Supplementary Figures

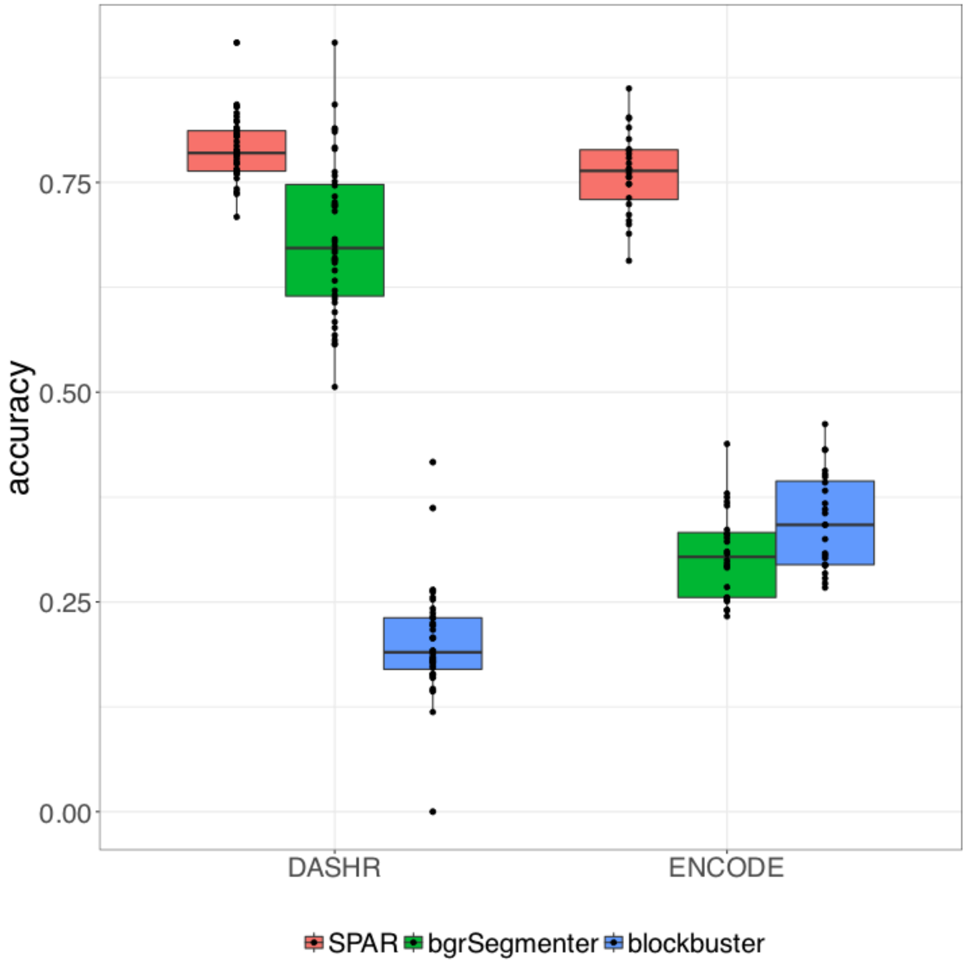

Figure S1. Small RNA loci detection accuracy on DASHR and ENCODE datasets for SPAR and other methods (bgrSegmenter, blockbuster). Accuracy of calling miRNA loci is estimated using mature miRNA annotations from mirBase as the reference set. For each DASHR and ENCODE dataset, accuracy is computed as [#(correct 5’ start sites) / #(expressed miRNAs)].

Supplementary Methods

Annotation resources

The annotation information for miRNAs is based on miRBase (v19 for GRCh19/hg19, v21 for GRCh38/hg38 and GRCm38/mm10 genomes) (Kozomara & Griffiths-Jones, 2014); snRNA, snoRNA, scRNA and rRNA annotations are from UCSC Genome Browser (Tyner et al., 2017) and GENCODE (Harrow et al., 2012); tRNA information is based on tRNAdb (Jühling et al., 2009); and piRNA annotation for both human and mouse is derived from NCBI (Pruitt, Tatusova, Klimke, & Maglott, 2009) piRNA sequences by mapping to reference genomes. The current SPAR annotation also includes tRNA fragment (tRF) annotations (Leung et al., 2016) that are created based on 5p and 3p 50 nt sequences upstream and downstream of the known tRNA genes as well as obtained from tRFdb (Kumar, Mudunuri, Anaya, & Dutta, 2015). Long non-coding RNA annotations are obtained from LNCipedia 4.1 (Volders et al., 2015) for human genome and NONCODE v5.0 (Zhao et al., 2016) for mouth genome. Annotations for mRNA genes and repeat elements are obtained from UCSC Genome Browser (Tyner et al., 2017) (knownGene, kgXref, RMSK tables).

SPAR pipeline

The SPAR pipeline involves the following steps:

1. Call peaks corresponding to smRNA sites (*ab initio* annotation-free segmentation)
2. Annotate called peaks by using integrated RNA annotation from DASHR
3. Construct genome-wide tracks with called peaks and raw read coverage information
4. Integrate called peaks with DASHR and ENCODE reference expression information
5. Conservation, genomic location, sequence analyses

SPAR peak caller

The inputs to the algorithm are mapped reads from the small RNA-seq experiment (BAM) or genome-wide read coverage profile (BigWig). The parameters are the minimum peak height (10 reads by default), and the minimum fold change in read depth (2 by default) for detecting peak starts. The minimum peak width / read length (15 nts by default) and the maximum peak width / read length (44 nts by default) are other optional parameters that can be used to select the desired RNA length range for analysis.

SPAR peak calling algorithm (Leung et al., 2016) identifies peaks with evidence of specific processing patterns, i.e. mature RNA products (e.g., low 5p read entropy). To do this, the peak calling algorithm scans the genomic sequence and identifies the start of the peak by finding two adjacent positions with at least a 2-fold increase in the number of mapped reads. Similarly, the corresponding end of the peak is found by looking for adjacent positions with at least a 2-fold decrease in the number of mapped reads. Additionally, the detected peaks need to have at least 10 reads by default.

After identifying peaks (the mature sncRNA locations), we then quantified the number of reads falling within these regions as expression for each sncRNA. The raw expression values provided in the SPAR output are weighted by the number of hits as 1/h, where h is the number of places that the read has been aligned to.

To enable comparison across tissues, we take into account the library size information for each of the sequencing experiments and report the read count in 'reads per million' (RPM), since this is the most commonly used normalization method to account for differences in library size across different experiments (Anders et al., 2013).

Annotation of mature RNA products

The peaks with evidence of specific processing identified in the previous step are overlapped with SPAR annotations ((Leung et al., 2016); also see Annotation resources).
Each peak is assigned to its mature sncRNA product class (mir-3p, mir-5p, etc) or precursor sncRNA gene (miRNAprimary, snRNA, snoRNA, etc).
Annotation algorithm uses hierarchical/prioritized assignment starting from annotated mature products (mir-3p, mir-5p, mir-5p3pno, tRFs) and snRNA, snoRNA genes followed by miRNA primary genes, piRNAs, and rRNA genes. Remaining un-assigned loci (i.e. loci without any overlaps with annotated sncRNA genes or mature products) are classified into un-annotated category.
For all sncRNA loci, SPAR computes multiple features describing 1) processing / cleavage patterns; 2) co-localization with non-small RNA genes and other genomic elements such as repeat elements, promoters, exons, introns, lncRNA exons, lncRNA introns; and 3) evolutionary conservation scores derived from the UCSC phastCons conservation tracks for human and mouse(Siepel et al., 2005; Tyner et al., 2017).

Conservation computation The evolutionary conservation scores were derived from the UCSC 100-way phastCons conservation track (Siepel et al., 2005; Tyner et al., 2017). We computed conservation for each sncRNA locus as a mean of per-nucleotide conservation values along the locus, $\frac{1}{n}\sum_i c_i$, where c_i is conservation score (phastCons probability) for position i, or 0 if missing. The resulting mean conservation score, as well as minimum and maximum per-nucleotide conservation values along the locus are reported by SPAR for each sncRNA locus.

Specific processing computation For each sncRNA locus, specificity of 5' RNA cleavage is computed as

$$(1 - H^{5'})$$

where $H^{5'} = -\sum_{1 \leq a \leq n_{5'}} p_a \log_2 p_a / \log_2 n_{5'}$, normalized entropy of 5' reads ends across $n_{5'}$ alternative 5' positions, and p_a is proportion of reads with 5' ends at position a. Values of $(1 - H^{5'})$ closer to 1 are indicative of specific processing/cleavage as opposed to random, non-specific processing. Additionally, the position of most common 5' read end among the reads mapping to the locus, and relative proportion of the reads with the 5' end coinciding with that position are reported for each sncRNA locus by SPAR. Similarly, 3' specificity, 3' most common cleavage position, and proportion are computed and reported for each sncRNA locus.

SPAR evaluation

We systematically evaluated the accuracy of SPAR in detecting smRNA sites using ENCODE and DASHR smRNA-seq datasets. To compute accuracy, we used miRNA loci from mirBase as the reference set. We also analyzed the accuracy of the baseline peak detectors (bgrSegmenter (Habegger et al., 2011), blockbuster (Langenberger et al., 2009)) on the same data.

References

Anders, S., McCarthy, D. J., Chen, Y. S., Okoniewski, M., Smyth, G. K., Huber, W., & Robinson, M. D. (2013). Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nature Protocols*, *8*(9), 1765–1786. https://doi.org/10.1038/nprot.2013.099

Habegger, L., Sboner, A., Gianoulis, T. A., Rozowsky, J., Agarwal, A., Snyder, M., & Gerstein, M. (2011). RSEQtools: a modular framework to analyze RNA-Seq data using compact, anonymized data summaries. *Bioinformatics*, *27*(2), 281–283. https://doi.org/10.1093/bioinformatics/btq643

Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., … others. (2012). *GENCODE: the reference human genome annotation for {The} {ENCODE} {Project}*. Genome Res.

Jühling, F., Mörl, M., Hartmann, R. K., Sprinzl, M., Stadler, P. F., & Pütz, J. (2009). *tRNAdb 2009: compilation of {tRNA} sequences and {tRNA}, genes*. Nucleic Acids Res.

Kozomara, A., & Griffiths-Jones, S. (2014). MiRBase: Annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Research*, *42*(D1). https://doi.org/10.1093/nar/gkt1181

Kumar, P., Mudunuri, S. B., Anaya, J., & Dutta, A. (2015). tRFdb: A database for transfer RNA fragments. *Nucleic Acids Research*, *43*(D1), D141–D145. https://doi.org/10.1093/nar/gku1138

Langenberger, D., Bermudez-Santana, C., Hertel, J., Hoffmann, S., Khaitovich, P., & Stadler, P. F. (2009). Evidence for human microRNA-offset RNAs in small RNA sequencing data. *Bioinformatics*, *25*(18), 2298–2301. https://doi.org/10.1093/bioinformatics/btp419

Leung, Y. Y., Kuksa, P. P., Amlie-Wolf, A., Valladares, O., Ungar, L. H., Kannan, S., … Wang, L.-S. (2016). DASHR: Database of Small human noncoding RNAs. *Nucleic Acids Research*, *44*(D1). https://doi.org/10.1093/nar/gkv1188

Pruitt, K. D., Tatusova, T., Klimke, W., & Maglott, D. R. (2009). NCBI reference sequences: Current status, policy and new initiatives. *Nucleic Acids Research*, *37*(SUPPL. 1). https://doi.org/10.1093/nar/gkn721

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., … Haussler, D. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research*, *15*(8), 1034–1050. https://doi.org/10.1101/gr.3715005

Tyner, C., Barber, G. P., Casper, J., Clawson, H., Diekhans, M., Eisenhart, C., … James Kent, W. (2017). The UCSC Genome Browser database: 2017 update. *Nucleic Acids Research*, *45*(D1), D626–D634. https://doi.org/10.1093/nar/gkw1134

Volders, P. J., Verheggen, K., Menschaert, G., Vandepoele, K., Martens, L., Vandesompele, J., & Mestdagh, P. (2015). An update on LNCipedia: A database for annotated human lncRNA sequences. *Nucleic Acids Research*, *43*(D1), D174–D180. https://doi.org/10.1093/nar/gku1060

Zhao, Y., Li, H., Fang, S., Kang, Y., wu, W., Hao, Y., … Chen, R. (2016). NONCODE 2016: an informative and valuable data source of long non-coding RNAs. *Nucleic Acids Research*, *44*(D1), D203–D208. https://doi.org/10.1093/nar/gkv1252