

ITAS: integrated transcript annotation for small RNA

Bezuglov V.^{1,2*}, Stupnikov A.^{3,4}, Skakov I.², Shtratnikova V.¹, Pilsner J.R.⁵,
Suvorov A.^{1,6}, Sergeyev O.^{1*}

¹ Belozersky Institute of Physico-Chemical Biology, Lomonosov Moscow State University, Moscow, Russia

² Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University, Moscow, Russia

³ Moscow Institute of Physics and Technology, Moscow, Russia

⁴ National Medical Research Center for Endocrinology, Moscow, Russia

⁵ Department of Obstetrics and Gynecology, Wayne State University School of Medicine, Detroit, USA

⁶ Department of Environmental Health Sciences, University of Massachusetts, Amherst, MA, USA

* vitaly1530@gmail.com, olegsergeyev1@yandex.ru

Key words: Transcript annotation, Differential Gene Expression, small RNA, microRNA, piRNA, tRNA-derived small RNA, RNA-seq, small RNA fragments, transcriptomics

Motivation and Aim: Transcriptomics analysis of various small RNA (sRNA) biotypes is a new and rapidly developing field. Annotations for microRNAs, tRNAs, tRNA-derived (tsRNA), piRNAs and rRNAs, contain information on transcripts sequences and loci that is vital for downstream analyses. Several databases have been established to provide this type of data for specific RNA biotypes. However, these sources often contain data in different formats, which makes bulk analysis of several sRNA biotypes in a single pipeline challenging. Information on some transcripts may be incomplete or conflicting with other entries. To overcome these challenges, we aimed to introduce ITAS, or Integrated Transcript Annotation for Small RNA, a filtered, corrected and integrated transcript annotation containing information on several types of small RNAs including tsRNAs for several species (*Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Drosophila melanogaster*, *Caenorhabditis elegans*), which was tested in several case studies for human-derived data against existing alternative databases.

Methods and Algorithms: Files with transcripts sequences in fasta format (referred to as fasta sequences) and annotation of transcripts in bed-format (loci) were obtained from corresponding databases: GtRNAdb Data Release 19 (June 2021) for mature tRNA sequences, piRNAdb v1.7.6 for piRNAs, miRBase Release 22.1 for microRNAs, UCSC for rRNA, MINTbase v2.0 (Human) and tRFdb for tsRNA. For some loci additional UCSC liftOver procedure followed to transform loci to the most recent genome version. To overcome the incompleteness of databases or conflicting data for some transcripts sequences in fasta-format from the reference genome (getfasta sequences) which correspond to the annotation bed-file were obtained by bedtools getfasta version 2.27.1, and fasta sequences were mapped on the reference genome of current version by hisat2. Then consensus database for each sRNA type was built, and transcripts correction was proceeded with the help of R instruments. Transcripts with conflicting data due to inbase and interbase (mature microRNA, piRNA, rRNA, tRNA) loci conflicts were detected with the help of bedtools intersect. Human sperm RNA-seq data from three publicly available datasets [1–3] were used for case studies. The differential expression analysis was done by DESeq2 package in relation to factors analyzed in original studies.

Results: The most popular databases of sRNA were analyzed. The following steps were invoked to approach outlined problems while generating the ITAS: (1) missing data for the incomplete entries were retrieved and filled in; (2) the problem of multiple loci per transcript was addressed; (3) transcripts with conflicting fasta-delivered and locus-delivered data were

identified, corrected if possible, and filtered out otherwise; (4) in-database and inter-databases loci-wise intersecting entries were identified and filtered out; (5) inter-databases loci-wise intersecting entries were identified and filtered out.

The identified problems were corrected where possible and entries with severe conflicts were filtered out. The statistics for Human annotations correction for different biotypes of small RNA is provided in Table 1. The conducted case studies using human sperm RNA-seq data [1–3] demonstrated the advantages of ITAS. Mapping of reads to ITAS which unifies in a single gtf format database transcripts from all source databases allowed for the identification of more significant transcripts as compared with the “map and remove” remove’ approach. In particular for sRNAs expression analysis, all cases revealed that ITAS based genome alignment approach identified more significant transcripts than the “map and remove” pipeline with default databases (11 vs 66 for [1], 43 vs 212 for [2], 26 vs 242 for [3]).

Table 1. Statistics on correction events in Human RNA transcript entries: cases when both locus and sequence were present (Loci & Seq, no correction), only locus or only sequence (Loci only, Seq only, sequence or locus retrieved from genome); cases that required extending entry’s locus, sequence or both (Ext Loci & Seq, Loci & Ext Seq, Ext Loci & Ext Seq); cases with transcripts loci intersections within same database (Inbase conflicts) and intersections between different databases (Interbases conflicts)

RNA type	Database	Loci & Seq	Loci only	Seq only	Ext Loci & Seq	Loci & Ext Seq	Ext Loci& Ext Seq	Inbase conflict	Inter-bases
mature microRNA	miRBase	1291	1227	1	0	0	0	140	21
piRNA	piRNadb	42 2017	0	1	0	0	0	388 826	1268
tRNA	GtRNadb	225	176	2	0	0	0	0	214
rRNA	UCSC	1591	0	186	0	0	0	37	126
tRNA-derived	MINTbase	0	0	8120	115 040	33	8	*	*

* No intersection events were considered for tsRNAs, as fragments of the same tRNA naturally have intersecting loci.

Conclusion: We have identified several issues during the inference of existing databases containing information on sRNA transcripts sequences and annotation for five species. Some transcripts had missing information on their sequences or loci; for others, their genome locus-retrieved sequence and database provided-sequence were not matching. Transcripts had both in-database and inter-databases intersecting loci with other entries. To address these drawbacks we established ITAS, a filtered, corrected, and integrated database for 5 species in a widely used gtf format, which was tested in several case studies for human-derived data against existing alternative databases.

Acknowledgements: The study was supported by the Russian Science Foundation grant No. 18-15-00202, <https://rscf.ru/project/18-15-00202>.

References

1. Donkin I., Versteyhe S., Ingerslev L.R. et al. Obesity and bariatric surgery drive epigenetic variation of spermatozoa in humans. *Cell Metabolism*. 2016;23(2):369-378.
2. Ingerslev L.R., Donkin I., Fabre O., Versteyhe S., Mechta M., Pattamaprapanont P., Mortensen B., Krarup N.T., Barrès R. Endurance training remodels sperm-borne small RNA expression and methylation at neurological gene hotspots. *Clin Epigenetics*. 2018;10:12.
3. Hua M., Liu W., Chen Y., Zhang F., Xu B., Liu S., Chen G., Shi H., Wu L. Identification of small non-coding RNAs as sperm quality biomarkers for *in vitro* fertilization. *Cell Discov*. 2019;5:20.