Inferring microRNA-mRNA causal regulatory relationships from expression data

Supplementary File 1

Thuc Duy Le[1,*], Lin Liu[1], Anna Tsykin[2], Gregory J Goodall[2,3,4], Bing Liu[5], Bing-Yu Sun[6], and Jiuyong Li[1,*]

[1] School of Information Technology and Mathematical Sciences, University of South Australia, Mawson Lakes, SA 5095, AU

[2] Centre for Cancer Biology, SA Pathology, Adelaide, SA 5000, AU

[3] School of Molecular and Biomedical Science, University of Adelaide, Adelaide, SA 5005, AU

[4] Department of Medicine, University of Adelaide, Adelaide, SA 5005, AU

[5] Children's Cancer Institute Australia, Randwick, NSW 2301, AU

[6] Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei, Anhui 230031, China

In this file, we describe the data used in the paper, provide the supported result, and detail the contents of the Supplementary files.

1. Data

The dataset used in the paper is the NCI-60 dataset for EMT. The dataset includes the miRNA expression profiles for the NCI-60 panel of 60 cancer cell lines from Søkilde et al. (Sø kilde et al., 2011), available from http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE26375 .

The mRNA expression profiles for NCI60 were downloaded from ArrayExpress http://www.ebi.ac.uk/arrayexpress, accession number E-GEOD-5720. Cell lines categorised as epithelial (11 samples) and mesenchymal (36 samples) were used for this work.

Epithelial-to-mesenchymal transition (EMT) is part of the processes of tissue remodeling during embryonic development and wound healing (Savagner, 2001), and during carcinogenesis (Dvorak, 1986) when cancer cells undergo a change and transform into a more invasive tumor (Savagner, 2001; Fuchs et al., 2002).

After EMT induction, cells lose their epithelial features characterised by the high E-cadherin expression level, and acquire mesenchymal characteristics, including Vimentin filaments and a flattened phenotype. By expressing proteases, cells become more invasive, and they can pass through the underlying basement membrane and migrate. These are crucial steps in the multistep process of metastasis (Park, 2008).

The 60 cell lines of the drug screening panel of human cancer cell lines acquired from the National Cancer Institute (NCI60) represent nine different types of cancers (http://dtp.nci.nih.gov, and they can be genetically divided into two major clusters: epithelial and mesenchynal.

Park et al. (Park, 2008) determined the expression levels of E-cadherin and Vimentin in 59 of the NCI60 cells, using Western blot analysis followed by densitometry (Fig. 1A). These 59 NCI60 cell lines are then sorted from the highest to the lowest E-cadherin/Vimentin ratio (Fig. 1B). Cells with very high E-cadherin expression and no detectable Vimentin expression will be placed in the epithelial group. Likewise, the mesenchymal group includes cell with high Vimentin and no detectable E-cadherin expression. Cells that have either both markers expressed at the same level, or none of these two markers expressed will be placed in the undefined group.

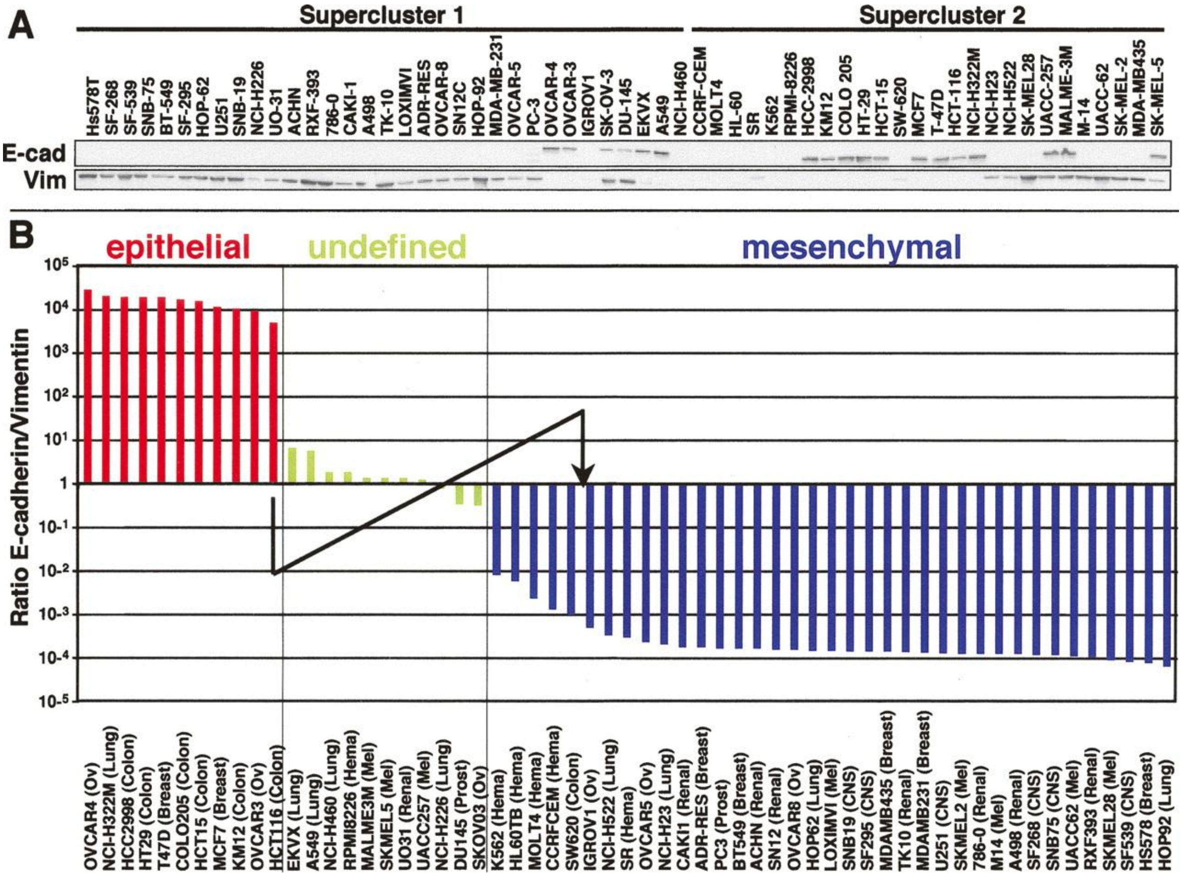


Figure1. **A.** The E-cadherin and Vimentin expression levels of cells in NCI-60. **B.** Samples are grouped into 3 categories based on the ratio of E-cadherin/Vimentin: epithelial, undefined, and mesenchymal.

mRNA analysis

We download the raw data (CEL files) from Arrayexpress and use Bioconductor in R to process the data. We then use the RMA (Robust Multichip Average) method to pre-process the data. The RMA method consists of three steps: background adjustment, quantile normalisation, and summarisation. We use the default setting of rma() command that implemented the RMA method in Biocondutor (http://www.bioconductor.org/). Further details of the RMA method can be found in (Bolstad et al. 2003, Rafael et al. 2003, Irizarry et al. 2003).

In the final step we use empirical Bayes, moderated *t*-statistics implemented in the *limma* package from Bioconductor (Smyth, 2005) to identify the differentially expressed genes. Linear models were fitted to the data, and comparisons of interest were extracted as contrasts. To control the false discovery rate we use the Benjamini & Hochberg method (Benjamini & Hochberg, 1995) implemented in *limma.* There are 1635 mRNA probes identified to be differentially expressed with p-value < 0.05 (adjusted p-value).

miRNA analysis

The dataset includes the miRNA expression profiles for the NCI-60 panel of 60 cancer cell lines from Søkilde et al. (Søkilde et al., 2011). Søkilde et al describe the low-level analyses as the following.

"All low-level analyses, such as importing and pre-processing of the data, were done in the R environment by using the *limma* package. After excluding flagged spots from the analysis, the "normexp" background correction method, plus offset = 50, was applied, after which intensities were $\log_2$ transformed and quantile normalized as implemented in *limma*. Both $\log_2$ intensities (single-channel analysis) and $\log_2$ ratios (dual channel analysis) of 4 intraslide replicates were averaged. All the expression data were deposited in the Rosetta Resolver (Rosetta Biosoftware) data management system."

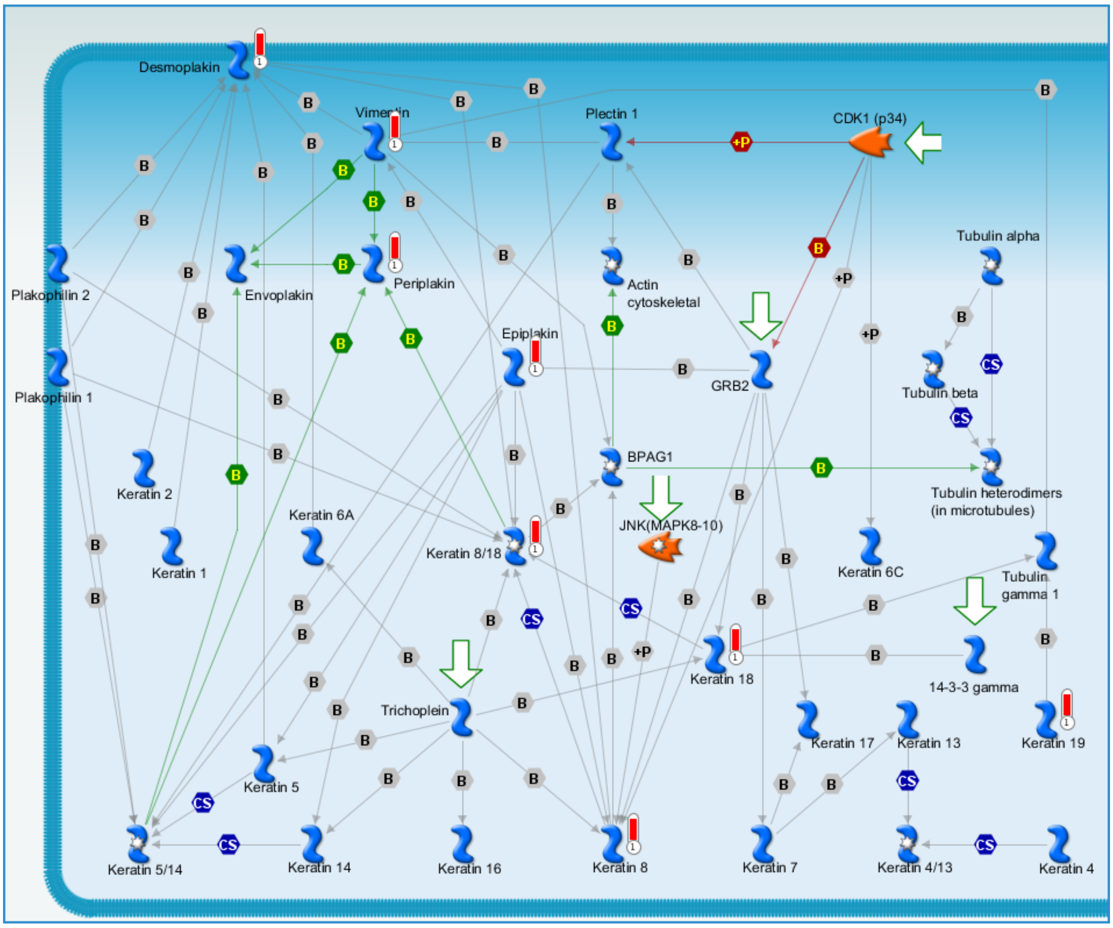
To identify miRNAs that were differentially expressed, we again use the *limma* package to perform the test with Benjamini & Hochberg correction method. There are 43 miRNAs differentially expressed with the adjusted p-values<0.05.

Inferring miRNA-mRNA causal relationships

The differentially expressed mRNAs and miRNAs are integrated into a dataset with 47 samples and 1678 variables. Please refer to the supplementary files 5 and 6 for details of inferring miRNA-mRNA causal effects.

2. Results

As mentioned in the paper, the following is the pathway of *cytoskeleton remodeling_keratin filaments*. The red bars represent the genes in the top 150 list.



3. Supplementary data files

3.1. Supplementary data file 1

Supplementary data file 1 shows the differentially expressed miRNAs and mRNAs from different conditions, Epithelial and Mesenchymal. Using *limma* package from bio-conductor (Smyth, 2005), we are able to identify 1635 mRNA probes and 43 miRNA probes differentially expressed with *p*-value<0.05 (adjusted *p*-value).

3.2. Supplementary data file 2

Supplementary data file 2 indicates the causal effect values that miRNAs have on mRNAs. The higher the values are, the stronger the causal effects. The positive/negative numbers represent the up/down regulations accordingly. We provide the causal effect results for all miRNAs in Sheet1, and in Sheet 2 are the extracted causal effect values for miR-200 family. The results in Sheet 2 will then be used to validate against experimental results. Sheet 3 is the causal effects between each miRNA and all mRNAs. The results in Sheet 3 are ranked based on the absolute values of the causal effects. This would be the resources for further experimental validation or research.

3.3. Supplementary data file 3

Supplementary data file 3 reports the experimental results. The coded headers are explained below:

PG_231_200a_1: Expression values of genes in the MDA-MB-231 sample transfected with miR-200a.

PG_231_200b_1 and **PG_231_200b_2**: Expression values of genes in the MDA-MB-231 sample transfected with miR-200b.

PG_231neg_1 and **PG_231neg_2:** Expression values of genes in the MDA-MB-231 that are not transfected with miR-200a and miR-200b (controlled samples).

3.4. Supplementary data file 4

Supplementary data file 4 shows the validation results. The genes in the top 20, 50, and 100 that have been confirmed by controlled experiments are reported.

3.5. Supplementary data file 5

Supplementary data file 5 is the R script (miRCausality.R) to infer the causal effects that miRNAs have on mRNAs. You should save this file into your local machine, and then use the commands and instructions (in the comments) in the Supplementary data file 6 to run the program.

3.6. Supplementary data file 6

While the Supplementary data file 5 is the R script that you should not modify, the Supplementary data file 6 details all the steps to run the program where you can change the parameters.

References

Sø kilde, R., Kaczkowski, B., and Podolska, A. (2011). Global microRNA Analysis of the NCI-60 Cancer Cell Panel. Mol Cancer Ther, 10, 375–384.

Savagner, P.(2001) Leaving the neighborhood: Molecular mechanisms involved during epithelial–mesenchymal transition. Bioessays 23:912–923.

Dvorak, H.F.(1986) Tumors: Wounds that do not heal. Similarities between tumor stroma generation and wound healing. N. Engl. J. Med. 315:1650–1659.

Fuchs, I.B., Lichtenegger, W., Buehler, H., Henrich, W., Stein, H., Kleine-Tebbe, A., Schaller, G.(2002) The prognostic significance of epithelial–mesenchymal transition in breast cancer. Anticancer Res. 22:3415–3419.

Park, S.M., Gaur, A.B., Lengyel, E. and Peter, M.E. (2008). The miR-200 family determines the epithelial phenotype of cancer cells by targeting the E-cadherin repressors ZEB1 and ZEB2. Genes Dev., 22, 894–907.

Smyth, G. K. (2005). Limma : Linear Models for Microarray Data. Bioinformatics and Computational Biology Solutions using R and Bioconductor, pages 397–420.

Bolstad, B.M., Irizarry R. A., Astrand M., and Speed, T.P. (2003), A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Bias and Variance. *Bioinformatics* 19(2):185-193.

Rafael. A. Irizarry, Benjamin M. Bolstad, Francois Collin, Leslie M. Cope, Bridget Hobbs and Terence P. Speed (2003), Summaries of Affymetrix GeneChip probe level data *Nucleic Acids Research* 31(4):e15.

Irizarry, RA, Hobbs, B, Collin, F, Beazer-Barclay, YD, Antonellis, KJ, Scherf, U, Speed, TP (2003) Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data. *Biostatistics* .Vol. 4, Number 2: 249-264.

Benjamini Y, Hochberg Y: Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. Series B (Methodological) 1995, 57:289–300.