WORLD JOURNAL OF PHARMACEUTICAL RESEARCH
SJIF Impact Factor 5.045
Volume 3, Issue 4, 1037-1046. Research Article ISSN 2277 – 7105

IDENTIFICATIN OF HUMAN LARGE INTERGENIC NONCODING RNAS

Shruthi Prakash*, Anita.P.M, Kusum Paul

Department of Biotechnology,The Oxford College of Engineering , Bommanhalli , Bangalore-560068, Karnataka ,India.

Article Received on
05 April 2014,

Revised on 28 April 2014,
Accepted on 21 May 2014

***Correspondence for Author**
Shruthi Prakash
Department of Biotechnology, The Oxford College of Engineering , Bommanhalli , Bangalore-560068, Karnataka , India

ABSTRACT

Large intergenic noncoding RNAs (lincRNAs) are emerging as key regulators of diverse cellular processes.Determining the function of individual lincRNAs remains a challenge. Recent advances in RNA sequencing (RNA-seq). Here, we presentan integrative NGS approach to define a reference catalog of >3000 human lincRNAs.Our catalog unifies previouslyexisting annotation sources with transcripts we assembled from RNA-seq data collected from RNA-seq reads across 2 tissue and cell types. We have aligned and assembled the reads to the reference genome version 19 (Hg19).We found that lincRNA expression is strikinglytissue-specific compared with coding genes, and that lincRNAs are typically coexpressed with their neighboringgenes, Our integrated, comprehensive, yet conservative reference catalog of humanlincRNAs reveals the global properties of lincRNAs and will facilitate experimental studies and further functionalclassification of these genes.

Keywords: long noncoding RNAs; RNA sequencing; lincRNAs.

INTRODUCTION

Over the past decade, evidence from numerous highthroughputgenomic platforms reveals that even thoughless than 2% of the mammalian genome encodes proteins,a significant fraction can be transcribed into differentcomplex families of non-coding RNAs (ncRNAs) [1-4].Other than microRNAs and other families of small noncodingRNAs, long non-coding RNAs (lncRNAs, >200nt)are emerging as potent regulators of gene expression [5].Originally identified from two homosapien cell types using chromatin state maps as a subtypeof lncRNAs, long intergenic non-coding RNAs (lincRNAs),Recent studies demonstrate the functionalsignificance of lincRNAs. However, it remains challengingtask to identify all the

lincRNAs existent in various biologicalprocesses and systems.Whole transcriptome sequencing, known as RNA-Seq, offers the rapid comprehensive discovery of novel genes and transcripts [6]. With the de novo assembly software such as Cufflinks [7] a large set of novel assemblies can be obtained from RNA-Seq data. Several programs have been used to facilitate the cataloguing of lincRNAs from RNA-Seq assemblies.In particular, recent studies have focused on large intergenic noncoding RNAs (lincRNAs) [8-11].,which do not overlap annotated protein-coding regions, as this facilitates experimental manipulation and computational analysis.Recent work has suggested various functions and molecularmechanisms for lincRNAs [12,13], including the regulation of epigenetic marks and gene expression[14,15].Other studies have inferred and tested the functional role of lincRNAs in processes such as pluripotency and p53 response pathways by associating the expression of lincRNAs with those of protein-coding genes[16]. Despite these intriguing studies of individual lincRNAs, generalizing these findings to thousands of lincRNAs remains a substantial challenge. Collectively, lincRNAs are likely to reflect different families with distinct roles.

A first requirement toward functional categorization is a systematic catalog of lincRNA transcripts and their expression across tissues. In practice, however, researchers studying human lincRNAs are faced with an excessive set of noncoding transcripts of varying or unknown reliability that may not be well defined[10] and have little or no expression data [17], or with very small sets of experimentally validated on[18].Transcripts in current annotations of the human transcriptome from the GENCODE/HAVANA[17]or theUniversity of California at Santa Cruz (UCSC) Genome Browser [19] are valuable resources, but it is hard to evaluate their biological characteristics in the absence of expression levels and further processing. Recent advances lincRNAsequencing (RNA-seq)[21] and computational methods for transcriptome reconstruction [20] now provide an opportunity to comprehensively annotate and characterize lincRNA transcripts. Indeed, an initial application of this approach in three Human cell types characterized the gene structure of more than a thousand human lincRNAs, most of which were not previously identified[20].

Here, we present an integrative approach to define areference set of lincRNAs that unifies existing annotationsources with transcripts reconstructed from RNA-seq reads collected human tissueWe developed a conservative, broadly applicable pipeline to identify transcripts that are sufficiently expressedand have a negligible potential to encode proteins.We used

these features to test some of the proposed rolesand characteristics of lincRNAs in a global and systematicway. For example, we found that lincRNAs—at all expressionlevels—are expressed in a highly tissue-specificmanner—much more so than protein-coding genes. Weobserved no significant enrichment of correlated coexpressionbetween lincRNAs and their neighboring genesbeyond that expected for any two neighboring proteincodinggenes.We identified expressed orthologous transcriptsin another vertebrate species for 993 humanlincRNAs. An additional set of around 4000 other transcripts withhigh evolutionary conservation but ambiguous codingpotential may function as noncoding RNAs or as smallpeptides. Finally, we highlight 270 lincRNAs that residewithin intergenic regions previously associated with specific specific diseases/traits.

MATERIALS AND METHODS

RNA-seq data sets

We used RNA-Seq H1 embryonic stem cells of Homo sapiensdata sample. This tissue that were sequenced using Illumina and obtained from the Sequence Read Archive.

lincRNA classification pipeline

Once the data sample has been obtained from the database.Since the reads of SRA database is non text format we are going to convert the reads to readable which is of text format.To do so We are using a Software called Sratoolkit.2.3.4.3-win64 which converts to text format.

Command to convert sra format to fastq format:fastq-dump.2.3.2 --split-files -A filename ./filename.sra

Quality Checking And Trimming

Once the reads has been converted in to a text form.our next step is to check for the quality of the data.The quality of the data can be checked by using FastQC software based on the Phred score value for ILLUMINA platform.For ILLUMINA platform the standard phred score value is 20.For the data to be good the reads should be greater than 20,if values are lesser than 20 then we should go for trimming using the software Fastx_toolkit.0.0.14which trims the values which lies below 20 and doesnot affect for our analysis.

Command for Trimming of data:fastx_trimmer –Q33 –t 4 I input file –o output file

Indexing And Allignment To The Reference Genome

Our next step is to align and index our reads to reference genome that is human genome version 19 (Hg19).We are using BOWTIE AND TOPHAT to index and alignment respectively.

Commands to use bowtie and Tophat

For Bowtie:Bowtie2 –built –f (genome file index file)

For Tophat: Tophat –o output filename –G path of GTF file path of reference genome.

Assemblly Of Alligned Transcripts

Once the transcripts are alligned to our reference genome our next step is to do with assembly using cufflinks tools.Assembles transcripts, estimates their abundances, and tests for differential expression and regulation in RNA-Seq samples. It accepts aligned RNA-Seq reads and assembles the alignments into a parsimonious set of transcripts. Cufflinks then estimates the relative abundances of these transcripts based on how many reads support each one, taking into account biases in library preparation protocols.

Command for cufflinks

Cufflink –o outputfile –g path of gtf file/gtf file accepted_hits.bam(Tophat output result).

Cuffcompare

Cufflinks includes a program that you can use to help analyze the transfrags you assemble. The program cuffcompare helps you:

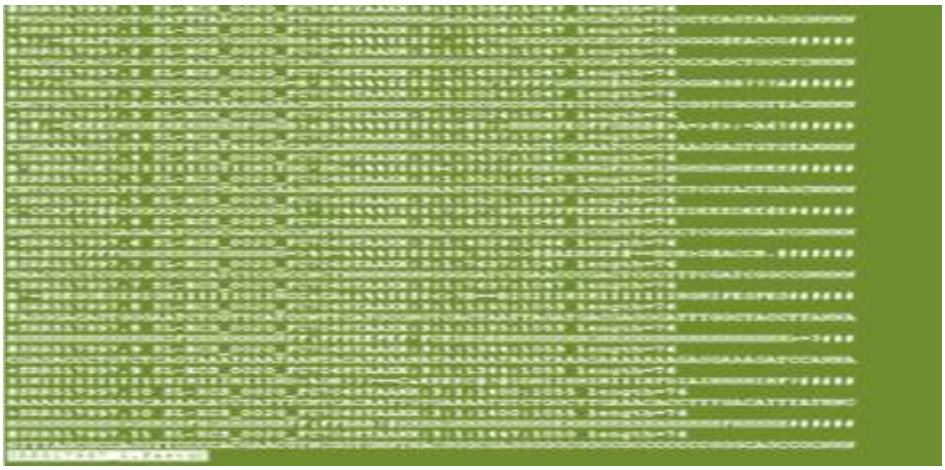
- Compare your assembled transcripts to a reference annotation.
- Track Cufflinks transcripts across multiple experiments.

Command for cuffcompare

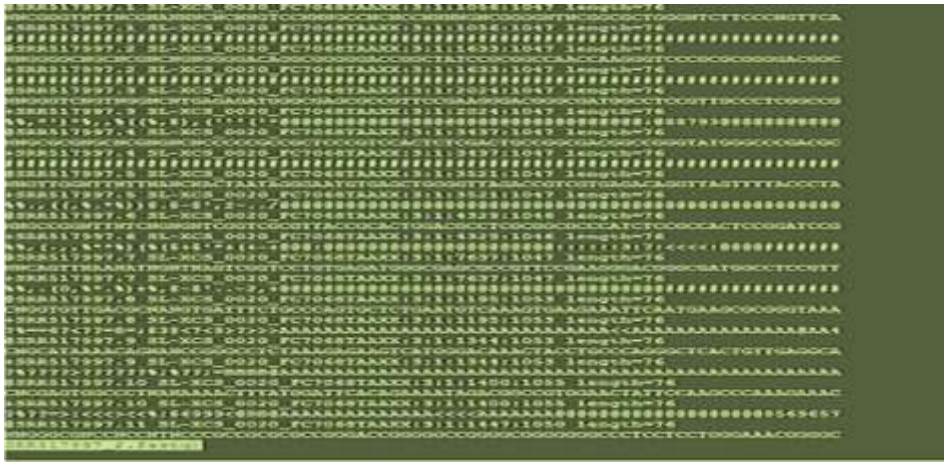
Cuffcompare –G path of gtf files/genes.gtf cufflinks gtf file

RESULTS

The output file for conversion of sra format to fastq format

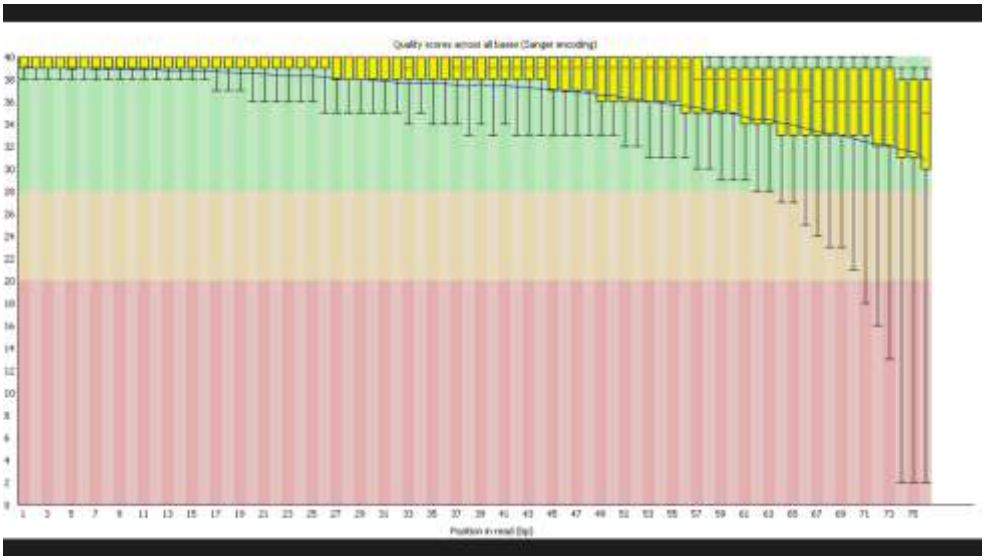


Read 1

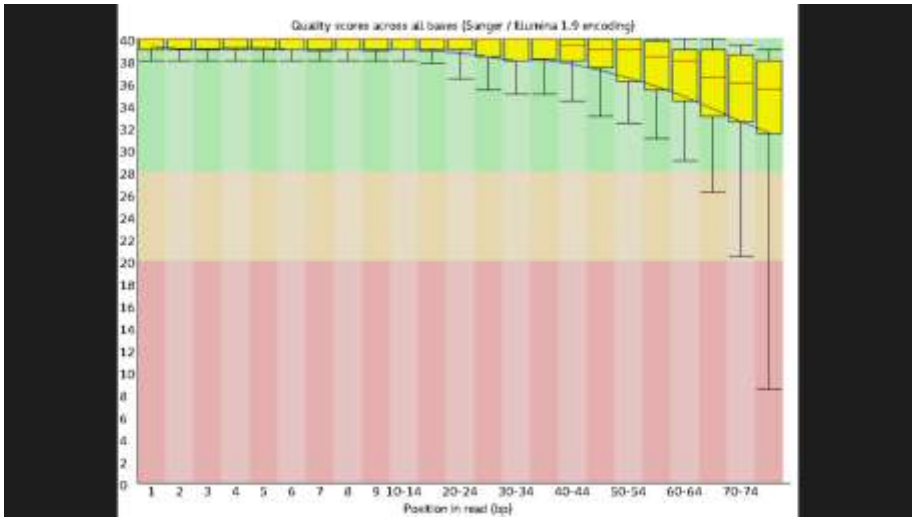


Read 2

Output File For Quality Checking

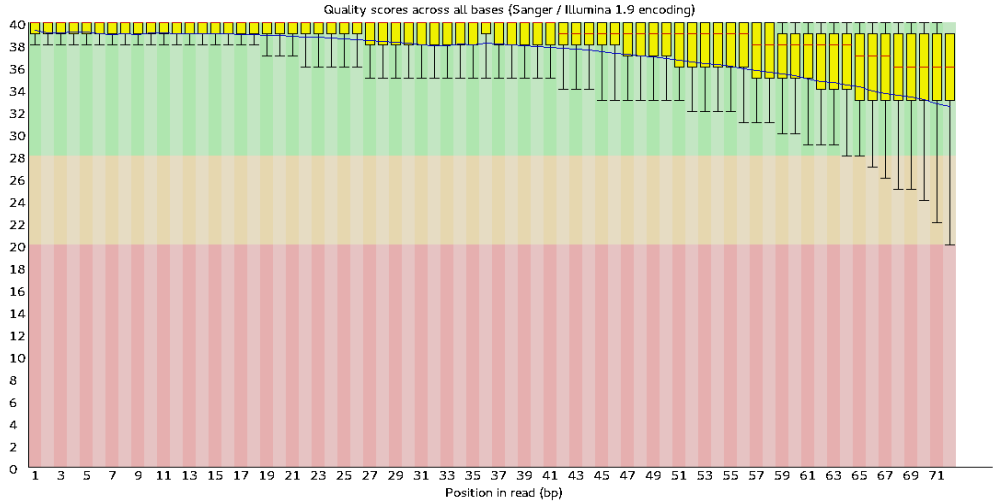


READ 1

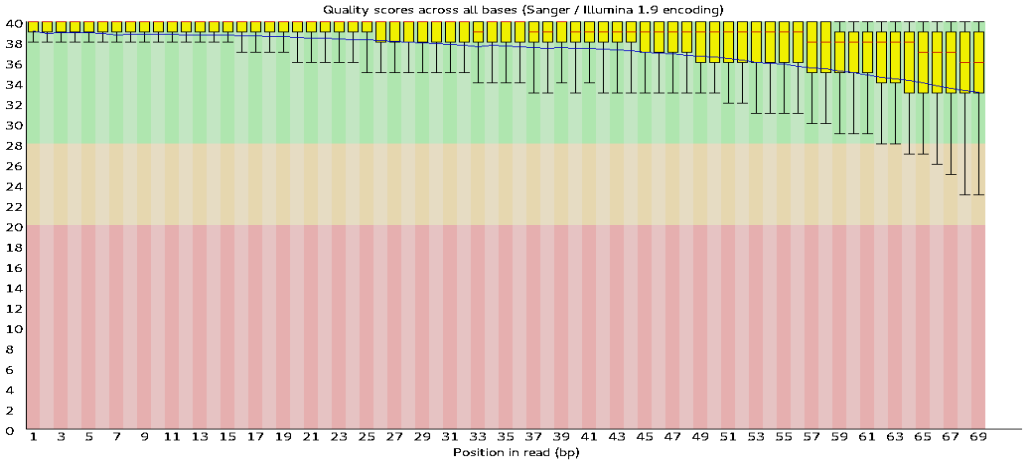


READ 2

Quality Trimmer

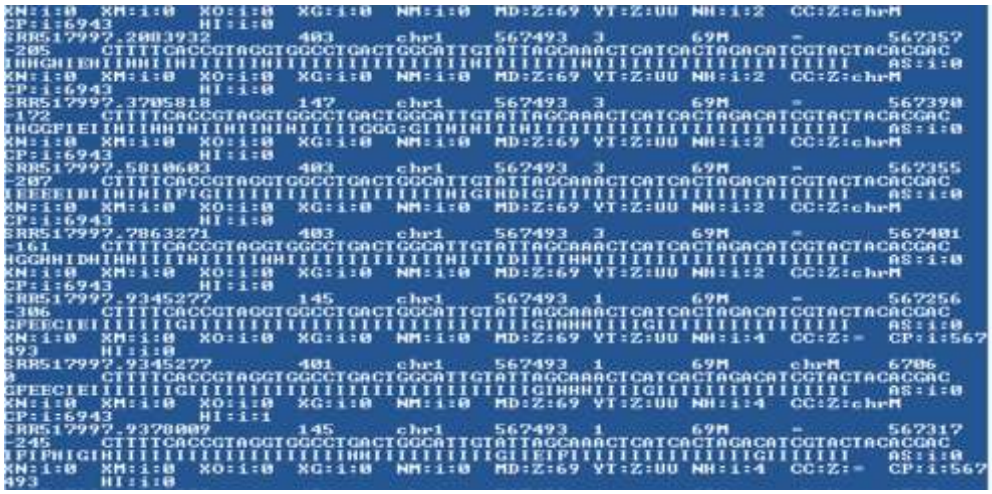


READ 1



READ 2

Index And Allignment Output File



Tophat output :Which contains the sequence

Cufflink Output

We have obtained 53843 transcripts which contains both coding RNA and non coding RNA out of that 837 are noncoding RNA and whose fpkm value is greater than 1.Then we filtered out based on class codes then we have got19 linc RNAs which are known to be novel transcripts.

DISCUSSION

We generated a reference catalog of human lincRNAsbased on integrating RNA-seq data from tissues of the transcripts in our catalog are novel andare now identified for the first time using RNA-seq. Weannotated each lincRNA with a broad range of structural,expression, and evolutionary features, shedding new lighton their global properties and testing or generalizing previoushypotheses.lincRNAs are remarkably tissue-specific compared with protein-coding genes. This possibility was previouslyraised [12,28]based ondifferential expression patterns in specific biologicalsystems and has several implications. First, researchersstudying a particular system may benefit from RNA-seqprofiling followed by de novo assembly in that system.Second, it is consistent with the hypothesis that somelincRNAs interact with chromatinmodulators and providetheir target specificity. Third, itmay indicate that lincRNAs could serve as specific fine-tuners. Fourth, the low level oflincRNA expression in a complex tissue such as the embryonic stemcells may in fact be a by-product of their expression in only afew specific cells. Future targeted perturbations of tissuespecificlincRNAs defined in our studymay elucidate their role in tissue-specific processes.Could many lincRNAs act as enhancer elements, promoting the transcription of their neighboring codinggenes? Recent studies have demonstrated that severallincRNAs have enhancer-like functions[24,25]. While our coexpression analysis is consistentwiththis notion, it is insufficient to suggest a globaltrend in which lincRNAs act as enhancers and Substantial progress has been recently made toward theessential goal of annotating long noncoding RNA loci.Our study presents an integrative yet conservative computationalapproach to mapping lincRNA transcripts thatcan be used for mapping new transcripts in other species.This is critical to overcome major barriers for future experiments(e.g., cloning, expression profiling, gain of function,and loss of function), as well as for the interpretationof genetic association studies. Indeed, 414 lincRNAs in ourcatalog stand out as located within intergenic regions associated. Future work will be necessaryto identify RNA sequence domains that relate tofunction [26,27]], and tofurther classify lincRNAs into families and it can be connected to networkpath analysis. Our panorama oflincRNA properties will greatly advance these goals.

CONCLUSION

Present analysis based on paired end linc RNA data sample of homo sapiens ,we have identified 4383 linc RNA transcripts and 19 linc RNA transcripts are known to be novel transcripts.The work can be connected to network path analysis for the further analysis . It will be greater contribution to structural biology and molecular biophysics researchers.Future work will be necessary to identify RNA sequence domains that relate to function.

REFERENCES

1.Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R,Ravasi T, Lenhard B, Wells C, et al: The transcriptional landscape of themammalian genome. Science 2005, 309(5740):1559-1563.

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

3. Kapranov P, Drenkow J, Cheng J, Long J, Helt G, Dike S, Gingeras TR:Examples of the complex architecture of the human transcriptomerevealed by RACE and high-density tiling arrays. Genome Res 2005,15(7):987-997.

4. Timmers HT, Tora L: The spectacular landscape of chromatin and ncRNAsunder the Tico sunlight. EMBO Rep 11(3):147-149.

5. Mercer TR, Dinger ME, Mattick JS: Long non-coding RNAs: insights intofunctions. Nat Rev Genet 2009, 10(3):155-159.

6.Wang Z, Gerstein M, Snyder M: RNA-Seq: a revolutionary tool fortranscriptomics. Nat Rev Genet 2009, 10(1):57-63.

7.Roberts A, Pimentel H, Trapnell C, Pachter L: Identification of noveltranscripts in annotated genomes using RNA-Seq. Bioinformatics27(17):2325-2329.

8.Ponjavic J, Ponting CP, Lunter G..Functionality or transcriptional noise? Evidence for selection within longnoncoding RNAs(2007). Genome Res 17: 556–565.

9.Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D,Huarte M, Zuk O, Carey BW, Cassady JP, et al. 2009.Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature 458:223–227.

10.Khalil AM, Guttman M, Huarte M, Garber M, Raj A, Rivea Morales D, Thomas K, Presser A, Bernstein BE, vanOudenaarden A, et al. 2009. Many human large intergenicnoncoding RNAs associate with chromatin-modifying complexesand affect gene expression. Proc Natl Acad Sci 106:11667–11672.

11. Orom UA, Derrien T, Beringer M, Gumireddy K, Gardini A, Bussotti G, Lai F, Zytnicki M, Notredame C, Huang Q, et al.2010. Long noncoding RNAs with enhancer-like function inhuman cells. Cell 143: 46–58.
12.Mercer TR, Dinger ME, Mattick JS. 2009. Long non-coding RNAs: insights into functions. Nat Rev Genet 10: 155–159.
13.Ponting CP, Oliver PL, Reik W. 2009. Evolution and functions oflong noncoding RNAs. Cell 136: 629–641.
14.Rinn JL, Kertesz M, Wang JK, Squazzo SL, Xu X, Brugmann SA, Goodnough LH, Helms JA, Farnham PJ, Segal E, et al. 2007.Functional demarcation of active and silent chromatindomains in human HOX loci by noncoding RNAs. Cell129: 1311–1323.
15.Nagano T, Mitchell JA, Sanz LA, Pauler FM, Ferguson-Smith AC, Feil R, Fraser P. 2008. The Air noncoding RNA epigeneticallysilences transcription by targeting G9a to chromatin.Science 322: 1717–1720.
16.Loewer S, Cabili MN, Guttman M, Loh YH, Thomas K, Park IH, Garber M, Curran M, Onder T, Agarwal S, et al. 2010. Large intergenic non-coding RNA-RoR modulates reprogramming of human induced pluripotent stem cells. Nat Genet 42:1113–1117.
17.Harrow J, Denoeud F, Frankish A, Reymond A, Chen CK, ChrastJ, Lagarde J, Gilbert JG, Storey R, Swarbreck D, et al. 2006.GENCODE: producing a reference annotation for ENCODE.Genome Biol 7: S4. doi: 10.1186/gb-2006-7-s1-s4.
18.Amaral PP, Clark MB, Gascoigne DK, Dinger ME, Mattick JS. 2010. lncRNAdb: a reference database for long noncodingRNAs. Nucleic Acids Res 39: D146–D151. doi: 10.1093/nar/gkq1138.
19.Hsu F, Kent WJ, Clawson H, Kuhn RM, Diekhans M, Haussler D. 2006. The UCSC known genes. Bioinformatics 22: 1036– 1046.
20.Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X, Fan L, Koziol MJ, Gnirke A, Nusbaum C, et al. 2010. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nat Biotechnol 28: 503–510.
21.Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes byRNA-seq. Nat Methods 5: 621–628.
22.Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. 2010. Transcriptassembly and quantification by RNA-seq reveals unannotatedtranscripts and isoform switching during cell differentiation. Nat Biotechnol 28: 511–515.

23.Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. 2009. Potential etiologic and functionalimplications of genome-wide association loci forhuman diseases and traits. Proc Natl Acad Sci 106: 9362–9367.

24.Orom UA, Derrien T, Beringer M, Gumireddy K, Gardini A, Bussotti G, Lai F, Zytnicki M, Notredame C, Huang Q, et al. 2010. Long noncoding RNAs with enhancer-like function in human cells. Cell 143: 46–58.

25.Wang KC, Yang YW, Liu B, Sanyal A, Corces-Zimmerman R, Chen Y, Lajoie BR, Protacio A, Flynn RA, Gupta RA, et al. 2011. A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. Nature 472: 120–124.

26.Zhao J, Sun BK, Erwin JA, Song JJ, Lee JT. 2008. Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. Science 322: 750–756.

27.Kanhere A, Viiri K, Araujo CC, Rasaiyaah J, Bouwman RD, Whyte WA, Pereira CF, Brookes E, Walker K, Bell GW, et al. 2010. Short RNAs are transcribed from repressed polycomb target genes and interact with polycomb repressive complex- 2. Mol Cell 38: 675–688.

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