Supplementary Material: Improving Deconvolution Methods in Biology through Open Innovation Competitions: An Application to the Connectivity Map

March 16, 2021

A Supplementary Material

A.1 Clustering by method and perturbation type

The goal of this section is to provide a high-level overview of the differences in the deconvolution data (DECONV) and the corresponding differential expression (DE) generated for each solution.

We used a t-SNE projection of the samples generated by UNI and by applying a deconvolution algorithm to DUO data. t-SNE was run twice and separately for the deconvolution (DECONV) and differential-expression (DE) data; on the holdout datasets only. In both cases, the corresponding UNI or ground-truth (GT) data was included as well.

Results show that the DECONV data clusters well by perturbation type (Fig. 1, **a**), reflecting the different nature of the shRNA and compound treatments. Deconvolution algorithms in the same class also tend to have similar values (Fig. 1, **c**). For example, the points for the decision tree regressor (DTR) algorithms overlap substantially in the projection, as do the k-means and Gaussian mixture model (GMM) algorithms. However, after the transformation to DE data, the t-SNE projection is more homogenous, with no particular clustering by perturbation and algorithm type (Fig. 1, **b**).

While it can be hard to interpret the clustering at the DECONV level without further analysis, the lack of clustering by algorithm in DE space suggests that there are no global biases introduced by algorithm type. This is reassuring because it means that each approach could be considered a viable solution and the outputs of each are likely to be interoperable in the downstream analysis.

[Figure 1 about here.]

A.2 List of compound treatments

This table shows componud perturbagen names (`pert_iname`), unique id (`pert_id`), time of treatment (`pert_itime`), dose (`pert_idose`), and number of replicates (`num_replicates`).

[Table 1 about here.]

A.3 List of Short-hairpin (shRNA) treatments

This table shows shRNA perturbagen names (`pert_iname`), unique id (`pert_id`), and number of replicates (`num_replicates`).

[Table 2 about here.]

A.4 Data availability and implementation

- The contest data are available in the Clue.io data library https://clue.io/data/CT#CT_DPEAK.
- The source codes of the solutions along with Docker containers that include all the dependencies needed to run the codes are available in the CMap Github repository https://github.com/cmap/gene_deconvolution_challenge
- A Docker container used for converting the deconvolution data to differential expression values is available in the Docker Hub https://hub.docker.com/r/cmap/sig_2to4_tool.
- A collection of scripts in the language R used to generate tables and figures are available in the CMap Github repository https://github.com/cmap/deconv.

A.5 Scoring function

This appendix describes the scoring function used in the contest to evaluate the performance of the competitors' submissions.

Submissions were scored based on a scoring function that combines measures of accuracy and computational speed. Accuracy measures were obtained by comparing the contestant's predictions, which were derived from DUO data, to the equivalent UNI ground truth data generated from the same samples.

The scoring function combines two measures of accuracy: correlation and AUC, which are applied to deconvoluted ($DECONV$) data and one to differential expression (DE) data, respectively.

DE is derived from DECONV by applying a series of transformations (parametric scaling, quantile normalization, and robust z-scoring) that are described in detail in Subramanian *et al.* (2017). The motivation for scoring DE data in addition to $DECONV$ is because it is at this level where the most biologically interesting gene expression changes are observed. Of particular interest is obtaining significant improvement in the detection of, so called, "extreme modulations." These are genes that notably up- or down-regulated by perturbation and hence exhibit an exceedingly high (or low) DE values relative to a fixed threshold.

The first accuracy component is based on the Spearman rank correlation between the predicted $DECONV$ data and the corresponding UNI ground truth data.

For a given dataset p, let $M_{\mathrm{DUO},p}$ and $M_{\mathrm{UNI},p}$ denote the matrices of the estimated gene intensities for $G = 976$ genes (rows) and $S = 384$ experiments (columns) under DUO and UNI detection. Compute the Spearman rank correlation matrix, ρ, between the rows of these matrices and take the median of the diagonal elements of the resulting matrix (i.e., the values corresponding to the matched experiments between UNI and DUO) to compute the median correlation per dataset,

$$\mathrm{COR}_p = median(diag(\rho(M_{\mathrm{DUO},p}, M_{\mathrm{UNI},p}))).$$

The second component of the scoring function is based on the Area Under the receiver operating characteristic Curve (AUC) that uses the competitor's DE values at various thresholds to predict the UNI's DE values being higher than 2 ("high") or lower than -2 ("low").

For a given dataset p, let $\mathrm{AUC}_{p,c}$ denote the corresponding area under the curve where $c = \{\mathrm{high}, \mathrm{low}\}$; then, compute the arithmetic mean of the area under the curve per class to obtain the corresponding score per dataset:

$$\mathrm{AUC}_p = (\mathrm{AUC}_{p,\mathrm{high}} + \mathrm{AUC}_{p,\mathrm{low}})/2.$$

These accuracy components were integrated into a single aggregate scores:

$$\mathrm{SCORE} = \mathrm{SCORE}_{\max} \cdot (\max(\mathrm{COR}_p, 0))^2 \cdot \mathrm{AUC}_p \cdot \exp(-T_{\mathrm{solution}}/(3 \cdot T_{\mathrm{benchmark}})),$$

where T_{solution} is the run time for deconvoluting the data in each plate, and $T_{\mathrm{benchmark}}$ is the deconvolution time required by the benchmark dpeak implementation.

A.6 Extreme modulations

[Table 3 about here.]

References

Subramanian,A. *et al.* (2017) A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, **171**, 1437–1452.

Table 1: *list of compound treatments*

pert_iname	pert_id	pert_itime	pert_idose	num_replicates
abiraterone(cb-7598)	BRD-K50071428	24 h	10 um	11
acalabrutinib	BRD-K64034691	24 h	10 um	11
afatinib	BRD-K66175015	24 h	10 um	11
artesunate	BRD-K54634444	24 h	10 um	11
azithromycin	BRD-K74501079	24 h	10 um	11
diprolene	BRD-K58148589	24 h	10 um	11
CGS-21680	BRD-A81866333	24 h	10 um	11
chelidonine	BRD-K32828673	24 h	10 um	11
clobetasol	BRD-K84443303	24 h	10 um	11
digoxin	BRD-A91712064	24 h	10 um	11
disulfiram	BRD-K32744045	24 h	10 um	10
emetine hcl	BRD-A77414132	24 h	10 um	10
eplerenone	BRD-K19761926	24 h	10 um	11
epothilone-a	BRD-K71823332	24 h	10 um	9
flumetasone	BRD-K61496577	24 h	10 um	11
fluocinolone	BRD-K94353609	24 h	10 um	11
genipin	BRD-K28824103	24 h	10 um	11
hydrocortisone	BRD-K93568044	24 h	10 um	10
hyoscyamine	BRD-K40530731	24 h	10 um	11
indirubin	BRD-K17894950	24 h	10 um	10
L-745870	BRD-K05528470	24 h	10 um	10
nTZDpa	BRD-K54708045	24 h	10 um	11
oligomycin-a	BRD-A81541225	24 h	10 um	11
PRIMA1	BRD-K15318909	24 h	10 um	11
RITA	BRD-K00317371	24 h	10 um	11
spironolactone	BRD-K90027355	24 h	10 um	11
tanespimycin	BRD-K81473043	24 h	10 um	11
tretinoin	BRD-K71879491	24 h	10 um	10
UB-165	BRD-A14574269	24 h	10 um	11
ursolic-acid	BRD-K68185022	24 h	10 um	11
WAY-161503	BRD-A62021152	24 h	10 um	11
ZM-39923	BRD-K40624912	24 h	10 um	11

Table 2: *list of shRNA treatments*

pert_iname	pert_id	num_replicates
ABCB6	TRCN0000060320	4
ADI1	TRCN0000052275	4
ALDOA	TRCN0000052504	4
ANXA7	TRCN0000056304	4
ARHGAP1	TRCN0000307776	4
ASAH1	TRCN0000029402	4
ATMIN	TRCN0000141397	4
ATP2C1	TRCN0000043279	4
B3GNT1	TRCN0000035909	4
BAX	TRCN0000033471	4
BIRC5	TRCN0000073718	4
BLCAP	TRCN0000161355	4
BLVRA	TRCN0000046391	4
BNIP3L	TRCN0000007847	4
CALU	TRCN0000053792	4
CCDC85B	TRCN0000242754	4
CCND1	TRCN0000040038	4
CD97	TRCN0000008234	4
CHMP4A	TRCN0000150154	4
CNOT4	TRCN0000015216	4
DDR1	TRCN0000000618	4
DDX10	TRCN0000218747	4
DECR1	TRCN0000046516	4
DNM1L	TRCN0000001097	3
ECH1	TRCN0000052455	4
EIF4EBP1	TRCN0000040206	4
EMPTY_VECTOR	TRCN0000208001	15
ETFB	TRCN0000064432	4
FDFT1	TRCN0000036327	4
GALE	TRCN0000049461	4
GFP	TRCN0000072181	16
GRN	TRCN0000115978	4
GTPBP8	TRCN0000343727	4
HDGFRP3	TRCN0000107348	4
HIST1H2BK	TRCN0000106710	4
IKBKAP	TRCN0000037871	4
INPP4B	TRCN0000230838	4
INSIG1	TRCN0000134159	4
ITFG1	TRCN0000343702	3
JMJD6	TRCN0000063340	4
LBR	TRCN0000060460	4
LGMN	TRCN0000029255	4

pert_iname	pert_id	num_replicates
LPGAT1	TRCN0000116066	4
LSM6	TRCN0000074719	4
MAPKAPK2	TRCN0000002285	4
MAPKAPK3	TRCN0000006154	4
MAPKAPK5	TRCN0000000684	4
MIF	TRCN0000056818	4
MRPL12	TRCN0000072655	4
NT5DC2	TRCN0000350758	4
NUP88	TRCN0000145079	4
PARP2	TRCN0000007933	4
PLCB3	TRCN0000000431	4
POLE2	TRCN0000233181	4
PPIE	TRCN0000049371	4
PRKAG2	TRCN0000003146	4
PSMB10	TRCN0000010833	4
PTPN6	TRCN0000011052	4
RAB11FIP2	TRCN0000322640	4
RALB	TRCN0000072956	4
RHEB	TRCN0000010425	3
RNF167	TRCN0000004100	4
RPN1	TRCN0000072588	4
SLC25A4	TRCN0000044967	4
SNX11	TRCN0000127684	4
STK25	TRCN0000006270	4
STUB1	TRCN0000007525	4
STXBP1	TRCN0000147480	4
SYPL1	TRCN0000059926	4
TATDN2	TRCN0000049828	4
TM9SF3	TRCN0000059371	4
TMEM110	TRCN0000127960	4
TMEM50A	TRCN0000129223	4
trcn0000014632	TRCN0000014632	4
trcn0000040123	TRCN0000040123	4
trcn0000220641	TRCN0000220641	4
trcn0000221408	TRCN0000221408	4
trcn0000221644	TRCN0000221644	4
TSKU	TRCN0000005222	4
UGDH	TRCN0000028108	4
USP14	TRCN0000007428	4
USP6NL	TRCN0000253832	4
VAT1	TRCN0000038193	4
VDAC1	TRCN0000029126	4
WIPF2	TRCN0000029833	4
YME1L1	TRCN0000073864	4

pert_iname	pert_id	num_replicates
ZW10	TRCN0000155335	4

Table 3: *Detection of extreme modulation of gene expression*

		Extremely modulated genes		
rank	method	True positives	False positives	Total positives
1	random forest regressor	37002	26826	63828
2	Gaussian mixture model	**37410**	30137	67547
3	modified k-means	34975	**26376**	**61351**
4	ConvNet	36328	35832	72160
5	Gaussian mixture model	36731	37375	**74106**
6	modified k-means	**34944**	27345	62289
7	boosted tree regressor	35498	31664	67162
8	modified k-means	35654	27362	63016
9	other	35640	**37447**	73087
BM	k-means	35111	37050	72161

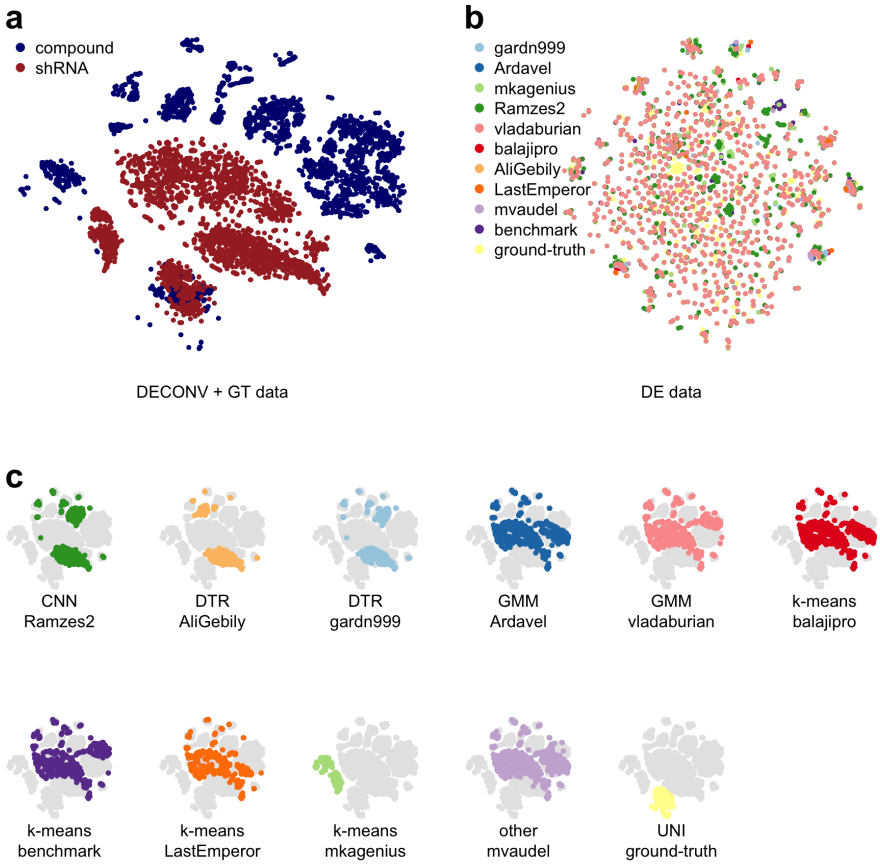

Figure 1: *t-SNE projection of deconvolution and differential expression data; both include the raw and differential expression data for the ground truth (UNI). Projections are colored and subset to generate the following panels: a, ground truth and deconvolution data colored by perturbation type (compound or shRNA treatment); b, differential expression data colored by competitor's name with benchmark and ground-truth as well; c, deconvolution data colored by competitor's name, with abbreviations for the algorithm type (Convolutional Neural Network, Decision Tree Regressor, Gaussian Mixed Model, K-means, other)*