Reply to evolutionary flux of canonical microRNAs and mirtrons in Drosophila

Jian Lu[1], **Yang Shen**[2], **Richard W Carthew**[3], **San Ming Wang**[4], and **Chung-I Wu**[1,2,5]

[1]Department of Ecology and Evolution, University of Chicago, Chicago, Illinois, USA

[2]State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-sen (Zhongshan) University, Guangzhou, China

[3]Department of Biochemistry, Molecular Biology, and Cell Biology, Northwestern University, Evanston, Illinois, USA

[4]Center for Functional Genomics, Division of Medical Genetics, Department of Medicine, Evanston Northwestern Healthcare Research Institute, Northwestern University Feinberg School of Medicine, Evanston, Illinois, USA

[5]CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, China

It has been known for some time that there are many weakly expressed and fast-evolving microRNAs (miRNAs)[1-6]. After analyzing more than 100,000 reads of small RNAs obtained from *Drosophila* heads, we suggested that most of these miRNAs were born and then died with the evolutionary dynamics of neutrally evolving sequences[7]. The birth and death rates of miRNAs were estimated to be about 12 and 11.7 genes per Myr, respectively, resulting in a fairly modest net gain of only 0.3 genes per Myr.

Berezikov *et al.* revised the estimate of net gain to be about 1 gene per Myr but did not provide an estimate of the birth and death rates separately[8]. Instead, they argued that our estimates for both the birth and death rates were too high by claiming that many of the miRNAs in our analysis were not miRNAs at all but were merely degraded products of mRNAs.

Their main concern was that many of the newly born miRNAs in our observation were singletons. In their expanded data, many of these singletons were missing and some were accompanied by sequences of similar length in the same hairpin, both of which suggested RNA degradation. Being mindful of the limitation of low coverage in our study[7], we have collected 18 million small RNA reads in *Drosophila* heads by sequencing of oligonucleotide ligation and detection (SOLiD) sequencing. We analyzed only reads that appeared at least 50 times in the arm of a hairpin and for which the accurate processing rate of the 5′ end of the miRNAs was 90%. The new dataset shows that the conclusion of Lu *et al.*[7] was not biased by low coverage, as suggested by Berezikov *et al.*[8].

Correspondence should be addressed to C.-I.W. (ciwu@chicago.edu)..

What, then, may be the reasons for the discrepancy between the analysis of Berezikov et al.[8] and our analysis[7]? Due to space limitations, we shall only give a brief account and will present a more thorough comparison on our website (http://pondside.uchicago.edu/wulab/microRNA/).

First, Berezikov et al.[8] defined miRNAs much more narrowly than we believe is reasonable. They did not consider hairpins on exons—a legitimate source of miRNAs, as many known miRNAs share sequences with exons[9-11]. In addition, they used criteria derived from conserved miRNAs to screen out candidates. These criteria include a much lower rate of evolution in the arms of the hairpin than in the loop and also the narrow size distribution of mature miRNAs. If miRNAs are defined by conservation, then there would be few that are evolutionarily transient.

Second, different sequencing platforms are known to yield fairly different results[2,12]. By sequencing platform, we mean more than just sequencing chemistry; rather, we include the entire protocols, from upstream library preparation to base callings, recommended by the manufacturers of Roche 454 GS-FLX, IlluminaGA or SOLiD-ABI DNA sequencers. Lu et al.[7], Berezikov et al.[8] and our new study (unpublished) used the 454, GA and SOLiD methods, respectively. Even the most abundant miRNAs were recorded with surprisingly large disparity by different methods, and the rare miRNAs were not detected by all methods. Our SOLiD datasets indeed contain many genes with multiple reads that were absent in the GA data Berezikov et al. used[8]. In another accompanying paper (Zhou et al., unpublished), we show that GA and SOLiD sequencing also perform very differently in SNP detection.

Third, Berezikov et al.[8] incorrectly used the argument of proportionality, which further confounded the interpretation of rare miRNAs. When there exists a large number of weakly expressed genes, a tenfold increase in coverage might increase the observed number of all rare genes tenfold, but each new gene discovery is likely to be different from the previously observed ones[13]. The same rare genes should not be expected to occur ten times as often, as Berezikov et al. asserted[8].

Fourth, the variation in miRNA expression within the same species should not be ignored. In humans, with GA sequencing, we have observed 300 conserved miRNAs that could be detected in fewer than 6 of the 12 human kidney libraries (Lu et al., unpublished data). Because the level of genetic variation in *D. melanogaster* is 20 times greater than that in humans, we expect the level of miRNA expression polymorphism to be at least as large. Berezikov et al.[8] and Lu et al.[7] used different fly strains in their surveys.

The account above shows that low-frequency small RNAs (including miRNAs) are often seen in some studies (such as Lu et al.[7] and our new SOLiD data) but not others (such as Berezikov et al.[8]). The hasty conclusion that RNA degradation must be at issue seems inappropriate.

Finally, we believe that the debate should shift to the rate of net gene gain (birth rate – death rate). After all, there is no disagreement about the conclusion that the transient genes, miRNAs or not, do not appear to be biologically important. We suspect that our original estimate of 0.3 genes per Myr[7] might be an underestimate. Because of the moderate number

of reads from a single platform (454), many miRNAs were not observed or were observed in one species only. As a result, birth rate might have been underestimated and death rate overestimated. We will provide a revised estimate on the basis of the new and expanded dataset elsewhere.

References

1. Berezikov E, et al. Nat. Genet. 2006; 38:1375–1377. [PubMed: 17072315]

2. Berezikov E, et al. Genome Res. 2006; 16:1289–1298. [PubMed: 16954537]

3. Bentwich I, et al. Nat. Genet. 2005; 37:766–770. [PubMed: 15965474]

4. Liang H, Li WH. Mol. Biol. Evol. 2009; 26:1195–1198. [PubMed: 19299536]

5. Kloosterman WP, et al. Nucleic Acids Res. 2006; 34:2558–2569. [PubMed: 16698962]

6. Fahlgren N, et al. PLoS One. 2007; 2:e219. [PubMed: 17299599]

7. Lu J, et al. Nat. Genet. 2008; 40:351–355. [PubMed: 18278047]

8. Berezikov E, Liu N, Flynt A, Lai CE. Nat. Genet. 2010; 42:6–9. [PubMed: 20037610]

9. Das S. Mol. Biol. Evol. 2009; 26:1179–1189. [PubMed: 19246621]

10. Kim VN, Han J, Siomi MC. Nat. Rev. Mol. Cell Biol. 2009; 10:126–139. [PubMed: 19165215]

11. Han J, et al. Cell. 2009; 136:75–84. [PubMed: 19135890]

12. Linsen SE, et al. Nat. Methods. 2009; 6:474–476. [PubMed: 19564845]

13. Balwierz P, et al. Genome Biol. 2009; 10:R79. [PubMed: 19624849]