differentiation in an abstract at the Cold Spring Harbor Conference on Dynamic Organization of Nuclear Function (2008). There is even older evidence, although not mapped to specific chromosomal locations, showing an increase in H3K9me2 in differentiated cells compared to undifferentiated ES cells[7]. Furthermore, Bing Ren and colleagues have confirmed our observation of large heterochromatin domains of hundreds of kilobases in size arising in differentiated ES cells from regions with bumps of a few kilobases in undifferentiated ES cells, albeit in human cells and with different heterochromatin markers (ref. 8 and B. Ren, personal communication). They also showed that partially methylated domains (PMDs), in which DNA are less methylated in fibroblasts compared to human ES cells[8], are enriched for expanded heterochromatin blocks in fibroblasts but not in ES cells. Interestingly, the LOCKs we defined in differentiated ES cells largely overlap the PMDs (**Supplementary Fig. 3**), even given that the mapping is cross-species.

Bo Wen[1,2], Hao Wu[1,3], Yoichi Shinkai[4], Rafael A Irizarry[1,3] & Andrew P Feinberg[1,2]

[1]Center for Epigenetics and [2]Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA. [3]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, USA. [4]Institute for Virus Research, Kyoto University, Sakyo-ku, Kyoto, Japan. Correspondence should be addressed to: A.P.F. (afeinberg@jhu.edu).

Note: Supplementary information is available on the Nature Genetics website.

1. Wen, B., Wu, H., Shinkai, Y., Irizarry, R.A. & Feinberg, A.P. *Nat. Genet.* **41**, 246–250 (2009).
2. Conley, B.J. *et al. Curr. Protoc. Cell Biol.* Chapter 23 (2005).
3. Roach, M.L. & McNeish, J.D. *Methods Mol. Biol.* **185**, 1–16 (2002).
4. *Stem Cells: Scientific Progress and Future Research Directions* (US Department of Health and Human Services, Washington, DC, USA, 2001).
5. Kalmar, T. *et al. PLoS Biol.* **7**, e1000149 (2009).
6. Guelen, L. *et al. Nature* **453**, 948–951 (2008).
7. Dai, B. & Rasmussen, T.P. *Stem Cells* **25**, 2567–2574 (2007).
8. Lister, R. *et al. Nature* **462**, 315–322 (2009).

© 2010 Nature America, Inc. All rights reserved.

Evolutionary flux of canonical microRNAs and mirtrons in *Drosophila*

To the Editor:

Next-generation sequencing technologies generate vast catalogs of short RNA sequences from which to mine microRNAs (miRNAs), which are ~21–24-nucleotide regulatory RNAs derived from RNase III–mediated cleavages of hairpin transcripts. However, such data must be vetted to appropriately categorize miRNA precursors and interpret their evolution. A recent study annotated hundreds of miRNAs in three *Drosophila* species on the basis of singleton reads of heterogeneous length[1]. Our multimillion-read datasets indicated that most of these putative miRNAs were not produced by RNase III cleavage and that they comprised many mRNA degradation fragments. We instead identified a distinct and smaller set of new miRNAs supported by high-confidence cloning signatures, which included a high proportion of evolutionarily nascent mirtrons. Our data support a much lower rate for the emergence of lineage-specific miRNAs than was previously inferred[1], with a net flux of ~1 miRNA per million years of drosophilid evolution.

Conserved miRNA genes are differentiated from bulk hairpins in that their terminal loops diverge more quickly than their stems[2]. However, species-specific miRNAs cannot be confidently identified by using solely computational methods, as hundreds of thousands of *Drosophila*[1,3–5] and human loci[6] are plausible as miRNA hairpins. Instead, we and others have turned to next-generation sequencing to identify recently evolved miRNAs, which lack support from evolutionary signatures (for example, **Supplementary Table 1**). Such deep sequence data often reveal heterogeneous size and read patterns with respect to predicted hairpins (**Fig. 1** and **Supplementary Fig. 1**), indicating that only a subset of hairpins with reads are substrates of Dicer-driven biogenesis pathways. In particular, it is not possible to determine whether a predicted hairpin associated with a single-cloned short RNA is indeed an endogenous substrate of RNase III cleavage (**Fig. 1**).

Lu and colleagues reported ~900 putative novel miRNAs sequenced from three *Drosophila* species—*D. melanogaster* (*Dme*), *D. simulans* (*Dsi*) and *D. pseudoobscura* (*Dps*)—including ~400 annotated under 'high-stringency' criteria[1]. They concluded that evolutionarily transient miRNA genes are continually born and lost, with only a small proportion of miRNAs fixed across drosophilid radiation. Inspection of these annotations showed that 35 *Dme*, 47 *Dsi* and 30 *Dps* 'novel' miRNAs corresponded to orthologs of 50 distinct genes whose cloning and evolutionary characteristics had been previously described[4,5,7] (miRBase 10.1 and **Supplementary Tables 2–4**). Another locus comprising multiple tandem hairpins corresponded to hairpin RNA hp-CG4068, which generates endogenous small interfering RNAs (endo-siRNAs)[8]. We sought to understand the nature of the remaining hundreds of miRNA candidates, whose abundant numbers were previously used to estimate a birthrate of ~12 miRNAs per Myr of drosophilid evolution[1].

We mapped ~15 million *Dme* reads from diverse developmental stages and tissues, including ~1 million from adult heads[4,9]. Compared to their frequency among ~16,000 reads from adult *Dme* heads[1], we expected our data to contain ~60-fold more reads for genuine miRNAs and likely more, given that many are expressed in multiple stages and tissues. This was true for the 35 *Dme* miRBase 10.1 loci designated 'novel' by Lu and colleagues[1]. These 'novel' loci were represented by 1,247 reads in their data (~34 reads per locus, although 6 loci were cloned only 2–3 times and 12 were singletons) but by ~320,000 reads in our data (~8,800 reads per locus). The remaining 23 non-miRBase loci were severely under-represented in our data, with 9 cloned 1–6 times and 9 that were not recovered at all (**Supplementary Table 2**).

For non-miRBase loci cloned in our dataset, the reads mapped incoherently across the predicted hairpin and/or adjacent genomic regions (**Fig. 1** and **Supplementary Fig. 1**). They also showed broadly heterogeneous sizes, contrasting with the restricted lengths of genuine *Drosophila* miRNAs (**Fig. 2**). Although some loci were conserved, the most abundant reads mapped to a ribosomal RNA (rRNA; *Lu-mir-2018*) and two small nuclear RNAs (snoRNAs; *Lu-mir-2324* and *Lu-mir-2213*); 16 out of the 20 remaining loci derived from mRNAs (**Supplementary Table 2**). Therefore, instances of conservation were attributable to protein-coding or functional RNA status and not to evol-

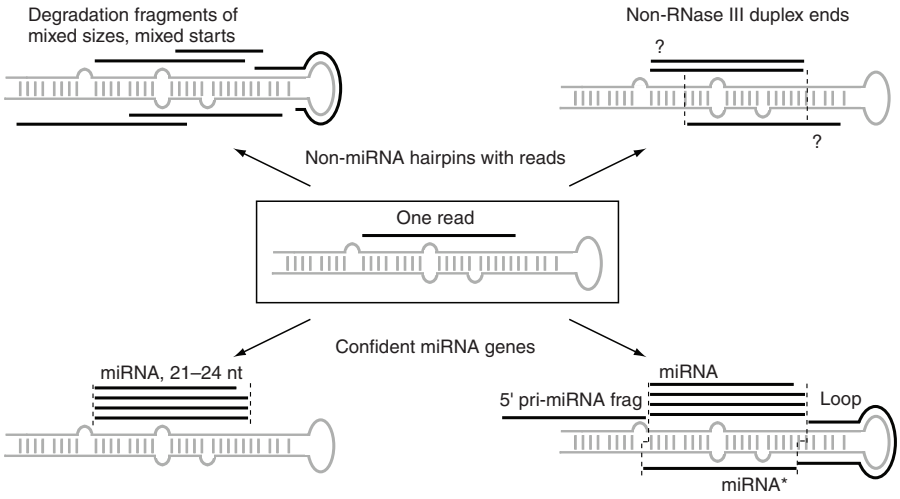

Figure 1 Putative miRNA loci annotated on the basis of single reads and plausible hairpin structures (center box) show distinct patterns when more reads are available. Reads may be distributed throughout the inferred hairpin, have heterogeneous sizes and/or pair as duplexes lacking 3′ overhangs (top left and right); reads with any of these characteristics cannot be annotated as miRNAs. High-confidence miRNAs have multiple cloned 21–24 nucleotide reads with relatively fixed 5′ ends (bottom left). With sufficient sequencing, it is usually possible to identify the duplex partner miRNA* species, as well as other byproducts of miRNA biogenesis such as terminal loops or species flanking the pre-miRNA hairpin (bottom right).

utionary dynamics characteristic of genuine miRNAs (**Supplementary Fig. 1a**,**b**). Similar analysis revealed that hundreds of new *Dsi* and *Dps* miRNA candidates[1] mapped to syntenic exons of *Dme* protein-coding transcripts (**Supplementary Tables 3–6**), with reads spanning the 18–28-nucleotide window used for cloning (**Fig. 2**). We conclude that the prior miRNA annotations[1] included a high proportion of RNA fragments derived from the degradation of diverse mRNAs and some noncoding RNAs (ncRNAs).

We therefore wished to gauge miRNA flux using independent small-RNA data. We and others annotated 147 miRNA loci (including 14 mirtrons) from ~1 million *Dme* reads[4,5,7], but >17 million additional reads[9,10] yielded only 14 new miRNA loci and the high-confidence antisense locus *Dme-mir-307-as* (**Supplementary Tables 7** and **8**). Because of this sequencing depth, we could assign confident miRNA cloning patterns to novel loci, and most had star reads despite their evolutionary transience (**Supplementary Figs. 1c** and **2**). Curiously, 5 out of 14 were mirtrons, a high proportion consistent with the hypothesis that mirtrons generally evolve more quickly than canonical miRNAs[11,12]. Four miRBase loci that did not meet confident read criteria are discussed in the **Supplementary Text** and **Supplementary Figure 3**.

We next mapped 3,712,683 and 3,318,524 small RNAs from mixed embryos of *Dsi* and *Dps*, respectively, and 3,442,645 reads from adult *Dps* heads (**Supplementary Table 1**). These data comprise 50–270 times the data earlier used to estimate miRNA diversity[1] and provided an appropriate basis for annotating the miRNAs in these other species without needing to consider their evolutionary features. Our datasets contained abundant reads for previously rare or uncloned *Dsi* and *Dps* orthologs of miRBase 10.1 loci (**Supplementary Tables 9–12**), consistent with the expectation that genuine miRNAs are recovered proportionally to sequencing depth. These reads yielded 11 new *Dsi* miRNAs, including 5 mirtrons (2 of which were orthologous to novel *Dme* mirtrons *mir-2489* and *mir-2494*) and >88 distinct novel *Dps* miRNAs, including 17 mirtrons (**Supplementary Tables 9–12** and **Supplementary Figs. 4** and **5**; see also **Supplementary Text** for discussion of potentially duplicate *Dps* loci). Among these, the overlap with the annotations of Lu and colleagues was minimal: only 4 out of 261 *Dsi* loci and 19 out of 598 *Dps* loci[1] overlapped between their annotations and ours. Conversely, nearly 300 of their reported *Dsi* and *Dps* miRNAs had 0 reads in our data, and ~100 had fewer than 5 reads (**Supplementary Tables 3** and **4**). Therefore, deep sequencing failed to validate most of the previously reported miRNAs[1], and the minimal overlap in annotated loci highlights that the differences were not due to the application of more 'conservative' compared to more 'lenient' cutoffs to a common set of hairpins.

Although the rates of miRNA flux amongst different species of *Drosophila* might be expected to be reasonably similar, Lu and colleagues annotated vastly different numbers of species-specific miRNAs in *Dme*, *Dsi* and *Dps*[1]. This does not seem likely to be a consequence of their different sampling depths in these species, as all of their datasets were smaller by a factor of 100 than those analyzed in the present study. Our annotations from

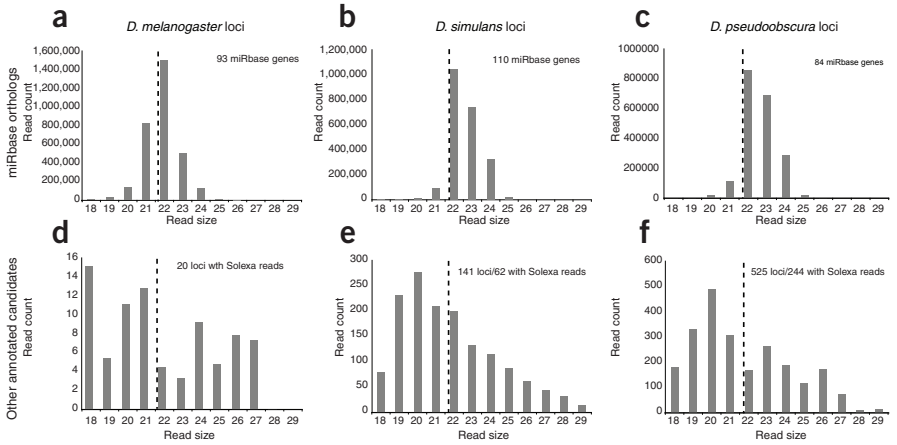

Figure 2 Size comparison of miRBase *Dme*, *Dsi* and *Dps* miRNAs and other miRNA candidates annotated by Lu and colleagues[1]. (**a**–**f**) We used Solexa data from diverse *Dme* samples and *Dsi* or *Dps* embryos to assess the distribution of read sizes from annotated loci that were orthologous to miRBase 10.1 genes (**a**–**c**) or lacked miRBase orthologs (**d**–**f**). The top panels indicate that genuine *Drosophila* miRNAs produce a characteristic range of 21–24-nucleotide RNAs, with preference for 22 nucleotides (dashed reference lines). The other candidate miRNAs, nearly all of which were annotated on the basis of single reads[1], showed broadly heterogeneous sizes in our larger datasets; note that we did not recover any reads for many of these loci.

© 2010 Nature America, Inc. All rights reserved.

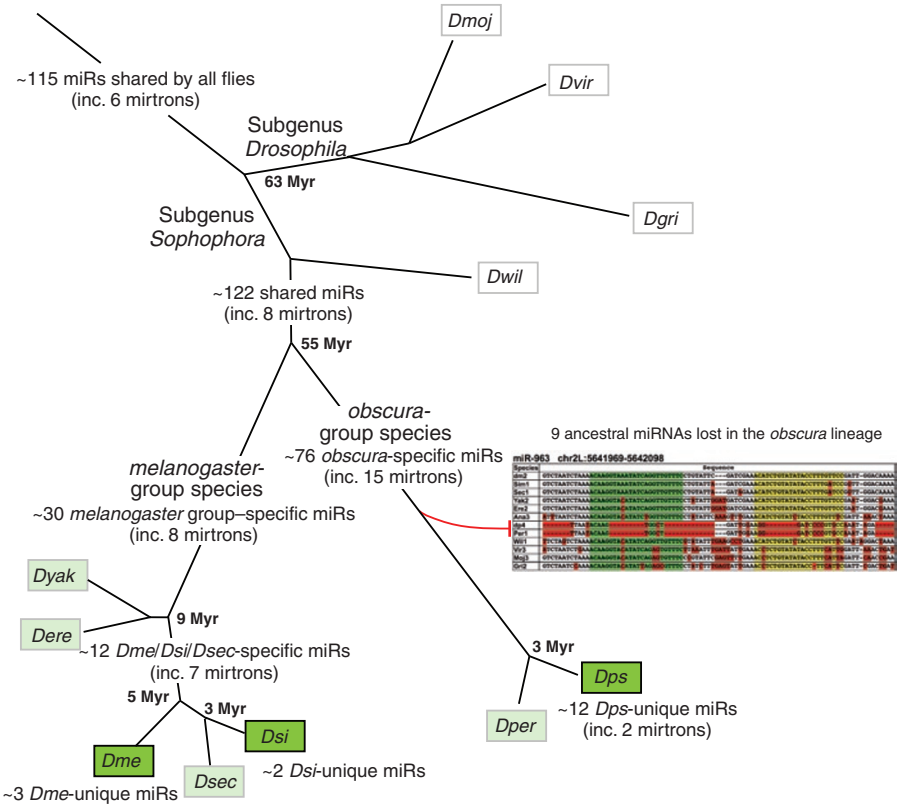

Figure 3 Flux of drosophilid miRNA genes assessed using multimillion-read datasets in three species. Small RNAs were cloned from the species in dark green; detailed orthology of novel miRNAs annotated in this study was determined with respect to species, shown in light green. Because not all loci are necessarily present in all of the species in a given branch, some values are designated as approximate. For example, the *Dps* and *Dper* genomes coordinately lack orthologs of nine miRNA genes present in the sophophoran and/or proto-drosophilid ancestor (**Supplementary Fig. 7**); these orthologs are considered to have died in the *obscura* lineage. Among the dozen *Dme*- or *Dsi*-cloned miRNAs for which aligning sequences were found only in their closest sister species, only a few have cloned small-RNA evidence from multiple species thus far (for example, the highly species-restricted miR-2489 was cloned from both *Dme* and *Dsi*). We do not exclude that some of these miRNAs may actually prove to be unique to a single species. Note that mirtrons comprise a small fraction of the deeply conserved set of miRNAs, but they comprise a much higher fraction of lineage-restricted miRNAs in various drosophilid genomes.

multimillion-read datasets instead yielded numbers of new genes that were consistent with the relative ancestries of these species. We recovered few new miRNAs in the highly related *Dme* and *Dsi* sister species but many more in the distant *Dps* species (Fig. 3); most newly identified *Dps* genes were conserved only in its related sister *D. persimilis* (*Dper*). The overall flux in the miRNA repertoire was consistent: 45–47 miRNAs cloned from *Dme* or *Dsi* have no *obscura*-group homologs, whereas 88 miRNAs were cloned from *Dps* for which no *melanogaster*-group homologs exist. Assuming ~55 Myr of divergence between these clades as before[1], this puts the rate of drosophilid miRNA flux at 0.82–1.6 genes per Myr, far less than the ~12 genes per Myr earlier proposed[1]. Notably, the tally of species-restricted mirtrons relative to canonical miRNAs was disproportionately high in all three species (**Fig. 3** and **Supplementary Figs. 2**, **4** and **5**). Therefore, mirtrons and canonical miRNAs show distinct evolutionary dynamics for emergence and fixation, even though they generate functionally identical regulatory RNAs.

The net rate of miRNA flux is a combination of genes born and genes lost, but distinguishing birth from death is challenging. For example, the ~70 miRNAs shared by *Dps* and *Dper* for which no orthologs exist in any *melanogaster*-group genomes might have been 'born' in the ancestor to the *obscura* lineage or 'died' in the ancestor to the *melanogaster* lineage (**Fig. 3**). In addition, the poorer state of the *Dsi* genome assembly obfuscates whether it truly lost some genes (nine pan-drosophilid miRNAs have gaps or errors in *DroSim1*, **Supplementary Fig. 6**). However, we could confidently judge that nine miRNAs distributed in four operons died in the *obscura* group, as they were ancestrally conserved but absent from both *Dps* and *Dper* (Fig. 3 and **Supplementary Fig. 7**). Conversely, the small number of *Dme*, *Dsi* and *Dps* miRNAs lacking aligned sequences in any other sequenced species are good candidates for 'newly born' miRNAs. Their identification supports the concept that substrates occasionally arise *de novo* from neutral evolution of transcripts with hairpin character[1].

Nascent miRNAs might have cleavage registers that are more imprecise than those for well-conserved miRNAs, but the biogenesis of miRNAs via RNAse III enzymes indicates that duplexes of appropriate size should be cloned with sufficient sequencing, as observed in our data (**Figs. 1** and **2** and **Supplementary Figs. 1–5**). Similar to what was done in previous analyses[1], we assigned singleton reads to hundreds of candidate hairpins (see URL section), and these loci evolved neutrally with respect to hairpin character. However, as few of these loci are likely to be bona fide substrates for Dicer-driven miRNA biogenesis (**Fig. 1**), their evolution is not generally germane to the evolution of genuine miRNAs.

In principle, there may exist hairpin loci that mostly generate short species via generic RNA catabolism, but for which a fraction of reads derive from RNase III cleavages. The evolutionary dynamics of this population should prove relevant for understanding the birth of miRNA genes. However, experimental evidence beyond deep sequencing is necessary to unequivocally demonstrate their processing by Drosha and Dicer. Because the majority of animal euchromatin is actively transcribed[13,14], deep sequencing is expected to recover small RNAs constituting degradation fragments from many incidental hairpins. This is the case even when using protocols that select for 5′ phosphates (and presumably against degradation fragments) because endogenous kinases can phosphorylate arbitrary short RNAs[15]. The existence of exceptionally diverse populations of Piwi-interacting RNAs (piRNAs) and endo-siRNAs[6] further highlights the fact that non-miRNA reads can be abundant in total RNA libraries. In conclusion, confident annotation of miRNAs from deep sequence yields unified rates of canonical miRNA and mirtron evolution among the drosophilids and provides evidence for only a limited set of species-specific miRNAs in this genus.

Eugene Berezikov[1,2,5], Na Liu[3,5], Alex S Flynt[3],

© 2010 Nature America, Inc. All rights reserved.

Emily Hodges[4], Michelle Rooks[4], Gregory J Hannon[4] & Eric C Lai[3]

[1]Hubrecht Institute, Royal Netherlands Academy of Arts and Sciences & University Medical Center Utrecht, Utrecht, The Netherlands. [2]InteRNA Genomics, Bilthoven, The Netherlands. [3]Sloan-Kettering Institute, Department of Developmental Biology, New York, New York, USA. [4]Howard Hughes Medical Institute, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, USA. [5]These authors contributed equally to this work. Correspondence should be addressed to E.B. (e.berezikov@hubrecht.eu) or E.C.L. (laie@mskcc.org).

Note: Supplementary information is available on the Nature Genetics website.

Accession numbers. The three small-RNA datasets from *Dsi* embryos and *Dps* embryos and heads were submitted to NCBI GEO under series GSE13677.

URLs. Additional supplementary supporting material is available at http://www.internagenomics.com/public/dros0811.

ACKNOWLEDGEMENTS

K. Okamura assisted with library amplification. *Dsi* and *Dps* Solexa sequencing was performed at the BC Genome Sciences Centre. This work was supported by VIDI grant and the European Commission Sixth Framework Programme Integrated Project SIROCCO (LSHG-CT-2006-037900) to E.B., Howard Hughes Medical Institute support to G.J.H., and grants from the V Foundation for Cancer Research, the Sidney Kimmel Cancer Foundation, the Alfred Bressler Scholars Fund and the US National Institutes of Health (R01-GM083300 and U01-HG004261) to E.C.L.

AUTHOR CONTRIBUTIONS

The study was designed by E.C.L. and E.B. A.S.F. and E.C.L. prepared small-RNA libraries. E.H., M.R. and G.J.H. performed Dme sequencing. N.L. and E.B. analyzed small-RNA data. E.C.L. and E.B. wrote the paper.

COMPETING INTERESTS STATEMENT

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/naturegenetics/.

1. Lu, J. *et al. Nat. Genet.* **40**, 351–355 (2008).
2. Lai, E.C. *Curr. Biol.* **13**, R925–R936 (2003).
3. Lai, E.C., Tomancak, P., Williams, R.W. & Rubin, G.M. *Genome Biol.* **4**, R42.1–R42.20 (2003).
4. Ruby, J.G. *et al. Genome Res.* **17**, 1850–1864 (2007).
5. Stark, A. *et al. Genome Res.* **17**, 1865–1879 (2007).
6. Bentwich, I. *et al. Nat. Genet.* **37**, 766–770 (2005).
7. Sandmann, T. & Cohen, S.M. *PLoS One* **2**, e1265 (2007).
8. Okamura, K. & Lai, E.C. *Nat. Rev. Mol. Cell Biol.* **9**, 673–678 (2008).
9. Chung, W.J., Okamura, K., Martin, R. & Lai, E.C. *Curr. Biol.* **18**, 795–802 (2008).
10. Seitz, H., Ghildiyal, M. & Zamore, P.D. *Curr. Biol.* **18**, 147–151 (2008).
11. Okamura, K. *et al. Cell* **130**, 89–100 (2007).
12. Ruby, J.G., Jan, C.H. & Bartel, D.P. *Nature* **448**, 83–86 (2007).
13. Manak, J.R. *et al. Nat. Genet.* **38**, 1151–1158 (2006).
14. Kapranov, P. *et al. Science* **316**, 1484–1488 (2007).
15. Aravin, A.A. *et al. Dev. Cell* **5**, 337–350 (2003).

© 2010 Nature America, Inc. All rights reserved.

Lu *et al.* reply:

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].

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 conclu-