MicroRNA Sequencing

Acute Lymphoblastic Leukemia Phase 2 (ALL P2) – miRNA Profiling

*Protocols Performed at British Columbia Cancer Agency).

MicroRNA-seq library construction

Small RNAs, containing microRNA (miRNA), in the flow-through material following mRNA purification on a MultiMACS separator (Miltenyi Biotec, Germany) are recovered by ethanol precipitation. MiRNA-seq libraries are constructed using a 96-well plate-based protocol developed at the BC Cancer Agency, Genome Sciences Centre. Briefly, an adenylated single-stranded DNA 3' adapter is selectively ligated to miRNAs using a truncated T4 RNA ligase2 (NEB Canada, cat. M0242L). An RNA 5' adapter is then added, using a T4 RNA ligase (Ambion USA, cat. AM2141) and ATP. Next, first strand cDNA is synthesized using Superscript II Reverse Transcriptase (Invitrogen, cat.18064 014), and serves as the template for PCR. Index sequences (6 nucleotides) are introduced at this PCR step to enable multiplexed pooling of miRNA libraries. PCR products are pooled, then size-selected on an in-house developed 96-channel robot to enrich the miRNA containing fraction and remove adapter contaminants. Each size-selected indexed pool is ethanol precipitated and quality checked on an Agilent Bioanalyzer DNA 1000 chip and quantified using a Qubit fluorometer (Invitrogen, cat. Q32854). Each pool is then diluted to a target concentration for cluster generation and loaded into a single lane of a HiSeq 2000 flow cell for sequencing with a 31-bp main read (for the insert) and a 7-bp read for the index.

miRNA/hg19 alignment:

Illumina miRNA sequencing reads were aligned to the hg19 reference using BWA version 0.5.7. This reference contains chromosomes 1-22, X, Y, MT, 20 unlocalized scaffolds and 39 unplaced scaffolds. Duplicated reads were marked with Picard Tools.

miRNA preprocessing, alignment and annotation

Briefly, the sequence data are separated into individual samples based on the index read sequences, and the reads undergo an initial QC assessment. Adapter sequence is then trimmed off, and the trimmed reads for each sample are aligned to the NCBI GRCh37-lite reference genome.

Routine QC assesses a subset of raw sequences from each pooled lane for the abundance of reads from each indexed sample in the pool, the proportion of reads that possibly originate from adapter dimers (i.e. a 5' adapter joined to a 3' adapter with no intervening biological sequence) and for the proportion of reads that map to human miRNAs. Sequencing error is estimated by a method originally developed for SAGE (Khattra et al., 2007).

Libraries that pass this QC stage are preprocessed for alignment. While the size-selected miRNAs vary somewhat in length, typically they are ~21 bp long, and so are shorter than the 31-bp read length. Given this, each read sequence extends some distance into the 3' sequencing adapter. Because this non-biological sequence can interfere with aligning the read to the

reference genome, 3’ adapter sequence is identified and removed (trimmed) from a read. The adapter-trimming algorithm identifies as long an adapter sequence as possible, allowing a number of mismatches that depends on the adapter length found. A typical sequencing run yields several million reads; using only the first (5’) 15 bases of the 3’ adapter in trimming makes processing efficient, while minimizing the chance that an miRNA read will match the adapter sequence.

After each read has been processed, a summary report is generated containing the number of reads at each read length. Any trimmed read that is shorter than 15bp is discarded; remaining reads are submitted for alignment to the reference genome. BWA (Li and Durbin, 2009) alignment(s) for each read are checked with a series of three filters. A read with more than 3 alignments is discarded as too ambiguous. Only perfect alignments with no mismatches are used. Reads that fail the Illumina basecalling chastity filter are retained, while reads that have soft-clipped CIGAR strings are discarded.

For reads retained after filtering, each coordinate for each read alignment is annotated using a reference databases, and requiring a minimum 3-bp overlap between the alignment and an annotation. If a read has more than one alignment location, and the annotations for these are different, we use a priority list to assign a single annotation to the read, as long as only one alignment is to a miRNA. When there are multiple alignments to different miRNAs, the read is flagged as cross-mapped (de Hoon et al., 2010), and all of its miRNA annotations are preserved, while all of its non-miRNA annotations are discarded. This ensures that all annotation information about ambiguously mapped miRNAs is retained, and allows annotation ambiguity to be addressed in downstream analyses. Note that we consider miRNAs to be cross-mapped only if they map to different miRNAs, not to functionally identical miRNAs that are expressed from different locations in the genome. Such cases are indicated by miRNA miRBase names, which can have up to 4 separate sections separated by "-", e.g. hsa-mir-26a-1. A difference in the final (e.g. ‘-1’) section denotes functionally equivalent miRNAs expressed from different regions of the genome, and we consider only the first 3 sections (e.g. ‘hsa-mir-26a’) when comparing names. As long as a read maps to multiple miRNAs for which the first 3 sections of the name are identical (e.g. hsa-mir-26a-1 and hsa-mir-26a-2), it is treated as if it maps to only one miRNA, and is not flagged as cross-mapped.

The minimum depth of sequencing required to detect the miRNAs that are expressed in one sample is 1,000,000 reads per library mapped to miRBase annotations.

Finally, for each sample, the reads that correspond to particular miRNAs are summed and normalized to a million miRNA-aligned reads to generate the quantification files.

MicroRNA Sequencing Analysis

*Protocols performed at British Columbia Cancer Agency

miRNA NMF methods

We identified groups of samples with similar abundance profiles using unsupervised non-negative matrix factorization (NMF) consensus clustering of reads-per-million (RPM) data for the 25% most-variant 5p or 3p miRBase v20 mature strands. We generated a heatmap for the discriminatory miRNAs that had the highest scores in each of the four NMF metagenes (Gaujoux and Seoighe 2010) as follows. We reordered columns (samples) in a RPM-normalized abundance matrix to match the NMF result. We $\log_2$-transformed and median-centered the rows (miRs), and then hierarchically clustered the rows using an absolute centered correlation distance metric and average linkage (de Hoon 2004, Saldanha 2004). 5p and 3p mature strand names were assigned using miRBase v20. We generated covariate association P-values with R's Fisher exact test.

miRNA-Seq-Differential expression

We used SAMseq (samr v2.0, R 2.15.0) two-class unpaired analyses with an FDR threshold of 0.05 to identify miRs that were differentially expressed. Each run generated a pair of files: miRs 'up' and 'down'. We filtered each file by removing miRs with median expression less than 50 RPKM in both of the input sample groups, and miRs for which the Wilcoxon BH adjusted P-value was greater than 0.05; then ranked the filtered results by a median-based fold change, and generated a figure showing up to 10 of the largest fold changes in each direction.