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.