RESEARCH ARTICLE

Human Mutation

OFFICIAL JOURNAL
HGVS
HUMAN GENOME VARIATION SOCIETY
www.hgvs.org

Genome-Wide Analysis of Human SNPs at Long Intergenic Noncoding RNAs

Geng Chen,[1†] Chengxiang Qiu,[1,2†] Qipeng Zhang,[1] Bing Liu,[2*] and Qinghua Cui[1,3,4*]

[1]*Department of Biomedical Informatics, Peking University School of Basic Medical Sciences, Beijing, China;* [2]*LIAMA Center for Computational Medicine, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China;* [3]*Institute of Systems Biomedicine, Peking University, Beijing, China;* [4]*MOE Key Lab of Molecular Cardiovascular Sciences, Peking University, Beijing, China*

Communicated by Michael Dean

Received 3 July 2012; accepted revised manuscript 2 October 2012.
Published online 28 November 2012 in Wiley Online Library (www.wiley.com/humanmutation). DOI: 10.1002/humu.22239

ABSTRACT: Long intergenic noncoding RNAs (lincRNAs) represent a large portion of the noncoding genes in mammals and other eukaryotes but remains among the least well-understood of genetic factors to date. Here, we systematically analyzed the human SNPs of lincRNAs at a genome level. We found a significantly lower SNP density in lincRNA regions than both their upstream and downstream flanking regions. Functional regions show lower SNP density than other regions in lincRNAs. We revealed that lincRNAs with higher expression levels and broader expression spectrum have significantly lower SNP density. Moreover, we identified lincRNAs that are under recent positive selection and revealed that these lincRNAs show distinct SNP density, expression level, and tissue specificity. Importantly, we identified a genetic variant (rs7990916:T>C) under recent positive selection at a brain-specific lincRNA that significantly affects the structure of normal brain. Analysis of brain magnetic resonance images showed that individuals with CC genotype have significant bigger regional gray matter volume than individuals with TT genotype. Moreover, the genotype of this SNP shows different distribution in normal elders, mild cognitive impairment, and Alzheimer disease subjects, suggesting that this lincRNA may have a role in physiology and pathophysiology of human brain.

Hum Mutat 34:338–344, 2013.

KEY WORDS: long intergenic noncoding RNA; SNP; Alzheimer disease; brain

Introduction

One surprising result of the sequencing of the human genome is that only a small part (~1%) of the human genomic sequence encodes proteins [Birney et al., 2007]. Furthermore, analysis of human and mouse transcriptomes revealed that protein-coding sequences only account for a small portion of the genome transcripts [Bertone et al., 2004; Birney et al., 2007; Cheng et al., 2005; Kapranov et al., 2007]. Currently, it is known that the noncoding transcripts have a broad spectrum of noncoding RNA molecules including miRNAs and long intergenic noncoding RNAs (lincRNAs). miRNAs have shown their critical importance in a number of biological processes [Bartel, 2004] and a broad spectrum of human disease [Lu, 2008]. However, miRNAs only represent a small portion of the noncoding genes and the majority of noncoding RNAs are lincRNAs [Kapranov et al., 2007].

The lincRNAs are defined as the intergenic noncoding RNAs with more than 200 nucleotides. It is known that lincRNAs are poorly conserved across species [Cabili et al., 2011; Chodroff et al., 2010; Church et al., 2009]. For example, only ~14% of the mouse long noncoding RNAs have human orthologs [Church et al., 2009]. A recent study on a larger set of lincRNAs revealed that ~12% of human lincRNAs have orthologous transcripts in other species [Cabili et al., 2011]. Therefore, some researchers may argue against their functionality. However, increasing studies have shown that lincRNAs have important functions and therefore have roles in disease although they tend to be less conserved across species and often show low expression levels and high tissue specificity [Mercer et al., 2008; Ponting et al., 2009]. For example, by inducing a repressive chromatin state, Hox transcript antisense RNA (*HOTAIR*) silences transcription across 40 kb of the HOXD locus in trans [Rinn et al., 2007], and the reprogramming of chromatin state by *HOTAIR* can promote cancer metastasis [Tian et al., 2010]. Recent evidences indicated that lincRNAs can also act as scaffolds to recruit protein partners in unique combinations [Adkins et al., 2010; Reich et al., 2003]. The lincRNAs *Xist* and *Jpx* are key players in mediating *X* chromosome inactivation [Tian et al., 2010]. The above evidences have shown that lincRNAs have a huge complexity and diversity in functions; however, there is still a huge gap in the understanding of the lincRNAs, which still remain among the least-understood biological molecules.

SNPs represent the most frequent genetic changes in the human genome, and SNPs occurring in functional regions may be associated with phenotype changes and disease susceptibility [Reich et al., 2003]. Therefore, SNPs occurring in lincRNAs may affect lincRNA functions and subsequently result in phenotype changes. Indeed, increasing evidences have reported that SNPs at lincRNAs are

Additional Supporting Information may be found in the online version of this article.

†These authors contributed equally to this work.

*Correspondences to: Bing Liu, LIAMA Center for Computational Medicine, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China. E-mail: bliu@nlpr.ia.ac.cn; Qinghua Cui, Department of Biomedical Informatics, Peking University School of Basic Medical Sciences, Beijing 100191, China. E-mail: cuiqinghua@hsc.pku.edu.cn

Contract grant sponsors: National Basic Research program of China (2012CB517500); Natural Science Foundation of China (31000585, 81000582).

associated with a number of phenotypes and diseases, such as birth and body weight [Adkins et al., 2010; Petry et al., 2005; Zhang et al., 2006], Wilms' tumors [Yuan et al., 2005], myocardial infarction [Ishii et al., 2006], nonmuscle-invasive bladder cancer [Verhaegh et al., 2008], DNA methylation [Braunschweig et al., 2011], prostate cancer [Jin et al., 2011], and photoperiod-sensitive male sterility [Ding et al., 2012]. However, a global view of the human SNPs at lincRNAs remains largely limited. For miRNAs, another class of noncoding RNAs, systematic analysis on SNPs at miRNAs and their target sites have already revealed a number of findings [Bao et al., 2007; Bhartiya et al., 2011; Chen and Rajewsky, 2006; Duan et al., 2009; Gong et al., 2011; Hariharan et al., 2009; Hiard et al., 2010; Iwai and Naraba, 2005; Landi et al., 2008; Lu, 2008; Saunders et al., 2007; Yu et al., 2007; Zhang, 2008; Ziebarth et al., 2011], such as low SNP density of miRNA genes and target sites [Chen and Rajewsky, 2006; Saunders et al., 2007], aberrant allele frequencies of the SNPs in cancer [Yu et al., 2007], and low SNP incidence in disease associated miRNAs [Lu, 2008]. These studies do great helps to the understanding of miRNA function and roles in disease. We believe that such an analysis of SNPs at human lincRNAs will be helpful to the understanding of their functions and roles in diseases.

To have a global view of polymorphisms at human lincRNAs, here we conducted a genome-wide analysis of human SNPs in lincRNAs. We revealed the patterns of SNP density of lincRNAs as well as their correlation with expression level and tissue specificity of lincRNAs. We identified a number of candidate lincRNAs that are under recent positive selection and uncovered their distinct features in SNP density, expression level, and tissue specificity. Moreover, we identified a SNP that is correlated with brain structure and is associated with AD susceptibility.

Materials and Methods

Identifying SNPs in lincRNAs and their Flanking Regions

We obtained the genomic coordinate data (hg19) of human lincRNAs ($n = 21630$) from UCSC [Karolchik et al., 2003] and downloaded the human SNP data from the latest version of dbSNP (build 135). Then, we mapped SNPs that occur in lincRNAs. For comparison, we also identified SNPs in three upstream (and three downstream) regions with the same size as the given lincRNAs. We next calculated SNP densities for lincRNAs and their flanking regions. We obtained genomic coordinate data (hg19) of human protein-coding genes from UCSC [Karolchik et al., 2003] and human miRNAs from miRBase [Griffiths-Jones, 2004].

Expression Data of Human lincRNAs

We downloaded the expression profiles of lincRNAs across 22 tissues and cell lines from UCSC [Karolchik et al., 2003], which was estimated based on RNA-Seq [Trapnell et al., 2010]. We then added the expression levels in all tissues of each lincRNA as the expression level of this lincRNA. On the basis of the lincRNA expression data, we then calculated the tissue-specific index for each lincRNA using the program developed by Lu (2008). We take the lincRNAs with tissue specificity index ≥ 0.7 as tissue-specific lincRNAs.

Identifying Candidate lincRNAs Under Recent Positive Selection

We downloaded integrated Haplotype Score (iHS), a statistic used to detect evidence of recent positive selection at a locus, from Haplotter [Voight et al., 2006]. Haplotter presented iHS data for three populations, ASN (combined Japanese from Tokyo, Japan, and Han Chinese from Beijing, China), CEU (Utah residents with Northern and Western European ancestry from the CEPH collection), and YRI (Yorubans from Ibadan, Nigeria). We retrieved the SNPs with extreme iHS scores (iHS ≥ 2.5 or iHS ≤ -2.5) and then mapped these SNPs to lincRNAs. These lincRNAs are integrated to this study as candidate lincRNAs that are under recent positive selection.

Identifying the Association of One Candidate SNP (rs7990916:T>C) and Regional Gray Matter Volume of Human Brain in Alzheimer Disease Neuroimaging Initiative Database

Genotype and Brain Imaging Data used here were obtained from the Alzheimer Disease Neuroimaging Initiative (ADNI) database (www.loni.ucla.edu/ADNI) in October 2011. Structural magnetic resonance imaging (MRI) scans were acquired from 1.5 T scanners at multiple sites across the United States and Canada. MRI protocols ensured comparability across a variety of scanners (GE, Siemens, or Philips). Original scans and preprocessed images, and genome-wide genotype data are available at http://adni.loni.ucla.edu/. To investigate the association of SNPs at lincRNAs with brain function and structure, we focus on the SNPs that are brain-specific lincRNAs and are under recent positive selection in Europe population. As a result, we focused on one SNP (rs7990916:T>C) that has an absolute iHS value of 2.51 and whose lincRNAs has a brain tissue specificity of 0.996. By optimized-VBM (Voxel based morphometry) method, we computed the voxel-vised gray matter volume in the whole brain and then yielded regional cortical gray matter volumes by segmenting the whole brain into 116 brain anatomic regions (Supp. Table S1) based on the Automated Anatomical Labeling atlas. We then investigate the association of this SNP and regional cortical gray matter volume in normal subjects. We also investigate the distribution of the genotype of this SNP in normal elders, MCI, and AD subjects.

Statistical Analysis

All statistical analysis including Wilcoxon test, *t* test, and correlation analysis are performed on R (http://cran.r-project.org/), a free statistical software.

Results

Low Polymorphism in Human lincRNAs

There are 21,630 human lincRNAs in UCSC and more than 54 million human SNPs in dbSNP (build 135). By mapping SNPs onto the genomic coordinates of lincRNAs, we identified 2,618,623 SNPs in 6,637 lincRNAs. As a result, the lincRNAs shows a SNP density of $\approx$6.3 SNPs per kb. As a comparison, flanking regions exhibit a higher SNP density (Fig. 1A; Table 1). Moreover, the flanking regions that are more far from the lincRNAs tend to show higher SNP density. It is known that some lincRNAs host small RNAs, for example miRNAs [Jalali et al., 2012]. We also found that 66 SNPs are located in 27 miRNAs of the lincRNAs that host miRNAs.

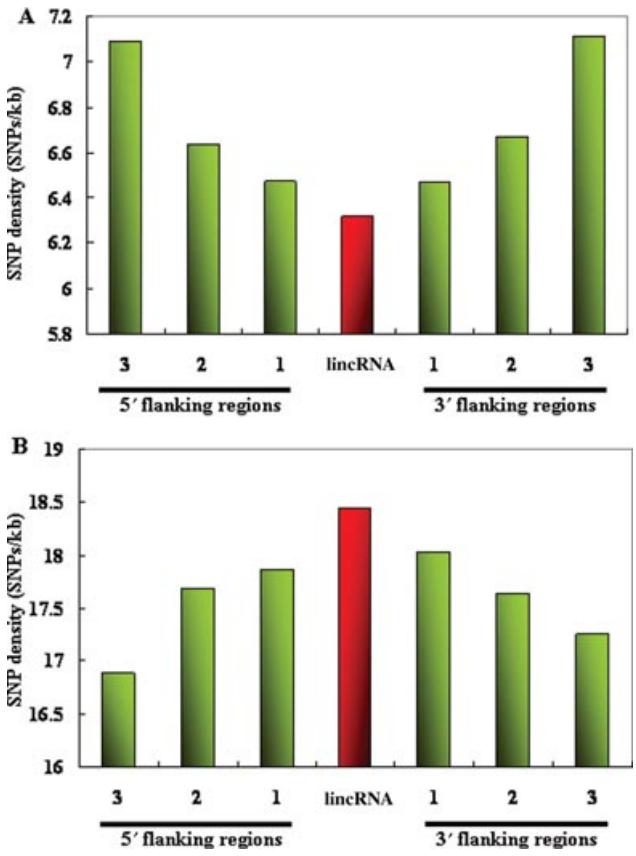

Figure 1. SNPs in human lincRNAs. **A:** SNP density in all available human lincRNAs and flanking regions. Flanking regions 1–3 represent successive nonoverlapping windows that are equal to the size of the given lincRNAs immediately adjacent to the lincRNAs. **B:** SNP density in human lincRNAs that are under recent positive selection.

Functional Motifs Show Lower SNP Density than Other Regions in Human lincRNAs

It is expected functional motifs of lincRNAs tend to show lower SNP density than other regions. In a recent study, Jayaraj et al. (2012) revealed that lincRNAs also have G-quadruplexes, one type of functional motif through computational analysis. Therefore, it is interesting to compare SNP densities of G-quadruplexes regions and other regions in lincRNAs. To do this, we identified 535 lincRNAs that have G-quadruplexes motifs using the pipeline provided by Scaria et al. (2006). We then analyzed SNP distribution of these regions and other regions in lincRNAs. As a result, we found significant difference between the two types of regions. Regions of G-quadruplexes show lower SNP density than other regions in lincRNAs (11.9 SNPs/kb vs. 17.3 SNPs/kb, $P = 5.4 \times 10^{-4}$). This result suggests that functional motifs in lincRNAs generally show lower SNP densities than other regions in lincRNAs. Furthermore, considering that exon regions can represent more functional content compared with the intron regions, we compared the SNP densities of exon regions and intron regions of lincRNAs at the genome-wide level. As expected, the exon regions show lower SNP density than the intron regions (5.85 SNPs/kb vs. 6.35 SNPs/kb, P value $= 5.5 \times 10^{-10}$), suggesting that exons may be under greater evolutionary constraint than introns.

Polymorphism in Human lincRNAs is Correlated with the Expression of lincRNAs

The expression profile of a gene can provide clues for understanding its functions. For example, house-keeping genes often show high and broad expression spectrum, suggesting they are critically important. For lincRNAs, they normally show low expression level and high tissue specificity [Ponting et al., 2009]. However, the relationship between their expression and polymorphism remains unknown. To dissect the patterns behind lincRNA expression and polymorphism, here we first classified lincRNAs into 10 groups with approximately the same number of lincRNAs according to their expression level. We next calculated SNP density for each group of lincRNs. As a result, the SNP density of lincRNAs is negatively correlated with the expression level of lincRNAs ($R = -0.99$, $P < 2.2 \times 10^{-16}$; Fig. 2). Highly expressed lincRNAs tend to have lower SNP density, whereas lowly expressed lincRNAs tend to have higher SNP density. For example, the SNP density of the lincRNAs whose expression level is lower than 100 is 11.1 SNPs/kb, which is nearly fivefold of the SNP density (2.3 SNPs/kb) of the lincRNA whose expression level is higher than 10,000. Moreover, the flanking regions show similar patterns with lincRNAs (Fig. 2). Interestingly, lincRNAs even have more similar SNP density with flanking regions when compared with lincRNAs in other groups, suggesting that lincRNAs show functional relationship with neighbor regions in a large scale. From Figure 2, we can see that the SNP density of the lincRNAs is sharply increased when the expression level decreases less than 1,000 (Fig. 2). For protein-coding genes, it is known that gene tissue specificity tends to be positively correlated its evolutionary rate [Winter et al., 2004]. It remains unknown whether lincRNAs have the similar consistency as protein-coding genes. We observed a negative correlation between expression level and tissue specificity ($R = -0.85$, $P < 2.2 \times 10^{-16}$), suggesting that lincRNAs with higher tissue specificity tend to have higher SNP density. Indeed, we observed a significant positive correlation between tissue specificity and SNP density ($R = 0.91$, $P = 4.7 \times 10^{-4}$; Supp. Fig. S1). For example, the SNP density of the lincRNAs whose tissue

Table 1. Summary of the SNPs in lincRNAs and their Flanking Regions (5′:1 Stands for the 5′ Flanking Region 1, and so on)

	5′ Flanking regions				3′ Flanking regions		
	3	2	1	lincRNA regions	1	2	3
Num_SNP[a]	2937077	2749992	2680580	2618623	2678329	2763190	2945766
Num_entry[b]	7267	6887	6676	6634	6716	6910	7361
SNP_density[c]	6.469991	6.637528	7.089086	6.320448	6.464558	6.669383	7.110058
P value[d]	4.40E–10	5.02E–3	0.3781	n/a	0.0855	5.55E–4	2.71E–13

[a] Num_SNP indicates the number of SNPs occurring corresponding regions.
[b] Num_entry indicates the number of lincRNAs genes or their flanking regions that have at least one SNP.
[c] SNP_density indicates the number of SNPs per kb in corresponding regions.
[d] P value indicates the P value of SNP density comparison between lincRNA region and each flanking region (Wilcoxon test).

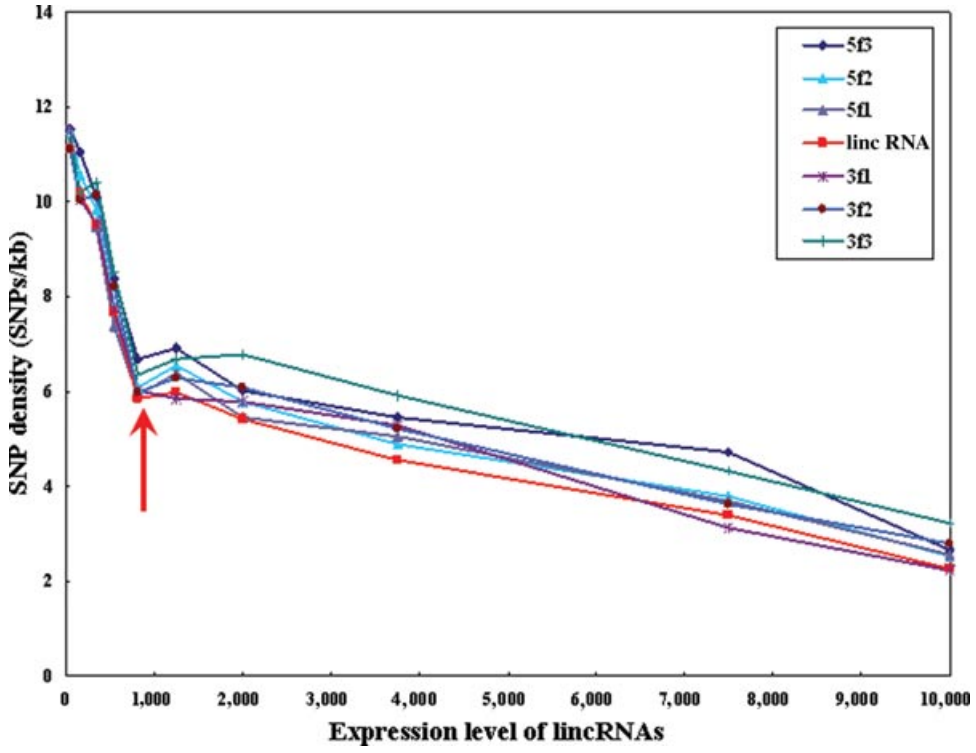

Figure 2. Correlations between expression level and SNP density for lincRNAs and flanking regions.

specificity index is higher than 0.95 is 7.8 SNPs/kb, which is 2.6-folds of the SNP density (3.0 SNPs/kb) of the lincRNA whose tissue specificity index is less than 0.1. We further compared the lincRNAs with expression level >1,000 (highly expressed lincRNAs) and those with expression level ≤1000 (lowly expressed lincRNAs). Dramatically, 71.3% (7,206/10,109) of the lowly expressed lincRNAs are tissue specific, whereas only 2.9% (292/10,125) of the highly expressed lincRNAs are tissue specific. Moreover, 96.1% (7,206/7,498) of the tissue-specific lincRNAs are lowly expressed. Two groups have significantly differences in abundance of tissue-specific lincRNAs ($P < 2.2 \times 10^{-16}$, OR = 83.6).

Human lincRNAs Under Recent Positive Selection Show Distinct Features

Most of the human polymorphisms are neutral and do not cause functional changes but some polymorphism may lead to functional differences. As a result, the advantageous SNPs may benefit population survival and therefore may be under recent positive selection. The iHS is a powerful static for the testing of recent positive selection. We obtained iHS scores for all available candidate SNPs from Haplotter [Voight et al., 2006]. As suggested, a SNP with |iHS| value ≥2.5 is considered to be under recent positive selection [Voight et al., 2006]. Although most of the SNPs have a |iHS| value <2.5, a number of SNPs exhibit high |iHS| values (|iHS| value ≥ 2.5) in the three populations (1,004 SNPs in 311 lincRNAs for ASN; 1,229 SNPs in 344 lincRNAs for CEU; 1,224 SNPs in 443 lincRNAs for YRI), suggesting that these locus may be under recent positive selection. Allele frequencies of these loci tend to show differences across populations. Indeed, we observed a higher fraction of different distribution of alleles in SNPs under recent positive selection compared to SNPs with low |iHS| score (|iHS| score ≤ 0.01) ($P < 2.2 \times 10^{-16}$, Chi-squared test). We next analyzed the SNP density for the lincRNAs that are under recent positive selection in at least two populations. Interestingly, the SNP density in these lincRNAs is higher than that in flanking regions (Fig. 1B), a converse pattern with the global SNP density.

For all three populations, we revealed a negative correlation between incidence of lincRNAs under recent positive selection and expression level ($R = -0.98$, $P < 2.2 \times 10^{-16}$ for ASN; $R = -0.98$, $P < 2.2 \times 10^{-16}$ for ASN; $R = -0.98$, $P < 2.2 \times 10^{-16}$ for CEU; $R = -0.96$, $P < 2.2 \times 10^{-16}$ for YRI). The lincRNAs with lower expression level tend to be more under recent positive selection (Supp. Fig. S2). Moreover, we found differences in lincRNAs specifically expressed in different tissues (Supp. Fig. S3). For human lincRNAs, testes and brain are the top two tissues having the most tissue-specific lincRNAs, and represent 74.8% and 4.6% of all the tissue-specific lincRNAs, respectively. In all three populations, we found that brain-specific lincRNAs tend to be more under recent positive selection than average (Fig. 3A), however, fewer testes-specific lincRNAs tend to be under recent positive selection (Fig. 3B), significant difference exists between brain-specific lincRNAs and testes-specific lincRNAs ($P = 0.004$, t test). This suggests that specific tissue may affect recent positive selection of lincRNAs.

A SNP under Recent Positive Selection is Associated with Regional Gray Matter Volume and the Susceptibility of AD

We choose the SNPs that are under recent positive selection in Europe population and whose lincRNAs are brain specific and have an expression level ≥1,000 to scan the ADNI database. As a result, we identified a brain-structure-related SNP (rs7990916:T>C), which is located in the lincRNA, TCONS_00021856 (Fig. 4A). This lincRNA has a total length of 148,213 bp and has two exons, which are 130 and 627 bp in length, respectively (Fig. 4A). Using BLAST tool, we found

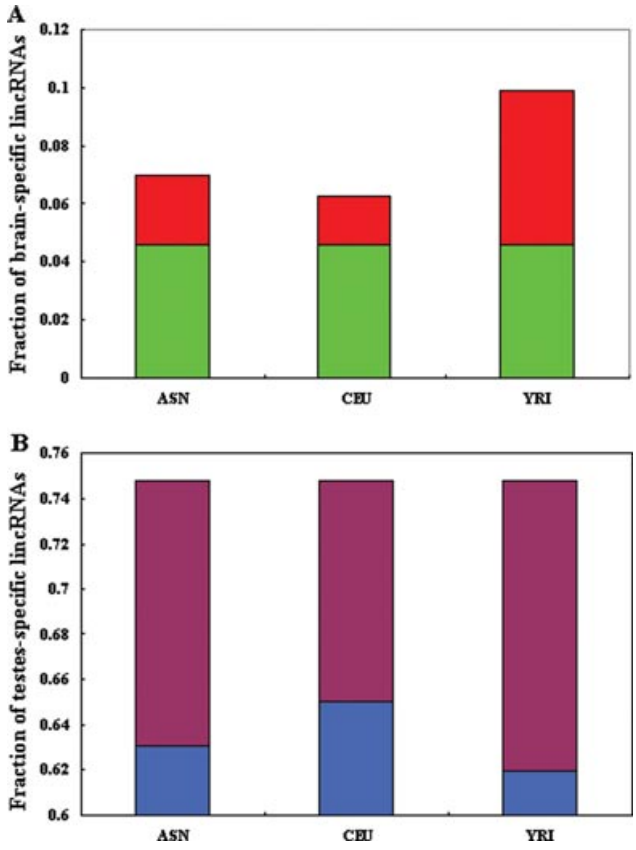

Figure 3. Fraction of brain-specific lincRNAs (**A**) and testes-specific lincRNAs (**B**) to total tissue-specific lincRNAs. For (**A**), the green bars represent the fraction of brain-specific lincRNAs to total tissue-specific lincRNAs; the whole bars represent the fraction of brain-specific lincRNAs to total tissue-specific lincRNAs for the lincRNAs that are under recent positive selection. For (**B**), the whole bars represent the fraction of testes-specific lincRNAs to total tissue-specific lincRNAs; the blue bars represent the fraction of testes-specific lincRNAs to total tissue-specific lincRNAs under recent positive selection for the lincRNAs that are under recent positive selection.

highly conserved sequences in chimpanzee, gorilla, orangutan, gibbon, rhesus, and marmoset, but did not find conserved sequences in other species, indicating that this lincRNA is specific to primate. Moreover, we revealed that the first exon of this lincRNA is perfectly matched with three expressed sequence tags (ESTs) (Fig. 4B). They are DA225206.1 from the brain tissue, DA306307.1 from the hippocampus tissue, and DA346981.1 from the substantia nigra tissue. All three ESTs are from brain-related tissues, suggesting that this lincRNA tend to be associated with brain structure and function.

We further performed analysis of the association of this SNP and brain structure. Individuals with CC have significantly bigger gray matter volume than individuals with TT for 40 brain regions ($P < 0.05$, false discovery rate corrected) (Fig. 5A and Supp. Table S2). Moreover, these 40 brain regions mainly fall into the temporal cortex (Fig. 5B). In addition, the gray matter volume of some regions in frontal cortex, parietal cortex, occipital cortex, and cerebellum cortex are also affected by this genotype (Fig. 5B, Supp. Table S2). These findings suggest that this SNP may be associated with the cortical development of memory-related brain regions. For example, medial temporal cortex and posterior cingulate cortex are known to be associated with memory [Jeneson and Squire, 2011] and play roles in pathophysiology of AD [Schmidt-Wilcke et al., 2009]. This further suggests that this SNP may be involved in the pathophysiology of AD. To test this hypothesis, we further analyzed the genotype distribution of this SNP in normal population, and patients with MCI and AD based on ADNI. As expected, the distribution of the genotype frequency of this SNP in populations of NC (normal control), MCI, and AD is significantly different ($P = 0.005$, Chi-square test), suggesting that the lincRNA hosting this SNP is associated with the pathophysiology of MCI and AD (Supp. Table S3).

Validation of the Main Finding Using an Unbiased SNP Dataset

The SNP data used for the above analysis is from dbSNP135. This dataset is the most comprehensive dataset for SNP but integrated SNP data from various projects. Therefore, the data in dbSNP135 may be biased. To investigate whether the main findings are resulted

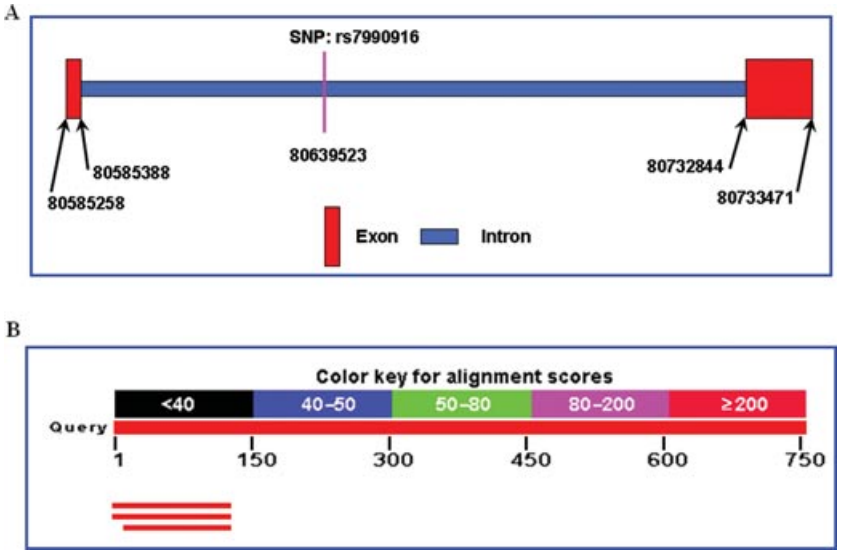

Figure 4. The gene structure of lincRNA *TCONS_00021856* and the location of the SNP *rs7990916:T>C* (**A**); BLAST of lincRNA *TCONS_00021856* with human EST database identified three brain-derived ESTs matched with the first exon of this lincRNA (**B**).

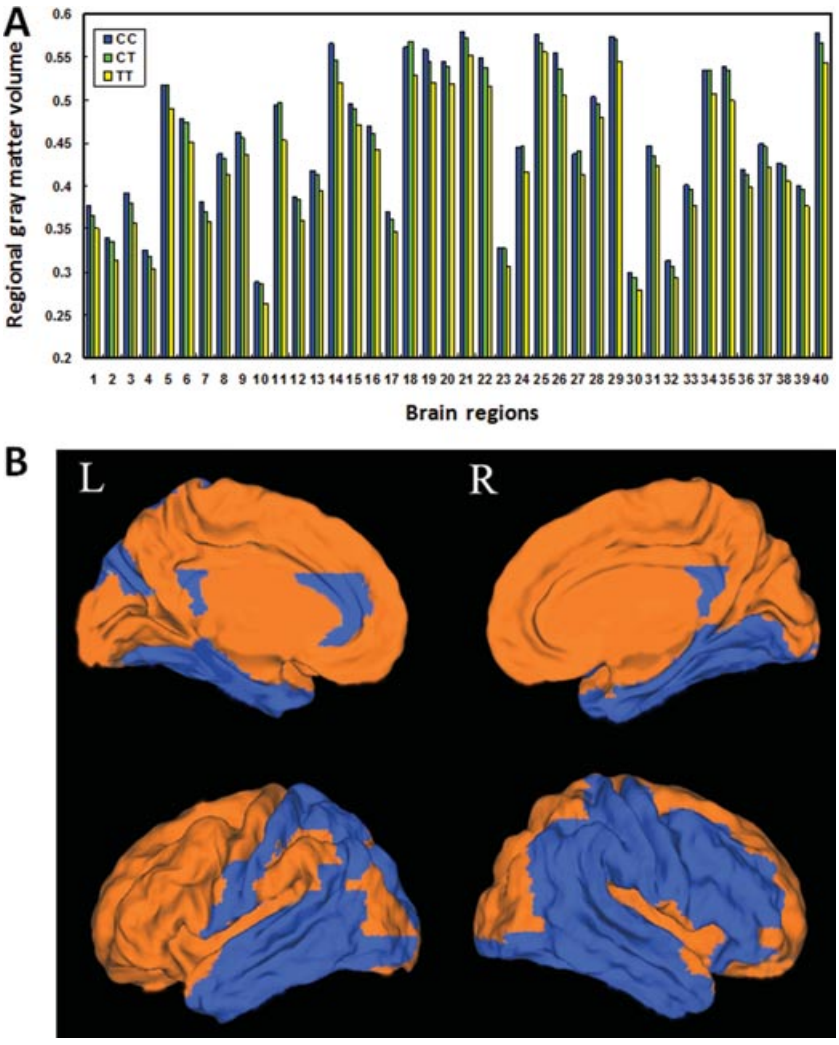

Figure 5. Genotypes of the SNP (rs7990916:T>C) are significantly associated with regional gray matter volume of 40 brain regions (**A**), which are highlighted in blue color in the brain map (**B**). The affected brain regions are mapped on the cerebral cortex using CARET software. The top figures are shown in medial views and the bottom figures are shown in lateral views.

from the potential bias of the data, we reperformed the main analysis based on an unbiased SNP dataset, the SNP data from an Asian individual (Yanhuang, YH) [Wang et al., 2008]. As a result, the main findings keep unchanged (Supp. Results). The YH SNP density in human lincRNAs and flanking regions is shown in Supp. Figure S4. The correlation between expression level and YH SNP density for lincRNAs is shown in Supp. Figure S5. The correlation between tissue specificity and YH SNP density for lincRNAs is shown in Supp. Figure S6. The YH SNP density in human lincRNAs that are under recent positive selection is shown in Supp. Figure S7.

Discussion

Increasing evidences have shown that lincRNAs are functional, however, lincRNAs are among one of the most unknown genetic factors to date. One of the major reasons is that currently data for lincRNAs remains greatly limited. The currently available data of SNP, genomic coordinate, and expression profile of lincRNAs give us an opportunity to have a global view of the polymorphism of human lincRNAs, which will provide clues for their function and associated disease.

We revealed that SNPs occurring in lincRNAs is relatively lower than flanking regions, suggesting that the regions of lincRNAs are under more stringent functional constraint. Functional regions of lincRNAs show lower SNP density than other regions in lincRNAs. Moreover, we found significant correlations between SNP density and expression level/tissue specificity for lincRNAs, suggesting that expression profile has a significant effect on lincRNA polymorphism. Interestingly, lincRNA expression level leads to a greater difference in polymorphism of lincRNAs than the flanking regions, suggesting that lincRNAs may be more functionally related with their flanking regions other than lincRNAs with significantly different expression profiles. We identified a number of candidate SNPs and lincRNAs that have extreme high |iHS| scores, suggesting that these lincRNAs may be under recent positive selection. We also revealed a correlation between lincRNA recent positive selection and their expression level. We further found that the majority of the lincRNAs under recent positive selection are tissue specific. Moreover, we identified a SNP that is associated with brain gray matter volume of normal population and the MCI and AD, suggesting that SNPs at

lincRNAs could have effects on the structure of human's brain and is associated with the pathophysiology of neurodegeneration disease.

In summary, combing data of SNP, lincRNA genomic coordinates, and lincRNA expression profile, we uncovered a number of patterns of polymorphisms in human lincRNAs. The findings are helpful to the understanding of lincRNA origin, evolution, expression, function, and associated disease. We believe that more findings for lincRNAs will be observed when more data become available in the future.

Acknowledgment

We thank Vinod Scaria for kindly providing us the pipeline to predict G-quadruplexes of lincRNAs.

References

Adkins RM, Somes G, Morrison JC, Hill JB, Watson EM, Magann EF, Krushkal J. 2010. Association of birth weight with polymorphisms in the IGF2, H19, and IGF2R genes. Pediatr Res 68(5):429–434.

Bao L, Zhou M, Wu L, Lu L, Goldowitz D, Williams RW, Cui Y. 2007. PolymiRTS database: linking polymorphisms in microRNA target sites with complex traits. Nucleic Acids Res 35(Database issue):D51–D54.

Bartel DP. 2004. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell 116(2):281–297.

Bertone P, Stolc V, Royce TE, Rozowsky JS, Urban AE, Zhu X, Rinn JL, Tongprasit W, Samanta M, Weissman S, Gerstein M, Snyder M. 2004. Global identification of human transcribed sequences with genome tiling arrays. Science 306(5705):2242–2246.

Bhartiya D, Laddha SV, Mukhopadhyay A, Scaria V. 2011. miRvar: a comprehensive database for genomic variations in microRNAs. Hum Mutat 32(6):E2226–E2245.

Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE, Kuehn MS, Taylor CM, 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.

Braunschweig MH, Owczarek-Lipska M, Stahlberger-Saitbekova N. 2011. Relationship of porcine IGF2 imprinting status to DNA methylation at the H19 DMD and the IGF2 DMRs 1 and 2. BMC Genet 12:47.

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

Chen K, Rajewsky N. 2006. Natural selection on human microRNA binding sites inferred from SNP data. Nat Genet 38(12):1452–1456.

Cheng J, Kapranov P, Drenkow J, Dike S, Brubaker S, Patel S, Long J, Stern D, Tammana H, Helt G, Sementchenko V, Piccolboni A, et al. 2005. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. Science 308(5725):1149–1154.

Chodroff RA, Goodstadt L, Sirey TM, Oliver PL, Davies KE, Green ED, Molnar Z, Ponting CP. 2010. Long noncoding RNA genes: conservation of sequence and brain expression among diverse amniotes. Genome Biol 11(7):R72.

Church DM, Goodstadt L, Hillier LW, Zody MC, Goldstein S, She X, Bult CJ, Agarwala R, Cherry JL, DiCuccio M, Hlavina W, Kapustin Y, et al. 2009. Lineage-specific biology revealed by a finished genome assembly of the mouse. PLoS Biol 7(5):e1000112.

Ding J, Lu Q, Ouyang Y, Mao H, Zhang P, Yao J, Xu C, Li X, Xiao J, Zhang Q. 2012. A long noncoding RNA regulates photoperiod-sensitive male sterility, an essential component of hybrid rice. Proc Natl Acad Sci USA 109(7):2654–2659.

Duan S, Mi S, Zhang W, Dolan ME. 2009. Comprehensive analysis of the impact of SNPs and CNVs on human microRNAs and their regulatory genes. RNA Biol 6(4):412–425.

Gong J, Tong Y, Zhang HM, Wang K, Hu T, Shan G, Sun J, Guo AY. 2011. Genome-wide identification of SNPs in microRNA genes and the SNP effects on microRNA target binding and biogenesis. Hum Mutat 33(1):254–263.

Griffiths-Jones S. 2004. The microRNA Registry. Nucleic Acids Res 32(Database issue):D109–D111.

Hariharan M, Scaria V, Brahmachari SK. 2009. dbSMR: a novel resource of genome-wide SNPs affecting microRNA mediated regulation. BMC Bioinformatics 10:108.

Hiard S, Charlier C, Coppieters W, Georges M, Baurain D. 2010. Patrocles: a database of polymorphic miRNA-mediated gene regulation in vertebrates. Nucleic Acids Res 38(Database issue):D640–D651.

Ishii N, Ozaki K, Sato H, Mizuno H, Saito S, Takahashi A, Miyamoto Y, Ikegawa S, Kamatani N, Hori M, Saito S, Nakamura Y, et al. 2006. Identification of a novel non-coding RNA, MIAT, that confers risk of myocardial infarction. J Hum Genet 51(12):1087–1099.

Iwai N, Naraba H. 2005. Polymorphisms in human pre-miRNAs. Biochem Biophys Res Commun 331(4):1439–1444.

Jalali S, Jayaraj GG, Scaria V. 2012. Integrative transcriptome analysis suggest processing of a subset of long non-coding RNAs to small RNAs. Biol Direct 7(1):25.

Jayaraj GG, Pandey S, Scaria V, Maiti S. 2012. Potential G-quadruplexes in the human long non-coding transcriptome. RNA Biol 9(1):81–86.

Jeneson A, Squire LR. 2011. Working memory, long-term memory, and medial temporal lobe function. Learn Mem 19(1):15–25.

Jin G, Sun J, Isaacs SD, Wiley KE, Kim ST, Chu LW, Zhang Z, Zhao H, Zheng SL, Isaacs WB, Xu J. 2011. Human polymorphisms at long non-coding RNAs (lncRNAs) and association with prostate cancer risk. *Carcinogenesis* 32(11):1655–1659.

Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, Willingham AT, Stadler PF, Hertel J, Hackermuller J, Hofacker IL, Bell I, Cheung Ea, et al. 2007. RNA maps reveal new RNA classes and a possible function for pervasive transcription. Science 316(5830):1484–1488.

Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, Lu YT, Roskin KM, Schwartz M, Sugnet CW, Thomas DJ, Weber RJ, Haussler D, et al. 2003. The UCSC Genome Browser Database. Nucleic Acids Res 31(1):51–54.

Landi D, Gemignani F, Barale R, Landi S. 2008. A catalog of polymorphisms falling in microRNA-binding regions of cancer genes. DNA Cell Biol 27(1):35–43.

Lu M, Zhang Q, Deng M, Miao J, Guo Y, Gao W, Cui Q. 2008. An analysis of human microRNA and disease associations. PLoS ONE 3(10):e3420; doi:10.1371/journal.pone.0003420.

Mercer TR, Dinger ME, Sunkin SM, Mehler MF, Mattick JS. 2008. Specific expression of long noncoding RNAs in the mouse brain. Proc Natl Acad Sci USA 105(2):716–721.

Petry CJ, Ong KK, Barratt BJ, Wingate D, Cordell HJ, Ring SM, Pembrey ME, Reik W, Todd JA, Dunger DB. 2005. Common polymorphism in H19 associated with birthweight and cord blood IGF-II levels in humans. BMC Genet 6:22.

Ponting CP, Oliver PL, Reik W. 2009. Evolution and functions of long noncoding RNAs. Cell 136(4):629–641.

Reich DE, Gabriel SB, Altshuler D. 2003. Quality and completeness of SNP databases. Nat Genet 33(4):457–458.

Rinn JL, Kertesz M, Wang JK, Squazzo SL, Xu X, Brugmann SA, Goodnough LH, Helms JA, Farnham PJ, Segal E, Chang HY. 2007. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell 129(7):1311–1323.

Saunders MA, Liang H, Li WH. 2007. Human polymorphism at microRNAs and microRNA target sites. Proc Natl Acad Sci USA 104(9):3300–3305.

Scaria V, Hariharan M, Arora A, Maiti S. 2006. Quadfinder: server for identification and analysis of quadruplex-forming motifs in nucleotide sequences. Nucleic Acids Res 34(Web Server issue):W683–W685.

Schmidt-Wilcke T, Poljansky S, Hierlmeier S, Hausner J, Ibach B. 2009. Memory performance correlates with gray matter density in the ento-/perirhinal cortex and posterior hippocampus in patients with mild cognitive impairment and healthy controls—a voxel based morphometry study. Neuroimage 47(4):1914–1920.

Tian D, Sun S, Lee JT. 2010. The long noncoding RNA, Jpx, is a molecular switch for X chromosome inactivation. Cell 143(3):390–403.

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28(5):511–515.

Verhaegh GW, Verkleij L, Vermeulen SH, den Heijer M, Witjes JA, Kiemeney LA. 2008. Polymorphisms in the H19 gene and the risk of bladder cancer. Eur Urol 54(5):1118–1126.

Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006. A map of recent positive selection in the human genome. PLoS Biol 4(3):e72.

Wang J, Wang W, Li R, Li Y, Tian G, Goodman L, Fan W, Zhang J, Li J, Zhang J, Guo Y, Feng B, et al. 2008. The diploid genome sequence of an Asian individual. Nature 456(7218):60–65.

Winter EE, Goodstadt L, Ponting CP. 2004. Elevated rates of protein secretion, evolution, and disease among tissue-specific genes. Genome Res 14(1):54–61.

Yu Z, Li Z, Jolicoeur N, Zhang L, Fortin Y, Wang E, Wu M, Shen SH. 2007. Aberrant allele frequencies of the SNPs located in microRNA target sites are potentially associated with human cancers. Nucleic Acids Res 35(13):4535–4541.

Yuan E, Li CM, Yamashiro DJ, Kandel J, Thaker H, Murty VV, Tycko B. 2005. Genomic profiling maps loss of heterozygosity and defines the timing and stage dependence of epigenetic and genetic events in Wilms' tumors. Mol Cancer Res 3(9):493–502.

Zhang Q, Lu, M, Cui Q. 2008. SNP analysis reveals an evolutionary acceleration of the human-specific microRNAs. Nature Precedings. Accessed, at: http://hdl.handle.net/10101/npre.2008.2127.1. Accessed 1 November 2012.

Zhang W, Maniatis N, Rodriguez S, Miller GJ, Day IN, Gaunt TR, Collins A, Morton NE. 2006. Refined association mapping for a quantitative trait: weight in the H19-IGF2-INS-TH region. Ann Hum Genet 70(Pt 6):848–856.

Ziebarth JD, Bhattacharya A, Chen A, Cui Y. 2011. PolymiRTS Database 2.0: linking polymorphisms in microRNA target sites with human diseases and complex traits. Nucleic Acids Res 40(Database issue):D216–D221.