G&I Genomics & Informatics

eISSN 2234-0742
Genomics Inform 2017;15(4):170-177
https://doi.org/10.5808/GI.2017.15.4.170

ORIGINAL ARTICLE

HOTAIR Long Non-coding RNA: Characterizing the Locus Features by the *In Silico* Approaches

Mohammadreza Hajjari*, Saghar Rahnama

Department of Genetics, Shahid Chamran University of Ahvaz, Ahvaz 61336-3337, Iran

HOTAIR is an lncRNA that has been known to have an oncogenic role in different cancers. There is limited knowledge of genetic and epigenetic elements and their interactions for the gene encoding *HOTAIR*. Therefore, understanding the molecular mechanism and its regulation remains to be challenging. We used different *in silico* analyses to find genetic and epigenetic elements of *HOTAIR* gene to gain insight into its regulation. We reported different regulatory elements including canonical promoters, transcription start sites, CpGIs as well as epigenetic marks that are potentially involved in the regulation of *HOTAIR* gene expression. We identified repeat sequences and single nucleotide polymorphisms that are located within or next to the CpGIs of *HOTAIR*. Our analyses may help to find potential interactions between genetic and epigenetic elements of *HOTAIR* gene in the human tissues and show opportunities and limitations for researches on *HOTAIR* gene in future studies.

Keywords: bioinformatics, CpG Islands, epigenetics, gene expression, *HOTAIR*

Introduction

It has been estimated that about 1.5% of human genomic DNA can be annotated as protein coding sequences [1]. So, more than 98% of the human genome does not encode protein [2, 3]. However, a large proportion of the genome transcribes non-coding RNAs such as miRNAs and long non-coding RNAs (lncRNAs) [4, 5]. LncRNAs have important roles in different cellular and molecular mechanisms [6]. These long RNAs regulate the activity and position of epigenetic machinery during cell function and segregation [7]. In fact, some of the lncRNAs can recruit catalytic activity of chromatin-modifying proteins [8]. Dysregulation of lncRNAs has been also reported in cancer initiation and progression. However, the molecular mechanism and regulation of these RNAs have been remained to be unknown [9, 10].

Rinn *et al*. [11] identified *HOTAIR* lncRNA with a 2.2 kb length. *HOTAIR* gene is located in a region between *HOX11* and *HOX12* on chromosome 12q13.3 [12-16]. *HOTAIR* lncRNA binds to both polycomb repressive complex 2 (PRC2) and lysine specific demethylase 1 (LSD1) complexes, through its 5'-3' domains and directs them to *HOXD* gene cluster as well as other genes in order to increase gene silencing by coupling the histone H3K27 trimethylation and H3K4 demethylation [17, 18].

HOTAIR is an oncogene RNA that is known to have potential role in several cancers. Its overexpression is reported in different solid tumors such as breast, gastric, and colorectal tumors [19, 20]. The oncogenic role of *HOTAIR* is reported in different mechanisms such as cell proliferation, invasion, aggression, and metastasis of the tumor cells as well as inhibition of apoptosis [3, 21-25]. In spite of different reports on the potential oncogenic role of *HOTAIR*, the molecular regulation of this gene needs to be revealed by more studies.

Since the genetic and epigenetic complexities of the *HOTAIR* locus have not been characterized yet, we aimed to provide an integration data to highlight different compositional features of *HOTAIR* gene. The potential model may help to design future studies to reveal the molecular mechanisms of this lncRNA. In this study, we highlighted and described a number of features in *HOTAIR* locus, which may be involved in regulation of this gene. The integrated report is derived from the *in silico* approaches through different databases and software.

Received August 4, 2017; Revised September 4, 2017; Accepted September 18, 2017

***Corresponding author:** Tel: +98-6133338965, Fax: +98-6133337009, E-mail: Mohamad.hajari@gmail.com, m-hajari@scu.ac.ir

Methods

Different databases and bioinformatics software were used. Then, the data were reanalyzed and integrated in order to provide a potential model for describing the genetic and epigenetic features of the *HOTAIR* locus. Table 1 shows list of the *in silico* tools used in this study and the methodology is represented as a flowchart (Fig. 1). In our analyses, the desired sequence was mostly defined as a sequence that spans from 2 kb upstream of annotated transcription start site (TSS) of *HOTAIR* to the end of the gene. The selection was based on the previous studies defining putative promoter regions from −2 kb to +1 kb of the TSS [26]. Some data were analyzed through Encyclopedia of DNA Elements (ENCODE) project cited in University of California, Santa Cruz (UCSC) genome browser. Encode is a genome-wide consortium project with the aim of cataloging all functional elements in the human genome through related experimental conditions. In addition, all of the software was run with default parameters and criteria. The description of each software and database as well as their criteria of the analyses are described in below.

Table 1. Softwares and databases utilized in this article

Type of analysis		Usage	Software/database	Reference/address
Genetic features	Epigenetic features			
O	O	Finding different transcripts	Ace view UCSC Ensembl	https://www.ncbi.nlm.nih.gov/ieb/research/acembly/ http://genome.ucsc.edu/ http://www.ensembl.org
O	-	Promoter detection	HMM Promoter scan Promoter 2.0 Ensembl Ace view	http://genome.ucsc.edu/ https://www-bimas.cit.nih.gov/molbio/proscan/ http://www.cbs.dtu.dk/services/Promoter/ http://www.ensembl.org https://www.ncbi.nlm.nih.gov/ieb/research/acembly/
O	-	Alternative transcription start sites	Eponine SwitchGear	http://genome.ucsc.edu/ http://genome.ucsc.edu/
O	-	CpGIs detection	UCSC Bona fides CGIs CpGProD Weizmann evolutionary CGIs	http://genome.ucsc.edu/ http://epigraph.mpi-inf.mpg.dedownloadCpG_islands revisited http://doua.prabi.fr/software/cpgprod http://genome.ucsc.edu/
O	-	DNase I hypersensitivity peak clusters	UCSC	http://genome.ucsc.edu/
-	O	CpGIs methylation status	ENCODE	http://genome.ucsc.edu/
O	O	Gene expression analysis	Ace view GTEX RNA-SEQ	https://www.ncbi.nlm.nih.gov/ieb/research/acembly/ http://genome.ucsc.edu/
O	O	Finding CTCF	ENCODE	http://genome.ucsc.edu
O	-	Finding motifs	MEME Mast program	http://MEME-Suite.org/ http://MEME-Suite.org/
O	O	Transcription factor binding sites	PreMode	http://genomequebec.mcgill.ca/PReMod//
O	-	Detection of enhancers	HMM	http://genome.ucsc.edu/
O	-	Finding repeated sequences	Repeat masker Tandem repeat by TRF	http://genome.ucsc.edu/ http://genome.ucsc.edu/
O	-	Single nucleotide polymorphism	dbSNP	http://genome.ucsc.edu/
-	O	Detection of histone marks	UCSC	http://genome.ucsc.edu/

UCSC, University of California, Santa Cruz; HMM, Hidden Markov Model; CGI, CpG Island; ENCODE, Encyclopedia of DNA Elements; TRF, tandem repeat finder.

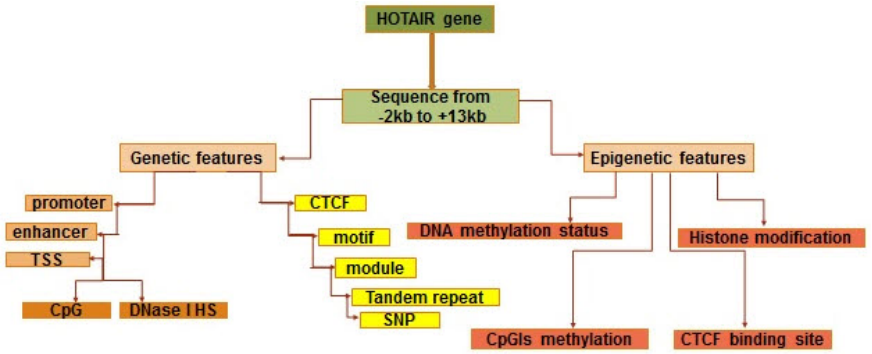

Fig. 1. The flowchart of the methods used in the study. TSS, transcription start site; SNP, single nucleotide polymorphism.

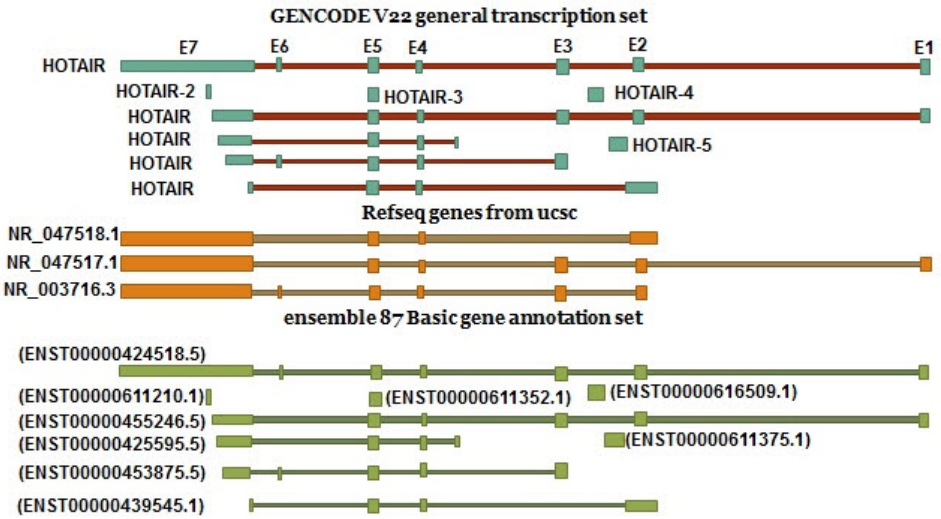

Fig. 2. Transcript variants of *HOTAIR* gene derived from the GENCODE, Ensemble and Refseq.

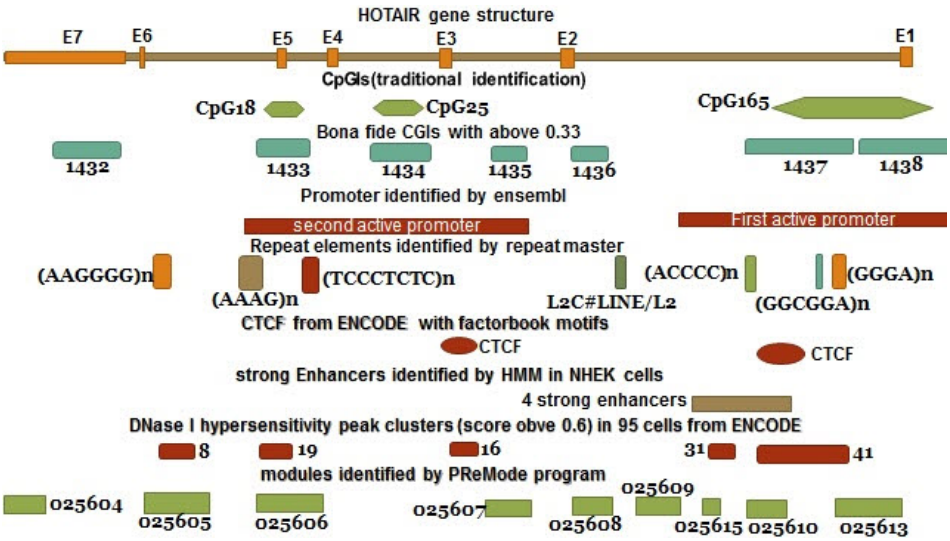

Fig. 3. Integrated regulatory elements of *HOTAIR* gene structure. The schematic diagram shows a summary of results from different databases and software which are described in the text.

Results

HOTAIR gene is transcribed into different RNA isoforms by alternative compositional features

According to the Ace view database, 11 distinct GT-AG introns are identified in the *HOTAIR* gene. This results in seven different transcripts, six of which are created through alternative splicing (https://www.ncbi.nlm.nih.gov/ieb/research/acembly/). Different variants were found in GENECODE V22 and Ensembl. According to the Refseq, there are three transcript variants for this gene (NR_047518.1, NR_047517.1 and NR_003716.3) (Fig. 2).

Since it seems that alternative transcripts of *HOTAIR* are due to alternative promoters, TSSs, alternative polyadenylation sites, and alternative splicing, we tried to find different promoters, TSSs, polyadenylation, and splice sites in the *HOTAIR* gene.

We found alternative promoters and polyadenylation sites in the *HOTAIR* locus (https://www.ncbi.nlm.nih.gov/ieb/research/acembly/). According to the Ensembl, there are two active promoters in this gene (Fig. 3). Also, Chromatin state segmentation using Hidden Markov Model (HMM) [27] identified these two active promoters as well as enhancers in the *HOTAIR* gene in some cell lines. The HMM is a probabilistic model representing probability distributions over sequences of observations. Supplementary Table 1 which is based on UCSC hg19, shows the positions of the active promoters of *HOTAIR* locus in Ensemble and HMM.

Promoter prediction with different tools recognized alternative promoters throughout this gene. Promoter scan program was run with the default promoter cutoff score. This program predicts promoters based on the degree of homologies with eukaryotic RNA pol II promoter sequences (https://www-bimas.cit.nih.gov/molbio/proscan/) [28]. Different TSSs were also found in the *HOTAIR* gene by different programs and software including Eponine, Switchgear, and Promoter 2 [29]. The Eponine program provides a probabilistic method for detecting TSSs. The Switchgear algorithm uses a scoring metric based largely on existing transcript evidence. Promoter2 takes advantage of a combination of principles that are common to neural networks and genetic algorithms. The positions of found TSSs compared to other features are shown in the Supplementary Table 1.

CpG islands were found to be overlapped with active promoters and DNase I hypersensitivity sites

According to the UCSC browser, bona fide CpGIs, Weizmann Evolutionary, and CpG ProD program, there were different CpG Islands (CGIs) in the *HOTAIR* gene. These CpGIs are shown in the Fig. 4. UCSC genome browser identifies CGIs of human genome based on the regions of DNA with average (G+C) content greater than 50%, length greater than 200 bp and a moving average CpG O/E greater than 0.6 [30, 31]. "Bona fide" identifies functional CpGIs by linking genetic and epigenetic information [32]. Weizmann evolutionary (WE) predicts highly conserved CGIs through their classification of evolutionary dynamics (http://genome.ucsc.edu/) [33]. "CpG ProD" program identifies CpGIs-overlapping with promoters in the large genomic regions under analysis and shows these CpGIs with length longer than other CpGIs [34]. Then, we tried to find any overlap between CpGIs and other regulatory elements. Two TSSs (CHR-12-P0397-R1, CHR12-P0397-R2) were found within CpG165 (annotated in UCSC genome browser) and 1437 (derived

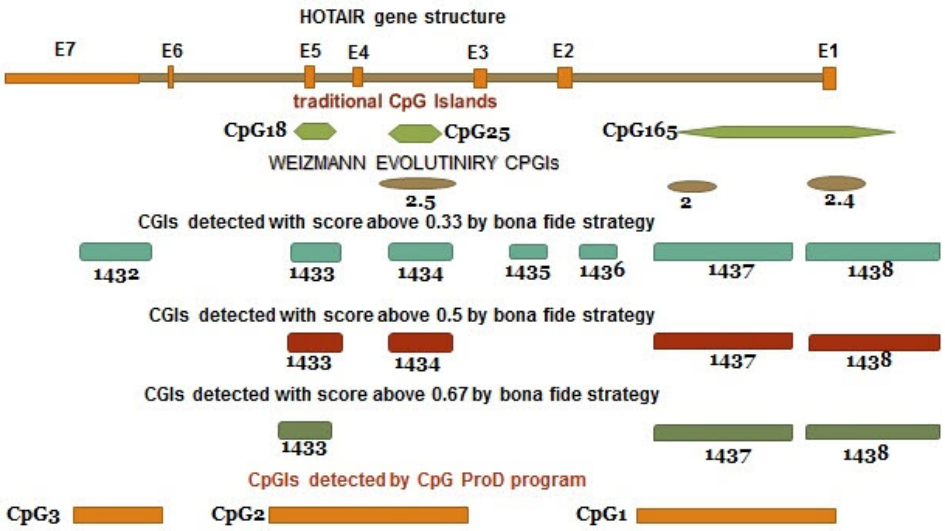

Fig. 4. CpG Islands in the *HOTAIR* gene. The data are derived from databases and prediction software. CGI, CpG Island.

Table 2. The positions of regulatory sequences which are near or within CpG165 of *HOTAIR*

Position of CpG165	Promoter (active)	Other CpGIs	Tandem repeat (strand+)	CTCF	Strong enhancer	DNase I hypersensitivity	Module and TSSs
CpG165: 54366816-54369103	HSMM cells: 54365934-54370733	Bona fide 1437: 54366623-54367999	(GGCGGA)n: 54367601-54367637	54366799-54367314	NHEK cells: 4 strong enhancers: 54365934-54367133	41: 54366785-54367814	025610: 54366634-54366977
	NHEK cells: 54367139-54369133	CpG2 (WE): 54366684-54366909	(GGGA)n: 54367731-54367801			NHEK cells DNase I hotspot: 75095: 54366045-54370999	025613: 54367707-54368584
	First active promoter based on ensembl: 54365691-54370092	CpG1 (CpGProD): 54366456-54368740	GAGGGAGG GAGCGAGA: 54367742-54367783				TSSs: CHR12-P0397-R1: 54366912-54366912
		CpG2.4 (WE): 54368334-54368964					CHR12-P0397-R2: 54367584-54367584
		Bona fide 1438: 54368166-54369840					

Positions are based on UCSC hg19.
TSS, transcription start site; WE, Weizmann evolutionary.

from bona fide CGIs). The CpGIs were mostly overlapped with the active promoter regions (Fig. 3, Supplementary Table 1). We focused on CpG165 and found some regulatory elements which are within or near to this CpG (Table 2).

In addition, several DNase I hypersensitivity hotspots were found to be overlapped with CpGIs in some cell lines (Supplementary Table 1). We found the DNase I hypersensitivity peak clusters of *HOTAIR* gene in 95 cells with score greater than 0.6 by using UCSC genome browser. DNase I hypersensitivity peak cluster 19 is located within CpG1433 and mostly overlaps with CpG18. Also, DNase I hypersensitivity peak cluster 41 is located within CpG1437 and mostly overlaps with CpG165 and partially overlaps with CpG2 (WE) (Fig. 3, Supplementary Table 1).

Furthermore, we detected specific CpG dinucleotides methylation status within or near the predicted CpGIs in some cell lines by using ENCODE (Supplementary Table 2). This track identifies specific CpG dinucleotides methylation status by Infinium human methylation 450 bead array platform and classifies the methylation status into four groups: (1) not available (score = 0), (2) unmethylated ($0 <$ score ≤ 200), (3) partially methylated ($200 <$ score < 600), and (4) methylated (score ≥ 600) (http://genome.ucsc.edu/).

CTCF and transcription factor binding sites are overlapped with CpGIs and TSSs

GTEx RNA-seq strategy indicates that *HOTAIR* has variable expression in different tissues and its most expression level is in the artery-tibial tissue (data not shown). We found two putative regions for CTCF binding sites in the *HOTAIR* locus by ENCODE with factorbook motifs, one of which is located within CpG1437 (bona fide CpGIs) and mostly overlaps with CpG165 (Table 2, Fig. 3). This track determines regions of transcription factor binding sites taken from a comprehensive chip-seq experiments identified by ENCODE and factorbook pool (http://genome.ucsc.edu/). We predicted sequences of motifs and positions of these motifs in the *HOTAIR* locus by using MEME and MAST programs (Supplementary Table 3). MEME program searches the motifs from downloaded sequences through using complementary strengths of probabilistic and discrete models (http://MEME-Suite.org/) [35, 36]. The program was run with default parameters and normal mode of motif discovery. Mast program searches specific sequences based on predicted motifs by MEME program and exactly matches these sequences with the motifs sequences (http://MEME-Suite.org/) [37].

We found nine sequences of modules depending on their

transcription factor binding sites in the *HOTAIR* locus by PReMode program [38, 39]. We observed some of these elements overlapped with the predicted CpGIs and TSSs (Fig. 3, Supplementary Table 1). In addition, we determined that some of these modules have common transcription factors (data not shown).

Some polymorphisms such as tandem repeats exist within the regulatory elements

Repeat Masker found several repeats sequences overlapped with regulatory elements of the *HOTAIR* locus such as CpGIs (Fig. 3, Supplementary Table 1) and motifs (Supplementary Table 3). Repeat master investigates query sequences and generates a detailed annotation of available repeats in these sequences and shows dispersed repeats and low complexity DNA sequences (http://genome.ucsc.edu/). In addition, tandem repeat finder, which analyzes simple tandem repeats, predicted one simple tandem repeat (GAGGGAGGGAGCGAGA) within this gene (Supplementary Table 1) (http://genome.ucsc.edu/) [40]. In addition, we found some simple nucleotide polymorphisms within regulatory sequences of *HOTAIR* gene (Supplementary Table 4).

Discussion

Studies have shown that aberrant epigenetic modifications including aberrant DNA methylation and histone modification are significantly involved in the dysregulation of genes with their potential roles in cancers [41]. However, identification of the exact elements of *HOTAIR* as well as their interaction has not been discovered yet. This study was aimed to find and highlight different regulatory elements by data integration. We identified putative regulatory elements that contribute to the regulation of *HOTAIR* expression by *in silico* analyses. Identification of these elements suggests new understanding of *HOTAIR* expression and might help to design future studies on this lncRNA which has oncogenic role in different cancers [42-45].

First, we tried to show different isoforms of *HOTAIR* RNA transcribed through alternative mechanisms. Since a recent study suggested the important role of *HOTAIR* domains in its function [46], we propose studying the molecular roles of different RNA isoforms in future researches. Then, in order to find alternative and potential features involved in generation of RNA isoforms, we checked the putative TSSs, promoters, and polyadenylation sites. We found different features, which are potentially involved in alternative transcription of *HOTAIR* gene.

Considering the potential involvement of methylation beyond CGI-promoters in human cancer, we focused on potential CGIs of *HOTAIR*. According to the fact that function of DNA methylation seems to be varied with context, we tried to find any relation between the CGIs and other compositional features such as TSSs, promoters, enhancers, DNase I hypersensitivity sites, and CTCF binding sites. Alterations in DNA methylation are known to cooperate with genetic elements and to be involved in human carcinogenesis. The results showed different CpGIs in the *HOTAIR* locus and determined their epigenetic status through integration analysis. The methylation status of these CGIs needs to be revealed in future researches. The methylation analysis will be so important because we currently know that most CGIs located in TSSs are not methylated. However, CGI methylation of the TSS is associated with long-term silencing. In addition, CGIs in gene bodies are sometimes methylated in a tissue-specific manner [47]. It has been reported that methylation of a CTCF-binding site may block the binding of CTCF. Altogether, different CpGIs overlapped with genetic elements seem to have important roles in controlling *HOTAIR*.

Some repeat sequences and single nucleotide polymorphisms exist within or next to the predicted CpGIs. We think that repeat number variations may effect on methylation status of regulatory regions of *HOTAIR* gene. Different studies reported some associations between polymorphisms of *HOTAIR* and cancers risks. The examples are the association between rs920778 [48], rs4759314 [49], and rs12826786 [25] and gastric cancer, rs7958904 and colorectal cancer [50], rs920788 and breast cancer [51], rs4759314 and rs7958904 in epithelial ovarian cancer [52]. We found that some SNPs are located within regulatory regions and so may effect on the gene expression. Also, since the repeat sequences of *HOTAIR* gene might contribute to the methylation status of regulatory regions, we highlighted the overlaps between these sequences and the predicted CpGIs.

Due to the overlap with active promoter, strong enhancer, CTCF binding site, DNase I hypersensitive sites, SNPs, and repeat sequences, CpG165 seems to be more important compared to other CpGIs for generation of the long RNA isoform. However, according to the Fig. 3, considering the overlap with other structural features, other CpGIs within the gene structure also seems to be involved in gene regulation. This integration model should be checked and validated in future experimental works.

Altogether, it seems that alternative transcripts of *HOTAIR* originate from interactions between genetic and epigenetic elements. Our data provide strong evidence based on the databases and *in silico* prediction that specific sequence motifs may potentially be involved in DNA methylation states of various set of CGIs in different tissues including normal and tumors. Our study suggests that the combinatorial binding of specific transcription factors plays a

major role in regulation of *HOTAIR* expression. Future work that aims to provide detailed maps of epigenome in normal and diseased states is crucial to our understanding of *HOTAIR* role in cancer pathogenesis.

ORCID: Mohammadreza Hajjari: http://orcid.org/0000-0003-3838-0259; Saghar Rahnama: http://orcid.org/0000-0002-2068-5436

Authors' contribution

Conceptualization: MH
Formal analysis: SR, MH
Methodology: MH
Visualization: MH, SR
Writing – original draft: SR, MH
Review and edit: MH

Acknowledgments

This study was conducted and supported as a project in Shahid Chamran Universty of Ahvaz.

Supplementary materials

Supplementary data including four tables can be found with this article online at http://www.genominfo.org/src/sm/gni-15-170-s001.pdf.

References

1. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, *et al*. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* 2012;22:1760-1774.
2. Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, Willingham AT, *et al*. RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* 2007;316:1484-1488.
3. Hajjari M, Salavaty A. HOTAIR: an oncogenic long non-coding RNA in different cancers. *Cancer Biol Med* 2015;12:1-9.
4. Fatica A, Bozzoni I. Long non-coding RNAs: new players in cell differentiation and development. *Nat Rev Genet* 2014;15:7-21.
5. Yu X, Li Z. Long non-coding RNA *HOTAIR*: a novel oncogene (review). *Mol Med Rep* 2015;12:5611-5618.
6. Khandelwal A, Malhotra A, Jain M, Vasquez KM, Jain A. The emerging role of long non-coding RNA in gallbladder cancer pathogenesis. *Biochimie* 2017;132:152-160.
7. Rinn JL. lncRNAs: linking RNA to chromatin. *Cold Spring Harb Perspect Biol* 2014;6:a018614.
8. Mercer TR, Mattick JS. Structure and function of long non-coding RNAs in epigenetic regulation. *Nat Struct Mol Biol* 2013;20:300-307.
9. Li CH, Chen Y. Targeting long non-coding RNAs in cancers: progress and prospects. *Int J Biochem Cell Biol* 2013;45:1895-1910.
10. Ishibashi M, Kogo R, Shibata K, Sawada G, Takahashi Y, Kurashige J, *et al*. Clinical significance of the expression of long non-coding RNA *HOTAIR* in primary hepatocellular carcinoma. *Oncol Rep* 2013;29:946-950.
11. Rinn JL, Kertesz M, Wang JK, Squazzo SL, Xu X, Brugmann SA, *et al*. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* 2007;129:1311-1323.
12. Loewen G, Zhuo Y, Zhuang Y, Jayawickramarajah J, Shan B. lincRNA *HOTAIR* as a novel promoter of cancer progression. *J Can Res Updates* 2014;3:134-140.
13. Bhan A, Mandal SS. Estradiol-induced transcriptional regulation of long non-coding RNA, *HOTAIR*. *Methods Mol Biol* 2016;1366:395-412.
14. He S, Liu S, Zhu H. The sequence, structure and evolutionary features of *HOTAIR* in mammals. *BMC Evol Biol* 2011;11:102.
15. Zhang J, Zhang P, Wang L, Piao HL, Ma L. Long non-coding RNA *HOTAIR* in carcinogenesis and metastasis. *Acta Biochim Biophys Sin (Shanghai)* 2014;46:1-5.
16. Schorderet P, Duboule D. Structural and functional differences in the long non-coding RNA *hotair* in mouse and human. *PLoS Genet* 2011;7:e1002071.
17. Tsai MC, Manor O, Wan Y, Mosammaparast N, Wang JK, Lan F, *et al*. Long noncoding RNA as modular scaffold of histone modification complexes. *Science* 2010;329:689-693.
18. Meredith EK, Balas MM, Sindy K, Haislop K, Johnson AM. An RNA matchmaker protein regulates the activity of the long noncoding RNA *HOTAIR*. *RNA* 2016;22:995-1010.
19. Ma MZ, Li CX, Zhang Y, Weng MZ, Zhang MD, Qin YY, *et al*. Long non-coding RNA *HOTAIR*, a c-Myc activated driver of malignancy, negatively regulates miRNA-130a in gallbladder cancer. *Mol Cancer* 2014;13:156.
20. Wu Y, Zhang L, Wang Y, Li H, Ren X, Wei F, *et al*. Long non-coding RNA *HOTAIR* involvement in cancer. *Tumour Biol* 2014;35:9531-9538.
21. Berrondo C, Flax J, Kucherov V, Siebert A, Osinski T, Rosenberg A, *et al*. Expression of the long non-coding RNA *HOTAIR* correlates with disease progression in bladder cancer and is contained in bladder cancer patient urinary exosomes. *PLoS One* 2016;11:e0147236.
22. Kim HJ, Lee DW, Yim GW, Nam EJ, Kim S, Kim SW, *et al*. Long non-coding RNA *HOTAIR* is associated with human cervical cancer progression. *Int J Oncol* 2015;46:521-530.
23. Li J, Yang S, Su N, Wang Y, Yu J, Qiu H, *et al*. Overexpression of long non-coding RNA *HOTAIR* leads to chemoresistance by activating the Wnt/beta-catenin pathway in human ovarian cancer. *Tumour Biol* 2016;37:2057-2065.
24. Chiyomaru T, Fukuhara S, Saini S, Majid S, Deng G, Shahryari V, *et al*. Long non-coding RNA *HOTAIR* is targeted and regulated by miR-141 in human cancer cells. *J Biol Chem* 2014;289:12550-12565.
25. Guo W, Dong Z, Bai Y, Guo Y, Shen S, Kuang G, *et al*. Associations between polymorphisms of *HOTAIR* and risk of gastric

cardia adenocarcinoma in a population of north China. *Tumour Biol* 2015;36:2845-2854.
26. Marino-Ramirez L, Spouge JL, Kanga GC, Landsman D. Statistical analysis of over-represented words in human promoter sequences. *Nucleic Acids Res* 2004;32:949-958.
27. Pedersen AG, Baldi P, Brunak S, Chauvin Y. Characterization of prokaryotic and eukaryotic promoters using hidden Markov models. *Proc Int Conf Intell Syst Mol Biol* 1996;4:182-191.
28. Prestridge DS. Predicting Pol II promoter sequences using transcription factor binding sites. *J Mol Biol* 1995;249:923-932.
29. Down TA, Hubbard TJ. Computational detection and location of transcription start sites in mammalian genomic DNA. *Genome Res* 2002;12:458-461.
30. Gardiner-Garden M, Frommer M. CpG islands in vertebrate genomes. *J Mol Biol* 1987;196:261-282.
31. Boucher CA, King SK, Carey N, Krahe R, Winchester CL, Rahman S, *et al*. A novel homeodomain-encoding gene is associated with a large CpG island interrupted by the myotonic dystrophy unstable (CTG)n repeat. *Hum Mol Genet* 1995;4: 1919-1925.
32. Bock C, Walter J, Paulsen M, Lengauer T. CpG island mapping by epigenome prediction. *PLoS Comput Biol* 2007;3:e110.
33. Hajjari M, Khoshnevisan A, Lemos B. Characterizing the retinoblastoma 1 locus: putative elements for Rb1 regulation by *in silico* analysis. *Front Genet* 2014;5:2.
34. Ponger L, Mouchiroud D. CpGProD: identifying CpG islands associated with transcription start sites in large genomic mammalian sequences. *Bioinformatics* 2002;18:631-633.
35. Wang Z, Fan H, Yang HH, Hu Y, Buetow KH, Lee MP. Comparative sequence analysis of imprinted genes between human and mouse to reveal imprinting signatures. *Genomics* 2004;83:395-401.
36. Hajjari M, Behmanesh M, Jahani MM. *In silico* finding of putative cis-acting elements for the tethering of polycomb repressive complex2 in human genome. *Bioinformation* 2014; 10:187-190.
37. Janssen CS, Phillips RS, Turner CM, Barrett MP. Plasmodium interspersed repeats: the major multigene superfamily of malaria parasites. *Nucleic Acids Res* 2004;32:5712-5720.
38. Jeziorska DM, Jordan KW, Vance KW. A systems biology approach to understanding cis-regulatory module function. *Semin Cell Dev Biol* 2009;20:856-862.
39. Ferretti V, Poitras C, Bergeron D, Coulombe B, Robert F, Blanchette M. PReMod: a database of genome-wide mammalian cis-regulatory module predictions. *Nucleic Acids Res* 2007;35:D122-D126.
40. Benson G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* 1999;27:573-580.
41. Suzuki H, Maruyama R, Yamamoto E, Niinuma T, Kai M. Relationship between noncoding RNA dysregulation and epigenetic mechanisms in cancer. In: *The Long and Short Non-coding RNAs in Cancer Biology* (Song E, ed.). Singapore: Springer, 2016. pp. 109-135.
42. Deng J, Yang M, Jiang R, An N, Wang X, Liu B. Long non-coding RNA *HOTAIR* regulates the proliferation, self-renewal capacity, tumor formation and migration of the cancer stem-like cell (CSC) subpopulation enriched from breast cancer cells. *PLoS One* 2017;12:e0170860.
43. Kim K, Jutooru I, Chadalapaka G, Johnson G, Frank J, Burghardt R, *et al*. *HOTAIR* is a negative prognostic factor and exhibits pro-oncogenic activity in pancreatic cancer. *Oncogene* 2013;32:1616-1625.
44. Nakagawa T, Endo H, Yokoyama M, Abe J, Tamai K, Tanaka N, *et al*. Large noncoding RNA *HOTAIR* enhances aggressive biological behavior and is associated with short disease-free survival in human non-small cell lung cancer. *Biochem Biophys Res Commun* 2013;436:319-324.
45. Borley J, Brown R. Epigenetic mechanisms and therapeutic targets of chemotherapy resistance in epithelial ovarian cancer. *Ann Med* 2015;47:359-369.
46. Loewen G, Jayawickramarajah J, Zhuo Y, Shan B. Functions of lncRNA *HOTAIR* in lung cancer. *J Hematol Oncol* 2014;7:90.
47. Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat Rev Genet* 2012;13:484-492.
48. Pan W, Liu L, Wei J, Ge Y, Zhang J, Chen H, *et al*. A functional lncRNA *HOTAIR* genetic variant contributes to gastric cancer susceptibility. *Mol Carcinog* 2016;55:90-96.
49. Du M, Wang W, Jin H, Wang Q, Ge Y, Lu J, *et al*. The association analysis of lncRNA *HOTAIR* genetic variants and gastric cancer risk in a Chinese population. *Oncotarget* 2015;6:31255-31262.
50. Xue Y, Gu D, Ma G, Zhu L, Hua Q, Chu H, *et al*. Genetic variants in lncRNA *HOTAIR* are associated with risk of colorectal cancer. *Mutagenesis* 2015;30:303-310.
51. Yan R, Cao J, Song C, Chen Y, Wu Z, Wang K, *et al*. Polymorphisms in lncRNA *HOTAIR* and susceptibility to breast cancer in a Chinese population. *Cancer Epidemiol* 2015;39:978-985.
52. Wu H, Shang X, Shi Y, Yang Z, Zhao J, Yang M, *et al*. Genetic variants of lncRNA *HOTAIR* and risk of epithelial ovarian cancer among Chinese women. *Oncotarget* 2016;7:41047-41052.

SUPPLEMENTARY INFORMATION

HOTAIR Long Non-coding RNA: Characterizing the Locus Features by the *In Silico* Approaches

Mohammadreza Hajjari*, Saghar Rahnama

Department of Genetics, Shahid Chamran University of Ahvaz, Ahvaz 61336-3337, Iran

http://www.genominfo.org/src/sm/gni-15-170-s001.pdf

Supplementary Table 1. The positions of regulatory sequences in the *HOTAIR* locus

Position	Promoter (active)	CpGIs	Tandem repeat (strand+)	CTCF	Enhancer	DNase I hypersensitivity	Module and TSSs
54354858–54356533	No	No	No	No	No	No	**025604:** 54355865–54356303 **TSS: (chr12.11801):** 54354858–54354859
54356534–54359334	**HSMM cells:** 54359134–54359334	**Bonafide1432:** 54357217–54357921 **CpG3 (CpGProD):** 54357032-54358001 **CpG2 (CpGProD):** 54359256–54359334	**(AAGGGG)n:** 54358177–54358291	No	**HSMM cells:** 12 Weak enhancers **HMEC cells:** 20 Weak enhancers	**8:** 54358245–54358454 **HSMM cells DNase** I **hotspot: 66574:** 54357302–54358901	**025605:** 54358063–54358978
54359335–54362492	**HSMM cells:** 54359335–54361533 **NHEK cells:** 54362134–54362333 **Second active promoter based on Ensembl:** 54359491–54362492	**CpG18:** 54359659–54359906 **Bonafide1433:** 54359598–54360005 **CpG2.5 (WE):** 54360184–54360883 **CpG25:** 54360375–54360660 **Bonafide1434:** 54360202–54360827 **Bonafide1435:** 54362119–54362323 **CPG2 (CpGProD):** 54359334–54360945	**(AAAG)n:** 54359478–54359704 **(TCCCTCTC)n:** 54359986–54360112	54361413–54361642	**HMEC cells:** 19 Weak enhancers	**19:** 54359645–54359854 **16:** 54361465–54361654 **HSMM cells DNase I hotspot**: **66575**: 54359619–54360710 **66576:** 54361172–54361793	**025606:** 54359632–54360527 **025607:** 54361760–54362456 **TSS: CHR12-M0409-R1:** 54361133–54361133
54362493–54363334	No	**Bona fide 1436:** 54362691–54362900	No	No	**HSMM cells:** 6 Weak enhancers	No	**025608:** 54362765–54363139
54363335–54364965	No	No	**L2C (**within strand-**):** 54363655–54363707	No	**HSMM cells:** 7 Weak enhancers	No	**025609:** 54364519–54364965
54364966–54370999	**HSMM cells:** 54365934–54370733 **NHEK cells:** 54367139–54369133 **First active promoter based on Ensembl:** 54365691–54370092	**Bona fide 1437:** 54366623–54367999 **CpG2 (WE):** 54366684–54366909 **CpG165:** 54366816–54369103 **CpG1 (CpGProD):** 54366456–54368740 **CpG2.4 (WE):** 54368334–54368964 **Bona fide 1438:** 54368166–54369840	**(ACCCC)n:** 54366647–54366670 **(GGCGGA)n:** 54367601–54367637 **(GGGA)n:** 54367731–54367801 **GAGGGAGGGAGCGAGA**: 54367742–54367783	54366799–54367314	**HEpG2 cells:** 6 Weak enhancers **HMEC cells:** 13 Weak enhancers **HSMM cells:** 20 Weak enhancers **NHEK cells:** 7 Weak enhancers, **4 Strong enhancers**: 54365934–54367133	**31:** 54366145–54366374 **41:** 54366785–54367814 **HSMM cells DNase I hotspot: 66579:** 54365947–54366518 **NHEK cells DNase I hotspot: 75095:** 54366045–54370999	**025615:** 54366091–54366249 **025610:** 54366634–54366977 **025613:** 54367707–54368584 **TSSs: CHR12-P0397-R1:** 54366912–54366912 **CHR12-P0397-R2:** 54367584–54367584

Exact position of this gene is chr12:54356092-54368740. For easiness, genomic region under analysis is divided into smaller portions. Positions are based on UCSC hg19.
TSS, transcription start site; WE, Weizmann evolutionary.

Supplementary Table 2. Specific CpG dinucleotides methylation status identified from different cell lines in ENCODE

Cell line	Position					
	54357408–54357772 Within CpG1432	54359712–54359797 Within CpG1433 and CpG18	54360263–54360837 Within CpG1434, CpG25 and CpG2.5 (WE)	54363055–54366424 Near to CpG1436 and after it	54366760–54367822 Within CpG1437 and CpG2 (WE)	54368203–54368640 Within CpG165 and CpG2.4 (WE)
GM12878	Unmethylated	Mostly unmethylated	Partially methylated	Different	Partially* methylated	Different
H1-heSC	Mostly unmethylated	Unmethylated	Unmethylated	Partially Methylated	Mostly unmethylated	Mostly unmethylated
K562	Unmethylated	Mostly unmethylated	Partially[a] methylated	Mostly unmethylated	Different	Partially[b] methylated
Hela-S3	methylated	Methylated	Methylated	Mostly methylated	Mostly unmethylated	Unmethylated
HepG2	Different	Partially methylated	Partially methylated	Different	Partially methylated	Mostly unmethylated
HuVEC	Unmethylated	Unmethylated	Unmethylated	Partially[c] methylated	Mostly unmethylated	Unmethylated

Table shows different cell lines (first column) and positions of specific CpG dinucleotides within or near to the predicted CpGIs (first row) in *HOTAIR* gene.
ENCODE, Encyclopedia of DNA Elements; WE, Weizmann evolutionary.

[a]One specific C nucleotide is unmethylated; other specific C nucleotides are partially methylated.
[b]One specific C nucleotide is methylated; other specific C nucleotides are partially methylated.
[c]Two specific C nucleotides are unmethylated; other specific C nucleotides are partially methylated.

Supplementary Table 3. The motifs sequences identified by MEME and Mast programs in *HOTAIR*

Motifs	Width	Best possible match (strand -)	p-value	Position
1	47	GCGAAAAAGGACCAAGAGGGCGAGACGAGGGAAGAGACCTAGAGAGA	0.00032	Chr12: 54357805–54357852 (within CpG1432)
2	40	TTTACTCTTTCTTTTCTCTCTTTCTTCCTCTCTTTTTTTT	0.00121	Chr12: 54360742–54360782 (within CpG143, CpG2.5(WE))
3	39	CCCTCTCCCTTTCCCTCCCTCTCCCTTCCCCCTCCCTTT	0.00048	Chr12: 54367746–54367785 (within CpG1437, CpG165)

The motifs sequences are predicted from antisense strand of *HOTAIR* locus and a specified p-value of the motifs are applied by Mast program.
WE, Weizmann evolutionary.

Supplementary Table 4. Simple nucleotide polymorphisms in *HOTAIR*

Name (SNP)	Function	Summary	Reference allele	Strand	Class	Position
rs1838169	nc-transcript variant	G>C/G	G	-	Single	Chr12:54357495–54357495 **(within CpG 1432)**
rs7958904	nc-transcript variant	G>G/C	C	+	Single	Chr12:54357552–54357552 **(within CpG 1432)**
rs17840857	nc-transcript variant	A/C/G/T	G	+	Single	Chr12:54357757–54357757 **(within CpG 1432)**
rs111434707	nc-transcript variant	-/G	G	+	Deletion	Chr12:54357757–54357757 **(within CpG 1432)**
rs200062983	nc-transcript variant	C>C/T	C	+	Single	Chr12:54357761–54357761 **(within CpG 1432)**
rs35951424	Intron variant	-/A	A	+	Deletion	Chr12:54357997–54357997 **(within HSMM cells DNase I hotspot:66574)**
rs201719283	Intron variant-Splice donor variant	-/C	C	+	Deletion	Chr12:54358048–54358014 **(within HSMM cells DNase I hotspot:66574)**
rs71227278	Intron variant nc-transcript variant	->TTAA	-	+	Insertion	Chr12:54358048–54358047 **(within HSMM cells DNase I hotspot:66574)**
rs58072355	Intron variant	A>A/G	A	+	Single	Chr12:54358443–54358443 **(within DNase I hypersensitivity peak clusters 8)**
rs139645979	Intron variant	-/ACGCACAAG	ACGCACAAG	+	Deletion	Chr12:54358629–54358629 **(within HSMM cells DNase I hotspot:66574)**
rs10783616	Intron variant	C>C/G	C	+	Single	Chr12:54359220–54359220 **(within active promoter of HSMM cells)**
rs10783617	Intron variant	G>G/T	G	+	Single	Chr12:54359387–54359387 **(within active promoter of HSMM cells)**
rs376812530	Intron variant	-/GAAG	-	+	Insertion	Chr12:54359525–54359525 **(within tandem repeat (AAAG)n)**
rs76084431	Intron variant	C>C/T	C	+	Single	Chr12:54359946–54359946 **(within CpG 1433)**
rs920778	Intron variant	C>C/T	C	-	Single	Chr12:54360232-54360232 **(within CpG1434 and CpG2.5(WE))**
rs920777	Intron variant	C>C/T	T	-	Single	Chr12:54360429–54360429 **(within CpG 25, CpG1434 and CpG2.5(WE))**
rs74089839	Intron variant	A>A/T	A	+	Single	Chr12:54360561–54360561 **(within CpG 25, CpG1434 and CpG2.5(WE))**
rs11301759	Intron variant	-/C	C	+	Deletion	Chr12:54360613–54360613 **(within CpG 25 ,CpG1434 andCpG2.5(WE))**
Rs1899663	Intron variant	G>G/T	G	-	Single	Chr12:54360994–54360994 **(within first active promoter based on Ensembl)**
rs4759314	Intron variant	A>A/G	G	+	Single	Chr12:54361835–54361835 **(within module025607)**
rs17105613	Intron variant	C>C/T	T	+	Single	Chr12:54362194–54362194 **(within CpG 1435)**
rs73313155	nc-transcript variant	C>C/T	C	+	Single	Chr12:54362432–54362432 **(within module025607)**
rs73313156	Intron variant	G>A/G	A	+	Single	Chr12:54362915–54362915 **(within module025608)**
rs5798292	Intron variant	-/G	G	+	Deletion	Chr12:54366274–54366274 **(within 4 strong Enhancers of NHEK cells)**
rs12427129	Intron variant	C>C/T	C	+	Single	Chr12:54367690–54367690 **(within CpG 165 and CpG1437)**
rs74089843	Intron variant	T>A/T	T	+	Single	Chr12:54368227–54368227 **(within CpG 165)**
rs78894992	Intron variant	G>A/G	A	+	Single	Chr12:54368400–54368400 **(within CpG 165 and CpG2.4(WE))**
rs75547142	Intron variant	C>C/T	C	+	Single	Chr12:54368560–54368560 **(within CpG 165 and CpG2.4(WE))**

Simple nucleotide polymorphisms (SNPs) were recognized by "dbSNP 147"and positions are based on UCSC hg19.