THE
EMBO
JOURNAL

Manuscript EMBO-2014-88411

Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation

Ariel A. Bazzini, Timothy G. Johnstone, Romain Christiano, Sebastian D. Mackowia, Benedikt Obermayer , Elizabeth S. Fleming ,Charles E. Vejnar, Miler T. Lee , Nikolaus Rajewsky, Tobias C. Walther and Antonio J. Giraldez

Corresponding authors:

Antonio J. Giraldez, Yale University School of Medicine
Ariel A. Bazzini, Yale University School of Medicine
Nikolaus Rajewsky, Max-Delbrück-Center for Molecular Medicine

Review timeline:

Submission date:	06 March 2014
Editorial Decision:	10 March 2014
Revision received:	13 March 2014
Accepted:	14 March 2014

Transaction Report:

(Note: With the exception of the correction of typographical or spelling errors that could be a source of ambiguity, letters and reports are not edited. The original formatting of letters and referee reports may not be reflected in this compilation.)

Editor: Thomas Schwarz-Romond

Transfer Note 09 October 2013

PLEASE NOTE that this manuscript was transferred from a different journal and the arbitrating referee assessing suitability for The EMBO Journal had access to both the original anonymous comments as well as the point by point response provided by the authors.

Editorial Staff
The EMBO Journal

1st Editorial Decision 10 March 2014

Thank you very much for submitting your manuscript on the genome-wide identification of small ORFs employing two complementary approaches for consideration to The EMBO Journal.

Taking advantage of the transmitted referee comments that originate from peer-review at another scientific title, I consulted with an external scientist for relatively rapid arbitration/scientific integrity of the dataset. As you will see from the enclosed comments, this expert essentially supports publication of your paper. We will thus be delighted to formally accept the paper for publication in The EMBO Journal upon a few editorial requirements.

-Please respond to the remarks of the referee by adding relevant statements to the discussion to your earliest convenience.

-Please check once more quality and completeness of the provided figures.

-Please provide via the online system all required details (affiliation, E-mail addresses, etc.) of all relevant authors/co-authors.

-Please also provide a short synopsis (minimal 2 up to 4 'bullet point') that highlight the major novelty/advance provided by your study (via separate E-mail OR textfile).

-Please note that we will also aim to feature the manuscript with one of our 'Have you seen?' as to integrate your results into the context of most recent developments in this area of research.

We are very much looking forward to hear from you re final amendments to the manuscript files to ensure rapid production/publication.

REFEREE REPORT:

Referee #1:

Evidence has been emerging for functions associated with novel small peptides. Several of these have been identified in transcripts previously thought to be long non-coding RNAs. Recently developed methods for ribosome footprinting have provided a means to systematically identify RNAs with protein coding capacity, whether they contain long ORFs. This approach has provided access to transcripts that can encode short peptides, which might otherwise have been considered to be non-coding RNAs

In this report Bazzini & Johnstone et al. describe a method to identify small translated ORFs that combines ribosome footprinting, with computational analysis to detect ribosome phasing as a means to identify small translated ORFs. The ends of Zebrafish RNA fragments protected by ribosome binding showed a phasing relationship to the reading frame of the ORFs in mRNAs. This was not seen in the input RNA fragments. Evidence of phasing provided a starting point to develop a computational method "ORFscore" to identify ORFs. ORFscore was able to correctly identify 99% of all Zebrafish refseq transcripts.

The method is in part based on the co-occurrence of multiple phased ribosome bound fragments within a given ORF sequence. Shorter ORFs should therefore pose a greater challenge, and indeed they report correct identification of 86% of known ORFS <100aa.

Applying ORFscore to 2450 transcripts without known ORFs identified 303 transcripts with novel small ORFs (termed smORFS). After removing redundancy they identified 190 new ORFs between 20-100aa, and 89 ORFs >100aa. This set overlaps with ~50% of a set of previously predicted short ORFs in zebrafish. Six of the 190 new smORFs were validated by mass spectrometry on embryo lysates, along with 98 previously annotated smORFs. It would be helpful to the reader to have an explicit statement on the proportion of new and previously known smORFs detected by MS. The comment on the effect of peptide length and coverage is appreciated.

The authors, appropriately, consider the possibility that phased ribosome binding might not be sufficient to predict translation of a stable peptide. They describe a computational micropeptide detection method, micPDP, which uses evolutionary conservation for smORF detection. 63 smORFs were detected, of which 23 were also found by the ORFscore approach. This difference is attributed in part to lack of sequence conservation in some of the smORFs detected by ribosome footprinting. Detection by the two methods appears to provide a high confidence set of smORF prediction. Comparison of micPDP predictions with ORFscore analysis of previously published ribosome footprinting data also showed limited (albeit statistically significant) overlap between the two methods.

Overall the data are sound and the analysis supports the new peptide predictions as described. Understandably, the discussion aims to provide explanations that might reconcile the limited overlap between the methods. A few words discussing over- and under-prediction and the importance of experimental validation might not be amiss.

I have read the confidential reviews provided by the authors from another journal, and find no outstanding issues of substance that would preclude publication. The manuscript is suitable for publication.

1st Revision - authors' response 13 March 2014

Evidence has been emerging for functions associated with novel small peptides. Several of these have been identified in transcripts previously thought to be long non-coding RNAs. Recently developed methods for ribosome foot printing have provided a means to systematically identify RNAs with protein coding capacity, whether they contain long ORFs. This approach has provided access to transcripts that can encode short peptides, which might otherwise have been considered to be non-coding RNAs In this report Bazzini & Johnstone et al. describe a method to identify small translated ORFs that combines ribosome foot printing, with computational analysis to detect ribosome phasing as a means to identify small translated ORFs. The ends of Zebrafish RNA fragments protected by ribosome binding showed a phasing relationship to the reading frame of the ORFs in mRNAs. This was not seen in the input RNA fragments. Evidence of phasing provided a starting point to develop a computational method "ORFscore" to identify ORFs. ORFscore was able to correctly identify 99% of all Zebrafish refseq transcripts.

The method is in part based on the co-occurrence of multiple phased ribosome bound fragments within a given ORF sequence. Shorter ORFs should therefore pose a greater challenge, and indeed they report correct identification of 86% of known ORFS <100aa.

Applying ORFscore to 2450 transcripts without known ORFs identified 303 transcripts with novel small ORFs (termed smORFS). After removing redundancy they identified 190 new ORFs between 20-100aa, and 89 ORFs >100aa. This set overlaps with ~50% of a set of previously predicted short ORFs in zebra fish. Six of the 190 new smORFs were validated by mass spectrometry on embryo lysates, along with 98 previously annotated smORFs. It would be helpful to the reader to have an explicit statement on the proportion of new and previously known smORFs detected by MS. The comment on the effect of peptide length and coverage is appreciated.

We appreciate this comment. To make it clear, we have defined 302 translated smORFs using ORFscore that were previously annotated as coding ORFs.

From those (302) we have detected 98 peptides by MS (~32%). On the other hand, we defined 190 novel sORF and from those we detected 6 (~3%). We understand that for the reader might be useful to have the proportion of new and previously known smORFs detected by MS.

However, as the detection highly depends on length and charge of the peptide and the new smORFs are shorter.. We have now included the proportion of smORFs identified by MS.

The authors, appropriately, consider the possibility that phased ribosome binding might not be sufficient to predict translation of a stable peptide. They describe a computational micropeptide detection method, micPDP, which uses evolutionary conservation for smORF detection. 63 smORFs were detected, of which 23 were also found by the ORFscore approach. This difference is attributed in part to lack of sequence conservation in some of the smORFs detected by ribosome foot printing. Detection by the two methods appears to provide a high confidence set of smORF prediction. Comparison of micPDP predictions with ORFscore analysis of previously published ribosome foot printing data also showed limited (albeit statistically significant) overlap between the two methods.

Overall the data are sound and the analysis supports the new peptide predictions as described. Understandably, the discussion aims to provide explanations that might reconcile the limited overlap between the methods. A few words discussing over- and under-prediction and the importance of experimental validation might not be amiss.

We appreciate the reviewer comments. We think that both approaches are parallel but support each other. Both methods use completely different features: while the ORFscore relies on the experimentally gathered ribosome profiling data, the micPDP result depends on conservation (reflecting selective pressure acting on that genomic region). Thus, it is hard to estimate over- or under estimation using a comparison between the methods. For some species such Human, the genome alignment should not be a big problem, compared to fish where the number of alignable genomes is limited. In the case of ribosome profiling, it is very difficult to define a true yet comparable negative set, as non-canonical translation has been discovered essentially all transcript regions (including UTRs). In the case of Ribosome profiling data, our data (fish) is extremely good with high codon resolution. While using ORFscore did detect translated regions using previously published ribosome foot printing data, depending on the quality and depth of ribosome profiling the results may vary. However, it must be noted that the experimental validation of the computationally predicted micropeptides based on conservation adds a strong layer of confidence to pursue the function of those selected smORFs in the near future. We have added the following text to the discussion: "Further, the larger number of species in the human genome alignment allowed us to score smORF coding potential more comprehensively than in zebra fish. Still, smORFs might have different codon usage and conservation patterns compared to canonical protein coding genes and may therefore be only incompletely captured by our computational pipeline. Also, smORFs encoding for lineage-specific or fast-evolving peptides, or smORFs with purely regulatory function, will be missed since our comparative method is tailored towards the identification of smORFs with conservation of their encoded amino acid sequence over larger evolutionary distances. The conceptual differences between experimental and computational approaches lend strong support to jointly identified smORFs but translate into an inability to use one method to estimate false positive or false negative rate of the other."

I have read the confidential reviews provided by the authors from another journal, and find no outstanding issues of substance that would preclude publication. The manuscript is suitable for publication.