Supplementary Information

The miRCandRef bioinformatics pipeline. The algorithm MiRCandRef was developed for assembling high-quality filtered next-generation sequence (NGS) data into relatively short genomic contigs (crystal-contigs) that match predicted miRNA candidate loci. In brief, miRCandRef includes 4 steps (Supplementary Figure 1). Initially, total genomic DNA reads are blasted against a smallRNA dataset to identify all genomic NGS reads that match miRNA-candidate sequences. The second step executes the perl script "*get_miRNA_reads_and_mates.pl*" and creates fasta files for all genomic reads and their mates (optional: no mates) that match with 100% sequence identity the full length query-smallRNA-sequences (or its reverse complement). Subsequently, the perl-script "*run_velvet.pl*" automatically executes a velvet (Zerbino and Birney 2008) assembly for each generated fasta-file. Parameters such as e.g., k-mer size, insertion size, and expected coverage can be adjusted to particular needs. The assembled contigs for each fasta file will be compared for their respective lengths. The longest contigs will be kept and included in a multi-fasta file. Finally, the obtained contigs will be further processed by a bash-executable script that utilizes custom scripts and the fastx-toolkit to format the contigs so they can be used as input for miRDeep2*, a modified version of miRDeep2 (Friedländer, et al. 2011), and for clustering the redundant sequences using CD-Hit (Fu, et al. 2012). The final product of miRCandRef is a multi-fasta file of assembled and clustered miRNA crystal-contigs that can be used as a reference for miRNA prediction software like miRDeep2 instead of a properly assembled genome.

Modifications to miRDeep2. In the current approach, we integrated miRDeep2, the state-of-the-art tool for predicting miRNA from smallRNA NGS data, and reference genomes (Friedländer, et al. 2011). In contrast, miRCandRef delivers non-standard references from redundant assemblies. The reasoning for obtaining these assemblies is that miRNA loci include at least 3 different smallRNA species; these are the mature miRNA, the star, and the loop sequences (Supplementary Figure 2) that are usually represented in different NGS reads. Each of these NGS reads will serve as starting point for read assemblies that might differ in their length. The resulting redundancy hampers a comprehensive and accurate miRNA prediction by miRDeep2, because mapping of smallRNA reads to multiple queries is restricted to five. Thus, miRDeep2 could only be integrated into the bioinformatics pipeline used in this study after slight changes to the source code of miRDeep2. The number of independent hits was set to 2,000 (sub *parse_mappings* -i 2000). The upper length limit for precursors to be excised and tested by miRDeep2 was reset to 100 nucleotides. This was necessary because miRDeep2 excises precursor candidates outside a -20 and +70 nucleotide window around a read stack. However, significantly longer hairpins have been described (Hendrix, et al. 2010; Ruby, et al. 2007) and were also detected in *G. salaris* (Supplementary File 1). For such loci, the default miRDeep2 would not test for folding and mapping to the typical -2nt offset required for proper Dicer-function.

Processing of miRDeep2* results/criteria for miRNA prediction. For *G. salaris* miRDeep2* automatically identified known miRNAs based on a reference file that contained all miRNAs from *Schmidtea mediterranea*, *Schistosoma japonicum*, and *Echinococcus granulosus* included in miRBase. Predicted miRDeep2* miRNA loci were processed in order to summarize redundant predictions using custom perl scripts. All unique hairpins were assessed if they were represented by at least 10 smallRNA reads, at least 1 miRNA* read, and if they consisted of proper hairpins with a 2nt mature/star-sequence offset, as required by the protein Dicer.

Friedländer MR, Mackowiak SD, Li N, Chen W, Rajewsky N 2011. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. Nucleic acids research. doi: 10.1093/nar/gkr688

Fu L, Niu B, Zhu Z, Wu S, Li W 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28: 3150-3152. doi: 10.1093/bioinformatics/bts565

Hendrix D, Levine M, Shi W 2010. miRTRAP, a computational method for the systematic identification of miRNAs from high throughput sequencing data. Genome Biol 11: R39. doi: 10.1186/gb-2010-11-4-r39

Ruby JG, Stark A, Johnston WK, Kellis M, Bartel DP, Lai EC 2007. Evolution, biogenesis, expression, and target predictions of a substantially expanded set of Drosophila microRNAs. Genome Res 17: 1850-1864. doi: 10.1101/gr.6597907

Zerbino DR, Birney E 2008. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Research 18: 821-829. doi: DOI 10.1101/gr.074492.107