Regression Analysis of Combined Gene Expression Regulation in Acute Myeloid Leukemia

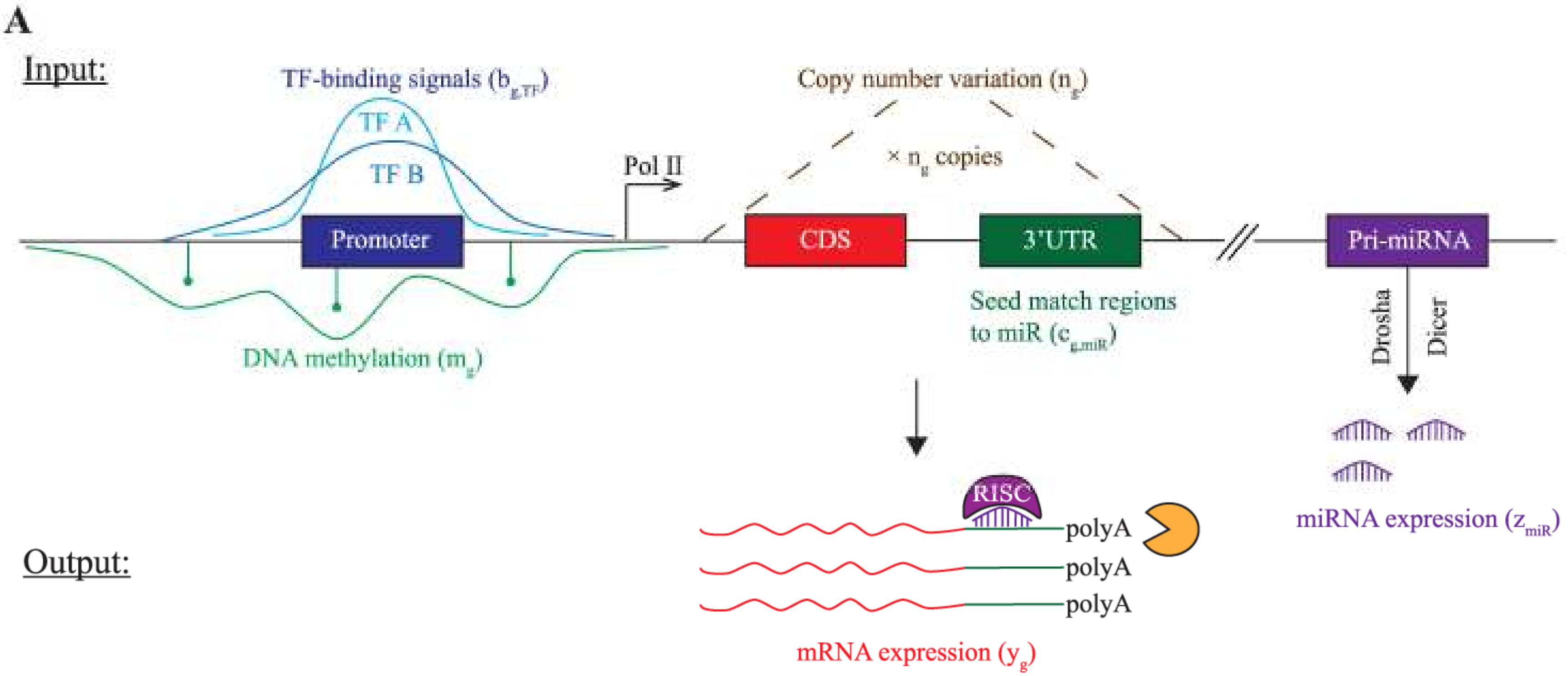
Yue Li , Minggao Liang, Zhaolei Zhang

Main Contribution

- Using the TF data from ENCODE, and CNV, DM, miRNA expression signals from TCGA.
- A two stage regression model.

Main Contribution

- Comparing to *Integrated modeling of transcriptional drivers*, it uses collected TF data instead of infered measurement.



B

Stage 1: Estimate sample-specific TF and miR activities ($\alpha_{TF,t}$, $\alpha_{miR,t}$) in sample t:

$$[y_{g,t}]_{N\times 1} \approx \alpha_0 + \alpha_{CNV,t}[n_{g,t}]_{N\times 1} + \alpha_{DM,t}[m_{g,t}]_{N\times 1} + [b_{g,TF}]_{N\times K}\times[\alpha_{TF,t}]_{K\times 1} + [c_{g,miR}]_{N\times M}\times([\alpha_{miR,t}]_{M\times 1}[z_{miR,t}]_{M\times 1})$$

Stage 2: Estimate TF-gene and miRNA-mRNA interactions ($W_{TF,g}$, $W_{g,miR}$) for gene g across all samples:

$$[y_{g,t}]_{1\times T} \approx w_0 + w_{g,CNV}[n_{g,t}]_{1\times T} + w_{g,DM}[m_{g,t}]_{1\times T} + [w_{g,TF}]_{1\times K^*}\times[\alpha_{TF,t}]_{K^*\times T} + [w_{g,miR}]_{1\times M^*}\times[\alpha_{miR,t}]_{M^*\times T}$$

Stage one

In the first stage, we estimate *sample-specific TF* and *miRNA* activities ($\alpha_{TF,t}$, $\alpha_{miR,t}$) in sample *t* with α_0 being the intercept, and $\alpha_{CNV,t}$ and $\alpha_{DM,t}$ being the respective offsets for *CNV* and *DM*:

$$
\begin{aligned}
y_{g,t} \approx \alpha_0 + \alpha_{CNV,t} n_{g,t} + \alpha_{DM,t} m_{g,t} + \sum_{TF \in \{1,\ldots,K\}} b_{g,TF} \alpha_{TF,t} \\
+ \sum_{miR \in \{1,\ldots,M\}} \alpha_{miR,t} c_{g,miR} z_{miR,t}
\end{aligned}
$$

where $b_{g,TF}$ is the binding score of TF on gene *g*, $C_{g,miR}$ is the number of conserved target sites on the 3 UTR of the target gene *g* for *miR* , which is obtained as sequence-based information from TargetScan

Stage two

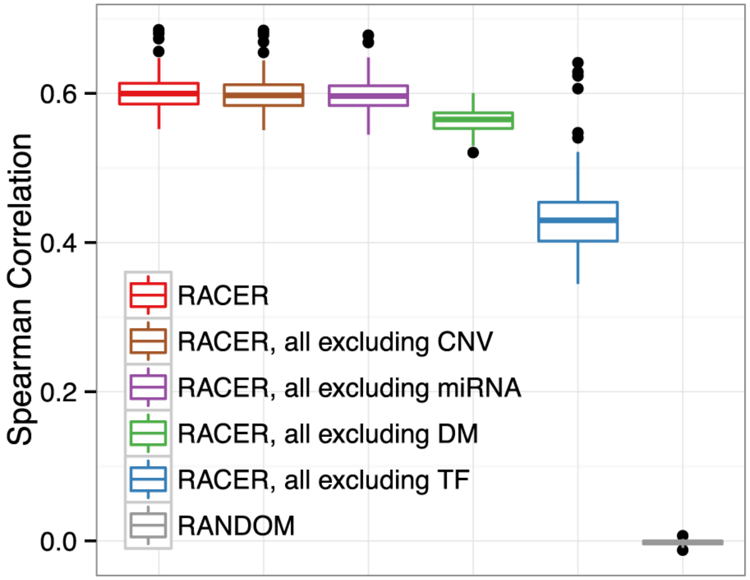
In the second stage, using the estimated $\alpha_{TF,t}$ and $\alpha_{miR,t}$ in stage one, they infer for each gene *g* its association with the candidate TF ($w_{g,TF}$) and miR regulators

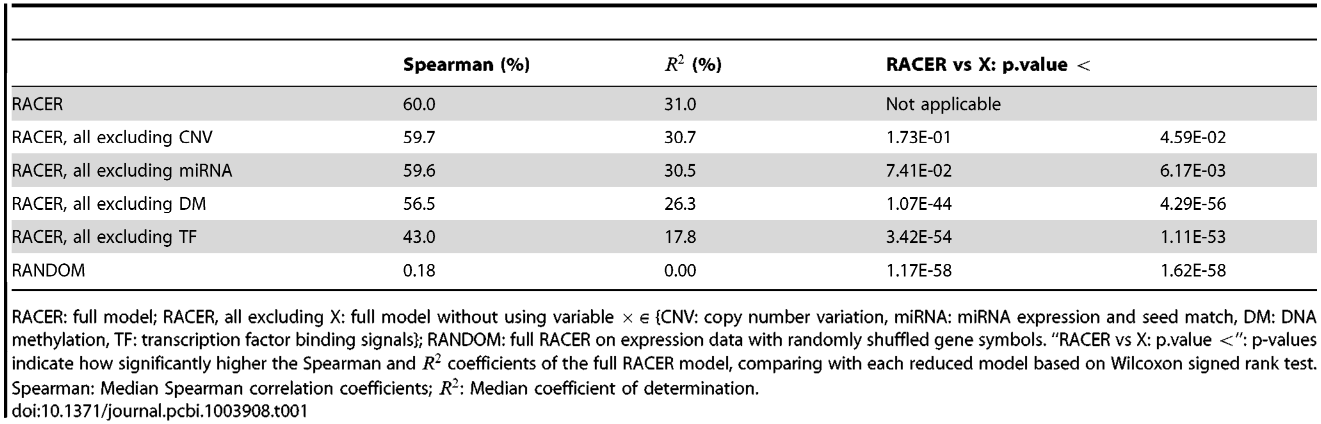
($w_{g,miR}$) *across all of the T samples*:

$$y_{g,t} \approx w_0 + w_{g,CNV} n_{g,t} + w_{g,DM} m_{g,t} + \sum_{TF \in \{1,\ldots,K^*\}} w_{g,TF} \alpha_{TF,t} \\ + \sum_{miR \in \{1,\ldots,M^*\}} w_{g,miR} \alpha_{miR,t}$$

where M* and K* are the respective number of selected TFs and miRNAs with nonzero binding signals $b_{g,TF} > 0$ and conserved target sites $C_{g,miR} > 0$ for gene.

- Comparing with the findings in glioblastoma, however, where CNV played a major role in explaining gene expression, they suggest that the moderate effect of CNV observed here may be AML-specific, i.e., it is unlikely that CNV will have the same effect in other diseases. Indeed, recent studies have shown that many of the AML genomes lack structural abnormalities, implying that the disease complexity may likely reside at the transcriptional and epigenetic level.



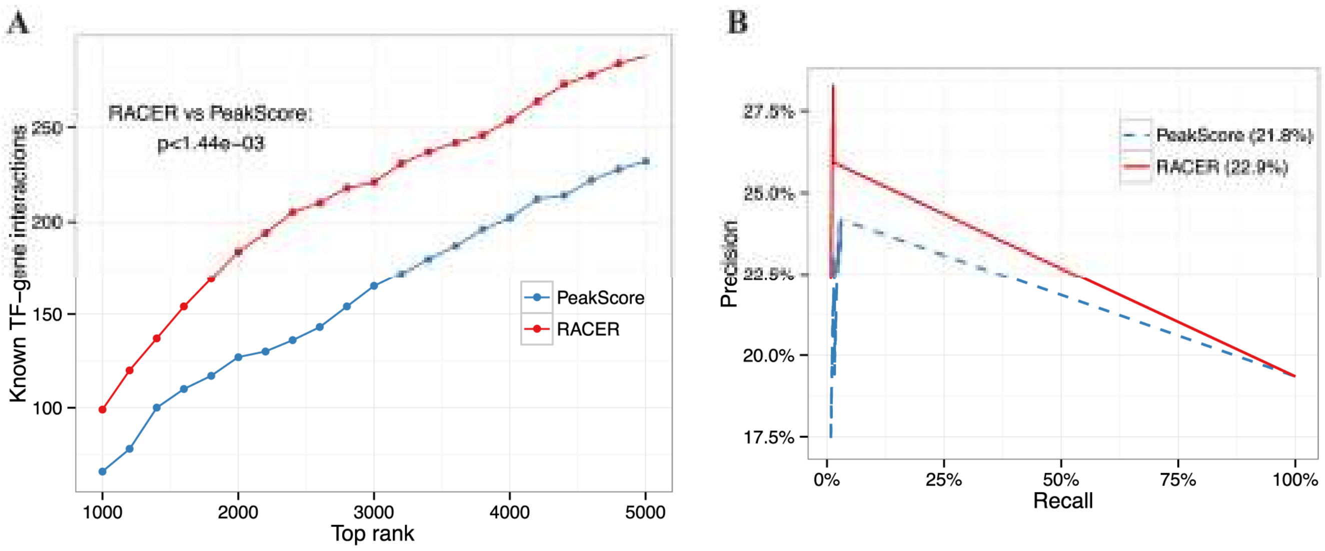


	Spearman (%)	R^2 (%)	RACER vs X: p.value $<$	
RACER	60.0	31.0	Not applicable	
RACER, all excluding CNV	59.7	30.7	1.73E-01	4.59E-02
RACER, all excluding miRNA	59.6	30.5	7.41E-02	6.17E-03
RACER, all excluding DM	56.5	26.3	1.07E-44	4.29E-56
RACER, all excluding TF	43.0	17.8	3.42E-54	1.11E-53
RANDOM	0.18	0.00	1.17E-58	1.62E-58

RACER: full model; RACER, all excluding X: full model without using variable $\times \in$ {CNV: copy number variation, miRNA: miRNA expression and seed match, DM: DNA methylation, TF: transcription factor binding signals}; RANDOM: full RACER on expression data with randomly shuffled gene symbols. "RACER vs X: p.value $<$": p-values indicate how significantly higher the Spearman and R^2 coefficients of the full RACER model, comparing with each reduced model based on Wilcoxon signed rank test. Spearman: Median Spearman correlation coefficients; R^2: Median coefficient of determination.

doi:10.1371/journal.pcbi.1003908.t001

Power analysis



Feature selection

$$F(1, N - M - K + 1) = \frac{(RSS_{\mathrm{RACER,\ all\ excluding\ regulatorX}} - RSS_{\mathrm{RACER}})}{RSS_{\mathrm{RACER}} / (N - M - K + 1)}$$

Regulator	F-statistic	FDR	Enriched pathways or biological processes	Hits	Gene set	Enrichment FDR
PHF8	1565.63	0	misfolded or incompletely synthesized protein catabolic process (GO:0015693)	8	8	0
			DNA repair (GO:0006903)	77	168	1.18E-02
			REACTOME SIGNALING BY WNT	42	65	4.79E-06
			DNA repair (GO:0006903)	61	168	1.73E-04
Max	112.82	8.20E-24	REACTOME DNA REPAIR	44	112	6.91E-04
			KEGG BASE EXCISION REPAIR	17	35	6.06E-02
MAZ	64.14	2.34E-13	ST TUMOR NECROSIS FACTOR PATHWAY	15	29	4.82E-02
ZBTB7A	50.29	1.96E-10	REACTOME P38MAPK EVENTS	7	13	7.27E-02
PU1	31.50	2.30E-06	SA PTEN PATHWAY	6	17	3.11E-02
CCNT2	29.32	5.89E-06	REACTOME CDK MEDIATED PHOSPHORYLATION AND REMOVAL OF CDC6	22	48	4.96E-02
			REACTOME SIGNALING BY WNT	28	65	2.39E-02
hsa-miR-506	28.73	6.84E-06	REACTOME SYNTHESIS OF PC	4	18	2.03E-02
YY1	19.60	4.54E-04	DNA repair (GO:0006903)	50	168	1.92E-02
			REACTOME SIGNALING BY WNT	29	65	7.60E-05
CEBPB	14.11	6.53E-03	DNA repair (GO:0006903)	19	168	1.36E-02
			REACTOME P53 INDEPENDENT G1 S DNA DAMAGE CHECKPOINT	9	51	4.76E-02
hsa-miR-548p	13.33	9.29E-03	ST ERK1 ERK2 MAPK PATHWAY	8	32	4.91E-03
			KEGG CHRONIC MYELOID LEUKEMIA	11	73	4.15E-02
ELF1	10.30	4.45E-02	ST TUMOR NECROSIS FACTOR PATHWAY	16	29	3.56E-02

doi:10.1371/journal.pcbi.1003908.t002