Extensive and choreographed transcription of noncoding RNAs within cell cycle promoters

Tiffany Hung, Yulei Wang, Michael F. Lin, Ashley Koegel, Yojiro Kotake, Gavin Grant, Hugo M. Horlings, Nilay Shah, Christopher Umbricht, Pei Wang, Yu Wang, Seung K. Kim, Marc van de Vijver, Saraswati Sukumar, Michael L. Whitfield, Manolis Kellis, Yue Xiong, David J. Wong and Howard Y. Chang

1 Program in Epithelial Biology, Stanford University School of Medicine, Stanford, California 94305, USA.
2 Department of Developmental Biology, Stanford University School of Medicine, Stanford, California 94305, USA.
3 Life Technologies, Foster City, California 94404, USA.
4 The Broad Institute, Cambridge, Massachusetts 02142, USA.
5 Department of Biochemistry and Biophysics, Lineberger Comprehensive Cancer Center, Program in Molecular Biology and Biotechnology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA.
6 Department of Genetics, Dartmouth Medical School, Hanover, New Hampshire 03755, USA.
7 Department of Pathology, Academic Medical Center, Meibergdreef 9, 1105AZ, Amsterdam, The Netherlands.
8 Sidney Kimmel Comprehensive Cancer Center,
9 Department of Surgery, Johns Hopkins University School of Medicine, Baltimore, Maryland 21231, USA.
10 Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA.
11 Howard Hughes Medical Institute, Stanford University School of Medicine, Stanford, CA 94305.

*These authors equally contributed to this work.

Corresponding authors: D.J.W. (davewong@stanford.edu) and H.Y.C. (howchang@stanford.edu)
SUMMARY

Transcription of long noncoding RNAs (lncRNAs) within gene regulatory elements can modulate gene activity in response to external stimuli, but the scope and functions of such noncoding transcription are not known. Here we use an ultra-high density array that tiles the promoters of 56 cell cycle genes to interrogate 108 samples representing diverse perturbations. We identify 216 transcribed regions that encode putative lncRNAs--many of which have periodic expression during the cell cycle, show altered expression in human cancers, and are regulated in expression by specific oncogenic stimuli, stem cell differentiation, or DNA damage. DNA damage induces five lncRNAs from the CDKN1A promoter, and one such lncRNA, which we named PANDA (P21-Associated ncRNA, DNA damage Activated), is induced in a p53-dependent but p21-independent manner. Depletion of PANDA induced expression of pro-apoptotic genes and markedly sensitized human fibroblasts to apoptosis by doxorubicin. These findings suggest potentially widespread roles for promoter lncRNAs in cell growth control.
INTRODUCTION

Mammalian genomes are more pervasively transcribed than previously expected (Bertone et al., 2004; Calin et al., 2007; Carninci et al., 2005). In addition to protein coding genes, many types of non-coding RNAs (ncRNAs) are transcribed. Small regulatory ncRNAs, including small interfering RNAs (siRNAs), microRNAs, and Piwi-associated RNAs (piRNAs), function in genome defense and post-transcriptional regulation (He and Hannon, 2004; Hutvagner and Simard, 2008; Mattick and Makunin, 2005). Near transcriptional start sites (TSS), divergent transcription by RNA polymerase can generate small ncRNAs ranging from 20 to 200 nucleotides, which have been variously named promoter associated sRNAs (PASRs), transcription-initiation RNAs (tiRNAs), and TSS-associated RNAs (TSSa-RNAs) (Core et al., 2008; Kapranov et al., 2007; Seila et al., 2008; Taft et al., 2009). However, it remains uncertain if these ncRNAs are functional or just represent byproducts of RNA polymerase infidelity (Ponjavic et al., 2007; Struhl, 2007). Long ncRNAs (lncRNAs) vary in length from several hundred bases to tens of kilobases; they may be located in isolation from protein coding genes (long intergenic ncRNAs, or lincRNAs), or they may be interspersed nearby and within protein coding genes (Guttman et al., 2009; Katayama et al., 2005). Moreover, recent evidence suggest that active enhancer elements are also transcribed to produce ncRNAs (De Santa et al., 2010; Kim et al.).

Although evidence for function of ncRNAs as a group is lacking, several lncRNAs have been implicated in transcriptional regulation. Two prime examples are in the genomic loci of cell cycle genes. In the cyclin D1 (CCND1) promoter, an ncRNA transcribed two kilobases upstream of the CCND1 gene is induced by ionizing radiation
and regulates transcription of \textit{CCND1} in \textit{cis} by forming a ribonucleoprotein repressor complex (Wang et al., 2008). This ncRNA binds to and allosterically activates the RNA-binding protein TLS (translated in liposarcoma), which inhibits histone acetyltransferases, resulting in repression of \textit{CCND1} transcription. A second example is an antisense \textit{p15} ncRNA gene (\textit{p15AS} or \textit{ANRIL}) that overlaps with the \textit{p15} coding sequence, and \textit{p15AS} expression is increased in human leukemias with an inverse correlation with \textit{p15} expression (Pasmant et al., 2007; Yu et al., 2008). \textit{p15AS} can transcriptionally silence \textit{p15} directly as well as through induction of heterochromatin formation. Many well studied IncRNAs, such as those involved in dosage compensation and imprinting, regulate gene expression in \textit{cis} (Lee, 2009), but other lincRNAs, such as \textit{HOTAIR} and linc-p21, can regulate the activity of distantly located genes in \textit{trans} (Gupta et al., ; Huarte, 2010; Rinn et al., 2007). Inspired by these examples, we hypothesized that the genomic loci of cell cycle genes may harbor other functional ncRNAs that have yet to be discovered.

In this study, we create an ultrahigh-resolution tiling microarray to interrogate the transcriptional and epigenetic landscape around the TSSs of 56 cell cycle genes, including all cyclins, cyclin-dependent kinases (CDKs), and cyclin-dependent kinase inhibitors (CDKIs). We analyze a diverse collection of cells and tissues samples that interrogate distinct perturbations in cell growth control. Our results reveal a map of extensive and choreographed noncoding transcription, and identify a specific set of IncRNAs that function in the DNA damage response.
RESULTS

**Extensive and regulated noncoding transcription near cell cycle genes**

To systematically discover functional ncRNAs in the regulatory region of human cell cycle genes, we created a tiling array that interrogates at 5-nucleotide resolution across 25kb of the 9p21 locus [which encompasses \textit{CDKN2A} (p16), \textit{p14ARF}, and \textit{CDKN2B} (p15)], as well as from 10kb upstream to 2kb downstream of each TSS from 53 cell cycle genes to include all known cyclins, CDKs, and CDKIs (Fig 1A, Table S1). These genes are also critical for fundamental biological processes such as senescence, self renewal, DNA damage response, and tumor formation (Hall and Peters, 1996; Johnson and Walker, 1999; Sherr and Roberts, 1999). Thus, we hybridized 54 pairs of polyadenylated RNAs from various human cells that were altered or perturbed through cell cycle synchronization, DNA damage, differentiation stimuli, oncogenic stimuli, or carcinogenesis (Table S2).

A peak calling algorithm searched for statistically significant signals above background and detected contiguous regions (peaks) of at least 50 basepairs. We then compiled statistically significant transcripts from all 108 channels of the 54 arrays, clustered all transcripts that overlapped by a minimum of 50 bases, and identified clusters that were present in at least 10% of the samples. Averaging the signal intensity across all probes in a peak produced a quantitative estimate of transcript abundance. Despite possible 3’ bias due to poly-adenylated RNA selection, our procedure detected exon 1 transcription from the majority of cell cycle coding genes (41 of the 56), demonstrating that this custom tiling array can detect previously reported transcribed regions. In each
individual sample, we detected an average of 73 of the 216 transcribed regions (range 14-189) that did not overlap with known exons of the 56 cell cycle genes (Fig S1; example of the CCNE1 locus in human fetal lung fibroblasts shown in Fig 1B). Across all 108 samples, we identified a total of 216 discrete transcribed regions (Table S3). The average transcript length was 234 nucleotides (range 50-1494). 171 of the 216 (79%) novel transcribed regions were located 5’ of the TSS of the cell cycle genes (“upstream”), 40 of the 216 (19%) were located within introns (“intronic”), and 5 of the 216 (2%) were located downstream of the 3’ end of CDKN2A.

Genes actively transcribed by RNA polymerase II are marked by trimethylation of histone H3 on lysine 4 (H3K4me3) and lysine 36 of histone H3 (H3K36me3), which reflect gene starts and bodies, respectively (Rando and Chang, 2009). These chromatin marks can be used to identify noncoding transcription (Guttman et al., 2009). In a subset of our samples, we determined whether the 216 transcribed regions were similarly marked for active transcription by performing chromatin immunoprecipitation followed by hybridization to our custom tiling array (ChIP-chip). This analysis confirmed that the chromatin state at a majority of the newly defined transcripts were enriched in both H3K4me3 and H3K36me3 (Fig 1B and 1C). Using EpiGRAPH analysis to query our transcripts against approximately 900 published genomic attributes (Bock et al., 2009), the 216 putative transcribed regions are enriched for H3K4me3 (p<10^{-9}) and RNA polymerase II binding (p<10^{-7}), providing further evidence that these genomic regions are actively transcribed.

To determine whether the 216 transcripts may encode previously unknown protein-coding exons or non-coding RNAs, we used the codon substitution frequency
(CSF) analysis to assess for characteristic evolutionary signatures of protein-coding sequences across 21 sequenced mammalian genomes (Methods). As expected, the transcribed regions that coincided with annotated exons had high CSF scores. However, over 86% of the novel transcribed regions had CSF scores well below the threshold of known protein-coding genes and resemble known ncRNAs (Fig 1D and Table S3), suggesting that most of the novel regions do not have protein coding potential. BLAST analysis confirmed that the majority of the transcripts are not known protein coding genes (Table S3). Furthermore, none of the transcripts intersect known pre-miRNAs, C/D box small nucleolar RNAs, H/ACA box snoRNAs, and scaRNAs as annotated in the UCSC genome browser. Thereafter, we refer to these transcribed regions as long non-coding (lnc)RNAs. We aligned the RNA hybridization signals at all 56 protein-coding loci of all 108 samples relative to their TSS (Fig 1E). As expected, we found a peak immediately downstream of the TSS corresponding to exon 1 of the protein coding gene. In addition, we found enrichment of noncoding transcription in the region 4 to 8 kilobases upstream of the TSS. Thus, unlike the previously described PASRs, tiRNAs, and TSSaRNAs, which are primarily located within 100 bp of the TSS, the majority of these ncRNAs are longer and are not clustered immediately around the TSS.

**Expression patterns of ncRNAs suggest specific biological functions**

Next, we examined the biological conditions that regulate expression of these ncRNAs in order to infer possible biological functions. We assembled a matrix of the expression changes of the 216 novel transcribed regions across all 54 perturbations and hierarchically clustered the genes and samples (Fig 2A). Of the 216 novel transcribed
regions, 92 (43%) had at least a 2 fold change in expression detected on the tiling array in at least one of the perturbations, suggesting that a large subset may have functional roles. The samples that had the most transcripts with at least 2 fold expression change were the embryonic stem cells (ESC) relative to day 152 fetal pancreas (40 of 216) and invasive ductal breast carcinomas relative to normal (as many as 35 of 216), suggesting that a subset of these lncRNAs may play a role in self-renewal and carcinogenesis(Fig 2A). Interestingly, lncRNA expression profiles of keratinocytes with knockdown of p63, which inhibits keratinocyte differentiation, clustered with that of ESC, suggesting that these ncRNAs may have a role in the undifferentiated state. Expression patterns from five keratinocyte samples that were transduced with the oncogene MYC alone or in combination with other oncogenes relative to controls clustered together, demonstrating that MYC has a dominant effect on ncRNA expression. MYC-RAS-IκBα transduced human keratinocytes activate an ESC-like mRNA gene expression program and acquire properties of cancer stem cells (Wong et al., 2008). Notably, the lncRNA expression profile of MYC-RAS-IκBα cells clustered with that of ESC (Fig 2), suggesting a shared lncRNA signature for embryonic and cancer stem cells. In contrast, the E2F3-RAS-IκBα transduced keratinocyte, which do not express the ESC-like mRNA gene expression program, had an inverse pattern of expression for the majority of lncRNAs. In addition, 8 primary human invasive ductal breast carcinomas split into 2 different groups based on their lncRNA profiles: 4 of the cancers clustered with the ES cells and MYC-RAS-IκBα tumors and the other 4 clustered with the E2F3- RAS-IκBα tumors, suggesting that these tumor models mimic the expression pattern of not only mRNAs but also these lncRNAs in bona fide human cancers.
The 216 lncRNAs are divided into 3 main clusters based on their expression pattern across all samples (Fig 2). Notably, cluster 1 is composed of lncRNAs that are strongly induced in ES cells, keratinocytes with p63-knockdown, and Myc-Ras-IkB tumors relative to differentiated cells and GFP-Ras-IkB tumors, which we interpret to be a “stemness cluster” (Fig 2B). Interestingly, each cluster is composed of many of the ncRNAs from the same genomic locus, suggesting that multiple adjacent ncRNAs are either coordinately regulated in a shared response or are spliced together as exons of one transcript. High correlation of the dynamic expression patterns of these ncRNAs and different biological and cellular conditions suggest that these ncRNAs may be functional in the cell cycle, self renewal, and cancer.

**A gene co-expression map infers trans regulatory mechanisms and biological functions**

Multiple ncRNAs, including p15AS and the lncRNA upstream of CCND1, have been shown to regulate the transcription of the nearby coding gene. To determine whether gene-proximal lncRNAs are typically correlated with the expression of the nearest mRNA, we conducted whole genome expression arrays on 17 samples that were also examined on our tiling array, and calculated pair-wise Pearson correlations between the expression patterns of each cell cycle promoter lncRNAs vs. every mRNA genome-wide. Surprisingly, there was no significant correlation or anti-correlation between most of the 216 lncRNAs and the nearby protein-coding mRNA, suggesting that most of the lncRNAs may not function in cis to activate or repress nearby mRNA expression (Fig 3A). In contrast, we found that the median correlation between two ncRNAs of the same
locus was positive, supporting our hypothesis that neighboring ncRNAs may be coordinately regulated, positively regulate each other, and/or are exons of the same transcript (Fig 3B).

Given that expression of the 216 ncRNAs do not correlate with the mRNA in cis, we further explored the genes and pathways that they may regulate, using a guilt-by-association approach (Guttman et al., 2009). For each IncRNA, we define a co-expression gene set as the group of mRNAs that are positively or negatively correlated with that IncRNA across the 17 samples ($R>0.5$ or $R<0.5$, respectively) (Fig S2). We then constructed a gene module map (Segal et al., 2004) of the association of each ncRNA co-expression gene set vs. the Gene Ontology Biological Processes, and performed biclustering to identify ncRNAs that are associated with distinct Gene Ontology terms (Fig 3C). This analysis revealed multiple sets of ncRNAs that are associated with biological processes including cell cycle, DNA recombination, ribonucleoprotein complex biogenesis and assembly, RNA splicing, and response to DNA damage. Thus, despite having no correlation in expression to their neighboring protein-coding gene, the expression patterns of these ncRNAs are still strongly related to the cell cycle. We constructed a similar module map with curated gene sets of metabolic and signaling pathways as well as biological and clinical states from the Molecular Signatures Database (MSigDB c2 collection) (Subramanian et al., 2005). This module map confirmed the enrichment for cell-cycle related sets (e.g. Cell Cycle Brentani, Cell Cycle KEGG). In addition, enriched modules included several poor prognosis breast cancer gene sets (BRCA ER negative, BRCA prognosis negative, BRCA1 overexpressed
up), DNA damage related gene sets (UVA/UVB), several oncogenic signatures (Ras, Myc), and stem cell gene sets (Hematopoietic stem cell, Neural Stem Cell) (Fig S3).

**Validation of ncRNA expression in cell cycle, ES cell differentiation, cancer and DNA damage response**

To validate these inferred functional associations, we designed quantitative RT-PCR assays for 60 of the 216 novel transcribed regions (43 upstream and 7 intronic) to obtain a more quantitative measure of these ncRNAs across different conditions. Expression in HeLa cells synchronized in cell cycle progression by double thymidine block demonstrate that most of the ncRNAs have periodic expression with peaked expression at different phases of the cell cycle (Fig 4A) (Whitfield et al., 2002). Next, comparison of human ES cells and fetal pancreas at days 76 and 152 demonstrated that a majority of these ncRNAs are regulated during differentiation (Fig 4B). In addition, unsupervised clustering of lncRNA expression patterns in 5 metastatic breast cancers and 5 normal mammary tissues readily distinguished the 5 metastatic breast cancers from the normal mammary tissues (Fig 4C). Some of the lncRNAs, including upst:CCNL1::-2767 and int:CDKN1A:+885, are repressed in the metastatic breast cancers relative to normal mammary tissues, whereas others, including upst:CDKN1A:-4845, upst:CDKN2B:-2817, and int:ARF:+4517, are induced. Thus, the majority of these lncRNAs has periodic expression in the cell cycle, and is differentially expressed in different states of cell differentiation and cancer progression.

Our co-expression maps predicted associations of several ncRNAs with DNA damage response pathways (Fig 3C and Fig S3). In support of this finding, doxorubicin-
treated human fetal lung fibroblasts showed at least 2-fold change in 12 of the 216 ncRNAs on the tiling array (Fig 2). Interestingly, 2 of those 12 ncRNAs were located 5’ of the TSS of the canonical p53 target gene CDKN1A (upst:CDKN1A: -1210 and upst:CDKN1A: -4845), and similar to the CDKN1A mRNA, were induced by doxorubicin (Fig 5A). In addition, a third ncRNA at the CDKN1A locus, upst:CDKN1A:-800, was also induced by doxorubicin, but was not included in the 216 ncRNAs because it was only expressed in one of the 108 samples, the doxorubicin-treated fibroblasts. In order to confirm whether these ncRNAs may be responsive to DNA damage, we measured expression changes of 60 ncRNAs predicted in the DNA damage pathway (as well as upst:CDKN1A:-800) by quantitative RT-PCR in human fetal lung fibroblasts treated with doxorubicin, over a 24 hour time course. Most of the ncRNAs were either significantly induced or repressed by doxorubicin, and all 5 of the tested ncRNAs surrounding the CDKN1A TSS were induced, including the 3 that were previously detected on the tiling array (Fig 5B). Notably, several ncRNAs upstream of CDKN1A are induced more rapidly and with substantially higher magnitude than CDKN1A upon DNA damage. Upst:CDKN1A:-4845 is induced up to 40 fold upon DNA damage (Fig 5C). These variations in expression patterns within the same locus suggest that the ncRNAs in the CDKN1A locus may play distinct roles in the DNA damage response from the CDKN1A protein, p21.

**PANDA: a long noncoding RNA involved in the DNA-damage response**

To investigate the functional relevance of these ncRNAs at the CDKN1A locus, we selected upst:CDKN1A:-4845, hereafter termed *PANDA* (P21 Associated NcRNA
DNA damage Activated) for further analysis. *PANDA* is located approximately 5 kilobases upstream of the *CDKN1A* TSS and coincides with a cluster of previously annotated ESTs. Although the *PANDA* locus intersects a computationally predicted pseudogene of *LAP3*, quantitative RT-PCR demonstrated that *PANDA* was specifically induced by DNA damage, whereas *LAP3* expression did not significantly change, confirming that the change in expression detected by the tiling array was not due to cross-hybridization with *LAP3* (*Fig S4*). Furthermore, the CSF score of *PANDA*, 9.3, indicated very low protein coding potential compared to *LAP3* (CSF range 117-1343 for its 13 exons). Finally, rapid amplification of 5′ and 3′ complementary DNA ends (RACE) revealed a 1.5 kilobase transcript that is divergently transcribed from *CDKN1A*, antisense of the predicted *LAP3* pseudogene (*Fig 5A and Fig S5*). Thus, *PANDA* is a 5′-capped and polyadenylated non-spliced ncRNA that is transcribed antisense to *CDKN1A*.

Since p53 is a positive regulator of *CDKN1A* during the DNA damage response, we asked whether p53 also regulates *PANDA* expression. ChIP-chip analysis confirmed the p53 binding site immediately upstream of the *CDKN1A* TSS (*Fig 5A*) (Wei et al., 2006). *PANDA* and *CDKN1A* are diametrically situated 2.5kb from this intervening p53 binding site, which supports the possibility of p53 co-regulation. Indeed, siRNA-mediated knockdown of p53 prior to DNA damage inhibited the induction of *PANDA* by 70% 24-hours post-DNA damage (*Fig 5D*), similar to its effect on *CDKN1A*. In contrast, RNA interference of *CDKN1A* had no effect on *PANDA* expression, indicating that *PANDA* is not a linked transcript of *CDKN1A* nor is *PANDA* expression dependent on p21.
Next, we addressed whether *PANDA* affects the DNA damage response. We transduced human fetal fibroblasts with custom siRNAs targeting *PANDA* and then applied doxorubicin for 24 hours following the knockdown (Fig 6A). Global gene expression analysis showed that 224 genes were induced and 193 genes were repressed at least 2-fold by *PANDA* knockdown (Fig. 6B and Table S4). Genes induced by *PANDA* knockdown are significantly enriched for those involved in apoptosis, such as Gene Ontology terms *cell death* (p<0.04) and *apoptosis* (p<0.03) (Fig 6B). Quantitative RT-PCR confirmed the induction of several canonical apoptosis genes after *PANDA* depletion, including *FAS, APAF1, LRDD*, and *BIK* (Fig 6C). On the other hand, expression of neither *CDKN1A* itself nor *TP53* was affected by *PANDA* depletion (Fig. 6D), suggesting that *PANDA* is a p53 effector that acts independently of p21<sub>CDKN1A</sub>.

DNA damage in human fibroblasts triggers p53-dependent G1 arrest, but not apoptosis (Agarwal et al., 1995; Di Leonardo et al., 1994). Consistent with this finding, doxorubicin treatment in cells exposed to control siRNA had little to no apoptosis as measured by TUNEL. In contrast, *PANDA* knockdown resulted in five to seven-fold increased TUNEL-positive cells (Fig 6E and F). Immunoblot analysis of PARP, a caspase substrate and marker of apoptosis, revealed PARP cleavage only in *PANDA* depleted cells (Fig 6G). Thus, *PANDA* knock down sensitized fibroblasts to DNA damaged-induced apoptosis. In contrast, six additional siRNAs targeting other transcripts within the CDKN1A promoter had no effect on apoptosis (data not shown). Altogether, these data suggest that in parallel with p53-mediated induction of CDKN1A for cell cycle arrest, p53-mediated induction of *PANDA* delimits apoptosis.
DISCUSSION

A regulatory network of noncoding transcription in cell cycle promoters

Recent studies have revealed that a surprisingly large fraction of mammalian genomes is transcribed. In addition to small noncoding RNAs, long noncoding RNAs can be produced from gene promoters, enhancers, as well as stand-alone intergenic loci (De Santa et al., 2010; Guttman et al., 2009; Katayama et al., 2005). New approaches are needed that not only identify ncRNAs, but also provide insight into their potential biological function. Using an ultra-high resolution tiling array, we interrogated the transcriptional landscape at cell cycle promoters in 108 samples that represent diverse perturbations. The ability to interrogate numerous and diverse biological samples in a rapid and economical fashion is advantageous for at least two reasons. First, many of the noncoding transcripts are induced only in highly specific conditions, and may have been missed if only a few conditions were surveyed. Of the 216 new noncoding transcribed regions we identified, on average only 73 of these are transcribed in any one biological sample. Second, comparison of ncRNA profiles amongst these diverse samples highlighted unexpected similarities in cell cycle promoter states among distinct perturbations. For instance, we identified a similarity of promoter states among ES cells, tumors induced by MYC, and epithelial progenitors depleted of differentiation regulator p63. Likewise, authentic human tumors can be classified based on the similarity of their promoter states to that of cells with defined oncogenic perturbation.

Noncoding transcription through regulatory elements may affect gene activity in a variety of ways. The act of transcription may open compacted chromatin over regulatory sequences, or compete with transcription factor binding (so called transcriptional
interference). In addition, the ncRNA product may modulate neighboring gene expression *in cis* (Kanhere et al.,; Lee, 2009), affect distantly located genes *in trans* (Rinn et al., 2007), or even serve as a target for regulation by small regulatory RNAs (Han et al., 2007; Schwartz et al., 2008). Because these different mechanisms predict distinct relationships between levels of ncRNAs and cognate mRNAs, we compared ncRNA and mRNA expression profiles across our samples. We found that most promoter ncRNAs are neither positively nor negatively correlated in expression with their neighboring mRNA, but are rather correlated in expression with genes located elsewhere in the genome. The genes co-expressed (and presumably co-regulated) with promoter ncRNAs function in specific biological pathways, including cell cycle, DNA damage response, stem cell differentiation, and have been associated with cancer prognosis. Quantitative RT-PCR analysis further validated that many of these ncRNAs are periodically expressed in the cell cycle, are regulated in response to DNA damage, ES cell differentiation, and are differentially expressed in human cancers. These findings suggest that cell cycle ncRNAs may participate in gene regulation *in trans*. In addition, noncoding transcription of cell cycle promoters may be a form of regulatory anticipation or feedback to modulate the chromatin state of cell cycle promoters.

**Parallel functions of coding and noncoding RNAs driven by cell cycle promoters**

Our results suggest that the human genome is organized into genomic units that code for multiple transcripts that function in the same biological pathways (*Fig. 7*). 49 of 56 cell cycle protein-coding gene loci had at least one detected ncRNA and an average of four ncRNAs within 10 kilobases upstream and 2 kilobases downstream of the TSS. At
the CDKN1A promoter, five ncRNAs, similar to the CDKN1A mRNA itself, are induced by DNA damage. One of these ncRNAs, which we named PANDA, is a non-spliced 1.5 kilobase ncRNA that is transcribed antisense to CDKN1A and is induced with slower kinetics than that of CDKN1A. siRNA-mediated knockdown of p53 inhibited the induction of PANDA by DNA damage, demonstrating that PANDA induction is p53-dependent. In contrast, depletion of CDKN1A or depletion of PANDA had no effect on each other’s response to DNA damage, indicating that their induction by p53 occurs in parallel. PANDA inhibits the expression of apoptotic genes and delimit DNA-damage induced apoptosis. Intriguingly, Huarte et al. recently identified a distinct long intergenic noncoding RNA located 15 kilobases upstream of CDKN1A, named lincRNA-p21, that is induced by p53 and mediates p53-dependent gene repression (Huarte et al., in press).

Thus, the regulatory sequence upstream of the CDKN1A gene drives the expression of multiple coding and noncoding transcripts that cooperate to regulate the DNA damage response. These findings provide a vivid example that shows the blurring boundary between “genes” and “regulatory sequences” (Mattick, 2003).

Our study provides an initial catalogue of lncRNAs in cell cycle promoters that may play diverse functions. At a minimum, promoter ncRNA expression provides a convenient means of tracking the chromatin state of promoters, which may be of use in cancer biology and regenerative medicine. Future studies are needed to pinpoint the functions of these and likely other ncRNAs emanating from regulatory sequences.
Acknowledgements

We thank J. Rinn for communicating unpublished results, M. Guttman and A. Regev for discussions, L. Attardi for careful reading of the manuscript, and P. Khavari for reagents. Authors Yulei Wang and Yu Wang are employees of Life Technologies. This work was supported by grants from NIH/NIAMS (K08-AR054615 to D.J.W.), NIH/NCI (R01-CA118750 to H.Y.C.), Juvenile Diabetes Research Foundation (S.K. K. and H.Y.C.) and American Cancer Society (H.Y.C.). H.Y.C. is an Early Career Scientist of the Howard Hughes Medical Institute.
FIGURE LEGENDS

Figure 1: Identification of ncRNAs near and within cell cycle genes.

(A) Flow chart of strategy for systematic discovery of cell cycle ncRNAs.

(B) Representative tiling array data. RNA hybridization intensity and H3K36me3 and H3K4me3 ChIP-chip signal relative to input at the CCNE1 locus in human fetal lung fibroblasts. Predicted transcripts shown in red boxes. Known mRNA exons in black boxes. Peak Calling: Each bar represents a significant peak from one of the 108 array channels.

(C) Chromatin state at transcribed regions. Average ChIP-chip signal relative to input calculated across transcriptional peaks expressed in human fetal lung fibroblasts +/- doxorubicin treatment.

(D) Codon substitution frequency (CSF) analysis. Graph of average evolutionary CSF of exons of coding genes and predicted transcripts. CSF<10 represents no protein coding potential.

(E) Transcriptional landscape of cell cycle promoters. All 56 cell cycle promoters were aligned at the TSS and average RNA hybridization signal was calculated across the 12 kilobase window. This process was repeated with all 49 samples. Output represents a 150 basepair running window of average transcription signal across all 56 promoters and all 54 arrays.

See also: Supplementary Table S1 and Supplementary Figure S1.

Figure 2: ncRNA expression across diverse cell cycle perturbations.

A) Hierarchical clustering of 216 predicted ncRNAs across 54 arrays, representing 108 conditions. Red indicates that the cell cycle perturbation induced transcription of the
ncRNA. Green indicates that the cell cycle perturbation repressed transcription of the ncRNA. Black indicates no significant expression change.

B) Zoom in view of ncRNAs in cluster 1.

See also: Supplementary Table S2 and S3.

**Figure 3: Functional associations of ncRNAs.**

(A) ncRNA expression patterns do not correlate with that of the mRNAs in cis. Histogram of Pearson correlations between each of the 216 ncRNAs and the cis mRNA across 17 samples. Median=.0239.

(B) ncRNA expression patterns have positive correlation with neighboring ncRNA transcripts. Histogram of Pearson correlations between each of the 216 ncRNAs and nearby transcripts on the same locus across 108 samples. Mode=0.4.

(C) Genes co-expressed with ncRNAs are enriched for functional groups in cell cycle and DNA damage response. Module map of ncRNA gene sets (columns) versus Gene Ontology Biological Processes gene sets (rows) across 17 samples (p<0.05, FDR<0.05). A yellow entry indicates that the GO gene set is positively associated with the ncRNA gene set. A blue entry indicates that the GO gene set is negatively associated with the ncRNA gene set. Black entry indicates no significant association. Representative enriched GO gene sets listed.

See also: Supplementary Figure S2 and S3.

---

**Figure 4: Validated expression of ncRNAs in cell cycle progression, ES cell differentiation, and human cancers.** Custom Taqman probes were generated and used to interrogate independent biological samples for ncRNA expression.
(A) Periodic expression of lncRNAs (blue) during synchronized cell cycle progression. Cell cycle phases are confirmed by FACS and expression of genes with known periodic peaks in the cell cycle (orange).

(B) Regulated expression of lncRNAs in human ES cells vs. fetal pancreas. (C) Differential expression of lncRNAs in normal breast epithelium vs. breast cancer.

**Figure 5: lncRNAs at CDKN1A locus are induced by DNA damage.**

(A) Top: map of all detected transcripts at the CDKN1A promoter. Middle two tracks: Example of RNA hybridization intensity in control or 24 hour doxorubicin treated (200ng/ml) human fetal lung fibroblasts. Note, not all DNA damage inducible transcripts are observed in one single time point. Bottom track: p53 ChIP-chip signal relative to input confirmed the p53 binding site immediately upstream of the CDKN1A TSS upon DNA damage. RACE clone of upst:CDKN1A:-4425 closely matches predicted transcript on tiling array. See also: Supplementary Table S4.

(B) Quantitative RT-PCR of lncRNAs shows coordinate induction or repression across a 24 hour time course of doxorubicin treatment. A cluster of lncRNAs transcribed from the CDKN1A locus are induced.

(C) Expression of transcripts from the CDKN1A locus over a 24 hour time course after doxorubicin-treatment of normal human fibroblasts.

(D) Doxorubicin induction of PANDA requires p53 but not CDKN1A. Mean ± s.d. are shown.

See also: Supplementary Figure S6.
Figure 6: *PANDA* ncRNA inhibits apoptotic response to DNA damage.

(A) siRNA knockdown of PANDA in the presence of DNA damage with doxorubicin. Custom siRNAs specifically target *PANDA* with no discernable effect on the *LAP3* mRNA.

(B) Heat map of gene expression changes with siPANDA relative to control siRNA at 24 hours of doxorubicin treatment.

(C) Quantitative RT-PCR of canonical apoptosis pathway genes reveals induction with siPANDA relative to control siRNA at 28 hours of doxorubicin treatment.

(D) Quantitative RT-PCR of *CDKN1A* and *TP53* reveal no expression changes with siPANDA relative to control siRNA.

(E) TUNEL immunofluorescence of control and siPANDA cells at 28 hours of doxorubicin treatment.

(F) Quantification of 3 independent TUNEL assays.

(G) Western blot of PARP cleavage in control and PANDA siRNA cells treated at 28 hours of doxorubicin treatment. Mean ± s.d. are shown in all bar graphs.

Figure 7. Coding and noncoding transcripts at *CDKN1A* locus coordinate DNA damage response.
METHODS

Tiling array design and RNA Hybridization

A custom tiling array (Roche Nimblegen) was designed at 5 basepair resolution across 25kb of the 9p21 region (which encompasses CDKN2a, p14ARF, and CDKN2b), as well as from 10kb upstream to 2kb downstream of each TSS from 53 other cell cycle genes including cyclins, CDKs, and CDKIs (Table S1). In addition, the HOXA and HOXD loci were placed on the array as a control. Briefly, RNA was amplified (MessageAmp Kit, Ambion), reverse transcribed (Retroscript Kit, Ambion), labeled, and hybridized according to the standard Nimblegen protocol.

Peak Calling

Robust multichip average (RMA) normalized single channel data from each array was subjected to peak calling using the Nimblescan program (Roche Nimblegen) with a window size = 50. Peaks with a peak score > 10 were considered significant transcriptional units. Peak calls from all 55 array samples were clustered using Galaxy (Carninci et al., 2005; Taylor et al., 2007), and only transcripts present in a minimum of 10% of the samples were considered for further analysis. Transcripts were annotated as following – “genomic location (upstream of TSS of cell cycle protein-coding gene = upst; exon of cell cycle protein-coding gene = exon; intron of cell cycle protein-coding gene = int; downstream of cell cycle protein coding gene = dst)” : “gene symbol of nearest mRNA” : “distance from TSS”.

Measuring Protein Coding Potential
To assess the coding potential of the novel transcribed regions, we evaluated the evolutionary signatures in their alignments with orthologous regions in 20 other sequenced placental mammalian genomes using a new version of the Codon Substitution Frequencies (CSF) method (Lin et al., 2007; Lin et al., 2008), which has also been applied to assess novel transcribed regions in mouse (Guttman et al., 2009). The new version of CSF is methodologically more advanced than the original, but it operates very similarly in that it produces a score for any region in the genome considering all codon substitutions observed within its alignment, based on the relative frequency of similar substitutions in known coding and non-coding regions. Briefly, the new version performs a statistical comparison between two empirical codon models (Kosiol et al., 2007), one estimated from alignments of known coding regions and the other based on non-coding regions, and reports a likelihood ratio that quantifies whether the protein-coding model is a better explanation, while controlling for the overall level of sequence conservation (MF Lin, I Jungreis & M Kellis, in prep). Following the previous approach in mouse (Guttman et al., 2009), we tested whether each transcribed region contains even a short window with significant protein-coding evolutionary signatures, by taking the maximally scoring 30 codon window in any reading frame to represent the region.

**Module Map analysis**

We generated a module map of the ncRNAs versus the protein-coding genes by computing the Pearson correlations for all pairwise combinations based on expression across 17 different samples. This map was clustered and visualized using the program Genomica (http://genomica.weizmann.ac.il/). For each ncRNA, we then defined gene sets of the protein coding genes that had a Pearson correlation that was greater than or
less than 0.5 with that ncRNA. To determine functional associations, we then generated a module map of these ncRNA gene sets with Gene Ontology Biological Processes gene sets (Fig 3C) and with curated gene sets of metabolic and signaling pathways and biological and clinical states from the Molecular Signatures Database (MSigDB c2 collection) (Fig S4) (Subramanian et al., 2005). P-value of enrichment was determined by the hypergeometric distribution, and a false discovery rate (FDR) calculation was used to account for multiple hypothesis testing (p<0.05, FDR<0.05).

**Tissue samples and cells**

Informed consent was obtained for tissue donation as well as approval from institutional review boards. Human primary breast tumors from the Netherlands Cancer Institute (van de Vijver et al., 2002), and normal breast tissues and metastatic breast tumors from the Johns Hopkins University Rapid Autopsy Program (Gupta et al., 2010) are as described. Human fetal pancreata were obtained from the Birth Defects Research Laboratory, University of Washington (Seattle, WA). Staged fetal pancreata were processed within 24 hours of receipt, minced, washed and processed for RNA isolation using standard methods. Human fetal lung fibroblasts were obtained from the Coriell repository (Coriell AG04546 and AG04393) and cultured in 10% FBS (Hyclone), 1% penicillin-streptomycin (Gibco) at 37C in 5% CO2.

**PANDA Cloning and sequence analysis**

3’ and 5’ RACE was performed using the FirstChoice RLM-RACE Kit (Ambion). RNA was extracted from 200ng/ml doxorubicin (Sigma) treated human fetal lung fibroblasts and RLM-RACE was performed according to the standard manufacturer’s protocol. The
following primers were used for RACE of PANDA: Fwd1: 
5’CAGAACTTGGCAGATGGAG3’, Rev1: 
5’TGATATGAACTCGTTTACTATGAC3’, Fwd2: 
TGACACATTTAACCCGAAG, Rev2: CCCCCAAGCTACATCTATGACA. 
Rev3: 5’CGTCTCCATCATGCCAAGTT3’, 
Rev4: 5’CATAGAGTTCCACCCGACATG3’.

RT-PCR

Total RNA was extracted from cells using the Trizol reagent (Invitrogen) and the RNeasy Mini Kit (Qiagen) and genomic DNA was eliminated using Turbo DNA-free (Ambion). RT-PCR using 50-100 ng of total RNA was performed using the One-Step RT-PCR Master Mix (Applied Biosystems) using the following Taqman Gene Expression Assays: PANDA (custom Taqman), TP53 (Hs99999147_m1), LAP3 (Rh02870758_m1), APAF1 (Hs00559441_m1), LRDD (Hs00388035_m1), FAS (Hs00163653_m1), BIK (Hs00154189_m1), and CDKN1A (Hs01121168_m1). Expression values were normalized to GAPDH (Hs99999905_m1).

Strand specific RT-PCR for PANDA was performed using the One Step RT-PCR Master Mix SYBR Green (Stratagene) with the following primers: Fwd (5’-TGCACACATTTAACCCGAAG-3’), Rev (5’-CCCCAAGCTACATCTATGACA-3’).

TaqMan® custom ncRNA Assays

A panel of TaqMan® custom ncRNA assays was developed targeting 60 of the 219 novel transcribed regions using “Single-exon” design mode. The transcript specificity and genome specificity of all TaqMan assays were verified using a position
specific alignment matrix to predict potential cross-reactivity between designed assays and genome-wide non-target transcripts or genomic sequences. For gene expression profiling of these ncRNAs across different conditions, cDNAs were generated from 50ng of total RNA using the High Capacity cDNA Reverse Transcription Kit (Life Technologies, Foster City, CA). The resulting cDNA was subjected to a 14-cycle PCR amplification followed by real-time PCR reaction using the manufacturer’s TaqMan® PreAmp Master Mix Kit Protocol (Life Technologies, Foster City, CA). Two replicates were run for each gene for each sample in a 384-well format plate on 7900HT Fast Real-Time PCR System (Life Technologies, Foster City, CA). PPIA was used as an endogenous control for normalization across different samples.

**Antibodies**

The following antibodies were used for Chromatin Immunoprecipitation Assays: anti-H3K4me3 (Abcam ab8580), anti-H3K35me3 (Abcam ab9050), anti-p53 (Abcam ab28). Western blots were performed using anti-PARP (Cell Signal 9542) and anti-B-tubulin (Abcam ab6046).

**RNA Interference**

Human fetal lung fibroblasts were transfected with 50 nM of onTargetPLUS siRNAs (Dharmacon) targeting $PANDA$ (duplex A= 5’-AAUGUGUGCAGCUAACAGAUU-3’ and 5’-GAGAUUUGCAGCACACAUU-3’, duplex B= 5’-GGGCAUGUUUUCACAGGCUU-3’ and 5’-GAGAUUUGCAGCAACACAUU-3’, duplex C= 5’-AAUGUGUGCAGCUAACAGAUU-3’ and 5’-
GGGCAUGUUUUCACAGAGGUU-3’). Non-targeting onTargetPLUS siRNAs (Dharmacon, D-001810-10) were used as a control.

**TUNEL**

TUNEL assays were performed using the In Situ Cell Death Detection Kit, TMR Red (Roche). Human fetal lung fibroblasts were cultured on chamber slides (Lab-Tek), treated with 200ng/ml doxorubicin (Sigma) for 28 hours, fixed with methanol at -20C for 10 minutes, and incubated with the TUNEL labeling mixture for one hour at 37C. Slides were then washed with PBS and mounted in Prolong® Gold antifade reagent with DAPI (Invitrogen) and imaged at 20x magnification.
REFERENCES


Hung et al., p.30


Custom high density tiling arrays

Samples:
- ES cells and differentiated cells
- Cells transduced with oncogenes
- Cell cycle-arrested cells
- Cell cycle-synchronized cells
- Cancer and normal tissue

RNA peaks on arrays

RNA peak ratios vs conditions

ncRNAs vs mRNAs

siRNA assays of ncRNAs

Fig 1B

Fig 2, 4

Fig 3

Fig 5, 6

Fig 1D

Fig 2, 4

Fig 3

Fig 5, 6

Fig 1E

Fig 2, 4

Fig 3

Fig 5, 6
Figure 2

A 216 transcriptional units

Breast carcinomas
Doxorubicin-treated fibroblasts
E2F3-Ras-IkB tumors

Ca²⁺-differentiated keratinocytes
Fibroblasts in low serum

Breast carcinomas
Myc-Ras-IkB tumors
Human ES cells

p63 knockdown in keratinocytes

Myc overexpression in keratinocytes

B

Hung et al, Figure 2
Figure 3

A

B

C

Hung et al., Figure 3
**Figure 4**

A. Heatmap depicting the expression pattern of protein-coding genes over time (0-14 hours) post-double thymidine block. The genes include E2F1, CCNA2, BUB1B, and RAD21.

B. Heatmap showing the expression of 60 ncRNAs in ES and fetal pancreas samples, with fold change values ranging from -2.0 to +2.0.

C. Heatmap comparing normal breast and metastatic ductal carcinoma, with fold change values ranging from -8.0 to +8.0.
Figure 5

A. Predicted and experimental binding sites for p53 on the CDKN1A promoter.

B. Heatmap showing the expression of 61 ncRNAs at different hours post-doxorubicin treatment.

C. Graph showing the fold change in RNA expression for different promoter regions.

D. Bar graph showing the fold change in siRNA/siCTRL for different genes.
Figure 6

A

RNA (Fold)

Dox

Ctrl   A   B   C

siPANDA

B

Apoptosis (p=0.03)
Programmed Cell Death (p=0.04)
RNA processing (=0.03)
DNA packaging (p=0.0001)
Nucleosome Assembly (p=0.01)
Protein-DNA complex assembly (p=0.01)

C

FAS

Fold ∆ Dox/Ctrl

Ctrl   A   B   C

siPANDA

LRDD

Fold ∆ Dox/Ctrl

Ctrl   A   B   C

siPANDA

APAF1

Fold ∆ Dox/Ctrl

Ctrl   A   B   C

siPANDA

BIK

Fold ∆ Dox/Ctrl

Ctrl   A   B   C

siPANDA

CDKN1A

Fold ∆ Dox/Ctrl

Ctrl   A   B   C

siPANDA

TP53

Fold ∆ Dox/Ctrl

Ctrl   A   B   C

siPANDA

E

siCTRL

DAPI

TUNEL

MERGE

siPANDA

F

% TUNEL Positive

Ctrl   A   B   C

siPANDA

G

Parp

Tubulin

116 kDa

85 kDa

Hung et al., Figure 6
CDKN1A

Cell cycle arrest

p21

PANDA

linc-p21

p53-mediated gene repression

Apoptosis

Cell cycle arrest

Hung et al., Figure 7

Figure 7