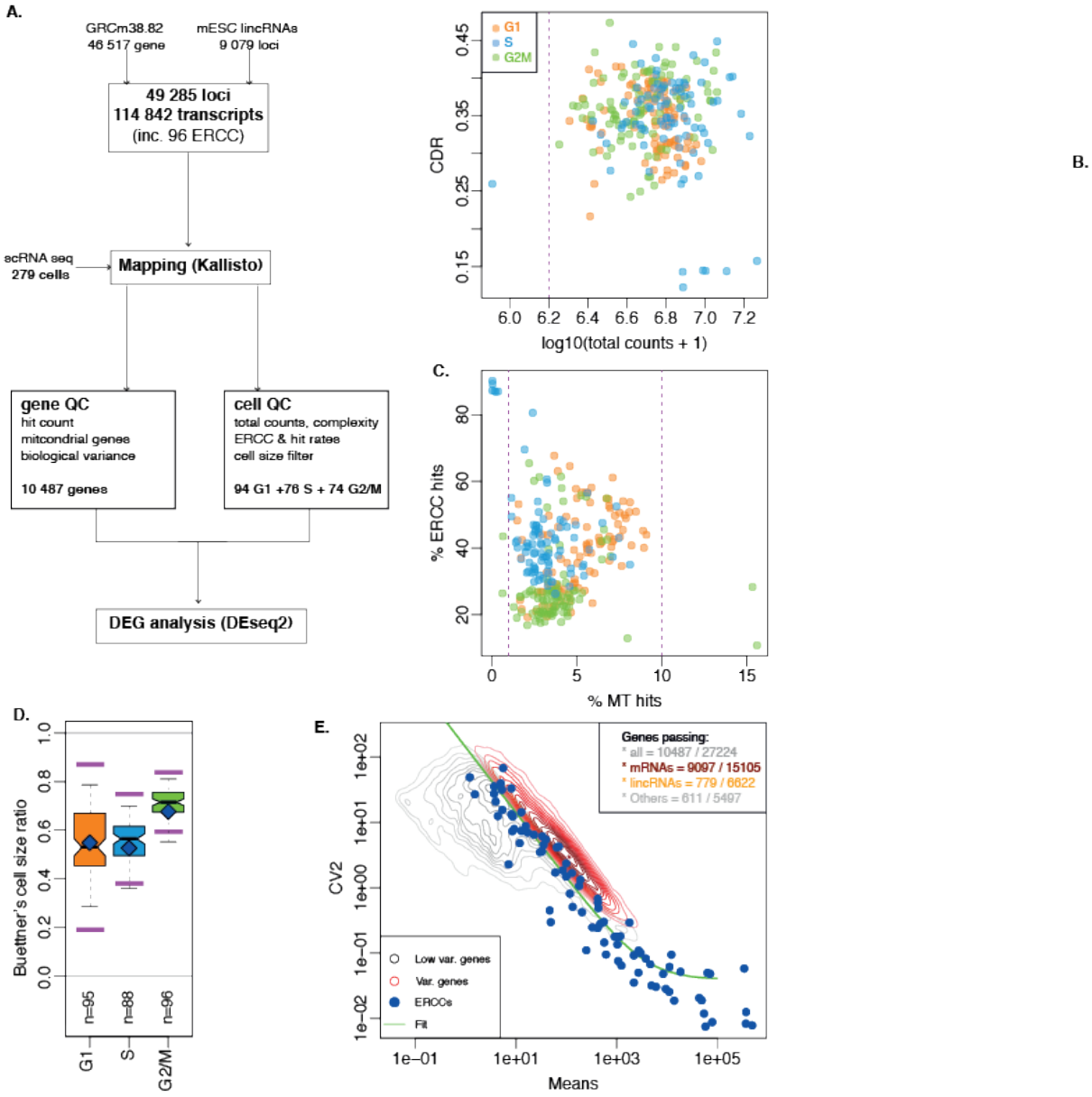
We used E-CRISPR (Heigwer et al., 2014) to design 2 guide RNA (gRNA) sequences located within 1 kb of lincCC1 and lincCC2 promoters respectively. We synthetized oligos containing these sequences and BbsI restriction sites. Annealed oligos were inserted into the pKLV-U6gRNA(BbsI)-PGKpuro2ABFP vector (Ochiai et al., 2015).
We seeded $1.5*10^5$ E14 cells/well in 6-well plates and allowed cells to grow overnight. We used lipofectamine 2000 (Thermo Fischer) to transfect 1000 ng of gRNA expression vector and 1000 ng of pAC95-pmax-dCas9-Vp160-2A-Neo (Cheng et al., 2013) vector into mESC cultures overnight according to manufacturer's instructions. We replaced growth medium 8 hours after transfection to minimize toxicity. Seventy-two hours post transfection one sample was collected for RNA to ensure the efficiency of the over-expression by qPCR; in parallel the other cells were pulsed with 10 uM Edu for 30 minutes in the dark at 37ºC. Cells were washed with 1% BSA in PBS and trypsinized cells. Following a second wash with 1% BSA in PBS, cells were fixed at room temperature, protected from the light for 15 minutes in 100 uL of Click-iT fixative solution. After fixation, cells were washed in 1% BSA in PBS, resuspended in 100 uL of 1x Click-iT saponin-based permeabilization and wash reagent, and incubated at room temperature in the dark for 15 minutes. Following incubation, 500 uL of Click-iT reaction cocktail containing Alexa Fluor 488 Fluorescent Dye Azide were added and samples were incubated at room temperature in the dark for 30 minutes. Reaction was quenched by resuspending cells in 1 mL of 1x Click-iT saponin-based permeabilization reagent. To stain the samples for DNA content, 1 uL of Fx Cycle FarRed staining solution (Thermo Fischer F10348) was added to each sample, and Pure Link RNase A (Thermo Fischer 12091039) was added at a final concentration of 100 ng/mL. The samples were incubated at 4ºC in the dark for 30 minutes, analyzed by flow cytometry analysis. Cells stained for DNA incorporation and content were analyzed using a 10 color/3 laser Beckman Coulter Gallios Analyzer. Blue (405 nm) and red (640 nm) excitation lasers were used for excitation of the Alexa Fluor 488 Fluorescent Dye Azide and Fx FarCycle Red respectively. Emission was detected using channels FL-09 and

FL-06 for Alexa Fluor 488 and Fx FarCycle Red respectively. Two biological replicates, with 3 technical replicates each, were tested for both candidate lincRNAs using two different guide RNAs (gRNA1 and gRNA2). We excluded flow cytometry data for one of CC1's gRNA1 experiments due to issues with RNA collection preventing us from confirming candidate overexpression. 30-60k events were flowed per sample, except for the first experiment where 15-30k events were flowed.

Flow data was analyzed in FlowJo (v10.2). Flow events were gated on Cells and Singlets, and spillover was compensated. Low incorporation 2n and 4n populations, as well as a medium-to-high incorporation population, were easily identifiable in the DNA incorporation vs content scatterplot, and gated respectively as G1, G2/M and S (and ungated). This stage-level gating was performed by two gaters, blind to the direction and amplitude of target lincRNA expression modulation. Stage population fractions were comparable between gaters (Pearson correlations between stage fractions 0.65-0.97 depending on stage & experiment), and thus all results presented here are those of one gater.

Sample-level summaries of the flow cytometry results were generated using FlowJo (v10.2). Summaries were imported into R. Samples were quality-controlled based on their percentages of cells/singlets (indicative of cell health), and MFIs (indicative of staining strength). Experiments for which the associated qPCRs did not validate over-expression of the targets were excluded. Ultimately, 27 samples from 2 experiments and 5 treatment groups (scrambled, two different guides for lincCC1 and two for lincCC2) were retained. For each stage, we ran a linear mixed model (restricted log-likelihood maximisation, R package nlme version 3.1-137) to investigate the impact of treatment on the percentage of cells in that stage. We used the log2-transformed percentage as outcome, target lincRNA (scrambled, lincCC1, lincCC2) as a fixed effect, and biological replicates as a random intercept.

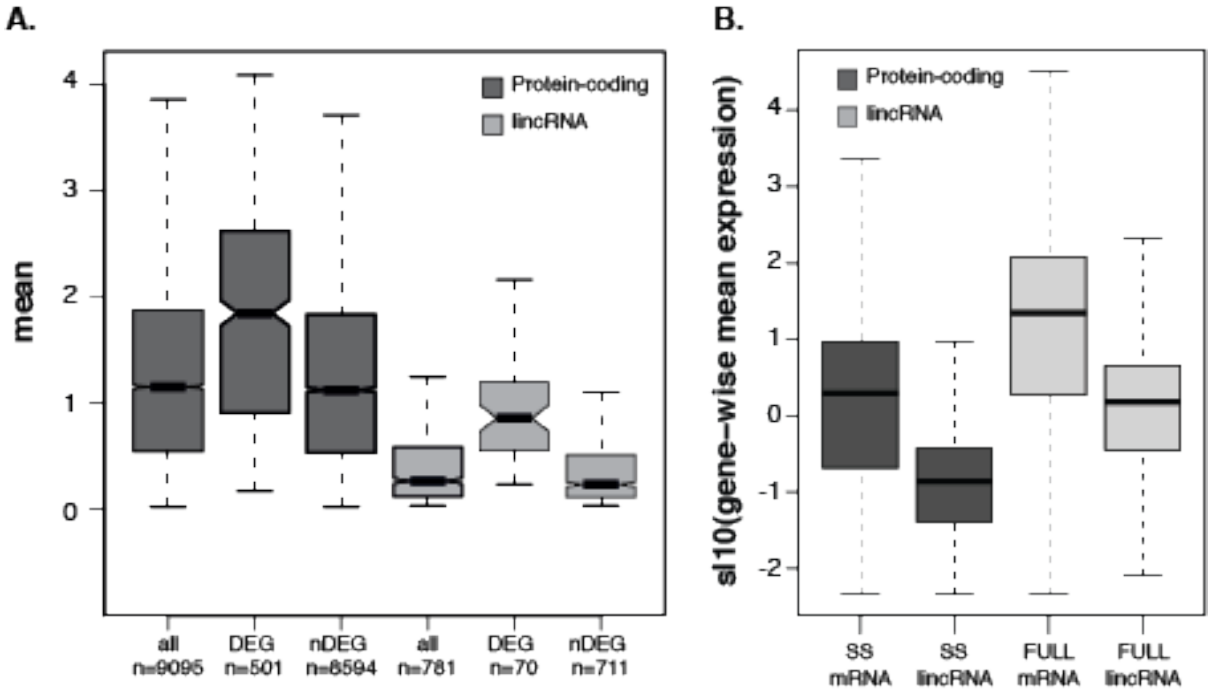
SUPPLEMENTARY FIGURES



Supplementary Figure S1 . Schematics of the scRNA sequencing analysis, related to Figure 1.

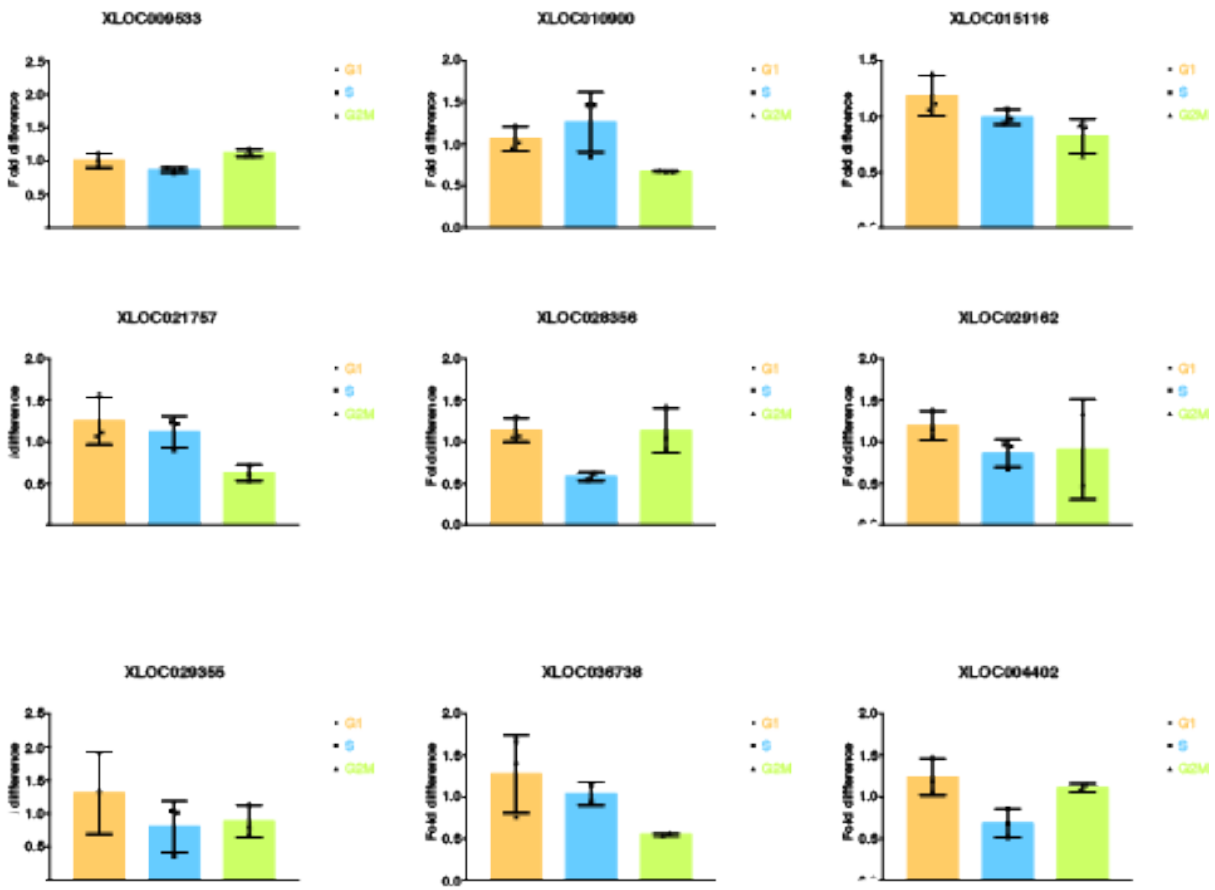
A) Schematic representation of the differential expression analysis pipeline. **B)** Scatterplot showing for each sample (cell) the total hit (mapped read) counts & "Cell Detection Rate", i.e. fraction of genes detected by at least 1 hit in each cell. Low library quality exclusion thresholds are indicated by dashed magenta lines. Cells are coloured by cell cycle stage. **C)** Scatterplot showing for each cell the percent of hits to mitochondrial (MT) genes & to ERCC spike-ins. Low library quality thresholds are indicated by dashed magenta lines. **D)** Distributions of Buettner's cell size ratio (Buettner et al., 2015) for cells in each stage. Magenta lines show medians +/- 2 median absolute deviations; cells falling outside of these windows are excluded. **E)**

Technical noise filter (Brennecke et al., 2013) as implemented in scLVM (Buettner et al., 2015): each gene is represented by its cross-sample mean expression (x-axis) and coefficient of variation (squared: CV2, y-axis). Blue dots correspond to ERCC spike-ins, while the green line represents the non-linear CV2 ~ mean model that was fit to the spike-ins, and then used to test which genes show significantly higher-than-technical variability. Genes in black do not pass the significance test, those in red do. Counts of genes passing the test are indicated in the top-right hand corner.



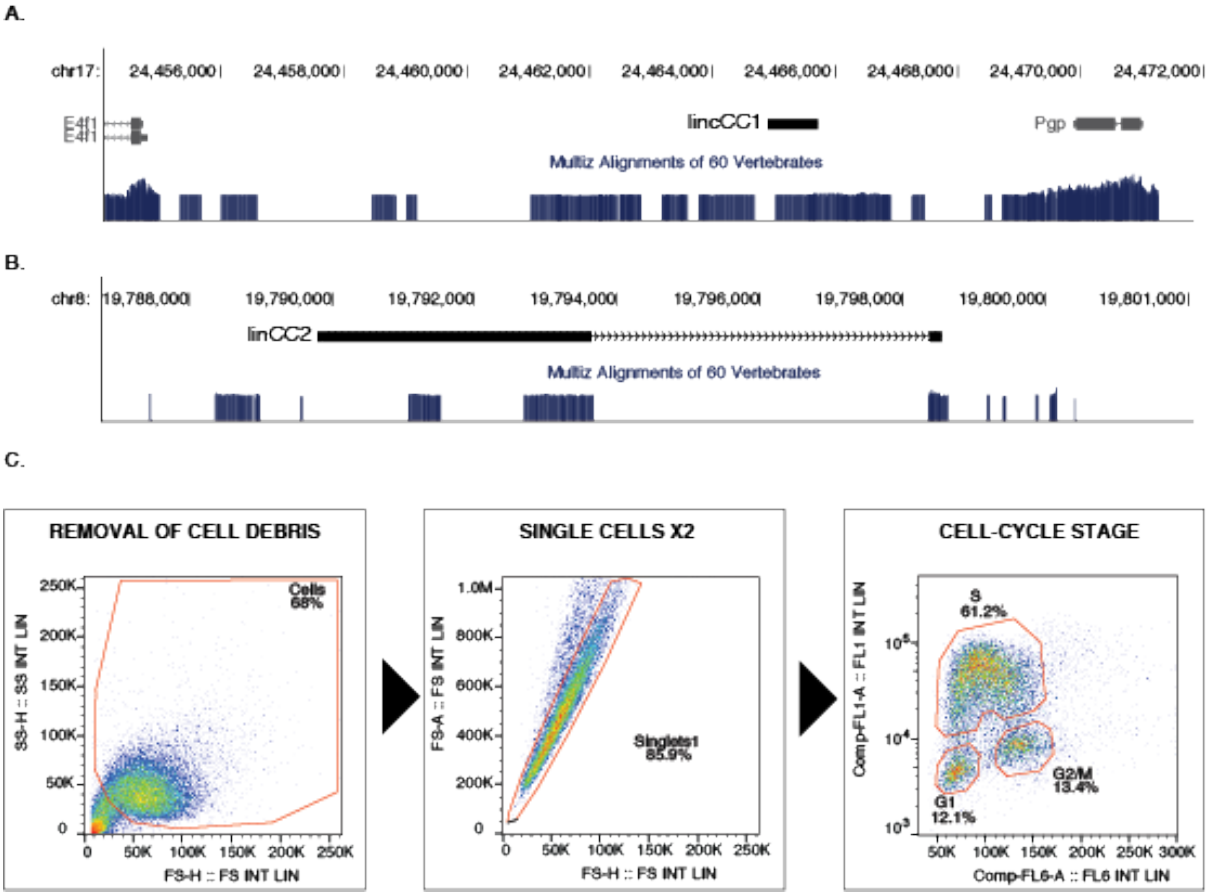
Supplementary Figure S2. Cell-cycle stage specific expression, related to Figure 1.

A) boxplots of cross-sample mean expression (normalised shifted log10 counts) for genes of various classes: all=tested for differential expression. DE=differentially-expressed. NDE: non-differentially expressed. mRNA: protein-coding genes. lincRNA: lincRNA loci. **B)** Boxplots of shifted log10 mean loci expression for protein-coding genes (mRNA, dark grey) and lincRNAs (light grey), in both the sub-sampled (1/14th) data (SS) and FULL data from cell cycle staged single-cell RNAseq dataset (Buettner et al., 2015).



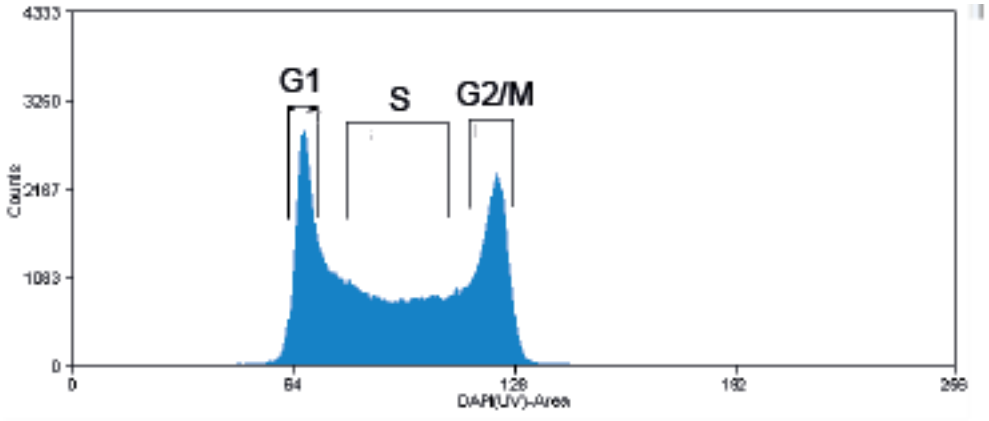
Supplementary Figure S3. RT-qPCR analysis of selected lincRNA's expression in mESC sorted by cell cycle stage based on DNA content, related to Figure 1.

Fold difference in relative expression for 9 predicted cell-cycle differentially expressed lincRNAs.



Supplementary Figure S4. Characteristics of candidate CC-lincRNA, related to Figure 4.

Genome browser view of the genomic location (mm10) of **A)** lincCC1 and **B)** lincCC2. Blue bars indicate conservation in multiZ alignments (60 vertebrates). **C)** gating strategy to identify cells from distinct cell cycle stages. Panel 1 shows the size exclusion gate based on FSC-H & SSC-H for removing debris. Panels 2 shows the FSC-H vs FSC-A doublet exclusion gate; a similar gate was implemented for SSC-H vs SSC-A. Panel 3 shows the partitioning of cells into 3 stages based on DNA content (measured on FL6-A) and EdU incorporation (measured on FL1-A).



Supplementary Figure S5. Representative DNA content histogram for mESC, related to Figure 4.

Histogram of cellular DNA content measures as stained by DAPI, showing gates for delineating clearly-defined G1, S or G2/M stages.

REFERENCES

BRAY, N. L., PIMENTEL, H., MELSTED, P. & PACHTER, L. 2016. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol,* 34**,** 525-7.

BRENNECKE, P., ANDERS, S., KIM, J. K., KOLODZIEJCZYK, A. A., ZHANG, X., PROSERPIO, V., BAYING, B., BENES, V., TEICHMANN, S. A., MARIONI, J. C. & HEISLER, M. G. 2013. Accounting for technical noise in single-cell RNA-seq experiments. *Nat Methods,* 10**,** 1093-5.

BUETTNER, F., NATARAJAN, K. N., CASALE, F. P., PROSERPIO, V., SCIALDONE, A., THEIS, F. J., TEICHMANN, S. A., MARIONI, J. C. & STEGLE, O. 2015. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat Biotechnol,* 33**,** 155-60.

CHEN, X., XU, H., YUAN, P., FANG, F., HUSS, M., VEGA, V. B., WONG, E., ORLOV, Y. L., ZHANG, W., JIANG, J., LOH, Y. H., YEO, H. C., YEO, Z. X., NARANG, V., GOVINDARAJAN, K. R., LEONG, B., SHAHAB, A., RUAN, Y., BOURQUE, G., SUNG, W. K., CLARKE, N. D., WEI, C. L. & NG, H. H. 2008. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell,* 133**,** 1106-17.

CHENG, A. W., WANG, H., YANG, H., SHI, L., KATZ, Y., THEUNISSEN, T. W., RANGARAJAN, S., SHIVALILA, C. S., DADON, D. B. & JAENISCH, R. 2013. Multiplexed activation of endogenous genes by CRISPR-on, an RNA-guided transcriptional activator system. *Cell Res,* 2**3,** 1163-71.

CONSORTIUM, E. P. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature,* 489**,** 57-74.

DOBIN, A., DAVIS, C. A., SCHLESINGER, F., DRENKOW, J., ZALESKI, C., JHA, S., BATUT, P., CHAISSON, M. & GINGERAS, T. R. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics,* 29**,** 15-21.

HEGER, A., WEBBER, C., GOODSON, M., PONTING, C. P. & LUNTER, G. 2013. GAT: a simulation framework for testing the association of genomic intervals. *Bioinformatics,* 29**,** 2046-8.

HEIGWER, F., KERR, G. & BOUTROS, M. 2014. E-CRISP: fast CRISPR target site identification. *Nat Methods,* 11**,** 122-3.

JAAKKOLA, M. K., SEYEDNASROLLAH, F., MEHMOOD, A. & ELO, L. L. 2017. Comparison of methods to detect differentially expressed genes between single-cell populations. *Brief Bioinform,* 18**,** 735-743.

KRYUCHKOVA-MOSTACCI, N. & ROBINSON-RECHAVI, M. 2017. A benchmark of gene expression tissue-specificity metrics. *Brief Bioinform,* 18**,** 205-214.

LE, S., JOSSE, J. & HUSSON, F. 2008. FactoMineR: An R package for multivariate analysis. *Journal of Statistical Software,* 25**,** 1-18.

LI, B. & DEWEY, C. N. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics,* 12**,** 323.

LOVE, M. I., HUBER, W. & ANDERS, S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol,* 15**,** 550.

LUN, A. T., MCCARTHY, D. J. & MARIONI, J. C. 2016. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Res,* 5**,** 2122.

OCHIAI, H., SUGAWARA, T. & YAMAMOTO, T. 2015. Simultaneous live imaging of the transcription and nuclear position of specific genes. *Nucleic Acids Res,* 43**,** e127.

REIMAND, J., ARAK, T., ADLER, P., KOLBERG, L., REISBERG, S., PETERSON, H. & VILO, J. 2016. g:Profiler-a web server for functional interpretation of gene lists (2016 update). *Nucleic Acids Res,* 44**,** W83-9.

SANTOS, A., WERNERSSON, R. & JENSEN, L. J. 2015. Cyclebase 3.0: a multi-organism database on cell-cycle regulation and phenotypes. *Nucleic Acids Res,* 43**,** D1140-4.

SONESON, C., LOVE, M. I. & ROBINSON, M. D. 2015. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Res,* 4**,** 1521.

SUN, N., YU, X., LI, F., LIU, D., SUO, S., CHEN, W., CHEN, S., SONG, L., GREEN, C. D., MCDERMOTT, J., SHEN, Q., JING, N. & HAN, J. J. 2017. Inference of differentiation time for single cell transcriptomes using cell population reference data. *Nat Commun,* 8**,** 1856.

TAN, J. Y., SIREY, T., HONTI, F., GRAHAM, B., PIOVESAN, A., MERKENSCHLAGER, M., WEBBER, C., PONTING, C. P. & MARQUES, A. C. 2015. Extensive microRNA-mediated crosstalk between lncRNAs and mRNAs in mouse embryonic stem cells. *Genome Res,* 25**,** 655-66.

TAN, J. Y., VANCE, K. W., VARELA, M. A., SIREY, T., WATSON, L. M., CURTIS, H. J., MARINELLO, M., ALVES, S., STEINKRAUS, B., COOPER, S., NESTEROVA, T., BROCKDORFF, N., FULGA, T., BRICE, A., SITTLER, A., OLIVER, P. L., WOOD, M. J., PONTING, C. P. & MARQUES, A. C. 2014. Cross-talking noncoding RNAs contribute to cell-specific neurodegeneration in SCA7. *Nat Struct Mol Biol,* 21**,** 955-961.

XU, H., BAROUKH, C., DANNENFELSER, R., CHEN, E. Y., TAN, C. M., KOU, Y., KIM, Y. E., LEMISCHKA, I. R. & MA'AYAN, A. 2013. ESCAPE: database for integrating high-content published data collected from human and mouse embryonic stem cells. *Database (Oxford),* 2013**,** bat045.