Chapter 19

NONCODEv4: Annotation of Noncoding RNAs with Emphasis on Long Noncoding RNAs

Yi Zhao, Jiao Yuan, and Runsheng Chen

Abstract

The rapid development of high-throughput sequencing technologies and bioinformatics algorithms now enables detection and profiling of a large number of noncoding transcripts. Long noncoding RNAs (lncRNAs), which are longer than 200 nucleotides, are accumulating with important roles involved in biological processes and tissue physiology. In this chapter, we describe the use of NONCODEv4, a database that provide a comprehensive catalog of noncoding RNAs with particularly detailed annotations for lncRNAs. NONCODEv4 stores more than half million transcripts, of which more than 200,000 are lncRNAs. NONCODEv4 raises the concept of lncRNA genes and explores their expression and functions based on public transcriptome data. NONCODEv4 also integrated a series of online tools and have a web interface easy to use. NONCODEv4 is available at http://www.noncode.org/ http://www.bioinfo.org/noncode.

Key words Sequencing, lncRNA, lncRNA gene, Expression, Function

1 Introduction

Noncoding RNAs (ncRNAs) participate in many biological processes such as translation [1], RNA splicing [2], and DNA replication [3]. It has been recognized that the number of ncRNAs is much larger than expected. Although long noncoding RNAs (lncRNAs) which refer to ncRNAs longer than 200 nucleotides [4] form a group, they have diverse functions and mechanisms [4]. The continuously developed high-throughput sequencing technology has resulted in an explosion of transcriptome data [5, 6]. Through exploring these data resource, NONCODEv4 identified a large number of lncRNAs from both qualitative and quantitative perspective [7].

In the situation that more and more interest has been focused on lncRNAs, NONCODEv4 provides a platform that would felicitate and benefit researches into lncRNAs for both traditional biologists and bioinformaticians. In order to provide a reliable list of

Yi Feng and Lin Zhang (eds.), *Long Non-Coding RNAs: Methods and Protocols*, Methods in Molecular Biology, vol. 1402,
DOI 10.1007/978-1-4939-3378-5_19,

ncRNAs with detailed annotation, NONCODEv4 integrated information following a pipeline described below:

1. Collecting ncRNA sequences from literature, the latest version of specialized databases [8, 9] and the third version of NONCODE [10] following redundancy elimination, exclusion of protein coding transcripts, and mapping to various genome. Currently, NONCODEv4 contains more than half million ncRNA sequences from more than 1000 organisms. Of them, the majority comes from model organism: human, mouse, nematode, and fly. The genomic location for some short sequences that could not be mapped to unique position is not shown.
2. Defining lncRNA genes and classifying lncRNA genes into four categories according to their position relationship with protein coding genes. A lncRNA gene is region of genome which encodes overlapping lncRNAs. Currently, NONCODEv4 defines lncRNA genes for only human and mouse. 56,018 and 46,475 lncRNA genes of human and mouse are further classified into four categories: intergenic, antisense, sense non-exonic, and sense exonic [11].
3. Constructing expression pattern of lncRNA transcripts and lncRNA genes respectively based on public RNA-Seq data from different tissues of human and mouse. Expression profile is represented as a bar graph.
4. Predicting functions of lncRNA genes based on expression correlation between lncRNA genes and protein coding genes inferred from public RNA-Seq data.

All data curated in NONCODEv4 are available for browse and download. There is also an online pipeline named "iLncRNA" incorporated in order to provide an access to deal with their own raw RNA-Seq data for users to identify novel lncRNAs. NONCODEv4 also encourages users to submit their own discoveries of novel lncRNAs to facilitate future updates. Other tools have been incorporated to make NONCODEv4 more friendly to use, such as Genome Browser by which users could check neighboring genes and isoforms of the lncRNA they are interested in and an ID conversion tool which quickly convert accessions of NONCODE to those of other databases.

2 Materials

NONCODEv4 is available online via the URL http://www.noncode.org/. A Web browser is needed by a workstation of UNIX, Windows, or Macintosh with an Internet connection. In addition, decompression software is required since files provided for download is compressed.

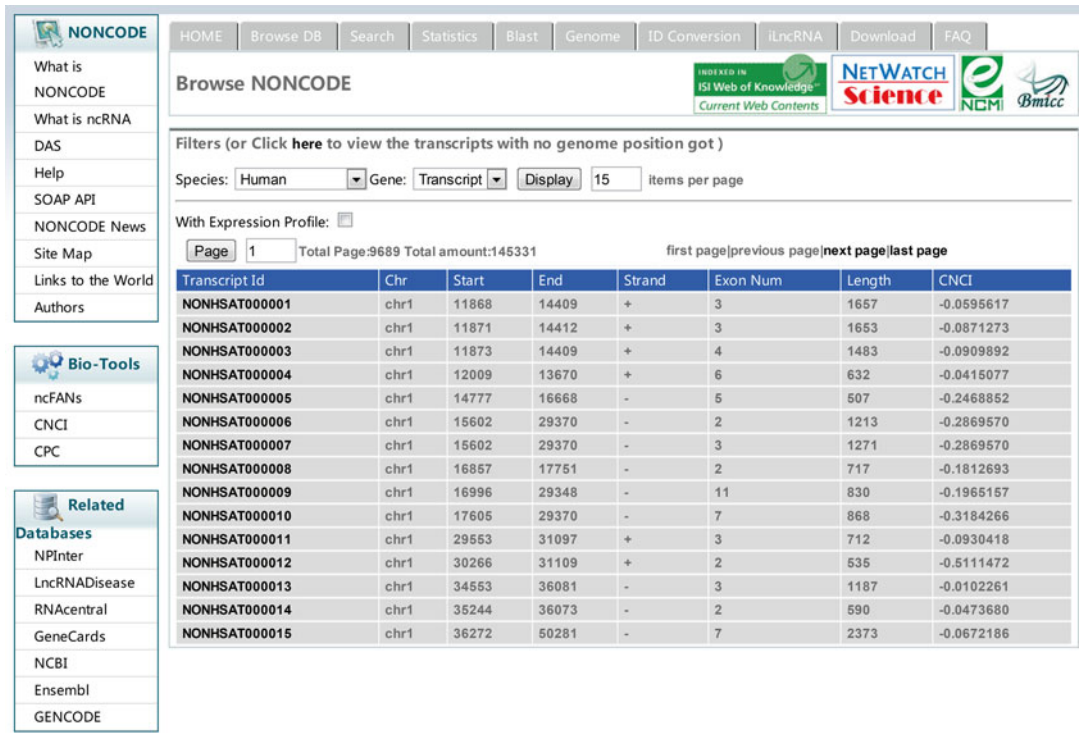
3 Methods

The methods presented in this chapter describes how to use the NONCODEv4 Web interface to obtain information for a specific lncRNA (Subheading 3.1), how to navigate the record of a lncRNA of interest (Subheading 3.2), and how to submit data to NONCODEv4 server (Subheading 3.3).

3.1 Browsing Information for a Specific lncRNA

Access to the data content of NONCODEv4 may be performed by browsing the list of all lncRNAs which provides a quick overview about the dataset. The following steps describe how to browse the list of all lncRNAs and the detailed information:

1. Click "Browse DB" on the home page to open the browse page. A user who is interested in specific organism might first select it from the list of species (Fig. 1). By default, lncRNA transcripts are listed. It could be switched to the list of lncRNA genes by selection through the check box followed by clicking "Display". An accession designated by NONCODEv4, genomic coordinates, exon number, length and CNCI score suggesting coding potential are all listed in the shown table.



Transcript Id	Chr	Start	End	Strand	Exon Num	Length	CNCI
NONHSAT000001	chr1	11868	14409	+	3	1657	-0.0595617
NONHSAT000002	chr1	11871	14412	+	3	1653	-0.0871273
NONHSAT000003	chr1	11873	14409	+	4	1483	-0.0909892
NONHSAT000004	chr1	12009	13670	+	6	632	-0.0415077
NONHSAT000005	chr1	14777	16668	-	5	507	-0.2468852
NONHSAT000006	chr1	15602	29370	-	2	1213	-0.2869570
NONHSAT000007	chr1	15602	29370	-	3	1271	-0.2869570
NONHSAT000008	chr1	16857	17751	-	2	717	-0.1812693
NONHSAT000009	chr1	16996	29348	-	11	830	-0.1965157
NONHSAT000010	chr1	17605	29370	-	7	868	-0.3184266
NONHSAT000011	chr1	29553	31097	+	3	712	-0.0930418
NONHSAT000012	chr1	30266	31109	+	2	535	-0.5111472
NONHSAT000013	chr1	34553	36081	-	3	1187	-0.0102261
NONHSAT000014	chr1	35244	36073	-	2	590	-0.0473680
NONHSAT000015	chr1	36272	50281	-	7	2373	-0.0672186

Fig. 1 A screenshot of browse page. The species of interest might be selected by *check box*. It is optional to browse lncRNA transcript list or lncRNA gene list. ncRNAs have no exact genomic coordinates might be browsed by clicking "here" *above* the *check box*

2. Click on the accession listed on the first column of the table to launch a detailed page providing further information on that lncRNA transcript or gene (Figs. 2 and 3).
3. The detailed information for a specific lncRNA transcript includes five sections (Fig. 2):
   (a) In the section for general information, clicking the "NONCODE Gene ID" will link to the webpage of the lncRNA gene encoding the lncRNA transcript described at present page.
   (b) In the second section, the full length of sequence is provided in a fasta format [12].
   (c) In the third section, the expression pattern of the lncRNA transcript across different tissues is given by both numerical value and bar graph.
   (d) In the section for isoforms, the accessions of other lncRNA transcripts encoded by the same lncRNA gene are listed.
   (e) In the section for data resources, accessions of Ensembl, RefSeq, or NONCODE v3 might be listed to indicate the origin of the lncRNA transcript.
4. The detailed information contained in the page of lncRNA gene (Fig. 3) includes four sections. The first three sections are similar to those of lncRNA transcript. Notice that the category of the lncRNA gene is classified into (intergenic, antisense, sense non-exonic, or sense exonic) is provided in the section for general information. The last section listed predicted function of the lncRNA gene by the software ncFANs [13].

3.2 Navigating to the Record for a Specific lncRNA

From the browse page, it might not be easy to quick navigate to the right webpage for the lncRNA of interest. Navigating to the exact webpage for a specific lncRNA might be performed via three options summarized below.

1. Click "Search" on the home page to open the search page, which enables users to search ncRNAs by keywords or accessions. Single word or multiple words separated by whitespace might be entered into the blank box, "HOTAIR" for example (Fig. 4). Multiple types of terms, including accessions from NONCODEv4, NONCODEv3, RefSeq, and Ensembl, lncRNA name, and other keywords, are supported. Click "Search" to view the search result. Selecting and clicking an accessions from the list in the result page would direct a new page for browsing annotations for the lncRNA transcript.
2. For a ncRNA of which none information except sequence is known, search based on keywords might not work. Sequence alignment [14] is a solution for this situation. Click "Blast" on the home page to open the blast page, which enables users to

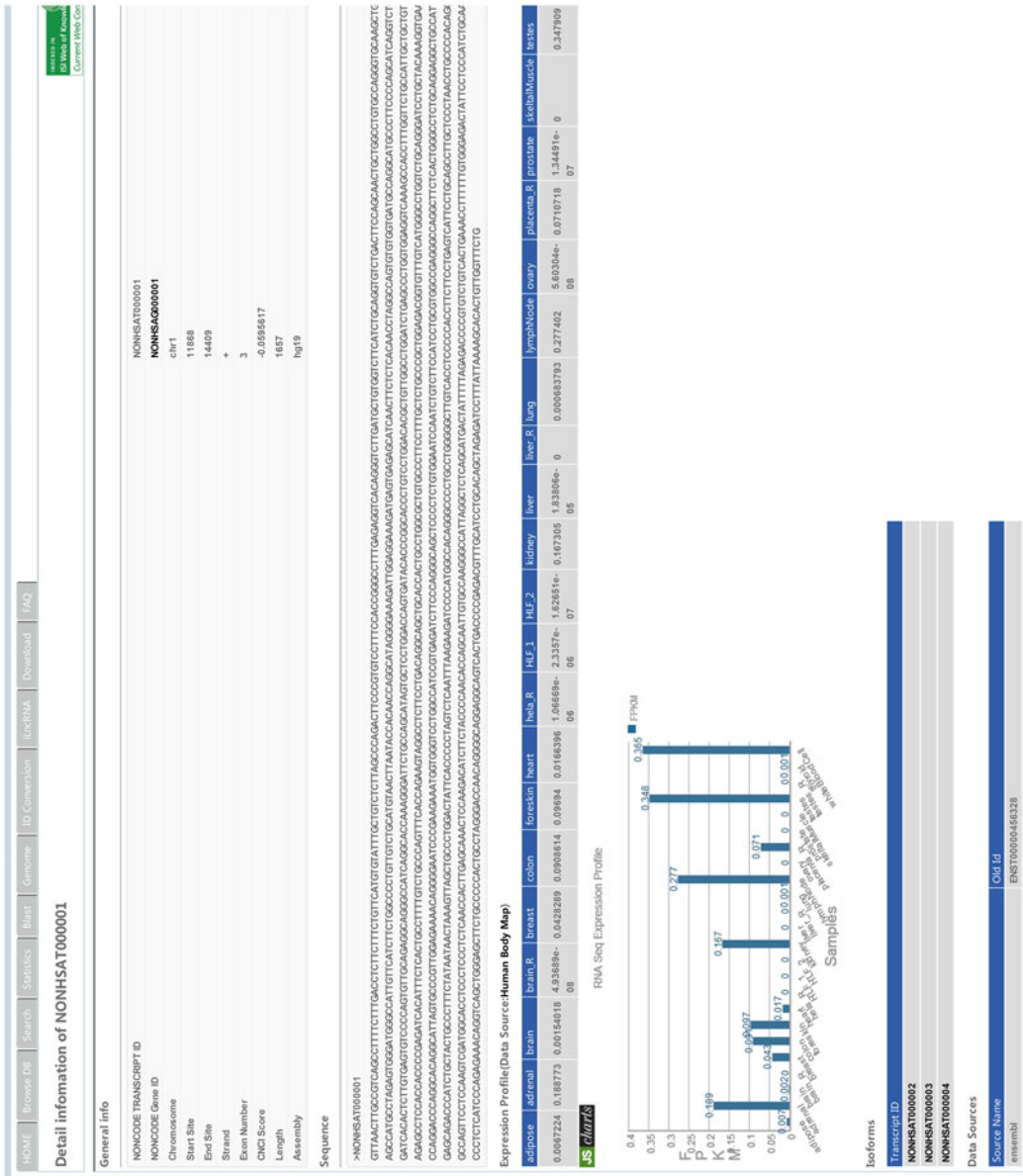


Fig. 2 A sample screenshot of detail information of a lncRNA transcript. Annotation includes five sections: general information, sequence, expression profile, isoforms, and data sources. A link to the webpage for annotation of the lncRNA gene encoding the lncRNA transcript describe at present webpage is included in the section of general information

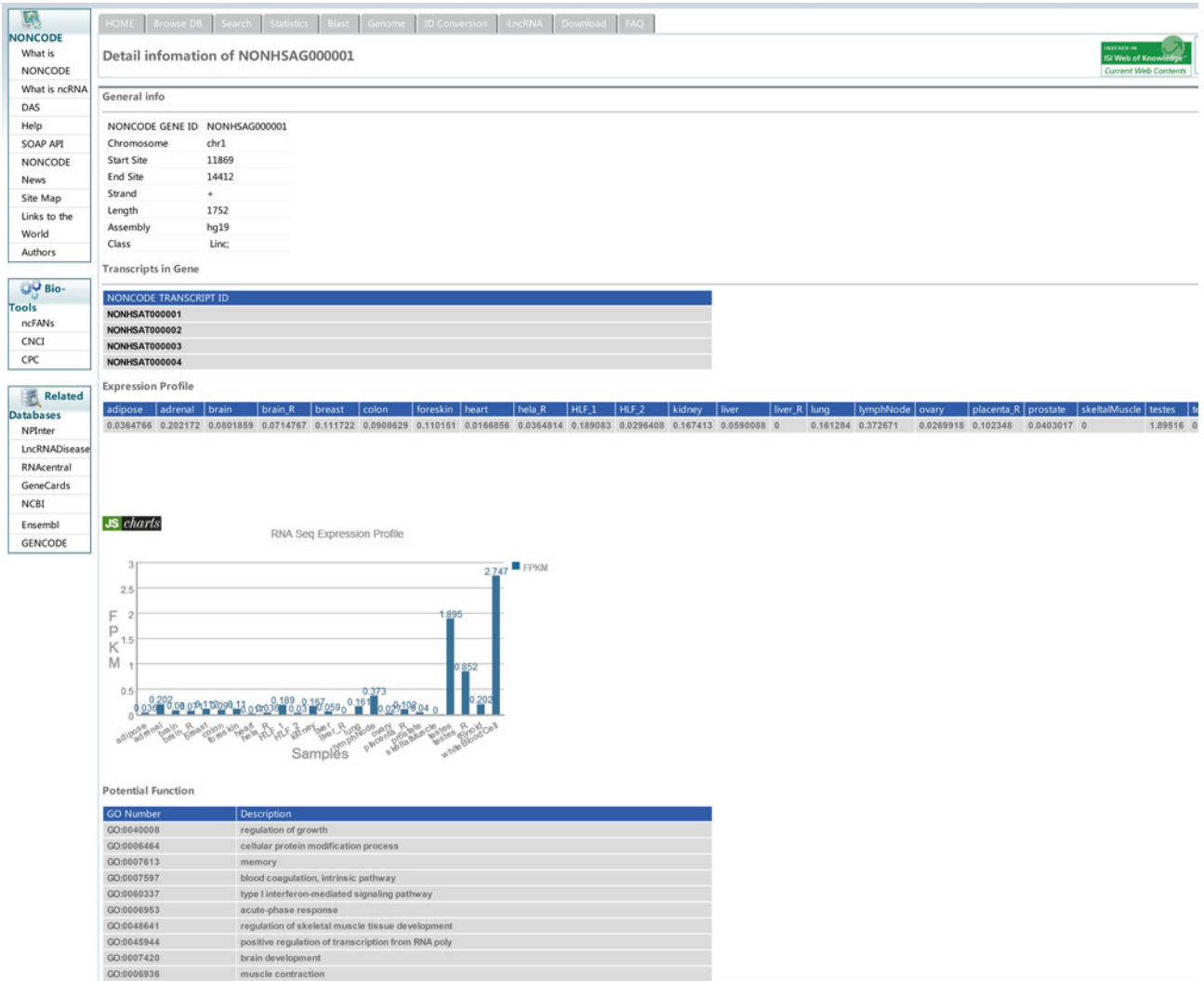


Fig. 3 A sample screenshot of detail information of a lncRNA gene. Annotation includes four sections: general information, transcripts, expression profile, and potential function. The category into which the lncRNA gene is classified is shown in the section of general information

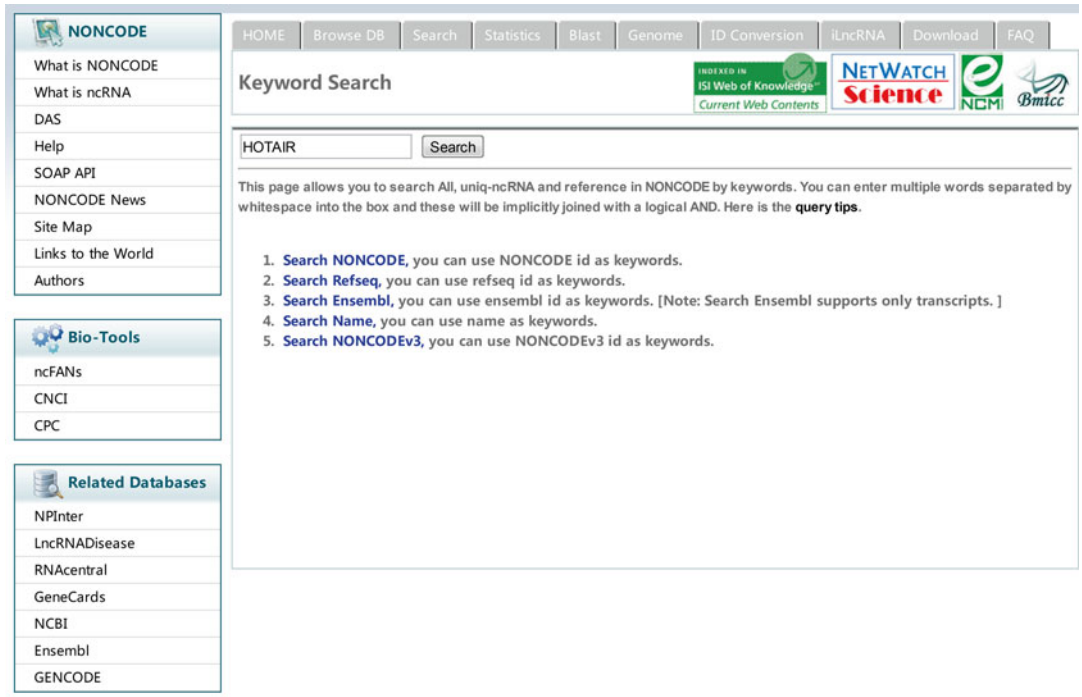


Fig. 4 A sample screenshot of search page. It is recommended for users to follow the query tips listed to determine their keywords

search ncRNAs by sequences (Fig. 5). Enter the sequences in fasta format into the blank box or upload the file containing sequences in fasta format from local disk. Click "Search" button to start the search. Parameters for BLAST might be adjusted through check boxes below. In the result page of BLAST output, the top NONCODEv4 accessions with highest score might match the input sequence. If there are no match records with score high enough according to the length of input sequence, then the sequence might not be curated in NONCODEv4.

3. Another option is navigating to the record for a specific lncRNA based on its genomic coordinates. NONCODEv4 provides Genome Browser, a visualization tool, to quickly navigate to a specific genome region. Click "Genome" on the home page to open the Genome Browser page. Select the genome of interest as described above, and then type the genomic coordinate of the lncRNA of interest into the "genomic position" text box. Click "submit" button. In the display page (Fig. 6), all lncRNA transcripts and genes overlapping the submitted genomic coordinates are shown, as well as annotation from other tracks within the region. Clicking a track item within the browser launches a detailed page providing further information on that item. The width of the displayed coordinate range could be

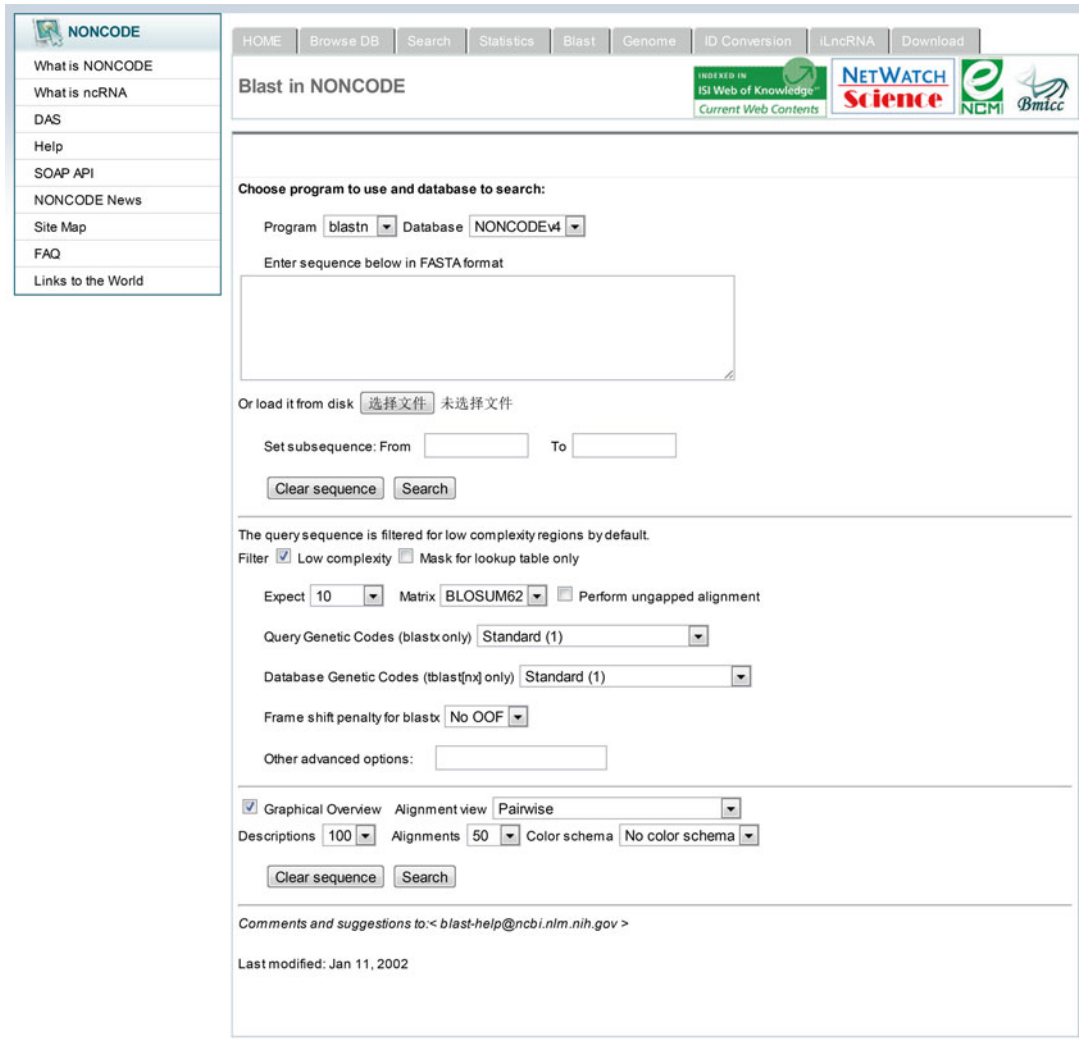


Fig. 5 A screenshot of BLAST page. Query sequence could either be entered into the *blank box* in format or uploaded in a file from local disk. Parameters for blast could be adjusted

adjusted by clicking the "zoom out" or "zoom in" button. The genomic coordinates could also be shifted to the left or right by clicking "move left" or "move right" button. Custom tracks might be uploaded to compare it with tracks on NONCODEv4 server. Optionally, it could be custom determined which track should be shown of hidden through the track option track.

3.3 Submitting Raw Data to NONCODEv4 Server

There is continuing emergence of high-throughput sequencing data. NONCODEv4 provides an online pipeline named "iLncRNA" to help users with identification of novel lncRNAs based on their own deep-sequencing data.

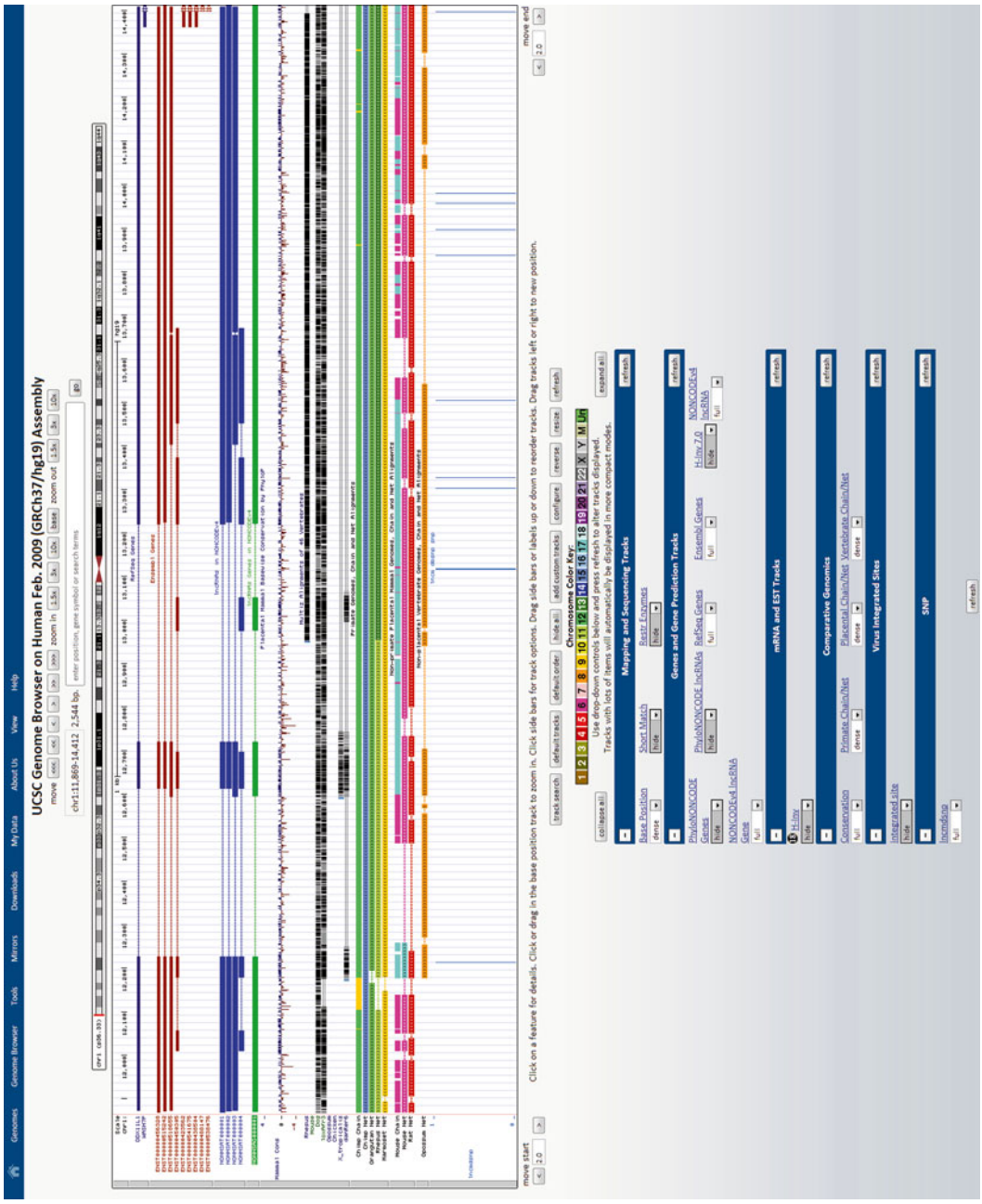


Fig. 6 The Genome Browser displaying the chr1:11,869-14,412 region in the human genome (UCSC hg19). The Genome Browser provides an integrated view by integrating annotations from NONCODEv4, RefSeq, Ensembl, and UCSC. The display region might be adjusted by the navigation buttons at the *top* of the image, either shifted or scaled

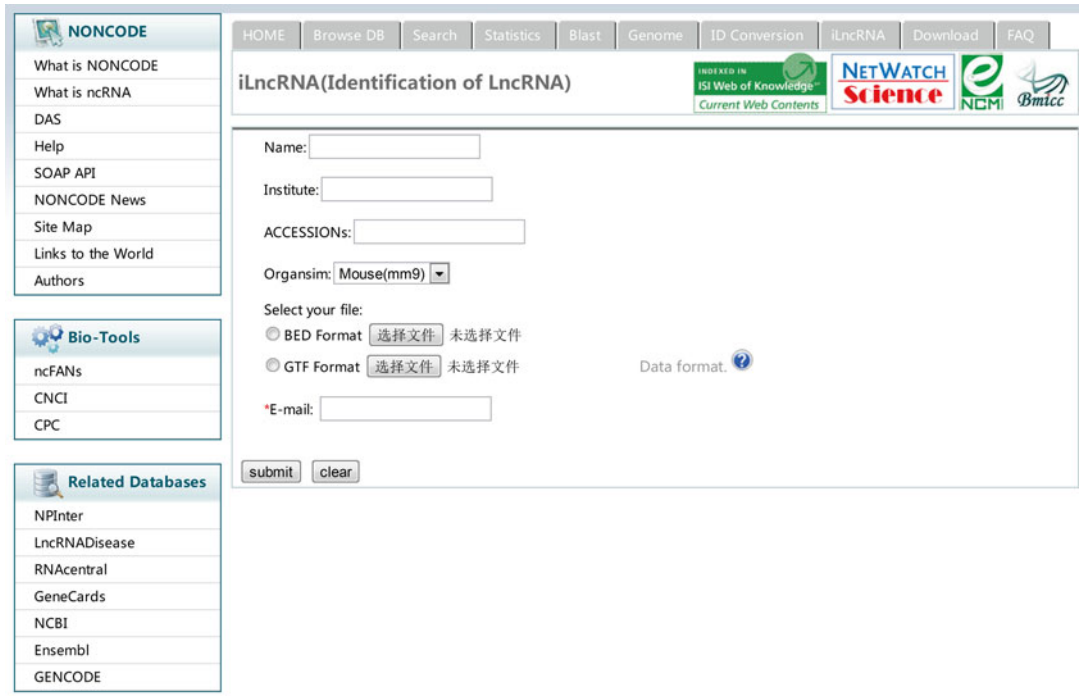


Fig. 7 A screenshot of iLncRNA webpage. The species of interest should be selected. User supplied files should be either bed format or gtf format. An e-mail address is needed for receiving message when analysis is done

1. Click "iLncRNA" on the home page to open the webpage for iLncRNA pipeline (Fig. 7).
2. Enter information including name, institute and accessions of the user into the corresponding boxes. Select organism (human or mouse) in the check box and upload the file in bed format or gtf format generated by assembly software or other tools. User's e-mail is necessary for receiving feedback when the analysis is done.
3. After reading user supplied files, the iLncRNA server would first extract sequences with length more than 200 nucleotides. Then transcripts completely matching known protein coding transcripts from RefSeq or pseudogenes from Ensembl would be discarded. CNCI [15] would be used to judge whether the sequences which have passed the previous filtration have coding potential. Sequences with no coding potential are kept for further annotation by Cuffcompare [16]. Sequences which do not completely match lncRNA transcripts curated in NONCODEv4 would be classified as novel lncRNAs. The result would be sent to the email address which user filled in the check box.

4 Notes

1. NONCODEv4 is under periodically update. There would be statement of the latest modification highlighted in red on the home page. In turn, it is encouraged for users to report their problems regarding to usage of NONCODEv4 or interpreting annotations made by NONCODEv4 to the group working for NONCODEv4 by e-mail.
2. It is supported to search accessions from Ensembl or RefSeq on the search webpage (*see* Subheading 3.1). Besides, NONCODEv4 provided an online tool named "ID Conversion". It enables quick conversion between accessions of NONCODEv4 and other resources including Ensembl, RefSeq, and NONCODEv3. Batch conversion is supported.
3. All data of NONCODEv4 are stored in relational tables of MySQL database. Accessions of ncRNAs in NONCODEv4 are designated systematically. Take human as an example, lncRNA transcripts of human are designated with accessions from NONHSAT000001 to NONHSAT148172. The prefix of "NON" stands for "noncoding". The following "HSA" stands for "Homo sapiens". Similarly, it should be replaced by other letters for other organisms, such as "MMU" for "Mus musculus". The next letter "T" stands for "transcript", which should be replaced by "G" in accessions of lncRNA genes. By default, the numeric string with which NONCODEv4 accessions end is according to the order of transcripts or genes sorted by chromosome. Additionally, it is a little different for ncRNAs with unclear genomic coordinates, with "NOBED" in the middle. For example, NONHSANOBEDT000001 denotes a ncRNA sequence form human which could not be uniquely mapped to human genome.

Acknowledgment

This work was supported by National High-tech Research and Development Projects 863 [2012AA020402, 2012AA022501], National Key Basic Research and Development Program 973 [2009CB825401], Training Program of the Major Research plan of the National Natural Science Foundation of China [91229120], and National Natural Science Foundation of China [31371320]. Funding for open access charge: Training Program of the Major Research plan of the National Natural Science Foundation of China [91229120].

References

1. Himeno H, Kurita D, Muto A (2014) tmRNA-mediated trans-translation as the major ribosome rescue system in a bacterial cell. Front Genet 5:66
2. Jones TA, Otto W, Marz M, Eddy SR, Stadler PF (2009) A survey of nematode SmY RNAs. RNA Biol 6:5–8
3. Christov CP, Gardiner TJ, Szuts D, Krude T (2006) Functional requirement of noncoding Y RNAs for human chromosomal DNA replication. Mol Cell Biol 26:6993–7004
4. Mercer TR, Dinger ME, Mattick JS (2009) Long non-coding RNAs: insights into functions. Nat Rev Genet 10:155–159
5. Croucher NJ, Thomson NR (2010) Studying bacterial transcriptomes using RNA-seq. Curr Opin Microbiol 13:619–624
6. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-seq. Nat Methods 5:621–628
7. Xie C, Yuan J, Li H, Li M, Zhao G, Bu D, Zhu W, Wu W, Chen R, Zhao Y (2014) NONCODEv4: exploring the world of long non-coding RNA genes. Nucleic Acids Res 42:D98–D103
8. Flicek P, Ahmed I, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S et al (2013) Ensembl 2013. Nucleic Acids Res 41:D48–D55
9. Pruitt KD, Tatusova T, Brown GR, Maglott DR (2012) NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. Nucleic Acids Res 40:D130–D135
10. Bu D, Yu K, Sun S, Xie C, Skogerbo G, Miao R, Xiao H, Liao Q, Luo H, Zhao G et al (2012) NONCODE v3.0: integrative annotation of long noncoding RNAs. Nucleic Acids Res 40:D210–D215
11. Picardi E, D'Erchia AM, Gallo A, Montalvo A, Pesole G (2014) Uncovering RNA editing sites in long non-coding RNAs. Front Bioeng Biotechnol 2:64
12. Lipman DJ, Pearson WR (1985) Rapid and sensitive protein similarity searches. Science 227:1435–1441
13. Liao Q, Xiao H, Bu D, Xie C, Miao R, Luo H, Zhao G, Yu K, Zhao H, Skogerbo G et al (2011) ncFANs: a web server for functional annotation of long non-coding RNAs. Nucleic Acids Res 39:W118–W124
14. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410
15. Sun L, Luo H, Bu D, Zhao G, Yu K, Zhang C, Liu Y, Chen R, Zhao Y (2013) Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. Nucleic Acids Res 41(17):e166
16. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc 7:562–578