Genomics Data 2 (2014) 325–327

Contents lists available at ScienceDirect

Genomics Data

journal homepage: http://www.journals.elsevier.com/genomics-data/

Data in Brief

Genome-wide analysis of thapsigargin-induced microRNAs and their targets in NIH3T3 cells

Jody Groenendyk [a], Xiao Fan [b], Zhenling Peng [b], Yaroslav Ilnytskyy [c], Lukasz Kurgan [b], Marek Michalak [a,*]

[a] Department of Biochemistry, University of Alberta, Edmonton, Alberta T6G 2S7, Canada
[b] Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Alberta T6G 2V4, Canada
[c] Department of Biological Sciences, University of Lethbridge, Lethbridge, Alberta T1K 3M4, Canada

ARTICLE INFO

Article history:
Received 24 September 2014
Received in revised form 30 September 2014
Accepted 1 October 2014
Available online 7 October 2014

Keywords:
Endoplasmic reticulum stress
Unfolded protein response
MicroRNA
Bioinformatics

ABSTRACT

Disruption of the endoplasmic reticulum (ER) homeostasis is the cause of ER stress. We performed microRNA (miRNA) analysis (deep sequencing) to search for coping responses (including signaling pathways) induced by disrupted ER Ca^{2+} homeostasis. Our focus was on a specific branch of UPR namely the bi-functional protein kinase/endoribonuclease inositol-requiring element 1α (IRE1α). Activated IRE1α undergoes autophosphorylation and oligomerization, leading to the activation of the endoribonuclease domain and splicing of the mRNA encoding XBP1 specific transcription factor. This processing changes the coding reading frame, producing a potent transcription factor termed XBP1s. We utilized the XBP1 splicing luciferase reporter to screen for modulators of the IRE1α branch of the unfolded protein response (UPR). Here, we describe a detailed experimental design and bioinformatics analysis of ER Ca^{2+} depletion (thapsigargin treated)-induced microRNA (deep sequencing) profile. The data can be access at the Gene Expression Omnibus (GEO), the National Center for Biotechnology Information (NCBI), reference number GSE57138.

Specifications	
Organism/cell line/tissue	NIH-3T3 mouse embryonic fibroblast cells
Sequencer or array type	GAIIx (36 cycles, single-end)
Data format	Raw data: fastq files; analyzed data: txt files
Experimental factors	Wild-type NIH-3T3 cells versus NIH-3T3 cells treated with 500 nM thapsigargin for 24 h
Experimental features	RNA-seq and bioinformatics analysis
Sample source location	Edmonton, Alberta, Canada

Direct link to deposited data

Deposited data can be found at: http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE57138.

Experimental design, materials and methods

Isolation of mRNA

NIH-3T3 cells were cultured as previously published [1,2] and total RNA was harvested using TriZol (Invitrogen) as per manufacturer's protocol. Purified mRNA was measured spectrophotometrically and diluted to 1 μg/μl. Samples were frozen and sent to Biosys Inc. (Lethbridge, AB) for deep sequencing.

Sequencing of miRNA and quality control

Short reads in fastq format were demultiplexed and assembled using BclToFastq.pl script from Illumina CASAVA 1.8.1 software pipeline. For further analysis short reads were then transferred to desktop workstation with the following parameters:

Processor: IntelCorei7 CPU930 @ 2.80 GHz × 8; RAM: 8.8 Gb; operational system: 64 bit, Ubuntu 11.04 (Natty Narwahl). First, read quality was examined using FastQC program (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc).

Command: fastqc<reads.fastq>

FastQC analysis revealed high base quality in all of the libraries, which exceeded 30 on Phred scale (less than 1/1000 chance of a base being wrong). GC and sequence content were obviously skewed as compared to theoretical nucleotide distributions for vertebrate genome; however such skewed nucleotide distribution is expected in small RNA libraries characterized by low sequence diversity and a limited number of sequences occupying large fraction of the sequence pool. FastQC also matches short reads to known adapter sequences in order to detect the extent of adapter contamination, in our case only a tiny fraction (~1%) of the libraries matched RNA PCR Primer Index 1 from Illumina's small RNA library construction

* Corresponding author at: Department of Biochemistry, University of Alberta, Edmonton, Alberta T6G 2H7, Canada. Tel.: +1 780 492 2256; fax: +1 780 492 0886.
E-mail address: marek.michalak@ualberta.ca (M. Michalak).

http://dx.doi.org/10.1016/j.gdata.2014.10.002

kit. The rest of the over-represented sequences could not be matched to known adapters and PCR primers, which pointed at their cellular origin.

After high sequence quality and absence of significant adapter contamination were established, we proceeded to trim the parts of adapter sequences remaining in short reads. Common length of miRNA is 21–24 nucleotides, yet the length of the short read retained after demultiplexing of 36 cycle sequencing run is 29 (7 bases were allocated to the barcode). Therefore several nucleotides at the 3′ end of the short read belong to the sequencing adapter and have to be trimmed to enable efficient alignment to the reference (genome or small RNA collection). Adapters were trimmed at the 3′end with the Btrim program (PMID: 21651976). We considered sequences with at least 18 nt and leading N base was trimmed at the 5′ end.

Command: btrim -p<pattern.txt>-3 -f 1 -t -l 18<reads.fastq>-o<trimmed_reads.fasq> (see btrim manual for details).

FastQC was used to examine the characteristics of the sequencing libraries after trimming and to verify its efficiency. The length distribution of the sequencing libraries after adapter trimming showed that the absolute majority of the reads were in the range of 21–22 nucleotides consistent with expected miRNA length. About 20–30% of the reads were discarded by Btrim due to the minimum length of 18 nucleotide requirement. The total number of reads retained after trimming was between 2.8 and 6.0 millions.

In order to verify that the sequencing reads originate from the murine cells and align primarily to miRNAs, trimmed reads were mapped to the mouse genome (UCSC, mm9) and to the collection of mouse miRNA stem loop sequences downloaded from miRBase. The alignment was performed using bowtie [3] with one mismatch allowed.

Command: bowtie -v 1 -best -p 4 -S<bowtie_index><trimmed_reads.fastq><sam>

The absolute majority of the sequencing reads 96–98% could be mapped to the mouse genome with one mismatch allowed, similarly most of the reads 93–95% were successfully mapped to miRNA stem-loop sequences confirming that the sequencing library indeed originates from murine cells and mostly represents miRNAs. Lastly, unique reads were collapsed using Raw_data_parse program from miRExpress suite (PMID: 19821977).

Command: Raw_data_parse -i<trimmed_reads.fastq>-o<trimmed_reads.merged>

Resulting merged files are tab delimited text files with 2 columns: unique sequence and opposite column with the number of times given sequence were counted in the library (read count).

Initial annotation of known miRNAs

The collapsed reads were uploaded into miRanalyzer web-based pipeline (http://bioinfo2.ugr.es/miRanalyzer/miRanalyzer.php; PMID: 21515631) and matched to known mature miRNA (miRBase version 16), RFAM database (version 15) of known non-coding RNAs and known gene transcripts. MiRanalyzer uses bowtie short read aligner to match short sequences against the genome, known mature miRNA sequences, star miRNA sequences, miRNA generating hairpins, star sequences, which are not in miRBase and, finally, the library of other non-coding RNAs (RFAM). The corresponding sequences were uploaded from miRanalyzer website (http://bioinfo2.ugr.es/miRanalyzer/standalone.html). We used miRanalyzer with default search parameters to detect known miRNAs. The miRanalyzer's output is saved in the miRanalyzer folder with detailed information about mapping to known miRNA. Known miRNAs were divided into mature, maturestar (star sequences), maturestarunobs (star sequences not in miRBase) and hairpin. For each of the libraries there are files with unique and ambiguous mappings.

Differential expression analysis

Differential expression analysis was performed based on unique alignments to known miRNAs (mature_unique.txt file). Mature_unique.txt is the tab delimited file with the following format:

name	mature miRNA ID from miRBase;
#unique reads	number of unique reads mapped;
readCount	number of reads mapped;
norm_expressed_all	normalized to all reads;
norm_expressed_mapped	normalized to mapped reads.

miRNA expression profiling was performed using edgeR bioconductor package (PMID: 20478825). Raw read counts (readCount column in mature_unique.txt file) of uniquely mapped miRNAs were loaded into R version 2.13.0. Differentially expressed miRNAs were detected using bioconductor package edgeR 2.2.5 following the instructions provided in user manual. First, raw count data contained in mature_unique.txt files was loaded into edgeR and represented as a count matrix where row names are mature miRNA ids and columns contain read counts for every sample.

Commands: R # start R in the directory with mature_unique.txt files

> library(edgeR) # load the library

> d<- readDGE(c("CT1.txt", "CT2.txt", "T1.txt", "T2.txt"), column = c(1,3), group = c("ct", "ct", "tr", "tr")) # create count matrix, specify experimental groups

> d<- d[rowSums(d$counts) > = 5,] # remove miRNAs with low expression (blow 5 counts summed between samples)

> d<- calcNormFactors(d) # calculate normalization factors and perform TMM normalization

> d<- estimateCommonDisp(d) # estimate common dispersion

> de.com<- exactTest(d) # detect differentially expressed miRNAs, FDR is calculated using the Hochberg–Benjamini procedure by default

> de.num<- sum(de.com$table$FDR < 0.1) # get number of differentially expressed miRNAs (FDR < 0.1)

> de<- rownames(topTags(de.com, n = de.num)$table) # get the ids of differentially expressed miRNAs

> # Build MA plot, with differentially expressed miRNAs shown as red dots and straight blue lines drawn at 1 and −1 to denote 2 fold change in expression on the log scale

> plotSmear(de.com, de.tags = de, cex = 0.6)

> abline(h = c(−1,1), col="dodgerblue", lwd = 2)

> dev.off()

> write.table(de.com$table, file="diff_exp.txt", quote = F, row.names = F, col.names = T, sep="\t") # write ou the result to tab delimited file

We used TMM normalization and common dispersion (using tagwise dispersion yielded the same result). FDR was calculated according to the Hochberg–Benjamini procedure (PMID: 2218183).

Comprehensive annotation of known and putative miRNAs

The short RNAs that were generated based on the protocol described in the "Sequencing of miRNA" section were processed to find putative (novel) miRNAs and a more complete list of known miRNAs. First, these short reads were filtered. The reads that did not match the NCBI's mouse genome (using build 37.2 from http://www.ncbi.nlm.nih.gov/projects/genome/guide/mouse/) were discarded. This was based on the sequence alignment with the Bowtie program [4] assuming perfect match. In the second filtration step, the reads that matched repetitive DNAs from Repbase [5] (uploaded from http://www.girinst.org/server/RepBase/) and non-coding RNAs (tRNAs, rRNAs, snRNAs, and snoRNAs) from Rfam [6] (using build 10.0 from http://rfam.janelia.org/) were removed. Second, known miRNAs were tagged using release 17 of miRBase [7] and set aside. Third, the remaining reads were processed to find putative miRNAs. They were aligned to mRNAs using RefSeq database [8]; the matching reads were tagged as mRNA-matching. They were also aligned to the expressed sequence tags (ESTs) using dbEST [9]; the matching sequences were tagged as EST-matching. Two putative miRNA precursor sequences of the mRNA-matching, ETS-matching and the remaining short reads (one with 10 nt upstream and 70 nt downstream, assuming that miRNA is at the 5′ arm of the RNA hairpin; the with 70 nt upstream and 10 nt downstream, assuming that miRNA is at the 3′ arm) were processed by the MIREAP program (http://sourceforge.net/projects/mireap/) to analyze whether they have hairpin structure. The hairpin-like reads were folded with the RNAfold program [10] to select those with minimum free energy below −25 Kcal/mol. The protocol that utilizes MIREAP and RNAfold is consistent with related works [11–13], and performance of MIREAP was recently favorably evaluated in ([14]). The remaining short RNAs, including both miRNAs and miRNA star, were clustered to group similar reads. Each cluster corresponding to one putative miRNA and its sequence was set as the most frequent/abundant sequence in the cluster. The abundance was computed as a sum of abundance values of reads in a given cluster.

The known and putative miRNAs and miRNA star were combined together and those with the abundance (number of counts) below 5 were removed. Similar to our initial analysis, the Bioconductor package edgeR [15] was applied to find miRNAs that were differentially expressed. We utilized TMM normalization and tagwise dispersion.

Annotation of targets for selected differentially expressed microRNAs

The miRNAs were sorted by the adjusted p-values generated when annotating differentially expressed miRNAs. Known and putative target genes were computed for the miRNAs with the adjusted p-values < 0.5. The experimentally validated target genes were collected using miRecords database [16]. Since the number of experimental annotations was relatively low, we also used three target predictors: TargetScan [17,18], DIANAmicroT [19], and RepTar [20]. RepTar does not predict targets for novel miRNAs and thus we used the remaining two predictors for the putative miRNAs. Targets that were predicted by multiple methods were considered to be more reliable.

Acknowledgments

This study was supported by the Canadian Institutes of Health Research grants MOP-15291, MOP-15415, MOP-53050 to M.M.

References

[1] J. Groenendyk, et al., Interplay between the oxidoreductase PDIA6 and microRNA-322 controls the response to disrupted endoplasmic reticulum calcium homeostasis. Sci. Signal. 7 (329) (2014) ra54.
[2] J. Groenendyk, M. Michalak, Disrupted WNT signaling in mouse embryonic stem cells in the absence of calreticulin. Stem Cell Rev. 10 (2) (2014) 191–206.
[3] B. Langmead, et al., Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10 (3) (2009) R25.
[4] B. Langmead, et al., Searching for SNPs with cloud computing. Genome Biol. 10 (11) (2009) R134.
[5] J. Jurka, et al., Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res 110 (1–4) (2005) 462–467.
[6] S. Griffiths-Jones, Annotating non-coding RNAs with Rfam. Curr Protoc Bioinformatics (2005), http://dx.doi.org/10.1002/0471250953.bi1205s9 (9:12.5:12.5.1–12.5.12.).
[7] A. Kozomara, S. Griffiths-Jones, miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic Acids Res. 39 (Database issue) (2011) D152-7.
[8] K.D. Pruitt, et al., NCBI Reference Sequences: current status, policy and new initiatives. Nucleic Acids Res. 37 (Database issue) (2009) D32–D36.
[9] M.S. Boguski, T.M. Lowe, C.M. Tolstoshev, dbEST—database for "expressed sequence tags". Nat. Genet. 4 (4) (1993) 332–333.
[10] I.L. Hofacker, Vienna RNA secondary structure server. Nucleic Acids Res. 31 (13) (2003) 3429–3431.
[11] Z. Yin, et al., Genome-wide profiling of miRNAs and other small non-coding RNAs in the Verticillium dahliae-inoculated cotton roots. PLoS One 7 (4) (2012) e35765.
[12] Chen X, et al., Identification and characterization of microRNAs in raw milk during different periods of lactation, commercial fluid, and powdered milk products. Cell Res. 20 (10) (2010) 1128–1137.
[13] J. Huang, et al., Genome-wide identification of Schistosoma japonicum microRNAs using a deep-sequencing approach. PLoS One. 4 (12) (2009) e8206.
[14] Y. Li, et al., Performance comparison and evaluation of software tools for microRNA deep-sequencing data analysis. Nucleic Acids Res. 40 (10) (2012) 4298–4305.
[15] M.D. Robinson, D.J. McCarthy, G.K. Smyth, edgeR: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26 (1) (2010) 139–140.
[16] F. Xiao, et al., miRecords: an integrated resource for microRNA-target interactions. Nucleic Acids Res. 37 (Database issue) (2009) D105-10.
[17] M.R. Lewis, Radiolabeled RGD peptides move beyond cancer: PET imaging of delayed-type hypersensitivity reaction. J. Nucl. Med. 46 (1) (2005) 2–4.
[18] A. Grimson, et al., MicroRNA targeting specificity in mammals: determinants beyond seed pairing. Mol. Cell 27 (1) (2007) 91–105.
[19] M. Maragkakis, et al., DIANA-microT web server: elucidating microRNA functions through target prediction. Nucleic Acids Res. 37 (Web Server issue) (2009) W273-6.
[20] N. Elefant, et al., RepTar: a database of predicted cellular targets of host and viral miRNAs. Nucleic Acids Res. 39 (Database issue) (2011) D188–D194.