

README file - 2010-03-15**Requirements:**

- Linux OS. The pipeline was tested on an Ubuntu 9.04 64 bit system running on an Intel Quad core 2.4MHz with 4Gb of memory.
- Bowtie or Eland aligner with appropriate genome index
- perl v5.10.0
- gzip v1.3.12
- Ensembl perl API installed (optional)

What does E-miR do:

- Generates an expression matrix for all know non-coding RNAs detected in the input data
- Report transcripts per million (correct for sequence depth) and square root transformed expression levels (variance stabilization)
- Generate Bed and Wig files for data visualization in the UCSC browser.

What does E-miR not do:

- Build the Eland or Bowtie genome indexes to be used in the analyses.
- Predict novel miRNA transcripts or other novel non-coding RNA transcripts
- Statistical inference of differential expression between (groups of) samples

The EmiR microRNA analysis pipeline consists of the following perl scripts:

- EmiR_Bowtie.pl & EmiR_Eland.pl [main pipeline files]
- 1_filter_chr.pl & 1_prep_wig_file.pl [automatically called by EmiR for generating Wig files]
- Retrieve_non-coding_RNA_annotation.pl [retrieves RNA annotation via the Ensembl perl API]

Preparation for the Eland version

The 'par_file_eland.info' file should be edited. This is a file listing info like adapter truncation, aligner executable and genome index, RNA annotation files and data processing. The info needed is listed per line. Do not edit the first element of each line and do not change the name of the file.

adapter

Leading nucleotide sequence of the 3' adapter to be used for adapter identification and removal. For the modified SREK-to-Illumina protocol this sequence is CGCCTTGGCCGTACAGCAG. Using the first 8 nucleotides [CGCCTTGG] is sufficient to ensure proper adapter removal without aberrant truncation in the RNA insert sequence.

allow_mismatches

[yes or no]: if yes, one mismatch in the identification of the 3' adapter is allowed

eland_location:

Full path to the eland executable, followed by a '/', e.g., '/home/henk/Eland_exec/'. The pipeline expects the eland executables for the different lengths to be in the eland_xx format, where xx is the specific nucleotide length of the sequences to be aligned.

eland_genome_location:

Full path to the folder containing the eland genome index files to be used in the alignment. The script expects there to be files on one genome in this folder, e.g., '/home/henk/Eland_exec/Gga_genome/'

transcript_annotation

This is a tab delimited file listing transcript annotation in the following column order: Ensembl Gene ID, Ensembl Transcript ID, Strand, Transcript Start (bp), Transcript End (bp), Chromosome Name, Description, Associated Gene Name, Transcript Biotype.

mature_star_annotation

This file lists the positions of the 3p and/or 5p (mature/star) miRNA transcripts resulting from Dicer cleavage of the precursor miRNA hairpin. Positions are relative to the begin of the precursor (!), not genomic locations. miRNA transcript precursors should be listed by their Ensembl transcript identifiers. Note that not for all miRNAs positions for the 3p and 5p products are in the Ensembl database.

```
# e.g., ENSGALT00000028949 13-34
#      ENSGALT00000028949 50-70
#      ENSGALT00000028950 16-37
```

The files for 'transcript_annotation' and 'mature_star_annotation' can be generated by the 'Retrieve_non-coding RNA_annotation_2009-12-22.pl' accessory perl script that connects to the Ensembl perl API to retrieve annotation on known non-coding RNA transcripts. These files can directly be used in the E-miR pipeline. This script can be executed by running

```
'perl Retrieve_non-coding RNA_annotation_2009-12-22.pl <species>'
```

with species being human, mouse, chicken etc. In order to run this file, the Ensembl perl API needs to be installed. Output files are <species>_non-coding RNA_annotation and <species>_mature-star_positions. Custom transcripts of interest can be added to the transcript_annotation files manually in the proper format.

onlyU0U1

[yes or no]: If yes, only sequences uniquely aligned to the genome index with no or one mismatch will be extracted from the Eland output files before processing them. This will speed up the pipeline. However, information on sequences aligned to Repeats, uniquely aligned with two mismatches or not aligned to the genome will not be reported in the output file.

seqsets_out

[yes or no]: During the annotation process, a file can be generated that lists all unique sequences that are annotated to each of the RNA transcripts that were included in the Annotation files. If 'yes', this data is reported in the 'Matched_sequence-sets_Eland' file.

data_format

[seq_count or FASTQ or SCARF]: Indicate the data format to be analyzed. Only one type of data format can be processed at a time!

The 'seq_count' format indicates a tab delimited file with on each line the sequence and the number of times that sequence was found in the data, i.e.,

```
AAGGTGCATCTAGTGCAGATAGCGCCTTGCCCGTA      35037
CGGCTGGGAGCCGCCCTTGCCCGTACAGCAGGGTGT      28319
```

The FASTQ and SCARF are the standard output formats for most Next Generation sequencers. Both the sequence reads and the quality scores are included.

FASTQ format example:

```
@HWI-EAS384:2:1:2:1673#TATATG/1
CTGCCCTTCCAATCATTTTACCCTTTTCGCCCTT
+HWI-EAS384:2:1:2:1673#TATATG/1
a_J[ ]`IRb\HN\SY^PXNLHDPH_UUG]SVY
```

SCARF format example:

```
HWI-EAS384:3:1:4:778#0/1:CTCACACAGAAATCGCTCCGNCGCCTTGNNNNNNN:_aZa`aaa`a^a`_[^`a]BBBBBBBBBBBBBBBB
HWI-EAS384:3:1:4:1312#0/1:CCCGCTAACTCAGTCGGTANAGCATGANNNNNNN:^^`aaaaaa`_a^`_a`UBBBBBBBBBBBBBBBBB
```

The current version of the EmiR pipeline ignores the quality scores!

Preparation for the Bowtie version

The par_file_bowtie.info differs from the Eland aligner version on two fields.

processors: For Bowtie alignment the number of processors to be used can be specified here. The 'onlyU0U1' option is not available in the Bowtie version because it reports only the sequences aligned with no or one mismatch.

The other fields contain identical info.

Running the pipeline

When the 'par_file_eland/bowtie.info' files have been prepared open a command line to the location where the perl scripts, info and data files are stored and type:

```
perl EmiR_Bowtie.pl <as> <many> <data> <files> <as> <you> <want>
or EmiR_Eland.pl <as> <many> <data> <files> <as> <you> <want>
```

with spaces in between the data files, and press enter

e.g., for the demo data: perl EmiR_Bowtie.pl EMS-multi_10K EMS-single_10K HTs-multi_10K HTs-single_10K

EmiR pipeline output

Several files are generated by the EmiR pipeline.

session_Info_Eland/Bowtie_run. Contains:

- General info from the par_file_bowtie/eland.info file.
- For each input datafile the sum of uniquely aligned sequences with no or one mismatch and the scaling factors to calculate to expression values into transcripts per million.
- A short table listing alignment and annotation characteristics (columns) for each input sample (rows).

Column description:

total counts: total number of sequence reads

rejected counts: total number of sequence reads that were shorter than 15 nt after 3' adapter removal

accepted counts: total number of sequence reads with lengths between 15 and 32 nt after 3' adapter removal. These are aligned to the genome.

NoMatch: number of sequences that could not be mapped to the genome *

QCfail: no matching done: QC failure (too many Ns in the sequence) *

Repeats: number of sequences mapped to repeat regions *

U0: number of sequences mapped to the genome without mismatches

U1: number of sequences mapped to the genome with one mismatch

U2: number of sequences mapped to the genome with two mismatches *

*: listed in Eland version only

The remaining columns list the total number of transcripts annotated to each of the different RNA transcript biotypes represented in the Annotation files, e.g., miRNA, rRNA, snoRNA, Mt_tRNA, misc_RNA, tRNA, snRNA, etc.

UCSC visualization files for each input data file (gzipped): Can be uploaded at:

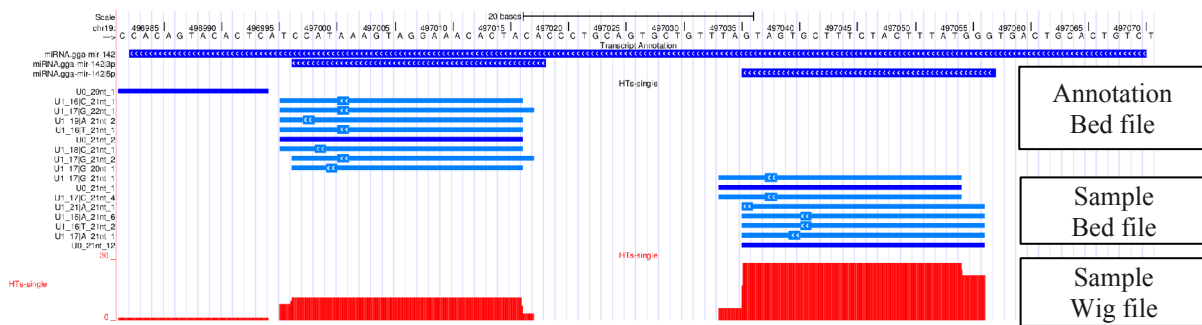
<http://genome.ucsc.edu/cgi-bin/hgGateway>

***Bed.gz:** Visualization of ALL individual transcript sequences aligned to the genome. In the Bed files, red and blue bars represent transcripts mapped to the forward and reverse strand, respectively. Dark red and light red represent transcripts aligned to the genome with no and one mismatch, respectively (similar for light and dark blue). The position of the mismatch is indicated at the bold section of the bar. At the left side of the graph specific info on the transcript is shown, e.g., U1_16|A_21nt_6, indicating this transcript had one mismatch, which is located at position 16, an Adenine was sequenced, the whole transcript is 21 nt long and was found 6 times in the sample.

***Wig.gz:** Expression density visualization of the data. These files are generated from transcripts per million transformed data. The height of the bar represents the expression level of the transcripts and can be compared across samples.

Annotation_BED_file_Eland.gz: A bed file containing all RNA transcripts from the Annotation files. At the left of the window the transcript identity is shown.

An example of all three visualization types is show in the UCSC browser screen shot below:



Matched_table_summary_Eland/Bowtie:

This table lists all expressed non-coding RNA transcripts from the annotation files that were detected in the sample data. The first column hold the transcript identifier, e.g., ENSGALT00000042423|3p~|chr7|-|1330071|1330092|miRNA|gga-mir-1559|sense, indicates :

EnsemblTranscript ID | dicer product | chromosome | strand | begin | end | biotype | transcript name | (anti)sense

The separate 3p and 5p dicer cleavage products are separately represented in the table by 3p or 5p in the identifier. In cases where the miRbase has only one of the Dicer cleaved miRNA transcripts, the complementary transcripts were inferred from the hairpin structure. These unlisted transcripts are indicated by a '3p~' and '5p~' in the identifier.

When there is an '|n|' in the identifier and its a microRNA, the data listed in that row is compiled from all transcripts annotated to the entire miRNA precursor transcript.

Expression for other non-coding RNA transcripts, like snoRNA and tRNA, are also included in the table. These ALL have the '|n|' in the identifier. The second column holds the genome location

For each of the input files, seven columns of data are included:

unique	unique number of reads annotated to this miRNA transcript
counts	sum of the number of times this miRNA transcript was found
U0-counts	same as 'counts' but then only the sum of perfect matches only
highest_count	most abundant miRNA isomir
highest_seq	which seq was the most abundant one
tpm scaled	the 'counts' value normalized/scaled to sequences per million
sqrt	square root of the scaled value. This stabilizes variance.

Matched_sequence-sets_Eland:

Lists all unique sequences, tab delimited, that are annotated to each of the RNA transcripts that were included in the Annotation files. The identifier is identical to the one in column1 of the Matched_table_summary_Eland/Bowtie file.

tar.gz archives:

These archives contain processed data files for each of the input data files

[OUTPUT 1 trunc Archived.tar.gz:](#)

Sequence and counts after 3' adapter removal, tab delimited

[OUTPUT 2 Not_accepted_sequences Archived.tar.gz](#)

Sequences that were shorter than 15 nt after truncation and sequences that did not align to the genome with no one one mismatch.

[OUTPUT 3 Mapped_seqs_tpm Archived.tar.gz](#)

Sequences aligned to the genome with no or one mismatch, followed by chromosome, strand, begin position, end, position, length, alignment type and expression level in tpm. Alignment type U1_6A indicates one mismatch [U1], at position 6 found an Adenine. For sequences aligned without mismatches U0 is listed.

[OUTPUT 5 not_RNA-matched_tpm Archived.tar.gz](#)

Sequences that were aligned to the genome with no or one mismatch but were not annotated to any of the non-coding RNA transcripts included in the Annotation files. Format is identical to the '3_Mapped_seqs_tpm_Archived' files

[OUTPUT 6 Insert_histogram Archived.tar.gz](#)

A table of the number of unique sequences and the sum of sequences for the length of the sequence after 3' adapter removal.