miRNA expression data

Carles Hernandez-Ferrer, Marta Vives, Mariona Bustamante, Juan Ramon González

28/06/2017

1. Introduction

MiRNA are short single-stranded non coding RNA sequences (~22 nucleotids) that regulate gene expression at post-transcriptional level.[1,2] Due to their regulatory capabilities it is thought that they are potential biomarkers for environmental exposure.
For that reason, the analysis of whole blood miRNAs is included in HELIX project, whose aim is to define the exposome during early life, associating it with children health and identifying biomarkers which explain the action mechanisms.[3]

2. Final datasets

The table below shows the final sample size after the quality control.

Table 1. *Final sample size of the miRNA datasets*

	helix_extra	helix	N
1A	0	120	120
1B	0	123	123
1X	14	821	835
N	14	1064	1078

A. mirna_subcohort_notfiltr_inclsex_v1 and **mirna_panel_notfiltr_inclsex_v1**
All QCed HELIX samples and all miRNAs (N miRNA=2,549, N samples=1,078)
Subcohort: N=955 1X + 1A
Panel: N=243 1A + 1B

B. **mirna_subcohort_notfiltr_v1** and **mirna_panel_notfiltr_v1**
All QCed HELIX samples and all miRNAs in autosomic chromosomes (N miRNA=2,380, N samples=1,078)
Subcohort: 955 1X + 1A
Panel: 243 1A + 1B

C. mirna_subcohort and **mirna_panel**
All QCed HELIX samples and miRNAs with call rate>70% in autosomic chromosomes (N miRNA=330, N samples=1,078)
Subcohort: N=955 1X + 1A
Panel: N=243 1A + 1B

<u>miRNA annotation variables:</u>
- SystematicName: mature miRNA name in Agilent array (variable "Name" in miRbase v21)
- ID: unique mirbase.accession.No for mature miRNA (mirbase.accession.No is not unique)
- MirbaseAccessionNo: mirbase.accession.No for mature miRNA (variable "Alias" in miRbase v21)
- DerivesFrom: mirbase.accession.No for the precurssor miRNA
- Gene_symbol: miRNA official gene symbol
- ActiveSequence: active sequence of the miRNA
- Length: lenght of the sequence
- StrandAgilent: strand orientation according to Agilent annotation for mature miRNA (GRCh37/hg19)
- ChromosomeAgilent: chromosome according to Agilent annotation for mature miRNA (GRCh37/hg19). It has missings.

- StartAgilent: genome start position according to Agilent annotation for mature miRNA / or probe sequence (GRCh37/hg19). It has missings.
- StopAgilent: genome final position according to Agilent annotation for mature miRNA / or probe sequence (GRCh37/hg19). It has missings.
- StrandMirbase: strand orientation according to miRBase v21 annotation for mature miRNA (GRCh38). Not consistent with strand.x.
- GRCh38Mirbase: position according to miRBase v21 annotation for mature miRNA (GRCh38)
- hg19Mirbase: position according to miRBase v21 annotation for mature miRNA (translation from GRCh38 to hg19 with UCSC)

3. Methods

RNA extraction and quality control

RNA was extracted from 1,690 HELIX samples using the MagMAX for Stabilized Blood Tubes RNA Isolation Kit (TermoFisher). They were extracted in two rounds, the first one including HELIX samples (N=1,382) and the second one including 308 extra samples from three HELIX cohorts (extra HELIX samples). The quality of RNA was evaluated with a 2100 Bioanalyzer (Agilent) and the concentration with a NanoDrop 1000 UV-Vis Spectrophotometer. We obtained 1,304 samples with good RNA quality (1,087 in the first round and 217 in the second round). Samples classified as good RNA quality had a RIN >5, a similar RNA integrity pattern in the visual inspection (bioanalyzer) and a concentration >10 ng/ul. Mean values for the RIN, concentration (ng/ul) and Nanodrop 260/230 ratio were: 7.05, 109.07 and 2.15.

miRNA laboratory processing

Expression miRNA levels of 1,087 samples with good RNA quality were analysed using the SurePrint Human miRNA Microarray rel. 21 (Agilent) at the Genomics Core facility at the Centre for Genomic Regulation (CRG). Samples were randomized by sex and cohort and samples of the same subject (panel study) were processed in the same batch and array. Batches consisted of 24 samples which were hybridized onto 3 slides (8 samples per slide). One control RNA (a mixture of RNAs from several human tissues: universal miRNA reference kit) was included in 2/3 of the batches (N=39 control samples).

Samples were processed following Agilent's recommendations. Briefly, RNA samples were concentrated or evaporated in order to reach the required concentration using a vacuum equipment (SpeedVac). The miRNA Complete Labeling and Hyb kit generates fluorescently-labeled miRNA with a sample input of 100 ng of total RNA. This method involves the ligation of one Cyanine 3-pCp molecule to the 3' end of a RNA molecule. Agilent SurePrint G3 Human miRNA microarrays were hybridized following the Agilent Microarray Hybridization Chamber User Guide. Raw data and GeneView files were extracted with the Feature Extraction software (Agilent). Samples that did not pass the laboratory quality control parameters were repeated (N=52). miRNAs were annotated with the Annotation_70156 version from Agilent and with additional information from mirBase v21 (http://www.mirbase.org/).

Normalization and quality control of the miRNA expression data

We performed an initial quality control which included: check of the Agilent quality control parameters, check of the number of miRNAs in duplicated samples, and calculation of the sample and miRNA call rate. Agilent considers that miRNAs are not detected when the expression signal is not different from the background or the standard error of the different probes is >3 times higher than the expression signal. At a 70% call rate 360 miRNAs were detected in HELIX samples. Seventy-two samples with low quality were excluded (6.1%). None sample had sex discrepancies. Final sample size was 1,078.

We tested three normalization methods (90th percentile, 90th percentile excluding undetected miRNAs and LVS) using the same control RNA sample included in 39 arrays. The LVS method identifies a subset of miRNAs with the smallest array-to-array variation which then are used to normalize the miRNA expression values among samples (default proportion of constant miRNA expression was set to 0.7). The least variant set (LVS) method[6] with background correction using the Normexp method in limma package[7] slightly outperformed the others and was selected for the normalization of HELIX samples. The LVS method was applied on the HELIX samples that had passed the QC. For the identification of housekeeping miRNAs a random sample of 50 HELIX samples was used. After normalization, miRNAs with a call rate <70% and miRNAs in sexual chromosomes were filtered out. The final dataset consisted of 1,078 samples and 330 miRNAs.

4. References

1. Esteller, M. Non-coding RNAs in human disease. *Nat Rev Genet* **12,** 861–874 (2011).
2. Pritchard, C. C., Cheng, H. H. & Tewari, M. MicroRNA profiling: approaches and considerations. *Nat. Rev. Genet.* **13,** 358–369 (2012).
3. Hou, L., Wang, D. & Baccarelli, A. Environmental chemicals and microRNAs. *Mutat. Res. - Fundam. Mol. Mech. Mutagen.* **714,** 105–112 (2011).
4. Mestdagh, P. *et al.* Evaluation of quantitative miRNA expression platforms in the microRNA quality control (miRQC) study. *Nat. Methods* **11,** 809–815 (2014).
5. Akhtar, M. M., Micolucci, L., Islam, M. S., Olivieri, F. & Procopio, A. D. Bioinformatic tools for microRNA dissection. *Nucleic Acids Res.* **44,** 24–44 (2016).
6. Suo, C., Salim, A., Chia, K.-S., Pawitan, Y. & Calza, S. Modified least-variant set normalization for miRNA microarray. *RNA* **16,** 2293–2303 (2010).
7. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43,** e47 (2015).
8. Leek, J. T. & Storey, J. D. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.* **3,** 1724–1735 (2007).