MARIO tools

Supplementary methods for

Mapping RNA-RNA interactome and RNA structures in vivo by MARIO

MARIO tools is a package of command-line tools for analyses of MARIO data. It is written in Python and R and is version controlled by GitHub. The full documentation is at The MARIO-tools software is available at http://mariotools.ucsd.edu. The pipeline takes pair-end sequencing reads as input (Supplementary Figure 9A). The oligonucleotide sequences of the RNA linker and the sample barcodes used for multiplexed sequencing should also be provided to the pipeline. The main outputs include: 1. a parsed cDNA library, including the list of chimeric cDNAs in the form of RNA1-Linker-RNA2 (see the final product in Supplementary Figures 1 and 9C), 2. the genomic locations of RNA1 and RNA2 of every chimeric cDNA (Supplementary Figure 9D), 3. interacting RNA pairs inferred from statistical enrichment of chimeric cDNAs (Supplementary Figure 9E). The analysis steps are as follows. Detailed documentation of MARIO tools is available at

1. Removing PCR duplicates

The forward read (Read1 in Supplementary Figure 9A) contains a 4nt sample barcode and a 6nt random barcode at the 5' end. A read pair was classified as a PCR duplicate of another read pair and is therefore discarded if the two read pairs had identical sequences and contained identical barcodes (10nt). The tool 'remove_dup_PE.py' provides this function, and generates a fastq/fasta file containing the non-duplicated reads, and reports the number of duplicates removed.

2. Assigning multiplexed sequencing reads into corresponding experimental samples

The tool 'split_library_pairend.py' assigns each pair-end read into a sample by matching the sample barcode in each read with those in the list of sample barcodes (a user input text file), generates a fastq/fasta file for the reads assigned to each sample, as well as a fastq/fasta file for the unassigned reads.

3. Recovering the cDNAs in the sequencing library

This step identifies the overlapping regions of the two ends of every read pair, if any. It also recovers the entire sequences of the cDNAs in the sequencing library, whenever possible.

1. If an overlap existed, this read pair was sequenced from a cDNA between 100bp and 200bp (not counting the lengths of P5 and P7) (Type 2, Supplementary Figure 32). In this case the entire sequence of the cDNA was completely covered by concatenating the forward read (Read1) with the non-overlapping region of the reverse read (Read2).
    a. If the cDNA was shorter than 100bp, we verified the presence of the P5 and the P7 primers at the two ends of the cDNA (Type 1). The ones did not contain P5 or P7 were discarded (Type 4).
2. Without an overlap, the read pair was sequenced from a cDNA longer than 200bp, whose sequence can only be partially recovered (Type 3, Supplementary Figure 32).

This function is achieved by 'recoverFragment.py', which uses local alignment to identify the overlapping regions. When the overlap was small (15bp or less) compared to read length (100bp on each end), local alignment could be insensitive. To overcome this insensitivity, 'recoverFragment.py' collects the read pairs without identifiable overlaps after the first alignment (ALIGN1, Supplementary Figure 32), truncates each read into one third of its length (retaining 33bp at the 3' of each read), and repeats local alignment (ALIGN4).

4. Parsing the chimeric cDNAs

This step categorizes the cDNAs based on their configurations (Supplementary Figure 9C). This takes the completely (Type 1 and Type 2, Supplementary Figure 32) and partially recovered (Type 3) cDNA sequences, as well as the linker sequence as inputs. It identifies the location of the linker in the cDNA, and generates five categories of cDNAs based the locations of the linker sequence, including:

1. No linker. Any Type 1 or Type 2 cDNA that does not contain the linker sequence belongs to this category. This category can be further classified into three subsets, including:
    a. Barcode only. The entire cDNA was the 10nt barcode (4nt sample barcode + 6nt random barcode), most likely results of contamination of the unligated RT primers.
    b. Single RNA. The entire cDNA was a continuous fraction of an RNA.
    c. RNA1-RNA2. These were likely produced from a proximity ligation prior to the linker ligation.

Four linker-containing categories, including:

2. RNA1-Linker-RNA2. These were generated from the desirable chimeric RNAs. Any linker-free Type 3 cDNA, whose two reads were completed aligned two distinct RNA genes, was put into this category as well. We required that both RNA1 and RNA2 sides contained at least 5bp sequences.
3. Linker-RNA2. A linker was successfully ligated to the 5' end of an RNA, but it was not succeeded by a proximity ligation.
4. RNA1-Linker. A linker was ligated to the 3' end of an RNA. This was likely generated from RNAs or RNA fragments with a 3'-OH group, or cutting off the other RNA (RNA2) from the RNA1-Linker-RNA2 chimeras during the 2nd fragmentation step.
5. LinkerOnly. The entire cDNA was a barcode and a linker sequence.

This step outputs the list of cDNAs belonged to the RNA1-Linker-RNA2 category.

5. Mapping to the genome

Hereafter, all analyses were based on the RNA1-Linker-RNA2 type of read pairs. First, any cDNA containing less than 15bp on either the RNA1 or RNA2 side of linker was discarded, because it is unlikely to uniquely map a 15bp or less sequence to the genome in the mapping step. Then the two RNA fragments on each side of the linker (RNA1 and RNA2) were separately mapped to the mouse genome mm9/NCBI37 using Bowtie version 0.12.7 [1], and parameters -f -n 1 -l 15 -e 200 -p 9 –S. This step, implemented in 'Stitch-seq_Aligner.py' outputs the read pairs where both RNA1 and RNA2 were uniquely mapped to the genome.

We tested a potentially more sensitive mapping method using Bowtie2 [2]'s "--sensitive-local" mode, with parameters "-D 15 -R 2 -N 0 -L 20 -i S,1,0.75". This "multiseed alignment" used 20bp seeds, allowing for 0 mismatches in any seed, 9bp intervals ($ceil(1 + 0.75\times\sqrt{100})$) between seeds, up to 15 consecutive seed extension attempts, and up to 2 times of "re-seeding". It turned out that this alternative strategy identified slightly fewer unique alignments than Bowtie 0.12.7. We therefore passed the Bowtie 0.12.7 results into the next steps.

6. Identifying interacting RNA pairs

The annotations were retrieved from Ensembl (release 67, mouse NCBIM37), including the genes of mRNAs, lincRNAs, rRNAs, snRNAs, snoRNAs, miRNAs, misc_RNAs, tRNAs, and transposons. The different genomic copies of the same transposon were considered as different genes in this analysis. The reads mapped to rRNAs were removed from further analysis. The number of uniquely aligned reads (from either RNA1 or RNA2 of the RNA1-Linker-RNA2 type) were counted on every gene. Any gene with a read count less than 5 was filtered out. Next, the association between any two genes was tested with Fisher's exact test. The null hypothesis was that gene A and gene B independently contributed to the

sequencing reads. The alternative hypothesis was that their contributions to read counts were associated. We denoted c_A, c_B as the read counts for gene A and gene B, respectively, and $I_{A,B}$ as the read counts of co-appearance, where the two genes co-appeared on the same read pair. A Fisher's exact test was carried out on each gene pair, with $I_{A,B}, c_A, c_B, \overline{c_A}, \overline{c_B}$ as the test statistics, where $\overline{c_A}$ ($\overline{c_B}$) was the read counts on other genes besides gene A (gene B). Both p-values and FDRs (Benjamini-Hochberg procedure [3]) were calculated for every gene pair. This step outputs gene pairs with FDR < 0.05, where FDR was estimated by Benjamin & Hochberg method. This step was implemented in 'Select_strongInteraction_RNA.py' which outputs strong interacting RNA pairs with information of their interaction regions, number of supporting pairs, p-value of significance, FDR and fold changes.

7. Identifying RNA interaction sites

We defined the RNA interaction site as a continuous RNA segment that frequently contributed to RNA-RNA interactions. RNA interaction sites were inferred from MARIO data as continuous RNA segments with multiple overlapping reads and frequent co-appearance (proximity ligation) with other RNAs. First, any continuous RNA segment covered by 5 or more uniquely aligned reads was identified as a candidate interaction site. Second, the association between any two candidate sites were tested with Fisher's exact test. The null hypothesis was that candidate sites A and gene B independently contributed to the sequencing reads. The alternative hypothesis was that their contributions to read counts were associated. We denote c_A, c_B as the read counts for candidate sites A and B, respectively, and $I_{A,B}$ as the read counts of co-appearance, where the two sites co-appeared on the same read pair. A Fisher's exact test was carried out on each site pair, with $I_{A,B}, c_A, c_B, \overline{c_A}, \overline{c_B}$ as the test statistics, where $\overline{c_A}$ ($\overline{c_B}$) was the read counts on other candidate sites besides A (B). Both p-values and FDRs (Benjamini-Hochberg procedure) were calculated for every pair of candidate sites. The candidate sites exhibiting significant associations (FDR < 0.05) were regarded as RNA interaction sites. This step was automated in 'Select_strongInteraction_pp.py' which outputs the identified RNA interaction sites.

The tool 'Plot_interaction.py' was developed for visualizing RNA interaction sites and the ligation events of these sites (Supplementary Figure 10A-B). Given any two genomic regions as input, for example the locations of two genes, this tool displays all the supporting read pairs in the form of RNA1-Linker-RNA2, where RNA1 and RNA2 were aligned to each of the two genomic locations. The linker of each RNA pair was plotted as well. This tool also plots RNA interaction sites in the input regions, if any, as well as the identified interactions between these sites.

The tool 'Plot_Circos.R' provides a global view of the RNA-RNA interactome (Supplementary Figure 10C). It plots the entire genome as a circle, and any RNA-RNA interaction as a curved line connecting two contributing genes. The interactions involving different types of RNAs are coded with different colors. The densities of RNA1 and RNA2 read fragments are displayed along with every chromosome as inner circles.

Reference

1. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology 10.
2. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9: 357-359.
3. Benjamini Y, Hochberg Y (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society 57: 289-300.