A flexible framework for the analysis of short RNA sequencing data

Pierre-Luc Germain, Gretchen van Steenwyk, Isabelle Mansuy - Laboratory of Neuroepigenetics, Universität & ETH Zürich, Brain Research Institute, Zürich

Introduction & background

Short non-coding RNAs have been shown to be altered in different models of transgenerational epigenetic inheritance[1-3], sometimes functionally contributing to the phenotype[4-5], and have been implicated in a large number of biological processes and pathological conditions. However, their study is hampered by a number of technical issues, including the bioinformatic analysis of short RNA sequencing data. Among issues associated to the latter are the following:

1) First, analysis methods generally are not simultaneously optimized towards all known short RNA types (e.g. miRNA, piRNA, tRNA fragments, etc). On top of multiplying the work needed for an extensive analysis of the data, this can potentially create misassignment mistakes.
2) Second, most current methods either do not deal adequately with post-transcriptional modifications (for genome-based methods), or if they do (transcript-based methods), do not deal with unannotated features.
3) Third, current methods do not adequately account for the hierarchical organization of the features one might want to quantify/test.
4) Finally, there is still no consensus on the most appropriate normalization method for such data.

Because of these shortcomings, we developed a new analysis framework that addresses these issues using alternative nested equivalence classes over a customized annotation. We present this approach and package, and show how it can be used to redress biases in the quantification of both specific RNA as well as large RNA classes. We showcase its application in the context of the study of transgenerational epigenetic inheritance.

Example 1: the issue of post-transcriptional modifications:

tRNA-iMet-CAT-4 is transcribed from chrX, and as other tRNAs receives post-transcriptionally the 3' addition of the nucleotides CCA:

... GATTGAAACCATCCTCTGCTT ⋯➤ aligns to multiple locations on the genome

▼

... GATTGAAACCATCCTCTGCTT**CCA**⋯➤ does not align to the genome

to deal with this, researchers have often simply trimmed 3' CCAs before alignment, however this can often results in the read becoming ambiguous when instead it initially wasn't, as in the case above.

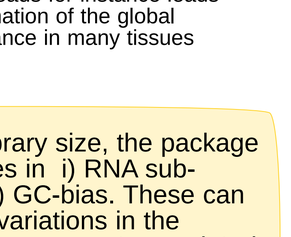
One solution to this issue is to build a custom genome that is complemented with known, post-transcriptionally modified transcripts (see also [6])

Example 2: the issue of ambiguity between related features

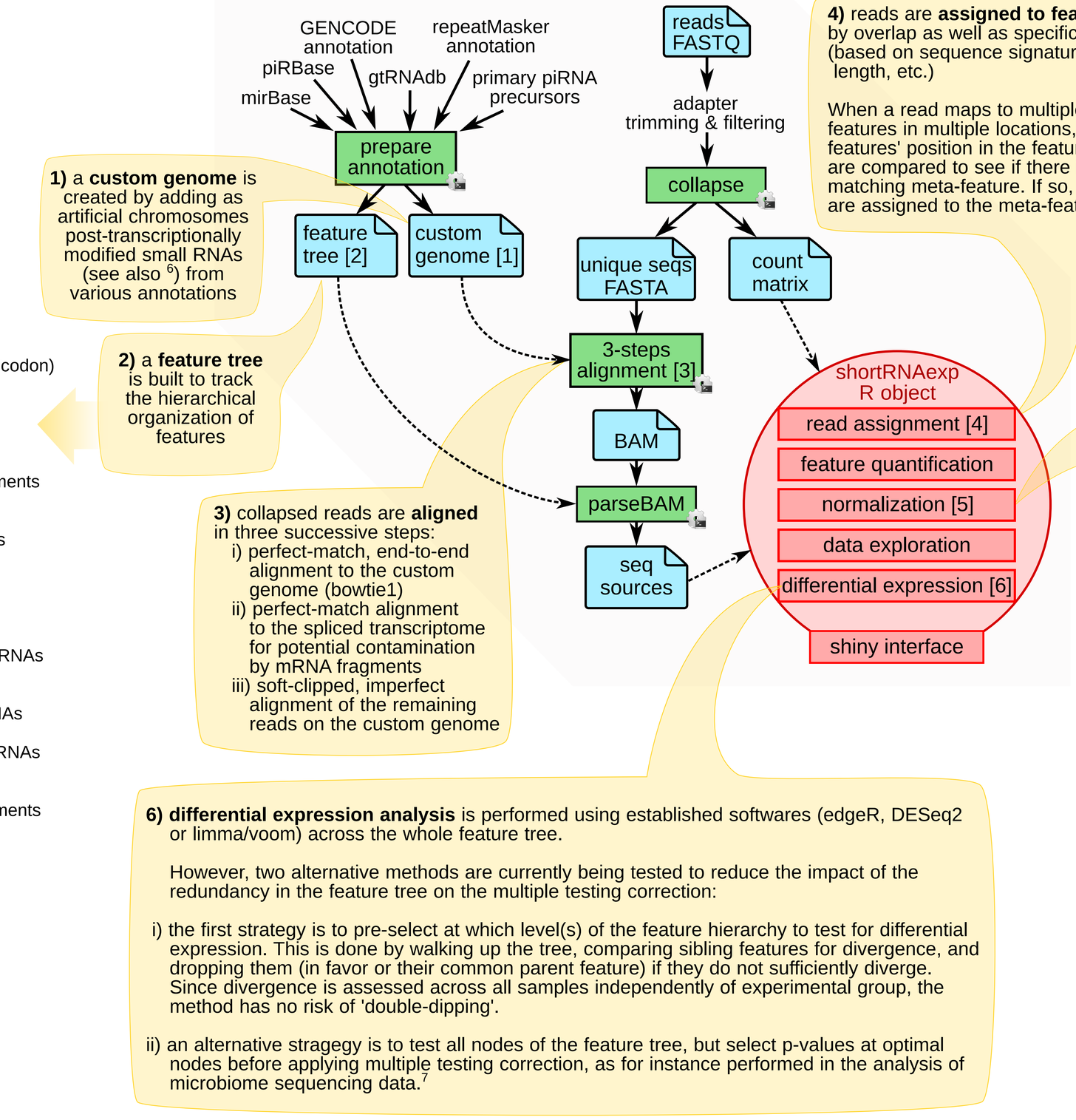
tRNA typically have several copies across the genome - for instance tRNA-Ala-AGC has 23 nearly-identical copies. While a genome alignment will make most reads from such features ambiguous (i.e. multimapping), from a functional point of view it is irrelevant from which exact location they came from.

This issue becomes even more critical with tRNA fragments, which often have conserved sequences across different tRNAs.

One way to address this issue is to aggregate reads into functional equivalence classes, i.e. higher level than specific genes/transcripts



Overview of the analysis pipeline



1) a **custom genome** is created by adding as artificial chromosomes post-transcriptionally modified small RNAs (see also [6]) from various annotations

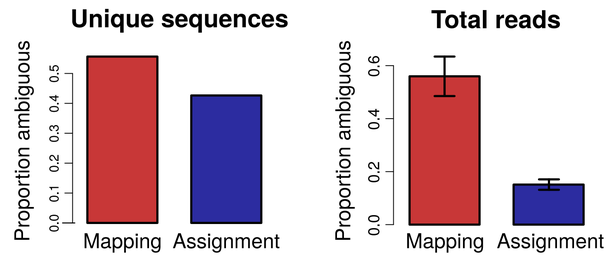
2) a **feature tree** is built to track the hierarchical organization of features

3) collapsed reads are **aligned** in three successive steps:
i) perfect-match, end-to-end alignment to the custom genome (bowtie1)
ii) perfect-match alignment to the spliced transcriptome for potential contamination by mRNA fragments
iii) soft-clipped, imperfect alignment of the remaining reads on the custom genome

4) reads are **assigned to features** by overlap as well as specific rules (based on sequence signatures, length, etc.)

When a read maps to multiple features in multiple locations, the features' position in the feature tree are compared to see if there is any matching meta-feature. If so, reads are assigned to the meta-feature.

This leads to a dramatic decrease in the proportion of reads that are **ambiguous** (and hence unusable for differential expression analysis) :



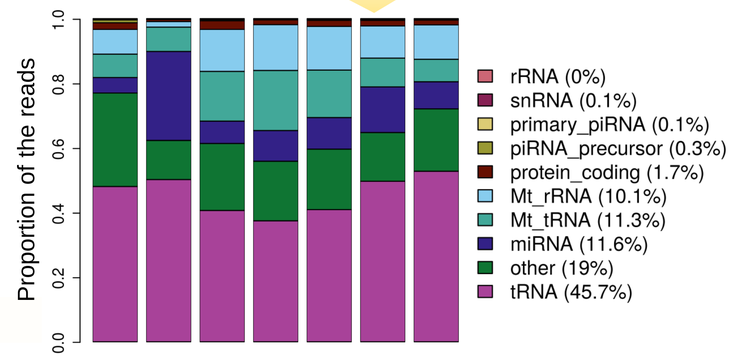
(example from sperm short RNAseq[2])

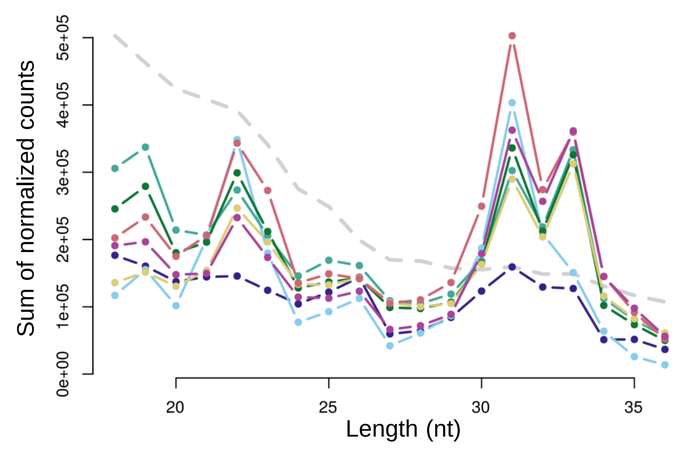
Ignoring these 'ambiguous' reads for instance leads to a ~10-fold underestimation of the global tRNA 5' fragments abundance in many tissues

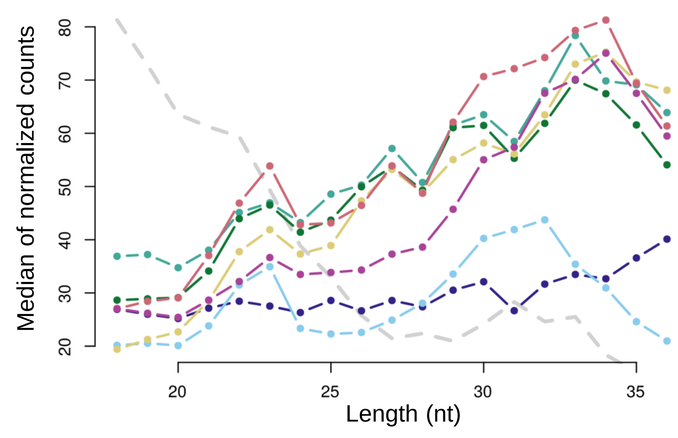
5) in addition to normalizing for library size, the package looks for variations across samples in i) RNA sub-populations, iii) length bias, and ii) GC-bias. These can easily be introduced by technical variations in the extraction or degradation, and if they are not associated with the experimental condition, they can be corrected for either in the normalization (e.g. through conditional quantile normalization) or through covariates added to the model during differential expression analysis

Feature tree

- --tRNA fragments
  - °--tRNA families (anticodon)
    - °--tRNA genes
      - --5' fragments
        - °--5' halves
      - --3' fragments
        - °--3' halves
      - °--internal fragments
- --miRNAs
  - °--miRNA families
    - °--specific miRNAs
      - °--isomirs
- --piRNAs
  - --primary
    - °--clusters
      - °--precursors
        - °--specific piRNAs
  - °--secondary
    - °--clusters
      - °--specific piRNAs
- --other sncRNA
- --fragments from long RNAs
  - °--RNA biotypes
    - °--gene
      - °--specific fragments
- °--unknown
  - °--specific fragments







(example of variations in length bias in samples from sperm short RNAseq[2])

6) differential expression analysis is performed using established softwares (edgeR, DESeq2 or limma/voom) across the whole feature tree.

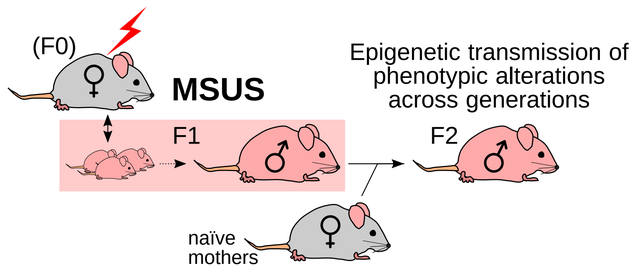
However, two alternative methods are currently being tested to reduce the impact of the redundancy in the feature tree on the multiple testing correction:

i) the first strategy is to pre-select at which level(s) of the feature hierarchy to test for differential expression. This is done by walking up the tree, comparing sibling features for divergence, and dropping them (in favor or their common parent feature) if they do not sufficiently diverge. Since divergence is assessed across all samples independently of experimental group, the method has no risk of 'double-dipping'.

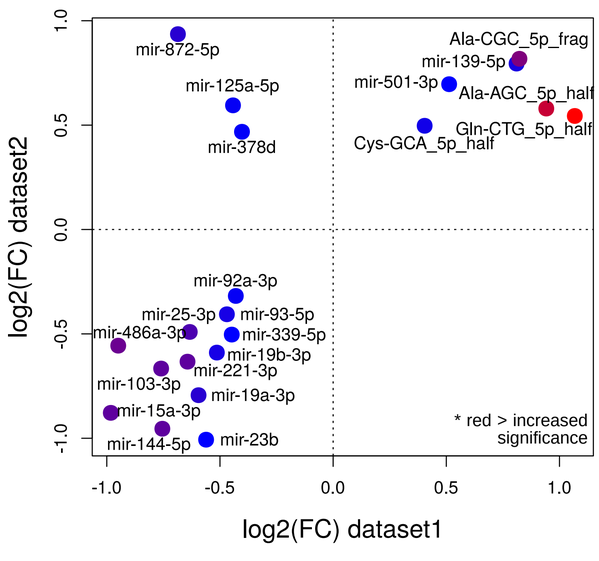
ii) an alternative stragegy is to test all nodes of the feature tree, but select p-values at optimal nodes before applying multiple testing correction, as for instance performed in the analysis of microbiome sequencing data.[7]

Example application

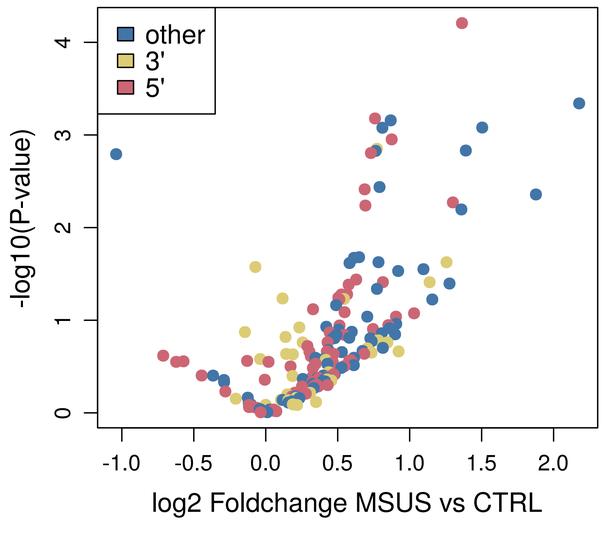
The **unpredictable maternal separation combined with unpredictable maternal stress (MSUS)** model, developed by the Mansuy lab, is the only germline-dependent transgenerational model of early trauma that exists, and one of the most repeatedly reproduced and richly characterized model of transgenerational epigenetic inheritance. Daily for the first two weeks of life, pups are randomly separated from their mother and the latter subjected to stress. The exposed pups show various behavioral and metabolic alterations in adulthood, many of which are also visible in the following generations. Preliminary results have suggested that transmission involves a reprogramming of the germline by circulating factors. We therefore investigated the short non-coding RNA landscape in serum and sperm of MSUS animals.

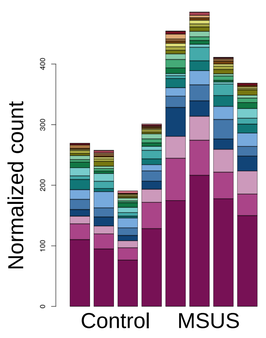


Alterations in serum small non-coding RNAs found across two MSUS experiments



Identification of consistent alterations in tRNA fragments in MSUS serum and sperm





References

1. Rechavi et al., *Cell* 2011.
2. Gapp et al., *Nature Neuroscience* 2014.
3. Bohacek and Mansuy, *Nature Reviews Genetics* 2015.
4. Grandjean et al., *Sci. Rep.* 2015.
5. Chen et al., *Science* 2016.
6. Hoffmann et al., *Bioinformatics* 2017.
7. Tang et al., *Bioinformatics* 2017.

Universität Zürich[UZH]