Supplementary Material for 'A Non-negative Factorization Method for Detecting Modules in Heterogeneous Omics Multi-modal Data'

Zi Yang [1] and George Michailidis [1]

[1]Department of Statistics, University of Michigan, Ann Arbor, MI 48109, US

S1. Derivation of the iNMF algorithm. We show here that the multiplicative updates used to solve iNMF ensure that the objective function $\mathcal{F}(W, H, V)$ is monotonically decreasing:

$$\mathcal{F}(W, H, V) = \sum_k \|X_k - (W + V_k)H_k\|_F^2 + \lambda \sum_k \|V_k H_k\|_F^2.$$

All quantities X_k, W, H_k, V_k are as defined in the main article. For convenience, we use H and V to denote $\{H_1, \ldots, H_K\}$ and $\{V_1, \ldots, V_K\}$, respectively.

1. The bulk of the proof involves auxilliary functions and some algebraic manipulation, but an application of duality theory reveals some useful relations. The corresponding dual problem of iNMF is:

$$\begin{aligned} \max_{\Theta} \inf_{W,H,V} \quad & \mathcal{F}(W, H, V) + \operatorname{tr}(\Phi W^T) + \sum_k \operatorname{tr}(\Psi_k H_k^T) + \sum_k \operatorname{tr}(\Xi_k V_k^T) \qquad (1) \\ \text{subject to:} \quad & \Phi \geq 0, \Psi_k \geq 0, \Xi_k \geq 0, k = 1, \ldots, K, \end{aligned}$$

where $\Theta = \{\Phi, \Psi_1, \ldots, \Psi_K, \Xi_1, \ldots, \Xi_K\}$ are matrices whose elements are the Lagrangian multipliers for the elements of $\{W, H_1, \ldots, H_K, V_1, \ldots, V_K\}$, respectively. By definition, we have $\Phi \in \mathbb{R}^{N \times D}$, $\Psi_k \in \mathbb{R}^{D \times M_k}$, and $\Xi_k \in \mathbb{R}^{N \times D}$ for all $k = 1, \ldots, K$.

From the first order conditions of the Lagrangian function in (1), we may solve for the Lagrangian multipliers:

$$\begin{aligned} \Phi &= 2 \sum_k \left(X_k H_k^T - (W + V_k) H_k H_k^T \right) \\ \Psi_k &= 2 \left((W + V_k)^T X_k - (W + V_k)^T (W + V_k) H_k - \lambda V_k^T V_k H_k \right), \ \ k = 1, \ldots, K \\ \Xi_k &= 2 \left(X_k H_k^T - (W + V_k) H_k H_k^T - \lambda V_k H_k H_k^T \right), \ \ k = 1, \ldots, K. \end{aligned}$$

By the complementary slackness property, we have the following relations at the optimal solution for all indices (i, j):

$$\begin{aligned} & W_{ij} \sum_k \left(X_k H_k^T - (W + V_k) H_k H_k^T \right)_{ij} = 0 \\ & (H_k)_{ij} \left((W + V_k)^T X_k - (W + V_k)^T (W + V_k) H_k - \lambda V^T V H_k \right)_{ij} = 0, \ \ k = 1, \ldots, K \\ & (V_k)_{ij} \left(X_k H_k^T - (W + V_k) H_k H_k^T - \lambda V_k H_k H_k^T \right)_{ij} = 0, \ \ k = 1, \ldots, K. \end{aligned}$$

These relations lead to our multiplicative updates after some algebraic manipulation.

2. The last portion of the proof involves auxiliary functions, defined below:

   Definition. *$G(h, h')$ is an auxiliary function for $F(h)$ if the following are satisfied:*

   $$G(h, h') \geq F(h) \ \forall h,$$

   $$G(h, h) = F(h).$$

   Auxiliary functions have the following property:

   Lemma 1. *If G is an auxiliary function for F, and $h^{(t+1)} = \arg\min_h G(h, h^{(t)})$, then*

   $$F(h^{(t+1)}) \leq F(h^{(t)}).$$

   Proof. $F(h^{(t+1)}) \leq G(h^{(t+1)}, h^{(t)}) \leq G(h^{(t)}, h^{(t)}) = F(h^{(t)})$. □

   If G is easier to minimize than F, then we may take repeated iterations of $h^{(t+1)} = \arg\min_h G(h, h^{(t)})$ instead of directly dealing with F.

3. Each of the iNMF updates may be derived with an appropriate auxiliary function. We outline here only the derivation for V_k update, but the updates for W, H_k are similarly derived. Since the updates are performed element-wise, it is enough to show that the update $(V_k)_{ij}^{(t+1)}$ satisfies:

   $$\mathcal{F}((V_k)_{ij}^{(t+1)}) \leq \mathcal{F}((V_k)_{ij}^{(t)}). \tag{2}$$

   The first two derivatives of $\mathcal{F}$ with respect to $(V_k)_{ij}$ are:

   $$\begin{aligned} \mathcal{F}'_{ij} &= \mathcal{F}'((V_k)_{ij}) = \left(-2X_k H_k^T + 2(W + V_k)H_k H_k^T + 2\lambda V_k H_k H_k^T\right)_{ij} \\ \mathcal{F}''_{ij} &= \mathcal{F}''((V_k)_{ij}) = 2(1 + \lambda)\left(H_k H_k^T\right)_{jj}. \end{aligned}$$

   Lemma 2. *The function:*

   $$\mathcal{G}(h, (V_k)_{ij}) = \mathcal{F}((V_k)_{ij}) + \mathcal{F}'((V_k)_{ij})(h - (V_k)_{ij}) + \frac{\left((W + V_k + \lambda V_k)H_k H_k^T\right)_{ij}}{(V_k)_{ij}}(h - (V_k)_{ij})^2,$$

   is an auxiliary function for $\mathcal{F}$.

Proof. $\mathcal{G}((V_k)_{ij}, (V_k)_{ij}) = \mathcal{F}((V_k)_{ij})$ is easy to see. To show that $G(h, (V_k)_{ij}) \geq \mathcal{F}(h)$, we write out the Taylor expansion of $\mathcal{F}$ at $(V_k)_{ij}$:

$$\mathcal{F}(h) = \mathcal{F}((V_k)_{ij}) + \mathcal{F}'((V_k)_{ij})(h - (V_k)_{ij}) + (1+\lambda)\left(H_k H_k^T\right)_{jj}(h - (V_k)_{ij})^2.$$

Thus, it is sufficient to show that:

$$\frac{\left((W + V_k + \lambda V_k)H_k H_k^T\right)_{ij}}{(V_k)_{ij}} \geq (1+\lambda)\left(H_k H_k^T\right)_{jj}.$$

By nonnegativity of the matrix factors, we have:

$$\begin{aligned}
\frac{\left((W + V_k + \lambda V_k)H_k H_k^T\right)_{ij}}{(V_k)_{ij}} &\geq (1+\lambda)\frac{\left(V_k H_k H_k^T\right)_{ij}}{(V_k)_{ij}} \\
&= (1+\lambda)\frac{\sum_l (V_k)_{il}\left(H_k H_k^T\right)_{lj}}{(V_k)_{ij}} \\
&= (1+\lambda)\left(H_k H_k^T\right)_{jj}.
\end{aligned}$$

□

Combining Lemmas 1 & 2, we have that the update:

$$(V_k)_{ij}^{(t+1)} = \arg\min_h G(h, (V_k)_{ij}^{(t)}),$$

guarantees (2). But this minimizer can be expressed as:

$$\begin{aligned}
\arg\min_h G(h, (V_k)_{ij}^{(t)}) &= (V_k)_{ij} - (V_k)_{ij}\frac{\mathcal{F}'((V_k)_{ij})}{2\left((W + V_k + \lambda V_k)H_k H_k^T\right)_{ij}} \\
&= (V_k)_{ij} - (V_k)_{ij}\frac{\left(-2X_k H_k^T + 2(W + V_k)H_k H_k^T + 2\lambda V_k H_k H_k^T\right)_{ij}}{2\left((W + V_k + \lambda V_k)H_k H_k^T\right)_{ij}} \\
&= (V_k)_{ij}\frac{\left(X_k H_k^T\right)_{ij}}{\left((W + V_k + \lambda V_k)H_k H_k^T\right)_{ij}},
\end{aligned}$$

which is exactly our iNMF update for $(V_k)_{ij}$.

S2. Intuition for the tuning selection procedure.

We discuss here the intuition behind the stopping threshold $R_I^{(\lambda)} - R_J > 2(R_J - R_S)$ from the tuning selection procedure. Let $X_k, k = 1, \ldots, K$ be observationally linked data sets, and let X_k^S, X_k^J, X_k^I be the approximating solutions of sNMF, jNMF, and iNMF:

$$X_k^S = W_k^S H_k^S, \quad X_k^J = W^J H_k^J, \quad X_k^I = (W^I + V_k^I)H_k^I.$$

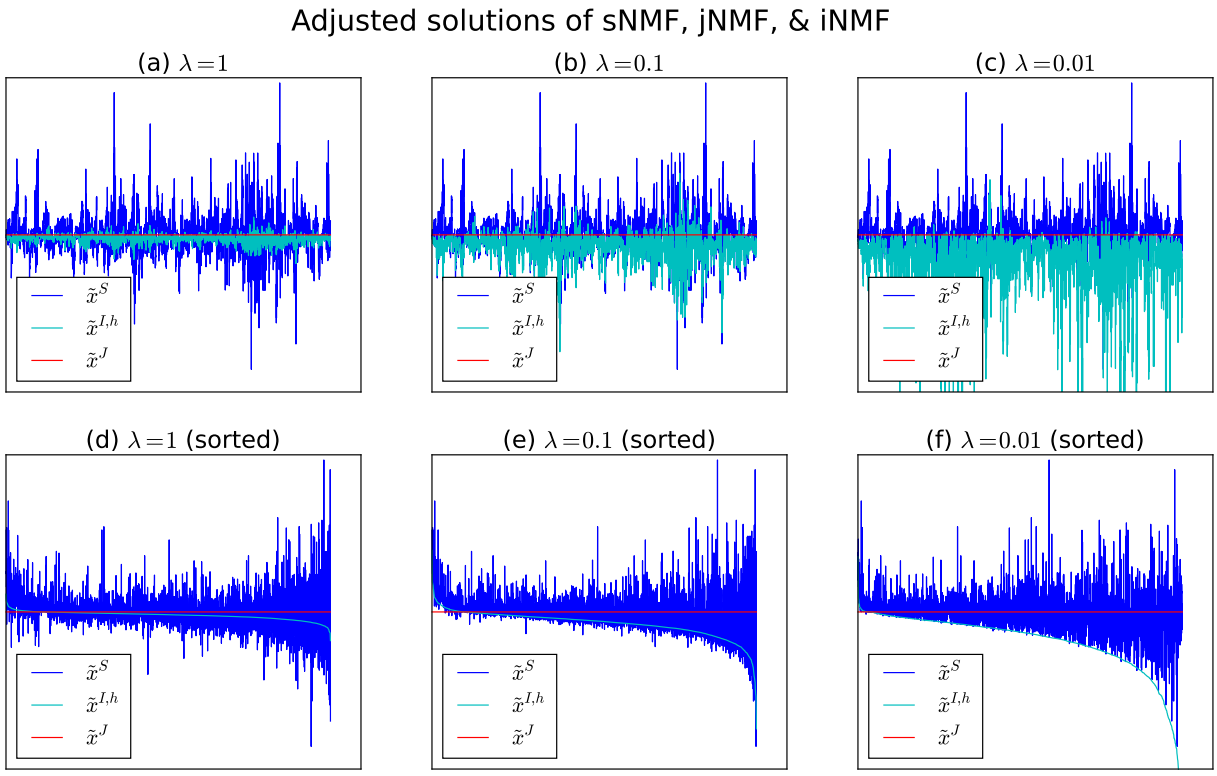

Figure S1: Adjusted sNMF, jNMF, and iNMF solutions with different λ choices for iNMF, computed from generated data ($\sigma_u, \sigma_s, \sigma_h = (0.01, 0.2, 0.01)$). First row: unsorted; second row: sorted with respect to the adjusted iNMF solution.

Suppose that we adjust these solutions entry-wise with respect to the jNMF solution:

$$\tilde{x}^S_{k,ij} = x^S_{k,ij} - x^J_{k,ij}, \quad \tilde{x}^J_{k,ij} = x^J_{k,ij} - x^J_{k,ij}, \quad \tilde{x}^{I,h}_{k,ij} = (W^I H^I_k)_{k,ij} - x^J_{k,ij}.$$

Note that for iNMF we consider only the homogeneous portion. We will omit subscripts for the sake of brevity.

Figure S1 plots the entries of the adjusted solutions $\tilde{x}^S, \tilde{x}^{I,h}, \tilde{x}^J$ computed from simulated data (see Supplementary Section S3) over different choices of λ for iNMF. Naturally, $\tilde{x}^S$ (sNMF) are centered around $\tilde{x}^J$ (jNMF). Also, $\tilde{x}^{I,h}$ (iNMF, homogeneous) generally lie below $\tilde{x}^J$ (jNMF), since the other heterogeneous portion of iNMF is nonnegative. As the choice of λ shrinks, the iNMF solution becomes less homogeneous and $\tilde{x}^{I,h}$ becomes less resembling of $\tilde{x}^J$. When λ is small enough, iNMF begins to overfit the data. Our tuning selection procedure selects $\lambda = 0.1$ for this particular example, which in fact leads to optimal performance.

When we sort the adjusted solutions by $\tilde{x}^{I,h}$, we see some interesting relations. At the optimal λ (Figure S1e), the iNMF homogeneous solutions $\tilde{x}^{I,h}$ lie slightly above the minimum of the distribution of the sNMF solutions $\tilde{x}^S$. If the level of $\tilde{x}^{I,h}$ had been higher (Figure S1d), then the full iNMF solution would deviate from the sNMF solution, and hence yield a poor approximation of the data. If the level of $\tilde{x}^{I,h}$ had been lower (Figure S1f), then the

approximation accuracy of iNMF will be slightly improved at the expense of losing detection of the joint signal. In principle, the optimal iNMF solution must (1) achieve good fit on the data and (2) maximize the homogeneous portion used to achieve that fit.

Now consider the distributions of the unsorted adjusted solutions (Figure S1a-c). What is notable about the iNMF solutions $\tilde{x}^{I,h}$ under optimal $\lambda = 0.1$ (Figure S1b) is that its distribution appears to match the lower half of the distribution of $\tilde{x}^{S}$. Similar to before, this is a distinguishing feature of an optimal iNMF solution. Another way of describing this is to say that the deviation between the iNMF (under optimal λ) and jNMF solutions is roughly twice the deviation between the jNMF and sNMF solutions. In fact, our stopping threshold $R_I^{(\lambda)} - R_J > 2(R_J - R_S)$ takes advantage of precisely this relation. As the selection procedure iteratively evaluates choices of λ, it effectively tunes the relative magnitude of the iNMF homogeneous solution to a level that matches that of the optimal solution.

In summary, selecting the optimal λ is akin to finding the iNMF solution with the most appropriate level of deviation from the jNMF and sNMF solutions. What remains is to decide how to quantify this deviation for each data source. We use the (unsquared) Frobenius norm of the residuals for this task, summed across sources:

$$R_{S,k} = \|X_k - X_k^S\|_F, \quad R_{J,k} = \|X_k - X_k^J\|_F, \quad R_{I,k} = \|X_k - W^I H_k^I\|_F.$$

Note the unconventional definition of the iNMF residual with respect to the homogeneous part only. Since the sNMF and jNMF are minimizers of their respective objective functions, we have $R_{S,k} \leq R_{J,k} \leq R_{I,k}$.

Our primary reasons for using the unsquared residuals are that (1) they are on approximately the same scale as the solutions and (2) they are more robust than comparing the solutions directly, particularly due to the relative non-identifiability of NMF-type solutions with respect to scale and rotation. Also, our tuning selection procedure suggests searching across a decreasing list of λ until the threshold is exceeded. This is slightly more conservative (to avoid overfitting) than finding the λ such that $R_I^{(\lambda)} - R_J$ is closest to $2(R_J - R_S)$, although the latter is an option.

S3. Data generation for the simulation study. We outline here our method of generating data sets containing multi-dimensional modules with various types of perturbations.

1. Generate a joint block diagonal support:
   (a) Set $W_{N\times D}$ and $(H_k)_{D\times M_k}, k = 1, \ldots, K$ to be binary and block diagonal (D blocks) so that their products WH_k align with the desired data and module dimensions.
   (b) Independently assign each nonzero entry in W and H_k a random value according to Beta$(2, 2) * 2$ (this is arbitrary).
   (c) Multiply to obtain the matrices $WH_k, k = 1, \ldots, K$.
2. Introduce heterogeneous perturbations:

(a) Set $(V_k)_{N\times D}$ to be zero matrices, and consider the D^2 regions whose rows and columns align with the D blocks (modules) in W.

(b) In each of the D^2 regions, introduce a heterogeneous perturbation (with independent probability σ_h) by assigning either the top or lower half (with equal probability) to be ones.

(c) Independently assign to each nonzero entry of $(V_k)_{N\times D}$ a random value according to $\text{Beta}(2,2) * 2$

(d) Add the products $V_k H_k$ to the previous results to obtain $X_k = (W + V_k)H_k$ (the data sets should resemble the ones in Scenario 2 of Figure 1a).

3. Introduce scattered and uniform error:

(a) For each entry in X_k, with independent probability σ_s, either replace a positive value with zero, or replace a zero with a randomly generated $(\text{Beta}(2,2) * 2)^2$ value.

(b) For each entry in X_k, with independent probability σ_u, add a random $\text{Unif}(-\sigma_u, \sigma_u)$ value, and take the absolute magnitude as the new entry.

S4. Normalization for iNMF. In dealing with multiple data sources, integrative methods must find a way to represent the information from each source in a balanced way. In iNMF, we may consider attaching weights c_k to each data matrix to control the level of influence of each source over the analysis:

$$\mathcal{F}(W, H, V) = \sum_k \|c_k X_k - (W + V_k)H_k\|_F^2 + \lambda \sum_k \|V_k H_k\|_F^2.$$

Of course, this is equivalent to scaling each data set X_k by a factor of c_k. Here, we explore how one should approach choosing these normalization coefficients.

In our application, we normalized with respect to the within-source variance of each data set (i.e. $c_k = 1/\text{std}(X_k)$). This accounts for the inherent levels of variation within the sources, but not the numbers of variables (about a 19:20:1 ratio). Therefore, we also consider here normalizing with respect to the sum-of-squares of each data set (i.e. $c_k = 1/\sqrt{\text{SS}(X_k)}$). Table S1 shows the validation results from repeating the analysis under this alternative normalization. Compared with those of the previous normalization, the GE clusters are less concordant with the reference while the ME clusters are more concordant. The scores for the DM clusters remain roughly consistent, likely due to these clusters having poor concordance to begin with.

In principle, the normalization weights should be chosen to address discrepancies in the variability of data and the number of variables in each source. However, the integrative value of a data source may depend on many other factors such as the reliability of each source, the relevance of each source to the research purpose, and the clarity of each source's signal.

		I			P		
		DM	GE	ME	DM	GE	ME
Null clusters	mean	61	58	44	49	50	65
	st.dev.	4	7	2	5	8	1
$\lambda_s = 1$	jNMF	58	23	36	50	77	74
	iNMF	56	49	24	58	62	85
$\lambda_s = 0.1$	jNMF	58	26	30	50	77	79
	iNMF	55	43	27	58	69	82
$\lambda_s = 0.01$	jNMF	62	30	26	50	77	82
	iNMF	46	31	27	67	69	82
$\lambda_s = 10^{-3}$	jNMF	45	56	30	67	54	79
	iNMF	48	48	27	67	62	82
$\lambda_s = 10^{-4}$	jNMF	56	31	41	58	69	68
	iNMF	54	62	27	58	54	82

Table S1: Impurity (I) and purity (P) scores (in percentages) of empirical clusters obtained from jNMF and iNMF with respect to three reference clusters, under sum-of-squares normalization. Shading indicates significantly (≥ 2 sd) higher concordance compared to both the alternative method and the null distribution.

Therefore, dimensionality and variability should not completely dictate the normalization. As we have seen, applying the sum-of-squares normalization does not necessarily produce a more concordant joint approximation of modules, possibly due to differences in signal strength and fidelity between the sources.

As a general rule, dimensionality and data variability should guide the choice of normalization, but the nature of the sources themselves should also be taken into account. In our application, our follow-up analysis takes place in the space of genes, so it was natural to use the standard deviation normalization which produced more concordant GE results.

In any case, it is recommended to check the robustness of the findings under different normalizations. Under the sum-of-squares normalization, iNMF produced the most concordant results under $\lambda_s = 0.01$ (Table S1). Using this result, we applied the same procedure as used before to obtain the visualization in Figure S2. Apart from minor discrepancies, all four modules (I/P/D/M) are distributed in roughly the same topological regions as before. The empirical memberships of the genes in these pathways appear stable between the two normalizations.

S5. Reference variable clusters.
GE reference cluster:

- CXCL11, CXCL10, CXCR3
- HMGA2, SOX11, MCM2, PCNA
- MUC16, MUC1, SLPI
- FAP, ANGPTL2, ANGPTL1

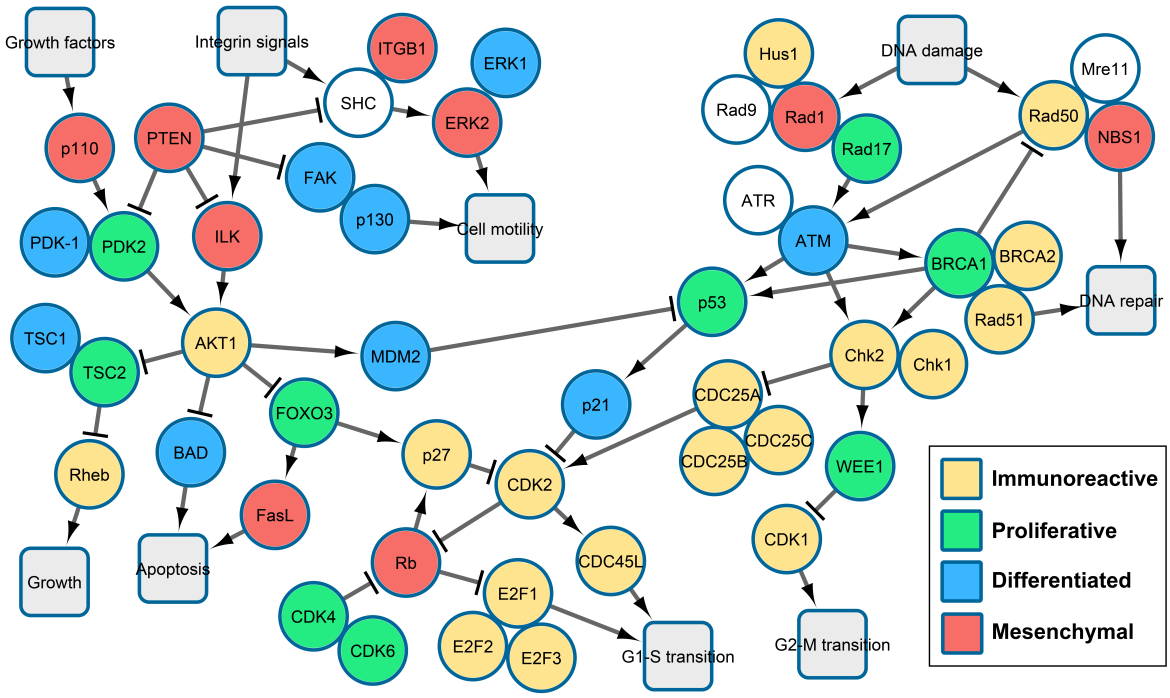

Figure S2: Module memberships of genes (from iNMF with alternative sum-of-squares normalization) arranged according to pathways derived from BioCarta and relevant literature.

DM reference cluster:

- cg08046471, cg01288089, cg08843314
- cg03251079, cg20088964, cg08432727, cg20008332, cg10691006, cg15057726, cg02689825, cg04562739, cg25984124
- cg06420088, cg07399355, cg17257175, cg24512973, cg12966875, cg23889010
- cg08826839, cg09427311, cg11213150, cg07044282

ME reference cluster:

- hsa-miR-19a, hsa-miR-19b, hsa-miR-136, hsa-miR-376c, hsa-miR-483-5p, hsa-miR-572, hsa-miR-575, hsa-miR-638, hsa-miR-671-5p, hsa-miR-769-5p, hsa-miR-923, hsa-miR-1225-5p
- hsa-miR-15b, hsa-miR-98, hsa-miR-135b, hsa-miR-146a, hsa-miR-148a, hsa-miR-148b, hsa-miR-150, hsa-miR-221*, hsa-miR-342-5p, hsa-miR-361-3p, hsa-miR-362-3p, hsa-miR-374a, hsa-miR-374b, hsa-miR-450a, hsa-miR-454, hsa-miR-502-5p, hsa-miR-505, hsa-miR-532-3p, hsa-miR-582-5p, hsa-miR-625, hsa-miR-652, hsa-miR-660