Identify functional miRNA-mRNA regulatory modules with correspondence latent dirichlet allocation

Supplementary File 1

Bing Liu[1,*], Lin Liu[1], Anna Tsykin[2,3], Gregory J. Goodall[2,4], Jeffrey E. Green[5], Min Zhu[5], Chang Hee Kim[6], and Jiuyong Li[1]

[1]School of Computer and Information Science, University of South Australia, Mawson Lakes, SA 5095, Australia.

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

[3]School of Molecular & Biomedical Science, The University of Adelaide, Adelaide, SA 5005, Australia

[4]Department of Medicine, The University of Adelaide, Adelaide, SA 5005, Australia

[5]Laboratory of Cancer Biology and Genetics, National Cancer Institute, Bethesda, MD 20892, USA

[6]Laboratory of Molecular Technology, NCI-FCRDC, Frederick, MD 21702, USA

In this file, we provide supplementary information discussed in the Method and Results sections.

1. Methods

Given *K* functional modules, the generative procedure without the putative target constraint for each sample d is illustrated by the following sampling process:

1. Choose $\overline{\theta_d}|\alpha\sim\mathrm{Dir}(\alpha)$
2. Choose $\overline{\varphi_k}|\beta\sim\mathrm{Dir}(\beta)$
3. Choose $\overline{\omega_k}|\gamma\sim\mathrm{Dir}(\gamma)$
4. For each $r_{d,n}$, $n\in\{1,\cdots,N_d\}$:
    a. Choose module $z_{d,n}|\overline{\theta_d}\sim\mathrm{Mult}(\theta_d)$
    b. Choose a $r_{d,n}|\{z_{d,n},\varphi_{1:K}\}\sim\mathrm{Mult}(\varphi_{z_{d,n}})$
5. For each $g_{d,m}$, $m\in\{1,\cdots,M_d\}$:
    a. Choose miRNA index $y_{d,m}|N\sim\mathrm{Unif}\{1,\cdots,N_d\}$
    b. Choose $g_{d,m}|\{y_{d,m},\bar{z},\omega_{1:K}\}\sim\mathrm{Mult}(\omega_{z_{y_{d,m}}})$

Using Eq. (2), (3), (4), and (5), the Gibbs sampling procedure can be designed as Algorithm 1. The algorithm includes three stages: initialization, sampling, and reading out parameters.

Algorithm 1: Gibbs sampling for FMRM discovery

Initialization

Assign zeros to all count variables, $n_d^{(k)}$, n_d, $n_k^{(v)}$, n_k, $m_k^{(t)}$, m_k

foreach $d\in\{1,\cdots,D\}$ do

foreach miRNA $r_{d,n}, n \in \{1, \cdots, N_d\}$ do
sample FMRM index $z_{d,n} = k \sim Mult(1/K)$
increment sample-FMRM count: $n_d^{(k)} + 1$
increment sample-FMRM sum: $n_d + 1$
increment FMRM-miRNA count: $n_k^{(v)} + 1$
increment FMRM-miRNA sum: $n_k + 1$
end for
foreach mRNA $g_{d,m}$ $m \in [1, M_d]$ do
sample index for FMRM index $y_{d,m} = x \sim Uniform\{1, \cdots, N_d\}$
assign the FMRM $k = z_{d,y_{d,m}}$ to mRNA $g_{d,m}$
increment FMRM-mRNA count: $m_k^{(t)} + 1$
increment FMRM-mRNA sum: $m_k + 1$
end for

Gibbs sampling over burn-in period and sampling period

while not converge or not reach iteration limit do
foreach $d \in \{1, \cdots, D\}$ do
foreach miRNA $r_{d,n}, n \in \{1, \cdots, N_d\}$ do
*for the current assignment k to a miRNA term v indexed by miRNA $r_{d,n}$: decrement counts and sums: $n_d^{(k)} - 1, n_d - 1, n_k^{(v)} - 1, n_k - 1$
*sample index $x_{d,n} = x \sim Uniform\{1, \cdots, M_d\}$ for mRNA $g_{d,x_{d,n}}$, which corresponding to miRNA $r_{d,n}$
*for the current assignment of k to a term t for mRNA $g_{d,x_{d,n}}$: decrement counts and sums: $m_k^{(t)} - 1, m_k - 1$
*multinomial sampling according to Eq. (2), sample topic index $\hat{k} \sim p(z_{d,n} | Z_{-(d,n)}, Y_d, R_d, G_d)$
*use the new assignment of $z_{d,n} = \hat{k}$ to the miRNA term v for miRNA $r_{d,n}$ and increment counts and sum: $n_d^{(\hat{k})} + 1, n_d + 1, n_{\hat{k}}^{(v)} + 1, n_{\hat{k}} + 1$
*use the new assignment of $z_{d,n} = \hat{k}$ to the term t for mRNA $g_{d,x_{d,n}}$ and increment counts and sum: $m_{\hat{k}}^{(t)} + 1, m_{\hat{k}} + 1$
end for
end for

Check convergence and read out parameters

If converged and L sampling iterations since last read out then
read out parameter set Θ , Φ, and Ω according to (3)
end if
end while

2. Results

2.1. Materials and experiment data sets

The data sets are from Min Zhu et al. (Zhu, et al., 2010). The mRNA expression data were profiled with mouse genome 430A 2.0 GeneChip (Affymetrix) and scanned on Affymetrix GeneChip scanner 3000. The microRNA microarray chip (LMT_miRNA_v2 microarray) was designed using the Sanger miR9.0 database (http://microrna.sanger.ac.uk) and manufactured by Agilent Technologies as custom-synthesized 8 x 15k microarrays. The array contains 1667 unique mature miRNA sequences across all species, among them, 334 unique miRNAs were for mouse. Each mature miRNA is represented by + and – (reverse complement) strand sequences, and each with 4 replicate probes.

miRNA gene expression data normalization

The gProcessSignal values of probes designed for mouse miRNAs were feature extracted using the GE2 protocol (www.agilent.com) with exclusion of internal control probes, non-mouse probes, and all negative strand probes. A global median normalization procedure was applied to the gProcessSignal values of the selected probes across all arrays.

mRNA expression data normalization

mRNA array data were normalized using GC-RMA of Partek Genomic Suite (www.partek.com). The normalized data was further filtered using MAS5 detection calls for probes designated as "P" (present) or "M" (Marginal) in less than 3 samples of all data population. Basal-luminal differential miRNAs and model-specific miRNA signatures were derived as described above.

Sample information

The sample names and their corresponding mouse model class and tumor subtypes:

Sample Name	Mouse Model	Model Class	Tumor Subtype
X503_BT	Brca1-/- p53 503 BT	BRCA_p53	Basal
X4176_BT	Brca1-/- p53 4176 BT	BRCA_p53	Basal
X627_BT	Brca1-/- p53 T627 BT	BRCA_p53	Basal
X572_BT	Brca1-/- p53 572 BT	BRCA_p53	Basal
X53447_BT	Brca1-/-p53447BT	BRCA_p53	Basal
C3Tag_2	C3TAg 2	C3TAg	Basal
C3Tag_4	C3TAg 4	C3TAg	Basal
C3Tag_5	C3TAg 5	C3TAg	Basal
C3Tag_7	C3TAg 7	C3TAg	Basal
C3Tag_8	C3TAg 8	C3TAg	Basal
P53_1570	p53-/- 1570R_PN1b	p53	Basal
P53_2979	p53-/- 2979R_PN1b	p53	Basal
P53_5354	p53-/- 5354L_PN10	p53	Basal
P53_5809	p53-/- 5809R_PN2(254c)	p53	Basal
P53_5817	p53-/- 5817_PN2(254c)	p53	Basal
P53_5851	p53-/- 5851L_PN2(254c)	p53	Basal
P53_8546	p53-/- 8546R_PN1b	p53	Basal
Cmyc_043508	C-Myc Tumor 043508	C-Myc	Luminal
Cmyc_04004022	C-Myc Tumor 04004022	C-Myc	Luminal
Cmyc_04005648	C-Myc Tumor 04005648	C-Myc	Luminal
Cmyc_04004021	C-Myc Tumor 04004021	C-Myc	Luminal
H2N_Founder_A	MMTV-H2N Founder A	H2N	Luminal
H2N_53	MMTV-H2N 53	H2N	Luminal
H2N_1	MMTV-H2N 1	H2N	Luminal
H2N_61	MMTV-H2N 61	H2N	Luminal

Sample Name	Mouse Model	Model Class	Tumor Subtype
H2N_64	MMTV-H2N 64	H2N	Luminal
Hras_1.4	MMTV-Haras An #1-4	Hras	Luminal
Hras_3.4	MMTV-Haras An #3-4	Hras	Luminal
Hras_5.4	MMTV-Haras An #5-4	Hras	Luminal
Hras_4.4	MMTV-Haras An #4-4	Hras	Luminal
Hras_2.4	MMTV-Haras An #2-4	Hras	Luminal
PyMT_436	PyMT 436	MMTV_PyMT	Luminal
PyMT_437	PyMT 437	MMTV_PyMT	Luminal
MMTV.11567	MMTV-PymT #11567	MMTV_PyMT	Luminal
MMTV.11568	MMTV-PymT #11568	MMTV_PyMT	Luminal
MMTV.11570	MMTV-PymT #11570	MMTV_PyMT	Luminal
MMTV.5.1	MMTV-PymT #5-1	MMTV_PyMT	Luminal
Wnt_4675	MMTV-Wnt 4675	MMTV-Wnt	Luminal
Wnt_4676	MMTV-Wnt 4676	MMTV-Wnt	Luminal
Wnt_4635	MMTV-Wnt 4635	MMTV-Wnt	Luminal
Wnt_4677	MMTV-Wnt 4677	MMTV-Wnt	Luminal
FVB_M1_2	FVB pregnant M2-1	NormalMammary	Normal
FVB_M1_4	FVB pregnant M4-1	NormalMammary	Normal
FVB_M1_1	FVB pregnant M1-1	NormalMammary	Normal
FVB_M1_3	FVB pregnant M3-1	NormalMammary	Normal
FVB_M1_5	FVB pregnant M5-1	NormalMammary	Normal

2.2. Target reconstruction

To investigate whether the miRNAs and mRNAs in each module are identified by chance, we randomly selected a group of miRNAs and a group of mRNAs from MicroCosm with the same numbers as those in the identified modules, and queried how many pairs they could be linked by MicroCosm. The distribution of the number of matched pairs was estimated by a simulation which was executed for 10,000 times. The estimated distribution (Figure S1) shows that the numbers of target relationship of the randomly chosen miRNAs and mRNAs are significantly different from those of the identified miRNAs and mRNAs in each module (p-value < 0.05). It indicates that the miRNAs and mRNAs in each module are not identified by chance.

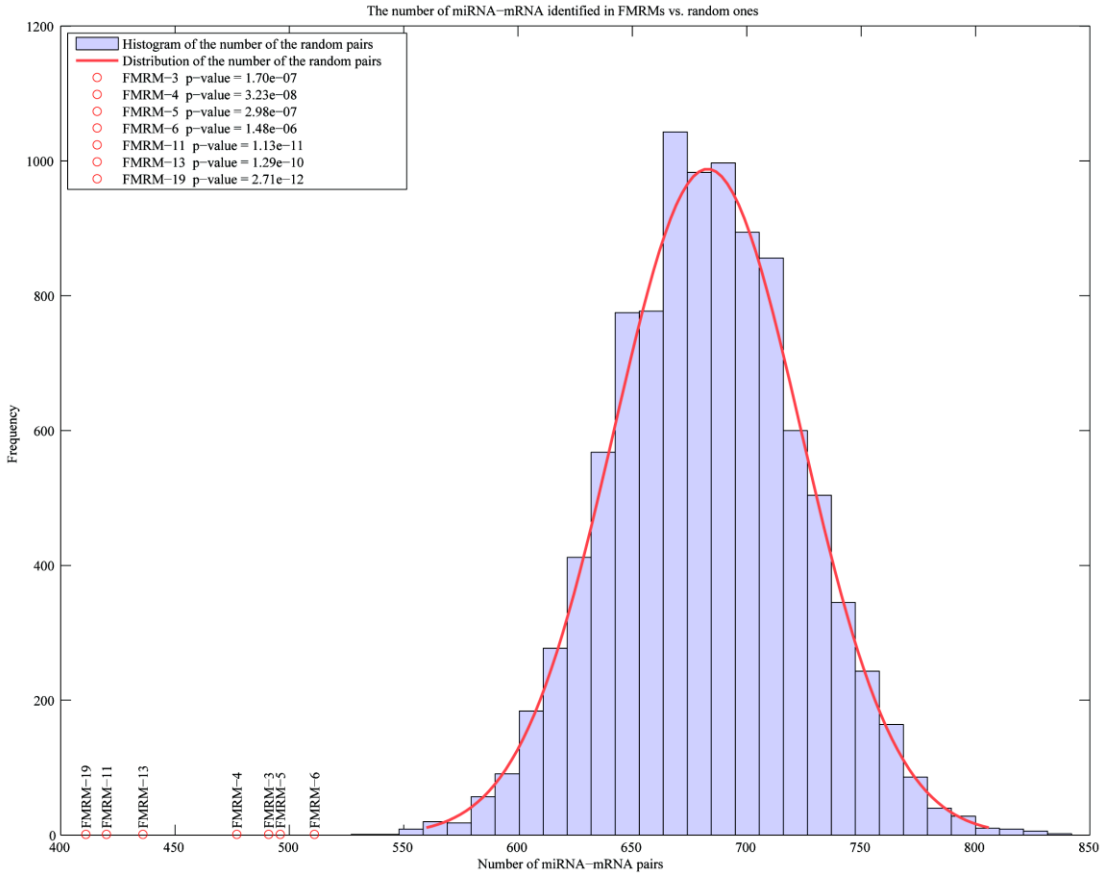

Fig. S1. Comparison of the numbers of miRNA-mRNA pairs in the identified modules with the ones from the random matching. The distribution of the number of matched target pairs is estimated by a simulation which was executed for 10,000 times. It indicates that the miRNAs and mRNAs identified in each module are not identified by chance.

2.3. Validation of identified mRNAs in the FMRMs

It is expected that the miRNA target genes are also relevant to the specific biological processes. In our results, 18 genes have been identified by Adelaide et al. (Adelaide, et al., 2007) as in Table S1.

Table S1. Validation of identified mRNAs in the FMRMs

Gene	Expression	Associated Type	Module
Cct3	Over	Basal	1
Upf2	Over	Basal	1
Eif4a1	Under	Luminal	2,18,20
Ccdc77	Over	Basal	3
Rbm4b	Over	Luminal	3
Hspa14	Over	Basal	4
Rbx1	Under	Luminal	5,18
Ppap2a	Under	Basal	7
Tpd52	Over	Luminal	8
Tulp3	Over	Basal	9

Gpm6a	Under	Basal	10
Gdap1	Over	Luminal	10
Gspt1	Over	Luminal	11
Rbx1	Over	Luminal	12
Npy1r	Under	Basal	16
Rpl13	Under	Luminal	18
Cox4i1	Under	Luminal	19
Arfgef1	Over	Luminal	20

Adelaide et al. (Adelaide, et al., 2007) suggest the existence of potential oncogenes and tumor suppressor genes differentially associated with the basal and luminal subtype. In our results, 18 genes identified in FMRMs are consistent with their results.

The networks which are explicitly associated with cancers and within the top five networks of each module are given in Table S2. The detailed networks are also given in Figure S2 to S6.

Table S2. Associated network functions of FMRMs

FMRM#	Associated Network Function	Ref Figure	Score
3	Cancer, Cellular Compromise, DNA Replication, and Repair	Fig. S2	12
3	Gene Expression, Cancer, Immunological Disease	Fig. S3	12
4	Cellular Growth and Proliferation, Cancer, Dermatological Diseases and Conditions	Fig. S4	14
4	Cellular Assembly and Organization, Cancer, Cellular Development		2
13	Cancer, Cell Cycle, DNA Replication, Recombination, and Repair	Fig. S5	18
13	Cancer, Cell Morphology, Cellular Development	Fig. S6	12
19	Cancer, Cell-To-Cell Signaling and Interaction, Cellular Function and Maintenance		2
19	Amino Acid Metabolism, Cancer, Cell Morphology		2

The networks participated by the genes identified in FMRMs are highly related to cancers. The networks associated with cancers that are explicitly within the top five networks of each module are listed. It is worth noting that many mouse genes have been filtered out as we specifically target human breast cancers.

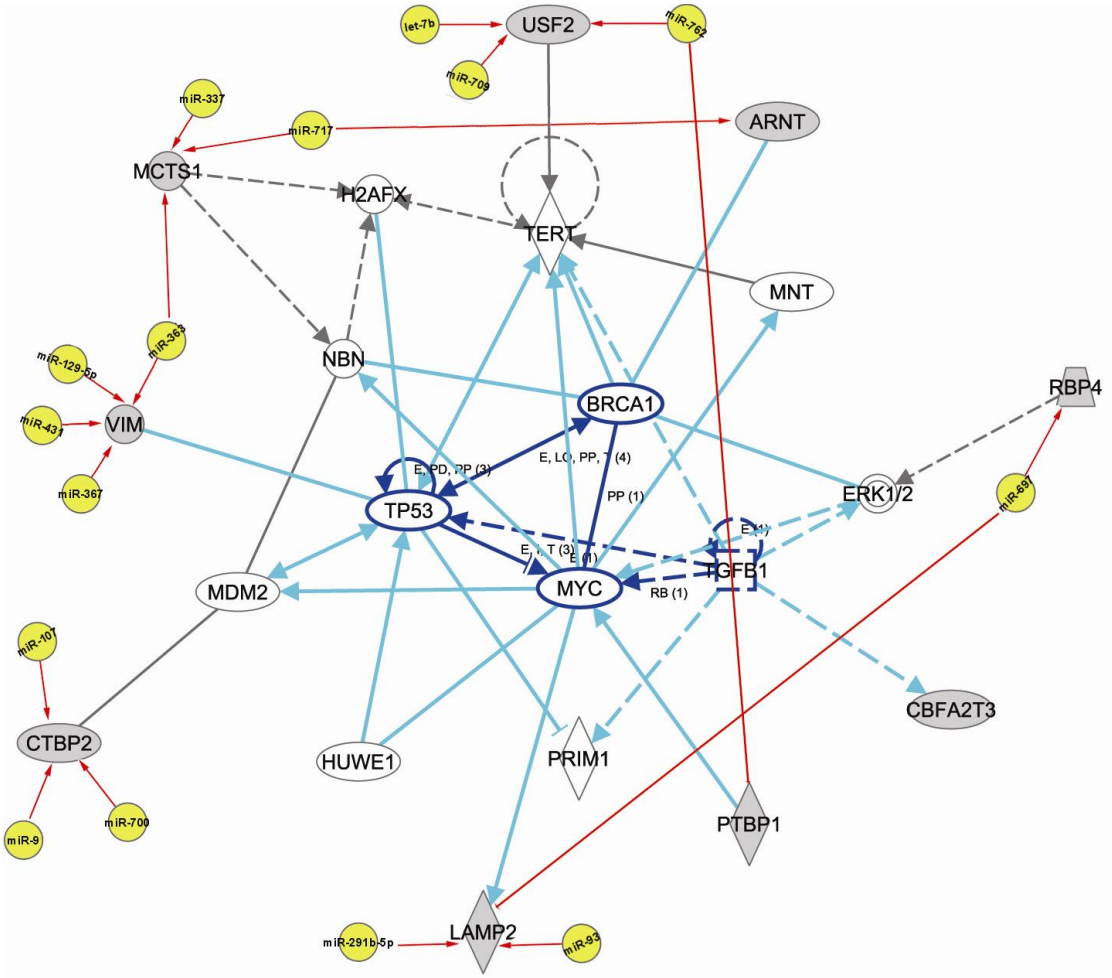

Fig. S2. Module-3 associated network functions. Cancer, Cellular Compromise, DNA Replication, and Repair

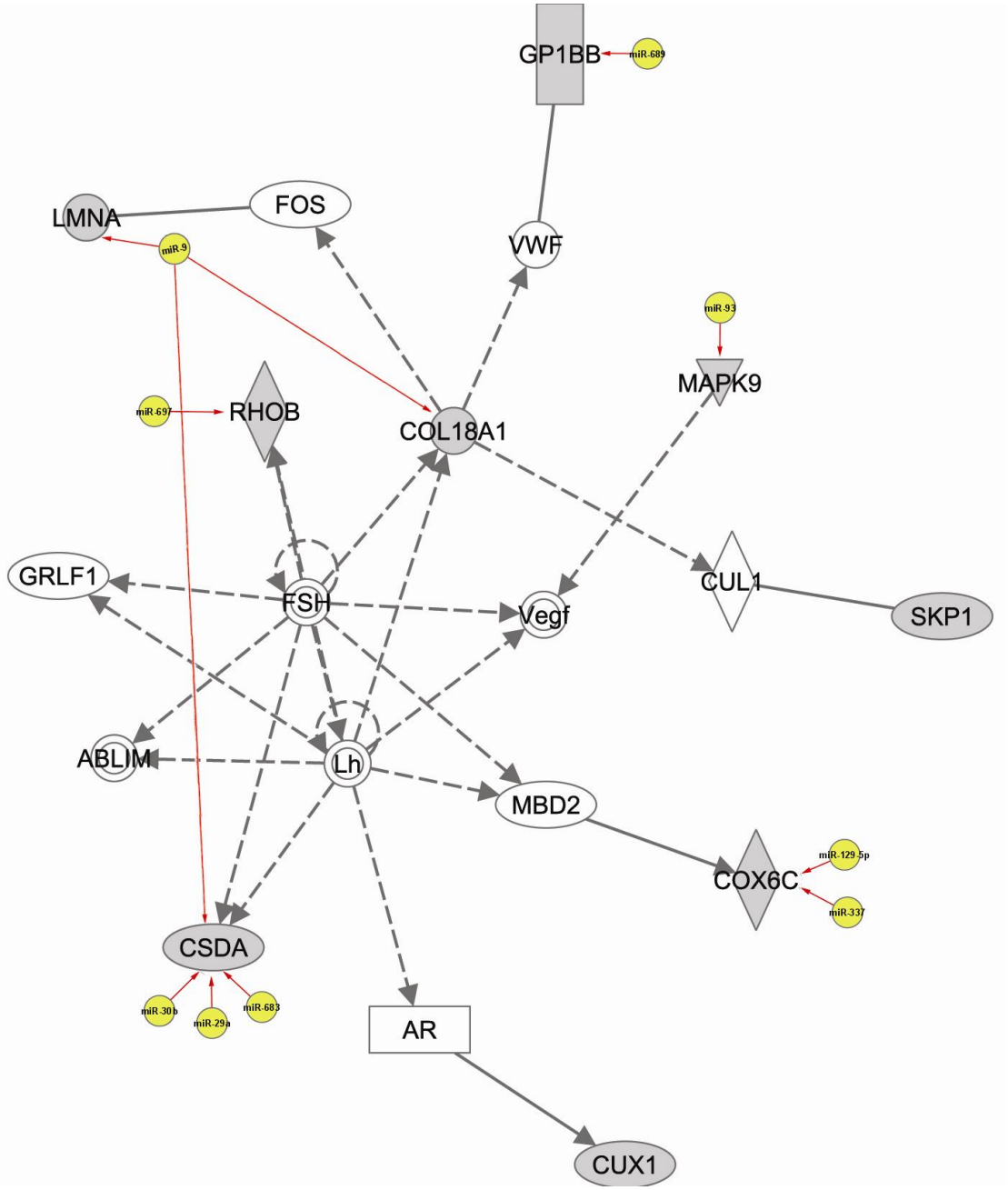

Fig. S3. Module-3 associated network functions. Gene Expression, Cancer, Immunological Disease

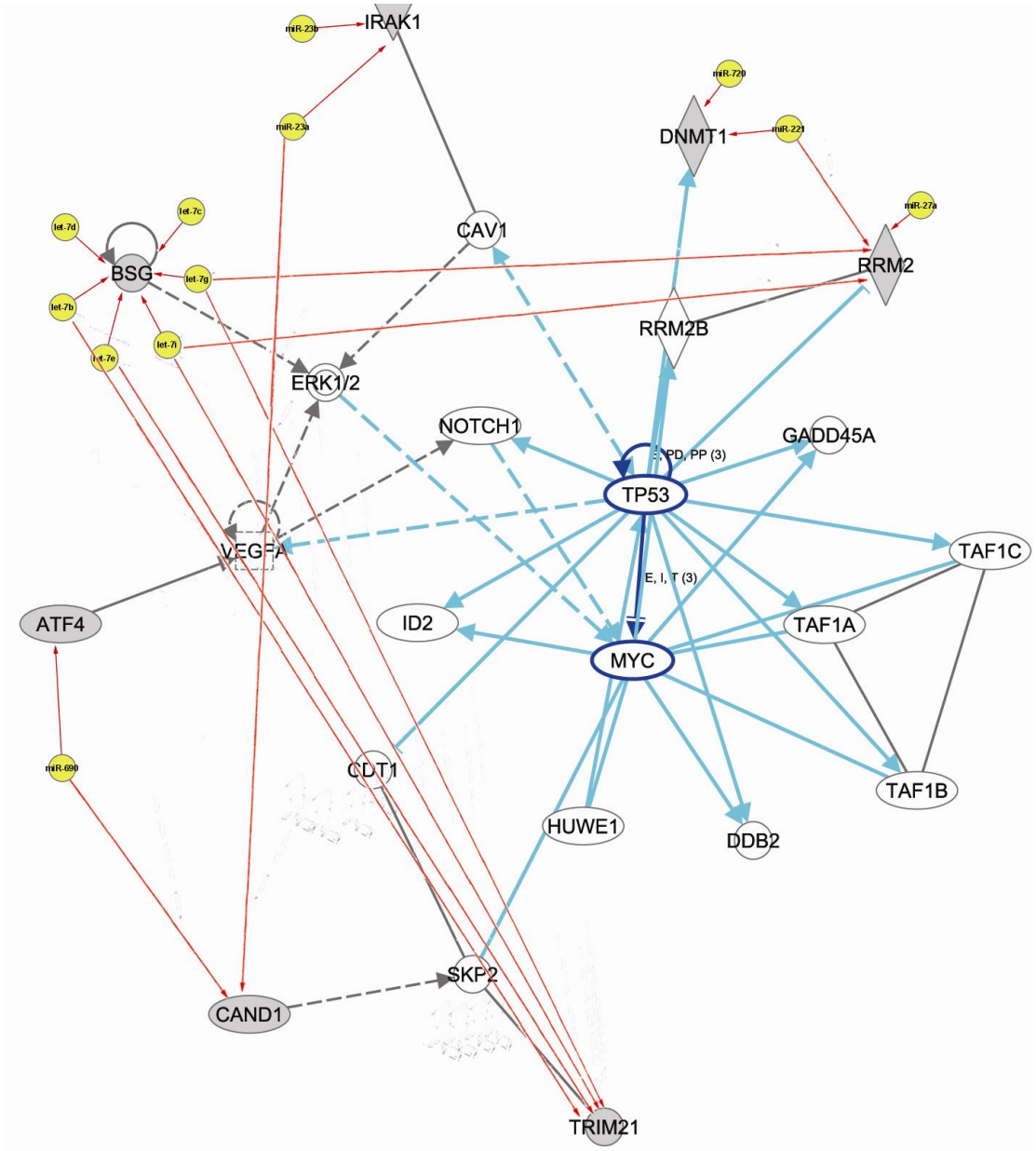

Fig. S4. Module-4 associated network functions. Cellular Growth and Proliferation, Cancer, Dermatological Diseases and Conditions

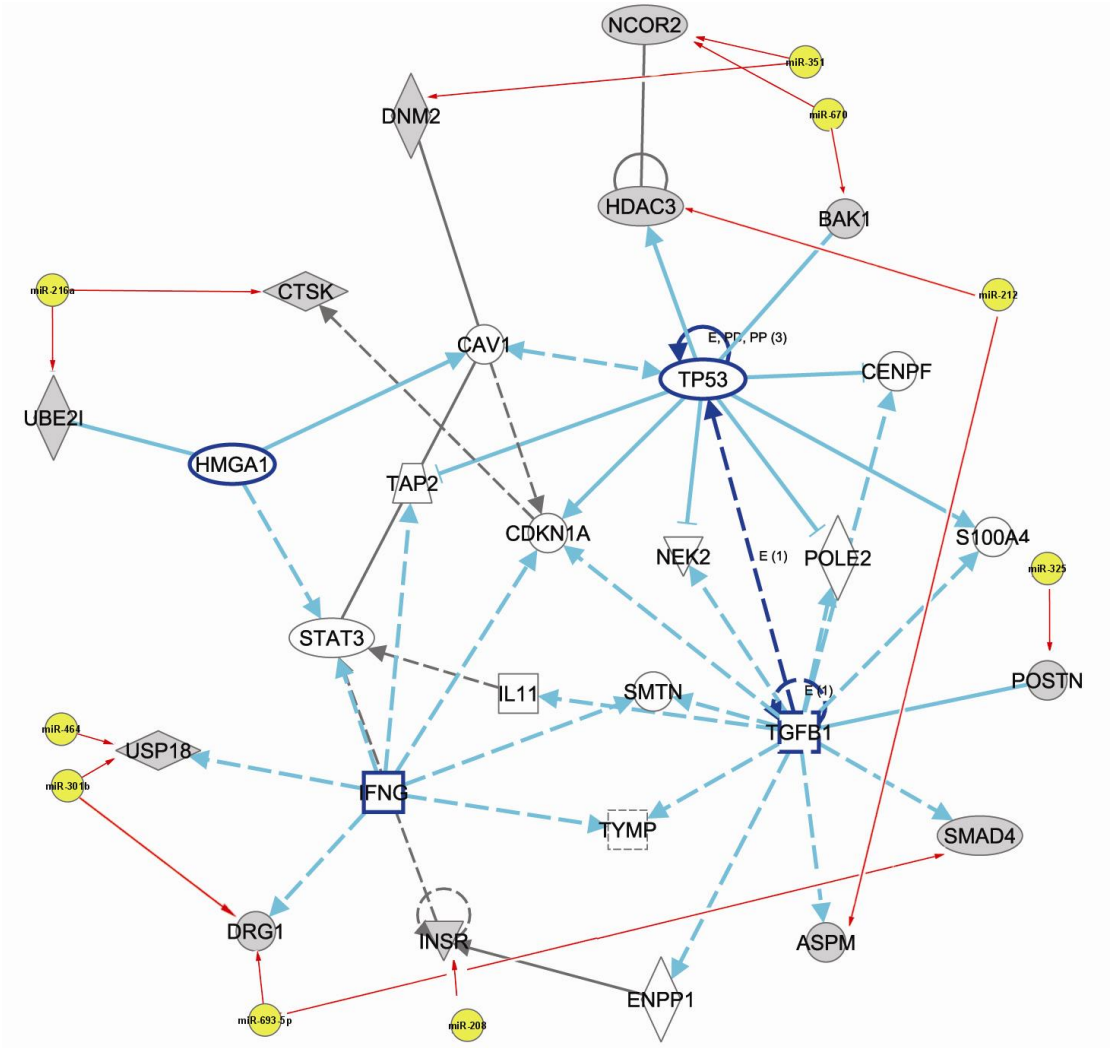

Fig. S5. Module-13 associated network functions. Cancer, Cell Cycle, DNA Replication, Recombination, and Repair

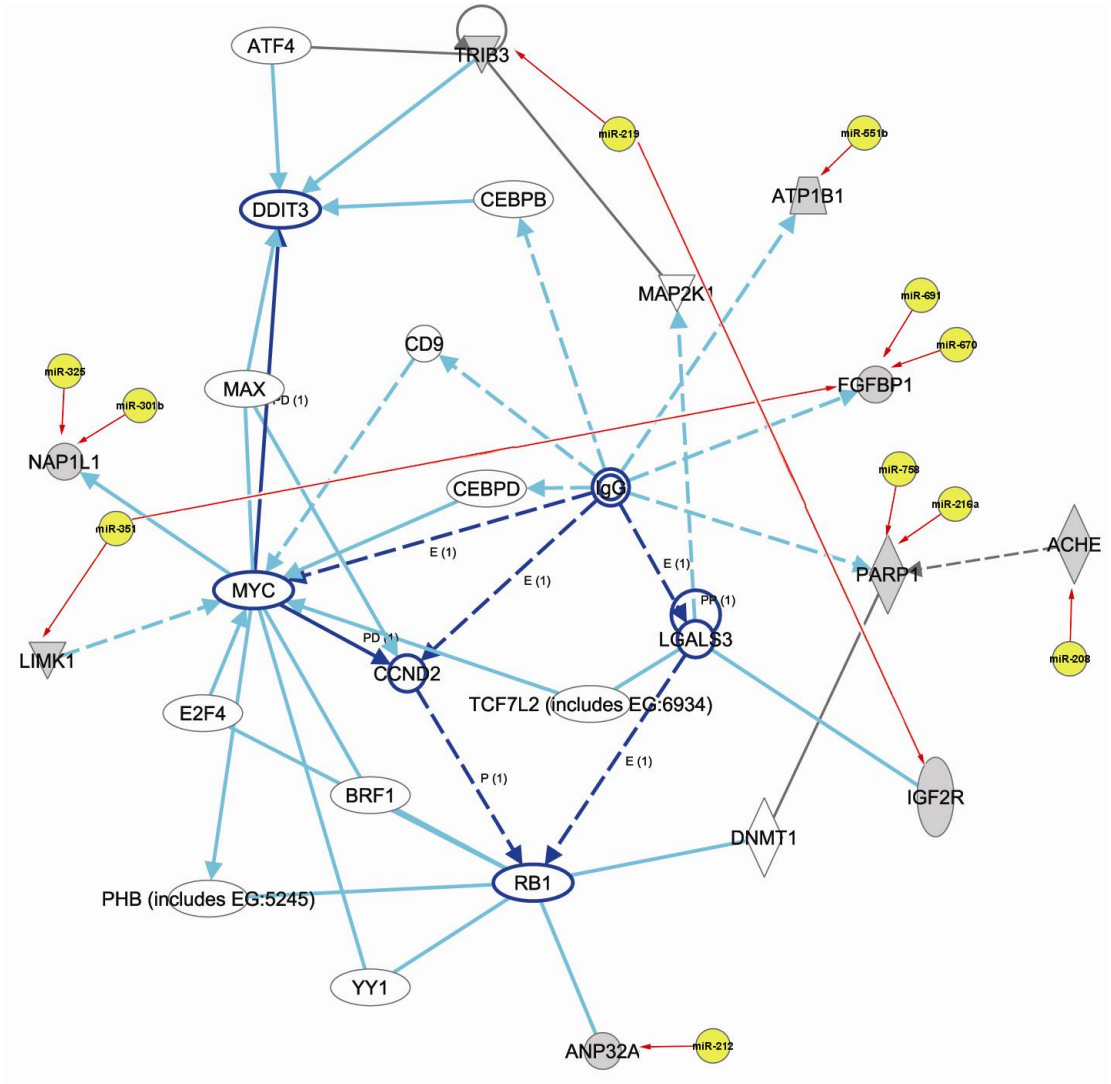

Fig. S6. Module-13 associated network functions. Cancer, Cell Morphology, Cellular Development

Reference

Adelaide, J., Finetti, P., Bekhouche, I., Repellini, L., Geneix, J., Sircoulomb, F., Charafe-Jauffret, E., Cervera, N., Desplans, J., Parzy, D., Schoenmakers, E., Viens, P., Jacquemier, J., Birnbaum, D., Bertucci, F. and Chaffanet, M. (2007) Integrated Profiling of Basal and Luminal Breast Cancers, *Cancer Res*, **67**, 11565-11575.

Zhu, M., Yi, M., Kim, C.H., Deng, C., Li, Y., D.Medina, Hunter, K., Stephen, R. and Green, a.J. (2010) Comprehensive genomic profiling identifies miRNA signatures associated with mouse mammary tumor subtypes, *In Preparation*.