Electronic Supplementary Material (ESI) for Molecular BioSystems.
This journal is © The Royal Society of Chemistry 2015

Compound signature detection on LINCS L1000 big data

Chenglin Liu[1,2,3], Jing Su[2,*], Fei Yang[2], Kun Wei[2], Jinwen Ma[1] and Xiaobo Zhou[2,*]

[1] School of Mathematical Sciences and LMAM, Peking University, Beijing 100871, China

[2] Center for Bioinformatics and Systems Biology, Department of Diagnostic Radiology and Comprehensive Cancer Center of Wake Forest University, Wake Forest School of Medicine, Winston-Salem, NC, 27157, USA

[3] School of Life Sciences & Technology, Shanghai Jiaotong University Shanghai 200240, China

* To whom correspondence should be addressed. Tel: 336-713-1789; Fax: 336-713-5891; Email: JS jsu@wakehealth.edu and XZ xizhou@wakehealth.edu

Method 1: The GMM Model.

To deconvolute the overlapped peaks of the distributions of the L1000 analyte fluorescent intensities, we assumed that the fluorophore intensities of each analyte type (corresponding to a specific mRNA type) subjected to a Gaussian distribution. The distribution of the mixture of analytes Gene^{H}(*i*) and Gene^{L}(*i*) corresponding to the expression levels of Gene^{H} and Gene^{L}, respectively, should subject to a two Gaussian mixture, with the proportion of 1.25 to 0.75:

$$f(x) = p\mathrm{N}(\mu_L, \sigma_L) + (1-p)\mathrm{N}(\mu_H, \sigma_H)$$

where $f(x)$ is the cumulative distribution function of the analytes Gene^{H}(*i*) and Gene^{L}(*i*) of the same analyte color *i*, *x* is the fluorescent intensity value (i.e., the measure of the gene expression levels) of beads, and $N(\cdot,\cdot)$ is the cumulative Gaussian distribution function of the fluorescent intensity of an analyte type. The parameter *p* was initialized as 0.375 based on the prior knowledge of bead portions. The objective function for GMM optimization was:

$$\underset{\mu_L,\mu_H,\sigma_L,\sigma_H}{\operatorname{argmin}} \left\| f(x) - \left(p\mathrm{N}(\mu_L, \sigma_L) + (1-p)\mathrm{N}(\mu_H, \sigma_H)\right)\right\|_F^2$$

Parameters μ_L, σ_H, μ_H, and σ_H were initialized by the fuzzy c-means clustering [10]. The expression levels of the two transcripts were determined by solving μ_L and μ_H through optimizing the objective function using the Nelder–Mead method[11].

Method 2: The EGEM score.

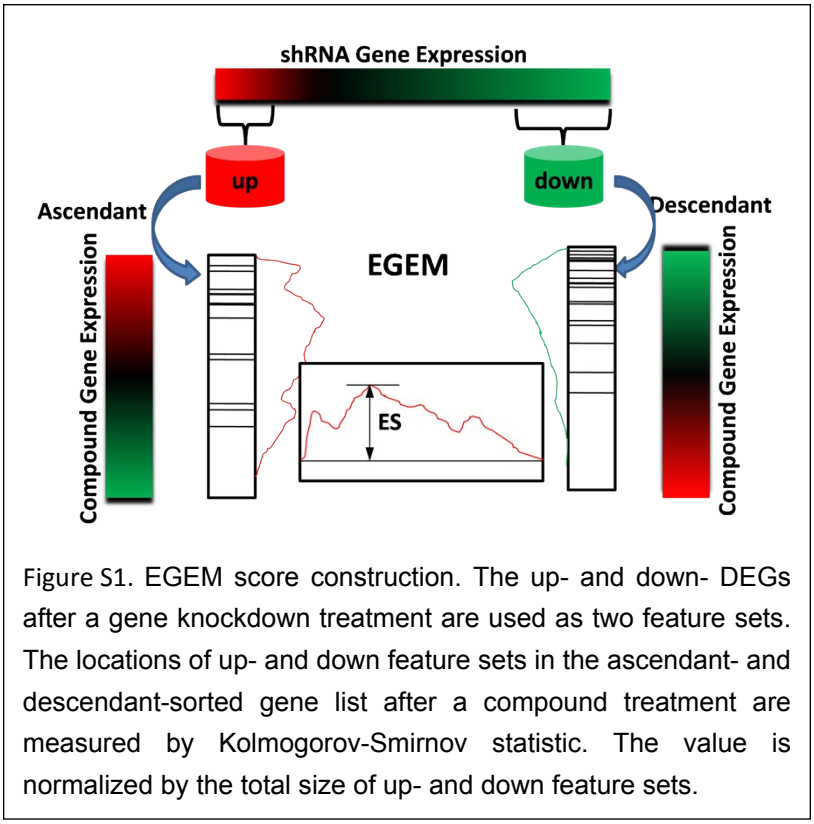

Figure S1. EGEM score construction. The up- and down- DEGs after a gene knockdown treatment are used as two feature sets. The locations of up- and down feature sets in the ascendant- and descendant-sorted gene list after a compound treatment are measured by Kolmogorov-Smirnov statistic. The value is normalized by the total size of up- and down feature sets.

We defined the EGEM score to describe the similarity between the treatments of a compound and an shRNA targeting a gene using the mutual enrichment of their resultant differential expressed landmark genes. The EGEM metric was derived from the rank-based gene set enrichment analysis (GSEA) [1] and the connectivity analysis[2]. Compound treatments could be taken as “phenotypes” and the differentially expressed genes (DEGs) of a single gene knocking down treatment as a “signature gene set” in the GSEA terminology. The EGEM metric enabled gene set enrichment analysis against the LINCS target gene reference library.

The construction of EGEM score was shown in Figure S1. *A signature gene set of a target gene*, which was composed of n DEGs after the knockdown of a target gene, among them t_{up} were up-regulated and t_{down} down-regulated. DEGs were detected according to the LFCs of the L1000 landmark genes using 1.5 IQR (interquartile range) as the threshold, which was robust against outliers. For a small molecule compound, two lists of landmark genes were used to represent the patterns of the compound-induced L1000 gene expression changes, one (p_{up}) sorted ascendantly and the other (p_{down}) descendantly based on according to the LFCs.

, such, $p_1(j)$ and $p_2(j)$ were the positions of the *j*th up- and down-regulated DEGs, respectively, in their corresponding probe gene lists. The EGEM score was defined as

$$egem = \max_{i=1:t_1}\left(\frac{i}{t_1+t_2}-\frac{p_1(i)-i}{2n-t_1-t_2}\right)+\max_{j=1:t_2}\left(\frac{j}{t_1+t_2}-\frac{p_2(j)-j}{2n-t_1-t_2}\right)$$

The EGEM score ranges from -1 to 1. The absolute value of an EGEM score represents the enrichment degree. The positive or negative sign of an EGEM score indicates that the change of gene expression pattern due to knocking down the corresponding gene is similar or reversely similar to that induced by the drug treatment. The statistical significance of an EGEM score was determined by t-test against permutations of 100 times. The EGEM scores were kept only if the associated p-values were less than 0.05 and otherwise were set to zero.

We constructed an EGEM matrix by pairwisely calculating the EGEM score between each compound and each knockdown gene. We assumed that both the positive and negative EGEM scores followed normal distributions. We also assumed that the EGEM matrix was sparse by observing the fact that, among the 3,000 proteins, a compound usually only targets a limited number of them. Hence, we chose the EGEM scores with single-side p-values less than 0.05. Other scores were forced to zeroes.

Method 3: The NMF algorithm.

The basic NMF problem can be solved according to the multiplicative update rules proposed by Lee and Seung[3]:

Theorem 1 The Euclidean distance $\| A\| - WH$ is non-increasing under the update rules:

$$H_{rj} \leftarrow H_{rj} \frac{(W^T A)_{rj}}{(W^T WH)_{rj}}, W_{ir} \leftarrow W_{ir} \frac{(AH^T)_{ir}}{(WHH^T)_{ir}}. \quad \text{S1}$$

The Euclidean distance is invariant under these updates if and only if *W* and *H* are at a stationary point of the distance.

We solve the csNMF problem according to Eq. S1. Construct the following matrixes $\bar{A}$, $\bar{W}$ and $\bar{H}$ so that:

$$\bar{A} = \begin{pmatrix} A_s & O_{n\times k} & \sqrt{\lambda} W_s D \\ A_r & O_{n\times k} & \sqrt{\lambda} W_r D \\ O_{1\times m} & \sqrt{\eta\beta} e_{1\times k} & \sqrt{\beta\lambda} e_{1\times k} P \end{pmatrix}, \quad \bar{W} = \begin{pmatrix} W_s \\ W_r \\ \sqrt{\beta} e_{1\times k} \end{pmatrix}, \quad \bar{H} = \begin{pmatrix} H & \sqrt{\eta} I_k & \sqrt{\lambda} P \end{pmatrix},$$

where $e_{1\times k} \in R^{1\times k}$ is a row vector with all components equal to one and $0_{1\times n}$ is a zero vector, I_k is an identity matrix of size $k \times k$ and $0_{k\times m}$ is a zero matrix of size $k \times m$.

Then, the objective function can be written as:

$$f(\bar{W}, \bar{H}) = \frac{1}{2} \| \bar{A} - \bar{W}\bar{H} \|^2.$$

According to **Eq. S1**, the problem of csNMF can be solved by the multiplicative update rules:

$$\bar{H}_{rj} \leftarrow \bar{H}_{rj} \frac{(\bar{W}^T \bar{A})_{rj}}{(\bar{W}^T \bar{W}\bar{H})_{rj}},$$

that is:

$$H_{rj} \leftarrow H_{rj} \frac{(W_s^T A_1 + W_r^T A_2)_{rj}}{(W_s^T W_s H + W_r^T W_r H + \beta e_{k\times k} H)_{rj}}. \quad \text{S2}$$

$$\bar{W}_{ir} \leftarrow \bar{W}_{ir} \frac{(\bar{A}\bar{H}^T)_{ir}}{(\bar{W}\bar{H}\bar{H}^T)_{ir}},$$

that is:

$$W_{ir}^c \leftarrow W_{ir}^c \frac{(A^c H^T + \lambda W^c DP)_{ir}}{(W^c HH^T + \eta W^c + \lambda W^c PP)_{ir}}, \quad \text{S3}$$

where $W^c = W_s, W_r$.

Determine the optimal signatures. We constructed the adjacency matrix $C_M \in \{C_{t(W_S)}, C_{t(W_R)}, C_H\}$:

$$C_M(r,j) = \begin{cases} 1, \text{if } \dfrac{X(r,j) - mean(X(r,:))}{sd(X(r,:))} > threshould; \\ 0, \text{others} \end{cases} \quad \text{S4}$$

to determine the clustering results. Since the optimization might converge to local minimum, in order to decompose the EGEM matrix into k clusters, we permuted the original orders of the EGEM matrix,

repeated the optimization process for 30 times, and used the average of the calculated adjacency matrixes $C_{M,k}^{-}$ to determine the final clustering results (**Eq. S5**)

$$C_{M,k}^{-}(r,j) = \begin{cases} 1, \text{ if } \dfrac{\sum_{u=1}^{N} C_{M,k,u}(r,j)}{N} > 0.5; \\ 0, \text{ others} \end{cases} \quad \text{S5}$$

Determine the optimal signature number. We used the cophenetic correlation coefficient (CCC) method [4] to measure the stability of the clustering results and thus to determine the optimal cluster number. Briefly, the cosine similarity[5] between each adjacency matrix $C_{M,k,u}(r,j)$ and the average $C_{M,k}^{-}$. The cosine similarity was chosen over Pearson's correlation because it has been shown insensitive to zeroes [6] which were abundant in adjacency matrixes.

Adjacency matrix construction and signature number determination

After performing csNMF approach, EGEM matrix is decomposed into weight and coefficient matrixes. The next is to assigning the elements (compounds and genes) to different signatures. This fulfils by a adjacency matrix *C,* which is a 0-1 matrix of size *k x n*, $C_{i,j} = 1$ if element *j* is clustered to signature *i*, and $C_{i,j} = 0$ if not [7]. As to the adjacency matrixes C_M (*M* is *V, t(W₁), t(W₂)* of the csNMF results, *t(M)* is the transpose matrix of *M*), if the value of one element *j* is of the top *p* quartile of signature *i*, $C_{i,j}^{M} = 1$, otherwise $C_{i,j}^{M} = 0$. Since the objective function of csNMF is not convex in W_1, W_2 and *V*, the algorithm may converge to different local minimums on each run of optimization, based on different random initial conditions. Thus, different adjacency matrixes are obtained based on different initializations. However, it is estimated that 20-100 runs suffice for a stable average adjacency matrix $\overline{C}$ [4]. Hence, we randomly reorder the compounds and genes of EGEM matrix 30 runs, and perform csNMF. Only if one element related with one signature in at least 15 runs is one in the average adjacency matrix, otherwise, it is zero. That is to say,

$$\overline{C}_{i,j} = \begin{cases} 1 \; if \sum_{u=1..30} C_{ij}^{u} \geq 15 \\ 0 \; else \end{cases},$$

where u is the index of run times, $u \in [1:30]$.

Another crucial matter of csNMF is the chosen of signature number *k*, which needs to be determined prior to optimization. Suppose the clustering of *k* signatures is strong, the assignments of genes and compounds should vary little among different runs. Hence, we set a list of candidate signature numbers $k_r \in K$. Then, we compute the average consistence degree as to each signature number k_r. A consistence degree relating to a signature in the u^{th} run based on k_r signature number is defined as the cosine value between the adjacency matrix value of that signature of run C^u and those of the average adjacency matrix $\overline{C}$. The average consistence degree of k_r is the average consistence degree of all the signatures and all the 30 runs. Hence, the stronger clustering is the one with larger consistence degree, and the corresponding *k* is the best signature number among $k_r \in K$.

Additional Table S1: Algorithm for EGEM and csNMF

Algorithm: The Algorithm of compound signature discovery

1. Determine the up- and down- DEGs after the gene perturbations.
2. Construct EGEM matrix based on the DEGs after the gene perturbations and the gene expression patterns after the compound treatments.
3. Estimate the significance of EGEM scores and adjust the EGEM matrix.
4. Construct the PPI matrix according to the genes in the EGEM matrix.
5. Construct the objective function of csNMF, and provide the candidate number of signatures k as well as the replication number N.
6. while do
while do
Disorder the genes and compounds of the EGEM matrix randomly.
Solve the csNMF problem based on **Eq. S2** and **Eq. S3**.
Build the i^{th} connectivity matrix C_H , $C_{t(Ws)}$, $C_{t(wr)}$.
end while
Determine the signature detection results based on *k*.
Calculate the consistence degrees based on *k*.
end while
7. Determine the stable signatures of each k based on **Eq. S5**
8. Determine the optimal k and obtain the signature detection result.
9. The Biological analysis of each signature.

Additional Table S2 : Correlations of inhibitors of Sig 2.

Combination	(1,2)	(1,3)	(1,4)	(2,3)	(2,4)	(3,4)	randMean	randSD
Cov(target)	0.79	0.81	0.04	0.74	0.04	0.05	0.10	0.17
Cov(exp)	0.35	0.26	0.21	0.18	0.16	0.20	0.08	0.14
Cov(egem)	0.47	0.49	0.38	0.52	0.43	0.46	0.14	0.21

1. “ALW-II-38-3”; 2."ALW-II-49-7”; 3. “QL-XI-92”; 4. “CP724714”. Cov(target) was the correlation of the two inhibitors’ interacting rate to 450 kinases. Cov(exp) was the correlation of gene expression pattern after two inhibitors’ perturbagens. Cov(egem) was the correlation of EGEM scores of two inhibitors. randMean and randSD were the average and standard deviation of correlation of each two inhibitors combination.

Additional Figure S2: Comparison of EGEM-based and experimentally measured kinase inhibitor target similarity.

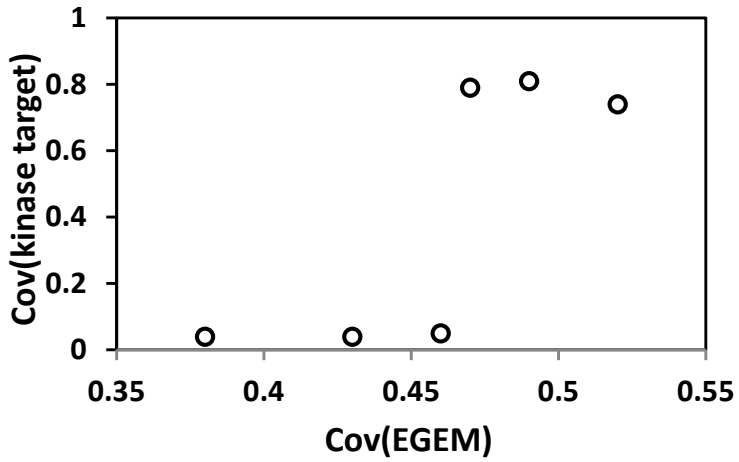

Figure S2. Comparison of EGEM-based and experimentally measured kinase inhibitor target similarity. Cov(EGEM): the similarity among kinase inhibitors ALW-II-38-3, ALW-II-49-7, QL-XI-92, and CP724714 predicted by the covariance of their EGEM scores. Cov(kinase target): the similarity among their direct kinase inhibitor assayed by KINOMEscan®.

Additional Table S3: The p-value of target gene GO similarities of each signature

	Sig 1	Sig 2	Sig 3	Sig 4	Sig 5	Sig 6
BP	0.15	0.28	0.71	**2.66e-5**	0.04	0.24
MF	0.18	**0.01**	0.90	**3.68e-3**	0.49	0.77
CC	0.11	**0.04**	0.72	**9.49e-3**	**0.02**	0.76

GO similarities were evaluated in three categories: the biological process (BP), the molecular function (MF), and the cellular component (CC). Statistically significant similarities (≤0.05) were labeled in bold font.

Additional Table S4: Univariable and multivariable survival analysis using compound signatures as well as conventional clinical features for chemotherapy.

	Univariable					Multivariable				
Variables	Hazard Ratio (95% CI)	Cl	Cu	Relative risk (z)	p-value	Hazard Ratio (95% CI)	Cl	Cu	Relative risk (z)	p-value
age	1	1	1.1	1.9	0.0678	1	0.98	1.1	1.1	0.25
size	1	1	1	1.8	0.101	1	0.99	1	0.68	0.49
pam50	1.2	0.44	3.5	0.26	0.343	0.052	0.0024	1.1	-1.9	0.06
subtype	0.92	0.29	3.4	-0.25	0.328	23	0.91	> 100	1.9	0.057
lymphnode	**2.2**	**0.99**	**4.7**	**1.9**	**0.0386**	2.8	0.74	10	1.5	0.13
ER	0.72	0.38	1.4	-1	0.307	0.33	0.1	1.1	-1.8	0.067
grade	1	0.23	4.6	0.025	0.979	0.41	0.041	4.1	-0.76	0.45
Sig1	0.99	0.53	1.9	-0.04	0.968	0.55	0.2	1.6	-1.1	0.27
Sig2	0.78	0.42	1.5	-0.77	0.443	0.98	0.32	3	-0.035	0.97
Sig3	0.95	0.5	1.8	-0.16	0.871	0.55	0.19	1.6	-1.1	0.28
Sig4	**2.3**	**1.2**	**4.5**	**2.5**	**0.00956**	**3.8**	**1.4**	**10**	**2.6**	**0.0092**
Sig5	**2**	**1.1**	**3.7**	**2.1**	**0.0331**	1.5	0.56	3.8	0.78	0.43
Sig6	1.2	0.64	2.3	0.59	0.558	1.8	0.63	5.4	1.1	0.27
Sig7	0.97	0.52	1.8	-0.088	0.93	0.96	0.27	3.4	-0.071	0.94
Sig8	1.2	0.62	2.2	0.5	0.616	0.46	0.13	1.6	-1.2	0.22

Additional Table S5: Univariable and multivariable survival analysis using compound signatures as well as conventional clinical features for Tamoxifen treatment.

	Univariable					Multivariable				
Variables	Hazard Ratio (95% CI)			Relative risk	p-value	Hazard Ratio (95% CI)			Relative risk	p-value
		Cl	Cu	(z)			Cl	Cu	(z)	
age	1	0.99	1	0.92	0.359	1	0.98	1	0.35	0.73
size	1	1	1	6.9	< 0.0001	1	1	1	4	< 0.0001
pam50	1	0.51	2.2	-0.34	< 0.0001	2.1	0.92	4.8	1.8	0.076
subtype	0.81	0.42	1.6	-1.2	< 0.0001	0.22	0.067	0.7	-2.6	0.011
lymphnode	2.4	1.7	3.3	5	< 0.0001	1.7	1.1	2.6	2.4	0.018
ER	0.39	0.22	0.71	-3.1	0.00651	0.52	0.24	1.1	-1.7	0.097
grade	5.4	2.3	13	3.9	< 0.0001	2.4	0.99	6	1.9	0.054
Sig1	0.88	0.63	1.2	-0.73	0.466	0.9	0.57	1.4	-0.47	0.64
Sig2	1.1	0.78	1.5	0.5	0.616	1.6	0.96	2.6	1.8	0.069
Sig3	1.1	0.8	1.6	0.67	0.501	1.3	0.85	2	1.2	0.21
Sig4	1.2	0.83	1.6	0.89	0.372	1.4	0.89	2.1	1.4	0.15
Sig5	0.93	0.66	1.3	-0.44	0.657	0.84	0.55	1.3	-0.82	0.41
Sig6	0.98	0.7	1.4	-0.14	0.888	0.81	0.53	1.2	-0.95	0.34
Sig7	0.97	0.69	1.4	-0.17	0.866	1.4	0.84	2.3	1.3	0.2
Sig8	0.84	0.6	1.2	-1	0.303	0.63	0.39	1	-2	0.05

Additional Figure S3: Association between breast cancer compound signatures and clinical traits.

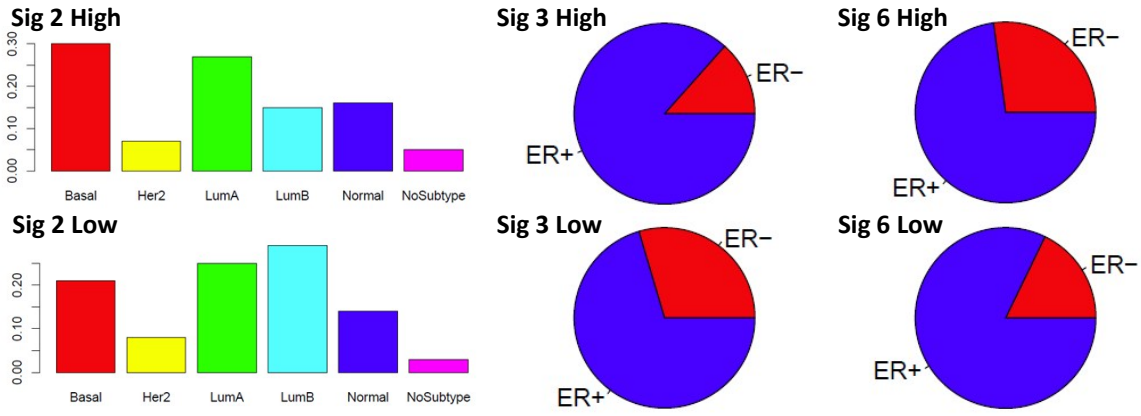

Figure S3. Association between breast cancer compound signatures and clinical traits.

REFERENCES

1. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America. 2005;102(43):15545-50.
2. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. science. 2006;313(5795):1929-35.
3. Lee DD, Seung HS, editors. Algorithms for non-negative matrix factorization. Advances in neural information processing systems; 2001.
4. Brunet J-P, Tamayo P, Golub TR, Mesirov JP. Metagenes and molecular pattern discovery using matrix factorization. Proceedings of the national academy of sciences. 2004;101(12):4164-9.
5. Cheetham AH, Hazel JE. Binary (presence-absence) similarity coefficients. Journal of Paleontology. 1969:1130-6.
6. Leydesdorff L. Similarity measures, author cocitation analysis, and information theory. Journal of the American Society for Information Science and Technology. 2005;56(7):769-72.
7. Mejía-Roa E, Carmona-Saez P, Nogales R, Vicente C, Vázquez M, Yang XY, et al. bioNMF: a web-based tool for nonnegative matrix factorization in biology. Nucleic Acids Research. 2008;36(suppl 2):W523-W8.