



# Non-coding RNA: what is functional and what is junk?

Alexander F. Palazzo\* and Eliza S. Lee

Department of Biochemistry, University of Toronto, Toronto, ON, Canada

## Edited by:

Karim Mekhail, University of Toronto, Canada

## Reviewed by:

Deng-Ke Niu, Beijing Normal University, China

Pedro Miura, University of Nevada, Reno, USA

## \*Correspondence:

Alexander F. Palazzo, Department of Biochemistry, University of Toronto, 1 King's College Circle, MSB Room 5336, Toronto, ON M5S 1A8, Canada  
e-mail: alex.palazzo@utoronto.ca

The genomes of large multicellular eukaryotes are mostly comprised of non-protein coding DNA. Although there has been much agreement that a small fraction of these genomes has important biological functions, there has been much debate as to whether the rest contributes to development and/or homeostasis. Much of the speculation has centered on the genomic regions that are transcribed into RNA at some low level. Unfortunately these RNAs have been arbitrarily assigned various names, such as “intergenic RNA,” “long non-coding RNAs” etc., which have led to some confusion in the field. Many researchers believe that these transcripts represent a vast, uncharted world of functional non-coding RNAs (ncRNAs), simply because they exist. However, there are reasons to question this Panglossian view because it ignores our current understanding of how evolution shapes eukaryotic genomes and how the gene expression machinery works in eukaryotic cells. Although there are undoubtedly many more functional ncRNAs yet to be discovered and characterized, it is also likely that many of these transcripts are simply junk. Here, we discuss how to determine whether any given ncRNA has a function. Importantly, we advocate that in the absence of any such data, the appropriate null hypothesis is that the RNA in question is junk.

**Keywords:** Junk RNA, Junk DNA, non-coding RNA, evolution, genome biology

## INTRODUCTION

Starting with the discovery of transfer RNA and ribosomal RNA in the 1950s, non-coding RNAs (ncRNAs) with biological roles have been known for close to 60 years. Even in the late 1970s and early 1980s the existence of other functional ncRNAs was known, including RNase P (Stark et al., 1978), snRNAs (Yang et al., 1981), and 7SL [the RNA component of the signal recognition particle (Walter and Blobel, 1982)]. Later, ncRNAs that serve to regulate chromosome structure, such as Xist, were discovered (Brockdorff et al., 1992). Since then, the number of new and putative functional ncRNAs has greatly expanded (for reviews see Wilusz et al., 2009; Wang and Chang, 2011; Ulitsky and Bartel, 2013; Rinn and Guttman, 2014). Interest in this field was further stimulated by the finding that almost all of the mammalian genome is transcribed at some level (Carninci et al., 2005; Birney et al., 2007; Djebali et al., 2012), with some individuals speculating that much of this pervasive transcription is likely functional (Mattick et al., 2010; Ecker et al., 2012; Pennisi, 2012). This idea was epitomized by the ENCODE consortium, which claimed to have assigned “biochemical functions for 80% of the genome” (ENCODE Project Consortium et al., 2012). Others have disagreed, pointing out that the vast majority of these novel transcripts are present at low levels, and that the term “function” had been misappropriated (Eddy, 2012; Doolittle, 2013; Graur et al., 2013; Niu and Jiang, 2013; Palazzo and Gregory, 2014). Despite these criticisms, the idea that the pervasive transcription of the human genome plays some role in homeostasis and/or development persists, with one group even proclaiming that they had “refuted the specific claims that most of the observed transcription across the human genome is random” (Mattick and Dinger, 2013).

At present, the distinction between functional ncRNAs and junk RNA appears to be quite vague. There has been, however, some effort to differentiate between these two groups, based on various criteria ranging from their expression levels and splicing to conservation. Ultimately these efforts have failed to bring consensus to the field.

A similar problem has plagued the investigation of whether transposable elements (TEs), which make up a significant proportion of most vertebrate genomes, have been exapted for the benefit of the host organism. Although some have claimed that many TEs are functional, a few groups have offered a much more balanced view that is in line with our current understanding of molecular evolution (de Souza et al., 2013; Elliott et al., 2014).

In this article we explain several concepts that researchers must keep in mind when evaluating whether a given ncRNA has a function at the organismal level. Importantly, the presence of low abundant non-functional transcripts is entirely consistent with our current understanding of how eukaryotic gene expression works and how the eukaryotic genome is shaped by evolution. With this in mind, researchers should take the approach that an uncharacterized non-coding RNA likely has no function, unless proven otherwise. This is the null hypothesis. If a given ncRNA has supplementary attributes that would not be expected to be found in junk RNA, then this would provide some evidence that this transcript may be functional.

## THE AMOUNT OF VARIOUS RNA SPECIES IN THE TYPICAL EUKARYOTIC CELL

As is evident from a number of sources, almost all of the human genome is transcribed. However, one must not confuse the number of different *types* of transcripts with their *abundance* in a typical

cell. Many of the putative functional ncRNAs are present at very low levels and thus unlikely to be of any importance with respect to cell or organismal physiology. Importantly, the abundance of an ncRNA species roughly correlates with its level of conservation (Managadze et al., 2011), which is a good proxy for function (Doolittle et al., 2014; Elliott et al., 2014; however, see below); thus, determining the relative abundance of a given ncRNA in the relevant cell type is an important piece of information. However, one should keep in mind that if the ncRNA has catalytic activity or if it acts as a scaffold to regulate chromosomal architecture near its site of transcription, the RNA may not need to be present at very high levels to be able to perform its task.

At steady state, the vast majority of human cellular RNA consists of rRNA (~90% of total RNA for most cells, see **Table 1** and **Figure 1**). Although there is less tRNA by mass, their small size results in their molar level being higher than rRNA (**Figure 1**). Other abundant RNAs, such as mRNA, snRNA, and snoRNAs are present in aggregate at levels that are about 1–2 orders of magnitude lower than rRNA and tRNA (**Table 1** and **Figure 1**). Certain small RNAs, such as miRNA and piRNAs can be present at very high levels; however, this appears to be cell type dependent.

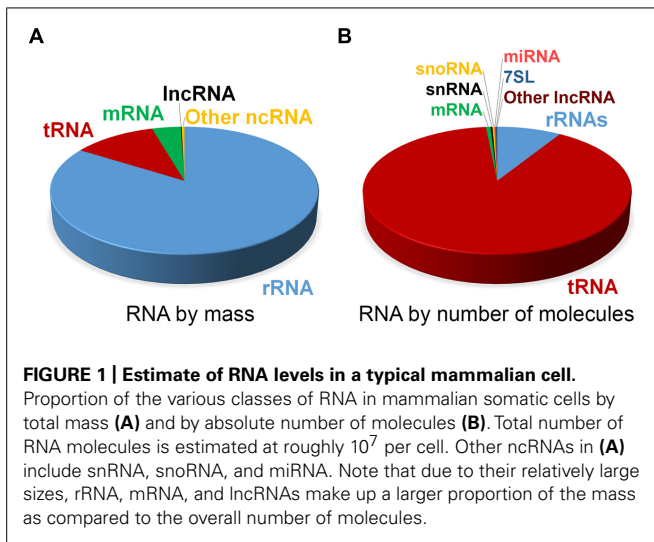
By general convention, most other ncRNAs longer than 200 nucleotides, regardless of whether or not they have a known function, have been lumped together into a category called “long non-coding RNAs” (lncRNAs). As a whole, these are present at levels that are two orders of magnitude less than total mRNA (**Table 1**). Although the estimated number of different types of human lncRNAs has ranged from 5,400 to 53,000 (**Table 2**), only a small fraction have been found to be present at levels high enough to suggest that they have a function. According to ENCODE’s own estimates, fewer than 1,000 lncRNAs are present at greater than one copy per cell in the typical human tissue culture cell line (Djebali et al., 2012; Palazzo and Gregory, 2014), although some other estimates have determined that the levels may be substantially higher (Hangauer et al., 2013). One caveat with the data collected thus far is that some of these lncRNAs may have a very restricted expression pattern; therefore until the relevant cell type is tested, we may not be in a position to judge whether it is expressed at a sufficient level to provide evidence of functionality. It is also worthwhile noting that certain annotated lncRNAs may actually encode short functional peptides (Ingolia et al., 2011, 2014; Magny et al., 2013; Bazzini et al., 2014), although

**Table 1 | Estimates of total RNA content in mammalian cells.**

Type	Percent of total RNA by mass	Molecules per cell	Average size (kb)	Total weight picograms/cell	Notes	Reference
rRNAs	80 to 90	3–10 × 10 <sup>6</sup> (ribosomes)	6.9	10 to 30		Blobel and Potter (1967), Wolf and Schlessinger (1977), Duncan and Hershey (1983)
tRNA	10 to 15	3–10 × 10 <sup>7</sup>	<0.1	1.5 to 5	About 10 tRNA molecules /ribosome	Waldron and Lacroute (1975)
mRNA	3 to 7	3–10 × 10 <sup>5</sup>	1.7	0.25 to 0.9		Hastie and Bishop (1976), Carter et al. (2005)
hnRNA (pre-mRNA)	0.06 to 0.2	1–10 × 10 <sup>3</sup>	10*	0.004 to 0.03	Estimated at 2–4% of mRNA by weight	Mortazavi et al. (2008), Menet et al. (2012)
Circular RNA	0.002 to 0.03	3–20 × 10 <sup>3</sup>	~0.5	0.0007 to 0.005	Estimated at 0.1–0.2% of mRNA**	Salzman et al. (2012), Guo et al. (2014)
snRNA	0.02 to 0.3	1–5 × 10 <sup>5</sup>	0.1–0.2	0.008 to 0.04		Kiss and Filipowicz (1992), Castle et al. (2010)
snoRNA	0.04 to 0.2	2–3 × 10 <sup>5</sup>	0.2	0.02 to 0.03		Kiss and Filipowicz (1992), Cooper (2000), Castle et al. (2010)
miRNA	0.003 to 0.02	1–3 × 10 <sup>5</sup>	0.02	0.001 to 0.003	About 10 <sup>5</sup> molecules per 10 pg total RNA	Bissels et al. (2009)
7SL	0.01 to 0.2	3–20 × 10 <sup>4</sup>	0.3	0.005 to 0.03	About 1–2 SRP molecules/100 ribosomes	Raue et al. (2007), Castle et al. (2010)
Xist	0.0003 to 0.02	0.1–2 × 10 <sup>3</sup>	2.8	0.0001 to 0.003		Buzin et al. (1994), Castle et al. (2010)
Other lncRNA	0.03 to 0.2	3–50 × 10 <sup>3</sup>	1	0.002 to 0.03	Estimated at 1–4% of mRNA by weight	Mortazavi et al. (2008), Ramsköld et al. (2009), Menet et al. (2012)

\*The size for the average unspliced pre-mRNA is 17 kb; however, most pre-mRNAs are partially spliced at any given time, and the average size of hnRNA is estimated at 10 kb (Salditt-Georgieff et al., 1976).

\*\*Based on the finding that 1–2% of all mRNA species generate circular RNA, which is present at 10% of the level of the parental mRNA.



in general lncRNAs are poorly translated (Bánfai et al., 2012; Guttman et al., 2013; Hangauer et al., 2013). Finally, it is also worth pointing out that a significant fraction of these lncRNAs may actually be misannotated untranslated regions of known mRNAs (Miura et al., 2013).

Other short ncRNAs have been lumped into several groups, depending on their attributes. For example, several regions of the human genome that are believed to be enhancers, are transcribed into short enhancer RNAs (eRNAs). These are thought to act as scaffolds that regulate the 3D architecture of chromosomes in the vicinity of their transcription site (Lai et al., 2013). eRNAs are typically present at even lower levels than lncRNAs (Djebali et al., 2012; Andersson et al., 2014); however, if these play a localized structural role, then they would be expected to be present at only a few copies per cell.

There are still other more exotic species of RNAs (Cech and Steitz, 2014), including circular RNAs (Wilusz and Sharp, 2013). Due to their lack of free 5' or 3' ends, circular RNAs are quite stable and some can accumulate to levels that are comparable to mRNAs (Salzman et al., 2012; Jeck et al., 2013; Memczak et al., 2013). However, it is likely that besides a few examples, circular RNAs represent a minute fraction of the total pool of cellular RNAs (see Table 1).

In addition to all of the mentioned species, ENCODE and other groups have found transcripts that map to the rest of the genome termed “intergenic RNA” (Djebali et al., 2012). Most of these transcripts are present at levels that are significantly below one copy per cell (Djebali et al., 2012; Palazzo and Gregory, 2014). Again this arbitrary division of ncRNAs has led to much confusion. It is unclear why these transcripts are considered to be intergenic if they are also functional (as in 80% of the genome is functional); after all, if a region of DNA that is transcribed into a functional product is called a gene, then the term intergenic would automatically imply that these regions have no function.

Regardless of these concerns, it is clear that most of the ncRNAs in question (lncRNAs, eRNAs, circular RNAs, intergenic RNAs, etc.) are typically present at very low levels when compared to known functional RNAs. These observations are consistent with

the idea that the eukaryotic genome produces a vast amount of spurious transcripts.

## WHERE DO ALL THESE ncRNAs COME FROM?

As of spring 2014, the LNCipedia website<sup>1</sup> (Volders et al., 2013) has compiled a list of ~21,000 human lncRNAs, with an average length of about 1 kb (Table 2). These would originate from <1% of the human genome. Needless to say, this is a very small fraction of the total. Even if we compiled all of the putative lncRNAs using the most optimistic analysis (Managadze et al., 2013), all the putative lncRNAs would still be transcribed from at most 2% of the genome (Table 2). Thus far, only a small minority of lncRNAs have been shown to be important for organismal development, cell physiology, and/or homeostasis. As of December 2014, the lncRNA Database<sup>2</sup>, a repository of lncRNAs “curated from evidence supported by the literature,” lists only 166 biologically validated lncRNAs in humans (Quek et al., 2014). Additionally there are so called eRNAs, which according to FANTOM5 come from an additional 43,000 loci. However, at an average length of ~250 nucleotides they would be made from ~0.34% of the human genome (Andersson et al., 2014). Again, these are very small numbers.

In summary, our best candidates for novel functional ncRNAs (lncRNAs, eRNAs) arise from only a minute fraction of the genome. Again it appears that the vast majority of the genome that falls outside of these loci is transcribed into junk RNA that is present at very low levels at steady state.

## BIOCHEMICAL SUPPORT FOR JUNK RNA

It is important to recognize that the pervasive transcription associated with the human genome is entirely consistent with our understanding of biochemistry. Although RNA polymerases prefer to start transcription at promoter regions, they do have a low probability of initiating transcription on any accessible DNA (Struhl, 2007; Tisseur et al., 2011). Indeed it has been observed that most nucleosome-free DNA is transcribed *in vivo* (Cheung et al., 2008) and that many random pieces of DNA can promote transcription by recruiting transcription factors [TFs; see figure S4 in White et al. (2013)].

Of course eukaryotic cells limit the amount of inappropriate transcription by packaging intergenic regions into heterochromatin. This shields the DNA from both RNA polymerases and TFs which can bind to DNA and activate adjacent cryptic transcriptional start sites. The formation of these heterochromatic regions is largely dictated by a complicated array of DNA elements that initiate and restrict chromatin packing. However, there is quite a bit of data that supports the notion that heterochromatin formation is not always strictly regulated or enforced. For example, it has been shown that many heterochromatic regions are transcriptionally active, albeit at a low level (Moazed, 2009), suggesting that either heterochromatin is periodically loosened, or that under certain circumstances RNA polymerases can transcribe these tightly packed regions. Another line of evidence that suggests that heterochromatin formation is not strictly regulated comes from the

<sup>1</sup><http://www.lncipedia.org/>

<sup>2</sup><http://www.lncrnadb.org/>

**Table 2 | Estimate number of human ncRNAs from various sources.**

Source	Number of types	Percent of human genome	Estimated average size (kb)	Reference
Mammalian lincRNome	53647	2.3	1.4	Managadze et al. (2013)
LNCipedia (as of March 2014)*	21487	0.67	1	Volders et al. (2013)
FPKM > 1 lincRNAs	35585	1.1**	NA	Hangauer et al. (2013)
Gencode v7 catalog of human ncRNAs	9277	0.29**	Median of 0.6***	Derrien et al. (2012)
LncRNAs – Gencode v21 (as of 2014 November)	15877	0.50**	NA	www.gencodegenes.org/stats.html
Jia et al. (2010)	5446	0.17**	NA	Jia et al. (2010)
Cabili – low confidence	8195	0.26	1	Cabili et al. (2011)
Cabili – high confidence	4273	0.26	1	Cabili et al. (2011)
Small ncRNAs – Gencode v21 (as of 2014 November)****	9534	0.045	0.15	www.gencodegenes.org/stats.html
eRNAs	43011	0.34	0.25	Andersson et al. (2014)

\*Splice variants were excluded.

\*\*Assumes an average length of 1 kb.

\*\*\*Median size of 592 nucleotides but with a significant fraction at higher sizes.

\*\*\*\*Contains snRNAs, snoRNAs, rRNAs, mitochondrial tRNAs and rRNAs, miRNAs, and “miscellaneous RNAs.”

investigation of TF binding sites. In particular, it has been observed that most TF binding sites which are occupied by TF proteins are not conserved between highly related species (Paris et al., 2013) and that many TF binding events have little to no impact on the expression of nearby genes (Li et al., 2008; Biggin, 2011; Lickwar et al., 2012; Paris et al., 2013). In other words, many putative TF binding sites are created and destroyed by neutral evolution and do not appear to contribute to the expression of functional parts of the genome. These TF binding sites are nonetheless accessible to TF proteins, and thus are not found in heterochromatin.

From the above discussion it is clear that there are many sources for cryptic transcription in eukaryotic genomes. Consistent with this idea, it was found that nascent RNA polymerase II transcripts from mouse liver cells generate a fair amount of transcripts that map to unannotated genomic regions (Menet et al., 2012). When these nascent transcripts were analyzed by next generation sequencing, the number of reads that mapped to intergenic regions (i.e., unannotated parts of the genome) was equal to those mapping to known exonic regions (Menet et al., 2012). Thus it appears that transcription in mammalian cells is quite non-specific.

Although many newly synthesized transcripts likely originate from non-functional parts of the genome, these RNAs are present at very low levels at steady state (Ramsköld et al., 2009; van Bakel et al., 2010; Menet et al., 2012), suggesting that they are rapidly degraded. It is likely that various quality control mechanisms degrade any RNA that lacks features that are overrepresented in either protein-coding mRNAs or other functional ncRNAs (Palazzo and Akef, 2012; Palazzo et al., 2013). This idea is supported by numerous studies that have documented that the level of these spurious transcripts increase when RNA degradation machinery is either depleted or inhibited (Wyers et al., 2005; Davis and Ares, 2006; Thiebaut et al., 2006; Chekanova et al., 2007; Milligan et al., 2008; Preker et al., 2008; Vasiljeva et al., 2008; Neil et al., 2009; Xu et al., 2009; Tisseur et al., 2011; Tan-Wong et al., 2012). It

is also supported by the fact that features found in mRNAs (e.g., strong splice sites, polyadenylation sites, etc.) act to stabilize the RNA (Lu and Cullen, 2003; Palazzo et al., 2007; Akef et al., 2013).

Thus it appears that transcription in eukaryotes is very messy, but that much of the junk RNA is removed by quality control mechanisms. This view is completely in line with what is known about the biochemistry underlying eukaryotic gene expression.

## EVOLUTIONARY SUPPORT FOR JUNK RNA

Ultimately to understand how TF binding sites, heterochromatin domains, and transcriptional start sites are created and destroyed within the genome, one needs to take into consideration certain concepts that have been derived from the field of population genetics.

One of the most fundamental discoveries in population genetics came from the work of Kimura, Ohta, King and Jukes. They showed that the ability of natural selection to weed out slightly deleterious mutations depends on the size of the breeding population in a given species (Kimura, 1968, 1984; King and Jukes, 1969; Ohta, 1973). The higher the number of individuals, the more powerful natural selection is at identifying slightly deleterious mutations and eliminating them. Due to certain aspects of population dynamics, the effective population size is far smaller than the number of individuals [for a more detailed discussion see (Lynch, 2007)]. For modern humans, the effective population size has been calculated to be 10,000 throughout most of its history, which is typical for mammals (Charlesworth, 2009). Indeed there exists an inverse linear correlation between the effective population size and how deleterious a mutation has to be before it can be effectively eliminated from a population by natural selection. In the absence of selection pressure, some neutral and slightly deleterious mutations will reach fixation due solely to genetic drift [for an extensive examination of this process see (Lynch, 2007)]. It is also important to realize that this relationship also applies to slightly beneficial mutations – there is an inverse correlation

between the effective population size and how beneficial a mutation has to be before it can be effectively selected for by natural selection. Thus when one observes some genetic alteration, it is critical that we keep in mind how the alteration affects the fitness of the organism and whether this change can be acted on by selection (either positively or negatively) given the size of the population.

Given that the displacement of a few nucleosomes can promote transcription initiation (Cheung et al., 2008), that TF binding sites and transcriptional start sites are made up of small degenerate sequences (Stewart et al., 2012), and that many random pieces of DNA can activate transcription (White et al., 2013), we would expect that a large number of random mutations would create fortuitous transcriptional start sites. Importantly, natural selection will be powerless to prevent the appearance of these sites, as long as the resulting RNA is not too deleterious to the organism. Conversely, a transcriptional event needs to provide a substantial advantage before natural selection can act to preserve this alteration in future generations. Most of the data on eukaryotic genomes support the view that the fixation of most genomic alterations are due to drift, while few can be ascribed to positive selection (Lynch, 2007).

Thus the presence of a certain level of junk RNA is not only compatible with our understanding of evolution, but would be expected. Nevertheless, it still remains unclear how much junk RNA a eukaryote could tolerate before natural selection would begin to eliminate it.

### THE DANGERS OF HYPERADAPTATIONISM

The overreliance on adaptationist “just-so stories” in the field of evolutionary biology has been openly criticized since the 1970s. Famously, Gould and Lewontin (1979) compared such thinking to the ideology espoused by Pangloss, the fictional professor from Voltaire’s novel *Candide* who used just-so stories to prove that we lived in the best of all possible worlds. Unfortunately hyperadaptationism, or the belief that the vast majority of traits found in an organism (including its DNA) are present due to some selective force, has plagued much of molecular biology as well (Sarkar, 2014). The proclamation that a biochemical activity is equivalent to function (ENCODE Project Consortium et al., 2012) is just another example of this ideology. Using this logic we would state that any transcribed DNA is functional, but would this mean that the transcript (or transcriptional process) is functional by virtue of its mere existence? To resolve this paradox, we would either have to state that (1) although the DNA is functional, its output, the RNA (or the act of transcription) is not; or (2) that all RNAs are *de facto* functional. Obviously both of these nonsensical conclusions have their roots in hyperadaptationist thinking and an abuse of the concept of biological function. To resolve this, we need to install a more rigorous definition of function. However, this can only be accomplished if we properly define the null hypothesis.

### THROWING DOWN THE GAUNTLET: THE HYPOTHETICAL EXAMPLE OF A NON-FUNCTIONAL ncRNA

To determine the degree to which a process is adaptive, it is important to establish how the exact same events would evolve

by non-adaptive mechanisms. Selection should only be invoked when non-adaptive explanations do not suffice. This viewpoint has been used to determine the contribution of selection to alternative splicing, RNA editing and in determining the lengths of UTRs and introns (Lynch, 2007; Huang and Niu, 2008; Wang et al., 2014; Xu and Zhang, 2014). Here, we would like to introduce the example of a hypothetical non-functional ncRNA as a useful null hypothesis. Again, adaptation (and hence function) should only be invoked if an ncRNA has more attributes than our hypothetical non-functional ncRNA. Using principles of biochemistry and population genetics, we will describe its attributes.

### EXPRESSION LEVELS

This putative non-functional ncRNA would be present at levels that would not be a burden to the cell. There are three considerations to take into account when considering the level at which a ncRNA is present.

First the mere presence of the ncRNA may act as a burden. The typical mammalian tissue culture cell has on the order of 500,000 mRNA molecules. Other RNAs with unknown function (i.e., “intergenic” RNA and lncRNA) are at levels between 1 and 4% those of mRNA (Mortazavi et al., 2008; Ramsköld et al., 2009; Menet et al., 2012) and thus present on the order of about 10,000 total copies per cell (Table 1). Therefore if a hypothetical ncRNA were present at 10 copies per cell at steady state, they would increase the pool of intergenic/lncRNAs by 0.1%, and would increase the total pool of RNA by a negligible amount (Figure 1).

Second, there is a cost to synthesizing the RNA. One study that investigated the energetics of synthesizing long introns has estimated that for an mRNA that is expressed at a level of 30 copies per cell and a half-life of 1 h (resulting in the generation of 360 new RNA molecules/cell per day), an intron would have to be roughly 83,000 nucleotides long for it to be a significant burden, given the effective population size of humans (Huang and Niu, 2008). Using these figures, we can estimate that in humans a non-functional ncRNA that is 1 kb in length and is ubiquitously expressed throughout the body would have to be synthesized at a rate of almost 30,000 copies per cell per day before it would be eliminated by natural selection. Of course if the ncRNA was spliced from a longer transcript, this number would be less.

Third, the ncRNA may have some associated activity that may be deleterious. Most often the major concern is whether it will be translated into short random peptides (see point 3, below). Although ncRNAs are poorly translated, most studies have found that they can be engaged by the ribosome at low levels (Guttman et al., 2013). This can be further mitigated by subcellular localization (see below). Thus as long as the putative ncRNA does not have some activity that negatively impacts some cellular function or the organism in general, our guess would be that if a ncRNA was present even at a level of 10 copies per cell, this small increase in the ncRNA burden would be tolerable (i.e., not deleterious enough to be subjected to negative selection).

### EXPRESSION PROFILES

We might imagine that our putative non-functional RNA was transcribed due to the fortuitous action of one or more TF

binding events. As described above, it is likely that many such sites exist in the mammalian genome as the number of active transcriptional start sites exceed the number of protein-coding genes by an order of magnitude (Carninci et al., 2006). Since the majority of TFs are expressed in a developmentally or spatially regulated manner, it follows that our hypothetical ncRNA will also be expressed in a manner that appears to be under some sort of precise regulatory control. Some researchers have tried to claim tissue-specific expression patterns provide some proof of functionality (Ponting et al., 2009; Hangauer et al., 2013; Mattick and Dinger, 2013); however, such a restricted expression of the ncRNA is entirely consistent with a lack of function.

#### DISTRIBUTION OF THE ncRNA IN THE CELL

In determining how our putative non-functional RNA would be distributed intracellularly, there are several facts to take into account. First, if this RNA were to be exported to the cytoplasm, it is reasonable to believe that it would be a substrate for the translational machinery, as long as the RNA is free of extensive secondary structures. Second, this RNA would be translated into a random polypeptide. Unlike nucleic acids, unstructured polypeptides have a high tendency to aggregate and activate cellular stress (West et al., 1999; Chi et al., 2003). Lastly, a single RNA molecule can be used to generate many polypeptides, thus amplifying any potential deleterious effects. For these reasons, we believe that non-functional RNAs are much more likely to promote cellular stress if they are present in the cytoplasm where they can be translated by ribosomes. Indeed, it is likely that the nucleo-cytoplasmic division evolved in part to prevent ribosomes from translating misprocessed mRNAs and aberrant RNA transcripts (Martin and Koonin, 2006; Akef et al., 2013; Palazzo and Gregory, 2014). This may be the reason that features associated with mRNAs tend to promote their nuclear export (Palazzo and Akef, 2012), while problems during translation will promote the degradation of the RNA by processes such as non-sense mediated decay (Baker and Parker, 2004). These reasons may explain why most lncRNAs are nuclear (Derrien et al., 2012; Djebali et al., 2012) and not significantly translated (Guttman et al., 2013). By this same logic we would expect that our putative ncRNA would not likely be present in the cytoplasm, although we do not yet have any hard data about what level of cytoplasmic ncRNA would be tolerable.

From this discussion it makes sense that our non-functional ncRNA would be nuclear, but what about its localization to a specific sub-nuclear compartment? Again, some have used localization to sub-nuclear loci as proof of functionality (Mattick et al., 2010; Kapusta and Feschotte, 2014). In experiments performed in our lab we have documented how reporter RNAs with an essentially random sequence are indeed localized to discrete nuclear foci. In some cases these colocalize with known nuclear structures, such as nuclear speckles (Akef et al., 2013); in other instances these RNAs form discrete nuclear puncta that are of unknown nature (Lee and Palazzo, unpublished observations). These observations suggest that even sub-nuclear compartmentalization cannot be used as evidence to support functionality for any ncRNA.

#### PROCESSING

We would expect that the non-functional ncRNA would lack strong processing signals, as these regions would be expected to be under strong purifying selection only in functional spliced transcripts. For example in most mRNAs, introns are not only flanked by splicing donor and acceptor sites but are also defined by location of intronic and exonic splicing elements (Blencowe, 2000; Wang et al., 2012). However, to our knowledge, no one has systematically studied the splicing of randomly generated RNAs. Despite this we can still estimate the prevalence of splicing signals computationally. For example, the occurrence of consensus donor and acceptor splice sites in essentially random human DNA sequences are one every 3 and 10 kb, respectively (Shepard et al., 2009). Because the spliceosome requires suboptimal sequences to initiate splicing, it is likely that the actual number of potential donor and acceptor sites is much higher. Thus if the primary transcript of our non-functional RNA is long enough, it will probably be spliced to a certain extent. As for smaller transcripts, a small but significant number are also likely to be spliced. However, since splicing helps to stabilize the RNA (Palazzo and Akef, 2012), it is likely that a non-functional ncRNA would only be present at detectable levels by virtue of the fact that it is spliced. In other words, although a lack of processing would lead to the instability of many functionless RNAs, we would expect that a small minority of junk RNAs would be spliced and hence stabilized, and it is precisely these ncRNAs that would be under investigation.

Polyadenylation signals are also likely to be present in our putative junk RNA. These sites are quite abundant – to the extent that many of these sites are present in introns but are normally suppressed by the action of the spliceosome. These cryptic 3' cleavage sites become quite heavily used in cells with reduced U1 snRNA levels (Kaida et al., 2010). As with splicing, polyadenylation promotes mRNA stability (Akef et al., 2013); thus many junk RNAs that would be present at detectable levels are likely present by the very fact that they are polyadenylated. In summary, the fact that a given ncRNA is spliced and polyadenylated is entirely consistent with it not having any function.

Certain groups, such as the HUGO Gene Nomenclature Committee (Wright and Bruford, 2011), have defined lncRNAs as being “spliced, capped and polyadenylated,” with the clear implication that these processes are more likely to be found in functional RNAs than stable junk RNA. We disagree with this view on three counts. First, some non-functional RNAs may be processed [as described above, and by others (Ulitsky and Bartel, 2013)]. Second, many known functional ncRNAs lack all of these processing steps, one example being 7SL (Walter and Blobel, 1982; Ullu and Weiner, 1984). Third, although the goal of this nomenclature is presumably to identify functional non-coding RNAs, as is implied by the term “lncRNA,” these groups never come out and categorically state whether they consider these RNAs functional (although we assume that they do). If the term lncRNA does not imply function, then what exactly does it mean? Is it a meaningless term?

The important distinction between functional and non-functional RNAs is that processing signals are under a high level of selection pressure in the former but not in the latter. Thus,

although non-functional ncRNAs may be processed, they will likely have weak signals. Interestingly, introns in lncRNAs tend to be spliced post-transcriptionally, while those in mRNAs tend to be removed co-transcriptionally (Derrien et al., 2012; Tilgner et al., 2012). This may suggest that many lncRNA introns have weak signals due to a lack of selection pressure. Of course a paucity of hard data about the processing of random RNA polymers prevents us from making firmer conclusions. Perhaps studies such as the random genome project (Eddy, 2013) would help us identify how often spliced non-functional ncRNAs would occur purely by chance from a given stretch of DNA.

## CONSERVATION

It has been demonstrated for the last 50 years that sequence conservation is a reliable indicator of function. In line with this thinking, many commentators have declared that conservation should be the only criterion for identifying functional genomic loci (Doolittle et al., 2014). In agreement with this, we would expect that our non-functional ncRNA would accumulate mutations at a rate consistent with genetic drift. Indeed some groups have tried to restrict their definition of lncRNAs by using conservation (Guttman et al., 2009).

There are, however, various circumstances that may give the appearance of conservation. For example, the transcribed loci may also contain some conserved functional element, such as a critical TF binding site. If the region of conservation is confined to a pseudogene or TE sequences, one may simply be detecting these entities, which are typically non-functional.

The other problem with relying exclusively on sequence conservation to define functionality is that we know of many genomic loci which have sequence-independent roles. In many cases these regions serve as spacers. Thus natural selection may conserve the presence of any sequence, but not a precise sequence. For example, 5'UTRs and introns need to have a minimal length in order to promote robust translation initiation (Kozak, 1991) and splicing (Wieringa et al., 1984), respectively. Other examples include centromeric-associated repeats, which serve as sequence-independent scaffolds for kinetochore assembly (Torras-Llort et al., 2009). It is also possible that certain ncRNAs may act as a sequence-independent scaffold for protein-binding, as likely is the case of the regulation of HP1 by transcripts produced from heterochromatic regions of the *Saccharomyces cerevisiae* genome (Keller et al., 2012). Some evidence exists supporting the idea that certain eRNAs may recruit the Mediator complex to form DNA-loops, and this may require very little sequence specificity in the RNA itself (Lai et al., 2013; Andersson et al., 2014; Shibayama et al., 2014). Other times, the act of transcription, and not the resulting ncRNA, may play a role in regulating the expression of nearby genes. Presumably, the initiation of these putative regulatory transcription events are due to the activity of transcriptional start sites and/or other critical *cis*-acting elements that do display some degree of conservation. However, in practice, these promoters may be hard to identify solely by sequence analysis.

There has also been much talk about human-specific functional ncRNAs, which have generated considerable interest since they could potentially help explain differences between us and

related species (Wu et al., 2013). Although these ncRNAs would not be conserved between species, they could in principle be distinguished from non-functional ncRNA by the analysis of numerous human genomes. We would predict that non-functional ncRNA would diverge between individuals within the species at a rate comparable with genetic drift. In contrast, loci producing functional ncRNAs would be conserved. This calculation would depend on when the region in question became fixed and how fast it spread in the population. Unfortunately determining these parameters is not straight forward as it requires a large number of human genomes to be sequenced. Further, complicating the issue is the possibility that the ncRNA locus in question might be located near a genomic region that was under positive selection. The spread of neutral loci by riding on the coattails of nearby positive mutations is known as hitchhiking or draft and may be quite common (Gillespie, 2000). For these reasons, sorting lineage-specific functional ncRNA genes from non-functional ncRNAs is not trivial. Even when one turns to protein-coding genes, many of those that were once thought to be human-specific may not code for proteins after all and may indeed be non-functional (Ezkurdia et al., 2014). It is useful to keep in mind that if our ability to spot lineage-specific coding genes is problematic and fraught with error, the identification of functional human-specific ncRNAs would be even more difficult.

## CAUSAL ROLES

As stated above, certain commentators have championed selection as the primary arbiter of whether a genomic locus is functional. These same individuals have dismissed any evidence that is based on causal roles, which is defined as “the way(s) in which a component contributes to a stated capacity of some predefined system of which it is a part: what it in fact does” (Doolittle et al., 2014). The problem with defining functionality with causal roles, according to these commentators, is that this concept can be easily misappropriated. For example, a given genetic locus may be transcribed (i.e., caused the production of an RNA), but this event may not necessarily contribute to the fitness of the organism. Only if this activity was important, then natural selection would act to conserve it. Thus in the absence of any evidence of selection, regions of the genome that display some sort of causal role are *likely* not functional. This is not an absolute statement. As we point out in the previous section, certain functional RNAs may have a critical role that is sequence-independent. In other circumstances, the act of transcription, and not the ncRNA (or presumably its sequence), plays some critical role.

In light of these problems, the question clearly becomes, can a non-functional ncRNA be distinguished from one that is functional, simply on the basis of an experiment that demonstrates a “causal role”? In our opinion the answer is yes, as long as the appropriate causal role is chosen. By definition, elimination of functional ncRNAs should affect homeostasis, development or other important biological processes that would impact the fitness of the organism. In contrast, other causal role events that could potentially be associated with non-functional ncRNAs would be insufficient to qualify as evidence of functionality.

There are some problems with relying on causal roles to determine function, in that it is not always clear whether an activity

could occur by chance with an RNA with a random sequence. For example, if the overexpression of an ncRNA promotes oncogenesis, would this provide evidence of functionality? This hypothetical ncRNA could simply be sequestering an RNA binding protein that has a pro-apoptotic function, and in this instance this type of evidence would be weak. If, on the other hand, the ncRNA in question acted as a ribozyme that generated free radicals which caused DNA damage, this would then be much stronger evidence, as this activity would not be expected from a random RNA. Other evidence, such as the association of lncRNAs with certain protein complexes [e.g., the polycomb repressive complex (Khalil et al., 2009)], seems unclear. How often would such an association occur with a random non-functional nuclear RNA?

Ultimately, the ideal experiment is to determine whether the elimination of an ncRNA affects a biological process that is required for the proper development or homeostasis of the organism. This has become more feasible with the advent of CRISPR/Cas9 technology (Doudna and Charpentier, 2014). One serious problem with this approach is that the elimination of a given ncRNA may only have a small impact on the biological process being assessed and thus results in a small reduction in fitness, for example, reducing the number of offspring by 0.1%. Such small effects would be hard to detect in a laboratory setting but would be strongly selected against in the wild, and would indicate that the RNA has a function. In this case it might be beyond our current experimental abilities to obtain causal evidence for certain functional ncRNAs.

### BUILDING A CASE FOR FUNCTION

To date, projects such as ENCODE, LNCipedia and the HUGO Gene Nomenclature Committee have distinguished lncRNAs from junk RNA primarily based on expression levels and RNA-processing. In contrast, we believe that researchers need to evaluate whether any putative functional ncRNAs have properties that are beyond what one would expect from a non-functional ncRNA, given our knowledge of biochemistry, genomic evolution and current empirical data. Evidence for function can consist of expression levels that are very high (i.e., imposing a significant cost on the organism), a high degree of conservation, and/or experimental evidence that the ncRNA is required for some important biological process. Importantly, ncRNAs should be evaluated on a case-by-case basis. In the absence of sufficient evidence, a given ncRNA should be *provisionally labeled as non-functional*. Subsequently, if the ncRNA displays features/activities beyond what one would expect for the null hypothesis, then we can reclassify the ncRNA in question as being functional.

### CONCLUSION

It is clear that the human genome contains a large number of functional ncRNAs. Indeed it is likely that the list of biologically validated ncRNAs, as listed in the LncRNA Database (Quek et al., 2014), will continue to grow. As others have pointed out, even if 10% of current lncRNAs prove to be functional, this would represent a wealth of new biology. However, given our current understanding of biochemistry and evolution, it is likely that most of the RNAs generated from the low levels

of pervasive transcription, and likely a substantial number of currently annotated “lncRNAs,” are non-functional.

### ACKNOWLEDGMENTS

We would like to thank L. Moran and J. Wan for feedback on the manuscript. This work was supported by a grant from the Canadian Institutes of Health Research to Alexander F. Palazzo (FRN 102725). The funding sponsors had no role in the design of the study, in the collection, analyses, or interpretation of data, in the writing of the manuscript, and in the decision to publish the results.

### REFERENCES

- Akef, A., Zhang, H., Masuda, S., and Palazzo, A. F. (2013). Trafficking of mRNAs containing ALREX-promoting elements through nuclear speckles. *Nucleus* 4, 326–340. doi: 10.4161/nucl.26052
- Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., et al. (2014). An atlas of active enhancers across human cell types and tissues. *Nature* 507, 455–461. doi: 10.1038/nature12787
- Baker, K. E., and Parker, R. (2004). Nonsense-mediated mRNA decay: terminating erroneous gene expression. *Curr. Opin. Cell Biol.* 16, 293–299. doi: 10.1016/j.ceb.2004.03.003
- Bánfai, B., Jia, H., Khatun, J., Wood, E., Risk, B., Gundling, W. E., et al. (2012). Long noncoding RNAs are rarely translated in two human cell lines. *Genome Res.* 22, 1646–1657. doi: 10.1101/gr.134767.111
- Bazzini, A. A., Johnstone, T. G., Christiano, R., Mackowiak, S. D., Obermayer, B., Fleming, E. S., et al. (2014). Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *EMBO J.* 33, 981–993. doi: 10.1002/embj.201488411
- Biggin, M. D. (2011). Animal transcription networks as highly connected, quantitative continua. *Dev. Cell* 21, 611–626. doi: 10.1016/j.devcel.2011.09.008
- Birney, E., Stamatoyannopoulos, J. A., Dutta, A., Guigó, R., Gingeras, T. R., Margulies, E. H., et al. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447, 799–816. doi: 10.1038/nature05874
- Bissels, U., Wild, S., Tomiuk, S., Holste, A., Hafner, M., Tuschl, T., et al. (2009). Absolute quantification of microRNAs by using a universal reference. *RNA* 15, 2375–2384. doi: 10.1261/rna.1754109
- Blencowe, B. (2000). Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. *Trends Biochem. Sci.* 25, 106–110. doi: 10.1016/S0968-0004(00)01549-8
- Blobel, G., and Potter, V. R. (1967). Studies on free and membrane-bound ribosomes in rat liver. I. Distribution as related to total cellular RNA. *J. Mol. Biol.* 26, 279–292. doi: 10.1016/0022-2836(67)90297-5
- Brockdorff, N., Ashworth, A., Kay, G. F., McCabe, V. M., Norris, D. P., Cooper, P. J., et al. (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. doi: 10.1016/0092-8674(92)90519-I
- Buzin, C. H., Mann, J. R., and Singer-Sam, J. (1994). Quantitative RT-PCR assays show Xist RNA levels are low in mouse female adult tissue, embryos and embryoid bodies. *Development* 120, 3529–3536.
- Cabili, M. N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., et al. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.* 25, 1915–1927. doi: 10.1101/gad.17446611
- Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M. C., Maeda, N., et al. (2005). The transcriptional landscape of the mammalian genome. *Science* 309, 1559–1563. doi: 10.1126/science.1112014
- Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., et al. (2006). Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet.* 38, 626–635. doi: 10.1038/ng1789
- Carter, M. G., Sharov, A. A., VanBuren, V., Dudekula, D. B., Carmack, C. E., Nelson, C., et al. (2005). Transcript copy number estimation using a mouse whole-genome oligonucleotide microarray. *Genome Biol.* 6:R61. doi: 10.1186/gb-2005-6-7-r61



- Castle, J. C., Armour, C. D., Löwer, M., Haynor, D., Biery, M., Bouzek, H., et al. (2010). Digital genome-wide ncRNA expression, including SnoRNAs, across 11 human tissues using polyA-neutral amplification. *PLoS ONE* 5:e11779. doi: 10.1371/journal.pone.0011779
- Cech, T. R., and Steitz, J. A. (2014). The noncoding RNA revolution—trashing old rules to forge new ones. *Cell* 157, 77–94. doi: 10.1016/j.cell.2014.03.008
- Charlesworth, B. (2009). Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. *Nat. Rev. Genet.* 10, 195–205. doi: 10.1038/nrg2526
- Chekanova, J. A., Gregory, B. D., Reverdatto, S. V., Chen, H., Kumar, R., Hooker, T., et al. (2007). Genome-wide high-resolution mapping of exosome substrates reveals hidden features in the *Arabidopsis* transcriptome. *Cell* 131, 1340–1353. doi: 10.1016/j.cell.2007.10.056
- Cheung, V., Chua, G., Batada, N. N., Landry, C. R., Michnick, S. W., Hughes, T. R., et al. (2008). Chromatin- and transcription-related factors repress transcription from within coding regions throughout the *Saccharomyces cerevisiae* genome. *PLoS Biol.* 6:e277. doi: 10.1371/journal.pbio.0060277
- Chi, E. Y., Krishnan, S., Randolph, T. W., and Carpenter, J. F. (2003). Physical stability of proteins in aqueous solution: mechanism and driving forces in non-native protein aggregation. *Pharm. Res.* 20, 1325–1336. doi: 10.1023/A:1025771421906
- Cooper, G. M. (2000). *The Cell: A Molecular Approach*. Washington, DC; Sunderland, MA: ASM Press; Sinauer Associates.
- Davis, C. A., and Ares, M. Jr. (2006). Accumulation of unstable promoter-associated transcripts upon loss of the nuclear exosome subunit Rrp6p in *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci. U.S.A.* 103, 3262–3267. doi: 10.1073/pnas.0507783103
- Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., et al. (2012). The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res.* 22, 1775–1789. doi: 10.1101/gr.132159.111
- de Souza, F. S. J., Franchini, L. F., and Rubinstein, M. (2013). Exaptation of transposable elements into novel cis-regulatory elements: is the evidence always strong? *Mol. Biol. Evol.* 30, 1239–1251. doi: 10.1093/molbev/mst045
- Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., et al. (2012). Landscape of transcription in human cells. *Nature* 489, 101–108. doi: 10.1038/nature11233
- Doolittle, W. F. (2013). Is junk DNA bunk? A critique of ENCODE. *Proc. Natl. Acad. Sci. U.S.A.* 110, 5294–5300. doi: 10.1073/pnas.1221376110
- Doolittle, W. F., Brunet, T. D. P., Linnquist, S., and Gregory, T. R. (2014). Distinguishing between “function” and “effect” in genome biology. *Genome Biol. Evol.* 6, 1234–1237. doi: 10.1093/gbe/evu098
- Doudna, J. A., and Charpentier, E. (2014). Genome editing. The new frontier of genome engineering with CRISPR-Cas9. *Science* 346:1258096. doi: 10.1126/science.1258096
- Duncan, R., and Hershey, J. W. (1983). Identification and quantitation of levels of protein synthesis initiation factors in crude HeLa cell lysates by two-dimensional polyacrylamide gel electrophoresis. *J. Biol. Chem.* 258, 7228–7235.
- Ecker, J. R., Bickmore, W. A., Barroso, I., Pritchard, J. K., Gilad, Y., and Segal, E. (2012). Genomics: ENCODE explained. *Nature* 489, 52–55. doi: 10.1038/489052a
- Eddy, S. R. (2012). The C-value paradox, junk DNA and ENCODE. *Curr. Biol.* 22, R898–R899. doi: 10.1016/j.cub.2012.10.002
- Eddy, S. R. (2013). The ENCODE project: missteps overshadowing a success. *Curr. Biol.* 23, R259–R261. doi: 10.1016/j.cub.2013.03.023
- Elliott, T. A., Linnquist, S., and Gregory, T. R. (2014). Conceptual and empirical challenges of ascribing functions to transposable elements. *Am. Nat.* 184, 14–24. doi: 10.1086/676588
- ENCODE Project Consortium, Bernstein, B. E., Birney, E., Dunham, I., Green, E. D., Gunter, C., et al. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74. doi: 10.1038/nature11247
- Ezkurdia, I., Juan, D., Rodriguez, J. M., Frankish, A., Diekhans, M., Harrow, J., et al. (2014). Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes. *Hum. Mol. Genet.* 23, 5866–5878. doi: 10.1093/hmg/ddu309
- Gillespie, J. H. (2000). Genetic drift in an infinite population. The pseudohitchhiking model. *Genetics* 155, 909–919.
- Gould, S. J., and Lewontin, R. C. (1979). The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme. *Proc. R. Soc. Lond. Biol. Sci.* 205, 581–598. doi: 10.1098/rspb.1979.0086
- Graur, D., Zheng, Y., Price, N., Azevedo, R. B. R., Zufall, R. A., and Elhaik, E. (2013). On the immortality of television sets: “function” in the human genome according to the evolution-free gospel of ENCODE. *Genome Biol. Evol.* 5, 578–590. doi: 10.1093/gbe/evt028
- Guo, J. U., Agarwal, V., Guo, H., and Bartel, D. P. (2014). Expanded identification and characterization of mammalian circular RNAs. *Genome Biol.* 15:409. doi: 10.1186/s13059-014-0409-z
- Guttman, M., Amit, I., Garber, M., French, C., Lin, M. F., Feldser, D., et al. (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* 458, 223–227. doi: 10.1038/nature07672
- Guttman, M., Russell, P., Ingolia, N. T., Weissman, J. S., and Lander, E. S. (2013). Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. *Cell* 154, 240–251. doi: 10.1016/j.cell.2013.06.009
- Hanguer, M. J., Vaughn, I. W., and McManus, M. T. (2013). Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs. *PLoS Genet.* 9:e1003569. doi: 10.1371/journal.pgen.1003569
- Hastie, N. D., and Bishop, J. O. (1976). The expression of three abundance classes of messenger RNA in mouse tissues. *Cell* 9, 761–774. doi: 10.1016/0092-8674(76)90139-2
- Huang, Y.-F., and Niu, D.-K. (2008). Evidence against the energetic cost hypothesis for the short introns in highly expressed genes. *BMC Evol. Biol.* 8:154. doi: 10.1186/1471-2148-8-154
- Ingolia, N. T., Brar, G. A., Stern-Ginossar, N., Harris, M. S., Talhouarne, G. J. S., Jackson, S. E., et al. (2014). Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. *Cell Rep.* 8, 1365–1379. doi: 10.1016/j.celrep.2014.07.045
- 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. doi: 10.1016/j.cell.2011.10.002
- Jeck, W. R., Sorrentino, J. A., Wang, K., Slevin, M. K., Burd, C. E., Liu, J., et al. (2013). Circular RNAs are abundant, conserved, and associated with ALU repeats. *RNA* 19, 141–157. doi: 10.1261/rna.035667.112
- 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. doi: 10.1261/rna.1951310
- Kaida, D., Berg, M. G., Younis, I., Kasim, M., Singh, L. N., Wan, L., et al. (2010). U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. *Nature* 468, 664–668. doi: 10.1038/nature09479
- Kapusta, A., and Feschotte, C. (2014). Volatile evolution of long noncoding RNA repertoires: mechanisms and biological implications. *Trends Genet.* 30, 439–452. doi: 10.1016/j.tig.2014.08.004
- Keller, C., Adaixo, R., Stunnenberg, R., Woolcock, K. J., Hiller, S., and Bühler, M. (2012). HPI5wi6 mediates the recognition and destruction of heterochromatic RNA transcripts. *Mol. Cell* 47, 215–227. doi: 10.1016/j.molcel.2012.05.009
- Khalil, A. M., Guttman, M., Huarte, M., Garber, M., Raj, A., Rivea Morales, D., et al. (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. doi: 10.1073/pnas.0904715106
- Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature* 217, 624–626. doi: 10.1038/217624a0
- Kimura, M. (1984). *The Neutral Theory of Molecular Evolution*. Cambridge: Cambridge University Press.
- King, J. L., and Jukes, T. H. (1969). Non-Darwinian evolution. *Science* 164, 788–798. doi: 10.1126/science.164.3881.788
- Kiss, T., and Filipowicz, W. (1992). Evidence against a mitochondrial location of the 7-2-MRP RNA in mammalian cells. *Cell* 70, 11–16. doi: 10.1016/0092-8674(92)90528-K
- Kozak, M. (1991). A short leader sequence impairs the fidelity of initiation by eukaryotic ribosomes. *Gene Expr.* 1, 111–115.
- Lai, F., Orom, U. A., Cesaroni, M., Beringer, M., Taatjes, D. J., Blobel, G. A., et al. (2013). Activating RNAs associate with Mediator to enhance chromatin architecture and transcription. *Nature* 494, 497–501. doi: 10.1038/nature11884

- Li, X., MacArthur, S., Bourgon, R., Nix, D., Pollard, D. A., Iyer, V. N., et al. (2008). Transcription factors bind thousands of active and inactive regions in the *Drosophila* blastoderm. *PLoS Biol.* 6:e27. doi: 10.1371/journal.pbio.0060027
- Lickwar, C. R., Mueller, F., Hanlon, S. E., McNally, J. G., and Lieb, J. D. (2012). Genome-wide protein-DNA binding dynamics suggest a molecular clutch for transcription factor function. *Nature* 484, 251–255. doi: 10.1038/nature10985
- Lu, S., and Cullen, B. R. (2003). Analysis of the stimulatory effect of splicing on mRNA production and utilization in mammalian cells. *RNA* 9, 618–630. doi: 10.1261/rna.5260303
- Lynch, M. (2007). *The Origins of Genome Architecture*. Sunderland MA: Sinauer Associates.
- Magny, E. G., Pueyo, J. I., Pearl, F. M. G., Cespedes, M. A., Niven, J. E., Bishop, S. A., et al. (2013). Conserved regulation of cardiac calcium uptake by peptides encoded in small open reading frames. *Science* 341, 1116–1120. doi: 10.1126/science.1238802
- Managadze, D., Lobkovsky, A. E., Wolf, Y. I., Shabalina, S. A., Rogozin, I. B., and Koonin, E. V. (2013). The vast, conserved mammalian lincRNome. *PLoS Comput. Biol.* 9:e1002917. doi: 10.1371/journal.pcbi.1002917
- Managadze, D., Rogozin, I. B., Chernikova, D., Shabalina, S. A., and Koonin, E. V. (2011). Negative correlation between expression level and evolutionary rate of long intergenic noncoding RNAs. *Genome Biol. Evol.* 3, 1390–1404. doi: 10.1093/gbe/evr116
- Martin, W., and Koonin, E. V. (2006). Introns and the origin of nucleus-cytosol compartmentalization. *Nature* 440, 41–45. doi: 10.1038/nature04531
- Mattick, J. S., and Dinger, M. E. (2013). The extent of functionality in the human genome. *HUGO J.* 7:2. doi: 10.1186/1877-6566-7-2
- 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. doi: 10.1016/j.tig.2009.11.002
- Memczak, S., Jens, M., Elefsinioti, A., Torti, F., Krueger, J., Rybak, A., et al. (2013). Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature* 495, 333–338. doi: 10.1038/nature11928
- Menet, J. S., Rodriguez, J., Abruzzi, K. C., and Rosbash, M. (2012). Nascent-Seq reveals novel features of mouse circadian transcriptional regulation. *Elife* 1:e00011. doi: 10.7554/eLife.00011
- Milligan, L., Decourty, L., Saveanu, C., Rappsilber, J., Ceulemans, H., Jacquier, A., et al. (2008). A yeast exosome cofactor, Mpp6, functions in RNA surveillance and in the degradation of noncoding RNA transcripts. *Mol. Cell. Biol.* 28, 5446–5457. doi: 10.1128/MCB.00463-08
- Miura, P., Shenker, S., Andreu-Agullo, C., Westholm, J. O., and Lai, E. C. (2013). Widespread and extensive lengthening of 3'UTRs in the mammalian brain. *Genome Res.* 23, 812–825. doi: 10.1101/gr.146886.112
- Moazed, D. (2009). Small RNAs in transcriptional gene silencing and genome defence. *Nature* 457, 413–420. doi: 10.1038/nature07756
- 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. doi: 10.1038/nmeth.1226
- Neil, H., Malabat, C., d'Aubenton-Carafa, Y., Xu, Z., Steinmetz, L. M., and Jacquier, A. (2009). Widespread bidirectional promoters are the major source of cryptic transcripts in yeast. *Nature* 457, 1038–1042. doi: 10.1038/nature07747
- Niu, D.-K., and Jiang, L. (2013). Can ENCODE tell us how much junk DNA we carry in our genome? *Biochem. Biophys. Res. Commun.* 430, 1340–1343. doi: 10.1016/j.bbrc.2012.12.074
- Ohta, T. (1973). Slightly deleterious mutant substitutions in evolution. *Nature* 246, 96–98. doi: 10.1038/246096a0
- Palazzo, A. F., and Akef, A. (2012). Nuclear export as a key arbiter of “mRNA identity” in eukaryotes. *Biochim. Biophys. Acta* 1819, 566–577. doi: 10.1016/j.bbagr.2011.12.012
- Palazzo, A. F., and Gregory, T. R. (2014). The case for junk DNA. *PLoS Genet.* 10:e1004351. doi: 10.1371/journal.pgen.1004351
- Palazzo, A., Mahadevan, K., and Tarnawsky, S. (2013). ALREX-elements and introns: two identity elements that promote mRNA nuclear export. *Wiley Interdiscip. Rev. RNA* 4, 523–533. doi: 10.1002/wrna.1176
- Palazzo, A. F., Springer, M., Shibata, Y., Lee, C.-S., Dias, A. P., and Rapoport, T. A. (2007). The signal sequence coding region promotes nuclear export of mRNA. *PLoS Biol.* 5:e322. doi: 10.1371/journal.pbio.0050322
- Paris, M., Kaplan, T., Li, X. Y., Villalta, J. E., Lott, S. E., and Eisen, M. B. (2013). Extensive divergence of transcription factor binding in *Drosophila* embryos with highly conserved gene expression. *PLoS Genet.* 9:e1003748. doi: 10.1371/journal.pgen.1003748
- Pennisi, E. (2012). Genomics. ENCODE project writes eulogy for junk DNA. *Science* 337, 1159, 1161. doi: 10.1126/science.337.6099.1159
- Ponting, C. P., Oliver, P. L., and Reik, W. (2009). Evolution and functions of long noncoding RNAs. *Cell* 136, 629–641. doi: 10.1016/j.cell.2009.02.006
- Preker, P., Nielsen, J., Kammler, S., Lykke-Andersen, S., Christensen, M. S., Mapendano, C. K., et al. (2008). RNA exosome depletion reveals transcription upstream of active human promoters. *Science* 322, 1851–1854. doi: 10.1126/science.1164096
- Quek, X. C., Thomson, D. W., Maag, J. L. V., Bartonicek, N., Signal, B., Clark, M. B., et al. (2014). lincRNAdb v2.0: expanding the reference database for functional long noncoding RNAs. *Nucleic Acids Res.* doi: 10.1093/nar/gku988 [Epub ahead of print].
- Ramsköld, D., Wang, E. T., Burge, C. B., and Sandberg, R. (2009). An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS Comput. Biol.* 5:e1000598. doi: 10.1371/journal.pcbi.1000598
- Raue, U., Oellerer, S., and Rospert, S. (2007). Association of protein biogenesis factors at the yeast ribosomal tunnel exit is affected by the translational status and nascent polypeptide sequence. *J. Biol. Chem.* 282, 7809–7816. doi: 10.1074/jbc.M611436200
- Rinn, J., and Guttman, M. (2014). RNA Function. RNA and dynamic nuclear organization. *Science* 345, 1240–1241. doi: 10.1126/science.1252966
- Salditt-Georgieff, M., Jelinek, W., Darnell, J. E., Furuichi, Y., Morgan, M., and Shatkin, A. (1976). Methyl labeling of HeLa cell hnRNA: a comparison with mRNA. *Cell* 7, 227–237. doi: 10.1016/0092-8674(76)90022-2
- Salzman, J., Gawad, C., Wang, P. L., Lacayo, N., and Brown, P. O. (2012). Circular RNAs are the predominant transcript isoform from hundreds of human genes in diverse cell types. *PLoS ONE* 7:e30733. doi: 10.1371/journal.pone.0030733
- Sarkar, S. (2014). The genomic challenge to adaptationism. *Br. J. Philos. Sci.* doi: 10.1093/bjps/axu002 [Epub ahead of print].
- Shepard, S., McCreary, M., and Fedorov, A. (2009). The peculiarities of large intron splicing in animals. *PLoS ONE* 4:e7853. doi: 10.1371/journal.pone.0007853
- Shibayama, Y., Fanucchi, S., Magagula, L., and Mhlanga, M. M. (2014). lncRNA and gene looping: what's the connection? *Transcription* doi: 10.4161/trns.28658 [Epub ahead of print].
- Stark, B. C., Kole, R., Bowman, E. J., and Altman, S. (1978). Ribonuclease P: an enzyme with an essential RNA component. *Proc. Natl. Acad. Sci. U.S.A.* 75, 3717–3721. doi: 10.1073/pnas.75.8.3717
- Stewart, A. J., Hannehalli, S., and Plotkin, J. B. (2012). Why transcription factor binding sites are ten nucleotides long. *Genetics* 192, 973–985. doi: 10.1534/genetics.112.143370
- Struhl, K. (2007). Transcriptional noise and the fidelity of initiation by RNA polymerase II. *Nat. Struct. Mol. Biol.* 14, 103–105. doi: 10.1038/nsmb0207-103
- Tan-Wong, S. M., Zaugg, J. B., Camblong, J., Xu, Z., Zhang, D. W., Mischo, H. E., et al. (2012). Gene loops enhance transcriptional directionality. *Science* 338, 671–675. doi: 10.1126/science.1224350
- Thiebaut, M., Kisseleva-Romanova, E., Rougemaille, M., Boulay, J., and Libri, D. (2006). Transcription termination and nuclear degradation of cryptic unstable transcripts: a role for the nrd1-nab3 pathway in genome surveillance. *Mol. Cell* 23, 853–864. doi: 10.1016/j.molcel.2006.07.029
- Tilgner, H., Knowles, D. G., Johnson, R., Davis, C. A., Chakraborty, S., Djebali, S., et al. (2012). Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. *Genome Res.* 22, 1616–1625. doi: 10.1101/gr.134445.111
- Tisseur, M., Kwapisz, M., and Morillon, A. (2011). Pervasive transcription – lessons from yeast. *Biochimie* 93, 1889–1896. doi: 10.1016/j.biochi.2011.07.001
- Torras-Llort, M., Moreno-Moreno, O., and Azorin, F. (2009). Focus on the centre: the role of chromatin on the regulation of centromere identity and function. *EMBO J.* 28, 2337–2348. doi: 10.1038/emboj.2009.174
- Ulitksy, I., and Bartel, D. P. (2013). lincRNAs: genomics, evolution, and mechanisms. *Cell* 154, 26–46. doi: 10.1016/j.cell.2013.06.020
- Ullu, E., and Weiner, A. M. (1984). Human genes and pseudogenes for the 7SL RNA component of signal recognition particle. *EMBO J.* 3, 3303–3310.

- 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
- Vasiljeva, L., Kim, M., Terzi, N., Soares, L. M., and Buratowski, S. (2008). Transcription termination and RNA degradation contribute to silencing of RNA polymerase II transcription within heterochromatin. *Mol. Cell* 29, 313–323. doi: 10.1016/j.molcel.2008.01.011
- Volders, P.-J., Helsen, K., Wang, X., Menten, B., Martens, L., Gevaert, K., et al. (2013). LNCipedia: a database for annotated human lncRNA transcript sequences and structures. *Nucleic Acids Res.* 41, D246–D251. doi: 10.1093/nar/gks915
- Waldron, C., and Lacroute, F. (1975). Effect of growth rate on the amounts of ribosomal and transfer ribonucleic acids in yeast. *J. Bacteriol.* 122, 855–865.
- Walter, P., and Blobel, G. (1982). Signal recognition particle contains a 7S RNA essential for protein translocation across the endoplasmic reticulum. *Nature* 299, 691–698. doi: 10.1038/299691a0
- Wang, K. C., and Chang, H. Y. (2011). Molecular mechanisms of long noncoding RNAs. *Mol. Cell* 43, 904–914. doi: 10.1016/j.molcel.2011.08.018
- Wang, M., Zhang, P., Shu, Y., Yuan, F., Zhang, Y., Zhou, Y., et al. (2014). Alternative splicing at GYNNGY 5' splice sites: more noise, less regulation. *Nucleic Acids Res.* 42, 13969–13980. doi: 10.1093/nar/gku1253
- Wang, Y., Ma, M., Xiao, X., and Wang, Z. (2012). Intronic splicing enhancers, cognate splicing factors and context-dependent regulation rules. *Nat. Struct. Mol. Biol.* 19, 1044–1052. doi: 10.1038/nsmb.2377
- West, M. W., Wang, W., Patterson, J., Mancias, J. D., Beasley, J. R., and Hecht, M. H. (1999). De novo amyloid proteins from designed combinatorial libraries. *Proc. Natl. Acad. Sci. U.S.A.* 96, 11211–11216. doi: 10.1073/pnas.96.20.11211
- White, M. A., Myers, C. A., Corbo, J. C., and Cohen, B. A. (2013). Massively parallel in vivo enhancer assay reveals that highly local features determine the cis-regulatory function of ChIP-seq peaks. *Proc. Natl. Acad. Sci. U.S.A.* 110, 11952–11957. doi: 10.1073/pnas.1307449110
- Wieringa, B., Hofer, E., and Weissmann, C. (1984). A minimal intron length but no specific internal sequence is required for splicing the large rabbit beta-globin intron. *Cell* 37, 915–925. doi: 10.1016/0092-8674(84)90426-4
- Wilusz, J. E., and Sharp, P. A. (2013). Molecular biology. A circuitous route to noncoding RNA. *Science* 340, 440–441. doi: 10.1126/science.1238522
- Wilusz, J. E., Sunwoo, H., and Spector, D. L. (2009). Long noncoding RNAs: functional surprises from the RNA world. *Genes Dev.* 23, 1494–1504. doi: 10.1101/gad.1800909
- Wolf, S. F., and Schlessinger, D. (1977). Nuclear metabolism of ribosomal RNA in growing, methionine-limited, and ethionine-treated HeLa cells. *Biochemistry (Mosc.)* 16, 2783–2791. doi: 10.1021/bi00631a031
- Wright, M. W., and Bruford, E. A. (2011). Naming “junk”: human non-protein coding RNA (ncRNA) gene nomenclature. *Hum. Genomics* 5:90. doi: 10.1186/1479-7364-5-2-90
- Wu, P., Zuo, X., Deng, H., Liu, X., Liu, L., and Ji, A. (2013). Roles of long noncoding RNAs in brain development, functional diversification and neurodegenerative diseases. *Brain Res. Bull.* 97, 69–80. doi: 10.1016/j.brainresbull.2013.06.001
- Wyers, F., Rougemaille, M., Badis, G., Rousselle, J.-C., Dufour, M.-E., Boulay, J., et al. (2005). Cryptic pol II transcripts are degraded by a nuclear quality control pathway involving a new poly(A) polymerase. *Cell* 121, 725–737. doi: 10.1016/j.cell.2005.04.030
- Xu, G., and Zhang, J. (2014). Human coding RNA editing is generally nonadaptive. *Proc. Natl. Acad. Sci. U.S.A.* 111, 3769–3774. doi: 10.1073/pnas.1321745111
- Xu, Z., Wei, W., Gagneur, J., Perocchi, F., Clauder-Münster, S., Camblong, J., et al. (2009). Bidirectional promoters generate pervasive transcription in yeast. *Nature* 457, 1033–1037. doi: 10.1038/nature07728
- Yang, V. W., Lerner, M. R., Steitz, J. A., and Flint, S. J. (1981). A small nuclear ribonucleoprotein is required for splicing of adenoviral early RNA sequences. *Proc. Natl. Acad. Sci. U.S.A.* 78, 1371–1375. doi: 10.1073/pnas.78.3.1371

**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: 20 November 2014; accepted: 06 January 2015; published online: 26 January 2015.

Citation: Palazzo AF and Lee ES (2015) Non-coding RNA: what is functional and what is junk? *Front. Genet.* 6:2. doi: 10.3389/fgene.2015.00002

This article was submitted to *Genetics of Aging*, a section of the journal *Frontiers in Genetics*.

Copyright © 2015 Palazzo and Lee. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.