frontiers
RESEARCH TOPICS

GENOMIC "DARK MATTER": IMPLICATIONS FOR UNDERSTANDING HUMAN DISEASE MECHANISMS, DIAGNOSTICS, AND CURES

Hosted by
Philipp Kapranov

frontiers in
GENETICS

FRONTIERS COPYRIGHT STATEMENT

Cover image provided by Ibbl sarl, Lausanne CH

ISSN 1664-8714
ISBN 978-2-88919-053-9
DOI 10.3389/978-2-88919-053-9

ABOUT FRONTIERS

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

FRONTIERS JOURNAL SERIES

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing.

All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

DEDICATION TO QUALITY

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.

Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view.

By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

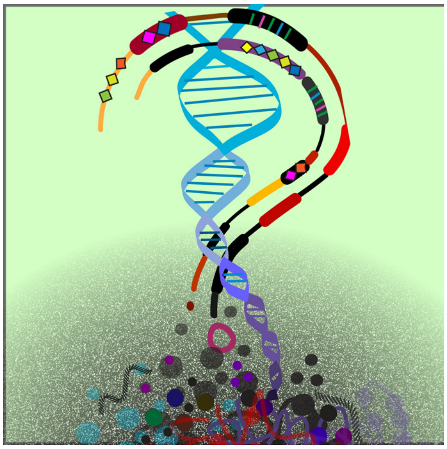
WHAT ARE FRONTIERS RESEARCH TOPICS?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area!

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: researchtopics@frontiersin.org

GENOMIC "DARK MATTER": IMPLICATIONS FOR UNDERSTANDING HUMAN DISEASE MECHANISMS, DIAGNOSTICS, AND CURES

Hosted By:
Philipp Kapranov, St. Laurent Institute, USA



The vast majority of the human genome has been historically ignored from the point of view of molecular mechanisms of disease, diagnostics and potential therapeutic targets. The predominant focus of disease research has traditionally been placed on the protein-coding regions of the human genome, which account for only ~4-5% of its total sequence complexity. This bias has an obvious underlying reason: protein-coding regions encode a crucial class of molecules in a cell, whose function and importance are well established. Furthermore, proteins are the predominant class of cellular molecules against which effective therapeutics can be designed. This bias pervades the design of analytical tools made to measure DNA, DNA-protein interactions, as well as procedures used to measure and annotate transcriptome expression. Microarrays for example, are often biased to the regions of genome known to encode exons or promoters of protein-coding mRNAs. Other aspects of our approach towards measuring expression of RNAs such as the typical choice of polyA+ RNA selection, enriched in mRNAs, for next generation sequencing also reinforces this bias. In summary, the 2-3% of the genome and RNAs made from it has dominated the conceptual thinking of academic and medical communities as well as industries that make devices that measure nucleic acids for research or diagnostic purposes and the pharmaceutical industry.

However, during the last decade a tide of data has gained sufficient momentum to suggest that the cell actually uses the remaining 97-98% of the genome to produce stable RNAs – the so-called "dark matter" RNA. The first reports to suggest this were based on tiling array technology and sequencing of ESTs, which while powerful, had their limitations: tiling arrays could not estimate the relative mass of the RNAs produced from the non-protein coding regions in a cell and the EST sequencing methods were not deep enough. The advent of next-generation sequencing, in particular, single-molecule sequencing has allowed us not only confirm the previous observations but also for the first time to estimate not only from where, but also how much non-exonic RNA is produced. Its fraction of the total transcriptome is quite significant, up to 2/3 of all RNA made in a human cell (http://www.biomedcentral.com/1741-7007/8/149). Moreover, the non-exonic RNAs are differentially expressed in disease: for example, between the primary tumors and metastatic derivatives. We believe that the logical next step from these observations is to ask three questions, perhaps some of the most important questions of our time in biomedical science: (1) do "dark matter" RNAs underlie mechanisms of human disease?; (2) Can they be used for diagnostics?; and (3) Can they be used as targets for therapeutics?.

We thus would like to propose a Research Topic in the Frontiers in Genetics/Frontiers in Non-Coding RNAs that is specifically dedicated to publishing manuscripts addressing these three questions.

Table of Contents

05 ***Genomic "Dark Matter": Implications for Understanding Human Disease Mechanisms, Diagnostics, and Cures***
Philipp Kapranov and Georges St. Laurent

07 ***Dark Matter RNA: Existence, Function, and Controversy***
Philipp Kapranov and Georges St. Laurent

16 ***Thousands of Novel Transcripts Identified in Mouse Cerebrum, Testis, and ES Cells Based on Ribo-Minus RNA Sequencing***
Wanfei Liu, Yuhui Zhao, Peng Cui, Qiang Lin, Feng Ding, Chengqi Xin, Xinyu Tan, Shuhui Song, Jun Yu and Songnian Hu

33 ***The Long Non-Coding RNAs: A New (P)layer in the "Dark Matter"***
Thomas Derrien, Roderic Guigó and Rory Johnson

39 ***Trans-splicing in Higher Eukaryotes: Implications for Cancer Development?***
Peter G. Zaphiropoulos

43 ***The Beginning of the Road for Non-Coding RNAs in Normal Hematopoiesis and Hematologic Malignancies***
Elisabeth F. Heuston, Kenya T. Lemon and Robert J. Arceci

52 ***Genomic "Dark Matter" in Prostate Cancer: Exploring the Clinical Utility of ncRNA as Biomarkers***
Ismael A. Vergara, Nicholas Erho, Timothy J. Triche, Mercedeh Ghadessi, Anamaria Crisan, Thomas Sierocinski, Peter C. Black, Christine Buerki and Elai Davicioni

62 ***Perspectives of Long Non-Coding RNAs in Cancer Diagnostics***
Eduardo M. Reis and Sergio Verjovski-Almeida

70 ***A Central Role for Long Non-Coding RNA in Cancer***
Sheetal A. Mitra, Anirban P. Mitra and Timothy J. Triche

79 ***Characterizing ncRNAs in Human Pathogenic Protists Using High-Throughput Sequencing Technology***
Lesley Joan Collins

89 ***A Network of Regulations by Small Non-Coding RNAs: The P-TEFb Kinase in Development and Pathology***
Hossein Ghanbarian, Valérie Grandjean, François Cuzin and Minoo Rassoulzadegan

95 ***Dark matter RNA: An Intelligent Scaffold for the Dynamic Regulation of the Nuclear Information Landscape***
Georges St. Laurent, Yiannis A. Savva and Philipp Kapranov

frontiers in
GENETICS

EDITORIAL
published: 29 May 2012
doi: 10.3389/fgene.2012.00095

Genomic "dark matter": implications for understanding human disease mechanisms, diagnostics, and cures

Philipp Kapranov* and Georges St. Laurent*

St. Laurent Institute, Cambridge, MA, USA
**Correspondence: philippk08@gmail.com; georgest98@yahoo.com*

WHAT IS GENOMIC "DARK MATTER?"

The realization that protein-coding genes use only a tiny fraction of the three billion base pairs that make up the human genome has given birth to perhaps the largest and most persistent question in modern genetics: of what use, if any, is the vast non-coding sequences that we all carry in each of our cells. Is it really non-functional "junk" DNA as referred to by some, or does it provide the basis for the blueprint for organismal complexity and cellular information processing, as argued by others? The massive expanse of the non-coding portion of the genome, combined with the current technological and analytical limitations inherent to its functional analysis, has resulted in a mass of conflicting ideas and conclusions. Collectively this has created an aura of mystery and doubt surrounding it, leading to the label of genomic "dark matter." In a manner analogous to the "dark matter" of the universe, it is something that we can neither easily detect nor understand, but that nonetheless exists and is open to careful experimental queries.

DOES IT HAVE FUNCTION?

Classical approaches, such as sequence conservation and mutagenesis, have been unable to address the question of the functionality in the non-coding realm. The simplest explanation for these observations is that the non-coding portion of the genome lacks function, but is instead a neutral passenger on the evolutionary journey. However, this leaves us with the intellectually unsatisfying conclusion that most of our genome exists for no reason. Function resonates well with an important aspect of the mysteries surrounding genomic "dark matter" – most of it is used to produce RNA. Moreover, this "dark matter" RNA is not just a minor fraction of the cell's RNA, but rather makes up a majority of it (not counting the ribosomal or mitochondrial RNAs). RNA production is a sign of a functional DNA sequence and even more so if such RNA is abundant. While encouraging, our knowledge of the "dark matter" RNA is still rather limited in large measure due to the fact that most RNA analysis endeavors so far have focused on polyA + RNA, while detection of "dark matter" transcripts requires total RNA presumably because they either tend to be polyA – or somehow lost during polyA-selection process. Thus, this realm is ripe for discoveries, with brain and embryonic tissues likely (based on analysis of protein-coding mRNAs) harboring some of the richest reservoirs of novel "dark matter" transcripts. It is also worth noting that exons of well-characterized protein-coding transcripts can be found in unusual arrangements linking for example very distant genes, potentially due to trans-splicing, adding to the repertoire of un-annotated transcripts whose function we only now begin to understand.

HOW COULD IT FUNCTION?

It is generally assumed that "dark matter" RNAs do not code for proteins and function via regulation of expression of other loci. Two very general themes of non-coding RNA-mediated regulation became prominent in the recent decade: modulation of chromatin state via association with chromatin modulator complexes and, the production of small RNAs from longer non-coding RNAs to regulate various layers of transcript expression. Discovery of RNA-mediated non-Mendelian inheritance of an epigenetic change in mammals uncovered a new tantalizing possibility for RNA function. It is worth noting that the basic assumption has been challenged recently by evidence suggesting existence of plethora of short peptides produced by the "dark matter" RNAs, even though they could conceivably be products of non-functional translation of bona fide non-coding RNAs. However, if the last decade has taught us anything, its that a given locus can produce a variety of different RNAs, that can be thought of as a "transcriptional forest" as coined by the FANTOM consortium researchers, and such RNAs could well have different functions in a cell.

DOES IT BOOST HUMAN NERVOUS SYSTEM COMPLEXITY?

The genomes of humans and flies have approximately the same information complexity – on the order of ~20K protein-coding genes, just a couple fold higher than that of yeast. Some additional reservoir of complexity should exist. The human nervous system contains widespread expression of "dark mater" RNAs, distributed in highly articulated intracellular and cell specific patterns. Over this decade, investigations have revealed a stream of more and more striking functions in the nervous system, including the recent demonstration that LINE1 transposons result in somatic diversity within the neurons of individual humans.

COULD IT HAVE IMPLICATIONS FOR HUMAN HEALTH?

Even if one were to assume the worst-case scenario where most of the "dark matter" RNA is not functional at all, we now know that it is highly cell-type specific and this opens a wide area of additional diagnostic biomarkers based on these RNAs. Indeed, first reports showing that profiling "dark matter" RNAs offers superior diagnostic and prognostic information compared to protein-coding genes are now starting to appear, particularly in the cancer field which is leading others in development of these RNAs as biomarkers. However, if some or all, of the "dark matter" RNA is indeed functional then we can only imagine the plethora of secrets that are locked within that can shed light on basic mechanisms of development, homeostasis, and disease that still await exploration. For example, its very curious that a (very) long non-coding RNA could be highly restricted to a particular type of cancer and begs a question of why that would be the case if it had no functional role in the disease. Illuminating these mechanisms may help approach the contemporary challenge of understanding molecular and cellular biology from a new, holistic perspective, potentially revealing novel key aspects of cellular systems integration pathways. And, this is not necessarily limited to human cells, as other organisms, including those medically important to our health also likely possess their "dark matter" RNAs and its mysteries. However, one cannot over-emphasize the fact that we are the very beginning of the process of understanding what kind of RNAs are produced by a cell and what functionality they may have. Still, the possibility that almost entirely unexplored treasure-trove of biological information is buried within our reach is too tantalizing to ignore.

Received: 19 April 2012; accepted: 09 May 2012; published online: 29 May 2012.
Citation: Kapranov P and St. Laurent G (2012) Genomic "dark matter": implications for understanding human disease mechanisms, diagnostics, and cures. Front. Gene. ***3****:95. doi: 10.3389/fgene.2012.00095*
This article was submitted to Frontiers in Non-Coding RNA, a specialty of Frontiers in Genetics.

frontiers in
GENETICS

HYPOTHESIS AND THEORY ARTICLE
published: 23 April 2012
doi: 10.3389/fgene.2012.00060

Dark matter RNA: existence, function, and controversy

Philipp Kapranov* and Georges St. Laurent*

St. Laurent Institute, Cambridge, MA, USA

Edited by:
Cheng-Han Huang, New York Blood Center, USA

Reviewed by:
Cheng-Han Huang, New York Blood Center, USA
Vladimir Benes, European Molecular Biology Laboratory, Germany
Eduardo M. Reis, Universidade de Sao Paulo, Brazil

****Correspondence:***
Philipp Kapranov and Georges St. Laurent, St. Laurent Institute, One Kendall Square Suite 200LL, Cambridge, MA 02139, USA.
e-mail: philippk08@gmail.com; georgest98@yahoo.com

The mysteries surrounding the ~97–98% of the human genome that does not encode proteins have long captivated imagination of scientists. Does the protein-coding, 2–3% of the genome carry the 97–98% as a mere passenger and neutral "cargo" on the evolutionary path, or does the latter have biological function? On one side of the debate, many commentaries have referred to the non-coding portion of the genome as "selfish" or "junk" DNA (Orgel and Crick, 1980), while on the other side, authors have argued that it contains the real blueprint for organismal development (Penman, 1995; Mattick, 2003), and the mechanisms of developmental complexity. Thus, this question could be referred to without much exaggeration as the most important issue in genetics today.

Keywords: dark matter RNA, genomics, transcriptome, non-coding, intronic RNA, vlinc, linc, gene

Historically, genetic approaches have very successfully determined the function of a variety of biologically important regions of a genome (usually called "genes"), based on necessary and sufficient linkage between relatively obvious alterations in a phenotype(s) and specific changes in nucleotide sequence. The vast majority of sequences identified in the genetic screens do correspond to protein-coding portions of the genome. For example, most of the changes associated with simple Mendelian genetic diseases harbor mutations in exons of protein-coding genes or in the sequences that prevent their proper assembly into mature transcripts (near splice junctions; Cooper et al., 1995). Thus, at face value at least, the non-coding portions of the genome do not really seem to represent a reservoir of biologically or medically relevant sequences.

However, this interpretation lacks intellectual closure, primarily because of its counter-intuitive conclusion that almost all of the DNA in every cell of our body has no function. Upon closer examination, a number of reasons exist to explain why the traditional genetics methods did not uncover the genotype–phenotype relationships in the non-coding portions of the genome (Mattick, 2009). For example, in addition to the simple fact that protein-coding regions have traditionally been the primary focus of forward genetic screens, alteration of non-coding, presumably regulatory regions, may impart more subtle phenotypes than coding regions, which cause catastrophic component damage. Non-coding regions could have a higher tolerance to sequence changes compared to protein-coding regions, or a higher redundancy within cellular machineries, functioning as a major substrate for evolutionary innovation and phenotypic radiation.

Answering the basic question of the functionality of the non-coding portion of the genome has shifted more toward molecular methods, specifically toward measuring the primary output of the genome, the RNA. At its core, the central premise behind these endeavors relies on the following concept: the only functional "products" of a DNA sequence that we can identify are copies of itself, either in the form of an RNA molecule or a DNA molecule. Copying of DNA into DNA ensures replication, cell division, and DNA repair, while copying of DNA into RNA transmits information into cellular actions. Even if a regulatory DNA sequence does not directly encode RNA – its function is still measured by the eventual production of RNA from somewhere in the genome. And, while cellular processes could affect the function of a sequence of DNA in many different ways, by either covalent modification of its bases or non-covalent interaction with a plethora of DNA binding proteins, RNA output remains the only known way for a cell to use DNA-encoded information. The central posit of this concept implies that if a sequence of DNA participates in the production of some RNA or affects the quantity or type of the RNA produced, then this sequence can be functional if the RNA product has a function.

This basic hypothesis has led to several whole-genome RNA mapping experiments done during the past decade – in effect, the first attempts at genome-wide "RNA Bookkeeping." These unbiased surveys of RNA relied on high-throughput technologies such as tiling arrays and various sequencing methods (Rinn et al., 2003; Bertone et al., 2004; Carninci et al., 2005, 2008; Kapranov et al., 2005, 2007a,b; Birney et al., 2007). In essence, the goal of all these experiments was to identify as many molecules of RNA or sites of transcription as possible in a given tissue, and catalog them into those whose localization to protein-coding regions of the genome could explain their function, and those whose localization could not. Surprisingly, the latter class grew into a pervasive and highly numerous collection (see below for more details) and became broadly dubbed as "dark matter" RNA (Johnson et al., 2005).

Originally, "dark matter" RNA referred simply to RNA produced from the regions of genome without known function, yet stable enough for detection (Johnson et al., 2005). Tiling arrays (Kapranov et al., 2003) can identify regions of genome that give rise to RNA by virtue of hybridization to probes evenly spaced throughout the non-repetitive portions of the genome. The resulting map of transcription specifies a series of RNA producing regions that could then be compared to the map of other genomic features, such as exons of protein-coding genes. The fraction of genomic sequence covered by such fragments located outside of the exons estimated the complexity of the "dark matter" RNA (Kapranov et al., 2002).

Typically, about 75% of all bases represented by all transcribed fragments detected by tiling arrays in any given human cell-line or tissue originated outside of exons of cytosolic polyA+ mRNAs, suggesting that "dark matter" transcription was prevalent in human cells (Kapranov et al., 2007a). As might be expected, this fraction was much higher for human nuclear RNA (Cheng et al., 2005). The FANTOM consortium has shown that the mouse genome could be pervasively transcribed, producing a very complex transcript architecture (Carninci et al., 2005). After combining all available microarray and sequence-based data from all biological sources, the ENCODE consortium estimated that ∼20% of all human genomic sequence might function to produce RNA (Birney et al., 2007). As a consequence of hybridization-based deconvolution of complex mixtures of nucleic acids, the signal thresholds of detection had to be set relatively high to prevent detection of spurious cross-hybridization. This resulted in one of the disadvantages of these experiments: a significant undercounting of transcribed elements. For example, using rapid amplification of cDNA ends (RACE), a more sensitive method for measurement of RNA output form specific loci, evidence of RNA production was found at 75% of randomly chosen human genomic sequences where RNA had not been detected by ENCODE consortium tiling arrays: see Supplementary Table 2 of Birney et al. (2007). This data suggested that the fraction of genome that gives to rise to RNA could far exceed the 20% figure. Indeed, when combining the regions of transcription detected by any method with the total length of all introns (always transcribed to give rise to the mature RNAs), ENCODE estimated that 93% of the human genome is transcribed (Birney et al., 2007). Thus, the matter of detecting the transcribed portion of the genome in stable RNA could depend largely on the sensitivity of the technology used.

However, these experiments have always suffered from criticism that the abundance of the "dark matter" RNAs in mammalian cells could be trivial, in part because of the sensitivity of the techniques used to detect and validate the "dark matter" transcription (van Bakel et al., 2010, 2011). Indeed, these studies were mostly aimed at giving an estimate of the fraction of genomic sequences represented in the "dark matter" RNA and thus tell us something about its complexity, but not about its relative mass (Clark et al., 2011). Perhaps "dark matter" had a very complex population of RNA, and yet represented nothing more than a trivial fraction of cellular RNA mass. Such a scenario might suggest that "dark matter" RNA resulted from non-consequential by-products of cellular processes, consistent with the overall "junk DNA" label given to the non-coding portions of the genome in general (Brosius, 2005; Struhl, 2007; van Bakel and Hughes, 2009; van Bakel et al., 2010). An opposite scenario, where the "dark matter" RNA population was indeed complex and constituted a significant mass of cellular RNA, would on the other hand, suggest that this RNA could indeed be an important and previously hidden component of the regulatory architecture controlling differentiation and development (Mattick, 2003, 2004, 2011; Kapranov et al., 2007b; St Laurent and Wahlestedt, 2007).

The advent of next generation sequencing technologies has allowed for a digital output based count of reads representing short (typically on the order of 25–100 bases) stretches of RNAs from which they were derived (Cloonan and Grimmond, 2008; Wang et al., 2009). By calculating the relative fraction of such reads, one can estimate the relative mass of "dark matter" RNA as a whole, or any specific RNA or transcribed region in the total mass of the assayed RNA population. Despite the apparent simplicity of this approach, the original estimates of the fraction of non-exonic reads in human or mouse RNAseq experiments varied significantly, from as little as 7% (Mortazavi et al., 2008) to as much as 40–50% (Cloonan et al., 2008; Morin et al., 2008). A subsequent report by van Bakel et al. (2010) attempted to directly estimate the relative mass of the "dark matter" RNA and came to the conclusion that it accounts for only 12% of the polyadenylated RNA mass in human or mouse cells. In addition, this report also stated that the same conclusions could be reached by the analysis of total RNA (depleted for rRNA). One common feature of all these reports was the usage of PCR amplification as a part of the RNA preparation for sequencing, which has the potential to alter the original profile of the population (Mamanova et al., 2010; Raz et al., 2011; Sam et al., 2011). Thus, unequivocal estimation of the relative mass of the "dark matter" RNA would require RNA profiling using a sequencing approach that does not rely on amplification. Such profiling performed using single-molecule sequencing of total rRNA-depleted RNA and polyA+ RNA (Kapranov et al., 2010) found that "dark matter" RNA represents a majority of the total non-ribosomal non-mitochondrial RNA most of the human cell-lines and tissues tested (Kapranov et al., 2010). In addition, total human RNA contained a much higher complexity than the polyA+ RNA, especially in terms of "dark matter" RNAs (Kapranov et al., 2010; Raz et al., 2011). This could also explain at least in part the failure of some of the earlier reports to detect a significant fraction of the "dark matter" RNA: not only did those reports rely on PCR amplification, but also they used an RNA fraction highly enriched for polyadenylated RNAs.

Interestingly, very long (100s of kbs) stretches of intergenic space in the human genome, previously considered as "gene desert" regions produced significant levels of RNA (Kapranov et al., 2010). Hundreds of such regions (named vlincs for very long intergenic non-coding regions) spanning ∼4% of intergenic space were detected in just nine different biological sources of RNA (seven tumors and two normal tissues) used in that report. This combined with the observation that most of the vlincs tend to be highly specific to a given biological source (Kapranov et al., 2010), suggests that profiling of the pool of total cellular RNA with hundreds or thousands of different biological samples would result in detection of RNA from a large fraction of intergenic space. This assertion is supported by in-depth analysis of selected genomic regions using methods to select and enrich for all transcripts derived from such regions followed by either tiling array analysis or deep sequencing (Kapranov et al., 2005; Mercer et al., 2011a). Such studies reveal that what appears to be a low signal from either a tiling array or RNAseq experiment obtained on a complex RNA population from a single cell, can in fact represent a complex population of low abundant transcripts (Kapranov et al., 2005; Mercer et al., 2011a). Low abundance could also imply expression restricted to a sub-set of cells in a given population (from a cell-culture or especially, a tissue sample), and thus should not immediately be relegated into the realm of biological noise. These observations are important to keep in mind when interpreting the results of RNAseq or microarray experiments, especially considering that most current RNAseq experiments produce far fewer reads than the estimated minimum of ∼70 million reads required to completely cover the transcriptome from an average human cell (Kapranov et al., 2010).

These results are consistent with those of the ENCODE consortium as far as pervasive transcription is concerned. However, they differ in the estimate of how much stable RNA would remain from that pervasive transcription. The ENCODE consortium suggests that only on the order of ∼20% of human genomic sequence ever exists as stable RNA based on compilation of all available experimental data from a large number of biological sources (Birney et al., 2007). However the logical extrapolation from Kapranov et al. (2010) would suggest that most of the genomic sequence likely exists in the RNA pool when profiling a significant number of tissues using total rRNA-depleted RNA, instead of the polyA+ fraction. The discrepancy may result from the fact the ENCODE, like other similar endeavors before and after, focused on the polyadenylated fraction of RNA, that is estimated to capture only 5–25% of the total mass of the non-ribosomal non-mitochondrial RNA in a human (Kapranov et al., 2010; Raz et al., 2011). Clearly, the dominance of polyA+ RNA as the source of RNA for RNAseq experiments has significantly undercounted the complexity of RNA present in a human cell. In fact, an oligo-dT column may also not necessarily capture all the polyadenylated RNAs in a sample. For example, one can imagine that, long polyadenylated RNA molecules may not bind efficiently due to structural interference, resulting in depletion from the polyA+-selected RNA pool. In fact, depletion of longer mRNAs in polyA+ RNA pool occurs (Raz et al., 2011).

Still, the wider question of functionality of non-polyadenylated RNA as a class has received very little attention, and still remains un-answered. The absence of a polyA-tail does not mean absence of function – clearly, most short non-coding RNA species are non-polyadenylated and functional, for example tRNAs, miRNAs, snRNAs, and other classes of short RNAs. Furthermore, the presence of complex non-adenylated RNA populations in mammalian cells has been established back in 1970s (Salditt-Georgieff et al., 1981) and this type RNA occurred even in the polysomal fraction and was shown to be used for protein production (van Ness et al., 1979; Katinakis et al., 1980). More recently, a reporter mRNA engineered to contain a miRNA in its 3′ UTR served as a target for cleavage by Drosha into a polyA− RNA, and then traveled to the cytosol to function as a template for protein production (Cai et al., 2004). Thus, absence of the polyA-tail does not preclude RNAs from having a function in the cell. However, we are still at the very beginning of the exploration of the functional properties of the vast complexity of novel and apparently non-polyadenylated RNA recently discovered.

Perhaps one of the greatest hurdles in accepting the biological relevance of "dark matter" transcription is the fact that a large proportion of it comes from intronic regions of already annotated genes. Based on single-molecule RNAseq data, it is estimated that the intronic "dark matter" RNA constitutes 70–80% of all mass of the human "dark matter" RNA (Kapranov et al., 2010). The report van Bakel et al. (2010) obtained a similarly high estimate of the fraction of the intronic RNA, but proposed that it simply represents un-processed pre-mRNA. This conclusion was further supported by the data presented in that report where the fraction of intronic RNA amounted only 5.8% of the mass of the human cell's total RNA not including the ribosomal and mitochondrial RNA (van Bakel et al., 2010). However, as mentioned above, this estimate could result from the choice of polyadenylated RNA used in that study, combined with the effect of PCR amplification. Single-molecule sequencing of total RNA revealed a much higher fraction of intronic RNAs in a human cell, on the order of 30–50% of non-ribosomal, non-mitochondrial RNA (Kapranov et al., 2010). The latter estimate should at least cause us to pause before any unambiguous acceptance of the trivial explanation above – as much as half of nuclear-encoded non-ribosomal RNA in the cell is probably not something one should dismiss outright as noise. In addition, different genes vary in terms of how much intronic RNA they produce, as do different introns of the same gene, and even different regions of the same intron (Kapranov et al., 2010). These observations are not consistent with noise expected from pre-mRNA en-route to splicing or excised introns en-route to degradation. Furthermore, intronic signal does not necessarily mean that it arises from excised introns or pre-mRNA. Since RNAseq does not provide information on the complete structure of an RNA molecule, we do not know what kind of transcripts make up the intronic signal observed in RNAseq experiments. In fact, it could represent different types of elements: alternative exons, exon isoforms of known transcripts, independent stand-alone transcripts, or excised introns (**Figure 1**). Moreover, one can imagine that any given gene could have a collection of such different types of novel transcripts buried in its introns (Mattick, 1994; Kapranov et al., 2005). Overall, it is fair to say that we are at the beginning of our understanding of role of intronic RNAs in a cell and we should maintain an open mind as to its functional importance (Clark et al., 2011).

Another class of sequences deserves special mention in the context of the transcriptional activity of genomic "dark matter" – the repetitive regions of a genome, which until very recently had been largely avoided by genome-wide RNA profiling studies for technical reasons. For example, tiling microarray designs typically exclude these regions (Kapranov et al., 2007a) because the signal from the probes cannot resolve into an attribute for a specific repeat element. However, the nucleotide-level precision of the next generation sequencing technologies allows mapping of reads with a relatively high specificity, even to repeat regions of the genome. This in turn allows for interrogation of RNAs produced from repeats. For example, one such study relied on mapping of CAGE tags that mark the 5′ ends of capped RNAs (Kodzius et al., 2006) to profile expression of different types of repeats in mammalian cells (Faulkner et al., 2009). Interestingly, a significant fraction of transcription in that study coincided with repeats.

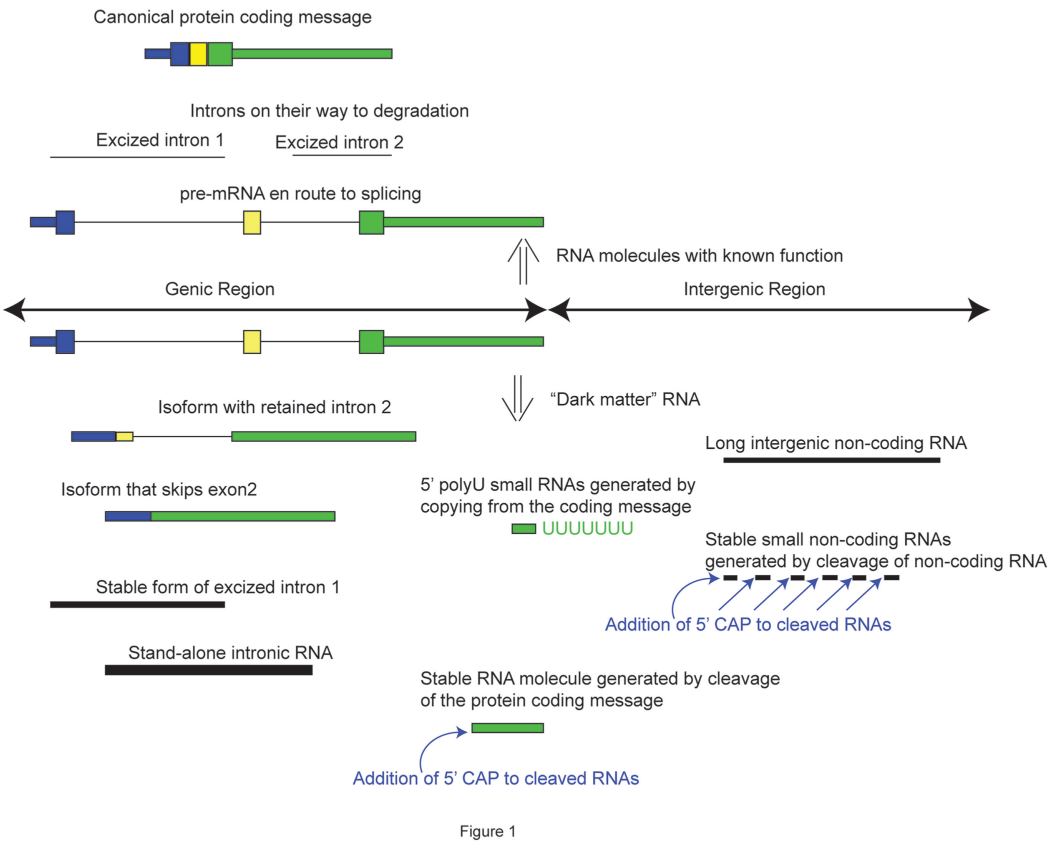

Canonical protein coding message
Introns on their way to degradation
Excized intron 1
Excized intron 2
pre-mRNA en route to splicing
RNA molecules with known function
Genic Region
Intergenic Region
"Dark matter" RNA
Isoform with retained intron 2
Long intergenic non-coding RNA
5' polyU small RNAs generated by copying from the coding message
UUUUUUU
Isoform that skips exon2
Stable small non-coding RNAs generated by cleavage of non-coding RNA
Stable form of excized intron 1
Addition of 5' CAP to cleaved RNAs
Stand-alone intronic RNA
Stable RNA molecule generated by cleavage of the protein coding message
Addition of 5' CAP to cleaved RNAs
Figure 1

FIGURE 1 | Types of RNA molecules derived from annotated and unannotated loci that constitute "dark matter" RNAs.

Different tissues express different levels and types of repetitive elements, with embryonic tissues having the highest levels of CAGE tags (Faulkner et al., 2009). Interestingly, that study also found that a certain class of repetitive elements, retrotransposons, might provide alternative or tissue-specific promoters for protein-coding genes (Faulkner et al., 2009), and a recent paper has shown that these sequences mobilize to effect somatic transposition events in the human brain (Baillie et al., 2011).

However, the last decade of transcriptome exploration also revealed additional dimensions of its complexity. The first added level of complexity arises from the fact that any given locus can be criss-crossed by different transcripts on both strands, described as "transcriptional forests" by the RIKEN researchers after a large scale effort aimed at sequencing full-length cDNAs from mammalian samples (Carninci et al., 2005). The transcriptional forests are common in the protein-coding loci, where the transcripts that form the complex lattices of overlapping transcription often borrow sequences from known exons and non-exonic regions; however, the function of most of the additional RNA isoforms, which are presumably context-specific, is not understood. For example, based on EST evidence, the GENCODE consortium has shown that a human protein-coding locus specifies on average 5.4 isoforms (Harrow et al., 2006). However, only 2.4 of those could encode a protein, while the function of the rest remains an enigma (Harrow et al., 2006).

Other studies have reached similar conclusions using RACE in combination with tiling arrays to profile the complexity of transcripts sharing exons of ~400 human protein-coding genes (Birney et al., 2007; Denoeud et al., 2007). More than 80% of all transcripts had alternative 5′ ends or novel exons (Denoeud et al., 2007). In-depth analysis of one human locus encoding the MeCP2 proteins using the RACE/array method revealed 15 new isoforms that have exons derived from intronic and intergenic sequences with often perfectly correct splice sites (Djebali et al., 2008). In most cases, however, additional isoforms identified either do not appear to change the open reading frame or do not encode proteins (Denoeud et al., 2007; Djebali et al., 2008), consistent with previous GENCODE results (Harrow et al., 2006).

The recent realization that RNA could be cleaved and capped at the newly formed 5′ end to produce a separate stable RNA species provides an additional conceptual dimension of the complexity of the mammalian cell's RNA population (Affymetrix/CSHL ENCODE Project, 2009; Otsuka et al., 2009; Mercer et al., 2010, 2011b). This opens a whole new realm of possibilities where the final, apparently mature and spliced RNA species may not represent the final and/or the only functional product. Conversely, shorter RNAs that would otherwise be considered as simple degradation products, may have function. One tantalizing possibility suggests that they might function in a manner similar to that observed in RNA-mediated inheritance, carried out by apparent

RNA degradation products loaded into germ line cells to mediate regulation of gene expression in the subsequent generation of mice (Rassoulzadegan et al., 2006).

Taken together, all of these added complexities suggest a complete reconsideration of the definition of RNA "dark matter." We would like to posit that it includes not just the RNAs that are made from the "dark matter" regions of the genome, but any RNA molecule whose function we do not understand (**Figure 1**). For example, an RNA molecule consisting solely of exons of an otherwise protein-coding transcript, but spliced into an RNA with no open reading frame, can be considered as an RNA molecule whose function we do not understand, even though it is assembled from individual sequences with known function. Likewise, an RNA molecule processed from a protein-coding gene or pseudogene and having lost its protein-coding capacity can be considered a "dark matter" transcript as long as we do not understand its function. The "dark matter" RNAs can therefore comprise both coding and non-coding RNAs, as long as their function currently remains unclear. While, for the most part, "dark matter" RNAs have features of non-coding RNAs (Carninci et al., 2005; Cheng et al., 2005; Djebali et al., 2008), it remains possible that some of the RNAs previously considered as non-coding do encode short peptides (Kondo et al., 2010). In fact, recent results based on profiling of sites in RNA molecules bound by ribosomes suggest that many mouse "dark matter" RNAs indeed encode short peptides (Ingolia et al., 2011). Undoubtedly, the prevalence and biological relevance of these peptides will remain a very interesting and important question for years to come.

If the entire genome is transcribed and represented as stable RNAs at least in some biological samples, then we should re-evaluate as a community our strategies in terms of annotating the "dark matter" RNAs. Despite the ongoing efforts to annotate the lncRNAs (Amaral et al., 2011; Cabili et al., 2011; Chen et al., 2011; Wang and Chang, 2011), the lists obtained from different experiments do not overlap significantly. For example, only ~19% of base pairs covered by the human vlinc regions in the intergenic space overlap those found by lncRNAs (Kapranov et al., 2010; also see **Figure 2A**). This suggests that current databases only scratch the surface of the immense complexity of the RNA population of human cells.

In retrospect, this is not surprising when one considers that current genomic annotations, such as the human GenBank mRNA track on the UCSC browser (Kent et al., 2002), depend primarily on sequenced full-length cDNAs, each one representing only a single-molecule of RNA. GenBank currently contains ~300K such

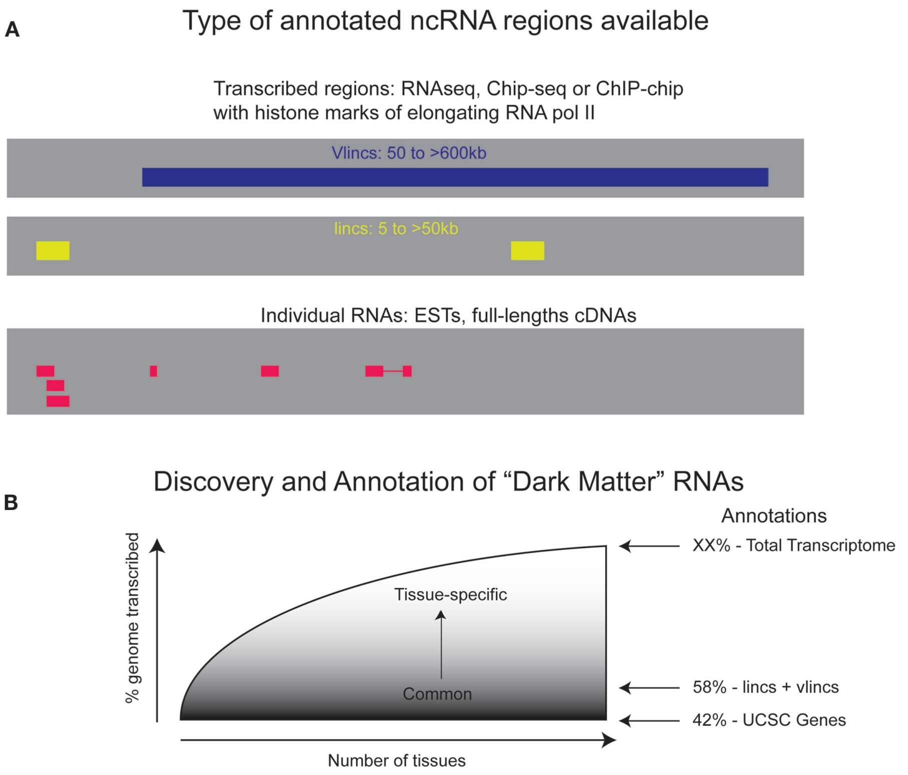

A
Type of annotated ncRNA regions available
Transcribed regions: RNAseq, Chip-seq or ChIP-chip with histone marks of elongating RNA pol II
Vlincs: 50 to >600kb
lincs: 5 to >50kb
Individual RNAs: ESTs, full-lengths cDNAs
B
Discovery and Annotation of "Dark Matter" RNAs
Annotations
XX% - Total Transcriptome
Tissue-specific
Common
58% - lincs + vlincs
42% - UCSC Genes
% genome transcribed
Number of tissues

FIGURE 2 | Coverage of the genome by "dark matter" RNAs. (A) Information currently available about the regions of dark matter transcription and the actual RNA molecules made from these region comes from various types of experiments and databases. There is relatively little overlap between these different databases suggesting that the actual extend of dark matter transcription is far greater than any one database suggests. **(B)** A theoretical curve showing expected results of the fraction of the genome that is transcribed as a function of the number of biological sources whose RNA is profiled. The coverage of transcribed genome by protein coding genes including their introns is 42% and lincRNAs bring it up to 58%. However, the full extent of the transcribed genome is expected to be much greater than that.

entries, which closely approximates estimates of the total number of polyadenylated RNA molecules contained in a single cell (~300K; Hastie and Bishop, 1976). Thus, based on these numbers, it is fair to say that all we know in terms of the complete sequences of RNAs from the human transcriptome represents just one cell's worth of polyadenylated RNA! Of the 300K human GenBank mRNA entries, ~88% are represented by unique cDNAs, pointing to the fact that many of the current gene models and annotations are based on a single (!!!) fully sequenced RNA molecule. This is reinforced by the recent application of targeted RNA sequencing, which revealed a plethora of new coding and non-coding transcripts, even from intensively studied human loci such as p53, HOX, and *sonic hedgehog* (SHH) that are either only expressed in a very limited number of cells in what was previously considered a homogenous culture, or where otherwise missed in the cDNA libraries (Kapranov et al., 2005; Mercer et al., 2011a). In addition, most of the annotated cDNAs have been characterized from the polyadenylated transcriptome, thus the non-polyadenylated fraction remains virtually un-uncharted from the point of view of full-length cDNA sequencing. Considering how many molecules of RNA a given human locus must make during the lifetime of an individual, evidently, this depth of knowledge only scratches the surface of RNA complexity.

Finally, we believe that understanding of the true extent and function of human transcription remains one of the most important philosophical and scientific questions of our time. Considering this, we suggest that the community should undertake a directed approach aimed at answering this question. We envision the profiling of a reasonably large number of carefully chosen samples based on total RNA depleted of rRNA, rather than polyA+, using amplification free RNAseq approaches. Given the high-tissue specificity of dark matter RNAs, samples would include at least 100–200 key tissues or cell-lines, rich in intergenic RNAs, such as Ewing Sarcomas. We expect the curve of detected dark matter transcripts to reach a plateau steeply – the big un-answered question so far is where this plateau will be and how much further the curve will rise as more samples are added (**Figure 2B**). RNAseq will yield regions of transcription, while additional methods will unravel the complexities of individual transcripts in each region of transcription. This could be accomplished by a site-directed methods similar to the one described by Djebali et al. (2008).

As our understanding of the function of the novel RNA expands, the domain of "dark matter" transcripts will shrink. Unfortunately, for the most part we cannot yet predict *in silico* which of these "non-canonical" RNA molecules are functional and what function they might fulfill, like we usually can for protein-coding mRNAs. This is probably the greatest challenge to our understanding and acceptance of this type of RNA – our general inability to predict what an RNA species might do when it does not have an obvious open reading frame. However, this should not stop us from exploring the function of these RNAs in biological or medical context. Even if a function of a given RNA molecule or transcribed region in a genome may not be known, its association with a disease should provide novel mechanistic insights, and novel diagnostic tools for the disease. The fact that "dark matter" RNAs tend to be highly specific to their biological source emphasizes the promise of this approach (Cheng et al., 2005; Kapranov et al., 2010). Surprisingly perhaps, it seems remarkably easy to detect phenotypes associated with siRNA-mediated knockdown or over-expression of non-coding RNAs, even in cell-culture, and to correlate these phenotypes with aberrant expression of the non-coding RNAs in disease states like cancer and neurological diseases (see, e.g., Mattick, 2009; Gupta et al., 2010; Askarian-Amiri et al., 2011; Gibb et al., 2011; Khaitan et al., 2011; Ulitsky et al., 2011). Moreover, one of the first examples of association of "dark matter" transcripts emanating from a family of repetitive regions with a particular type of cancer has been recently provided by Ting et al., 2011. Overall, it would not be surprising if "dark matter" transcripts would eventually occupy a central place in our conceptual understanding of the molecular events underlying human development and disease, and thereby enter the arsenal of therapeutic targets as prominently as those gene products whose function we currently understand.

GLOSSARY

CAGE: cap analysis of gene expression, a method based on selection of RNAs containing the 5′ CAP modification and obtaining short sequences or tags near the 5′ end of these RNAs. Typically, millions of tags are obtained in each experiment.

ENCODE: encyclopedia of DNA elements, an NHGRI-sponsored project aimed at empirically identifying functionally important element in the human genome sequence, http://www.genome.gov/10005107.

Genomic dark matter: usually refers to the portion of a genome that does not correspond to exons (coding or non-coding) of annotated mRNAs.

RACE: rapid amplification of cDNA Ends, a PCR-approach with "outward" positioned primers to amplify toward the 5′ and 3′ end of an RNA molecule from a point inside the molecule.

RNAseq: a method to quantify and profile RNA population in a cell based on massive sequencing of short (typically less than 100 bases) regions of a large number of RNA molecules. Typically, sequencing is conducted on cDNA, rather than RNA, thus cDNAseq would have been more appropriate. However, RNAseq is used for historical reasons. A note, since direct RNA sequencing is now possible, a *bona fide* RNAseq analysis should somehow be distinguished from cDNAseq.

Single-molecule sequencing: a method where a single-molecule of a nucleic acid is sequenced directly as opposed methods that obtain sequence signal from a population of molecules.

Tiling array: a microarray platform designed to interrogate genomic sequence with a certain resolution set by the distance between the probes. Opposite in concept to exon arrays where probes are designed only to the annotated regions of interest.

Vlinc: very long intergenic non-coding RNA region, identified based on continuous RNAseq signal that spans genomic regions of 50 kb or longer (often much longer) in the area of genome where no annotated gene has been found.

ACKNOWLEDGMENTS

We wish to thanks Dr. John Mattick for very helpful suggestions on the manuscript, and Drs. Robert Arceci, Tim Triche, Jason Farrar, Patrice Milos, John Thompson, and Claes Wahlestedt for helpful and encouraging discussions.

REFERENCES

Affymetrix/CSHL ENCODE Project. (2009). Post-transcriptional processing generates a diversity of 5'-modified long and short RNAs. *Nature* 457, 1028–1032.

Amaral, P. P., Clark, M. B., Gascoigne, D. K., Dinger, M. E., and Mattick, J. S. (2011). lncRNAdb: a reference database for long noncoding RNAs. *Nucleic Acids Res.* 39, D146–D151.

Askarian-Amiri, M. E., Crawford, J., French, J. D., Smart, C. E., Smith, M. A., Clark, M. B., Ru, K., Mercer, T. R., Thompson, E. R., Lakhani, S. R., Vargas, A. C., Campbell, I. G., Brown, M. A., Dinger, M. E., and Mattick, J. S. (2011). SNORD-host RNA Zfas1 is a regulator of mammary development and a potential marker for breast cancer. *RNA* 17, 878–891.

Baillie, J. K., Barnett, M. W., Upton, K. R., Gerhardt, D. J., Richmond, T. A., De Sapio, F., Brennan, P. M., Rizzu, P., Smith, S., Fell, M., Talbot, R. T., Gustincich, S., Freeman, T. C., Mattick, J. S., Hume, D. A., Heutink, P., Carninci, P., Jeddeloh, J. A., and Faulkner, G. J. (2011). Somatic retrotransposition alters the genetic landscape of the human brain. *Nature* 479, 534–537.

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

Birney, E., Stamatoyannopoulos, J. A., Dutta, A., Guigo, R., Gingeras, T. R., Margulies, E. H., Weng, Z., Snyder, M., Dermitzakis, E. T., Thurman, R. E., Kuehn, M. S., Taylor, C. M., Neph, S., Koch, C. M., Asthana, S., Malhotra, A., Adzhubei, I., Greenbaum, J. A., Andrews, R. M., Flicek, P., Boyle, P. J., Cao, H., Carter, N. P., Clelland, G. K., Davis, S., Day, N., Dhami, P., Dillon, S. C., Dorschner, M. O., Fiegler, H., Giresi, P. G., Goldy, J., Hawrylycz, M., Haydock, A., Humbert, R., James, K. D., Johnson, B. E., Johnson, E. M., Frum, T. T., Rosenzweig, E. R., Karnani, N., Lee, K., Lefebvre, G. C., Navas, P. A., Neri, F., Parker, S. C., Sabo, P. J., Sandstrom, R., Shafer, A., Vetrie, D., Weaver, M., Wilcox, S., Yu, M., Collins, F. S., Dekker, J., Lieb, J. D., Tullius, T. D., Crawford, G. E., Sunyaev, S., Noble, W. S., Dunham, I., Denoeud, F., Reymond, A., Kapranov, P., Rozowsky, J., Zheng, D., Castelo, R., Frankish, A., Harrow, J., Ghosh, S., Sandelin, A., Hofacker, I. L., Baertsch, R., Keefe, D., Dike, S., Cheng, J., Hirsch, H. A., Sekinger, E. A., Lagarde, J., Abril, J. F., Shahab, A., Flamm, C., Fried, C., Hackermuller, J., Hertel, J., Lindemeyer, M., Missal, K., Tanzer, A., Washietl, S., Korbel, J., Emanuelsson, O., Pedersen, J. S., Holroyd, N., Taylor, R., Swarbreck, D., Matthews, N., Dickson, M. C., Thomas, D. J., Weirauch, M. T., Gilbert, J., Drenkow, J., Bell, I., Zhao, X., Srinivasan, K. G., Sung, W. K., Ooi, H. S., Chiu, K. P., Foissac, S., Alioto, T., Brent, M., Pachter, L., Tress, M. L., Valencia, A., Choo, S. W., Choo, C. Y., Ucla, C., Manzano, C., Wyss, C., Cheung, E., Clark, T. G., Brown, J. B., Ganesh, M., Patel, S., Tammana, H., Chrast, J., Henrichsen, C. N., Kai, C., Kawai, J., Nagalakshmi, U., Wu, J., Lian, Z., Lian, J., Newburger, P., Zhang, X., Bickel, P., Mattick, J. S., Carninci, P., Hayashizaki, Y., Weissman, S., Hubbard, T., Myers, R. M., Rogers, J., Stadler, P. F., Lowe, T. M., Wei, C. L., Ruan, Y., Struhl, K., Gerstein, M., Antonarakis, S. E., Fu, Y., Green, E. D., Karaöz, U., Siepel, A., Taylor, J., Liefer, L. A., Wetterstrand, K. A., Good, P. J., Feingold, E. A., Guyer, M. S., Cooper, G. M., Asimenos, G., Dewey, C. N., Hou, M., Nikolaev, S., Montoya-Burgos, J. I., Löytynoja, A., Whelan, S., Pardi, F., Massingham, T., Huang, H., Zhang, N. R., Holmes, I., Mullikin, J. C., Ureta-Vidal, A., Paten, B., Seringhaus, M., Church, D., Rosenbloom, K., Kent, W. J., Stone, E. A., NISC Comparative Sequencing Program, Baylor College of Medicine Human Genome Sequencing Center, Washington University Genome Sequencing Center, Broad Institute, Children's Hospital Oakland Research Institute, Batzoglou, S., Goldman, N., Hardison, R. C., Haussler, D., Miller, W., Sidow, A., Trinklein, N. D., Zhang, Z. D., Barrera, L., Stuart, R., King, D. C., Ameur, A., Enroth, S., Bieda, M. C., Kim, J., Bhinge, A. A., Jiang, N., Liu, J., Yao, F., Vega, V. B., Lee, C. W., Ng, P., Shahab, A., Yang, A., Moqtaderi, Z., Zhu, Z., Xu, X., Squazzo, S., Oberley, M. J., Inman, D., Singer, M. A., Richmond, T. A., Munn, K. J., Rada-Iglesias, A., Wallerman, O., Komorowski, J., Fowler, J. C., Couttet, P., Bruce, A. W., Dovey, O. M., Ellis, P. D., Langford, C. F., Nix, D. A., Euskirchen, G., Hartman, S., Urban, A. E., Kraus, P., Van Calcar, S., Heintzman, N., Kim, T. H., Wang, K., Qu, C., Hon, G., Luna, R., Glass, C. K., Rosenfeld, M. G., Aldred, S. F., Cooper, S. J., Halees, A., Lin, J. M., Shulha, H. P., Zhang, X., Xu, M., Haidar, J. N., Yu, Y., Ruan, Y., Iyer, V. R., Green, R. D., Wadelius, C., Farnham, P. J., Ren, B., Harte, R. A., Hinrichs, A. S., Trumbower, H., Clawson, H., Hillman-Jackson, J., Zweig, A. S., Smith, K., Thakkapallayil, A., Barber, G., Kuhn, R. M., Karolchik, D., Armengol, L., Bird, C. P., de Bakker, P. I., Kern, A. D., Lopez-Bigas, N., Martin, J. D., Stranger, B. E., Woodroffe, A., Davydov, E., Dimas, A., Eyras, E., Hallgrímsdóttir, I. B., Huppert, J., Zody, M. C., Abecasis, G. R., Estivill, X., Bouffard, G. G., Guan, X., Hansen, N. F., Idol, J. R., Maduro, V. V., Maskeri, B., McDowell, J. C., Park, M., Thomas, P. J., Young, A. C., Blakesley, R. W., Muzny, D. M., Sodergren, E., Wheeler, D. A., Worley, K. C., Jiang, H., Weinstock, G. M., Gibbs, R. A., Graves, T., Fulton, R., Mardis, E. R., Wilson, R. K., Clamp, M., Cuff, J., Gnerre, S., Jaffe, D. B., Chang, J. L., Lindblad-Toh, K., Lander, E. S., Koriabine, M., Nefedov, M., Osoegawa, K., Yoshinaga, Y., Zhu, B., and de Jong, P. J. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447, 799–816.

Brosius, J. (2005). Waste not, want not – transcript excess in multicellular eukaryotes. *Trends Genet.* 21, 287–288.

Cabili, M. N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., and Rinn, J. L. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.* 25, 1915–1927.

Cai, X., Hagedorn, C. H., and Cullen, B. R. (2004). Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. *RNA* 10, 1957–1966.

Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M. C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C., Kodzius, R., Shimokawa, K., Bajic, V. B., Brenner, S. E., Batalov, S., Forrest, A. R., Zavolan, M., Davis, M. J., Wilming, L. G., Aidinis, V., Allen, J. E., Ambesi-Impiombato, A., Apweiler, R., Aturaliya, R. N., Bailey, T. L., Bansal, M., Baxter, L., Beisel, K. W., Bersano, T., Bono, H., Chalk, A. M., Chiu, K. P., Choudhary, V., Christoffels, A., Clutterbuck, D. R., Crowe, M. L., Dalla, E., Dalrymple, B. P., De Bono, B., Della Gatta, G., Di Bernardo, D., Down, T., Engstrom, P., Fagiolini, M., Faulkner, G., Fletcher, C. F., Fukushima, T., Furuno, M., Futaki, S., Gariboldi, M., Georgii-Hemming, P., Gingeras, T. R., Gojobori, T., Green, R. E., Gustincich, S., Harbers, M., Hayashi, Y., Hensch, T. K., Hirokawa, N., Hill, D., Huminiecki, L., Iacono, M., Ikeo, K., Iwama, A., Ishikawa, T., Jakt, M., Kanapin, A., Katoh, M., Kawasawa, Y., Kelso, J., Kitamura, H., Kitano, H., Kollias, G., Krishnan, S. P., Kruger, A., Kummerfeld, S. K., Kurochkin, I. V., Lareau, L. F., Lazarevic, D., Lipovich, L., Liu, J., Liuni, S., Mcwilliam, S., Madan Babu, M., Madera, M., Marchionni, L., Matsuda, H., Matsuzawa, S., Miki, H., Mignone, F., Miyake, S., Morris, K., Mottagui-Tabar, S., Mulder, N., Nakano, N., Nakauchi, H., Ng, P., Nilsson, R., Nishiguchi, S., Nishikawa, S., Nori, F., Ohara, O., Okazaki, Y., Orlando, V., Pang, K. C., Pavan, W. J., Pavesi, G., Pesole, G., Petrovsky, N., Piazza, S., Reed, J., Reid, J. F., Ring, B. Z., Ringwald, M., Rost, B., Ruan, Y., Salzberg, S. L., Sandelin, A., Schneider, C., Schönbach, C., Sekiguchi, K., Semple, C. A., Seno, S., Sessa, L., Sheng, Y., Shibata, Y., Shimada, H., Shimada, K., Silva, D., Sinclair, B., Sperling, S., Stupka, E., Sugiura, K., Sultana, R., Takenaka, Y., Taki, K., Tammoja, K., Tan, S. L., Tang, S., Taylor, M. S., Tegner, J., Teichmann, S. A., Ueda, H. R., van Nimwegen, E., Verardo, R., Wei, C. L., Yagi, K., Yamanishi, H., Zabarovsky, E., Zhu, S., Zimmer, A., Hide, W., Bult, C., Grimmond, S. M., Teasdale, R. D., Liu, E. T., Brusic, V., Quackenbush, J., Wahlestedt, C., Mattick, J. S., Hume, D. A., Kai, C., Sasaki, D., Tomaru, Y., Fukuda, S., Kanamori-Katayama, M., Suzuki, M., Aoki, J., Arakawa, T., Iida, J., Imamura, K., Itoh, M., Kato, T., Kawaji, H., Kawagashira, N., Kawashima, T., Kojima, M., Kondo, S., Konno, H., Nakano, K., Ninomiya, N., Nishio, T., Okada, M., Plessy, C., Shibata, K., Shiraki, T., Suzuki, S., Tagami, M., Waki, K., Watahiki, A., Okamura-Oho, Y., Suzuki, H., Kawai, J., Hayashizaki, Y., FANTOM Consortium, and RIKEN Genome Exploration Research Group and Genome Science Group (Genome Network Project Core Group). (2005). The transcriptional landscape of the mammalian genome. *Science* 309, 1559–1563.

Carninci, P., Yasuda, J., and Hayashizaki, Y. (2008). Multifaceted mammalian transcriptome. *Curr. Opin. Cell Biol.* 20, 274–280.

Chen, F., Evans, A., Gaskell, E., Pham, J., and Tsai, M. C. (2011). Regulatory RNA: the new age. *Mol. Cell* 43, 851–852.

Cheng, J., Kapranov, P., Drenkow, J., Dike, S., Brubaker, S., Patel, S., Long, J., Stern, D., Tammana, H., Helt, G., Sementchenko, V., Piccolboni, A., Bekiranov, S., Bailey, D. K., Ganesh, M., Ghosh, S., Bell, I., Gerhard, D. S., and Gingeras, T. R. (2005). Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science* 308, 1149–1154.

Clark, M. B., Amaral, P. P., Schlesinger, F. J., Dinger, M. E., Taft, R. J., Rinn, J. L., Ponting, C. P., Stadler, P. F., Morris, K. V., Morillon, A., Rozowsky, J. S., Gerstein, M. B., Wahlestedt, C., Hayashizaki, Y., Carninci, P., Gingeras, T. R., and Mattick, J. S. (2011). The reality of pervasive transcription. *PLoS Biol.* 9, e1000625; discussion e1001102. doi:10.1371/journal.pbio.1000625

Cloonan, N., Forrest, A. R., Kolle, G., Gardiner, B. B., Faulkner, G. J., Brown, M. K., Taylor, D. F., Steptoe, A. L., Wani, S., Bethel, G., Robertson, A. J., Perkins, A. C., Bruce, S. J., Lee, C. C., Ranade, S. S., Peckham, H. E., Manning, J. M., Mckernan, K. J., and Grimmond, S. M. (2008). Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat. Methods* 5, 613–619.

Cloonan, N., and Grimmond, S. M. (2008). Transcriptome content and dynamics at single-nucleotide resolution. *Genome Biol.* 9, 234.

Cooper, D. N., Krawczak, M., and Antonarakis, S. E. (1995). "The nature and mechanisms of human gene mutation," in *The Metabolic and Molecular Bases of Inherited Disease*, 7th Edn, eds C. Scriver, A. L. Beaudet, W. S. Sly, and D. Valle (New York: McGraw-Hill), 259–291.

Denoeud, F., Kapranov, P., Ucla, C., Frankish, A., Castelo, R., Drenkow, J., Lagarde, J., Alioto, T., Manzano, C., Chrast, J., Dike, S., Wyss, C., Henrichsen, C. N., Holroyd, N., Dickson, M. C., Taylor, R., Hance, Z., Foissac, S., Myers, R. M., Rogers, J., Hubbard, T., Harrow, J., Guigo, R., Gingeras, T. R., Antonarakis, S. E., and Reymond, A. (2007). Prominent use of distal 5' transcription start sites and discovery of a large number of additional exons in ENCODE regions. *Genome Res.* 17, 746–759.

Djebali, S., Kapranov, P., Foissac, S., Lagarde, J., Reymond, A., Ucla, C., Wyss, C., Drenkow, J., Dumais, E., Murray, R. R., Lin, C., Szeto, D., Denoeud, F., Calvo, M., Frankish, A., Harrow, J., Makrythanasis, P., Vidal, M., Salehi-Ashtiani, K., Antonarakis, S. E., Gingeras, T. R., and Guigo, R. (2008). Efficient targeted transcript discovery via array-based normalization of RACE libraries. *Nat. Methods* 5, 629–635.

Faulkner, G. J., Kimura, Y., Daub, C. O., Wani, S., Plessy, C., Irvine, K. M., Schroder, K., Cloonan, N., Steptoe, A. L., Lassmann, T., Waki, K., Hornig, N., Arakawa, T., Takahashi, H., Kawai, J., Forrest, A. R., Suzuki, H., Hayashizaki, Y., Hume, D. A., Orlando, V., Grimmond, S. M., and Carninci, P. (2009). The regulated retrotransposon transcriptome of mammalian cells. *Nat. Genet.* 41, 563–571.

Gibb, E. A., Brown, C. J., and Lam, W. L. (2011). The functional role of long non-coding RNA in human carcinomas. *Mol. Cancer* 10, 38.

Gupta, R. A., Shah, N., Wang, K. C., Kim, J., Horlings, H. M., Wong, D. J., Tsai, M. C., Hung, T., Argani, P., Rinn, J. L., Wang, Y., Brzoska, P., Kong, B., Li, R., West, R. B., Van De Vijver, M. J., Sukumar, S., and Chang, H. Y. (2010). Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* 464, 1071–1076.

Harrow, J., Denoeud, F., Frankish, A., Reymond, A., Chen, C. K., Chrast, J., Lagarde, J., Gilbert, J. G., Storey, R., Swarbreck, D., Rossier, C., Ucla, C., Hubbard, T., Antonarakis, S. E., and Guigo, R. (2006). GENCODE: producing a reference annotation for ENCODE. *Genome Biol.* 7(Suppl. 1), S41–S49.

Hastie, N. D., and Bishop, J. O. (1976). The expression of three abundance classes of messenger RNA in mouse tissues. *Cell* 9, 761–774.

Ingolia, N. T., Lareau, L. F., and Weissman, J. S. (2011). Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* 147, 789–802.

Johnson, J. M., Edwards, S., Shoemaker, D., and Schadt, E. E. (2005). Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments. *Trends Genet.* 21, 93–102.

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.

Kapranov, P., Cheng, J., Dike, S., Nix, D. A., Duttagupta, R., Willingham, A. T., Stadler, P. F., Hertel, J., Hackermuller, J., Hofacker, I. L., Bell, I., Cheung, E., Drenkow, J., Dumais, E., Patel, S., Helt, G., Ganesh, M., Ghosh, S., Piccolboni, A., Sementchenko, V., Tammana, H., and Gingeras, T. R. (2007a). RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* 316, 1484–1488.

Kapranov, P., Willingham, A. T., and Gingeras, T. R. (2007b). Genome-wide transcription and the implications for genomic organization. *Nat. Rev. Genet.* 8, 413–423.

Kapranov, P., Drenkow, J., Cheng, J., Long, J., Helt, G., Dike, S., and Gingeras, T. R. (2005). Examples of the complex architecture of the human transcriptome revealed by RACE and high-density tiling arrays. *Genome Res.* 15, 987–997.

Kapranov, P., Sementchenko, V. I., and Gingeras, T. R. (2003). Beyond expression profiling: next generation uses of high density oligonucleotide arrays. *Brief. Funct. Genomic. Proteomic.* 2, 47–56.

Kapranov, P., St Laurent, G., Raz, T., Ozsolak, F., Reynolds, C. P., Sorensen, P. H., Reaman, G., Milos, P., Arceci, R. J., Thompson, J. F., and Triche, T. J. (2010). The majority of total nuclear-encoded non-ribosomal RNA in a human cell is "dark matter" un-annotated RNA. *BMC Biol.* 8, 149. doi:10.1186/1741-7007-8-149

Katinakis, P. K., Slater, A., and Burdon, R. H. (1980). Non-polyadenylated mRNAs from eukaryotes. *FEBS Lett.* 116, 1–7.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., and Haussler, D. (2002). The human genome browser at UCSC. *Genome Res.* 12, 996–1006.

Khaitan, D., Dinger, M. E., Mazar, J., Crawford, J., Smith, M. A., Mattick, J. S., and Perera, R. J. (2011). The melanoma-upregulated long non-coding RNA SPRY4-IT1 modulates apoptosis and invasion. *Cancer Res.* 71, 3852–3862.

Kodzius, R., Kojima, M., Nishiyori, H., Nakamura, M., Fukuda, S., Tagami, M., Sasaki, D., Imamura, K., Kai, C., Harbers, M., Hayashizaki, Y., and Carninci, P. (2006). CAGE: cap analysis of gene expression. *Nat. Methods* 3, 211–222.

Kondo, T., Plaza, S., Zanet, J., Benrabah, E., Valenti, P., Hashimoto, Y., Kobayashi, S., Payre, F., and Kageyama, Y. (2010). Small peptides switch the transcriptional activity of Shavenbaby during *Drosophila* embryogenesis. *Science* 329, 336–339.

Mamanova, L., Andrews, R. M., James, K. D., Sheridan, E. M., Ellis, P. D., Langford, C. F., Ost, T. W., Collins, J. E., and Turner, D. J. (2010). FRT-seq: amplification-free, strand-specific transcriptome sequencing. *Nat. Methods* 7, 130–132.

Mattick, J. S. (1994). Introns: evolution and function. *Curr. Opin. Genet. Dev.* 4, 823–831.

Mattick, J. S. (2003). Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms. *Bioessays* 25, 930–939.

Mattick, J. S. (2004). RNA regulation: a new genetics? *Nat. Rev. Genet.* 5, 316–323.

Mattick, J. S. (2009). The genetic signatures of noncoding RNAs. *PLoS Genet.* 5, e1000459. doi:10.1371/journal.pgen.1000459

Mattick, J. S. (2011). The central role of RNA in human development and cognition. *FEBS Lett.* 585, 1600–1616.

Mercer, T. R., Dinger, M. E., Bracken, C. P., Kolle, G., Szubert, J. M., Korbie, D. J., Askarian-Amiri, M. E., Gardiner, B. B., Goodall, G. J., Grimmond, S. M., and Mattick, J. S. (2010). Regulated post-transcriptional RNA cleavage diversifies the eukaryotic transcriptome. *Genome Res.* 20, 1639–1650.

Mercer, T. R., Gerhardt, D. J., Dinger, M. E., Crawford, J., Trapnell, C., Jeddeloh, J. A., Mattick, J. S., and Rinn, J. L. (2011a). Targeted RNA sequencing reveals the deep complexity of the human transcriptome. *Nat. Biotechnol.* 30, 99–104.

Mercer, T. R., Wilhelm, D., Dinger, M. E., Solda, G., Korbie, D. J., Glazov, E. A., Truong, V., Schwenke, M., Simons, C., Matthaei, K. I., Saint, R., Koopman, P., and Mattick, J. S. (2011b). Expression of distinct RNAs from 3' untranslated regions. *Nucleic Acids Res.* 39, 2393–2403.

Morin, R., Bainbridge, M., Fejes, A., Hirst, M., Krzywinski, M., Pugh, T., Mcdonald, H., Varhol, R., Jones, S., and Marra, M. (2008). Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. *BioTechniques* 45, 81–94.

Mortazavi, A., Williams, B. A., Mccue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nat. Methods* 5, 621–628.

Orgel, L. E., and Crick, F. H. (1980). Selfish DNA: the ultimate parasite. *Nature* 284, 604–607.

Otsuka, Y., Kedersha, N. L., and Schoenberg, D. R. (2009). Identification of a cytoplasmic complex that adds a cap onto 5'-monophosphate RNA. *Mol. Cell. Biol.* 29, 2155–2167.

Penman, S. (1995). Rethinking cell structure. *Proc. Natl. Acad. Sci. U.S.A.* 92, 5251–5257.

Rassoulzadegan, M., Grandjean, V., Gounon, P., Vincent, S., Gillot, I., and Cuzin, F. (2006). RNA-mediated non-mendelian inheritance of an epigenetic change in the mouse. *Nature* 441, 469–474.

Raz, T., Kapranov, P., Lipson, D., Letovsky, S., Milos, P. M., and Thompson, J. F. (2011). Protocol dependence of sequencing-based gene expression measurements. *PLoS ONE* 6, e19287. doi:10.1371/journal.pone.0019287

Rinn, J. L., Euskirchen, G., Bertone, P., Martone, R., Luscombe, N. M., Hartman, S., Harrison, P. M., Nelson, F. K., Miller, P., Gerstein, M., Weissman, S., and Snyder, M. (2003). The transcriptional activity of human chromosome 22. *Genes Dev.* 17, 529–540.

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.

Sam, L. T., Lipson, D., Raz, T., Cao, X., Thompson, J., Milos, P. M., Robinson, D., Chinnaiyan, A. M., Kumar-Sinha, C., and Maher, C. A. (2011). A comparison of single molecule and amplification based sequencing of cancer transcriptomes. *PLoS ONE* 6, e17305. doi:10.1371/journal.pone.0017305

St Laurent, G. III, and Wahlestedt, C. (2007). Noncoding RNAs: couplers of analog and digital information in nervous system function? *Trends Neurosci.* 30, 612–621.

Struhl, K. (2007). Transcriptional noise and the fidelity of initiation by RNA polymerase II. *Nat. Struct. Mol. Biol.* 14, 103–105.

Ting, D. T., Lipson, D., Paul, S., Brannigan, B. W., Akhavanfard, S., Coffman, E. J., Contino, G., Deshpande, V., Iafrate, A. J., Letovsky, S., Rivera, M. N., Bardeesy, N., Maheswaran, S., and Haber, D. A. (2011). Aberrant overexpression of satellite repeats in pancreatic and other epithelial cancers. *Science* 331, 593–596.

Ulitsky, I., Shkumatava, A., Jan, C. H., Sive, H., and Bartel, D. P. (2011). Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. *Cell* 147, 1537–1550.

van Bakel, H., and Hughes, T. R. (2009). Establishing legitimacy and function in the new transcriptome. *Brief. Funct. Genomic. Proteomic.* 8, 424–436.

van Bakel, H., Nislow, C., Blencowe, B. J., and Hughes, T. R. (2010). Most "dark matter" transcripts are associated with known genes. *PLoS Biol.* 8, e1000371. doi:10.1371/journal.pbio.1000371

van Bakel, H., Nislow, C., Blencowe, B. J., and Hughes, T. R. (2011). Response to "the reality of pervasive transcription." *PLoS Biol.* 9, e1001102. doi:10.1371/journal.pbio.1001102

van Ness, J., Maxwell, I. H., and Hahn, W. E. (1979). Complex population of nonpolyadenylated messenger RNA in mouse brain. *Cell* 18, 1341–1349.

Wang, K. C., and Chang, H. Y. (2011). Molecular mechanisms of long noncoding RNAs. *Mol. Cell* 43, 904–914.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 10, 57–63.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 27 October 2011; accepted: 30 March 2012; published online: 23 April 2012.

Citation: Kapranov P and St. Laurent G (2012) Dark matter RNA: existence, function, and controversy. Front. Gene. ***3****:60. doi: 10.3389/fgene.2012.00060*

This article was submitted to Frontiers in Non-Coding RNA, a specialty of Frontiers in Genetics.

frontiers in
GENETICS

ORIGINAL RESEARCH ARTICLE
published: 26 December 2011
doi: 10.3389/fgene.2011.00093

Thousands of novel transcripts identified in mouse cerebrum, testis, and ES cells based on ribo-minus RNA sequencing

Wanfei Liu[1,2†], Yuhui Zhao[1,2†], Peng Cui[1,3†], Qiang Lin[1,2], Feng Ding[1,3], Chengqi Xin[1,2], Xinyu Tan[1], Shuhui Song[1]*, Jun Yu[1]* and Songnian Hu[1]*

[1] CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, China
[2] Graduate University of Chinese Academy of Sciences, Beijing, China
[3] Department of Pharmacology and Toxicology, Medical College of Wisconsin, Milwaukee, WI, USA

Edited by:
Philipp Kapranov, St. Laurent Institute, USA

Reviewed by:
Philipp Kapranov, St. Laurent Institute, USA
John Stanley Mattick, The University of Queensland, Australia
Piero Carninci, RIKEN, Japan

***Correspondence:**
Shuhui Song, CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, No.7, Beitucheng West Road, Chaoyang District, Beijing 100029, China.
e-mail: songshh@big.ac.cn;
Jun Yu, CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, No.7, Beitucheng West Road, Chaoyang District, Beijing 100029, China.
e-mail: junyu@big.ac.cn;
Songnian Hu, CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, No.7, Beitucheng West Road, Chaoyang District, Beijing 100029, China.
e-mail: husn@big.ac.cn

†Wanfei Liu, Yuhui Zhao and Peng Cui have contributed equally to this work.

The high-throughput next-generation sequencing technologies provide an excellent opportunity for the detection of less-abundance transcripts that may not be identifiable by previously available techniques. Here, we report a discovery of thousands of novel transcripts (mostly non-coding RNAs) that are expressed in mouse cerebrum, testis, and embryonic stem (ES) cells, through an in-depth analysis of rmRNA-seq data. These transcripts show significant associations with transcriptional start and elongation signals. At the upstream of these transcripts we observed significant enrichment of histone marks (histone H3 lysine 4 trimethylation, H3K4me3), RNAPII binding sites, and cap analysis of gene expression tags that mark transcriptional start sites. Along the length of these transcripts, we also observed enrichment of histone H3 lysine 36 trimethylation (H3K36me3). Moreover, these transcripts show strong purifying selection in their genomic loci, exonic sequences, and promoter regions, implying functional constraints on the evolution of these transcripts. These results define a collection of novel transcripts in the mouse genome and indicate their potential functions in the mouse tissues and cells.

Keywords: novel transcripts, non-coding RNA, ribo-minus RNA-seq, next-generation sequencing

INTRODUCTION

The mammalian transcriptomes are much more complex than what we have been anticipated according to the related research activities over the past decade. Recently, novel transcripts have been continuously identified in mammalian genomes. Bertone et al. (2004) found 10,595 novel transcribed sequences in human liver tissue. Carninci et al. (2005) demonstrated that the majority of the mammalian genome is transcribed and reported 16,247 new mouse protein-coding transcripts. The ENCODE pilot project reported that the human genome is pervasively transcribed and discovered the relationship between transcripts and chromatin accessibility features (Birney et al., 2007). According to the chromatin-state maps, about 1,600 large multi-exonic RNAs were identified by Guttman et al. (2009) in mouse. Cabili et al. (2011) presented an integrative approach and defined >8,000 human lincRNAs. Trapnell et al. (2010) got 3,724 previously un-annotated transcripts in mouse and 62% of them were supported by independent expression data or homologous genes in other species. These novel transcripts are called the "dark matter" RNAs, which include any RNAs whose functions are still unknown (Kapranov et al., 2010). Kapranov et al. (2010) concluded that the "dark matter" RNA can be greater than protein-encoding transcripts and a large number of long non-coding RNA reside in intergenic regions.

However, controversial opinions still exist. It has been suggested that most novel transcribed regions are associated with known neighboring gene models. For example, by mapping and quantifying mouse transcriptome using poly(A) selected RNA-seq data, 92% of novel transcription regions can be assigned to their neighboring genes in a recent study (Mortazavi et al., 2008). van Bakel et al. (2010) also concluded that most non-exonic transcribed sequence fragments (seqfrags) probably are indeed partial fragments of pre-mRNA with introns, new exons of known genes in intergenic sequences, or promoter- and terminator-associated transcripts. Clark et al. (2011) and van Bakel et al. (2011) have discussed possible mechanisms of the pervasive transcription and some of the arguments are focused on universality and functionality of these novel transcripts (Jarvis and Robertson, 2011). In addition, studies have suggested that non-coding RNAs are important in transcriptional and post-transcriptional regulations, chromatin-modification, development and diseases, such as cancers (Gupta et al., 2010; Mattick et al., 2010; Glass et al., 2011; Kogo et al., 2011) and indeed fundamental to eukaryotic evolution (Mattick, 2010).

Recently, RNA-seq methods, mRNA-based, or ribo-minus (rm) based on the next-generation sequencing technologies, are considered to be more accurate and comprehensive for transcriptome profiling (Wang et al., 2009). They are supreme over other transcriptomic methods, including expressed sequence tag (EST), serial analysis of gene expression (SAGE), and microarray, in dynamic range, sampling depth, and material processing. The methods allow researchers to acquire adequate amount of data to characterize novel transcripts, and moreover, when combined with other complementary data, such as those from cap analysis of gene expression (CAGE), histone modification, and RNAPII, as well as sequence conservation analysis, they provide stronger evidence for identifying novel transcription.

In this study, we used publicly available rmRNA-seq data from the mouse cerebrum, testis, and embryonic stem (ES) cells to excavate new transcripts and verify their existence in the mouse genome, with an anticipation that rmRNA-seq data are expected to contribute more coding and non-coding transcripts, which lack polyA tails (Cui et al., 2010). We built a pipeline to identify expressed regions and candidate exons in the entire genome to define novel transcripts through comparison to known transcripts and carried out a combined analysis on relevant public data, including CAGE (Kawaji et al., 2006), histone modifications (H3K4me3, H3K27me3, and H3K36me3) and RNAPII (Mikkelsen et al., 2007), and sequence conservation values (Fujita et al., 2011). We also examined potential functions of these novel transcripts according to their sequence structures and characteristics. We expect to provide useful insights into the "dark matter" of the mouse genome.

MATERIALS AND METHODS

DATASETS

The transcriptome profiling of mouse cerebrum, testis, and ES cells, as well as data for histone modifications (H3K4me3 and H3K27me3) of mouse cerebrum and testis were from NCBI SRA database, SRA039962 and SRX005943, which were produced by our group previously. We also retrieved ChIP-seq data of RNAPII, H3K4me3, H3K27me3, and H3K36me3 from mouse ES cells[1] (Mikkelsen et al., 2007) and 5′ CAGE tags from multiple mouse tissues published by the Fantom3 project[2] (Kawaji et al., 2006). In addition, we obtained conservation scores from the UCSC database[3] (Fujita et al., 2011).

EXON OR TRANSCRIPTION UNIT (TU) IDENTIFICATION BASED ON rmRNA-seq DATA

We built an efficient pipeline for TU identification (**Figure A1** in Appendix; File S3 in Supplementary Material). First, RNA sequencing reads were mapped to the mouse genome assembly (mm9) by using TopHat (Langmead et al., 2009; Trapnell et al., 2009) and the coverage files were created based on mapping results by using a custom-designed perl script. Second, according to the coverage files, we obtained average coverage of all Refgene introns and set a cutoff value of the coverage to exclude 95% of introns (3, 4, and 7 for cerebrum, ES cells, and testis, respectively). To define the expressed regions, we limited each region to have at least 55-bp consecutive length and all these positions must be equal or above the cutoff value. If the distance of adjacent expressed regions (exons) is equal or smaller than the length of 95 bp (95% intron lengths are larger than 95 bp), we combined the adjacent expressed regions into one. We also revised the boundaries of exons using the split read feature from TopHat. Third, we evaluated the accuracy of exon identification, calculated the average coverage for exons defined in Refgene introns and removed exons whose coverage below the cutoff value. Fourth, we annotated and removed certain exons by comparing our putative exons with several databases (UCSC, ENSEMBL, NONCODE, RNAdb, fRNAdb, Rfam, miRBase, tRNAdb, and ncRNAdb). Fifth, we constructed TUs for exons found in intergenic regions according to the distance between exons, RNAPII signals, and H3K36me3 signals. Sixth, to assess the accuracy of this method, we compared our TUs (by using all exons in intergenic regions) with the Fantom3 RNAs.

IDENTIFICATION OF ENRICHED INTERVALS OF ChIP-seq DATA

We defined H3K4me3- and H3K27me3-enriched intervals by using SICER program (v1.03; Zang et al., 2009). The parameters were set as follows: (1) 200-bp window, 200-bp gap, and 0.001 for False Discovery Rate of H3K4me3; (2) 200-bp window, 600-bp gap, and 0.001 for False Discovery Rate of H3K27me3. The sequencing reads from a pan-H3 experiment was used as a background control for H3K4me3 and H3K27me3. The H3K36me3 and RNAPII enrichment intervals were downloaded from the website at the Broad Institute (see text footnote 1; Mikkelsen et al., 2007). Chromatin states of exons or TUs were determined based on overlapping regions where H3K4me3, H3K27me3, H3K36me3, and RNAPII are all enriched.

CONSERVATION OF EXONS, TUs, AND THEIR PROMOTERS

To estimate sequence conservation of exons, TUs, and their promoters, we used conservation scores derived from an alignment of 29-vertebrate-to-mouse genomes from the UCSC database (Fujita et al., 2011). We calculated the conservation score in a 12-bp sliding window with a step length of 1 bp and selected the maximal value as the conservation score. The sequences that have higher conservation scores are more conservative than other sequences.

[1] ftp://ftp.broad.mit.edu/pub/papers/chipseq/

[2] http://fantom3.gsc.riken.jp/db/

[3] http://genome.ucsc.edu/

CORRELATION BETWEEN SENSE AND ANTISENSE GENE EXPRESSION

We extracted the information for sense–antisense gene pairs and calculated the RPKM value for the sense and antisense genes based on mapping results. We subsequently divided the sense–antisense gene pairs into two portions according to their expression ratios between two samples for the sense and antisense expression. The expression ratio is equal to the sample 2 expression divided by the sample 1 expression. We classified them as positive if the log10 value of both sense and antisense expression ratios are greater or less than zero. Otherwise, we classified them as negative. We correlated the positive and negative types of sense–antisense gene pairs using the expression ratio.

RESULTS

IDENTIFICATION OF ACTIVELY TRANSCRIBED REGIONS (EXONS)

We obtained rmRNA-seq data from the mouse cerebrum, testis, and ES cells which were generated based on the SOLiD sequencing platforms using a strand-specific rmRNA-seq method (Cui et al., 2010), and mapped rmRNA-seq reads onto the mouse genome assembly (mm9) using TopHat software (**Table A1** in Appendix). Based on the mapped reads, we assessed sequencing saturation according to the increase of read start-points with increasing mapped read (**Figure 1**). To define actively transcribed regions, we calculated the coverage per nucleotide position and used those positions whose coverage values are equal or larger than the cutoff values (3, 4, and 7 for cerebrum, ES cells, and testis, respectively; also see Materials and Methods for details; **Figure 2A**). Moreover, due to sequencing bias, we required that each region to have at least 55 bp consecutive sequence above the cutoff value of coverage and the distance of adjacent regions is larger than 95bp (**Figures 2C,B**). Consequently, we obtained 395,546, 465,149, and 194,996 putative exons in the total in the three libraries, respectively (**Table 1**; File S1 in Supplementary Material). For assessing the accuracy of exon identification, we compared the defined actively transcribed regions to Refgene exons (Karolchik et al., 2004), and found that most Refgene exons (∼94.12%) have been identified and that the aligned length is up to ∼88.71%. Furthermore, ∼93.81% RefSeq-defined exons are shown to be one-to-one matches (**Table A2** in Appendix). These statistics proved the viability of our pipeline for this analysis. Moreover, we found that different samples have different percentages of reads assembled into exons (**Table A3** in Appendix). We believe that such variability is related to read length, read coverage, and the cutoff value of the read coverage.

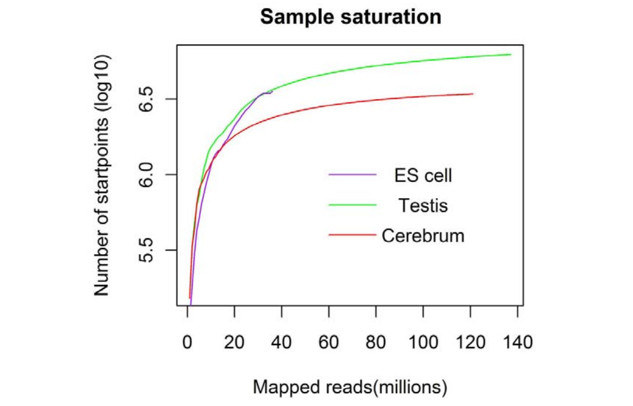


FIGURE 1 | The saturation curve for the number of start-points of mapped reads. *x*-Axis shows the number of the mapped reads and *y*-axis displays the start-points number (million) of mapped reads.

ANNOTATION OF NOVEL EXONS

To annotate novel transcripts, we first removed known exons according to the Refgene collection. We then removed all other known exons that have matches to other databases, such as the NCBI nr database (Johnson et al., 2008) based on sequence alignment using the BLAST software packages. The repeat regions of the mouse genome were avoided according to the repeat annotation at UCSC (Fujita et al., 2011). We also built a custom-designed ncRNA database through integrating several databases that include mouse ncRNA data in ENSEMBL (Flicek et al., 2011), UCSC (Fujita et al., 2011), NONCODE (He et al., 2008), RNAdb (Pang et al., 2007), fRNAdb (Kin et al., 2007), ncRNAdb (Szymanski et al., 2007), Rfam (Griffiths-Jones et al., 2005), miRBase (Griffiths-Jones et al., 2006), and tRNAdb (Juhling et al., 2009; **Table A4** in Appendix). Moreover, we filtered the newly identified exons of known genes using the split reads from the TopHat result. We also predicted the function of novel exons (unannotated) by comparing them to the Rfam database. Most Rfam-predicted exons are snoRNAs, but some are miRNAs, tRNAs, and snRNAs. Finally, we obtained three sets of putative novel exons (**Table 2**).

BUILDING NOVEL TRANSCRIPTION UNITS (TU) IN INTERGENIC REGION

Since well-defined actively transcribed regions exhibit obvious gene structure features, we tried to connect the neighboring active regions into the same transcription units (TUs). When we calculated the distance of adjacent actively transcribed regions, we found that there are two main peaks in the density plots and this feature can be used for building novel TUs (**Figure 3**). In addition, there is a small peak appeared around 100 bp in length, which is a characteristic of the minimal intron (∼100 bp in length) described in our previous publications (such as Zhu et al., 2010). The first major peak represents the distance of adjacent exons inside TUs and the second major peak is related to the distance of exons between adjacent TUs. Moreover, we downloaded the RNAPII and H3K36me3 data of ES cell, which were used to define the transcription start and the elongation of the transcripts, respectively. We finally constructed TUs for novel exons in intergenic regions according to the information from the distance between exons, RNAPII signals, and H3K36me3 signals, producing 17,931, 18,512, and 6,966 annotated TUs in cerebrum, testis, and ES cells, respectively (File S2 in Supplementary Material).

To evaluate our processing algorithm, we compared our TUs with the intergenic RNAs annotated by the Fantom3 project. As expected, the one-to-one matching rate is about 95.62%, but the aligned length is a little bit lower, ∼70.99% (**Table A5** in Appendix). The reason why the aligned length is not as high as the matching rate is that we may lose some exonic sequences due to their low coverage in the real data. It can be improved when more rmRNA-seq data are added. Nevertheless, the matching rate encouraged us to proceed.

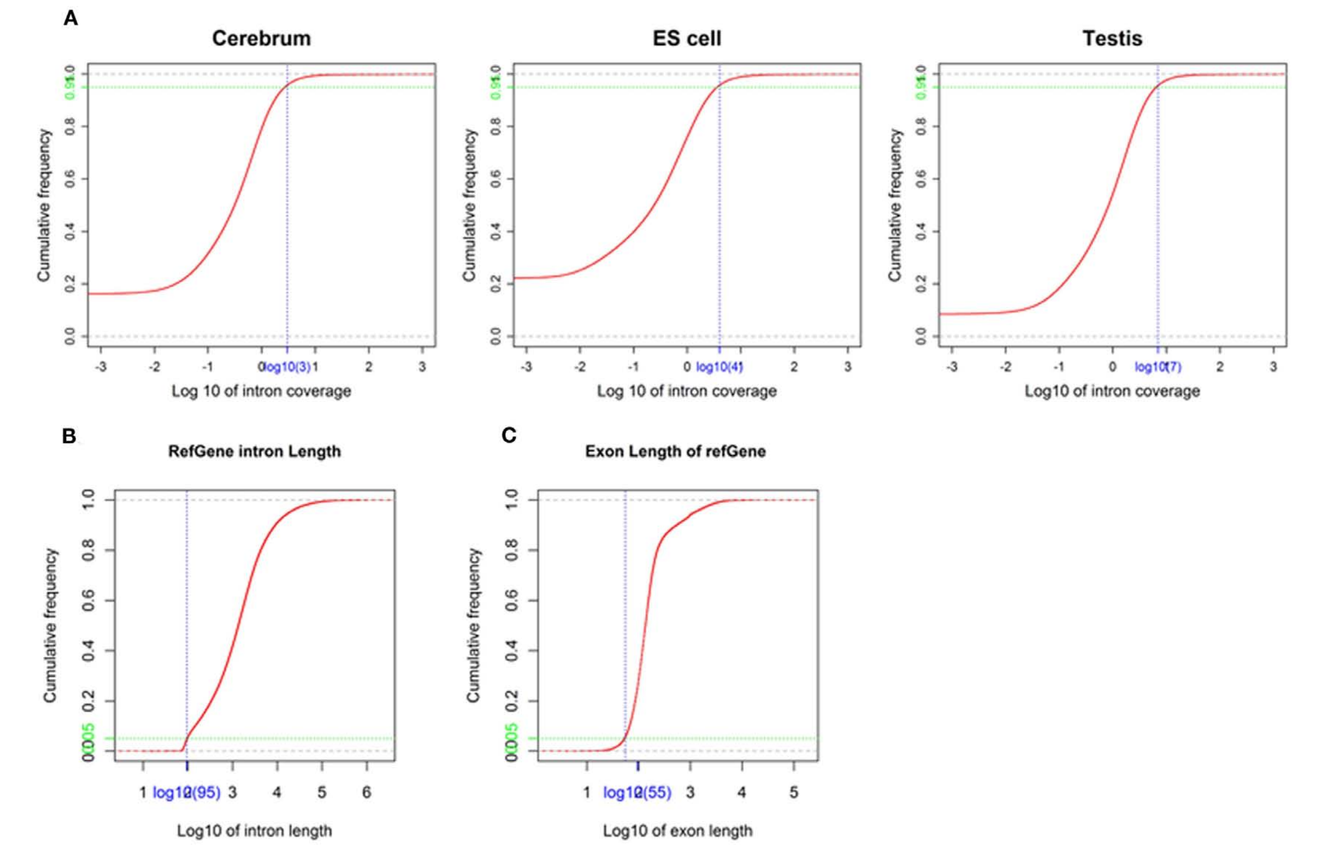


FIGURE 2 | Parameters used in exon identification. (A) The cutoff value of coverage in the mouse cerebrum, testis, and ES cells. The cutoff value (blue) of coverage (3, 4, and 7 for cerebrum, ES cells, and testis, respectively) is labeled on the *x*-axis and the corresponding accumulative frequency (0.95, colored in green) is labeled on the *y*-axis. **(B)** The minimal intron length used in exon identification. The value (blue) on *x*-axis is identified as minimal intron length (95) and the value (green) on *y*-axis is the corresponding accumulative frequency (0.05). **(C)** The minimal exon length used in exon identification. The value (blue) on the *x*-axis is identified as minimal exon length (55) and the value (green) on the *y*-axis is the corresponding accumulative frequency (0.05).

Table 1 | Summary of novel exons identified in our analysis.

Sample	Cerebrum	Testis	ES cell
Identified exons[1]	395,546 (105,657,702, 100%)[3]	465,149 (109,695,106, 100%)	194,996 (28,838,854, 100%)
Refgene exons	106,218 (25,924,734, 24.54%)	98,065 (38,325,083, 34.94%)	84,792 (18,396,077, 63.79%)
Refgene introns	233,775 (33,864,388, 32.95%)	243,879 (32,067,720, 29.23%)	75,426 (7,308,663, 25.34%)
Intergenic regions	69,971 (45,868,580, 43.41%)	135,644 (39,302,303, 35.83%)	43798 (3,134,114, 10.87%)
Refgene introns (filtered)[2]	33,053 (27,401,823, 25.93%)	28,931 (15,245,705, 13.9%)	10,011 (3,107,050, 10.77%)

[1] Because there are some overlaps among Refgene exons, Refgene introns, and intergenic regions due to gene alternatively spliced isoforms, the identified exons is less than the sum of Refgene exons, Refgene introns, and intergenic regions. [2] We removed the exons whose average coverage is below the cutoff value to reduce the errors of exon identification. [3] The numbers of identified regions are listed, and the numbers of reads and percentages of the region-specific reads over all reads are in the parentheses.

We subsequently compared our intergenic TUs with the intergenic vlinc regions identified by Kapranov et al. (2007) in human. The coordinates of the 580 vlinc RNA domains were transformed from the hg18 to the mm9 version of the mouse genome, and we converted 486 vlinc RNAs successfully. The total matched vlinc RNAs and the total one-to-one matched vlinc RNA are 316 and 278, respectively (**Table A6** in Appendix). This result implicated that many intergenic TUs are conserved among mammalian genomes. The one-to-one matching rate between intergenic TUs and vlinc RNAs is lower than what between intergenic TUs and the Fantom3 RNAs, and it may be resulted from expression regulation of intergenic TUs and the evolution of intergenic TUs among different species.

Table 2 | The exon annotation based on Refgene intron (filtered) and intergenic regions.

Sample	Cerebrum	Testis	ES cell
Total exons[1]	103,024 (73,270,403, 100%)[2]	164,575 (54,548,008, 100%)	53,809 (6,241,164, 100%)
nr	14,333 (7,613,882, 10.39%)	24,343 (19,326,301, 35.43%)	13,556 (1,803,672, 28.90%)
ncRNA	3,933 (2,372,724, 3.24%)	7,862 (4,096,291, 7.50%)	4,547 (97,395, 1.56%)
Repeat	19,116 (15,226,166, 20.78%)	34,015 (11,659,836, 21.38%)	13,109 (971,411, 15.56%)
New exons of known genes	1,073 (46,792, 0.06%)	1,404 (64,663, 0.12%)	1,101 (8,593, 0.14%)
Rfam prediction	207 (4,306,737, 5.90%)	219 (617,575, 1.13%)	104 (15,076, 0.24%)
Remaining	64,357 (43,704,102, 59.63%)	96,607 (18,783,342, 34.44%)	19,790 (3,345,017, 53.60%)

[1] Total number of exons equals to the sum of exons in Refgene intron (filtered) and intergenic regions. Because there are overlaps among Refgene introns and intergenic regions due to alternative spliced isoforms, the total number of exons found in Refgene introns and intergenic regions is more than the total number of annotated exons. [2] The numbers of identified regions are listed, and the numbers of reads and percentages of the region-specific reads over all reads are in the parentheses.

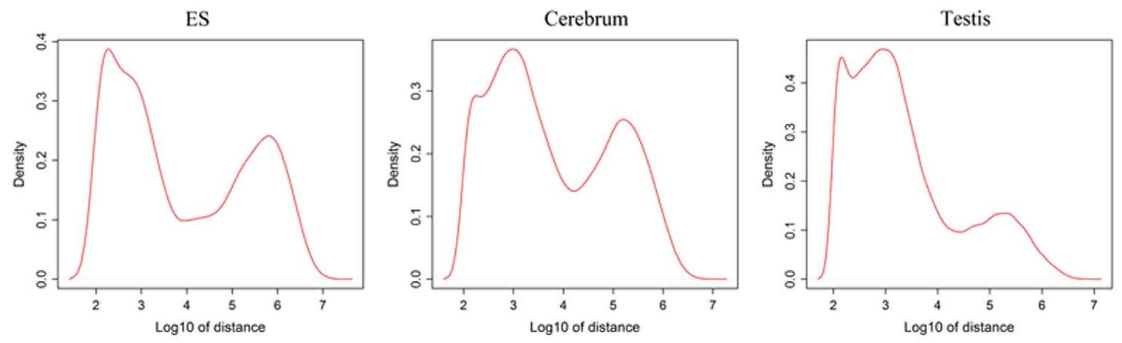


FIGURE 3 | Density distribution for distances between adjacent exons in intergenic regions. *x*-Axis shows the distance between adjacent exons and *y*-axis displays the density. There is a small peak appeared around 100 bp length and this peak implies the minimal introns. In addition, the first main peak represents the distance of general adjacent exons inside TUs and the second main peak is related to the distance of exons between adjacent TUs.

THE EVIDENCE OF NOVEL TUs IN INTERGENIC REGION

To define the function of novel TUs, we examined the distribution of 5′ CAGE tags (Kawaji et al., 2006) and RNAPII (Mikkelsen et al., 2007), histone modifications (H3K4me3, H3K27me3, and H3K36me3; Mikkelsen et al., 2007) around the TUs, and evaluated their sequence conservation value (Fujita et al., 2011). First, we found that there is a significant enrichment of 5′ CAGE tags at the TU start in all three samples (**Figure 4A**), suggesting that these TUs have 5′ 7-methylguanosine caps and possess transcriptional start sites. Moreover, we investigated the binding of RNAPII within upstream of these TUs using RNAPII data from mouse ES cells and observed an obvious enrichment of RNAPII around their TSS (**Figure 5A**), suggesting that the TUs have their own promoters for regulating transcriptional initiation. Second, based on ChIP-seq data for the three mouse samples, we examined H3K4me3, H3K27me3, and H3K36me3 statuses around the TUs (**Figures 4B–D**) and observed that H3K4me3 and H3K27me3 are enrichment at the upstream of the TUs and their densities are correlated well with gene expression. Moreover, H3K36me3 are also enriched across the TUs and marked the transcriptional elongation sites. These lines of evidence suggested that these novel TUs may be indeed independently transcribed in the samples. Finally, we investigated the sequence conservation of the novel TUs (see Materials and Methods) by calculating their conservation scores of the exonic sequences in comparison with Refgene protein exons and random sequences as controls. The conservation scores of the novel TU exons are highly similar to those of the Fantom3 RNA-defined exons (**Figure 5B**) and similar results were observed in the promoter conservation scores (**Figure 5C**). Results from both analyses suggest possible functionality of the novel TUs.

To illustrate the related characteristics of the novel TUs, we showed an intronic TU and an intergenic TU in **Figure A2** in Appendix. Both TUs are significantly expressed in the tissues and cell. Moreover, the exons of both TUs have homologous sequences according to their conservation scores (0 means no conservation and 1 means highly conserved).

CLASSIFICATION AND FUNCTION ANALYSIS OF NOVEL TUs AND EXONS

In order to explore the functionality of the novel TUs, we predicted their protein-coding capability based on PhyloCSF result (**Table 3**; Lin et al., 2011). We obtained their amino acid sequences (for single exon TUs) according to ORF prediction and aligned the amino acid sequences to the NCBI nr database. We found that most of these protein-coding TUs are similar to either known (such as ribosomal proteins and dehydrogenase) or hypothetical proteins (such as hypothetical and unnamed proteins). For

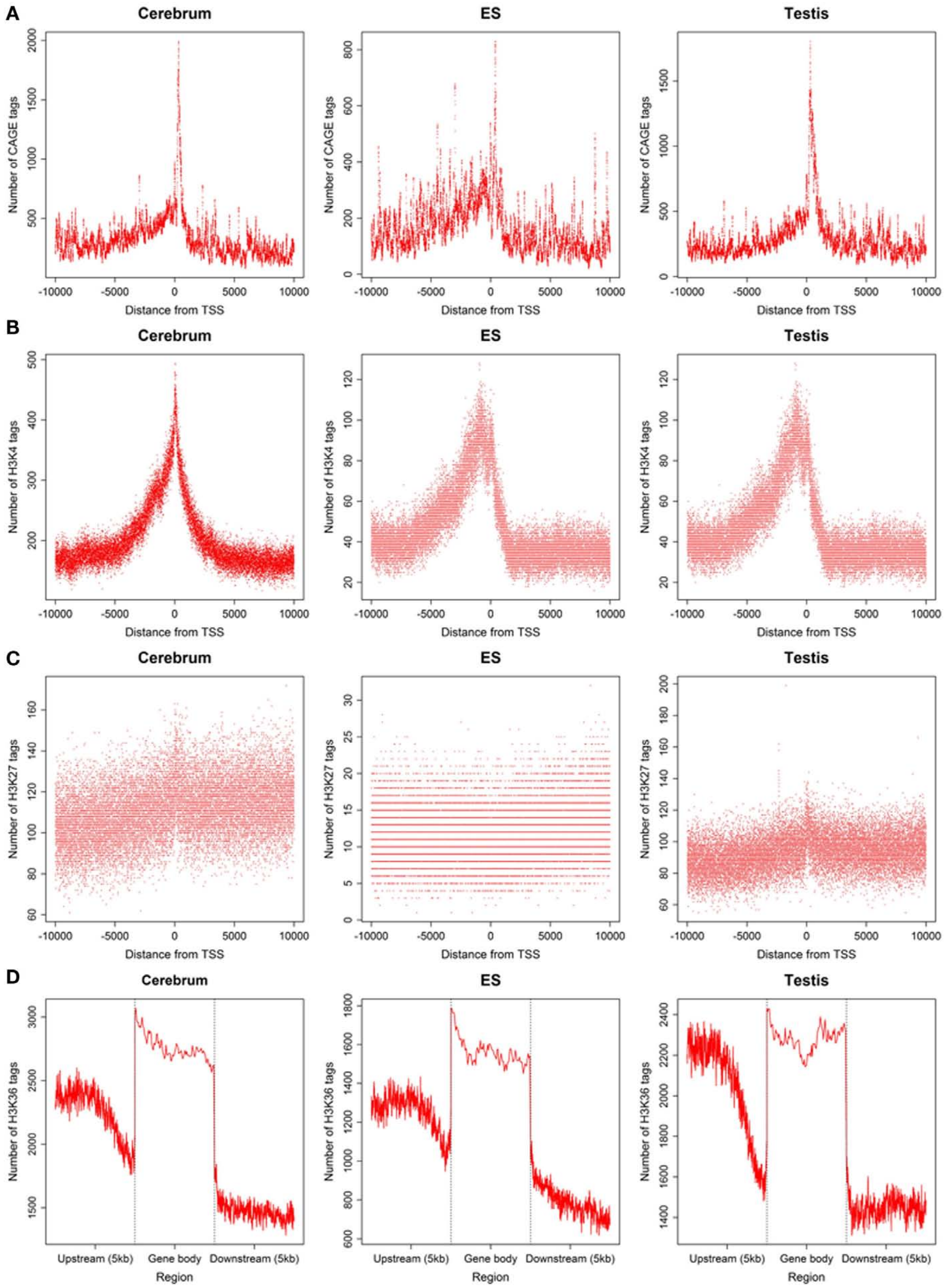


FIGURE 4 | 5′ CAGE and histone modification around novel TU–TSS or gene bodies in the mouse cerebrum, testis, and ES cell. (A–D) Profiles of 5′ CAGE, H3K4me3, H3K27me3, and H3K36me3.

exploring whether these protein-coding TUs are pseudogenes, we compared them to the Vega pseudogene annotations, and only about 8.85% of them are likely to be pseudogenes (**Table A7** in Appendix). Moreover, we selected two protein-coding transcripts and predicted their secondary structures. One of them is similar to mouse mCG1041001 protein and is predicted to be extracellularly located. The other is similar to mouse EG382421 protein and possesses nuclear localization sequence.

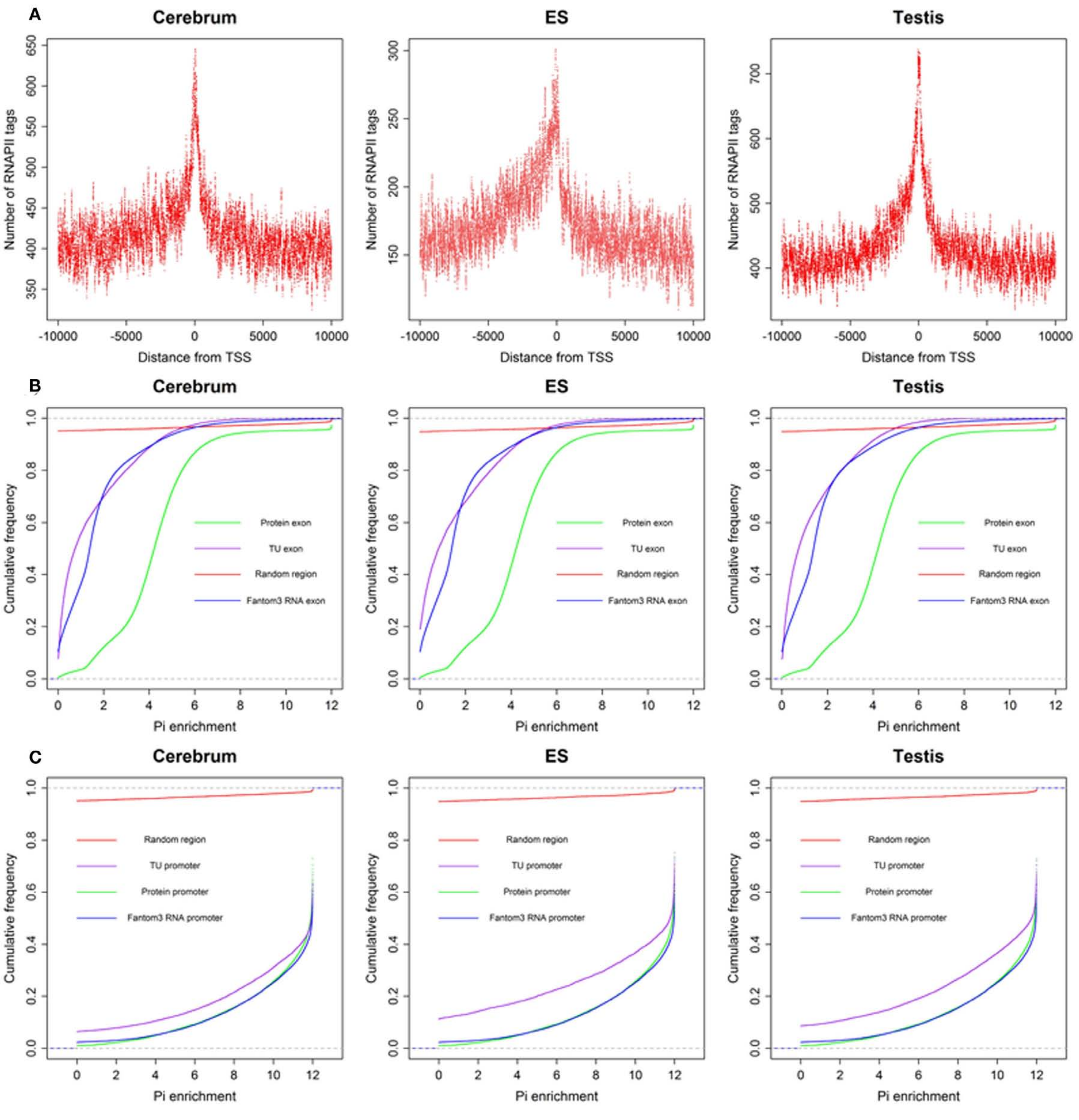


FIGURE 5 | The RNAPII around novel TU–TSS and sequence conservation of TU exon and promoters. (A) Profile of RNAPII, **(B)** cumulative distribution of sequence conservation for TU exon, protein exon, Fantom3 RNA exon, and random region, and **(C)** cumulative distribution of sequence conservation for TU promoter, protein promoter, Fantom3 RNA promoter, and random region.

We looked into the antisense regulation of the novel TUs. According to the PhyloCSF prediction, 65.93% of the novel TUs can be defined as non-coding RNAs due to lacking protein-coding characteristics (**Table 4**). To further examine them, we extracted the antisense RNAs by comparing the location of the TUs to known genes as putative cis-antisense RNAs ($n_{\text{cerebrum}} = 2{,}614$, $n_{\text{testis}} = 2{,}756$, and $n_{\text{ES}} = 732$) and their target genes ($n_{\text{cerebrum}} = 2{,}324$, $n_{\text{testis}} = 2{,}356$, and $n_{\text{ES}} = 689$). Since previous studies have suggested that sense–antisense gene pairs may play potential regulatory roles (Okada et al., 2008), we clustered the sense–antisense regulated genes using DAVID website (Huang da et al., 2009; Huang et al., 2009) and found that most

Table 3 | Summary of coding and non-coding exons and TUs.

Sample	Cerebrum	Testis	ES cell
New exons[1]	64,357	96,607	19,790
Intron exons	19,986	16,263	3,703
Coding exons[2]	3,911 (3,625,424)[6]	3,308 (579,368)	1,716 (73,420)
Non-coding exons[2]	15,973 (2,865,243)	12,866 (1,583,086)	1,952 (108,833)
Unknown exons[2]	102 (3,526,680)	89 (14,187)	35 (215)
Intergenic exons	44,525	80,489	16,153
Coding exons[3]	7,440 (4,073,104)	8,717 (535,205)	2,703 (28,257)
Non-coding exons[3]	36,363 (15,469,822)	69,193 (4,068,501)	12,690 (238,595)
Unknown exons[3]	722 (150,523)	2,579 (214,720)	760 (2,599)
Intergenic TUs	17,931	18,512	6,966
Coding TUs[3]	5,441 (4,100,203)	4,794 (1,144,421)	2,005 (54,654)
Non-coding TUs[3]	11,735 (11,077,673)	12,230 (2,212,637)	4,618 (193,861)
Unknown TUs[3]	426 (1,286,476)	698 (559,251)	281 (8,639)
Inconsistent TUs[3,4]	329 (174,288)	790 (814,202)	62 (8,232)
Modified non-coding TUs[5]	12,445 (11,077,673)	13,199 (2,212,637)	4,963 (193,861)

[1] Because there are overlaps among Refgene intronic and intergenic regions due to alternative spliced isoforms, the number of novel exons is more than the total of intronic exons and intergenic exons. [2] These exons are in known introns. [3] These exons reside in intergenic regions. [4] Inconsistent TUs means those have abnormal exon-patterns, such as non-coding-coding-non-coding-coding. [5] Because sometimes two adjacent TUs are combined into one in intergenic regions, we correct them manually to yield modified non-coding TUs. [6] The number of identified regions and their reads are outside and inside the parentheses, respectively.

Table 4 | The classification of ncRNAs in intronic and intergenic regions.

Sample	Cerebrum	Testis	ES cell
Intron RNAs	15,973 (2865243, 100%)[3]	12,866 (1,583,086, 100%)	1,952 (108,833, 100%)
Antisense RNAs[1]	141 (4,347, 0.15%)	156 (12,829, 0.81%)	19 (626, 0.58%)
Small ncRNAs[1]	14,196 (1,781,616, 62.18%)	10,483 (764,884, 48.32%)	1,467 (30,496, 28.02%)
Long ncRNAs[1]	1,636 (1,079,280, 37.67%)	2,227 (80,5373, 50.87%)	466 (77,711, 71.40%)
Intergenic RNAs	12,445 (11,077,673, 100%)	13,199 (2,212,637, 100%)	4,963 (193,861, 100%)
Antisense RNAs[2]	2,614 (792,653, 7.16%)	2,756 (440,655, 19.92%)	732 (27,289, 14.08%)
Small ncRNAs[2]	6,502 (9,716,663, 87.71%)	5,072 (624,312, 28.22%)	2,271 (27,232, 14.05%)
Long ncRNAs[2]	3,329 (568,357, 5.13%)	5,371 (1,147,670, 51.86%)	1,960 (139,340, 71.87%)

[1] These ncRNAs are intronic. [2] These ncRNAs are intergenic. [3] The numbers are those of the identified regions. The numbers in the parentheses are the read number of a region and its rate over all reads in the regions.

of these antisense regulated genes are associated with synapse, ion binding/transport, cell junction, cytoskeletal, membrane, and signal transduction in the three samples (S1–S3 in Supplementary Material). Moreover, we found that the genes in cerebrum and testis are related to cardiomyopathy, cancer, endocytosis, cell junction, and signal pathway (S4 and S5 in Supplementary Material). The expression levels of the sense–antisense transcripts are either positively or negatively correlated among different tissues and cell lines (Katayama et al., 2005; Okada et al., 2008). We also compared the sense–antisense expression in a pairwise fashion among the three samples (see Materials and Methods) and found that the antisense expression is either positively or negatively associated with the sense expression (**Figure 6**). This characteristic is in agreement with previous studies (Katayama et al., 2005; Okada et al., 2008).

We examined the novel TUs to see if some of them are actually non-coding RNAs. We divided the remaining (non-exonic) novel ncRNAs into long or small ncRNAs according to their sizes. About 43.87% of the remaining ncRNAs are larger than 200 bp in size, which were defined as long ncRNAs. There are 3,329, 5,371, and 1,960 novel long ncRNAs identified in the cerebrum, testis, and ES cells, respectively (**Table 4**). Comparing our long ncRNAs to lincRNAs identified by Guttman et al. (2009), we found 724 lincRNAs in our three samples, which are accounted for 43.48% of all lincRNAs. There are about 21% of lincRNAs found in each of our samples ($n_{\text{cerebrum}} = 359$, $n_{\text{testis}} = 391$, and $n_{\text{ES}} = 304$).

We defined the rest of the ncRNAs as small ncRNAs, ranging from 55 to 200 bp in length; the majority of these small ncRNAs (∼24.04%) are from 55 to 65 bp in size (**Figure 7A**). This size range of small ncRNAs is related to the insert size of the

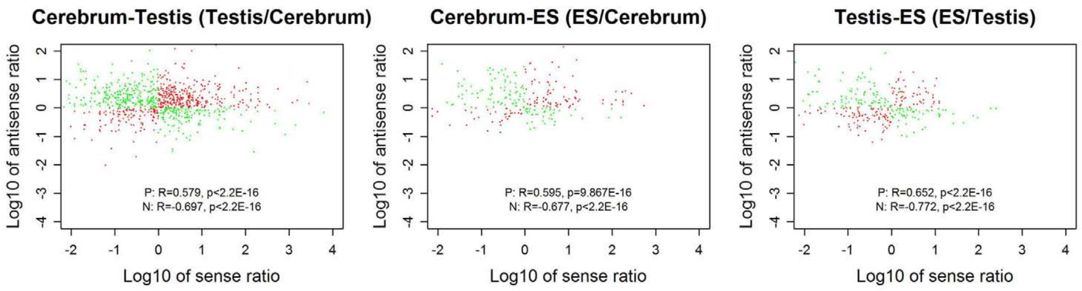


FIGURE 6 | The correlation between sense and antisense expression ratio in sense–antisense gene pairs. Red and green points represent the sense–antisense gene pairs in positive and negative types. "P" stands for the positive type and "N" stands for the negative type.

libraries and data processing parameters. First, we selected RNA fragments in a length range of 50–150 bp for analysis. Second, we filtered the small RNAs whose lengths are less than 55 bp and have overlapping sequences so that some of the smaller RNAs were eliminated in data processing procedures. For these small ncRNAs, we predicted their motifs using MEME software (Bailey and Elkan, 1994) and some conserved motifs were identified (**Figures 7B–D**), which were accounted for ∼20% of all small ncRNAs. To explore the relationship between conserved motifs and RNA structures, we calculated two distances: one is what between the RNA 5′ end and the motif start and the other is what between the motif end and RNA 3′ end; we did not observe any obvious patterns in the motif distribution (**Figure A3** in Appendix). We also compared the novel ncRNAs among the three samples and found that the ratios of the tissue- or cell-specific novel ncRNAs are larger than the ratio of known genes (**Figure A4** in Appendix). The biased distribution of the novel ncRNAs indicates their possible functional roles in different tissues or cell types. More ncRNA expression data from a broader tissue spectrum are certainly needed to decipher the functionality of the ncRNAs.

ACTIVELY TRANSCRIBED INTRONIC REGIONS

Based on the PhyloCSF prediction (**Table 3**), we identified about 79.52% ncRNA exons in the intronic regions of the cerebrum and testis, whereas only 52.71% ncRNA exons in the intronic regions of ES cells. Whether most of the predicted protein-coding exons are actually parts of known genes remains to be elucidated. For the analysis of ncRNAs in the intronic regions, we also divided them into three portions: antisense RNAs, small ncRNAs, and long ncRNAs (**Table 4**). Unlike ncRNAs in intergenic regions, most intronic ncRNAs are small ncRNAs: 88.88, 81.48, and 75.15% in the mouse cerebrum, testis, and ES cell, respectively. Since the intronic expressions are mostly weak and interfered by background expression, more efforts are to be devoted in the future for exploring their functions.

DISCUSSION

In this study, we attempted to identify novel transcripts using rmRNA-seq data from two mouse transcript-rich tissues, the cerebrum and testis, and ES cells. Compared to what generated from polyA-based mRNA-seq method, rmRNA-seq data are expected to harbor more novel transcripts that do not have the polyA tails typical for eukaryotic mRNA (Cui et al., 2010). In addition, we took the advantage of a strand-specific nature of the method, which is readily done using the SOLiD platform and allows us to define sense and antisense transcript pairs of the antisense regulated genes.

Using a custom-designed data processing pipeline, we carefully identified several to twenty thousands of novel TUs from different mouse tissues and cells and analyzed their distributions in both intronic and intergenic regions. We also used other supporting evidence from transcriptional initiation and epigenetic signals as well as one of the common evolutionary strategies – sequence conservation. These features helped us to argue for their functional roles in the tissues and cells. Our pipeline is able to recover ∼94.12% Refgene exons (average coverage is equal or larger than the cutoff value) from the dataset and the method is capable of driving mammalian transcriptome annotation to a completion if coupled with a protocol for characterizing even smaller RNAs, such as miRNAs.

The annotation of these novel UTs remains challenging. First, when aligning these TUs that are characteristic of amino acid sequences, such as single exons, to sequences in the NCBI nr database, we can readily annotate about 24% of the novel protein-coding TUs. Although some of them are annotated to be structural proteins, such as those similar to ribosomal proteins and housekeeping enzymes, most of them are actually matching to unknown proteins. Second, we identified a large number of ncRNAs, including antisense RNAs, small ncRNAs, and long ncRNAs. According to the analysis on the targeted genes of antisense RNAs, we found that they are associated with synapse, ion binding-transport, cell junction, cytoskeletal, membrane, and signal transduction. Surprisingly, these genes are enriched in disease related pathways, such as cardiomyopathy and cancer. We believe that such enrichment is largely an artifact due to the fields of intensive research activities. In addition, we found that antisense expression is either positively or negatively associated with sense expression of sense–antisense gene pairs. Furthermore, numerous long ncRNAs are identified in intergenic regions, providing a basis for future functional studies. Moreover, we found that the majority of small ncRNAs are in a length range of 55–65 bp in intergenic regions, which may represent a novel class of ncRNAs since conserved motifs were found among the sequences. In addition, most novel exons we found in intronic regions are small ncRNAs of the same size range.

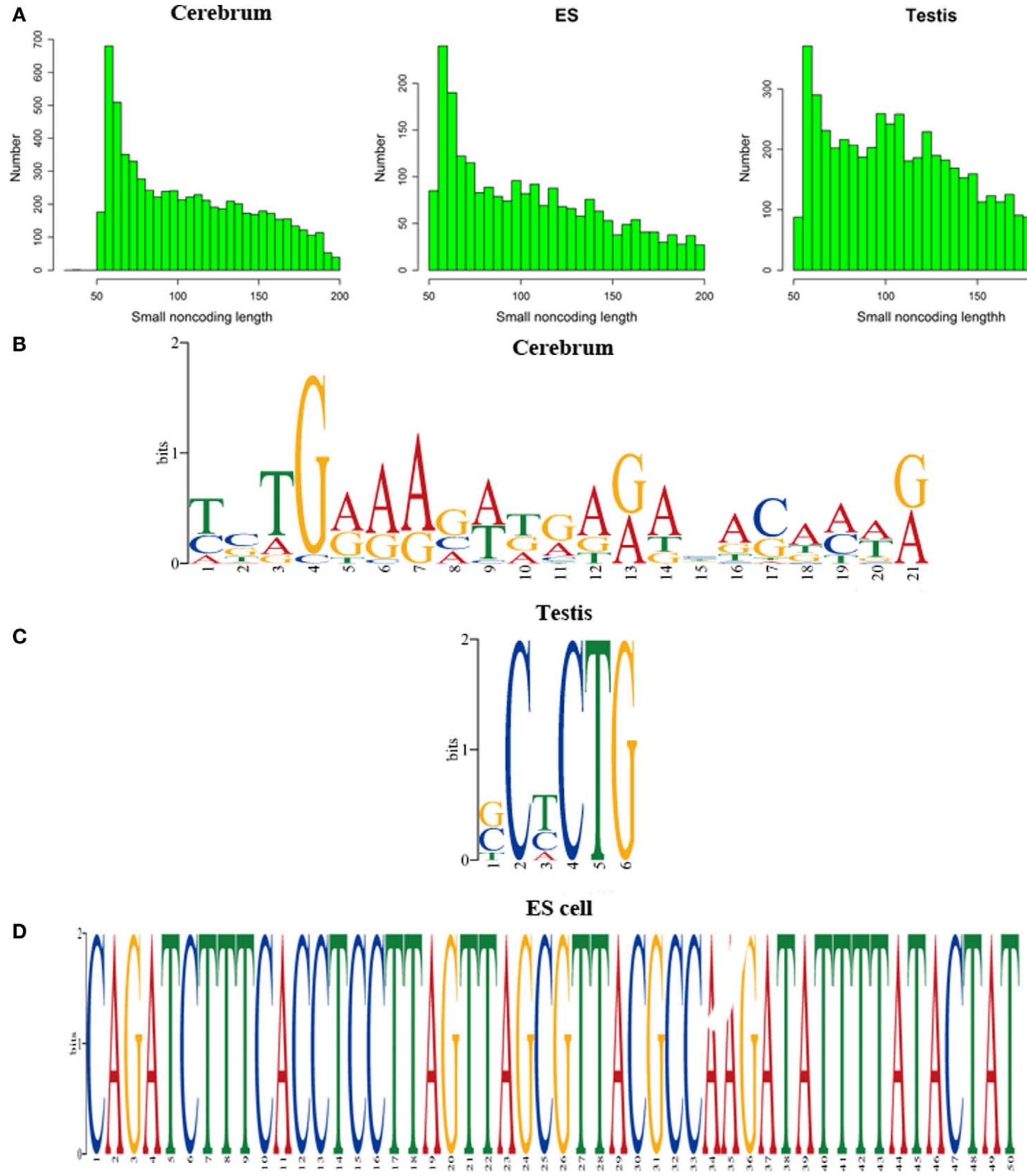


FIGURE 7 | Histograms and motif logo of small RNAs in intergenic regions. **(A)** the histogram of small RNA length, **(B)** a motif logo of small RNAs in cerebrum (16.13% of 65 bp small RNAs involved in this motif), **(C)** a motif logo of small RNA in testis (41.54% of 64 bp small RNAs involved in this motif), and **(D)** a motif logo of small RNAs in ES cell (25.97% of 56 bp small RNAs involved in this motif).

CONCLUSION

In this study, we identified a large number of novel exons and TUs using three strand-specific rmRNA-seq datasets. We also evaluated the universality and functionality of these novel TUs to demonstrate their features as actively transcribed genes based on an analysis that combines data from transcription start site, histone modification, RNAPII binding site, and sequence conservation. Our efforts in annotating these novel TUs revealed their possible functional features, resembling sequences of protein-coding and sense–antisense regulated genes as well as long and small ncRNAs. This study also provides a practical approach for the identification of most, if not all, genes of mammalian genomes.

ACKNOWLEDGMENTS

The study is supported by grants from Knowledge Innovation Program of the Chinese Academy of Sciences (KSCX2-EW-R-01-04), Natural Science Foundation of China (90919024), Natural Science Foundation of China (30900831), and the National Basic Research Program (973 Program) from the

Ministry of Science and Technology of the People's Republic of China (2011CB944100). The authors thank the anonymous reviewers for critical comments and helpful suggestions.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at http://www.frontiersin.org/non-coding_rna/10.3389/fgene.2011.00093/abstract

Table S1 | Functional classification of cerebrum-associated genes that have antisense RNA.

Table S2 | Functional classification of testis-associated genes that have antisense RNA.

Table S3 | Functional classification of ES cell-associated genes that have antisense RNA.

Table S4 | KEGG pathways of the cerebrum-associated genes that have antisense RNA.

Table S5 | KEGG pathways of testis-associated genes that have antisense RNA.

File S1 | Putatively identified exons in the mouse cerebrum, testis, and ES cells. These files have five columns. They are chromosome, strand, type ("modified" or "no-modified" for distinguishing whether it was revised using split reads), start position and end position.

File S2 | Putatively identified TUs in the mouse cerebrum, testis, and ES cells.

File S3 | The main step of identification of exons and TUs.

REFERENCES

Bailey, T. L., and Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 2, 28–36.

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

Birney, E., Stamatoyannopoulos, J. A., Dutta, A., Guigo, R., Gingeras, T. R., Margulies, E. H., Weng, Z., Snyder, M., Dermitzakis, E. T., Thurman, R. E., Kuehn, M. S., Taylor, C. M., Neph, S., Koch, C. M., Asthana, S., Malhotra, A., Adzhubei, I., Greenbaum, J. A., Andrews, R. M., Flicek, P., Boyle, P. J., Cao, H., Carter, N. P., Clelland, G. K., Davis, S., Day, N., Dhami, P., Dillon, S. C., Dorschner, M. O., Fiegler, H., Giresi, P. G., Goldy, J., Hawrylycz, M., Haydock, A., Humbert, R., James, K. D., Johnson, B. E., Johnson, E. M., Frum, T. T., Rosenzweig, E. R., Karnani, N., Lee, K., Lefebvre, G. C., Navas, P. A., Neri, F., Parker, S. C., Sabo, P. J., Sandstrom, R., Shafer, A., Vetrie, D., Weaver, M., Wilcox, S., Yu, M., Collins, F. S., Dekker, J., Lieb, J. D., Tullius, T. D., Crawford, G. E., Sunyaev, S., Noble, W. S., Dunham, I., Denoeud, F., Reymond, A., Kapranov, P., Rozowsky, J., Zheng, D., Castelo, R., Frankish, A., Harrow, J., Ghosh, S., Sandelin, A., Hofacker, I. L., Baertsch, R., Keefe, D., Dike, S., Cheng, J., Hirsch, H. A., Sekinger, E. A., Lagarde, J., Abril, J. F., Shahab, A., Flamm, C., Fried, C., Hackermuller, J., Hertel, J., Lindemeyer, M., Missal, K., Tanzer, A., Washietl, S., Korbel, J., Emanuelsson, O., Pedersen, J. S., Holroyd, N., Taylor, R., Swarbreck, D., Matthews, N., Dickson, M. C., Thomas, D. J., Weirauch, M. T., Gilbert, J., Drenkow, J., Bell, I., Zhao, X., Srinivasan, K. G., Sung, W. K., Ooi, H. S., Chiu, K. P., Foissac, S., Alioto, T., Brent, M., Pachter, L., Tress, M. L., Valencia, A., Choo, S. W., Choo, C. Y., Ucla, C., Manzano, C., Wyss, C., Cheung, E., Clark, T. G., Brown, J. B., Ganesh, M., Patel, S., Tammana, H., Chrast, J., Henrichsen, C. N., Kai, C., Kawai, J., Nagalakshmi, U., Wu, J., Lian, Z., Lian, J., Newburger, P., Zhang, X., Bickel, P., Mattick, J. S., Carninci, P., Hayashizaki, Y., Weissman, S., Hubbard, T., Myers, R. M., Rogers, J., Stadler, P. F., Lowe, T. M., Wei, C. L., Ruan, Y., Struhl, K., Gerstein, M., Antonarakis, S. E., Fu, Y., Green, E. D., Karaöz, U., Siepel, A., Taylor, J., Liefer, L. A., Wetterstrand, K. A., Good, P. J., Feingold, E. A., Guyer, M. S., Cooper, G. M., Asimenos, G., Dewey, C. N., Hou, M., Nikolaev, S., Montoya-Burgos, J. I., Löytynoja, A., Whelan, S., Pardi, F., Massingham, T., Huang, H., Zhang, N. R., Holmes, I., Mullikin, J. C., Ureta-Vidal, A., Paten, B., Seringhaus, M., Church, D., Rosenbloom, K., Kent, W. J., Stone, E. A., NISC Comparative Sequencing Program, Baylor College of Medicine Human Genome Sequencing Center, Washington University Genome Sequencing Center, Broad Institute; Children's Hospital Oakland Research Institute, Batzoglou, S., Goldman, N., Hardison, R. C., Haussler, D., Miller, W., Sidow, A., Trinklein, N. D., Zhang, Z. D., Barrera, L., Stuart, R., King, D. C., Ameur, A., Enroth, S., Bieda, M. C., Kim, J., Bhinge, A. A., Jiang, N., Liu, J., Yao, F., Vega, V. B., Lee, C. W., Ng, P., Shahab, A., Yang, A., Moqtaderi, Z., Zhu, Z., Xu, X., Squazzo, S., Oberley, M. J., Inman, D., Singer, M. A., Richmond, T. A., Munn, K. J., Rada-Iglesias, A., Wallerman, O., Komorowski, J., Fowler, J. C., Couttet, P., Bruce, A. W., Dovey, O. M., Ellis, P. D., Langford, C. F., Nix, D. A., Euskirchen, G., Hartman, S., Urban, A. E., Kraus, P., Van Calcar, S., Heintzman, N., Kim, T. H., Wang, K., Qu, C., Hon, G., Luna, R., Glass, C. K., Rosenfeld, M. G., Aldred, S. F., Cooper, S. J., Halees, A., Lin, J. M., Shulha, H. P., Zhang, X., Xu, M., Haidar, J. N., Yu, Y., Ruan, Y., Iyer, V. R., Green, R. D., Wadelius, C., Farnham, P. J., Ren, B., Harte, R. A., Hinrichs, A. S., Trumbower, H., Clawson, H., Hillman-Jackson, J., Zweig, A. S., Smith, K., Thakkapallayil, A., Barber, G., Kuhn, R. M., Karolchik, D., Armengol, L., Bird, C. P., de Bakker, P. I., Kern, A. D., Lopez-Bigas, N., Martin, J. D., Stranger, B. E., Woodroffe, A., Davydov, E., Dimas, A., Eyras, E., Hallgrímsdóttir, I. B., Huppert, J., Zody, M. C., Abecasis, G. R., Estivill, X., Bouffard, G. G., Guan, X., Hansen, N. F., Idol, J. R., Maduro, V. V., Maskeri, B., McDowell, J. C., Park, M., Thomas, P. J., Young, A. C., Blakesley, R. W., Muzny, D. M., Sodergren, E., Wheeler, D. A., Worley, K. C., Jiang, H., Weinstock, G. M., Gibbs, R. A., Graves, T., Fulton, R., Mardis, E. R., Wilson, R. K., Clamp, M., Cuff, J., Gnerre, S., Jaffe, D. B., Chang, J. L., Lindblad-Toh, K., Lander, E. S., Koriabine, M., Nefedov, M., Osoegawa, K., Yoshinaga, Y., Zhu, B., and de Jong, P. J. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447, 799–816.

Cabili, M. N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., and Rinn, J. L. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.* 25, 1915–1927.

Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M. C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C., Kodzius, R., Shimokawa, K., Bajic, V. B., Brenner, S. E., Batalov, S., Forrest, A. R., Zavolan, M., Davis, M. J., Wilming, L. G., Aidinis, V., Allen, J. E., Ambesi-Impiombato, A., Apweiler, R., Aturaliya, R. N., Bailey, T. L., Bansal, M., Baxter, L., Beisel, K. W., Bersano, T., Bono, H., Chalk, A. M., Chiu, K. P., Choudhary, V., Christoffels, A., Clutterbuck, D. R., Crowe, M. L., Dalla, E., Dalrymple, B. P., De Bono, B., Della Gatta, G., Di Bernardo, D., Down, T., Engstrom, P., Fagiolini, M., Faulkner, G., Fletcher, C. F., Fukushima, T., Furuno, M., Futaki, S., Gariboldi, M., Georgii-Hemming, P., Gingeras, T. R., Gojobori, T., Green, R. E., Gustincich, S., Harbers, M., Hayashi, Y., Hensch, T. K., Hirokawa, N., Hill, D., Huminiecki, L., Iacono, M., Ikeo, K., Iwama, A., Ishikawa, T., Jakt, M., Kanapin, A., Katoh, M., Kawasawa, Y., Kelso, J., Kitamura, H., Kitano, H., Kollias, G., Krishnan, S. P., Kruger, A., Kummerfeld, S. K., Kurochkin, I. V., Lareau, L. F., Lazarevic, D., Lipovich, L., Liu, J., Liuni, S., Mcwilliam, S., Madan Babu, M., Madera, M., Marchionni, L., Matsuda, H., Matsuzawa, S., Miki, H., Mignone, F., Miyake, S., Morris, K., Mottagui-Tabar, S., Mulder, N., Nakano, N., Nakauchi, H., Ng, P., Nilsson, R., Nishiguchi, S., Nishikawa, S., Piazza, S., Reed, J., Reid, J. F., Ring, B. Z., Ringwald, M., Rost, B., Ruan, Y., Salzberg, S. L., Sandelin, A., Schneider, C., Schönbach, C., Sekiguchi, K., Semple, C. A., Seno, S., Sessa, L., Sheng, Y., Shibata, Y., Shimada, H., Shimada, K., Silva, D., Sinclair, B., Sperling, S., Stupka, E., Sugiura, K., Sultana, R., Takenaka, Y., Taki, K., Tammoja, K., Tan, S. L., Tang, S., Taylor, M. S., Tegner, J., Teichmann, S. A., Ueda, H. R, van Nimwegen, E., Verardo, R., Wei, C. L., Yagi, K., Yamanishi, H., Zabarovsky, E., Zhu, S., Zimmer, A., Hide, W., Bult, C., Grimmond, S. M., Teasdale, R. D., Liu, E. T., Brusic, V., Quackenbush, J., Wahlestedt, C., Mattick, J. S., Hume, D. A., Kai, C., Sasaki, D., Tomaru, Y., Fukuda, S., Kanamori-Katayama, M., Suzuki, M., Aoki, J., Arakawa, T., Iida, J., Imamura, K., Itoh, M., Kato, T., Kawaji, H., Kawagashira, N., Kawashima, T., Kojima, M., Kondo, S., Konno, H., Nakano, K., Ninomiya, N., Nishio, T., Okada, M., Plessy, C., Shibata, K., Shiraki, T., Suzuki, S., Tagami, M., Waki, K., Watahiki, A., Okamura-Oho, Y., Suzuki, H., Kawai, J., Hayashizaki, Y., FANTOM Consortium, and RIKEN Genome Exploration Research Group and Genome Science Group (Genome Network Project Core Group). (2005). The transcriptional landscape of the mammalian genome. *Science* 309, 1559–1563.

Clark, M. B., Amaral, P. P., Schlesinger, F. J., Dinger, M. E., Taft, R. J., Rinn, J. L., Ponting, C. P., Stadler, P. F., Morris, K. V., Morillon, A., Rozowsky, J. S., Gerstein, M. B., Wahlestedt, C., Hayashizaki, Y., Carninci, P., Gingeras, T. R., and Mattick, J. S. (2011). The reality of pervasive transcription. *PLoS Biol.* 9, e1000625; discussion e1001102. doi:10.1371/journal.pbio.1000625

Cui, P., Lin, Q., Ding, F., Xin, C., Gong, W., Zhang, L., Geng, J., Zhang, B., Yu, X., Yang, J., Hu, S., and Yu, J. (2010). A comparison between ribominus RNA-sequencing and polyA-selected RNA-sequencing. *Genomics* 96, 259–265.

Flicek, P., Amode, M. R., Barrell, D., Beal, K., Brent, S., Chen, Y., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Gordon, L., Hendrix, M., Hourlier, T., Johnson, N., Kahari, A., Keefe, D., Keenan, S., Kinsella, R., Kokocinski, F., Kulesha, E., Larsson, P., Longden, I., Mclaren, W., Overduin, B., Pritchard, B., Riat, H. S., Rios, D., Ritchie, G. R., Ruffier, M., Schuster, M., Sobral, D., Spudich, G., Tang, Y. A., Trevanion, S., Vandrovcova, J., Vilella, A. J., White, S., Wilder, S. P., Zadissa, A., Zamora, J., Aken, B. L., Birney, E., Cunningham, F., Dunham, I., Durbin, R., Fernandez-Suarez, X. M., Herrero, J., Hubbard, T. J., Parker, A., Proctor, G., Vogel, J., and Searle, S. M. (2011). Ensembl 2011. *Nucleic Acids Res.* 39, D800–D806.

Fujita, P. A., Rhead, B., Zweig, A. S., Hinrichs, A. S., Karolchik, D., Cline, M. S., Goldman, M., Barber, G. P., Clawson, H., Coelho, A., Diekhans, M., Dreszer, T. R., Giardine, B. M., Harte, R. A., Hillman-Jackson, J., Hsu, F., Kirkup, V., Kuhn, R. M., Learned, K., Li, C. H., Meyer, L. R., Pohl, A., Raney, B. J., Rosenbloom, K. R., Smith, K. E., Haussler, D., and Kent, W. J. (2011). The UCSC Genome Browser database: update 2011. *Nucleic Acids Res.* 39, D876–D882.

Glass, C. K., Kaikkonen, M. U., and Lam, M. T. Y. (2011). Non-coding RNAs as regulators of gene expression and epigenetics. *Cardiovasc. Res.* 90, 430–440.

Griffiths-Jones, S., Grocock, R. J., Van Dongen, S., Bateman, A., and Enright, A. J. (2006). miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.* 34, D140–D144.

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, D121–D124.

Gupta, R. A., Shah, N., Wang, K. C., Kim, J., Horlings, H. M., Wong, D. J., Tsai, M. C., Hung, T., Argani, P., Rinn, J. L., Wang, Y., Brzoska, P., Kong, B., Li, R., West, R. B., Van De Vijver, M. J., Sukumar, S., and Chang, H. Y. (2010). Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* 464, 1071–1076.

Guttman, M., Amit, I., Garber, M., French, C., Lin, M. F., Feldser, D., Huarte, M., Zuk, O., Carey, B. W., Cassady, J. P., Cabili, M. N., Jaenisch, R., Mikkelsen, T. S., Jacks, T., Hacohen, N., Bernstein, B. E., Kellis, M., Regev, A., Rinn, J. L., and Lander, E. S. (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* 458, 223–227.

He, S., Liu, C., Skogerbo, G., Zhao, H., Wang, J., Liu, T., Bai, B., Zhao, Y., and Chen, R. (2008). NONCODE v2.0: decoding the non-coding. *Nucleic Acids Res.* 36, D170–D172.

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* 4, 44–57.

Huang da, W., Sherman, B. T., and Lempicki, R. A. (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* 37, 1–13.

Jarvis, K., and Robertson, M. (2011). The noncoding universe. *BMC Biol.* 9, 52. doi:10.1186/1741-7007-9-52

Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., Mcginnis, S., and Madden, T. L. (2008). NCBI BLAST: a better web interface. *Nucleic Acids Res.* 36, W5–W9.

Juhling, F., Morl, M., Hartmann, R. K., Sprinzl, M., Stadler, P. F., and Putz, J. (2009). tRNAdb 2009: compilation of tRNA sequences and tRNA genes. *Nucleic Acids Res.* 37, D159–D162.

Kapranov, P., Cheng, J., Dike, S., Nix, D. A., Duttagupta, R., Willingham, A. T., Stadler, P. F., Hertel, J., Hackermuller, J., Hofacker, I. L., Bell, I., Cheung, E., Drenkow, J., Dumais, E., Patel, S., Helt, G., Ganesh, M., Ghosh, S., Piccolboni, A., Sementchenko, V., Tammana, H., and Gingeras, T. R. (2007). RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* 316, 1484–1488.

Kapranov, P., St Laurent, G., Raz, T., Ozsolak, F., Reynolds, C. P., Sorensen, P. H., Reaman, G., Milos, P., Arceci, R. J., Thompson, J. F., and Triche, T. J. (2010). The majority of total nuclear-encoded non-ribosomal RNA in a human cell is 'dark matter' un-annotated RNA. *BMC Biol.* 8, 149. doi:10.1186/1741-7007-8-149

Karolchik, D., Hinrichs, A. S., Furey, T. S., Roskin, K. M., Sugnet, C. W., Haussler, D., and Kent, W. J. (2004). The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* 32, D493–D496.

Katayama, S., Tomaru, Y., Kasukawa, T., Waki, K., Nakanishi, M., Nakamura, M., Nishida, H., Yap, C. C., Suzuki, M., Kawai, J., Suzuki, H., Carninci, P., Hayashizaki, Y., Wells, C., Frith, M., Ravasi, T., Pang, K. C., Hallinan, J., Mattick, J., Hume, D. A., Lipovich, L., Batalov, S., Engstrom, P. G., Mizuno, Y., Faghihi, M. A., Sandelin, A., Chalk, A. M., Mottagui-Tabar, S., Liang, Z., Lenhard, B., and Wahlestedt, C. (2005). Antisense transcription in the mammalian transcriptome. *Science* 309, 1564–1566.

Kawaji, H., Kasukawa, T., Fukuda, S., Katayama, S., Kai, C., Kawai, J., Carninci, P., and Hayashizaki, Y. (2006). CAGE Basic/Analysis Databases: the CAGE resource for comprehensive promoter analysis. *Nucleic Acids Res.* 34, D632–D636.

Kin, T., Yamada, K., Terai, G., Okida, H., Yoshinari, Y., Ono, Y., Kojima, A., Kimura, Y., Komori, T., and Asai, K. (2007). fRNAdb: a platform for mining/annotating functional RNA candidates from non-coding RNA sequences. *Nucleic Acids Res.* 35, D145–D148.

Kogo, R., Shimamura, T., Mimori, K., Kawahara, K., Imoto, S., Sudo, T., Tanaka, F., Shibata, K., Suzuki, A., Komune, S., Miyano, S., and Mori, M. (2011). Long non-coding RNA HOTAIR regulates polycomb-dependent chromatin modification and is associated with poor prognosis in colorectal cancers. *Cancer Res.* 71, 6320–6326.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 10, R25.

Lin, M. F., Jungreis, I., and Kellis, M. (2011). PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* 27, i275–i282.

Mattick, J. (2010). Video Q&A: non-coding RNAs and eukaryotic evolution – a personal view. *BMC Biol.* 8, 67. doi:10.1186/1741-7007-8-67

Mattick, J. S., Taft, R. J., Pang, K. C., Mercer, T. R., and Dinger, M. (2010). Non-coding RNAs: regulators of disease. *J. Pathol.* 220, 126–139.

Mikkelsen, T. S., Ku, M., Jaffe, D. B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T. K., Koche, R. P., Lee, W., Mendenhall, E., O'donovan, A., Presser, A., Russ, C., Xie, X., Meissner, A., Wernig, M., Jaenisch, R., Nusbaum, C., Lander, E. S., and Bernstein, B. E. (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* 448, 553–560.

Mortazavi, A., Williams, B. A., Mccue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* 5, 621–628.

Okada, Y., Tashiro, C., Numata, K., Watanabe, K., Nakaoka, H., Yamamoto, N., Okubo, K., Ikeda, R., Saito, R., Kanai, A., Abe, K., Tomita, M., and Kiyosawa, H. (2008). Comparative expression analysis uncovers novel features of endogenous antisense transcription. *Hum. Mol. Genet.* 17, 1631–1640.

Pang, K. C., Stephen, S., Dinger, M. E., Engstrom, P. G., Lenhard, B., and Mattick, J. S. (2007). RNAdb 2.0 – an expanded database of mammalian non-coding RNAs. *Nucleic Acids Res.* 35, D178–D182.

Szymanski, M., Erdmann, V. A., and Barciszewski, J. (2007). Noncoding RNAs database (ncRNAdb). *Nucleic Acids Res.* 35, D162–D164.

Trapnell, C., Pachter, L., and Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25, 1105–1111.

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., Van Baren, M. J., Salzberg, S. L., Wold, B. J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* 28, 511–515.

van Bakel, H., Nislow, C., Blencowe, B. J., and Hughes, T. R. (2010). Most "dark matter" transcripts are associated with known genes. *PLoS Biol.* 8, e1000371. doi:10.1371/journal.pbio.1000371

van Bakel, H., Nislow, C., Blencowe, B. J., and Hughes, T. R. (2011). Response to "The Reality of Pervasive Transcription." *PLoS Biol.* 9. e1001102. doi:10.1371/journal.pbio.1001102

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 10, 57–63.

Zang, C., Schones, D. E., Zeng, C., Cui, K., Zhao, K., and Peng, W. (2009). A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. *Bioinformatics* 25, 1952–1958.

Zhu, J., He, F., Wang, D., Liu, K., Huang, D., Xiao, J., Wu, J., Hu, S., and Yu, J. (2010). A novel role for minimal introns: routing mRNAs to the cytosol. *PLoS ONE* 5, e10144. doi:10.1371/journal.pone.0010144

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 01 August 2011; accepted: 07 December 2011; published online: 26 December 2011.

Citation: Liu W, Zhao Y, Cui P, Lin Q, Ding F, Xin C, Tan X, Song S, Yu J and Hu S (2011) Thousands of novel transcripts identified in mouse cerebrum, testis, and ES cells based on ribo-minus RNA sequencing. Front. Gene. ***2****:93. doi: 10.3389/fgene.2011.00093*

This article was submitted to Frontiers in Non-Coding RNA, a specialty of Frontiers in Genetics.

APPENDIX

Table A1 | The sequence mapping summary of rmRNA-seq data.

Tissue/cell	Total reads	Multiple mapping reads	Unique mapping reads	Average coverage (whole genome)	Average coverage (identified exons)	Mapping percent
Cerebrum	428,434,624	124,991,301	120,041,080	2.19	69.46	29.17
Testis	497,996,641	144,583,797	136,348,798	2.37	55.97	29.03
ES cell	126,791,595	52,476,546	35,829,866	0.67	31.15	41.38

Table A2 | The evaluation of exon identification.

Sample	RefGene exons[1]	Aligned exons (percent)	Percent of aligned length	One-to-one percent
Cerebrum	93,947	87,501(93.14%)	86.13%	93.57
Testis	99,332	93,834(94.47%)	88.66%	95.82
ES cell	82,942	78,580(94.74%)	91.34%	96.80

[1]The number of RefGene exons whose coverage is equal or greater than the cutoff value of coverage.

Table A3 | The percentage of exon reads in all mapped reads.

Tissue/cell	Total reads	Identified exon reads	Exon reads percent
Cerebrum	124,991,301	105,657,702	84.53
Testis	144,583,797	109,695,106	75.87
ES cell	524,76,546	28,838,854	54.96

Table A4 | The summary of ncRNA database records.

Sub-database	Records
ENSEMBL	8,269
UCSC	1,432
NONCODE	107,090
RNAdb	38,227
fRNAdb	510,055
Rfam	4,253
miRBase	579
tRNAdb	433
ncRNAdb	31,136

Table A5 | The evaluation of TU building based on the Fantom3 RNAs with multiple exons.

Sample	Fantom3 RNA[1]	Percent of aligned length	One-to-one percent
Cerebrum	443	71.96	95.49
Testis	605	75.51	94.88
ES cell	143	65.51	96.50

[1]The number of intergenic Fantom3 RNAs aligned with our TUs and their RPKM value of expression is larger than cutoff value ($RPKM_{cereburm} = 0.4242$, $RPKM_{testis} = 0.8199$, and $RPKM_{ES} = 0.7360$).

Table A6 | The evaluation of TU building based on the vlinc RNA.

Sample	Vlinc RNA	Matching number	Matching percent	One-to-one number	One-to-one percent
Cerebrum	486	236	48.56	156	32.10
Testis	486	215	44.24	157	32.30
ES cell	486	95	19.55	62	12.76

Table A7 | The pseudogenes in novel intergenic TUs.

Sample	Protein-coding TUs in intergenic region	Matched pseudogenes	Matched percent
Cerebrum	5,441	329	6.05
Testis	4,794	278	5.80
ES cell	2,005	295	14.71

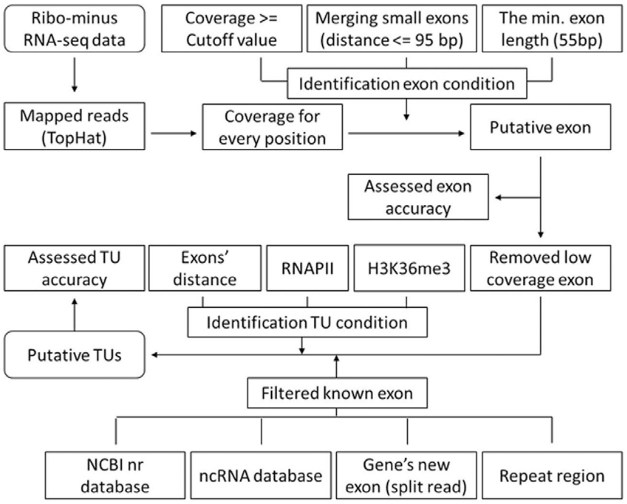


FIGURE A1 | A flowchart of gene identification process. We mapped the ribo-minus RNA-seq data using TopHat and created the coverage file for genome and identified exons according to the coverage of each position (>= cutoff value). Since 95% intron lengths are > or =95 bp, we merged small exons (distance < or =95 bp). Moreover, since 95% exon lengths are > or =55, we only keep the exons whose length is equal or larger than 55 bp to reduce false positives. We remove low coverage exons to reduce errors. We also filter known exons and build novel TUs on the basis of H3K36me3, RNAPII, and the different distance of adjacent exons between internal of TUs and adjacent TUs. We evaluate the accuracy of TU building by comparing our TUs with Fantom3 RNAs of intergenic regions.

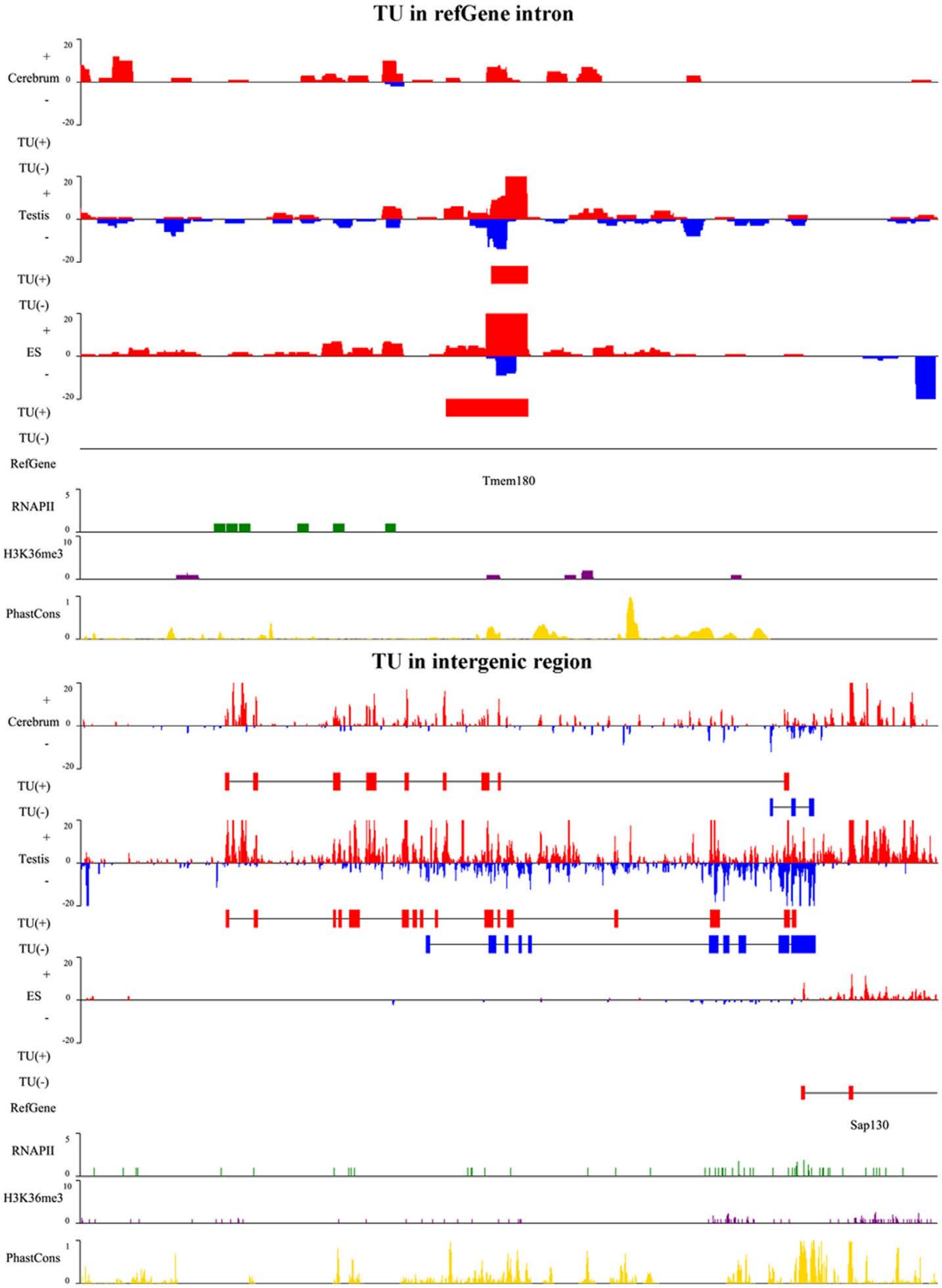


FIGURE A2 | A snapshot for TUs in Refgene intron and intergenic region. The upper panel is a TU in an intron of the transmembrane protein gene, Tmem180, and lower panel is a TU adjacent to Sap130 gene. SAP130 is a subunit of the histone deacetylase-dependent SIN3A co-repressor complex which acts as a transcriptional repressor. The TU in plus and minus strands is shown as red and blue horizontal bars, respectively. For each TU, we show RNA expression level (vertical bars in red and blue), identified TU, Refgene, RNAPII signal (green), H3K36me3 signal (purple), and conservation score (yellow).

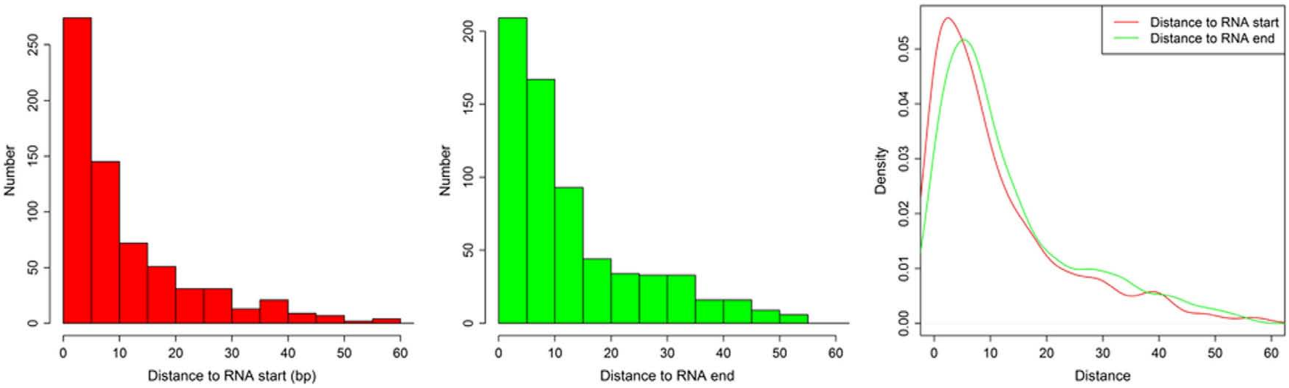


FIGURE A3 | The distances between (1) motif start and RNA 5′ end and (2) between motif end and RNA 3′ end. The histogram shows the distance between motif start and RNA 5′ end (left), the distance between motif end and RNA 3′ end (middle), and the density of both (right).

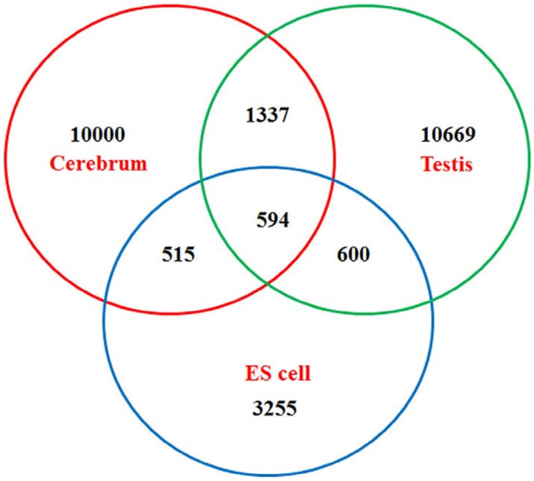


FIGURE A4 | Venn diagram of newly identified non-coding TUs among mouse the cerebrum, testis, and ES cells.

frontiers in
GENETICS

MINI REVIEW ARTICLE
published: 09 January 2012
doi: 10.3389/fgene.2011.00107

The long non-coding RNAs: a new (p)layer in the "dark matter"

Thomas Derrien[1]*, Roderic Guigó[1,2] and Rory Johnson[1]

[1] Bioinformatics and Genomics, Centre for Genomic Regulation, Universitat Pompeu Fabra, Barcelona, Spain
[2] Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, Barcelona, Spain

Edited by:
Philipp Kapranov, St. Laurent Institute, USA

Reviewed by:
Yohei Kirino, Cedars-Sinai Medical Center, USA
Chris Ponting, MRC Functional Genomics Unit, UK

***Correspondence:**
Thomas Derrien, Bioinformatics and Genomics Group, Centre for Genomic Regulation, Biomedical Research Park of Barcelona, C. Dr. Aiguader 88, Barcelona 08003, Spain.
e-mail: toma.derrien@gmail.com

The transcriptome of a cell is represented by a myriad of different RNA molecules with and without protein-coding capacities. In recent years, advances in sequencing technologies have allowed researchers to more fully appreciate the complexity of whole transcriptomes, showing that the vast majority of the genome is transcribed, producing a diverse population of non-protein coding RNAs (ncRNAs). Thus, the biological significance of non-coding RNAs (ncRNAs) have been largely underestimated. Amongst these multiple classes of ncRNAs, the long non-coding RNAs (lncRNAs) are apparently the most numerous and functionally diverse. A small but growing number of lncRNAs have been experimentally studied, and a view is emerging that these are key regulators of epigenetic gene regulation in mammalian cells. LncRNAs have already been implicated in human diseases such as cancer and neurodegeneration, highlighting the importance of this emergent field. In this article, we review the catalogs of annotated lncRNAs and the latest advances in our understanding of lncRNAs.

Keywords: non-coding RNAs, regulation, long non-coding RNA, epigenetics

THE CELL, AN RNA-DEPENDENT MACHINERY

Some of the most fundamental cellular processes rely on anciently conserved non-coding RNAs (ncRNAs). These include, for instance, the ribosomal RNAs which are assembled together to constitute ribosomes, the factories for translation of messenger RNAs (mRNAs) into proteins. Other ancient roles of ncRNAs include the transport of amino acids through ribosomes via the transfer RNAs (tRNAs) or the splicing of introns of pre-mRNA which is mediated in part by the snRNAs (small nuclear RNAs). More recently, the crucial role of ncRNA in post-transcriptional gene regulation has been highlighted by the discovery of microRNAs (miRNAs), which repress gene expression by targeting semi-complementary motifs in target mRNAs (Lee et al., 1993). Many additional classes of ncRNAs have been discovered in the last decade reinforcing the view that they are of central importance in the functioning of cells from all the branches of life (Amaral et al., 2008).

Amongst the various ncRNA classes, we know probably least about the long non-coding RNAs (lncRNAs). In particular, what is the total number of lncRNAs in mammalian genomes? Where are they localized? What is their significance in the context of evolution, and particularly in the evolution of complex processing in primate brains? Now that good catalogs of lncRNAs have become available, the most critical question is to address the functionality of these transcripts. This question is particularly acute given that we have no *a priori* methods for the prediction of lncRNA function based on sequence alone, in contrast to proteins where confident inferences on protein function can be made by simply analysis of the amino acid sequence. Given the sheer number of new unexplored lncRNA transcripts (∼15,000 at last count; Derrien et al., submitted), the field must move forward to address this question of function by using large-scale functional screens. Such moves are already underway, with groups such as Eric Lander's carrying out siRNA screens (Guttman et al., 2011). Large-scale analysis of protein-binding partners will also add another layer of valuable information to such annotation of lncRNA catalogs. Hopefully, advances in bioinformatic annotation of RNA structures (Torarinsson et al., 2006; Parker et al., 2011), and methods to predict functions based on this, will be developed. In this way, we might build up a richly annotated catalog of lncRNAs with functional predictions, that will enable us to integrate them into existing knowledge of the cell, and infer possible roles in human diseases.

Cis AND *trans* FUNCTIONS FOR lncRNAs

Until recently, only a handful of lncRNAs have been described in the literature. One of the earliest examples was XIST, a 19 kb non-protein-coding transcript which is responsible for the inactivation of one of the two X chromosome in placental females through DNA methylation (Brockdorff et al., 1992). Others examples of lncRNAs located in imprinted regions, such as Airn (Sleutels et al., 2002; Nagano et al., 2008), H19 (Gabory et al., 2009), NESPAS (Wroe et al., 2000), or Kcnq1ot1 (Mancini-Dinardo et al., 2006; Mohammad et al., 2010) are involved in the inactivation of gene expression via specific associations with chromatin-modifying complexes. More recently, the HOTAIR lncRNA was shown to epigenetically repress the HOXD locus via the recruitment of the PRC2 complex (Rinn et al., 2007). Strikingly, this study described a trans mechanism of action of a lncRNA located on human Chromosome 5 which modulates expression of multiple genes clustered on human Chromosome 4 (HOXD locus; Rinn et al., 2007). Supporting this hypothesis, two recent papers (Cabili et al., 2011; Guttman et al., 2011) showed that lncRNAs primarily affect gene expression in trans. The latter work used loss-of-function protocols to demonstrate that large intergenic ncRNAs (lincRNAs) both up- and down-regulate hundreds of genes expression in trans which support a primary role of lincRNAs in the circuitry controlling embryonic stem (ES) cell states (Guttman et al., 2011).

On the other hand, previous studies showed that some lncRNAs could also activate expression of protein-coding genes in their immediate genomic neighborhood. This cis-mechanism of action was demonstrated by Ørom and colleagues who used interference RNAs (siRNAs) to knock down candidate lncRNAs annotated as part of the GENCODE project (Harrow et al., 2006). The inactivation of some of these lncRNAs further triggers a down-regulation of protein-coding genes transcription located either in the same or opposite strand within 1 Mb from the lncRNA (Ørom et al., 2010) suggesting the latter functions as a transcriptional activator. Further supporting the cis-mechanism, a lincRNA called HOTTIP transcribed from the HOX A locus coordinates the transcription of several genes localized in cis at the 5′ of the HOXA locus (Wang et al., 2008). HOTTIP was shown to activate gene expression by recruiting the WDR5/MLL complex and thus depositing the activating histone modification H3K4me3. Finally, the distinction between activating lncRNAs and enhancers remains unclear. For instance, about 12,000 actively regulated enhancer were identified based on their bindings to the transcriptional co-activator p300/CBP in mouse neurons (Kim et al., 2010). Using ChipSeq analysis to define RNA polymerase II binding sites, the authors also reported that 25% of the enhancers co-localize with RNAPII sites suggesting that some enhancers are transcribed; they termed these transcripts eRNAs for enhancer RNAs (Kim et al., 2010). It will be important to functionally define whether such eRNAs are all required for enhancer function, or are simply a by-product of some non-functional transcription of enhancers by RNA PolII.

Similarly it will be important to define whether the activating lncRNAs (Ørom et al., 2010) are in fact a subset of eRNAs, or not.

While it is more likely that an lncRNA regulates the co-expression of nearby protein coding genes (as for tandemly duplicated genes, imprinted genes, or ubiquitously expressed genes), an interesting study demonstrate that modulating the expression of a particular locus will also trigger the modification of the expression of nearby transcripts by a mechanism known as «ripple of transcription»(Ebisuya et al., 2008). Taken together and similar to the behavior of protein-coding genes, lncRNAs seem to act both in cis and trans and are a key player of the regulation of gene expression.

LncRNAs IN HUMAN DISEASE

There is growing evidence that lncRNAs are involved in disease progression and especially cancers. For instance, recent work implies a non-coding RNA, lincRNA-p21, in the p53 response though the modulation of multiple p53 dependent gene expression in trans (Huarte et al., 2010). Another example is MEG3, which is thought to directly activate the tumor suppressor gene p53, although the mechanism has yet to be elucidated (Zhou et al., 2007). Finally, another long non-coding RNA, called ANRIL, located in the p15/CDKN2B–p16/CDKN2A–p14/ARF is genetically associated with diverse diseases such as diabetes, gliomas, coronary diseases, and basal cell carcinomas via genome-wide association studies (GWAS; Pasmant et al., 2010; Wapinski and Chang, 2011). More generally, given the lack of annotation of human lncRNAs, one could speculate on the impact of non-coding regions of the human genome in an answer to the "missing heritability" in GWAS studies (Manolio et al., 2009). Indeed, given that at least a half of the human genome is transcribed into RNA molecules (Carninci et al., 2005; ENCODE Project Consortium et al., 2007), it is now exciting to further characterize the 80% of disease-associated variants that are located outside of protein-coding genes (Manolio et al., 2009). Thus lncRNA represent a new frontier in human disease genomics. Presently no drugs against lncRNAs are available. It will be fascinating to observe whether it will be possible to specifically drug lncRNA pathways, perhaps through the use of specific modified small oligonucleotides. It is also worth mentioning that ncRNAs can be detected in human bodily fluids and hold great promise as biomarkers (Gaughwin et al., 2011).

RESOURCES FOR THE ANNOTATION OF LncRNAs

Similar to that of protein coding genes, resources for the global annotation of lncRNAs are needed in order to identify, classify and elucidate the roles of these transcripts within the cell machinery.

Particularly relevant is the effort from John Mattick's group to compile and centralize biologically meaningful information dedicated to lncRNA (Amaral et al., 2011). The lncRNA database (lncRNAdb) provides sequence, structural, and conservation evidence for mutli-species lncRNAs together with a list of lncRNAs that are experimentally known to interact with coding mRNAs.

In mouse in the early 2000s, the FANTOM consortium pioneered the genome-wide discovery of lncRNAs publishing a set of 34,030 lncRNAs based on cDNA sequencing (Maeda et al., 2006). More recently, Guttman and colleagues used chromatin signatures via ChIPSeq (Chromatin Immuno-Precipitation followed by high throughput Sequencing) to reveal ∼1,600 lincRNAs (Guttman et al., 2009). They further showed that some of these lincRNAs are functional and transcriptionally regulated by key transcription factors such Oct4 (Guttman et al., 2009). While expressed in a wide range of tissue, lincRNAs tend to be modestly conserved (Marques and Ponting, 2009) as shown by using a neutral indel model which exploits the patterns of substitutions and insertions or deletions (Lunter et al., 2006). The methodology employed by Guttman and colleagues has been applied to human thus leading to the identification of about ∼3,300 lincRNAs whose functional roles may include guidance of chromatin-modifying complexes to specific regions of the genome (Khalil et al., 2009). Very recently, the growing interest in lincRNAs led to the annotation of more than 8,000 lincRNA genes in human using a combination of computational methods and RNASeq experiments especially from the Human Body Map (HBM) project (Cabili et al., 2011; **Table 1**).

It is worth mentioning that many of the current RNASeq data (including HBM) mainly select RNA transcripts harboring a polyA tail at their 3′end (polyA+) and therefore offer little information on transcripts lacking polyA (polyA−). To tackle this issue, sequencing technologies such as single-molecule sequencing (SMS; Pushkarev et al., 2009) was used to estimate the abundance of ncRNAs by avoiding amplification and minimizing sample preparation (Kapranov et al., 2010). Interestingly, this

Table 1 | Description of human lncRNAs published catalogs.

References	Number of lncRNA elements	LncRNAs classes considered	Type of annotation	PolyA type	Experimental evidence
Khalil et al. (2009)	∼3,300	Intergenic	Bioinformatic predictions	PolyA+	(ChIPSeq) + expression array
Jia et al. (2010).	6,736	Genic + intergenic	Bioinformatic predictions + manual curation	PolyA+	Full-length cDNAs
Kapranov et al. (2010)	580	Intergenic	Bioinformatic predictions	PolyA+ PolyA−	Single-molecule sequencing (SMS) Helicos
Ørom et al. (2010)	3,019	Intergenic	Manual curation	Mainly polyA+	cDNA/ESTs + RNAseq
Cabili et al. (2011)	8,263	Intergenic	Bioinformatic predictions + manual curation	PolyA+	(ChIPSeq) + RNAseq
Derrien et al. (submitted)	9,277	Genic + intergenic	Manual curation	PolyA+ PolyA−	(ChIPSeq) + cDNA/ESTs + RNAseq + CAGE/diTAG

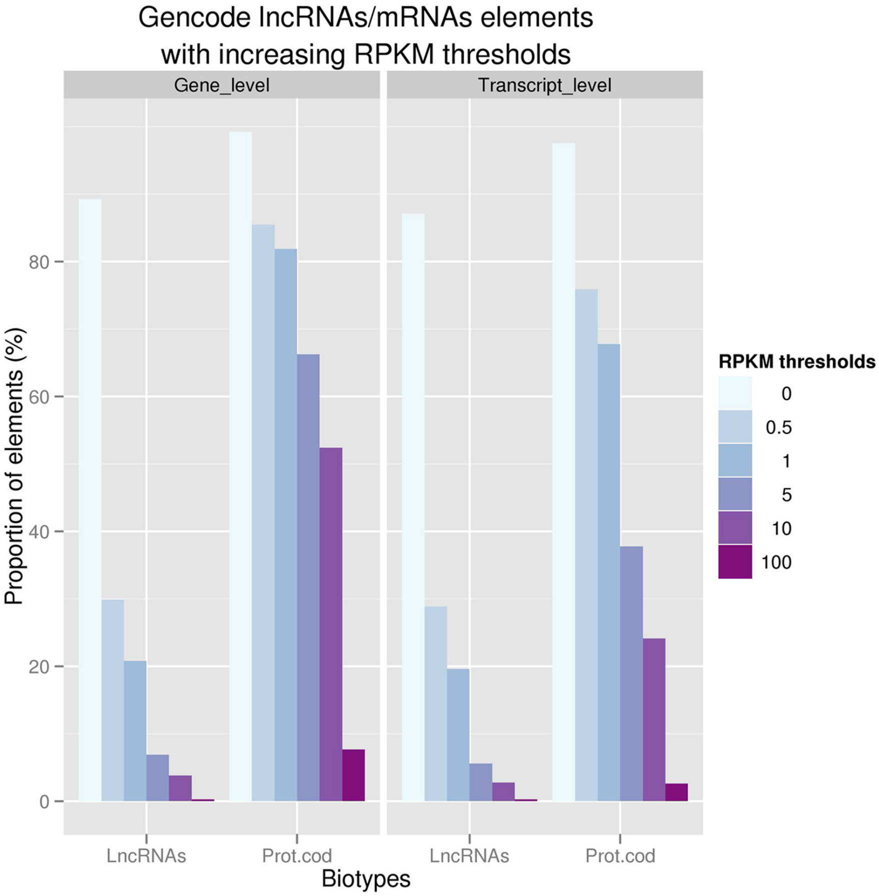


FIGURE 1 | Proportion of GENCODE polyA+ LncRNAs and protein coding at the gene ($n = 9{,}277$ and 18,063; respectively) and transcript levels with increasing thresholds of expression values (RPKM) in ENCODE RNASeq experiments.

studies revealed that "dark matter" transcription may represent the majority of the total (non-ribosomal and non-mitochondrial) RNA of a cell. In addition, it shed light on a new class of very long ncRNAs (min size ∼50 kb), abundantly expressed and localized in intergenic regions of the genome, the so-called vlincRNAs (very long intergenic ncRNAs). Focusing on the total RNA of a cell rather than the highly selected polyA+ transcripts seems to complement the latest catalog of lincRNAs (Cabili et al., 2011) since only 40% of these vlincRNAs overlap the lincRNA genes. We also recently showed that the GENCODE lncRNA set tends to have higher PolyA− representation compared to protein-coding mRNAs (Derrien et al., submitted). Although many studies have concentrated on the intergenic lncRNAs (the lincRNAs), this seriously underestimates the true number of lncRNA transcripts in the genome. Approximately one third (Derrien et al., submitted) to one half (Jia et al., 2010) of lncRNAs overlap protein-coding loci in some way – "genic" lncRNAs. It seems therefore essential to annotate lncRNAs both in intergenic and coding regions since (i) the exact boundaries of protein-coding genes is frequently subject to variations and reannotations (Denoeud et al., 2007; Gingeras, 2007) and thus could lead to the revision of a lincRNAs into a *bona-fide* lncRNAs, (ii) thousands of protein-coding genes harbor natural antisense transcripts belonging to the lncRNAs class (He et al., 2008; iii) numerous functional genic lncRNAs overlapping protein-coding genes have been experimentally validated, especially in disease states (Faghihi et al., 2008; Pasmant et al., 2011; Wapinski and Chang, 2011). A recent catalog of both genic and intergenic lncRNAs has been released based on genome-wide computational approach combined with intensive manual annotation. This led to the identification oh 6,736 lncRNA genes in human (Jia et al., 2010) among which 63% are localized within or in a close proximity (<10 kb) of known protein coding genes (Jia et al., 2010).

THE GENCODE CATALOG OF HUMAN lncRNAs

Most recently, the GENCODE annotation group has produced the most comprehensive, high-quality human lncRNA annotation to date. In order to identify all evidence-based functional gene features in the human genome, the GENCODE group (Harrow et al., 2006) within the ENCODE framework (ENCyclopedia Of DNA Elements; ENCODE Project Consortium et al., 2007) provides a high-quality collection of lncRNAs. GENCODE annotation involves manual curation, multiple computational analysis, and targeted experimental approaches, all together representing complementary methodologies for the complete identification of all human functional elements (coding and non-coding genes). At present, the GENCODE collection (Version 7) comprises 14,880 lncRNA transcripts arising from 9,277 distinct gene loci (Derrien et al., submitted).

In a recent study, we investigated whether these lncRNAs are under negative evolutionary selection, indicative of functionality (Derrien et al., submitted). Evolutionary scores were computed based both on the phastCons program (Siepel et al., 2005) and custom BLAST alignments within mammals in order to measure the conservation profiles of GENCODE lncRNAs in comparison with protein-coding transcripts and ancestral repeats (ARs), the latter representing a good proxy for measuring neutrally evolving sequences (Ponjavic et al., 2007). Overall, lncRNAs show moderate sequence conservation compared to coding transcripts. This lower sequence conservation may reflect the fact that functional RNA structures are more robust in the face of sequence mutations and insertions–deletions (indels), compared to the higher constraints inherent of protein-coding open reading frames. Nevertheless, lncRNAs and more especially, their promoters, showed statistically significant, non-random conservation, strongly suggesting a functional role for these ncRNAs. Interestingly, about one third of the 15,000 lncRNAs display a primate-specific pattern of conservation (Derrien et al., submitted).

Using whole transcriptome sequencing (RNAseq) of 16 human cell lines produced in the framework of the ENCODE consortium (ENCODE Project Consortium et al., 2007) and 16 tissues from the Human Body Map project (www.illumina.com), we showed that 94% of the GENCODE lncRNAs transcripts are expressed in at least one of these tissue/cell line studied. Strikingly, the level of expression of polyA+ lncRNAs is ~10–20 times lower than protein-coding transcripts reinforcing the need to use deep sequencing based technologies to identify these low expressed non-coding loci (**Figure 1**.). We also demonstrated that lncRNAs tend to be enriched in nucleus in comparison with mRNAs; this latter observation being consistent with the idea that many lncRNAs may be devoted to gene regulation in the nucleus. Finally, the question is raised as to whether lincRNAs could encode very small peptides as shown by Ingolia et al. (2011). However, there is still conflicting evidence about this hypothesis since a recent study which used comprehensive mass spectrometry data (MS) produced as part of the ENCODE project only found about a hundred of GENCODE lncRNA to be matched by small peptides (Banfai et al., submitted).

CONCLUSION

Over the past decade, the estimation of the proportion of "functional DNA" in the human genome has been constantly revised upward (Ponting and Hardison, 2011).

We now know that the human genome contains thousands of lncRNAs, both genic and intergenic. This new class of non-protein coding RNAs (ncRNAs) lack functional ORFs, are modestly conserved and seem to negatively and positively regulate protein coding gene expression, in cis and trans. Diverse mechanisms of action have been observed (see for reviews Ponting et al., 2009; Nagano and Fraser, 2011) suggesting that lncRNAs are a fundamental regulators of transcription. The classification of lncRNAs remains difficult, and we presently have only a vague idea of what sub-categories exist, and how we might use experimental or sequence information to distinguish between such categories. With the ongoing and increasing number of RNAseq experiments characterizing transcriptomes of multiples cell lines and human tissues (in particular within the ENCODE consortium), it is likely that the number of annotated lncRNAs will increase dramatically in the near future. Future studies will likely focus on identifying functional lncRNAs, and those involved in human disease processes.

ACKNOWLEDGMENTS

We would like to thank reviewers for the helpful comments.

REFERENCES

Amaral, P. P., Dinger, M. E., Mercer, T. R., and Mattick, J. S. (2008). The Eukaryotic genome as an RNA machine. *Science* 319, 1787–1789.

Amaral, P. P., Michael, B. C., Dennis, K. G., Marcel, E. D., and John, S. M. (2011). lncRNAdb: a reference database for long noncoding RNAs. *Nucleic Acids Res.* 39, D146–D151.

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.

Cabili, M. N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., and Rinn, J. L. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.* doi:10.1101/gad.17446611

Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M. C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C., Kodzius, R., Shimokawa, K., Bajic, V. B., Brenner, S. E., Batalov, S., Forrest, A. R., Zavolan, M., Davis, M. J., Wilming, L. G., Aidinis, V., Allen, J. E., Ambesi-Impiombato, A., Apweiler, R., Aturaliya, R. N., Bailey, T. L., Bansal, M., Baxter, L., Beisel, K. W., Bersano, T., Bono, H., Chalk, A. M., Chiu, K. P., Choudhary, V., Christoffels, A., Clutterbuck, D. R., Crowe, M. L., Dalla, E., Dalrymple, B. P., de Bono, B., Della Gatta, G., di Bernardo, D., Down, T., Engstrom, P., Fagiolini, M., Faulkner, G., Fletcher, C. F., Fukushima, T., Furuno, M., Futaki, S., Gariboldi, M., Georgii-Hemming, P., Gingeras, T. R., Gojobori, T., Green, R. E., Gustincich, S., Harbers, M., Hayashi, Y., Hensch, T. K., Hirokawa, N., Hill, D., Huminiecki, L., Iacono, M., Ikeo, K., Iwama, A., Ishikawa, T., Jakt, M., Kanapin, A., Katoh, M., Kawasawa, Y., Kelso, J., Kitamura, H., Kitano, H., Kollias, G., Krishnan, S. P., Kruger, A., Kummerfeld, S. K., Kurochkin, I. V., Lareau, L. F., Lazarevic, D., Lipovich, L., Liu, J., Liuni, S., McWilliam, S., Madan Babu, M., Madera, M., Marchionni, L., Matsuda, H., Matsuzawa, S., Miki, H., Mignone, F., Miyake, S., Morris, K., Mottagui-Tabar, S., Mulder, N., Nakano, N., Nakauchi, H., Ng, P., Nilsson, R., Nishiguchi, S., Nishikawa, S., Nori, F., Ohara, O., Okazaki, Y., Orlando, V., Pang, K. C., Pavan, W. J., Pavesi, G., Pesole, G., Petrovsky, N., Piazza, S., Reed, J., Reid, J. F., Ring, B. Z., Ringwald, M., Rost, B., Ruan, Y., Salzberg, S. L., Sandelin, A., Schneider, C., Schönbach, C., Sekiguchi, K., Semple, C. A., Seno, S., Sessa, L., Sheng, Y., Shibata, Y., Shimada, H., Shimada, K., Silva, D., Sinclair, B., Sperling, S., Stupka, E., Sugiura, K., Sultana, R., Takenaka, Y., Taki, K., Tammoja, K., Tan, S. L., Tang, S., Taylor, M. S., Tegner, J., Teichmann, S. A., Ueda, H. R., van Nimwegen, E., Verardo, R., Wei, C. L., Yagi, K., Yamanishi, H., Zabarovsky, E., Zhu, S., Zimmer, A., Hide, W., Bult, C., Grimmond, S. M., Teasdale, R. D., Liu, E. T., Brusic, V., Quackenbush, J., Wahlestedt, C., Mattick, J. S., Hume, D. A., Kai, C., Sasaki, D., Tomaru, Y., Fukuda, S., Kanamori-Katayama, M., Suzuki, M., Aoki, J., Arakawa, T., Iida, J., Imamura, K., Itoh, M., Kato, T., Kawaji, H., Kawagashira, N., Kawashima, T., Kojima, M., Kondo, S., Konno, H., Nakano, K., Ninomiya, N., Nishio, T., Okada, M., Plessy, C., Shibata, K., Shiraki, T., Suzuki, S., Tagami, M., Waki, K., Watahiki, A., Okamura-Oho, Y., Suzuki, H., Kawai, J., Hayashizaki, Y., FANTOM Consortium, RIKEN Genome Exploration Research Group and Genome Science Group (Genome Network Project Core Group). (2005). The transcriptional landscape of the mammalian genome. *Science* 309, 1559–1563.

Denoeud, F., Kapranov, P., Ucla, C., Frankish, A., Castelo, R., Drenkow, J., Lagarde, J., Alioto, T., Manzano, C., Chrast, J., Dike, S., Wyss, C., Henrichsen, C. N., Holroyd, N., Dickson, M. C., Taylor, R., Hance, Z., Foissac, S., Myers, R. M., Rogers, J., Hubbard, T., Harrow, J., Guigó, R., Gingeras, T. R., Antonarakis, S. E., and Reymond, A. (2007). Prominent use of distal 5′ transcription start sites and discovery of a large number of additional exons in ENCODE regions. *Genome Res.* 17, 746–759.

Ebisuya, M., Yamamoto, T., Nakajima, M., and Nishida, E. (2008). Ripples from neighbouring transcription. *Nat. Cell Biol.* 10, 1106–1113.

ENCODE Project Consortium, Birney, E., Stamatoyannopoulos, J. A., Dutta, R., Guigo, A., Gingeras, T. R., Margulies, E. H., Weng, Z., Snyder, M., Dermitzakis, E. T., Thurman, R. E., Kuehn, M. S., Taylor, C. M., Neph, S., Koch, C. M., Asthana, S., Malhotra, A., Adzhubei, I., Greenbaum, J. A., Andrews, R. M., Flicek, P., Boyle, P. J., Cao, H., Carter, N. P., Clelland, G. K., Davis, S., Day, N., Dhami, P., Dillon, S. C., Dorschner, M. O., Fiegler, H., Giresi, P. G., Goldy, J., Hawrylycz, M., Haydock, A., Humbert, R., James, K. D., Johnson, B. E., Johnson, E. M., Frum, T. T., Rosenzweig, E. R., Karnani, N., Lee, K., Lefebvre, G. C., Navas, P. A., Neri, F., Parker, S. C., Sabo, P. J., Sandstrom, R., Shafer, A., Vetrie, D., Weaver, M., Wilcox, S., Yu, M., Collins, F. S., Dekker, J., Lieb, J. D., Tullius, T. D., Crawford, G. E., Sunyaev, S., Noble, W. S., Dunham, I., Denoeud, F., Reymond, A., Kapranov, P., Rozowsky, J., Zheng, D., Castelo, R., Frankish, A., Harrow, J., Ghosh, S., Sandelin, A., Hofacker, I. L., Baertsch, R., Keefe, D., Dike, S., Cheng, J., Hirsch, H. A., Sekinger, E. A., Lagarde, J., Abril, J. F., Shahab, A., Flamm, C., Fried, C., Hackermüller, J., Hertel, J., Lindemeyer, M., Missal, K., Tanzer, A., Washietl, S., Korbel, J., Emanuelsson, O., Pedersen, J. S., Holroyd, N., Taylor, R., Swarbreck, D., Matthews, N., Dickson, M. C., Thomas, D. J., Weirauch, M. T., Gilbert, J., Drenkow, J., Bell, I., Zhao, X., Srinivasan, K. G., Sung, W. K., Ooi, H. S., Chiu, K. P., Foissac, S., Alioto, T., Brent, M., Pachter, L., Tress, M. L., Valencia, A., Choo, S. W., Choo, C. Y., Ucla, C., Manzano, C., Wyss, C., Cheung, E., Clark, T. G., Brown, J. B., Ganesh, M., Patel, S., Tammana, H., Chrast, J., Henrichsen, C. N., Kai, C., Kawai, J., Nagalakshmi, U., Wu, J., Lian, Z., Lian, J., Newburger, P., Zhang, X., Bickel, P., Mattick, J. S., Carninci, P., Hayashizaki, Y., Weissman, S., Hubbard, T., Myers, R. M., Rogers, J., Stadler, P. F., Lowe, T. M., Wei, C. L., Ruan, Y., Struhl, K., Gerstein, M., Antonarakis, S. E., Fu, Y., Green, E. D., Karaöz, U., Siepel, A., Taylor, J., Liefer, L. A., Wetterstrand, K. A., Good, P. J., Feingold, E. A., Guyer, M. S., Cooper, G. M., Asimenos, G., Dewey, C. N., Hou, M., Nikolaev, S., Montoya-Burgos, J. I., Löytynoja, A., Whelan, S., Pardi, F., Massingham, T., Huang, H., Zhang, N. R., Holmes, I., Mullikin, J. C., Ureta-Vidal, A., Paten, B., Seringhaus, M., Church, D., Rosenbloom, K., Kent, W. J., Stone, E. A., NISC Comparative Sequencing Program, Baylor College of Medicine Human Genome Sequencing Center, Washington University Genome Sequencing Center, Broad Institute, Children's Hospital Oakland Research Institute, Batzoglou, S., Goldman, N., Hardison, R. C., Haussler, D., Miller, W., Sidow, A., Trinklein, N. D., Zhang, Z. D., Barrera, L., Stuart, R., King, D. C., Ameur, A., Enroth, S., Bieda, M. C., Kim, J., Bhinge, A. A., Jiang, N., Liu, J., Yao, F., Vega, V. B., Lee, C. W., Ng, P., Shahab, A., Yang, A., Moqtaderi, Z., Zhu, Z., Xu, X., Squazzo, S., Oberley, M. J., Inman, D., Singer, M. A., Richmond, T. A., Munn, K. J., Rada-Iglesias, A., Wallerman, O., Komorowski, J., Fowler, J. C., Couttet, P., Bruce, A. W., Dovey, O. M., Ellis, P. D., Langford, C. F., Nix, D. A., Euskirchen, G., Hartman, S., Urban, A. E., Kraus, P., Van Calcar, S., Heintzman, N., Kim, T. H., Wang, K., Qu, C., Hon, G., Luna, R., Glass, C. K., Rosenfeld, M. G., Aldred, S. F., Cooper, S. J., Halees, A., Lin, J. M., Shulha, H. P., Zhang, X., Xu, M., Haidar, J. N., Yu, Y., Ruan, Y., Iyer, V. R., Green, R. D., Wadelius, C., Farnham, P. J., Ren, B., Harte, R. A., Hinrichs, A. S., Trumbower, H., Clawson, H., Hillman-Jackson, J., Zweig, A. S., Smith, K., Thakkapallayil, A., Barber, G., Kuhn, R. M., Karolchik, D., Armengol, L., Bird, C. P., de Bakker, P. I., Kern, A. D., Lopez-Bigas, N., Martin, J. D., Stranger, B. E., Woodroffe, A., Davydov, E., Dimas, A., Eyras, E., Hallgrímsdóttir, I. B., Huppert, J., Zody, M. C., Abecasis, G. R., Estivill, X., Bouffard, G. G., Guan, X., Hansen, N. F., Idol, J. R., Maduro, V. V., Maskeri, B., McDowell, J. C., Park, M., Thomas, P. J., Young, A. C., Blakesley, R. W., Muzny, D. M., Sodergren, E., Wheeler, D. A., Worley, K. C., Jiang, H., Weinstock, G. M., Gibbs, R. A., Graves, T., Fulton, R., Mardis, E. R., Wilson, R. K., Clamp, M., Cuff, J., Gnerre, S., Jaffe, D. B., Chang, J. L., Lindblad-Toh, K., Lander, E. S., Koriabine, M., Nefedov, M., Osoegawa, K., Yoshinaga, Y., Zhu, B., and de Jong, P. J. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447, 799–816.

Faghihi, M. A., Modarresi, F., Khalil, A. M., Wood, D. E., Sahagan, B. G., Morgan, T. E., Finch, C. E., St Laurent, G. III., Kenny, P. J., and Wahlestedt, C. (2008). Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of beta-secretase. *Nat. Med.* 14, 723–730.

Gabory, A., Ripoche, M.-A., Le Digarcher, A., Watrin, F., Ziyyat, A., Forné, T., Jammes, H., Ainscough, J. F., Surani, M. A., Journot, L., and Dandolo, L. (2009). H19 acts as a trans regulator of the imprinted gene network controlling growth in mice. *Development* 136, 3413–3421.

Gaughwin, P. M., Ciesla, M., Lahiri, N., Tabrizi, S. J., Brundin, P., and Björkqvist, M. (2011). Hsa-miR-34b is a plasma-stable microRNA that is elevated in pre-manifest Huntington's disease. *Hum. Mol. Genet.* 20, 2225–2237.

Gingeras, T. R. (2007). Origin of phenotypes: genes and transcripts. *Genome Res.* 17, 682–690.

Guttman, M., Amit, I., Garber, M., French, C., Lin, M. F., Feldser, D., Huarte, M., Zuk, O., Carey, B. W., Cassady, J. P., Cabili, M. N., Jaenisch, R., Mikkelsen, T. S., Jacks, T., Hacohen, N., Bernstein, B. E., Kellis, M., Regev, A., Rinn, J. L., and Lander, E. S. (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* 458, 223–227.

Guttman, M., Donaghey, J., Carey, B. W., Garber, M., Grenier, J. K., Munson, G., Young, G., Lucas, A. B., Ach, R., Bruhn, L., Yang, X., Amit, I., Meissner, A., Regev, A., Rinn, J. L., Root, D. E., and Lander, E. S. (2011). lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature* 1–11.

Harrow, J., Denoeud, F., Frankish, A., Reymond, A., Chen, C. K., Chrast, J., Lagarde, J., Gilbert, J. G., Storey, R., Swarbreck, D., Rossier, C., Ucla, C., Hubbard, T., Antonarakis, S. E., and Guigo, R. (2006). GENCODE: producing a reference annotation for ENCODE. *Genome Biol.* 7(Suppl. 1), S4.1–S9.

He, Y., Vogelstein, B., Velculescu, V. E., Papadopoulos, N., and Kinzler, K. W. (2008). The antisense transcriptomes of human cells. *Science* 322, 1855–1857.

Huarte, M., Guttman, M., Feldser, D., Garber, M., Koziol, M. J., Kenzelmann-Broz, D., Khalil, A. M., Zuk, O., Amit, I., Rabani, M., Attardi, L. D., Regev, A., Lander, E. S., Jacks, T., and Rinn, J. L. (2010). A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. *Cell* 142, 409–419.

Ingolia, N. T., Lareau, L. F., and Weissman, J. S. (2011). Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* 147, 789–802.

Jia, H., Osak, M., Bogu, G. K., Stanton, L. W., Johnson, R., and Lipovich, L. (2010). Genome-wide computational identification and manual annotation of human long noncoding RNA genes. *RNA* 16, 1478–1487.

Kapranov, P., St. Laurent, G., Raz, T., Ozsolak, F., Reynolds, C. P., Sorensen, P. H., Reaman, G., Milos, P., Arceci, R. J., Thompson, J. F., and Triche, T. J. (2010). The majority of total nuclear-encoded non-ribosomal RNA in a human cell is 'dark matter' unannotated RNA. *BMC Biol.* 8, 149. doi:10.1186/1741-7007-8-149

Khalil, A. M., Guttman, M., Huarte, M., Garber, M., Raj, A., Rivea Morales, D., Thomas, K., Presser, A., Bernstein, B. E., van Oudenaarden, A., Regev, A., Lander, E. S., and Rinn, J. L. (2009). Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc. Natl. Acad. Sci.* 106, 11667–11672.

Kim, T. K., Hemberg, M., Gray, J. M., Costa, A. M., Bear, D. M., Wu, J., Harmin, D. A., Laptewicz, M., Barbara-Haley, K., Kuersten, S., Markenscoff-Papadimitriou, E., Kuhl, D., Bito, H., Worley, P. F., Kreiman, G., and Greenberg, M. E. (2010). Widespread transcription at neuronal activity-regulated enhancers. *Nature* 465, 182–187.

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.

Lunter, G., Ponting, C. P., and Hein, J. (2006). Genome-wide identification of human functional DNA using a neutral indel model. *PLoS Comput. Biol.* 2, e5. doi:10.1371/journal.pcbi.0020005

Maeda, N., Kasukawa, T., Oyama, R., Gough, J., Frith, M., Engström, P. G., Lenhard, B., Aturaliya, R. N., Batalov, S., Beisel, K. W., Bult, C. J., Fletcher, C. F., Forrest, A. R., Furuno, M., Hill, D., Itoh, M., Kanamori-Katayama, M., Katayama, S., Katoh, M., Kawashima, T., Quackenbush, J., Ravasi, T., Ring, B. Z., Shibata, K., Sugiura, K., Takenaka, Y., Teasdale, R. D., Wells, C. A., Zhu, Y., Kai, C., Kawai, J., Hume, D. A., Carninci, P., and Hayashizaki, Y. (2006). Transcript annotation in FANTOM3: mouse gene catalog based on physical cDNAs. *PLoS Genet.* 2, e62. doi:10.1371/journal.pgen.0020062

Mancini-Dinardo, D., Steele, S. J., Levorse, J. M., Ingram, R. S., and Tilghman, S. M. (2006). Elongation of the Kcnq1ot1 transcript is required for genomic imprinting of neighboring genes. *Genes Dev.* 20, 1268–1282.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., Boehnke, M., Clark, A. G., Eichler, E. E., Gibson, G., Haines, J. L., Mackay, T. F., McCarroll, S. A., and Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature* 461, 747–753.

Marques, A. C., and Ponting, C. P. (2009). Catalogues of mammalian long noncoding RNAs: modest conservation and incompleteness. *Genome Biol.* 10, R124.

Mohammad, F., Mondal, T., Guseva, N., Pandey, G. K., and Kanduri, C. (2010). Kcnq1ot1 noncoding RNA mediates transcriptional gene silencing by interacting with Dnmt1. *Development* 137, 2493–2499.

Nagano, T., and Fraser, P. (2011). No-nonsense functions for long noncoding RNAs. *Cell* 145, 178–181.

Nagano, T., Mitchell, J. A., Sanz, L. A., Pauler, F. M., Ferguson-Smith, A. C., Feil, R., and Fraser, P. (2008). The air noncoding RNA epigenetically silences transcription by targeting G9a to chromatin. *Science* 322, 1717–1720.

Ørom, U. A., Derrien, T., Beringer, M., Gumireddy, K., Gardini, A., Bussotti, G., Lai, F., Zytnicki, M., Notredame, C., Huang, Q., Guigo, R., and Shiekhattar, R. (2010). Long noncoding RNAs with enhancer-like function in human cells. *Cell* 143, 46–58.

Parker, B. J., Moltke, I., Roth, A., Washietl, S., Wen, J., Kellis, M., Breaker, R., and Pedersen, J. S. (2011). New families of human regulatory RNA structures identified by comparative analysis of vertebrate genomes. *Genome Res.* 21, 1929–1943.

Pasmant, E., Laurendeau, I., Sabbagh, A., Parfait, B., Vidaud, M., Vidaud, D., and Bièche, I. (2010). The amazing story of ANRIL, a long noncoding RNA. *Med. Sci. (Paris)* 26, 564–566.

Pasmant, E., Sabbagh, A., Vidaud, M., and Bièche, I. (2011). ANRIL, a long, noncoding RNA, is an unexpected major hotspot in GWAS. *FASEB J.* 25, 444–448.

Ponjavic, J., Ponting, C. P., and Lunter, G. (2007). Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Res.* 17, 556–565.

Ponting, C. P., and Hardison, R. C. (2011). What fraction of the human genome is functional? *Genome Res.* 21, 1769–1776.

Ponting, C. P., Oliver, P. L., and Reik, W. (2009). Evolution and functions of long noncoding RNAs. *Cell* 136, 629–641.

Pushkarev, D., Neff, N. F., and Quake, S. R. (2009). Single-molecule sequencing of an individual human genome. *Nat. Biotechnol.* 27, 847–850.

Rinn, J. L., Kertesz, M., Wang, J. K., Squazzo, S. L., Xu, X., Brugmann, S. A., Goodnough, L. H., Helms, J. A., Farnham, P. J., Segal, E., and Chang, H. Y. (2007). Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* 129, 1311–1323.

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L. W., Richards, S., Weinstock, G. M., Wilson, R. K., Gibbs, R. A., Kent, W. J., Miller, W., and Haussler, D. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* 15, 1034–1050.

Sleutels, F., Zwart, R., and Barlow, D. P. (2002). The non-coding Air RNA is required for silencing autosomal imprinted genes. *Nature* 415, 810–813.

Torarinsson, E., Sawera, M., Havgaard, J. H., Fredholm, M., and Gorodkin, J. (2006). Thousands of corresponding human and mouse genomic regions unalignable in primary sequence contain common RNA structure. *Genome Res.* 16, 885–889.

Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S. F., Schroth, G. P., and Burge, C. B. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature* 456, 470–476.

Wapinski, O., and Chang, H. Y. (2011). Long noncoding RNAs and human disease. *Trends Cell Biol.* 21, 354–361.

Wroe, S. F., Kelsey, G., Skinner, J. A., Bodle, D., Ball, S. T., Beechey, C. V., Peters, J., and Williamson, C. M. (2000). An imprinted transcript, antisense to Nesp, adds complexity to the cluster of imprinted genes at the mouse Gnas locus. *Proc. Natl. Acad. Sci. U.S.A.* 97, 3342–3346.

Zhou, Y., Zhong, Y., Wang, Y., Zhang, X., Batista, D. L., Gejman, R., Ansell, P. J., Zhao, J., Weng, C., and Klibanski, A. (2007). Activation of p53 by MEG3 non-coding RNA. *J. Biol. Chem.* 282, 24731–24742.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 06 September 2011; accepted: 21 December 2011; published online: 09 January 2012.

Citation: Derrien T, Guigó R and Johnson R (2012) The long non-coding RNAs: a new (p)layer in the "dark matter". Front. Gene. **2***:107. doi: 10.3389/fgene.2011.00107*

This article was submitted to Frontiers in Non-Coding RNA, a specialty of Frontiers in Genetics.

frontiers in GENETICS

PERSPECTIVE ARTICLE
published: 26 December 2011
doi: 10.3389/fgene.2011.00092

Trans-splicing in higher eukaryotes: implications for cancer development?

Peter G. Zaphiropoulos*

Department of Biosciences and Nutrition, Karolinska Institutet, Huddinge, Sweden

Edited by:
Philipp Kapranov, St. Laurent Institute, USA

Reviewed by:
Zhaohui Wang, University of Texas Southwestern Medical Center, USA
Woan-Yuh Tarn, Academia Sinica, Taiwan

****Correspondence:***
Peter G. Zaphiropoulos, Department of Biosciences and Nutrition, Karolinska Institutet, 141 83 Huddinge, Sweden.
e-mail: pezaph@ki.se

Trans-splicing, the possibility of exons from distinct pre-mRNAs to join together, is still a concept in gene expression that is generally regarded of limited significance. However, recent work has provided evidence that in human tumors *trans*-splicing events may precede chromosomal rearrangements. In fact, it has been suggested that the *trans*-spliced molecules could act as "guides" that facilitate the genomic translocation. This perspective highlights the development of the ideas of *trans*-splicing in higher eukaryotes during the last 25 years, from a bizarre phenomenon to a biological event that is attaining stronger recognition.

Keywords: RNA, exon, evolution, gene, recombination

DEFINITION OF TERMS

Splicing: the joining of exons from primary RNAs.

Cis-splicing: the joining of exons from a primary RNA in a 5′ to 3′ order.

Exon scrambling: the joining of exons from a primary RNA in an order where 3′ exons are positioned upstream of 5′ exons.

Trans-splicing: the joining of exons from more than one primary RNA.

Exon repetition: the presence of repeats of exon(s) in an RNA.

Exon: a sequence that is retained in a processed RNA, after removal of the intervening sequences. Exons are flanked by the major GT/AG (CG) or the minor AT/AC dinucleotides (internal exon) or a single dinucleotide and the start/end of the processed RNA (terminal exon).

Spliced leader (SL) RNA: a short RNA sequence that is *trans*-spliced to many gene transcripts in certain lower organisms, including trypanosomes and nematodes.

JOINING OF EXONS FROM DISTINCT PRE-mRNAs – EVOLUTION OF THE CONCEPT OF *TRANS*-SPLICING

The earliest reports on splicing reactions between two different RNA substrates date from 1985 (Konarska et al., 1985; Solnick, 1985). In these pioneering *in vitro* experiments the efficiency of *trans*-splicing was found to be enhanced by sequence complementarity in the intronic regions of the two mRNA precursor molecules. This was followed by evidence that in the trypanosome *Trypanosoma brucei*, and the nematode *Caenorhabditis elegans* a single RNA sequence, the SL, is *trans*-spliced to many RNAs (Murphy et al., 1986; Sutton and Boothroyd, 1986; Krause and Hirsh, 1987). A few years later, the possibility that mammalian cells may actually be involved in RNA processes that include *trans*-splicing was elegantly demonstrated by Bruzik and Maniatis (1992), when the SL RNA of *C. elegans* was shown to be capable to *trans*-splice to the adenovirus exon 2 in COS cells *in vivo*. However, this proposal was met with a lot of skepticism. In fact, it has been suggested that even if mammalian cells have this capacity, such phenomena are not really occurring (Blumenthal, 1993). Certainly, the SL type of *trans*-splicing is not apparently taking place in higher eukaryotes. On the other hand, reports that eukaryotic exons may be joined in an order that deviates from their linear arrangement in the genome have started to accumulate since the early nineties, challenging the universality of *cis*-splicing. One early observation was that the order of exons in spliced RNAs could be reversed compared to that present in genomic DNA (Nigro et al., 1991; Cocquerelle et al., 1992). These "scrambled" RNAs were found at levels significantly lower compared to the corresponding "canonical" mRNAs, were mostly cytoplasmic and appeared to lack a polyA+ tail.

Moreover, additional reports highlighted the presence of abundant polyA+ mRNAs containing repetitions of certain exons, a phenomenon that can be rationalized by a *trans*-splicing process of independent pre-mRNA molecules (Caudevilla et al., 1998; Frantz et al., 1999). Furthermore, polyA+ mRNAs generated from gene loci present on opposite strands of a chromosome have also been reported, although in some cases, the expression level of such *trans*-spliced mRNAs was found to be quite low (Dorn et al., 2001; Labrador et al., 2001; Finta and Zaphiropoulos, 2002). Additionally, *trans*-splicing was suggested to have a role in the process of interallelic complementation in *Drosophila*, as this type of splicing was shown to also occur between different alleles (Horiuchi et al., 2003).

An elegant computational strategy was employed to detect *trans*-splicing events using non-linear exon splice junction probes on expressed sequences from the GenBank. This approach revealed 178 human genes that engage in splicing processes resulting in a change of the canonical 5′–3′ exon order (Dixon et al., 2005). Further analysis suggested that complementarity of intronic sequences has a role in promoting this non-linear splicing (Dixon et al., 2007). More recently, *trans*-splicing events that are mediated through sequence complementarity of independent transcripts have also been observed in *C. elegans* and the unicellular eukaryote *Giardia intestinalis*, which has only few *cis*-spliced introns (Fischer et al., 2008; Kamikawa et al., 2011).

A WORD OF CAUTION

As with any novel concept the suggestion that a directional *cis*-splicing process may not account for all spliced RNAs observed in higher eukaryotic species should be subjected to rigorous quality controls. This is especially important since the methods used to detect *trans*-splicing do so in an indirect way. In a typical assay, the RNA is subjected to reverse transcription, followed by PCR amplification. However, both polymerases can engage into template switching, resulting in an artifactual generation of hybrid molecules, with this event enhanced, but not fully dependent, by repeat sequences (Zaphiropoulos, 1998; Houseley and Tollervey, 2010). This may explain a number of reports, which claimed the widespread abundance of *trans*-splicing processes, with the canonical splice signals, GT and AG not being retained (Li et al., 2009). Even RNA protection analysis, which has been used to confirm the presence of *trans*-spliced RNAs, as an independent method that avoids the involvement of reverse transcription and PCR amplification, is not full proof, unless appropriate controls are employed (Kralovicova and Vorechovsky, 2005).

One way to strengthen the argument for the finding of *bona fide trans*-spliced RNAs and not *in vitro* recombinants generated during reverse transcription/PCR, apart from the presence of canonical splice signals, is to look for premature termination codons (PTCs). Absence of PTCs in the putative *trans*-spliced molecule, and consequently maintenance of a long open reading frame, would be in line with the quality control mechanism of nonsense mediated decay (Durand and Lykke-Andersen, 2011) that *trans*-spliced mRNAs, similarly to *cis*-spliced mRNAs, are subjected to.

TRANS-SPLICING AS A TEMPLATE FOR GENE TRANSLOCATION IN ENDOMETRIAL CANCER

An unanticipated proposal on the role of *trans*-splicing in cancer biology was put forward in 2008. Namely, normal endometrial stromal cells were shown to produce a *trans*-spliced RNA, which joins the first three exons of the *JAZF1* gene on chromosome 7 to the last 15 exons of the *JJAZ1* gene on chromosome 17, that is identical to the hybrid RNA produced by the (7;17; p15;q21) chromosomal rearrangement found in endometrial stromal tumors (Li et al., 2008). This unprecedented finding raised the possibility that endometrial stromal cells capable of this *trans*-splicing may be predisposed for the genomic translocation that characterizes endometrial stromal tumors. Additionally, a possible mechanism for this translocation could be that the *trans*-spliced RNA acts as a template that facilitates the genomic fusion (**Figure 1**), a process that is in line with the RNA mediated genome rearrangement events described in ciliates (Nowacki et al., 2008). One may hypothesize that the *trans*-spliced RNA intercalates through the "breathing" DNA duplexes of the two chromosomes bringing them in proximity, and this RNA–DNA base pairing promotes strand breaks and chromosome translocation. Alternatively, the capacity of transcripts from different chromosomes to *trans*-splice may be the result of chromosomal interactions that could be mediated, at least in part, through sequence complementarity, and this proximity of the gene loci may also enhance genomic rearrangements. As a way to discriminate between these alternatives, determining the impact of exogenously added, hybrid RNA in promoting chromosomal translocations would be helpful.

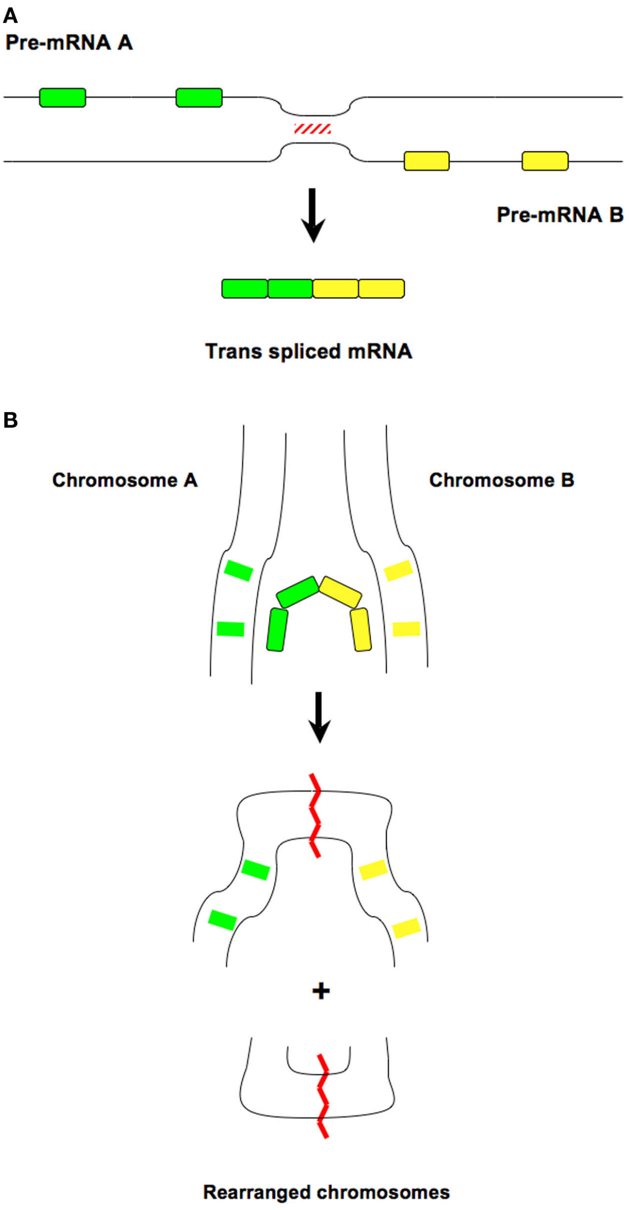


FIGURE 1 | Hypothetical scheme on the role of *trans*-splicing in promoting genomic rearrangement. (A) Two independent primary RNA transcripts are brought in proximity, possibly through base pair complementarity, allowing splicing events in *trans* to take place. Exons are shown by green and yellow boxes. The putative base pairing is highlighted by red lines. **(B)** The *trans*-spliced RNA may serve as a scaffold to facilitate genomic interactions that could lead to gene translocation. Exons are shown by green and yellow boxes. The position of genomic recombination in the rearranged chromosomes is highlighted by red lines.

In an additional reported case, a fusion transcript composed of the *SLC45A3* exon 1 joined to the *ELK4* exon 2 was found to be expressed in both benign prostate tissue and prostate cancer. Moreover, the levels of the fusion transcript did not correlate with alterations at the chromosomal level, raising the possibility that a mechanism for generation of the fusion transcript may be *trans*-splicing (Rickman et al., 2009). However, as both *SLC45A3* and the *ELK4* genes are positioned within 30 kb and in the same orientation on chromosome 1, an alternative interpretation is the presence of an extended bicistronic primary RNA out of which the "fusion" transcript is processed. Interestingly, more than a decade ago exon scrambling events on the *AML* (*MLL*) gene, which is frequently rearranged in human leukemias, that could only be partly interpreted as the result of genome duplications, were reported (Caligiuri et al., 1996; Caldas et al., 1998). Thus, the possibility of non-linear splicing processes that mimic genomic alterations had already been raised.

TRANS-SPLICING VERSUS ALTERNATIVE SPLICING – PARALLEL PATHS

Apart from certain well-documented cases of abundant *trans*-splicing events with functional implications (Gingeras, 2009), most reports on *trans*-splicing or exon scrambling are indicative of an infrequent process, with its biological significance being questioned. It is therefore possible that the majority of these non-*cis*-splicing events are products of an error prone RNA processing machinery, with limited functional consequences. However, such arguments are reminiscent of the evolution of the concept of alternative *cis*-splicing. Since its identification in the late seventies, alternative *cis*-splicing was thought for years to represent an oddity in gene expression. It is not until the last decade that it has clearly been demonstrated that this phenomenon characterizes almost all human genes (Pan et al., 2008; Wang et al., 2008). The deeper the extent of the transcriptome analysis, the higher the diversity of the identified alternative transcripts. It is therefore envisioned that with the advent of global deep sequencing technologies, which could directly sequence long RNAs at a single molecule level (Ozsolak et al., 2009), it may be possible to get convincing evidence on the pervasiveness of *trans*-splicing or exon scrambling, and their possible biological significance. Thus, the approximate 1% of human genes that engage in non-linear exon splicing, deduced from GenBank entries a few years ago (Dixon et al., 2005), is anticipated to increase. In line with this goal has been the effort to use pair end sequencing in *Drosophila* mRNAs, which identified 80 novel cases of *trans*-splicing between homologous alleles (McManus et al., 2010).

EVOLUTIONARY "TINKERING"

The concept of "bricoleur" by Jacob (2001) may be quite relevant in envisioning the biological implications of *trans*-splicing. An organism is likely to take advantage of all available "tools" in order to adapt to a constantly changing environment. As processed RNAs are composed of joined exons, more complex "tools" may be produced by a combinatorial use of exons that originate from two or more gene loci, providing a new means for expanding the diversity of the transcriptome and the proteome. Thus, similarly to alternative *cis*-splicing, which has been demonstrated to be more pronounced in higher than lower eukaryotes (Kim et al., 2007), *trans*-splicing may be a way for eukaryotic cells to take advantage of novel exon combinations that are not limited by a linear *cis* arrangement in the genome. Considering that more complex organisms do not differ so much from simpler ones, as far as the numbers of protein coding genes are concerned, it may be that significance attention should focus on the regulation of gene expression. Consequently, *trans*-splicing is to be regarded as a regulatory process that diversifies the output of exon containing genes.

FINAL NOTE

An elegant hypothesis on the evolutionary role of *trans*-splicing has recently been put forward by Blumenthal (2011). In the unicellular parasite *Giardia*, three convincing cases of *trans*-splicing mediated by base pair interactions of independent transcripts were reported (Kamikawa et al., 2011; Nageshan et al., 2011), resulting in the formation of mature mRNAs for heat shock protein 90 and dynein molecular motor protein β, which, in other organisms, are produced from single, *cis*-spliced gene loci. Thus, it is proposed that during evolution *trans*-spliced molecules, such as the ones described in *Giardia*, may have guided genomic rearrangements resulting in the formation of contiguous genes. This possibility is in line with the RNA mediated genomic rearrangement that occurs in the ciliate *Oxytricha* and the one suggested for the *JAZF1-JJAZ1* *trans*-spliced RNA in endometrial cancer (Li et al., 2008; Nowacki et al., 2008). Further analysis of the genome/transcriptome of other diplomonads and related organisms is anticipated to provide additional clues in this direction.

ACKNOWLEDGMENTS

I apologize to the many authors in the field, whose work is not cited due to the limited number of references. Our laboratory is supported by the Swedish Cancer Fund, the Swedish Research Council and the Swedish Childhood Cancer Foundation.

REFERENCES

Blumenthal, T. (1993). Mammalian cells can trans-splice. But do they? *Bioessays* 15, 347–348.

Blumenthal, T. (2011). Split genes: another surprise from *Giardia. Curr. Biol.* 21, R162–R163.

Bruzik, J. P., and Maniatis, T. (1992). Spliced leader RNAs from lower eukaryotes are trans-spliced in mammalian cells. *Nature* 360, 692–695.

Caldas, C., So, C. W., MacGregor, A., Ford, A. M., McDonald, B., Chan, L. C., and Wiedemann, L. M. (1998). Exon scrambling of MLL transcripts occur commonly and mimic partial genomic duplication of the gene. *Gene* 208, 167–176.

Caligiuri, M. A., Strout, M. P., Schichman, S. A., Mrózek, K., Arthur, D. C., Herzig, G. P., Baer, M. R., Schiffer, C. A., Heinonen, K., Knuutila, S., Nousiainen, T., Ruutu, T., Block, A. W., Schulman, P., Pedersen-Bjergaard, J., Croce, C. M., and Bloomfield, C. D. (1996). Partial tandem duplication of ALL1 as a recurrent molecular defect in acute myeloid leukemia with trisomy 11. *Cancer Res.* 56, 1418–1425.

Caudevilla, C., Serra, D., Miliar, A., Codony, C., Asins, G., Bach, M., and Hegardt, F. G. (1998). Natural trans-splicing in carnitine octanoyltransferase pre-mRNAs in rat liver. *Proc. Natl. Acad. Sci. U.S.A.* 95, 12185–12190.

Cocquerelle, C., Daubersies, P., Majérus, M. A., Kerckaert, J. P., and Bailleul, B. (1992). Splicing with inverted order of exons occurs proximal to large introns. *EMBO J.* 11, 1095–1098.

Dixon, R. J., Eperon, I. C., Hall, L., and Samani, N. J. (2005). A genome-wide survey demonstrates widespread non-linear mRNA in expressed sequences from multiple species. *Nucleic Acids Res.* 33, 5904–5913.

Dixon, R. J., Eperon, I. C., and Samani, N. J. (2007). Complementary intron sequence motifs associated with human exon repetition: a role for intragenic, inter-transcript interactions in gene expression. *Bioinformatics* 23, 150–155.

Dorn, R., Reuter, G., and Loewendorf, A. (2001). Transgene analysis proves mRNA trans-splicing at the complex mod(mdg4) locus in *Drosophila. Proc. Natl. Acad. Sci. U.S.A.* 98, 9724–9729.

Durand, S., and Lykke-Andersen, J. (2011). SnapShot: nonsense-mediated mRNA decay. *Cell* 145, 324–324.e2.

Finta, C., and Zaphiropoulos, P. G. (2002). Intergenic mRNA molecules resulting from trans-splicing. *J. Biol. Chem.* 277, 5882–5890.

Fischer, S. E., Butler, M. D., Pan, Q., and Ruvkun, G. (2008). Trans-splicing in *C. elegans* generates the negative RNAi regulator ERI-6/7. *Nature* 455, 491–496.

Frantz, S. A., Thiara, A. S., Lodwick, D., Ng, L. L., Eperon, I. C., and Samani, N. J. (1999). Exon repetition in mRNA. *Proc. Natl. Acad. Sci. U.S.A.* 96, 5400–5405.

Gingeras, T. R. (2009). Implications of chimaeric non-co-linear transcripts. *Nature* 461, 206–211.

Horiuchi, T., Giniger, E., and Aigaki, T. (2003). Alternative trans-splicing of constant and variable exons of a *Drosophila* axon guidance gene, lola. *Genes Dev.* 17, 2496–2501.

Houseley, J., and Tollervey, D. (2010). Apparent non-canonical trans-splicing is generated by reverse transcriptase in vitro. *PLoS ONE* 5, e12271. doi:10.1371/journal.pone.0012271

Jacob, F. (2001). Complexity and tinkering. *Ann. N. Y. Acad. Sci.* 929, 71–73.

Kamikawa, R., Inagaki, Y., Tokoro, M., Roger, A. J., and Hashimoto, T. (2011). Split introns in the genome of *Giardia intestinalis* are excised by spliceosome-mediated trans-splicing. *Curr. Biol.* 21, 311–315.

Kim, E., Magen, A., and Ast, G. (2007). Different levels of alternative splicing among eukaryotes. *Nucleic Acids Res.* 35, 125–131.

Konarska, M. M., Padgett, R. A., and Sharp, P. A. (1985). Trans splicing of mRNA precursors in vitro. *Cell* 42, 165–171.

Kralovicova, J., and Vorechovsky, I. (2005). Intergenic transcripts in genes with phase I introns. *Genomics* 85, 431–440.

Krause, M., and Hirsh, D. (1987). A trans-spliced leader sequence on actin mRNA in *C. elegans. Cell* 49, 753–761.

Labrador, M., Mongelard, F., Plata-Rengifo, P., Baxter, E. M., Corces, V. G., and Gerasimova, T. I. (2001). Protein encoding by both DNA strands. *Nature* 409, 1000.

Li, H., Wang, J., Mor, G., and Sklar, J. (2008). A neoplastic gene fusion mimics trans-splicing of RNAs in normal human cells. *Science* 321, 1357–1361.

Li, X., Zhao, L., Jiang, H., and Wang, W. (2009). Short homologous sequences are strongly associated with the generation of chimeric RNAs in eukaryotes. *J. Mol. Evol.* 68, 56–65.

McManus, C. J., Duff, M. O., Eipper-Mains, J., and Graveley, B. R. (2010). Global analysis of trans-splicing in *Drosophila. Proc. Natl. Acad. Sci. U.S.A.* 107, 12975–12979.

Murphy, W. J., Watkins, K. P., and Agabian, N. (1986). Identification of a novel Y branch structure as an intermediate in trypanosome mRNA processing: evidence for trans splicing. *Cell* 47, 517–525.

Nageshan, R. K., Roy, N., Hehl, A. B., and Tatu, U. (2011). Post-transcriptional repair of a split heat shock protein 90 gene by mRNA trans-splicing. *J. Biol. Chem.* 286, 7116–7122.

Nigro, J. M., Cho, K. R., Fearon, E. R., Kern, S. E., Ruppert, J. M., Oliner, J. D., Kinzler, K. W., and Vogelstein, B. (1991). Scrambled exons. *Cell* 64, 607–613.

Nowacki, M., Vijayan, V., Zhou, Y., Schotanus, K., Doak, T. G., and Landweber, L. F. (2008). RNA-mediated epigenetic programming of a genome-rearrangement pathway. *Nature* 451, 153–158.

Ozsolak, F., Platt, A. R., Jones, D. R., Reifenberger, J. G., Sass, L. E., McInerney, P., Thompson, J. F., Bowers, J., Jarosz, M., and Milos, P. M. (2009). Direct RNA sequencing. *Nature* 461, 814–818.

Pan, Q., Shai, O., Lee, L. J., Frey, B. J., and Blencowe, B. J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.* 40, 1413–1415.

Rickman, D. S., Pflueger, D., Moss, B., VanDoren, V. E., Chen, C. X., de la Taille, A., Kuefer, R., Tewari, A. K., Setlur, S. R., Demichelis, F., and Rubin, M. A. (2009). SLC45A3-ELK4 is a novel and frequent erythroblast transformation-specific fusion transcript in prostate cancer. *Cancer Res.* 69, 2734–2738.

Solnick, D. (1985). Trans splicing of mRNA precursors. *Cell* 42, 157–164.

Sutton, R. E., and Boothroyd, J. C. (1986). Evidence for trans splicing in trypanosomes. *Cell* 47, 527–535.

Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S. F., Schroth, G. P., and Burge, C. B. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature* 456, 470–476.

Zaphiropoulos, P. G. (1998). Non-homologous recombination mediated by *Thermus aquaticus* DNA polymerase I. Evidence supporting a copy choice mechanism. *Nucleic Acids Res.* 26, 2843–2848.

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 03 June 2011; paper pending published: 23 June 2011; accepted: 07 December 2011; published online: 26 December 2011.

Citation: Zaphiropoulos PG (2011) Trans-splicing in higher eukaryotes: implications for cancer development? Front. Gene. **2***:92. doi: 10.3389/fgene.2011.00092*

This article was submitted to Frontiers in Non-Coding RNA, a specialty of Frontiers in Genetics.

frontiers in
GENETICS

REVIEW ARTICLE
published: 28 December 2011
doi: 10.3389/fgene.2011.00094

The beginning of the road for non-coding RNAs in normal hematopoiesis and hematologic malignancies

Elisabeth F. Heuston†**, Kenya T. Lemon**† **and Robert J. Arceci** *

Department of Oncology, Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University School of Medicine, Baltimore, MD, USA

Edited by:
Philipp Kapranov, St. Laurent Institute, USA

Reviewed by:
Soheil Meshinchi, Fred Hutchinson Cancer Research Center, USA
TIm McCaffrey, The George Washington University Hospital, USA
Timothy J. Triche, USC Keck School of Medicine, USA

Correspondence:
Robert J. Arceci, Sydney Kimmel Comprehensive Cancer Center at Johns Hopkins, The Bunting and Blaustein Cancer Research Building, 1650 Orleans Street, Suite 207, Baltimore, MD 21231, USA.
e-mail: arcecro@jhmi.edu

† Elisabeth F. Heuston and Kenya T. Lemon have contributed equally to this work.

The field of non-coding RNAs (ncRNAs) encompasses a wide array of RNA classes that are indispensible for the regulation of cellular activities. However, de-regulation of these ncRNAs can also play key roles in malignant transformation and cancer cell behavior. In this article we survey a select group of microRNAs and long ncRNAs that appear to contribute in keys ways to the development of acute and chronic leukemias, as well as contribute to their diagnosis, prognosis, and potentially, their treatment.

Keywords: non-coding RNA, AML, ALL, CLL, CML

INTRODUCTION

Non-coding RNAs (ncRNAs) are regarded as regulators of cell cycle progression, proliferation, and fate. There are numerous classes of ncRNAs, including very long ncRNAs, PIWI-associated ncRNAs, and small interfering RNAs. However few have been described with respect to hematopoiesis. Here we review the current understanding of the role of microRNAs (miRNA) and long non-coding RNAs (lncRNAs) in the pathogenesis of acute myeloid leukemia (AML), acute lymphoblastic leukemia (ALL), chronic lymphocytic leukemia (CLL), and chronic myeloid leukemia (CML).

REGULATORY microRNAs IN LEUKEMIA

MicroRNAs, negative regulators of gene expression, are initially transcribed as long primary miRNAs by RNA polymerase II, and subsequently processed by the nuclear ribonuclease Drosha and the cytoplasmic dsRNA-specific endonuclease Dicer. Mature miRNAs, averaging 22 nucleotides, associate with the RNA induced silencing complex (RISC) that aids in generating their repressive functions. The most efficient miRNA targeting is achieved by mature miRNAs binding to the 3′-untranslated regions (UTR) of target messenger RNAs (mRNA) leading to translational inhibition or mRNA cleavage (Zhang et al., 2007). miRNAs also bind within coding regions and 5′UTRs, through which additional regulation of mRNA translation can be mediated (**Figure 1**; Ørom et al., 2008; Tay et al., 2008). miRNAs have garnered much interest due to their role as post-transcriptional regulators of genes involved in numerous physiologic and developmental processes including cellular proliferation, cell cycle progression, and apoptosis (reviewed in Lee and Ambros, 2001; Cheng et al., 2005; Hayashita et al., 2005; Ivanovska et al., 2008). Altered miRNA expression has been implicated in the pathology of leukemia, where some miRNAs are proposed to function either as tumor suppressor genes (TS) or oncogenes (**Table 1**; Esquela-Kerscher and Slack, 2006).

miRNAs IN AML

Acute myeloid leukemia is characterized by the abnormal proliferation and accumulation of immature myeloblasts in the bone marrow and blood. Diagnosis is based on the presence of malignant blasts in the bone marrow as well as distinctive translocations and gene mutations. AML represents a group of heterogeneous disorders with striking differences in survival based on cytogenetics, prognostic gene mutations, and age and has an average long-term overall survival (OS) rate of approximately 25–60% depending on the above characteristics in patients over age 55 (Heerema-McKenney and Arber, 2009; http://seer.cancer.gov, accessed September 24, 2011).

Golub et al. (1999) first demonstrated that expression profiling of protein-encoding genes could distinguish AML from ALL, pioneering the way for expression-based diagnostic approaches. The advent of miRNA expression profiling further distinguished AML from ALL and has been used to explore possible associations of miRNAs with regard to diagnosis, prognosis, and response to treatment (Calin and Croce, 2006; Mi et al., 2007). Garzon et al. (2008b) reported that the expression of *miR-29* family member miR-29b is often down-regulated in primary AML with respect to bone marrow CD34+ progenitors. miR-29b directly targets the 3′UTR of

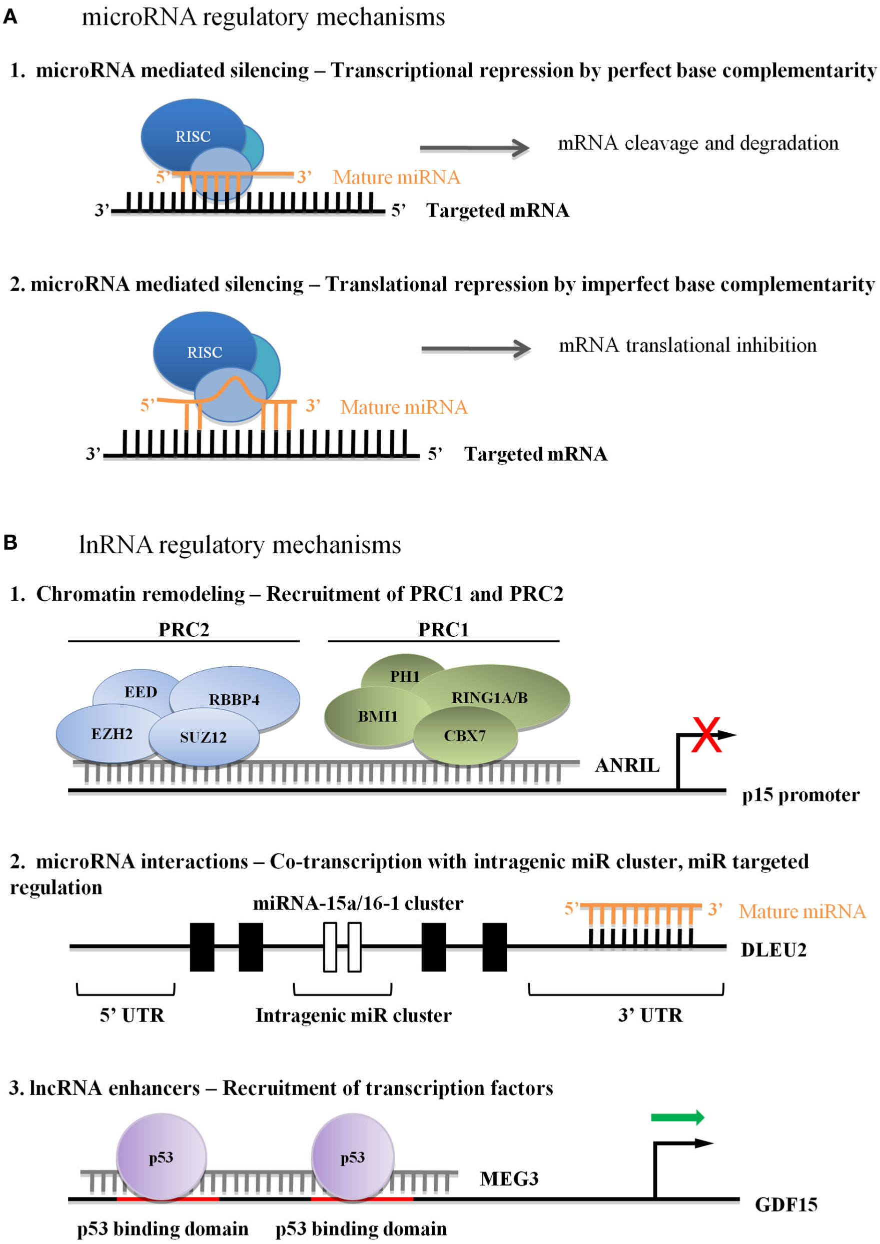

A
microRNA regulatory mechanisms
1. microRNA mediated silencing – Transcriptional repression by perfect base complementarity
RISC
mRNA cleavage and degradation
5'
3'
Mature miRNA
3'
5'
Targeted mRNA
2. microRNA mediated silencing – Translational repression by imperfect base complementarity
RISC
mRNA translational inhibition
5'
3'
Mature miRNA
3'
5'
Targeted mRNA
B
lnRNA regulatory mechanisms
1. Chromatin remodeling – Recruitment of PRC1 and PRC2
PRC2
PRC1
EED
RBBP4
EZH2
SUZ12
PH1
RING1A/B
BMI1
CBX7
ANRIL
p15 promoter
2. microRNA interactions – Co-transcription with intragenic miR cluster, miR targeted regulation
miRNA-15a/16-1 cluster
5'
3'
Mature miRNA
DLEU2
5' UTR
Intragenic miR cluster
3' UTR
3. lncRNA enhancers – Recruitment of transcription factors
p53
p53
MEG3
GDF15
p53 binding domain
p53 binding domain

FIGURE 1 | (A) Regulatory microRNA mechanisms in leukemia. **(A1)** Perfect base complementarity between miRNA and target mRNA can lead to target mRNA degradation. **(A2)** Imperfect base complementarity between miRNA and target mRNA can lead to mRNA translational repression. **(B)** Currently hypothesized mechanisms of lncRNA regulation in leukemia. **(B1)** Recruitment of PcGs leads to stable repression, although the mechanism of lncRNA association with the target sequence is unknown. **(B2)** lncRNAs can influence miRNA cluster transcriptional activity, and can be post-transcriptionally regulated by miRNAs. **(B3)** Activator lncRNAs enhance transcriptional or binding affinity of transcription factors, although the mechanism of lncRNA association with the target sequence is unknown.

DNA methyltransferases *DNMT3a* and *DNMT3b* and indirectly targets *DNMT1* via negative regulation of *SP1*, a transactivator of *DNMT1*. Down-regulation of the methyltransferases caused by forced expression of miR-29b leads to altered DNA methylation and subsequent repression of TS genes such as *p15*^{INK4b} and *ESR1* often observed in myeloid leukemogenesis (Garzon et al., 2009b). Restoration of miR-29b increases sensitivity to the hypomethylating agent, decitabine, in patients with AML over age 60 (Blum et al., 2010). Forced expression of miR-29b in AML cell lines and primary AML samples reduces cell growth and induces apoptosis, indicative of its TS activity in acute leukemia. Additional support was provided by the dramatic reduction of tumors in a xenograft leukemia model in response to miR-29b over-expression. Following miR-29b expression, gene expression analysis of the AML cell line K562 showed that miR-29b, along with miR-29a, target pathways involved with cell cycle, proliferation, and apoptosis (Garzon

Table 1 | Characteristics of discussed ncRNAs[a].

ncRNA(s)	ncRNA class	Target	Disease association	Clinical relevance	Citation
miR-29b	miRNA	DNMT3a and DNMT3b, DNMT1 (indirect)	AML	Predictive response to therapy; potentially therapeutic	Garzon et al. (2009b)
miR-126/126*	miRNA	N/A	AML	Diagnostic: CBF	Li et al. (2008)
miR-224, -368, -382	miRNA	N/A	AML	Diagnostic: *t*(15; 17)	Li et al. (2008)
miR-126, -126*, -224, -368, -382, -17-5p, -20a	miRNA	N/A	AML	Diagnostic: MLL-rearrangement	Li et al. (2008)
miR-21	miRNA	PTEN	AML	Diagnostic: *t*(6; 11) vs. *t*(9; 11), prognostic	Garzon et al. (2008a)
miR-29b	miRNA	TCL1	AML	Diagnostic: balanced 11q23 translocations; potentially therapeutic	Garzon et al. (2008a)
miR-155	miRNA	Unknown in AML	AML	Diagnostic: FLT3–ITD mutations	Garzon et al. (2008a)
miR-199a, -191	miRNA	N/A	AML	Prognostic	Garzon et al. (2008a)
miR-181	miRNA	TLR and IL1-β	AML	Prognostic	Marcucci et al. (2008)
(miR-128a, -128b) vs. (miR-223, let-7b]	miRNA	N/A	AML vs. ALL	Diagnostic	Mi et al. (2007)
miR-18a, -532, -218, -625, 193a, -638, -550, -663	miRNA	N/A	ALL	Prognostic	Zhang et al. (2009a)
miR-143	miRNA	MLL–AF4	ALL	Prognostic, potentially therapeutic	Dou et al. (2011)
miR-15a, -16-1	miRNA	Bcl2	CLL	Potentially therapeutic	Bandi et al. (2009b)
miR-29, -181	miRNA	TCL1	CLL	Prognostic, potentially therapeutic	Pekarsky et al. (2006)
miR-15a, -195, -221, -23b, -155, -223, -29a, -24, -29b, -146, -16, -16-2, -29c	miRNA	N/A	CLL	Diagnostic: expressed ZAP-70 and unmutated IgVH vs. no ZAP-70 and mutated IgVH	San Jose-Eneriz et al. (2009)
miR-7, -23a, -26a, -29a, -29c, -30b, -30c, -100, -126, -134, -141, -183, -196b, -199a, -224, -362, -422b, 520a, -191	miRNA	N/A	CML	Predictive response to therapy	San Jose-Eneriz et al. (2009)
ANRIL, p15AS	Antisense	p15	AML/ALL	Diagnostic, potentially prognostic	Yu et al. (2008); Iacobucci et al. (2011)
lincRNA-p21	lincRNA	BCR-ABL	Potentially CML	Potentially therapeutic	Notari et al. (2006); Du et al. (2010)
MEG3	lincRNA	p53, GDF15	MDS/AML	Prognostic	Benetatos et al. (2009)
Combined T-UCR and miR profile	UCR	N/A	CLL/AML	Diagnostic: tumor type; prognostic	Calin et al. (2007)
Dleu2	lincRNA	miR-15a/16	CLL	Diagnostic, potentially therapeutic	Migliazza et al. (2001); Lerner et al. (2009)
HOTAIRM1	Antisense	HOXA1, HOXA4, CD11b, CD18	Hematopoietic regulator	Potentially therapeutic	Zhang et al. (2009b)
EGO	Antisense	MBP, EDN	Hematopoietic regulator	Potentially therapeutic	Wagner et al. (2007)
lincRNA-a7	lncRNA	SCL/TAL1	Hematopoietic regulator	Potentially therapeutic	Ørom et al. (2010)

[a] *Non-coding RNAs having different diagnostic or therapeutic roles in different leukemia subtypes are listed separately.*

et al., 2009a). Together these data suggest that synthetic *miR-29b* oligonucleotides could potentially serve as a therapeutic approach in AML.

Just as numerous groups have identified mRNA signatures distinguishing AML subclasses, several groups have associated miRNA signatures with cytogenetic abnormalities and predictors of outcome in AML (Bullinger et al., 2004; Valk et al., 2004). Li et al. reported that miRNA expression signatures accurately discriminate between AMLs with the common translocations *t*(15; 17), *t*(8; 21), and inv(16), all of which represent favorable prognosis rearrangements that result in the disruption of core-binding factors (CBF). Expression of two (miR-126/126*), three (miR-224, -368, and -382), and seven (miR-126, -126*, -224, -368, -382, -17-5p, and -20a) miRNAs distinguishes CBF *t*(8; 21) and inv16, *t*(15; 17) and MLL-rearrangement AMLs, respectively, from one another (Li et al., 2008). Up-regulation of miR-21 distinguishes AMLs with *t*(6; 11) from those with *t*(9; 11), while down-modulation of miR-29 correlates with balanced 11q23 translocations. miR-155 shows increased expression in patients with AML characterized by FLT3–ITD mutations (Garzon et al., 2008a).

Expression patterns of miRNAs are also associated with prognosis. Marcucci et al. (2008) have associated a worse prognosis in patients with *t*(6; 11) who display increased expression of miR-21, an inhibitor of the TS PTEN. This group has also shown that patients with AML that express elevated miR-199a and -191 have a significantly lower OS and event-free survival (Garzon et al., 2008a). The miR-181 family has been shown to contribute to an aggressive AML phenotype through mechanisms associated with the activation of pathways controlled by toll-like receptors and interleukin-1β (Marcucci et al., 2008).

miRNAs IN ALL

Acute lymphoblastic leukemia represents a heterogeneous group of disorders that result in the malignant transformation of lymphoblasts at various stages of differentiation. Although more common in children, ALL also occurs in adults, with a shift from childhood-prevalent to adulthood-prevalent subtypes during adolescence. A diagnosis of ALL occurs when blood and bone marrow samples show a large number of abnormal lymphocyte blasts. Treatment options are based on ALL subtype and prognostic factors such as age.

Mi et al. (2007) have shown that expression signatures of as few as two miRNAs can discriminate ALL from AML; miR-128a and -128b were significantly higher in ALL, whereas miR-223 and let-7b were expressed at significantly higher levels in AML. This study also demonstrated that lineage discriminating miRNAs could differentiate ALL samples from AML samples even when both leukemias displayed the same translocation/fusion events such as *t*(11; 19)(q23; p13.3)/MLL–ENL (Mi et al., 2007). Zhang et al. identified an 8-miRNA-expression profile (miR-18a, -532, -218, -625, -193a, -638, -550, and -663) that differentiates good from poor steroid response in pediatric ALL. miRNA expression signatures were also able to identify relapse and non-relapse pediatric cases (Zhang et al., 2009a). miR-143 has been shown to negatively regulate the MLL–AF4 fusion protein, a product of the translocation *t*(4; 11)(q21; q23), which confers poorer prognosis compared to other MLL arrangements. Restoration of miR-143 in MLL–AF4-positive cells induced apoptosis by reducing MLL–AF4 fusion protein levels, suggesting that forced expression of miR-128 could serve as a therapeutic agent in MLL–AF4 ALL (Dou et al., 2011).

miRNAs IN CLL

Chronic lymphocytic leukemia, the most common leukemia found in adults, results from immature, malignant resting B cell lymphocytes overexpressing the anti-apoptotic B cell lymphoma 2 (Bcl2) protein and accumulating in the bone marrow and the blood (Cimmino et al., 2005). Of those diagnosed with CLL, most are over the age of 50. CLL exhibits clinical heterogeneity as marked by the observation that some patients present with aggressive leukemia requiring immediate treatment while others require no intervention for many years (Li et al., 2011).

The first demonstration of miRNA association with pathogenesis was discovered in CLL (Calin et al., 2002). The Croce group showed that miR-15a and -16 reside in the fragile chromosomal band 13q14, which is deleted in more than half of CLL cases, as well as some prostate tumor and retinoblastoma samples (Chen et al., 2001; Kivelä et al., 2003). They went on to show that miR-15a and -16 negatively regulate cell growth and cell cycle progression (Calin et al., 2002). Additionally, over-expression of these miRNAs has been shown to repress Bcl2 expression and induce apoptosis in a leukemic cell line model, indicating a potential therapeutic role for miR-15a and -16 in the treatment of Bcl2-overexpressing tumors (Bandi et al., 2009a).

Two markers of aggressive CLL include up-regulated 70-kDa zeta-associated protein (ZAP-70) and unmutated immunoglobulin variable genes (IgVH). When these factors are expressed at high levels they are associated with high levels of TCL1, an oncogene that co-activates the anti-apoptotic oncoprotein AKT and aids in regulatory pathways involved in cell survival and death. miR-29 and -181 have been shown to negatively regulate TCL1. Clinical CLL samples show miR-29 and -181 expression is inversely related to TCL1 expression, suggesting that these miRNAs could potentially serve as prognostic markers of CLL progression and as therapeutic agents in aggressive forms of TCL1-overexpressing CLL (Pekarsky et al., 2006). Calin et al. (2005) demonstrated that 13 miRNAs (miR-15a, -195, -221, -23b, -155, -223, -29a, -24, -29b, -146, -16, -16-2, and -29c) could discriminate between patients expressing ZAP-70 and unmutated IgVH and those not expressing ZAP-70 and mutated IgVH, thereby distinguishing aggressive from indolent CLL.

miRNAs IN CML

Chronic myeloid leukemia is a malignant clonal stem cell disorder characterized by an increase of mature granulocytes in the bone marrow and blood. It often expresses the constitutively active BCR-ABL tyrosine kinase formed by a translocation that links *Abl1* on chromosomal band 9q34 to a portion of *BCR* on chromosomal band 22q11. CML is often suspected on the basis an extremely elevated white blood cell count with maturation of white blood cells and few or no leukemic blasts. Treatment and prognosis depend primarily on the phase of CML, i.e., whether in chronic, accelerated, or blast crisis. The development of the BCR-ABL1 kinase inhibitor, imatinib mesylate, as well as other Bcr–Abl inhibitors, has significantly improved treatment and outcome of patients with CML (Kantarjian et al., 2002). Jose-Eneriz et al. (2009) identified 19 miRNAs (18 up-regulated: miR-7, -23a, -26a, -29a, -29c, -30b, -30c, -100, -126, -134, -141, -183, -196b, -199a, -224, -362, -422b, -520a, and 1 down-regulated: miR-191) that are differentially expressed between imatinib resistant and responder samples.

LONG NON-CODING RNAs IN LEUKEMIA

Long non-coding RNAs are ncRNAs greater than 200 nucleotides long, transcribed by RNA polymerase II or III, and can account for nearly 60% of all non-ribosomal and non-mitochondrial RNA in human cells (Kapranov et al., 2010). LncRNAs are involved in transcriptional silencing, chromatin remodeling, and gene activation (reviewed in Huarte and Rinn, 2010; Gibb et al., 2011) and originate from many chromosomal environments, including antisense to protein-coding genes, intergenic DNA (termed "lincRNAs" for long intergenic non-coding RNAs'), and from ultraconserved regions (Carninci et al., 2005). Excluding the latter, only a few lncRNA primary sequences are evolutionarily conserved (Guttman et al., 2009; Baker, 2011). In contrast to miRNAs, several cases have shown that conservation of secondary structure is more important to preserving ncRNA function than nucleotide sequence (Yap et al., 2010). Although lncRNAs in leukemia are not as extensively characterized as they are in some other tumor types, recent progress has identified several high profile targets that could alter how lncRNAs are viewed in terms of leukemia development, classification, and therapeutic targeting.

SUPPRESSORS OF TUMOR SUPPRESSORS ARE ONCOGENES

The INK4A–ARF–INK4B locus is deregulated in up to 40% of human cancers (Sherr, 1998; Kim and Sharpless, 2006). In addition to p15^{INK4b}, p14^{ARF}, and p16^{INK4a}, TS genes interact closely with p53 and Rb to regulate cell cycle progression (Bandi et al., 2009a). Pasmant et al. identified a polyadenylated lncRNA that was transcribed antisense to p15^{INK4b}. The full-length transcript, ANRIL (antisense ncRNAs in the INK4 locus) has several isoforms (Pasmant et al., 2007). The p15AS variant, isolated from two AML cell lines, was significantly up-regulated in 11 of 16 AML and ALL primary samples (Yu et al., 2008). The authors demonstrated that p15AS was responsible for Dicer-independent silencing of p15^{INK4b} by altering H3K9me2 and H3K4me2 levels at both endogenous and exogenous p15^{INK4b} promoters. EZH2 and SUZ12 were required for stable silencing of p15^{INK4b}, even after p15AS repression, indicating the likely recruitment of the polycomb repressive complex 2 (PRC2). Other reports demonstrated ANRIL-CBX7 interacts with either di- or tri-methylated H3K27, implying that ANRIL-mediated silencing works through both PRC1 and PRC2 complexes (**Figure 1**; Group et al., 2005; Kotake et al., 2011; Margueron and Reinberg, 2011). A study by Iacobucci et al. (2011) involving acute leukemia and normal peripheral blood cells showed a statistically significant association between an ANRIL nucleotide polymorphism and ALL phenotype, demonstrating the importance of regulated ANRIL expression in primary leukemia.

Acting downstream of ANRIL, lincRNA-p21 is a 3.1-kb transcript induced by p53 expression that represses cellular proliferation by both p53-dependent and independent mechanisms (Huarte et al., 2010). LincRNA-p21 contains two canonical p53 binding sites in a promoter distinct from its closest neighbor, CDNK1A. Independent knock-down of either p53 or lincRNA-p21 produces significantly overlapping gene set enrichments ($p < 10^{-200}$), implying at least a partial overlap of apoptosis-induction mechanisms. Interestingly, in p53$^{-/-}$ cell lines, lincRNA-p21 cannot direct proper localization and function of the pre-mRNA binding protein hnRNP-K (Huarte et al., 2010). Although these functional studies were done in mouse embryonic fibroblasts and human lung carcinoma cell lines, two reports demonstrate that BCR-ABL stimulates hnRNP-K expression and stability, subsequently promoting tumor progression (Notari et al., 2006; Du et al., 2010). Although the activity of lincRNA-p21 in acute or chronic leukemia is currently unknown, these data suggest that further work in CML is warranted.

p53-INTERACTING TUMOR SUPPRESSOR lncRNAs

A second p53-regulating lincRNA, MEG3, was highlighted in human malignancies when investigators reported high expression levels in normal gonadotrophs and severely reduced expression in tumor-derived gonadotroph cells (Zhang et al., 2003). Forcing re-expression of MEG3 in HCT-116 cells leads to p53 accumulation and inhibition of cellular proliferation, indicative of a high-level regulator of p53-dependent TS activities (Zhang et al., 2003; Zhou et al., 2007). Although MEG3 isoforms can contain several small open reading frames, they are not required for p53-mediated cellular activities (Zhou et al., 2007; Zhang et al., 2010a). Instead, MEG3 secondary structure is critical to maintaining function, including down-regulation of MDM2 expression and enhanced p53 binding to a specific subset of gene promoters, including GDF15 (**Figure 1**; Zhang et al., 2010a). Of interest, MEG3 can suppress cell growth in p53^{null} cells, indicating p53-independent activities as well (Zhou et al., 2007).

Expression of the MEG3-DLK1 locus is tightly regulated by two differentially methylated regions (DMRs), which are hypermethylated in a subset of solid tumors and suppress MEG3 expression (Kagami et al., 2010; Astuti et al., 2005). Benetatos et al. (2009) examined a cohort of 85 patients with either myelodysplastic syndrome (MDS) or AML. They found that 48% (20/42) of patients with AML displayed aberrant hypermethylation of the MEG3 promoter, which significantly correlated with decreased OS (HR 1.98, $p = 0.04$). In MDS, 35% (15/43) of patients displayed aberrant hypermethylation, a result that trended toward decreased survival but was not quite significant (HR 2.15, $p = 0.072$; Benetatos et al., 2009). An independent assessment of 40 AML samples by a second group confirmed aberrant methylation in the MEG3-associated DMRs in AML samples, but not in normal controls (Khoury et al., 2010). Importantly, neither study showed MEG3 hypermethylation to be associated with karyotype or disease subtype (Benetatos et al., 2009). Although additional studies will be needed to determine the functional role of MEG3 in leukemia, the importance of its interactions with MDM2, p53, and GDF15, as well as p16 in pituitary carcinomas, will likely demonstrate a role for MEG3 in leukemia (Zhang et al., 2010b).

ncRNA–ncRNA REGULATION: TRANSCRIBED ULTRACONSERVED REGIONS AND HOST GENES

Within the human genome, 481 transcribed genomic segments have been identified that are 100% conserved between orthologous regions in the human, mouse, and rat (Bejerano et al., 2004; Calin et al., 2007). Of these "transcribed ultraconserved regions"

(T-UCRs), 39% are contained within intergenic sequences and 43% are intronic; the remainder are exonic or exon-overlapping (Mestdagh et al., 2010). Like miRNAs, T-UCR locations are closely associated with genomic fragile sites and ubiquitously expressed T-UCRs are frequently associated with cancer-associated genomic regions (CAGRs; Calin et al., 2004, 2007). Microarray analysis has shown that T-UCRs differentially expressed in human malignancies are highly likely to be associated with CAGRs of that tumor type (Calin et al., 2007).

Transcribed ultraconserved regions expression has been used to predict disease outcome in both CLL and neuroblastoma samples (Calin et al., 2007). In a survey of 133 cancers and 22 normal tissues, a profile of 19 T-UCRs (8 up- and 11 down-regulated) could differentiate between normal, CLL, colorectal, and hepatocarcinoma samples (Calin et al., 2007). This study also showed that the expression of five T-UCRs (three intronic and two intergenic) could divide a CLL cohort into two prognostic groups, previously defined by low (favorable) vs. high (poor) ZAP-70 expression (Calin et al., 2007). Expression of these diagnostic T-UCRs negatively correlated with the previously defined CLL miRNA signature, suggesting a mechanism for miRNA regulation of these T-UCRs (Calin et al., 2005). Of the diagnostic T-UCR profile, three (uc.160, uc.346A, and uc.348) contain miRNAs recognition sites, including targeting by miR-155:: miR-24:: and miR-29 (Calin et al., 2005). Repressive activity of miR-155 was confirmed in an *in vitro* assay, while the negative correlations were observed between miR-155?uc.346A and miR-24?uc pairings; 160 were validated in the aforementioned diagnostic cohort (Calin et al., 2007). miR-155 over-expression was identified in CLL, while in AML its up-regulation was associated with expanded bone marrow granulocyte and monocyte proliferation (reviewed in Faraoni et al., 2009) and miR-24 loss-of-function has been linked to methotrexate resistance in HCT-116 cells (Mishra et al., 2009). Interestingly, miR-29a has also been shown to regulate MEG3 expression in hepatocarcinoma cell lines (Braconi et al., 2011). Although it is currently unknown whether miRNAs repress T-UCRs, or whether the altered expression of the T-UCRs affects the primary target of the leukemia-associated miRNAs, it is clear that careful regulation of these two ncRNA species is critical to understanding the disease state and for developing ncRNA-associated classification or ncRNA-directed targeted therapy. In addition to miRNA regulation, differential methylation of T-UCR-associated CpG islands (CGI) may also control expression. In a study of 15 leukemia cell lines and 64 primary leukemia samples, differential methylation was seen at three of these CGI (∼60% in cell lines and approximately 18% in primary samples), although the relevance of CGI-mediated hypermethylation to leukemia etiology remains unclear (Lujambio et al., 2010). A much larger cohort will need to be investigated before diagnostic implications can be verified.

Long non-coding RNAs targeting by miRNAs is only one example of ncRNA-ncRNA regulation. Deleted in leukemia 2 (Dleu2) is an lncRNA transcribed from 13q14 (Liu et al., 1997; Migliazza et al., 2001). The Myc-repressed Dleu2 transcript is a host gene for the miR-15a/miR-16-1 cluster, providing the primary transcript from which miR-15a and miR-16-1 are processed (**Figure 1**; Klein et al., 2010). While de-regulation of these two miRNAs has long been associated with CLL, deletion-mapping studies have demonstrated that the Dleu2 transcript is frequently disrupted in CLL cell line, and increased expression of this gene leads to reduced proliferation and clonogenicity (Lerner et al., 2009). Although further analysis is needed in order to validate these observations, Dleu2 serves as an important example of the intricate co-regulation of lncRNA and miRNAs.

HEMATOPOIETIC REGULATOR lncRNAs AS POTENTIAL ONCOGENIC GENES

Newly discovered lncRNAs are being characterized at a rapid pace. In hematopoiesis, the antisense lincRNA, HOTAIRM1, has recently been identified as an essential regulator of myeloid cell differentiation. This lincRNA is transcribed from within the HOXA cluster and regulates HOXA1, HOXA4, CD11b, and CD18 during retinoic acid-induced differentiation of an acute promyelocytic leukemia cell line (Zhang et al., 2009b). EGO is expressed during eosinophil development and is essential for major basic protein and eosinophil-derived neurotoxin mRNA expression (Wagner et al., 2007). Ørom et al. (2010) identified several lncRNA activators, one of which strongly induced expression of SCL/TAL1. Although these lncRNAs have not yet been associated with hematopoietic malignancies, such critical regulators of cell fate are likely to be identified as potent regulators of tumorigenicity.

CONCLUSION

It has become evident in recent years that the de-regulation of miRNAs and lncRNAs plays a critical role in malignant transformation, tumor cell behavior and, in particular, hematologic malignancies. These ncRNAs could prove to be increasingly useful in the development of much needed novel diagnostic, prognostic, and therapeutic strategies for acute and chronic leukemias.

ACKNOWLEDGMENTS

We thank Dr. Jason E. Farrar for his critical evaluation of this manuscript. This work was in part supported by grants from the United States Department of Defense, the NIH (CA120535 and a TARGET initiative through the Children's Oncology Group and CureSearch) as well as from the Lyles Parachini Fund, the Michael Corb Fund, and the Michael Garil Leukemia Survivors Fund and Program. Dr. Arceci is supported in part through the endowed King Fahd Professorship in Pediatric Oncology at Johns Hopkins.

REFERENCES

Astuti, D., Latif, F., Wagner, K., Gentle, D., Cooper, W. N., Catchpoole, D., Grundy, R., Ferguson-Smith, A. C., and Maher, E. R. (2005). Epigenetic alteration at the DLK1-GTL2 imprinted domain in human neoplasia: analysis of neuroblastoma, phaeochromocytoma and Wilms' tumour. *Br. J. Cancer* 92, 1574–1580.

Baker, M. (2011). Long noncoding RNAs: the search for function. *Nat. Methods.* 8, 379–383.

Bandi, N., Zbinden, S., Gugger, M., Arnold, M., Kocher, V., Hasan, L., Kappeler, A., Brunner, T., and Vassella, E. (2009a). miR-15a and miR-16 are implicated in cell cycle regulation in a Rb-dependent manner and are frequently deleted or down-regulated in non-small cell lung cancer. *Cancer Res.* 69, 5553–5559.

Bandi, N., Zbinden, S., Gugger, M., Arnold, M., Kocher, V., Hasan, L., Kappeler, A., Brunner, T., and Vassella, E. (2009b). miR-15a and miR-16 are implicated in cell cycle regulation in a Rb-dependent manner and are frequently deleted or down-regulated in non‚Äí

Small cell lung cancer. *Cancer Res.* 69, 5553–5559.

Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W. J., Mattick, J. S., and Haussler, D. (2004). Ultraconserved elements in the human genome. *Science* 304, 1321–1325.

Benetatos, L., Hatzimichael, E., Dasoula, A., Dranitsaris, G., Tsiara, S., Syrrou, M., Georgiou, I., and Bourantas, K. L. (2009). CpG methylation analysis of the MEG3 and SNRPN imprinted genes in acute myeloid leukemia and myelodysplastic syndromes. *Leuk. Res.* 34, 148–153.

Blum, W., Garzon, R., Klisovic, R. B., Schwind, S., Walker, A., Geyer, S., Liu, S., Havelange, V., Becker, H., Schaaf, L., Mickle, J., Devine, H., Kefauver, C., Devine, S. M., Chan, K. K., Heerema, N. A., Bloomfield, C. D., Grever, M. R., Byrd, J. C., Villalona-Calero, M., Croce, C. M., and Marcucci, G. (2010). Clinical response and miR-29b predictive significance in older AML patients treated with a 10-day schedule of decitabine. *Proc. Natl. Acad. Sci. U.S.A.* 107, 7473–7478.

Braconi, C., Kogure, T., Valeri, N., Huang, N., Nuovo, G., Costinean, S., Negrini, M., Miotto, E., Croce, C. M., and Patel, T. (2011). MicroRNA-29 can regulate expression of the long non-coding RNA gene MEG3 in hepatocellular cancer. *Oncogene* 30, 4750–4756.

Bullinger, L., DöHner, K., Bair, E., FröHling, S., Schlenk, R. F., Tibshirani, R., DöHner, H., and Pollack, J. R. (2004). Use of gene-expression profiling to identify prognostic subclasses in adult acute myeloid leukemia. *N. Engl. J. Med.* 350, 1605–1616.

Calin, G. A., and Croce, C. M. (2006). MicroRNA-cancer connection: the beginning of a new tale. *Cancer Res.* 66, 7390–7394.

Calin, G. A., Dumitru, C. D., Shimizu, M., Bichi, R., Zupo, S., Noch, E., Aldler, H., Rattan, S., Keating, M., Rai, K., Rassenti, L., Kipps, T., Negrini, M., Bullrich, F., and Croce, C. M. (2002). Frequent deletions and down-regulation of micro-RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia. *Proc. Natl. Acad. Sci. U.S.A.* 99, 15524–15529.

Calin, G. A., Ferracin, M., Cimmino, A., Di Leva, G., Shimizu, M., Wojcik, S. E., Iorio, M. V., Visone, R., Sever, N. I., Fabbri, M., Iuliano, R., Palumbo, T., Pichiorri, F., Roldo, C., Garzon, R., Sevignani, C., Rassenti, L., Alder, H., Volinia, S., Liu, C.-G., Kipps, T. J., Negrini, M., and Croce, C. M. (2005). A MicroRNA signature associated with prognosis and progression in chronic lymphocytic leukemia. *N. Engl. J. Med.* 353, 1793–1801.

Calin, G. A., Liu, C.-G., Ferracin, M., Hyslop, T., Spizzo, R., Sevignani, C., Fabbri, M., Cimmino, A., Lee, E. J., Wojcik, S. E., Shimizu, M., Tili, E., Rossi, S., Taccioli, C., Pichiorri, F., Liu, X., Zupo, S., Herlea, V., Gramantieri, L., Lanza, G., Alder, H., Rassenti, L., Volinia, S., Schmittgen, Thomasâ d., Kipps, T. J., Negrini, M., and Croce, C. M. (2007). Ultraconserved regions encoding ncRNAs are altered in human leukemias and carcinomas. *Cancer Cell* 12, 215–229.

Calin, G. A., Sevignani, C., Dumitru, C. D., Hyslop, T., Noch, E., Yendamuri, S., Shimizu, M., Rattan, S., Bullrich, F., Negrini, M., and Croce, C. M. (2004). Human microRNA genes are frequently located at fragile sites and genomic regions involved in cancers. *Proc. Natl. Acad. Sci. U.S.A.* 101, 2999–3004.

Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M. C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C., Kodzius, R., Shimokawa, K., Bajic, V. B., Brenner, S. E., Batalov, S., Forrest, A. R. R., Zavolan, M., Davis, M. J., Wilming, L. G., Aidinis, V., Allen, J. E., Ambesi-Impiombato, A., Apweiler, R., Aturaliya, R. N., Bailey, T. L., Bansal, M., Baxter, L., Beisel, K. W., Bersano, T., Bono, H., Chalk, A. M., Chiu, K. P., Choudhary, V., Christoffels, A., Clutterbuck, D. R., Crowe, M. L., Dalla, E., Dalrymple, B. P., De Bono, B., Gatta, G. D., Di Bernardo, D., Down, T., Engstrom, P., Fagiolini, M., Faulkner, G., Fletcher, C. F., Fukushima, T., Furuno, M., Futaki, S., Gariboldi, M., Georgii-Hemming, P., Gingeras, T. R., Gojobori, T., Green, R. E., Gustincich, S., Harbers, M., Hayashi, Y., Hensch, T. K., Hirokawa, N., Hill, D., Huminiecki, L., Iacono, M., Ikeo, K., Iwama, A., Ishikawa, T., Jakt, M., Kanapin, A., Katoh, M., Kawasawa, Y., Kelso, J., Kitamura, H., Kitano, H., Kollias, G., Krishnan, S. P. T., Kruger, A., Kummerfeld, S. K., Kurochkin, I. V., Lareau, L. F., Lazarevic, D., Lipovich, L., Liu, J., Liuni, S., Mcwilliam, S., Babu, M. M., Madera, M., Marchionni, L., Matsuda, H., Matsuzawa, S., Miki, H., Mignone, F., Miyake, S., Morris, K., Mottagui-Tabar, S., Mulder, N., Nakano, N., Nakauchi, H., Ng, P., Nilsson, R., Nishiguchi, S., Nishikawa, S., Nori, F., Ohara, O., Okazaki, Y., Orlando, V., Pang, K. C., Pavan, W. J., Pavesi, G., Pesole, G., Petrovsky, N., Piazza, S., Reed, J., Reid, J. F., Ring, B. Z., Ringwald, M., Rost, B., Ruan, Y., Salzberg, S. L., Sandelin, A., Schneider, C., Schönbach, C., Sekiguchi, K., Semple, C. A., Seno, S., Sessa, L., Sheng, Y., Shibata, Y., Shimada, H., Shimada, K., Silva, D., Sinclair, B., Sperling, S., Stupka, E., Sugiura, K., Sultana, R., Takenaka, Y., Taki, K., Tammoja, K., Tan, S. L., Tang, S., Taylor, M. S., Tegner, J., Teichmann, S. A., Ueda, H. R., van Nimwegen, E., Verardo, R., Wei, C. L., Yagi, K., Yamanishi, H., Zabarovsky, E., Zhu, S., Zimmer, A., Hide, W., Bult, C., Grimmond, S. M., Teasdale, R. D., Liu, E. T., Brusic, V., Quackenbush, J., Wahlestedt, C., Mattick, J. S., Hume, D. A., Kai, C., Sasaki, D., Tomaru, Y., Fukuda, S., Kanamori-Katayama, M., Suzuki, M., Aoki, J., Arakawa, T., Iida, J., Imamura, K., Itoh, M., Kato, T., Kawaji, H., Kawagashira, N., Kawashima, T., Kojima, M., Kondo, S., Konno, H., Nakano, K., Ninomiya, N., Nishio, T., Okada, M., Plessy, C., Shibata, K., Shiraki, T., Suzuki, S., Tagami, M., Waki, K., Watahiki, A., Okamura-Oho, Y., Suzuki, H., Kawai, J., Hayashizaki, Y., FANTOM Consortium, and RIKEN Genome Exploration Research Group and Genome Science Group (Genome Network Project Core Group). (2005). The transcriptional landscape of the mammalian genome. *Science* 309, 1559–1563.

Chen, C., Frierson, Jr. H. F., Haggerty, P. F., Theodorescu, D., Gregory, C. W., and Dong, J.-T. (2001). An 800-kb region of deletion at 13q14 in human prostate and other carcinomas. *Genomics* 77, 135–144.

Cheng, A. M., Byrom, M. W., Shelton, J., and Ford, L. P. (2005). Antisense inhibition of human miRNAs and indications for an involvement of miRNA in cell growth and apoptosis. *Nucleic Acids Res.* 33, 1290–1297.

Cimmino, A., Calin, G. A., Fabbri, M., Iorio, M. V., Ferracin, M., Shimizu, M., Wojcik, S. E., Aqeilan, R. I., Zupo, S., Dono, M., Rassenti, L., Alder, H., Volinia, S., Liu, C.-G., Kipps, T. J., Negrini, M., and Croce, C. M. (2005). miR-15 and miR-16 induce apoptosis by targeting BCL2. *Proc. Natl. Acad. Sci. U.S.A.* 102, 13944–13949.

Dou, L., Zheng, D., Li, J., Li, Y., Gao, L., Wang, L., and Yu, L. (2011). Methylation-mediated repression of microRNA-143 enhances MLL-AF4 oncogene expression. *Oncogene.* doi: 10.1038/onc.2011.248

Du, Q., Wang, L., Zhu, H., Zhang, S., Xu, L., Zheng, W., and Liu, X. (2010). The role of heterogeneous nuclear ribonucleoprotein K in the progression of chronic myeloid leukemia. *Med. Oncol.* 27, 673–679.

Esquela-Kerscher, A., and Slack, F. J. (2006). Oncomirs [mdash] microRNAs with a role in cancer. *Nat. Rev. Cancer* 6, 259–269.

Faraoni, I., Antonetti, F. R., Cardone, J., and Bonmassar, E. (2009). miR-155 gene: a typical multifunctional microRNA. *Biochim. Biophys. Acta* 1792, 497–505.

Garzon, R., Garofalo, M., Martelli, M. P., Briesewitz, R., Wang, L., Fernandez-Cymering, C., Volinia, S., Liu, C. G., Schnittger, S., Haferlach, T., Liso, A., Diverio, D., Mancini, M., Meloni, G., Foa, R., Martelli, M. F., Mecucci, C., Croce, C. M., and Falini, B. (2008a). Distinctive microRNA signature of acute myeloid leukemia bearing cytoplasmic mutated nucleophosmin. *Proc. Natl. Acad. Sci. U.S.A.* 105, 3945–3950.

Garzon, R., Volinia, S., Liu, C.-G., Fernandez-Cymering, C., Palumbo, T., Pichiorri, F., Fabbri, M., Coombes, K., Alder, H., Nakamura, T., Flomenberg, N., Marcucci, G., Calin, G. A., Kornblau, S. M., Kantarjian, H., Bloomfield, C. D., Andreeff, M., and Croce, C. M. (2008b). MicroRNA signatures associated with cytogenetics and prognosis in acute myeloid leukemia. *Blood* 111, 3183–3189.

Garzon, R., Heaphy, C. E. A., Havelange, V., Fabbri, M., Volinia, S., Tsao, T., Zanesi, N., Kornblau, S. M., Marcucci, G., Calin, G. A., Andreeff, M., and Croce, C. M. (2009a). MicroRNA 29b functions in acute myeloid leukemia. *Blood* 114, 5331–5341.

Garzon, R., Liu, S., Fabbri, M., Liu, Z., Heaphy, C. E. A., Callegari, E., Schwind, S., Pang, J., Yu, J., Muthusamy, N., Havelange, V., Volinia, S., Blum, W., Rush, L. J., Perrotti, D., Andreeff, M., Bloomfield, C. D., Byrd, J. C., Chan, K., Wu, L.-C., Croce, C. M., and Marcucci, G. (2009b). MicroRNA-29b induces global DNA hypomethylation and tumor suppressor gene reexpression in acute myeloid leukemia by targeting directly DNMT3A and 3B and indirectly DNMT1. *Blood* 113, 6411–6418.

Gibb, E. A., Brown, C. J., and Lam, W. L. (2011). The functional role of long non-coding RNA in human carcinomas. *Mol. Cancer* 10, 38.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286, 531–537.

Group, R. G. E. R., Genome Science, G., The, F. C., Katayama, S., Tomaru, Y., Kasukawa, T., Waki, K., Nakanishi, M., Nakamura, M., Nishida, H., Yap, C. C., Suzuki, M., Kawai, J., Suzuki, H., Carninci, P., Hayashizaki, Y., Wells, C., Frith, M., Ravasi, T., Pang, K. C., Hallinan, J., Mattick, J., Hume, D. A., Lipovich, L., Batalov, S., Engström, P. G., Mizuno, Y., Faghihi, M. A., Sandelin, A., Chalk, A. M., Mottagui-Tabar, S., Liang, Z., Lenhard, B., and Wahlestedt, C. (2005). Antisense Transcription in the mammalian transcriptome. *Science* 309, 1564–1566.

Guttman, M., Amit, I., Garber, M., French, C., Lin, M. F., Feldser, D., Huarte, M., Zuk, O., Carey, B. W., Cassady, J. P., Cabili, M. N., Jaenisch, R., Mikkelsen, T. S., Jacks, T., Hacohen, N., Bernstein, B. E., Kellis, M., Regev, A., Rinn, J. L., and Lander, E. S. (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* 458, 223–227.

Hayashita, Y., Osada, H., Tatematsu, Y., Yamada, H., Yanagisawa, K., Tomida, S., Yatabe, Y., Kawahara, K., Sekido, Y., and Takahashi, T. (2005). A polycistronic microRNA cluster, miR-17-92, is overexpressed in human lung cancers and enhances cell proliferation. *Cancer Res.* 65, 9628–9632.

Heerema-McKenney, A., and Arber, D. A. (2009). Acute myeloid leukemia. *Hematol. Oncol. Clin. North Am.* 23, 633–654.

Huarte, M., Guttman, M., Feldser, D., Garber, M., Koziol, M. J., Kenzelmann-Broz, D., Khalil, A. M., Zuk, O., Amit, I., Rabani, M., Attardi, L. D., Regev, A., Lander, E. S., Jacks, T., and Rinn, J. L. (2010). A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. *Cell* 142, 409–419.

Huarte, M., and Rinn, J. L. (2010). Large non-coding RNAs: missing links in cancer? *Hum. Mol. Genet.* 19, R152–R161.

Iacobucci, I., Sazzini, M., Garagnani, P., Ferrari, A., Boattini, A., Lonetti, A., Papayannidis, C., Mantovani, V., Marasco, E., Ottaviani, E., Soverini, S., Girelli, D., Luiselli, D., Vignetti, M., Baccarani, M., and Martinelli, G. (2011). A polymorphism in the chromosome 9p21 ANRIL locus is associated to Philadelphia positive acute lymphoblastic leukemia. *Leuk. Res.* 35, 1052–1059.

Ivanovska, I., Ball, A. S., Diaz, R. L., Magnus, J. F., Kibukawa, M., Schelter, J. M., Kobayashi, S. V., Lim, L., Burchard, J., Jackson, A. L., Linsley, P. S., and Cleary, M. A. (2008). MicroRNAs in the miR-106b family regulate p21/CDKN1A and promote cell cycle progression. *Mol. Cell. Biol.* 28, 2167–2174.

Jose-Eneriz, S. E., Roman-Gomez, J., Jimenez-Velasco, A., Garate, L., Martin, V., Cordeu, L., Vilas-Zornoza, A., Rodriguez-Otero, P., Calasanz, M. J., Prosper, F., and Agirre, X. (2009). MicroRNA expression profiling in imatinib-resistant chronic myeloid leukemia patients without clinically significant ABL1-mutations. *Mol. Cancer* 8, 69.

Kagami, M., O'sullivan, M. J., Green, A. J., Watabe, Y., Arisaka, O., Masawa, N., Matsuoka, K., Fukami, M., Matsubara, K., Kato, F., Ferguson-Smith, A. C., and Ogata, T. (2010). The IG-DMR and the *MEG3*-DMR at human chromosome 14q32.2: hierarchical interaction and distinct functional properties as imprinting control centers. *PLoS Genet* 6, e1000992. doi:10.1371/journal.pgen.1000992

Kantarjian, H., Sawyers, C., Hochhaus, A., Guilhot, F., Schiffer, C., Gambacorti-Passerini, C., Niederwieser, D., Resta, D., Capdeville, R., Zoellner, U., Talpaz, M., and Druker, B. (2002). Hematologic and cytogenetic responses to imatinib mesylate in chronic myelogenous leukemia. *N. Engl. J. Med.* 346, 645–652.

Kapranov, P., St Laurent, G., Raz, T., Ozsolak, F., Reynolds, C. P., Sorensen, P. H., Reaman, G., Milos, P., Arceci, R. J., Thompson, J. F., and Triche, T. J. (2010). The majority of total nuclear-encoded non-ribosomal RNA in a human cell is "dark matter" unannotated RNA. *BMC Biol.* 8, 149. doi:10.1186/1741-7007-8-149

Khoury, H., Suarez-Saiz, F., Wu, S., and Minden, M. D. (2010). An upstream insulator regulates DLK1 imprinting in AML. *Blood* 115, 2260–2263.

Kim, W. Y., and Sharpless, N. E. (2006). The regulation of INK4/ARF in cancer and aging. *Cell* 127, 265–275.

Kivelä, T., Tuppurainen, K., Riikonen, P., and Vapalahti, M. (2003). Retinoblastoma associated with chromosomal 13q14 deletion mosaicism. *Ophthalmology* 110, 1983–1988.

Klein, U., Lia, M., Crespo, M., Siegel, R., Shen, Q., Mo, T., Ambesi-Impiombato, A., Califano, A., Migliazza, A., Bhagat, G., and Dalla-Favera, R. (2010). The DLEU2/miR-15a/16-1 Cluster controls B cell proliferation and its deletion leads to chronic lymphocytic leukemia. *Cancer Cell* 17, 28–40.

Kotake, Y., Nakagawa, T., Kitagawa, K., Suzuki, S., Liu, N., Kitagawa, M., and Xiong, Y. (2011). Long non-coding RNA ANRIL is required for the PRC2 recruitment to and silencing of p15INK4B tumor suppressor gene. *Oncogene* 30, 1956–1962.

Lee, R. C., and Ambros, V. (2001). An Extensive Class of Small RNAs in *Caenorhabditis elegans. Science* 294, 862–864.

Lerner, M., Harada, M., Lovén, J., Castro, J., Davis, Z., Oscier, D., Henriksson, M., Sangfelt, O., Grandér, D., and Corcoran, M. M. (2009). DLEU2, frequently deleted in malignancy, functions as a critical host gene of the cell cycle inhibitory microRNAs miR-15a and miR-16-1. *Exp. Cell Res.* 315, 2941–2952.

Li, S., Moffett, H. F., Lu, J., Werner, L., Zhang, H., Ritz, J., Neuberg, D., Wucherpfennig, K. W., Brown, J. R., and Novina, C. D. (2011). MicroRNA expression profiling identifies activated B cell status in chronic lymphocytic leukemia cells. *PLoS ONE* 6, e16956. doi:10.1371/journal.pone.0016956

Li, Z., Lu, J., Sun, M., Mi, S., Zhang, H., Luo, R. T., Chen, P., Wang, Y., Yan, M., Qian, Z., Neilly, M. B., Jin, J., Zhang, Y., Bohlander, S. K., Zhang, D.-E., Larson, R. A., Le Beau, M. M., Thirman, M. J., Golub, T. R., Rowley, J. D., and Chen, J. (2008). Distinct microRNA expression profiles in acute myeloid leukemia with common translocations. *Proc. Natl. Acad. Sci. U.S.A.* 105, 15535–15540.

Liu, Y., Corcoran, M., Rasool, O., Ivanova, G., Ibbotson, R., Grander, D., Iyengar, A., Baranova, A., Kashuba, V., Merup, M., Wu, X., Gardiner, A., Mullenbach, R., Poltaraus, A., Hultstrom, A. L., Juliusson, G., Chapman, R., Tiller, M., Cotter, F., Gahrton, G., Yankovsky, N., Zabarovsky, E., Einhorn, S., and Oscier, D. (1997). Cloning of two candidate tumor suppressor genes within a 10 kb region on chromosome 13q14, frequently deleted in chronic lymphocytic leukemia. *Oncogene* 15, 2463–2473.

Lujambio, A., Portela, A., Liz, J., Melo, S. A., Rossi, S., Spizzo, R., Croce, C. M., Calin, G. A., and Esteller, M. (2010). CpG island hypermethylation-associated silencing of non-coding RNAs transcribed from ultraconserved regions in human cancer. *Oncogene* 29, 6390–6401.

Marcucci, G., Radmacher, M. D., Maharry, K., MróZek, K., Ruppert, A. S., Paschka, P., Vukosavljevic, T., Whitman, S. P., Baldus, C. D., Langer, C., Liu, C.-G., Carroll, A. J., Powell, B. L., Garzon, R., Croce, C. M., Kolitz, J. E., Caligiuri, M. A., Larson, R. A., and Bloomfield, C. D. (2008). MicroRNA expression in cytogenetically normal acute myeloid leukemia. *N. Engl. J. Med.* 358, 1919–1928.

Margueron, R., and Reinberg, D. (2011). The Polycomb complex PRC2 and its mark in life. *Nature* 469, 343–349.

Mestdagh, P., Fredlund, E., Pattyn, F., Rihani, A., Van Maerken, T., Vermeulen, J., Kumps, C., Menten, B., De Preter, K., Schramm, A., Schulte, J., Noguera, R., Schleiermacher, G., Janoueix-Lerosey, I., Laureys, G., Powel, R., Nittner, D., Marine, J. C., Ringner, M., Speleman, F., and Vandesompele, J. (2010). An integrative genomics screen uncovers ncRNA T-UCR functions in neuroblastoma tumours. *Oncogene* 29, 3583–3592.

Mi, S., Lu, J., Sun, M., Li, Z., Zhang, H., Neilly, M. B., Wang, Y., Qian, Z., Jin, J., Zhang, Y., Bohlander, S. K., Le Beau, M. M., Larson, R. A., Golub, T. R., Rowley, J. D., and Chen, J. (2007). MicroRNA expression signatures accurately discriminate acute lymphoblastic leukemia from acute myeloid leukemia. *Proc. Natl. Acad. Sci. U.S.A.* 104, 19971–19976.

Migliazza, A., Bosch, F., Komatsu, H., Cayanis, E., Martinotti, S., Toniato, E., Guccione, E., Qu, X., Chien, M., Murty, V. V. V., Gaidano, G., Inghirami, G., Zhang, P., Fischer, S., Kalachikov, S. M., Russo, J., Edelman, I., Efstratiadis, A., and Dalla-Favera, R. (2001). Nucleotide sequence, transcription map, and mutation analysis of the 13q14 chromosomal region deleted in B-cell chronic lymphocytic leukemia. *Blood* 97, 2098–2104.

Mishra, P. J., Song, B., Mishra, P. J., Wang, Y., Humeniuk, R., Banerjee, D., Merlino, G., Ju, J., and Bertino, J. R. (2009). *MiR-24* Tumor suppressor activity is regulated independent of p53 and through a target site polymorphism. *PLoS ONE* 4, e8445. doi:10.1371/journal.pone.0008445

Notari, M., Neviani, P., Santhanam, R., Blaser, B. W., Chang, J.-S., Galietta, A., Willis, A. E., Roy, D. C., Caligiuri, M. A., Marcucci, G., and Perrotti, D. (2006). A MAPK/HNRPK pathway controls BCR/ABL oncogenic potential by regulating MYC mRNA translation. *Blood* 107, 2507–2516.

Ørom, U. A., Derrien, T., Beringer, M., Gumireddy, K., Gardini, A., Bussotti, G., Lai, F., Zytnicki, M., Notredame, C., Huang, Q., Guigo, R., and Shiekhattar, R. (2010). Long noncoding RNAs with enhancer-like function in human cells. *Cell* 143, 46–58.

Ørom, U. A., Nielsen, F. C., and Lund, A. H. (2008). MicroRNA-10a binds the 52UTR of ribosomal protein mRNAs and enhances their translation. *Mol. Cell* 30, 460–471.

Pasmant, E., Laurendeau, I., Héron, D., Vidaud, M., Vidaud, D., and Bièche, I. (2007). Characterization of a germ-line deletion, including the entire INK4/ARF locus, in a melanoma-neural system tumor family: identification of ANRIL, an antisense noncoding RNA whose expression coclusters with ARF. *Cancer Res.* 67, 3963–3969.

Pekarsky, Y., Santanam, U., Cimmino, A., Palamarchuk, A., Efanov, A., Maximov, V., Volinia, S., Alder, H., Liu, C.-G., Rassenti, L., Calin, G. A., Hagan, J. P., Kipps, T., and Croce, C. M. (2006). Tcl1 expression in chronic lymphocytic leukemia is regulated by miR-29 and miR-181. *Cancer Res.* 66, 11590–11593.

San Jose-Eneriz, E., Roman-Gomez, J., Jimenez-Velasco, A., Garate, L., Martin, V., Cordeu, L., Vilas-Zornoza, A., Rodriguez-Otero, P., Calasanz, M. J., Prosper, F., and Agirre, X. (2009). MicroRNA expression profiling in imatinib-resistant chronic myeloid leukemia patients without clinically significant ABL1-mutations. *Mol. Cancer* 8, 69.

Sherr, C. J. (1998). Tumor surveillance via the ARF-p53 pathway. *Genes Dev.* 12, 2984–2991.

Tay, Y., Zhang, J., Thomson, A. M., Lim, B., and Rigoutsos, I. (2008). MicroRNAs to Nanog, Oct4 and Sox2 coding regions modulate embryonic stem cell differentiation. *Nature* 455, 1124–1128.

Valk, P. J. M., Verhaak, R. G. W., Beijen, M. A., Erpelinck, C. A. J., Van Doorn-Khosrovani, S. B. V. W., Boer, J. M., Beverloo, H. B., Moorhouse, M. J., Van Der Spek, P. J., Lã Wenberg, B., and Delwel, R. (2004). Prognostically useful gene-expression profiles in acute myeloid leukemia. *N. Engl. J. Med.* 350, 1617–1628.

Wagner, L. A., Christensen, C. J., Dunn, D. M., Spangrude, G. J., Georgelas, A., Kelley, L., Esplin, M. S., Weiss, R. B., and Gleich, G. J. (2007). EGO, a novel, noncoding RNA gene, regulates eosinophil granule protein transcript expression. *Blood* 109, 5191–5198.

Yap, K. L., Li, S., Muñoz-Cabello, A. M., Raguz, S., Zeng, L., Mujtaba, S., Gil, J., Walsh, M. J., and Zhou, M.-M. (2010). Molecular interplay of the noncoding RNA ANRIL and methylated histone H3 lysine 27 by polycomb CBX7 in transcriptional silencing of INK4a. *Mol. Cell* 38, 662–674.

Yu, W., Gius, D., Onyango, P., Muldoon-Jacobs, K., Karp, J., Feinberg, A. P., and Cui, H. (2008). Epigenetic silencing of tumour suppressor gene p15 by its antisense RNA. *Nature* 451, 202–206.

Zhang, B., Pan, X., Cobb, G. P., and Anderson, T. A. (2007). microRNAs as oncogenes and tumor suppressors. *Dev. Biol.* 302, 1–12.

Zhang, H., Luo, X.-Q., Zhang, P., Huang, L.-B., Zheng, Y.-S., Wu, J., Zhou, H., Qu, L.-H., Xu, L., and Chen, Y.-Q. (2009a). MicroRNA patterns associated with clinical prognostic parameters and cns relapse prediction in pediatric acute leukemia. *PLoS ONE* 4, e7826. doi:10.1371/journal.pone.0007826

Zhang, X., Lian, Z., Padden, C., Gerstein, M. B., Rozowsky, J., Snyder, M., Gingeras, T. R., Kapranov, P., Weissman, S. M., and Newburger, P. E. (2009b). A myelopoiesis-associated regulatory intergenic noncoding RNA transcript within the human HOXA cluster. *Blood* 113, 2526–2534.

Zhang, X., Rice, K., Wang, Y., Chen, W., Zhong, Y., Nakayama, Y., Zhou, Y., and Klibanski, A. (2010a). Maternally expressed gene 3 (MEG3) noncoding ribonucleic acid: isoform structure, expression, and functions. *Endocrinology* 151, 939–947.

Zhang, X., Zhou, Y., and Klibanski, A. (2010b). Isolation and characterization of novel pituitary tumor related genes: a cDNA representational difference approach. *Mol. Cell. Endocrinol.* 326, 40–47.

Zhang, X., Zhou, Y., Mehta, K. R., Danila, D. C., Scolavino, S., Johnson, S. R., and Klibanski, A. (2003). A pituitary-derived MEG3 isoform functions as a growth suppressor in tumor cells. *J. Clin. Endocrinol. Metab.* 88, 5119–5126.

Zhou, Y., Zhong, Y., Wang, Y., Zhang, X., Batista, D. L., Gejman, R., Ansell, P. J., Zhao, J., Weng, C., and Klibanski, A. (2007). activation of p53 by MEG3 non-coding RNA. *J. Biol. Chem.* 282, 24731–24742.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 01 October 2011; paper pending published: 31 October 2011; accepted: 07 December 2011; published online: 28 December 2011.

Citation: Heuston EF, Lemon KT and Arceci RJ (2011) The beginning of the road for non-coding RNAs in normal hematopoiesis and hematologic malignancies. Front. Gene. ***2****:94. doi: 10.3389/fgene.2011.00094*

This article was submitted to Frontiers in Non-Coding RNA, a specialty of Frontiers in Genetics.

frontiers in GENETICS

ORIGINAL RESEARCH ARTICLE
published: 22 February 2012
doi: 10.3389/fgene.2012.00023

Genomic "dark matter" in prostate cancer: exploring the clinical utility of ncRNA as biomarkers

Ismael A. Vergara[1†], Nicholas Erho[1†], Timothy J. Triche[1]*, Mercedeh Ghadessi[1], Anamaria Crisan[1], Thomas Sierocinski[1], Peter C. Black[2], Christine Buerki[1] and Elai Davicioni[1]

[1] GenomeDx Biosciences, Inc., Vancouver, BC, Canada
[2] Department of Urologic Sciences, University of British Columbia, Vancouver, BC, Canada

Edited by:
Philipp Kapranov, St. Laurent Institute, USA

Reviewed by:
David Ting, Massachusetts General Hospital Cancer Center, USA
Robert Arceci, Johns Hopkins, USA
Eduardo M. Reis, Universidade de Sao Paulo, Brazil

***Correspondence:**
Timothy J. Triche, GenomeDx Biosciences, Inc., 201-1595 West 3rd Avenue, Vancouver, BC, Canada V6J 1J8.
e-mail: tim@genomedx.com

† Ismael A. Vergara and Nicholas Erho have contributed equally to this work.

Prostate cancer is the most diagnosed cancer among men in the United States. While the majority of patients who undergo surgery (prostatectomy) will essentially be cured, about 30–40% men remain at risk for disease progression and recurrence. Currently, patients are deemed at risk by evaluation of clinical factors, but these do not resolve whether adjuvant therapy will significantly attenuate or delay disease progression for a patient at risk. Numerous efforts using mRNA-based biomarkers have been described for this purpose, but none have successfully reached widespread clinical practice in helping to make an adjuvant therapy decision. Here, we assess the utility of non-coding RNAs as biomarkers for prostate cancer recurrence based on high-resolution oligonucleotide microarray analysis of surgical tissue specimens from normal adjacent prostate, primary tumors, and metastases. We identify differentially expressed non-coding RNAs that distinguish between the different prostate tissue types and show that these non-coding RNAs can predict clinical outcomes in primary tumors. Together, these results suggest that non-coding RNAs are emerging from the "dark matter" of the genome as a new source of biomarkers for characterizing disease recurrence and progression. While this study shows that non-coding RNA biomarkers can be highly informative, future studies will be needed to further characterize the specific roles of these non-coding RNA biomarkers in the development of aggressive disease.

Keywords: prostate cancer, prognosis, microarrays, clinical progression, non-coding RNA

INTRODUCTION

Prostate cancer is a major public health concern, with over 240,000 newly diagnosed men in the United States alone (Siegel et al., 2011). This clinically heterogeneous disease ranges from indolent forms of cancer with good long term prognosis to life-threatening disease associated with only a couple of months of survival (Rubin et al., 2011). After initial diagnosis, one of the most successful treatments with curative intent is radical prostatectomy, i.e., the complete removal of the prostate gland. It is, however, known that patients who present with aggressive clinical features after surgery, such as positive surgical margins (SM), extracapsular extension (ECE), and seminal vesicle invasion (SVI) likely will require further therapy in order to delay the onset of life-threatening metastasis (Bolla et al., 2005; Thompson et al., 2009; Wiegel et al., 2009). The efficient delivery of such therapies after prostatectomy is currently hampered by a lack of predictive tools to assess the risk of clinically significant recurrence and progression.

Biochemical recurrence (BCR), defined as a detectable prostate specific antigen (PSA) level above a certain threshold or as a rising PSA level after surgery, is a widely used surrogate for disease progression and prostate cancer specific mortality (PCSM). Still, BCR has been deemed an unreliable surrogate since, even though BCR always precedes metastatic progression and PCSM, not every patient with BCR will experience metastatic disease (Simmons et al., 2007). Given this, numerous efforts using mRNA-based biomarkers as a tool to assess the risk of recurrence and progression have been described, but none have successfully reached widespread clinical practice (Sorensen and Orntoft, 2010). Recently, the clinical utility of micro RNAs (or miRNAs) as potential biomarkers for disease diagnosis and prognosis has been assessed (Schaefer et al., 2010; Sevli et al., 2010; Catto et al., 2011; Martens-Uzunova et al., 2011). miRNAs have shown altered expression in prostate cancer and were found to be involved in the regulation of key pathways such as androgen signaling and apoptosis (Catto et al., 2011). In general, recent evidence showing that a much larger fraction of normal and cancer transcriptomes are composed of non-coding RNAs (or ncRNAs) than previously anticipated (Kapranov et al., 2010) has driven researchers towards exploring the utility of not only short ncRNAs but also long ncRNAs as biomarkers. For example, Chung et al. (2011) identified *PRNCR1* (prostate cancer non-coding RNA 1) as a long intergenic ncRNA (or lincRNA) transcribed in the gene desert of the prostate cancer susceptibility locus 8q24. The same genomic region was found to be transcribed into *PCAT-1*, a lincRNA highly expressed in metastatic tissue specimens from prostate cancer patients (Prensner et al., 2011).

While there is increasing knowledge of the importance of ncRNAs in cancer, their clinical usefulness for diagnosis and prognosis is limited. To date, only one ncRNA is routinely used in the clinical setting in prostate cancer: prostate cancer antigen 3 (*PCA3*), a non-coding antisense transcript that is highly overexpressed in prostate cancer compared to benign tissue (Bussemakers et al., 1999). *PCA3* is used in a urinary-based diagnostic test for patient screening in conjunction with PSA serum testing and other clinical information (Day et al., 2011).

In this study, we perform high-resolution oligonucleotide microarray analysis of a publicly available dataset (Taylor et al., 2010) from different types of normal and cancerous prostate tissue. We find, by analysis of the entire set of exonic and non-exonic features, differentially expressed ncRNAs that accurately discriminate clinical outcomes such as BCR and metastatic disease.

MATERIALS AND METHODS

MICROARRAY AND CLINICAL DATA

The publically available genomic and clinical data was generated as part of the Memorial Sloan–Kettering Cancer Center (MSKCC) Prostate Oncogenome Project, previously reported by (Taylor et al., 2010). The Human Exon arrays for 131 primary prostate cancer, 29 normal adjacent, and 19 metastatic tissue specimens were downloaded from GEO Omnibus at http://www.ncbi.nlm.nih.gov/geoseries GSE21034. The patient and specimen details for the primary and metastases tissues used in this study are summarized in **Table 1**. For the analysis of the clinical data, the following ECE statuses were summarized to be concordant with the pathological tumor stage: inv-capsule: ECE−, focal: ECE+, established: ECE+.

MICROARRAY PRE-PROCESSING

Normalization and summarization

The normalization and summarization of the 179 microarray samples (cell line samples were removed) were done with the frozen Robust Multiarray Average (fRMA) algorithm using custom frozen vectors (McCall et al., 2010). These custom vectors were created using the vector creation methods described in (McCall and Irizarry, 2011) including all MSKCC samples. Quantile normalization and robust weighted average methods were used for normalization and summarization, respectively, as implemented in fRMA.

Table 1 | Summary of the clinical characteristics of the dataset used in this study.

	Primary tumor	Metastasis
N	131	19
Median age at Dx (years)	58	58
PRE-OP PSA (ng/ml)		
<10	108	7
≥10 < 20	16	1
≥20	6	9
NA	1	2
PATHOLOGICAL GLEASON SCORE		
≤6	41	0
7	74	2
≥8	15	7
NA	1	10
PATHOLOGICAL STAGE		
T2	85	1
T3	40	7
T4	6	2
NA	0	9

Sample subsets

The normalized and summarized data was partitioned into three groups. The first group contains the matched samples from primary localized prostate cancer tumors and normal adjacent tissues ($n = 58$; used for the normal vs. primary comparison). The second group contains all the samples from metastatic tumors ($n = 19$) and all the localized prostate cancer tumors that were not matched with normal adjacent tissues ($n = 102$; used for the primary tumor vs. metastasis comparison). The third group corresponds to all samples from metastatic tumors ($n = 19$) and all the normal adjacent tissues ($n = 29$; used for the normal vs. metastasis comparison).

Feature selection

Probe sets (or PSRs) annotated as "unreliable" by the xmapcore package (Yates, 2010; defined as one or more probes that do not align uniquely to the genome) as well as those defined as class 2 and class 3 cross-hybridizing by Affymetrix annotation were excluded from further analysis. The remaining PSRs were subjected to univariate analysis to identify those associated to features differentially expressed between the labeled groups (primary tumor vs. metastatic, normal adjacent vs. primary tumor, and normal vs. metastatic). For this analysis, features were selected as differentially expressed if their Holm–Bonferroni adjusted (Holm, 1979) t-test p-value was significant ($<$0.05). The t-test was applied as implemented in the *rowttests* function of the genefilter package.[1]

The multiple testing correction was applied using the *p.adjust* function of the stats package in R.

This multiple testing correction was performed for the exonic (353k PSRs) and non-exonic (931k PSRs) sets independently due to differences in cardinality of the PSR sets. Data A1 in Appendix provides the detailed steps for the generation of differentially expressed features.

Feature evaluation and model building

Classical multidimensional scaling (MDS, Pearson's distance) was used to evaluate the ability of the selected features to segregate primary tumor samples into clinically relevant clusters based on metastatic events and Gleason scores. MDS was applied as implemented in the *cmdscale* function of the stats package in R. The significance of the segregation in these two-dimensional MDS plots was assessed using permutational ANOVA as implemented within the vegan package in R[2].

A custom implementation of the k-nearest-neighbor (KNN) model ($k = 1$, Pearson's correlation distance metric) was trained on the normal and metastatic samples ($n = 48$) using only the features found to be differentially expressed between these two groups. Unmatched primary tumors were used as an independent set for validation.

[1] http://www.bioconductor.org/packages/2.3/bioc/html/genefilter.html

[2] http://cran.r-project.org/web/packages/vegan/index.html

Re-annotation of the human exon microarray probe sets

Affymetrix Human Exon 1.0 ST Arrays[3] have about 1.4 million probe sets, with most probe sets containing four probes each. In order to properly assess the nature of the probe sets found differentially expressed in this study, we re-annotated them using the xmapcore R package[4] (Yates, 2010) as follows: (i) exonic, if the probe set overlaps with the coding portion of a protein-coding exon or an untranslated region (UTR), and (ii) non-exonic if the probe set overlaps with an intron, an intergenic region, or a non-protein-coding transcript.

Annotation of non-coding transcripts was pursued using Ensembl Biomart available at http://www.ensembl.org

STATISTICAL ANALYSIS

Biochemical recurrence and metastatic disease progression end points are used as defined by the "BCR Event" and "Mets Event" columns of the supplementary material provided by (Taylor et al., 2010), respectively. Survival analysis for BCR was performed using the *survfit* function of the survival package[5]. Logistic regression for metastatic disease progression was performed using the *lrm* function of the rms package[6].

RESULTS

RE-ANNOTATION AND CATEGORIZATION OF CODING AND NON-CODING DIFFERENTIALLY EXPRESSED FEATURES

Previous transcriptome-wide assessments of differential expression using prostate tissues in the post-prostatectomy setting have been focused on protein-coding features (see Nakagawa et al., 2008 for a comparison of protein-coding gene-based panels). Recent evidence based on the characterization of transcriptomes from normal and cancerous tissues has shown that most of it is of non-coding nature (Kapranov et al., 2010). Human Exon Arrays provide a unique opportunity to explore the differential expression of non-coding parts of the genome, as 75% of their probe sets cover regions other than protein-coding sequences. In this study, we use the publicly available Human Exon Array data set from normal adjacent, localized primary tumors, and metastatic tissues generated as part of the MSKCC Prostate Oncogenome Project to explore the potential of non-coding regions in prostate cancer prognosis. Previous attempts on this dataset focused only on mRNA and gene-level analysis and concluded that expression analysis was inadequate for discrimination of outcome groups in primary tumors (Taylor et al., 2010). In order to assess the contribution of ncRNA probe sets in differential expression analysis between sample types, we re-assessed the annotation of all probe sets found to be differentially expressed according to their genomic location and categorized them into exonic and non-exonic (see Materials and Methods). Briefly, a probe set is classified as exonic if it falls in a region that encodes for a protein-coding transcript or an UTR; otherwise, it is annotated as non-exonic.

Based on the above categorization, we assessed the exonic and non-exonic sets for the presence of differentially expressed features for each possible pairwise comparison (i.e., primary vs. normal, normal vs. metastatic, and primary vs. metastatic). The majority of the differentially expressed features are labeled as exonic for a given pairwise comparison (81%, 81%, and 75% for normal-primary, primary-metastatic, and normal-metastatic comparisons, respectively; see Table S1 in Supplementary Material for the top 100 differentially expressed features for each pairwise comparison). For each category, the number of differentially expressed features is highest in normal vs. metastatic tissues, which is expected since the metastatic samples are a heterogeneous group that has likely undergone major genomic alterations through disease progression and through effects of therapy on the genome (**Figure 1**). Additional variation in expression may be due to contamination with metastatic site tissue as well as host cell-metastatic cell interactions for metastases that include distant lymph nodes (seven samples), bone (five samples), and brain (three samples). As expected, assessment of all gene loci with features found to be differentially expressed between normal and metastatic samples shows that those up-regulated in metastatic tissue compared to normal are enriched in cellular processes such as cell division, spindle check point, and cytokinesis, whereas those down-regulated are enriched in terms like cell adhesion, muscle contraction, neuron development, and urogenital system development (Table S2 in Supplementary Material).

For each category of exonic and non-exonic features there is a significant number that are specific to each pairwise comparison. For example, 21% of the exonic features are specific to the differentiation between normal tissue and primary tumors and 10% are specific to the primary tumor vs. metastatic comparison. The same proportions are observed for the non-exonic category, suggesting that different genomic regions may play a role in the progression from normal tissue to primary tumor and from primary tumor to metastatic tumor.

Within the non-exonic category, the majority of the features are "intronic" for all pairwise comparisons (see **Figures 2A–C**). Also, a large proportion of features correspond to intergenic regions. Still, hundreds of features lie within non-coding transcripts, as reflected by the "NC Transcript" segment in **Figure 2**. These non-coding transcripts found to be differentially expressed in each pairwise comparison were categorized using the "Transcript Biotype" annotation of Ensembl. For all pairwise comparisons the "processed transcript", "lincRNA", "retained intron", and "antisense" are the most prevalent (**Figures 2D–F**; see **Table 2** for a definition of each transcript type). Even though "processed transcript" and "retained intron" categories are among the most frequent ones, they have a very broad definition.

Previous studies have reported several long non-coding RNAs to play a role in prostate adenocarcinoma (Srikantan et al., 2000; Berteaux et al., 2004; Petrovics et al., 2004; Lin et al., 2007; Poliseno et al., 2010; Yap et al., 2010; Chung et al., 2011; Day et al., 2011). Close inspection of our data reveals that four of them (*PCGEM1*, *PCA3*, *MALAT1*, and *PTENP1*) have associated differentially expressed features in at least one pairwise comparison based on a 1.5 Median Fold Difference (MFD) threshold (**Table 3**). After adjusting the *p*-value for multiple testing however, only two ncRNA transcripts, *PCA3* and *MALAT1*, remain significant (**Table 3**). In addition, we found three differentially expressed microRNA-encoding transcripts in primary tumor vs. metastatic (*MIR143*, *MIR145*, and *MIR221*) and two in normal

[3] www.affymetrix.com/
[4] http://www.bioconductor.org/packages/2.6/bioc/html/xmapcore.html
[5] http://cran.r-project.org/web/packages/survival/index.html
[6] http://biostat.mc.vanderbilt.edu/twiki/bin/view/Main/RmS

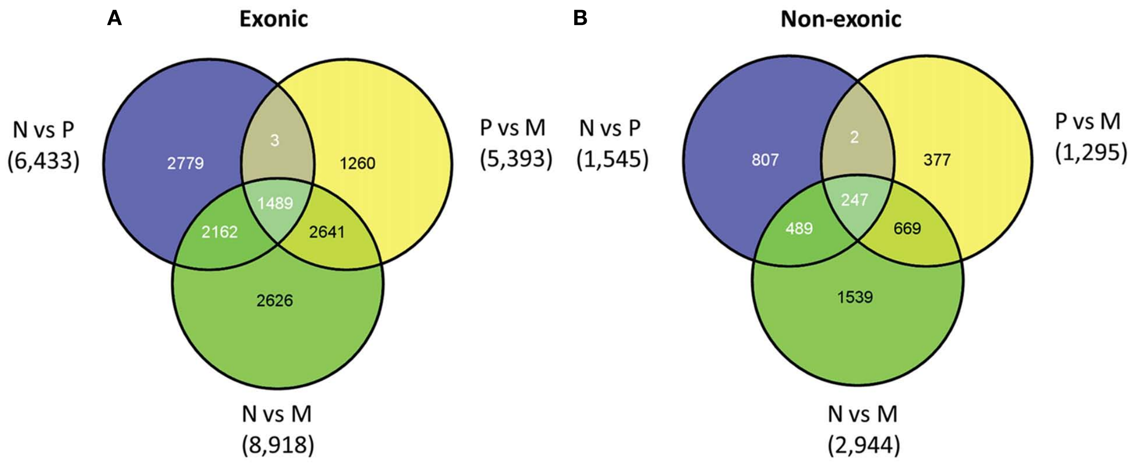


FIGURE 1 | Venn diagram of exonic (A) and non-exonic (B) features found differentially expressed in the following comparisons: normal vs. primary tumor tissue (N vs. P), primary tumor vs. metastatic tissue (P vs. M), and normal vs. metastatic tissue (N vs. M).

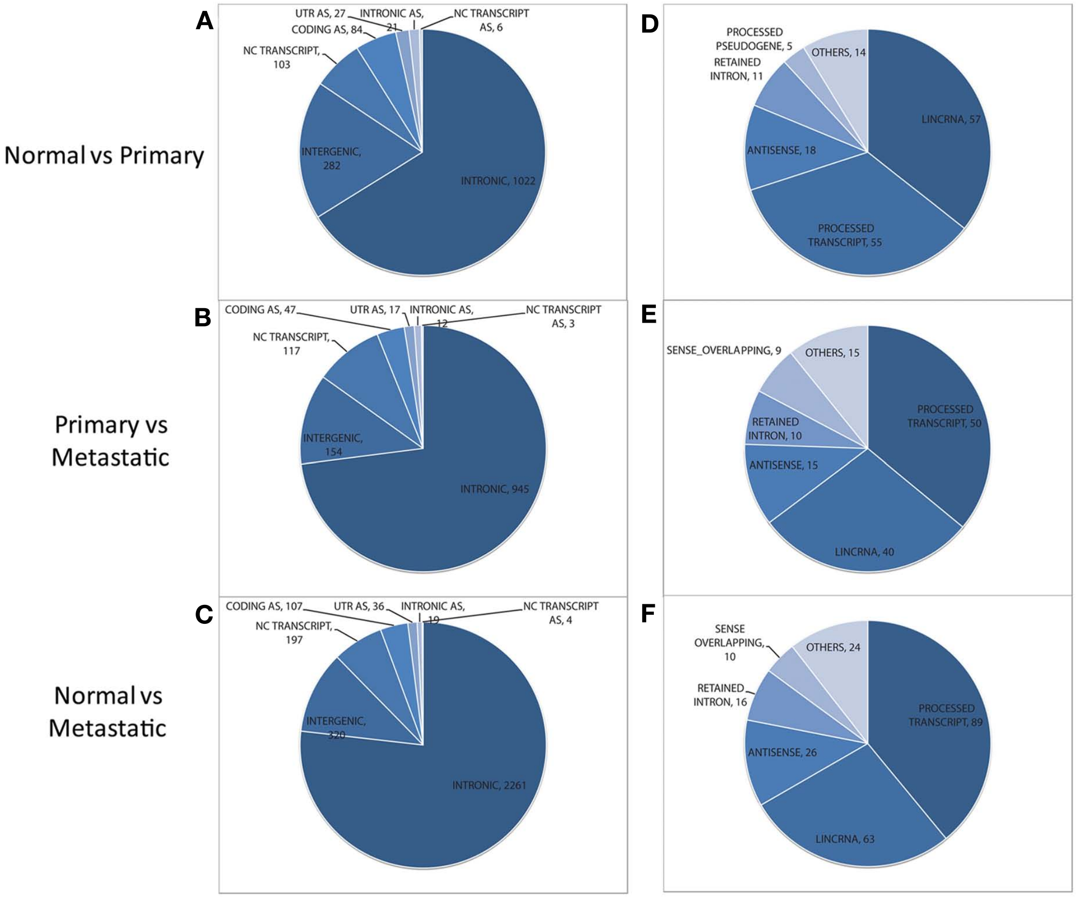


FIGURE 2 | Distribution of non-exonic features (left) and overlapping annotated non-coding transcripts (right) found to be differentially expressed between normal and primary tumor (A,D), primary tumor and metastatic tissue (B,E), and normal vs. metastatic tissue (C,F). Features in the NC TRANSCRIPT slice of each pie chart (left) are assessed for their overlap with non-coding transcripts to generate the distribution of transcripts (shown at the right for each pairwise comparison). AS, antisense. UTR, untranslated region; lincRNA, long intergenic ncRNA.

Table 2 | Definitions of Ensembl "Transcript Biotype" annotations for non-coding transcripts found differentially expressed.

Name	Definition
Processed transcript	Non-coding transcript that does not contain an ORF
Retained intron	Non-coding transcript containing intronic sequence
LincRNA	Large intergenic non-coding RNA, or long non-coding RNA, usually associated with open chromatin signatures such as histone modification sites
Antisense	Non-coding transcript believed to be an antisense product used in the regulation of the gene to which it belongs
Sense overlapping	Has a long non-coding transcript that contains a coding gene in its intron on the same strand
Processed pseudogene	Non-coding pseudogene produced by integration of a reverse transcribed mRNA into the genome

Table 3 | Long non-coding RNAs previously reported as differentially expressed in prostate cancer.

Gene type	Gene	Probe set ID	Comparison	*t*-Test *p*-value	MFD ratio	Reference
LncRNA	*ANRIL*	3165014	Primary vs. normal	<0.01	−1.17	Yap et al. (2010)
		3165015	Metastatic vs. normal	<0.01	1.49	
		3165015	Metastatic vs. primary	<0.02	1.33	
	H19	3359101	Primary vs. normal	<0.01	1.32	Berteaux et al. (2004)
		3359097	Metastatic vs. normal	<0.01	1.43	
		3359095	Metastatic vs. primary	<0.01	−1.12	
	PCA3	3175541	Primary vs. normal	<0.01	14.3	Bussemakers et al. (1999)
		3175545	Metastatic vs. normal	<0.01	4.46	
		3175541	Metastatic vs. primary	<0.01	−3.44	
	MALAT1	3335195	Primary vs. normal	<0.01	−1.64	Lin et al. (2007)
		3335195	Metastatic vs. normal	<0.01	−3.33	
		3335195	Metastatic vs. primary	<0.01	−2.63	
	PCGEM1	2520747	Primary vs. normal	<0.01	1.75	Srikantan et al. (2000)
		2520749	Metastatic vs. normal	<0.2	−1.58	
		2520749	Metastatic vs. primary	<0.01	−4.00	
	PTENP1	3203669	Primary vs. normal	<0.01	−1.34	Poliseno et al. (2010)
		3203666	Metastatic vs. normal	<0.6	−1.09	
		3203669	Metastatic vs. primary	<0.04	1.50	
miRNA	*MIR143*	2835118	Metastatic vs. primary	<0.01	−1.78	Clape et al. (2009)
	MIR145	2835126	Metastatic vs. primary	<0.01	−4.77	Zaman et al. (2010)
		2835126	Metastatic vs. normal	<0.01	−7.98	
	MIR221	4006597	Metastatic vs. primary	<0.01	−1.52	Porkka et al. (2007)
		4006597	Metastatic vs. normal	<0.01	−2.11	

MFD: median fold difference in this dataset in various comparisons. The MFD value is computed as the ratio of the median between the first tissue type and the second tissue type in the "Comparison" column. Gray cells indicate statistical significance after multiple testing correction. Genes PCAT-1 (Prensner et al., 2011) and PRNCR1 (Chung et al., 2011) are not included as there is no gene model associated.

vs. metastatic (*MIR145* and *MIR221*) that have been previously reported as differentially expressed in prostate cancer (Porkka et al., 2007; Clape et al., 2009; Zaman et al., 2010).

Therefore, in addition to the handful of known ncRNAs, our analysis detected many other ncRNAs in regions that have yet to be explored in prostate cancer and that may play a role in the progression of the disease from normal glandular epithelium to distant metastases.

ASSESSMENT OF CLINICALLY SIGNIFICANT PROSTATE CANCER RISK GROUPS

Using MDS we observed that both exonic and non-exonic subsets of features present a statistically significant segregation of primary tumors from patients that progressed to metastatic disease (**Figure A1** in Appendix), in contrast to the findings of Taylor et al. (2010). Similarly, we found the exonic and non-exonic subsets to discriminate high and low Gleason score samples (**Figure A2** in Appendix). In order to assess the prognostic significance of differentially expressed exonic and non-exonic features, we trained a KNN classifier for each group using features from the comparison of normal and metastatic tissue samples (see Materials and Methods). Next, we used unmatched primary tumors (i.e., removing those tumors that had a matched normal in the training subset) as an independent validation set for the KNN classifiers. Each primary tumor in the validation set was classified by KNN as either more similar to normal or metastatic tissue. Subsequent

Kaplan–Meier analysis of the classified primary tumor samples using BCR as end point showed that, as expected, primary tumors classified as belonging to the metastatic group had a higher rate of BCR (**Figure 3**). However, the KNN classifier trained on the exonic subset of features showed no statistically significant difference in BCR-free survival using a log-rank test ($p < 0.08$) whereas the difference was highly significant for the non-exonic KNN classifier ($p < 0.00003$).

Next, we used logistic regression analysis to determine the odds ratio of metastatic disease progression (i.e., castrate or non-castrate resistant clinical metastatic patients) for the exonic and non-exonic KNN classifiers. The univariable analysis shows that, while the exonic set is significant ($\text{OR} = 8.57$, $p < 0.04$), the non-exonic set had more than double the odds ratio ($\text{OR} = 18.13$, $p < 0.0003$). Multivariable logistic regression further revealed that, after adjusting for clinicopathological variables using the Kattan nomogram (Kattan et al., 1999), the non-exonic KNN classifier retains a significant odds ratio for predicting metastatic disease progression ($\text{OR} = 11.7$, $p < 0.003$) whereas the exonic KNN classifier does not ($\text{OR} = 9.8$, $p < 0.07$; **Table 4**). These results suggest that additional prognostic information can be obtained from analysis of non-exonic RNAs and that these may have the potential to be used as biomarkers along with individual clinical variables and nomograms to enhance the prediction of metastatic disease progression post-prostatectomy.

DISCUSSION

One of the key challenges in prostate cancer is clinical and molecular heterogeneity (Rubin et al., 2011); therefore this common disease provides an appealing opportunity for genomic-based personalized medicine to identify diagnostic, prognostic, or predictive biomarkers to assist in clinical decision making. There have been extensive efforts to identify biomarkers based on high-throughput molecular profiling such as protein-coding mRNA expression microarrays (Sorensen and Orntoft, 2010). While many different biomarkers signatures have been identified, none of them are actively being used in clinical practice. The major reason that no new biomarker signatures have widespread use in the clinic is because they fail to show meaningful improvement for prognostication over PSA testing or established pathological variables (e.g., Gleason).

In this study, we assessed the utility of ncRNAs, and particularly non-exonic ncRNAs as potential biomarkers to be used for patients who have undergone prostatectomy but are at risk for recurrent disease and hence further treatment would be considered. We identified thousands of exonic and non-exonic RNAs differentially expressed between different tissue specimens from

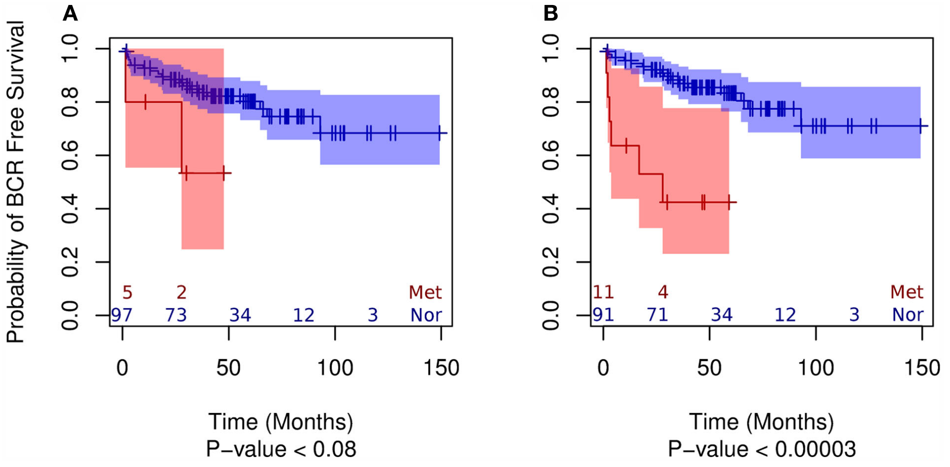


FIGURE 3 | Kaplan–Meier plots of the two groups of primary tumor samples classified by KNN ("normal-like" vs. "metastatic-like") using the BCR end point for exonic (A) and non-exonic (B) features.

Table 4 | Multivariable logistic regression analysis for prediction of the probability of metastatic disease progression.

Classifier	Exonic			Non-exonic		
Predictor	OR	OR CI (95%)	*P*-value	OR	OR CI (95%)	*P*-value
KNN-positive*	9.76	0.9–109.8	<0.07	11.7	1.7–80.8	<0.02
Nomogram§	14.8	2.4–92	<0.004	9.12	1.4–61.1	<0.03

Gray cells indicate statistical significance at the 5% significance level.

**KNN-positive: metastatic-like.*

§Greater than 50% probability of BCR was used as cut-off.

OR, odds ratio; CI, confidence interval.

the MSKCC Oncogenome Project. Of the non-exonic features, the majority fall within intronic regions. This further confirms the potential utility of intronic transcripts as biomarkers, given that previous studies have shown differential expression of these ncRNAs to correlate with Gleason score (Reis et al., 2004) and with tumor vs. benign prostate tissue types (Romanuik et al., 2009). In a more focused analysis of these feature subset groups (derived from comparison of normal adjacent to primary tumor and metastatic prostate cancer) three lines of evidence showed that the non-exonic feature subset contained substantial prognostic information as measured by its ability to discriminate two clinically relevant end points. First, we observed clustering of those primary tumor samples from patients that progressed to metastatic disease with true metastatic disease samples when using the non-exonic features. Second, Kaplan–Meier analysis showed that only the KNN classifier trained on the non-exonic feature set predicts risk groups (i.e., "normal-like" and "metastatic-like") with statistically significant differences in BCR-free survival. Finally, multivariable analysis showed that only the non-exonic KNN classifier had a statistically significant odds ratio of 11.7 for predicting metastatic disease progression in primary tumors after adjustment for Kattan nomogram.

Based on these three main results, we conclude that non-exonic RNAs contain previously unrecognized prognostic information that may be relevant in the clinic for the prediction of cancer progression post-prostatectomy. This goes in hand with the increasing evidence of ncRNAs being involved in metastasis, their key role in the regulation of protein-coding genes (Gibb et al., 2011) and their significantly higher tissue-specific expression compared to protein-coding genes (Cabili et al., 2011).

Perhaps the reason that previous efforts to develop new biomarker-based predictors of outcome in prostate cancer have not translated into the clinic is the focus on mRNA and proteins, therefore largely ignoring the wealth of information contained within the non-coding transcriptome. As more high-resolution data sets of the prostate cancer transcriptome become available (e.g., by new technologies such as RNA-Seq; Prensner et al., 2011) and as expression profiles of specific ncRNA transcripts are further validated, the results presented here can be further tested. While the clinical utility of these results require further validation on larger numbers of patients, they do show the potential of prognostic information encoded within ncRNAs, a part of the genome largely ignored in the immediate post-human genome project era.

These results add to the growing body of literature showing that the "dark matter" of the genome has potential to shed light on tumor biology, characterize aggressive cancer and improve in the prognosis and prediction of disease progression.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at http://www.frontiersin.org/Non-Coding_RNA/10.3389/fgene.2012.00023/abstract

Table S1 | Top 100 differentially expressed exonic and non-exonic features for each pairwise comparison. The features were ranked according to their adjusted *p*-value.

Table S2 | Gene ontology and pathway enrichment analysis for all, up-regulated and down-regulated features.

ACKNOWLEDGMENTS

We would like to thank Zaid Haddad and Benedikt Zimmermann for valuable input regarding this project. This research was supported in part by the National Research Council of Canada Industrial Research Assistance Program.

REFERENCES

Berteaux, N., Lottin, S., Adriaenssens, E., Van Coppenolle, F., Leroy, X., Coll, J., Dugimont, T., and Curgy, J. J. (2004). Hormonal regulation of H19 gene expression in prostate epithelial cells. *J. Endocrinol.* 183, 69–78.

Bolla, M., Van Poppel, H., Collette, L., Van Cangh, P., Vekemans, K., Da Pozzo, L., De Reijke, T. M., Verbaeys, A., Bosset, J. F., Van Velthoven, R., Marechal, J. M., Scalliet, P., Haustermans, K., and Pierart, M. (2005). Postoperative radiotherapy after radical prostatectomy: a randomised controlled trial (EORTC trial 22911). *Lancet* 366, 572–578.

Bussemakers, M. J., Van Bokhoven, A., Verhaegh, G. W., Smit, F. P., Karthaus, H. F., Schalken, J. A., Debruyne, F. M., Ru, N., and Isaacs, W. B. (1999). DD3: a new prostate-specific gene, highly overexpressed in prostate cancer. *Cancer Res.* 59, 5975–5979.

Cabili, M. N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., and Rinn, J. L. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.* 25, 1915–1927.

Catto, J. W., Alcaraz, A., Bjartell, A. S., De Vere White, R., Evans, C. P., Fussel, S., Hamdy, F. C., Kallioniemi, O., Mengual, L., Schlomm, T., and Visakorpi, T. (2011). MicroRNA in prostate, bladder, and kidney cancer: a systematic review. *Eur. Urol.* 59, 671–681.

Chung, S., Nakagawa, H., Uemura, M., Piao, L., Ashikawa, K., Hosono, N., Takata, R., Akamatsu, S., Kawaguchi, T., Morizono, T., Tsunoda, T., Daigo, Y., Matsuda, K., Kamatani, N., Nakamura, Y., and Kubo, M. (2011). Association of a novel long non-coding RNA in 8q24 with prostate cancer susceptibility. *Cancer Sci.* 102, 245–252.

Clape, C., Fritz, V., Henriquet, C., Apparailly, F., Fernandez, P. L., Iborra, F., Avances, C., Villalba, M., Culine, S., and Fajas, L. (2009). miR-143 interferes with ERK5 signaling, and abrogates prostate cancer progression in mice. *PLoS ONE* 4, e7542. doi:10.1371/journal.pone.0007542

Day, J. R., Jost, M., Reynolds, M. A., Groskopf, J., and Rittenhouse, H. (2011). PCA3: from basic molecular science to the clinical lab. *Cancer Lett.* 301, 1–6.

Gibb, E. A., Brown, C. J., and Lam, W. L. (2011). The functional role of long non-coding RNA in human carcinomas. *Mol. Cancer* 10, 38.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Statist.* 6, 65–70.

Kapranov, P., St Laurent, G., Raz, T., Ozsolak, F., Reynolds, C. P., Sorensen, P. H., Reaman, G., Milos, P., Arceci, R. J., Thompson, J. F., and Triche, T. J. (2010). The majority of total nuclear-encoded non-ribosomal RNA in a human cell is "dark matter" unannotated RNA. *BMC Biol.* 8, 149. doi:10.1186/1741-7007-8-149

Kattan, M. W., Wheeler, T. M., and Scardino, P. T. (1999). Postoperative nomogram for disease recurrence after radical prostatectomy for prostate cancer. *J. Clin. Oncol.* 17, 1499–1507.

Lin, R., Maeda, S., Liu, C., Karin, M., and Edgington, T. S. (2007). A large noncoding RNA is a marker for murine hepatocellular carcinomas and a spectrum of human carcinomas. *Oncogene* 26, 851–858.

Martens-Uzunova, E. S., Jalava, S. E., Dits, N. F., Van Leenders, G. J., Moller, S., Trapman, J., Bangma, C. H., Litman, T., Visakorpi, T., and Jenster, G. (2011). Diagnostic and prognostic signatures from the small non-coding RNA transcriptome in prostate cancer. *Oncogene.* doi: 10.1038/onc.2011.304. [Epub ahead of print].

McCall, M. N., Bolstad, B. M., and Irizarry, R. A. (2010). Frozen robust multiarray analysis (fRMA). *Biostatistics* 11, 242–253.

McCall, M. N., and Irizarry, R. A. (2011). Thawing Frozen Robust multi-array analysis (fRMA). *BMC Bioinformatics* 12, 369. doi:10.1186/1471-2105-12-369

Nakagawa, T., Kollmeyer, T. M., Morlan, B. W., Anderson, S. K., Bergstralh, E. J., Davis, B. J., Asmann, Y. W., Klee, G. G., Ballman, K. V., and Jenkins, R. B. (2008). A tissue biomarker panel predicting systemic progression after PSA recurrence post-definitive prostate cancer therapy. *PLoS ONE* 3, e2318. doi: 10.1371/journal.pone.0002318

Petrovics, G., Zhang, W., Makarem, M., Street, J. P., Connelly, R., Sun, L., Sesterhenn, I. A., Srikantan, V., Moul, J. W., and Srivastava, S. (2004). Elevated expression of PCGEM1, a prostate-specific gene with cell growth-promoting function, is associated with high-risk prostate cancer patients. *Oncogene* 23, 605–611.

Poliseno, L., Salmena, L., Zhang, J., Carver, B., Haveman, W. J., and Pandolfi, P. P. (2010). A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature* 465, 1033–1038.

Porkka, K. P., Pfeiffer, M. J., Waltering, K. K., Vessella, R. L., Tammela, T. L., and Visakorpi, T. (2007). MicroRNA expression profiling in prostate cancer. *Cancer Res.* 67, 6130–6135.

Prensner, J. R., Iyer, M. K., Balbin, O. A., Dhanasekaran, S. M., Cao, Q., Brenner, J. C., Laxman, B., Asangani, I. A., Grasso, C. S., Kominsky, H. D., Cao, X., Jing, X., Wang, X., Siddiqui, J., Wei, J. T., Robinson, D., Iyer, H. K., Palanisamy, N., Maher, C. A., and Chinnaiyan, A. M. (2011). Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. *Nat. Biotechnol.* 29, 742–749.

Reis, E. M., Nakaya, H. I., Louro, R., Canavez, F. C., Flatschart, A. V., Almeida, G. T., Egidio, C. M., Paquola, A. C., Machado, A. A., Festa, F., Yamamoto, D., Alvarenga, R., Da Silva, C. C., Brito, G. C., Simon, S. D., Moreira-Filho, C. A., Leite, K. R., Camara-Lopes, L. H., Campos, F. S., Gimba, E., Vignal, G. M., El-Dorry, H., Sogayar, M. C., Barcinski, M. A., Da Silva, A. M., and Verjovski-Almeida, S. (2004). Antisense intronic non-coding RNA levels correlate to the degree of tumor differentiation in prostate cancer. *Oncogene* 23, 6684–6692.

Romanuik, T. L., Ueda, T., Le, N., Haile, S., Yong, T. M., Thomson, T., Vessella, R. L., and Sadar, M. D. (2009). Novel biomarkers for prostate cancer including noncoding transcripts. *Am. J. Pathol.* 175, 2264–2276.

Rubin, M. A., Maher, C. A., and Chinnaiyan, A. M. (2011). Common gene rearrangements in prostate cancer. *J. Clin. Oncol.* 29, 3659–3668.

Schaefer, A., Jung, M., Mollenkopf, H. J., Wagner, I., Stephan, C., Jentzmik, F., Miller, K., Lein, M., Kristiansen, G., and Jung, K. (2010). Diagnostic and prognostic implications of microRNA profiling in prostate carcinoma. *Int. J. Cancer* 126, 1166–1176.

Sevli, S., Uzumcu, A., Solak, M., Ittmann, M., and Ozen, M. (2010). The function of microRNAs, small but potent molecules, in human prostate cancer. *Prostate Cancer Prostatic Dis.* 13, 208–217.

Siegel, R., Ward, E., Brawley, O., and Jemal, A. (2011). Cancer statistics, 2011: the impact of eliminating socioeconomic and racial disparities on premature cancer deaths. *CA Cancer J. Clin.* 61, 212–236.

Simmons, M. N., Stephenson, A. J., and Klein, E. A. (2007). Natural history of biochemical recurrence after radical prostatectomy: risk assessment for secondary therapy. *Eur. Urol.* 51, 1175–1184.

Sorensen, K. D., and Orntoft, T. F. (2010). Discovery of prostate cancer biomarkers by microarray gene expression profiling. *Expert Rev. Mol. Diagn.* 10, 49–64.

Srikantan, V., Zou, Z., Petrovics, G., Xu, L., Augustus, M., Davis, L., Livezey, J. R., Connell, T., Sesterhenn, I. A., Yoshino, K., Buzard, G. S., Mostofi, F. K., Mcleod, D. G., Moul, J. W., and Srivastava, S. (2000). PCGEM1, a prostate-specific gene, is overexpressed in prostate cancer. *Proc. Natl. Acad. Sci. U.S.A.* 97, 12216–12221.

Taylor, B. S., Schultz, N., Hieronymus, H., Gopalan, A., Xiao, Y., Carver, B. S., Arora, V. K., Kaushik, P., Cerami, E., Reva, B., Antipin, Y., Mitsiades, N., Landers, T., Dolgalev, I., Major, J. E., Wilson, M., Socci, N. D., Lash, A. E., Heguy, A., Eastham, J. A., Scher, H. I., Reuter, V. E., Scardino, P. T., Sander, C., Sawyers, C. L., and Gerald, W. L. (2010). Integrative genomic profiling of human prostate cancer. *Cancer Cell* 18, 11–22.

Thompson, I. M., Tangen, C. M., Paradelo, J., Lucia, M. S., Miller, G., Troyer, D., Messing, E., Forman, J., Chin, J., Swanson, G., Canby-Hagino, E., and Crawford, E. D. (2009). Adjuvant radiotherapy for pathological T3N0M0 prostate cancer significantly reduces risk of metastases and improves survival: long-term followup of a randomized clinical trial. *J. Urol.* 181, 956–962.

Wiegel, T., Bottke, D., Steiner, U., Siegmann, A., Golz, R., Storkel, S., Willich, N., Semjonow, A., Souchon, R., Stockle, M., Rube, C., Weissbach, L., Althaus, P., Rebmann, U., Kalble, T., Feldmann, H. J., Wirth, M., Hinke, A., Hinkelbein, W., and Miller, K. (2009). Phase III postoperative adjuvant radiotherapy after radical prostatectomy compared with radical prostatectomy alone in pT3 prostate cancer with postoperative undetectable prostate-specific antigen: ARO 96-02/AUO AP 09/95. *J. Clin. Oncol.* 27, 2924–2930.

Yap, K. L., Li, S., Munoz-Cabello, A. M., Raguz, S., Zeng, L., Mujtaba, S., Gil, J., Walsh, M. J., and Zhou, M. M. (2010). Molecular interplay of the noncoding RNA ANRIL and methylated histone H3 lysine 27 by polycomb CBX7 in transcriptional silencing of INK4a. *Mol. Cell* 38, 662–674.

Yates, T. (2010). *Xmapcore: Core Access to the Xmap Database*. Available at: http://xmap.picr.man.ac.uk

Zaman, M. S., Chen, Y., Deng, G., Shahryari, V., Suh, S. O., Saini, S., Majid, S., Liu, J., Khatri, G., Tanaka, Y., and Dahiya, R. (2010). The functional significance of microRNA-145 in prostate cancer. *Br. J. Cancer* 103, 256–264.

Conflict of Interest Statement: Ismael A. Vergara, Anamaria Crisan, Thomas Sierocinski, Christine Buerki, Nicholas Erho, Mercedeh Ghadessi, Timothy J. Triche, and Elai Davicioni are employes of GenomeDx Biosciences, Inc.

Received: 14 December 2011; paper pending published: 27 December 2011; accepted: 04 February 2012; published online: 22 February 2012.

Citation: Vergara IA, Erho N, Triche TJ, Ghadessi M, Crisan A, Sierocinski T, Black PC, Buerki C and Davicioni E (2012) Genomic "dark matter" in prostate cancer: exploring the clinical utility of ncRNA as biomarkers. Front. Gene. ***3****:23. doi: 10.3389/fgene.2012.00023*

This article was submitted to Frontiers in Non-Coding RNA, a specialty of Frontiers in Genetics.

APPENDIX

STEPS FOR THE DETECTION OF DIFFERENTIALLY EXPRESSED FEATURES

1) Download raw CEL files from http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc = GSE21034
2) Pre-process all exon arrays using the fRMA algorithm (McCall et al., 2010; McCall and Irizarry, 2011) with custom fRMA vectors created from the files obtained in step 1.fRMA can be obtained from http://www.bioconductor.org/packages/release/bioc/html/frma.html
3) Exclude all probe selection regions (PSRs) annotated as "unreliable" by the xmapcore package (Yates, 2010; one or more probes do not align uniquely to the genome) as well as those not defined as class 1 cross-hybridizing by Affymetrix annotation (http://www.affymetrix.com).
4) Classify each PSR as "exonic" if they overlap with protein-coding regions or UTRs according to the xmapcore package annotation, and as "non-exonic" if they do not (this can be achieved with the "coding.probe sets" and "utr.probe sets" functions).

   Then, the following steps need to be pursued separately for each pairwise comparison: (i) normal vs. primary, (ii) primary vs. metastatic, and (iii) normal vs. metastatic. For the normal vs. primary comparison, only matched samples were used and for the primary vs. metastatic comparison the matched samples were excluded.
5) Calculate the background expression level by taking the median of the Affymetrix defined anti-genomic PSRs (http://www.affymetrix.com/Auth/support/downloads/library_files/HuEx-1_0-st-v2.r2.zip; file HuEx-1_0-st-v2.r2.antigenomic.bgp). For each PSR, calculate the median expression level for each group. Filter PSRs where the median expression levels for both groups are below the background expression level.
6) Apply the rowttests function of the genefilter R package available at http://www.bioconductor.org/packages/2.3/bioc/html/genefilter.html in order to perform a *t*-test on each PSR.
7) Adjust the obtained *p*-values using the p.adjust function of the stats package in R for each group of PSRs (exonic and non-exonic) separately. Select the Holm–Bonferroni method for this purpose (Holm, 1979).
8) Filter out those PSRs that have an adjusted *p*-value higher than 0.05.

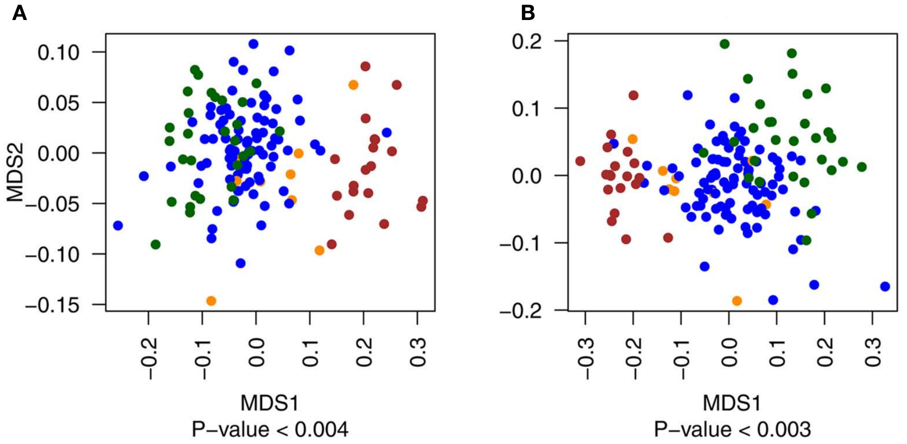


FIGURE A1 | Multidimensional scaling plots of the distribution of primary tumor samples with (yellow) and without (blue) metastatic events compared to metastatic (red) and normal (green) tissues for exonic (A) and non-exonic (B) features. Metastatic and normal data points are included in the figure for illustrative purposes only.

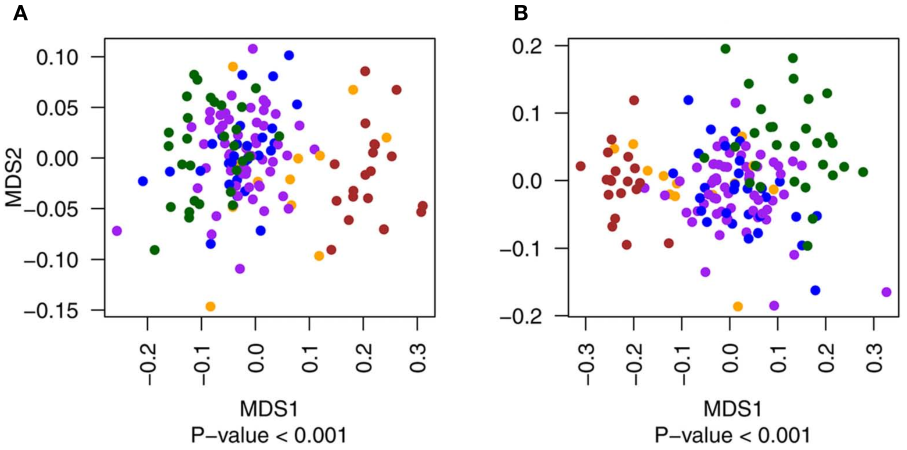


FIGURE A2 | Multidimensional scaling plots of the distribution of primary tumor samples with Gleason score of 6 (blue), 7 (purple), 8 and 9 (both in yellow) compared to metastatic (red) and normal (green) tissues for exonic (A) and non-exonic (B) features. Metastatic and normal data points are included in the figure for illustrative purposes only.

frontiers in
GENETICS

MINI REVIEW ARTICLE
published: 08 March 2012
doi: 10.3389/fgene.2012.00032

Perspectives of long non-coding RNAs in cancer diagnostics

Eduardo M. Reis* and Sergio Verjovski-Almeida*

Departamento de Bioquímica, Instituto de Química, Universidade de São Paulo, São Paulo, Brasil

Edited by:
Philipp Kapranov, St. Laurent Institute, USA

Reviewed by:
Georges St. Laurent, St. Laurent Institute, USA
Elai Davicioni, GenomeDx Biosciences Inc., Canada

****Correspondence:***
Eduardo M. Reis and Sergio Verjovski-Almeida, Departamento de Bioquímica, Instituto de Química, Universidade de São Paulo, 05508-900 São Paulo, Brasil.
e-mail: emreis@iq.usp.br; verjo@iq.usp.br

Long non-coding RNAs (lncRNAs) transcribed from intergenic and intronic regions of the human genome constitute a broad class of cellular transcripts that are under intensive investigation. While only a handful of lncRNAs have been characterized, their involvement in fundamental cellular processes that control gene expression highlights a central role in cell homeostasis. Not surprisingly, aberrant expression of regulatory lncRNAs has been increasingly documented in different types of cancer, where they can mediate both oncogenic or tumor suppressor effects. Interaction with chromatin remodeling complexes that promote silencing of specific genes or modulation of splicing factor proteins seem to be two general modes of lncRNA regulation, but it is conceivable that additional mechanisms of action are yet to be unveiled. LncRNAs show greater tissue specificity compared to protein-coding mRNAs making them attractive in the search of novel diagnostics/prognostics cancer biomarkers in body fluid samples. In fact, lncRNA prostate cancer antigen 3 can be detected in urine samples and has been shown to improve diagnosis of prostate cancer. We suggest that an unbiased screening of the presence of RNAs in easily accessible body fluids such as serum and urine might reveal novel circulating lncRNAs as potential biomarkers in many types of cancer. Annotation and functional characterization of the lncRNA complement of the cancer transcriptome will conceivably provide new venues for early diagnosis and treatment of the disease.

Keywords: long non-coding RNA, cancer, diagnostics, expression signature

Over the last decade, advances in genome-wide analyses of the eukaryotic transcriptome have revealed that most of the human genome is transcribed, generating a large repertoire of (>200 nt) long non-coding RNAs (lncRNAs) that map to intronic and intergenic regions (Birney et al., 2007; Dinger et al., 2009; Ponting et al., 2009; Kapranov et al., 2010). These include subsets of polyadenylated and non-polyadenylated transcripts that accumulate differently in the nucleus and cytoplasm of cells (Kapranov et al., 2007, 2010). The catalog of human lncRNAs has expanded dramatically just in the last several years; in fact, recently published deep RNA sequencing reveals that the range, depth, and complexity of the human transcriptome is far from fully characterized (Mercer et al., 2012) and expectations are that very soon the human lncRNA genes will outnumber protein-coding genes. The definition and naming of lncRNAs are currently evolving in the literature and different classes or categories of lncRNA have been described (Prensner and Chinnaiyan, 2011; Wright and Bruford, 2011). The categorization recently proposed by the HUGO Gene Nomenclature Committee (HGNC) is an ongoing project (Wright and Bruford, 2011), where lncRNAs were described as spliced, capped, and polyadenylated RNAs (Wright and Bruford, 2011); this clearly does not encompass all different lncRNAs that may be also unspliced and/or non-polyadenylated (Nakaya et al., 2007; Kapranov et al., 2010; Yang et al., 2011a). The rapid increase in the number of described lncRNAs along with the lack of uniform and systematic annotation nomenclature for the diverse and extensive amount of lncRNAs expressed in human tissues imposes a considerable limitation regarding the completeness of any database related to lncRNAs (Paschoal et al., 2012).

It is apparent that lncRNAs may act through diverse molecular mechanisms and play regulatory and structural roles in important biological processes (see Mattick, 2009 for a review). Presently, the mechanisms of action of only a few lncRNAs have been characterized in detail (Wang and Chang, 2011), and many of these lncRNAs have an altered expression in different types of human cancer (Huarte and Rinn, 2010; Gibb et al., 2011; Prensner and Chinnaiyan, 2011).

Cancers are the result of a process where somatic cells mutate and escape the controlled balance exerted by gene expression programs and cellular networks that maintain cellular homeostasis and normally prevent their unwanted expansion. Cancer cells differ from normal cells in many important characteristics, including loss of differentiation, increased invasiveness, and decreased drug sensitivity. Genes that affect these processes can be classified into two major groups: tumor suppressor genes and oncogenes. Tumor suppressor genes protect cells against mutations that initiate transformation. Conversely, oncogenes initiate the cellular transformation process when inappropriately activated. LncRNAs have been recently implicated as having tumor suppressor and oncogenic roles (Huarte and Rinn, 2010; Gibb et al., 2011; Prensner and Chinnaiyan, 2011).

GENERAL MECHANISMS OF lncRNA FUNCTION IMPLICATED IN CANCER

Long non-coding RNAs can activate cellular pathways that lead to tumorigenesis, in analogy to protein-coding oncogenes. The molecular mechanisms by which lncRNAs exert their biological functions has been extensively reviewed by (Wang and Chang, 2011). Here we highlight some examples of lncRNAs implicated in oncogenic functions.

One example of such an oncogenic lncRNA is metastasis-associated in lung adenocarcinoma transcript 1 (MALAT1), a nuclear-retained non-coding RNA that has been recently shown to regulate alternative splicing by modulating serine/arginine (SR) splicing factor phosphorylation (Tripathi et al., 2010). Increased expression of the lncRNA MALAT1 has been first observed in metastatic non-small cell lung cancer (Ji et al., 2003; see details in the next section), followed by endometrial stromal sarcoma of the uterus (Yamada et al., 2006), and more recently in six other types of cancer, including hepatocellular carcinoma, breast, pancreas, lung, colon, and prostate cancers (Lin et al., 2007). Recently, short hairpin RNA inhibition of MALAT1 in human cervical cancer cells was shown to suppress cell proliferation and invasion (Guo et al., 2010), whereas RNA interference-mediated silencing of MALAT1 reduced the *in vitro* migration of lung adenocarcinoma cells by influencing the expression of motility-related genes (Tano et al., 2010). Altogether, these findings reinforce the role of MALAT1 as an oncogenic lncRNA, and point to one of the possible modes of action of lncRNAs, namely through their interaction with and modulation of splicing factor proteins.

The mode of action of lncRNAs through the interaction with chromatin remodeling complexes may be a more general one, as it has been documented for two lncRNAs. One example is ANRIL (antisense non-coding RNA in the INK4 locus) that is altered in an estimated 30–40% of human tumors (Kim and Sharpless, 2006). Tumor suppressor p15^{INK4B} is silenced by its antisense ANRIL transcript (Yu et al., 2008); lncRNA ANRIL is required for the recruitment of polycomb PRC1 and PRC2 complexes to the INK4B locus and for silencing of p15^{INK4B} tumor suppressor gene (Yap et al., 2010; Kotake et al., 2011).

Another example is HOTAIR (HOX Antisense Intergenic RNA), a metastasis-associated gene located in the mammalian HOXC locus that reprograms chromatin state to promote cancer metastasis (Gupta et al., 2010). HOTAIR lncRNA interacts with Polycomb Repressor Complex PRC2, determining PRC2 localization and repression of the HOXD locus (Rinn et al., 2007). Recently it was found that HOTAIR serves as a scaffold for at least two distinct histone modification complexes. HOTAIR binds the PRC2 complex responsible for H3K27 methylation and also LSD1, a histone lysine demethylase that mediates enzymatic demethylation of H3K4Me2 (Tsai et al., 2010). Although the precise mechanism of HOTAIR activities remains to be elucidated, it is clear that HOTAIR participates in silencing of metastasis suppressor genes thus promoting metastasis, as discussed below.

LARGE-SCALE EXPRESSION PROFILING OF lncRNAs IN PATIENTS

Cancer gene profiling studies have had an enormous impact on understanding of the biology of cancers, pointing to the biological heterogeneity of specific cancer types, providing identification of novel oncogenes and tumor suppressors, and defining pathways that interact to drive the growth of individual cancers (Cowin et al., 2010). Large-scale genomic studies are underway, such as The Cancer Genome Atlas project that aims to catalog in each cancer type the changes in DNA copy number and methylation, as well as in small (19–25 nt) non-coding microRNA (miRNA) and protein-coding mRNA expression (Cancer_Genome_Atlas_Research_Network, 2008, 2011). Noteworthy, changes in expression of lncRNAs have not been analyzed in these large cohort studies.

Identification of lncRNAs correlated to cancer has benefited in the past decade from the development of a number of effective high-throughput expression analyses technologies as well as from the increasing realization that lncRNAs may play important roles in physiological and pathological processes in the cell (see **Table 1**). Early efforts to identify molecular cancer markers based on the screening of cDNA libraries enriched in tumor-specific transcripts have identified lncRNAs whose expression levels correlate to cancer. Using a differential display approach, the lncRNA DD3, later named prostate cancer antigen 3 (PCA3), was initially identified as overexpressed in prostate tumors relative to benign prostate hyperplasia and normal epithelium (Bussemakers et al., 1999). Further studies later indicated that PCA3 is a very specific prostate cancer gene whose mechanism of action is still not identified (Marks and Bostwick, 2008; Shappell, 2008).

Another report that used large-scale transcriptome analysis to look for differential gene expression in cancer and gave attention to a differentially expressed lncRNA, namely MALAT1, employed a subtractive hybridization approach to determine differences in gene expression between primary non-small cell lung cancer tumors of five patients that were cured by surgery and tumors of four patients that subsequently metastasized (Ji et al., 2003). In all, 26 transcripts were found more than once, and among them the novel lncRNA named MALAT1. Subsequently, 31 samples from stage I patients suffering from adenocarcinoma or squamous cell carcinoma were analyzed by qPCR, and the expression levels of MALAT1 were significantly higher in metastasizing adenocarcinomas compared to non-metastasizing ones ($p = 0.03$); interestingly, no significant differences in gene expression were found for squamous cell carcinomas ($n = 34$; Ji et al., 2003). These data provided evidence that the association of lncRNA MALAT1 with metastasis depended on the lung tumor's histology.

A large-scale gene expression approach specifically designed to look for lncRNAs correlated to cancer has employed hybridization of RNA derived from normal human breast epithelia, primary breast carcinomas, and distant metastases to ultra-dense tiling arrays covering the entire HOX gene loci (Gupta et al., 2010). The authors found that 233 transcribed regions in the HOX loci, comprising 170 lncRNAs and 63 HOX exons, were differentially expressed (Gupta et al., 2010), with a systematic variation in the expression of HOX lncRNAs among normal breast epithelia, primary tumor, and metastases. Dozens of HOX lncRNAs were expressed in normal breast but showed reduced expression in all cancer samples; conversely, a set of HOX lncRNAs was frequently expressed in primary tumors but not in metastases (Gupta et al., 2010). Notably, one such metastasis-associated lncRNA was

Table 1 | Long non-coding RNAs differentially expressed in human cancer.

LncRNAs	Description	Tumor type	[a]Comparison	Down in tumors	Up in tumors	Platform	Reference
PCA3 (DD3)	3.7 kb lncRNA overexpressed in prostate tumors. Used for diagnosis of prostate cancer in body fluids	Prostate cancer	prostatic tumors vs. non-neoplastic adjacent prostatic tissue ($n = 56$)	–	–	Differential display	Bussemakers et al. (1999)
MALAT1	8.5 kb lncRNA aberrantly expressed in lung cancer. Several other ncRNAs were identified in the same screening	Non-small cell lung cancer (NSCLC)	Recurrent ($n = 4$) vs. non-recurrent ($n = 5$) NSCLC patients	–	–	Subtractive hybridization	Ji et al. (2003)
HOTAIR	2.3 kb HOTAIR plus 170 other lncRNAs were identified as differentially expressed in breast cancer using tiling arrays covering the entire HOX gene *loci*	Breast cancer	Non-tumor ($n = 5$, pooled) primary tumors ($n = 8$) vs. metastases ($n = 6$). MALAT1 validated in additional 132 primary tumor samples with known followup	–	–	ultra-dense tiling oligo arrays, qRT-PCR	Gupta et al. (2010)
Non-coding transcribed ultra conserved regions (T-UCRs)	Transcripts >200 nt from regions 100% conserved between orthologous segments of the human, rat, and mouse genomes	Chronic lymphocytic leukemia (CLL)	CLL ($n = 50$) vs. CD5+ (normal, $n = 6$)	4	5	40-mer custom oligo arrays	Calin et al. (2007)
		Colorectal carcinoma (CRC)	CRC ($n = 78$) vs. normal colon mucosa ($n = 21$)	1	27		
		Hepatocellular carcinoma (HCC)	HCC ($n = 9$) vs. normal liver ($n = 4$)	4	1		
Long intronic non-coding RNAs	Transcripts >500 nt mapping to intronic regions of protein-coding genes	Prostate adenocarcinoma (PC)	[b]Low GS ($n = 6$) vs. high GS ($n = 5$) PC	21	3	Custom-designed double-stranded cDNA micro arrays	Reis et al. (2004)
		Renal cell carcinoma (RCC)	RCC vs. adjacent non-tumor tissues ($n = 6$)	6	1		Brito et al. (2008)
		Pancreatic ductal adenocarcinoma (PDAC)	PDAC ($n = 15$) vs. non-tumor tissues ($n = 17$)	23	1		Tahira et al. (2011)
			PDAC ($n = 15$) vs. metastasis ($n = 6$)	99	2		
Conserved long non-coding transcripts (NCTs)	Abundantly expressed non-coding transcripts that are >400 nt long, and which displayed a high degree of sequence conservation	Breast and ovarian cancers	Breast tumor ($n = 17$) vs. normal breast primary cell cultures ($n = 4$)	15	1	Affy 25-mer oligo arrays, and real rime RT-PCR	Perez et al. (2008)
			Ovarian tumor ($n = 20$) vs. normal ovary primary cell cultures ($n = 3$)	9	13		
Prostate cancer-associated ncRNA transcripts (PCATs)	Long (>250 bp) intergenic transcripts similar to lincRNAs (Guttman et al., 2009) that are differentially expressed in prostate cancer	Localized and metastatic prostate cancer	localized prostate cancer ($n = 47$) vs. benign adjacent prostate tissues ($n = 20$)	10	111	RNA-seq	Prensner et al. (2011)

[a] *Number of samples are shown in parentheses.*

[b] *GS, Tumor Gleason Score.*

HOTAIR, which had a unique association with patient prognosis (Gupta et al., 2010).

Oligoarrays were used to interrogate 481 ultra conserved regions (UCRs) in the human genome (Calin et al., 2007); UCRs are a subset of conserved sequences that are located in both intra- and intergenic regions and are absolutely conserved (100%) between orthologous regions of the human, rat, and mouse genomes (Bejerano et al., 2004). A total of 256 UCRs (53%) were identified as non-coding genomic regions (Bejerano et al., 2004). The authors investigated the expression of UCRs in a panel of 173 samples, including 133 human cancers [e.g., chronic lymphocytic leukemias (CLL), colorectal (CRC), and hepatocellular carcinomas (HCC)] and 40 corresponding normal tissues (Calin et al., 2007). Specific sets of UCRs were differentially expressed in distinct tumor types, and among them, 42 were non-coding UCRs (48% of the differentially expressed UCRs). This work demonstrated that the transcribed UCR expression profiles can be used to differentiate human cancers (Calin et al., 2007).

Custom-designed cDNA microarrays interrogating selected sets of protein-coding genes and lncRNAs from intronic/intergenic genomic regions were used for obtaining expression profiles from clinical samples of a number of cancer types (Reis et al., 2004; Brito et al., 2008; Tahira et al., 2011). In prostate cancer, RNA from 27 patient tumor samples with Gleason scores ranging from 5 to 10 were hybridized to these custom-designed microarrays (Reis et al., 2004). Among the 56 transcripts that were found to be significantly correlated to the degree of prostate tumor differentiation (Gleason score), 23 were lncRNAs mapping to intronic regions (Reis et al., 2004). Among the top twelve transcripts most significantly correlated to tumor differentiation, six were antisense intronic lncRNAs as shown by orientation-specific RT-PCR or northern blot analysis with strand-specific riboprobe (Reis et al., 2004).

Aberrant expression of intronic lncRNAs was studied in clear cell renal cell carcinoma (RCC) using matched samples of tumor and adjacent non-neoplastic tissue obtained from six patients (Brito et al., 2008). A set of 55 transcripts was significantly down-regulated in clear cell RCC relative to the matched non-tumor tissue; among the down-regulated transcripts, 49 mapped to untranslated or coding exons of protein-coding genes and 6 were lncRNAs mapped to intronic regions in genomic loci of protein-coding genes (Brito et al., 2008).

More recently, pancreatic ductal adenocarcinoma (PDAC) was studied, aiming at identifying gene expression profiles of protein-coding and lncRNAs correlated to pancreatic cancer and metastasis in 38 clinical samples of tumor and non-tumor pancreatic tissues (Tahira et al., 2011). Statistically significant expression signatures comprising protein-coding mRNAs and intronic/intergenic lncRNAs that correlate to PDAC or to pancreatic cancer metastasis were identified; interestingly, loci harboring intronic lncRNAs differentially expressed in PDAC metastases were enriched in genes associated to the MAPK pathway (Tahira et al., 2011).

Whole-genome tiling arrays were utilized to identify the expression of novel lncRNAs across the entire human genome (Perez et al., 2008). The authors hybridized RNA from normal lung cell cultures to the tiling arrays and found 495 transcriptionally active regions originated from non-protein-coding sequence (intergenic or intronic regions) and chose 15 candidate RNAs for subsequent real-time RT–PCR, northern blot, and sequencing experiments of which three were intronic lncRNAs (Perez et al., 2008). Altered expression of these lncRNAs was found in patient samples in both breast and ovarian cancers (Perez et al., 2008).

The first high-throughput sequencing of polyA+ RNA (RNA-seq) from a large cohort of 102 prostate tissues and cells lines has been recently reported (Prensner et al., 2011). The work has identified 121 unannotated prostate cancer-associated lncRNA transcripts (PCATs) and has characterized one lncRNA, PCAT-1, as a prostate-specific regulator of cell proliferation, showing that it is a target of PRC2 (Prensner et al., 2011). Patterns of PCAT-1 and PRC2 expression stratified patient tissues into molecular subtypes distinguished by expression signatures of PCAT-1-repressed target genes. These findings establish the utility of RNA-seq to identify disease-associated lncRNAs that may improve the stratification of cancer subtypes (Prensner et al., 2011).

Although the functional consequences of the deregulation of lncRNAs in cancer development are currently unknown, the studies discussed above indicate that this class of transcripts may play important functions in both normal and malignant tissues.

lncRNAs AS A DIAGNOSTIC TEST TOOL

Molecular markers of malignancy are important diagnostic and prognostic tools that help patient management in the oncology clinics. Cancer is a multi-factorial disease and for most types of malignancies an increase in the number of available assessment and management tools is desirable. Expression of the lncRNA MALAT1 has been identified by Kaplan–Meier analyses as a prognostic parameter for patient survival in stage I non-small cell lung cancer (Ji et al., 2003). MALAT1 has been subsequently validated as a marker for endometrial stromal sarcoma of the uterus (Yamada et al., 2006) and for HCC and a spectrum of five other human carcinomas (Lin et al., 2007). In addition, increased expression of MALAT1 has been recently shown to be an independent prognostic factor for HCC following liver transplantation (Lai et al., 2011).

Increased expression of lncRNA HOTAIR was shown to be associated with metastasis in breast cancer patients, having a unique association with patient prognosis (Gupta et al., 2010). Subsequently, HOTAIR expression levels was found to correlate with metastasis in colorectal carcinoma (Kogo et al., 2011), and to predict tumor recurrence in hepatocellular carcinoma (Yang et al., 2011b).

At present, few lncRNAs have been characterized as potential biomarkers in human fluids. Measurement of lncRNA PCA3 in patient urine samples has been shown to allow more sensitive and specific diagnosis of prostate cancer than the widely used PSA (prostate-specific antigen) serum levels (Fradet et al., 2004; Tinzl et al., 2004; Shappell, 2008). The potential of PCA3 urine assay to aid prostate cancer diagnosis and minimize unnecessary biopsies has been extensively documented, highlighting its advantages over PSA and pointing to future challenges for this new diagnostic biomarker (Lee et al., 2011).

The lncRNA HULC (highly upregulated in liver cancer) is highly expressed in HCC patients and can be detected in the blood by conventional PCR methods (Panzitt et al., 2007). It has been later shown that HULC lncRNA expression is not confined to HCC, but is also expressed in colorectal carcinomas that metastasize to the liver (Matouk et al., 2009).

Diagnosis and treatment follow up of complex multi-factorial diseases such as cancer could conceivably be improved by screening of a larger number of molecular biomarkers in easily accessible sample specimens. In fact, highly stable cell-free circulating nucleic acids (cfCNA), both RNA and DNA species, have been discovered in the blood, plasma, and urine of humans (Tong and Lo, 2006). At present, there is evidence of a good correlation between tumor-associated changes in genomic, epigenetic, or transcriptional patterns and alterations in cfCNA levels (Schwarzenbach et al., 2011), strongly pointing to the utility of this blood biomarker class as promising clinical tools.

The release of nucleic acids into the blood is thought to be related to the apoptosis and necrosis of cancer cells in the tumor microenvironment and is also the result of secretion. Circulating RNAs are detectable in the serum and plasma of cancer patients, being surprisingly stable in spite of the fact that high amounts of RNases circulate in the blood of cancer patients. This implies that RNA may be protected from degradation by its packaging into microparticles, which include exosomes, microvesicles, apoptotic bodies, and apoptotic microparticles (Orozco and Lewis, 2010). Microparticles are small, membranous vesicles that can contain DNA, RNA, miRNA, intracellular proteins, and extracellular surface markers from the parental cells; they can be secreted from intracellular multivesicular bodies or released from the surface of blebbing membranes (Cocucci et al., 2009; Orozco and Lewis, 2010). The detection and identification of RNA in serum and plasma can be carried out using microarray technologies or reverse transcription quantitative real-time PCR (O'Driscoll et al., 2008). The reported RNA content of microvesicles and exosomes thus far includes primarily small miRNAs and long protein-coding mRNAs (Record et al., 2011).

Recent advances in small non-coding miRNA expression profiling in human cancer and their potential as therapeutic targets and novel biomarkers have been reviewed (Farazi et al., 2011; Munker and Calin, 2011). The presence of small non-coding miRNAs in serum of cancer patients was first described for diffuse large B-cell lymphoma patients (Lawrie et al., 2008). Circulating miRNAs were subsequently used in assessing patients with prostate cancer (Mitchell et al., 2008) and at present circulating miRNAs have been characterized as potential biomarkers in over ten different cancers (Kosaka et al., 2010). Despite being promising biomarkers for cancer diagnosis and prognosis, there have been conflicting findings about circulating miRNAs from the same tumor reported from different studies (Kosaka et al., 2010). These discrepancies might be due to the lack of an established endogenous miRNA control to normalize for circulating miRNA levels, and also to the different extraction and quantification methods used among the studies (Kosaka et al., 2010). An effort to standardize results is warranted by putting forward recommendations for controlling pre-analytical variables, including the reduction of contaminant cellular miRNAs of hematopoietic origin in the isolation and quantization of cell-free circulating RNAs (Duttagupta et al., 2011).

We speculate that in addition to miRNAs and mRNAs the human serum might contain a considerable amount of lncRNAs that will eventually be detected by the use of unbiased high-throughput technologies such as genome tiling expression microarrays or RNA-seq deep-sequencing of serum samples. Such approaches should be subjected to the same controls regarding pre-analytical variables (Duttagupta et al., 2011), including the reduction of contaminant hematopoietic cells in the isolation and quantization of cell-free circulating lncRNAs.

Comparative studies of lncRNAs in serum from cancer patient large cohorts and from normal subjects will possibly reveal novel circulating lncRNAs as potential biomarkers in many types of cancers.

CONCLUSION AND PERSPECTIVES

Long non-coding RNA expression profiles in human cancers have highlighted the potential value of this class of non-coding RNAs as tumor markers in patient diagnosis and prognosis. The rapidly expanding catalog of lncRNAs holds promises that in the near future lncRNAs will become ever more important in cancer patient management. An analogy can be made with the impact of small miRNA profiling in many types of cancer (Braconi et al., 2011; Ferracin et al., 2011; Schetter and Harris, 2011; Wang and Sen, 2011), which has provided different experimental lines of evidence that deregulation of miRNAs not only results as consequence of cancer progression but also directly affects gene networks that promote tumor initiation and progression in a cause-effect manner (Lovat et al., 2011). As the catalog of lncRNAs grows, it will become important to elucidate the genetic networks and pathways regulated by the abnormally expressing lncRNAs in cancer cells as a means to understanding the role of these lncRNAs in the induction of malignant transformation.

ACKNOWLEDGMENTS

This research was supported by grants from Fundação de Amparo a Pesquisa do Estado de São Paulo, and from Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Brasil. Eduardo M. Reis and Sergio Verjovski-Almeida received established investigator fellowships from CNPq.

REFERENCES

Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W. J., Mattick, J. S., and Haussler, D. (2004). Ultraconserved elements in the human genome. *Science* 304, 1321–1325.

Birney, E., Stamatoyannopoulos, J. A., Dutta, A., Guigo, R., Gingeras, T. R., Margulies, E. H., Weng, Z., Snyder, M., Dermitzakis, E. T., Thurman, R. E., Kuehn, M. S., Taylor, C. M., Neph, S., Koch, C. M., Asthana, S., Malhotra, A., Adzhubei, I., Greenbaum, J. A., Andrews, R. M., Flicek, P., Boyle, P. J., Cao, H., Carter, N. P., Clelland, G. K., Davis, S., Day, N., Dhami, P., Dillon, S. C., Dorschner, M. O., Fiegler, H., Giresi, P. G., Goldy, J., Hawrylycz, M., Haydock, A., Humbert, R., James, K. D., Johnson, B. E., Johnson, E. M., Frum, T. T., Rosenzweig, E. R., Karnani, N., Lee, K., Lefebvre, G. C., Navas, P. A., Neri, F., Parker, S. C., Sabo, P. J., Sandstrom, R., Shafer, A., Vetrie, D., Weaver, M., Wilcox, S., Yu, M., Collins, F. S., Dekker, J., Lieb, J. D., Tullius, T. D., Crawford, G. E., Sunyaev, S., Noble, W. S., Dunham, I., Denoeud, F., Reymond, A., Kapranov, P., Rozowsky, J., Zheng, D., Castelo, R., Frankish, A., Harrow, J., Ghosh, S., Sandelin, A., Hofacker, I. L., Baertsch, R., Keefe, D., Dike, S., Cheng, J., Hirsch, H. A., Sekinger, E. A., Lagarde, J., Abril, J. F., Shahab, A., Flamm, C., Fried, C., Hackermuller, J., Hertel, J., Lindemeyer, M., Missal, K., Tanzer, A., Washietl, S., Korbel, J., Emanuelsson, O., Pedersen, J. S., Holroyd, N., Taylor, R., Swarbreck, D., Matthews, N., Dickson, M. C., Thomas, D. J., Weirauch, M. T., Gilbert, J., Drenkow, J., Bell, I., Zhao, X., Srinivasan, K. G., Sung, W. K., Ooi, H. S., Chiu, K. P., Foissac, S., Alioto, T., Brent, M., Pachter, L., Tress, M. L., Valencia, A., Choo, S. W., Choo, C. Y., Ucla, C., Manzano, C., Wyss, C., Cheung, E., Clark, T. G., Brown, J. B., Ganesh, M., Patel, S., Tammana, H., Chrast, J., Henrichsen, C. N., Kai, C., Kawai, J., Nagalakshmi, U., Wu, J., Lian, Z., Lian, J., Newburger, P., Zhang, X., Bickel, P., Mattick, J. S., Carninci, P., Hayashizaki, Y., Weissman, S., Hubbard, T., Myers, R. M., Rogers, J., Stadler, P. F., Lowe, T. M., Wei, C. L., Ruan, Y., Struhl, K., Gerstein, M., Antonarakis, S. E., Fu, Y., Green, E. D., Karaöz, U., Siepel, A., Taylor, J., Liefer, L. A., Wetterstrand, K. A., Good, P. J., Feingold, E. A., Guyer, M. S., Cooper, G. M., Asimenos, G., Dewey, C. N., Hou, M., Nikolaev, S., Montoya-Burgos, J. I., Löytynoja, A., Whelan, S., Pardi, F., Massingham, T., Huang, H., Zhang, N. R., Holmes, I., Mullikin, J. C., Ureta-Vidal, A., Paten, B., Seringhaus, M., Church, D., Rosenbloom, K., Kent, W. J., Stone, E. A., NISC Comparative Sequencing Program, Baylor College of Medicine Human Genome Sequencing Center, Washington University Genome Sequencing Center, Broad Institute, Children's Hospital Oakland Research Institute, Batzoglou, S., Goldman, N., Hardison, R. C., Haussler, D., Miller, W., Sidow, A., Trinklein, N. D., Zhang, Z. D., Barrera, L., Stuart, R., King, D. C., Ameur, A., Enroth, S., Bieda, M. C., Kim, J., Bhinge, A. A., Jiang, N., Liu, J., Yao, F., Vega, V. B., Lee, C. W., Ng, P., Shahab, A., Yang, A., Moqtaderi, Z., Zhu, Z., Xu, X., Squazzo, S., Oberley, M. J., Inman, D., Singer, M. A., Richmond, T. A., Munn, K. J., Rada-Iglesias, A., Wallerman, O., Komorowski, J., Fowler, J. C., Couttet, P., Bruce, A. W., Dovey, O. M., Ellis, P. D., Langford, C. F., Nix, D. A., Euskirchen, G., Hartman, S., Urban, A. E., Kraus, P., Van Calcar, S., Heintzman, N., Kim, T. H., Wang, K., Qu, C., Hon, G., Luna, R., Glass, C. K., Rosenfeld, M. G., Aldred, S. F., Cooper, S. J., Halees, A., Lin, J. M., Shulha, H. P., Zhang, X., Xu, M., Haidar, J. N., Yu, Y., Ruan, Y., Iyer, V. R., Green, R. D., Wadelius, C., Farnham, P. J., Ren, B., Harte, R. A., Hinrichs, A. S., Trumbower, H., Clawson H, Hillman-Jackson, J., Zweig, A. S., Smith, K., Thakkapallayil, A., Barber, G., Kuhn, R. M., Karolchik, D., Armengol, L., Bird, C. P., de Bakker, P. I., Kern, A. D., Lopez-Bigas, N., Martin, J. D., Stranger, B. E., Woodroffe, A., Davydov, E., Dimas, A., Eyras, E., Hallgrímsdóttir, I. B., Huppert, J., Zody, M. C., Abecasis, G. R., Estivill, X., Bouffard, G. G., Guan, X., Hansen, N. F., Idol, J. R., Maduro, V. V., Maskeri, B., McDowell, J. C., Park, M., Thomas, P. J., Young, A. C., Blakesley, R. W., Muzny, D. M., Sodergren, E., Wheeler, D. A., Worley, K. C., Jiang, H., Weinstock, G. M., Gibbs, R. A., Graves, T., Fulton, R., Mardis, E. R., Wilson, R. K., Clamp, M., Cuff, J., Gnerre, S., Jaffe, D. B., Chang, J. L., Lindblad-Toh, K., Lander, E. S., Koriabine, M., Nefedov, M., Osoegawa, K., Yoshinaga, Y., Zhu, B., and de Jong, P. J. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447, 799–816.

Braconi, C., Henry, J. C., Kogure, T., Schmittgen, T., and Patel, T. (2011). The role of microRNAs in human liver cancers. *Semin. Oncol.* 38, 752–763.

Brito, G. C., Fachel, A. A., Vettore, A. L., Vignal, G. M., Gimba, E. R., Campos, F. S., Barcinski, M. A., Verjovski-Almeida, S., and Reis, E. M. (2008). Identification of protein-coding and intronic noncoding RNAs down-regulated in clear cell renal carcinoma. *Mol. Carcinog.* 47, 757–767.

Bussemakers, M. J., Van Bokhoven, A., Verhaegh, G. W., Smit, F. P., Karthaus, H. F., Schalken, J. A., Debruyne, F. M., Ru, N., and Isaacs, W. B. (1999). DD3: a new prostate-specific gene, highly overexpressed in prostate cancer. *Cancer Res.* 59, 5975–5979.

Calin, G. A., Liu, C. G., Ferracin, M., Hyslop, T., Spizzo, R., Sevignani, C., Fabbri, M., Cimmino, A., Lee, E. J., Wojcik, S. E., Shimizu, M., Tili, E., Rossi, S., Taccioli, C., Pichiorri, F., Liu, X., Zupo, S., Herlea, V., Gramantieri, L., Lanza, G., Alder, H., Rassenti, L., Volinia, S., Schmittgen, T. D., Kipps, T. J., Negrini, M., and Croce, C. M. (2007). Ultraconserved regions encoding ncRNAs are altered in human leukemias and carcinomas. *Cancer Cell* 12, 215–229.

Cancer_Genome_Atlas_Research_Network. (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* 455, 1061–1068.

Cancer_Genome_Atlas_Research_Network. (2011). Integrated genomic analyses of ovarian carcinoma. *Nature* 474, 609–615.

Cocucci, E., Racchetti, G., and Meldolesi, J. (2009). Shedding microvesicles: artefacts no more. *Trends Cell Biol.* 19, 43–51.

Cowin, P. A., Anglesio, M., Etemadmoghadam, D., and Bowtell, D. D. (2010). Profiling the cancer genome. *Annu. Rev. Genomics Hum. Genet.* 11, 133–159.

Dinger, M. E., Amaral, P. P., Mercer, T. R., and Mattick, J. S. (2009). Pervasive transcription of the eukaryotic genome: functional indices and conceptual implications. *Brief. Funct. Genomic. Proteomic.* 8, 407–423.

Duttagupta, R., Jiang, R., Gollub, J., Getts, R. C., and Jones, K. W. (2011). Impact of cellular miRNAs on circulating miRNA biomarker signatures. *PLoS ONE* 6, e20769. doi:10.1371/journal.pone.0020769

Farazi, T. A., Spitzer, J. I., Morozov, P., and Tuschl, T. (2011). miRNAs in human cancer. *J. Pathol.* 223, 102–115.

Ferracin, M., Querzoli, P., Calin, G. A., and Negrini, M. (2011). MicroRNAs: toward the clinic for breast cancer patients. *Semin. Oncol.* 38, 764–775.

Fradet, Y., Saad, F., Aprikian, A., Dessureault, J., Elhilali, M., Trudel, C., Masse, B., Piche, L., and Chypre, C. (2004). uPM3, a new molecular urine test for the detection of prostate cancer. *Urology* 64, 311–315; discussion 315–316.

Gibb, E. A., Brown, C. J., and Lam, W. L. (2011). The functional role of long non-coding RNA in human carcinomas. *Mol. Cancer* 10, 38.

Guo, F., Li, Y., Liu, Y., Wang, J., and Li, G. (2010). Inhibition of metastasis-associated lung adenocarcinoma transcript 1 in CaSki human cervical cancer cells suppresses cell proliferation and invasion. *Acta Biochim. Biophys. Sin. (Shanghai)* 42, 224–229.

Gupta, R. A., Shah, N., Wang, K. C., Kim, J., Horlings, H. M., Wong, D. J., Tsai, M. C., Hung, T., Argani, P., Rinn, J. L., Wang, Y., Brzoska, P., Kong, B., Li, R., West, R. B., Van De Vijver, M. J., Sukumar, S., and Chang, H. Y. (2010). Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* 464, 1071–1076.

Guttman, M., Amit, I., Garber, M., French, C., Lin, M. F., Feldser, D., Huarte, M., Zuk, O., Carey, B. W., Cassady, J. P., Cabili, M. N., Jaenisch, R., Mikkelsen, T. S., Jacks, T., Hacohen, N., Bernstein, B. E., Kellis, M., Regev, A., Rinn, J. L., and Lander, E. S. (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* 458, 223–227.

Huarte, M., and Rinn, J. L. (2010). Large non-coding RNAs: missing links in cancer? *Hum. Mol. Genet.* 19, R152–R161.

Ji, P., Diederichs, S., Wang, W., Boing, S., Metzger, R., Schneider, P. M., Tidow, N., Brandt, B., Buerger, H., Bulk, E., Thomas, M., Berdel, W. E., Serve, H., and Muller-Tidow, C. (2003). MALAT-1, a novel noncoding RNA, and thymosin beta4 predict metastasis and survival in early-stage non-small cell lung cancer. *Oncogene* 22, 8031–8041.

Kapranov, P., Cheng, J., Dike, S., Nix, D. A., Duttagupta, R., Willingham, A. T., Stadler, P. F., Hertel, J., Hackermuller, J., Hofacker, I. L., Bell, I., Cheung, E., Drenkow, J., Dumais, E., Patel, S., Helt, G., Ganesh, M., Ghosh, S., Piccolboni, A., Sementchenko, V., Tammana, H., and Gingeras, T. R. (2007). RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* 316, 1484–1488.

Kapranov, P., St. Laurent, G., Raz, T., Ozsolak, F., Reynolds, C. P., Sorensen, P. H., Reaman, G., Milos, P., Arceci, R. J., Thompson, J. F., and Triche, T. J. (2010). The majority of total nuclear-encoded non-ribosomal RNA in a human cell is "dark matter" unannotated RNA. *BMC Biol.* 8, 149. doi:10.1186/1741-7007-8-149

Kim, W. Y., and Sharpless, N. E. (2006). The regulation of INK4/ARF in cancer and aging. *Cell* 127, 265–275.

Kogo, R., Shimamura, T., Mimori, K., Kawahara, K., Imoto, S., Sudo, T., Tanaka, F., Shibata, K., Suzuki, A., Komune, S., Miyano, S., and Mori, M. (2011). Long noncoding RNA HOTAIR regulates polycomb-dependent chromatin modification and is associated with poor prognosis in colorectal cancers. *Cancer Res.* 71, 6320–6326.

Kosaka, N., Iguchi, H., and Ochiya, T. (2010). Circulating microRNA in body fluid: a new potential biomarker for cancer diagnosis and prognosis. *Cancer Sci.* 101, 2087–2092.

Kotake, Y., Nakagawa, T., Kitagawa, K., Suzuki, S., Liu, N., Kitagawa, M., and

Xiong, Y. (2011). Long non-coding RNA ANRIL is required for the PRC2 recruitment to and silencing of p15(INK4B) tumor suppressor gene. *Oncogene* 30, 1956–1962.

Lai, M. C., Yang, Z., Zhou, L., Zhu, Q. Q., Xie, H. Y., Zhang, F., Wu, L. M., Chen, L. M., and Zheng, S. S. (2011). Long non-coding RNA MALAT-1 overexpression predicts tumor recurrence of hepatocellular carcinoma after liver transplantation. *Med. Oncol.* doi: 10.1007/s12032-011-0004-z. [Epub ahead of print].

Lawrie, C. H., Gal, S., Dunlop, H. M., Pushkaran, B., Liggins, A. P., Pulford, K., Banham, A. H., Pezzella, F., Boultwood, J., Wainscoat, J. S., Hatton, C. S., and Harris, A. L. (2008). Detection of elevated levels of tumour-associated microRNAs in serum of patients with diffuse large B-cell lymphoma. *Br. J. Haematol.* 141, 672–675.

Lee, G. L., Dobi, A., and Srivastava, S. (2011). Prostate cancer: diagnostic performance of the PCA3 urine test. *Nat. Rev. Urol.* 8, 123–124.

Lin, R., Maeda, S., Liu, C., Karin, M., and Edgington, T. S. (2007). A large noncoding RNA is a marker for murine hepatocellular carcinomas and a spectrum of human carcinomas. *Oncogene* 26, 851–858.

Lovat, F., Valeri, N., and Croce, C. M. (2011). MicroRNAs in the pathogenesis of cancer. *Semin. Oncol.* 38, 724–733.

Marks, L. S., and Bostwick, D. G. (2008). Prostate cancer specificity of PCA3 gene testing: examples from clinical practice. *Rev. Urol.* 10, 175–181.

Matouk, I. J., Abbasi, I., Hochberg, A., Galun, E., Dweik, H., and Akkawi, M. (2009). Highly upregulated in liver cancer noncoding RNA is overexpressed in hepatic colorectal metastasis. *Eur. J. Gastroenterol. Hepatol.* 21, 688–692.

Mattick, J. S. (2009). The genetic signatures of noncoding RNAs. *PLoS Genet.* 5, e1000459. doi:10.1371/journal.pgen.1000459

Mercer, T. R., Gerhardt, D. J., Dinger, M. E., Crawford, J., Trapnell, C., Jeddeloh, J. A., Mattick, J. S., and Rinn, J. L. (2012). Targeted RNA sequencing reveals the deep complexity of the human transcriptome. *Nat. Biotechnol.* 30, 99–104.

Mitchell, P. S., Parkin, R. K., Kroh, E. M., Fritz, B. R., Wyman, S. K., Pogosova-Agadjanyan, E. L., Peterson, A., Noteboom, J., O'Briant, K. C., Allen, A., Lin, D. W., Urban, N., Drescher, C. W., Knudsen, B. S., Stirewalt, D. L., Gentleman, R., Vessella, R. L., Nelson, P. S., Martin, D. B., and Tewari, M. (2008). Circulating microRNAs as stable blood-based markers for cancer detection. *Proc. Natl. Acad. Sci. U.S.A.* 105, 10513–10518.

Munker, R., and Calin, G. A. (2011). MicroRNA profiling in cancer. *Clin. Sci.* 121, 141–158.

Nakaya, H. I., Amaral, P. P., Louro, R., Lopes, A., Fachel, A. A., Moreira, Y. B., El-Jundi, T. A., Da Silva, A. M., Reis, E. M., and Verjovski-Almeida, S. (2007). Genome mapping and expression analyses of human intronic noncoding RNAs reveal tissue-specific patterns and enrichment in genes related to regulation of transcription. *Genome Biol.* 8, R43.

O'Driscoll, L., Kenny, E., Mehta, J. P., Doolan, P., Joyce, H., Gammell, P., Hill, A., O'Daly, B., O'Gorman, D., and Clynes, M. (2008). Feasibility and relevance of global expression profiling of gene transcripts in serum from breast cancer patients using whole genome microarrays and quantitative RT-PCR. *Cancer Genomics Proteomics* 5, 94–104.

Orozco, A. F., and Lewis, D. E. (2010). Flow cytometric analysis of circulating microparticles in plasma. *Cytometry A* 77, 502–514.

Panzitt, K., Tschernatsch, M. M., Guelly, C., Moustafa, T., Stradner, M., Strohmaier, H. M., Buck, C. R., Denk, H., Schroeder, R., Trauner, M., and Zatloukal, K. (2007). Characterization of HULC, a novel gene with striking up-regulation in hepatocellular carcinoma, as noncoding RNA. *Gastroenterology* 132, 330–342.

Paschoal, A. R., Maracaja-Coutinho, V., Setubal, J. C., Simões, Z. L. P., Verjovski-Almeida, S., and Durham, A. M. (2012). Non-coding transcription characterization and annotation: a guide and web resource for non-coding RNA databases. *RNA Biol.* 9. [Epub ahead of print].

Perez, D. S., Hoage, T. R., Pritchett, J. R., Ducharme-Smith, A. L., Halling, M. L., Ganapathiraju, S. C., Streng, P. S., and Smith, D. I. (2008). Long, abundantly expressed non-coding transcripts are altered in cancer. *Hum. Mol. Genet.* 17, 642–655.

Ponting, C. P., Oliver, P. L., and Reik, W. (2009). Evolution and functions of long noncoding RNAs. *Cell* 136, 629–641.

Prensner, J. R., and Chinnaiyan, A. M. (2011). The emergence of lncRNAs in cancer biology. *Cancer Discov.* 1, 391–407.

Prensner, J. R., Iyer, M. K., Balbin, O. A., Dhanasekaran, S. M., Cao, Q., Brenner, J. C., Laxman, B., Asangani, I. A., Grasso, C. S., Kominsky, H. D., Cao, X., Jing, X., Wang, X., Siddiqui, J., Wei, J. T., Robinson, D., Iyer, H. K., Palanisamy, N., Maher, C. A., and Chinnaiyan, A. M. (2011). Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. *Nat. Biotechnol.* 29, 742–749.

Record, M., Subra, C., Silvente-Poirot, S., and Poirot, M. (2011). Exosomes as intercellular signalosomes and pharmacological effectors. *Biochem. Pharmacol.* 81, 1171–1182.

Reis, E. M., Nakaya, H. I., Louro, R., Canavez, F. C., Flatschart, A. V., Almeida, G. T., Egidio, C. M., Paquola, A. C., Machado, A. A., Festa, F., Yamamoto, D., Alvarenga, R., Da Silva, C. C., Brito, G. C., Simon, S. D., Moreira-Filho, C. A., Leite, K. R., Camara-Lopes, L. H., Campos, F. S., Gimba, E., Vignal, G. M., El-Dorry, H., Sogayar, M. C., Barcinski, M. A., Da Silva, A. M., and Verjovski-Almeida, S. (2004). Antisense intronic non-coding RNA levels correlate to the degree of tumor differentiation in prostate cancer. *Oncogene* 23, 6684–6692.

Rinn, J. L., Kertesz, M., Wang, J. K., Squazzo, S. L., Xu, X., Brugmann, S. A., Goodnough, L. H., Helms, J. A., Farnham, P. J., Segal, E., and Chang, H. Y. (2007). Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* 129, 1311–1323.

Schetter, A. J., and Harris, C. C. (2011). Alterations of microRNAs contribute to colon carcinogenesis. *Semin. Oncol.* 38, 734–742.

Schwarzenbach, H., Hoon, D. S., and Pantel, K. (2011). Cell-free nucleic acids as biomarkers in cancer patients. *Nat. Rev. Cancer* 11, 426–437.

Shappell, S. B. (2008). Clinical utility of prostate carcinoma molecular diagnostic tests. *Rev. Urol.* 10, 44–69.

Tahira, A. C., Kubrusly, M. S., Faria, M. F., Dazzani, B., Fonseca, R. S., Maracaja-Coutinho, V., Verjovski-Almeida, S., Machado, M. C., and Reis, E. M. (2011). Long noncoding intronic RNAs are differentially expressed in primary and metastatic pancreatic cancer. *Mol. Cancer* 10, 141.

Tano, K., Mizuno, R., Okada, T., Rakwal, R., Shibato, J., Masuo, Y., Ijiri, K., and Akimitsu, N. (2010). MALAT-1 enhances cell motility of lung adenocarcinoma cells by influencing the expression of motility-related genes. *FEBS Lett.* 584, 4575–4580.

Tinzl, M., Marberger, M., Horvath, S., and Chypre, C. (2004). DD3PCA3 RNA analysis in urine – a new perspective for detecting prostate cancer. *Eur. Urol.* 46, 182–186; discussion 187.

Tong, Y. K., and Lo, Y. M. (2006). Diagnostic developments involving cell-free (circulating) nucleic acids. *Clin. Chim. Acta* 363, 187–196.

Tripathi, V., Ellis, J. D., Shen, Z., Song, D. Y., Pan, Q., Watt, A. T., Freier, S. M., Bennett, C. F., Sharma, A., Bubulya, P. A., Blencowe, B. J., Prasanth, S. G., and Prasanth, K. V. (2010). The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. *Mol. Cell* 39, 925–938.

Tsai, M. C., Manor, O., Wan, Y., Mosammaparast, N., Wang, J. K., Lan, F., Shi, Y., Segal, E., and Chang, H. Y. (2010). Long noncoding RNA as modular scaffold of histone modification complexes. *Science* 329, 689–693.

Wang, J., and Sen, S. (2011). MicroRNA functional network in pancreatic cancer: from biology to biomarkers of disease. *J. Biosci.* 36, 481–491.

Wang, K. C., and Chang, H. Y. (2011). Molecular mechanisms of long noncoding RNAs. *Mol. Cell* 43, 904–914.

Wright, M. W., and Bruford, E. A. (2011). Naming "junk": human non-protein coding RNA d gene nomenclature. *Hum. Genomics* 5, 90–98.

Yamada, K., Kano, J., Tsunoda, H., Yoshikawa, H., Okubo, C., Ishiyama, T., and Noguchi, M. (2006). Phenotypic characterization of endometrial stromal sarcoma of the uterus. *Cancer Sci.* 97, 106–112.

Yang, L., Duff, M. O., Graveley, B. R., Carmichael, G. G., and Chen, L. L. (2011a). Genomewide characterization of non-polyadenylated RNAs. *Genome Biol.* 12, R16.

Yang, Z., Zhou, L., Wu, L. M., Lai, M. C., Xie, H. Y., Zhang, F., and Zheng, S. S. (2011b). Overexpression of long non-coding RNA HOTAIR predicts tumor recurrence in hepatocellular carcinoma patients following liver transplantation. *Ann. Surg. Oncol.* 18, 1243–1250.

Yap, K. L., Li, S., Munoz-Cabello, A. M., Raguz, S., Zeng, L., Mujtaba, S., Gil, J., Walsh, M. J., and Zhou, M. M. (2010). Molecular interplay of the noncoding RNA ANRIL and methylated histone H3 lysine 27 by polycomb CBX7 in transcriptional silencing of INK4a. *Mol. Cell* 38, 662–674.

Yu, W., Gius, D., Onyango, P., Muldoon-Jacobs, K., Karp, J., Feinberg, A. P., and Cui, H. (2008). Epigenetic silencing of tumour suppressor gene p15 by its antisense RNA. *Nature* 451, 202–206.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 21 December 2011; paper pending published: 14 January 2012; accepted: 20 February 2012; published online: 08 March 2012.

Citation: Reis EM and Verjovski-Almeida S (2012) Perspectives of long non-coding RNAs in cancer diagnostics. Front. Gene. ***3****:32. doi: 10.3389/fgene.2012.00032*

This article was submitted to Frontiers in Non-Coding RNA, a specialty of Frontiers in Genetics.

REVIEW ARTICLE
published: 15 February 2012
doi: 10.3389/fgene.2012.00017

A central role for long non-coding RNA in cancer

Sheetal A. Mitra*[1,2], *Anirban P. Mitra*[2,3] and *Timothy J. Triche*[1,2,3]

[1] Department of Pathology and Laboratory Medicine, Children's Hospital Los Angeles, Los Angeles, CA, USA
[2] Center for Personalized Medicine, Children's Hospital Los Angeles, Los Angeles, CA, USA
[3] Department of Pathology, Keck School of Medicine of the University of Southern California, Los Angeles, CA, USA

Edited by:
Philipp Kapranov, St. Laurent Institute, USA

Reviewed by:
Robert Arceci, Johns Hopkins, USA
Tim McCaffrey, The George Washington University Hospital, USA

****Correspondence:***
Timothy J. Triche, Department of Pathology and Laboratory Medicine, Children's Hospital Los Angeles, 4650 Sunset Blvd., SRT 1016, MS 133, Los Angeles, CA 90027, USA.
e-mail: triche@usc.edu

Long non-coding RNAs (ncRNAs) have been shown to regulate important biological processes that support normal cellular functions. Aberrant regulation of these essential functions can promote tumor development. In this review, we underscore the importance of the regulatory role played by this distinct class of ncRNAs in cancer-associated pathways that govern mechanisms such as cell growth, invasion, and metastasis. We also highlight the possibility of using these unique RNAs as diagnostic and prognostic biomarkers in malignancies.

Keywords: non-coding RNA, tumor, diagnosis, prognosis, development

INTRODUCTION

Common knowledge until recently would suggest that only about 1% of the genome produces biologically meaningful RNA transcripts, specifically those that encode for proteins (Guttman et al., 2009). In the past decade, however, numerous papers have appeared that clearly document widespread transcription across most of the genome. However, the biological relevance of this transcription has been a matter of much debate, as witnessed by numerous original peer-reviewed articles, reviews, letters to the editor, and rebuttals. In particular, van Bakel and Hughes believe this pervasive transcription is simply transcriptional "noise," while Mattick and Kapranov, among others, are of the strong belief these transcripts are functionally relevant (Kapranov, 2009; Mercer et al., 2009; van Bakel and Hughes, 2009). An in-depth review of published data in combination with unpublished observations from our work in this field has led to support the latter interpretation, based on a strong association of patterns of non-coding RNA (ncRNA) transcription with diagnosis and prognosis in cancer. We believe these data can only be interpreted as non-random, biologically meaningful patterns that point to a general functional role for ncRNA over much of the genome. Further, the character and pattern of transcription is itself of considerable interest, as there is no parallel among annotated, protein-encoding genes for many of the transcripts we have identified in several different types of cancer.

The findings reported here are consonant with several historical observations. First, as early as 1980, in a landmark publication, "The RNA World," Watson, Crick, and others present compelling evidence that RNA was the first nucleic acid associated with life on this planet (Atkins et al., 2011). Several lines of evidence converge on the idea that DNA, and subsequently protein-encoding RNA, only appeared eons later. Prior to that time, RNA subserved the functions necessary for life currently associated with DNA (organismal memory) and protein (enzymatic cleavage, regulation of transcription, and many others). This "RNA First" hypothesis is widely accepted now, but the implications for genome-wide pervasive RNA transcription have only recently garnered attention. Most students today are still taught the central dogma of molecular biology, namely that DNA encodes RNA transcripts that are transcribed from the genome and translated into protein after cleavage and migration from the nucleus to the cytoplasm, where the ribosome utilizes the messenger RNA (mRNA) strand as a specific template to produce protein. With the vast amounts and variety of ncRNA transcripts found in every cell at all stages of development, this central dogma does not fully capture the role of RNA as a regulatory molecule independent of protein (Pauli et al., 2011; Suh and Blelloch, 2011).

This review provides evidence that these non-coding transcripts as a group are not only functional, but may well be some of the most basic and ancient functional RNAs of all. Certainly they are strikingly different than mRNA, in both structure and function, yet certain features are shared with mRNA, including, in some cases, "exons" and "introns," poly-adenylation, and alternate splice variants. Conversely, unlike coding genes, ncRNAs occur in poorly conserved regions of the genome, in stark contrast to the highly conserved regions associated with coding genes.

Beyond these easily documented features, relatively little is known about the general structure, function, and transcriptional control of ncRNAs, and even less about their potential functions as a group or singly. Unlike coding genes, there is no easily documented protein product linked to the RNA sequence (Clamp et al., 2007). However, numerous individual examples have been documented, with very different structure and function. On the one hand, *H19*, a long recognized ncRNA, shows significant sequence conservation and is known to control *IGF2* on the opposite paternal chromosome by epigenetic control mechanisms (Feil et al., 1994; Juan et al., 2000). On the other hand, *XIST* shows almost no sequence conservation between species yet consistently silences one X chromosome in females (Hendrich et al., 1993; Panning et al., 1997). In yet another well documented example, a ncRNA transcribed from the HOX locus, *HOTAIR*, has been shown to bind to the polycomb repressor complex 2 (PRC2) and affect expression of over 300 coding gene targets as part of a general reprogramming of breast cancer cells from a locally aggressive epithelial phenotype to an invasive and metastatic "mesenchymal" phenotype (Gupta et al., 2010). Clearly there are specific examples of functional ncRNA. There remains the larger question though of whether ncRNAs are predominantly functional, and if so, how this might be determined. In the following examples, we present evidence that ncRNAs are in fact strongly associated with cancer diagnosis and prognosis, functions that can hardly be ascribed to genome-wide random transcription.

CLASSES OF ncRNA TRANSCRIPTION

Non-coding RNAs are an integral part of the mammalian transcriptome. Once described as "dark matter," these underestimated molecules can play important functional and structural roles in the cell (Kapranov et al., 2010; Qureshi and Mehler, 2011). Based on size, these RNA can be grouped into three major classes: small ncRNAs, which include microRNA (miRNA), PIWI-interacting RNA (piRNA), endogenous short interfering RNA (siRNA), and other non-coding transcripts of less than 200 nucleotides (nt); long ncRNAs (lncRNA) that are greater than 200 nt and arise from intergenic regions or are organized around protein-coding regions; and very long ncRNA (vlncRNA) that can stretch through hundreds of kilobases, often across intergenic regions. While miRNAs post-transcriptionally regulate mRNA through the RNA-induced silencing complex, piRNA and siRNA are implicated in maintaining genomic integrity by silencing of transposable elements in cells. lncRNAs are involved in various levels of genomic regulation and related fundamental epigenetic processes: genomic imprinting, dosage compensation, and chromatin modifications. These can assist in subcellular transport, recruitment of transcription factors, and RNA processing and editing by forming ribonucleoprotein complexes.

CHARACTERISTICS OF ncRNA

POLY-ADENYLATED RNA VERSUS TOTAL RNA TRANSCRIPTION

Non-coding genomic DNA, some of which is genetically functional, has increased proportionally with genomic size and complexity (Taft et al., 2007). The human genome has more "dark matter" when compared to that of *Drosophila*. RNA polymerase II transcribes both protein-coding and non-coding transcripts. While transcription terminates at a poly-adenylation site for most protein-coding genes, there is substantial evidence that a fraction of ncRNAs do not necessarily end with a poly-A signal. This alternate termination of transcripts is sometimes associated with RNA-binding proteins Nrd1, Nab3, and Sen1 as is seen with non-poly-adenylated end of small nucleolar RNAs (snoRNAs) in *S. cerevisiae* (Creamer et al., 2011). Antisense *asOct4-pg5* or the brain-associated *BC200* are examples of functional lncRNA that are not poly-adenylated (Chen et al., 1997; Hawkins and Morris, 2010). ncRNA comprise approximately 30% of the total poly-A fraction, while they account for approximately 50–60% of total RNA with or devoid of ribosomal RNA, thus suggesting that a significant amount of these ncRNAs are non-poly-adenylated (Kapranov et al., 2010). The lack of poly-A tails has caused these transcripts to be underrepresented in cDNA libraries, SAGE, differential display, and microarrays which typically employ a 3′ poly-A labeling method.

CONSERVED VERSUS NOT CONSERVED

While protein-coding genes are under high constraint, this is not the case with all ncRNAs. Recent studies have shown the emerging importance of lncRNAs as regulators of essential cellular functions that involve a great number of protein interactions (Guttman et al., 2009). Increased system complexity that enables highly skilled functions increases evolutionary pressure on regulators of this dynamic signaling network (Mattick, 2003). These RNAs are predicted to undergo more rapid evolution than pre-existing proteins or the *de novo* evolution of a unique set of signaling molecules (Ponjavic et al., 2007). Guttman et al. (2010) compared orthologous sequences of lncRNAs among 29 mammals and showed that their conservation is far greater than random genomic sequences or introns. On the other hand, Babak et al. (2005) found poor conservation between intergenic genomic transcripts and proposed that they may thus be non-functional. However, one of the non-coding elements of the genome, referred to as ultra-conserved elements (UCEs), is highly conserved. These regions span at least 200 base pairs in length and maintain 100% identity with no insertions or deletions between human, mouse, and rat genomes (Bejerano et al., 2004). Other than the exonic components, there are about 38.7% UCEs that are intergenic while another 42.6% are intronic (Mestdagh et al., 2010). The average distances observed among them (approximately 10 Mb) suggest that they are unlikely to function as exons of a gene. Some of these non-coding UCEs are transcribed (T-UCEs) and maintain evolutionary constraints.

FUNCTIONAL VERSUS NON-FUNCTIONAL

Less than 1% of lncRNAs have been associated with a function. Their cell- and tissue-specific expression that changes in response to external factors such as stress and other environmental signals implies that their presence is dependent on the need of the cell. Many of these lncRNAs have binding sites for transcription factors Sp1, c-Myc, p53, and Creb, thus suggesting different levels of regulation (Cawley et al., 2004; Euskirchen et al., 2004). Their involvement varies from transcriptional to post-transcriptional regulation to translational control. There is evidence that some of these are essential for development (Rosenbluh et al., 2011; Han et al., 2012). For example, *Mirg*, a maternal ncRNA from the Dlk–Dio3 imprinted cluster, is expressed in different tissues at different time during murine embryonic development (Han et al., 2012).

GENERAL PATTERNS OF ncRNA EXPRESSION IN NORMAL TISSUES AND CANCER

The concept of a functional genome is being rewritten with the discovery of ncRNA. The abundance of these transcripts in cancer suggests their role in tumor pathogenesis. ncRNAs are abundant during embryogenesis (van Leeuwen and Mikkers, 2010; Pauli et al., 2011) and reactivation or non-suppression of some of these fetal lncRNAs may critically regulate pluripotency and uninhibited cellular growth, thus giving rise to adult or developmental cancers. For example, the *H19* lncRNA is expressed during vertebrate embryogenesis but is downregulated after birth in most tissues except for cartilage and skeletal muscle (Lustig et al., 1994). However, loss of imprinting and overexpression of *H19* in many cancers such as those of esophagus, liver, colon, and bladder cause it to function as an oncogene and promote tumor development (Hibi et al., 1996; Barsyte-Lovejoy et al., 2006; Matouk et al., 2007). Similarly, normal adult tissues express lncRNAs at various levels with lymph nodes and gall bladder reportedly having the most distinct lncRNAs (Gibb et al., 2011). Comparisons between normal and cancerous tissues revealed differential expression of at least 200 lncRNAs. The chromosome distribution of lncRNAs did not correlate with either protein-coding genes or miRNAs. Kapranov et al. (2010) also showed that in Ewing sarcoma, a childhood cancer, 43–63% of all non-ribosomal, non-mitochondrial RNAs by mass were non-exonic RNAs, and 24–37% of these were detected in intergenic regions. This study also suggested the presence of a vlncRNA of approximately 650 kb on chromosome 7 that was exclusively present in Ewing sarcoma and not in the leukemia cell line K562, normal brain, or liver. Similarly, another 300 kb intergenic region on chromosome 21 in the K562 cell line was not detected in Ewing sarcoma, suggesting that certain ncRNAs may be present in specific cancers.

ROLE OF ncRNAs IN TUMOR PATHOGENESIS: ONCOGENES OR TUMOR SUPPRESSORS

ncRNAs have been detected in cancer by various techniques including expression microarrays, tiling arrays, next generation sequencing, and methylation analysis (Cheung et al., 2010; Gupta et al., 2010; Sang et al., 2010; Trapnell et al., 2010). These approaches have led to the identification of several lncRNAs whose expression and epigenetic state are significantly associated with cancerous tissues.

Like protein-coding genes, ncRNAs may function as tumor oncogenes or tumor suppressors. Some T-UCEs are frequently located at fragile sites and cancer-associated genomic regions (CAGRs) such as minimal regions of amplification and of loss of heterozygosity, while others are known to act as oncogenes in cancer cells (Rossi et al., 2008). Functional analysis involving siRNAs identified *uc.73A* as a promoter of cell survival by evading cellular apoptosis in colorectal cancer (Calin et al., 2007). Enrichment analyses confirmed that UCEs are contained in genes involved in RNA processing and RNA binding (Licastro et al., 2010). They bear resemblance to enhancer-like sequences and are involved in transcription.

Protein-coding genes are known to be associated with antisense transcripts, and perturbation of these can alter protein expression that promotes cancer development (He et al., 2008). Antisense transcripts *ANRIL* and *p21*/*CDKN1A*-associated transcript repress tumor suppressor loci and promote cancer (Morris et al., 2008). Aberrant gene expression causes changes in chromatin structure leading to genomic instability that can give rise to uncontrollable growth and an invasive cellular phenotype. Therefore, proteins that control chromatin organization including polycomb repressor complexes, PRC1 and PRC2, and members of the trithorax family constitute key players in the molecular pathogenesis of cancer. Selective binding of lncRNAs, *HOTAIR* and *ANRIL*, with PRC1 and PRC2 to execute histone modifications at specific loci thus strongly supports the idea that lncRNAs may function as ideal regulators for epigenetic transcriptional repression (Gupta et al., 2010; Kotake et al., 2011). *ANRIL* and associated factors play critical roles in repression of the *INK4b–ARF–INK4a* locus that encodes for three critical tumor suppressors, $p15^{INK4b}$, $p14^{ARF}$ ($p19^{ARF}$ in mice), and $p16^{INK4a}$, which play central roles in cell-cycle inhibition, senescence, and stress-induced apoptosis (Pasmant et al., 2007; Yap et al., 2010; Kotake et al., 2011).

Long ncRNAs may also act as tumor suppressors. They may inhibit cell-cycle progression in response to DNA damage due to stress and environmental factors. lncRNA ncRNA$_{CCND1}$ is induced during DNA damage from the *CCND1* promoter (Wang et al., 2008). This lncRNA recruits the TLS protein to the *CCND1* promoter where it binds to histone acetyltransferases CBP/p300 and in turn inhibits *CCND1* transcription thus affecting cell-cycle progression. Some lncRNAs may inhibit growth in cancer cells. *MEG3*, a lncRNA that is expressed in many normal tissues but not in human cancer cell lines, may function as a tumor suppressor as its ectopic expression in cancer cells suppressed their growth (Zhang et al., 2003).

ASSOCIATIONS OF lncRNAs WITH CANCER

Genome-wide association studies of cancer susceptibility have identified single nucleotide polymorphisms (SNPs) in some of the transcribed regions of the non-coding portions of the human genome (Manolio et al., 2008). T-UCEs differentially expressed in human cancers are located in CAGRs that are specifically associated with that type of cancer (Calin et al., 2007). These could be candidate players for cancer susceptibility. For example, differential expressions of *uc.349A* and *uc.352* between normal and leukemic CD5-positive cells have been linked to susceptibility to familial chronic lymphocytic leukemia (Ng et al., 2007). Consistent with these findings, Yang et al. (2008) have reported that two SNPs in UCEs (rs9572903 and rs2056116) are associated with familial breast cancer risk. Recently, Pasmant et al. (2011) have also shown that modulation of *ANRIL* levels in patients with neurofibromatosis mediates susceptibility to plexiform neurofibromas. SNP rs2151280 located in *ANRIL* locus was statistically significantly associated with number of plexiform neurofibromas in these patients.

ncRNAs AND CANCER DIAGNOSIS

The differences in lncRNA profiling between normal and cancer cells may or may not be a mere secondary effect of cancerous transformations. Several lncRNAs can control transcriptional alteration, as seen with *ANRIL* and its interaction with PRC proteins that leads to repression of *INK4b* locus, a change observed in most cancers (Kotake et al., 2011). In other cases, altered expression of these RNAs may show a strong association with tumor progression, and thus can be used as classification markers for these malignancies. Most lncRNAs are expressed in various types of cancers; however, some have been associated with specific tumor types. A striking example is that of three lncRNAs in prostate cancer: *PCGEM1*, *DD3*, and *PCNCR1* (Bussemakers et al., 1999; Petrovics et al., 2004; Chung et al., 2011). These lncRNAs either promote tumorigenicity or are associated with susceptibility to prostate adenocarcinoma. These unique lncRNAs could therefore potentially be used for prostate cancer diagnosis. The malignant cells have a unique spectrum of expressed UCEs when compared with the corresponding normal cells, suggesting that variations in T-UCE expression are involved in the malignant process. Moreover, distinct T-UCE signatures were differentially expressed in leukemias and carcinomas, and thus may offer a novel strategy for cancer diagnosis and prognosis (Calin et al., 2007).

Our experience with childhood tumors have led us to believe that ncRNAs play key roles in defining tumor subtypes (Bajaj et al., 2011). We have performed several exploratory analyses in pediatric tumors that provide evidence of unique non-coding intergenic regions that are characteristic of tumor types. One such preliminary analysis depicted in **Figure 1** involved 40 unique primary tumors from patients with *PAX–FKHR* fusion-positive rhabdomyosarcoma ($n = 10$), fusion-negative rhabdomyosarcoma ($n = 10$), Ewing family of tumors (EFT, $n = 5$), osteosarcoma ($n = 5$), neuroblastoma ($n = 5$), and Wilms' tumors ($n = 5$). An unsupervised nearest shrunken centroid model, a class prediction procedure that identifies transcripts that best characterize tumor subtypes, was used to analyze whole-transcriptome expression profiling data obtained from these tumors using Affymetrix Human Exon 1.0 ST microarrays. This procedure eliminates classifier transcripts from the prediction signature as the shrinkage parameter (Δ) increases, thereby creating highly class-specific profiles (Tibshirani et al., 2002). This revealed the presence of several classifier coding and non-coding transcripts, represented as probe set regions (PSRs) on the top histogram of **Figure 1A**, that were able to categorize tumors in the training (aqua line) and test (gold line) sets with 100 and 95% accuracy at $\Delta = 5.6$, respectively. Examination of features contained in the centroid classes revealed the presence of a 250-kb stretch of non-coding transcript (locus marked by dashed black box in **Figure 1C**), a putative vlncRNA, which was unique to EFTs (tumor class 3 in **Figure 1C**). This tumor subgroup uniquely showed marked overexpression of this genomic stretch that does not code for any known proteins (aqua trace in **Figure 1D**); none of the other childhood tumors examined in this cohort appeared to express this vlncRNA at levels comparable to EFT. This demonstrates that the presence of such transcripts, if found on a larger scale with similar discriminatory power, may be extremely helpful in diagnosing such tumor types. In addition, it also suggests that such non-coding transcripts may play a role in the genesis and maintenance of these malignancies.

ncRNAs AND CANCER PROGNOSIS

Differential expressions of protein-coding genes and small ncRNAs between cancers have been used as a valuable tool to generate signatures that can reliably predict disease outcomes (Martens-Uzunova et al., 2012). A panel of 10 biomarkers that included 8 protein-coding genes and 2 miRNAs, *miR-519d* and *miR-647,* could significantly predict clinical recurrence in prostate cancer following radical prostatectomy (Long et al., 2011).

With recent growing evidence of similar expression patterns of lncRNAs in cancers, these transcripts may be profiled as prognostic candidates. A similar strategy may be adopted to develop lncRNA-dependent gene signatures that may predict disease outcomes and response to treatments. The lncRNA *MALAT1* is upregulated in many solid tumors and is associated with cancer metastasis and recurrence. In hepatocellular carcinoma, *MALAT1* levels corresponded to advanced disease stage and were inversely related to disease-free survival after liver transplantation (Lai et al., 2012). Similarly, an expression profile based on 28 T-UCEs in 14 patients with neuroblastoma was able to significantly distinguish between short-term and long-term survivors (Scaruffi et al., 2009).

Our group's efforts in identifying non-coding transcripts that are associated with outcome have focused on childhood tumors. In one such analysis shown in **Figure 2**, we initially analyzed Affymetrix Human Exon 1.0 ST array-derived whole-transcriptome expression profiling data on primary EFT samples from 40 patients at surgical resection with long subsequent follow-up. Thirteen (32.5%) patients eventually metastasized (depicted in red in **Figures 2A–D**). An unsupervised nearest shrunken centroid model was used to identify coding and non-coding features that could categorize these tumors based on their probability of eventually metastasizing. At $\Delta = 1.2$, several features were identified that could categorize the tumors into two groups based on risk of metastasis with 92.5 and 70% accuracies in the training and test sets, respectively (**Figure 2A**).

To further investigate the biological implications of ncRNA features that could predict tumor metastasis, expression profiles on two EFT cell lines, CHLA-9 and CHLA-10, were analyzed using a similar nearest shrunken centroid model. At $\Delta = 6.0$, the selected features were able to classify samples in the training and test sets with 100% accuracy (**Figure 2C**). When this set of classifier features was compared to those obtained from the analysis of the above EFT samples, a unique 26 kb intergenic non-coding transcript was identified on chromosome 2 (dashed black box in **Figures 2B,D**). The expression of this transcript was seemingly protective in nature – its expression was highest in primary tumors that did not metastasize, and lower in those primary tumors that eventually metastasized (**Figure 2E**). Following this trend, its expression was comparably lower in CHLA-9, a cell line generated from the primary tumor of an EFT patient, and lowest in CHLA-10, a cell line generated from a subsequent metastatic tumor in the same patient (Batra et al., 2004). Such observations provide credence to the argument that non-coding transcripts play crucial roles in the modulation of tumor behavior and can be used as markers in the primary malignancy to determine long-term prognosis.

ncRNAs: BRIDGING NORMAL TISSUE DEVELOPMENT AND ONCOGENESIS

The data and studies presented here offer compelling evidence that transcription of ncRNAs in cancer is tightly linked to key biological processes, from differentiation to metastasis. The parallel with normal tissue differentiation during fetal development is striking and reminiscent of another well documented phenomenon in cancer: to reprise the expression of fetal antigens during oncogenesis. Given the documented higher levels of ncRNA transcription during normal tissue development, it should be no surprise that ncRNA levels in cancer are elevated compared to normal tissue development. Many parallels between oncogenesis and development are well known, such that oncogenesis is often viewed as a poorly executed mimicry of normal tissue development. Environmental influences may allow embryonic expression of lncRNAs in adult tissues that alter gene expression, thereby increasing cancer

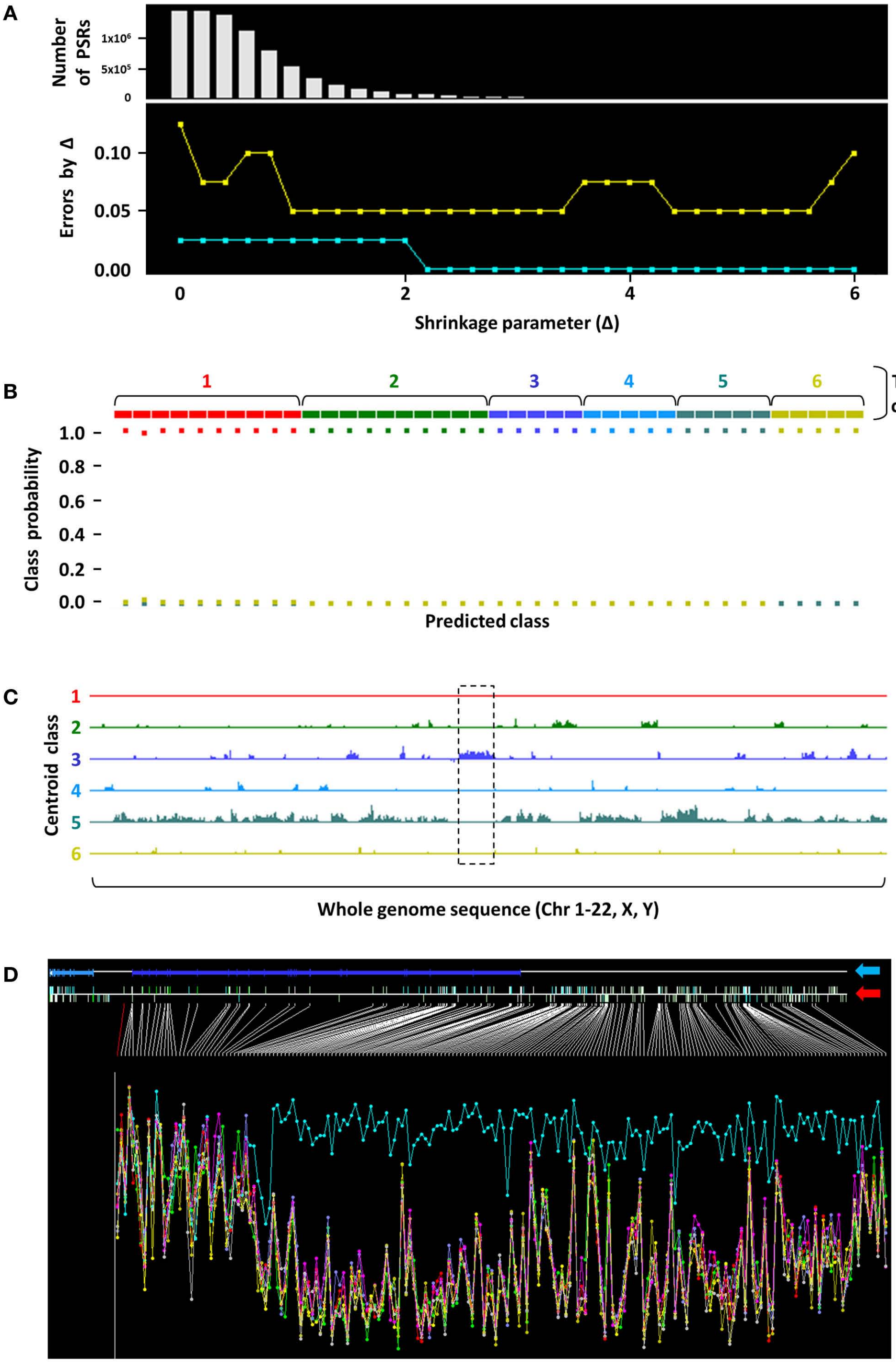


FIGURE 1 | Nearest shrunken centroid analysis to identify a putative EFT-specific vlncRNA. (A) Nearest shrunken centroid modeling was performed on 40 unique primary childhood tumors. Shrinkage parameter (*X*-axis) Δ = 5.6 was selected as the threshold where the fewest number of PSRs (*Y*-axis, top panel) were required to categorize tumors in the training (aqua line) and test (gold line) sets with 0 and 5% error, respectively (*Y*-axis, bottom panel). **(B)** Classification performance of training set samples is shown, where probability of samples belonging to each color-coded tumor class (1, *PAX–FKHR* fusion-positive rhabdomyosarcoma; 2, fusion-negative rhabdomyosarcoma; 3, EFT; 4, osteosarcoma; 5, neuroblastoma; 6, Wilms' tumors) was predicted with 100% accuracy at Δ = 5.6. Note that only squares of the like color are found at the 100% probability level in each true class. **(C)** Whole-genome plot of positions of the diagnostic PSRs (*X*-axis) that characterize the respective tumor groups versus their expression levels (*Y*-axis). A 250-kb stretch corresponding to a putative vlncRNA region (dashed black box) was observed as being uniquely overexpressed in EFT. **(D)** When zoomed in at this genomic segment (blue arrow points to the RefSeq annotation; red arrow indicates positions of PSRs across the region), evidence of significant overexpression of this transcript in EFTs (aqua trace) was clear compared to other childhood tumor types. Height of the *Y*-axis corresponds to the logarithm of PSR expression levels, and samples are aggregated into their respective tumor groups.

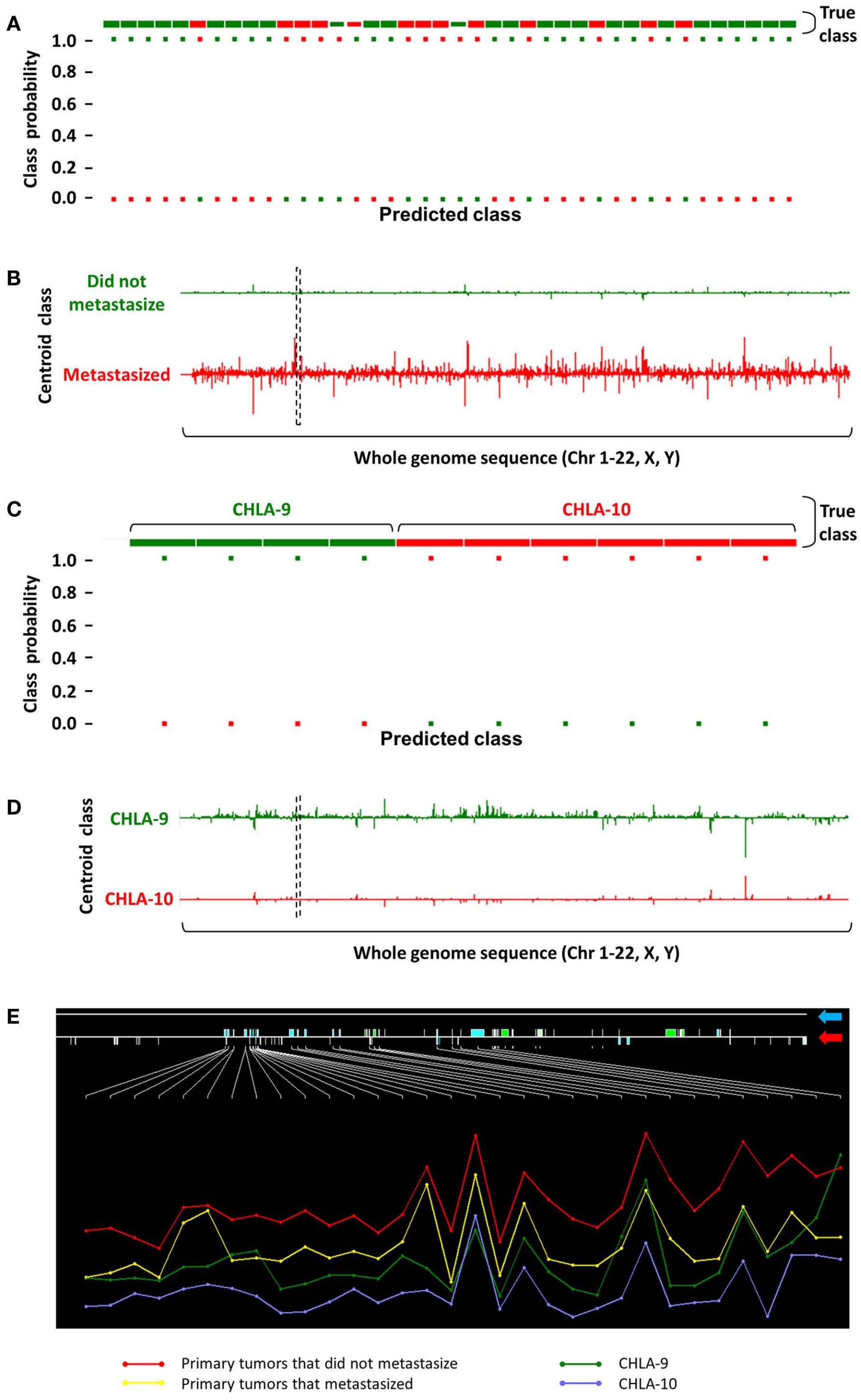


FIGURE 2 | Identification of a non-coding transcript showing differential expression in EFTs with respect to metastasis. (A) Classification performance of a nearest shrunken centroid model is shown, where 40 primary EFTs were categorized based on their eventual metastatic fate (green, did not metastasize; red, eventually metastasized) in the training set with 92.5% accuracy at $\Delta = 1.2$. **(B)** PSRs identified by this analysis that distinguish between non-metastasized versus metastasized groups are plotted over a whole-genome sequence, where height of the *Y*-axis over and under the baseline corresponds to their log fold change. **(C)** A similar nearest shrunken centroid analysis on CHLA-9 and CHLA-10 achieved 100% classification accuracy at $\Delta = 6.0$. **(D)** Comparing the PSR profiles between both nearest shrunken centroid models resulted in the identification of a common 26 kb intergenic non-coding transcript [dashed black box in **(B)** and **(D)**]. **(E)** A zoomed in inspection of this genomic segment (blue arrow points to the RefSeq annotation; red arrow indicates positions of PSRs across the region) showed that the transcript was highly expressed in tumors that never metastasized, moderately expressed in tumors that eventually metastasized and CHLA-9, and showed low expression in CHLA-10. Height of the *Y*-axis corresponds to the logarithm of PSR expression levels, and samples are aggregated into their respective tumor groups.

susceptibility. The chromatin-interacting ncRNA *KCNQ1OT1* causes imprinting of *CDKN1C* gene in embryonic tissues (Lewis et al., 2004). *CDKN1C* gene expression is suppressed in breast cancers by estrogen through epigenetic mechanisms involving the highly expressed *KCNQ1OT1* gene (Rodriguez et al., 2011). It is therefore not unreasonable to deduce that ncRNA expression is of fundamental importance, to the extent that ncRNA expression may well control coding RNA expression, using the latter to execute complex and fundamental programs responsible for organismal development. From a combined viewpoint, therefore, ncRNA is primary and coding RNA is secondary. The fact that a ncRNA gene like *HOTAIR* can orchestrate the expression of over 300 coding genes via complex formation with PRC2 and epigenetic regulation, leading to altered tumor cell differentiation and behavior, is entirely consistent with this concept. It will not be surprising, therefore, if a general pattern of ncRNA control of coding gene expression emerges from the many current studies on ncRNAs.

Beyond simple primary–secondary control mechanisms, it also appears that ncRNA itself is likely tightly regulated in an interactive network (Sumazin et al., 2011). This model of self-regulating RNA networks is intuitively attractive, as it allows for a degree of subtle control via multiple interacting regulatory networks that is essential to account for the development of higher organisms such as humans. The observation that ncRNA expression levels are highest in developing brain is consonant with this concept. The challenge going forward will be to unravel and understand these complex interactions. The reward will almost certainly be a far more sophisticated understanding of how biology works, and by extension, how it is perturbed in cancer.

CONCLUSION

This review provides some evidence of the multifaceted roles of lncRNAs in cancer. It underscores the importance of the functional existence of these transcripts that are proving to be much more than "transcriptional noise." Understanding their biological relevance in normal development may provide an insight into their perturbed functions in cancer. This will allow use of these enigmatic molecules as diagnostic or predictive biomarkers. They may be further developed into cancer-specific RNA targets to improve treatment sensitivity for various malignancies.

REFERENCES

Atkins, J. F., Gesteland, R. F., and Cech, T. R. (2011). *RNA Worlds: From Life's Origins to Diversity in Gene Regulation.* Woodbury, NY: Cold Spring Harbor Laboratory Press.

Babak, T., Blencowe, B. J., and Hughes, T. R. (2005). A systematic search for new mammalian noncoding RNAs indicates little conserved intergenic transcription. *BMC Genomics* 6, 104. doi:10.1186/1471-2164-6-104

Bajaj, S. V., Wai, D. H., Buckley, J. D., Kapranov, P., Lawlor, E. R., and Triche, T. J. (2011). A large non-coding RNA that is characteristic of Ewing sarcoma family of tumors. *Paper Presented at 102nd Annual Meeting of the American Association for Cancer Research*, Orlando, FL: American Association for Cancer Research.

Barsyte-Lovejoy, D., Lau, S. K., Boutros, P. C., Khosravi, F., Jurisica, I., Andrulis, I. L., Tsao, M. S., and Penn, L. Z. (2006). The c-Myc oncogene directly induces the H19 noncoding RNA by allele-specific binding to potentiate tumorigenesis. *Cancer Res.* 66, 5330–5337.

Batra, S., Reynolds, C. P., and Maurer, B. J. (2004). Fenretinide cytotoxicity for Ewing's sarcoma and primitive neuroectodermal tumor cell lines is decreased by hypoxia and synergistically enhanced by ceramide modulators. *Cancer Res.* 64, 5415–5424.

Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W. J., Mattick, J. S., and Haussler, D. (2004). Ultraconserved elements in the human genome. *Science* 304, 1321–1325.

Bussemakers, M. J., van Bokhoven, A., Verhaegh, G. W., Smit, F. P., Karthaus, H. F., Schalken, J. A., Debruyne, F. M., Ru, N., and Isaacs, W. B. (1999). DD3: a new prostate-specific gene, highly overexpressed in prostate cancer. *Cancer Res.* 59, 5975–5979.

Calin, G. A., Liu, C. G., Ferracin, M., Hyslop, T., Spizzo, R., Sevignani, C., Fabbri, M., Cimmino, A., Lee, E. J., Wojcik, S. E., Shimizu, M., Tili, E., Rossi, S., Taccioli, C., Pichiorri, F., Liu, X., Zupo, S., Herlea, V., Gramantieri, L., Lanza, G., Alder, H., Rassenti, L., Volinia, S., Schmittgen, T. D., Kipps, T. J., Negrini, M., and Croce, C. M. (2007). Ultraconserved regions encoding ncRNAs are altered in human leukemias and carcinomas. *Cancer Cell* 12, 215–229.

Cawley, S., Bekiranov, S., Ng, H. H., Kapranov, P., Sekinger, E. A., Kampa, D., Piccolboni, A., Sementchenko, V., Cheng, J., Williams, A. J., Wheeler, R., Wong, B., Drenkow, J., Yamanaka, M., Patel, S., Brubaker, S., Tammana, H., Helt, G., Struhl, K., and Gingeras, T. R. (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, W., Bocker, W., Brosius, J., and Tiedge, H. (1997). Expression of neural BC200 RNA in human tumours. *J. Pathol.* 183, 345–351.

Cheung, H. H., Lee, T. L., Davis, A. J., Taft, D. H., Rennert, O. M., and Chan, W. Y. (2010). Genome-wide DNA methylation profiling reveals novel epigenetically regulated genes and non-coding RNAs in human testicular cancer. *Br. J. Cancer* 102, 419–427.

Chung, S., Nakagawa, H., Uemura, M., Piao, L., Ashikawa, K., Hosono, N., Takata, R., Akamatsu, S., Kawaguchi, T., Morizono, T., Tsunoda, T., Daigo, Y., Matsuda, K., Kamatani, N., Nakamura, Y., and Kubo, M. (2011). Association of a novel long non-coding RNA in 8q24 with prostate cancer susceptibility. *Cancer Sci.* 102, 245–252.

Clamp, M., Fry, B., Kamal, M., Xie, X., Cuff, J., Lin, M. F., Kellis, M., Lindblad-Toh, K., and Lander, E. S. (2007). Distinguishing protein-coding and noncoding genes in the human genome. *Proc. Natl. Acad. Sci. U.S.A.* 104, 19428–19433.

Creamer, T. J., Darby, M. M., Jamonnak, N., Schaughency, P., Hao, H., Wheelan, S. J., and Corden, J. L. (2011). Transcriptome-wide binding sites for components of the *Saccharomyces cerevisiae* non-poly(A) termination pathway: Nrd1, Nab3, and Sen1. *PLoS Genet.* 7, e1002329. doi:10.1371/journal.pgen.1002329

Euskirchen, G., Royce, T. E., Bertone, P., Martone, R., Rinn, J. L., Nelson, F. K., Sayward, F., Luscombe, N. M., Miller, P., Gerstein, M., Weissman, S., and Snyder, M. (2004). CREB binds to multiple loci on human chromosome 22. *Mol. Cell. Biol.* 24, 3804–3814.

Feil, R., Walter, J., Allen, N. D., and Reik, W. (1994). Developmental control of allelic methylation in the imprinted mouse Igf2 and H19 genes. *Development* 120, 2933–2943.

Gibb, E. A., Vucic, E. A., Enfield, K. S., Stewart, G. L., Lonergan, K. M., Kennett, J. Y., Becker-Santos, D. D., MacAulay, C. E., Lam, S., Brown, C. J., and Lam, W. L. (2011). Human cancer long non-coding RNA transcriptomes. *PLoS ONE* 6, e25915. doi:10.1371/journal.pone.0025915

Gupta, R. A., Shah, N., Wang, K. C., Kim, J., Horlings, H. M., Wong, D. J., Tsai, M. C., Hung, T., Argani, P., Rinn, J. L., Wang, Y., Brzoska, P., Kong, B., Li, R., West, R. B., van de Vijver, M. J., Sukumar, S., and Chang, H. Y. (2010). Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* 464, 1071–1076.

Guttman, M., Amit, I., Garber, M., French, C., Lin, M. F., Feldser, D., Huarte, M., Zuk, O., Carey, B. W., Cassady, J. P., Cabili, M. N., Jaenisch, R., Mikkelsen, T. S., Jacks, T., Hacohen, N., Bernstein, B. E., Kellis, M., Regev, A., Rinn, J. L., and Lander, E. S. (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* 458, 223–227.

Guttman, M., Garber, M., Levin, J. Z., Donaghey, J., Robinson, J., Adiconis, X., Fan, L., Koziol, M. J., Gnirke, A., Nusbaum, C., Rinn, J. L., Lander, E. S., and Regev, A. (2010). Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat. Biotechnol.* 28, 503–510.

Han, Z., He, H., Zhang, F., Huang, Z., Liu, Z., Jiang, H., and Wu, Q. (2012). Spatiotemporal expression pattern of Mirg, an imprinted non-coding gene, during mouse embryogenesis. *J. Mol. Histol.* 43, 1–8.

Hawkins, P. G., and Morris, K. V. (2010). Transcriptional regulation of Oct4 by a long non-coding RNA antisense to Oct4-pseudogene 5. *Transcription* 1, 165–175.

He, Y., Vogelstein, B., Velculescu, V. E., Papadopoulos, N., and Kinzler, K. W. (2008). The antisense transcriptomes of human cells. *Science* 322, 1855–1857.

Hendrich, B. D., Brown, C. J., and Willard, H. F. (1993). Evolutionary conservation of possible functional domains of the human and murine XIST genes. *Hum. Mol. Genet.* 2, 663–672.

Hibi, K., Nakamura, H., Hirai, A., Fujikake, Y., Kasai, Y., Akiyama, S., Ito, K., and Takagi, H. (1996). Loss of H19 imprinting in esophageal cancer. *Cancer Res.* 56, 480–482.

Juan, V., Crain, C., and Wilson, C. (2000). Evidence for evolutionarily conserved secondary structure in the H19 tumor suppressor RNA. *Nucleic Acids Res.* 28, 1221–1227.

Kapranov, P. (2009). Studying chromosome-wide transcriptional networks: new insights into disease? *Genome Med.* 1, 50.

Kapranov, P., St Laurent, G., Raz, T., Ozsolak, F., Reynolds, C. P., Sorensen, P. H., Reaman, G., Milos, P., Arceci, R. J., Thompson, J. F., and Triche, T. J. (2010). The majority of total nuclear-encoded non-ribosomal RNA in a human cell is 'dark matter' unannotated RNA. *BMC Biol.* 8, 149. doi:10.1186/1741-7007-9-86

Kotake, Y., Nakagawa, T., Kitagawa, K., Suzuki, S., Liu, N., Kitagawa, M., and Xiong, Y. (2011). Long non-coding RNA ANRIL is required for the PRC2 recruitment to and silencing of p15INK4B tumor suppressor gene. *Oncogene* 30, 1956–1962.

Lai, M. C., Yang, Z., Zhou, L., Zhu, Q. Q., Xie, H. Y., Zhang, F., Wu, L. M., Chen, L. M., and Zheng, S. S. (2012). Long non-coding RNA MALAT-1 overexpression predicts tumor recurrence of hepatocellular carcinoma after liver transplantation. *Med. Oncol.* (in press).

Lewis, A., Mitsuya, K., Umlauf, D., Smith, P., Dean, W., Walter, J., Higgins, M., Feil, R., and Reik, W. (2004). Imprinting on distal chromosome 7 in the placenta involves repressive histone methylation independent of DNA methylation. *Nat. Genet.* 36, 1291–1295.

Licastro, D., Gennarino, V. A., Petrera, F., Sanges, R., Banfi, S., and Stupka, E. (2010). Promiscuity of enhancer, coding and non-coding transcription functions in ultraconserved elements. *BMC Genomics* 11, 151. doi:10.1186/1471-2164-11-151

Long, Q., Johnson, B. A., Osunkoya, A. O., Lai, Y. H., Zhou, W., Abramovitz, M., Xia, M., Bouzyk, M. B., Nam, R. K., Sugar, L., Stanimirovic, A., Williams, D. J., Leyland-Jones, B. R., Seth, A. K., Petros, J. A., and Moreno, C. S. (2011). Protein-coding and microRNA biomarkers of recurrence of prostate cancer following radical prostatectomy. *Am. J. Pathol.* 179, 46–54.

Lustig, O., Ariel, I., Ilan, J., Lev-Lehman, E., De-Groot, N., and Hochberg, A. (1994). Expression of the imprinted gene H19 in the human fetus. *Mol. Reprod. Dev.* 38, 239–246.

Manolio, T. A., Brooks, L. D., and Collins, F. S. (2008). A HapMap harvest of insights into the genetics of common disease. *J. Clin. Invest.* 118, 1590–1605.

Martens-Uzunova, E. S., Jalava, S. E., Dits, N. F., van Leenders, G. J., Moller, S., Trapman, J., Bangma, C. H., Litman, T., Visakorpi, T., and Jenster, G. (2012). Diagnostic and prognostic signatures from the small non-coding RNA transcriptome in prostate cancer. *Oncogene* (in press).

Matouk, I. J., DeGroot, N., Mezan, S., Ayesh, S., Abu-lail, R., Hochberg, A., and Galun, E. (2007). The H19 non-coding RNA is essential for human tumor growth. *PLoS ONE* 2, e845. doi:10.1371/journal.pone.0000845

Mattick, J. S. (2003). Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms. *Bioessays* 25, 930–939.

Mercer, T. R., Dinger, M. E., and Mattick, J. S. (2009). Long non-coding RNAs: insights into functions. *Nat. Rev. Genet.* 10, 155–159.

Mestdagh, P., Fredlund, E., Pattyn, F., Rihani, A., Van Maerken, T., Vermeulen, J., Kumps, C., Menten, B., De Preter, K., Schramm, A., Schulte, J., Noguera, R., Schleiermacher, G., Janoueix-Lerosey, I., Laureys, G., Powel, R., Nittner, D., Marine, J. C., Ringnér, M., Speleman, F., and Vandesompele, J. (2010). An integrative genomics screen uncovers ncRNA T-UCR functions in neuroblastoma tumours. *Oncogene* 29, 3583–3592.

Morris, K. V., Santoso, S., Turner, A. M., Pastori, C., and Hawkins, P. G. (2008). Bidirectional transcription directs both transcriptional gene activation and suppression in human cells. *PLoS Genet.* 4, e1000258. doi:10.1371/journal.pgen.1000258

Ng, D., Toure, O., Wei, M. H., Arthur, D. C., Abbasi, F., Fontaine, L., Marti, G. E., Fraumeni, J. F. Jr., Goldin, L. R., Caporaso, N., and Toro, J. R. (2007). Identification of a novel chromosome region, 13q21.33-q22.2, for susceptibility genes in familial chronic lymphocytic leukemia. *Blood* 109, 916–925.

Panning, B., Dausman, J., and Jaenisch, R. (1997). X chromosome inactivation is mediated by Xist RNA stabilization. *Cell* 90, 907–916.

Pasmant, E., Laurendeau, I., Heron, D., Vidaud, M., Vidaud, D., and Bieche, I. (2007). Characterization of a germ-line deletion, including the entire INK4/ARF locus, in a melanoma-neural system tumor family: identification of ANRIL, an antisense noncoding RNA whose expression coclusters with ARF. *Cancer Res.* 67, 3963–3969.

Pasmant, E., Sabbagh, A., Masliah-Planchon, J., Ortonne, N., Laurendeau, I., Melin, L., Ferkal, S., Hernandez, L., Leroy, K., Valeyrie-Allanore, L., Parfait, B., Vidaud, D., Bièche, I., Lantieri, L., Wolkenstein, P., Vidaud, M., and NF France Network. (2011). Role of noncoding RNA ANRIL in genesis of plexiform neurofibromas in neurofibromatosis type 1. *J. Natl. Cancer Inst.* 103, 1713–1722.

Pauli, A., Rinn, J. L., and Schier, A. F. (2011). Non-coding RNAs as regulators of embryogenesis. *Nat. Rev. Genet.* 12, 136–149.

Petrovics, G., Zhang, W., Makarem, M., Street, J. P., Connelly, R., Sun, L., Sesterhenn, I. A., Srikantan, V., Moul, J. W., and Srivastava, S. (2004). Elevated expression of PCGEM1, a prostate-specific gene with cell growth-promoting function, is associated with high-risk prostate cancer patients. *Oncogene* 23, 605–611.

Ponjavic, J., Ponting, C. P., and Lunter, G. (2007). Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Res.* 17, 556–565.

Qureshi, I. A., and Mehler, M. F. (2011). Non-coding RNA networks underlying cognitive disorders across the lifespan. *Trends. Mol. Med.* 17, 337–346.

Rodriguez, B. A., Weng, Y. I., Liu, T. M., Zuo, T., Hsu, P. Y., Lin, C. H., Cheng, A. L., Cui, H., Yan, P. S., and Huang, T. H. (2011). Estrogen-mediated epigenetic repression of the imprinted gene cyclin-dependent kinase inhibitor 1C in breast cancer cells. *Carcinogenesis* 32, 812–821.

Rosenbluh, J., Nijhawan, D., Chen, Z., Wong, K. K., Masutomi, K., and Hahn, W. C. (2011). RMRP is a non-coding RNA essential for early murine development. *PLoS ONE* 6, e26270. doi:10.1371/journal.pone.0026270

Rossi, S., Sevignani, C., Nnadi, S. C., Siracusa, L. D., and Calin, G. A. (2008). Cancer-associated genomic regions (CAGRs) and noncoding RNAs: bioinformatics and therapeutic implications. *Mamm. Genome* 19, 526–540.

Sang, X., Zhao, H., Lu, X., Mao, Y., Miao, R., Yang, H., Yang, Y., Huang, J., and Zhong, S. (2010). Prediction and identification of tumor-specific noncoding RNAs from human UniGene. *Med. Oncol.* 27, 894–898.

Scaruffi, P., Stigliani, S., Moretti, S., Coco, S., De Vecchi, C., Valdora, F., Garaventa, A., Bonassi, S., and Tonini, G. P. (2009). Transcribed-ultra conserved region expression is associated with outcome in high-risk neuroblastoma. *BMC Cancer* 9, 441. doi:10.1186/1471-2407-9-441

Suh, N., and Blelloch, R. (2011). Small RNAs in early mammalian development: from gametes to gastrulation. *Development* 138, 1653–1661.

Sumazin, P., Yang, X., Chiu, H. S., Chung, W. J., Iyer, A., Llobet-Navas, D., Rajbhandari, P., Bansal, M., Guarnieri, P., Silva, J., and Califano, A. (2011). An extensive microRNA-mediated network of RNA-RNA interactions regulates established oncogenic pathways in glioblastoma. *Cell* 147, 370–381.

Taft, R. J., Pheasant, M., and Mattick, J. S. (2007). The relationship between non-protein-coding DNA and eukaryotic complexity. *Bioessays* 29, 288–299.

Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl. Acad. Sci. U.S.A.* 99, 6567–6572.

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* 28, 511–515.

van Bakel, H., and Hughes, T. R. (2009). Establishing legitimacy and function in the new transcriptome. *Brief. Funct. Genomic. Proteomic.* 8, 424–436.

van Leeuwen, S., and Mikkers, H. (2010). Long non-coding RNAs: guardians of development. *Differentiation* 80, 175–183.

Wang, X., Arai, S., Song, X., Reichart, D., Du, K., Pascual, G., Tempst, P., Rosenfeld, M. G., Glass, C. K., and Kurokawa, R. (2008). Induced ncRNAs allosterically modify RNA-binding proteins in cis to inhibit transcription. *Nature* 454, 126–130.

Yang, R., Frank, B., Hemminki, K., Bartram, C. R., Wappenschmidt, B., Sutter, C., Kiechle, M., Bugert, P., Schmutzler, R. K., Arnold, N., Weber, B. H., Niederacher, D., Meindl, A., and Burwinkel, B. (2008). SNPs in ultraconserved elements and familial breast cancer risk. *Carcinogenesis* 29, 351–355.

Yap, K. L., Li, S., Munoz-Cabello, A. M., Raguz, S., Zeng, L., Mujtaba, S., Gil, J., Walsh, M. J., and Zhou, M. M. (2010). Molecular interplay of the noncoding RNA ANRIL and methylated histone H3 lysine 27 by polycomb CBX7 in transcriptional silencing of INK4a. *Mol. Cell* 38, 662–674.

Zhang, X., Zhou, Y., Mehta, K. R., Danila, D. C., Scolavino, S., Johnson, S. R., and Klibanski, A. (2003). A pituitary-derived MEG3 isoform functions as a growth suppressor in tumor cells. *J. Clin. Endocrinol. Metab.* 88, 5119–5126.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 02 December 2011; paper pending published: 14 December 2011; accepted: 28 January 2012; published online: 15 February 2012.

Citation: Mitra SA, Mitra AP and Triche TJ (2012) A central role for long non-coding RNA in cancer. Front. Gene. ***3****:17. doi: 10.3389/fgene.2012.00017*

This article was submitted to Frontiers in Non-Coding RNA, a specialty of Frontiers in Genetics.

frontiers in
GENETICS

REVIEW ARTICLE
published: 27 December 2011
doi: 10.3389/fgene.2011.00096

Characterizing ncRNAs in human pathogenic protists using high-throughput sequencing technology

Lesley Joan Collins*

Institute of Fundamental Sciences, Massey University, Palmerston North, New Zealand

Edited by:
Philipp Kapranov, St. Laurent Institute, USA

Reviewed by:
Toshifumi Inada, Tohoku University, Japan
Michiaki Hamada, The University of Tokyo, Japan

***Correspondence:**
Lesley Joan Collins, Massey University, Palmerston North, New Zealand.
e-mail: l.j.collins@massey.ac.nz

ncRNAs are key genes in many human diseases including cancer and viral infection, as well as providing critical functions in pathogenic organisms such as fungi, bacteria, viruses, and protists. Until now the identification and characterization of ncRNAs associated with disease has been slow or inaccurate requiring many years of testing to understand complicated RNA and protein gene relationships. High-throughput sequencing now offers the opportunity to characterize miRNAs, siRNAs, small nucleolar RNAs (snoRNAs), and long ncRNAs on a genomic scale, making it faster and easier to clarify how these ncRNAs contribute to the disease state. However, this technology is still relatively new, and ncRNA discovery is not an application of high priority for streamlined bioinformatics. Here we summarize background concepts and practical approaches for ncRNA analysis using high-throughput sequencing, and how it relates to understanding human disease. As a case study, we focus on the parasitic protists *Giardia lamblia* and *Trichomonas vaginalis,* where large evolutionary distance has meant difficulties in comparing ncRNAs with those from model eukaryotes. A combination of biological, computational, and sequencing approaches has enabled easier classification of ncRNA classes such as snoRNAs, but has also aided the identification of novel classes. It is hoped that a higher level of understanding of ncRNA expression and interaction may aid in the development of less harsh treatment for protist-based diseases.

Keywords: ncRNA, high-throughput sequencing, miRNA, siRNA, snoRNA, *Giardia, Trichomonas*

INTRODUCTION

The discovery and analysis of ncRNAs has become an important step in the understanding of the genomics behind human disease. Genomics in humans has in the past tended to concentrate on the small percentage (1–2%) of genomic space coding for proteins. Since the vast majority of human ncRNAs lie in non-coding regions including introns and intergenic spaces, there is a need for fast and flexible methods of ncRNA identification. ncRNA classes in general, and in particular microRNAs (miRNAs) and short interfering RNAs (siRNAs) are of great interest in disease studies. In some cases miRNAs have been implicated in various cancers with altered expression levels appearing to be associated with the genetic alterations seen in malignancies (Ryther et al., 2003). miRNAs and siRNAs to be important effectors in host–pathogen interaction networks between humans and their viruses (Aurrecoechea et al., 2009), most of which use RNA interference processes. RNA interference (RNAi) has also been raised as an option for medical treatment of human diseases including cancer (Garlapati et al., 2011), viruses (Khaliq et al., 2010; Haasnoot and Berkhout, 2011), and transplantation (Zhang et al., 2011b).

RNA interference in general, is a process where small RNAs (e.g., miRNAs and siRNAs) are used by a protein macromolecule, to target and then to cleave transcribed mRNAs, hence "interfering" with the expression of a targeted gene. There are a number of different pathways for this interference (Collins and Penny, 2009; Batista and Marques, 2011; Ketting, 2011), with three main proteins or their like, being typically required. These proteins are Dicer, Argonaute, and RNA-dependent RNA polymerase (RdRp). Finding homologs to these proteins in protist species is usually the first step in determining that RNAi exists in that species. However, as can be seen in species such as *Giardia lamblia* and *Entamoeba histolytica*, some of these proteins may not contain all the domains we expect to find (Macrae et al., 2006; Carlton et al., 2007; Batista and Marques, 2011). It is thus, very likely that protist RNAi pathways will differ from their well-studied multicellular counterparts, and that understanding these differences will enable a far more efficient use of RNA as a molecular tool.

RNA interference has been used to understand gene expression levels and the changes that occur at different stages of disease, different stages of life cycle or development, and differences in environmental conditions (typically with miRNA studies). Genes can also be specifically "knocked-out" using gene silencing studies to investigate the effects of particular parts of metabolic pathways. This has been done in some protists, in particular Kinetoplasts where RNAi has been used as a tool in genomic studies for *Trypanosoma brucei* (reviewed in Kolev et al., 2011), and to a lesser extent in *Leishmania braziliensis* subgenus *Viannia* (Lye et al., 2011). The difficulty in protist RNAi research is that the small RNAs that are used in RNAi (i.e., miRNAs and siRNAs), are not easily isolated and characterized.

Genomic-wide sequencing is also furthering studies on how a host species reacts to pathogens in immune and preventative responses. One non-protist example is where high-throughput sequencing was used to characterize miRNA levels and identify novel miRNAs involved in avian influenza virus (AIV) infection of chicken (Aurrecoechea et al., 2009). In this study, sequences were matched not only to genomic sequences but to mature miRNA sequences previously lodged in miRbase (Finn et al., 2006), allowing for insertions and deletions of 1–4 nt. Profiling analysis compared infected and non-infected tissue to identify miRNAs that changed expression upon infection. Mapping of the sequences also revealed that many miRNAs are grouped in clusters on the chicken chromosomes and up- or down-regulated together. Results from this study suggest that different miRNA regulation mechanisms may exist on host response to virus infection with some genes up regulated to aid host immune response and down regulation of targets to aid inhibition of virus replication. Different tissues may express different levels of miRNAs. For example in the Wang et al. study 377 miRNAs were identified from chicken lung tissue but only 149 miRNAs were identified from tracheae. Clearly this type of study will soon extend to the host response to protists. The techniques for analysis will be similar but will require a greater understanding of the typical features of the different miRNA classes in the protist of study.

There are many different classes of ncRNAs found in protists (**Table 1**), and only some of these such as miRNAs and siRNAs and sometimes small nucleolar RNAs (snoRNAs) are involved in RNAi. Other ncRNAs such as tRNAs and rRNAs are relatively easy to characterize because they look familiar to those already studied, but there are classes such as snoRNAs, that are harder to find and classify because either their sequence or their action, is novel. Previously, there were two main approaches to ncRNA identification (**Figure 1**). The first, the "traditional" approach, involved the isolation of expressed RNA in a designated size range, cloning, sequencing then finally, Northern blotting to confirm size and conformational isotopes from RNA modifications. This approach was costly both in laboratory expenditure and time, and was not very practical on a genomic scale. The other approach took a sequenced genome and computationally screened it for candidate ncRNAs, using mathematical models based on the sequence and structural characteristics of a class of ncRNA. This second approach, although it could be applied on a genomic scale, often produced masses of candidates that would then have to be experimentally tested by the first approach. Another issue with the computational approach is that only a single class could be searched at a time, and one had to know what that class looked like both in sequence and secondary structure in order to find it. Permitting more flexibility however, results in more false positives, a circumstance that can quickly overload the computer and its user. High-throughput sequencing permits the genome-wide sequencing of ncRNAs from expressed RNA (the power of the first approach), and for rapid comparison to known classes (the power of the second approach), as well as the characterization of novel ncRNAs (**Figure 1**). The disadvantage of this type of sequencing is that it demands a different type of computational analysis than previously used with ncRNAs (see later).

Short interfering or silencing RNAs (siRNAs) are similar to miRNAs but are produced from double stranded precursors instead of folded single-stranded precursors. What makes this mechanism of great interest is that the gene silencing is highly specific, and also highly potent in where only a few copies of an small RNA (22–25 nt long) can demonstrate wide ranging affects. In plants, they have been highly investigated for their role in virus response (Ridanpaa et al., 2003; Pantaleo, 2011), but in humans and mice they are under intense study for therapeutic medicine

Table 1 | Summary of ncRNA discovery in human pathogenic protists.

Protist	Lineage	Disease	ncRNAs+	RNAi proteins	Functional RNAi
Giardia lamblia	Diplomonad	Giardiasis	miRNA, siRNA	Dicer-like, Argonaute, RdRp (Macrae et al., 2006)	Not proven natively with miRNA but long dsRNA shown to control specific gene down regulation (Rivero et al., 2010)
Trichomonas vaginalis	Parabasalid	Trichomoniasis	miRNA, siRNA	Dicer-like, Argonaute, RdRp (Carlton et al., 2007)	Yes (Lin et al., 2009)
Plasmodium falciparum	Apicomplexa	Malaria	–	Absence of proteins with a PAZ or piwi domain (Baum et al., 2009)	No (Baum et al., 2009) dsRNA triggering down regulation (see review in Kolev et al., 2011)
Entamoeba histolytica	Amebozoa	Amoebic dysentery	siRNA	Argonaute, Dicer-like*, RdRp	Yes (Reviewed in Zhang et al., 2011a)
Trypanosoma spp.	Kinetoplastid	*T. brucei* (sleeping sickness), *T. cruzi* (Chagas disease)	siRNA, miRNA (*T. gondii*)	Argonaute, Dicer-like (not in *T. cruzi*)	Yes in *T. brucei* but not in *T. cruzi* (reviewed in Lye et al., 2011)
Leishmania spp.	Kinetoplastid	Leishmaniasis	siRNA	Argonaute, Dicer-like	Some species (reviewed in Lye et al., 2011)

+These protists all contain RNase P RNA, RNase MRP RNA, tRNAs, rRNAs, snRNAs, and snoRNAs; ∗atypical protein structure.

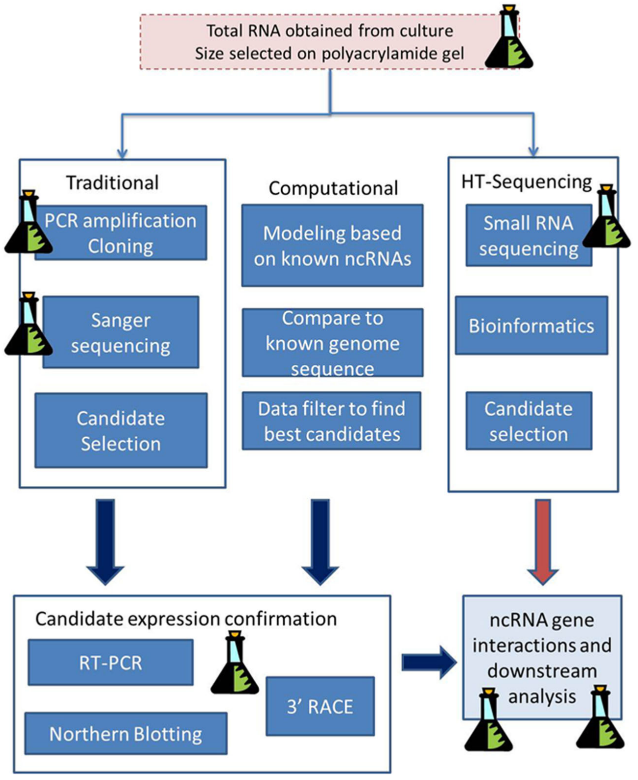

Total RNA obtained from culture
Size selected on polyacrylamide gel
Traditional
PCR amplification Cloning
Sanger sequencing
Candidate Selection
Computational
Modeling based on known ncRNAs
Compare to known genome sequence
Data filter to find best candidates
HT-Sequencing
Small RNA sequencing
Bioinformatics
Candidate selection
Candidate expression confirmation
RT-PCR
3' RACE
Northern Blotting
ncRNA gene interactions and downstream analysis

FIGURE 1 | Genomic approaches to ncRNA identification. Both the traditional laboratory approach and the more recent high-throughput sequencing approach begin with the isolation of total RNA from culture, followed by size selection of the RNA by excising a given band from a polyacrylamide gel. Under the traditional approach the excised RNA is cloned then sequenced by Sanger Sequencing to obtain candidate miRNA sequences. With High-throughput sequencing the RNA is sequenced directly without cloning and bioinformatics is used to select the best candidates. Computational approaches do not begin with biological samples but instead use mathematical models based on known ncRNAs to search an already sequenced genome. Both the traditional and computational approaches require that candidate gene expression be confirmed by additional laboratory work. Key: Molecular biology stages are represented by the flask icon. All other stages use RNA genomic and bioinformatics procedures.

(recently reviewed in Burnette et al., 2005; Khaliq et al., 2010; Vaishnaw et al., 2011). In an example, a combination of host and viral genes has been used as a siRNA-based treatment for hepatitis C virus (HCV; Ashfaq et al., 2011). Developing RNAi based treatments for HCV are an important are of research, since there is currently no vaccine available for HCV due to a high degree of strain variation. Another factor is that the current drug treatment (with a pegylated interferon α/ribavirin combination) is costly, has significant side effects and is not always effective (Ashfaq et al., 2011). RNAi offers new and less harsher types of treatment especially for viral diseases, but there are still challenges in this area in systematic siRNA delivery and distribution to appropriate tissue (Vaishnaw et al., 2011).

Small nucleolar RNAs typically have roles in the modification of other ncRNAs such as rRNA, small nuclear RNAs (snRNAs), and possibly even mRNAs (Ridanpaa et al., 2003; Bompfunewerer et al., 2005; Gardner et al., 2010; Khanna and Stamm, 2010). snoRNAs are between 60 and 150 nt long and fall into two main classes based on conserved sequence motifs, H/ACA (sometimes called SNORAs), and C/D (sometimes called SNORDs). snoRNAs may have a high potential for use as markers for diseases either by having mutated sequences or differential expression. In one example some snoRNAs were discovered to have a higher expression in some lung cancer cells than in non-cancerous cells, and thus have a potential as markers for the early detection of this cancer (Liao et al., 2010).

The examples above focus on how ncRNAs from the host can be used to study gene expression from healthy, diseased, and sometimes treated tissue. A second type of study looks at the ncRNAs from the pathogen itself to understand where the pathogen could be vulnerable and hence open to new treatment options. This is where there is less work published since until now the discovery and characterization of ncRNAs in most pathogens was slow and laborious. Until very recently ncRNAs in prokaryotes (often called "small RNAs") were not commonly thought of as being important in pathogenic studies. The characterization of the CRISPR system and small RNA pathways (e.g., Hfq-binding sRNAs) has made us more aware that an RNA-based backbone exists just as much in prokaryotes as in eukaryotes (for review see Biggs and Collins, 2011; Collins and Biggs, 2011). Eukaryotic pathogens (e.g., nematodes, yeast, and protists) have received a little more attention but lag behind our understanding of host (typically human and mice) ncRNAs (review by Batista and Marques, 2011).

RNA interference in the protist *T. brucei* (causative agent of sleeping sickness) was characterized early on as functional (Ngo et al., 1998), only months after it was demonstrated in the nematode *Caenorhabditis elegans* (Fire et al., 1998; reviewed in Kolev et al., 2011). However, RNAi mechanisms can be lost from a lineage, as demonstrated in some yeast (Drinnenberg et al., 2009, 2011) and some trypanosomatids (Lye et al., 2011). Thus, even though RNAi is considered to be an ancient system that was likely to be present in the last common ancestor of eukaryotes (Collins and Chen, 2009), it does not follow that it will be still be present. Further studies (reviewed in Kolev et al., 2011) have revealed that RNAi in trypanosomatids loss in multiple lineages are in most cases correlated with the lack of mobile elements (Kolev et al., 2011; Lye et al., 2011). The lack of RNAi but presence of mobile element-like sequences in one lineage of trypanosomatids *T. cruzi* falls against that trend (Kolev et al., 2011) indicating that there is much more to be learned about the evolution of RNAi mechanisms.

One of the issues contributing to the slow progress in understanding protist RNAi is that many of these pathogens (and especially the protists) do not yet have well annotated genomes. Others may have genomes sequenced (some fungi and nematodes) but genes are annotated based on sequence similarity to the few very well known genomes. ncRNA genes change rapidly and mis-annotation is common. Very small genes such as those for miRNAs and siRNAs can be extremely hard to characterize based on sequence similarity. Hence, the arrival of high-throughput sequencing has enabled these ncRNAs to be tackled in a slightly easier manner, but it has meant the incorporation of more bioinformatics into these projects. High-throughput sequencing of pathogens and especially protists in practice uses the power of computational biology combined with the expression from real RNA. With this technology we can look at miRNAs, siRNAs, snoRNA, and even longer ncRNAs from a pathogenic protist and potentially link some of them to host responses. However, first we have to characterize the ncRNA classes from our species of choice. Here we will use the two pathogenic protists, *G. lamblia* and *Trichomonas vaginalis* as examples to highlight the issues and possible solutions with high-throughput small RNA sequencing. It should be noted however, that these issues and solutions are not confined to protist genomics, but are also generally applicable to other species including prokaryotes.

ncRNAs FROM PATHOGENIC PARASITIC PROTISTS

Pathogenic protists are responsible for a host of human diseases that affect millions worldwide, but not surprisingly ncRNA research in these pathogens has lagged behind protein-based research. Over the last decade, there are two protist species, *G. lamblia* and *T. vaginalis* that are of interest in the characterization of their RNAs because of their evolutionary distance from other eukaryotes (Collins and Penny, 2005, 2009; Collins and Chen, 2009; Chen et al., 2011). *G. lamblia* is a Diplomonad anaerobic protist that infects humans and other mammals. When ingested, the cysts hatch into trophozoites in the small intestine causing diarrhea and growth hindrance in children. It is a significant pathogen for the immune-compromised and those in developing countries affected by malnutrition. The most common treatment for giardiasis is Metronidazole (Flagyl), which unfortunately has unpleasant side effects including nausea and dizziness, but more importantly has potential carcinogenic properties. Metronidazole is not approved by the FDA in the USA for treatment in human medicine for this reason.

Trichomonas vaginalis is an anaerobic Parabasalid protist that causes the sexually transmitted disease trichomoniasis in humans. Despite its name, infections are common in men but are typically asymptomatic. In women trichomoniasis is symptomatic as an STD, but it can also lead to adverse pregnancy outcomes and be associated with an increased risk of human immunodeficiency virus (HIV) transmission (Cudmore et al., 2004). *Trichomonas* differs from *Giardia* in that it does not have a cyst stage, so infection is directly by the trophozoites being transferred from patient to patient. Treatment of trichomoniasis also includes Metronidazole, but studies have shown at least 5% of cases are resistant to this drug (Cudmore et al., 2004).

Other pathogenic protists include *Plasmodium falciparum* (malaria), *E. histolytica* (amebic dysentery) and *T. brucei* (sleeping sickness) and *T. cruzi* (Chagas disease). Although drug treatments for all these diseases are available there is still a need for further development, especially for *Giardia* and *Trichomonas* where problems with treatment persist. Thus, the use of ncRNAs and RNAi is especially applicable to eukaryotic pathogens such as *Giardia* and *Trichomonas* as both a tool and a potential treatment option.

High-throughput sequencing has been used to assemble protist genomes. *Giardia* and *Trichomonas* were "completed" before this technology appeared meaning that their assembly was slow and most is still in the form of large pieces (supercontigs). This should not put any researcher off using such genomes since genomes like this are very usable for high-throughput small RNA studies. From a number of studies, all using these annotated genomes, we can see that *Giardia* and *Trichomonas* like their distant multicellular human host contain a rich collection of ncRNAs (Chen et al., 2007, 2008, 2009). These include RNase P (Piccinelli et al., 2005), RNase MRP (Chen et al., 2011) snoRNAs (Yang et al., 2005; Chen et al., 2007, 2011), spliceosomal snRNAs (Chen et al., 2008), miRNAs (Saraiya and Wang, 2008; Chen et al., 2009; Zhang et al., 2009), and antisense transcripts (Teodorovic et al., 2007). Studies on *Trichomonas* ncRNAs show that the currently known ncRNAs also exhibit typical features of eukaryotes (Piccinelli et al., 2005; Simoes-Barbosa et al., 2008; Lin et al., 2009; Smith and Johnson, 2011) including RNase P (Piccinelli et al., 2005), RNase MRP (Piccinelli et al., 2005), snRNAs (Simoes-Barbosa et al., 2008), and some snoRNAs (Chen et al., 2007, 2009, 2011). However, some classes of ncRNAs (especially the snoRNAs) contain some features that are not typical. It is high-throughput sequencing that offers the opportunity to investigate these novel classes of ncRNA.

HIGH-THROUGHPUT SEQUENCING AND ANALYSIS

High-throughput sequencing of small RNAs requires an RNA sample of high quality and reasonably high concentration. An issue with many protists is that they are not culturable so RNA is sometimes extracted from patient samples. Therefore, obtaining enough clean and uncontaminated RNA for genomic sequencing is sometimes not easy or possible unless from a culturable strain. Additionally, lab strains carry the risk that they may contain genetic differences from their clinical relatives. Advances in sequencing protocols have meant that the amount of RNA required for sequencing (genomic, transcriptomic, and to some extent small RNA) is reduced, and it is hoped that with further improvements in protocols, more protists will be sequenced. Once a sample of RNA is obtained, then the next stage is to decide which ncRNAs are going to be sequenced. A sample of total RNA contains some RNAs that can drown out any underrepresented ncRNAs in samples so mRNAs, rRNAs, and tRNAs must be removed as much as possible. A typical procedure is to run the sample on a polyacrylamide gel, then excise a band corresponding to the size for sequencing (e.g., miRNAs, siRNAs:19–30 nt, snoRNAs:50–200 nt; Chen et al., 2009). Gel isolation, although contributing to the reduction in the amount of RNA available for sequencing, offers flexibility and selectivity in the class of ncRNA being examined. Overall, small RNA sequencing experiments are individualistic with all the advantages and disadvantages that come with the use of developing technology.

The actual sequencing of small RNA samples is typically performed by the operator of the technology, whether at a service provider or in-house facility. There are currently choices of sequencing platforms for high-throughput sequencing, and highly likely that there will be even more choice in the near future. For small RNA sequencing the platform selected will depend on the length of the ncRNA to be sequenced. Some platforms such as the Roche 454[1] produce long sequences (400–1000 nt) which is really only suitable if trying to sequence very long ncRNAs. Small RNAs can be done on this platform but it is not optimized for this type of work. Presently available platforms that generate short sequences are the Illumina systems (HiSeq, MiSeq, and Genome Analyzer[2]) and the SOLiD system[3].Because they enable sequences of short length to be obtained (36–50 nt) these systems are ideal for sequencing small ncRNAs such as miRNAs and siRNAs and as you will see later, can also be used to sequence ncRNAs of longer lengths. For a review of these platforms and their use in small RNA studies in general please see McCormick et al. (2011). There are also different RNA-based protocols available, including those that combine mRNA and ncRNA sequencing (i.e., without need for the mRNA to have polyA tails; Yang et al., 2011). This is very useful for prokaryotic sequencing and also for combining mRNA and ncRNA data in a single run. However, "small RNA sequencing" is the typical protocol used for human ncRNA analysis. The field of high-throughput sequencing is very dynamic with platforms and protocols constantly being added and upgraded. It is therefore best to consult with the platform operator on what is available, prior to submitting samples.

If the idea is to compare miRNAs and their expression levels across different conditions (i.e., ncRNA gene expression) then there is one further consideration. There can be affects in running samples in different partitions of the sequencing apparatus (e.g., in different lanes or partitions on a "flowcell"). To avoid an excess of statistical analysis to take "lane-effect" into account, it is suggested that both samples are run in the same partition or partitions and barcoded to aid sorting after the sequencing. This is now common practice for digital gene expression analysis with mRNAs.

The analysis of data from high-throughput sequencing is quite unlike other genomic analysis and can be daunting to the newcomer. Unfortunately the analysis of ncRNA data is one of the lesser-described protocols in high-throughput sequencing meaning that there are fewer automated pipelines, and the software is primarily at a command-line level. Only now are guidelines to small RNA analysis being published (e.g., McCormick et al., 2011) and this is primarily for mammalian and plant research. What follows in this section is a general approach to the analysis of short (<25 nt) and medium/long (>50 nt) ncRNA sequencing data from high-throughput sequencing (**Figure 2**).

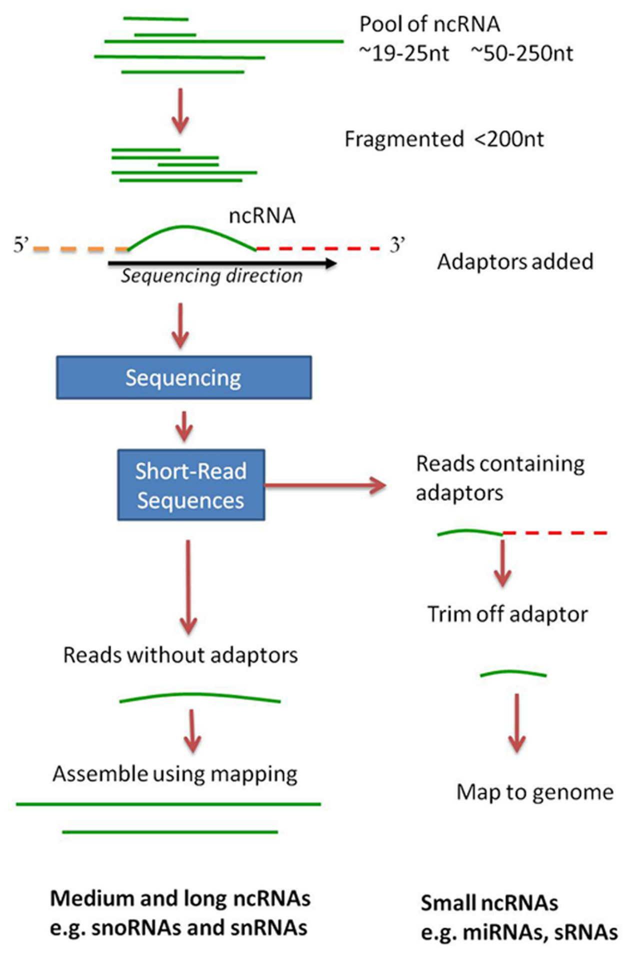


FIGURE 2 | A general approach for using high-throughput sequencing data to search for small and larger ncRNAs. Total RNA can contain ncRNA sequences of different sizes. Small ncRNAs such as miRNAs and siRNAs will produce data containing the ncRNA sequence plus adaptor sequence. Once the adaptor is trimmed off, the sequence is ready for mapping and further genomics. When longer ncRNAs are sequenced, they are fragmented to a predetermined size by the sequencing process. After sequencing they can be assembled by mapping (see text) into their longer sequences. This approach was successfully used for the analysis of ncRNAs from *Giardia lamblia* (see text for details).

The output from high-throughput sequencing is typically a file of short sequences (often termed short-reads or reads) accompanied by a quality score for each nucleotide in each sequence. This is termed FASTQ format with four lines for each sequence (Cock et al., 2010). However, because of the high sensitivity of this type of sequencing, the "raw" data will also contain sequences, including sequencing primers and contaminants, which occur in all high-sequencing datasets. Data filtering is therefore the first step to analysis and for small RNA sequencing it can permit the separation of reads into those likely to be from small ncRNAs (miRNAs and siRNAs) and those from medium ncRNAs. Sequencing adaptors and sequencing primers will occur in small RNA sequencing datasets. Adaptors are the short sequences that are added to the ends of the RNA fragment during the sequencing process. Additionally, during RNA work (due to the fragments being very short) we can also get adaptor dimers forming and these will show up in the results. Adaptor sequence information is available from the sequencing vendors and can be removed from the data using text-mining or sequence manipulation software. One example is the FASTX-toolkit[4] that contains scripts not only to remove adaptor sequences, but can also remove sequences containing too many N's and those sequences of lower quality. Other quality assessment tools include many commercial software

[1] www.my454.com
[2] www.illumina.com
[3] www.appliedbiosystems.com
[4] hannonlab.cshl.edu/fastx_toolkit/index.html

packages, and freeware such as FastQC[5], which can help guide data-filtering requirements.

Biologically ncRNAs may be of different lengths in the cell, and this can cause numerous issues for downstream analysis where computer programs prefer the data length to be consistent. It is therefore particularly recommended that sequences are trimmed to a common nucleotide length before mapping. For work with protist small RNA data such as that from *Giardia* and *Trichomonas* (used in studies by Chen et al., 2007, 2008, 2009, 2011), this basic approach of filtering sequences prior to further analysis was used. It was found that by filtering those sequences with adaptors, and those without, into different datasets, ncRNAs of different length ranges (small and medium) could be characterized (**Figure 3**).

Mapping (or matching) of the short-read sequences to a reference genome is the easiest approach for finding ncRNAs from protists. One approach is to map sequences to ncRNAs already characterized. Since there are still only a relatively few classes of ncRNAs characterized for many protists this approach can yield limited results but is worth doing to characterize medium length ncRNAs such as snoRNAs, snRNAs, RNase P, and RNase MRP. Smaller ncRNA classes such as miRNAs and siRNAs may have too many sequence differences over their short length to permit accurate mapping directly to known sequences, but mapping of miRNAs is possible to their longer pre-miRNA precursors, and siRNAs to potential target sequences. Mapping tools such as BWA (Li and Durbin, 2009), Bowtie (Langmead et al., 2009), and SOAP2 (Aurrecoechea et al., 2009), as well as commercial options are available for mapping small RNA short-read sequences. However, all of this software was designed for dealing with sequences longer than 22 nt, so software parameters must be changed for mapping short ncRNAs such as miRNAs and siRNAs. One of the key parameters to adjust is the "seed length" for the alignments, which is where the initial contact between the read and the reference genome is made. This should be set to about 17–18. Allowing more mismatches, e.g., up to $n = 3$ mismatches per seed (the standard is $n = 2$) may only be beneficial if the species being sequenced is different from the reference genome. However, this can cause spurious "false" mappings and typically $n = 2$ should suffice. Another useful parameter is in the reporting of the results. Often when a short-read maps to multiple places, only one place is reported and this is typically assigned by random. Since we commonly have miRNAs and siRNAs match both at their place in the genome and at their target position in the genome, multiple matches are normal, thus reporting settings should be adjusted for the reporting of all hits.

In practice, miRNAs have been characterized from protists using a mixture of traditional, computational, and high-throughput sequencing approaches. Some miRNAs from *Giardia* have been shown to be derived from snoRNAs (Saraiya and Wang, 2008; Kolev and Ullu, 2009), with two of these, miR2 and miR4 consequently shown to regulate the expression of variant surface proteins (VSPs), involved in resistance to host intestinal proteases (Prucca et al., 2008; Garlapati et al., 2011). A computational study

[5] www.bioinformatics.bbsrc.ac.uk/projects/fastqc

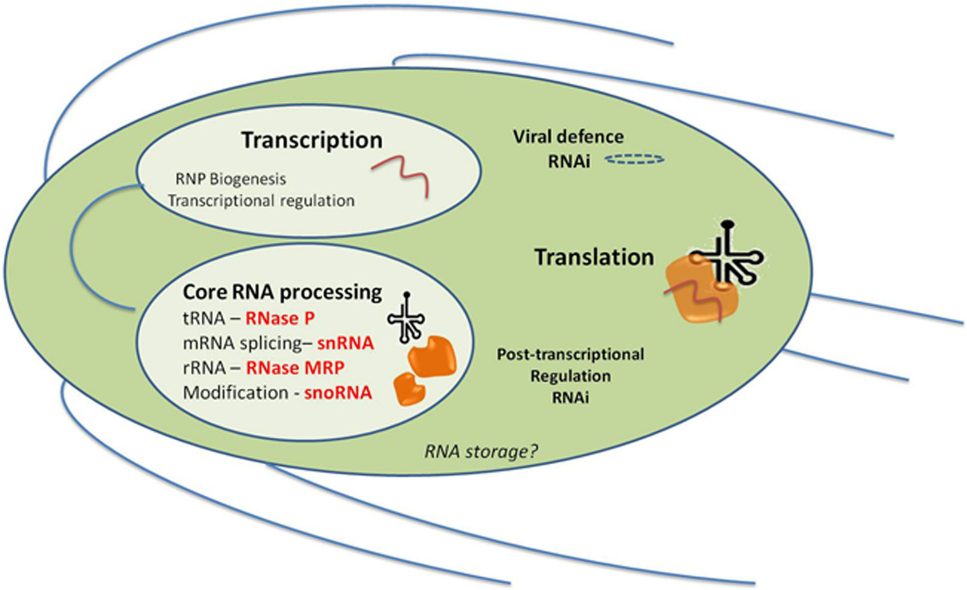


FIGURE 3 | ncRNAs processes characterized in *Giardia lamblia* and their likely cellular locations in the trophozoite stage, and in relation to the major RNA processes of Transcription and Translation. *Giardia* has two nuclei which appear identical and replicate at the same time and all of the processes within Transcription, RNP Biogenesis, Transcriptional regulation, and Core RNA processing could be expected to be found in both nuclei. RNAi may also be important in viral defense since some siRNAs found in high-throughput sequencing do map to the stable *Giardiavirus* (unpublished results). It is not known yet whether *Giardia* has RNA storage granules but their presence in *Trypanosoma* does not preclude them here. A similar diagram can be visualized for *Trichomonas* except that it only has one nucleus.

of miRNAs from *Giardia* (Zhang et al., 2009) identified 50 mature miRNA candidates, some of which also targeted VSP genes. Using high-throughput sequencing these VSP targeting miRNA candidates were also found (Chen et al., 2007, 2009) as well as many other miRNAs conforming to the expected size range of 25–27 nt (Macrae et al., 2006; Chen et al., 2007).

In *Trichomonas*, miRNAs have been characterized using traditional (Lin et al., 2009), and high-throughput sequencing approaches (Chen et al., 2009). Here, the use of miRNA-based software such as miRanda[6], aided in the selection of strong miRNA candidates. In practice, although high-throughput sequencing is sequencing biologically expressed RNAs (unlike computational analysis), there may also be some mRNA degradation products and contaminants that may slip through filtering. Studies in *Giardia* and *Trichomonas* have highlighted that confirmation procedures from another source (including miRNA-based software), is still required. Expression levels of some *Trichomonas* miRNAs have also been examined showing that one miRNA tva-miR-001 has a significantly lower expression level in the ameboid stage (Lin et al., 2009).

Most of the small RNA work to date in *Giardia* and *Trichomonas* has focused on miRNAs but siRNAs are also present (Chen et al., 2009). Some siRNAs appear to map in *Giardia* to a long-tandem repeat RNAs (Girep-1–5) which show a high degree of sequence similarity with a number of VSP genes (Chen et al., 2009) indicating yet another connection between RNAi and VSP gene selection. In addition, both *Giardia* and *Trichomonas* contain stable RNA viruses and high-throughput sequencing has resulted in some reads mapping to these viruses (unpublished results).

Small RNAs from other protists have also been characterized using traditional or computational methods. In *T. cruzi*, where there standard RNAi proteins have not been found, small RNAs derived from tRNAs have been characterized which are usually recruited to specific cytoplasmic granules (Garcia-Silva et al., 2010). *T. brucei* does contain standard RNAi pathways and computational studies have uncovered miRNAs that target another type of antigenic variation variant surface glycoproteins (VSGs; Rudra et al., 2007). Some miRNA candidates have also been characterized in *E. histolytica* using a computational study (De et al., 2006), but research here has focused on using dsRNA for gene silencing (reviewed by Kolev et al., 2011; Zhang et al., 2011a). Studies of protist miRNAs to date are showing that RNAi has an important role in antigenic variation and hence the parasite's survival in its host. Learning more about this system could enable more effective strategies to prevent and treat a range of protist diseases.

To characterize medium length ncRNAs, such as snoRNAs, RNase P, and RNase MRP RNAs, it can be useful to generate "contigs" (overlapping consensus fragments) from the mapping data. *De novo* assembly tools such as Velvet (Zerbino and Birney, 2008) and Abyss (Simpson et al., 2009) which assemble fragments without any prior alignment to a reference genome, are again primarily designed for working with longer sequences, and in practice did not perform well with data from *Giardia* and *Trichomonas* data (Chen et al., 2007, 2008, 2009, 2011). However, with careful parameter choice and a bit of experimentation, it is not inconceivable that these tools could be useful in the assembly of long (>200 nt long) ncRNAs, such as those discovered to be crucial for human and mouse epigenetics. An easier way to generate consensus contigs can be to use the results from mapping and convert them to contig sequences using software such as the mpileup function of SAMtools (Li et al., 2009), or the now depreciated tools in Maq (Li et al., 2008). The use of a reference genome to guide the assembly of contigs means that areas separated by low coverage of reads can be joined for further analysis. This technique was used to find medium length ncRNAs such as RNase P and RNase MRP RNAs from *Giardia* (Chen et al., 2011). Although the RNase P RNA has been previously identified from *Giardia*, the closely related RNase MRP RNA was found by forming larger contigs from short-reads then using the INFERNAL RNA comparison software (Nawrocki et al., 2009) to compare these contigs to ncRNAs from Rfam. Using this method RNase P and MRP RNAs were either characterized, confirmed for had ambiguous genomic positions clarified (Chen et al., 2011). Comparative genomes has been used to characterize snoRNAs from trypanosomatids including *Leishmania* (Liang et al., 2007) and *T. brucei* (Uliel et al., 2004; Barth et al., 2008; Gupta et al., 2010).

Specialist software such as snoScan (Schattner et al., 2005) can be used to characterize snoRNAs from both classes C/D box and H/ACA. Often it can be necessary to change parameters within this software to permit changes in expected secondary structure. C/D box snoRNAs direct 2′-O methylation, and are relatively easy to identify based on conserved sequence elements (stem-loop features) and complementary binding to the target RNAs. H/ACA box snoRNAs direct pseudouridylation, and often exhibit more variable features due to their shorter length of conserved elements and discontinuous complementary target binding regions. A combined experimental and computation approach using high-throughput sequencing enabled both of these snoRNAs classes to be characterized in *G. lamblia* (Chen et al., 2007). Here, snoScan was modified to search for snoRNAs in the *Giardia* WB genome but these adjustments caused a loosening of parameters, and hence, an increase in false positives being reported. Positive hits were compared to high-throughput sequencing results to filter through the large number of false positives. When the sequencing results were combined with information from potential ribosomal pseudouridylation sites, other snoRNAs were characterized from both *Giardia* and *Trichomonas* (Chen et al., 2011).

There are many snoRNAs even from well characterized species (i.e., humans) that do not have identified targets. These are termed orphan snoRNAs (Bazeley et al., 2008) and it is likely that such snoRNAs will be found in protists, but perhaps not as closely associated with splicing (*Giardia* and *Trichomonas* have few introns). However, without potential binding sites to check results against, much more laboratory work will be required to characterize these orphan snoRNAs, even those potential candidates found by high-throughput sequencing.

CONCLUDING REMARKS

High-throughput sequencing has opened the door on more research into the use of ncRNAs either as tools for investigating

[6] www.microrna.org

protist biology, markers for disease detection or progression, or as potential avenues for treatment. Although there is a long way to go to catch up with ncRNA analysis from host species (e.g., human and mouse), the genomic sequencing of many pathogenic protists is already permitting genome-wide screens of ncRNAs such as miRNAs and siRNAs. Studies of protist miRNAs to date are showing that RNAi has an important role in antigenic variation and hence the parasite's survival in its host. Learning more about these systems could enable more effective strategies to prevent and treat a range of protist diseases. Other areas of active research are now looking at the integration of regulatory RNA (miRNA and siRNA) data with protein gene expression sequencing data (i.e., RNA-seq, or mRNAseq), to characterize how miRNAs control their targets, and are themselves controlled, in different environments. In effect, this is a combination of miRNA expression and target expression, all coming from the same sample.

The use of high-throughput sequencing to uncover and characterize ncRNAs has both the biological relevance of traditional laboratory approaches and the genome-wide scale of the computational approaches. However, it does require the understanding of both the biological and computational aspects of RNA analysis. Although software both for mapping, assembly and sequence manipulation was written for longer mRNAs and genomic sequencing, it can be applied to short ncRNAs such as miRNAs and siRNAs with careful parameter adjustment. It is likely that in the next few years we will see further development of the small RNA sequencing protocols that are available especially as they rise in importance in the medical research world. What is needed to meet this rise is a general upskilling of molecular researchers to deal with the increased bioinformatics that this new technology brings, and further development of software pipelines to make it easier to adapt RNA software to non-mammalian and non-plant species. Protist biology is very different and their ncRNA systems are delivering us many surprises (Collins and Penny, 2009). It is clear that genome biology of host and pathogens can no longer exclude the analysis of non-coding sequences.

ACKNOWLEDGMENTS

Many thanks go to David Penny for his support and critical reading of this manuscript. This work was supported by the New Zealand Marsden Fund.

REFERENCES

Ashfaq, U. A., Yousaf, M. Z., Aslam, M., Ejaz, R., Jahan, S., and Ullah, O. (2011). siRNAs: potential therapeutic agents against hepatitis C virus. *Virol. J.* 8, 276.

Aurrecoechea, C., Brestelli, J., Brunk, B. P., Carlton, J. M., Dommer, J., Fischer, S., Gajria, B., Gao, X., Gingle, A., Grant, G., Harb, O. S., Heiges, M., Innamorato, F., Iodice, J., Kissinger, J. C., Kraemer, E., Li, W., Miller, J. A., Morrison, H. G., Nayak, V., Pennington, C., Pinney, D. F., Roos, D. S., Ross, C., Stoeckert, C. J. Jr., Sullivan, S., Treatman, C., and Wang, H. (2009). GiardiaDB and TrichDB: integrated genomic resources for the eukaryotic protist pathogens *Giardia lamblia* and *Trichomonas vaginalis*. *Nucleic Acids Res.* 37, D526–D530.

Barth, S., Shalem, B., Hury, A., Tkacz, I. D., Liang, X. H., Uliel, S., Myslyuk, I., Doniger, T., Salmon-Divon, M., Unger, R., and Michaeli, S. (2008). Elucidating the role of C/D snoRNA in rRNA processing and modification in *Trypanosoma brucei*. *Eukaryot. Cell* 7, 86–101.

Batista, T. M., and Marques, J. T. (2011). RNAi pathways in parasitic protists and worms. *J. Proteomics* 74, 1504–1514.

Baum, J., Papenfuss, A. T., Mair, G. R., Janse, C. J., Vlachou, D., Waters, A. P., Cowman, A. F., Crabb, B. S., and De Koning-Ward, T. F. (2009). Molecular genetics and comparative genomics reveal RNAi is not functional in malaria parasites. *Nucleic Acids Res.* 37, 3788–3798.

Bazeley, P. S., Shepelev, V., Talebizadeh, Z., Butler, M. G., Fedorova, L., Filatov, V., and Fedorov, A. (2008). snoTARGET shows that human orphan snoRNA targets locate close to alternative splice junctions. *Gene* 408, 172–179.

Biggs, P. J., and Collins, L. J. (2011). RNA networks in prokaryotes I: CRISPRs and riboswitches. *Adv. Exp. Med. Biol.* 722, 209–220.

Bompfunewerer, A. F., Flamm, C., Fried, C., Fritzsch, G., Hofacker, I. L., Lehmann, J., Missal, K., Mosig, A., Muller, B., Prohaska, S. J., Stadler, B. M. R., Stadler, P. F., Tanzer, A., Washietl, S., and Witwer, C. (2005). Evolutionary patterns of non-coding RNAs. *Theory Biosci.* 123, 301–369.

Burnette, J. M., Miyamoto-Sato, E., Schaub, M. A., Conklin, J., and Lopez, A. J. (2005). Subdivision of large introns in *Drosophila* by recursive splicing at non-exonic elements. *Genetics* 170, 661–674.

Carlton, J. M., Hirt, R. P., Silva, J. C., Delcher, A. L., Schatz, M., Zhao, Q., Wortman, J. R., Bidwell, S. L., Alsmark, U. C. M., Besteiro, S., Sicheritz-Ponten, T., Noel, C. J., Dacks, J. B., Foster, P. G., Simillion, C., Van De Peer, Y., Miranda-Saavedra, D., Barton, G. J., Westrop, G. D., Müller, S., Dessi, D., Fiori, P. L., Ren, Q., Paulsen, I., Zhang, H., Bastida-Corcuera, F. D., Simoes-Barbosa, A., Brown, M. T., Hayes, R. D., Mukherjee, M., Okumura, C. Y., Schneider, R., Smith, A. J., Vanacova, S., Villalvazo, M., Haas, B. J., Pertea, M., Feldblyum, T. V., Utterback, T. R., Shu, C.-L., Osoegawa, K., De Jong, P. J., Hrdy, I., Horvathova, L., Zubacova, Z., Dolezal, P., Malik, S.-B., Logsdon, J. M., Henze, K., Gupta, A., Wang, C. C., Dunne, R. L., Upcroft, J. A., Upcroft, P., White, O., Salzberg, S. L., Tang, P., Chiu, C.-H., Lee, Y.-S., Embley, T. M., Coombs, G. H., Mottram, J. C., Tachezy, J., Fraser-Liggett, C. M., and Johnson, P. J. (2007). Draft genome sequence of the sexually transmitted pathogen *Trichomonas vaginalis*. *Science* 315, 207–212.

Chen, X. S., Collins, L. J., Biggs, P. J., and Penny, D. (2009). High throughput genome-wide survey of small rnas from the parasitic protists *Giardia intestinalis* and *Trichomonas vaginalis*. *Genome Biol. Evol.* 2009, 165–175.

Chen, X. S., Penny, D., and Collins, L. J. (2011). Characterization of RNase MRP RNA and novel snoRNAs from *Giardia intestinalis* and *Trichomonas vaginalis*. *BMC Genomics* 12, 550. doi:10.1186/1471-2164-12-550

Chen, X. S., Rozhdestvensky, T. S., Collins, L. J., Schmitz, J., and Penny, D. (2007). Combined experimental and computational approach to identify non-protein-coding RNAs in the deep-branching eukaryote *Giardia intestinalis*. *Nucleic Acids Res.* 35, 4619–4628.

Chen, X. S., White, W. T., Collins, L. J., and Penny, D. (2008). Computational identification of four spliceosomal snRNAs from the deep-branching eukaryote *Giardia intestinalis*. *PLoS ONE* 3, e3106. doi:10.1371/journal.pone.0003106

Cock, P. J., Fields, C. J., Goto, N., Heuer, M. L., and Rice, P. M. (2010). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res.* 38, 1767–1771.

Collins, L., and Penny, D. (2005). Complex spliceosomal organization ancestral to extant eukaryotes. *Mol. Biol. Evol.* 22, 1053–1066.

Collins, L. J., and Biggs, P. J. (2011). RNA networks in prokaryotes II: tRNA processing and small RNAs. *Adv. Exp. Med. Biol.* 722, 221–230.

Collins, L. J., and Chen, X. S. (2009). Ancestral RNA: the RNA biology of the eukaryotic ancestor. *RNA Biol.* 6, 495–502.

Collins, L. J., and Penny, D. (2009). The RNA infrastructure: dark matter of the eukaryotic cell? *Trends Genet.* 25, 120–128.

Cudmore, S. L., Delgaty, K. L., Hayward-Mcclelland, S. F., Petrin, D. P., and Garber, G. E. (2004). Treatment of infections caused by metronidazole-resistant *Trichomonas vaginalis*. *Clin. Microbiol. Rev.* 17, 783–793.

De, S., Pal, D., and Ghosh, S. K. (2006). *Entamoeba histolytica*: computational identification of putative microRNA candidates. *Exp. Parasitol.* 113, 239–243.

Drinnenberg, I. A., Fink, G. R., and Bartel, D. P. (2011). Compatibility with killer explains the rise of RNAi-deficient fungi. *Science* 333, 1592.

Drinnenberg, I. A., Weinberg, D. E., Xie, K. T., Mower, J. P., Wolfe, K. H., Fink, G. R., and Bartel, D. P. (2009). RNAi in budding yeast. *Science* 326, 544–550.

Finn, R. D., Mistry, J., Schuster-Bockler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., Moxon, S., Marshall, M., Khanna, A., Durbin, R., Eddy, S. R., Sonnhammer, E. L., and Bateman, A. (2006). Pfam: clans, web tools and services. *Nucleic Acids Res.* 34, D247–D251.

Fire, A., Xu, S., Montgomery, M. K., Kostas, S. A., Driver, S. E., and Mello, C. C. (1998). Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans. Nature* 391, 806–811.

Garcia-Silva, M. R., Frugier, M., Tosar, J. P., Correa-Dominguez, A., Ronalte-Alves, L., Parodi-Talice, A., Rovira, C., Robello, C., Goldenberg, S., and Cayota, A. (2010). A population of tRNA-derived small RNAs is actively produced in *Trypanosoma cruzi* and recruited to specific cytoplasmic granules. *Mol. Biochem. Parasitol.* 171, 64–73.

Gardner, P. P., Bateman, A., and Poole, A. M. (2010). SnoPatrol: how many snoRNA genes are there? *J. Biol.* 9, 4.

Garlapati, S., Saraiya, A. A., and Wang, C. C. (2011). A la autoantigen homologue is required for the internal ribosome entry site mediated translation of giardiavirus. *PLoS ONE* 6, e18263. doi:10.1371/journal.pone.0018263

Gupta, S. K., Hury, A., Ziporen, Y., Shi, H., Ullu, E., and Michaeli, S. (2010). Small nucleolar RNA interference in *Trypanosoma brucei*: mechanism and utilization for elucidating the function of snoRNAs. *Nucleic Acids Res.* 38, 7236–7247.

Haasnoot, J., and Berkhout, B. (2011). RNAi and cellular miRNAs in infections by mammalian viruses. *Methods Mol. Biol.* 721, 23–41.

Ketting, R. F. (2011). The many faces of RNAi. *Dev. Cell* 20, 148–161.

Khaliq, S., Khaliq, S. A., Zahur, M., Ijaz, B., Jahan, S., Ansar, M., Riazuddin, S., and Hassan, S. (2010). RNAi as a new therapeutic strategy against HCV. *Biotechnol. Adv.* 28, 27–34.

Khanna, A., and Stamm, S. (2010). Regulation of alternative splicing by short non-coding nuclear RNAs. *RNA Biol.* 7, 480–485.

Kolev, N. G., Tschudi, C., and Ullu, E. (2011). RNA interference in protozoan parasites: achievements and challenges. *Eukaryot. Cell* 10, 1156–1163.

Kolev, N. G., and Ullu, E. (2009). snoRNAs in *Giardia lamblia*: a novel role in RNA silencing? *Trends Parasitol.* 25, 348–350.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 10, R25.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079.

Li, H., Ruan, J., and Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* 18, 1851–1858.

Liang, X. H., Hury, A., Hoze, E., Uliel, S., Myslyuk, I., Apatoff, A., Unger, R., and Michaeli, S. (2007). Genome-wide analysis of C/D and H/ACA-like small nucleolar RNAs in Leishmania major indicates conservation among trypanosomatids in the repertoire and in their rRNA targets. *Eukaryot. Cell* 6, 361–377.

Liao, J., Yu, L., Mei, Y., Guarnera, M., Shen, J., Li, R., Liu, Z., and Jiang, F. (2010). Small nucleolar RNA signatures as biomarkers for non-small-cell lung cancer. *Mol. Cancer* 9, 198.

Lin, W. C., Li, S. C., Lin, W. C., Shin, J. W., Hu, S. N., Yu, X. M., Huang, T. Y., Chen, S. C., Chen, H. C., Chen, S. J., Huang, P. J., Gan, R. R., Chiu, C. H., and Tang, P. (2009). Identification of microRNA in the protist *Trichomonas vaginalis. Genomics* 93, 487–493.

Lye, L. F., Owens, K., Shi, H., Murta, S. M., Vieira, A. C., Turco, S. J., Tschudi, C., Ullu, E., and Beverley, S. M. (2011). Retention and loss of RNA interference pathways in trypanosomatid protozoans. *PLoS Pathog.* 6, e1001161. doi:10.1371/journal.ppat.1001161

Macrae, I. J., Zhou, K., Li, F., Repic, A., Brooks, A. N., Cande, W. Z., Adams, P. D., and Doudna, J. A. (2006). Structural basis for double-stranded RNA processing by Dicer. *Science* 311, 195–198.

McCormick, K. P., Willmann, M. R., and Meyers, B. C. (2011). Experimental design, preprocessing, normalization and differential expression analysis of small RNA sequencing experiments. *Silence* 2, 2.

Nawrocki, E. P., Kolbe, D. L., and Eddy, S. R. (2009). Infernal 1.0: inference of RNA alignments. *Bioinformatics* 25, 1335–1337.

Ngo, H., Tschudi, C., Gull, K., and Ullu, E. (1998). Double-stranded RNA induces mRNA degradation in *Trypanosoma brucei. Proc. Natl. Acad. Sci. U.S.A.* 95, 14687–14692.

Pantaleo, V. (2011). Plant RNA silencing in viral defence. *Adv. Exp. Med. Biol.* 722, 39–58.

Piccinelli, P., Rosenblad, M. A., and Samuelsson, T. (2005). Identification and analysis of ribonuclease P and MRP RNA in a broad range of eukaryotes. *Nucleic Acids Res.* 33, 4485–4495.

Prucca, C. G., Slavin, I., Quiroga, R., Elías, E. V., Rivero, F. D., Saura, A., Carranza, P. G., and Luján, H. D. (2008). Antigenic variation in *Giardia lamblia* is regulated by RNA interference. *Nature* 456, 750–754.

Ridanpaa, M., Ward, L. M., Rockas, S., Sarkioja, M., Makela, H., Susic, M., Glorieux, F. H., Cole, W. G., and Makitie, O. (2003). Genetic changes in the RNA components of RNase MRP and RNase P in Schmid metaphyseal chondrodysplasia. *J. Med. Genet.* 40, 741–746.

Rivero, M. R., Kulakova, L., and Touz, M. C. (2010). Long double-stranded RNA produces specific gene downregulation in *Giardia lamblia. J. Parasitol.* 96, 815–819.

Rudra, D., Mallick, J., Zhao, Y., and Warner, J. R. (2007). Potential interface between ribosomal protein production and pre-rRNA processing. *Mol. Cell. Biol.* 27, 4815–4824.

Ryther, R. C., Mcguinness, L. M., Phillips, J. A. III, Moseley, C. T., Magoulas, C. B., Robinson, I. C., and Patton, J. G. (2003). Disruption of exon definition produces a dominant-negative growth hormone isoform that causes somatotroph death and IGHD II. *Hum. Genet.* 113, 140–148.

Saraiya, A. A., and Wang, C. C. (2008). snoRNA, a novel precursor of microRNA in *Giardia lamblia. PLoS Pathog.* 4, e1000224. doi:10.1371/journal.ppat.1000224

Schattner, P., Brooks, A. N., and Lowe, T. M. (2005). The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Res.* 33, W686–W689.

Simoes-Barbosa, A., Meloni, D., Wohlschlegel, J. A., Konarska, M. M., and Johnson, P. J. (2008). Spliceosomal snRNAs in the unicellular eukaryote *Trichomonas vaginalis* are structurally conserved but lack a 5′-cap structure. *RNA* 14, 1617–1631.

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J., and Birol, I. (2009). ABySS: a parallel assembler for short read sequence data. *Genome Res.* 19, 1117–1123.

Smith, A., and Johnson, P. (2011). Gene expression in the unicellular eukaryote *Trichomonas vaginalis. Res. Microbiol.* 162, 646–654.

Teodorovic, S., Walls, C. D., and Elmendorf, H. G. (2007). Bidirectional transcription is an inherent feature of *Giardia lamblia* promoters and contributes to an abundance of sterile antisense transcripts throughout the genome. *Nucleic Acids Res.* 35, 2544–2553.

Uliel, S., Liang, X. H., Unger, R., and Michaeli, S. (2004). Small nucleolar RNAs that guide modification in trypanosomatids: repertoire, targets, genome organisation, and unique functions. *Int. J. Parasitol.* 34, 445–454.

Vaishnaw, A. K., Gollob, J., Gamba-Vitalo, C., Hutabarat, R., Sah, D., Meyers, R., De Fougerolles, T., and Maraganore, J. (2011). A status report on RNAi therapeutics. *Silence* 1, 14.

Yang, C. Y., Zhou, H., Luo, J., and Qu, L. H. (2005). Identification of 20 snoRNA-like RNAs from the primitive eukaryote, *Giardia lamblia. Biochem. Biophys. Res. Commun.* 328, 1224–1231.

Yang, L., Duff, M. O., Graveley, B. R., Carmichael, G. G., and Chen, L. L. (2011). Genomewide characterization of non-polyadenylated RNAs. *Genome Biol.* 12, R16.

Zerbino, D. R., and Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* 18, 821–829.

Zhang, H., Pompey, J. M., and Singh, U. (2011a). RNA interference in *Entamoeba histolytica*: implications for parasite biology and gene silencing. *Future Microbiol.* 6, 103–117.

Zhang, Z. X., Min, W. P., and Jevnikar, A. M. (2011b). Use of RNA interference to minimize ischemia reperfusion injury. *Transplant. Rev. (Orlando)*. PMID: 22000663. [Epub ahead of print].

Zhang, Y. Q., Chen, D. L., Tian, H. F., Zhang, B. H., and Wen, J. F. (2009). Genome-wide computational identification of microRNAs and their targets in the deep-branching eukaryote *Giardia lamblia*. *Comput. Biol. Chem.* 33, 391–396.

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 05 November 2011; paper pending published: 28 November 2011; accepted: 07 December 2011; published online: 27 December 2011.

Citation: Collins LJ (2011) Characterizing ncRNAs in human pathogenic protists using high-throughput sequencing technology. Front. Gene. ***2****:96. doi: 10.3389/fgene.2011.00096*

This article was submitted to Frontiers in Non-Coding RNA, a specialty of Frontiers in Genetics.

frontiers in
GENETICS

REVIEW ARTICLE
published: 28 December 2011
doi: 10.3389/fgene.2011.00095

A network of regulations by small non-coding RNAs: the P-TEFb kinase in development and pathology

Hossein Ghanbarian*[1,2†], *Valérie Grandjean*[1,2], *François Cuzin*[1,2] and *Minoo Rassoulzadegan*[1,2]

[1] *INSERM U636, Nice, France*
[2] *Laboratoire de Génétique du Développement Normal et Pathologique, Université de Nice-Sophia Antipolis, Nice, France*

Edited by:
Philipp Kapranov, St. Laurent Institute, USA

Reviewed by:
Zhaohui Wang, University of Texas Southwestern Medical Center, USA
Philipp Kapranov, St. Laurent Institute, USA
Joseph Nadeau, Institute for Systems Biology, USA

****Correspondence:***
Minoo Rassoulzadegan, INSERM U636, Université de Nice-Sophia Antipolis, Campus Valrose, 06108 Nice Cedex 2, France.
e-mail: minoo@unice.fr

†Present address:
Hossein Ghanbarian, Department of Molecular Biology and Genetic Engineering, Stem Cell Technology Research Center, Tehran 15856364473, Iran.

Part of the heterodimeric P-TEF-b element of the Pol II transcription machinery, the cyclin-dependent kinase 9 plays a critical role in gene expression. Phosphorylation of several residues in the polymerase is required for elongation of transcript. It determines the rates of transcription and thus, plays a critical role in several differentiation pathways, best documented in heart development. The synthesis and activity of the protein are tightly regulated in a coordinated manner by at least three non-coding RNAs. First, its kinase activity is reversibly inhibited by formation of a complex with the 334 nt 7SK RNA, from which it is released under conditions of stress. Then, heart development requires a maximal rate of synthesis during cardiomyocyte differentiation, followed by a decrease in the differentiated state. The latter is insured by microRNA-mediated translational inhibition. In a third mode of RNA control, increased levels of transcription are induced by small non-coding RNA molecules with sequences homologous to the transcript. Designated paramutation, this epigenetic variation, stable during development, and hereditarily transmitted in a non-Mendelian manner over several generations, is thought to be a response to the inactivation of one of the two alleles by an abnormal recombination event such as insertion of a transposon.

Keywords: mice, cardiac hypertrophy, epigenetic, paramutation, heredity, 7SK, Cdk9, non-coding RNA

INTRODUCTION

Understanding the controls of gene expression by non-coding RNAs, the major part of the "dark matter" of the genome (Kapranov et al., 2010; Mattick et al., 2010), is becoming a major goal in developmental biology, physiology, and pathology. Several mechanisms of such control have been uncovered – and the list is certainly not complete. Our current work is focused on a network involving at least three distinct small non-coding (snc) RNAs, which work synergistically to adjust the levels of synthesis and activity of a pleiotropic protein, the cyclin-dependent kinase 9 (Cdk9), part of the heterodimeric complex designated P-TEFb. Its serine–threonine kinase activity plays an essential role in RNA polymerase II transcription and is currently recognized as a critical actor in a number of physiological and pathological processes. It is therefore understandable that both synthesis and activity of Cdk9 are tightly regulated. The control of its translation, critical in heart differentiation, is exerted by several microRNAs, primarily miR-1 and miR-133. Its enzymatic activity is independently regulated by formation of a complex with a 332 nt non-coding RNA, 7SK, which promotes the reversible integration of protein inhibitors in P-TEFb. A third avenue of regulation is via activation of transcription by sequence-related oligoribonucleotides that we reported for *Cdk9* and two other genes (Rassoulzadegan et al., 2006; Wagner et al., 2008; Grandjean et al., 2009).

Cdk9, A PROTEIN AT A NODAL POINT IN DEVELOPMENT AND DIFFERENTIATION

The Cdk9 protein is associated with a cyclin from either the T or the K families in the heterodimer P-TEFb, which plays a key role in the Polymerase II transcription machinery (Kohoutek, 2009). The Cdk9 component functions as the catalytic domain, while the cyclin constitutes the regulatory subunit (Malumbres and Barbacid, 2005). P-TEFb is required after binding of the initiation complex to a promoter – the phosphorylation of a C-terminal domain of the polymerase by P-TEFb allows elongation of the nascent transcript. Levels of transcription are therefore determined by the recruitment of P-TEFb to promoters, in turn dependent on its interactions with other proteins such as the bromodomain protein Brd4 (Yang et al., 2005). Cdk9 is also part of the p300/GATA4 complex required for the expression of cardiac specific genes such as *Nkx2.5*, *Anf*, and *ß-Myh* (Kaichi et al., 2011 and references therein). P-TEFb interacts with Myosin skeletal muscle differentiation (Simone et al., 2002) and interactions were documented with a plethora of transcription factors (Kohoutek, 2009). Phosphorylation of substrates other than RNA Pol II are also likely to be important, for instance that of several sites in p53 (Radhakrishnan and Gartel, 2006). A role in the activation of quiescent B cells was reported (De Falco et al., 2008), and a general function has been proposed in transcription coupled alternative splicing (Barboric et al., 2009). While the P-TEFb complex as a whole is a functional kinase unit, its activity appears to be essentially modulated by variations of the synthesis and/or activity of the Cdk9 moiety. Several RNA-mediated regulatory circuits have been identified. Our knowledge is still, however, incomplete, as illustrated by the paucity of data on the differential expression of the two isoforms encoded at the *Cdk9* locus and on their respective functions. In addition to the ubiquitous 42 kDa protein, a 55-kDa species transcribed from a distinct upstream promoter includes an additional N-terminal sequence of 117 aminoacids (Giacinti et al., 2008). Differential expression of the two proteins has been reported in various cell types and conditions (reviewed in Romano and Giordano, 2008). We observed that the longer isoform, initially not expressed in ES cells, is transcribed in the course of their differentiation into beating cardiac cells (Hossein Ghanbarian et al., unpublished).

Variations in Cdk9 expression, hence on P-TEFb activity, are associated with various diseases. Among them, cardiac hypertrophy was the most extensively documented. Cardiomyocyte differentiation normally involves an increase in cell size that requires a general activation of transcription together with that of the heart-specific genes. The kinase is required in normal differentiation and its overexpression results in cardiac hypertrophy (Sano et al., 2002; Wagner et al., 2008). Additional pathological developments are associated with abnormal Cdk9 expression (Kohoutek, 2009). One well-documented instance is viral immunodeficiency. A complex of Cdk9 with the Tat proteins of HIV1 and HIV2 viruses enhances their interaction with Tar RNA and thus, the transcription of the viral genome (Wei et al., 1998). In human gastric tumors (He et al., 2008), high levels of P-TEFb activity are generated by a mutation deleting part of LARP7, a protein that participates in inhibition of Cdk9 in the complex with 7SK RNA (see below). In addition to the immunodeficiency family, other viruses of clinical interest are dependent on the Cdk9 kinase, which therefore appears as a likely target for the development of new pharmacological strategies (reviewed by Romano and Giordano, 2008; Wang and Fischer, 2008).

MULTIPLE REGULATIONS BY SMALL NON-CODING RNAs

Given its essential functions, it is not unexpected that both the synthesis and activity of the protein are under tight regulatory controls. Three distinct circuits enacted by sncRNAs illustrate their power and versatility. We will briefly review the first one, translation control by microRNAs for which excellent reviews are available (Kohoutek, 2009; Sayed and Abdellatif, 2011), and then the reversible inhibition of enzymatic activity in the 7SK complex. We will discuss in more detail the properties and possible function of the inherited epigenetic increase in *Cdk9* transcription in the “Cdk9* paramutants” (Wagner et al., 2008).

DOWN REGULATION BY microRNA: THE CASE OF miR-1

Two of the microRNAs detected in heart (Callis and Wang, 2008) have been the subject of intensive studies, miR-1 and miR-133. miR-1, most important in heart development, targets the 3′ untranslated region of the *Cdk9* transcript. In undifferentiated ES cells, miR-1 is either absent or barely detectable. When cardiac differentiation is induced by culture in suspension (Boheler et al., 2002), expression progressively increases. Ectopic expression in the initial ES cultures inhibits differentiation while reducing expression of Cdk9 (Takaya et al., 2009). Forced expression of miR-1 in embryonic cardiomyocytes inhibits their proliferation and increases differentiation (Zhao et al., 2005). Down regulation of Cdk9 is clearly critical to keep cardiac growth within the physiological limits, and the microRNA also targets *Hand2* mRNA, which encodes a transcription factor required for ventricular myocyte expansion. miR-1 and miR-133 play opposite roles in cardiomyocyte growth and apoptosis. Unlike miR-1, which is apoptotic and targets the anti-apoptotic heat shock proteins HSP60 and HSP70, miR-133 is anti-apoptotic and represses caspase-9, a regulator of mitochondria-mediated apoptosis (Xu et al., 2007). It is expected that more miRNAs will be added to the growing list of cardiac regulators (Callis and Wang, 2008).

REGULATION OF P-TEFb ENZYMATIC ACTIVITY: THE 7SK COMPLEX

A fraction of the Cdk9 protein is stored in a reversibly inactivated form. Inhibition is achieved in a complex with the evolutionary conserved 7SK sncRNA (Nguyen et al., 2001; Yang et al., 2001). 7SK bridges the protein to either one of two proteins, Hexim1 and Hexim2, which inhibit the kinase activity (Yik et al., 2003; Byers et al., 2005). In addition, two other proteins, MePCE and LARP7 are required for the stabilization of the complex by inhibiting degradation of the 7SK RNA (Krueger et al., 2008). Complex formation is fully reversible: its dissociation results in a burst of Cdk9 activity under conditions of stress, UV irradiation, mechanical load, and pharmacological treatments by “hypertrophic agonists” such as endothelin-1, phenylephrine, calcineurin (Nguyen et al., 2001; Sano et al., 2002; Espinoza-Derout et al., 2009). As one additional enzyme was found associated with 7SK complexes (He et al., 2006), one may consider a more general type of regulation by the means of readily reversible inhibition of enzyme activity in complexes with inhibitory proteins mediated by non-coding RNAs.

LONG TERM TRANSCRIPTIONAL ACTIVATION BY HOMOLOGOUS OLIGORIBONUCLEOTIDES AND MICRORNAs

A third mode of regulation of *Cdk9* expression directed by sncRNAs is epigenetic transcriptional activation during heart development. It results in a dramatic heart hypertrophy syndrome stably inherited over several generations (Wagner et al., 2008), designated “paramutation” by analogy with the phenomenon of plant hereditary epigenetic variations (Chandler, 2007). We previously reported a heritable epigenetic modification of the *Kit* gene induced by microinjecting mouse fertilized eggs with oligoribonucleotides with two types of sequences: (1) corresponding to short stretches of the *Kit* transcript and (2) corresponding to certain endogenous microRNAs (Rassoulzadegan et al., 2006). In a quest for other examples, we practiced a series of injections of other microRNAs and transcript fragments, among them oligoribonucleotides with a sequence from *Cdk9* mRNA and the miR-1 microRNA. Their injection into one-cell embryos followed by reimplantation in foster mothers resulted in increased expression of *Cdk9* in cardiomyocyte precursors at the E18.5 embryonic stage (Wagner et al., 2008).

All the treated embryos developed into newborns and adults showed increased heart size: 1.5- to 2-fold that of the controls, with a disrupted organization of the wall ultrastructure and TUNEL-positive cells. All these features are characteristic of the human hypertrophic cardiomyopathy, a severe condition with a poor prognosis and, remarkably, a frequent familial distribution. In the mouse, the cardiac phenotype was efficiently transmitted to the offspring in three generations of crosses with normal partners. Increased P-TEFb activity evidenced by Pol II phosphorylation resulted from elevated levels of *Cdk9* mRNA resulting from increased transcriptional activity. Levels of the Cyclin T1 and 7SK RNAs remained unchanged.

Two remarkable features of the epigenetic variation are worth noting. First, the *Cdk9* gene remains subject to the same developmental regulations as in the wild type controls. Overexpression, as seen in the cardiomyocyte precursors (E18.5) and normal development, affects only the level of the short isoform exclusively made in the wild type. The other remarkable feature, common to the different instances of paramutation so far analyzed, is the mode of inheritance described for other epigenetic variations as "rheostat transmission" (Beaudet and Jiang, 2002). The quantitative phenotypic modulation, in this case heart size, appears to be set separately for every individual at some early time in development. In other words, mice with large hearts will generate in the same litter animals with very large, large, more than average, and nearly normal heart sizes. Conversely, mice with nearly normal or normal hearts will generate the same assortment of phenotypes, including the very large ones (Wagner et al., 2008). This pattern of phenotypic distribution is more evocative of a continuous modulation such as that of a rheostat than of the on–off "switch" of most genetic controls.

LONG TERM ACTIVATION OF *Cdk9* IN EMBRYONIC STEM CELLS

With the aim to limit the complexity of the experimental approach, we tested whether the same oligoribonucleotides would activate transcription of *Cdk9* in embryonic stem cells in culture (ES lines) and affect their differentiation into cardiac precursors (Hossein Ghanbarian and Minoo Rassoulzadegan, manuscript in preparation). In order to be able to compare the cell culture and *in vivo* results, we made use of a mouse ES cell line that we have previously tested for being able to successfully incorporate in the embryo and give rise to whole spectrum of differentiated cell types. The first question was whether transcription of the genes paramutated in the mouse would be increased to the same extent upon RNA transfer by electroporation in ES cells and this turned out to be the case. We were then able to compare these results with the other instance of transcriptional activation by homologous RNAs, recently reported in human cell lines in culture (Place et al., 2008; Morris, 2009; Huang et al., 2010). A critical question was the stability of the activated state. Unlike the long term effects generated in the mouse, the activated state in human cells was maintained for only a limited number of cell doublings. Results in the ES cells show a 1.5- to 2-fold activation of *Cdk9* transcription within the first days in the transfected culture. Cotransfer of a selectable marker allowed us to isolate clones maintaining elevated (1.5- to 3-fold) transcription rates, eventually reverting to the wild type level, but only after their propagation in culture for 15–20 cell generations. The field of RNA paramutation will thus now benefit from the availability of a cell culture system to hopefully reach a mechanistic explanation of RNA induction, an objective that was obviously more difficult to attain with the limited amounts of material offered by early mouse embryos. Several lines of research are now started on the cell culture system. One of them is a point by point comparison with the short term induction of transcription described in human cells. The two studies involved different genes and the possible induction of *Kit* in human cells, for instance, or paramutation of *Cdh1* (E-cadherin), inducible in the mouse (Huang et al., 2010), would allow more significant comparisons. An important mechanistic aspect is that in human cells, induction requires the Argonaute system and mouse cells can now be tested in this respect. One most intriguing aspect is the fact that, in ES cells as in the mouse, oligoribonucleotides with fragments of the sense sequence start the efficient and highly specific paramutation process. Recognition by base pairing would be the most obvious explanation, leading to the assumption that natural antisense transcripts are present in the cell. This class of non-coding RNAs has been identified for various loci and their regulatory roles is currently a matter of attention in several laboratories (reviewed by Werner and Swan, 2010).

The modified ES clones showed another property of interest (Hossein Ghanbarian and Minoo Rassoulzadegan, manuscript in preparation). The cells do not appear as even partially differentiated. They express characteristic ES cell markers such as *Nanog* (Silva et al., 2009) and none of the genes characteristic of cardiomyocytes. Still, the epigenetic variation had conferred on the cells the capacity to differentiate faster and more efficiently into the cardiac lineage. This was clearly ascertained when cardiac differentiation was induced in the cultures (Boheler et al., 2002). Briefly, cells are first grown in suspended aggregates, designated embryoid bodies ("EB"), then EBs separately replated as individual small cultures. Over the next days, cardiac marker genes are expressed with a defined time course, until eventually the appearance of rhythmically beating cardiac cells. They appeared at a much earlier time and at higher frequencies in cultures of the Cdk9 paramutant cells than in the control and the same was true of the expression of cardiac markers such as the cardiac myosin genes. Interestingly expression of *Cdk9* in the differentiated cardiomyocytes had returned to the value of the controls, as in the adult heart of the paramutant mice (Wagner et al., 2008). Directing differentiation of pluripotent stem cells to a differentiated state will be critical for their intended use in therapeutic and regenerative medicine. If generalized to other differentiation states, paramutation might offer one possible approach to this goal. A similar result, based on a different RNA-based technology, has been recently reported (Warren et al., 2010).

DISTINCT REGULATIONS FOR DIFFERENT PHYSIOLOGICAL SITUATIONS

An overall picture of *Cdk9* regulation by sncRNAs, albeit still incomplete and requiring further refinement, is emerging from these results. Under different physiological conditions, expression levels are set by distinct RNA-mediated regulatory pathways summarized in **Figure 1**. During cardiac differentiation, involving both cell growth and cell division, a general activation of transcription is achieved only for a limited period, with continuing

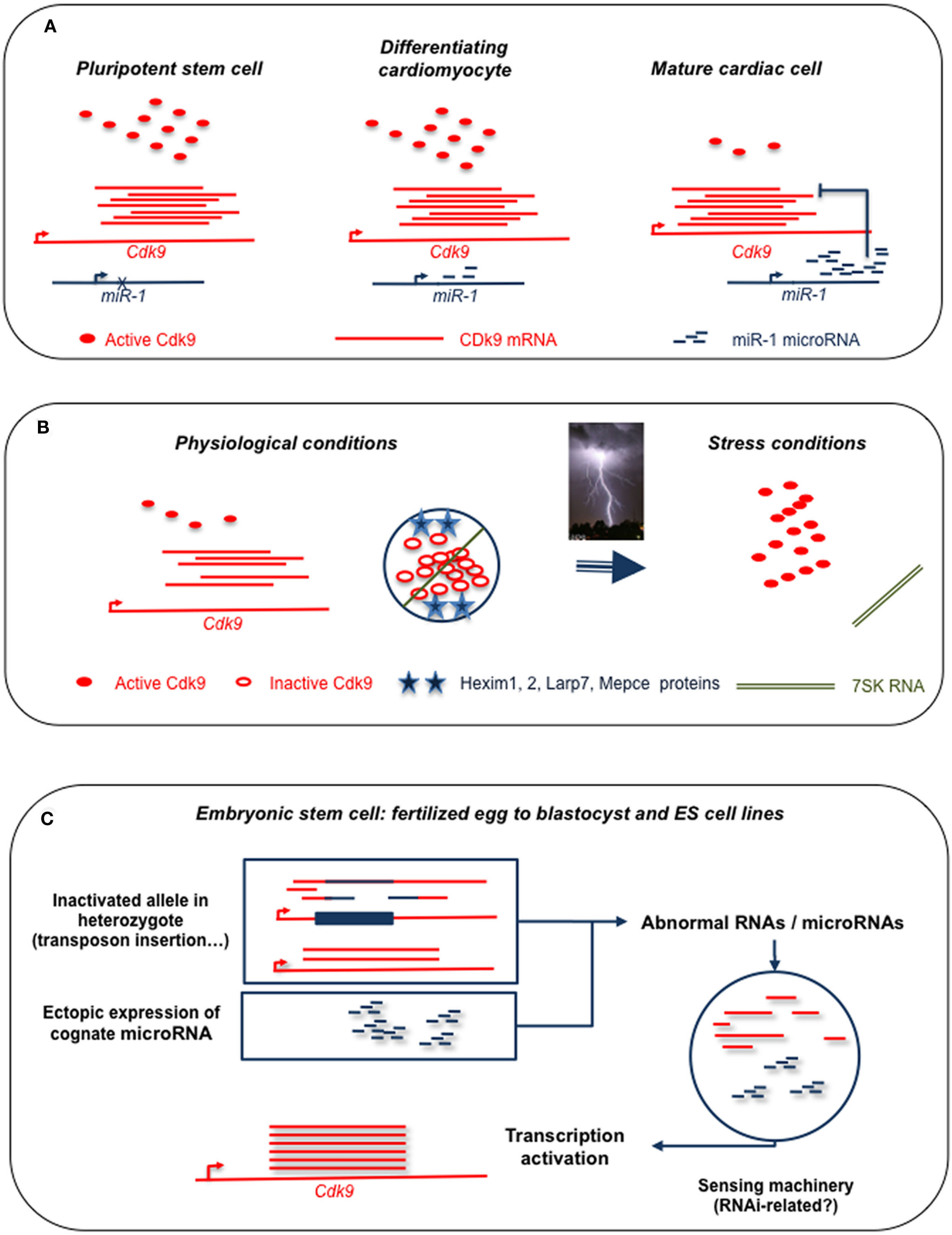

A
Pluripotent stem cell
Differentiating cardiomyocyte
Mature cardiac cell
Cdk9
miR-1
Active Cdk9
CDk9 mRNA
miR-1 microRNA
B
Physiological conditions
Stress conditions
Cdk9
Active Cdk9
Inactive Cdk9
Hexim1, 2, Larp7, Mepce proteins
7SK RNA
C
Embryonic stem cell: fertilized egg to blastocyst and ES cell lines
Inactivated allele in heterozygote (transposon insertion…)
Ectopic expression of cognate microRNA
Abnormal RNAs / microRNAs
Transcription activation
Sensing machinery (RNAi-related?)
Cdk9

FIGURE 1 | Schematic representation of sncRNA-mediated regulations of Cdk9 expression. (A) Down regulation by microRNA miR-1 of Cdk9 translation in differentiated cardiomyocytes following the high levels of expression required during the differentiation process. **(B)** Storage of inactive protein in a complex with 7SK RNA and inhibitory proteins, release of active kinase in stress conditions. **(C)** Summary of our current observations on the initiation of paramutation. Observed either in heterozygote genotypes with one allele disrupted by insertion as in the case of Kit* paramutants (Rassoulzadegan et al., 2006) or upon accumulation of a cognate microRNA, miR-1 in the case of Cdk9. Abnormal RNAs have to be detected by a sensing system in the embryonic cell. Suggestion that the sensor involves the RNAi machinery is based on our current data showing a requirement for Argonaut proteins. Transcription is then upregulated, possibly by a mechanism related to the RNA activation process reported in human cells (Place et al., 2008; Morris, 2009; Huang et al., 2010). As shown for miR-124 and the Sox9* paramutation (Grandjean et al., 2009), heritability is explained by the transfer of the inducing RNAs by oocyte and sperm (not shown).

cell division leading to overgrowth and eventually to hypertrophy. Co-induction of miR-1 and other cardiac markers promotes a progressive decline in Cdk9 synthesis. This is clearly only part of the story and future work will surely make this simple scheme more complex. The additional angle of RNA-mediated regulation is to make the storage of the protein in a reversibly inactivated state respond to a distinct requirement of the system, namely to provide an immediate burst of active enzyme under stress conditions.

Then, the possible "raison d'être" of the long term, hereditary form of regulation of gene expression induced by short RNAs with fragments of the sense sequence can only be at this stage a matter of speculation. Coming back to the initial description of the paramutation of the *Kit* gene (Rassoulzadegan et al., 2006), it is important to note that the modification was detected in the progeny of heterozygotes carrying a gene disrupted by exogenous sequences. Increased expression of the intact *Kit*$^{+}$ allele in germ cells was stably maintained during development and inherited in serial crosses. Abnormal transcript fragments generated from the disrupted allele appeared as the likely inducers of the epigenetic change that led to the change in gene expression. Induction by the cognate miRNAs (miR-221 and -222 for *Kit*, miR-1 for *Cdk9*, miR-124 for *Sox9*) may result from the induction by the microRNAs of degradation of the transcript into the inducing RNA fragments in cells in which the miRNA is not normally present. Alternatively, the ectopically expressed microRNAs may be acting as inducers by themselves. Triggered by abnormal forms of the transcript, paramutation of the wild type locus may be, in teleonomic terms and under natural circumstances, a way to compensate for the inactivation of the other allele by a non-homologous recombination event such as insertion of a transposon.

One important question is whether paramutation could be considered as a general property of a mouse gene – or its establishment a general property of a miRNA. In addition to the consideration that in this case, it should have been recognized a long time ago, several observations clearly point to a negative answer. No clear phenotype was generated in our hands by a series of oligoribonucleotides tested in the microinjection assay and/or in transfected ES cells. It is also significant in this respect that miR-1 efficiently induced, both in the mouse and in ES cells, the long term overexpression of *Cdk9*, but not of *Hand2*, a recognized target of the microRNA (Wagner et al., 2008). One may then have to consider the process as characteristic of a subset of genes. It may be noteworthy that all three genes for which paramutation events have been detected, *Kit*, *Cdk9*, and *Sox9*, are critical for mouse development to the point that homozygous negative mutants are embryonic lethals. On a purely speculative basis, we would thus be tempted to consider that we are dealing with a rescue mechanism, offered to a limited number of loci, that compensates the defective state of one locus by increasing the expression of the allelic wild type gene.

Finally, one may wonder whether the biological processes underlying the observations of "paramutation" are identical in plants and in the mouse. Strictly speaking, they both follow Chandler's definition as "RNA-mediated instructions passed across generations" (Chandler, 2007). The two processes appear, however, distinct in their most basic aspects – as distinct as mouse is from maize. Primarily, the plant paramutation phenomenon is a silencing process and the current thinking in the field considers RNA inducers, whose role was deduced from the requirements for plant-specific RNA polymerases, as guides for a DNA methylation process (Simon and Meyers, 2011). In the mouse, as discussed in the above sections, we and others reported, to the contrary, a transcriptional activation induced by sequence-related small RNAs – a clearly distinct process with, at the present time, no established relationship to DNA methylation (our unpublished data). Thus, a case can be made that a possible confusion is introduced by the use of the same word, "paramutation," to describe the two phenomena, a confusion for which we are ourselves responsible for a significant part. It remains however that they have in common is the notion of that heredity may not be circumscribed to the strict domain of Mendelism, with possible important implications in multiple biological fields including pathology, developmental, and evolutionary biology.

REFERENCES

Barboric, M., Lenasi, T., Chen, H., Johansen, E. B., Guo, S., and Peterlin, B. M. (2009). 7SK snRNP/P-TEFb couples transcription elongation with alternative splicing and is essential for vertebrate development. *Proc. Natl. Acad. Sci. U.S.A.* 106, 7798–7803.

Beaudet, A. L., and Jiang, Y. H. (2002). A rheostat model for a rapid and reversible form of imprinting-dependent evolution. *Am. J. Hum. Genet.* 70, 1389–1397.

Boheler, K. R., Czyz, J., Tweedie, D., Yang, H. T., Anisimov, S. V., and Wobus, A. M. (2002). Differentiation of pluripotent embryonic stem cells into cardiomyocytes. *Circ. Res.* 91, 189–201.

Byers, S. A., Price, J. P., Cooper, J. J., Li, Q., and Price, D. H. (2005). HEXIM2, a HEXIM1-related protein, regulates positive transcription elongation factor b through association with 7SK. *J. Biol. Chem.* 280, 16360–16367.

Callis, T. E., and Wang, D. Z. (2008). Taking microRNAs to heart. *Trends. Mol. Med.* 14, 254–260.

Chandler, V. L. (2007). Paramutation: RNA-mediated instructions passed across generations. *Cell* 23, 641–645.

De Falco, G., Leucci, E., Onnis, A., Bellan, C., Tigli, C., Wirths, S., Cerino, G., Cocco, M., Crupi, D., De Luca, A., Lanzavecchia, A., Tosi, P., Leoncini, L., and Giordano, A. (2008). Cdk9/Cyclin T1 complex: a key player during the activation/differentiation process of normal lymphoid B cells. *J. Cell. Physiol.* 215, 276–282.

Espinoza-Derout, J., Wagner, M., Salciccioli, L., Lazar, J. M., Bhaduri, S., Mascareno, E., Chaqour, B., and Siddiqui, M. A. (2009). Positive transcription elongation factor b activity in compensatory myocardial hypertrophy is regulated by cardiac lineage protein-1. *Circ. Res.* 104, 1347–1354.

Giacinti, C., Musaro, A., De Falco, G., Jourdan, I., Molinaro, M., Bagella, L., Simone, C., and Giordano, A. (2008). Cdk9-55: a new player in muscle regeneration. *J. Cell. Physiol.* 216, 576–582.

Grandjean, V., Gounon, P., Wagner, N., Martin, L., Wagner, K. D., Bernex, F., Cuzin, F., and Rassoulzadegan, M. (2009). The miR-124-Sox9 paramutation: RNA-mediated epigenetic control of embryonic and adult growth. *Development* 136, 3647–3655.

He, N., Jahchan, N. S., Hong, E., Li, Q., Bayfield, M. A., Maraia, R. J., Luo, K., and Zhou, Q. (2008). A La-related protein modulates 7SK snRNP integrity to suppress P-TEFb-dependent transcriptional elongation and tumorigenesis. *Mol. Cell* 29, 588–599.

He, W. J., Chen, R., Yang, Z., and Zhou, Q. (2006). Regulation of two key nuclear enzymatic activities by the 7SK small nuclear RNA. *Cold Spring Harb. Symp. Quant. Biol.* 71, 301–311.

Huang, V., Qin, Y., Wang, J., Wang, X., Place, R. F., Lin, G., Lue, T. F., and Li, L. C. (2010). RNAa is conserved in mammalian cells. *PLoS ONE* 5, e8848. doi:10.1371/journal.pone.0008848

Kaichi, S., Takaya, T., Morimoto, T., Sunagawa, Y., Kawamura, T., Ono, K., Shimatsu, A., Baba, S., Heike, T., Nakahata, T., and Hasegawa, K. (2011). Cyclin-dependent kinase 9 forms a complex with GATA4 and is involved in the differentiation of mouse ES cells into cardiomyocytes. *J. Cell. Physiol.* 226, 248–254.

Kapranov, P., St Laurent, G., Raz, T., Ozsolak, F., Reynolds, C. P., Sorensen, P. H., Reaman, G., Milos, P., Arceci, R. J., Thompson, J. F., and Triche, T. J. (2010). The majority of total nuclear-encoded non-ribosomal RNA in a human cell is 'dark matter' un-annotated RNA. *BMC Biol.* 8, 149. doi:10.1186/1741-7007-8-149

Kohoutek, J. (2009). P-TEFb- the final frontier. *Cell Div.* 4, 19.

Krueger, B. J., Jeronimo, C., Roy, B. B., Bouchard, A., Barrandon, C., Byers, S. A., Searcey, C. E., Cooper, J. J., Bensaude, O., Cohen, E. A., Coulombe, B., and Price, D. H. (2008). LARP7 is a stable component of the 7SK snRNP while P-TEFb, HEXIM1 and hnRNP A1 are reversibly associated. *Nucleic Acids Res.* 36, 2219–2229.

Malumbres, M., and Barbacid, M. (2005). Mammalian cyclin-dependent kinases. *Trends Biochem. Sci.* 30, 630–641.

Mattick, J. S., Taft, R. J., and Faulkner, G. J. (2010). A global view of genomic information – moving beyond the gene and the master regulator. *Trends Genet.* 26, 21–28.

Morris, K. V. (2009). RNA-directed transcriptional gene silencing and activation in human cells. *Oligonucleotides* 19, 299–306.

Nguyen, V. T., Kiss, T., Michels, A. A., and Bensaude, O. (2001). 7SK small nuclear RNA binds to and inhibits the activity of CDK9/cyclin T complexes. *Nature* 414, 322–325.

Place, R. F., Li, L. C., Pookot, D., Noonan, E. J., and Dahiya, R. (2008). MicroRNA-373 induces expression of genes with complementary promoter sequences. *Proc. Natl. Acad. Sci. U.S.A.* 105, 1608–1613.

Radhakrishnan, S. K., and Gartel, A. L. (2006). CDK9 phosphorylates p53 on serine residues 33, 315 and 392. *Cell Cycle* 5, 519–521.

Rassoulzadegan, M., Grandjean, V., Gounon, P., Vincent, S., Gillot, I., and Cuzin, F. (2006). RNA-mediated non-mendelian inheritance of an epigenetic change in the mouse. *Nature* 441, 469–474.

Romano, G., and Giordano, A. (2008). Role of the cyclin-dependent kinase 9-related pathway in mammalian gene expression and human diseases. *Cell Cycle* 7, 3664–3668.

Sano, M., Abdellatif, M., Oh, H., Xie, M., Bagella, L., Giordano, A., Michael, L. H., Demayo, F. J., and Schneider, M. D. (2002). Activation and function of cyclin T-Cdk9 (positive transcription elongation factor-b) in cardiac muscle-cell hypertrophy. *Nat. Med.* 8, 1310–1317.

Sayed, D., and Abdellatif, M. (2011). MicroRNAs in development and disease. *Physiol. Rev.* 91, 827–887.

Silva, J., Nichols, J., Theunissen, T. W., Guo, G., Van Oosten, A. L., Barrandon, O., Wray, J., Yamanaka, S., Chambers, I., and Smith, A. (2009). Nanog is the gateway to the pluripotent ground state. *Cell* 138, 722–737.

Simon, S. A., and Meyers, B. C. (2011). Small RNA-mediated epigenetic modifications in plants. *Curr. Opin. Plant Biol.* 14, 148–155.

Simone, C., Stiegler, P., Bagella, L., Pucci, B., Bellan, C., De Falco, G., De Luca, A., Guanti, G., Puri, P. L., and Giordano, A. (2002). Activation of MyoD-dependent transcription by cdk9/cyclin T2. *Oncogene* 21, 4137–4148.

Takaya, T., Ono, K., Kawamura, T., Takanabe, R., Kaichi, S., Morimoto, T., Wada, H., Kita, T., Shimatsu, A., and Hasegawa, K. (2009). MicroRNA-1 and MicroRNA-133 in spontaneous myocardial differentiation of mouse embryonic stem cells. *Circ. J.* 73, 1492–1497.

Wagner, K. D., Wagner, N., Ghanbarian, H., Grandjean, V., Gounon, P., Cuzin, F., and Rassoulzadegan, M. (2008). RNA induction and inheritance of epigenetic cardiac hypertrophy in the mouse. *Dev. Cell* 14, 962–969.

Wang, S., and Fischer, P. M. (2008). Cyclin-dependent kinase 9: a key transcriptional regulator and potential drug target in oncology, virology and cardiology. *Trends Pharmacol. Sci.* 29, 302–313.

Warren, L., Manos, P. D., Ahfeldt, T., Loh, Y. H., Li, H., Lau, F., Ebina, W., Mandal, P. K., Smith, Z. D., Meissner, A., Daley, G. Q., Brack, A. S., Collins, J. J., Cowan, C., Schlaeger, T. M., and Rossi, D. J. (2010). Highly efficient reprogramming to pluripotency and directed differentiation of human cells with synthetic modified mRNA. *Cell Stem Cell* 7, 618–630.

Wei, P., Garber, M. E., Fang, S. M., Fischer, W. H., and Jones, K. A. (1998). A novel CDK9-associated C-type cyclin interacts directly with HIV-1 Tat and mediates its high-affinity, loop-specific binding to TAR RNA. *Cell* 92, 451–462.

Werner, A., and Swan, D. (2010). What are natural antisense transcripts good for? *Biochem. Soc. Trans.* 38, 1144–1149.

Xu, C., Lu, Y., Pan, Z., Chu, W., Luo, X., Lin, H., Xiao, J., Shan, H., Wang, Z., and Yang, B. (2007). The muscle-specific microRNAs miR-1 and miR-133 produce opposing effects on apoptosis by targeting HSP60, HSP70 and caspase-9 in cardiomyocytes. *J. Cell. Sci.* 120, 3045–3052.

Yang, Z., Yik, J. H., Chen, R., He, N., Jang, M. K., Ozato, K., and Zhou, Q. (2005). Recruitment of P-TEFb for stimulation of transcriptional elongation by the bromodomain protein Brd4. *Mol. Cell* 19, 535–545.

Yang, Z., Zhu, Q., Luo, K., and Zhou, Q. (2001). The 7SK small nuclear RNA inhibits the CDK9/cyclin T1 kinase to control transcription. *Nature* 414, 317–322.

Yik, J. H., Chen, R., Nishimura, R., Jennings, J. L., Link, A. J., and Zhou, Q. (2003). Inhibition of P-TEFb (CDK9/Cyclin T) kinase and RNA polymerase II transcription by the coordinated actions of HEXIM1 and 7SK snRNA. *Mol. Cell* 12, 971–982.

Zhao, Y., Samal, E., and Srivastava, D. (2005). Serum response factor regulates a muscle-specific microRNA that targets Hand2 during cardiogenesis. *Nature* 436, 214–220.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 04 October 2011; paper pending published: 01 November 2011; accepted: 07 December 2011; published online: 28 December 2011.

Citation: Ghanbarian H, Grandjean V, Cuzin F and Rassoulzadegan M (2011) A network of regulations by small non-coding RNAs: the P-TEFb kinase in development and pathology. Front. Gene. ***2****:95. doi: 10.3389/fgene.2011.00095*

This article was submitted to Frontiers in Non-Coding RNA, a specialty of Frontiers in Genetics.

frontiers in
GENETICS

HYPOTHESIS AND THEORY ARTICLE
published: 26 April 2012
doi: 10.3389/fgene.2012.00057

Dark matter RNA: an intelligent scaffold for the dynamic regulation of the nuclear information landscape

Georges St. Laurent*[1,2] **, Yiannis A. Savva*[3] *and Philipp Kapranov*[2] **

[1] Immunovirology – Biogenesis Group, University of Antioquia, Medellin, Colombia
[2] St. Laurent Institute, Cambridge, MA, USA
[3] Department of Molecular Biology, Cell Biology and Biochemistry, Brown University, Providence, RI, USA

Edited by:
Claes Wahlestedt, University of Miami Miller School of Medicine, USA

Reviewed by:
Peng Jin, Emory University School of Medicine, USA
Antonio Sorrentino, Exiqon A/S, Denmark

****Correspondence:***
Georges St. Laurent and Philipp Kapranov, St. Laurent Institute, One Kendall Square, Suite 200LL, Cambridge, MA 02139, USA.
e-mail: georgest98@yahoo.com; philippk08@gmail.com

Perhaps no other topic in contemporary genomics has inspired such diverse viewpoints as the 95+% of the genome, previously known as "junk DNA," that does not code for proteins. Here, we present a theory in which dark matter RNA plays a role in the generation of a landscape of spatial micro-domains coupled to the information signaling matrix of the nuclear landscape. Within and between these micro-domains, dark matter RNAs additionally function to tether RNA interacting proteins and complexes of many different types, and by doing so, allow for a higher performance of the various processes requiring them at ultra-fast rates. This improves signal to noise characteristics of RNA processing, trafficking, and epigenetic signaling, where competition and differential RNA binding among proteins drives the computational decisions inherent in regulatory events.

Keywords: dark matter RNA, non-coding RNA, vlinc RNA, RNA binding protein, biological signaling, RNA editing, RNA processing, molecular scaffold

INTRODUCTION

The emerging picture of the nucleus portrays a multifaceted environment where RNA processing events occur with accuracy, precision, and high resolution. Since diffusion cannot account for the speed and coordination of the molecular events occurring within its matrix, the nucleus must depend on precisely articulated macromolecular architectures and active transport mechanisms to achieve adequate throughput and signal to noise performance (Lanctot et al., 2007; Misteli, 2007). In addition to the execution of baseline processing of RNA, the extensive network of RNA interaction machineries must respond to incoming physiological signaling, such as stress and cues from the physical environment (McKee and Silver, 2007; Sharma and Lou, 2011), by making rapid but precise changes at decision points, while at the same time maintaining robustness of the overall network. In effect, the entire nuclear space is a finely tuned RNA processing machine, designed to maintain accuracy in the dynamic and reversible regulation of myriads of transcriptome processing events simultaneously. Since the expansion of transcriptome processing increases the computational plasticity (Herbert and Rich, 1999) and the information processing capacity of biological networks (Mattick, 2007; St. Laurent and Wahlestedt, 2007), several authors argue that biological complexity itself has RNA complexity at its core (Licatalosi and Darnell, 2010).

Considering only current knowledge of these networks, and without extrapolating to as yet undiscovered regulatory intricacies, their performance already gracefully exceeds that of systems biology models and mechanisms. Its diversity of specific functions, and the finely tuned regulation of those functions in response to physiological signals, suggests the existence undiscovered mechanisms and network design principles at work to maintain robustness of the RNA output of a cell. In fact, recent studies of disease mechanisms suggest that humans can tolerate little loss of signal to noise performance in the nucleus. Healthy physiological function depends on the precision, reliability, and accuracy of the nuclear RNA processing machine, as processing errors in RNA molecules often lead to serious diseases (Garcia-Blanco et al., 2004; Cooper et al., 2009; Venables et al., 2009; Licatalosi and Darnell, 2010; Ward and Cooper, 2010; Jia et al., 2012).

It is in this context that we would like to consider the genomic "dark matter," one of the major mysteries of the post-genome era. Perhaps no other topic in contemporary genomics has inspired such diverse viewpoints as the 95+% of the genome, previously known as "junk DNA," that does not code for proteins. Reports of pervasive transcription of these vast "dark matter" regions, combined with frequent identification of families of long or very non-coding RNAs (lncRNAs) originating from them, have opened new chapters of both discovery as well as controversy. The observation that the percentage of "dark matter" genomic sequence correlates monotonically with organismal complexity, for every species sequenced to date (Taft et al., 2007), has inspired theories proposing a central role for these regions in the information processing of complex organisms (Mattick, 2007; St. Laurent and Wahlestedt, 2007). Yet, while an increasing number of specific interactions between lncRNAs and other biological molecules have demonstrated functions for a number of dark matter transcripts (Wang and Chang, 2011), a global concept of function has not yet emerged. In effect, the original reports of pervasive transcription (Kapranov et al., 2002, 2007b; Carninci et al., 2005, 2008; Katayama et al., 2005) of the mammalian genome have faded somewhat, with focus instead on separately developed lists of lncRNAs detected in specific experiments or filtered by certain properties that hint at functionality (Willingham et al., 2005; Guttman et al., 2009; Khalil et al., 2009; Wai et al., 2010; Askarian-Amiri et al., 2011; Khaitan et al., 2011). These lists of lncRNAs usually only cover a few percent of the genome, representing only a small fraction of the original pervasiveness of dark matter and typically, sample intergenic space as introns of known genes are usually assumed to represent pre-mRNAs. For example, our recent work has shown the presence of numerous very long transcribed regions of intergenic genomic space not currently covered by the lincRNA annotations (Kapranov et al., 2010) and has shown that introns of mouse genes produce stable RNAs regulated separately from the mature protein-coding RNAs (St. Laurent et al., submitted). Partly due to this uncertainty, some authors have cast doubt on the importance of dark matter transcripts, labeling them transcriptional noise (Brosius, 2005; Struhl, 2007; van Bakel and Hughes, 2009; Robinson, 2010; van Bakel et al., 2010), or even arguing that they largely represent "fragments of known pre-mRNAs" (van Bakel et al., 2010). Even the existence of much of the dark matter RNA implied by the early reports of pervasive transcription has stirred recent controversy (van Bakel et al., 2010). On balance, a common view in the field holds that while there is a collection of lncRNAs with specific interactions and functions, they exist among a larger collection of dark matter transcriptional noise.

As the controversy surrounding the function of "dark matter" RNA continues, a number of recent studies provided more comprehensive datasets, through the implementation of improved methodologies to confirm its existence and, more importantly, to measure its relative mass. A recent investigation designed to capture and measure non-exonic signals, revealed surprisingly that dark matter RNAs actually comprise a majority of non-ribosomal non-mitochondrial RNAs in human cells (Kapranov et al., 2010). We also know that the nucleus is rich in dark matter RNA (Cheng et al., 2005). Since the majority of protein-coding RNAs reside in the cytoplasm, the fraction of dark matter RNA is likely to be many folds higher in the nucleus than that of protein-coding RNAs.

Considering the vital importance of maintaining the performance of nuclear processing of all types, the nuclear molecular machineries would not tolerate the accumulation of large amounts of non-functional RNA molecules. Any significant population of such molecules would at best represent a large input of noise into the fine-tuned computational machinery of nuclear processing, not likely to benefit the performance of the nucleus or the cell as a unit. In practical terms, if dark matter had no biological function, the high performance and signal to noise ratios of the nuclear RNA processing machineries would logically conflict with the high levels of dark matter now documented in human cells (Kapranov et al., 2010). In other words, the currently emerging picture of the nucleus contains a paradox: a nuclear micro-environment simultaneously populated by high concentrations of precision RNA processing machineries, and by an astonishing level of noise from dark matter RNA. How can the nucleus precisely regulate such highly accurate processing events in tens of thousands of transcripts simultaneously, while ignoring the massive amount of inherent noise from dark matter RNA existing in the same nuclear space?

To resolve this apparent paradox, and to provide a mechanism for global function of dark matter RNA, in this article we present a theory in which dark matter RNA plays a role in the generation of a landscape of spatial micro-domains coupled to the information signaling matrix of the nuclear landscape. Within and between these micro-domains, dark matter RNAs additionally function to tether RNA interacting proteins and complexes of many different types, and by doing so, allow for a higher performance of the various processes requiring them at ultra-fast rates. This improves signal to noise characteristics of RNA processing, trafficking, and epigenetic signaling, where competition and differential RNA binding among proteins drives the computational decisions inherent in regulatory events.

THE SYSTEM WIDE PERFORMANCE CHARACTERISTICS OF NUCLEAR RNA PROCESSING MACHINERIES

It is estimated that an average human cells contains 300,000 mRNAs (Hastie and Bishop, 1976), each containing on average 10 exons, a start site, and a poly A+ tail. Thus, every such average molecule had to go through at least 18 splicing reactions (selection of splice donor and acceptor sites) plus selections of the start site and the polyadenylation site. In total, a minimum of 6M processing events had to occur to generate this diversity. This does not take into account (i) all subsequent base modification such as RNA editing, N6-methyladenosine, 5′-cap, (ii) subsequent cleavage events, or (iii) transportation of these RNA molecules to their sites of function or into well demarcated nuclear storage for later use. Nor does it account for the polyA− RNA population that exceeds that of the polyA+ by several folds. Also, if one were to include the ribosomal RNA, that represents ~95% of all cellular RNA (Raz et al., 2011), which is also processed and modified, then the minimal order of the number of cellular processing events needed to accommodate the real complexity of RNAs within a single nucleus is likely to be in the tens of millions.

As a vital step in transcript processing, RNA editing offers further insight into the high level of orchestration of nuclear RNA processing machineries. Adenosine deaminase acting on RNA (ADAR) mediates adenosine to inosine (A-to-I) RNA editing in dsRNA molecules, which often results in distinct downstream physiological outcomes for the edited RNAs. ADAR RNA editing frequently targets coding regions of mRNAs that encode ion channels and other components of the synaptic release machinery (Hoopengardner et al., 2003; Seeburg and Hartner, 2003). Intronic non-coding sequences with extensive complementarity to upstream or downstream exons containing the adenosine destined to be edited can form simple exon–intron hairpin structures (Higuchi et al., 1993; Burns et al., 1997; Hanrahan et al., 2000; Wang et al., 2000) or more complex RNA secondary structures such as a pseudoknot (Reenan, 2005). RNA editing in mRNAs often generates protein products that are not encoded by the literal genomic information, since upon translation the ribosomal machinery interprets inosines as guanosines (Basillo et al., 1962) resulting in amino acid substitutions. Various studies in different genetic model organisms suggest that RNA editing of mRNAs can result in profound changes in protein function (Rosenthal and Bezanilla, 2002; Bhalla et al., 2004; Ingleby et al., 2009).

Execution of this type of modification requires great deal of precision from the RNA processing machinery in terms of identification of RNA molecules to be edited, sites of editing within these molecules and also in the degree of editing at any given site. Editing could be separated into "pinpoint" and "prolific." The former one results in editing at specific sites in specific RNA molecules. In *Drosophila* for example, the nervous system editing sites generally demonstrate a high level of conservation across 12 fly genomes, representing 85 million years of evolutionary divergence (Hoopengardner et al., 2003). This high level of conservation includes sites that code for levels of transcript editing in the adult fly as low as a few percent, demonstrating physiological sensitivity for this form of transcript processing. In addition, some RNAs that form extensive dsRNA structures, such as non-coding transcripts, sense–antisense RNAs bound to each other, and exogenous RNAs can serve as ADAR substrates destined for prolific editing (Bass, 2002; Nishikura, 2006), resulting in up to 50% A-to-I conversions (Nishikura et al., 1991; Polson and Bass, 1994). Choice of such substrates is also controlled as not every RNA molecule that can form dsRNA will be edited and not every adenosine in molecules that are substrates for ADAR is edited. The fate of such inosine-rich RNA molecules is different from the ones subject to "pinpoint" editing. They can in fact have at least two fates: retention within the nuclear compartment through dependent localization by p54nrb/Vigilin (Zhang and Carmichael, 2001; Wang et al., 2005) and cytoplasmic degradation by Tudor-SN (Scadden, 2005).

Furthermore, the ADAR information processing pathway is sensitive to environmental stimuli in addition to stress responses. Editing analysis of K+ channel mRNAs between Arctic and tropical octopus species revealed substantial differences in editing levels, which are mediated by temperature variations (Garrett and Rosenthal, 2012). In humans, the three ADAR genes can undergo alternative splicing to produce over a dozen isoforms with heterogeneous RNA target specificities. The inflammatory cascade results in a dramatic induction of many of these ADAR isoforms, resulting in a widespread increase of edited RNAs during mammalian inflammation (Yang et al., 2003a,b). Since intronic sequences form dsRNAs with coding regions to serve as ADAR substrates, editing must precede splicing. During these circumstances a regulatory mechanism must exist to ensure an accurate coordination of an extensive network of RNA processing machines to operate with high fidelity to generate dynamic responses upon internal and external stimuli.

In addition to the plethora of transcript variation discussed above, the RNAs produced subsequently traffic into predetermined subcellular localizations. Many transcripts interact with sets of trafficking proteins to migrate to specific nuclear locations such as interchromatin granules (ICGs) or speckles (Spector and Lamond, 2011), for further processing in response to transient physiological signals. Transcripts can also undergo complex cleavage events, followed by 5′ capping, in response to little understood signals and circumstances (Affymetrix/CSHL ENCODE Project, 2009; Mercer et al., 2010). CTN RNA represents an intriguing example where both of these mechanisms are combined. Within minutes of amino acid deprivation or similar cellular stress, signals transduced into the nucleus result in cleavage of the sequestered CTN RNA, and the release and transport to the cytoplasmic translation machinery of the amino acid transporter for which the cleaved RNA product codes (Prasanth et al., 2005). In some cases, cleavage events themselves produce small RNAs, whose activities feedback into splicing decisions, as in the example of the HBII52 snoRNA, which is cleaved from intronic RNA templates in the SNURF-SNRPN locus, and interacts with the serotonin 2C mRNA to regulate its alternative splicing (Kishore and Stamm, 2006).

Considering all of these regulatory layers together, millions of RNA processing events have to happen with accuracy and precision to generate the complexity of RNA present in a cell at any given moment. Many of these events require computation-like decision making as multiple alternative outcomes are available to a cell. In some cases, a single locus can produce hundreds, or even thousands of alternative products of RNA processing. The *Drosophila* Dscam locus for example, can produce 37,000 distinct isoforms from one "gene" (Wojtowicz et al., 2004). High throughput studies using RNA-seq revealed that 94% of human genes undergo alternative splicing in some tissue (Wang et al., 2008). In light of this output volume, such widespread reliance on alternative splicing points to the magnitude of the regulatory challenge facing nuclear splicing machineries. Since the RNA signals that code for splicing events contain relatively low sequence complexity, and frequently diverge from consensus sequences (Egecioglu and Chanfreau, 2011), they provide only modest energetic and informatic vectors to support the accuracy and reliability of high volume splicing output. As a result, achieving a correct splicing decision at a given site usually depends on a precise sequence of combinatorial events, composed of multiple protein and RNA elements, and even chromatin adaptor systems (Luco et al., 2011), acting both in competition (Witten and Ule, 2011), and in cooperation (Hertel, 2008; Xiao and Lee, 2010).

Performance related challenges would face any system designed to produce such a wide array of molecular outputs. Yet, even with these challenges, the systems performance of nuclear RNA processing appears to be surprisingly high. A recent investigation of cell-to-cell variability of alternative splicing determined that non-transformed cells maintained very low splicing isoform variability between individual cells, and concluded that mammalian cells minimize fluctuations in mRNA isoform ratios by tightly regulating the splicing machinery (Waks et al., 2011). Evidence increasingly supports precise and finely tuned regulation of transcriptome processing events as a rule in the nucleus. For example, a growing number of reports describe links between perturbations in splicing (Cooper et al., 2009; Ward and Cooper, 2010), or transcript localization (Faghihi et al., 2008) and diseases. This trend underlines the importance of high precision and accuracy in RNA based machineries under healthy physiological conditions.

Thus, using as yet little known organizational principles, nuclear RNA processing machineries not only produce processed and modified RNAs with high efficiency and accuracy, but implement a large scale integration of dynamic physiological signals, which then drive precise regulatory control and plasticity in response to a myriad of signaling events. From this perspective, the nuclear RNA processing machineries, and the dynamic structural environment surrounding them, must orchestrate their tasks with high accuracy and precision. They must recognize and distinguish processing motifs in RNA structural signals with high sensitivity and yet reject sub-optimal motifs with a high selectivity. Furthermore, the catalytic processes involved in the processing steps must occur with little or no errors. Once formed, the products must enter downstream trafficking pathways, and RNA whose presence is no longer required must be rapidly degraded to avoid introducing noise into earlier steps in the processing pathways due to RNA-waste accumulation. Finally, the entire multilayer system must maintain sufficient plasticity to quickly respond to thousands of potential information signals from outside the nucleus to generate appropriate alterations in processing steps at determined loci in response to changing physiological signals.

While the mystery of how all of this occurs within the space of the nucleus remains unresolved, it depends on the choreography of combinatorial interactions between transcripts and hundreds of RNA binding proteins (RBPs). RBPs interact with complex combinations of primary and secondary structure signals in RNA, and function in both cooperative and competitive types of architectures, as documented in splicing regulation (Darnell, 2006; Sharma and Black, 2006; Ule and Darnell, 2007; Licatalosi et al., 2008; Hallegger et al., 2010), and more recently in chromatin signaling (Tsai et al., 2010; Zhao et al., 2010). The first RBPs to interact with a given nascent transcript can influence the subsequent folding steps of that RNA, and thereby change the downstream distribution of protein interactions for that RNA. The process of differential recognition of nascent RNA information signals by the correct RBP must occur at a pace complementary to that of transcription as well as subsequent processing or chromatin signaling. To maintain plasticity for the accurate transduction of environmental signals, the RNA–protein interaction landscape must somehow achieve an extraordinary coupling between computational, catalytic, and structural elements.

EMERGING FEATURES UNDERLYING THE HIGH PERFORMANCE OF NUCLEAR RNA PROCESSING MACHINERIES

As investigations continue to reveal the depth and performance of nuclear RNA processing functions, the challenge for systems biologists grows more daunting. Current systems biology modeling cannot account for the precision, accuracy, or signal to noise ratios achieved by RNA processing machineries. Nevertheless, the transcriptome–proteome interface in the nucleus contains a vast store of dynamical information. To consistently make effective use of this information, the nuclear systems network architecture must have a number of key design features, including maintenance of reversibility, temporal coherence (the timing and velocity of information processing between network layers), and the ability to resolve logical conflicts over the spatial extent of the networks that comprise the system. To help explain how nuclear RNA processing networks harness the power of that information, a number of concepts have emerged.

DYNAMIC SCAFFOLDING MAXIMIZES INFORMATION FLOW

Biological molecules in the nuclear space exist in a constrained environment where diffusion occurs relatively slowly (Albert et al., 2012). Thus, biochemical kinetics can represent hurdle for adequate performance of complex multistep processing pathways, as the entire interdependent system must minimize bottlenecks and flow imbalances. In order to overcome these physical limitations, RNA processing machineries must rely on highly articulated spatial domains, where local environments transduce information efficiently. The concept of global scaffolding can create these performance enhancing interaction topologies. For example, an RNA scaffold can increase the local concentration of an RBP, such as Nova 1, and a corresponding increase in the signal to noise performance of Nova's influence on splice site selection within that particular spatial domain (**Figure 1A**). Increasing the local concentration of these factors also permits improved interaction kinetics with less ΔG, making the interactions more reversible (**Figure 1B** and also below).

Structural features in the nucleus that enhance transcriptional control and RNA processing contain a large amount of spatially and temporally coded information content. The nucleolus for example appears to depend on RNA secondary structure signals for its effective formation, as the absence of these RNA secondary structures resulted in complete disarray of the nucleolus (Peng and Karpen, 2007). The reversibility of such events in nuclear architecture means that elements responsible for their formation also encode sufficient information to detect external signals and respond with disassembly and transport of components to other spatial domains or downstream processing pathways (Spector and Lamond, 2011).

COMPETITION AND COMPUTATION AT THE TRANSCRIPTOME–PROTEOME INTERFACE

With their unique combination of primary, secondary, and tertiary structure, RNA offers a multiplicity of ways to code for biological information. At the core of this system is a language of RNA–protein, and RNA–RNA/DNA recognition implemented by RNA's unique ability to couple analog and digital signals (St. Laurent and Wahlestedt, 2007). The efficient transduction of that information often depends on its timely recognition by the appropriate RBPs present in the immediate vicinity of an elongating primary transcript. As the transcript emerges from Pol II, it begins to fold. That folding is also influenced by the RBPs that are supposed to interact with it. They influence which of many folding paths that the RNA can take. If the correct RBPs are not right there to quickly associate with the RNA, then the RNA could take another folding path, which would in turn lead to a different set of downstream events, as in the case of Nova splicing proteins and their influence on upstream or downstream splice site choice. So the presence or absence of a given distribution of RBPs in the vicinity of a nascent RNA chain will influence a series of "memory states" that then modulate other processing events downstream. With such a large space of potential RNA–protein interactions, and the requirement for dynamic reversibility of many of their associated signaling events, the system faces a major challenge to achieve an adequate signal to noise ratio for effective function.

Active competition for recognition site on nascent RNA signals directly addresses these problems. Splicing regulation makes abundant use of competing RBPs to enhance the sensitivity, specificity,

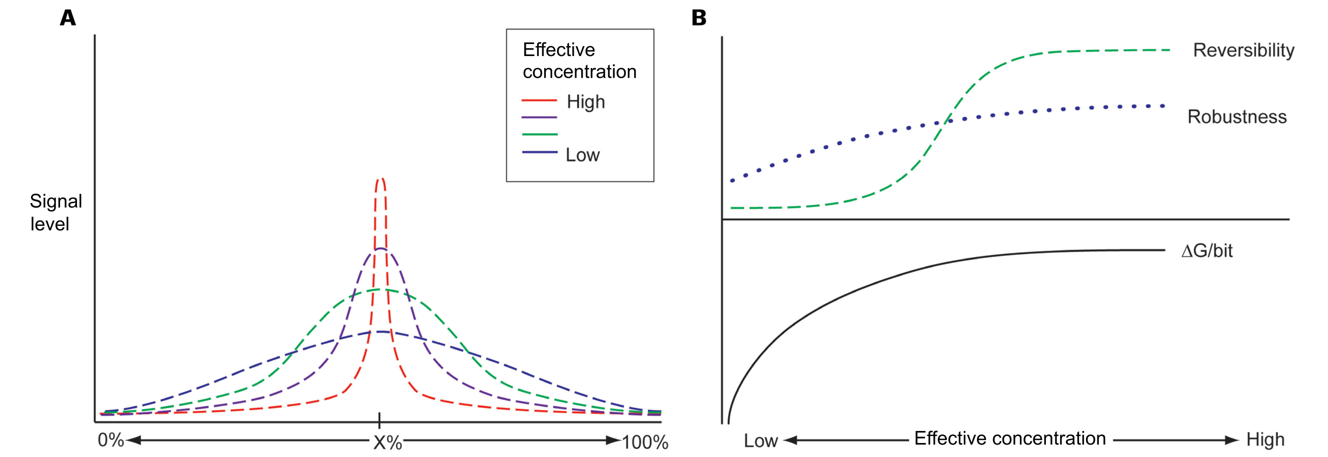


FIGURE 1 | (A) Information currently available about the regions of dark matter transcription and the actual RNA molecules made from these region comes from various types of experiments and databases. There is relatively little overlap between these different databases suggesting that the actual extend of dark matter transcription is far greater than any one database suggests. **(B)** A theoretical curve showing expected results of the fraction of the genome that is transcribed as a function of the number of biological sources whose RNA is profiled. The coverage of transcribed genome by protein coding genes including their introns is 42% and lincRNAs bring it up to 58%. However, the full extent of the transcribed genome is expected to be much greater than that.

and regulatory control of splice site decision commitment (Ule and Darnell, 2006; Chen and Manley, 2009). Examples include PTB protein which antagonizes Nova (Polydorides et al., 2000) at overlapping recognition sites, establishing a sensitive switch between two splicing choices. Interesting examples from spliceosome quality control also demonstrate the importance of reversibility, such as involvement of ATPase Prp16p in both forward and discard splicing pathways (Koodathingal et al., 2010).

Competition may also drive accurate computation in the small RNA regulatory pathways, with duplex regions competing for recognition by ADAR vs Drosha/Dicer, with contrasting outcomes depending on which protein prevails (Nishikura, 2006). Similarly, many epigenetic signaling events may be mediated by competitive interactions between lncRNAs and protein components of signaling machineries (Lee, 2011). All of these regulatory mechanisms require effective concentrations of interacting proteins to achieve adequate signal to noise ratios. The Lin28–let7 miRNA interaction provides an interesting example of specificity that would be difficult to achieve with low protein concentrations (Nam et al., 2011; Piskounova et al., 2011).

REVERSIBILITY AND FEEDBACK LOOPS

Erasure of information presents a challenge for any complex system (Lloyd, 2001). In biological systems, thermodynamic constraints make the cost of information innately high, and yet its value can oscillate from vital to worthless or even harmful in seconds once the message or a signal encoded in it is transduced. The dynamics of this "volatile market" reality make erasure of biological information a high priority in any system, but especially in the nucleus where many network pathways converge. While DNA retains the permanent information, a large majority of the dynamical information exists within the transcriptome, as combinatorial accumulations of RNA–protein and RNA–RNA/DNA interactions.

Not surprisingly, reversibility is a key feature of information coding at the transcriptome–proteome interface. The conformational flexibility of RNAs, especially ncRNAs whose secondary structures are not constrained by coding regions, and the dynamic changes in their structure that can occur in response to protein binding and environmental signals provide not only increased symbolic information density, but contribute to the reversibility of RNA–protein interactions. Proteins that bind RNA also tend to contain natively unstructured regions. This could be the basis for structural articulation (i.e., the incorporation of information containing motifs and elements into nuclear scaffolding structures) that improves precise temporal and spatial choreography of RNA processing machineries. For example, interactions between RNA and their cognate proteins often involve natively unstructured regions in the protein, and similarly flexible structures in the RNA (Leulliot and Varani, 2001). These regions of evolutionarily coded local disorder contribute useful properties for information processing. Precisely orienting them within an articulated regional structure increases their sensitivity, specificity, and reversibility, thereby contributing directly to the throughput and precision of nuclear machineries. When these regions form a stepwise interaction with their RNA target, the entropy of the complex is decreased, thereby producing an "entropic spring" effect, which enhances reversibility when the interaction is no longer required (Tompa and Csermely, 2004). Together with reversibility of individual interactions within RNA–protein interaction networks, frequent feedback loops support the reversibility of these networks. These features operate cooperatively to facilitate the timely erasure of information, and the finely tuned response of RNA processing machineries to changes in signaling and various environmental conditions.

Thus, a central part of our argument maintains that the performance and throughput of nuclear RNA processing machineries requires the functional coupling of well-articulated spatial and temporal landscapes in order to maximize the flow of biological information through the components of RNA processing networks.

THE DARK MATTER INTELLIGENT SCAFFOLD

IMPLICATIONS OF THE PREPONDERANCE OF CELLULAR DARK MATTER RNA IN MAMMALIAN CELLS

Several years ago, John Mattick presented the concept of the nucleus as an "RNA machine" (Amaral et al., 2008), arguing that much of the information processing in the nucleus occurs through RNA intermediates, and that ncRNA overcomes the prohibitive regulatory overhead associated with saturated protein–protein regulatory interactions in this environment (Gagen and Mattick, 2005). From the point of view of information theory, this implies that the RNA content of the nucleus functions in a manner roughly equivalent to an information channel, and that the "channel capacity" (information throughput) of this system depends on the available degrees of freedom of the combined population of RNA molecules contained in the system. Consequently, RNA quality control and degradation machinery must actively pursue the elimination of non-functional RNAs that would represent noise to the "RNA machine." Yet, with the recent discovery that dark matter RNA makes up the majority of cellular RNA by mass (Kapranov et al., 2010; and an even greater majority in the nucleus), it appears that this enigmatic class of RNA does not represent noise in the RNA based information channel of the nucleus, and instead likely comprises an integral part of the information channel itself.

The information containing structural features of dark matter RNAs and their ability to interact with the nuclear proteome appear similar to their coding counterparts. If the primary sequence patterns and secondary structure motifs that determine protein interactions occur with similar densities in both classes of RNAs, then they must both exist in the nucleus in complex with proteins. If dark matter RNAs represented noise or spurious transcription, their predominant mass would compromise signal to noise ratio performance of RNA processing machineries in the entire nucleus, as they depend on the information derived from such interactions. Instead, its high concentration suggests that dark matter RNA functions at the core of the multilayer nuclear "RNA machine" (Amaral et al., 2008).

DARK MATTER RNA ESTABLISHES A DYNAMIC AND REVERSIBLE MICRO-PARTITIONING OF NUCLEAR SPACE

The large amount of dark matter RNA in the nucleus, establishes the basis for the "intelligent scaffold" concept. Each dark matter RNA acts either in *cis* or in *trans*, depending on its own information content (complex combinations of primary sequence motifs and secondary structures), and the proteins with which it interacts. Long dark matter RNAs can form several types of interactions with DNA, and other RNAs, inside spatial domains of chromatin. These can involve direct interactions between RNA and DNA, similar to what occurs between pRNAs transcribed from regions in between rRNA genes, and the T0 element in rRNA promoters (Mayer et al., 2006; Schmitz et al., 2010). Proteins can also mediate the interactions, as recently demonstrated for XIST and transcription factor YY1 (Jeon and Lee, 2011). Alternatively, proteins or RNAs can use co-transcriptional targeting where the transcript is tethered during transcription by RNA polymerase, similar to the mechanism of the TAR RNA targeting by HIV TAT protein (Brady and Kashanchi, 2005). Transcriptional targeting may also occur with the short ncRNAs transcribed from the 5′ ends of many human genes (Wei et al., 1998; Kapranov et al., 2007a; Kanhere and Jenner, 2012). Dark matter RNAs have all three of these mechanisms available to mediate their interactions with DNA and other RNAs, providing a large combinatorial basis for the formation of flexible complexes that drive spatial and computational integration.

As these RNAs accumulate into spatial micro-domains surrounding one or more genomic loci, they establish a region of nuclear space under their influence, which in turn attracts a variety of molecules. The RNAs can interact with many proteins, and other large and small RNAs, often with relatively low affinities, which results in a temporally and spatially distributed macromolecular landscape around that locus (see **Figure 2**). Since these molecules function primarily in transcriptome processing and epigenetic regulation, the dark matter guided landscapes would facilitate the structural and computational operations of both systems, as well as catalyze crosstalk between them. In this manner dark matter RNAs can effectively establish finely tuned concentration gradients of epigenetic signaling and RNA processing proteins (and small RNAs) for efficient operation of these systems. An intriguing example of this has recently been described as a "molecular cage" for PRC1 complexes. The "molecular cage" apparently uses a combination of methylated H3K27 moieties and low affinity binding sites on nascent lncRNAs to increase the local concentration of PRC1 for chromatin signaling (Beisel and Paro, 2011).

The intelligent scaffold mechanism facilitates the accumulation of higher concentrations of RBPs (and small RNAs) within chromatin regions, as well as the micro-partitioning of these regions at an optional resolution for RNA processing, epigenetic signaling, and transcript expression regulation. Macromolecules within these micro-domains can disassociate from their low affinity binding sites in these dark matter rich micro-regions, as they find higher affinity sites in nascent strands emerging from RNA Pol II transcription. Abundant sites of alternative localization in dark matter equates with more effective differential recognition of RNA motifs by competing RBPs, and increased reversibility of signal transduction in regulatory events. The key here is that signal to noise ratio is not driven only by the size of ΔG, but by the ratio of ΔG "protein A" to ΔG "protein B" or the ratio of ΔG site1 of protein A on the "target" nascent strand RNA molecule to ΔG site2 of the same protein A on the "repository" dark matter RNA molecule. This is shown as "Biological Information Content" on **Figure 2**. If both ΔGs are large compared to their difference, then the signal is low and the noise is high. A recent experiment that used RNAi knockdown to reduce the expression levels of splicing regulator SRSF1 confirmed the importance of high concentrations of RNA processing proteins to maintain adequate signal to noise ratios. Lowered concentrations of SRSF1 markedly increased the variance of splicing isoform ratios of the target transcript, measured in populations of single cells (Waks et al., 2011).

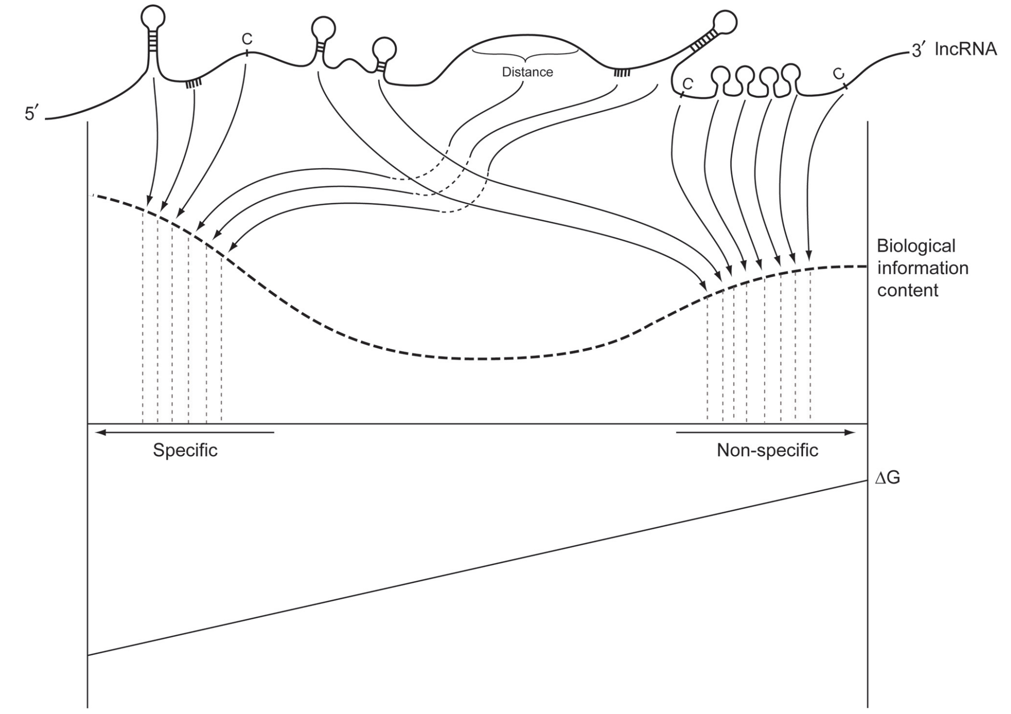


FIGURE 2 | The information content of a hypothetical dark matter RNA. Combinations of primary sequence and secondary structure form high affinity interaction sites with high information content (left side). These specific interactions have a large ΔG. At the other end of the spectrum, the same RNA can have a large number of relatively non-specific interaction sites that nevertheless have biological information content and functional significance. Their absence would result in subtle loss of signal to noise characteristics across many affected pathways that could be unrelated to the RNA.

Adjacent micro-partitions could favor higher concentrations of some proteins over others, due to the heterogeneous distributions of low affinity binding sites along the lengths of dark matter RNA molecules in each micro-partition. The result, depicted in **Figure 3**, shows varying levels of sequestration of RNA processing components, depending on the systems performance requirements of each component. Overall, higher concentrations of effector components equate with faster kinetics and more finely tunable regulation, which in turn improve signal to noise ratios and system performance.

The temporal and spatial dynamics of intelligent scaffolds permit integration of signals from many levels of biological information processing. Changes in the intelligent scaffolding environment of a three-dimensional chromatin micro-region can impact the dynamics of transcriptional folding, processing, localization, and degradation of transcripts as well as chromatin signaling (see **Figure 4**). For example, dark matter cleavage events can quickly change the structure of the micro-domain by sweeping away large numbers of proteins, RNAs, and scaffold, and at the same time generate small RNAs, or expose regions of RNA complementarity to small RNAs, as described in the recent theory of competing endogenous RNAs (ceRNAs) by the Pandolfi group (Salmena et al., 2012). Cleavage of very long dark matter RNAs, for example those coming from the vlinc regions (Kapranov et al., 2010), could occur even with their RBPs still attached. Cleaved RNAs could then function as lncRNAs. Small RNAs could also interact with sites in tethered vlincs, thereby acting as a sink, or by blocking sites that would otherwise be occupied by other signaling molecules. Under some circumstances, combinations of these other events could serve as the signal to trigger cleavage of the vlincs, which could then form a rapid feed-forward circuit as the cascade of cleavage continues in the entire micro-domain.

CONCLUSION: THE FOREST ENRICHES THE FUNCTIONALITY OF THE TREES

While specific interactions drive the bulk of molecular information processing in biological systems, in the RNA based regulatory networks of the nucleus the performance characteristics of specific interactions are determined by the surrounding micro-environment. The dark matter RNA plays a key role in implementing the dynamic responsiveness of that surrounding micro-environment. Considering its importance, the concept of functions for dark matter RNAs should embrace a continuum, from those that arise from highly specific interactions, to those at the other end of the spectrum that involve lower affinity and less specificity, but nevertheless contribute to the synergistic attributes of the surrounding micro-environment. Those attributes permit the specific interactions, and facilitate their coordination and integration.

Evaluating dark matter RNAs in this fashion provides a context and explanation for the relatively low level of conservation of these RNAs, as many informational elements either do not require conservation, or require only functional conservation. As demonstrated for a growing list of ncRNAs, functionality does not require

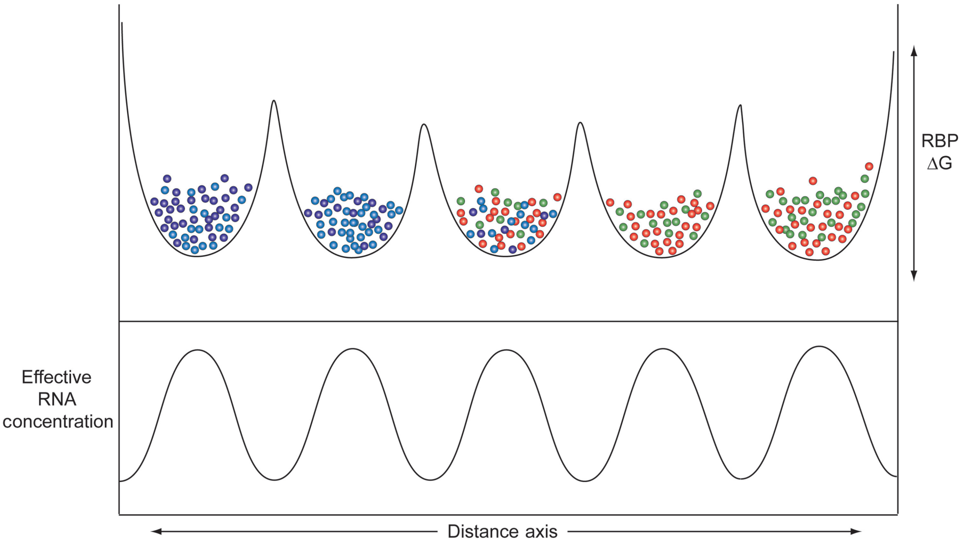


FIGURE 3 | Distribution of different domains with varying ΔG affinities (*Y*-axis) for different RBPs (balls of different colors) along the length of a hypothetical series of micro-domains each composed of different (very) long non-coding RNAs (*X*-axis). The heterogeneous landscape of dark matter RNAs creates pockets of enrichment of these different RBPs.

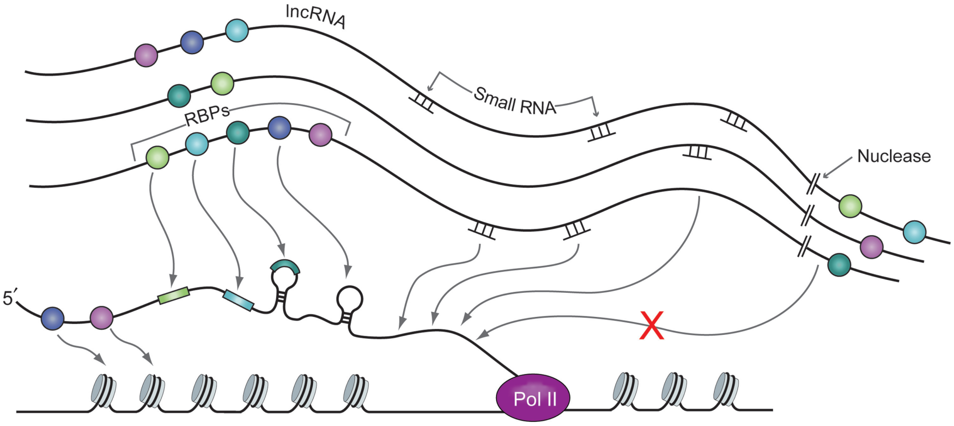


FIGURE 4 | A nascent RNA strand being synthesized by RNA Pol II interacts with RBPs (balls of various colors), small RNAs and lincRNAs. Interaction between all these molecules is made possible by close proximity of these molecules in the nuclear micro-domains. RBPs and small RNAs are bound to (v)lincRNAs with relatively low specificity and the latter present them in exact architectural and temporal environment to the nascent strand that possess specific motifs for the former. Nuclease action cleaves the non-coding RNA template and thus changes the structure of the scaffold complex. This in turn can change the kinetics of the interaction between the RBPs and the small RNAs bound to it and the sequence of their presentation and interaction with the nascent strand of RNA.

conservation, at least not in the same way that is known to occur for protein-coding sequences (Pang et al., 2006). The theory predicts increasing concentrations of dark matter complexed with RNA interacting proteins in complex organisms, and helps explain the direct correlation of organismal complexity with the genomic percentage of non-coding regions in all genomes sequenced to date (Taft et al., 2007). It also suggests expansion of regions of RNA interacting regions in proteomes of organisms as evolutionary complexity increases.

The dark matter intelligent scaffold concept focuses on the level of coupling between computation and spatial articulation. The theory holds that large increases in biological complexity required ever increasing levels of coupling between computation and structure, as a key driver of that complexity, and ultimately a measure of organismal fitness. Dark matter RNA was recruited to perform this function, to dynamically bridge these two ostensibly orthogonal dimensions, because its flexible structural and computation features endow it with special qualities to serve as a molecular intermediate in the coding, processing, and distribution of information.

ACKNOWLEDGMENTS

We wish to thank Erica Dumais for expert artistic assistance with figure preparation and Juan Pablo Lizarazo for logistical support.

REFERENCES

Affymetrix/CSHL ENCODE Project. (2009). Post-transcriptional processing generates a diversity of 5′-modified long and short RNAs. *Nature* 457, 1028–1032.

Albert, B., Leger-Silvestre, I., Normand, C., and Gadal, O. (2012). Nuclear organization and chromatin dynamics in yeast: biophysical models or biologically driven interactions? *Biochim. Biophys. Acta.* doi: 10.1016/j.bbagrm.2011.12.010 [Epub ahead of print].

Amaral, P. P., Dinger, M. E., Mercer, T. R., and Mattick, J. S. (2008). The eukaryotic genome as an RNA machine. *Science* 319, 1787–1789.

Askarian-Amiri, M. E., Crawford, J., French, J. D., Smart, C. E., Smith, M. A., Clark, M. B., Ru, K., Mercer, T. R., Thompson, E. R., Lakhani, S. R., Vargas, A. C., Campbell, I. G., Brown, M. A., Dinger, M. E., and Mattick, J. S. (2011). SNORD-host RNA Zfas1 is a regulator of mammary development and a potential marker for breast cancer. *RNA* 17, 878–891.

Basillo, C., Wahba, A., Lengyel, P., Speyer, J., and Ochoa, S. (1962). Synthetic polynucleotides and the amino acid code, V. *Proc. Natl. Acad. Sci. U.S.A.* 48, 613–616.

Bass, B. L. (2002). RNA editing by adenosine deaminases that act on RNA. *Annu. Rev. Biochem.* 71, 817–846.

Beisel, C., and Paro, R. (2011). Silencing chromatin: comparing modes and mechanisms. *Nat. Rev. Genet.* 12, 123–135.

Bhalla, T., Rosenthal, J. J., Holmgren, M., and Reenan, R. (2004). Control of human potassium channel inactivation by editing of a small mRNA hairpin. *Nat. Struct. Mol. Biol.* 11, 950–956.

Brady, J., and Kashanchi, F. (2005). Tat gets the "green" light on transcription initiation. *Retrovirology* 2, 69.

Brosius, J. (2005). Waste not, want not – transcript excess in multicellular eukaryotes. *Trends Genet.* 21, 287–288.

Burns, C. M., Chu, H., Rueter, S. M., Hutchinson, L. K., Canton, H., Sanders-Bush, E., and Emeson, R. B. (1997). Regulation of serotonin-2C receptor G-protein coupling by RNA editing. *Nature* 387, 303–308.

Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M. C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C., Kodzius, R., Shimokawa, K., Bajic, V. B., Brenner, S. E., Batalov, S., Forrest, A. R., Zavolan, M., Davis, M. J., Wilming, L. G., Aidinis, V., Allen, J. E., Ambesi-Impiombato, A., Apweiler, R., Aturaliya, R. N., Bailey, T. L., Bansal, M., Baxter, L., Beisel, K. W., Bersano, T., Bono, H., Chalk, A. M., Chiu, K. P., Choudhary, V., Christoffels, A., Clutterbuck, D. R., Crowe, M. L., Dalla, E., Dalrymple, B. P., De Bono, B., Della Gatta, G., Di Bernardo, D., Down, T., Engstrom, P., Fagiolini, M., Faulkner, G., Fletcher, C. F., Fukushima, T., Furuno, M., Futaki, S., Gariboldi, M., Georgii-Hemming, P., Gingeras, T. R., Gojobori, T., Green, R. E., Gustincich, S., Harbers, M., Hayashi, Y., Hensch, T. K., Hirokawa, N., Hill, D., Huminiecki, L., Iacono, M., Ikeo, K., Iwama, A., Ishikawa, T., Jakt, M., Kanapin, A., Katoh, M., Kawasawa, Y., Kelso, J., Kitamura, H., Kitano, H., Kollias, G., Krishnan, S. P., Kruger, A., Kummerfeld, S. K., Kurochkin, I. V., Lareau, L. F., Lazarevic, D., Lipovich, L., Liu, J., Liuni, S., Mcwilliam, S., Madan Babu, M., Madera, M., Marchionni, L., Matsuda, H., Matsuzawa, S., Miki, H., Mignone, F., Miyake, S., Morris, K., Mottagui-Tabar, S., Mulder, N., Nakano, N., Nakauchi, H., Ng, P., Nilsson, R., Nishiguchi, S., Nishikawa, S., Nori, F., Ohara, O., Okazaki, Y., Orlando, V., Pang, K. C., Pavan, W. J., Pavesi, G., Pesole, G., Petrovsky, N., Piazza, S., Reed, J., Reid, J. F., Ring, B. Z., Ringwald, M., Rost, B., Ruan, Y., Salzberg, S. L., Sandelin, A., Schneider, C., Schönbach, C., Sekiguchi, K., Semple, C. A., Seno, S., Sessa, L., Sheng, Y., Shibata, Y., Shimada, H., Shimada, K., Silva, D., Sinclair, B., Sperling, S., Stupka, E., Sugiura, K., Sultana, R., Takenaka, Y., Taki, K., Tammoja, K., Tan, S. L., Tang, S., Taylor, M. S., Tegner, J., Teichmann, S. A., Ueda, H. R., van Nimwegen, E., Verardo, R., Wei, C. L., Yagi, K., Yamanishi, H., Zabarovsky, E., Zhu, S., Zimmer, A., Hide, W., Bult, C., Grimmond, S. M., Teasdale, R. D., Liu, E. T., Brusic, V., Quackenbush, J., Wahlestedt, C., Mattick, J. S., Hume, D. A., Kai, C., Sasaki, D., Tomaru, Y., Fukuda, S., Kanamori-Katayama, M., Suzuki, M., Aoki, J., Arakawa, T., Iida, J., Imamura, K., Itoh, M., Kato, T., Kawaji, H., Kawagashira, N., Kawashima, T., Kojima, M., Kondo, S., Konno, H., Nakano, K., Ninomiya, N., Nishio, T., Okada, M., Plessy, C., Shibata, K., Shiraki, T., Suzuki, S., Tagami, M., Waki, K., Watahiki, A., Okamura-Oho, Y., Suzuki, H., Kawai, J., Hayashizaki, Y; FANTOM Consortium; RIKEN Genome Exploration Research Group and Genome Science Group (Genome Network Project Core Group). (2005). The transcriptional landscape of the mammalian genome. *Science* 309, 1559–1563.

Carninci, P., Yasuda, J., and Hayashizaki, Y. (2008). Multifaceted mammalian transcriptome. *Curr. Opin. Cell Biol.* 20, 274–280.

Chen, M., and Manley, J. L. (2009). Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. *Nat. Rev. Mol. Cell Biol.* 10, 741–754.

Cheng, J., Kapranov, P., Drenkow, J., Dike, S., Brubaker, S., Patel, S., Long, J., Stern, D., Tammana, H., Helt, G., Sementchenko, V., Piccolboni, A., Bekiranov, S., Bailey, D. K., Ganesh, M., Ghosh, S., Bell, I., Gerhard, D. S., and Gingeras, T. R. (2005). Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science* 308, 1149–1154.

Cooper, T. A., Wan, L., and Dreyfuss, G. (2009). RNA and disease. *Cell* 136, 777–793.

Darnell, R. B. (2006). Developing global insight into RNA regulation. *Cold Spring Harb. Symp. Quant. Biol.* 71, 321–327.

Egecioglu, D. E., and Chanfreau, G. (2011). Proofreading and spellchecking: a two-tier strategy for pre-mRNA splicing quality control. *RNA* 17, 383–389.

Faghihi, M. A., Modarresi, F., Khalil, A. M., Wood, D. E., Sahagan, B. G., Morgan, T. E., Finch, C. E., St. Laurent, G. III, Kenny, P. J., and Wahlestedt, C. (2008). Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of beta-secretase. *Nat. Med.* 14, 723–730.

Gagen, M. J., and Mattick, J. S. (2005). Inherent size constraints on prokaryote gene networks due to "accelerating" growth. *Theory Biosci.* 123, 381–411.

Garcia-Blanco, M. A., Baraniak, A. P., and Lasda, E. L. (2004). Alternative splicing in disease and therapy. *Nat. Biotechnol.* 22, 535–546.

Garrett, S., and Rosenthal, J. J. (2012). RNA editing underlies temperature adaptation in K+ channels from polar octopuses. *Science* 335, 848–851.

Guttman, M., Amit, I., Garber, M., French, C., Lin, M. F., Feldser, D., Huarte, M., Zuk, O., Carey, B. W., Cassady, J. P., Cabili, M. N., Jaenisch, R., Mikkelsen, T. S., Jacks, T., Hacohen, N., Bernstein, B. E., Kellis, M., Regev, A., Rinn, J. L., and Lander, E. S. (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* 458, 223–227.

Hallegger, M., Llorian, M., and Smith, C. W. (2010). Alternative splicing: global insights. *FEBS J.* 277, 856–866.

Hanrahan, C. J., Palladino, M. J., Ganetzky, B., and Reenan, R. A. (2000). RNA editing of the *Drosophila* para Na(+) channel transcript. Evolutionary conservation and developmental regulation. *Genetics* 155, 1149–1160.

Hastie, N. D., and Bishop, J. O. (1976). The expression of three abundance classes of messenger RNA in mouse tissues. *Cell* 9, 761–774.

Herbert, A., and Rich, A. (1999). RNA processing and the evolution of eukaryotes. *Nat. Genet.* 21, 265–269.

Hertel, K. J. (2008). Combinatorial control of exon recognition. *J. Biol. Chem.* 283, 1211–1215.

Higuchi, M., Single, F. N., Kohler, M., Sommer, B., Sprengel, R., and Seeburg, P. H. (1993). RNA editing of AMPA receptor subunit GluR-B: a base-paired intron–exon structure determines position and efficiency. *Cell* 75, 1361–1370.

Hoopengardner, B., Bhalla, T., Staber, C., and Reenan, R. (2003). Nervous system targets of RNA editing identified by comparative genomics. *Science* 301, 832–836.

Ingleby, L., Maloney, R., Jepson, J., Horn, R., and Reenan, R. (2009). Regulated RNA editing and functional epistasis in Shaker potassium channels. *J. Gen. Physiol.* 133, 17–27.

Jeon, Y., and Lee, J. T. (2011). YY1 tethers Xist RNA to the inactive X nucleation center. *Cell* 146, 119–133.

Jia, Y., Mu, J. C., and Ackerman, S. L. (2012). Mutation of a U2 snRNA gene causes global disruption of alternative splicing and neurodegeneration. *Cell* 148, 296–308.

Kanhere, A., and Jenner, R. G. (2012). Noncoding RNA localisation mechanisms in chromatin regulation. *Silence* 3, 2.

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.

Kapranov, P., Cheng, J., Dike, S., Nix, D. A., Duttagupta, R., Willingham, A. T., Stadler, P. F., Hertel, J., Hackermuller, J., Hofacker, I. L., Bell, I., Cheung, E., Drenkow, J., Dumais, E., Patel, S., Helt, G., Ganesh, M., Ghosh, S., Piccolboni, A., Sementchenko, V., Tammana, H., and Gingeras, T. R. (2007a). RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* 316, 1484–1488.

Kapranov, P., Willingham, A. T., and Gingeras, T. R. (2007b). Genome-wide transcription and the implications for genomic organization. *Nat. Rev. Genet.* 8, 413–423.

Kapranov, P., St. Laurent, G., Raz, T., Ozsolak, F., Reynolds, C. P., Sorensen,

P. H., Reaman, G., Milos, P., Arceci, R. J., Thompson, J. F., and Triche, T. J. (2010). The majority of total nuclear-encoded non-ribosomal RNA in a human cell is 'dark matter' un-annotated RNA. *BMC Biol.* 8, 149. doi: 10.1186/1741-7007-8-149

Katayama, S., Tomaru, Y., Kasukawa, T., Waki, K., Nakanishi, M., Nakamura, M., Nishida, H., Yap, C. C., Suzuki, M., Kawai, J., Suzuki, H., Carninci, P., Hayashizaki, Y., Wells, C., Frith, M., Ravasi, T., Pang, K. C., Hallinan, J., Mattick, J., Hume, D. A., Lipovich, L., Batalov, S., Engstrom, P. G., Mizuno, Y., Faghihi, M. A., Sandelin, A., Chalk, A. M., Mottagui-Tabar, S., Liang, Z., Lenhard, B., and Wahlestedt, C. (2005). Antisense transcription in the mammalian transcriptome. *Science* 309, 1564–1566.

Khaitan, D., Dinger, M. E., Mazar, J., Crawford, J., Smith, M. A., Mattick, J. S., and Perera, R. J. (2011). The melanoma-upregulated long noncoding RNA SPRY4-IT1 modulates apoptosis and invasion. *Cancer Res.* 71, 3852–3862.

Khalil, A. M., Guttman, M., Huarte, M., Garber, M., Raj, A., Rivea Morales, D., Thomas, K., Presser, A., Bernstein, B. E., Van Oudenaarden, A., Regev, A., Lander, E. S., and Rinn, J. L. (2009). Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc. Natl. Acad. Sci. U.S.A.* 106, 11667–11672.

Kishore, S., and Stamm, S. (2006). The snoRNA HBII-52 regulates alternative splicing of the serotonin receptor 2C. *Science* 311, 230–232.

Koodathingal, P., Novak, T., Piccirilli, J. A., and Staley, J. P. (2010). The DEAH box ATPases Prp16 and Prp43 cooperate to proofread 5′ splice site cleavage during pre-mRNA splicing. *Mol. Cell* 39, 385–395.

Lanctot, C., Cheutin, T., Cremer, M., Cavalli, G., and Cremer, T. (2007). Dynamic genome architecture in the nuclear space: regulation of gene expression in three dimensions. *Nat. Rev. Genet.* 8, 104–115.

Lee, J. T. (2011). Gracefully ageing at 50, X-chromosome inactivation becomes a paradigm for RNA and chromatin control. *Nat. Rev. Mol. Cell Biol.* 12, 815–826.

Leulliot, N., and Varani, G. (2001). Current topics in RNA-protein recognition: control of specificity and biological function through induced fit and conformational capture. *Biochemistry* 40, 7947–7956.

Licatalosi, D. D., and Darnell, R. B. (2010). RNA processing and its regulation: global insights into biological networks. *Nat. Rev. Genet.* 11, 75–87.

Licatalosi, D. D., Mele, A., Fak, J. J., Ule, J., Kayikci, M., Chi, S. W., Clark, T. A., Schweitzer, A. C., Blume, J. E., Wang, X., Darnell, J. C., and Darnell, R. B. (2008). HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature* 456, 464–469.

Lloyd, S. (2001). Measures of complexity: a nonexhaustive list. *IEEE Control Syst.* 21, 7–8.

Luco, R. F., Allo, M., Schor, I. E., Kornblihtt, A. R., and Misteli, T. (2011). Epigenetics in alternative pre-mRNA splicing. *Cell* 144, 16–26.

Mattick, J. S. (2007). A new paradigm for developmental biology. *J. Exp. Biol.* 210, 1526–1547.

Mayer, C., Schmitz, K. M., Li, J., Grummt, I., and Santoro, R. (2006). Intergenic transcripts regulate the epigenetic state of rRNA genes. *Mol. Cell* 22, 351–361.

McKee, A. E., and Silver, P. A. (2007). Systems perspectives on mRNA processing. *Cell Res.* 17, 581–590.

Mercer, T. R., Dinger, M. E., Bracken, C. P., Kolle, G., Szubert, J. M., Korbie, D. J., Askarian-Amiri, M. E., Gardiner, B. B., Goodall, G. J., Grimmond, S. M., and Mattick, J. S. (2010). Regulated post-transcriptional RNA cleavage diversifies the eukaryotic transcriptome. *Genome Res.* 20, 1639–1650.

Misteli, T. (2007). Beyond the sequence: cellular organization of genome function. *Cell* 128, 787–800.

Nam, Y., Chen, C., Gregory, R. I., Chou, J. J., and Sliz, P. (2011). Molecular basis for interaction of let-7 microRNAs with Lin28. *Cell* 147, 1080–1091.

Nishikura, K. (2006). Editor meets silencer: crosstalk between RNA editing and RNA interference. *Nat. Rev. Mol. Cell Biol.* 7, 919–931.

Nishikura, K., Yoo, C., Kim, U., Murray, J. M., Estes, P. A., Cash, F. E., and Liebhaber, S. A. (1991). Substrate specificity of the dsRNA unwinding/modifying activity. *EMBO J.* 10, 3523–3532.

Pang, K. C., Frith, M. C., and Mattick, J. S. (2006). Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. *Trends Genet.* 22, 1–5.

Peng, J. C., and Karpen, G. H. (2007). H3K9 methylation and RNA interference regulate nucleolar organization and repeated DNA stability. *Nat. Cell Biol.* 9, 25–35.

Piskounova, E., Polytarchou, C., Thornton, J. E., Lapierre, R. J., Pothoulakis, C., Hagan, J. P., Iliopoulos, D., and Gregory, R. I. (2011). Lin28A and Lin28B inhibit let-7 microRNA biogenesis by distinct mechanisms. *Cell* 147, 1066–1079.

Polson, A. G., and Bass, B. L. (1994). Preferential selection of adenosines for modification by double-stranded RNA adenosine deaminase. *EMBO J.* 13, 5701–5711.

Polydorides, A. D., Okano, H. J., Yang, Y. Y., Stefani, G., and Darnell, R. B. (2000). A brain-enriched polypyrimidine tract-binding protein antagonizes the ability of Nova to regulate neuron-specific alternative splicing. *Proc. Natl. Acad. Sci. U.S.A.* 97, 6350–6355.

Prasanth, K. V., Prasanth, S. G., Xuan, Z., Hearn, S., Freier, S. M., Bennett, C. F., Zhang, M. Q., and Spector, D. L. (2005). Regulating gene expression through RNA nuclear retention. *Cell* 123, 249–263.

Raz, T., Kapranov, P., Lipson, D., Letovsky, S., Milos, P. M., and Thompson, J. F. (2011). Protocol dependence of sequencing-based gene expression measurements. *PLoS ONE* 6, e19287. doi: 10.1371/journal.pone.0019287

Reenan, R. A. (2005). Molecular determinants and guided evolution of species-specific RNA editing. *Nature* 434, 409–413.

Robinson, R. (2010). Dark matter transcripts: sound and fury, signifying nothing? *PLoS Biol.* 8, e1000370. doi: 10.1371/journal.pbio.1000370

Rosenthal, J. J., and Bezanilla, F. (2002). Extensive editing of mRNAs for the squid delayed rectifier K+ channel regulates subunit tetramerization. *Neuron* 34, 743–757.

Salmena, L., Poliseno, L., Tay, Y., Kats, L., and Pandolfi, P. P. (2012). A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? *Cell* 146, 353–358.

Scadden, A. D. (2005). The RISC subunit Tudor-SN binds to hyper-edited double-stranded RNA and promotes its cleavage. *Nat. Struct. Mol. Biol.* 12, 489–496.

Schmitz, K. M., Mayer, C., Postepska, A., and Grummt, I. (2010). Interaction of noncoding RNA with the rDNA promoter mediates recruitment of DNMT3b and silencing of rRNA genes. *Genes Dev.* 24, 2264–2269.

Seeburg, P. H., and Hartner, J. (2003). Regulation of ion channel/neurotransmitter receptor function by RNA editing. *Curr. Opin. Neurobiol.* 13, 279–283.

Sharma, A., and Lou, H. (2011). Depolarization-mediated regulation of alternative splicing. *Front. Neurosci.* 5:141. doi: 10.3389/fnins.2011.00141

Sharma, S., and Black, D. L. (2006). Maps, codes, and sequence elements: can we predict the protein output from an alternatively spliced locus? *Neuron* 52, 574–576.

Spector, D. L., and Lamond, A. I. (2011). Nuclear speckles. *Cold Spring Harb. Perspect. Biol.* 3, a000646. doi: 10.1101/cshperspect.a000646.

St. Laurent, G. III, and Wahlestedt, C. (2007). Noncoding RNAs: couplers of analog and digital information in nervous system function? *Trends Neurosci.* 30, 612–621.

Struhl, K. (2007). Transcriptional noise and the fidelity of initiation by RNA polymerase II. *Nat. Struct. Mol. Biol.* 14, 103–105.

Taft, R. J., Pheasant, M., and Mattick, J. S. (2007). The relationship between non-protein-coding DNA and eukaryotic complexity. *Bioessays* 29, 288–299.

Tompa, P., and Csermely, P. (2004). The role of structural disorder in the function of RNA and protein chaperones. *FASEB J.* 18, 1169–1175.

Tsai, M. C., Manor, O., Wan, Y., Mosammaparast, N., Wang, J. K., Lan, F., Shi, Y., Segal, E., and Chang, H. Y. (2010). Long noncoding RNA as modular scaffold of histone modification complexes. *Science* 329, 689–693.

Ule, J., and Darnell, R. B. (2006). RNA binding proteins and the regulation of neuronal synaptic plasticity. *Curr. Opin. Neurobiol.* 16, 102–110.

Ule, J., and Darnell, R. B. (2007). Functional and mechanistic insights from genome-wide studies of splicing regulation in the brain. *Adv. Exp. Med. Biol.* 623, 148–160.

van Bakel, H., and Hughes, T. R. (2009). Establishing legitimacy and function in the new transcriptome. *Brief. Funct. Genomic Proteomic* 8, 424–436.

van Bakel, H., Nislow, C., Blencowe, B. J., and Hughes, T. R. (2010). Most "dark matter" transcripts are associated with known genes. *PLoS Biol.* 8, e1000371. doi: 10.1371/journal.pbio.1000371

Venables, J. P., Klinck, R., Koh, C., Gervais-Bird, J., Bramard, A., Inkel, L., Durand, M., Couture, S., Froehlich, U., Lapointe, E., Lucier, J. F., Thibault, P., Rancourt, C., Tremblay, K., Prinos, P., Chabot, B., and Elela, S. A. (2009). Cancer-associated regulation of alternative splicing. *Nat. Struct. Mol. Biol.* 16, 670–676.

Wai, D. H., Wu, D. U., Wing, M. R., Arceci, R. J., Reynolds, C. P., Sorensen, P. H., Reaman, G. H., Milos, P. M.,. Lawlor, E. R, Buckley,

J. D., Kapranov, P., and Triche, T. J. (2010). Large intergenic noncoding RNAs associated with Ewing sarcoma family of tumors. *Proc. Am. Assoc. Cancer Res.* Abstract #4087.

Waks, Z., Klein, A. M., and Silver, P. A. (2011). Cell-to-cell variability of alternative RNA splicing. *Mol. Syst. Biol.* 7, 506.

Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S. F., Schroth, G. P., and Burge, C. B. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature* 456, 470–476.

Wang, K. C., and Chang, H. Y. (2011). Molecular mechanisms of long noncoding RNAs. *Mol. Cell* 43, 904–914.

Wang, Q., O'Brien, P. J., Chen, C. X., Cho, D. S., Murray, J. M., and Nishikura, K. (2000). Altered G protein-coupling functions of RNA editing isoform and splicing variant serotonin2C receptors. *J. Neurochem.* 74, 1290–1300.

Wang, Q., Zhang, Z., Blackwell, K., and Carmichael, G. G. (2005). Vigilins bind to promiscuously A-to-I-edited RNAs and are involved in the formation of heterochromatin. *Curr. Biol.* 15, 384–391.

Ward, A. J., and Cooper, T. A. (2010). The pathobiology of splicing. *J. Pathol.* 220, 152–163.

Wei, P., Garber, M. E., Fang, S. M., Fischer, W. H., and Jones, K. A. (1998). A novel CDK9-associated C-type cyclin interacts directly with HIV-1 Tat and mediates its high-affinity, loop-specific binding to TAR RNA. *Cell* 92, 451–462.

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). A strategy for probing the function of noncoding RNAs finds a repressor of NFAT. *Science* 309, 1570–1573.

Witten, J. T., and Ule, J. (2011). Understanding splicing regulation through RNA splicing maps. *Trends Genet.* 27, 89–97.

Wojtowicz, W. M., Flanagan, J. J., Millard, S. S., Zipursky, S. L., and Clemens, J. C. (2004). Alternative splicing of *Drosophila* Dscam generates axon guidance receptors that exhibit isoform-specific homophilic binding. *Cell* 118, 619–633.

Xiao, X., and Lee, J. H. (2010). Systems analysis of alternative splicing and its regulation. *Wiley Interdiscip. Rev. Syst. Biol. Med.* 2, 550–565.

Yang, J. H., Luo, X., Nie, Y., Su, Y., Zhao, Q., Kabir, K., Zhang, D., and Rabinovici, R. (2003a). Widespread inosine-containing mRNA in lymphocytes regulated by ADAR1 in response to inflammation. *Immunology* 109, 15–23.

Yang, J. H., Nie, Y., Zhao, Q., Su, Y., Pypaert, M., Su, H., and Rabinovici, R. (2003b). Intracellular localization of differentially regulated RNA-specific adenosine deaminase isoforms in inflammation. *J. Biol. Chem.* 278, 45833–45842.

Zhang, Z., and Carmichael, G. G. (2001). The fate of dsRNA in the nucleus: a p54(nrb)-containing complex mediates the nuclear retention of promiscuously A-to-I edited RNAs. *Cell* 106, 465–475.

Zhao, J., Ohsumi, T. K., Kung, J. T., Ogawa, Y., Grau, D. J., Sarma, K., Song, J. J., Kingston, R. E., Borowsky, M., and Lee, J. T. (2010). Genome-wide identification of polycomb-associated RNAs by RIP-seq. *Mol. Cell* 40, 939–953.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 19 February 2012; accepted: 28 March 2012; published online: 26 April 2012.

Citation: St. Laurent G, Savva YA and Kapranov P (2012) Dark matter RNA: an intelligent scaffold for the dynamic regulation of the nuclear information landscape. Front. Gene. ***3****:57. doi: 10.3389/fgene.2012.00057*

This article was submitted to Frontiers in Non-Coding RNA, a specialty of Frontiers in Genetics.