Hou et al. BMC Genomics (2016) 17:931
DOI 10.1186/s12864-016-3287-9

BMC Genomics

SOFTWARE Open Access

AnnoLnc: a web server for systematically annotating novel human lncRNAs

Mei Hou[1], Xing Tang[1,3], Feng Tian[1,2], Fangyuan Shi[1], Fenglin Liu[1] and Ge Gao[1*]

Abstract

Background: Long noncoding RNAs (lncRNAs) have been shown to play essential roles in almost every important biological process through multiple mechanisms. Although the repertoire of human lncRNAs has rapidly expanded, their biological function and regulation remain largely elusive, calling for a systematic and integrative annotation tool.

Results: Here we present AnnoLnc (http://annolnc.cbi.pku.edu.cn), a one-stop portal for systematically annotating novel human lncRNAs. Based on more than 700 data sources and various tool chains, AnnoLnc enables a systematic annotation covering genomic location, secondary structure, expression patterns, transcriptional regulation, miRNA interaction, protein interaction, genetic association and evolution. An intuitive web interface is available for interactive analysis through both desktops and mobile devices, and programmers can further integrate AnnoLnc into their pipeline through standard JSON-based Web Service APIs.

Conclusions: To the best of our knowledge, AnnoLnc is the only web server to provide on-the-fly and systematic annotation for newly identified human lncRNAs. Compared with similar tools, the annotation generated by AnnoLnc covers a much wider spectrum with intuitive visualization. Case studies demonstrate the power of AnnoLnc in not only rediscovering known functions of human lncRNAs but also inspiring novel hypotheses.

Keywords: Annotation, LncRNAs, Long noncoding RNAs, Transcriptome, Web server

Background

Long noncoding RNAs (lncRNAs) are operationally defined as RNA transcripts that are 1) longer than 200 nt and 2) do not encode proteins [1]. With high-throughput screening and follow-up experimental validation, several studies show that lncRNAs play essential roles in almost every important biological process, including imprinting [2], cell cycles [3], tumorigenesis [4] and pluripotency maintenance [5] through multiple mechanisms, such as guides, scaffolds, and decoys, as well as chromatin architecture organizers [6, 7].

In recent years, the repertoire of human lncRNAs has rapidly expanded. Approximately 50% of human lncRNAs in the GENCODE catalog were identified in the past five years (15,512 in GENCODE v7 increased to 28,031 in GENCODE v24) [8, 9]. A recent study identified more than 30,000 additional unannotated human lncRNAs genes [10]. However, the functional roles of lncRNAs remain largely elusive: less than 1% of identified human lncRNAs have been experimentally investigated [11], driving the need for computational methods.

Several studies have proposed methods for *in silico* prediction of the function of novel lncRNAs. The "guilt-by-association" strategy is the most widely used approach [12]. A dedicated web server, ncFAN, was developed to predict lncRNA functions based on enriched functional terms of coding genes in the same co-expression module [13, 14]; the algorithm was improved by taking protein-protein interaction into account [15]. Moreover, several attempts have been made to characterize molecules interacting with a given lncRNA [16–20]. The large number and immense functional diversity of lncRNAs call for an integrative annotation tool that incorporates broader spectrum of annotations [6]. Hence, we have developed AnnoLnc, a one-stop annotation portal for novel human lncRNAs with rich annotation and user-friendly interface (see Table 1 for a detailed comparison with similar tools).

* Correspondence: gaog@mail.cbi.pku.edu.cn
[1]State Key Laboratory of Protein and Plant Gene Research, College of Life Sciences, Center for Bioinformatics, Peking University, Beijing 100871, P.R. China
Full list of author information is available at the end of the article

Table 1 Comparison of AnnoLnc with similar tools

		AnnoLnc	ncFAN	lncPro	lncTar	LongTarget
Annotation	Genomic location	√	-	-	-	-
	RNA secondary structure	√	-	-	-	-
	Expression	RNA-Seq	Microarray	-	-	-
	Regulation	√	-	-	-	-
	Molecular interaction	Protein/RNA	-	Protein	RNA	DNA
	Network	Co-expression	Bi-color	-	-	-
	Function annotation term catalog	GO	GO + KEGG	-	-	-
	Associated traits/diseases	√	-	-	-	-
	Evolution	√	-	-	-	-
Interface	Text summarization	√	-	-	-	-
	Integrative visualization	√	-	-	-	-
	Web-based	√	√	√	-	√
	Mobile-friendly	√	-	-	-	-
	Web Service APIs	√	-	-	-	-

Feature comparison with existing tools suggests AnnoLnc as the most comprehensive annotation tool with rich supports on user interface

Implementation

Designed as a flexible platform, AnnoLnc consists of multiple annotation modules (see Fig. 1 for the architecture of AnnoLnc).

Genomic location

When a lncRNA is submitted online, AnnoLnc first identifies its genomic coordinate and splicing structure by aligning the input sequence to the human reference genome hg19 with Blat [21]. When a single sequence is aligned in multiple places, genome-wide best alignments are identified by standard pslSort and pslReps. In case of false-positive junction sites caused by mismatches or small indels, putative exons shorter than 20 bp (as well as putative introns shorter than 40 bp) are discarded. The derived coordinates are further compared with annotated human gene models compiled from lncRNAdb [11] and GENCODE [8] (refer to the "Pre-calculated expression profiles of known transcripts" section), and a direct link to the corresponding database entry is provided for hits. Moreover, a link to the UCSC genome browser [22] shows the genomic context of each lncRNA.

Secondary structure

The structure of an RNA molecule is essential for its biological functions. For each input lncRNA, RNAfold v2.0.7 in the ViennaRNA package [23] is employed to predict the secondary structures, with the option "–noLP" enabled to avoid undesirable isolated base pairs. When multiple candidates are available, the one with minimum free energy is kept (as recommended by the authors of the ViennaRNA package) and rendered online as an interactive plot.

Biological functions of secondary structures lead to local stability and bring evolutionary constraints onto the sequences of lncRNAs [24]. To help users identify functional motifs, AnnoLnc allows users to color each base in the structure plot by its corresponding entropy or conservation score (Fig. 3a and b).

Expression profile and co-expression-based functional annotation

A transcript's expression pattern also provides important hints about its functionality [12]. For each input lncRNA, the expression profile is online estimated based on 64 RNA-Seq datasets covering 34 normal samples (16 adult healthy tissues and one embryonic stem cell line with two replicates of each) and 30 cancer cell lines (10 common cancers), then presented in interactive charts (Fig. 3d). Specifically, we mapped the reads of 34 normal samples to human genome hg19 by TopHat (v1.4.1.1). Bam files of 30 cancer samples were downloaded directly from CGHub. (see Additional file 1: Table S3 and S4 for the number of mapped reads and CGHub IDs, respectively.) To improve the response time, the expression of known lncRNAs (including lncRNAs in GENCODE v19 and lncRNAdb v2) was pre-calculated and loaded into the global cache. For novel lncRNAs, we adopt the LocExpress method to perform on-the-fly expression estimation accurately and efficiently.

Pre-calculated expression profiles of known transcripts

We generated a gene model (GM) gtf file (http://annolnc.cbi.pku.edu.cn/about/annolnc_gene_model_v1.gtf.gz) covering human lncRNAs in lncRNAdb v2.0 [11] based on GENCODE [8] v19. First, we downloaded human lncRNA sequences in the lncRNAdb and obtained

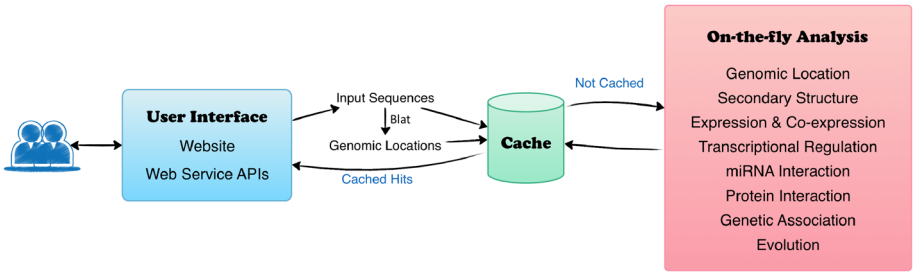


Fig. 1 The architecture of the AnnoLnc web server. Users can submit RNA sequences on the AnnoLnc website or by Web Service APIs. First, AnnoLnc tries to map input sequences to the human genome and, if possible, obtain genomic locations. Then, AnnoLnc searches the global cache for possible hits based on both sequences and aligned splicing structures. Cached results are returned directly if a hit is detected, and novel sequences and loci are sent to on-the-fly analysis modules

transcript structures as described under the "Genomic location" section. These transcript structures were compared with GENCODE v19 by Cuffcompare (v2.1.1). If the code was "=" or "c", the lncRNA was replaced by the known transcript; otherwise it was considered a "novel transcript" and merged into GENCODE v19. The expression of all annotated transcripts in the GM file was pre-calculated by StringTie (v1.0.4) with the options "-e -b", and then normalized by the geometric method in normal and cancer samples separately.

On-the-fly expression estimation of input transcripts

Taking advantage of the local nature of RNA-Seq, we developed a novel quantification method called LocExpress for real-time estimation of the expression level based on pre-mapped reads. Briefly, LocExpress takes full advantage of the locality of RNA-Seq data, and makes the abundance calls increasingly. For a novel transcript, LocExpress will first infer its minimum spanning bundle (MSB), and make the expression call based on reads within the MSB only. Then, the estimated relative abundance is further adjusted and normalized, and reported in canonical FPKM unit. (Refer to Hou et al. [25] for more details). For the normal sample set, the FPKM of two replicates of each tissue/cell line are averaged to report to users.

Co-expression analysis

To help users identify co-regulated partners of the input lncRNA, AnnoLnc reports co-expressed genes based on normal samples and cancer samples. An expression-based functional prediction is further performed by identifying statistically significant enriched Gene Ontology (GO) terms based on co-expressed protein-coding genes. Adjusted p values for the multiple-testing issue are reported as well (Fig. 3e) [12, 26].

Specifically, 34 normal samples and 30 cancer samples were treated separately. To avoid the duplicated GO annotation for isoforms, we first obtained expression profiles at the gene level by adding the FPKM of all transcripts of each gene in the GM file. Then, we filtered these genes as described below, resulting in 29,798 genes in the normal sample set and 25,449 genes in the cancer sample set.

1) FPKM filter. The sum of FPKM in all samples should be not less than 1.
2) Tissue-specific filter. The tissue-specific score is calculated by the "getsgene" function in the R package rsgcc. If a gene has a score larger than 0.85, it is not considered fit for the co-expression analysis.

For submitted transcripts that pass the above filters, the Pearson correlations with genes are calculated. Then, highly correlated genes are reported by a "gradually decreased" criterion to remove putative false positives and retain true positives. If there are more than 10 genes with $r \geq 0.9$, GO enrichment analysis is performed with these genes directly. If not, we determine whether there are 10 genes above the cutoff of 0.8. This process continues until the cutoff arrives at 0.7. Negatively correlated genes are identified in a similar manner. GO enrichment analysis for these correlated genes is further conducted with the R package GOstats, and significantly enriched GO terms (adjusted p value ≤ 0.01, users can also change the cutoff instantly at the result page) are reported as putative functional assignments of the input transcript.

Transcriptional regulation

Transcriptional factors (TF) largely determine the expression level of lncRNAs. AnnoLnc integrated 498 ChIP-Seq datasets covering 159 (TFs) in 45 cell lines (see Additional file 1: Table S5 for more details). Uniform peak files generated by the ENCODE project were downloaded from http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeAwgTfbsUniform/. AnnoLnc locates the binding sites of 159 TFs in the input lncRNA locus, and reports these binding sites based on their relative location to the lncRNA locus, such as "upstream transcriptional start site (TSS)" (defined as

5Kb upstream), "overlap with TSS", "inside the lncRNA loci", "overlap with transcriptional end site (TES)" (defined as 1Kb downstream) and "downstream TES" (Additional file 1: Table S1a). Moreover, we support the *ClosestGene* method as being suggested by Sikora-Wohlfeld et al. [27] which could be enabled by the option "Assign peaks to the closest gene" at the result page.

miRNA interaction

Interacting with miRNAs, lncRNAs can be post-transcriptionally regulated or act as decoys [28]. AnnoLnc provides predicted miRNA family partners of lncRNAs by TargetScan v6.0 [29]. To reduce the potential false positive rate, we run the prediction on 87 highly conserved miRNA families (Additional file 1: Table S6) derived from miRcode [30] (http://www.mircode.org/download.php). Then, conservation scores in primate, mammal and vertebrate clades for each identified site are calculated as described in Jeggari et al. [30]. For example, 10 species in primates are used in the TargetScan prediction, and if a binding site is identified in eight species, the conservation score in primates is 8/10=0.8. In mammals and vertebrates, scores are calculated in the same manner except that "mammals" are "non-primate mammals" (26 species) and "vertebrates" are "non-mammal vertebrates" (10 species). To further highlight high-confidence sites, predicted sites are screened based on a pre-compiled 61 AGO CLIP-Seq dataset (Additional file 1: Table S7, see the "Calling RNA-protein interactions based on CLIP-Seq data" section for more details about CLIP-Seq analysis), and hit sites are considered "CLIP supported".

Protein interaction

lncRNAs can interact with multiple proteins, as guides and/or scaffolds, to perform their functions [31]. For each lncRNA, AnnoLnc reports proteins partners based on both CLIP-Seq data and *in silico* prediction.

Calling RNA-protein interactions based on CLIP-Seq data

CLIP-Seq is one of the most widely used high-throughput methods to detect RNA-protein interactions experimentally [32]. AnnoLnc screens putative protein partners for an input lncRNA in 112 CLIP-Seq datasets covering 51 RNA binding proteins (RBPs) other than AGO. In case of methodology bias introduced by heterogeneous analysis pipelines, all the CLIP-Seq data were reanalyzed locally with a uniform pipeline. Finally, protein partners, cell types, treatments and corresponding p values reported by the analysis pipeline are reported to users.

Briefly, the raw data of CLIP-Seq datasets were downloaded from the Sequence Read Archive (SRA) (see Additional file 1: Table S8 for a full list). we first trimmed the adapter by FASTX Clipper, and only reads longer than 15 nt were kept and mapped to human genome hg19 by the algorithm BWA-backtrack (v0.7.10-r789) [33] with the options "-n 1 -i 0" (allow one alignment error). Then, only unique mapped reads were kept. To improve precision, we used stringent criteria for site calling with PIPE-CLIP v1.0.0 [34]; FDR cutoffs for both enriched clusters and reliable mutations were set as 0.05 (cross-linking sites in HITS-CLIP data identified by deletion, insertion and substitution were combined).

To evaluate the performance of our pipeline, we downloaded raw reads of wild-type FET proteins (FUS, EWSR1 and TAF15) from DDBJ (SRA025082) and performed the analysis described above. For comparison with reported results (Supplementary Data 1, [35]), cross-linking sites identified by both methods were mapped to RefSeq IDs. Our pipeline shows fairly high precision (0.95 for FUS, 0.96 for EWSR1, and 0.91 for TAF). Meanwhile, we evaluated on a HITS-CLIP dataset for the DGCR8 protein [36]. We downloaded raw reads of all four samples (D8.1, D8.2, T7.1 and T7.2) from GEO (GSE39086) and analyzed them as described above (D8.1 data was excluded because PIPE-CLIP failed to generate cross-linking sites with a "model failed to converge" error). Comparison with the original results downloaded from http://regulatorygenomics.upf.edu/Data/DGCR8/ also shows good precision (0.89 for D8.2, 0.74 for T7.1, and 0.78 for T7.2).

Ab initio prediction of lncRNA-protein interaction

AnnoLnc conducts *in silico* prediction across the entire human proteome for each lncRNA by lncPro [17]. We downloaded 99,459 human protein sequences from Ensembl, filtered 1,917 sequences that could not be processed by lncPro (containing "*", "X", "U" or length not within 30–30,000 AA), ultimately obtaining 97,542 protein sequences. For efficiency, we modified the source code of lncPro to pre-calculate all protein features in batch. To improve specificity, we further derived the statistical significance of the interaction scores reported by lncPro based on empirical NULL distribution (Additional file 2: Figure S1) generated by random shuffling. Interactions with a p value ≤ 0.01 are considered to be significant. Then, the predicted protein partners, interaction scores and empirical p values are reported. To make the results more intuitive, Ensembl IDs are finally converted into HGNC gene symbols. If multiple Ensembl IDs are mapped to one gene symbol, the score with the smallest p value are reported.

Genetic association

Large-scale genetic association studies enable detection of multiple phenotypic traits that lncRNAs may associate with [37]. By integrating the NHGRI GWAS Catalog [38] (downloaded from the UCSC genome browser), AnnoLnc links an input lncRNA to diseases/traits based

on strong linkage blocks defined by linkage disequilibrium (LD) values in multiple populations. Specifically, AnnoLnc first scans all SNPs within the transcript region (5 Kb upstream to 1 Kb downstream of each input transcript). Using one of these SNPs as an example, a SNP is linked with a tagSNP if it is within the haplotype region (defined as $r^2 > 0.5$, ftp://ftp.ncbi.nlm.nih.gov/hapmap/ld_data/2009-04_rel27/) tagged by the tagSNP reported in the NHGRI GWAS Catalog. Then, this linked SNP, corresponding tagSNPs, traits/diseases, *p* values, significance (defined as *p* value ≤ 5e-8), LD values, populations from which these LD values are derived, as well as supporting PubMed IDs, are reported by AnnoLnc (Additional file 1: Table S2).

Evolution

The evolutionary signature is an important hint as to biological function. For each submitted lncRNA, we incorporated the 46 way phyloP score (primates, mammals/placentals and vertebrates) from the UCSC Genome Browser, and the derived allele frequency (DAF) [39] of the YRI population (Yoruba in Ibadan, Nigeria) from http://compbio.mit.edu/human-constraint/ for every position (if has corresponding scores) of both the exon and promoter region (1 Kb upstream). To obtain an overall view, AnnoLnc calculates the mean scores for the exon and promoter regions, and organized the scores into interactive bar charts.

Because many lncRNAs are partially conserved, we also report conserved elements predicted by phastCons [40] in different clades with the length and score, which is an indicator of conservation. The phastCons conserved elements were downloaded from ftp://hgdownload.cse.ucsc.edu/goldenPath/hg19/database. The score reported to users is the LOD score. Conserved elements shorter than 20 bp are omitted from the table. These conserved blocks can help users identify functional elements combined with other annotation results, especially in the integrated view (Fig. 3f).

AnnoLnc website

The AnnoLnc website runs on the Tomcat server. The backend is based on Java Servlet and MySQL database. In the frontend, some JavaScript libraries are used to facilitate accessibility. Bootstrap is used for the mobile friendly layout. jQuery is used for Ajax. DataTables is used to show tables and Highcharts for interactive charts. The display of the interactive SVG plot is enabled by "svg-pan-zoom", available at https://github.com/ariutta/svg-pan-zoom.

Results and Discussion

User Interface

AnnoLnc is designed to be intuitive. The most common operations (such as submitting sequences and obtaining annotation results) can be performed with just a few clicks. As showed in Figs. 1 and 2, users can submit RNA sequences in fasta format at the "Home" page of

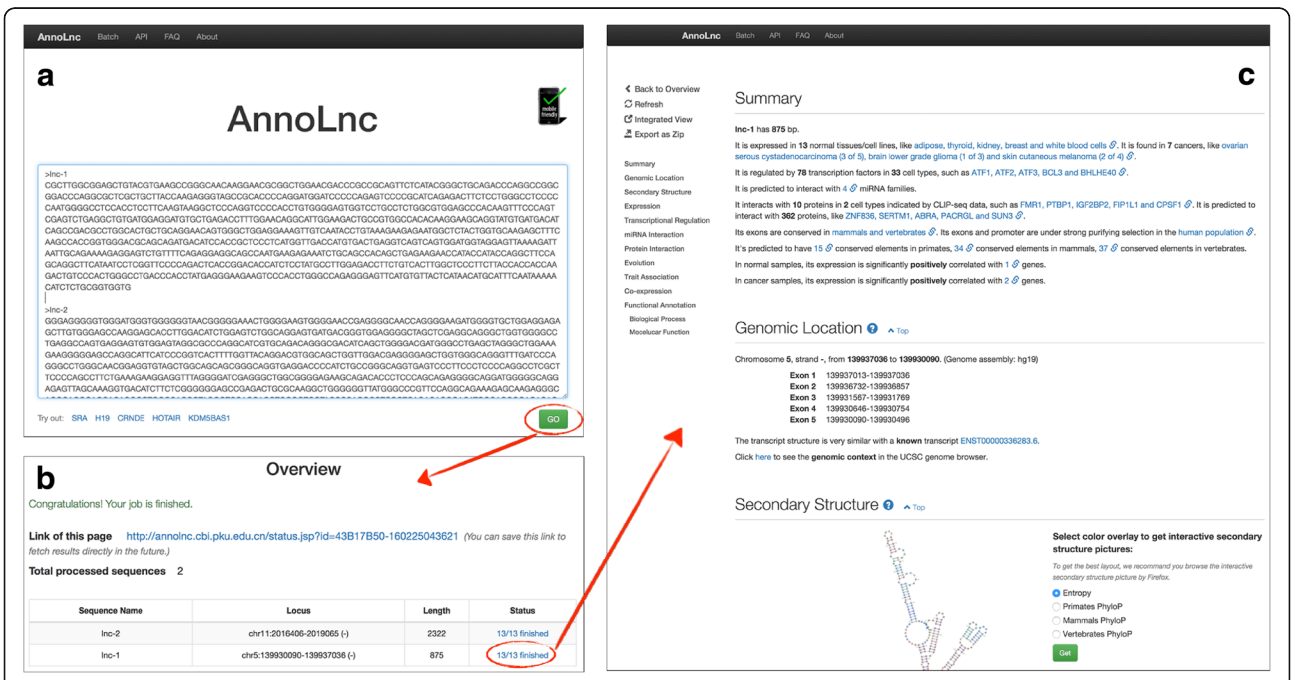
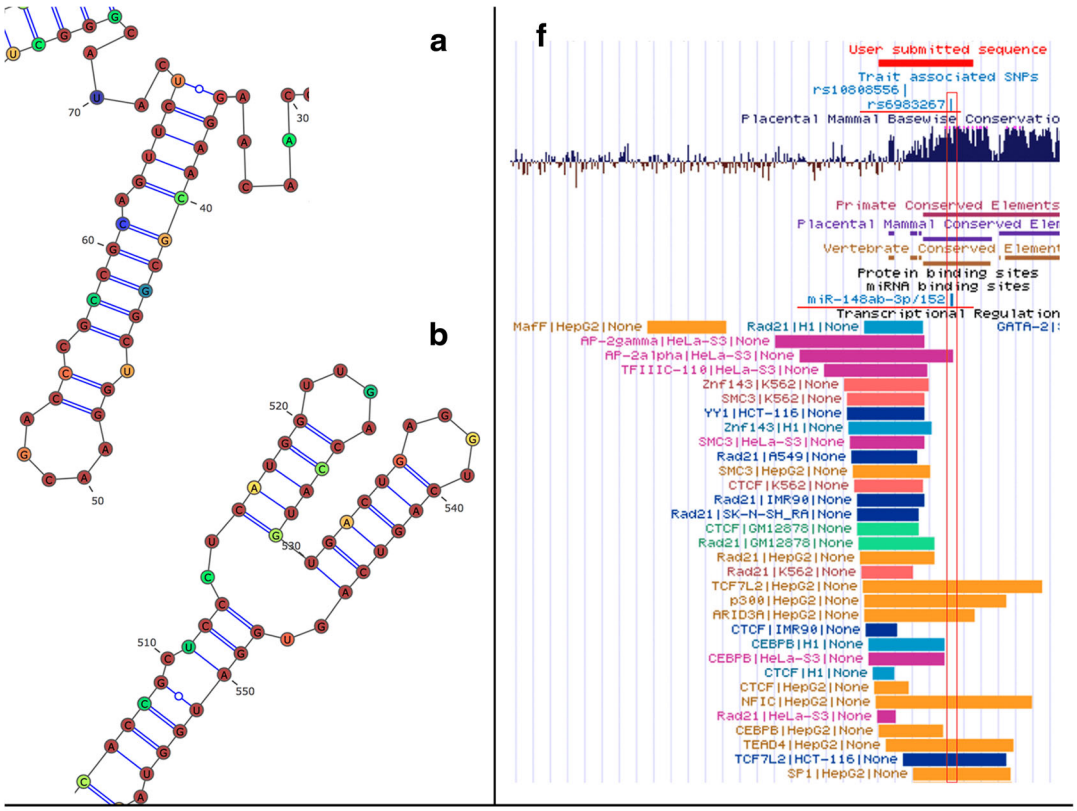


Fig. 2 The web interface of AnnoLnc. Only a few clicks are required from submission of sequences to receipt of annotation results. **a** The "Home" page. Users can submit lncRNA sequences here to start the on-the-fly analysis. **b** The "Overview" page of run status and basic information of processed lncRNAs. **c** An example of the annotation result page. Most annotations are presented in this page with a sidebar on the left for quick jump



Summary

c

H19 has **2322** bp.

It is expressed in **14** normal tissues/cell lines, like skeletal muscle, thyroid, adipose, breast and testes. It is found in **6** cancers, like ovarian serous cystadenocarcinoma (2 of 5), lung squamous cell carcinoma (3 of 6) and skin cutaneous melanoma (3 of 4).

It is regulated by **23** transcription factors in **10** cell types, such as AP-2alpha, AP-2gamma, Bach1, CEBPB and c-Myc.

It is predicted to interact with 18 miRNA families. Among those, **18** are supported by CLIP-seq data, such as miR-130ac/301ab/301b/301b-3p/454/721/4295/3666, miR-19ab, miR-148ab-3p/152, miR-140/140-5p/876-3p/1244 and miR-138/138ab.

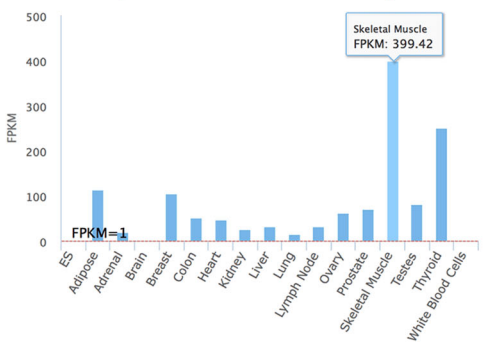
It interacts with **4** proteins in **2** cell types indicated by CLIP-seq data, such as EIF4A3, HNRNPU, PTBP1 and FUS. It is predicted to interact with **3133** proteins, like CNOT3, C6orf25, FDFT1, ATP6AP1 and STX17.

It's predicted to have 4 conserved elements in primates, 14 conserved elements in mammals, 14 conserved elements in vertebrates.

In normal samples, its expression is significantly **positively** correlated with 59 genes, which indicates it is involved in biological processes like muscle system process, and has molecular functions like structural constituent of muscle.

In cancer samples, its expression is significantly **positively** correlated with 11 genes.

d



e

Positively Correlated

Normal Samples

Search:

GO Term	Description	p Value
GO:0003012	muscle system process	1.34e-3
GO:0006936	muscle contraction	3.48e-3
GO:0033275	actin-myosin filament sliding	3.81e-3
GO:0030049	muscle filament sliding	3.81e-3
GO:0048747	muscle fiber development	3.81e-3
GO:0061061	muscle structure development	4.39e-3

Showing 1 to 6 of 6 entries Previous 1 Next

Fig. 3 (See legend on next page.)

(See figure on previous page.)
Fig. 3 The case studies for AnnoLnc. **a-b** The case study of lncRNA *SRA*. In the interactive secondary structure plot with vertebrate phyloP scores as the color overlay, two sub-structures are very easy to be identified because most bases are colored red. **a** is a hairpin region that corresponds to the most conserved H2 sub-structure highlighted by Novikova et al. [42]. **b** is a three-way junction hairpin region that is very similar to the conserved regions H15, H16 and H17 verified by Novikova et al. [42]. **c-e** The case study of lncRNA *H19*. **c** is a summary of the annotation results, which helps users quickly grasp the essentials. **d** is the expression profile of *H19* in normal samples. It has the highest expression in "skeletal muscle". **e** is the predicted GO terms based on positively correlated coding genes in normal samples. The terms are all muscle related. **f** The integrative view of lncRNA *CCAT2* in the UCSC genome browser for annotations at the transcript level. It is easy to determine that rs6983267 is within the seed binding site of the miRNA family miR-148ab-3p/152

AnnoLnc website (Fig. 2a) or by Web Service APIs. AnnoLnc tries to map input sequences to the human genome and, if possible, obtain genomic locations and report them in the "Overview" page (Fig. 2b). Some tune-ups are made to improve the user experience. AnnoLnc first searches the global cache for possible hits based on both sequences and aligned splicing structures. Cached results are returned directly if a hit is detected, and novel sequences and loci are sent to on-the-fly analysis modules (Fig. 1). When users fetch results, annotations generated by each module will be well summarized and integrated into intuitive web pages (Fig. 2c).

The default web interface is implemented based on responsive design, which enables the optimal view for both desktop PCs and mobile devices. To further help users quickly grasp the essentials from abundant annotations generated by various modules, AnnoLnc provides a concise summary text in plain English for each input lncRNA at the top of the annotation result page by abstraction-based summarization, with inline links available for checking original results when necessary (Fig. 3c). Furthermore, AnnoLnc supports exporting transcript-level annotations (including transcript structure, TF binding sites, miRNA binding sites, protein binding sites and SNPs) at the locus of an input lncRNA onto the UCSC genome browser as pre-tuned custom tracks (Fig. 3f).

In addition to the browser-based interactive analysis, AnnoLnc provides a "batch mode" that allows users to upload multiple sequences together and fetch annotations as a ZIP. AnnoLnc also offers a set of JSON-based Web Service APIs (Table 2) to help advanced users run the analysis and fetch results programmatically, enabling an easy integration of AnnoLnc into downstream analysis pipelines (see http://annolnc.cbi.pku.edu.cn/api.jsp for more detailed instruction as well as the demo code).

Case studies by AnnoLnc

The noncoding form of the steroid receptor RNA activator (*SRA*, AF092038, http://annolnc.cbi.pku.edu.cn/cases/SRA) has been reported to function as a noncoding RNA by Lanz et al. [41] and is the first lncRNA that has experimentally derived secondary structure, which was derived by Novikova et al. [42]. In the interactive secondary structure plot with vertebrate phyloP score as color overlay, it is easy to identify two conserved regions. One is a hairpin region from base 30 to 72 (Fig. 3a). With approximately 75% of bases colored red, this conserved sub-structure is clearly distinguishable from others. In fact, this region corresponds to the most conserved H2 sub-structure highlighted by Novikova et al. [42]. Site-directed mutagenesis of this region reduced the co-activation performance of *SRA* by 40% [43], suggesting the importance of lncRNA secondary structure on its function [44]. The other distinct region is a three-way junction hairpin sub-structure from base 506 to 555 with 78% colored red (Fig. 3b). This region is very similar to the conserved regions H15, H16 and H17 verified by Novikova et al. [42].

H19 (http://annolnc.cbi.pku.edu.cn/cases/H19) is the first identified imprinting lncRNA [45, 46]. Consistent with the work by Dey et al. [47], AnnoLnc shows that *H19* has the highest expression in "skeletal muscle" (Fig. 3d) and is associated with muscle-related function terms such as "muscle fiber development" (GO:0048747, Fig. 3e). Moreover, the "transcriptional regulation" module reports that *H19* is regulated by multiple known cell proliferation- and cell cycle-related TFs, including c-Myc, Max, Maz, and E2F6 in cancer cell lines (Additional file 1: Table S1a), confirming its previously reported tumorigenesis function [48]. In addition, AnnoLnc identified 18 CLIP-Seq-supported binding miRNA families (Additional file 1: Table S1b), and several miRNAs have already been verified experimentally, such as miR-138 in colorectal cancer [49] and miR-17-5p in HeLa cells and myoblasts [50].

In addition to confirming previous reports, the integrative annotations provided by AnnoLnc help users to generate new hypotheses. For example, lncRNA *CCAT2* (http://annolnc.cbi.pku.edu.cn/cases/CCAT2) has been reported to promote colorectal cancer (CRC) growth and metastasis [51] (also see Additional file 1: Table S2 for associated diseases), and risk allele G of rs6983267 within the *CCAT2* transcript is associated with up-regulated expression of this lncRNA [51]. Integrating miRNA annotation with the variant track (Fig. 3f), SNP rs6983267 is found to be just within the seed binding site of miRNA family miR-148ab-3p/152, suggesting that SNP rs6983267 might weaken the binding of miR-148ab-3p/152 and increases the transcript level of *CCAT2*.

Table 2 The introduction to Web Service APIs provided by AnnoLnc

Name	URL	Parameters	Return	HTTP Method
Upload	http://annolnc.cbi.pku.edu.cn/service/upload	*file*: a file containing lncRNA sequences (less than 500) in fasta format. *token*: a unique ID for authorized submission of batch job through the Web Service*. *email*: your email address used to apply the token.	The job ID.	POST
Info	http://annolnc.cbi.pku.edu.cn/service/info	*id*: the job ID.	The run info of this job, especially the run status for each sequence. Note that you can only get the download URL after all analyses are finished. Please check the run status first before fetching the download URL.	GET, POST
Fetch	http://annolnc.cbi.pku.edu.cn/service/fetch	*submitID*: the job ID. [*seqName*]: the name of a lncRNA which you want to download its annotation results.	The URL of the annotation results in a ZIP package.	GET, POST

* see http://annolnc.cbi.pku.edu.cn/api.jsp#apply-for-token for more details

Conclusions

To the best of our knowledge, AnnoLnc is the only online web server to systematically annotate novel human lncRNAs. The annotation generated by AnnoLnc covers a much wider range of perspectives with intuitive visualization and summarization. Several case studies have shown the power of AnnoLnc to systematically annotate lncRNAs, as well as inspire novel hypotheses for follow-up experimental studies. Employing Web Service APIs, AnnoLnc is friendly for not only interactive users, but also programmers.

Availability and requirements

Project name: AnnoLnc

Project home page: http://annolnc.cbi.pku.edu.cn

Operating system: AnnoLnc can be accessed from any platform by using modern Web browsers (recommended but not limited to the latest version of Safari, Chrome and Firefox).

Programming languages: Java, Python, R, Shell and JavaScript.

Any restrictions to use by non-academics: For non-academic use, please contact annolnc@mail.cbi.pku.edu.cn.

Additional files

Additional file 1: Table S1a. The annotation result of "trancriptional regulation" of lncRNA *H19*. **Table S1b.** The annotation result of "miRNA interaction" of lncRNA *H19*. **Table S2.** The annotation result of "trait association" of lncRNA *CCAT2*. **Table S3.** RNA-Seq datasets of normal samples. **Table S4.** RNA-Seq datasets of cancer samples. **Table S5.** ChIP-Seq datasets. **Table S6.** miRNA families to run targetScan prediction**. Table S7.** CLIP Seq datasets of AGO. **Table S8.** CLIP-Seq datasets of RNA binding protein. (XLS 190 kb)

Additional file 2: Figure S1. The empirical NULL distribution of interaction scores generated by random shuffling (calculated by lncPro). The red line is for the cutoff *p* value (0.01). (DOCX 63 kb)

Abbreviations

DAF: Derived allele frequency; GM: Gene model; GO: Gene ontology; LD: Linkage disequilibrium; LncRNA: Long noncoding RNA; MSB: Minimum spanning bundle; TES: Transcriptional end site; TF: Transcriptional factor; TSS: Transcriptional start site

Acknowledgements

Part of the analysis was performed on the Computing Platform of the Center for Life Sciences of Peking University.

Funding

This work was supported by funds from the China 863 Program (2015AA020108), the Seeding Grant for Medicine and Life Sciences of Peking University (2014-MB-13), and the State Key Laboratory of Protein and Plant Gene Research. The research of G.G. was supported in part by the National Program for Support of Top-notch Young Professionals.

Availability of data and material

Not applicable.

Authors' contributions

GG conceived and supervised this work. MH and XT designed the architecture and wrote the core code of most analysis modules. FT helped the RNA-Seq and CLIP-Seq analysis, as well as all the lncPro related analysis. FYS helped arrange the meta data of CLIP-Seq samples. FLL tried the pilot build of the website. MH designed the AnnoLnc website and wrote all the code of it. GG and MH wrote the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

Author details

[1]State Key Laboratory of Protein and Plant Gene Research, College of Life Sciences, Center for Bioinformatics, Peking University, Beijing 100871, P.R. China. [2]Peking-Tsinghua Center for Life Sciences, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing 100871, P.R. China. [3]Present address: Department of Hematology, St. Jude Children's Research Hospital, Memphis, TN, USA.

Received: 20 March 2016 Accepted: 10 November 2016
Published online: 16 November 2016

References

1. Rinn JL, Chang HY. Genome regulation by long noncoding RNAs. Annu Rev Biochem. 2012;81:145–66.
2. Lee JT, Bartolomei MS. X-inactivation, imprinting, and long noncoding RNAs in health and disease. Cell. 2013;152(6):1308–23.
3. Kitagawa M, Kitagawa K, Kotake Y, Niida H, Ohhata T. Cell cycle regulation by long non-coding RNAs. Cell Mol Life Sci. 2013;70(24):4785–94.
4. Park JY, Lee JE, Park JB, Yoo H, Lee SH, Kim JH. Roles of Long Non-Coding RNAs on Tumorigenesis and Glioma Development. Brain tumor research and treatment. 2014;2(1):1–6.
5. Ng SY, Johnson R, Stanton LW. Human long non-coding RNAs promote pluripotency and neuronal differentiation by association with chromatin modifiers and transcription factors. Embo J. 2012;31(3):522–33.
6. Jalali S, Kapoor S, Sivadas A, Bhartiya D, Scaria V. Computational approaches towards understanding human long non-coding RNA biology. Bioinformatics. 2015;31(14):2241–51.
7. Trimarchi T, Bilal E, Ntziachristos P, Fabbri G, Dalla-Favera R, Tsirigos A, Aifantis I. Genome-wide mapping and characterization of Notch-regulated long noncoding RNAs in acute leukemia. Cell. 2014;158(3):593–606.
8. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, et al. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. 2012;22(9):1760–74.
9. Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG, et al. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res. 2012;22(9):1775–89.
10. Iyer MK, Niknafs YS, Malik R, Singhal U, Sahu A, Hosono Y, Barrette TR, Prensner JR, Evans JR, Zhao S, et al. The landscape of long noncoding RNAs in the human transcriptome. Nature genetics. 2015;47(3):199-208.
11. Quek XC, Thomson DW, Maag JL, Bartonicek N, Signal B, Clark MB, Gloss BS, Dinger ME. lncRNAdb v2.0: expanding the reference database for functional long noncoding RNAs. Nucleic Acids Res. 2015;43(Database issue):D168–173.
12. Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, Huarte M, Zuk O, Carey BW, Cassady JP, et al. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature. 2009; 458(7235):223–7.
13. Liao Q, Liu C, Yuan X, Kang S, Miao R, Xiao H, Zhao G, Luo H, Bu D, Zhao H, et al. Large-scale prediction of long non-coding RNA functions in a coding-non-coding gene co-expression network. Nucleic Acids Res. 2011;39(9):3864–78.
14. Liao Q, Xiao H, Bu D, Xie C, Miao R, Luo H, Zhao G, Yu K, Zhao H, Skogerbo G, et al. ncFANs: a web server for functional annotation of long non-coding RNAs. Nucleic Acids Res. 2011;39(Web Server issue):W118–124.
15. Guo X, Gao L, Liao Q, Xiao H, Ma X, Yang X, Luo H, Zhao G, Bu D, Jiao F, et al. Long non-coding RNAs function annotation: a global prediction method based on bi-colored networks. Nucleic Acids Res. 2013;2:e35.
16. Agostini F, Zanzoni A, Klus P, Marchese D, Cirillo D, Tartaglia GG. catRAPID omics: a web server for large-scale prediction of protein-RNA interactions. Bioinformatics. 2013;29(22):2928–30.
17. Lu Q, Ren S, Lu M, Zhang Y, Zhu D, Zhang X, Li T. Computational prediction of associations between long non-coding RNAs and proteins. BMC Genomics. 2013;14:651.
18. Li J, Ma W, Zeng P, Wang J, Geng B, Yang J, Cui Q: LncTar: a tool for predicting the RNA targets of long noncoding RNAs. Briefings in bioinformatics. 2014;16(5):806-812.
19. Suresh V, Liu L, Adjeroh D, Zhou X: RPI-Pred: predicting ncRNA-protein interaction using sequence and structural information. Nucleic acids research. 2015;gkv020.
20. He S, Zhang H, Liu H, Zhu H. LongTarget: a tool to predict lncRNA DNA-binding motifs and binding sites via Hoogsteen base-pairing analysis. Bioinformatics. 2015;31(2):178–86.
21. Kent WJ. BLAT-the BLAST-like alignment tool. Genome Res. 2002;12(4): 656–64.
22. Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ. The UCSC Table Browser data retrieval tool. Nucleic Acids Res. 2004; 32(Database issue):D493–496.
23. Lorenz R, Bernhart SH, Honer Zu Siederdissen C, Tafer H, Flamm C, Stadler PF, Hofacker IL. ViennaRNA Package 2.0. Algorithms for molecular biology : AMB. 2011;6:26.
24. Smith MA, Gesell T, Stadler PF, Mattick JS. Widespread purifying selection on RNA structure in mammals. Nucleic Acids Res. 2013;41(17):8220–36.
25. Hou M, Tian F, Jiang S, Kong L, Yang D, Gao G: LocExpress: a web server for efficiently estimating expression of novel transcripts. *BMC genomics* 2016, Supplement for InCoB 2016, in press.
26. Tang X, Hou M, Ding Y, Li Z, Ren L, Gao G. Systematically profiling and annotating long intergenic non-coding RNAs in human embryonic stem cell. BMC Genomics. 2013;14 Suppl 5:S3.
27. Sikora-Wohlfeld W, Ackermann M, Christodoulou EG, Singaravelu K, Beyer A. Assessing computational methods for transcription factor target gene identification based on ChIP-seq data. PLoS Comput Biol. 2013;9(11): e1003342.
28. Yoon JH, Abdelmohsen K, Gorospe M. Functional interactions among microRNAs and long noncoding RNAs. Semin Cell Dev Biol. 2014;34:9–14.
29. Friedman RC, Farh KK, Burge CB, Bartel DP. Most mammalian mRNAs are conserved targets of microRNAs. Genome Res. 2009;19(1):92–105.
30. Jeggari A, Marks DS, Larsson E. miRcode: a map of putative microRNA target sites in the long non-coding transcriptome. Bioinformatics. 2012;28(15):2062–3.
31. Wang KC, Chang HY. Molecular mechanisms of long noncoding RNAs. Mol Cell. 2011;43(6):904–14.
32. McHugh CA, Russell P, Guttman M. Methods for comprehensive experimental identification of RNA-protein interactions. Genome Biol. 2014;15(1):203.
33. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754–60.
34. Chen B, Yun J, Kim MS, Mendell JT, Xie Y. PIPE-CLIP: a comprehensive online tool for CLIP-seq data analysis. Genome Biol. 2014;15(1):R18.
35. Hoell JI, Larsson E, Runge S, Nusbaum JD, Duggimpudi S, Farazi TA, Hafner M, Borkhardt A, Sander C, Tuschl T. RNA targets of wild-type and mutant FET family proteins. Nat Struct Mol Biol. 2011;18(12):1428–31.
36. Macias S, Plass M, Stajuda A, Michlewski G, Eyras E, Caceres JF. DGCR8 HITS-CLIP reveals novel functions for the Microprocessor. Nat Struct Mol Biol. 2012;19(8):760–6.
37. Cheetham SW, Gruhl F, Mattick JS, Dinger ME. Long noncoding RNAs and the genetics of cancer. Br J Cancer. 2013;108(12):2419–25.
38. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorff L, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 2014;42(Database issue):D1001–1006.
39. Genomes Project C, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA. A map of human genome variation from population-scale sequencing. Nature. 2010;7319:1061–73.
40. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 2005;15(8):1034–50.
41. Lanz RB, McKenna NJ, Onate SA, Albrecht U, Wong J, Tsai SY, Tsai MJ, O'Malley BW. A steroid receptor coactivator, SRA, functions as an RNA and is present in an SRC-1 complex. Cell. 1999;97(1):17–27.
42. Novikova IV, Hennelly SP, Sanbonmatsu KY. Structural architecture of the human long non-coding RNA, steroid receptor RNA activator. Nucleic Acids Res. 2012;40(11):5034–51.
43. Lanz RB, Razani B, Goldberg AD, O'Malley BW. Distinct RNA motifs are important for coactivation of steroid hormone receptors by steroid receptor RNA activator (SRA). Proc Natl Acad Sci U S A. 2002;99(25):16081–6.
44. Mercer TR, Mattick JS. Structure and function of long noncoding RNAs in epigenetic regulation. Nat Struct Mol Biol. 2013;20(3):300–7.
45. Brannan CI, Dees EC, Ingram RS, Tilghman SM. The product of the H19 gene may function as an RNA. Mol Cell Biol. 1990;10(1):28–36.
46. Bartolomei MS, Zemel S, Tilghman SM. Parental imprinting of the mouse H19 gene. Nature. 1991;351(6322):153–5.
47. Dey BK, Pfeifer K, Dutta A. The H19 long noncoding RNA gives rise to microRNAs miR-675-3p and miR-675-5p to promote skeletal muscle differentiation and regeneration. Genes Dev. 2014;28(5):491–501.
48. Guo G, Kang Q, Chen Q, Chen Z, Wang J, Tan L, Chen JL. High expression of long non-coding RNA H19 is required for efficient tumorigenesis induced by Bcr-Abl oncogene. FEBS Lett. 2014;588(9):1780–6.
49. Liang WC, Fu WM, Wong CW, Wang Y, Wang WM, Hu GX, Zhang L, Xiao LJ, Wan DC, Zhang JF, et al. The lncRNA H19 promotes epithelial to mesenchymal transition by functioning as miRNA sponges in colorectal cancer. Oncotarget. 2015;6(26):22513–25.
50. Imig J, Brunschweiger A, Brummer A, Guennewig B, Mittal N, Kishore S, Tsikrika P, Gerber AP, Zavolan M, Hall J. miR-CLIP capture of a miRNA

targetome uncovers a lincRNA H19-miR-106a interaction. Nat Chem Biol. 2015;11(2):107–14.
51. Ling H, Spizzo R, Atlasi Y, Nicoloso M, Shimizu M, Redis RS, Nishida N, Gafa R, Song J, Guo Z, et al. CCAT2, a novel noncoding RNA mapping to 8q24, underlies metastatic progression and chromosomal instability in colon cancer. Genome Res. 2013;23(9):1446–61.