Cell

TUF Love for "Junk" DNA

Aarron T. Willingham[1] and Thomas R. Gingeras[1,*]
[1]Affymetrix, Inc., Santa Clara, CA 95051, USA
*Contact: tom_gingeras@affymetrix.com
DOI 10.1016/j.cell.2006.06.009

The widespread occurrence of noncoding (nc) RNAs—unannotated eukaryotic transcripts with reduced protein coding potential—suggests that they are functionally important. Study of ncRNAs is increasing our understanding of the organization and regulation of genomes.

Over the past five years, researchers working with various organisms and using multiple technologies to explore genomewide gene expression have converged on the same surprising conclusion: transcription is widespread throughout the genome and many-fold higher than existing genome annotations would predict. The burgeoning number of these transcripts of unknown function, or TUFs (Cheng et al., 2005), highlights a remarkably complex transcriptional architecture that includes alternative splice isoforms for almost all protein-coding genes, widespread transcription of antisense RNAs, and abundant noncoding RNAs (ncRNAs) with important biological functions. By some estimates, TUFs could rival protein-coding transcripts in number (Cawley et al., 2004). Such transcriptional diversity may explain how the relatively similar numbers of protein-coding genes estimated for fruit fly (13,985; BDGP release 4), nematode worm (21,009; Wormbase release 150), and human (23,341; NCBI release 36) result in the remarkable phenotypic differences observed among these species.

Widespread Transcription

Although recent evidence indicates that the complexity of transcripts produced by the human genome and the underlying transcriptional architecture is striking, these observations are consistent with much earlier reports. Decades ago, several studies discovered evidence for widespread transcription (see Table S1 available with this article online). In the 1970s, studies involving sea urchin embryos found that heterogeneous nuclear RNA had 10-fold more nucleotide complexity than cytoplasmic mRNA associated with polysomes. Studies of lampbrush chromosomes in newt oocytes estimated that empirically measured transcription was an order of magnitude greater than necessary to produce mRNAs for the oocyte. An abundance of nonpolyadenylated polysome-associated RNAs (~30% of cytoplasmic RNAs) was first identified in HeLa cells and subsequently observed in a variety of plant and animal cells. Furthermore, a 10-fold greater complexity of nuclear versus cytoplasmic polyadenylated (polyA) RNA was also found in human cells, and 5′ cap structures outnumber polyA segments 3:1 in hamster ovary cells. These original studies arrived at the common conclusion that the complexity of transcripts made by most organisms seemed to be inexplicably larger than expected. Further insight into the nature of this transcriptional diversity awaited advancements in our understanding of primary structure of genomes and development of new analytical technologies.

Genomic tiling arrays represent an unbiased and sensitive tool for global studies of transcription. Such a technology offers independence from current limited genome annotations while enabling detection of rare transcripts. Using genome tiling microarrays, widespread transcription along chromosomes and across the human genome has been observed, including significant antisense products, and may be an order of magnitude greater than annotation-based predictions (reviewed in Johnson et al., 2005). In 2002, in a systematic analysis of transcription across human chromosomes 21 and 22, our group observed about an order of magnitude more transcriptional activity than could be accounted for by predicted protein-coding genes, suggesting that a significant portion of transcribed cytoplasmic polyA RNA may indeed be noncoding. A subsequent human genome-wide mapping effort of polyA RNA from human liver reached similar conclusions and identified >10,000 new transcripts, many of which have homology to genomic sequences in other mammalian species. Tiling array-based whole-genome mapping in the model plant *Arabidopsis* found that >50% of observed transcription was intergenic and that ~30% of annotated genes had associated antisense transcription, some of which was tissue-specific. Furthermore, about 20% of annotated pseudogenes were expressed, suggesting that examples of pseudogene-mediated regulation of gene activity may be common (Hirotsune et al., 2003). In *Drosophila*, ~40% of probes in intronic and intergenic areas detected RNA expression, much of which changed in a developmentally coordinated manner. Furthermore, alternative splicing was observed in ~40% of known genes, yielding over 5000 new splice forms. Recently, even the small, well-characterized yeast genome has yielded a more complex transcriptome than expected, with overlapping transcription and differential expression levels even within the same gene (David et al., 2006). Together, these studies provide several observations about transcriptomes: (1) the widespread incidence of unannotated transcripts

with limited protein-coding capacity, often expressed at low levels; (2) a large degree of overlapping transcription, evidenced in part by the presence of abundant antisense transcription; and (3) most coding genes have alternative splice forms.

Combining chromatin immunoprecipitation (ChIP) and tiling microarrays has allowed unbiased, often high-resolution mapping of transcription factor binding sites across the genome, and in the process has lent considerable unbiased empirical support to the existence of widespread unannotated transcription. The binding sites of the NF-κB family member RelA/p65 were mapped across human chromosome 22, with nearly 28% of the binding sites shown to lie >50 kb away from known protein-coding genes (Martone et al., 2003). Mapping of DNA binding sites for Sp1, cMyc, and p53 transcription factors on human chromosomes 21 and 22 revealed that almost half of the mapped binding sites potentially correlate with ncRNAs, and a significant subset of the noncoding transcripts showed transcriptional responsiveness to retinoic acid (Cawley et al., 2004). Recently, a whole-genome map of RNA polymerase II preinitiation complex binding sites in human primary fibroblasts identified ~10,000 active promoters, with approximately 13% corresponding to unannotated loci (Kim et al., 2005). Several factors could account for the substantially lower percent of observed binding sites that may correlate with noncoding transcription, such as the highly conservative thresholds used to identify binding sites and the low occupancy by RNA polymerase for low abundance transcripts. Taken together, these results suggest that widespread transcription of TUFs is initiated and regulated by molecular mechanisms similar to those modulating protein-coding RNAs.

Widespread unannotated transcription has been confirmed by other experimental means, including mapping 5′ ends of transcripts using cap analysis of gene expression (CAGE), 3′ ends with serial analysis of gene expression (SAGE), massively parallel signature sequencing (MPSS), and high-throughput full-length cDNA cloning and sequencing methodologies (reviewed in Mattick and Makunin, 2006). These sequencing-based approaches provide strand specificity for transcripts and can detect transcripts originating from repetitive regions of the genome, whereas tiling arrays often have repetitive regions excluded and may use labeling protocols that are strand-insensitive. Expression analysis by SAGE tags found at least 15,000 uncharacterized 3′ termini used in the genome, suggesting that new isoforms and genes numbered in the many thousands. In the functional annotation of the mouse genome (FANTOM) project, large-scale sequencing and manual annotation of full-length cDNAs characterized 102,281 transcripts, including 32,129 protein-coding transcripts (2222 of which encode new proteins) and 34,030 ncRNAs (Carninci et al., 2005). This study also concluded that the total number of transcripts is at least an order of magnitude larger than current estimates of genes present in the mouse genome. Furthermore, only about 40% of characterized ncRNA sequences were identified in previous FANTOM cDNA collections, suggesting that the number of unannotated transcripts could continue to grow. Sequencing of 3 million human CAGE tags also confirmed that human cells have a similar level of transcriptional diversity. Analysis of human gene expression by MPSS found that >65% of signature sequences do not overlap with annotated transcripts; rather, 38% map to introns, 21% are antisense to known exons, and 5% map to intergenic areas. Intergenic transcripts are expressed at low levels on arrays, and it may be that the cloning and sequencing methodologies of MPSS require greater analysis to detect transcripts with such low abundance.

Antisense transcription is also observed in cDNA sequencing, corroborating findings in microarray studies. Transcription from the opposite strand of a protein-encoding locus produces RNAs that may hybridize with DNA or RNA and could interfere with transcription, translation, or mRNA stability of the "sense" product. Systematic cDNA cloning of small RNAs from *C. elegans* identified 700 distinct endogenous siRNAs, possibly processed from larger antisense transcripts, precisely complementary to protein-coding regions from more than 500 different genes; such small RNAs may act as siRNAs in plants (reviewed in Sontheimer and Carthew, 2005). Genomewide surveys of expressed mouse and human sequences and cDNA sequencing experiments identified a surprisingly large number of transcripts antisense to protein-coding genes. This suggests that the majority of sense transcripts (e.g., 72% of mouse genes) may have an antisense partner (Katayama et al., 2005).

Noncoding RNAs and Their Functions

A key question hangs like an ominous cloud over these observations of widespread transcription: are these transcripts biologically functional, or are they the transcriptional noise of a less than precise set of biological processes? Recent experiments in mice in which megabase "gene desert" regions have been deleted underscore the relevance of this question. Deletion of 1.5 Mb and 0.8 Mb genomic intervals, which together contain 1243 noncoding sequences conserved between rodent and primate, resulted in viable mice with no obvious deleterious phenotypes (Nobrega et al., 2004). However, if history is our guide, then the answer to this question may be complex (see Figure 1).

It has taken more than 35 years to identify and characterize seven functional classes of ncRNAs: ribosomal (r), transfer (t), small nuclear (sn), antisense (AS), small nucleolar (sno), micro (mi) and Piwi-interacting (pi). With the exception of perhaps the rRNA and tRNA functional classes, the other ncRNA classes are believed to

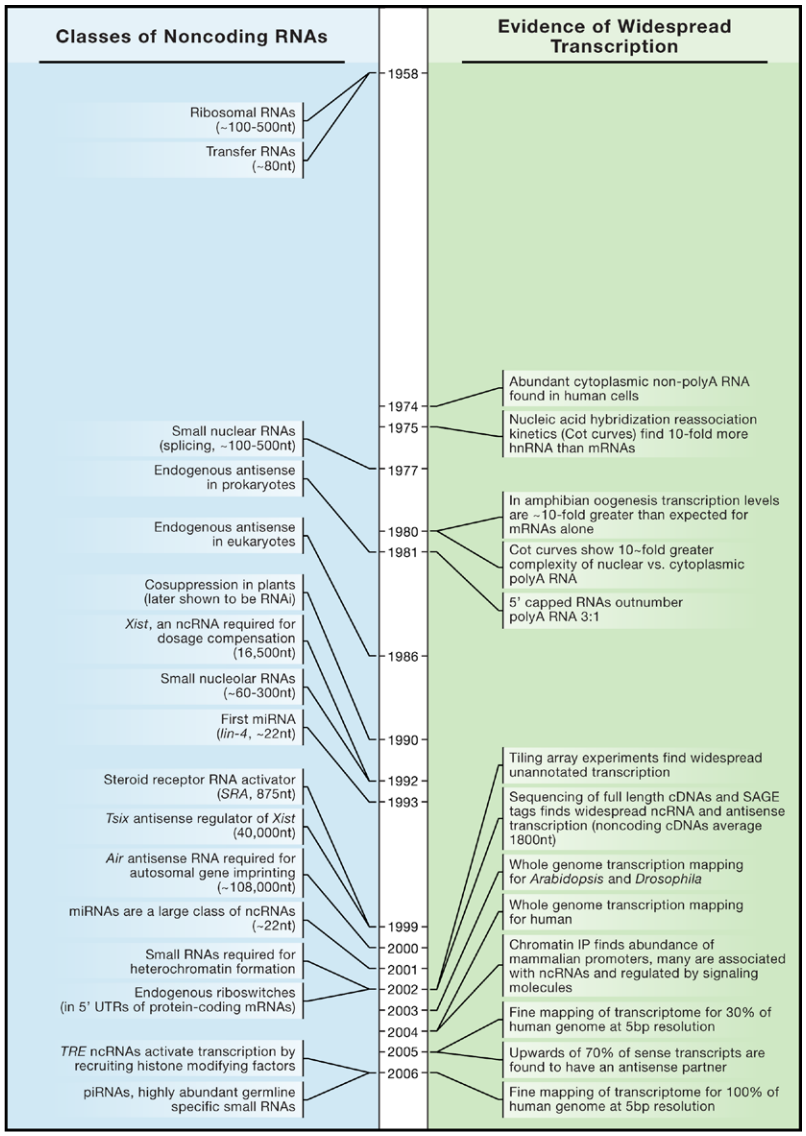

Figure 1. Discovery Timeline of Noncoding RNAs

This timeline highlights the discoveries of the different functional classes of noncoding RNAs (ncRNAs) and the emergence of evidence for widespread transcription throughout the genome. Widespread transcription produces of a plethora of unannotated RNA species, most of which are characterized by reduced protein coding potential. Examples of the many regulatory roles of ncRNAs in cellular processes are provided. Diversity in the size and function of ncRNAs coupled with the abundance of noncoding transcription suggest that many ncRNAs still await discovery. (See Table S1 for full citations for timeline events).

be incomplete. Many other functions for noncoding transcripts have been identified, including transcriptional activation, gene silencing, imprinting, dosage compensation, translational silencing, modulation of protein function, and binding as riboswitches to regulatory metabolites (reviewed in Kiss, 2002; Mattick and Makunin, 2006; Zamore and Haley, 2005).

One of the best characterized emerging classes of ncRNAs are the microRNAs (miRNAs) cloned over a decade ago in *C. elegans* and now recognized as a large conserved family of ~22-nucleotide regulatory RNAs essential for a variety of cellular processes (reviewed in Esquela-Kerscher and Slack, 2006; Zamore and Haley, 2005). The differential expression patterns of miRNAs determine cell fate and correct differentiation during development (for example, through regulation of Notch signaling and Hox gene expression). MicroRNAs can act as tumor suppressors and oncogenes (for example, by regulating Bcl2, Ras, Myc, and E2F), and they can regulate cellular proliferation and apoptosis. Surprisingly, miRNA expression profiles appear to reflect more accurately the developmental lineage and differentiation state of tumors than do mRNA profiles. Hundreds of miRNAs are estimated to be present in the human genome, and computational analysis suggests that more than 20%–30% of human genes are regulated by miRNAs. Microarray experiments support this view, revealing miRNA-mediated downregulation of large numbers of target mRNAs. In addition, miRNAs suppress initiation of protein translation, promote mRNA degradation and turnover, and initiate transcriptional silencing. However, the function of the vast majority of miRNAs is as yet unknown.

Small nucleolar RNAs (snoRNAs), another abundant group of RNAs, reveal an ever-increasing retinue of cellular functions for ncRNAs (reviewed in Kiss, 2002). Originally appreciated for their processing of rRNAs and involvement in the posttranscriptional 2′-O-methylation and pseudouridylation of target rRNAs and snRNAs, the identification of many "orphan" snoRNAs lacking complementarity to rRNAs, snRNAs, and tRNAs suggests a broad range of RNA substrates for snoRNAs. Furthermore, snoRNAs are known to accumulate (and potentially function) outside of the nucleolus. Given that most snoRNAs are processed from the introns of other genes, their expression is inextricably linked to transcription of their host gene. For example, a brain-specific snoRNA called HBII-52 modulates alternate splicing of the transcript encoding the serotonin receptor. Patients with Prader-Willi syndrome who lack HBII-52 have different serotonin

receptor splice forms, resulting in an alteration in serotonin efficacy (Kishore and Stamm, 2006).

The interaction of ncRNAs with proteins is another way in which ncRNAs can exert their effects in the cell. For example, steroid receptor RNA activator (SRA) is a transcriptional coactivator whose action requires chromatin remodeling and recruitment of histone deacetylases (Zhao et al., 2004). SRA facilitates establishment of the transcriptional preinitiation complex by interacting with nuclear receptors, probably forming a scaffold for the binding of coactivators (such as the DEAD box protein p72/p68) and corepressors (such as SHARP). A screen for evolutionarily conserved functional ncRNAs identified a noncoding repressor of the NFAT transcription factor (NRON) (Willingham et al., 2005). The demonstrated interaction of NRON with nuclear import factors immediately suggests that NRON in a complex with importin-β specifically downregulates the nuclear import of NFAT, a hypothesis supported by studies of NFAT translocation. Indeed, given the complicated networks of nuclear-cytoplasmic transport and the seemingly limited number of available importin-β family members, such ncRNA-mediated regulation of specific cargo proteins may emerge as a common biological strategy for dealing with the complexity of intracellular trafficking. Activation of the heat shock transcription factor (HSF1) is mediated by an RNP complex, which includes the translation elongation factor eEF1A and a new ncRNA called HSR1 (Shamovsky et al., 2006). Conserved from rodents to humans, HSR1 is essential for heat shock response activation and may act as an RNA thermosensor.

The identification of a specific class of small RNAs with a potential role in development has very recently been described. The Argonaute Piwi subfamily proteins have been shown to be important in germline development (Cox et al., 1998). Girard et al. (2006) and Aravin et al. (2006) have now identified and characterized a highly abundant class of small RNAs called piRNAs and showed them to be associated with the Piwi proteins MIWI (Girard et al., 2006) and MILI (Aravin et al., 2006) in murine testis. The slightly longer lengths (26–31 nucleotides) and testes-specific expression of these piRNAs perhaps hint at the start of a growing number of developmental stage-specific or tissue-specific members of new classes of small RNAs.

Finally, the burgeoning number of putative ncRNA genes has been cited as a mechanism for increasing the complexity of gene regulatory networks and, together with alternative gene splicing, significantly increases the variety and complexity of the transcriptome. The finding that the rising ratio of noncoding to protein-coding DNA correlates with increasing organismal complexity supports this notion (Mattick and Makunin, 2006). With the limited number of currently characterized ncRNAs acting in so many cellular processes and the demonstrated amount of unannotated transcription present in the genome, one can easily appreciate that ncRNAs probably function in more cellular pathways than their protein-coding brethren.

Implications of Noncoding Transcription

Noncoding RNAs possess several properties that make them attractive as regulatory factors. For example, miRNAs are 1000-fold smaller than most mRNAs. With their 20 nucleotides, miRNAs can perform the task of a protein domain composed of 100 amino acids. Considering the effects of compensatory base pair changes on structure and target hybridization, such RNAs could evolve much more readily than protein domains. The strong innate affinity of RNA for both DNA and RNA permits a diversity of stable mismatch-based secondary structure motifs such as stems, bulges, and loops. These elements may be stronger determinants of an RNA's function than the primary sequence itself, and would help to explain the lack of evolutionary conservation observed among many ncRNAs. Evolutionary conservation of even well-characterized and biologically important ncRNAs appears to be inconsistent at best. The let7 miRNA is conserved from worms to humans (Zamore and Haley, 2005), yet the dosage compensation ncRNA XIST appears to be rapidly evolving even among rodents (Nesterova et al., 2001). Furthermore, of the ~34,000 putative ncRNAs identified in mouse, only 3%–4% or so have limited human sequence conservation (Carninci et al., 2005). Only about 7%–20% (depending on the study) of human unannotated transcripts appear to be conserved over most of their lengths with their mouse counterparts; thus, the majority are species-specific (Johnson et al., 2005). Perhaps the criteria used to determine evolutionary conservation should be reevaluated. For example, it could be that the lengths of the conserved sequences are short and discontinuous, resembling islands of conservation within a larger nonconserved sequence. If such patches of conservation turn out to be biologically important for unannotated ncRNAs (as they are for miRNAs), the number of such conserved regions in the genome would escalate considerably.

Given the tolerance of RNA for mismatches and its probable reliance on secondary structure for bioactivity, ncRNAs are likely to have a greater degree of plasticity than mRNAs. Often multiple miRNAs must be deleted to obtain a phenotype (Abbott et al., 2005)—an illustration of functional redundancy and regulatory complexity that may extend to other ncRNAs. Noncoding RNAs have been implicated in a number of diseases, including B cell neoplasia, lung cancer, autism, DiGeorge syndrome, prostate cancer, and schizophrenia (see RNAdb, http://research.imb.uq.edu.au/rnadb/). Considering the abundance of empirically observed ncRNAs, it is highly likely that many more will be linked to human diseases. In addition, miRNAs linked to cancer may prove valuable therapeutic targets (Esquela-Kerscher and Slack, 2006).

There are several strategies for investigating the function of new noncoding transcripts. Starting with a large number of putative mouse ncRNAs identified by the FANTOM project, Ravasi et al. used expression analysis tools such as microarrays, PCR, and Northern blots combined with lipopolysaccharide treatment to validate the regulated expression of a subset of these ncRNAs (Ravasi et al., 2006). In a separate study but starting with the same set of mouse ncRNAs, Schultz and colleagues identified a subset of 512 ncRNAs apparently conserved during human evolution; they then used siRNAs to disrupt these conserved ncRNAs and screened for phenotypes using a panel of cellular assays (Willingham et al., 2005). Such detailed expression analyses and siRNA-mediated phenotypic screenings highlight a path for the followup characterization of newly mapped transcriptional "dark matter." Disruption of an essential RNA processing pathway such as Dicer cleavage or rRNA maturation should have dramatic transcriptional repercussions, which could be detected with tiling arrays. Such an experiment was conducted in yeast by depleting Rpp1, an essential protein required for both RNase P-mediated maturation of tRNAs and RNase MRP, which processes rRNA. This study revealed 74 new ncRNAs, many of which were antisense to known protein coding genes (Samanta et al., 2006). Future efforts could focus on: (1) combined use of RNase footprinting and tiling arrays for genomewide identification of "protected" RNAs; (2) purification of RNA-associated proteins such as helicases and identification of their associated RNAs by tiling array; (3) classification of subcellular and structural fractionations of RNAs using tiling arrays; and (4) development of higher-density tiling arrays that allow high-resolution genomewide surveys of transcription, permitting investigation of rare tissues, primary cells derived from tissues, developmental time-courses, and other low abundance samples.

How much of the nonredundant genome is transcribed? Based on published data, estimates range from 10% to 60%. However, this may be an underestimate given the limited number of cells and differentiation states surveyed thus far. Furthermore, array-based transcript mapping relies on conservative thresholds which select for the highest 2%–5% of probes, yielding a low false positive rate of a few percent. Given the low expression levels of many TUFs, these thresholds are likely to be underestimates of the true amount of unannotated transcription. Indeed, unpublished data from our lab suggest that 75% of RACE products selected from random regions of the human genome and hybridized to microarrays reveal the presence of complex transcripts. Analyzing the distribution of transposable elements that are counter-selected when they occur within functional transcriptional units (FTUs) implies that >50% of the genome consists of FTUs, and one-third of these are likely to be ncRNAs (Semon and Duret, 2004). Weighing these factors together, we suggest that all of the non-repeat portions of the human genome are transcribed. This may seem an excessive estimate, yet recent data in yeast imply that more that 85% of its genome is transcribed (David et al., 2006). Furthermore, large-scale cDNA sequencing and annotation in the mouse has shown that 62% of the mouse genome is transcribed (Carninci et al., 2005), and there are estimates that 90% of the human genome is transcribed (Wong et al., 2001). The overall architecture of transcribed portions of the genome is highly complex. Indeed, the landscape of most transcriptomes is a lattice-like network of overlapping transcription in which the same genomic sequences often serve as portions of separately regulated transcripts, making the boundaries and indeed the concept of the term gene less useful than it once was.

Supplemental Data

The Supplemental Data for this article, including Table S1, can be found online at http://www.cell.com/cgi/content/full/125/7/1215/DC1/.

ACKNOWLEDGMENTS

We thank Philip Kapranov for helpful discussions.

REFERENCES

Abbott, A.L., Alvarez-Saavedra, E., Miska, E.A., Lau, N.C., Bartel, D.P., Horvitz, H.R., and Ambros, V. (2005). Dev. Cell *9*, 403–414.

Aravin, A., Gaidatzis, D., Pfeffer, S., Lagos-Quintana, M., Landgraf, P., Iovino, N., Morris, P., Brownstein, M.J., Kuramochi-Miyagawa, S., Nakano, T., et al. (2006). Nature. Published online June 4, 2006. 10.1038/nature04916.

Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M.C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C., et al. (2005). Science *309*, 1559–1563.

Cawley, S., Bekiranov, S., Ng, H.H., Kapranov, P., Sekinger, E.A., Kampa, D., Piccolboni, A., Sementchenko, V., Cheng, J., Williams, A.J., et al. (2004). Cell *116*, 499–509.

Cheng, J., Kapranov, P., Drenkow, J., Dike, S., Brubaker, S., Patel, S., Long, J., Stern, D., Tammana, H., Helt, G., et al. (2005). Science *308*, 1149–1154.

Cox, D.N., Chao, A., Baker, J., Chang, L., Qiao, D., and Lin, H. (1998). Genes Dev. *12*, 3715–3727.

David, L., Huber, W., Granovskaia, M., Toedling, J., Palm, C.J., Bofkin, L., Jones, T., Davis, R.W., and Steinmetz, L.M. (2006). Proc. Natl. Acad. Sci. USA *103*, 5320–5325.

Esquela-Kerscher, A., and Slack, F.J. (2006). Nat. Rev. Cancer *6*, 259–269.

Girard, A., Sachidanandam, R., Hannon, G.J., and Carmell, M.A. (2006). Nature. Published online June 4, 2006. 10.1038/nature04917.

Hirotsune, S., Yoshida, N., Chen, A., Garrett, L., Sugiyama, F., Takahashi, S., Yagami, K., Wynshaw-Boris, A., and Yoshiki, A. (2003). Nature *423*, 91–96.

Johnson, J.M., Edwards, S., Shoemaker, D., and Schadt, E.E. (2005). Trends Genet. *21*, 93–102.

Katayama, S., Tomaru, Y., Kasukawa, T., Waki, K., Nakanishi, M., Nakamura, M., Nishida, H., Yap, C.C., Suzuki, M., Kawai, J., et al. (2005). Science *309*, 1564–1566.

Kim, T.H., Barrera, L.O., Zheng, M., Qu, C., Singer, M.A., Richmond, T.A., Wu, Y., Green, R.D., and Ren, B. (2005). Nature *436*, 876–880.

Kishore, S., and Stamm, S. (2006). Science *311*, 230–232.

Kiss, T. (2002). Cell *109*, 145–148.

Martone, R., Euskirchen, G., Bertone, P., Hartman, S., Royce, T.E., Luscombe, N.M., Rinn, J.L., Nelson, F.K., Miller, P., Gerstein, M., et al. (2003). Proc. Natl. Acad. Sci. USA *100*, 12247–12252.

Mattick, J.S., and Makunin, I.V. (2006). Hum. Mol. Genet. *15* (*Suppl 1*), R17–R29.

Nesterova, T.B., Slobodyanyuk, S.Y., Elisaphenko, E.A., Shevchenko, A.I., Johnston, C., Pavlova, M.E., Rogozin, I.B., Kolesnikov, N.N., Brockdorff, N., and Zakian, S.M. (2001). Genome Res. *11*, 833–849.

Nobrega, M.A., Zhu, Y., Plajzer-Frick, I., Afzal, V., and Rubin, E.M. (2004). Nature *431*, 988–993.

Ravasi, T., Suzuki, H., Pang, K.C., Katayama, S., Furuno, M., Okunishi, R., Fukuda, S., Ru, K., Frith, M.C., Gongora, M.M., et al. (2006). Genome Res. *16*, 11–19.

Samanta, M.P., Tongprasit, W., Sethi, H., Chin, C.S., and Stolc, V. (2006). Proc. Natl. Acad. Sci. USA *103*, 4192–4197.

Semon, M., and Duret, L. (2004). Trends Genet. *20*, 229–232.

Shamovsky, I., Ivannikov, M., Kandel, E.S., Gershon, D., and Nudler, E. (2006). Nature *440*, 556–560.

Sontheimer, E.J., and Carthew, R.W. (2005). Cell *122*, 9–12.

Willingham, A.T., Orth, A.P., Batalov, S., Peters, E.C., Wen, B.G., Aza-Blanc, P., Hogenesch, J.B., and Schultz, P.G. (2005). Science *309*, 1570–1573.

Wong, G.K., Passey, D.A., and Yu, J. (2001). Genome Res. *11*, 1975–1977.

Zamore, P.D., and Haley, B. (2005). Science *309*, 1519–1524.

Zhao, X., Patton, J.R., Davis, S.L., Florence, B., Ames, S.J., and Spanjaard, R.A. (2004). Mol. Cell *15*, 549–558.

Supplemental Data

TUF Love for "Junk" DNA

Aarron T. Willingham and Thomas R. Gingeras

Table S1.

timeline	date	category	references
abundant cytoplasmic poly(A)- RNA found in human cells (subsequently observed in a variety of plant and animal cells)	1974	widespread transcription	(Milcarek et al., 1974; Snider and Morrison-Bogorad, 1992)
nucleic acid hybridization reassociation kinetics (Cot curves) find 10-fold more hnRNA than mRNA	1975	widespread transcription	(Hough et al., 1975)
in amphibian oogenesis, transcription levels are ~10-fold greater than expected for mRNAs alone	1980	widespread transcription	(Varley et al., 1980)
Cot curves show 10-fold greater complexity of nuclear vs. cytoplasmic poly(A)+ RNA	1980	widespread transcription	(Holland et al., 1980)
5` capped RNAs outnumber poly(A)+ 3-to-1	1981	widespread transcription	(Salditt-Georgieff et al., 1981)
tiling array experiments find widespread unannotated transcription	2002	widespread transcription	(Kapranov et al., 2002)
sequencing of full-length cDNAs and SAGE tags finds widespread ncRNAs and antisense transcription (noncoding cDNAs average 1800nt)	2002	widespread transcription	(Chen et al., 2002; Okazaki et al., 2002; Saha et al., 2002)
whole genome transcription mapping for *Arabidopsis* and *Drosophila*	2003-2004	widespread transcription	(Stolc et al., 2004; Yamada et al., 2003)
whole genome transcription mapping for human	2004	widespread transcription	(Bertone et al., 2004)
chromatin immunoprecipitation (ChIP) studies of mammalian promoters find many ncRNAs are regulated by transcription factors and signaling molecules	2004	widespread transcription	(Cawley et al., 2004)
fine-mapping of transcriptome for 30% of human genome at 5bp resolution	2005	widespread transcription	(Cheng et al., 2005)
upwards of 70% of sense transcripts are found to have an antisense partner	2005	widespread transcription	(Katayama et al., 2005)
fine-mapping of transcriptome for 100% of human genome at 5bp resolution	2006	widespread transcription	manuscript in preparation
ribosomal RNAs (~100-5000nt)	1958	ncRNAs	(Crick, 1958)
transfer RNAs (~80nt)	1958	ncRNAs	(Hoagland et al., 1958)
small nuclear RNAs (splicing) (~100-500nt)	1977	ncRNAs	(Benecke and Penman, 1977; Goldstein et al.,

			1977)
endogenous antisense in prokaryotes	1981	ncRNAs	(Rosen et al., 1981; Tomizawa and Itoh, 1981)
endogenous antisense in eukaryotes	1986	ncRNAs	(Adeniyi-Jones and Zasloff, 1985; Nepveu and Marcu, 1986; Spencer et al., 1986; Williams and Fried, 1986)
cosuppression in plants (later shown to be RNAi-based)	1990	ncRNAs	(Napoli et al., 1990; van der Krol et al., 1990)
Xist, a ncRNA required for dosage compensation (16,500nt)	1992	ncRNAs	(Brockdorff et al., 1992; Brown et al., 1992)
small nucleolar RNAs (~60-300nt)	1992	ncRNAs	(Leverette et al., 1992)
first miRNA (lin-4) (~22nt)	1993	ncRNAs	(Lee et al., 1993)
steroid receptor RNA activator (SRA) (875nt)	1999	ncRNAs	(Lanz et al., 1999)
Tsix, antisense regulator of Xist (40,000nt)	1999	ncRNAs	(Lee et al., 1999)
Air, antisense RNA required for autosomal gene imprinting (~108,000nt)	2000	ncRNAs	(Lyle et al., 2000)
miRNAs found to be a large class of ncRNAs (~22nt)	2001	ncRNAs	(Lee and Ambros, 2001)
small RNAs required for heterochromatin formation	2002	ncRNAs	(Reinhart and Bartel, 2002; Volpe et al., 2002)
endogenous riboswitches (in 5` UTR of protein-coding mRNAs)	2002	ncRNAs	(Mironov et al., 2002; Winkler et al., 2002)
TRE ncRNAs activate transcription by recruiting histone modifying factors	2006	ncRNAs	(Sanchez-Elsner et al., 2006)
piRNAs, an abundant testes specific class of small RNAs with a role in germline development	2006	ncRNAs	(Aravin et al., 2006; Girard et al., 2006)

Additional noncoding RNA data available from RNAdb (Pang et al., 2005) and Rfam (Griffiths-Jones et al., 2005) online databases.

References:

Adeniyi-Jones, S., and Zasloff, M. (1985). Transcription, processing and nuclear transport of a B1 Alu RNA species complementary to an intron of the murine alpha-fetoprotein gene. Nature *317*, 81-84.

Aravin, A., Gaidatzis, D., Pfeffer, S., Lagos-Quintana, M., Landgraf, P., Iovino, N., Morris, P., Brownstein, M. J., Kuramochi-Miyagawa, S., Nakano, T., *et al.* (2006). A novel class of small RNAs bind to MILI protein in mouse testes. Nature *doi:10.1038/nature04916*.

Benecke, B. J., and Penman, S. (1977). A new class of small nuclear RNA molecules synthesized by a type I RNA polymerase in HeLa cells. Cell *12*, 939-946.

Bertone, P., Stolc, V., Royce, T. E., Rozowsky, J. S., Urban, A. E., Zhu, X., Rinn, J. L., Tongprasit, W., Samanta, M., Weissman, S., *et al.* (2004). Global identification of human transcribed sequences with genome tiling arrays. Science *306*, 2242-2246.

Brockdorff, N., Ashworth, A., Kay, G. F., McCabe, V. M., Norris, D. P., Cooper, P. J., Swift, S., and Rastan, S. (1992). The product of the mouse Xist gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus. Cell *71*, 515-526.

Brown, C. J., Hendrich, B. D., Rupert, J. L., Lafreniere, R. G., Xing, Y., Lawrence, J., and Willard, H. F. (1992). The human XIST gene: analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. Cell *71*, 527-542.

Cawley, S., Bekiranov, S., Ng, H. H., Kapranov, P., Sekinger, E. A., Kampa, D., Piccolboni, A., Sementchenko, V., Cheng, J., Williams, A. J., *et al.* (2004). Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. Cell *116*, 499-509.

Chen, J., Sun, M., Lee, S., Zhou, G., Rowley, J. D., and Wang, S. M. (2002). Identifying novel transcripts and novel genes in the human genome by using novel SAGE tags. Proc Natl Acad Sci U S A *99*, 12257-12262.

Cheng, J., Kapranov, P., Drenkow, J., Dike, S., Brubaker, S., Patel, S., Long, J., Stern, D., Tammana, H., Helt, G., *et al.* (2005). Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. Science *308*, 1149-1154.

Crick, F. H. (1958). On protein synthesis. Symp Soc Exp Biol *12*, 138-163.

Girard, A., Sachidanandam, R., Hannon, G. J., and Carmell, M. A. (2006). A germline-specific class of small RNAs binds mammalian Piwi proteins. Nature *doi:10.1038/nature04917*.

Goldstein, L., Wise, G. E., and Ko, C. (1977). Small nuclear RNA localization during mitosis. An electron microscope study. J Cell Biol *73*, 322-331.

Griffiths-Jones, S., Moxon, S., Marshall, M., Khanna, A., Eddy, S. R., and Bateman, A. (2005). Rfam: annotating non-coding RNAs in complete genomes. Nucleic Acids Res *33 Database Issue*, D121-124.

Hoagland, M. B., Stephenson, M. L., Scott, J. F., Hecht, L. I., and Zamecnik, P. C. (1958). A soluble ribonucleic acid intermediate in protein synthesis. J Biol Chem *231*, 241-257.

Holland, C. A., Mayrand, S., and Pederson, T. (1980). Sequence complexity of nuclear and messenger RNA in HeLa cells. J Mol Biol *138*, 755-778.

Hough, B. R., Smith, M. J., Britten, R. J., and Davidson, E. H. (1975). Sequence complexity of heterogeneous nuclear RNA in sea urchin embryos. Cell *5*, 291-299.

Kapranov, P., Cawley, S. E., Drenkow, J., Bekiranov, S., Strausberg, R. L., Fodor, S. P., and Gingeras, T. R. (2002). Large-scale transcriptional activity in chromosomes 21 and 22. Science *296*, 916-919.

Katayama, S., Tomaru, Y., Kasukawa, T., Waki, K., Nakanishi, M., Nakamura, M., Nishida, H., Yap, C. C., Suzuki, M., Kawai, J., *et al.* (2005). Antisense transcription in the mammalian transcriptome. Science *309*, 1564-1566.

Lanz, R. B., McKenna, N. J., Onate, S. A., Albrecht, U., Wong, J., Tsai, S. Y., Tsai, M. J., and O'Malley, B. W. (1999). A steroid receptor coactivator, SRA, functions as an RNA and is present in an SRC-1 complex. Cell *97*, 17-27.

Lee, J. T., Davidow, L. S., and Warshawsky, D. (1999). Tsix, a gene antisense to Xist at the X-inactivation centre. Nat Genet *21*, 400-404.

Lee, R. C., and Ambros, V. (2001). An extensive class of small RNAs in Caenorhabditis elegans. Science *294*, 862-864.

Lee, R. C., Feinbaum, R. L., and Ambros, V. (1993). The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell *75*, 843-854.

Leverette, R. D., Andrews, M. T., and Maxwell, E. S. (1992). Mouse U14 snRNA is a processed intron of the cognate hsc70 heat shock pre-messenger RNA. Cell *71*, 1215-1221.

Lyle, R., Watanabe, D., te Vruchte, D., Lerchner, W., Smrzka, O. W., Wutz, A., Schageman, J., Hahner, L., Davies, C., and Barlow, D. P. (2000). The imprinted antisense RNA at the Igf2r locus overlaps but does not imprint Mas1. Nat Genet *25*, 19-21.

Milcarek, C., Price, R., and Penman, S. (1974). The metabolism of a poly(A) minus mRNA fraction in HeLa cells. Cell *3*, 1-10.

Mironov, A. S., Gusarov, I., Rafikov, R., Lopez, L. E., Shatalin, K., Kreneva, R. A., Perumov, D. A., and Nudler, E. (2002). Sensing small molecules by nascent RNA: a mechanism to control transcription in bacteria. Cell *111*, 747-756.

Napoli, C., Lemieux, C., and Jorgensen, R. (1990). Introduction of a Chimeric Chalcone Synthase Gene into Petunia Results in Reversible Co-Suppression of Homologous Genes in trans. Plant Cell *2*, 279-289.

Nepveu, A., and Marcu, K. B. (1986). Intragenic pausing and anti-sense transcription within the murine c-myc locus. Embo J *5*, 2859-2865.

Okazaki, Y., Furuno, M., Kasukawa, T., Adachi, J., Bono, H., Kondo, S., Nikaido, I., Osato, N., Saito, R., Suzuki, H., *et al.* (2002). Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature *420*, 563-573.

Pang, K. C., Stephen, S., Engstrom, P. G., Tajul-Arifin, K., Chen, W., Wahlestedt, C., Lenhard, B., Hayashizaki, Y., and Mattick, J. S. (2005). RNAdb--a comprehensive mammalian noncoding RNA database. Nucleic Acids Res *33*, D125-130.

Reinhart, B. J., and Bartel, D. P. (2002). Small RNAs correspond to centromere heterochromatic repeats. Science *297*, 1831.

Rosen, J., Ryder, T., Ohtsubo, H., and Ohtsubo, E. (1981). Role of RNA transcripts in replication incompatibility and copy number control in antibiotic resistance plasmid derivatives. Nature *290*, 794-797.

Saha, S., Sparks, A. B., Rago, C., Akmaev, V., Wang, C. J., Vogelstein, B., Kinzler, K. W., and Velculescu, V. E. (2002). Using the transcriptome to annotate the genome. Nat Biotechnol *20*, 508-512.

Salditt-Georgieff, M., Harpold, M. M., Wilson, M. C., and Darnell, J. E., Jr. (1981). Large heterogeneous nuclear ribonucleic acid has three times as many 5' caps as polyadenylic acid segments, and most caps do not enter polyribosomes. Mol Cell Biol *1*, 179-187.

Sanchez-Elsner, T., Gou, D., Kremmer, E., and Sauer, F. (2006). Noncoding RNAs of trithorax response elements recruit Drosophila Ash1 to Ultrabithorax. Science *311*, 1118-1123.

Snider, B. J., and Morrison-Bogorad, M. (1992). Brain non-adenylated mRNAs. Brain Res Brain Res Rev *17*, 263-282.

Spencer, C. A., Gietz, R. D., and Hodgetts, R. B. (1986). Overlapping transcription units in the dopa decarboxylase region of Drosophila. Nature *322*, 279-281.

Stolc, V., Gauhar, Z., Mason, C., Halasz, G., van Batenburg, M. F., Rifkin, S. A., Hua, S., Herreman, T., Tongprasit, W., Barbano, P. E., *et al.* (2004). A gene expression map for the euchromatic genome of Drosophila melanogaster. Science *306*, 655-660.

Tomizawa, J., and Itoh, T. (1981). Plasmid ColE1 incompatibility determined by interaction of RNA I with primer transcript. Proc Natl Acad Sci U S A *78*, 6096-6100.

van der Krol, A. R., Mur, L. A., Beld, M., Mol, J. N., and Stuitje, A. R. (1990). Flavonoid genes in petunia: addition of a limited number of gene copies may lead to a suppression of gene expression. Plant Cell *2*, 291-299.

Varley, J. M., Macgregor, H. C., and Erba, H. P. (1980). Satellite DNA is transcribed on lampbrush chromosomes. Nature *283*, 686-688.

Volpe, T. A., Kidner, C., Hall, I. M., Teng, G., Grewal, S. I., and Martienssen, R. A. (2002). Regulation of heterochromatic silencing and histone H3 lysine-9 methylation by RNAi. Science *297*, 1833-1837.

Williams, T., and Fried, M. (1986). A mouse locus at which transcription from both DNA strands produces mRNAs complementary at their 3' ends. Nature *322*, 275-279.

Winkler, W., Nahvi, A., and Breaker, R. R. (2002). Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression. Nature *419*, 952-956.

Yamada, K., Lim, J., Dale, J. M., Chen, H., Shinn, P., Palm, C. J., Southwick, A. M., Wu, H. C., Kim, C., Nguyen, M., *et al.* (2003). Empirical analysis of transcriptional activity in the Arabidopsis genome. Science *302*, 842-846.