# **The flexible sRNA alignment pipeline**

**Rationale**

There is no standard sRNA-seq alignment and assignment pipeline. Key issues with sRNA-seq data compared with mRNA-seq:

* The short nature of miRNAs and sequence homology within a miRNA family means multi-mapping sequences are common (Figure 1); this is especially true if they arise from repetitive sequences.
* miRNAs can be found within sequences of other genes; for example, mirtrons lie within introns.
* isomiRs and nucleotide editing of miRNAs mean miRNA sequences can mismatch to the canonical miRNA sequence listed in miRbase.

|  |
| --- |
|  |

Figure 1 The overlap between different classes of RNAs at the sRNA-df level in a clinical dataset. Many sequences multimap to multiple RNAs. A. The proportion of sequences within an RNA class that do not map to another class. B. Upset plot, showing overlaps at a finer level.

### Considerations for an sRNA alignment and assignment pipeline:

#### Alignment:

Two approaches can be used: align to the genome and then assign using a program such as feature counts and annotation files (e.g. GTF files), which may or may not be restricted to the sRNAome. Alternatively, align to the sRNAome directly.

#### Dealing with multi-mapping reads mapping equally well to several loci

General reference for this section is Deschamps-Francoeur et al1. Options for dealing with multi-mapping reads, along with advantages and disadvantages, are summarised in Table 1.

*Should reads mapping to different RNA classes be dealt with differently?*

* Yes—miRNAs should be prioritised as they are biologically functional units with more evidence for existence compared with other putative sRNA-dfs (which could represent degradation products) and can exist within other genes: solution—use a hierarchical method to prioritise assignments based on full-length RNA species
* No—in the absence of certainty, assumptions should not be made about which full-length RNA a sequence derives from, particularly reads that map to non-miRNA loci.

Table 1 Options for dealing with multi-mapping reads

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Option** | **Explanation** | **Programs implementing this approach** | **Advantages** | **Disadvantages** |
| Ignore | Remove multi-mapping reads | Option in algorithms within star and HTSeq-count | Simple | - Not appropriate for sRNA seq data as multi-mapping sequences comprise a significant proportion of reads, and some features are inherently multi-mapping; therefore, they would be completely discarded |
| abundance-based approach | Assign to the most abundant read | A similar approach is implemented in Cufflinks2 | - Removes redundancy, meaning fewer statistical tests are done, potentially improving power in differential expression analyses.  - Based on Bayesian thinking—if there are genes, one is very abundant, and the other is less abundant, then an ambiguously mapping read is more likely to come from the abundant reads. | - An assumption is made about where that reads maps to  - Alternative loci are not represented in the final expression matrix |

Table 2 (continued)

2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| fractional approach | Assign reads to all possible loci but divide the frequency of each read by the number of times it maps. | Commonly used method implemented in the feature counts in the subread package3. | - This does not assume mapping and allows all possible loci to be represented in the final expression table. | - Many more features are present in the final dataset than were probably sequenced, exacerbating the multiple testing issue  - There is redundant information between genes that share reads. This leads to excessive tests, worsening the multiple testing issue with potentially little gain, particularly if none of those loci is differentially expressed (as it does not matter where their sequences map).  - If a gene is differentially expressed, it is not immediately clear whether it is genuinely differentially expressed or whether reads from another locus could drive differential expression.  - Count data are not integer, which some differential expression packages require  - odd dispersion values can arise, which some differential expression packages need to model |

Table 3 (continued)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| sequence level approach | Perform gene expression at the sequence level |  | - It removes the ambiguity around loci mapping. This can work in sRNA-seq data because the whole RNA is sequenced, meaning identical sequences can be picked up multiple times. This could not be implemented for mRNA data because random, unique, partial fragments of mRNA genes are sequenced.  - It enables differential expression at the isomiR / sRNA "iso-feature" level, which could hold biological meaning. | - Excessive features included in the differential expression analysis—this causes issues with computational time and memory. It can exacerbate the multiple testing problem without bringing many benefits as there can be redundancy between isomiRs/ sequences from the same gene as their expression may be highly correlated |
| gene\_union approach  (sometimes referred to as a gene cluster approach) | A sequence, which maps to several genes, e.g., gene A & gene B, gets assigned to a new feature called "geneA\_gene B" | Similar approach used in seqCluster4 and mmQuant5.  This approach is rarely used in RNA sequencing but common in mass spec proteomics data processing pipelines, e.g. Max Quant software | - All potential loci are represented in the gene expression table. Feature names retain information on mapping certainty, so if feature A\_B is differentially expressed, it is clear that this sequence could arise from either or both genes.  - It may reduce the number of features tested compared with the fractional approach.  - It allows greater specificity about, which loci may be differentially expressed without resorting to differential expression at the sequence level.  - Potentially a compromise between the specificity of the sequence level approach, the reduction in redundancy of the abundance-based approach, and feature retainment of the fractional approach. |  |

### Final pipeline

There is no standard (or perfect) way to deal with multi-mapping reads. This can lead researchers to remap sRNA expression data using multiple approaches, which is time-consuming. A new, flexible way of mapping sRNAs was therefore created. This algorithm aimed to:

* Assign reads using the main assignment options within one programme so the user can easily compare the outputs without remapping.
* Assign miRNA sequences to mature reads rather than hairpins if the read only extended slightly into the hairpin region.

Figure 2 contains an overview of the final alignment pipeline. Adapter-trimmed FastQ files are input to the pipeline, and files from the same samples are merged. Short sequences and low-complexity sequences are discarded using fqtrim6. Next, reads are mapped to pre-miRNAs and mature miRNAs with 2MM allowed. In the case of mature miRNAs, soft clipping is permitted at either end, allowing reads that extend beyond the canonical mature miRNA sequence to map to the mature miRNA.

|  |
| --- |
| **Diagram  Description automatically generated** |

Figure 2 Overview of alignment and assignment pipeline.

Next, all reads are mapped to the sRNAome, including pre-miRs again. The sRNAome bowtie indexes were built by merging FASTA files from:

miRbase7 (miRNAs)

SnOPY8 (snoRNAs)

trRNA\_db9 (tRNAs)

RefSeq10 (snRNAs, yRNAs, vault RNAs, lnc RNAs, ribosomal RNA)

These databases were chosen empirically because they assigned the most reads using a clinical dataset. piRNAs are excluded because no high-quality piRNA repositories exist. miscRNAs are excluded as they are a superset of more specific classes (Figure 1).3

Sequences are mapped to the sRNAome using Bowtie11. Up to 2 mismatches are allowed, and the best-strata option is implemented. The best-strata option only retains assignment(s) with the least number of mismatches. Unaligned sequences are then mapped using Bowtie2, which allows gapped alignment12. Bowtie2 does not have a best-strata alignment option. Therefore, options in the Bowtie2 alignment function are tweaked, and assignments filtered using information in the CIGAR string so that the final output is analogous to Bowtie’s best-strata option.

Unaligned sequences are then mapped to the protein-coding and anti-sense transcriptome (the reverse complement sequences of all genes) using the same process. Reverse complement sequences were mapped as endo-siRNAs, and piRNAs biogenesis can generate reverse complement sequences. Unmapped reads were then aligned to the genome (GRCh38) using prebuilt Bowtie and Bowtie 2 indexes13,14. The output of Bowtie and Bowtie2 is .sam files. These files were processed into “.sum” files containing each unique sequence's sum total (Figure 3).

|  |
| --- |
| Text  Description automatically generated |

Figure 3 Example of a .sum file:

### Assignment

With a hierarchical assignment approach, multi-class mapping reads are assigned to only one sRNA class using the following order: Mature miRNA>miRNA hairpin>yRNA>tRNA> snRNA>snoRNA>vault>viral miRNA hairpins>lncRNA>rRNA. Prior to assignment, sequences that map to a mature miRNA but not a miRNA's hairpin are discarded.

The final output files are gene expression matrices collapsed using the following options:

* Abundance based approach
* Abundance-based approach with hierarchical RNA class prioritisation (used for this thesis)
* Fractional counts
* Gene union counts

Each collapse option produces two expression matrices—one with ribosomal rRNAs retained and the other ribodepleted. The pipelines have been made user-friendly so that only folder names containing sample files need to be supplied to the alignment and assignment program. User options are contained in Table 4. The output is a folder containing gene expression matrices for each collapse option. A multi.reduced.RDS file is also created, which contains a record of how each sequence was assigned in each collapse option. This file can be used as a reference when reviewing differentially expressed sRNAs if the user wants to double-check where else reads may map to.

Table 4 Options that can be implemented in assignment

|  |  |
| --- | --- |
| **Option** | **Method** |
| Discard reads that map more than a user threshold defined cutoff. | Removes highly multi-mapping reads that arise from repetitive sequences |
| Collapse genome mapping reads. | This maps non-sRNAome mapping reads to 100bp sliding windows along the genome |
| Generate gene expression profiles (also known as presence plots). | Draws sample-wise expression profiles along the gene. Allows users to see whether reads map with a random uniform distribution or at discrete loci (see Figure 4 for examples). The former suggests the feature could be a degradation product, and the latter suggests specific cleavage by an endonuclease, possibly indicating that the feature has been processed systematically and thus has a biological function. |

|  |
| --- |
| A picture containing text, indoor, map, screenshot  Description automatically generated |

Figure 4 Example Gene expression profiles for an example gene taken from a clinical dataset. The x-axis represents the length along the transcript. Nucleotide 1 = the first place that at least one read maps to. The y-axis represents the number of times each nucleotide in the RNA sequences was identified amongst all reads for a given sample. The expression profile for each sample is plotted in a different colour.

1. Deschamps-Francoeur, G., Simoneau, J. & Scott, M. S. Handling multi-mapped reads in RNA-seq. *Comput Struct Biotechnol J* **18**, 1569–1576 (2020).

2. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology 2010 28:5* **28**, 511–515 (2010).

3. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).

4. Pantano, L., Estivill, X. & Martí, E. A non-biased framework for the annotation and classification of the non-miRNA small RNA transcriptome. *Bioinformatics* **27**, 3202–3203 (2011).

5. Zytnicki, M. mmquant: How to count multi-mapping reads? *BMC Bioinformatics* **18**, 1–6 (2017).

6. Pertea, G. gpertea/fqtrim. Preprint at https://doi.org/10.5281/ZENODO.1185412 (2018).

7. Griffiths-Jones, S., Grocock, R. J., van Dongen, S., Bateman, A. & Enright, A. J. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res* **34**, D140–D144 (2006).

8. Yoshihama, M., Nakao, A. & Kenmochi, N. SnOPY: A small nucleolar RNA orthological gene database. *BMC Res Notes* **6**, 1–5 (2013).

9. Chan, P. P. & Lowe, T. M. GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes. *Nucleic Acids Res* **44**, D184–D189 (2016).

10. O’Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **44**, D733–D745 (2016).

11. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**, 1–10 (2009).

12. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357 (2012).

13. Bowtie2 prebuilt indexes. https://bowtie-bio.sourceforge.net/bowtie2/manual.shtml.

14. Bowtie prebuilt indexes. https://bowtie-bio.sourceforge.net/manual.shtml.