Transcriptomics data. description

Carles Hernández, Marta Vives, Inés Quintana, Mariona Bustamante, Juan Ramon González

Version 1 (25/04/2017)

Summary

The transcriptome is the set of all RNA molecules, including mRNA, rRNA, tRNA, and other non-coding RNA produced in one cell or in a population of cells {Mattick, 2010 #20}. The transcriptome is dynamic and cell and time specific. Previous studies have linked coding RNA to environmental exposures, and only few studies presented data on non-coding RNA such as long non-coding RNAs {Karlsson, 2016 #92}.

In HELIX, we have analysed the coding and non-coding transcriptome from whole blood.

Final datasets

1. Final sample size

The final dataset consists of 1289 samples distributed as follows:

Period	HELIX		Total
	no	yes	
1X	151	889	1040
1A	0	121	121
1B	1	127	128
Total	152	1137	1289

Subcohort.all	1161
Subcohort.HELIX	1010
Panel.paired	105

Overlap with methylation data:

- 1137 HELIX samples: 1047 with methylation data (90 missing)
- 1010 HELIX subcohort samples: 921 with methylation data (89 missings)
- 1289 HELIX and extra HELIX samples: 1060 with methylation data (229 missing)
- 1161 HELIX and extra HELIX subcohort samples: 933 with methylation data (228 missings)

2. Final probe size

Two datasets were created:

- **Not filtered:** with the exception of control probes, probes not annotated to a chromosome, and sexual probes (N final = 64842)

All transcripts

	Coding	Non coding	Total
Assigned to a gene	26300	5733	32033
Not assigned to a gene	16574	16235	32809
Total	42874	21968	64842

*2809 without information of locus type

*Unique number of genes: 25114 out of the 32033

genes Small transcripts (<200 nt)

	Coding	Non coding	Total
Assigned to a gene	20055	4622	24677
Not assigned to a gene	14187	13508	27695
Total	34242	18130	52372

*Link to a miRNA gene (MIR) : 1103

Filtered: control probes, probes not annotated to a chromosome, sexual and call rate <70% (N final = 35841)

All transcripts

	Coding	Non coding	Total
Assigned to a gene	18604	3449	22053
Not assigned to a gene	7100	6688	13788
Total	25704	10137	35841

*Unique number of genes: 17841 out of the 22053 genes

Small transcripts (<200 nt)

	Coding	Non coding	Total
Assigned to a gene	745	4	749
Not assigned to a gene	3254	4	3259
Total	3999	8	4007

*Link to a miRNA gene : 522

3.Final ExpressionSets

Four ExpressionSet files containing normalized gene expression levels were created:

- transcriptome_subcohort_v1.RData: 1,161 1X + 1A samples and 35,841 probes
- transcriptome_panel_v1.RData: 249 1A+1B samples and 35,841 probes
- transcriptome_subcohort_notfiltr_v1.RData: 1,161 1X+1A samples and 64,842 probes
- transcriptome_panel_notfiltr_v1.RData: 249 1A+1B samples and 64,842 probes

ExpressionSets with all probes (except for control probes) are:
- 20170419_transcriptome_panel_notfitr_inclX.RData
- 20170419_transcriptome_subcohort_notfitr_inclX.RData

An ExpressionSet is an R object that contains:
- values of the gene expression levels
- annotation of the omics features based on Affymetrix annotation + 2 additional variables created by HELIX researchers
- metadata of the samples (laboratory variables (plate) and biological variables)

For more information on ExpressionSet visit https://www.bioconductor.org/packages/devel/bioc/vignettes/Biobase/inst/doc/ExpressionSetIntroduction.pdf.

Summary of 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.

Transcriptomics

Gene expression, including coding and non-coding transcripts was assessed with the Affymetrix Human Transcriptome Array 2.0 ST arrays (HTA 2.0). Amplified and biotynylated sense-strand DNA targets were generated from total RNA of the 1,304 samples with good RNA quality. Affymetrix HTA 2.0 arrays were hybridized according to Affymetrix recommendations using the Manual Target preparation for GeneChip Whole Transcript (WT) expression arrays (Affymetrix) and the Affymetrix labeling and hybridization kits. Samples were processed in two different rounds (HELIX and extra HELIX samples) at the University of Santiago de Compostela (USC). In each round, several batches of 24-48 samples were processed. Samples were randomized within each batch taking into account sex and cohort. Samples from the same subject (panel study) were processed in the same batch. Two different types of control RNA samples (HeLa or FirstChoice® Human Brain Reference RNA) were included in each batch, but they were hybridized only in the first batches. No HELIX duplicates were included in the study. Raw data were extracted with the AGCC software (Affymetrix) and stored into CEL files. Ten samples failed during the laboratory process (7 not enough cRNA or ss-cDNA, 2 low fluorescence, and 1 with an artefact in the CEL file). One extra sample was eliminated as it did not had the ethic consent. One thousand two hundred and ninety three samples (1,083 HELIX + 210 extra HELIX) passed the laboratory quality control.

Normalization and quality control of the trancriptomics data

Data was normalized by Affymetrix with the GCCN (SST-RMA) algorithm at the gene and transcript level. Annotation to transcripts clusters was done with the ExpressionConsole software using the HTA-2_0 Transcript Cluster Annotations Release na36. After normalization several quality control checks were performed and four samples with discordant sex were excluded. Control probes and probes, probes in sexual chromosomes or probes without chromosome information were excluded. Probes with a DABG (Detected Above Background) p

value < 0.05 were considered to have an expression level different from the background, and they were defined as detected. Probes with a call rate <70% were excluded from the analysis. Although we only observed one cluster of samples in the Principal Component Analysis (PCA), there was some grouping of samples within the cluster by cohort and by technical variables. To control for this potential confounding by technical bias, the surrogate variable analysis (SVA) method {Leek, 2007 #93} was applied during the statistical analysis.

References

1. Buckberry, S., Bent, S. J., Bianco-Miotto, T. & Roberts, C. T. MassiR: A method for predicting the sex of samples in gene expression microarray datasets. *Bioinformatics* **30,** 2084–2085 (2014).

2. Leek, J. T. & Storey, J. D. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.* **3,** 1724–1735 (2007).

3. Gaujoux, R. & Seoighe, C. CellMix: a comprehensive toolbox for gene expression deconvolution. *Bioinformatics* **29,** 2211–2212 (2013).

4. Abbas, A. R., Wolslegel, K., Seshasayee, D., Modrusan, Z. & Clark, H. F. Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. *PLoS One* **4,** e6098 (2009).