Extreme Information: Biocomplexity of Interlocking Genome Languages

Jeffrey P. Tomkins*

Abstract

People most often think of the genome as containing only the embedded protein-coding information carried in the DNA of chromosomes. However, there are a variety of other codes and language systems active in the genome that are only now beginning to be deciphered. This paper will discuss the amazing internetworked biocomplexity of these language systems that interactively control the way the genome functions. The systems that will be discussed are gene structure complexities, RNA transcript splicing codes, the microRNA binding code, circular RNAs, dual-use codons, antisense transcripts, and epigenetic language systems. The now debunked myth of junk DNA will also be briefly addressed in light of the ENCODE project and new research in genome-wide COT-1 DNA functionality. The interworking and interdependence of these complex and dynamic language systems unequivocally points towards an omnipotent and wise Creator.

Introduction

When people contemplate the language of the genome, they typically consider only the information encoded in the long strings of letters that symbolize the four nucleobases consisting of adenine (A), cytosine (C), thymine (T), and guanine (G). Indeed, even at this level, the complexity of DNA language systems is remarkable and dynamic.

Researchers have shown that the linear code in DNA/RNA responsible for synthesizing proteins engages all of the features required for a model of universal information and language/code (Gitt, 2011). These four distinguishing attributes include (1) cosyntics—an abstract code with syntactic rules; (2) semantics—which provides meaning (e.g., triplets of 3 bases in codons corresponding to amino acids); (3) pragmatics—the information expresses specific calls to action (e.g., stop/start sites, splice sites, protein-binding sites, processing signals, cellular address sites, etc.); and (4) apobetics—the information encodes a final purpose to be achieved (e.g., patterning, function, and replication of cells, organs, and whole organisms).

The universal information contained in the human genome consists of about 3 billion DNA letters (base pairs) in just one genome equivalent—6 billion when you consider both the maternal and paternal sets of chromosomes. Because DNA is a double-stranded molecule, the bases on one strand predict those on the opposite strand due to complementary base pairing (A pairs with T and C pairs with G). Despite this constraint, different information is encoded on both strands, running in

* Jeffrey P. Tomkins, Institute for Creation Research, Dallas, TX, jtomkins@icr.org

Accepted for publication January 13, 2015

opposite directions. Thus, you could expand the total amount of actual linear information in the human genome to about 12 billion bases. And within these letters are encoded a dizzying array of regulatory features, RNA products, and proteins.

Just as a computer system is composed of multiple software programs encoded by a wide variety of programming languages—all interacting together with the hardware of the system—so is the genome, only at a much higher level of complexity that is only beginning to be understood. This review will briefly highlight some of the better-understood language systems that interactively operate in the genome, demonstrating the many interlocking universal information systems.

The Unit of DNA Language—The Gene?

Perhaps the best place to start in explaining the diversity of complex information found in the genome is to begin the discussion of what is commonly referred to as genes.

According to the earliest genetic ideas that prevailed during the early part of the twentieth century, the classical view was that a gene was considered to be the smallest indivisible unit of transmission, recombination, mutation, and function, with all of these criteria being interdependent (Portin, 2002). For example, you cannot observe genetic recombination without transmission, and you cannot observe transmission without function (based on a phenotype). In light of these ideas, the term "gene" was introduced by Johannsen, who desired that it be free of any physical or chemical constraints and treated as an intact heritable unit that could be analyzed statistically (Johannsen, 1909). For several reviews on the history of the "classical view" of the gene beginning with Mendel's work, see Portin (2002) and Gerstein et al. (2007).

No sooner than the classical view that genes were distinct, single-unit heritable entities on chromosomes had matured in its paradigm in the late 1930s, the concept rapidly began to break down. This started with the discoveries of intragenic recombination in Drosophila in the early 1940s (Lewis, 1941; Oliver, 1940). A wide variety of other studies in a diversity of eukaryotes soon followed (Gerstein et al., 2007; Portin, 2002). Thus, the idea that genes were distinct, indivisible units was debunked using classical genetic tools of study prior to the onset of the use of advanced molecular technologies.

Interestingly, at the same time that elaborate genetic studies in eukaryotes were showing that genes did not always exist as distinct, single-unit entities, a wealth of biochemical studies in the late 1940s through the 1960s using bacteriophage and *E. coli* seemed to indicate that one gene controlled the synthesis of one messenger RNA molecule, which in turn encoded the synthesis of one polypeptide, an idea that some have termed the "neoclassical view" of the gene (Portin, 2002). In the genomics community, this now archaic paradigm is most often termed the "protein-centric" view of the gene (Gerstein et al., 2007).

With the advent of the use of new tools in molecular biology, the neoclassical, or protein-centric, view of the gene also broke down rapidly about as soon as it came to fruition. This was initially driven by the discovery in the late 1960s and early 1970s of restriction enzymes that cut DNA at specific sites (Portin, 2002). These new tools subsequently allowed for DNA segments to be dissected and cloned (Cohen et al., 1973), mapped (Southern, 1975), and eventually sequenced (Sanger et al., 1977). As a consequence of many early studies in the 1970s and 1980s using these tools, it slowly began to be realized that the protein-centric concept of the gene was grossly oversimplified. In fact, according to the most recent discoveries, the boundaries of what can be called a distinct gene unit, especially in eukaryotes, are becoming increasingly hard to define, along with a gene's complete set of known functions (Gerstein et al., 2007; Portin, 2009).

Immediately following the first drafts of the human genome (Lander et al., 2001; Venter et al., 2001), large-scale funding was directed toward deciphering the functional information contained therein in a project termed ENCODE (Encyclopedia of DNA Elements), which relied heavily on studying RNA transcripts produced in a variety of cell lines and tissues (Birney et al., 2007). In their first report, which targeted a test sample of just 1% of the total genome, the researchers stated, "First, our studies provide convincing evidence that the genome is pervasively transcribed, such that the majority of its bases can be found in primary transcripts, including non-protein-coding transcripts, and those that extensively overlap one another" (Birney et al., 2007, p. 799). In the second tier of ENCODE-related research, which targeted the entire human genome along with forays into other animal genomes using highly advanced high-throughput genomic technologies, it was unequivocally shown that entire genomes are a continuum of pervasive and overlapping transcription (Djebali et al., 2012a; Dunham et al., 2012; Liu et al., 2013).

These recent discoveries have also revealed that genes are not like single entities at all but instead are a mixture of genes within genes (nested genes) and genes that overlap each other (Clark et al., 2013; Portin, 2009; Sanna et al., 2008). See Figure 1 for a depiction of the various gene structures discussed in this section.

A nested gene is defined as having its entire coding sequence within the chromosomal region demarcated by the start and stop codons of a larger gene (Figure 1B). It should also be noted that nested genes are distinct from alternatively

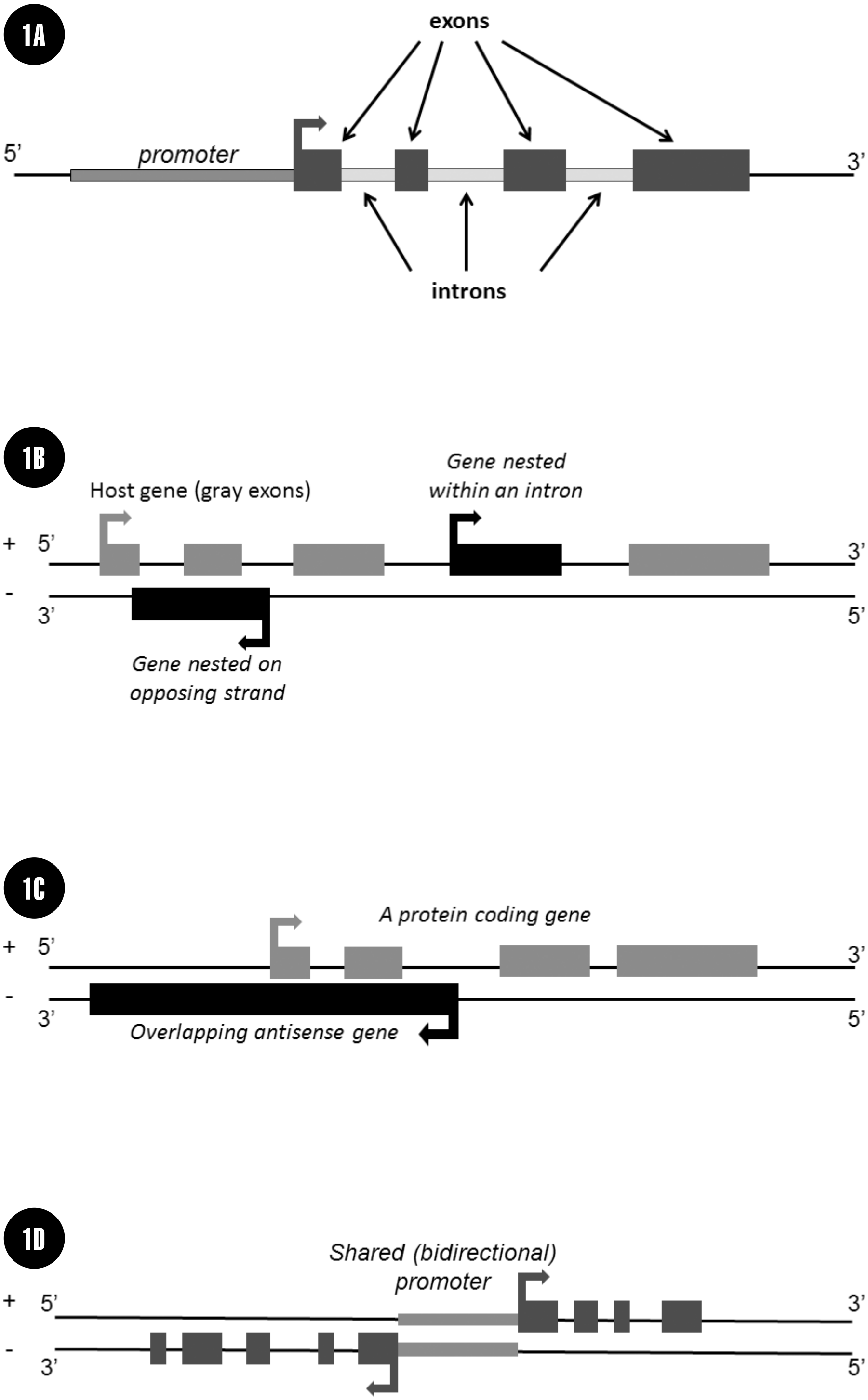
1A
exons
5'
promoter
3'
introns
1B
Host gene (gray exons)
Gene nested within an intron
+ 5'
3'
- 3'
5'
Gene nested on opposing strand
1C
A protein coding gene
+ 5'
3'
- 3'
5'
Overlapping antisense gene
1D
Shared (bidirectional) promoter
+ 5'
3'
- 3'
5'

Figure 1. (A) The basic structure of a eukaryotic gene representing the "genes in pieces" concept. Arrow in first exon represents direction of transcription. (B) Depiction of two types of nested genes—one running in the same direction as the host gene within an intron, and one on the opposing strand. (C) Depiction of an overlapping gene—a protein coding gene and a corresponding antisense gene on the opposing strand. (D) Configuration of two neighboring genes on separate strands sharing the same bidirectional promoter.

spliced transcripts (discussed below) in that the nested gene and its host gene do not share transcriptional start sites. In most cases, the start site and promoter of the nested gene are located inside one of the introns of the host gene and are encoded on the opposite strand. In other cases the entire nested gene is situated in the same strand orientation of its host gene—typically inside an intron (Kumar, 2009; Lee and Chang, 2013). In fact, a recent study in Drosophila has shown that nearly 10% of its genes are organized in nested structures and that despite their integrated configuration, nested genes were less likely to display correlated expression and biological function than were neighboring non-overlapping genes (Lee and Chang, 2013).

Overlapping genes are now known to be common in both prokaryotes and eukaryotes (Sanna et al., 2008; Veeramachaneni et al., 2004). They are defined as two separate, distinct genes that overlap each other either on the same strand or, more commonly, on the opposite strand (Figure 1C). In the case where they are on opposite strands, they are now commonly referred to as sense-antisense pairs (Wood et al., 2013). In a study of 13,484 genes shared between human and mouse, about 10% of the genes were overlapping—mostly on different strands (Sanna et al., 2008). However, in a more recent study, it was shown that while human and mouse shared similar sections of certain genes (called *homologous*), the gene landscapes and sense-antisense pair configurations were completely different for these seemingly evolutionarily conserved regions (Wood et al., 2013).

Another variant of shared language between genes that are close to each other in the genome is revealed in the finding that some regulatory control regions, called *promoters*, can be shared, often in a coregulatory fashion (Figure 1D). The presence of pervasive bidirectional promoters was first documented after the initial analysis of the first draft of the human genome, which showed the positioning of many genes arranged in a head-to-head divergent configuration—typically in the opposite strand configuration (Adachi and Lieber, 2002). Now it is known that greater than 10% of the total number of genes are controlled by bidirectional promoters in the human genome (Wang et al., 2013). One of the key findings with bidirectional promoters in humans is that their genes not only share similar categories of cell process involvement, but that they are highly enriched in functional classes important to DNA repair and genome maintenance (Trinklein et al., 2004; Wakano et al., 2012). This is why aberrations in these regulatory regions have been associated with cancer (Wang et al., 2013).

Yet one more characteristic blurring gene boundaries is the fact that some of the regulatory sequences controlling a gene can be located inside neighboring genes; researchers have determined that genes dynamically interact with each other in "gene neighborhoods" much more than previously believed (Lemay et al., 2012; Portin, 2009). One of the key regulatory features controlling genes is called *enhancer elements*. These are short, 50 to 150 base motifs that bind regulatory proteins and can be up to a million bases away from the gene they regulate (Dickel et al., 2013; Gerstein et al., 2012). Enhancers work as distinct modules that drive gene expression at particular time points and in particular tissues by integrating inputs (e.g., transcription factors, transcription activators, cell type information) in elaborate three-dimensional looping of the chromosomes, connecting them spatially in the genome to the transcriptional apparatus in the promoter of the gene (Dickel et al., 2013; Sakabe et al., 2012; see Figure 2). In fact, the arrangement of enhancers around a gene and the binding of transcription factors to them is itself a type of language referred to as the enhancer code (Weatheritt and Babu, 2013).

And finally, the most damaging concept for the protein-centric view of the gene is that compared to protein-coding genes, over twice as many genes in the human genome produce functional long, noncoding RNAs termed "lncRNA" (Hangauer et al., 2013; Managadze et al., 2013), and approximately two thirds of RNA binding proteins associate with nonprotein-coding transcripts (Gerstberger et al., 2014). These lncRNAs have the same types of promoters as protein-coding genes and often share bidirectional promoters with them, being situated on opposing strands (Sigova et al., 2013). In addition, lncRNAs also have the same type of intron and exon structures as protein-coding genes and undergo the same types of capping, splicing, and three-prime tail modifications as protein-coding genes (Rinn and Chang, 2012; Tomkins, 2014). Amazingly, these lncRNA genes are turning out to be the key factors in what controls and regulates protein-coding genes and in what specifically characterizes the transcriptomes of different kinds of cells, organs, and tissues (Clark et al., 2013; Ulitsky and Bartel, 2013). Characterization of lncRNA between different taxa reveals strong patterns of sequence discontinuity, which is proving to be an intractable problem for the evolutionary paradigm (Necsulea et al., 2014; Tomkins, 2014; Washietl et al., 2014).

The Splicing Code

In bacteria, the best-studied organisms in the early days of molecular biology, a typical gene corresponded to a single protein, although many bacterial genes that are functionally related are often linked together and transcribed under the control of a single regulatory region (Osbourn and Field, 2009). However, as discoveries increased in multicellular plants and animals, it was found that their "genes were in pieces," according to the memorable phrase of Walter Gilbert, who first popularized the amazing gene structure discoveries occurring at the time (Gilbert, 1978). This phrase referred to the fact that in the genomes of plants, animals, protists, and fungi, their protein-coding regions (called *exons*) are interrupted by noncoding sequences called *introns* (Figure 1A). In fact, the mystery of how this whole process of genes in pieces allegedly evolved is still one of the leading problems in explaining the evolution of genomes from bacteria and archaea that typically lack these features (Koonin et al., 2013).

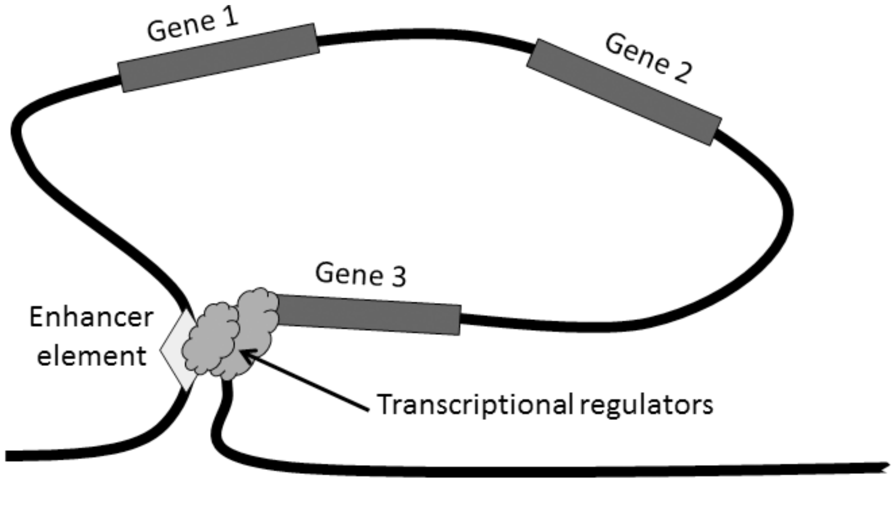

Figure 2. Simplified depiction of how a distant enhancer/regulatory element would interact with the promoter region of gene.

Not only are most plant and animal genes in pieces, but the exons can also be alternatively spliced together to form entirely new RNA products from the same gene (Figure 3). And if it is a gene that encodes the information needed for a protein, different proteins (called *isoforms*) with different functions can be produced—all from the same gene. In fact, the mRNAs and their resulting protein isoforms produced by the alternative processing of primary RNA transcripts can differ markedly in structure, function, cellular localization, and other properties (Wang et al., 2008).

This whole phenomenon, called *alternative splicing*, applies to both protein-coding and most long noncoding RNA genes. Also noted in Figure 3 are alternative promoter start sites for transcription that can affect what exons are included in the final transcripts and its overall final architecture. The use of alternative gene promoters associated with transcript variability is one the key drivers of gene output specificity based on cell-type, tissue-type, and developmental regulation (Gupta et al., 2010; Singer et al., 2008). In fact, it is common for genes in humans to have up to 10 different promoters, with some genes having 20 or more (Singer et al., 2008).

In humans, protein-coding exons represent less than 2% of the total genomic sequence, while introns occupy about 24% (Venter et al., 2001). It is the presence of this genes-in-pieces type of structure and its resulting capacity for alternative splicing that greatly expands the gene-coding lexicon in a language system termed "the splicing code" (Barash et al., 2010; Reddy et al., 2012). Obviously, a process like alternative splicing must be highly regulated and controlled. In fact, many human diseases, including cancer, are linked to a misregulation of alternative splicing, underscoring how important it is for this process to be tightly regulated (Orengo and Cooper, 2007; Ward and Cooper, 2010).

So, what is the language system or code used to determine which exons are included, skipped, doubled, or excluded from a final transcript variant? The ascertaining of such a code has been very difficult, and it is clear that many factors

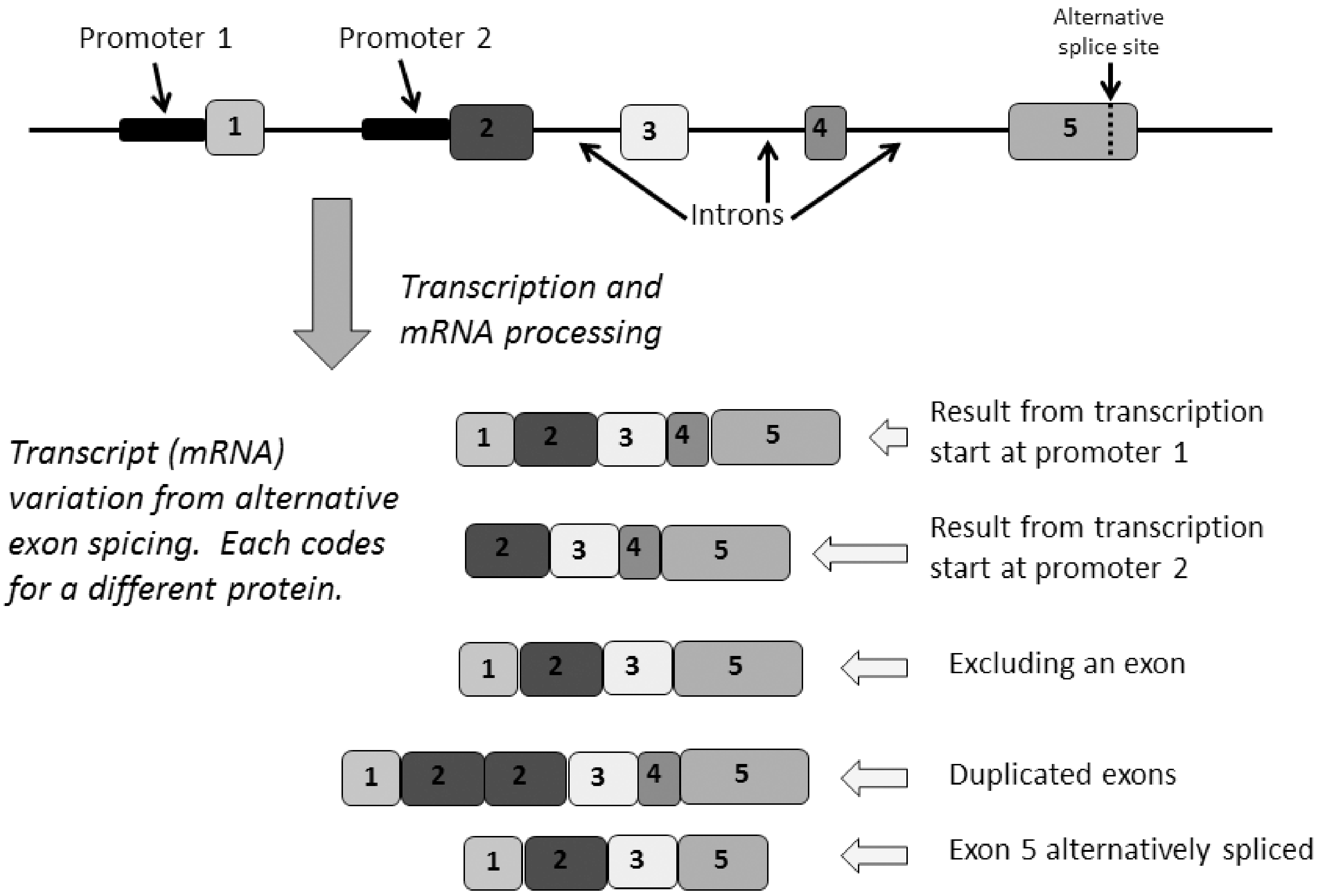
Promoter 1
Promoter 2
Alternative splice site
1
2
3
4
5
Introns
Transcription and mRNA processing
Transcript (mRNA) variation from alternative exon spicing. Each codes for a different protein.
1 2 3 4 5
Result from transcription start at promoter 1
2 3 4 5
Result from transcription start at promoter 2
1 2 3 5
Excluding an exon
1 2 2 3 4 5
Duplicated exons
1 2 3 5
Exon 5 alternatively spliced

Figure 3. Depiction of different types of exon splice variants observed in the alternative splicing of mRNAs.

must be considered together. Researchers now believe they can predict splicing events for most genes in the human genome about 80% of the time using this code (Barash et al., 2010). One of the first things the researchers investigating the splicing code took into account was tissue type, which they placed in four broad groups: central nervous system, muscle, digestive system, and embryonic. Then they factored in about 2,000 different types of regulatory DNA sequence features found in and around genes that provide key signals needed to regulate the splicing events. Finally, they included important details about transcript structure, such as exon/intron lengths, probabilities of secondary structures forming in the RNA transcripts, and whether the transcripts contained variable signals for early termination of making a protein (called *translation*).

Another issue related to the amazing informational diversity that can be derived by the alternative spicing of exons in genes is the ability of some genes to produce chimeric transcripts with other genes—called *fusion genes* or *chimeric RNA transcripts*. This can occur in several different ways (Akiva et al., 2006; Parra et al., 2006). In one scenario, two genes that are located in tandem on the same chromosome can be transcribed together as one large mRNA and then processed as a single RNA. In another scenario, genes located on completely different chromosomes can have their exons spliced together to form completely new transcripts. At first it was questioned as to whether these mRNAs could produce viable functional proteins, but now it is proven that they do (Frenkel-Morgenstern et al., 2012). And just like transcripts from single genes, chimeric mRNAs are also alternatively spliced and the misregulation of this process is associated with cancer and disease (Djebali et al., 2012b; Greger et al., 2014; Hernandez-Torres et al., 2013).

What sort of complicated apparatus actually does all of the dicing and splicing of RNA transcripts in the nucleus? The answer is perhaps one of the most remarkable and complex machines in the cell—the spliceosome. The spliceosome is comprised of a large group of proteins that reads each RNA transcript copied from a gene and then splices it into the correct variants needed at that specific time. In animals, there are actually several types/variants of these spliceosomes with the major spliceosomal complex consisting of about 200 different proteins (Valadkhan and Jaladat, 2010; Wahl et al., 2009). In other words, to code for just the main protein apparatus involved in splicing, at least 200 different genes are required. In addition, the whole process of splicing occurs while the RNA is being transcribed (called *cotranscriptional*) such that the complex machinery of both transcription and splicing are dynamically connected and interacting with each other to produce just the right final product required for the cell (Bentley, 2014). In addition, these sites of transcription and splicing occur at unique locations in the nucleus called *transcription factories* (Figure 4), where the chromosomes are three-dimensionally maneuvered into position (Davidson et al., 2013; Van Bortle and Corces, 2012). The functional nature of these transcription factories helps explain how transcript fusions occur among genes situated on completely different chromosomes.

MicroRNA Binding Code

Yet another language system that interacts with the RNA products of the genome to control how genes are expressed is called the *microRNA binding code* (Salmena et al., 2011; Taulli et al., 2013). MicroRNAs (miRNA) are encoded by a wide variety of miRNA genes all over

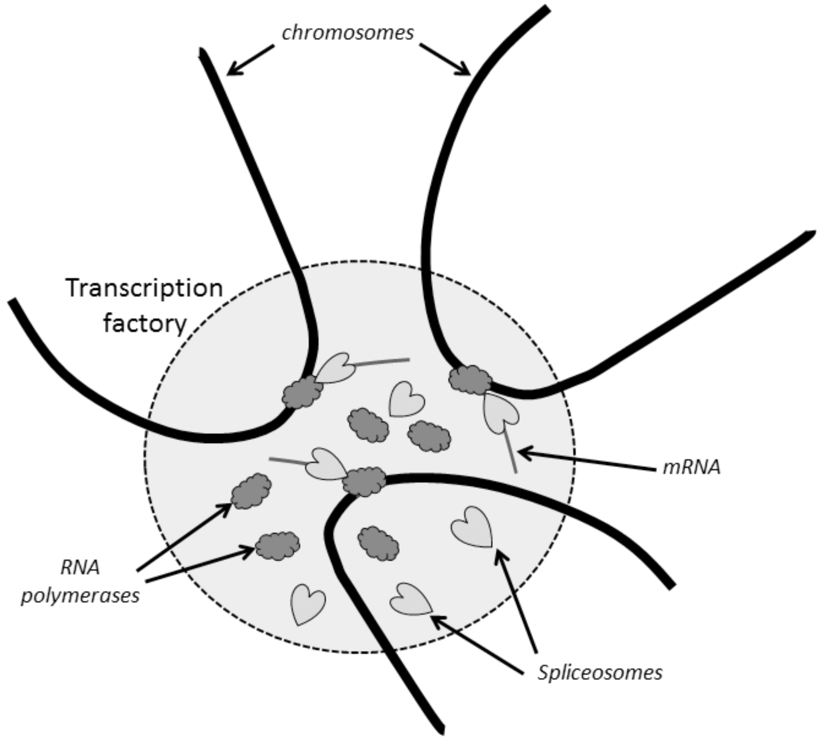

Figure 4. Simplified depiction of a transcription factory with three co-regulated genes from different chromosomes and/or chromosomal regions producing RNA products.

the human genome, some of which are nested inside protein-coding genes within their intron regions. According to the most recent count, there are about 2,555 different miRNA genes in the human genome (mirbase.org). After the RNA transcript from a miRNA gene is fully processed, it is about 22 to 25 nucleotides in length. It is then exported out of the nucleus, where it is further processed and combined with proteins to form a functional micro-machine that regulates the translation of mRNAs in the cell's cytoplasm (Pasquinelli, 2012).

These miRNAs function in gene regulation mostly by binding to specifically encoded sites in the noncoding tails of gene transcripts called *three-prime untranslated regions*. These highly specific sites are complementary with the sequence of the miRNA and are called *miRNA binding sites* or *miRNA response elements* (MRE; see Figure 5). The presence and specific ordering of these sites in the gene transcript is a specific type of RNA-based code that was previously unknown to scientists and is now only beginning to be understood. In fact, some are calling this emerging code the "Rosetta Stone" of a new cellular language that will help further unlock the mysteries of gene regulation in the cell (Salmena et al., 2011; Taulli et al., 2013).

Protein-coding gene transcripts that are ferried out of the nucleus for translation by the protein-making machinery will often have multiple and variable MREs that are in part determined by the process of alternative splicing in the nucleus. Just like the structure of exons in a transcript is determined through alternative splicing, so is the structure and content of MREs. In fact, while most MREs appear to reside in the tails of the gene transcript, to a lesser extent, other MREs are also found in the protein-coding exons (a type of dual code). All of this creates the formation of a specific sequence of sites for miRNA binding to occur, which also leads to various outcomes and types of regulation in the production of proteins.

Because many different genes will share a certain subset of MREs, it is believed that a complicated scenario of competitive binding occurs that helps to buffer and modulate the production of proteins. Many genes that are involved in this competitive binding are typically involved in the same types of cellular processes and are statistically highly co-expressed together (Pasquinelli, 2012; Taguchi, 2013; Wang et al., 2011). Thus, genes that share similar MREs are often coordinately and dynamically controlled together, and the interaction of their shared MREs has been termed "cross talk" (see Figure 5). Because miRNAs have been shown to be involved in nearly every cellular process studied, the importance of the miRNA binding code is key to understanding how the cell works.

Codes in Circles

Another amazing aspect of RNA transcript splicing of genes (both protein-coding and noncoding) is that exons are not only used to form mRNAs of variable content, but they are also selectively spliced into RNA circles. These exons of circular RNAs (circRNAs) are arranged in different orientations than the exons in linear mRNAs, which are produced from the same gene; this is accomplished through a process called *back splicing* (Vicens and Westhof, 2014). And these circRNAs are a fairly recent addition to the ever-increasing list of important functional noncoding RNAs, and now thousands of them have been reported in a variety of animal cells in recent years (Jeck et al., 2013; Memczak et al., 2013; Salzman et al., 2012). In fact, a recent study in humans identified over 7,000 different circRNAs that were estimated to account for at least 10% of the transcripts originating from the genes that were studied (Guo et al., 2014). And in addition to exons being circularized as part of the complexity of splicing, recent research has also shown that introns are circularized (Vicens and Westhof, 2014; Zhang et al., 2013).

While biologists have known of the existence of circular transcripts for over twenty years (Nigro et al., 1991), they were originally misdiagnosed as being nothing more than genetic accidents of aberrant mRNA splicing (Cocquerelle et al., 1993). Now it is known that some of these circRNAs act as an important functional component of the miRNA-mRNA posttranscriptional regulatory network, working as molecular sponges that sequester miRNAs in the cytoplasm (Hansen et al., 2013; Memczak et al., 2013; Vicens and Westhof, 2014). While most circRNAs have not been functionally characterized, researchers have speculated that in addition to acting as miRNA sponges, they are also involved in protein or RNA transport, assembled into functional RNA-protein complexes, or act directly as regulatory RNAs in the

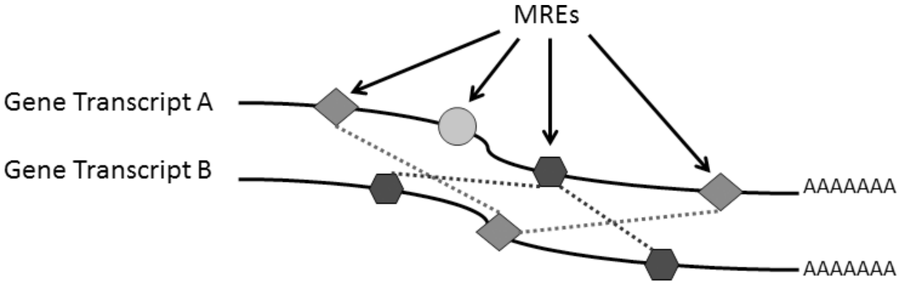

Figure 5. Depiction of miRNA response elements (MREs) in two different mRNA transcripts and "cross-talk" (dashed lines) between shared binding sites.

genome (Memczak et al., 2013; Vicens and Westhof, 2014).

The idea that circRNAs are functional is largely based on the fact that they are expressed in highly tissue-specific patterns (Guo et al., 2014; Memczak et al., 2013) and are generated interactively with their linear mRNA counterparts during splicing at regulated levels (Ashwal-Fluss et al., 2014). Another interesting point linked to the functional complexity of circular splicing involves the recent discovery of specific circRNA binding to the metazoan protein *muscleblind*, a factor that functions in the nucleus as part of the splicing apparatus (Ashwal-Fluss et al., 2014).

Further bolstering the idea of functionality is the fact that most circRNAs are actively transported out of the nucleus to specific destinations in the cytoplasm, where they are stably maintained due to their lack of free ends (Jeck et al., 2013; Memczak et al., 2013). Linear mRNAs are spliced, capped, adenylated (poly-A tail additions), and have specific transport factors and systems that recognize these features, bind to them, and ferry them out of nuclear pores to specific sites in the cytoplasm at ribosomes (Hocine et al., 2010; Wente and Rout, 2010). In contrast, circRNAs are circularized during splicing with no capping or three-prime tail modifications. Because circRNAs have a completely different posttranscriptional structure than linear mRNAs, this implies a specifically tailored recognition, cellular addressing, and transport system accommodating their unique features.

In contrast to exon-generated circRNAs, the circular RNAs encoded by introns function in the nucleus, not the cytoplasm (Zhang et al., 2013). In these newly discovered circular intronic RNAs (ciRNA), optimization of their levels in the nucleus was shown to enhance the transcription of the gene from which they were derived. This was proven when researchers were able to perturb the action of ciRNAs in cells by inhibiting their function and observing the effect on gene expression. It was also discovered that ciRNAs promoted optimal gene function by associating with the RNA polymerase II transcriptional machinery. Interestingly, these ciRNAs were also expressed specific to cell type, emphasizing that they are selectively controlled and functional.

One last piece of amazing evidence for functionality involves a recent discovery that both nuclear and cytoplasmic circular intronic RNAs are passed on to offspring in *Xenopus* (frog) oocytes (Talhouarne and Gall, 2014). This implies a role for these molecules in RNA-mediated inheritance and epigenetics—a newly emerging research field studying epigenetic regulation associated with heritable noncoding RNAs (Liebers et al., 2014).

Dual-Use Messages in Codons

One of the most amazing discoveries of the past few years has been that of dual-purpose codes in the same section of DNA within genes that code for proteins. The same stretch of DNA sequence containing different languages and having multiple purposes that are interpreted by complex cellular machinery in different ways is utterly defying evolutionary predictions.

In a gene's exons, three consecutive DNA letters form what is called a *codon*, and each codon corresponds to a specific amino acid in a protein. Long sets of codons in genes contain the protein-making information that ends up being translated into entire proteins that may be hundreds of amino acids in length.

It has been widely demonstrated that the protein-coding exon regions of genes contain a variety of signals (e.g., splice sites, miRNA binding sites) other than just the information delineating amino acids. It was also recently demonstrated in a genome-wide study that transcription factors commonly clamp onto exons inside genes (Neph et al., 2012). The previous belief was that transcription factors mostly latched onto the controlling regions (promoters) in front of genes—sections of the gene that do not actually code for protein. This finding was somewhat of a mystery because researchers originally thought that transcription-factor binding codes and the protein template codes containing the codons operated independently of each other.

In addition, more recent research is showing that these codes actually work both separately and together. They contain dual meanings (languages) for different types of cell machinery embedded in the same section of DNA. While one set of codons specifies the order of amino acids for a protein, the very same sequence of DNA letters also specifies where transcription factors are to bind to the gene to make the RNA transcript that codes for a protein (Stergachis et al., 2013). In fact, the researchers determined that about 14% of the codons inside 87% of human genes are occupied target sites for transcription factors. As a result of this new discovery, these dual-function code sites in exons have been labeled "duons."

The implications for the preponderance of dual codes providing yet another hurdle for evolutionary models to overcome immediately became obvious to the scientific community. Several researchers in a review recognize this in asking: "How widespread is the phenomenon of 'regulatory' codes that overlap the genetic code, and how do they constrain the evolution of protein sequences?" (Weatheritt and Babu, 2013, p. 1325).

Another interesting aspect of codons is that of apparent redundancy where the first two bases in the codon are non-negotiable but the third base can vary. For example, the codons GGU, GGC, GGA, and GGG all encode the same amino acid called *glycine*. When scientists first discovered this phenomenon, they called the variation in the third base "wobble" and simply relegated the variability as being redundant. In other words, they assumed that all different codon variants for a given amino acid were functionally equivalent.

When a mRNA transcript copy of a gene is ferried out of the nucleus and used to make a protein at cellular machinery sites called *ribosomes*, periodic pausing occurs during the process while the protein is being produced and directed out of a tunnel in the protein-making apparatus (O'Brien et al., 2014; O'Brien et al., 2010). The specific sequence and rate of pausing is critical to the folding of the protein into its proper three-dimensional shape, which occurs during the process of exiting the ribosome. Many different types of cellular machines aid in this folding process, including the ribosome tunnel itself. Because the translation (making of a protein from an RNA transcript) and the folding of the protein are linked together, the processes are called *cotranslational*.

Amazingly, researchers in a new study have shown that the variability in the third base of codons is not redundancy at all but a specific type of cellular language interpreted at the ribosome, telling it when to pause and how to regulate the rate at which the protein is being made, which ultimately has an effect on the folding of the protein into its functional, three-dimensional shape (D'Onofrio and Abel, 2014). Therefore, not only does a codon provide the information for which amino acid to add in the making of a protein, but it provides important information needed on how to regulate its folding. The researchers state, "These dual interpretations enable the assembly of the protein's primary structure while also providing important folding controls via pausing of the translation process."

What was once thought only to be meaningless redundancy has now been proven to be exactly the opposite. In fact, the researchers also stated, "The functionality of codonic [*sic*] redundancy denies the ill-advised label of 'degeneracy'" (D'Onofrio and Abel, 2014). The authors of the report also marveled at such ingenuity and unwittingly state their findings within the context of sophisticated intelligent design. They say, "Redundancy in the primary genetic code allows for additional independent codes. Coupled with the appropriate interpreters and algorithmic processors, multiple dimensions of meaning, and function can be instantiated into the same codon string." This type of jargon essentially describes a highly complex, interpretive, computerlike machine—something designed and engineered by a super-intelligent mind—certainly not the result of random processes.

Antisense Genome Languages

Antisense messages in the genome are obtained by the transcription of the double-stranded DNA from the strand opposite to that of the sense transcript of either protein-coding or nonprotein-coding genes (Grinchuk et al., 2010; Khorkova et al., 2014; Pelechano and Steinmetz, 2013). These antisense genes have their own promoters and are alternatively transcribed and spliced like protein-coding genes. Antisense transcripts operate by binding to sense-coding transcripts in the genome. The field of study surrounding gene regulation by antisense transcription is particularly intriguing because the genomic arrangement of the information directly indicates that the transcripts produced from opposing DNA strands act on each other in a regulatory manner (Grinchuk et al., 2010; Khorkova et al., 2014; Pelechano and Steinmetz, 2013). The amazing design of antisense RNAs inherently gives them unique properties that are different than other types of gene regulators (such as transcription factors) for integrating multiple kinds of regulatory signals, establishing on–off switches, and even "rewiring" and fine-tuning entire gene regulatory networks (Grinchuk et al., 2010; Khorkova et al., 2014; Pelechano and Steinmetz, 2013). Their methods of action are currently shown to include chromatin remodeling leading to differences in epigenetic states, regulatory masking of signals in RNA transcripts, assistance in alternative splicing of mRNAs, and regulation of the translation of mRNAs in the cytoplasm (Khorkova et al., 2014; Li and Ramchandran, 2010; Pelechano and Steinmetz, 2013).

The patterns of antisense expression in genes across the genome is highly complex and varies based on cellular processes, cell type, and environmental conditions and is even affected by neighboring genes (Khorkova et al., 2014; Pelechano and Steinmetz, 2013; Wood et al., 2013). Antisense genes have the same control features as other genes and can even share promoter regions with both protein-coding and noncoding RNA genes. They are also regulated in complex networks with these other types of genes. Thus, this is a yet another type of integrated language system that effectively acts in reverse to the normal forward sense messages contained in the genome.

Epigenetic Language Systems

Epigenetic changes involve the addition of chemical tags in an organism's genome without actually changing the genetic code. Both the DNA nucleotides and the proteins that DNA is wrapped around (called *histones*) can be chemically tagged by different types of molecules that ultimately determine how genes are turned on and off. Thus, the epigenetic regulation of the genome can produce marked differences in growth, development, physiology, and adaptive traits without actually being related to changes in the DNA sequence itself (Liebl et al., 2013; Skinner et al., 2014; Zhu et al., 2013). What is even more amazing is that these changes can also be inherited over multiple generations.

One of the most widely studied types of epigenetic modifications is that of DNA methylation, the process whereby a methyl group is added to cytosine residues. As a general principle, areas of the genome that are heavily methylated tend to be more genetically inactive, while areas that lack methylation are more transcriptionally active (Jones, 2012). For example, the promoter regions of actively expressed genes are typically significantly less methylated than those of inactive genes. However, this is not necessarily a hard and fast rule, as the relationship between methylation and transcription must be analyzed within a more complex information-rich combinatorial context. Indeed, while the promoter of the gene (where transcription factors bind) is often less methylated, the main body of the gene behind the promoter can be heavily methylated and both states are associated with increased transcription (Jones, 2012; Petty and Pillus, 2013). In fact, varying levels of methylation in the gene body are now also being shown to influence and play a role in the splicing code discussed above.

Plant and animal genomes are tightly wound around clusters of eight proteins (called *octamers*) that are referred to as core histones. Collectively, these DNA/histone clusters are called *nucleosomes* and form the basic unit of chromatin (a reference to DNA and its physically combined structure with proteins and RNA). These nucleosomes are densely packed and achieve an astounding 10,000 to 20,000-fold compaction needed to fit a human-sized genome into the tiny volume of a nucleus (Zentner and Henikoff, 2013). What is even more amazing is that these histone proteins that form the core of nucleosomes are highly dynamic and configurable features in the genome that control the access and activity of proteins involved in gene expression, DNA replication, and other regulatory activities.

Histone configuration is modified by adding methyl, acetyl, and crotonyl chemical groups to lysines, and the phosphorylation of serines and threonines in the amino acid histone tails that stick out from the nucleosomes (Cedar and Bergman, 2009; Zentner and Henikoff, 2013). In addition, protein variants of ubiquitins and SUMO (small ubiquitin-like modifier) can also be attached to these histone tails by a variety of sumoylation and ubiquitin-related enzymes (Cubeñas-Potts and Matunis, 2013; Pinder et al., 2013). Because all these types of histone modifications are comprised of different variants, many different chromatin configurations can be achieved to present a highly complex epigenetic language that controls and fine-tunes how the DNA is expressed and made available to the transcriptional machinery in the genome.

At present, well over 100 different histone modifications have been characterized with more being discovered at a rapid rate (Zentner and Henikoff, 2013). When one considers the combinatorial capacity of the histone code alone, and the necessity of systems to interpret, write, and replicate it over multiple cell divisions, developmental states, and organismal generations, the complexity involved is clearly mind-boggling, to say the least.

Both DNA methylation and histone modifications represent separate epigenetic languages that work in unison to regulate access of the regulatory and transcriptional machinery in the genome that extracts, copies, and utilizes information encoded in the DNA. The amazing complexity of these epigenetic language systems are interactively layered over the other languages present within the encoded bases of DNA.

Why Junk Is Bunk

Invariably, when the discussion of languages and information in the genome is brought up, the question of junk DNA arises. In the present sphere of debate, the term "junk DNA" is used in a broad sense to refer to any DNA sequence that does not function in development, physiology, or some other organismal trait (Palazzo and Gregory, 2014). Although not well documented, the idea originated as an answer to Haldane's dilemma, which proposed that only a small part of the genome could possibly be functional (Nei, 2013). While the first use of the term "junk DNA" is often attributed to Susumu Ohno in the early 1970s (who actually just used it in reference to pseudogenes), the idea of junk in the genome was popular throughout the 1960s (Graur, 2013; Palazzo and Gregory, 2014).

The whole idea of junk DNA is based on the concept of neutral evolution, which predicts that a large proportion of the human genome should be littered with freely evolving DNA that is not constrained and therefore nonfunctional. This is called the "neutral model" theory, and it seemed to provide solutions to selection problems uncovered in theoretical studies of population genetics (Kimura, 1983; Nei, 2013). Indeed, early discoveries in genomics seemed to lend support to this idea, as scientists found that only about 2% of the entire human genome codes directly for proteins [exons] (International Human Genome Sequencing, 2004). However, this statistic is deceiving because when the entire length of a protein-coding gene is considered (including introns and regulatory sequence), over 40% of the human genome is covered by protein-coding genes (Palazzo and Gregory, 2014).

Now we also realize there are many other genes that produce noncoding RNAs (discussed above), which are thought to outnumber protein-coding genes by at least two-to-one (Managadze et al., 2013). And variation in these types of long, intergenic, noncoding RNA genes has a significant impact on human health and disease, underscoring their functional importance despite the fact that most have not yet been functionally characterized (Chen et al., 2013; Kumar et al., 2013).In addition to long, noncoding RNA genes, the other primary candidate for being labeled "junk DNA" has been the highly repetitive regions that, after years of study, seemed to have no discernible function, despite the fact they are found to be actively transcribed into RNA. This fraction of the genome is labeled COT-1, getting its name from early studies in DNA reassociation kinetics (Britten et al., 1974). The procedure involves heating or chemically treating genomic DNA until it denatures and becomes single stranded, and then allowing it to reassociate—with the more repetitive fractions coming together first. The COT-1 fraction of the genome would be that initial fraction that reassociates the most rapidly.

In a recent study, it was shown that these COT-1 RNA molecules are literally painted across the euchromatic areas—the functionally active regions of chromosomes (Hall et al., 2014). Hence, these molecules are called *euchromatin-associated RNAs*, or *ecRNAs*; and in contrast to the previously characterized *Xist* RNAs that shut down chromosome activity, specifically on the X chromosome (Engreitz et al., 2013), ecRNAs do exactly the opposite: they promote an active local environment of genetic functionality. In simplest terms, they help create an RNA matrix surrounding the chromosomes that promotes gene function and transcriptional stability.

Amazingly, these ecRNAs even persist in experiments when the underlying chromosomes are destroyed with DNases. Clearly, they are very stable and an important part of the chromosomal matrix in the cell nucleus. In fact, when the ecRNAs themselves are destroyed using RNase, the chromosomes rapidly condense and collapse. If it were not for the presence of the ecRNAs, chromosome stability and genome function would not even be possible. Furthermore, the ecRNAs specifically associate with the chromosomal segments from which they are derived, exhibiting regional specificity.Much of the research that has debunked the idea of junk DNA has come from the ENCODE project (as discussed earlier), which began in 2003 as an expansion of the Human Genome project. While the human genome had been largely sequenced by 2004 (International Human Genome Sequencing, 2004), researchers knew very little about what it all meant, except for about 21,500 protein-coding genes they initially identified—and most of those they knew very little about. After the first round of ENCODE research (Birney et al., 2007), it soon became apparent that the human genome was pervasively transcribed, an idea that led to the realization that the eukaryotic genome is "an RNA machine" (Amaral et al., 2008). This fact is continuing to be confirmed not only in humans, but also across the spectrum of metazoan life (Djebali et al., 2012a; Liu et al., 2013; Managadze et al., 2013).

The second phase of ENCODE funding resulted in 30 different research papers being published in 2012 and was no less spectacular in its discoveries than the first tier. In the lead research paper, the authors wrote, "These data enabled us to assign biochemical functions for 80% of the genome, in particular outside of the well-studied protein-coding regions" (Dunham et al., 2012, p. 57). In a media interview, Tom Gingeras, one of the senior scientists on the ENCODE project, said, "Almost every nucleotide is associated with a function of some sort or another, and we now know where they are, what binds to them, what their associations are, and more" (Yong, 2012).The areas of study in the ENCODE project are diverse and cover all the different genome codes that are discussed in this paper. In reality, the work of discovery for ENCODE-related researchers has only just begun. The inner workings of the genome are more complex than researchers ever imagined they would be. A brief summary of the most pertinent ENCODE findings are listed below:

- Over 80 percent of the human genome is actively involved in at least one or more biochemical reactions associated with gene regulation in at least one type of cell. Nearly all of the genome lies within close proximity to some sort of genetic regulatory event and, therefore, very little of the genome can be considered unnecessary or nonfunctional.
- The human genome can be classified into seven different, broadly categorized, genetically active states that enhance gene expression, mapped to 399,124 different regions.
- Although the human genome may contain only ~21,000 genes, scientists found 70,292 areas called *gene promoters* associated with the protein-coding areas of genes. This finding confirms the idea that genes are like molecular Swiss Army knives, providing a diversity of products and outcomes depending on how they are operated and controlled.
- Gene expression is controlled by a broad array of regulatory proteins, chemical marks in the DNA (epigenetic factors), gene promoter features (specific DNA sites), and enhancer sequences that are sometimes located thousands and millions of bases from a gene or set of genes. All of these features operate in concert with other genes and regulatory features in irreducibly complex and intricately coordinated networks.
- ENCODE-related genetic variation plays a large role in the observed variability among humans, perhaps more so than the variation observed within protein-coding regions. Many heritable human diseases are associated with variations or mutations in ENCODE regions and not in the actual protein-coding regions.

Summary

For the genome to function in all its complexity, many different codes and languages are used, and they all mesh and work interactively with one another. In addition, all of these language systems follow the rule of information theory discussed at the beginning of this paper and thus necessitate an information provider. In fact, the effective, interlocking and internetworking of these highly complex language systems speaks directly to a Creator of infinite wisdom and capabilities.

We are only beginning to decipher the true complexity of these different genetic languages; and as research progresses, it is likely more languages and codes will be revealed, and the codes we are now aware of will likely grow more complex in their mechanisms and scope. Taken together, the genome is an irreducibly complex network of interacting dynamic codes and languages that undeniably speak of an omnipotent and all-wise Creator as described in the Bible.

References

Adachi, N., and M.R. Lieber. 2002. Bidirectional gene organization: a common architectural feature of the human genome. *Cell* 109(7): 807–809.

Akiva, P., A. Toporik, S. Edelheit, Y. Peretz, A. Diber, R. Shemesh, A. Novik, R. Sorek. 2006. Transcription-mediated gene fusion in the human genome. *Genome Res* 16(1): 30–36.

Amaral, P.P., M.E. Dinger, T.R. Mercer, J.S. Mattick. 2008. The eukaryotic genome as an RNA machine. *Science* 319(5871): 1787–1789.

Ashwal-Fluss, R., M. Meyer, N.R. Pamudurti, A. Ivanov, O. Bartok, M. Hanan, N. Evantal, S. Memczak, N. Rajewsky, S. Kadener. 2014. CircRNA biogenesis competes with pre-mRNA splicing. *Molecular Cell* 56(1): 55–66.

Barash, Y., J.A. Calarco, W. Gao, Q. Pan, X. Wang, O. Shai, B.J. Blencowe, B.J. Frey. 2010. Deciphering the splicing code. *Nature* 465(7294): 53–59.

Bentley, D.L. 2014. Coupling mRNA processing with transcription in time and space. *Nature Reviews Genetics* 15(3): 163–175.

Birney, E., J.A. Stamatoyannopoulos, A. Dutta, R. Guigo, T.R. Gingeras, E.H. Margulies, Z. Weng, M. Snyder, E.T. Dermitzakis, R.E. Thurman et al. 2007. Identification and analysis of functional elements in 1% of the human genome by the Encode Pilot Project. *Nature* 447(7146): 799–816.

Britten, R.J., D.E. Graham, and B.R. Neufeld. 1974. Analysis of repeating DNA sequences by reassociation. *Methods Enzymol* 29:363–418.

Cedar, H., and Y. Bergman. 2009. Linking DNA methylation and histone modification: patterns and paradigms. *Nature Reviews Genetics* 10(5): 295–304.

Chen, G., C. Qiu, Q. Zhang, B. Liu, and Q. Cui. 2013. Genome-wide analysis of human snps at long intergenic noncoding RNAs. *Human Mutatation* 34(2): 338–344.

Clark, M.B., A. Choudhary, M.A. Smith, R.J. Taft, and J.S. Mattick. 2013. The dark matter rises: the expanding world of regulatory RNAs. *Essays in Biochemistry* 54: 1–16.

Cocquerelle, C., B. Mascrez, D. Hetuin, and B. Bailleul. 1993. Mis-splicing yields circular RNA molecules. *FASEB Journal* 7(1): 155–160.

Cohen, S.N., A.C. Chang, H.W. Boyer, and R.B. Helling. 1973. Construction of biologically functional bacterial plasmids in vitro. *Proceedings of the National Academy of Sciences USA* 70(11): 3240–3244.

Cubeñas-Potts, C., and M.J. Matunis. 2013. Sumo: a multifaceted modifier of chromatin structure and function. *Developmental Cell* 24(1): 1–12.

Davidson, S., N. Macpherson, and J.A. Mitchell. 2013. Nuclear organization of RNA polymerase ii transcription. *Biochemistry and Cell Biology* 91(1): 22–30.

Dickel, D.E., A. Visel, L.A. Pennacchio. 2013. Functional anatomy of distant-acting mammalian enhancers. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 368(1620): 20120359.

Djebali, S., C.A. Davis, A. Merkel, A. Dobin, T. Lassmann, A. Mortazavi, A. Tanzer, J. Lagarde, W. Lin, F. Schlesinger et al. 2012a. Landscape of transcription in human cells. *Nature* 489(7414): 101–108.

Djebali, S., J. Lagarde, P. Kapranov, V. Lacroix, C. Borel, J.M. Mudge, C. Howald, S. Foissac, C. Ucla, J. Chrast et al. 2012b. Evidence for transcript networks composed of chimeric RNAs in human cells. *PLoS One* 7(1): e28213.

D'Onofrio, D.J., and D.L. Abel. 2014. Redundancy of the genetic code enables translational pausing. *Frontiers in Genetics* 5:140.

Dunham, I., A. Kundaje, S.F. Aldred, P.J. Collins, C.A. Davis, F. Doyle, C.B. Epstein, S. Frietze, J. Harrow, R. Kaul et al. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489(7414): 57–74.

Engreitz, J.M., A. Pandya-Jones, P. McDonel, A. Shishkin, K. Sirokman, C. Surka, S. Kadri, J. Xing, A. Goren, E.S. Lander et al. 2013. The xist lncRNA exploits three-dimensional genome architecture to spread across the x chromosome. *Science* 341(6147): 1237973.

Frenkel-Morgenstern, M., V. Lacroix, I. Ezkurdia, Y. Levin, A. Gabashvili, J. Prilusky, A. Del Pozo, M. Tress, R. Johnson, R. Guigo et al. 2012. Chimeras taking shape: potential functions of proteins encoded by chimeric RNA transcripts. *Genome Research* 22(7): 1231–1242.

Gerstberger, S., M. Hafner, and T. Tuschl. 2014. A census of human RNA-binding proteins. *Nature Reviews Genetics* 15(12): 829–845.

Gerstein, M.B., C. Bruce, J.S. Rozowsky, D. Zheng, J. Du, J.O. Korbel, O. Emanuelsson, Z.D. Zhang, S. Weissman, and M. Snyder. 2007. What is a gene, post-encode? History and updated definition. *Genome Research* 17(6): 669–681.

Gerstein, M.B., A. Kundaje, M. Hariharan, S.G. Landt, K.K. Yan, C. Cheng, X.J.

Mu, E. Khurana, J. Rozowsky, R. Alexander et al. 2012. Architecture of the human regulatory network derived from encode data. *Nature* 489(7414): 91–100.

Gilbert, W. 1978. Why genes in pieces? *Nature* 271(5645): 501.

Gitt, W. 2011. *Without Excuse* Creation Book Publishers, Atlanta GA.

Graur, D. 2013. The origin of junk DNA: a historical whodunnit. http://judgestarling.tumblr.com/post/64504735261/the-origin-of-the-term-junk-dna-a-historical.

Greger, L., J. Su, J. Rung, P.G. Ferreira, C. Geuvadis, T. Lappalainen, E.T. Dermitzakis, and A. Brazma. 2014. Tandem RNA chimeras contribute to transcriptome diversity in human population and are associated with intronic genetic variants. *PLoS One* 9(8): e104567.

Grinchuk, O.V., P. Jenjaroenpun, Y.L. Orlov, J. Zhou, and V.A. Kuznetsov. 2010. Integrative analysis of the human cis-antisense gene pairs, miRNAs and their transcription regulation patterns. *Nucleic Acids Research* 38(2): 534–547.

Guo, J.U., V. Agarwal, H. Guo, and D.P. Bartel. 2014. Expanded identification and characterization of mammalian circular RNAs. *Genome Biology* 15(7): 409.

Gupta, R., P. Wikramasinghe, A. Bhattacharyya, F.A. Perez, S. Pal, and R.V. Davuluri. 2010. Annotation of gene promoters by integrative data-mining of chip-seq pol-ii enrichment data. *BMC Bioinformatics* 11 Supplement 1: S65.

Hall, L.L., D.M. Carone, A.V. Gomez, H.J. Kolpa, M. Byron, N. Mehta, F.O. Fackelmayer, and J.B. Lawrence. 2014. Stable c0t-1 repeat RNA is abundant and is associated with euchromatic interphase chromosomes. *Cell* 156(5): 907–919.

Hangauer, M.J., I.W. Vaughn, and M.T. McManus. 2013. Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs. *PLoS Genetics* 9(6): e1003569.

Hansen, T.B., T.I. Jensen, B.H. Clausen, J.B. Bramsen, B. Finsen, C.K. Damgaard, and J. Kjems. 2013. Natural RNA circles function as efficient microRNA sponges. *Nature* 495(7441): 384–388.

Hernandez-Torres, F., A. Rastrojo, and B. Aguado. 2013. Intron retention and transcript chimerism conserved across mammals: Ly6g5b and csnk2b-ly6g5b as examples. *BMC Genomics* 14:199.

Hocine, S., R.H. Singer, and D. Grunwald. 2010. RNA processing and export. *Cold Spring Harbor Perspectives in Biology* 2(12): a000752.

International Human Genome Sequencing. 2004. Finishing the euchromatic sequence of the human genome. *Nature* 431(7011): 931–945.

Jeck, W.R., J.A. Sorrentino, K. Wang, M.K. Slevin, C.E. Burd, J. Liu, W.F. Marzluff, and N.E. Sharpless. 2013. Circular RNAs are abundant, conserved, and associated with alu repeats. *RNA* 19(2): 141–157.

Johannsen, W. 1909. Elemente der exakten erblichkeitslehre. *Jena: Gustav Fischer*: 143–145.

Jones, P.A. 2012. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature Reviews Genetics* 13(7): 484–492.

Khorkova, O., A.J. Myers, J. Hsiao, and C. Wahlestedt. 2014. Natural antisense transcripts. *Human Molecular Genetics* 23(R1): R54–63.

Kimura, M. 1983. *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge, UK.

Koonin, E.V., M. Csuros, and I.B. Rogozin. 2013. Whence genes in pieces: reconstruction of the exon-intron gene structures of the last eukaryotic common ancestor and other ancestral eukaryotes. *Wiley Interdisciplinary Reviews: RNA* 4(1): 93–105.

Kumar, A. 2009. An overview of nested genes in eukaryotic genomes. *Eukaryotic Cell* 8(9): 1321–1329.

Kumar, V., H.J. Westra, J. Karjalainen, D.V. Zhernakova, T. Esko, B. Hrdlickova, R. Almeida, A. Zhernakova, E. Reinmaa, U. Vosa et al. 2013. Human disease-associated genetic variation impacts large intergenic non-coding RNA expression. *PLoS Genetics* 9(1): e1003201.

Lander, E.S., L.M. Linton, B. Birren, C. Nusbaum, M.C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh et al. 2001. Initial sequencing and analysis of the human genome. *Nature* 409(6822): 860–921.

Lee, Y.C., and H.H. Chang. 2013. The evolution and functional significance of nested gene structures in drosophila melanogaster. *Genome Biology and Evolution* 5(10): 1978–1985.

Lemay, D.G., W.F. Martin, A.S. Hinrichs, M. Rijnkels, J.B. German, I. Korf, and K.S. Pollard. 2012. G-nest: a gene neighborhood scoring tool to identify co-conserved, co-expressed genes. *BMC Bioinformatics* 13:253.

Lewis, E.B. 1941. Another case of unequal crossing-over in drosophila melanogaster. *Proceedings of the National Academy of Sciences USA* 27(1): 31–35.

Li, K., and R. Ramchandran. 2010. Natural antisense transcript: A concomitant engagement with protein-coding transcript. *Oncotarget* 1(6): 447–452.

Liebers, R., M. Rassoulzadegan, and F. Lyko. 2014. Epigenetic regulation by heritable RNA. *PLoS Genetics* 10(4): e1004296.

Liebl, A.L., A.W. Schrey, C.L. Richards, and L.B. Martin. 2013. Patterns of DNA methylation throughout a range expansion of an introduced songbird. *Integrative and Comparative Biology* 53(2): 351–358.

Liu, G., J.S. Mattick, and R.J. Taft. 2013. A meta-analysis of the genomic and transcriptomic composition of complex life. *Cell Cycle* 12(13): 2061–2072.

Managadze, D., A.E. Lobkovsky, Y.I. Wolf, S.A. Shabalina, I.B. Rogozin, and E.V. Koonin. 2013. The vast, conserved mammalian lincRNome. *PLoS Computational Biology* 9(2): e1002917.

Memczak, S., M. Jens, A. Elefsinioti, F. Torti, J. Krueger, A. Rybak, L. Maier, S.D. Mackowiak, L.H. Gregersen, M. Munschauer et al. 2013. Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature* 495(7441): 333–338.

Necsulea, A., M. Soumillon, M. Warnefors,

A. Liechti, T. Daish, U. Zeller, J.C. Baker, F. Grutzner, and H. Kaessmann. 2014. The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature* 505(7485): 635–640.

Nei, M. 2013. *Mutation-Driven Evolution*. Oxford University Press, Oxford, UK.

Neph, S., J. Vierstra, A.B. Stergachis, A.P. Reynolds, E. Haugen, B. Vernot, R.E. Thurman, S. John, R. Sandstrom, A.K. Johnson et al. 2012. An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* 489(7414): 83–90.

Nigro, J.M., K.R. Cho, E.R. Fearon, S.E. Kern, J.M. Ruppert, J.D. Oliner, K.W. Kinzler, and B. Vogelstein. 1991. Scrambled exons. *Cell* 64(3): 607–613.

O'Brien, E.P., P. Ciryam, M. Vendruscolo, and C.M. Dobson. 2014. Understanding the influence of codon translation rates on cotranslational protein folding. *Accounts of Chemical Research* 47(5): 1536–1544.

O'Brien, E.P., S.-T.D. Hsu, J. Christodoulou, M. Vendruscolo, and C.M. Dobson. 2010. Transient tertiary structure formation within the ribosome exit port. *Journal of the American Chemical Society* 132(47): 16928–16937.

Oliver, C.P. 1940. A reversion to wild-type associated with crossing-over in drosophila melanogaster. *Proceedings of the National Academy of Sciences USA* 26(7): 452–454.

Orengo, J.P., and T.A. Cooper. 2007. Alternative splicing in disease. *Advances in Experimental Medicine and Biology* 623: 212–223.

Osbourn, A.E., and B. Field. 2009. Operons. *Cellular and Molecular Life Sciences* 66(23): 3755–3775.

Palazzo, A.F., and T.R. Gregory. 2014. The case for junk DNA. *PLoS Genetics* 10(5): e1004351.

Parra, G., A. Reymond, N. Dabbouseh, E.T. Dermitzakis, R. Castelo, T.M. Thomson, S.E. Antonarakis, and R. Guigó. 2006. Tandem chimerism as a means to increase protein complexity in the human genome. *Genome Research* 16(1): 37–44.

Pasquinelli, A.E. 2012. MicroRNAs and their targets: recognition, regulation and an emerging reciprocal relationship. *Nature Reviews Genetics* 13(4): 271–282.

Pelechano, V., and L.M. Steinmetz. 2013. Gene regulation by antisense transcription. *Nature Reviews Genetics* 14(12): 880–893.

Petty, E., and L. Pillus. 2013. Balancing chromatin remodeling and histone modifications in transcription. *Trends in Genetics* 29(11): 621–629.

Pinder, J.B., K.M. Attwood, and G. Dellaire. 2013. Reading, writing, and repair: the role of ubiquitin and the ubiquitin-like proteins in DNA damage signaling and repair. *Frontiers in Genetics* 4:45.

Portin, P. 2002. Historical development of the concept of the gene. *Journal of Medicine and Philosophy* 27(3): 257–286.

Portin, P. 2009. The elusive concept of the gene. *Hereditas* 146(3): 112–117.

Reddy, A.S., M.F. Rogers, D.N. Richardson, M. Hamilton, and A. Ben-Hur. 2012. Deciphering the plant splicing code: experimental and computational approaches for predicting alternative splicing and splicing regulatory elements. *Frontiers in Plant Science* 3:18.

Rinn, J.L., and H.Y. Chang. 2012. Genome regulation by long noncoding RNAs. *Annual Review of Biochemistry* 81:145–166.

Sakabe, N.J., D. Savic, and M.A. Nobrega. 2012. Transcriptional enhancers in development and disease. *Genome Biology* 13(1): 238.

Salmena, L., L. Poliseno, Y. Tay, L. Kats, and P.P. Pandolfi. 2011. A ceRNA hypothesis: the rosetta stone of a hidden RNA language? *Cell* 146(3): 353–358.

Salzman, J., C. Gawad, P.L. Wang, N. Lacayo, and P.O. Brown. 2012. Circular RNAs are the predominant transcript isoform from hundreds of human genes in diverse cell types. *PLoS One* 7(2): e30733.

Sanger, F., S. Nicklen, and A.R. Coulson. 1977. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences USA* 74(12): 5463–5467.

Sanna, C.R., W.H. Li, and L. Zhang. 2008. Overlapping genes in the human and mouse genomes. *BMC Genomics* 9:169.

Sigova, A.A., A.C. Mullen, B. Molinie, S. Gupta, D.A. Orlando, M.G. Guenther, A.E. Almada, C. Lin, P.A. Sharp, C.C. Giallourakis et al. 2013. Divergent transcription of long noncoding RNA/mRNA gene pairs in embryonic stem cells. *Proceedings of the National Academy of Sciences USA* 110(8): 2876–2881.

Singer, G.A.C., J. Wu, P. Yan, C. Plass, T.H.M. Huang, and R.V. Davuluri. 2008. Genome-wide analysis of alternative promoters of human genes using a custom promoter tiling array. *BMC Genomics* 9: 349.

Skinner, M.K., C. Gurerrero-Bosagna, M.M. Haque, E.E. Nilsson, J.A. Koop, S.A. Knutie, and D.H. Clayton. 2014. Epigenetics and the evolution of Darwin's finches. *Genome Biology and Evolution* 6(8): 1972–1989.

Southern, E.M. 1975. Detection of specific sequences among DNA fragments separated by gel electrophoresis. *Journal of Molecular Biology* 98(3): 503–517.

Stergachis, A.B., E. Haugen, A. Shafer, W. Fu, B. Vernot, A. Reynolds, A. Raubitschek, S. Ziegler, E.M. LeProust, J.M. Akey et al. 2013. Exonic transcription factor binding directs codon choice and affects protein evolution. *Science* 342(6164): 1367–1372.

Taguchi, Y.-H. 2013. MicroRNA-mediated regulation of target genes in several brain regions is correlated to both microRNA-targeting-specific promoter methylation and differential microRNA expression. *BioData Mining* 6(1): 11.

Talhouarne, G.J., and J.G. Gall. 2014. Lariat intronic RNAs in the cytoplasm of xenopus tropicalis oocytes. *RNA* 20(9): 1476–1487.

Taulli, R., C. Loretelli, and P.P. Pandolfi. 2013. From pseudo-ceRNAs to circ-cerRNAs: a tale of cross-talk and competition. *Nat Structural & Molecular Biology* 20(5): 541–543.

Tomkins, J. 2014. Comparison of the transcribed intergenic regions of the human genome to chimpanzee. *Creation Research Society Quarterly* 50:212–221.

Trinklein, N.D., S.F. Aldred, S.J. Hartman, D.I. Schroeder, R.P. Otillar, and R.M. Myers. 2004. An abundance of bidirectional promoters in the human genome. *Genome Research* 14(1): 62–66.

Ulitsky, I., and D.P. Bartel. 2013. LincRNAs: Genomics, evolution, and mechanisms. *Cell* 154(1): 26–46.

Valadkhan, S. and Y. Jaladat. 2010. The spliceosomal proteome: at the heart of the largest cellular ribonucleoprotein machine. *Proteomics* 10(22): 4128–4141.

Van Bortle, K., and V.G. Corces. 2012. Nuclear organization and genome function. *Annual Review of Cell and Developmental Biology* 28:163–187.

Veeramachaneni, V., W. Makalowski, M. Galdzicki, R. Sood, and I. Makalowska. 2004. Mammalian overlapping genes: the comparative perspective. *Genome Research* 14(2): 280–286.

Venter, J.C., M.D. Adams, E.W. Myers, P.W. Li, R.J. Mural, G.G. Sutton, H.O. Smith, M. Yandell, C.A. Evans, R.A. Holt et al. 2001. The sequence of the human genome. *Science* 291(5507): 1304–1351.

Vicens, Q., and E. Westhof. 2014. Biogenesis of circular RNAs. *Cell* 159(1): 13–14.

Wahl, M.C., C.L. Will, and R. Luhrmann. 2009. The spliceosome: design principles of a dynamic rnp machine. *Cell* 136(4): 701–718.

Wakano, C., J.S. Byun, L.J. Di, and K. Gardner. 2012. The dual lives of bidirectional promoters. *Biochimica and Biophysica Acta* 1819(7): 688–693.

Wang, E.T., R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S.F. Kingsmore, G.P. Schroth, and C.B. Burge. 2008. Alternative isoform regulation in human tissue transcriptomes. *Nature* 456(7221): 470–476.

Wang, G., K. Qi, Y. Zhao, Y. Li, L. Juan, M. Teng, L. Li, Y. Liu, Y. Wang. 2013. Identification of regulatory regions of bidirectional genes in cervical cancer. *BMC Medical Genomics* 6 Suppl 1: S5.

Wang, Y., X. Li, and H. Hu. 2011. Transcriptional regulation of co-expressed microRNA target genes. *Genomics* 98(6): 445–452.

Ward, A.J., and T.A. Cooper. 2010. The pathobiology of splicing. *Journal of Pathology* 220(2): 152–163.

Washietl, S., M. Kellis, and M. Garber. 2014. Evolutionary dynamics and tissue specificity of human long noncoding RNAs in six mammals. *Genome Research* 24:616–628.

Weatheritt, R.J., and M.M. Babu. 2013. Evolution. The hidden codes that shape protein evolution. *Science* 342(6164): 1325–1326.

Wente, S.R., and M.P. Rout. 2010. The nuclear pore complex and nuclear transport. *Cold Spring Harbor Perspectives in Biology* 2(10): a000562.

Wood, E.J., K. Chin-Inmanu, H. Jia, and L. Lipovich. 2013. Sense-antisense gene pairs: Sequence, transcription, and structure are not conserved between human and mouse. *Frontiers in Genetics* 4: 183.

Yong, E. 2012. Encode: the rough guide to the human genome. *Discover Magazine*, http://blogs.discovermagazine.com/notrocketscience/2012/09/05/encode-the-rough-guide-to-the-human-genome/#.VLwmuy6zkVM.

Zentner, G.E., and S. Henikoff. 2013. Regulation of nucleosome dynamics by histone modifications. *Nature Structural & Molecular Biology* 20(3): 259–266.

Zhang, Y., X.O. Zhang, T. Chen, J.F. Xiang, Q.F. Yin, Y.H. Xing, S. Zhu, L. Yang, and L.L. Chen. 2013. Circular intronic long noncoding RNAs. *Molecular Cell* 51(6): 792–806.

Zhu, J., M. Adli, J.Y. Zou, G. Verstappen, M. Coyne, X. Zhang, T. Durham, M. Miri, V. Deshpande, P.L. De Jager et al. 2013. Genome-wide chromatin state transitions associated with developmental and environmental cues. *Cell* 152(3): 642–654.