Gene and cell line efficiency of CRISPR computed by tensor decomposition in genome-wide CRISPR-Cas9 knockout screens

Y-H Taguchi[1*] and Turki Turki[2]

[1*]Department of Physics, Chuo University, 1-13-27 Kasuga, Bunkyo-ku, Tokyo, 112-8551, Japan.
[2]Department of Computer Science, King Abdulaziz University, Jeddah, 21589, Saudi Arabia.

*Corresponding author(s). E-mail(s): tag@granular.com;
Contributing authors: tturki@kau.edu.sa;

Abstract

Genome-wide CRISPR-Cas9 knockout screens is often used to evaluate gene functions experimentally. Nevertheless, the efficacy of individual sgRNAs that target unique gene varies and is hard to integrate. In this study, tensor decomposition (TD) was used to integrate multiple sgRNAs as well as sgRNA profiles simultaneously. As a result, TD can discriminate between essential and non-essential genes with the competitive performances of one of SOTA, JACKS, which previously outperformed various SOTA. In addition to this, in spite of that TD is nothing but the simple linear algebra, TD can achieve the performance even without control samples without which JACKS even cannot be performed. Moreover, since raw and logarithmic values can achieve the similar performances through TD for the largest dataset among those tested, taking logarithmic values as has been done frequently is questioned. In conclusion, TD is the ever first method that can integrate multiple sgRNAs and profiles simultaneously and can achieve comparative performances with JACKS, one of SOTA.

Keywords: unsupervised learning, tensor decomposition, CRISPR, sgRNA, genome wide analysis, novel AI application in CRISPR-Cas9

1 Introduction

Genome-wide CRISPR-Cas9 knockout screens is useful method to evaluate the function of individual genes [1]. In contrast to the gene expression analysis, by which we can only estimate the activity of individual genes indirectly since the amount of expression does not always reflect functionality of genes, genome-wide CRISPR-Cas9 knockout screens can measure functionality of genes more directly since the functions of knock outed genes are really missing. On the other hand, since efficacy of individual sgRNA is not stable, multiple sgRNAs that target individual genes must be prepared. Then integrating the outputs from individual sgRNAs must be integrated in some way. How to integrate multiple sgRNAs into one is problematic, since it definitely affects the final results.

To address this problems, various sophisticated methods were proposed. MAGeCK (Model-based Analysis of Genome-wide CRISPR-Cas9 Knockout) computes gene-level score with assuming that the outputs can obey negative signed binomial distribution [2]. MAGeCK is also improved with learning ENCODE data set (MAGeCK MLE - Maximum Likelihood Estimation) [3]. BAGEL (Bayesian Analysis of Gene EssentialityLethality) is the Bayesian classifier that learns essential and non-essential genes as reference and necessity of individual genes are scored. BAGEL is specialized to identify essneial genes by negative selection [4]. CERES (Cancer Essentiality analysis by RNAi and CRISPR screens) can correct the effects of copy number variation and score the dependency of individual genes with integrating multiple sgRNAs [5]. STARS (Score Tracking of sgRNA enrichment) can score genes based upon the most condensed genes; STARS is usually used for the positive selection [6]. JACKS (Joint Analysis of CRISPR/Cas9 Knockout Screens) can integrate screen data including multiple samples and conditions and evaluate the activity and necessity of individual sgRNA more precisely [7].

Besides these CRISPR specific methods, tensor decomposition (TD) is known to have ability to integrate multiple outputs [8], regardless to what the multiple outputs are. TD can integrate not only gene expression but also various features including DNA methylation, histone modification, miRNAs, and protein-protein interaction. Thus, it is not surprising even if TD can successfully integrate the outputs from mutiple sgRNAs that target the unique gene. In addition to the integration of multiple sgRNAs, TD also can integrate multiple sgRNA profiles themselves at the sametime whereas none of above mentioned methods can do this; individual profiles must be analyzed separately one by one and only their outcomes can be integrated.

To know if TD can successfully integrate the outputs from multiple sgRNAs that target the unique gene as well as multiple sgRNA profiles simultaneously, we applied TD to multiple datasets of sgRNA profiles to which multiple methods are tested [7] in the paper where JACKS was proposed. As a result, TD successfully discriminates essential genes from non-essential genes as much as JACKS that previously outperforms various conventional methods [7], although TD is forced to analyze datasets without control samples, without which other method even cannot be performed. It is a great advantage of TD over various conventional methods that made use of complicated advanced modeling.

2 Results

Figure 1 shows the outline of the analysis flow.

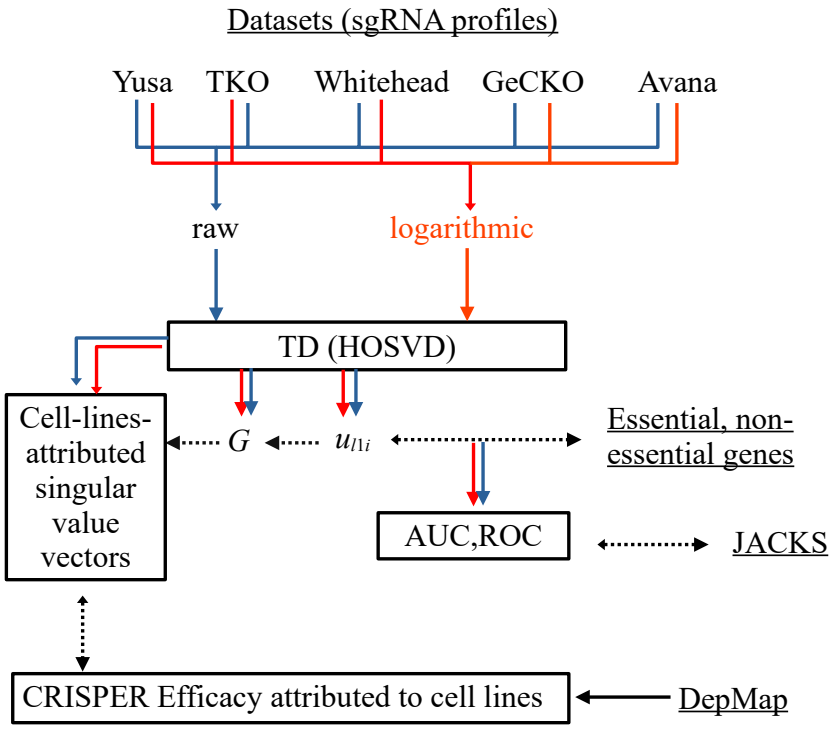

Fig. 1 Analysis flow in this study. Starting from five datasets composed of sgRNA profiles, they are formatted as tensors with raw and logarithmic values. TD (tensor decomposition) is applied to tensors and we get G (core tensors) as well as singular value vectors attributed to cell lines, genes, and sgRNAs. Singular value vectors attributed to genes (denoted as $u_{\ell_1 i}$) are evaluated by discriminating between essential and non-essential genes. Their performances are compared with those of JACKS. The cell-lines-attributed singular value vectors are selected based upon singular value vectors attributed to genes through Gs. The selected cell-lines-attributed singular value vectors are evaluated and some of them are further compared with CRISPR efficacy attributed to cell lines from DepMap data base. Blue and red lines are analysis flows for raw and logarithmic values, respectively. Dotted arrows show the comparisons to evaluate the performances of TD. Data sources of terms underlines are from outside ones.

2.1 AUC, ROC and comaprison with JACKS

At first, we tried to identify which singular value vector attributed to genes, $u_{\ell_1 i}$, can discriminate essential genes from non-essential genes to measure AUC best. Table 1 lists the result; the performances in Table 1 are generally good, since AUC is mostly

Table 1 The "best" singular value vectors attributed to genes, $u_{\ell_1 i}$, that can discriminate essential genes from non-essential genes. For Whitehead datasets, since there are two singular value vectors attributed to genes that have similar good performances, we included both in the table.

Cell lines	raw or logarithmic	ℓ_1	AUC
Yusa V1.0	raw	7	0.79
	logarithmic	6	0.77
TKv1	raw	1	0.68
	logarithmic	5	0.77
Whitehead	raw	11	0.81
	logarithmic	2	0.79
	logarithimc	12	0.78
GeCKOv2	raw	5	0.77
	logarithmic	5	0.81
Avana	raw	5	0.79
	logarithmic	5	0.79

as large as 0.8. To compare these performances with those by the previous methods, we have compared ours with those by JACKS that outperformed numerous earlier proposed methods (Table 2) [7]. Obviously, the performances are quite comparative.

Table 2 AUC for the discrimination between essential genes (positive set) and non-essential genes (negative set) achieved by JACKS

cell lines	Yusa V1.0	TKv1	Whitehead	GeCKOv2	Avana
AUC	0.83	0.82	0.85	0.79	0.83

In addition to this, the comparison of ROCs themselves reveals that JACKS and TD do not only achieve quite similar AUC (Tables 1 and 2) but also share similar shapes of ROC, too (Fig. 2). Since the same amount of AUC does not always mean that the shapes of ROC are similar, the fact that TD and JACKS share the shapes of ROC themselves mean that TD and JACKS work in quite similar mechanisms. It is a bit remarkable since TD is nothing but the consequence of simple linear algebra whereas JACKS employed advanced and sophisticated Bayesian method and TD does not even include control samples, without which JACKS is impossible to perform, for some of five datasets (i.e., GeCKOv2 and Avana).

In conclusion, TD can achieve comparative performances toward one of SOTA methods despite of simple linear algebra as well as missing controls.

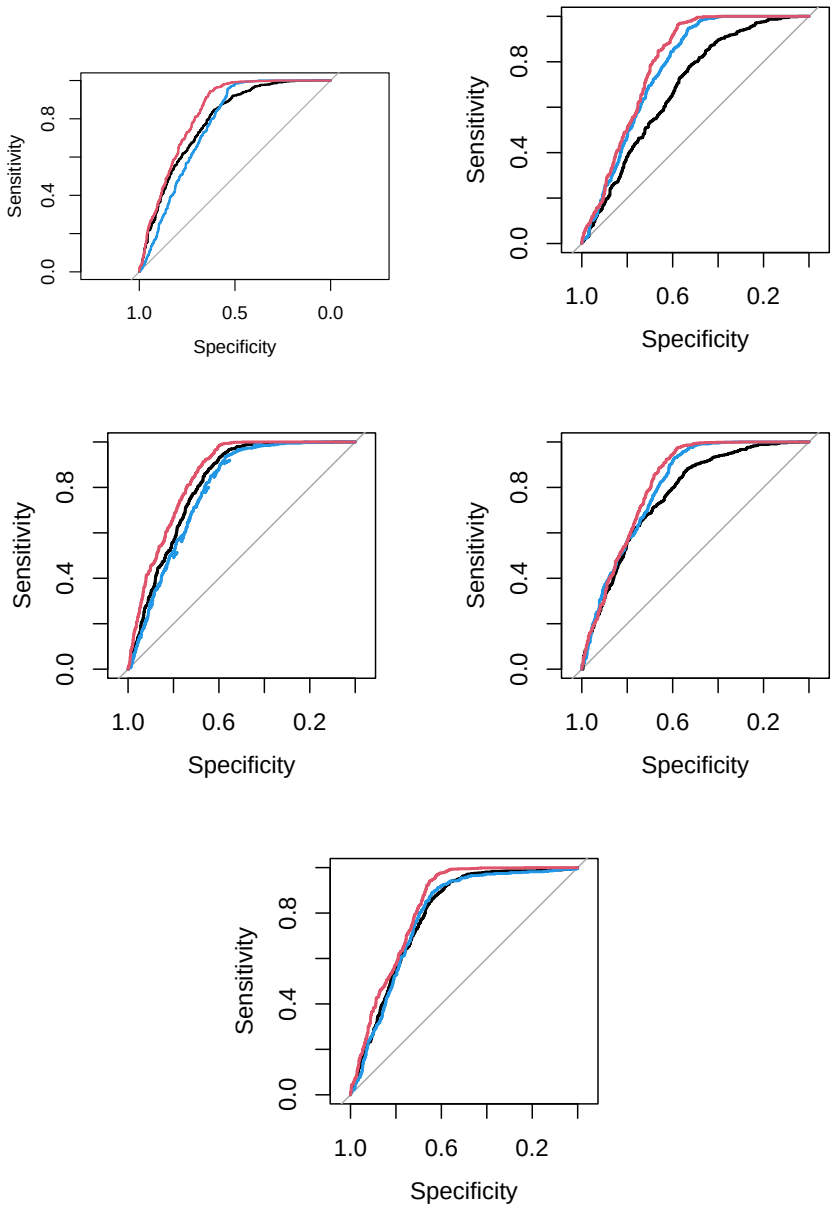

Fig. 2 ROC for the discrimination between essential genes (positive set) and non-essential genes (negative set) achieved by TD and JACKS. The corresponding AUC values are in Tables 1 and 2. Black: TD with raw value, blue: TD with logarithmic values, red: JACKS. Top left: Yusa V1.0, top right: TKOv1, mid left: Whitehead, mid right: GeCKOv2, bottom Avana.

2.2 Singular value vectors attributed to cell lines

Although TD successfully achieved the comparative performance toward one of SOTA, JACKS, there are many things to discuss. First of all, to clarify what decides which singular value vectors attributed to gene perform the best one is important. As can be seen in Table 1, ℓ_1s associated with the best performance heavily fluctuate. For example, the twelveth singular value vectors attributed to genes, u_{12i} can achieved the almost similar amount of AUC with the first one, in the Whitehead dataset. Generally speaking, as ℓ_1 increases, contribution to the original tensor decreases. Thus, the larger ℓ_1 less likely have biological meanings because of noisy nature. Why can such large ℓ_1 have biological meaning for the Whitehead dataset?

Table 3 Summary of the selected singular value vectors. At first, ℓ_1s (attributed to genes) are selected based upon AUC (Table 1). Then Gs with the largest absolute values are selected (Figs. 3, 4, 6, 8, and 9). ℓ_2, ℓ_3, and ℓ_4 (attributed to genes and sgRNA) are selected based on the selected Gs.

	Cell lines				
Singular value vectors	Yusa V1.0	TKv1	Whitehead	GeCKOv2	Avana
	Raw values				
Selected G	$G(7,1,4)$	$G(1,1,1)$	$G(11,1,2)$	$G(5,1,2,1)$	$G(5,3,1,1)$
Attributed to genes Selected ℓ_1	$\ell_1 = 7$	$\ell_1 = 1$	$u_{\ell_1 i}$ $\ell_1 = 11$	$\ell_1 = 5$	$\ell_1 = 5$
Attributed to cell lines Selected ℓ_3 or ℓ_2	$\ell_3 = 4$	$\ell_3 = 1$	$u_{\ell_3 k}$ $\ell_3 = 2$	$\ell_3 = 2$	$u_{\ell_2 j}$ $\ell_2 = 3$
Attributed to sgRNAs Selected ℓ_2 or ℓ_4	$\ell_2 = 1$	$\ell_2 = 1$	$u_{\ell_2 j}$ $\ell_2 = 1$	$\ell_2 = 1$	$u_{\ell_4 m}$ $\ell_4 = 1$
	Logarithmic values				
Selected G	$G(6,1,2)$	$G(5,1,2)$	$G(2,1,2)$, $G(12,1,3)$	$G(5,1,2,1)$	$G(5,2,1,1)$
Attributed to genes Selected ℓ_1	$\ell_1 = 6$	$\ell_1 = 5$	$u_{\ell_1 i}$ $\ell_1 =$ 2, 12	$\ell_1 = 5$	$\ell_1 = 5$
Attributed to cell lines Selected ℓ_3 or ℓ_2	$\ell_3 = 2$	$\ell_3 = 2$	$u_{\ell_3 k}$ $\ell_3 =$ 2, 3	$\ell_3 = 2$	$u_{\ell_2 j}$ $\ell_2 = 2$
Attributed to sgRNAs Selected ℓ_2 or ℓ_4	$\ell_2 = 1$	$\ell_2 = 1$	$u_{\ell_2 j}$ $\ell_2 = 1$	$\ell_2 = 1$	$u_{\ell_4 m}$ $\ell_m = 1$

To interpret the singular value vectors attributed to genes, it is better to find which singular value vectors attributed to cell lines are related to the selected singular value vectors attributed to genes (Table 3). For this purpose, it is easier to inspect which $G(\ell_1\ell_2\ell_3)$ or $G(\ell_1\ell_2\ell_3\ell_4)$ has larger absolute values with fixed ℓ_1. Then by inspecting singular value vectors attributed to cell lines with the specified G, we can know why the specific ℓ_1 is associated with the best performance. In the following, we discuss this point for one by one among the five datasets, since the interpretation is not common.

2.2.1 Yusa V1.0

At first, we consider Yusa V1.0. Upper two panels in Figure 3 show that the absolute values of G in descending order as well as boxplot of $u_{\ell_3 k}$s identified based upon Gs. For raw values and logarithmic values, $G(7,1,4)$ and $G(6,1,2)$ has the largest absolute values with the fixed $\ell_1 = 7$ and $\ell_1 = 6$, respectively. When investigating

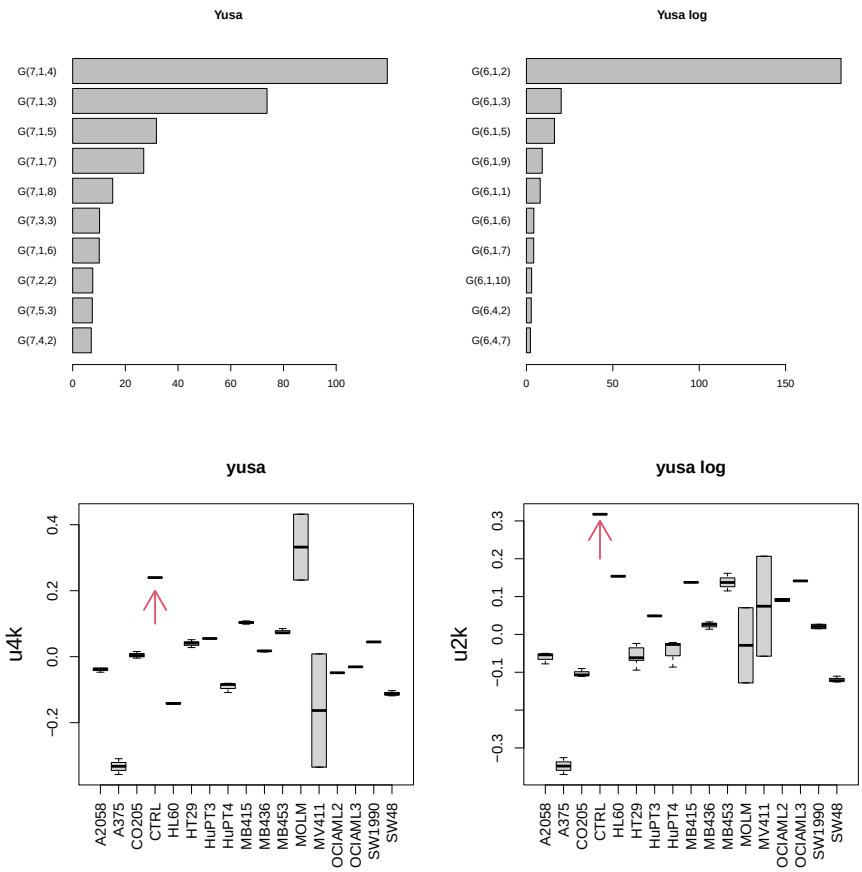

Fig. 3 Upper row: The absolute values of $G(\ell_1\ell_2\ell_3)$ in the descending order for Yusa V1.0 dataset. $\ell_3 = 4$ and $\ell_3 = 2$ are selected for raw values (left) and logarithmic vales (right), respectively. Lower raw: The singular value vectors attributed to cell lines and associated with the largest absolute G with the fixed ℓ_1. Left: u_{4k} for raw values and right: u_{3k} for logarithmic values. Red arrows denote controls.

corresponding singular value vectors attributed to cell lines, u_{4k} for raw values and u_{2k} for logarithmic values, we can know that control profiles are apart from others if we observe the two lower panels in Fig. 3 (no other singular value vectors attributed to cell lines exhibit the separation of controls from other cell lines as large as these ones, not shown here). Thus, the reason why some specific singular value vectors attributed

to genes are associated with the largest AUC is because they are associated with the cell-lines-attributed singular value vectors that have the largest separation between controls and others.

2.2.2 TKOv1

The similar observations stand for other datasets, too. For TKOv1, $G(1,1,1)$ for raw values with $\ell_1 = 1$ and $G(5,1,2)$ for logarithmic values with $\ell_1 = 5$ have the largest absolute values, respectively (Fig. 4). This means that u_{1k} is expected to be associated

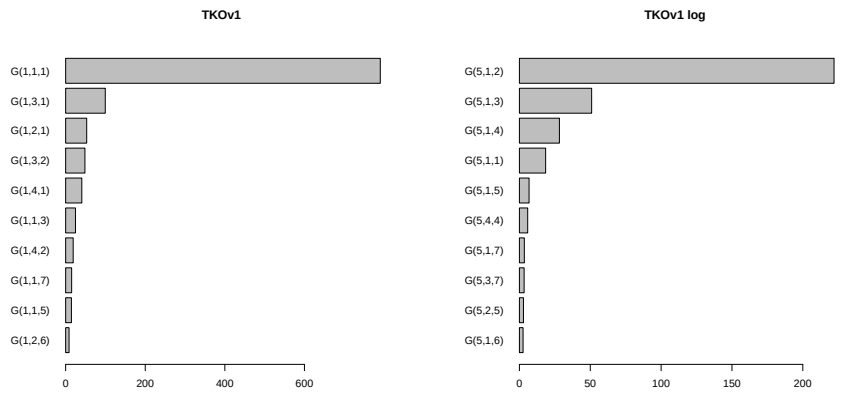

Fig. 4 The absolute values of $G(\ell_1\ell_2\ell_3)$ in the descending order in TKOv1 dataset. $\ell_1 = 1$ and $\ell_1 = 5$ are selected for raw values (left) and logarithmic vales (right).

with the selected u_{1i} for raw values whereas u_{2k} is expected to be associated with the selected u_{5i}. Thus, we should expect that u_{1k} for raw values and u_{2k} for logarithmic values have clear separation between T0 (in some sense, it is regarded to be control, since T0 is at the time point where the treatment is performed) and others. In cotrast to the expectation, no $u_{\ell_3 k}$s attributed to cell lines exhibit clear separation between T0 and others for raw values whereas u_{2k} attributed to cell lines exhibit clear separation between T0 and others for logarithmic values as expected (Fig. 5). This possibly coincides with the observation that raw values for TKOv1 dataset has the lowest AUC, 0.68 (Table 1). This means that the success of TD again depends upon if we can occasionally get cell-lines-attributed singular value vectors associated with the separation between control and others.

2.2.3 Whitehead

The Whitehead dataset also has similar outcomes. This data sets are composed of pairs between initial and final. $G(11,1,2)$ for raw values with $\ell_1 = 1$, $G(2,1,2)$ for logarithmic values with $\ell_1 = 2$, and $G(12,1,3)$ for logarithmic values with $\ell_1 = 12$ have the largest absolute values (Fig. 6). Thus, $\ell_3 = 2$ is selected for raw values and $\ell_3 = 2,3$ are selected for logarithmic values. As can be seen in Fig. 7, u_{2k} for

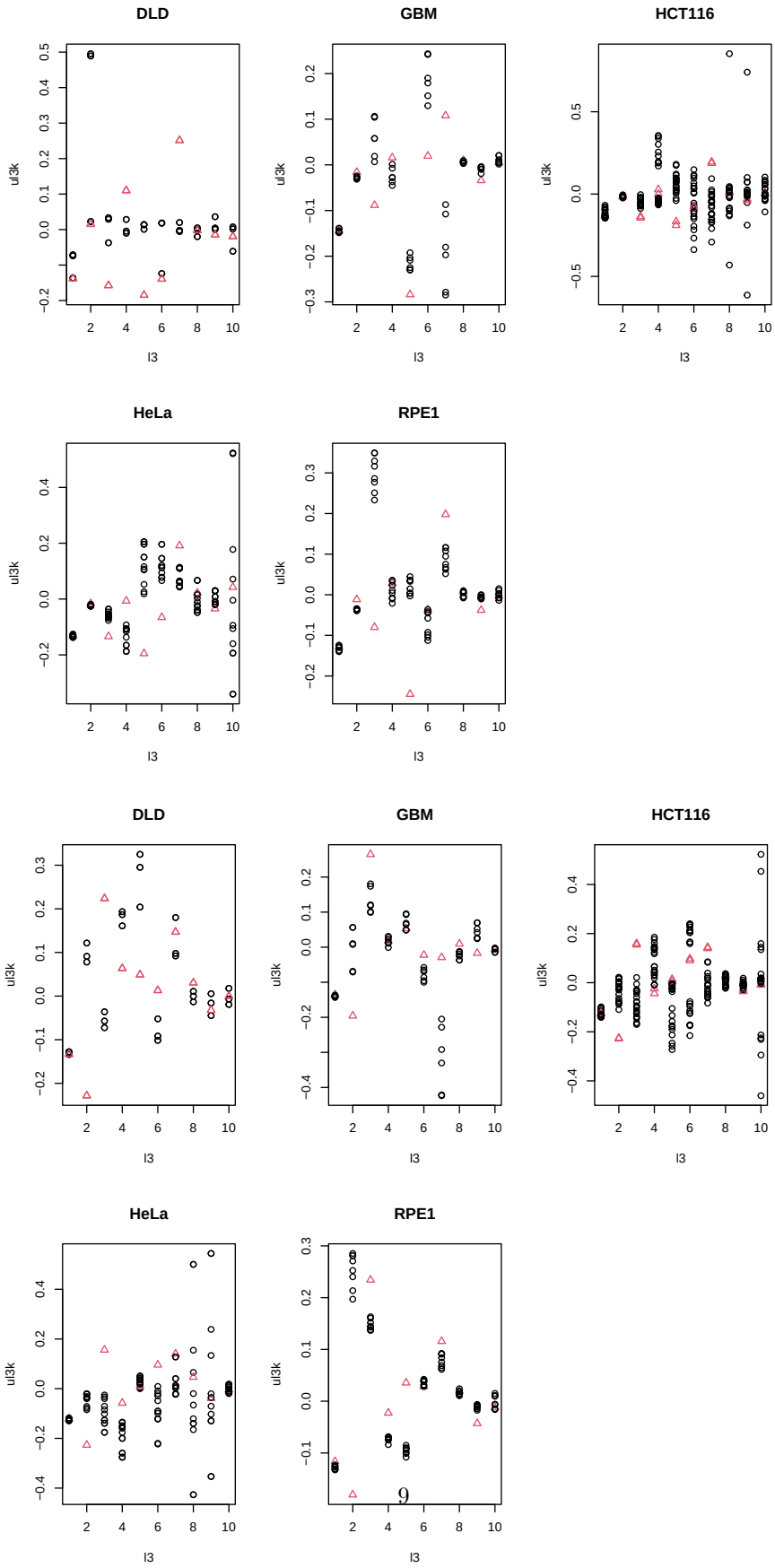

Fig. 5 The singular value vectors attributed to cell lines, $u_{\ell_3 k}$, for five cell lines in TKOv1 data sets. The first and second rows: for raw values and the third and fourth rows: for logarithmic values. Red triangles: T0 and black circles: others.

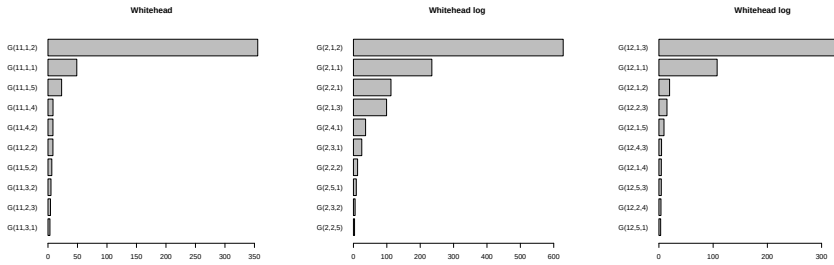

Fig. 6 The absolute values of $G(\ell_1\ell_2\ell_3)$ in the descending order in Whitehead dataset. Left: for raw values with $\ell_1 = 11$ ($\ell_3 = 1$ is selected), middle: for logarithmic values with $\ell_1 = 2$ ($\ell_3 = 2$ is selected), and right: for logarithmic values with $\ell_1 = 12$ ($\ell_3 = 3$ is selected).

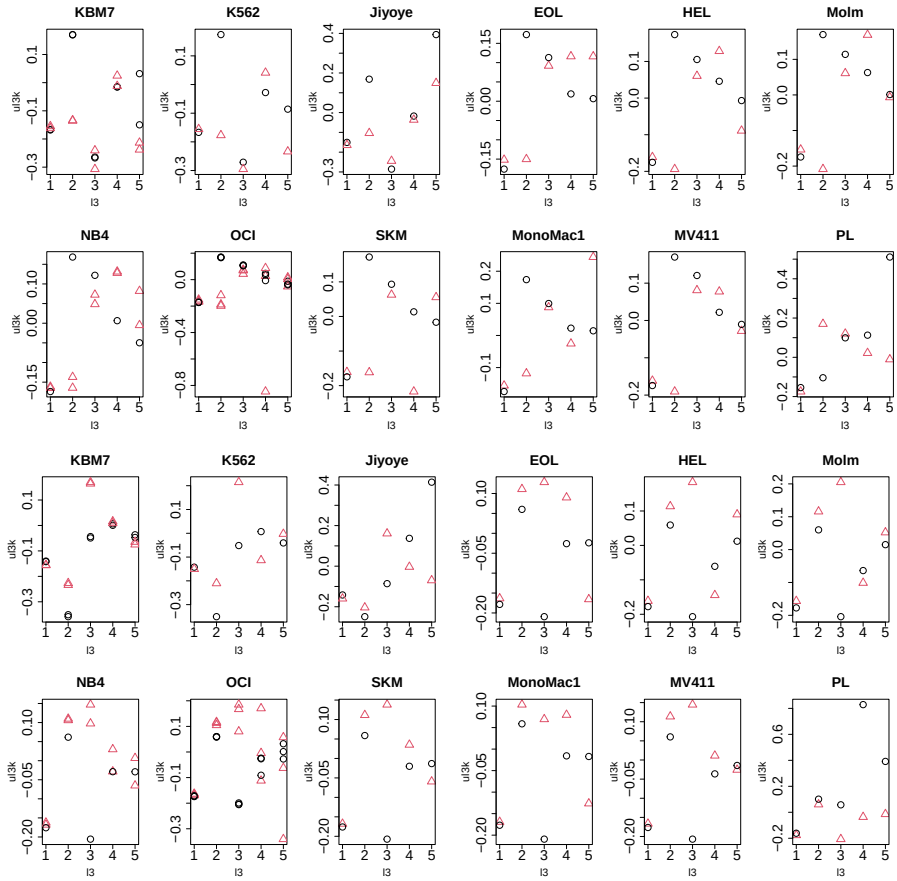

Fig. 7 The singular value vectors attributed to cell lines, $u_{\ell_3 k}$, for twelve cell lines in Whitehead data sets. the first and second rows: for raw values and the third and fourth rows: for logarithmic values. Red triangles: final and black circles: initial.

raw values and u_{2k} as well as u_{3k} for logarithmic values are associated with large separation between initial and final. Since initial can be regarded to be control in some sense, again large AUC is available when singular value vectors attributed to genes are associated with the cell-lines-attributed singular value vectors that are associated with large separation between control and others.

2.2.4 GeCKOv2

The above results that large AUC is associated with the cell-lines-attributed singular value vectors that are associated with large separation between control and others are quite reasonable. Nevertheless, inclusion of controls is not always required condition to have large AUC as can be seen below. As denoted in the Methods, controls are dropped during the screening process (at least three biological replicates) for GeCKOv2 data sets. Thus, GeCKOv2 data sets can get large AUC in spite of missing controls. To consider this, we tried to interpret the cell-lines-attributed singular value vectors associated with the gene-attributed singular value vectors associated with large AUC without dependence upon control. u_{2k} was selected (Upper two panels in Fig. 8) as cell-lines-attributed singular value vectors associated with the largest absolute $G(\ell_1\ell_2\ell_3\ell_4)$ with the selected singular value vectors attributed genes, u_{5i}, that has large AUC (Table 1). We have downloaded CRISPR efficacy attributed to cell lines from DepMap database and found that u_{2k}s are highly correlated with the downloaded CRISPR efficacy attributed to cell lines (Lower two panels in Fig. 8).

2.2.5 Avana

This also stands for Avana dataset. u_{3j} and u_{2j} were selected for raw and logarithmic values, respectively (Upper two panels in Fig. 9) as cell-lines-attributed singular value vectors associated with the largest absolute $G(\ell_1\ell_2\ell_3\ell_4)$ with the selected singular value vectors attributed genes, u_{5i}, that has large AUC (Table 1). High AUC is associated with the high correlation between CRESIPER Efficacy and the cell-lines-attributed singular value vectors, u_{3j} and u_{2j} (Lower two panels in Fig. 9). Avana dataset also lacks control.

Thus it is obvious that TD can achieve high AUC with identifying cell-lines-attributed singular value vectors associated with CRISPR Efficacy, even without controls.

3 Methods

3.1 sgRNA profiles and tensor format

We have employed as many as five datasets composed of sgRNA profiles tested in JACKS paper [7] and they are formatted as tensors for the downward analysis. All sgRNA profiles files were retrieved from Data.zip available at https://www.doi.org/10.6084/m9.figshare.6002438.

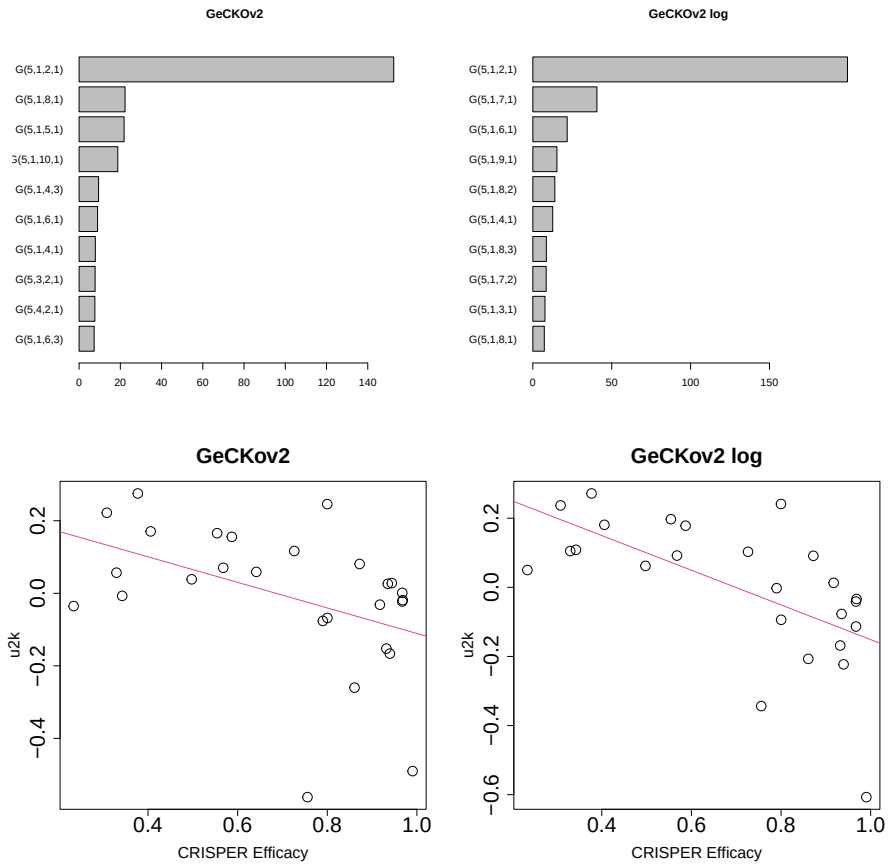

Fig. 8 Upper row: The absolute values of $G(\ell_1\ell_2\ell_3)$ in the descending order for GeCKOv2 dataset. $\ell_3 = 2$ is selected for raw values (left) and logarithmic vales (right). Lower row: Scatter plot between u_{2k} and CRISPR Efficacy. Correlation coefficients are -0.45 (left, raw values) and -0.62 (right, logarithmic values), respectively. Red lines are regression line.

3.1.1 Dataset 1: Yusa V1.0

The file extracted from Data.zip and used for the analysis is yusa_raw_v10.tab. Yusa V1.0 [9, 10] is composed of sixteen cell lines as well as one control. Since the number of replicates associated with individual cell lines varies, in total as many as 53 profiles are included. Although the number of sgRNAs that target individual genes also varies, we consider as many as 17,019 genes associated with at least five sgRNAs. When more than five sgRNAs target one genes, randomly selected five sgRNAs are considered. As a result, we have a tensor $x_{ijk} \in \mathbb{R}^{17019\times5\times53}$ that represent the number of jth sgRNAs that target ith gene of kth profile.

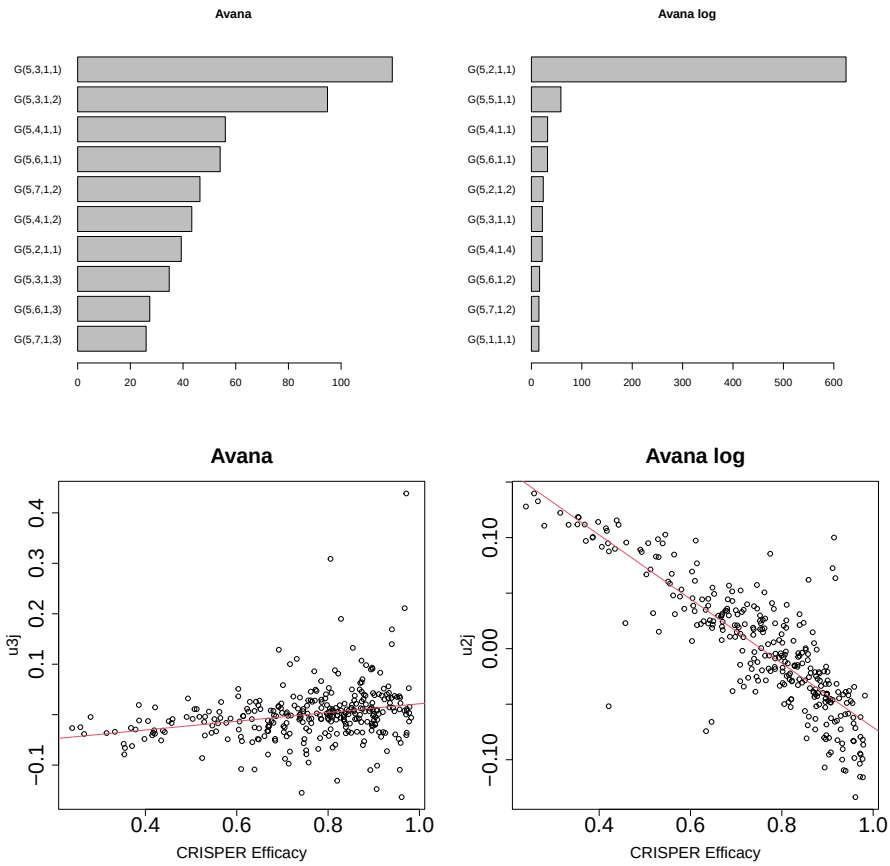

Fig. 9 Upper row: The absolute values of $G(\ell_1\ell_2\ell_3)$ in the descending order for Avana dataset. $\ell_2 = 3$ and $\ell_2 = 2$ are selected for raw values (left) and logarithmic vales (right), respectively. Lower row: Scatter plot between u_{3j} (left, raw values) or u_{2j} (right, logarithmic values) and CRISPR Efficacy. Correlation coefficients are 0.25 (left, raw values) and -0.84 (right, logarithmic values), respectively. Red lines are regression line.

3.1.2 Dataset 2: TKOv1

The file extracted from Data.zip and used for the analysis is tko_counts.txt. TKOv1 [11] is composed of six cell lines. Since the number of replicates associated with individual cell lines varies, in total as many as 61 profiles are included. Although the number of sgRNAs that target individual genes also varies, we consider as many as 14,770 genes associated with at least four sgRNAs. When more than four sgRNAs target one genes, randomly selected four sgRNAs are considered. As a result, we have a tensor $x_{ijk} \in \mathbb{R}^{14770\times4\times61}$ that represent the number of jth sgRNAs that target ith gene of kth profile. Remarkably, all cell lines are associated with at least one T0 sample (i.e., not treated one) within replicates.

3.1.3 Dataset 3: Whitehead

The file extracted from Data.zip and used for the analysis is Wang2015_2017_merged_counts.txt. Whitehead [12, 13] is composed of twelve cell lines. Since the number of replicates associated with individual cell lines varies, in total as many as 37 profiles are included. Although the number of sgRNAs that target individual genes also varies, we consider as many as 18,467 genes associated with at least ten sgRNAs. When more than ten sgRNAs target one genes, randomly selected ten sgRNAs are considered. As a result, we have a tensor $x_{ijk} \in \mathbb{R}^{18647\times10\times37}$ that represent the number of jth sgRNAs that target ith gene of kth profile. Remarkably, all cell lines are associated with pairs of replicates composed of initials and finals where initials can be regarded as controls (not treated).

3.1.4 Dataset 4: GeCKOv2

The file extracted from Data.zip and used for the analysis is Achilles_raw_GeckoV2.tab. GeCKOv2 [14, 15] is composed of thirty six cell lines including control. Since the number of replicates associated with individual cell lines varies, in total as many as 129 profiles are included. We consider as many as 29 cell lines associated with at least four replicates (after this screening, controls were dropped). Although the number of sgRNAs that target individual genes also varies, we consider as many as 18,862 genes associated with at least four sgRNAs. When more than four sgRNAs target one genes, randomly selected four sgRNAs are considered. As a result, we have a tensor $x_{ijkm} \in \mathbb{R}^{18862\times4\times29\times4}$ that represent the number of jth sgRNAs that target ith gene of mth biological replcate of kth cell line.

3.1.5 Dataset 5: Avana

The file extracted from Data.zip and used for the analysis is Avana_sgrna_raw_readcounts_matched.csv, the corresponding gene names of which are from Avana_sgrnamapping.csv. Avana [16] is composed of 335 cell lines without control. Since the number of replicates associated with individual cell lines varies, in total as many as 808 profiles are included. We consider as many as 317 cell lines associated with at least two replicates. Although the number of sgRNAs that target individual genes also varies, we consider as many as 17,390 genes associated with at least four sgRNAs. When more than four sgRNAs target one genes, randomly selected four sgRNAs are considered. As a result, we have a tensor $x_{ijkm} \in \mathbb{R}^{17390\times317\times2\times4}$ that represents the number of mth sgRNAs that target ith gene of kth biological replcate of jth cell line.

3.2 TD

TD was performed by HOSVD [8]. Tensors, $x_{ijk} \in \mathbb{R}^{I\times J\times K}$ or $x_{ijkm} \in \mathbb{R}^{I\times J\times K\times M}$, is decomposed as

$$x_{ijk} = \sum_{\ell_1=1}^{I}\sum_{\ell_2=1}^{J}\sum_{\ell_3=1}^{K} G(\ell_1\ell_2\ell_3)u_{\ell_1 i}u_{\ell_2 j}u_{\ell_3 k} \tag{1}$$

or

$$x_{ijkm} = \sum_{\ell_1=1}^{I} \sum_{\ell_2=1}^{J} \sum_{\ell_3=1}^{K} \sum_{\ell_4=1}^{M} G(\ell_1\ell_2\ell_3\ell_4) u_{\ell_1 i} u_{\ell_2 j} u_{\ell_3 k} u_{\ell_4 m}. \tag{2}$$

where $G(\ell_1\ell_2\ell_3) \in \mathbb{R}^{I\times J\times K}$ or $G(\ell_1\ell_2\ell_3\ell_4) \in \mathbb{R}^{I\times J\times K\times M}$ is a core tensor that represents the contribution of $u_{\ell_1 i} u_{\ell_2 j} u_{\ell_3 k}$ or $u_{\ell_1 i} u_{\ell_2 j} u_{\ell_3 k} u_{\ell_4 m}$ to x_{ijk} or x_{ijkm}, respective ly.

One should note that tensors are standardized as follows before being decomposed

$$\sum_{i=1}^{I} x_{ijk} = \sum_{i=1}^{I} x_{ijkm} = 0 \tag{3}$$

$$\sum_{i=1}^{I} x_{ijk}^2 = \sum_{i=1}^{I} x_{ijkm}^2 = I. \tag{4}$$

In addition to this, we also consider the alternative version that is logarithmically transformed before the standardization, eqs. (3) and (4),

$$x_{ijk} \leftarrow \log_{10}(x_{ijk} + c) \tag{5}$$

$$x_{ijkm} \leftarrow \log_{10}(x_{ijkm} + c) \tag{6}$$

where $c = 1$ or $c = 2$, smaller one of which was employed in invididual datasets to satisfy $x_{ijk}+c > 0$ or $x_{ijkm}+c > 0$. The reason why we used also logarithmically transformed value is because various conventional methods employed the logarithmically transformed values.

3.3 Retrieval of the lists of essential and non-essential genes

Genes used to evaluate performances of JACKS and TD are essential and non-essential genes [17]; there is the statement that the numbers of essential and non-essential genes are 360 and 927, respectively. The 927 non-essential genes are retrieved from the fifth column of dataset S1 file. Since there is no explicit list of the 360 essential genes, we used the 823 genes assoicated with "Num Obs" larger than or equal to 3, in the "essential-genes-top12screens" sheet in data set S4. Here, "Num Obs" is the number of cell lines (of the top 12) in which the gene is classified as essential.

3.4 Selection of singular value vectors used

At first, we try to find which singular value vectors attributed to genes can discriminate between essential genes and non-essential genes best. After that, we investigated G to find which singular value vectors attributed to cell lines are mostly related to the selected singular value vectors attributed to cell lines. The selected singular value vectors attributed to cell lines enable us to interpret the meaning of the selected singnular value vectors attributed to genes that can discriminate between essential genes and non-essential genes best.

3.5 Comparison with JACKS

The JACKS results were retrieved from JACKS.zip also available at https://www.doi.org/10.6084/m9.figshare.6002438. Files used are those named as "cell line names"_gene_pval_JACKS_results.txt ("cell line names" is "avana", "gecko2", "whitehead", "tko", or "yusa_v10"). Before being compared, P-values in datasets are logarithmic-transformed to recover raw values. The reason why we did not use raw values (i.e. files named as "cell line names"_gene_JACKS_results.txt) is because logarithmic-transformed P-values are more robust (i.e. free from the standardization etc). Since JACKS does not provide cell line wide AUC, we compute mean score along cell lines in each of five datasets and use mean value to compute AUC.

3.6 CRISPR efficacy of cell lines

The file CRISPRInferredModelEfficacy.csv, which includes CRISPR efficiency of cell lines, is retrieved from DepMap [18]. "Achilles-Avana-2D" column of it is used to evaluate the biological reliability of singular value vectors attributed to cell lines (for GeCKOvs and Avana). If the correlation between CRISPR efficiency of cell lines and singular value vectors attributed to cell lines is large enough, we can judge that the singular value vectors attributed to cell lines is biologically reliable.

4 Discussion

First of all, we would like to emphasize that TD is the ever first method that can integrate multiple sgRNA profiles from the beginning. In other SOTA introduced in Introduction, i.e., MAGeCK, BAGEL, CERES, STARS, and JACKS, individual sgRNA profiles are once analyzed one by one and their outcomes are integrated. Nevertheless, TD can integrate multiple sgRNA profiles from the beginning. No analysis processes are separately applied to individual profiles. This outstanding ability that can integrate multiple profiles from the beginning possibly possibly lets TD to achieve the comparative performance with JACKS, in spite of that TD is nothing but a simple linear algebra.

Second, it is remarkable that the consequence of TD which is nothing but a simple linear algebra can compete with one of SOTA, JACKS, that is thought to be highly advanced method. Especially, for the Avana data set that have the largest number of cell lines, the performance of TD is quite similar to JACKS even if we consider not AUC value but ROC curve itself (Fig. 2). Moreover, for Avana, raw and logarithmic values also achieve quite similar performances (Fig. 2). This also questions the reliability to take logarithmic values of the number of sRNA. If raw (not logarithmic) values can perform as much as logarithmic values when they try to discriminate between essential and non-essential genes, what justifies the usage of logarithmic values? It might be better to reconsider the definition of quantitative CRISPR efficiency.

TD also has advantages to compensate the distinct efficiency of individual sgRNAs that target the same genes. One should notice that the first singular value vectors attributed to sgRNA is always associated with the largest absolute values of G (Table 3). For the Yusa data sets, $u_{\ell_2 j}$ is attributed to sgRNA. In Fig. 3, $G(7, 1, 4)$ and

$G(6,1,2)$ with $\ell_2 = 1$ have the largest absolute values. For the TKOv1 dataset, $u_{\ell_2 j}$ is also attributed to sgRNA. In Fig. 4, $G(1,1,1)$ and $G(5,1,2)$ with $\ell_2 = 1$ have the largest absolute values. For the Whitehead data set, $u_{\ell_2 j}$ is again attributed to sgRNA. In Fig. 6, $G(11,1,2)$, $G(2,1,2)$, and $G(12,1,3)$ with $\ell_2 = 1$ have the largest absolute values. For the GeCKOv2 data sets, $u_{\ell_2 j}$ is attributed to sgRNA. In Fig. 8, $G(5,1,2,1)$ with $\ell_2 = 1$ has the largest absolute values. For Avana dataset, $u_{\ell_4 m}$ is attributed to sgRNA. In Fig. 9, $G(5,3,1,1)$ and $G(5,2,1,1)$ with $\ell_4 = 1$ has the largest absolute values. Since the first singular value vectors usually corresponds to the simple average, the fact that the first singular value vectors attributed to sgRNA is always employed means that TD is free from the sophisticated consideration of sgRNA efficiency. In other word, TD is the robust method that does not have to consider difference of CRISPR efficiency between multiple sgRNAs. This might be the reason why TD can compete with JACKS that had to consider the difference of sgRNA efficiency.

This also reasons why larger ℓ_1 is often attributed to singular value vectors attributed to genes, $u_{\ell_1 i}$, with large AUC. Since there are J (for datasets other than Avana) and M (for Avana dataset) singular value vectors attributed to sgRNAs and J or M singular value vectors attributed to sgRNAs must represent all dependence upon j (for datasets other than Avana) or m (for Avana dataset), if the first singular value vectors attributed to sgRNAs represent uniform one (i.e., u_{1j} or u_{1m} is independent of j or m), the next singular value vectors attributed to sgRNAs that represent uniform one must be $j > J$ or $m > M$. This results in $u_{\ell_1 i}$ with larger ℓ_1 is associated with the next singular value vectors attributed to sgRNAs that represent uniform one, because large AUC is associated with singular value vectors attributed to sgRNAs and independent of j or m.

TD also has disadvantages that genes and cell lines without large enough sgRNAs and biological replicates are discarded. Although we have also some method to deal with the cases where the different replicates are associated with [19], we did not employ it for the simplicity. In the case we need more sophisticated treatment, we might be forced to consider such cases.

5 Conclusion

In this paper, we applied TD to five datasets composed of sgRNA profiles. TD could integrate sgRNAs and profiles simultaneously and could discriminate between essential genes and non-essential genes as well as JACKS, one pf SOTA. Since TD is nothing but the consequence of simple linear algebra, it is remarkable for TD to compete with JACKS.

Supplementary information.

Acknowledgements. Acknowledgements are not compulsory. Where included they should be brief. Grant or contribution numbers may be acknowledged.

Please refer to Journal-level guidance for any specific requirements.

Declarations

Some journals require declarations to be submitted in a standardised format. Please check the Instructions for Authors of the journal to which you are submitting to see if you need to complete this section. If yes, your manuscript must contain the following sections under the heading 'Declarations':

Ethics approval and consent to participate

Not applicable

Competing interests

The authors have no conflicts of interest to declare.

Funding

This study was supported by KAKENHI Grant Number 24K15168.

Data availability

We have gotten all the data analysed in this study from https://www.doi.org/10.6084/m9.figshare.6002438.

Code availability

The sample source code is available at https://github.com/tagtag/TDbasedUFE_sgRNA

Author contribution

Y. H. T. planned the study and performed the analyses. Y.-H.T. and T.T. evaluated the results and wrote and reviewed the manuscript. All the authors have read and agreed to the published version of the manuscript. The conceptualization, data curation, analysis were performed by Y. H. T.

References

[1] Shalem, O., Sanjana, N.E., Zhang, F.: High-throughput functional genomics using CRISPR–Cas9. Nature Reviews Genetics **16**(5), 299–311 (2015) https://doi.org/10.1038/nrg3899

[2] Li, W., Xu, H., Sizemore, T., Love, M.I., Zhang, F., Meyer, C.A., Liu, J.S., Brown, M., Liu, X.S.: MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. Genome Biology **15**(12), 554 (2014) https://doi.org/10.1186/s13059-014-0554-4

[3] Li, W., Koster, J., Xu, H., Chen, C.-H., Sizemore, T., Liu, J.S., Liu, X.S.: Quality control, modeling, and visualization of CRISPR screens with MAGeCK-VISPR. Genome Biology **16**(1), 271 (2015) https://doi.org/10.1186/s13059-015-0843-6

[4] Hart, T., Moffat, J.: BAGEL: a computational framework for identifying essential genes from pooled library screens. BMC Bioinformatics **17**(1), 164 (2016) https://doi.org/10.1186/s12859-016-1015-8

[5] Meyers, R.M., Bryan, J.G., McFarland, J.M., Weir, B.A., Sizemore, A.E., Xu, H., Dharia, N.V., Montgomery, P.G., Cowley, G.S., Pantel, R., Brügmann, A.L.M., Kim, E., Vilfort, Z., Tsherniak, A., Vazquez, F., Adstamongkonkul, P., Mermel, C.H., Doench, J.G., Root, D.E., Ching, T.H.N., Boehm, J.S., Hahn, W.C.: Computational correction of copy number effect in CRISPR–Cas9 screens. Nature Genetics **49**(12), 1779–1784 (2017) https://doi.org/10.1038/ng.3980

[6] Doench, J.G., Hartenian, E., Brown, K.R., Root, D.E., Lander, E.S.: Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. Nature Biotechnology **32**(12), 1262–1267 (2014) https://doi.org/10.1038/nbt.3026

[7] Allen, F., Behan, F., Khodak, A., Iorio, F., Yusa, K., Garnett, M.J.: JACKS: joint analysis of CRISPR/Cas9 knockout screens. Genome Research **29**(3), 464–471 (2019) https://doi.org/10.1101/gr.238710.118

[8] Taguchi, Y.-h.: Unsupervised Feature Extraction Applied to Bioinformatics: A PCA Based and TD Based Approach, 2nd edn. Unsupervised and Semi-Supervised Learning. Springer, Switzland (2024)

[9] Tzelepis, K., Koike-Yusa, H., De Braekeleer, E., Li, Y., Metzakopian, E., Dovey, O.M., Mupo, A., Grinkevich, V., Li, M., Mazan, M., Gozdecka, M., Ohnishi, S., Cooper, J., Patel, M., McKerrell, T., Chen, B., Domingues, A.F., Gallipoli, P., Teichmann, S., Ponstingl, H., McDermott, U., Saez-Rodriguez, J., Huntly, B.J.P., Iorio, F., Pina, C., Vassiliou, G.S., Yusa, K.: A crispr dropout screen identifies genetic vulnerabilities and therapeutic targets in acute myeloid leukemia. Cell Reports **17**(4), 1193–1205 (2016) https://doi.org/10.1016/j.celrep.2016.09.079 . doi: 10.1016/j.celrep.2016.09.079

[10] Iorio, F., Behan, F.M., Gonçalves, E., Bhosle, S.G., Chen, E., Shepherd, R., Beaver, C., Ansari, R., Pooley, R., Wilkinson, P., Harper, S., Butler, A.P., Stronach, E.A., Saez-Rodriguez, J., Yusa, K., Garnett, M.J.: Unsupervised correction of gene-independent cell responses to CRISPR-Cas9 targeting. BMC Genomics **19**(1), 604 (2018) https://doi.org/10.1186/s12864-018-4989-y

[11] Hart, T., Chandrashekhar, M., Aregger, M., Steinhart, Z., Brown, K.R., MacLeod, G., Mis, M., Zimmermann, M., Fradet-Turcotte, A., Sun, S., Mero, P., Dirks, P., Sidhu, S., Roth, F.P., Rissland, O.S., Durocher, D., Angers, S., Moffat, J.: High-resolution crispr screens reveal fitness genes and genotype-specific cancer liabilities. Cell **163**(6), 1515–1526 (2015) https://doi.org/10.1016/j.cell.2015.11.015 . doi: 10.1016/j.cell.2015.11.015

[12] Wang, T., Birsoy, K., Hughes, N.W., Krupczak, K.M., Post, Y., Wei, J.J., Lander,

E.S., Sabatini, D.M.: Identification and characterization of essential genes in the human genome. Science **350**(6264), 1096–1101 (2015) https://doi.org/10.1126/science.aac7041 https://www.science.org/doi/pdf/10.1126/science.aac7041

[13] Wang, T., Yu, H., Hughes, N.W., Liu, B., Kendirli, A., Klein, K., Chen, W.W., Lander, E.S., Sabatini, D.M.: Gene essentiality profiling reveals gene networks and synthetic lethal interactions with oncogenic ras. Cell **168**(5), 890–90315 (2017) https://doi.org/10.1016/j.cell.2017.01.013 . doi: 10.1016/j.cell.2017.01.013

[14] Sanjana, N.E., Shalem, O., Zhang, F.: Improved vectors and genome-wide libraries for crispr screening. Nature Methods **11**(8), 783–784 (2014) https://doi.org/10.1038/nmeth.3047

[15] Aguirre, A.J., Meyers, R.M., Weir, B.A., Vazquez, F., Zhang, C.-Z., Ben-David, U., Cook, A., Ha, G., Harrington, W.F., Doshi, M.B., Kost-Alimova, M., Gill, S., Xu, H., Ali, L.D., Jiang, G., Pantel, S., Lee, Y., Goodale, A., Cherniack, A.D., Oh, C., Kryukov, G., Cowley, G.S., Garraway, L.A., Stegmaier, K., Roberts, C.W., Golub, T.R., Meyerson, M., Root, D.E., Tsherniak, A., Hahn, W.C.: Genomic copy number dictates a gene-independent cell response to crispr/cas9 targeting. Cancer Discovery **6**(8), 914–929 (2016) https://doi.org/10.1158/2159-8290.CD-16-0154 . This article is highlighted in the In This Issue feature, p. 803

[16] Meyers, R.M., Bryan, J.G., McFarland, J.M., Weir, B.A., Sizemore, A.E., Xu, H., Dharia, N.V., Montgomery, P.G., Cowley, G.S., Pantel, S., Goodale, A., Lee, Y., Ali, L.D., Jiang, G., Lubonja, R., Harrington, W.F., Strickland, M., Wu, T., Hawes, D.C., Zhivich, V.A., Wyatt, M.R., Kalani, Z., Chang, J.J., Okamoto, M., Stegmaier, K., Golub, T.R., Boehm, J.S., Vazquez, F., Root, D.E., Hahn, W.C., Tsherniak, A.: Computational correction of copy number effect improves specificity of crispr–cas9 essentiality screens in cancer cells. Nature Genetics **49**(12), 1779–1784 (2017) https://doi.org/10.1038/ng.3984

[17] Hart, T., Brown, K.R., Sircoulomb, F., Rottapel, R., Moffat, J.: Measuring error rates in genomic perturbation screens: gold standards for human functional genomics. Molecular Systems Biology **10**(7), 733 (2014) https://doi.org/10.15252/msb.20145216 https://www.embopress.org/doi/pdf/10.15252/msb.20145216

[18] Tsherniak, A., al.: Defining a cancer dependency map. Cell **170**(3), 564–576 (2017) https://doi.org/10.1016/j.cell.2017.06.010

[19] Taguchi, Y.-h., Turki, T.: A tensor decomposition-based integrated analysis applicable to multiple gene expression profiles without sample matching. Scientific Reports **12**(1), 21242 (2022) https://doi.org/10.1038/s41598-022-25524-4