Database update

NPInter v3.0: an upgraded database of noncoding RNA-associated interactions

Yajing Hao[1,2,3,†], Wei Wu[1,2,3,†], Hui Li[3,4], Jiao Yuan[1,2,3], Jianjun Luo[1,2], Yi Zhao[4,*] and Runsheng Chen[1,2,*]

[1]Key Laboratory of RNA Biology, [2]Beijing Key Laboratory of Noncoding RNA, Institute of Biophysics, Chinese Academy of Sciences, Beijing, 100101, China, [3]University of Chinese Academy of Sciences, Beijing, 100049, China, [4]Bioinformatics Research Group, Key Laboratory of Intelligent Information Processing, Advanced Computing Research Center, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190, China

*Corresponding author: Tel: +86-10-64888543; Fax: +86-10-64877837; Email: rschen@ibp.ac.cn, Correspondence may also be addressed to Yi Zhao. Tel: +86-10-62601010; Fax: +86-10-62601356; Email: biozy@ict.ac.cn

[†]These authors are considered as Joint first authors.

Abstract

Despite the fact that a large quantity of noncoding RNAs (ncRNAs) have been identified, their functions remain unclear. To enable researchers to have a better understanding of ncRNAs' functions, we updated the NPInter database to version 3.0, which contains experimentally verified interactions between ncRNAs (excluding tRNAs and rRNAs), especially long noncoding RNAs (lncRNAs) and other biomolecules (proteins, mRNAs, miRNAs and genomic DNAs). In NPInter v3.0, interactions pertaining to ncRNAs are not only manually curated from scientific literature but also curated from high-throughput technologies. In addition, we also curated lncRNA–miRNA interactions from *in silico* predictions supported by AGO CLIP-seq data. When compared with NPInter v2.0, the interactions are more informative (with additional information on tissues or cell lines, binding sites, conservation, co-expression values and other features) and more organized (with divisions on data sets by data sources, tissues or cell lines, experiments and other criteria). NPInter v3.0 expands the data set to 491,416 interactions in 188 tissues (or cell lines) from 68 kinds of experimental technologies. NPInter v3.0 also improves the user interface and adds new web services, including a local UCSC Genome Browser to visualize binding sites. Additionally, NPInter v3.0 defined a high-confidence set of interactions and predicted the functions of lncRNAs in human and mouse based on the interactions curated in the database. NPInter v3.0 is available at http://www.bioinfo.org/NPInter/.

Database URL: http://www.bioinfo.org/NPInter/

Introduction

Over the past decade, numerous noncoding RNAs (ncRNAs) have been identified in human (1), mouse (2) and other organisms (3–5) due to the advances in high-throughput sequencing (6). Emerging evidence has suggested that, except for the well-recognized ncRNAs such as rRNAs (7), tRNAs (8) and small nuclear RNAs (9), other regulatory ncRNAs, such as miRNAs (10), siRNAs (11), piRNAs (12), and the recently rapidly expanding class of long noncoding RNAs (lncRNAs) play key roles in a range of biological processes, including genomic imprinting, disease metastasis, cell pluripotency and differentiation, and many others (13–15).

ncRNAs are known to function by interfacing with diverse classes of biomolecules. For example, miRNAs associate with Argonaute proteins to form miRNA-induced silencing complexes to regulate the expression of mRNA targets (16). The lncRNA, Xist, physically interacts with different factors to initiate and maintain the processes of X chromosome silencing (17). Therefore, identifying a more complete spectrum of ncRNAs interacting partners will significantly deepen the understanding of how ncRNAs modulate biological processes. Towards this end, many recent molecular experimental approaches combined with high-throughput sequencing or mass spectrometry were carried out to identify these interactions, such as protein-centric approaches, crosslinking and immunoprecipitation followed by deep sequencing (CLIP-seq) (18), RNA-centric approaches, Chromatin isolation by RNA purification followed by high-throughput sequencing (ChIRP-seq) (19), and others (20–22).

With the widespread application of these new high-throughput technologies and the explosive data accumulation of interactions between RNA and other biomolecules, we initiated a project to build a data repository and platform for cataloguing their interactions (NPInter (23)), and successfully updated to version 2 (24) which expanded the data collection and introduced tools for data visualization. However, the large amount of new research, particularly studies on CLIP-seq, has largely overwhelmed the collection of ncRNAs' interactions in NPInter v2.0. Thus, NPInter have been upgraded to version 3.0 to collect substantially more interactions from the literature, high-throughput sequencing, and *in silico* predictions supported by high-throughput sequencing data. In addition, ncRNAs were given accession IDs from NONCODE (25–28), RefSeq (29), Ensembl (30), and miRBase (31) while protein-coding molecules were assigned from UniProt (32), UniGene and RefSeq. Binding site information was appended to interactions discovered by CLIP-seq with conservation scores. Gene expression correlation scores were also added to the descriptions of the interactions by co-expression analysis. Owing to the fact that the number of interactions had become quite large, NPInter v3.0 also provided a high-confidence set of interactions and reorganized interactions according to different aspects such as the source of the data, tissues or cell lines, experiments and other factors. Moreover, we predicted the functions of lncRNAs in human and mouse based on the interactions curated in the database. Furthermore, we designed a new website and integrated a Genome Browser service, which greatly improved the interface and user experience. All data are available on the download page.

An overview of NPInter v3.0 updates is shown in Figure 1.

Data collection and annotation

The workflow of updating NPInter v3.0 is depicted in Figure 2. The interactions curated in NPInter v3.0 were mainly obtained from three different processing pipelines. We then re-annotated the molecules using specific IDs, removed redundant interactions and categorized interactions based on different standards. Meanwhile, we calculated gene co-expression scores between interacting molecules, and predicted lncRNAs' functions. The detailed procedure is thoroughly explained in the following sections.

Interactions curated from CLIP-seq data sets

In order to obtain all the *bona fide* interactions from high-throughput sequencing technologies, we collected all the available processed data from the CLIPdb database (33) and the Gene Expression Omnibus (GEO) (34) using keywords: RIP, CLIP, HITS-CLIP, PAR-CLIP and CLASH. After selection, we retained 111 associated datasets, including 18 Ago CLIP-seq datasets. We then converted the genomic coordinates to specific genome versions (hg19, mm9 and ce10) using the UCSC LiftOver Tool (35). Binding site locations were derived from the same article, same condition and same protein using the IntersectBed from BedTools (36). Then, we compared the union of the binding sites stored in BED format with the NONCODE v4.0 database, which is one of the most comprehensive reference databases of ncRNAs, and assigned NONCODE IDs to binding sites within ncRNAs. In addition, we provided a binding probability score per interaction calculated through LncPro (37) with default parameters, which effectively discriminated interacting and noninteracting lncRNA-protein pairs based on amino acid and nucleotide sequences. Furthermore, to assess the evolutionary conservation of each interaction-binding site, we first downloaded pre-computed sequence conservation scores (using the PhastCons (38) algorithm) across 46 vertebrate species for human, 30 vertebrate
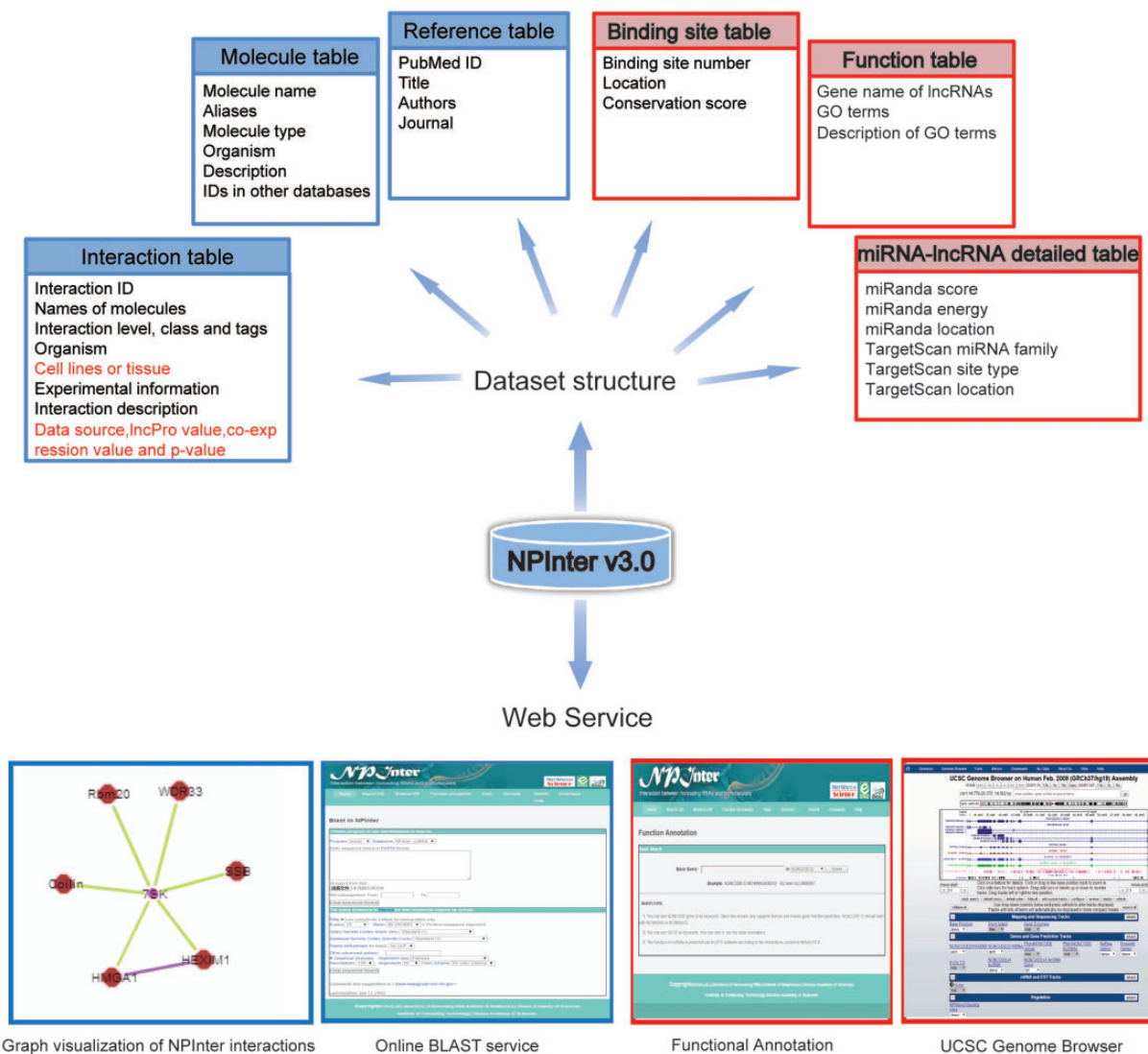
Figure 1 An overview of the NPInter v3.0 database. Improvements in this updated version are highlighted with a red border or by a red color. 191 × 183 mm (300 × 300 DPI).

species for mouse, and 7 yeast species for yeast provided from the UCSC database, and then calculated an average PhastCons score (39) per binding site. The PhastCons program used a hidden Markov model-based method that estimated the probability that each nucleotide belonged to a conserved element, based on multiple alignments of selected species. The average PhastCons scores ranged from 0 to 1 where a value >0.1 showed some conservation between the species we considered, while a value >0.5 was considered highly conserved (40).

miRNA–lncRNA interactions predicted by miRanda and TargetScan overlapped with Ago CLIP-seq datasets

ncRNAs from NONCODE v4.0 and miRNAs labeled as 'broadly conserved' or 'conserved' in TargetScan Release 7.0 (41) were acquired. AGO CLIP-seq datasets were retrieved from the GEO and CLIPdb databases. The conserved miRNA-target sites in ncRNAs were predicted using both TargetScan and miRanda (42) with the default parameters. The ncRNAs containing target sites that overlapped with any AGO CLIP cluster were considered as CLIP-supported interacting molecules. The resulting interactions were annotated as described. Moreover, the visualization of clusters in human and mouse were implemented in the Genome Browser and conservation scores were calculated as well.

Interactions curated from literature mining and NPInter v2.0

In addition to data from the former version of the NPInter database, new datasets were obtained from the literature.
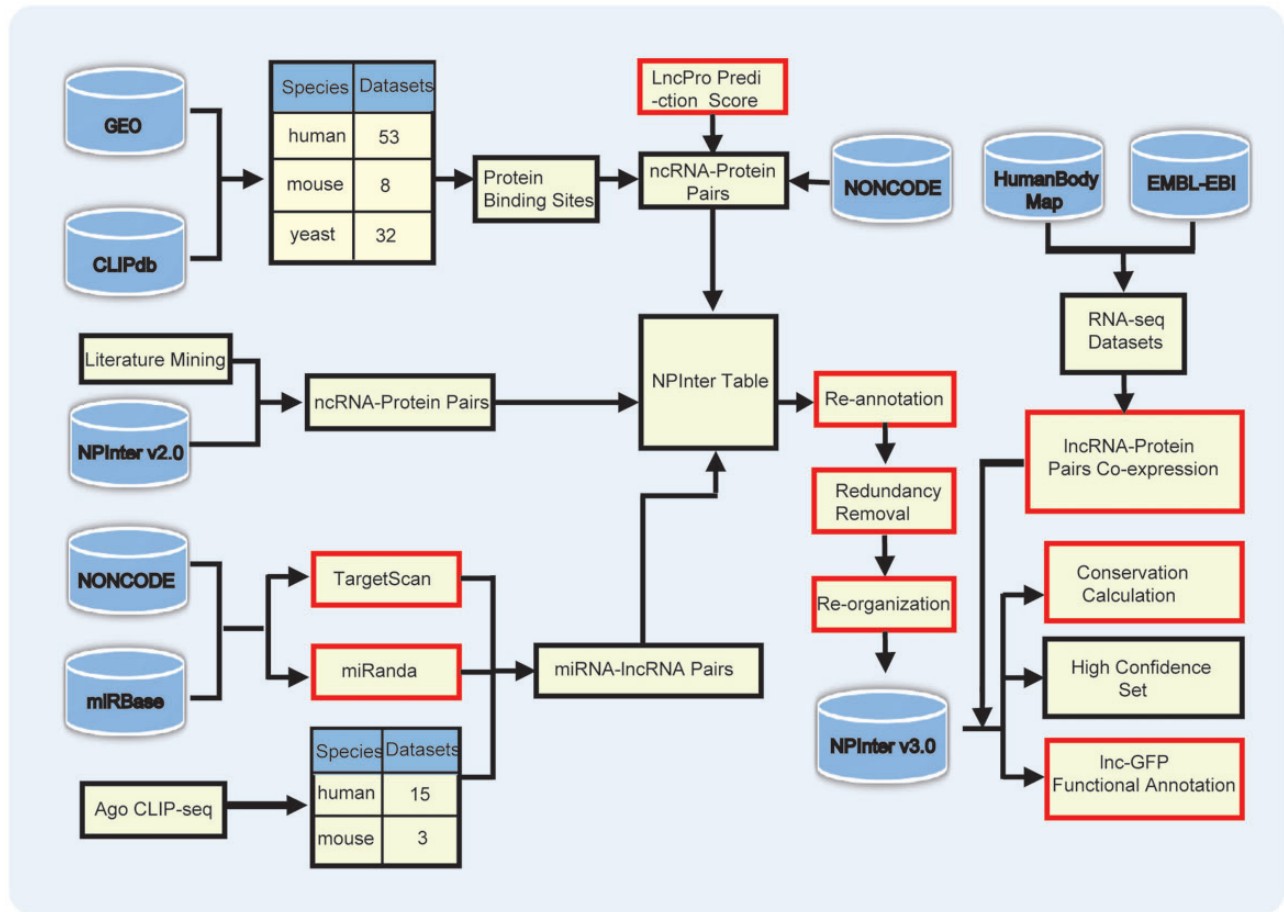
Figure 2 Workflow to collect interactions in NPInter v3.0. Red rectangles indicate that these steps need computational processing. Refer to main text for details. 169 × 123 mm (300 × 300 DPI).

First, we used a set of key words (listed in Supplementary Materials) to retrieve literature published since 12 April 2013 from PubMed and collected 991 articles. After selecting reports on new ncRNA-associated interactions, we extracted detailed information manually. Only interactions with strong support from experimental evidence were kept for further consideration.

Re-annotation, redundancy removal, re-organization

After obtaining the interactions from the three main data sources mentioned earlier, we used NONCODE IDs or miRBase IDs to annotate ncRNAs and their interacting partners. Protein-related molecules were assigned with UniProt IDs, RefSeq IDs or UniGene IDs according to the type of molecules (proteins, protein-coding RNAs or DNA, respectively). Owing to the fact that the experiments and tissues (or cell lines) used to explore the interactions were varied and came from different papers using different descriptions, we manually uniformed the descriptions of tissues (or cell lines) and experiments. Next, the whole

dataset was put through a filtration step to eliminate redundancy. Interactions with differences derived from various organisms, tissue or cell lines, or references were considered as independent records. As a final result, NPInter v3.0 provided a nonredundant and comprehensive resource of the ncRNA interactome. As different data sources had different confidences, we re-organized the interactions and defined 6 different data sources as described in Table 1. We added a data source per interaction in which different sources represented different probabilities describing a functional linkage between two molecules. In addition, we defined a high confidence set of interactions, which was obtained from literature mining and supported by low-throughput experiments, and represented the highest quality of interaction in the database. However, we did not sort data credibility according to the data sources, as the confidence of the data sources should be decided by the users, who should carefully compare them for any specific task at hand. In order to give guidance to users who want to balance different levels of accuracy, we provided three advices: First, the highest confidence data set was obtained from literature mining. Second, for the remaining five data

Table 1 The type of data sources in NPInter v3.0

Data sources	Description	Number of interactions
Literature mining	The interactions were obtained from literature mining.	8130
High-throughput data combined with LncPro prediction	The interactions were obtained from high-throughput data and were supported by LncPro prediction.	96 244
High-throughput data	The interactions were obtained from high-throughput data but were not supported by LncPro prediction.	252 317
miRanda and TargetScan with Ago CLIP data	The interactions were predicted by miRanda and TargetScan combined with Ago CLIP-seq data support.	33 366
miRanda with Ago CLIP data	The interactions were predicted by miRanda but were not predicted by TargetScan. The interactions were also supported by Ago CLIP-seq data.	39 447
TargetScan with Ago CLIP data	The interactions were predicted by TargetScan but were not predicted by miRanda. The interactions were also supported by Ago CLIP-seq data.	61 912

sources listed in Table 1, the second was more confident than the third as the interactions obtained from the second data source were supported by high-throughput data and also supported by the predictive software. Third, the interactions which were predicted by miRanda and TargetScan combined with the Ago CLIP-seq data support were more reliable than the fifth and sixth terms. In summary, users can choose the most appropriate data based on their specific task at hand, according to the detailed description provided for each data source.

Co-expression analysis of interactions

Gene co-expression values approximately evaluate co-regulation level and indicate the strength of the functional correlation between two genes (43, 44). We added gene expression correlation scores to the descriptions of interactions in order to help users to determine which interactions are more reliable. Higher co-expression scores in the interactions indicate that the interactions are more reliable if all the interactions were curated from the same data source. The co-expression scores in NPInter v3.0 were computed using an improved pipeline which will be described below, and made use of confident RNA-seq data from a single study to avoid batch effects. We downloaded the public RNA-seq data of human and mouse. The human data set was obtained from Human BodyMap 2.0 (ENA archive: ERP000546) across 16 tissues, while the mouse data set was retrieved from the EMBL-EBI database (ENA archive: ERP000591) across 6 different tissues. Next, we used Tophat (45) to map the reads to the reference genome (hg19 or mm9) and then calculated the FPKM of each reference gene using cuffnorm (46). The reference files for lncRNA genes and protein-coding genes were obtained from the NONCODE v4.0 and RefSeq databases, respectively. Utilizing the expression profiles of all reference genes, we then calculated the Pearson correlation coefficient of any two genes in each interaction curated in NPInter to represent the co-expression value. Finally, the co-expression values, as well as the p value, of each interaction were listed in the database. And one example was provided in the Supplementary Materials to illustrate the co-expression analysis.

Functional annotation of lncRNAs

To enable researchers to have a better understanding of ncRNAs' functions, we predicted the functions of lncRNAs through lnc-GFP (47) with the default parameters, a bi-colored network-based global function predictor according to the interactions curated in NPInter. The protein-protein interactions from the STRING v10 database (48) were integrated into the bi-colored network. A total of 8,710 lncRNA genes in NPInter v3.0 have been annotated with potential functions with a suitable parameter setting. Owing to the number of interactions, this version only supports human and mouse gene function prediction.

Database content and structure

To date, the number of interactions in NPInter v3.0 increased to 491,416 in 22 species (Table 2) informed by 793 published articles, while the previous version of NPInter released in 2013 only contained 201,107 interactions from 18 species. Each updated interaction entry contains basic information including interaction ID, names of the two interacting molecules, interaction level, interaction class, tags, organism, tissue or cell type, experiment description, the interaction description, the data source and the co-expression value. In addition, the number of supporting CLIP reads is also provided for interactions from miRNA–lncRNA prediction with AGO CLIP-seq

Table 2 The statistics of interactions in different species and tissues (or cell lines) in NPInter v3.0

Species	Number of interactions	Tissues or cell lines	Number of interactions
H. sapiens	346 644	HEK293 cells	81 838
M. musculus	143 645	Mouse brain	63 461
S. cerevisiae	571	Human brain	33 165
Agrobacterium tumefaciens	208	HeLa cells	32 456
Escherichia coli	102	MDA-MB-231	33 168
Caenorhabditis elegans	65	Embryonic Stem Cell	26 471
Drosophila melanogaster	58	Cerebrums	22 062
Kaposi sarcoma-associated herpesvirus	41	CD4+ T cells	8928
Others	37	others	49 585

data, and the LncPro value is also provided for the interaction from high-throughput data. Levels were defined according to the types of interacting molecules such as 'RNA–Protein', 'RNA–RNA' and 'RNA–DNA'. Tags of interactions were added according to the same definitions of NPInter v2.0. In order to improve usability to users, we added information on the relevant tissues or cell lines (Table 2), data source and co-expression values for each entry and assigned Ensembl IDs and RefSeq IDs to ncRNAs. For records with binding site information, every binding site is linked to the local UCSC Genome Browser, and an average PhastCons score is calculated.

NPInter v3.0 consists of six major tables:

i. Interaction table: the interaction table provided detailed information for entries. Take interaction 'ncRI-3001387' as an example: this interaction between HOTAIR and PCBP1 was discovered by RIP in gastric tissue of human with a description 'A direct interaction between the HOTAIR and PCBP1 was confirmed through RNA immunoprecipitation coupled with quantitative real-time PCR'. Furthermore, this interaction was tagged as 'ncRNA-Protein binding' and divided into class 'binding' at a 'RNA-protein; level. The interaction was curated from literature mining, and the co-expression value and *P*-value of this interaction from RNA-Seq data was −0.064 and 0.77, respectively.

ii. Molecule table: in this table, we described the name, aliases, molecule type, biotype, organism, a simple description and IDs from relevant databases for every molecule involved in interactions. In terms of ncRNAs, IDs in Ensembl, RefSeq, NONCODE v4.0 or miRBase are available while IDs in UniProt, UniGene and RefSeq are available for proteins or protein-related molecules.

iii. Reference table: the reference table listed general publication information and the MEDLINE standard article code (PMID) of literature recorded in the NPInter database.

iv. Binding site table: for interactions with binding sites generated from sequencing data, we provided the detailed position (chromosome, chromosome start and chromosome end) and PhastCons score for each site.

v. Function table: in the function table, we described the lncRNA genes and their predicted functions in three columns, the gene name of the lncRNAs, GO terms and the description of the GO terms.

vi. miRNA–lncRNA interaction detailed information table: in this table, we provided the detailed information of interactions predicted from software coupled with AGO CLIP-seq data. For example, the interaction 'ncRI-3356692' is predicted by miRanda and TargetScan which is also supported by AGO CLIP reads. The number of supporting reads is 28, the interaction score from miRanda is 157, the interaction energy predicted by miRanda is −17.14 and the interaction region from miRanda is 'chr1: 568999-569024'. In addition, the miRNA family from TargetScan was 'let-7-5p/98-5p/miR-4458/4500', the interaction type was '7mer-m8' and the interaction region predicted by TargetScan was 'chr1: 569017-569023'.

Service update

The web interface of the NPInter v3.0 database has been redesigned and now provides a user-friendly web site to browse and search interactions. In addition, NPInter v3.0 incorporates Cytoscape for users to visualize interactions. An online BLAST service has been integrated as well, enabling users to search entries by sequence. Furthermore, local UCSC Genome Browser has been added in this update.

Integration with a UCSC Genome Browser

As the UCSC Genome Browser (http://genome.ucsc.edu/) has been widely used, a local UCSC Genome Browser was built for *Homo sapiens*, *Mus musculus* and *Saccharomyces cerevisiae* in this new version of NPInter. The NPInter v3.0

track in the Genome Browser displays protein binding sites in these species. Associated tracks like NONCODE v4.0 lncRNA, NONCODE v4.0 lncRNA Gene, RefSeq Genes and Ensembl Genes are also shown in dense mode. Other general tracks, such as Conservation, are retrieved from the UCSC Genome Browser Database and users can change the display mode as they desire.

Conclusion and future directions

Aiming to become a valuable and cutting edge resource for researchers who focus on exploring ncRNAs' functions and molecular mechanisms, NPInter v3.0 has significantly increased the number of records and the amount of detailed information per interaction compared with the former version. NPInter v3.0 provides more detailed information about each individual entry: e.g. users can easily retrieve the basic information about an interaction, the data source where this interaction was curated from, the co-expression values, the exact interacting position, and tissue or cellular locations where this interaction occurred. In addition, in the new version, for records with binding site information, every binding site is linked to the local UCSC Genome Browser as well as an average PhastCons score. Moreover, we also assigned Ensembl IDs and RefSeq IDs to ncRNAs, in addition to the NONCODE ID. As for the interactions between lncRNAs and proteins, we obtained all the interactions supported by the high-throughput data in previous version. In NPInter v3.0, we not only curated the interactions supported by the new high-throughput data but also calculated an interaction score per interaction through LncPro software. The interactions whose scores are equal to or >50 are more reliable compared with the interactions whose scores are <50. In the previous version, we just obtained the interactions between miRNAs and lncRNAs from literature mining. In NPInter v3.0, we not only obtained the interactions from literature mining but also curated the interactions predicted by miRanda or TargetScan combined with Ago CLIP-seq data support. In addition, NPInter v3.0 not only updates the existing tools including BLAST and graph visualization of interactions to the new version, and also integrates Genome Browser services. Although there are some similar databases, NPInter v3.0 contains a more informative and organized data set with its own unique features. Take the RAID (49) and doRiNA (50) databases as examples: RAID only collects interactions in human and discards data generated by high-throughput techniques, while doRiNA mainly focuses on miRNAs without including other kinds of ncRNAs. NPInter contains interactions covering multiple species and ncRNAs, especially lncRNAs.

Furthermore, NPInter is a member of our platform for ncRNAs together with NONCODE, CNCI (51) and ncFANs (52). Consequently, NPInter v3.0 could present the most comprehensive transcriptome-wide map of interactions on ncRNAs for the scientific community.

As the amount of high-throughput sequencing data from a variety of species, tissues, cell lines and RNA-binding proteins increases, the authors will maintain and update the database. Importantly, although NPInter v3.0 has already improved the user interface and added new web services, we are planning to improve the performance of our computer servers, through expanding the memory and upgrading the processors, to provide a better user experience.

Supplementary data

Supplementary data are available at *Database* Online.

References

1. Alexander,R.P., Fang,G., Rozowsky,J. *et al.* (2010) Annotating non-coding regions of the genome. *Nat. Rev. Genet.*, 11, 559–571.
2. Yue,F., Cheng,Y., Breschi,A. *et al.* (2014) A comparative encyclopedia of DNA elements in the mouse genome. *Nature*, 515, 355–364.
3. Tan,M.H., Au,K.F., Yablonovitch,A.L. *et al.* (2013) RNA sequencing reveals a diverse and dynamic repertoire of the Xenopus tropicalis transcriptome over development. *Genome Res.*, 23, 201–216.
4. Pauli,A., Valen,E., Lin,M.F. *et al.* (2012) Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. *Genome Res.*, 22, 577–591.
5. Nam,J.W. and Bartel,D.P. (2012) Long noncoding RNAs in C. elegans. *Genome Res.*, 22, 2529–2540.
6. Wang,Z., Gerstein,M. and Snyder,M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, 10, 57–63.

7. Lafontaine,D.L.J. (2015) Noncoding RNAs in eukaryotic ribosome biogenesis and function. *Nat. Struct. Mol. Biol.*, 22, 11–19.

8. Raina,M. and Ibba,M. (2014) tRNAs as regulators of biological processes. *Front. Genet.*, 5, 171.

9. Karijolich,J. and Yu,Y.T. (2010) Spliceosomal snRNA modifications and their function. *RNA Biol.*, 7, 192–204.

10. Leonardo,T.R., Schultheisz,H.L., Loring,J.F. *et al.* (2012) The functions of microRNAs in pluripotency and reprogramming. *Nat. Cell Biol.*, 14, 1114–1121.

11. Morris,K.V., Chan,S.W.L., Jacobsen,S.E. *et al.* (2004) Small interfering RNA-induced transcriptional gene silencing in human cells. *Science*, 305, 1289–1292.

12. Vagin,V.V., Sigova,A., Li,C.J. *et al.* (2006) A distinct small RNA pathway silences selfish genetic elements in the germline. *Science*, 313, 320–324.

13. Lee,J.T.and Bartolomei,M.S. (2013) X-inactivation, imprinting, and long noncoding rnas in health and disease. *Cell*, 152, 1308–1323.

14. Flynn,R.A. and Chang,H.Y. (2014) Long noncoding rnas in cell-fate programming and reprogramming. *Cell Stem Cell*, 14, 752–761.

15. Esteller,M. (2011) Non-coding RNAs in human disease. *Nat. Rev. Genet.*, 12, 861–874.

16. Jonas,S. and Izaurralde,E. (2015) NON-CODING RNA Towards a molecular understanding of microRNA-mediated gene silencing. *Nat. Rev. Genet.*, 16, 421–433.

17. Lee,J.T. (2012) Epigenetic regulation by long noncoding rnas. *Science*, 338, 1435–1439.

18. Konig,J., Zarnack,K., Luscombe,N.M. *et al.* (2012) Protein-RNA interactions: new genomic technologies and perspectives (vol 13, pg 77, 2012). *Nat. Rev. Genet.*, 13, 221–221.

19. Chu,C., Qu,K., Zhong,F.L. *et al.* (2011) Genomic maps of long noncoding RNA occupancy reveal principles of RNA-chromatin interactions. *Mol. Cell*, 44, 667–678.

20. McHugh,C.A., Chen,C.K., Chow,A. *et al.* (2015) The Xist lncRNA interacts directly with SHARP to silence transcription through HDAC3. *Nature*, 521, 232.

21. Chu,C., Zhang,Q.F.C., da Rocha,S.T. *et al.* (2015) Systematic discovery of Xist RNA binding proteins. *Cell*, 161, 404–416.

22. Minajigi,A., Froberg,J.E., Wei,C.Y. *et al.* (2015) A comprehensive Xist interactome reveals cohesin repulsion and an RNA-directed chromosome conformation. *Science*, 349, 282.

23. Wu,T., Wang,J., Liu,C.N. *et al.* (2006) NPInter: the noncoding RNAs and protein related biomacromolecules interaction database. *Nucleic Acids Res.*, 34, D150–D152.

24. Yuan,J., Wu,W., Xie,C.Y. *et al.* (2014) NPInter v2.0: an updated database of ncRNA interactions. *Nucleic Acids Res.*, 42, D104–D108.

25. Liu,C.N., Bai,B.Y., Skogerbo,G. *et al.* (2005) NONCODE: an integrated knowledge database of non-coding RNAs. *Nucleic Acids Res.*, 33, D112–D115.

26. He,S.M., Liu,C.N., Skogerbo,G. *et al.* (2008) NONCODE v2.0: decoding the non-coding. *Nucleic Acids Res.*, 36, D170–D172.

27. Bu,D.C., Yu,K.T., Sun,S.L. *et al.* (2012) NONCODE v3.0: integrative annotation of long noncoding RNAs. *Nucleic Acids Res.*, 40, D210–D215.

28. Xie,C.Y., Yuan,J., Li,H. *et al.* (2014) NONCODEv4: exploring the world of long non-coding RNA genes. *Nucleic Acids Res.*, 42, D98–D103.

29. Pruitt,K.D., Brown,G.R., Hiatt,S.M. *et al.* (2014) RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res.*, 42, D756–D763.

30. Cunningham,F., Amode,M.R., Barrell,D. *et al.* (2015) Ensembl 2015. *Nucleic Acids Res.*, 43, D662–D669.

31. Kozomara,A. and Griffiths-Jones,S. (2014) miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.*, 42, D68–D73.

32. Bateman,A., Martin,M.J., O'donovan,C. *et al.* (2015) UniProt: a hub for protein information. *Nucleic Acids Res.*, 43, D204–D212.

33. Yang,Y.C., Di,C., Hu,B. *et al.* (2015) CLIPdb: a CLIP-seq database for protein-RNA interactions. *Bmc Genomics*, 16, 51.

34. Barrett,T., Wilhite,S.E., Ledoux,P. *et al.* (2013) NCBI GEO: archive for functional genomics data sets-update. *Nucleic Acids Res.*, 41, D991–D995.

35. Rosenbloom,K.R., Armstrong,J., Barber,G.P. *et al.* (2015) The UCSC Genome Browser database: 2015 update. *Nucleic Acids Res.*, 43, D670–D681.

36. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26, 841–842.

37. Lu,Q.S., Ren,S.J., Lu,M. *et al.* (2013) Computational prediction of associations between long non-coding RNAs and proteins. *Bmc Genomics*, 14,

38. Siepel,A., Bejerano,G., Pedersen,J.S. *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, 15, 1034–1050.

39. Bazzini,A.A., Johnstone,T.G., Christiano,R. *et al.* (2014) Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *EMBO J.*, 33, 981–993.

40. Fan,R., Bonde,S., Gao,P. *et al.* (2012) Dynamic HoxB4-regulatory network during embryonic stem cell differentiation to hematopoietic cells. *Blood*, 119, e139–e147.

41. Agarwal,V., Bell,G.W., Nam,J.W. *et al.* (2015) Predicting effective microRNA target sites in mammalian mRNAs. *eLife*, 4,

42. Betel,D., Koppal,A., Agius,. *et al.* (2010) Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. *Genome Biol.*, 11, R90.

43. Marcotte,E.M., Pellegrini,M., Thompson,M.J. *et al.* (1999) A combined algorithm for genome-wide prediction of protein function. *Nature*, 402, 83–86.

44. Eisen,M.B., Spellman,P.T., Brown,P.O. *et al.* (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA*, 95, 14863–14868.

45. Kim,D., Pertea,G., Trapnell,C. *et al.* (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.*, 14, R36.

46. Trapnell,C., Williams,B.A., Pertea,G. *et al.* (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, 28, 511. U174.

47. Guo,X.L., Gao,L., Liao,Q. *et al.* (2013) Long non-coding RNAs function annotation: a global prediction method based on bi-colored networks. *Nucleic Acids Res.*, 41, e35.

48. Szklarczyk,D., Franceschini,A., Wyder,S. *et al*. (2015) STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res*., 43, D447–D452.

49. Zhang,X., Wu,D., Chen,L. *et al*. (2014) RAID: a comprehensive resource for human RNA-associated (RNA-RNA/RNA-protein) interaction. *RNA*, 20, 989–993.

50. Blin,K., Dieterich,C., Wurmus,R. *et al*. (2015) DoRiNA 2.0-upgrading the doRiNA database of RNA interactions in post-transcriptional regulation. *Nucleic Acids Res*., 43, D160–D167.

51. Sun,L., Luo,H.T., Bu,D.C. *et al*. (2013) Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. *Nucleic Acids Res*., 41, e166.

52. Liao, Q., Xiao, H., Bu, D.C. *et al*. (2011) ncFANs: a web server for functional annotation of long non-coding RNAs. *Nucleic Acids Res*., 39, W118–W124.