Creation Research Society Quarterly 2014. 50:212–221.

Comparison of the Transcribed Intergenic Regions of the Human Genome to Chimpanzee

Jeffrey P. Tomkins*

Abstract

The human genome is pervasively transcribed and produces a wide array of long noncoding RNAs that have been implicated in gene regulation, chromatin modification, nuclear organization, and scaffolding for functionally active protein complexes. Of particular interest in human origins is the long and very long intergenic noncoding RNAs transcribed from genomic regions outside protein coding genes. These are known as lincRNA and vlincRNA, respectively. LincRNA regions of the genome are more taxonomically restricted than protein coding segments and make logical candidates for research in genomic discontinuity. This report describes the comparative use of three different human lincRNA datasets and one vlincRNA dataset to the chimpanzee genome using the BLASTN algorithm. Short human lincRNA genomic regions (less than 600 bases) were about 75–79% similar to chimpanzee, while the larger lincRNA regions (greater than 600 bases) were about 71 to 74% similar. The human vlincRNA genomic regions were only 67% similar to chimpanzee. In contrast, all known human protein coding exons 300 to 599 bases in length, are 86% similar to chimpanzee.

Introduction

The human genome of about 3 billion bases is an incredible storehouse of complex genetic information. The most recent estimate of protein coding sequences indicates about 28,000 to 31,000 genes (Wijaya et al., 2013), which comprises less than 5% of the total genomic sequence if just the coding exons are considered. Despite the proportionally small amount of protein coding sequence, the genome is ubiquitously copied (transcribed) into RNA. In fact, the initial report of the ENCODE project listed this phenomenon as their number one finding and 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).

More recent research using a variety of new technologies has provided evidence of pervasive transcription for at

* Jeffrey P. Tomkins, Institute for Creation Research, Dallas, TX, jtomkins@icr.org
Accepted for publication March 17, 2014

least 84 to 93% of the human genome (Clark et al., 2011; Djebali et al., 2012). This high level of transcription initiates and/or occurs outside the boundaries of known protein coding exons and, when first characterized, was initially labeled the "dark matter of the genome" because of its relatively uncharacterized and mysterious nature (Johnson et al., 2005). This expressed genomic dark matter is now commonly and broadly referred to as noncoding RNA and has been shown to encode a wide variety of functional sequence categories that are generally divided into short and long noncoding RNAs (Kapranov and St Laurent, 2012; Clark et al., 2013; Geisler and Coller, 2013).

Long noncoding RNAs (lncRNA) are generally defined as non-protein-coding regions whose transcripts are longer than 200 bases (Rinn and Chang, 2012; Geisler and Coller, 2013). These lncRNAs are transcribed from intergenic regions, introns within genes, and also include anti-sense transcripts that partially overlap protein-coding genes (Rinn and Chang, 2012; Geisler and Coller, 2013). The major emerging role of many lncRNAs is that they combine with a diversity of proteins to form extensive networks of nuclear complexes that target, recruit, and help position various enzymatic activities to specific addresses across the genome (Khalil et al., 2009; Rinn and Chang, 2012; Mercer and Mattick, 2013). Such activities would include chromatin modification to either facilitate or repress transcription. On a broader genomic level, lncRNAs are also proving to be key players in DNA repair, chromosomal positioning in the nucleus, and overall genome stability and function (Ohsawa et al., 2013). Amazingly, research is also revealing that the expressed lncRNAs that act in organizing and modifying chromatin are themselves epigenetically modified to facilitate this activity through cytosine methylation of transcripts (Squires et al., 2012; Amort et al., 2013).

Some lncRNAs are also emerging, not only as repressors of gene activity but also as key players in initiating gene activity and transcription (Krishnan and Mishra, 2013). A related field of research is showing that lncRNAs can also play a wide variety of roles in post-transcriptional gene regulation (Yoon et al., 2013). One aspect in this regard involves stabilizing and promoting translation of mRNAs via base pairing. Another posttranscriptional role played by some lncRNAs is in modulating gene expression by acting as decoys for RNA binding proteins and microRNAs (miRNAs).

While lncRNAs are generally categorized as noncoding, recent studies have shown that some lncRNAs can be processed into miRNAs (He et al., 2008; Jalali et al., 2012) and small open reading frames (smORFs) that encode short functional peptides (Magny et al., 2013). In fact, the association of a subset of lncRNAs with ribosomes has been verified in several studies (Ingolia et al., 2011; Chew et al., 2013). Thus, there is strong evidence emerging that a subset of lncRNA regions has a multitranscriptional output where their products get incorporated into diverse regulatory mechanisms (Yoon et al., 2013).

In regard to cellular location, there is about a twofold enrichment for lncRNAs in the nucleus compared to the cytoplasm (Derrien et al., 2012). Of course, this begs the question as to what these large numbers of lncRNAs are doing in the cytoplasm. This remains largely unknown at this point, but once elucidated it will undoubtedly advance the number of subcategories that exist and the diverse roles they play in the cell.

Characteristics of lincRNAs

A subset of lncRNAs includes those found in regions completely outside protein-coding genes and known as long intergenic noncoding RNA (lincRNA). Like the other types of lncRNAs, they share many regulatory features and characteristics of protein-coding genes. These genelike features include (1) their functioning as discrete transcriptional units with intron-exon boundaries, (2) alternative transcription start sites, (3) five prime capping and three prime polyadenylation of transcripts, (4) alternative exon splicing during transcript processing, (5) genelike promoters and regulatory elements that include the binding of a wide array of known transcription factors, (6) histone marks associated with actively expressed genes, (7) the ability to be posttranscriptionally modulated by miRNAs and to produce back-spliced exonic circular RNAs to titrate miRNA levels (described below), and (8) functional specificity in diverse cellular processes, contexts, tissues, developmental states, and cell lines (Guttman et al., 2009; Loewer et al., 2010; Cabili et al., 2011; Guttman et al., 2011; Ulitsky et al., 2011; Derrien et al., 2012; Geisler and Coller, 2013; Jalali et al., 2013; Krishnan and Mishra, 2013; Memczak et al., 2013; Paraskevopoulou et al., 2013; Ulitsky and Bartel, 2013). Another key factor highlighting the importance of lincRNAs to human health is the fact that about 50% of all human disease-related single nucleotide polymorphisms (SNPs) are located within intergenic regions (Hindorff et al., 2009).

So what are the key regulatory differences between lincRNA genes and protein-coding genes? First, there are an estimated twofold greater number of lincRNA genes compared to protein-coding sequences (Managadze et al., 2013). Although most lincRNA genes produce polyadenylated transcripts like protein-coding mRNAs, a small fraction of them contain alternative and novel three-prime topologies (Ulitsky and Bartel, 2013). The lincRNA genes also produce far fewer circular RNA transcripts derived from backspliced exons (Memczak et al., 2013). Circular RNAs composed of exons act as miRNA sponges in the cytoplasm,

titrating miRNA levels and modulating their binding activity to mRNAs. The lncRNAs, including lincRNAs, have also been implicated in being controlled by miRNAs as well as acting as miRNA decoys for the transcripts of protein-coding genes (Alaei-Mahabadi and Larsson, 2013; Jalali et al., 2013; Paraskevopoulou et al., 2013).

Yet another difference is that lincRNAs generally have fewer exons (2 to 3 on average) and their exons are longer, usually due to longer first and last exons (Derrien et al., 2012; Ulitsky and Bartel, 2013) compared to protein-coding genes that on average have about 10.7 exons (Cabili et al., 2011). The expression levels of different lincRNA genes vary widely, but the median activity is generally about one-tenth of protein-coding genes (Sigova et al., 2013; Ulitsky and Bartel, 2013). The regions encompassing lincRNA genes, including their transcripts, tend to contain larger amounts of transposable element sequence and repeats—a fact that also coincides with the knowledge that lincRNAs tolerate more variability than protein-coding genes (Ulitsky and Bartel, 2013). Finally, the expression of lincRNA genes tends to be more variable between cellular processes, contexts, tissues, developmental states, and cell lines than protein-coding genes, which indicates higher levels of transcriptional specificity (Guttman et al., 2011; Managadze et al., 2013; Sigova et al., 2013; Ulitsky and Bartel, 2013).

lincRNAs Are Taxonomically Restricted

Of greatest importance to the issue of human origins and the idea of universal common ancestry in general is that lincRNAs make logical substrates to test models of common ancestry for the following reasons. Despite their many critical functional roles, evolutionists are forced to believe that lincRNAs evolved far more rapidly than protein-coding mRNAs based on their much lower levels of sequence conservation compared to protein-coding genes (Marques and Ponting, 2009; Ulitsky and Bartel, 2013; Necsulea et al., 2014; Washietl et al., 2014). For example, less than 6% of zebrafish lincRNAs have any detectable DNA sequence similarity with human or mouse lincRNAs (Ulitsky et al., 2011). Even within closely related taxa, such as rodents, only ~50% of the mouse lincRNAs (expressed in liver) have alignable counterparts in rat—compared to ~90% of protein-coding mRNAs (Kutter et al., 2012). When Managadze et al. (2013) recently compared a 53,649 human lincRNA dataset to a mouse data set of 43,638 lincRNAs, there was shared homology for only 32% of the dataset's transcripts (100 bases of overlap was required as a threshold to denote a set of transcripts as orthologous between taxa).

Differences between human and chimpanzee long noncoding RNAs were originally most notably characterized in what was termed "human accelerated regions" (HAR). These comprised several hundred regions over 100 bases in length that contained high levels of putative substitutions, but the areas only represented highly homologous sequences that were at least 96% identical (Pollard et al., 2006a; Pollard et al., 2006b). Even with these small differences, however, it was discovered that the secondary structures produced in these noncoding RNAs were markedly different between humans and chimpanzees (Beniaminov et al., 2008).

In an early study using high-throughout genomics, expression patterns of both protein-coding genes and intergenic regions were compared between humans and chimpanzees using human microarrays, which by nature excluded the hybridization of chimpanzee sequences not highly homologous to human (Khaitovich et al., 2006). Nevertheless, they found that about 50% of the homologous expressed sequences in brain, heart, testis, and lymphoblastoid cell lines that contributed to differences between humans and chimps were intergenic noncoding RNAs—emphasizing their equal importance in contributing to taxonomic expression differences (compared to protein-coding genes).

In another study, the brain transcriptomes were compared between human, chimpanzee, and macaque, using an early variant of RNA-seq technology that produced very short reads of only ~36 bases (Xu et al., 2010). While the researchers discovered that approximately 40 to 48% of expressed brain sequences in humans originated from intronic and intergenic regions, very little information was provided as to the exact amount of differences in numbers of unique transcripts that existed between humans and chimps. The repetitive nature of these short reads rich in transposable element features likely prohibited their effective assembly into discrete transcripts. However, the researchers were able to compare the expression patterns of homologous sequences, omitting the taxonomically restricted transcripts. For these homologous transcripts among humans and apes, they found that the intergenic regions were largely conserved in their brain expression patterns across taxa, but less so than protein-coding regions.

More recently, several reports have compared lncRNA expression in a wide variety of tissues between humans, primates, and other mammals of which lincRNAs were a subset group. In one study, it was found that only 47% of expressed human lncRNAs were conserved across primates (chimpanzees, gorillas, orangutans, macaques) and only 28% were found to have homologs across non-marsupial mammals—i.e., eutherians (Necsulea et al., 2014). The results led the authors to state that "lncRNA transcription evolves rapidly," reflecting their evolutionary assumption of common ancestry. Yet another interesting result of the study was that the promoter regions of lncRNA genes preferentially bound over twice as of-

ten to homeobox transcription factors than protein-coding genes. Homeobox transcription factors are key regulators functioning in development.

In the other recent study, expression of 1,898 human lincRNAs was evaluated in human, chimpanzee, macaque, cow, mouse, and rat (Washietl et al., 2014). For three of the tissues, they could only "find orthologous transcripts for 80% in chimpanzee, 63% in rhesus, 39% in cow, 38% in mouse and 35% in rat." They also state, "Remarkably, we find that approximately 20% of human lincRNAs are not expressed beyond chimpanzee and are undetectable even in rhesus." Compared to protein-coding genes, they also claim that these human lincRNAs are "faster-evolving within the human lineage," meaning that much of the lincRNA sequence appears suddenly with no evolutionary history in apes. Both the hypothesized rapid divergence of these functional sequences and the sudden "appearance" of lincRNA genes in separate lineages are intractable problems for the evolutionary paradigm.

While a variety of reports have illustrated the differences in lincRNA expression for limited sets of transcripts between humans and chimpanzees in various tissues, none have actually compared the lincRNA genomic regions in humans from which they are derived. At the time of this report, no comprehensive comparison of the transcribed intergenic regions of the human genome compared to chimpanzee, the alleged closest living relative to humans, exists. This is despite the fact that several studies have been recently completed in humans extensively characterizing these regions. In one report, researchers used RNA-seq technology to compile a catalog of over 8,000 human lincRNAs derived from 24 different cell lines and tissue types that were strikingly tissue specific in their expression patterns, compared to protein-coding genes (Cabili et al., 2011). Data produced in this study form the bulk of sequences available at the Broad Institute lincRNA catalog (broadinstitute.org). In a more recent study, researchers compiled an even larger list of more than 58,000 human lincRNAs that included sequences derived from many novel intergenic regions of the human genome expressed at very low levels and were thus missed in previous studies (Hangauer et al., 2013). In addition, at the UCSC genome browser, a compiled set of about 22,000 lincRNAs entries exist for version hg19 of the human genome.

Interestingly, a novel study on human intergenic expressed sequences was recently published in which the researchers characterized a class of "very long intergenic noncoding RNAs," which they termed vlincRNA (St Laurent et al., 2013). In this new study, 2,147 different vlincRNAs were discovered, sequenced, and assembled. These vlincRNAs only overlap with lincRNAs by about 10% and form a completely novel class of intergenic sequence estimated to cover about 10% of the entire human genome. The vlincRNAs are much longer than protein-coding genes and standard lincRNAs and are believed to play key roles in chromatin remodeling and nuclear architecture related to gene expression. When the vlincRNAs were evaluated in a variety of cell types, they were found to be associated with cell identity, developmental states, and cancer, thus illustrating their importance to human cell and tissue development and overall health.

Given the high level of importance that the transcribed intergenic regions of the human genome play in virtually all types of cells and tissues studied to date, combined with the high levels of taxonomically restricted expression patterns they exhibit (compared to protein-coding genes), they were chosen as targets for a comparative study with the chimpanzee genome. This was done to further clarify and define the issue of human-chimp DNA sequence similarity in the human origins debate.

Methods

The four different sources used to develop query datasets are as follows: the human lincRNA catalog at the Broad Institute of MIT and Harvard (broadinstitute.org/genome_bio/human_lincrnas/?q=lincRNA_catalog) which contained 14,402 entries and largely corresponds to the study published by Cabili et al. (2011), the complete lincRNA data set from Hangauer et al. (2013) containing 58,537 sequences, the UCSC lincRNA gene tracks (downloaded Dec., 2013), and the vlincRNA dataset from St Laurent III et al. (2013). Oddly, the 2013 dataset from Hangauer et al. was based on the hg18 version of the human genome last updated in 2006, while the other data sets used hg19 (the most recent version). All data sets except for the UCSC lincRNA tracts (which were downloaded using the UCSC table browser), were each originally obtained in BED file format as indicated in their respective publications or database sites, which included genome coordinates for each sequence. Perl scripts I had written extracted the genomic sequence for each coordinate from the UCSC genome browser en masse, corresponding to whatever version of the human genome was used to originally set the coordinates (hg18 or hg19), saving them as FASTA format files with header lines for each sequence containing the corresponding BED file data. Genomic sequences were also parsed into new FASTA files based on individual sequence lengths using a Perl script I had written for the purpose of creating optimized BLASTN datasets.

Human protein-coding exons from all chromosomes, 300 to 599 bases in length, were downloaded from the hg19 version of the human genome at ucsc.genome.edu, using appropriate parameters in the table browser feature. These were obtained in FASTA format and queried against the chimpanzee genome and the human genome as a comparative control using the BLASTN parameters described below.

The lincRNA regions less than 300 bases in length provided unreliable BLAST results when compared against the human genome as a control and were thus omitted from the analyses. The lincRNA regions between 300 and 599 bases in length were used directly for BLASTN analyses, while lincRNA and vlincRNA regions 600 bases and longer were subjected to sequence slicing using a Python script I had written and described previously (Tomkins, 2013). Basic statistical analyses for the genomic sequences in the human lincRNA and vlincRNA query sets were done using a Perl script I had written creating the data shown in Table I.

The most recent versions of the chimpanzee (CHIMP2.1.4.71), and the human genomes (GRCh37.71/hg19) were downloaded from ftp.ensembl.org/pub. Human genome version hg18, for control testing of the Hangauer et al. linc hg18 annotated lincRNA dataset was downloaded from hgdownload.soe.ucsc.genome.edu. The various genome assemblies were then used to make individual BLAST databases using the makeblastdb tool. Batch BLASTN jobs were deployed on UNIX and Linux servers as described previously (Tomkins, 2013). BLASTN results were outputted as CSV format text files and parsed and analyzed via an integrated set of Python and POSIX shell scripts I had written. BLASTN algorithm parameters were as follows: -word_size 11, -evalue 10, -max_target_seqs 1, -dust no, -soft_masking false, -ungapped. These optimized parameters were chosen largely on the results of Tomkins (2011) and Tomkins (2013) and also preliminary analyses performed in this study.

Results

Four different long intergenic noncoding DNA data sets were used for this project: (1) the human lincRNA catalog at the Broad Institute of MIT (broadinstitute.org/genome_bio/human_lincrnas/?q=lincRNA_catalog), which contained 14,402 entries and largely corresponds to the study published by Cabili et al. (2011); (2) the complete lincRNA data set from Hangauer et al. (2013) that was demarcated based on the coordinates of human genome version hg18 and comprises 58,537 sequences; (3) the lincRNA entries at the UCSC genome browser for version hg19 of the human genome; and (4) the vlincRNA dataset from St Laurent III et al. (2013). The MIT and St Laurent III et al. datasets were also based on hg19 version of the human genome. Sequence statistics for each of these datasets can be viewed in Table I. Individual entries in each dataset were composed of the entire contiguous lincRNA or vlincRNA genomic region minus the promoter. All genomic data was downloaded from the UCSC genome browser using the BED file genome coordinates provided in the supplementary information of each published paper or listed on the respective databases (see Methods section for details).

As a comparative reference for the lincRNA and vlincRNA regions, all human protein coding exons between 300 and 599 bases in length also were utilized via extraction from the UCSC genome table browser (version hg19). The protein-coding exons of the human genome are arguably the most similar in sequence identity to chimpanzee, whose alignable regions have been selectively used by evolutionists in a wide variety of comparative studies (Tomkins and Bergman, 2012).

In regard to comparing lincRNA sequence between taxa, the following problem was recently noted in a review by Ulitsky and Bartel (2013), in which they stated, "Existing approaches for comparing genomic sequences, which rely heavily on stretches of high sequence conservation, might be poorly suited for detecting homology between lincRNAs" (pp 34–35). Previous research using a wide variety of BLASTN algorithm parameters showed that alignments of human-chimpanzee genomic DNA broke down significantly after only several hundred bases on average, terminating the extension of the algorithm (Tomkins, 2011). To overcome this limitation, Tomkins (2013) devised a strategy of sequence slicing to produce multiple datasets comprised of differ-

Table I. Sequence characteristics of the human lincRNA and vlincRNA genomic regions for each data set used in this study.

Data set source	Type of intergenic sequence	Number of sequences	Mean length (bases)	Median length (bases)	Minimum length (bases)	Maximum length (bases)
Cabili et al./MIT (2011)	lincRNA	14,402	15,403	5,363	256	603,040
UCSC hg19 (Dec 8, 2013)	lincRNA	21,629	19,117	6672	256	690,433
Hangauer et al. (2013)	lincRNA	58,537	1,788	511	202	373,456
St Laurent III et al. (2013)	vlincRNA	2,762	130,566	83,866	50,002	1,104,100

ent slice sizes representing the original contiguous sequence. Each set of slices is then BLASTed against the target database and the optimal output of multiple experiments is selected as an accurate indicator of overall sequence similarity. This strategy effectively overcomes the limitations imposed by large insertions and deletions that disrupt pairwise BLASTN comparisons across large genomic regions. In addition, this strategy also overcomes lack of synteny (linear order of genomic features) for alleged rearrangements of sequence. This strategy was used successfully to determine the overall sequence similarity for individual chromosomes in the chimpanzee genome compared to their homologous human counterparts (Tomkins, 2013).

Preliminary studies with all lincRNA datasets showed that sequences between 300 and 599 bases in length could be effectively aligned without sequence slicing (data not shown). The lincRNA regions more than 600 bases in length were treated as a single large genomic file and sliced into a range of sub files. The most recent version of the chimpanzee genome downloaded from Ensembl.org (CHIMP2.1.4.71) was used as the target database. To evaluate the amount of sequence that may have been lost in the process of concatenating and slicing, query sets were also BLASTed against the version of the human genome from which they were derived (hg18 or hg19). The amount of sequence lost as a caveat of concatenation and slicing was minimal, (0.0 to 1.3%) and was factored back into the similarity estimates achieved.

Basic sequence statistics for each data set are listed in Table I. The human lincRNA genomic regions from the MIT and UCSC datasets were heavily enriched for larger transcripts—only about 3% were less than 600 bases in length. In contrast, the more extensive lincRNA data set of 58,537 genomic sequences from Hangauer et al. (2013) was heavily enriched for regions of the genome encoding shorter transcripts, and 57% of the sequences were less than 600 bases. The Hangauer et al. data purportedly also represents a large number of newly characterized transcripts expressed at very low levels in the cell.

The shorter lincRNA regions of the human genome (300 to 599 bases) were 75 to 79% similar to chimpanzee, depending on the dataset (Table II). Given that slightly over half of the Hangauer et al. dataset consisted of lincRNA regions less than 600 bases, the best estimate of short lincRNA region similarity would probably be represented by this data. Eleven percent of the short human lincRNA regions in this data set were completely missing in the chimpanzee genome. This same percentage was also reflected in the two other data sets as well.

The larger lincRNA regions had to be subjected to optimized sequence slicing to ascertain their overall similarity to chimpanzee. The identity of the top aligning sets of slices for these experiments indicated that these longer lincRNA encoding regions of the human genome are only 71 to 74% identical chimpanzee (Table II). Clearly, the longer types of human lincRNA regions of the genome are slightly less similar to chimpanzee than the shorter segments.

For the vlincRNA dataset representing the regions of the human genome transcribed into very long noncoding RNAs (50,000 and 1,104,100 bases in length), the DNA sequence identity compared to chimpanzee was only 67% for the optimal aligning set of subsequences. Much of the dissimilarity was due to large segments of the vlincRNA genes present in human and missing in

Table II. BLASTN results for each data set using the chimpanzee genome (Ensembl ver chimpv2.1.4.71).

Data set source	Type of intergenic sequence	Sequence identity	Optimal sequence slice and range tested (bases)*
Cabili et al. (2011)/MIT	lincRNA 300-599 bases	75.3%	—
UCSC hg19 (Dec 8, 2013)	lincRNA 300-599 bases	75.5%	—
Hangauer et al. (2013)	lincRNA 300-599 bases	78.8%	—
Cabili et al. (2011)/MIT	lincRNA 600+ bases	72.1%	250 (200-450)
UCSC hg19 (Dec 8, 2013)	lincRNA 600+ bases	71.0%	250 (200-450)
Hangauer et al. (2013)	lincRNA 600+ bases	73.9%	250 (200-400)
St Laurent III et al. (2013)	vlincRNA	67.0%	450 (250-500)
Protein coding exons	300-599 bases	86.5%	—

* Subset query files were based on 50 base increments (e.g. 200, 250, 300, etc).

chimpanzee. The optimum alignment length was 450 bases, and approximately 29% of these segments had no match in chimp. The vlincRNA regions of the human genome represent a completely separate class of intergenic expressed regions and only overlap with lincRNA regions by an estimated 10% (St Laurent et al., 2013). It is believed that their function is primarily associated with chromatin modifying scaffolds that regulate genome function and architecture.

As a comparative control, all known human protein-coding exons from version hg19 of the human genome between 300 and 599 bases in length were BLASTed against the chimpanzee genome. Overall, DNA similarity was only 86%—a number that includes the results that approximately 6.3% of human protein-coding exons in this size range are completely missing in the chimpanzee genome. The exons that did align were 91.8% identical on average. Overall, the noncoding transcribed intergenic regions of the human genome are about 7 to 19% less similar to chimpanzee than protein-coding exons. The general trend is that the shorter noncoding transcribed intergenic regions tend be more similar on average than the longer regions. The vlincRNA regions are the most dissimilar.

Summary and Discussion

For years, the standard axiom has promoted the idea that humans are 98% genetically identical to chimpanzees. However, this dogmatic statement about the DNA similarity between humans and chimps is based on cherry-picked data from short, aligned segments of high similarity and omits the regions that are vastly different. The leading human and chimpanzee DNA comparison studies published by evolutionists during the past decade were recently reviewed and critiqued (Tomkins and Bergman, 2012). In every single report, the researchers selected highly similar DNA sequence data and discarded other data because it would not readily align. In fact, when the DNA similarities from these papers were recalculated using omitted data for the alignments, markedly lower levels of similarity were found that varied between 70 and 86%. Even the rough draft of the chimpanzee genome published in 2005 provides an overall genomic similarity of only about 70 to 80% when the discarded non-similar data is included (Tomkins and Bergman, 2012; Tomkins, 2013).

Much of the reported human-chimp DNA similarity data is due in part to the inherent BLASTN algorithm restrictions associated with aligning chimpanzee genomic sequence onto human and vice versa. In a recent study, a wide variety of BLASTN algorithm parameters were tested using 40,000 740-base long segments of chimpanzee genomic DNA (preselected to be homologous to human by NCBI) that were queried against four different versions of the human genome (Tomkins, 2011). The algorithm parameter combinations that produced the longest alignments gave similarities of 86% and the algorithm stopped aligning after only a few hundred bases on average, due to extreme dissimilarity between the genomes.

The phenomenon of high levels of human-chimp genomic discontinuity was first noted by evolutionists in the initial stages of sequencing the chimpanzee genome. Researchers produced over 3 million bases of chimp genomic sequence (60 to 950 bases per read) and then BLASTed them against the human genome (Ebersberger et al., 2002). The report stated that only "About two thirds could be unambiguously aligned to DNA sequences in humans" (p. 1490). The researchers also set their BLASTN parameters to omit DNA less than 98% identical and did not report the amount of each read not aligning, just that only two-thirds of them did.

Clearly a more informative technique was required to compare the chimpanzee genome to that of humans to provide estimates of DNA similarity over long genomic distances. Specifically, a technique was needed to counteract the problem of the BLASTN algorithm breaking off the alignment extension in regions of low similarity. By digitally slicing entire chimp chromosomes into small pieces, Tomkins (2013) found that the BLASTN algorithm could effectively compare chimp DNA to human piece-by-piece by testing a range of sub-slice datasets and then selecting the highest sequence identity output. The same technique was used in this study to compare the transcribed intergenic regions of the human genome to chimpanzee.

Research is showing that the mysterious whereabouts of information underpinning organismal complexity is not entirely associated with just the basic protein-coding gene sets. Instead, much of this important information is located in the highly functional, noncoding portions of the genome; and as organismal complexity increases, so does the amount and complexity of transcribed intergenic noncoding RNA (Liu et al., 2013). The main points concerning the noncoding portions of genomes can be summarized as follows: (1) Any given human or animal genome is a complete storehouse of important information, and this fact negates the concept of "Junk DNA." (2) Protein-coding genes are largely a basic set of instructions within a complex and larger expressed repertoire of both regulatory and structural noncoding DNA sequence.

Related to these emerging concepts about noncoding DNA is the fact that the transcribed intergenic regions of the genome contain much higher levels of taxonomically restricted DNA sequence, compared to the exonic protein-coding segments (Ponjavic et al., 2007; Ulitsky et al., 2011; Managadze et al., 2013). Previous research comparing these intergenic noncoding regions of the human genome to chimpanzee is based on studies using selected tissue and cell line

transcriptomic data sets. While these studies compared only a small fraction of the human intergenic transcriptomes, it was found that noncoding transcripts were significantly more taxonomically restricted than protein-coding ones (Khaitovich et al., 2006; Xu et al., 2010; Necsulea et al., 2014; Washietl et al., 2014). At present, research exhaustively comparing the regions of the human genome producing long intergenic transcripts to chimpanzee has not been done.

This report describes the comparative use of three different human lincRNA datasets and one vlincRNA genomic dataset to the chimpanzee genome using the BLASTN algorithm under parameters previously shown to provide optimal alignments (Tomkins, 2013). Short human lincRNA regions (less than 600 bases) are about 75–79% similar to chimpanzee while the larger lincRNA regions (greater than 600 bases) are about 71 to 74% similar. The human vlincRNA genomic regions are only 67% similar to chimpanzee. To provide a comparative contrast, all human protein-coding exons 300 to 599 bases in length were also queried against the chimpanzee genome, and found to be 86% similar to chimpanzee. Overall, the noncoding transcribed intergenic regions of the human genome are about 7 to 19% less similar to chimpanzee than protein-coding exons.

One point of particular interest is that the long (greater than 600 bases) lincRNA and vlinc RNA regions were markedly different, and their putative function appears to be related to large-scale chromatin modification. The implications are that significant RNA-mediated chromosomal and nuclear architecture differences between humans and chimpanzees may also be an important contributor to functional genomic differences.

The DNA similarity results from this study fit well with a previous report in which the chimpanzee chromosomes were sequentially compared to human chromosomes using the same technique of sequence slicing (Tomkins, 2013). Not counting the Y-chromosome, chimpanzee chromosome similarities compared to human varied between 66 and 78%. Overall, the chimp genome was only 70% identical on average to human. In addition, these current results also correlate well with a recent study of 1,898 human lincRNA genes expressed in a variety of tissues in which only 80% had counterparts expressed in chimp tissue (Washietl et al., 2014).

The real genome-wide differences between chimps and humans are too vast to be explained by hypothetical evolutionary processes. The regions that are similar between chimps and humans are easily interpreted as repetitions of effective design themes associated with code reuse, a concept that is very familiar to software designers and engineers. DNA sequence comparisons that include all the relevant data clearly show that the human and chimpanzee genomes are not nearly identical but instead are as different as one might expect based on the clearly observed phenotypic discontinuities.

References

Alaei-Mahabadi, B., and E. Larsson. 2013. Limited evidence for evolutionarily conserved targeting of long non-coding RNAs by microRNAs. *Silence* 4(1): 4.

Amort T., M.F. Soulière, A. Wille, X-Y. Jia, H. Fiegl, H. Wörle, R. Micura, and A. Lusser. 2013. Long non-coding RNAs as targets for cytosine methylation. *RNA Biology* 10(6): 1003–1008.

Beniaminov, A., E. Westhof, and A. Krol. 2008. Distinctive structures between chimpanzee and human in a brain noncoding RNA. *RNA* 14(7): 1270–1275.

Birney, E., J.A. Stamatoyannopoulos, A. Dutta, R.Guigo, T. Gingeras, E. Margulies, Z. Weng, D. Snyder, E. Dermitzakis, R. 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.

Cabili, M.N., C. Trapnell, L. Goff, M. Koziol, B. Tazon-Vega, A. Regev, and J. Rinn. 2011. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev* 25(18): 1915–1927.

Chew, G., A. Pauli, J. Rinn, A. Regev, A. Schier, and E. Valen. 2013. Ribosome profiling reveals resemblance between long non-coding RNAs and 5' leaders of coding RNAs. *Development* 140(13): 2828–2834.

Clark, M., P. Amaral, F. Schlesinger, M. Dinger, R. Taft, J. Rinn, C. Ponting, P. Stadler, K. Morris, A. Morillon, et al. 2011. The reality of pervasive transcription. *PLoS Biol* 9(7): e1000625; discussion e1001102.

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

Derrien, T., R. Johnson, G. Bussotti, A. Tanzer, S. Djebali, H. Tilgner, G. Guernec, D. Martin, A. Merkel, D. Knowles, et al. 2012. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Research* 22(9): 1775–1789.

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

Ebersberger, I., D. Metzler, C. Schwarz, and S. Pääbo. 2002. Genomewide comparison of DNA sequences between humans and chimpanzees. *American Journal of Human Genetics* 70(6): 1490–1497.

Geisler, S., and J. Coller. 2013. RNA in unexpected places: long non-coding RNA functions in diverse cellular contexts. *Nature Reviews Molecular Cell Biology* 14(11): 699–712.

Guttman, M., I. Amit, M. Garber, C. French, M. Lin, D. Feldser, M. Huarte, O. Zuk, B. Carey, J. Cassady, et al. 2009. Chromatin signature reveals over a thousand highly conserved large non-coding

RNAs in mammals. *Nature* 458(7235): 223–227.

Guttman, M., J. Donaghey, B. Carey, M. Garber, J. Grenier, G. Munson, G. Young, A. Lucas, R. Ach, L. Bruhn, et al. 2011. lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature* 477(7364): 295–300.

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

He, S., H. Su, C. Liu, G. Skogerbo, H. He, D. He, X. Zhu, T. Liu, Y. Zhao, and R. Chen. 2008. MicroRNA-encoding long non-coding RNAs. *BMC Genomics* 9:236.

Hindorff, L., P. Sethupathy, H. Junkins, E. Ramos, J. Mehta, F. Collins, and T. Manolio. 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America* 106(23): 9362–9367.

Ingolia, N., L. Lareau, and J. Weissman. 2011. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* 147(4): 789–802.

Jalali, S., G. Jayaraj, and V. Scaria. 2012. Integrative transcriptome analysis suggest processing of a subset of long non-coding RNAs to small RNAs. *Biology Direct* 7:25.

Jalali, S., D. Bhartiya, M. Lalwani, S. Sivasubbu, and V. Scaria. 2013. Systematic transcriptome wide analysis of lncRNA-miRNA interactions. *PloS One* 8(2): e53823.

Johnson, J., S. Edwards, D. Shoemaker, and E. Schadt. 2005. Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments. *Trends in Genetics: TIG* 21(2): 93–102.

Kapranov, P., and G. St Laurent. 2012. Dark matter RNA: existence, function, and controversy. *Frontiers in Genetics* 3:60.

Khaitovich, P., J. Kelso, H. Franz, J. Visagie, T. Giger, S. Joerchel, E. Petzold, R. Green, M. Lachmann, and S. Paabo. 2006. Functionality of intergenic transcription: an evolutionary comparison. *PLoS Genetics* 2(10): e171.

Khalil, A., M. Guttman, M. Huarte, M. Garber, A. Raj, D. Rivea Morales, K. Thomas, A. Presser, B. Bernstein, A. van Oudenaarden, et al. 2009. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proceedings of the National Academy of Sciences of the United States of America* 106(28): 11667–11672.

Krishnan, J., and R. Mishra. 2013. Emerging trends of long non-coding RNAs in gene activation. *The FEBS Journal* 281:34–45.

Kutter, C., S. Watt, K. Stefflova, M. Wilson, A. Goncalves, C. Ponting, D. Odom, and A. Marques. 2012. Rapid turnover of long noncoding RNAs and the evolution of gene expression. *PLoS Genetics* 8(7): e1002841.

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

Loewer, S., M. Cabili, M. Guttman, Y. Loh, K. Thomas, I. Park, M. Garber, M. Curran, T. Onder, S. Agarwal, et al. 2010. Large intergenic non-coding RNA-RoR modulates reprogramming of human induced pluripotent stem cells. *Nature Genetics* 42(12): 1113–1117.

Magny, E.G., J. Pueyo, F. Pearl, M. Cespedes, J. Niven, S. Bishop, and J. Couso. 2013. Conserved regulation of cardiac calcium uptake by peptides encoded in small open reading frames. *Science* 341(6150): 1116–1120.

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

Marques, A., and C. Ponting. 2009. Catalogues of mammalian long noncoding RNAs: modest conservation and incompleteness. *Genome Biology* 10(11): R124.

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

Mercer, T., and J. Mattick. 2013. Structure and function of long noncoding RNAs in epigenetic regulation. *Nature Structural & Molecular Biology* 20(3): 300–307.

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

Ohsawa, R., J. Seol, and J. Tyler. 2013. At the intersection of non-coding transcription, DNA repair, chromatin structure, and cellular senescence. *Frontiers in Genetics* 4:136.

Paraskevopoulou, M., G. Georgakilas, N. Kostoulas, M. Reczko, M. Maragkakis, T. Dalamagas, and A. Hatzigeorgiou. 2013. DIANA-LncBase: experimentally verified and computationally predicted microRNA targets on long non-coding RNAs. *Nucleic Acids Research* 41(Database issue): D239–245.

Pollard, K., S. Salama, B. King, A. Kern, T. Dreszer, S. Katzman, A. Siepel, J. Pedersen, G. Bejerano, R. Baertsch, et al. 2006a. Forces shaping the fastest evolving regions in the human genome. *PLoS Genetics* 2(10): e168.

Pollard K., S. Salama, N. Lambert, M. Lambot, S. Coppens, J. Pedersen, S. Katzman, B. King, C. Onodera, A. Siepel, et al. 2006b. An RNA gene expressed during cortical development evolved rapidly in humans. *Nature* 443(7108): 167–172.

Ponjavic, J., C. Ponting, and G. Lunter. 2007. Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Research* 17(5): 556–565.

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

Sauvageau, M., L. Goff, S. Lodato, B. Bonev, A. Groff, C. Gerhardinger, D. Sanchez-Gomez, E. Hacisuleyman, E. Li, M.

Spence, et al. 2013. Multiple knockout mouse models reveal lincRNAs are required for life and brain development. *eLife* 2: e01749.

Sigova, A., A. Mullen, B. Molinie, S. Gupta, D. Orlando, M. Guenther, A. Almada, C. Lin, P. Sharp, 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 of the United States of America* 110(8): 2876–2881.

Squires, J., H. Patel, M. Nousch, T. Sibbritt, D. Humphreys, B. Parker, C. Suter, and T. Preiss. 2012. Widespread occurrence of 5-methylcytosine in human coding and non-coding RNA. *Nucleic Acids Research* 40(11): 5023–5033.

St Laurent III, G., D. Shtokalo, B. Dong, M. Tackett, X. Fan, S. Lazorthes, E. Nicolas, N. Sang, T. Triche, T. McCaffrey, et al. 2013. VlincRNAs controlled by retroviral elements are a hallmark of pluripotency and cancer. *Genome Biology* 14(7): R73.

Tomkins, J. 2011. Genome-wide DNA alignment similarity (identity) for 40,000 chimpanzee DNA sequences queried against the human genome is 86–89%. *Answers Research Journal* 4:233–241.

Tomkins, J. 2013. Comprehensive analysis of chimpanzee and human chromosomes reveals average DNA similarity of 70%. *Answers Research Journal* 6:63–69.

Tomkins, J., and J. Bergman. 2012. Genomic monkey business—estimates of nearly identical human-chimp DNA similarity reevaluated using omitted data. *Journal of Creation* 26:94–100.

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

Ulitsky, I., A. Shkumatava, C. Jan, H. Sive, and D. Bartel. 2011. Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. *Cell* 147(7): 1537–1550.

Washietl, S., M. Kellis, and M. Garber. 2014. Evolutionary dynamics and tissue specificity of human long noncoding RNAs in six mammals. *Genome Research*. doi/10.1101/gr.165035.113.

Wijaya, E., M. Frith, P. Horton, and K. Asai. 2013. Finding protein-coding genes through human polymorphisms. *PloS One* 8(1): e54210.

Xu, A., L. He, Z. Li, Y. Xu, M. Li, X. Fu, Z. Yan, Y. Yuan, C. Menzel, N. Li, et al. 2010. Intergenic and repeat transcription in human, chimpanzee and macaque brains measured by RNA-Seq. *PLoS Computational Biology* 6: e1000843.

Yoon, J., K. Abdelmohsen, and M. Gorospe. 2013. Posttranscriptional gene regulation by long noncoding RNA. *Journal of Molecular Biology* 425(19): 3723–3730.