Mining Mammalian Transcript Data for Functional Long Non-Coding RNAs

Amit N. Khachane*, Paul M. Harrison
Department of Biology, McGill University, Montreal, Quebec, Canada

Abstract

Background: The role of long non-coding RNAs (lncRNAs) in controlling gene expression has garnered increased interest in recent years. Sequencing projects, such as Fantom3 for mouse and H-InvDB for human, have generated abundant data on transcribed components of mammalian cells, the majority of which appear not to be protein-coding. However, much of the non-protein-coding transcriptome could merely be a consequence of ‘transcription noise’. It is therefore essential to use bioinformatic approaches to identify the likely functional candidates in a high throughput manner.

Principal Findings: We derived a scheme for classifying and annotating likely functional lncRNAs in mammals. Using the available experimental full-length cDNA data sets for human and mouse, we identified 78 lncRNAs that are either syntenically conserved between human and mouse, or that originate from the same protein-coding genes. Of these, 11 have significant sequence homology. We found that these lncRNAs exhibit: (i) patterns of codon substitution typical of non-coding transcripts; (ii) preservation of sequences in distant mammals such as dog and cow; (iii) significant sequence conservation relative to their corresponding flanking regions (in 50% cases, flanking regions do not have homology at all; and in the remaining, the degree of conservation is significantly less); (iv) existence mostly as single-exon forms (8/11); and, (v) presence of conserved and stable secondary structure motifs within them. We further identified orthologous protein-coding genes that are contributing to the pool of lncRNAs; of which, genes implicated in carcinogenesis are significantly over-represented.

Conclusion: Our comparative mammalian genomics approach coupled with evolutionary analysis identified a small population of conserved long non-protein-coding RNAs (lncRNAs) that are potentially functional across Mammalia. Additionally, our analysis indicates that amongst the orthologous protein-coding genes that produce lncRNAs, those implicated in cancer pathogenesis are significantly over-represented, suggesting that these lncRNAs could play an important role in cancer pathomechanisms.

Citation: Khachane AN, Harrison PM (2010) Mining Mammalian Transcript Data for Functional Long Non-Coding RNAs. PLoS ONE 5(4): e10316. doi:10.1371/journal.pone.0010316

Introduction

With the rapid development in high-throughput sequencing methods, one is now able to describe the mammalian transcriptome in great detail [1,2,3]. Not only is the mammalian transcriptome vast (comprising millions of RNA transcripts) [1], but also is quite unexpectedly diverse. For example, transcript lengths vary from 18 nucleotides (small interfering RNAs) to more than 15,000 nucleotides (in the case of macroRNAs or long non-protein-coding RNAs). Some protein-coding genes not only encode proteins but also contribute to the non-protein-coding RNA pool [4]. It is to be noted however that a significant proportion of the mammalian transcriptome could simply be ‘transcriptional noise’ [5,6,7,8]. A wealth of data is now available for the two most studied mammalian genomes (human and mouse), and the chief challenge is to mine this data effectively for functionally relevant sequences. In this study, we have mined the full-length mammalian transcript (cDNA) data sets from the H-Invitational [3] and Fantom3 [2] projects, to identify potentially functional long non-protein-coding RNAs (lncRNAs). Our rationale was that those lncRNAs (>= 200 nucleotides) that are expressed in human and mouse and preserved in distant relatives, plus that show features of primary sequence and secondary structure conservation, are likely to be functional. We were also interested in knowing whether lncRNAs are transcribed from orthologous protein-coding genes, and if so, from which ones. A positive finding would indicate the conserved role of such protein-coding genes in producing noncoding RNAs, and also would indicate probable functional categories of the lncRNAs.

Previously, we developed a computational pipeline to annotate ‘transcribed pseudogenes’ (tPgs), a class of long non-protein-coding RNAs that are homologous to protein-coding gene transcripts, but which harbor features indicative of a lack of protein-coding ability [9]. We discovered thousands of cases of transcribed pseudogene annotations in the human genome, and filtered the list to identify potential functional cases. In this paper, in a complementary
analysis, we have identified conserved non-

Long non-protein-coding RNAs (also termed ‘messenger-like’ or
‘messenger-RNA-like’ non-coding RNAs) usually bear features of
mRNAs, viz., 5’ capping, splicing and polyadenylation. However,
they do not code for any protein. Although some well-
characterized cases lack sequence conservation indicative of
possible lineage-specific adaptive evolution [5,8], a recent
experimental work using chromatin immunoprecipitation and
massively parallel sequencing (ChIP-Seq) identified several
>1500 ‘large, intervening ncRNAs’ that have some signatures of
evolutionary conservation [10], thus challenging the current
notion that lncRNA are not generally evolutionarily conserved.

Examples of well-known functional long non-protein-coding
RNAs include: Xist, and H19. Xist mediates X chromosome
silencing as part of heterogametic dosage compensation during
development [11,12]. H19 regulates expression of its neighboring
gene Igf2, during embryogenesis, and may act as a tumour
suppressor [13,14,15]. Recently, by means of comparative
genomics, conserved long non-protein-coding RNAs have been
identified [16], but authors have either ignored the regions that
overlap protein-coding genes, or considered smaller length human
transcripts (EST sequences) as a proxy for transcription in the
absence of full-length non-protein-coding transcripts. It is possible
that non-protein-coding sequences arise in part from protein-
coding genes, for example, comprising of only UTR regions, or
including retained introns, in their non-protein-coding transcripts.
We propose that such cases have to be included in the category of
long non-protein-coding RNAs, and that some cases cannot be
clearly classified as either alternative splicing or partially
overlapping lncRNAs. Another parameter we considered as
essential was the length of potential lncRNA transcripts. In the
present analysis, we used a lower bound of 200 nucleotides for the
operational definition of lncRNAs, as in earlier work [8,17,18].
This criterion was chosen on the basis of a suitable practical cut-off
during RNA purification steps to exclude small RNAs.

Results and Discussion
Identification of conserved and expressed lncRNAs
H-Inv and Fantom3 projects catering to the human and mouse
genomes, respectively, have generated thousands of sequence
reads constituting expressed complements of the genomes [1,2,3].
Mere expression however does not necessarily indicate function-
ality. Many of these transcripts may simply be ‘transcriptional
noise’ [5,6]. Expressed elements that are syntenically conserved in
phylogenetically divergent mammals are likely to be functional
across Mammalia. Although a lot of transcripts could potentially be
degradation products of UTRs or incompletely processed hnRNA
fragments [7], natural selection would ensure preservation of
biologically relevant genomic elements over millions years of
evolution. Therefore, we developed a pipeline to identify
potentially functional lncRNA candidates (fig. 1). We defined
putative lncRNAs as full-length transcripts \( \geq 200 \) nucleotides
that do not: (i) exclusively contain known protein-coding exons; (ii)
contain UTR plus protein-coding exons. We examined for
syntetic conservation between the human and mouse genomes
(see Methods for details). Additionally, we were also interested in
identifying lncRNAs that originate from orthologous genes. Such
genes may give hints to the function of lncRNAs. We found that
78 lncRNAs are syntenically conserved or originate from
orthologous genes (Table 1). Some of these have detectable
sequence similarity (Table 2). It is imperative that we find
previously characterized functional lncRNAs in the list. Indeed,

A significant proportion of putative functional lncRNAs
originate from cancer-related genes
We found that \( \sim 35\% (20/57) \) of the protein-coding genes that
overlap with the annotations of the identified lncRNAs are
implicated in the causation of diseases, particularly cancer
(Table 3). To assess for the possible enrichment of such genes
we proceeded as follows. We counted the number of lncRNA-
producing genes from our list that are listed in the ‘CGMIM’
database [21]. ‘CGMIM’ provides a list of all gene entries in
OMIM that referred to some type of cancer. \( \sim 18\% \) of the protein
coding genes that produce lncRNAs (10/57) have reference to
cancer (see Table 3), whereas only 9% of all human protein coding
genes (2147/23621) have reference to cancer. The above
difference is statistically significant (chi-square test, \( P \)-value:
0.047; hypergeometric probability \( P \)-value=0.018), suggesting that the
genes implicated in cancer causation have a higher tendency to
produce lncRNAs. It has been earlier found that ncRNAs have
altered expression/splicing in cancer cells [22,23]. Thus, we
believe that the identified lncRNAs could have potential roles in
oncogenesis, although of course, we cannot ascertain here whether
there is a ‘cause-and-effect’ relationship.

Putative functional lncRNAs typically bear single
non-coding exon
We performed an intron/exon analysis on the identified set of
putative functional lncRNAs to study the contribution of splicing
to their generation, thereby assessing the possible relationship
between lncRNA splicing and function. We found that a vast
majority (\( \sim 83\% \), 65 out of 78) of the above lncRNAs contains just
a single exon. This suggests that functional lncRNAs tend to have
a single exon, and may thus (although speculative) reflect
avoidance of unnecessary (complex) involvement of splicing
mechanism regulation in lncRNA generation.

Examples of potential functional lncRNAs include
cases that overlap Dicer and U2AF2. Dicer is an endoribonuclease
that cleaves double-stranded RNAs into shorter double-stranded
segments called small interfering RNAs (siRNAs) [24,25,26]. The
Figure 1. A schematic representation of the discovery pipeline for conserved expressed long non-protein-coding RNAs (IncRNAs).

doi:10.1371/journal.pone.0010316.g001
Role of some lncRNAs in post-transcriptional regulation

Long non-protein-coding RNAs are known to play a role in the post-transcriptional regulation of target genes [8]. We found two examples of lncRNAs (HIT000097026.8 and HIT000091723.8) that are transcribed in the antisense direction to the orientation of the UTR region of the protein-coding gene (in these cases, also, there are no other protein-coding exons that overlap on the other strand in these particular genomic regions). These lncRNAs could therefore act as negative regulators of gene expression by complementary binding to the UTRs of target mRNAs (fig. 3). A good example is that of an lncRNA gene expression by complementary binding to the UTRs of these mRNAs. These lncRNAs could therefore act as negative regulators of gene expression by complementary binding to the UTRs of target mRNAs (fig. 3). A good example is that of an lncRNA (HIT000091723.8) that are transcribed in the antisense direction to the orientation of the UTR region of the protein-coding gene (in these cases, also, there are no other protein-coding exons that overlap on the other strand in these particular genomic regions). These lncRNAs could therefore act as negative regulators of gene expression by complementary binding to the UTRs of target mRNAs (fig. 3). A good example is that of an lncRNA (HIT000091723.8) that are transcribed in the antisense direction to the orientation of the UTR region of the protein-coding gene (in these cases, also, there are no other protein-coding exons that overlap on the other strand in these particular genomic regions).

Based on the above findings, we suggest a general model for the development of certain cancer types [28]. Therefore, it is possible that the lncRNA may also be involved in carcinogenesis. We then investigated whether any of the long non-coding RNAs (>200 nts) encode thermodynamically stable and conserved secondary-structure motifs, a finding that could lend support to their functional role. For this, we used the program RNAz [31] to examine for the conservation of stable secondary structure motifs in orthologous sequences. RNAz calculates a “RNA class probability” or P-value based on structural conservation index and thermodynamic stability scores. Alignments with P>0.5 are classified as functional RNA. We found that nearly 45% of the identified lncRNAs, i.e., 5 of the 11 orthologous lncRNAs that have detectable homology, have conserved and stable secondary structure motifs (i.e., P-value>0.5). This further strengthens our case that these lncRNAs could represent biologically relevant sequences.

**Table 1.** General statistics for the 78 conserved lncRNAs.

| Category                        | Number of cases* |
|---------------------------------|------------------|
| Syntenically conserved          | With significant sequence homology: 11 |
|                                  | Without significant sequence homology: 67 |
| Genomic location                | Protein-coding region: 57 |
|                                  | Non-protein-coding region: 21 |
| Spliced forms                   | Spliced: 64 |
|                                  | Non-spliced: 14 |

*BLAST e-value set was to <1x10^-6. The protein-coding region annotations were taken from the ENSEMBL website (www.ensembl.org).

doi:10.1371/journal.pone.0010316.t001

**Table 2.** A summary of the analysis results for preservation, sequence conservation and occurrence of secondary structure motifs in mouse lncRNAs that have orthologous human counterparts (with BLAST homology).

| Fantom3 entries | H-Inv db entries (syntenic to mouse lncRNA and having BLAST homology, e-value: <0.01) | Preservation in other mammals* | Sequence identities between conserved lncRNAs and between orthologous flanking regions indicated in brackets | Conservation of secondary structure motifs |
|-----------------|-------------------------------------------------------------------------------------|--------------------------------|-------------------------------------------------------------------------------------------------|------------------------------------------|
| A230108N10      | HIT000394689.1                                                                      | HMDC                           | 80.3% (not conserved)                                                                           | no                                       |
| 0730030106      | HIT000294155.8                                                                      | HMDC                           | 64.3% (not conserved)                                                                           | yes                                      |
| A430070C22      | HIT000091723.8                                                                      | HMDC                           | 10.12% (not conserved)                                                                          | no                                       |
| 1600017P15      | HIT000389557.3                                                                      | HM                             | 72% (not conserved)                                                                             | no                                       |
| A130061G12      | HIT000294554.8                                                                      | HMDC                           | 25.8% (34.8%)                                                                                  | yes                                      |
| 2600002C05      | HIT000323535.8                                                                      | HMDC                           | 43.4% (not conserved)                                                                           | yes                                      |
| 9530073M10      | HIT000093538.1                                                                      | HMDC                           | 66.1% (40.15%)                                                                                 | no                                       |
| S43043311       | HIT000282711.8                                                                      | HMDC                           | 87.9% (17.2%)                                                                                  | yes                                      |
| S330421F07      | HIT000248175.9                                                                      | HMDC                           | 28.1% (7.2%)                                                                                   | yes                                      |
| 1110021C24      | HIT000292834.1                                                                      | HMDC                           | 48.1% (11.3%)                                                                                  | no                                       |
| G370125G16      | HIT000430538.1                                                                      | HMDC                           | 39.4% (not conserved)                                                                           | no                                       |

*H* refers to human, ‘M’ to mouse, ‘D’ to dog and ‘C’ to cow.

Note: For the calculation of sequence conservation, orthologous sequences to mouse lncRNAs were identified in the human genome using synteny maps and BLAST searches (e-value<0.01) and subjected to further evolutionary analysis.

doi:10.1371/journal.pone.0010316.t002

**Discussion**

Evidence for selection on the identified putative functional lncRNAs

We analysed for features of selection in orthologous lncRNAs that have detectable (significant) similarity between them. As lncRNAs from mouse and human do not completely overlap although they show significant homology, we used mouse lncRNAs as reference sequences, and deduced the orthologous human counterpart by BLASTing [29] mouse lncRNAs against the human genome. We compared the sequence identities of these deduced orthologous lncRNAs to their flanking regions. Buffer (intergenic) regions flanking mouse lncRNAs, of length equivalent to that of lncRNA, were selected and examined for the presence of similar counterparts in near syntenic locations in other mammals. These were then aligned using a global alignment algorithm [30]. From the results (Table 2), it is clear that many lncRNAs do not have conserved flanking regions or are not as significantly conserved as lncRNAs. This indicates that the identified lncRNAs are under selection, thus giving further support to their potential functionality.

Secondary structure analysis

We then investigated whether any of the long non-coding RNAs (>200 nts) encode thermodynamically stable and conserved secondary-structure motifs, a finding that could lend support to their functional role. For this, we used the program RNAz [31] to examine for the conservation of stable secondary structure motifs in orthologous sequences. RNAz calculates a “RNA class probability” or P-value based on structural conservation index and thermodynamic stability scores. Alignments with P>0.5 are classified as functional RNA. We found that nearly 45% of the identified lncRNAs, i.e., 5 of the 11 orthologous lncRNAs that have detectable homology, have conserved and stable secondary structure motifs (i.e., P-value>0.5). This further strengthens our case that these lncRNAs could represent biologically relevant sequences.
Genomic conservation in other mammals

Expression per se does not indicate functionality. Sequences of long noncoding RNAs that are present in distantly related mammals (non-coding RNA orthologs) indicate the presence of evolutionary pressure for their preservation. Such preservation indicates possible functionality. Out of the 11 in our list, we find that 9 are conserved in human, mouse, dog and cow. One of them is preserved in human, mouse and dog, whereas the remaining one is preserved only in human and mouse. This indicates that the identified lncRNAs have been conserved across mammalian speciation.

Table 3. List of lncRNAs associated with known genes implicated in cancer pathogenesis.

| H-Inv id     | Gene name | Cancer                          | References                  |
|--------------|-----------|---------------------------------|-----------------------------|
| HIT00067299.10 | brf1      | Lymphoma                       | CGMIM                       |
| HIT00064387.8 | cbfa2t2    | Leukemia                       | CGMIM                       |
| HIT00257890.10 | dicer1    | Breast, Melanoma, Ovarian      | CGMIM, [25]                 |
| HIT00277951.8 | elF4a2     | Leukemia                       | CGMIM                       |
| HIT00323535.8 | hnrpd1     | Leukemia                       | CGMIM                       |
| HIT00389429.2 | ppaard     | Colorectal                     | CGMIM, [37,38]              |
| HIT00327147.7 | slc12a2    | Colorectal                     | CGMIM                       |
| HIT00079026.8 | st7        | Brain, Breast, Colorectal, Prostate, Ovarian | CGMIM, [28,39] |
| HIT00067550.9 | st8ia1     | Brain, Melanoma                | CGMIM                       |
| HIT00276030.9 | xist       | Breast, Ovarian                | CGMIM                       |
| HIT00284226.9 | rab4a      | Oncogene                       | OMIM                        |
| HIT00024195.13 | akt3      | Melanoma                       | [40,41,42,43]               |
| HIT00383650.1 | ptch1      | Basal cell carcinoma           | OMIM                        |
| HIT00243731.8 | rab18      | reduced expression in Pituitary tumors and its overexpression reverts growth hormone hypersecretion | [44] |
| HIT00248175.9 | nav2/Helad1 | Colorectal carcinomas          | [45]                        |
| HIT000075518.7 | reg      | Breast cancer                  | [46]                        |
| HIT00071420.7 | dach1      | Prostate cancer, Breast cancer | [47,48]                     |
| HIT00389219.2 | rad51L1    | Pulmonary chondroid hamartoma, Uterine leiomyomas | [49,50]                     |
| HIT00089413.9 | tnfap2     | Acute promyelocytic leukemia    | [51]                        |
| HIT00330125.5 | nat1       | Non Hodgkin lymphoma, Urinary bladder cancer susceptibility, Colorectal adenoma susceptibility | [52,53,54]                 |

Note: ‘CGMIM’ database is accessible at http://www.bccrc.ca/ccr/CGMIM/ and ‘OMIM’ database at www.ncbi.nlm.nih.gov/omim/.
doi:10.1371/journal.pone.0010316.t003
Evolutionary analysis of codon substitution rates

A measure of selection pressure for protein-coding ability of genes is the ratio of non-synonymous to synonymous substitution rates (\( \frac{K_a}{K_s} \)). Values significantly \(< 1.0\) indicate purifying selection, whereas neutral selection theoretically yields a value of \( \sim 1.0\). We compared \( \frac{K_a}{K_s} \) values for the above 11 IncRNA ortholog pairs (termed \( \frac{K_a}{K_s} \text{_{lncRNA-ortho}} \)) with the corresponding \( \frac{K_a}{K_s} \) values for their parent/nearby genes (\( \frac{K_a}{K_s} \text{_{parent-ortho}} \)) (fig. 4). These \( \frac{K_a}{K_s} \) values were calculated for the longest ORFs from each IncRNA. Only 19 ORFs out of the 66 possible longest ORFs obtained following six-frame conceptual-translations, were found to have significant similarity to respective human counterparts. Although we considered best-case similarity between any two conceptually translated long open reading frames (see Materials and Methods), we found that codon substitution patterns do not support the hypothesis of protein-coding ability, as the \( \frac{K_a}{K_s} \) ratios for these alignments are mostly in the range 0.5–1.5.

Figure 3. A model for antisense regulation of target mRNA transcripts by IncRNAs. The following IncRNA sequences: HIT000079026.8 and HIT000091723.8, have complementary relationship to UTR of the following protein-coding transcripts: ENST00000393449 and ENST00000383790, respectively. doi:10.1371/journal.pone.0010316.g003

Figure 4. Assessment for protein-coding ability. Comparison between \( \frac{K_a}{K_s} \) values of long ORFs derived from six-frame conceptual translation for human-mouse IncRNAs and orthologous neighboring protein-coding genes.

doi:10.1371/journal.pone.0010316.g004
Conclusion
In this comparative study, we mined publicly available (experimental) data sets of mammalian full-length cDNAs for evolutionarily conserved lncRNAs. These represent novel genomic elements of likely functional relevance. Of course, it cannot be ruled out that some of these apparent lncRNAs are conserved to produce functional short peptides, such as was recently described for two mRNAs in Drosophila [32]. Because quite a number of lncRNAs arise from protein-coding regions, it is conceivable that they are involved in functional roles complementing to that of the parent protein-coding gene. In this vein, we have found that cancer-related genes are over-represented in the protein-coding genes that are contributing to the pool of lncRNAs. This therefore suggests that lncRNAs may play an important role in cancer pathomechanisms.

Materials and Methods

Collection of data
Full-length cDNA datasets for human and mouse were obtained from the H-InvDB (www.h-invitational.jp/) and Fantom3 (http://fantom3.gsc.riken.jp/) databases respectively. Complete genome sequences of mammals were obtained from http://www.ensembl.org (Ensembl release 47 for human genome; Ensembl release 48 for other mammals, namely, rhesus monkey, mouse, rat, cow and dog). Full-length cDNAs with length ≥ 200 nucleotides only were considered for further analysis, as analysis of small RNAs was not the focus of this study. To identify genomic locations of transcripts in mammals, cDNAs were mapped onto the respective genome using GMAP software [33] with match criteria of ≥99% sequence identity and ≥99% sequence coverage.

Identification of orthologous IncRNAs in various sequenced mammalian genomes
Orthologous counterparts to mouse IncRNAs are detected by the presence of a similar sequence at the syntenic position in the other mammalian genome. Based on this criterion, a search was carried out in the target mammal as indicated in the synteny maps, to locate orthologous IncRNAs. The following mammals were included in the analysis: human, monkey, mouse, rat, cow and dog. The pair-wise synteny map data for the various mammals were obtained from http://genome.ucsc.edu/ . For a schematic representation of the discovery pipeline for putative functional IncRNAs, see fig. 1.

Ka/Ks calculation
Although orthologous lncRNAs from mouse and human show significant similarity, they however do not completely overlap. Hence, we deduced the orthologous human IncRNAs counterpart by BLASTing [29] mouse lncRNAs against the human genome. Next, putative lncRNA sequences were conceptually translated in all six frames, and the longest ORF in each frame translation was identified. These long ORFs were then pairwise aligned to assess for possible homology at the protein sequence level using BlastP program of BLAST package [29]. Those showing significant pairwise BlastP homology were short-listed and were used for the calculation of Ka/Ks values using the PAL2NAL web server (www.bork.embl.de/pal2nal/), which integrates PAL2NAL tool [34] and the PAML 4 software package [35].

Secondary structure prediction
RNAz predicts structurally conserved and thermodynamically stable secondary structures (http://rna.tbi.univie.ac.at/cgi-bin/RNAz.cgi). We used the RNAz program with default parameters to check for conserved secondary structure motifs in the set of human-mouse IncRNA orthologs.

Author Contributions
Conceived and designed the experiments: ANK PH. Performed the experiments: ANK. Analyzed the data: ANK PH. Contributed reagents/materials/analysis tools: ANK. Wrote the paper: ANK PH.

References
1. Okazaki Y, Furuno M, Kasukawa T, Adachi J, Bono H, et al. (2002) Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature 418: 503–537.
2. Maeda N, Kasukawa T, Oyama R, Gough J, Frith M, et al. (2006) Transcript annotation in FANTOM3: mouse gene catalog based on physical cDNAs. PLoS Genet 2: e62.
3. Yamazaki C, Murakami K, Fujiy S, Taro Y, Harada E, et al. (2008) The H-Invitational Database (H-InvDB), a comprehensive annotation resource for human genes and transcripts. Nucleic Acids Res 36: D793–799.
4. (2009) Post-transcriptional processing generates a diversity of 3′-modified long and short RNAs. Nature 457: 1029–1032.
5. Brosius J (2003) Waste not, want not: transcript excess in multicellular eukaryotes. Trends Genet 21: 287–288.
6. Ponjavic J, Ponting CP, Lunter G (2007) Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. Genome Res 17: 536–543.
7. Nordstrom KJ, Mirza MA, Almen MS, Gloriam DE, Fredriksson R, et al. (2009) Critical evaluation of the FANTOM3 non-coding RNA transcripts. Genomics 94: 169–176.
8. Ponting CP, Oliver PL, Reik W (2009) Evolution and functions of long noncoding RNAs. Cell 136: 629–641.
9. Khachane AN, Harrison PM (2009) Assessing the genomic evidence for conserved transcribed pseudogenes under selection. BMC Genomics 10: 435.
10. Gutmann M, Amit I, Garber M, French C, Lin MF, et al. (2009) Chromatin signature reveals over a thousand highly conserved long non-coding RNAs in mammals. Nature 458: 223–227.
11. Kay GF, Barton SC, Surani MA, Rastan S (1994) imprinting and X chromosome counting mechanisms determine Xist expression in early mouse development. Cell 77: 639–650.
12. Duret L, Chureau C, Samain S, Weissenbach J, Avner P (2006) The Xist RNA gene evolved in eutherians by pseudogenization of a protein-coding gene. Science 312: 1653–1655.
13. Li YM, Franklin G, Cui HM, Svensson K, He XB, et al. (1998) The H19 transcript is associated with polysomes and may regulate IGF2 expression in mouse. J Biol Chem 273: 20247–20252.
14. Gabory A, Rigoche MA, Yoshimizu T, Dandolo I (2006) The H19 gene: regulation and function of a non-coding RNA. Cytogenet Genome Res 113: 181–193.
15. Yoshimizu T, Miroglio A, Ripoche MA, Gabory A, Vernucci M, et al. (2008) The H19 locus contributes to breast tumor suppression. Proc Natl Acad Sci U S A 105: 12417–12422.
16. Church DM, Goodstadt L, Hillier LW, Zody MC, Goldstein S, et al. (2009) Lineage-specific biology revealed by a finished genome assembly of the mouse. PLoS Biol 7: e1000112.
17. Kapranov P, Cheng J, Dike S, Nis DA, Doughty R, et al. (2007) RNA maps reveal new RNA classes and a possible function for pervasive transcription. Science 318: 1404–1408.
18. Mercer TR, Dinger ME, Mattick JS (2009) Long non-coding RNAs: insights into functions. Nat Rev Genet 10: 155–159.
19. Nam DK, Lee S, Zhou G, Cao X, Wang C, et al. (2002) Oligo(UT) primer generates a high frequency of truncated cDNAs through internal poly(A) priming during reverse transcription. Proc Natl Acad Sci U S A 99: 6152–6156.
20. Bates RT (2006) Structures of regulatory elements in mammalian genes. Curr Opin Struct Biol 16: 299–306.
21. Bajic VD, Kass B, Russow JS, Jones S, Brooks-Wilson A (2005) CGMM: automated text-mining of Online Mendelian Inheritance in Man (OMIM) to identify genetically-associated cancers and candidate genes. BMC Bioinformatics 6: 78.
22. Guillanti A, Iacono M, Pelucchi P, Kim N, Solda G, et al. (2009) A transcriptional sketch of a primary human breast cancer by 454 deep sequencing. BMC Genomics 10: 163.
23. Mattick JS (2009) The genetic signatures of noncoding RNAs. PLoS Genet 5: e1000459.
24. Bernstein E, Caudy AA, Hammond SM, Hannon GJ (2001) Role for a bidentate ribonuclease in the initiation step of RNA interference. Nature 409: 363–366.
25. Chocea S, Jerzéczova E, Chandrana U, Acquafondata M, McHale T, et al. (2006) Up-regulation of dicer, a component of the MicroRNA machinery, in prostate adenocarcinoma. Am J Pathol 169: 169–1820.

26. Marrara IJ, Zhou K, Li F, Repic A, Brooks AN, et al. (2006) Structural basis for double-stranded RNA processing by Dicer. Science 311: 195–196.

27. Black DL (2003) Mechanisms of alternative pre-messenger RNA splicing. Annu Rev Biochem 72: 291–336.

28. Zrenkhusen JC, Conti CJ, Green ED (2001) Mutational and functional analyses reveal that SFT is a highly conserved tumor-suppressor gene on human chromosome 7p11. Nat Genet 27: 392–396.

29. Alsul M, Ghif W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403–410.

30. Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48: 443–453.

31. Gruber AR, Neubock R, Hofacker IL, Washietl S (2007) The RNAz web server: prediction of thermodynamically stable and evolutionarily conserved RNA structures. Nucleic Acids Res 35: W335–338.

32. Xu L, Han C, Lim K, Wu T (2006) A novel positive feedback loop between peroxisome proliferator-activated receptor delta and prostaglandin E2 signaling pathways in human hepatocellular carcinoma. J Biol Chem 281: 33982–33996.

33. Wu TD, Watanabe CK (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics 21: 1859–1875.

34. Suyama M, Torrents D, Bork P (2006) PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res 34: W609–612.

35. Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13: 555–556.

36. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403–410.

37. Zenklusen JC, Conti CJ, Green ED (2001) Mutational and functional analyses reveal that SFT is a highly conserved tumor-suppressor gene on human chromosome 7p11. Nat Genet 27: 392–396.

38. Ishiiro H, Shimokawa T, Tsunoda T, Tanaka T, Fujii Y, et al. (2002) Isolation of HELAD1, a novel human helicase gene up-regulated in colorectal carcinomas. Oncogene 21: 6307–6318.

39. Fuhrin BS, Gao CL, Murphy GA, Shao H, Kimel T, et al. (2001) RERG is a novel ras-related, estrogen-regulated and growth-inhibitory gene in breast cancer. J Biol Chem 276: 42259–42267.

40. Chiosea S, Jelezcova E, Chandrana U, Acquafondata M, McHale T, et al. (2006) Up-regulation of dicer, a component of the MicroRNA machinery, in prostate adenocarcinoma. Am J Pathol 169: 169–1820.

41. Davies MA, Stemke-Hale K, Teller C, Calderone TL, Deng W, et al. (2008) A novel AKT3 mutation in melanoma tumours and cell lines. Br J Cancer 99: 1265–1268.

42. Tran MA, Gowda R, Sharma A, Park EJ, Adams J, et al. (2008) Targeting V600EB-Raf and Akt3 using nanoliposomal-small interfering RNA inhibits cutaneous melanocytic lesion development. Cancer Res 68: 7638–7649.

43. Sharma A, Sharma AK, Madhumantapula SV, Desai D, Huh SJ, et al. (2009) Targeting Akt3 signaling in malignant melanoma using isoselenocyanates. Clin Cancer Res 15: 1674–1685.

44. Moslehi R, Chatterjee N, Church TR, Chen J, Yeager M, et al. (2006) Cigarette smoking, N-acetyltransferase genes and the risk of advanced colorectal adenoma. Pharmacogenomics 7: 381–389.