Databases and ontologies

FoldAtlas: a repository for genome-wide RNA structure probing data

Matthew Norris1,*, Chun Kit Kwok2, Jitender Cheema1, Matthew Hartley1, Richard J. Morris1, Sharon Aviran3 and Yiliang Ding1,*

1John Innes Centre, Norwich Research Park, Norwich, UK, 2Department of Biology and Chemistry, City University of Hong Kong, Kowloon Tong, Hong Kong SAR, China and 3Department of Biomedical Engineering and Genome Center, UC Davis, Davis, CA, USA

*To whom correspondence should be addressed.

Abstract

Summary: Most RNA molecules form internal base pairs, leading to a folded secondary structure. Some of these structures have been demonstrated to be functionally significant. High-throughput RNA structure chemical probing methods generate millions of sequencing reads to provide structural constraints for RNA secondary structure prediction. At present, processed data from these experiments are difficult to access without computational expertise. Here we present FoldAtlas, a web interface for accessing raw and processed structural data across thousands of transcripts. FoldAtlas allows a researcher to easily locate, view, and retrieve probing data for a given RNA molecule. We also provide in silico and in vivo secondary structure predictions for comparison, visualized in the browser as circle plots and topology diagrams. Data currently integrated into FoldAtlas are from a new high-depth Structure-seq data analysis in Arabidopsis thaliana, released with this work.

Availability and Implementation: The FoldAtlas website can be accessed at www.foldatlas.com. Source code is freely available at github.com/mnori/foldatlas under the MIT license. Raw reads data are available under the NCBI SRA accession SRP066985.

Contact: yiliang.ding@jic.ac.uk or matthew.norris@jic.ac.uk.

Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

RNA structure plays an important role in all steps of gene expression and regulation (Mortimer et al., 2014; Sharp, 2009). Earlier studies inferred the secondary structures of individual RNA sequences using low throughput in vitro probing or in silico prediction approaches. More recently, genome-wide in vivo structure probing methods have emerged, allowing structures to be determined across the transcriptomes of living cells (Ding et al., 2014; Rouskin et al., 2014; Spitale et al., 2013; Talkish et al., 2014; Tang et al., 2015).

Chemical probing methods can be used to determine RNA secondary structure in living cells (Kwok et al., 2013; McGinnis and Weeks, 2014; Spitale et al., 2013). These methods include dimethyl sulfate (DMS) probing and selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE). In DMS probing, the N1 position of adenine and the N3 position of cytosine are methylated when the base is not involved in Watson–Crick base pairing. In SHAPE, all unpaired bases are modified. Chemically modified bases lead to stalling of reverse transcriptase. With reverse transcription, PCR, deep sequencing and normalization, reactivities can be assigned to individual RNA sequence positions. These reactivities describe the extent of exposure of a nucleotide to solution, and can be exploited as pseudo-free energy constraints for

© The Author 2016. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com
FoldAtlas displays the splicing configuration of the selected transcript, functionality across the genome. When a transcript is selected and processed structure probing data, along with pre-generated RNA secondary structure prediction. At present, raw and processed reactivity data are hard to access without computational expertise.

Here we introduce FoldAtlas, a repository and web interface for accessing genome-scale RNA structure probing data. We also provide visualization of data-constrained RNA structures across the genome. The data included with FoldAtlas are from a new high-depth Structure-seq DMS analysis in Arabidopsis thaliana, covering over 11,000 transcripts.

2 Results

FoldAtlas allows a researcher without computational expertise to select a transcript of interest and retrieve its corresponding raw and processed structure probing data, along with pre-generated RNA structure predictions. FoldAtlas is the first tool that provides this functionality across the genome. When a transcript is selected and loaded, the d3nome genome browser (Fig. 1A) released with this work, displays the splicing configuration of the selected transcript, along with alternative splice isoforms, where relevant. An overview of the normalized chemical reactivities (Fig. 1B) is also shown, which can be expanded to show detailed nucleotide-resolution chemical reactivities. The reactivities are generated as described in the Supplementary Results section of the Supplementary Material.

Tab delimited text files containing normalized chemical reactivities are available for download. We also provide corresponding raw read termination counts from 3 independent biological replicates, allowing the significance of structure probing data to be estimated by assigning errors to reactivities.

For each transcript, we include the 20 lowest free energy unconstrained in silico and data-constrained in vivo structures generated by using the Fold program, from version 5.7 of the RNAstructure package (Reuter and Mathews, 2010), with default slope and intercept parameters of 1.8 and -0.6 kcal/mol respectively. The structure prediction set includes the MFE structure alongside suboptimal low free energy structures. Differences and similarities between these structures are visualized using a Principal Components Analysis (PCA) view (Fig. 1C). PCA plots were generated using a previously described method (Halvorsen et al., 2010). Each structure can also be visualized using both circle plots (Fig. 1D) and structure diagrams (Fig. 1E) generated using the ViennaRNA package (Hofacker, 2013; Kerpedjiev et al., 2015; Lorenz et al., 2011). The corresponding MFE structures can be downloaded as tab-delimited text files.

The FoldAtlas chemical reactivity data are from a DMS chemical modification experiment in Arabidopsis thaliana. These data were generated by using a previously established Structure-seq method (Ding et al., 2014, 2015), but with two rounds of poly-A selection to enrich the proportion of mRNA. Detailed analysis of this experiment is provided in the Supplementary Results section of the Supplementary Material.

3 Conclusions and future work

FoldAtlas provides convenient access to in vivo RNA structure probing data across thousands of transcripts. The current release, 1.1, includes data from a high depth genome-scale probing experiment in Arabidopsis thaliana. To predict structure for a transcript, we generated up to 20 secondary structures using the RNAstructure Fold tool, and visualized the structure ensemble using PCA plots. In this work, our preference to use RNAstructure is due to the ability to specify experimental constraints, and is consistent with the approach taken in our earlier work (Ding et al., 2014). In future versions of FoldAtlas, we plan to also provide options to visualize structure predictions made using other methods, including SeaFold (Ouyang et al., 2013), and Vienna RNAfold, which now allows experimental constraints (Lorenz et al., 2011, 2015). We are also considering including SHAPE probing data, in vitro data, reactivities calculated using alternative normalization methods, data from other organisms, and data from other studies.

Acknowledgements

This research was supported in part by the NBIP Computing infrastructure for Science (CGS) group through the provision of a High Performance Computing Cluster.

Funding

This work was supported by the Biotechnology and Biological Sciences Research Council (BBSRC) [BB/L025000/1 to M.N. and Y.D. and BB/N022572/1 to M.N. Y.D and S.A.]; and the National Institutes of Health [HG006860 to S.A.].

Conflict of Interest: none declared.

References