




# Micro-Processor, an automated pipeline for annotation, characterization and expression analysis of miRNAs reads from Illumina sequencing

P Batzel<sup>1</sup>, BF Eames<sup>1,2</sup>, T Desvignes<sup>1</sup>, H-L Fei<sup>3</sup>, K Artinger<sup>3</sup>, DE Clouthier<sup>3</sup>, JH Postlethwait<sup>1</sup>

<sup>1</sup>  Institute of Neuroscience, University of Oregon, Eugene OR 97403; <sup>2</sup>  UNIVERSITY OF SASKATCHEWAN Department of Anatomy & Cell Biology, Saskatoon, SK Canada; <sup>3</sup>  University of Colorado Anschutz Medical Campus, School of Dental Medicine, Denver CO 80045

## Abstract

The development of high-throughput sequencing techniques has recently allowed biologists to generate massive amounts of gene expression sequence data in a relatively quick and inexpensive manner. With up to hundreds of millions of individual reads from each sample, extracting biological significance from lists of sequences remains a major issue. Furthermore, micro-RNA (miRNA) sequences can often differ from annotated reference sequences and reflect a biological process -- micro-RNA editing -- that can modulate the function of a specific miRNA. To overcome these problems that confront our studies of the action of miRNAs in midface development, we developed Micro-Processor, an automated pipeline for the annotation, characterization and expression analysis of miRNAs across multiple samples and various species.

Micro-Processor takes as input short reads produced by Illumina high-throughput sequencing, an annotated or unannotated reference genome, and a list of known miRNAs from the same or another species. First, this pipeline will make strand-specific groups of miRNA sequences and name them. Next, it identifies putative new miRNAs and their genome position for further analysis. The software normalizes each miRNA's read counts and quantifies differential expression analysis across input datasets. Finally, Micro-Processor allows the user to easily access the gathered information on strand-specific expression and various types of RNA editing process that putatively have major biological significance. The software has allowed us to compare miRNA expression across three tissues of the mouse midface and to identify tissue-specific miRNAs.

## miRNA Processing

RNA editing: the post-transcriptional chemical alteration of a base.

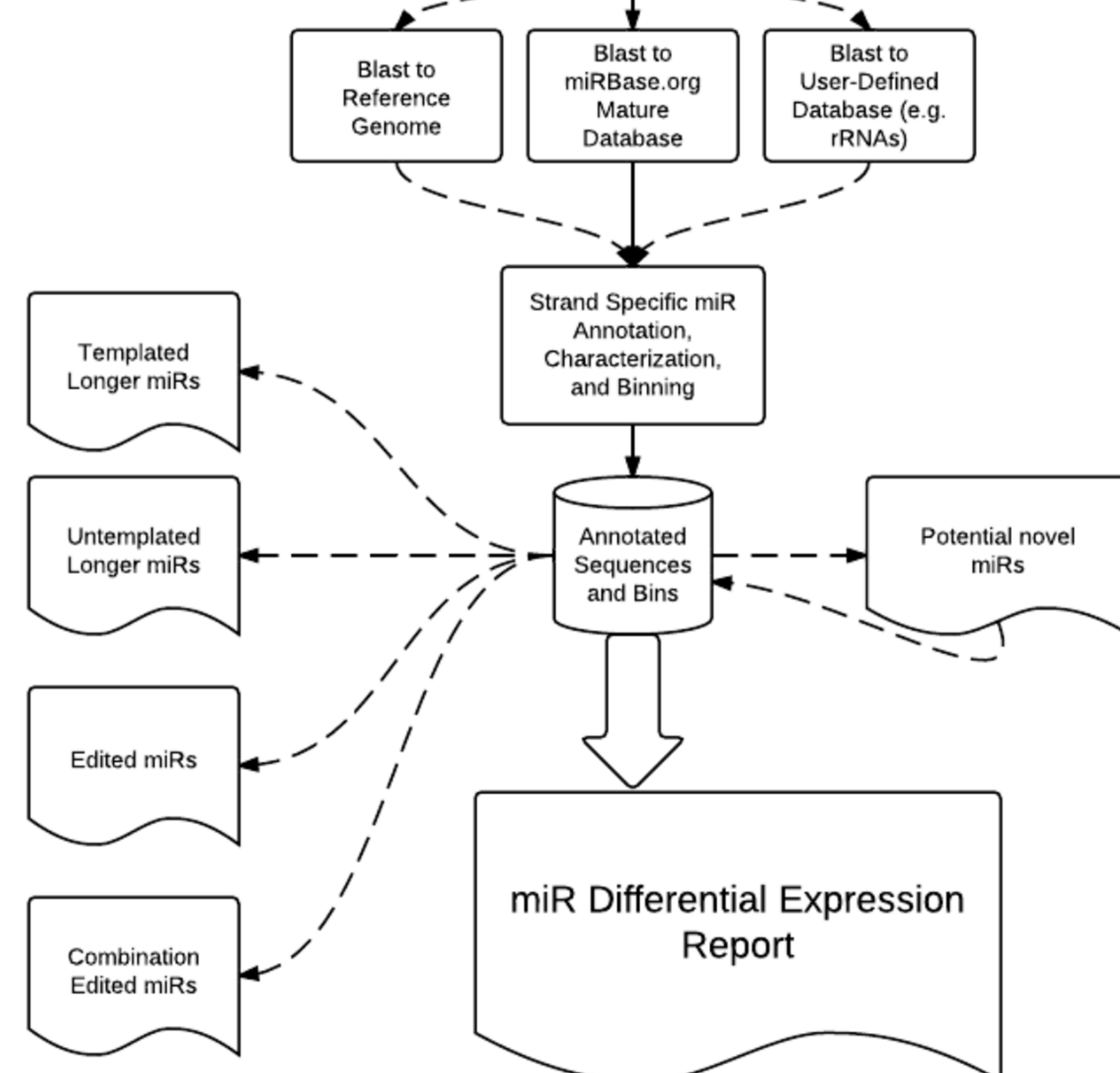
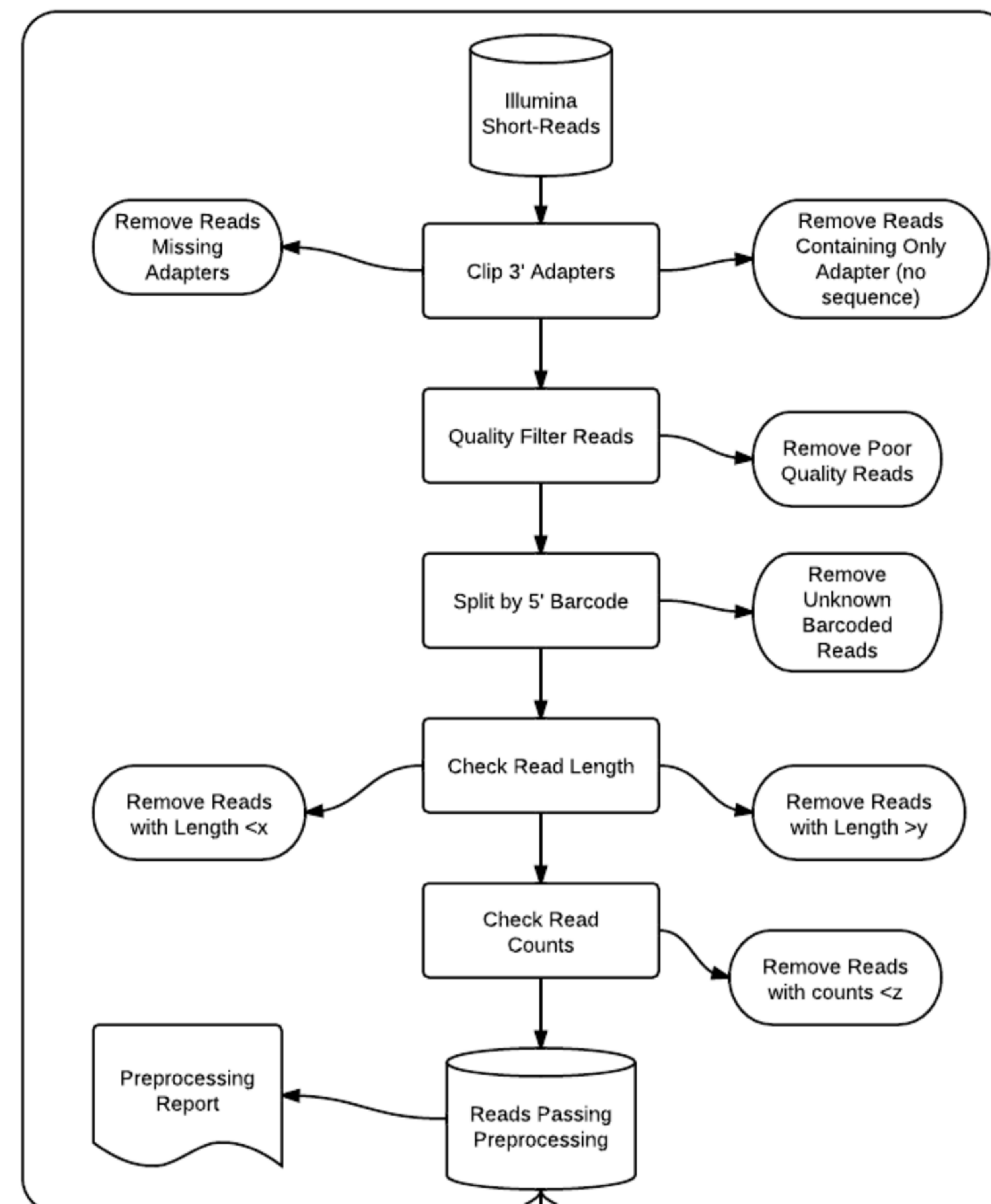
Examples of sequence modifications observed in micro-RNAs

Genomic sequence around a miR	<u>agcuagCUAGCUAGCUAGCUAGcuagcuagc</u>
Most common mature miR	<u>CUAGCUAGCUAGCUAGCUAG</u>
Templated longer miR	<u>CUAGCUAGCUAGCUAGCUAGcu</u>
Untemplated longer miR	<u>CUAGCUAGCUAGCUAGCUAGu</u>
Edited miR	<u>CUAGCUAGCU<u>U</u>CUAGCUAG</u>
And combinations!	<u>CUAGCUAGCU<u>U</u>CUAGCUAG<u>cu</u></u>

## Micro-Processor Input

```
@Example_Illumina_FastQ_Sequence1
ATGGACTGACTGACTGACTGACTGTCGTATGCCGCTTCTGCTTGAAAAA
+
__eeceeggggggfhdb^edghhhhhhhhhhh[XaffhfhgBBB
```

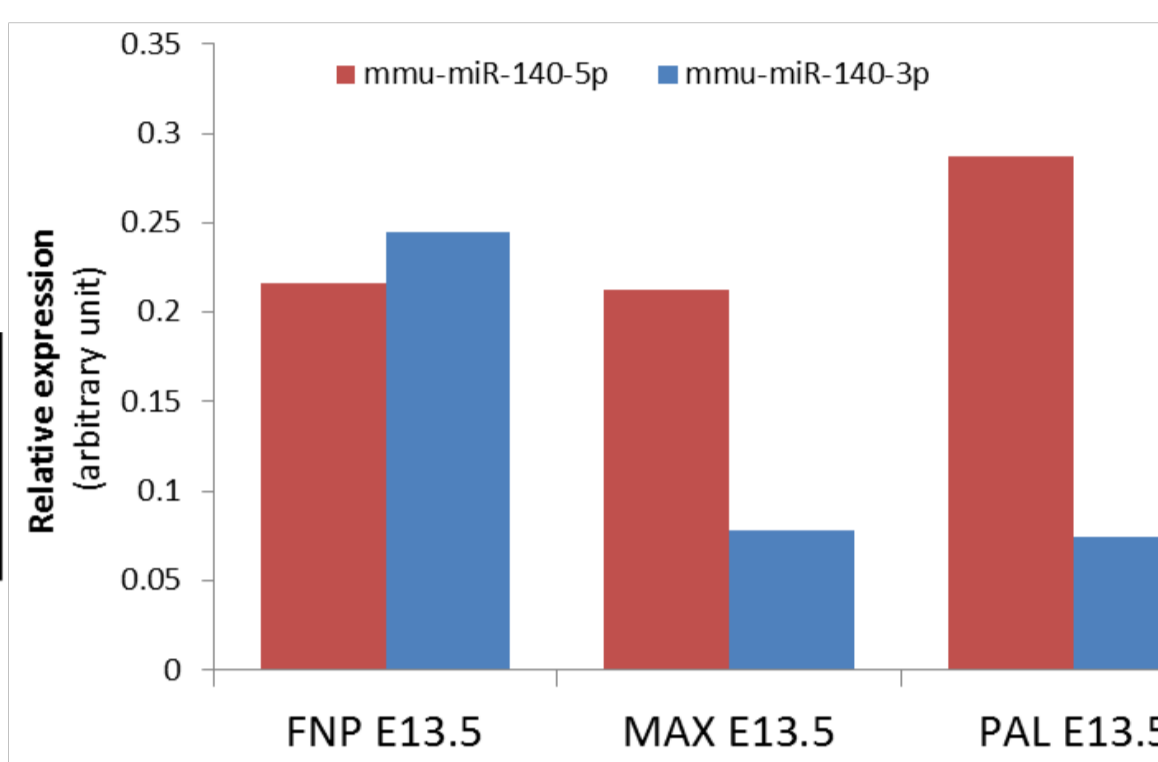
## Pipeline Flowchart



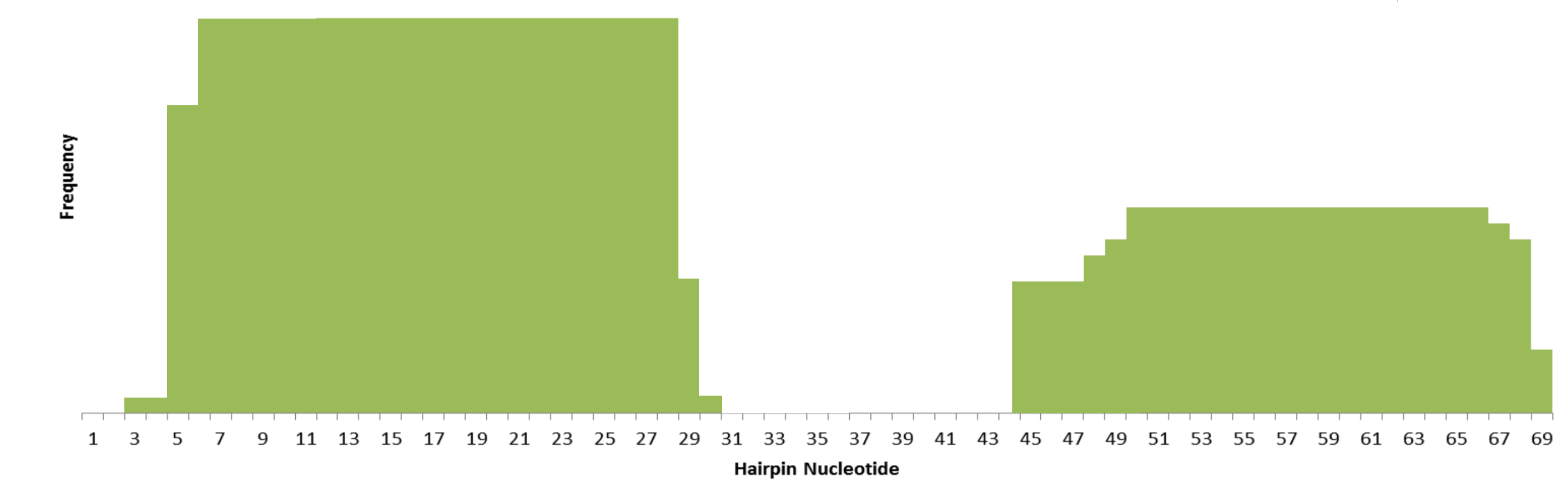
## Micro-Processor Output

### Micro-Processor Reveals Differential Expression Across Tissues

Sequence	Bin Name	FNP13 Count	FNP13 %	MAX13 Count	MAX13 %	PAL13 Count	PAL13 %	mmu Genomic Position
CAGTGGTTTACCCCTATGGTAG	mmu-miR-140-5p	32538	0.2162	24150	0.21279	38239	0.2872	8:110075149-110075170
TACCACAGGGTAGAACACCGG	mmu-miR-140-3p	36860	0.2448	8833	0.07776	9939	0.07454	8:110075188-110075208



Frequency of each base along the mmu-miR-140 hairpin



Alignment of processed Illumina reads on mmu-miR-140 hairpin (GSNAP+IGV)



## Conclusions

Micro-Processor allows the user to:

- Manage the stringency of analysis by adjusting quality filtering, read count threshold, and small RNA length
- Identify potential novel miRs or un-annotated strands
- Investigate the frequency of miRNA processing
- Compare expression levels between datasets

For the FaceBase project, Micro-Processor allows us to compare miRNA expression across three tissues of the mouse midface and to identify tissue-specific expression of certain miRNAs. These preliminary results are presented on the poster by Eames et al. "Comparative genomics reveals miRNAs that give identity to developing mid-facial tissues".

Micro-Processor is a user friendly pipeline for the annotation, characterization, and expression analysis of miRNAs across multiple samples and species, and is software of broad utility for any people studying miRNA through Illumina sequencing.

The Micro-Processor pipeline will be publicly released as a Python based script to run locally or as a web service.

Funding: Work was supported by grant U01-DE20076 from NIDCR to D.E.C, K.B.A and J.H.P.