Application Note

Genome-wide Profiling of microRNAs by Next-Generation Sequencing

Introduction

microRNAs (miRNA) are a class of small non-coding RNAs typically 21-23 nt long found in plants, animals, and some viruses (Figure 1). miRNAs play a pivotal role in RNA silencing and post-transcriptional regulation of gene expression. Discovered in the early 1990s, miRNA research has revealed (i) multiple roles for miRNAs in development (ii) disease-associated aberrant expression of miRNAs and (iii) the importance of miRNA in many other biological processes.

Next generation sequencing (NGS) technologies have become a powerful tool to study genome-wide miRNA expression patterns and have helped to identify disease associations, isoforms of miRNAs, and to discover previously uncharacterized miRNAs.

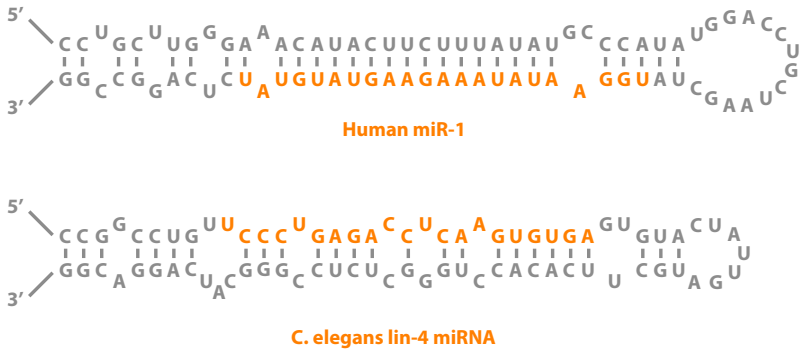


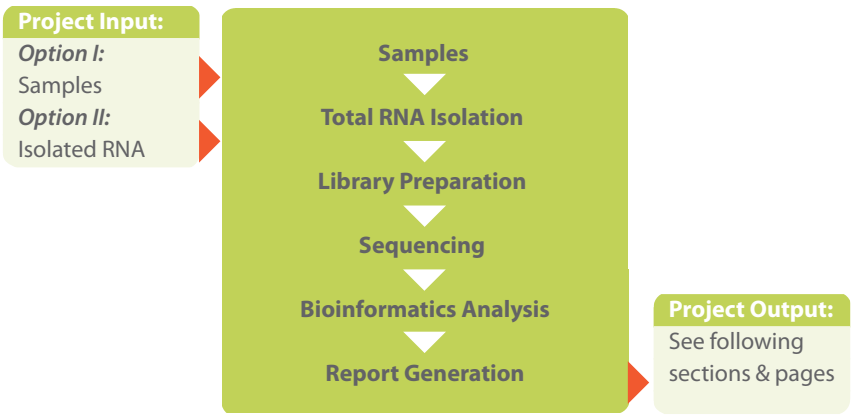
Figure 1. Typical structure of two precursor miRNAs showing the stem loop structure observed in precursor miRNAs (pre-miRNAs). The primary miRNA transcript which varies in between 500 and 3000 nt is processed by RNAse III and the dsRNA binding protein resulting in a 70-80 nt long pre-miRNA. The pre-miRNA is then actively transported from the nucleus to the cytoplasm where it is further processed by the protein Dicer resulting in mature 17-23 nt long miRNAs.

Microsynth Competences and Services

Experimental Design: As an expert in the area of miRNA-Seq, Microsynth is able to provide a one-stop service from experimental design consulting up to bioinformatics analysis. Should you not involve Microsynth in your experimental design, please consider the importance of the number of biological replicates. We usually advise to include at least 3 biological replicates per condition, to finally obtain statistical significance for your differential miRNA expression analysis.

RNA Isolation: Either you leave it up to Microsynth or you use a commercial kit to isolate total RNA used for the Illumina miRNA-Seq protocol.

Library Preparation and Sequencing: Following a quality check of your total RNA samples, Microsynth will perform miRNA enrichment. Illumina cDNA library is generated by reverse-transcription including specific sequencing adaptors with barcodes. Finally, the libraries are pooled and sequenced on the Illumina machine. The envisaged number of reads per library depends on the organism under study and the desired sensitivity. The usually required number of reads for higher eukaryotic species (e.g. human, rat, mouse) is approx. 5-15 Mio reads depending whether complex tissues or unique type of cells are analyzed.



Microsynth AG | Schützenstrasse 15 | 9436 Balgach | Switzerland
Phone: +41 71 722 83 33 | E-mail: info@microsynth.ch | Web: www.microsynth.ch

Bioinformatics Analysis: The analysis pipeline at Microsynth addresses three main questions: (i) what is the distribution of miRNAs and which of them are novel, (ii) which pathway is influenced in which way by the miRNAs and (iii) which of the miRNAs are differentially expressed. The first step of analysis is based on the sequence data itself. In short, sequence data is quality filtered and clustered for each condition of the experiment. A representative sequence of each cluster is then compared against the miRBase database using UBLAST to identify known miRNAs. Sequence clusters that did not result in a significant hit may be regarded as putative novel miRNAs. In the second step of the analysis the quality filtered reads are mapped against the reference genome using STAR. Then, HOMER is employed to find miRNA peaks and motifs and to exhaustively annotate them (e.g. proximity to genes and gene ontology). However, this in-depth annotation is only supported for a limited set of model organisms (e.g. human, mouse, zebra-fish). Finally, differentially expressed miRNAs are found using DESeq2.

Provided Output Files:

See examples below

Examples for Most Important Output Files Provided by Microsynth

query id	subject id	% identity	alignment length	mismatches	gap opens	q. start	q. end	s. start	s. end	evalue	bit score
NS500XXX:31:H15GRBGXX:4:13503:18671:3977;size=9;	dre-miR-430b-3p	100.0	23	0	0	1	23	1	23	9.5e-06	43.6
NS500XXX:31:H15GRBGXX:4:13503:18671:3977;size=9;	dre-miR-430b-3p	100.0	23	0	0	1	23	1	23	9.5e-06	43.6
NS500XXX:31:H15GRBGXX:1:11102:1527:1097;size=162;	dre-miR-19a-3p	100.0	23	0	0	1	23	1	23	9.5e-06	43.6
NS500XXX:31:H15GRBGXX:4:12512:6410:15697;size=6;	dre-miR-734	100.0	23	0	0	1	23	1	23	9.5e-06	43.6
NS500XXX:31:H15GRBGXX:2:11304:19107:1572;size=10;	dre-miR-734	100.0	23	0	0	1	23	1	23	9.5e-06	43.6

Figure 2. Exemplary extract of the comparison of cluster representative sequence (query id) against the miRBase database (subject id) using UBLAST. Clusters that show no significant blast hit can be regarded as putative novel miRNAs.

Rank	Motif	p-value	log p-value	q-value/FDR	% of Targets	% of Background	STD(Bg STD)	Best Match/Details	Motif File
1		1e-22	-5.173e+01	0.010	20.37%	0.07%	27.9bp (53.3bp)	NA More Information	motif file (matrix)
2		1e-10	-2.370e+01	0.060	7.41%	0.01%	43.4bp (18.7bp)	NA More Information	motif file (matrix)
3		1e-9	-2.264e+01	0.070	9.26%	0.05%	35.1bp (26.9bp)	NA More Information	motif file (matrix)
4		1e-9	-2.126e+01	0.055	55.56%	17.88%	40.0bp (54.6bp)	NA More Information	motif file (matrix)
5 *		1e-7	-1.762e+01	0.300	11.11%	0.31%	38.8bp (52.4bp)	NA More Information	motif file (matrix)

Figure 3. Exemplary extract of the de-novo motif search results. Given the p-value and the false discovery rate, motif 5 (*) is rated as putative false positive motif. Additionally total target and background sequences are listed and links to known motifs and gene ontology enrichment results are provided if present (not shown).

PeakID	Chr	Start	End	Strand	Peak Score	Focus Ratio/ Region Size	Annotation	Detailed Annotation	Distance to TSS	Nearest PromoterID	Entrez ID	Nearest Unigene	Nearest Refseq
chr9-1	chr9	53272860	53273010	+	1595122	0.997	Intergenic	(C)n|Simple_repeat|Simple_repea	-359284	NM_001076600	767660	Dr.89901	NM_001076600
chr14-1	chr14	13964023	13964173	-	746341.5	0.994	Intergenic	(GA)n|Simple_repeat|Simple_rep	7584	NM_201467	321544	Dr.150278	NM_201467
chr8-1	chr8	6469801	6469951	+	663414.6	0.965	Intergenic	(C)n|Simple_repeat|Simple_repea	529730	NM_001082937	100000097	Dr.140237	NM_001082937
chr1-1	chr1	34344610	34344760	+	170731.7	0.972	intron (NM_198140, intron 2 of 5)	intron (NM_198140, intron 2 of 5)	1698	NM_198140	30723	Dr.75674	NM_198140
chr19-1	chr19	15459446	15459596	+	165853.7	0.971	TTS (NM_199211)	TTS (NM_199211)	-4973	NM_213042	406315	Dr.132981	NM_213042
chr7-2	chr7	69525252	69525402	-	121951.2	0.962	intron (NM_198143, intron 2 of 4)	intron (NM_198143, intron 2 of 4)	-2337	NM_198143	378961	Dr.75569	NM_198143
chr14-2	chr14	32730323	32730473	+	92682.9	0.95	exon (NM_212598, exon 5 of 10)	exon (NM_212598, exon 5 of 10)	2189	NM_212598	321860	Dr.76840	NM_212598
chr4-1	chr4	13790454	13790604	-	92682.9	0.95	Intergenic	U4|snRNA|snRNA	1890	NM_001040252	561018	Dr.66020	NM_001030151
chrM-1	chrM	1872	2022	-	73170.7	0.938	NA	Intergenic	NA	NA			
chr7-4	chr7	26209441	26209591	-	68292.7	0.933	intron (NR_120316, intron 1 of 13)	intron (NR_120316, intron 1 of 13)	-344	NR_120316	325112	Dr.105998	NR_120316
chr2-1	chr2	7508129	7508279	-	68292.7	0.933	intron (NM_213451, intron 8 of 10)	intron (NM_213451, intron 8 of 1(	-8070	NM_213451	406760	Dr.8006	NM_213451
chr7-3	chr7	26212648	26212798	-	68292.7	0.778	intron (NR_120316, intron 10 of 13)	intron (NR_120316, intron 10 of 1:	-3551	NR_120316	325112	Dr.105998	NR_120316
chr3-2	chr3	53346553	53346703	+	39024.4	0.889	Intergenic	Intergenic	24034	NM_001004117	445494	Dr.15803	NM_001004117
chr5-1	chr5	34904482	34904632	+	29268.3	0.857	intron (NM_200047, intron 1 of 6)	intron (NM_200047, intron 1 of 6)	159	NM_200047	337010	Dr.53820	NM_200047

Figure 4. Exemplary extract of the peak annotation table generated during pathway analysis. Peak locations and known genomic features in close proximity are annotated.

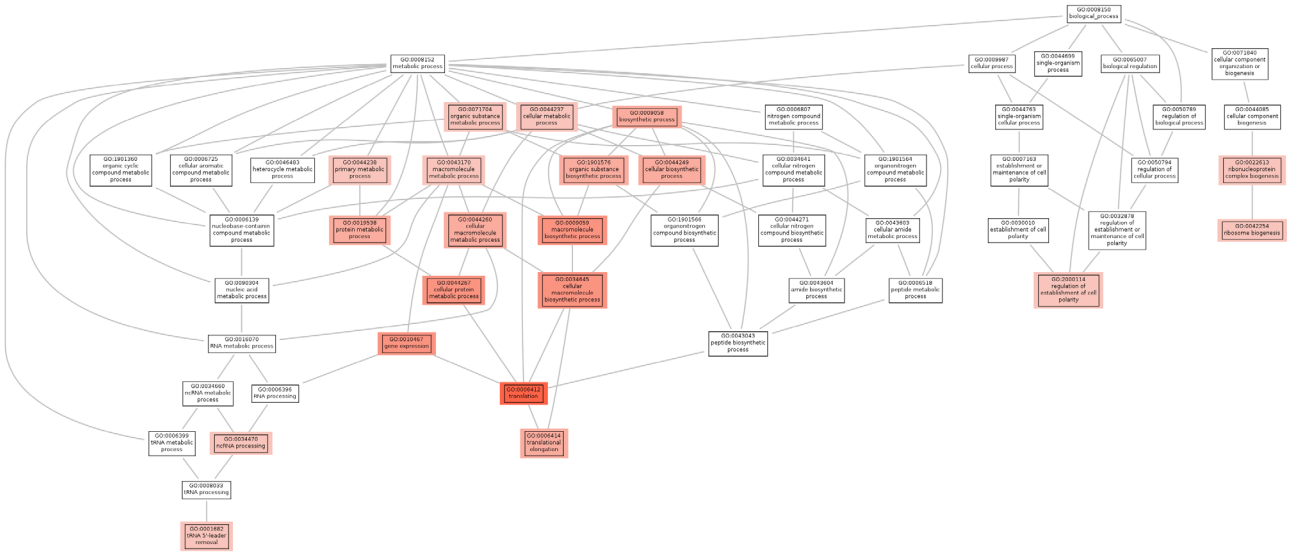


Figure 5. Along multiple detailed gene ontology tables the first 20 most significant terms are plotted in a connected network for a quick overlook. One such exemplary plot is shown above for the „biological process" gene ontology (shades of red denote strength of the respective significance).

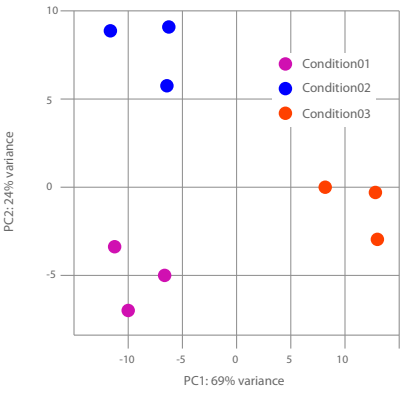


Figure 6. Differential miRNA expression for three different experimental conditions with each condition including three replicates. The expression data was submitted to principal component analysis (PCA) to show differences among replicates and conditions.

Related Topics

- siRNA synthesis service at Microsynth
- RNA-Sequencing at Microsynth
- CHiP Seq analysis pipeline at Microsynth

Further Reading

1. Griffiths, J.S., (2004) The microRNA Registry. Nucl. Acids Res. 32 (suppl 1): D109-D111.
2. Dobin et al. (2013) STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 29:15-21.
3. Love et al (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology 15: 550.
4. Edgar R.C. (2010) Search and clustering orders of magnitude faster than BLAST, Bioinformatics 26: 2460-2461.
5. Heinz et al. (2010) Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. Mol Cell 38: 576-589.
6. Mestdagh et al. (2014) Evaluation of quantitative miRNA expression platforms in the microRNRNA quality control (miRQC) study. Nat Methods. 11: 809-815.