Bioinformation by Biomedical Informatics Publishing Group *open access*
www.bioinformation.net **Software**

BatchGenAna: a batch platform for large-scale genomic analysis of mammalian small RNAs

Xiaomin Ying[1], You Jung Kim[2], Yiqing Mao[3], Ming Liu[4], Yanyan Hou[1], Hua Li[1], Xiaolei Wang[3], Yalin Zhao[1], Dongsheng Zhao[3], Jignesh M. Patel[2], Wuju Li[1]*

[1]Center of Computational Biology, Beijing Institute of Basic Medical Sciences, Beijing 100850, China; [2]Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109, USA; [3]Beijing Institute of Health Administration and Medicine Information, Beijing 100850, China; [4]Beijing Medical Library, Beijing 100039, China; Wuju Li - Email: liwj@bmi.ac.cn; Tel: +86 10 66931324; Fax: +86 10 68213039; *Corresponding author

received January 21, 2009; accepted March 06, 2009; published April 21, 2009

Abstract:
An increasing number of small RNAs have been discovered in mammals. However, their primary transcripts and upstream regulatory networks remain largely to be determined. Genomic analysis of small RNAs facilitates identification of their primary transcripts, and hence contributes to researches of their upstream regulatory networks. We here report a batch platform, BatchGenAna, which is specifically designed for large-scale genomic analysis of mammalian small RNAs. It can map and annotate for as many as 1000 small RNAs or 10,000 genomic loci of small RNAs at a time. It provides genomic features including RefSeq genes, mRNAs, ESTs and repeat elements in tabular and graphical results. It also allows extracting flanking sequences of submitted queries, specified genomic regions and host transcripts, which facilitates subsequent analysis such as scanning transcription factor binding sites in upstream sequences and poly(A) signals in downstream sequences. Besides small RNA fields, BatchGenAna can also be applied to other research fields, e.g. in silico analysis of target genes of transcription factors.

Availability: The platform is freely available at http://biosrv1.bmi.ac.cn/BatchGenAna.

Keywords: small RNA; genomic analysis; primary transcript

Background:
Small RNAs of less than 40 nucleotides (nt) in length, such as microRNAs (miRNAs), small interfering RNAs (siRNAs), scanRNAs (scnRNAs) and piwi-interacting RNAs (piRNAs), constitute a large family of tiny regulatory molecules. Among them, miRNAs, piRNAs and siRNAs are also discovered in mammals, which have diverse and important functions. Although our knowledge of mammalian small RNAs has advanced rapidly, the primary transcripts of most mammalian small RNAs remain to be determined. Uncovering primary transcripts of small RNAs is very important to our understanding of the biogenesis of small RNAs. It facilitates (a) identifying the regulatory regions such as transcription factor binding sites (TFBS) and hence discovering upstream regulators in the network, (b) detecting other sequence and structural motifs required in small RNA processing, and (c) providing essential information for small RNA knockout **[1]**. Genomic analysis is an efficient way to identify primary transcripts before experiments. For example, several studies have attempted to delineate the genomic boundaries of mammalian miRNAs by large-scale genomic analysis **[1-3]**.

Up to now, there are two ways for large-scale genomic analysis of mammalian small RNAs. One way is to download genome sequences, annotations and other data into local machines and to implement in-house programs to analyze the genomic features. Although this way is flexible, it requires significant effort and computer skills for use. Researchers are required to map sequences to genomes, to build local database on annotations and to develop computer programs for parsing, retrieving and displaying the results, which are very challenging for most molecular biology laboratories. The other way is to utilize public databases such as UCSC **[4],** NCBI **[5]** or Ensembl **[6].** These public databases align users' sequences (one or no more than 30 sequences at a time) against the specified genome, display one genome view at a time and extract flanking sequence for one sequence interactively. However, these interactive tools need much manual work. Doing large-scale genomic analysis with them would be exhausting. Moreover, researchers will have to record the overlapping ESTs and mRNAs manually if they want to further analyze tissue sources or other features of these transcripts. Recently, miRBase has been updated with genomic features of known miRNAs **[7]**, and piRNABank is built for human, mouse and rat piRNAs supporting searches for overlapping genes and repeat elements **[8]**. However, piRNABank does not provide other genomic features like EST matches and flanking sequence extraction for further analysis, and these two specific databases do not support genomic analysis for other classes of small RNAs or newly discovered small RNAs.

To facilitate large-scale genomic analysis of mammalian small RNAs, we have developed a batch platform, BatchGenAna, to provide batch mapping and annotating for as many as 1000 nucleotide sequences or 10,000 genomic loci of small RNAs at a time. BatchGenAna provides genomic features including RefSeq genes, mRNAs, ESTs and repeat elements, and produces both tabular and graphical results containing the selected genomic features. It also supports extraction of flanking sequences of submitted queries, specified genomic regions and host transcripts, which facilitates subsequent analysis such as scanning TFBS in upstream sequences and poly(A) signals in downstream sequences.

Bioinformation by Biomedical Informatics Publishing Group *open access*
www.bioinformation.net **Software**

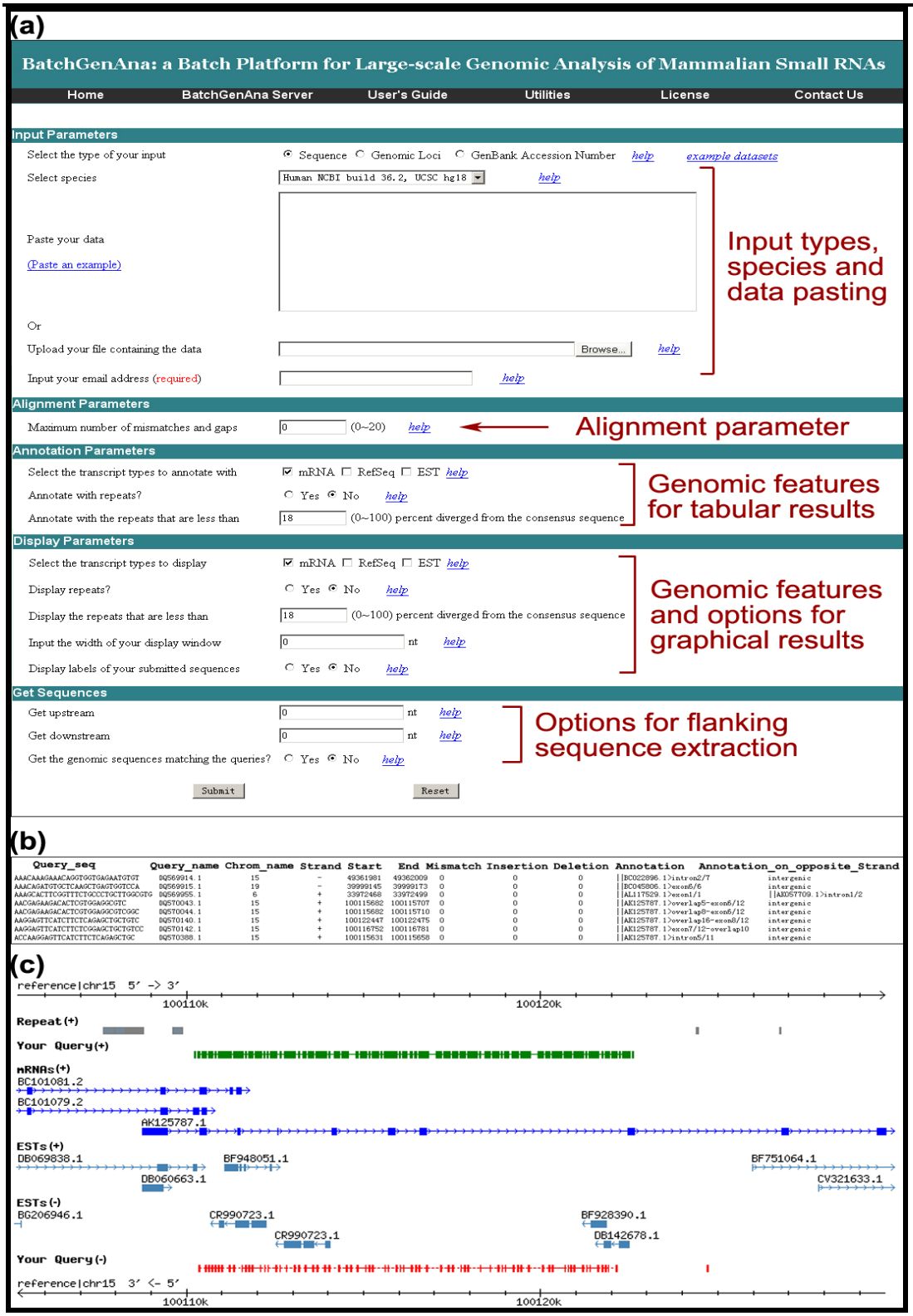

Figure 1: (a) A snapshot of BatchGenAna. (b) An example tabular result of human piRNAs for mRNAs. The genomic loci and the accession numbers of mRNAs overlapping submitted piRNAs are listed. (c) An example genome view of small RNA cluster. Small RNAs are denoted as green boxes (in plus strand) and red boxes (in minus strand). ESTs, mRNAs and repeat elements are displayed in different tracks with different colours.

Methodology of development:
The genome sequences are downloaded from NCBI **[5]**. The annotations of mRNAs, ESTs and RefSeq genes are downloaded from NCBI map viewer **[5]**. The repeat annotations of human, mouse and rat genome are downloaded from UCSC **[4]**. At the time of writing this manuscript, the human, mouse and rat genome assemblies and annotations are NCBI build 36.2, 37.1 and 4.1 respectively. To improve computational efficiency while achieving high sensitivity, BatchGenAna employs miBLAST **[9]** to search the sequences against the genome for sequences shorter than 40 nt. In our case, miBLAST is ~30 times faster than BLAST **[10]** for sequences shorter than 40 nt and has the same sensitivity. For those longer, BLAT is employed, since it is more accurate and much faster than popular existing tools for mRNA/DNA alignments when comparing vertebrate sequences **[11]**. A central MySQL database is used to store the downloaded data. The BatchGenAna interface was written in PHP. The program of sequence mapping, alignment result parsing and tabular/graphical result generating are implemented in C, Perl and Bioperl **[12]**. The current web service is running an Apache web server on a PC Linux box with quad Intel Xeon 3.2GHz processors and 4GB RAM. The operating system is Redhat Linux AS3.

Utility:
BatchGenAna provides a friendly interface for users to specify parameters for input type, species, alignment, annotation, display and flanking sequence extraction (Figure 1a). Users may input nucleotide sequences, genomic loci or GenBank accession numbers. The input type of Genomic loci is provided for users who have already obtained the genomic loci of small RNAs or small RNA clusters. The input type of GenBank accession number is provided for users to extract flanking sequences of the overlapping transcripts in a second run. Users can specify the genomic features they focus on, including ESTs, mRNAs, RefSeq genes and repeat elements. Users can also specify the window width of genome views and the upstream/downstream length for flanking sequences extraction. Since batch jobs are more time-consuming than immediate jobs, users are required to input their email addresses so that BatchGenAna can notify users instantly once their jobs are completed (Executing time of variable batch jobs is provided in the homepage).

BatchGenAna produces both tabular and graphical results. In tabular results, GenBank accession numbers of ESTs, mRNAs, RefSeq genes and the names/families of repeat elements are listed if they overlap with the submitted queries. Tabular results facilitate users to analyze tissue distribution and other features of their transcripts (Figure 1b). In graphical results, BatchGenAna displays genome views centered at submitted queries with selected genomic features. If the distance of two consecutive genomic loci of submitted queries is less than half of the display window width, these two genomic loci are plotted in the same view (Figure 1c). The detailed information and guidance is given in the webpage.

Future development:
Work is under way to incorporate CpG islands, TSS, polyA signals, TFBS and CAGE tags into BatchGenAna. We also plan to integrate annotations from NCBI, UCSC and Ensembl into BatchGenAna, since their genome annotations are somewhat different.

Acknowledgements:
We thank Wenjie Shu, Xiaochen Bo, Yuan Cao and Fei Li for their great help in web service implementation. This work is supported by grants of the National Natural Science Foundation of China (30500105 to XY, 30470411 to WL), the National Science Foundation (DBI-0543272 to JMP) and the National Institutes of Health (1-U54-DA021519-01A1 to JMP).

References:
[1] H. K. Saini et al., Proc Natl Acad Sci U S A. (2007) 104: 17719 [PMID: 17965236]
[2] Rodriguez et al., Genome Res (2004) 14: 1902 [PMID: 15364901]
[3] G. Jin et al., Mammalian Genome (2006) 17: 1033 [PMID: 17019647]
[4] D. Karolchik et al. Nucleic Acids Res (2008) 36: D773 [PMID: 18086701]
[5] D. L. Wheeler et al., Nucleic Acids Res (2008) 36: D13 [PMID: 18045790]
[6] P. Flicek et al., Nucleic Acids Res (2008) 36: D707 [PMID: 18000006]
[7] S. Griffiths-Jones et al., Nucleic Acids Res (2008) 36: D154 [PMID: 17991681]
[8] S. Lakshmi & S. Agrawal, Nucleic Acids Res (2008) 36: D173 [PMID: 17881367]
[9] Y. J. Kim et al., Nucleic Acids Res (2005) 33: 4335 [PMID: 16061938]
[10] S. F. Altschul et al., J Mol Biol (1990) 215: 403 [PMID: 2231712]
[11] W. J. Kent, Genome Res (2002) 12: 656 [PMID: 11932250]
[12] J. E. Stajich et al., Genome Res (2002) 12: 1611 [PMID: 12368254]

Edited by P. Kangueane
Citation: Ying et al. Bioinformation 3(8): 346-348 (2009)