VISAGE User's Guide

(Volumetric Interrogation of Synergy and Gene set Enrichment)

Introduction

The aim of this guide is to walk a non-programmer through the use of the VISAGE R script, hosted at (https://github.com/yaffelab/visage), in order to use cell line viability data collected on a panel of cell lines treated with multiple doses of a pair of drugs to calculate the synergy in each cell line, and identify genes and gene sets that are correlated with synergy.

If you use VISAGE, please cite the following publication:

VISAGE Reveals a Targetable Mitotic Spindle Vulnerability in Cancer Cells.
Patterson JC, Joughin BA, Prota AE, Mühlethaler T, Jonas OH, Whitman MA, Varmeh S, Chen S, Balk SP, Steinmetz MO, Lauffenburger DA, Yaffe MB.
Cell Syst. 2019
PMID: XXXXXXXX

Please contact visage_help@mit.edu with questions, comments, or if you find bugs, and we will help you.

Installing R and RStudio

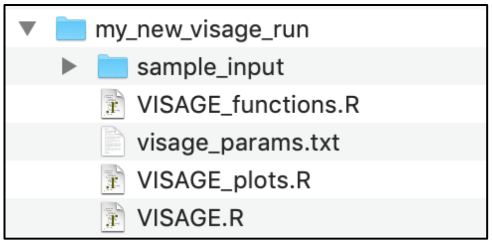
To use the R version of VISAGE in this tutorial, both R and RStudio must be installed on the user's machine. R and RStudio are free software available as described below.

1. Go to (https://www.r-project.org/) and follow the instructions there under "**Getting Started"** to download the most recent version of R for your operating system.

2. Go to (https://www.rstudio.com/) and click the **"Download RStudio"** button to download **RStudio Desktop** for your operating system.

RStudio can now be run as an interface to R, and is the environment that will be assumed for the rest of this guide. R may be used in other environments than RStudio, but this guide uses RStudio as the example.

Setting Up a New VISAGE Run



```
my_new_visage_run
    sample_input
    VISAGE_functions.R
    visage_params.txt
    VISAGE_plots.R
    VISAGE.R
```

Contents of a properly setup new VISAGE directory.

Make a directory in which you would like to run VISAGE. It can have any name, but we use "`my_new_visage_run`" as the example in this guide.

This directory needs to contain a number of specific files and a user input subdirectory, each of which will be discussed below:

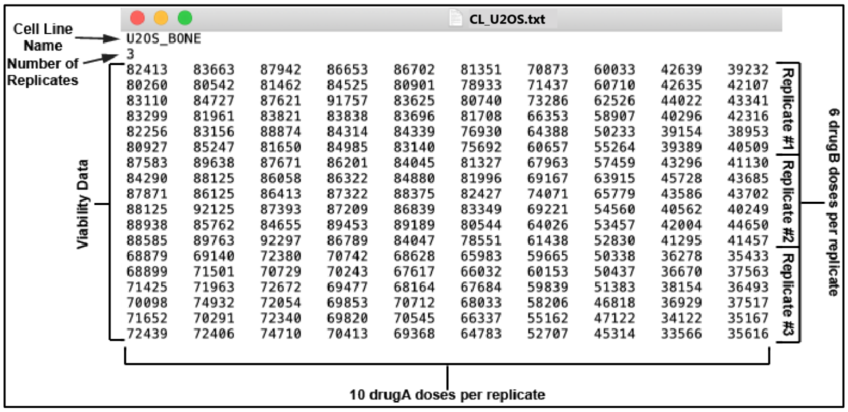
- Subdirectory containing user input:

All user-provided input data, including cell line viability data and gene expression data, are placed in a single subdirectory of your VISAGE run directory. Here we have called it "`sample_input`", and there is a "`sample_input`" directory available at GitHub containing the cell line viability data presented in the VISAGE manuscript. You may call it something else if you want, and instructions on telling the VISAGE script how to find this directory are included below in the section on the file "`visage_params.txt`".

Make sure the following files are present in your input subdirectory:

- **Cell line viability data**: For each cell line being analyzed, the input subdirectory must include a text file containing viability data. This file can have any name of the form `CL_`{identifier}`.txt`, where {identifier} is any unique name. We recommend this name be a short form of the name of the cell line (e.g. `CL_HeLa.txt`). The "`sample_input`" subdirectory of the VISAGE download from GitHub includes the viability data from the VISAGE manuscript as an example. The contents of these text files require very precise formatting as shown:

    - **Cell line name**: Line 1 must include the cell line name exactly as it is present in gene expression data and will be used in output files. It is not necessary that every cell line with viability data be present in gene expression data, and warnings will be printed by VISAGE listing which cell lines are not included in gene expression data.
    - **Number of replicates**: Line 2 must be the number of data replicates present in the rest of the file.
    - **Viability data:** From line 3 on, viability data is listed. Each column (separated by exactly one tab) represents a dose of `drugA`, and each row represents a dose of `drugB`. Each replicate is presented in a block immediately following the previous, as shown at right. Information on drug names and doses is given to VISAGE in the file "`visage_params.txt`", described below.

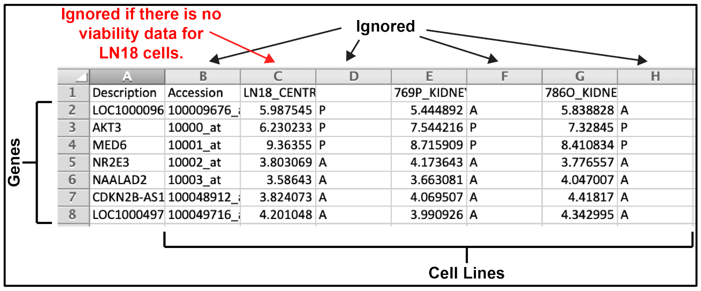


```
U2OS_BONE
3
82413   83663   87942   86653   86702   81351   70873   60033   42639   39232
80260   80542   81462   84525   80901   78933   71437   60710   42635   42107
83110   84727   87621   91757   83625   80740   73286   62526   44022   43341
83299   81961   83821   83838   83696   81708   66353   58907   40296   42316
82256   83156   88874   84314   84339   76930   64388   50233   39154   38953
80927   85247   81650   84985   83140   75692   60657   55264   39389   40509
87583   89638   87671   86201   84045   81327   67963   57459   43296   41130
84290   88125   86058   86322   84880   81996   69167   63915   45728   43685
87871   86125   86413   87322   88375   82427   74071   65779   43586   43702
88125   92125   87393   87209   86839   83349   69221   54560   40562   40249
88938   85762   84655   89453   89189   80544   64026   53457   42004   44650
88585   89763   92297   86789   84047   78551   61438   52830   41295   41457
68879   69140   72380   70742   68628   65983   59665   50338   36278   35433
68899   71501   70729   70243   67617   66032   60153   50437   36670   37563
71425   71963   72672   69477   68164   67684   59839   51383   38154   36493
70098   74932   72054   69853   70712   68033   58206   46818   36929   37517
71652   70291   72340   69820   70545   66337   55162   47122   34122   35167
72439   72406   74710   70413   69368   64783   52707   45314   33566   35616
```

Contents of a properly formatted cell line viability file.

Setting Up a New VISAGE Run (cntd.)

- **Gene expression data**: The input subdirectory must also contain a text file containing gene expression data on the cell lines of interest. Later, we will describe how to pass the name of this file to VISAGE in the "`visage_params.txt`" file.



	A	B	C	D	E	F	G	H
1	Description	Accession	LN18_CENTR		769P_KIDNE		786O_KIDNE	
2	LOC1000096	100009676_	5.987545	P	5.444892	A	5.838828	A
3	AKT3	10000_at	6.230233	P	7.544216	P	7.32845	P
4	MED6	10001_at	9.36355	P	8.715909	P	8.410834	P
5	NR2E3	10002_at	3.803069	A	4.173643	A	3.776557	A
6	NAALAD2	10003_at	3.58643	A	3.663081	A	4.047007	A
7	CDKN2B-AS1	100048912_	3.824073	A	4.069507	A	4.41817	A
8	LOC1000497	100049716_	4.201048	A	3.990926	A	4.342995	A

Partial contents of a gene expression data file.

This data must have a row for each gene with the gene name in the first column, and a column for each cell line with the cell line name in the first row.

- Rows with blank gene names are removed.
- Rows with duplicate gene names are kept and are later handled by GSEA.
- Extra columns not corresponding to a cell line with viability data are ignored. (These column titles might be cell lines not studied, auxiliary metadata regarding the gene (e.g., "`Accession`" in the image above), or blank.)

Note: If you open gene expression data directly with Microsoft Excel, it will incorrectly but automatically rename some genes (e.g., MARCH1, SEPT1) as if they are dates. If you must open a gene expression file in Excel, make sure to open it from the File menu in Excel and manually load the column containing gene names as "Text" instead of "General".

In the VISAGE manuscript, we used the shown file containing RMA-normalized, gene-centric microarray data from the Broad Institute CCLE (https://portals.broadinstitute.org/ccle). We cannot distribute this file with VISAGE, but it is available at (https://data.broadinstitute.org/ccle_legacy_data/mRNA_expression/CCLE_Expression_Entrez_2012-10-18.res) along with other expression data (free registration with the Broad Institute required). Note that this data is $\log_2$-transformed as part of RMA normalization. If you are using RNAseq data instead of microarray data, it is probably wise to use a log transformation or other variance-stabilizing transformation.

- `VISAGE.R, VISAGE_functions.R, VISAGE_plots.R`

These files are provided as part of the VISAGE download from GitHub and the user should not need to make any changes to these files. `VISAGE.R` is the script the user will run in RStudio. `VISAGE_functions.R` and `VISAGE_plots.R` contain some functions that VISAGE.R uses.

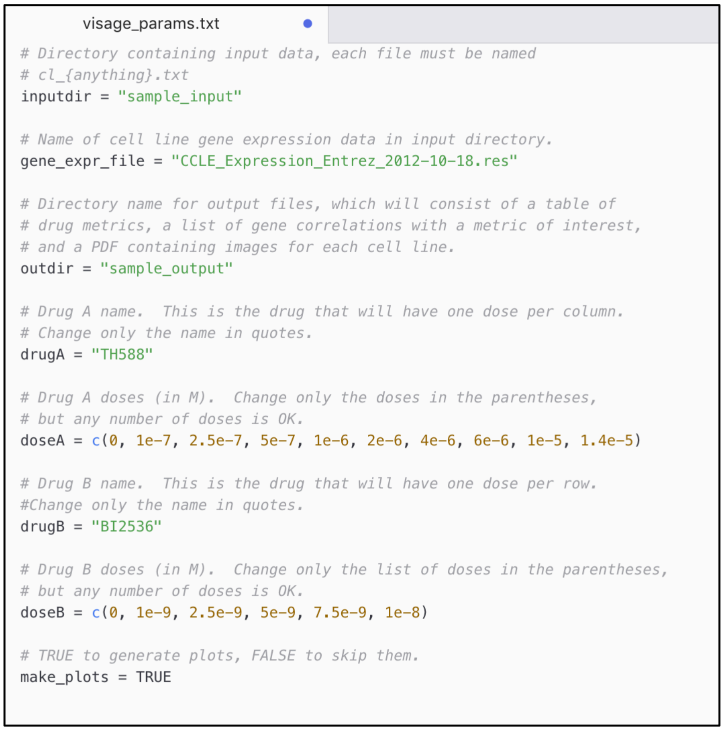
Setting Up a New VISAGE Run (cntd.)

- `visage_params.txt`

This file contains information that VISAGE requires to customize the run to your data and needs. Its contents are R code that will be run by VISAGE.R to set some parameter values. Your download of VISAGE includes this file, along with sample contents that allow you to analyze the cell lines and drugs discussed in the original VISAGE publication once you download gene expression data from the CCLE.

All lines beginning with the character "#" are comments and are not read by VISAGE.

Adjust the parameters of the `visage_params.txt` file as follows:



```
# Directory containing input data, each file must be named
# cl_{anything}.txt
inputdir = "sample_input"

# Name of cell line gene expression data in input directory.
gene_expr_file = "CCLE_Expression_Entrez_2012-10-18.res"

# Directory name for output files, which will consist of a table of
# drug metrics, a list of gene correlations with a metric of interest,
# and a PDF containing images for each cell line.
outdir = "sample_output"

# Drug A name.  This is the drug that will have one dose per column.
# Change only the name in quotes.
drugA = "TH588"

# Drug A doses (in M).  Change only the doses in the parentheses,
# but any number of doses is OK.
doseA = c(0, 1e-7, 2.5e-7, 5e-7, 1e-6, 2e-6, 4e-6, 6e-6, 1e-5, 1.4e-5)

# Drug B name.  This is the drug that will have one dose per row.
#Change only the name in quotes.
drugB = "BI2536"

# Drug B doses (in M).  Change only the list of doses in the parentheses,
# but any number of doses is OK.
doseB = c(0, 1e-9, 2.5e-9, 5e-9, 7.5e-9, 1e-8)

# TRUE to generate plots, FALSE to skip them.
make_plots = TRUE
```

Contents of a properly formatted `visage_params.txt` file.

- o **`inputdir`**: If the subdirectory that contains the cell line viability data and gene expression data you created has a name other than `sample_input`, change the name in quotation marks here, either as an absolute path or a path relative to the VISAGE run directory.

- o **`gene_expr_file`**: Put the name of the file, located in `inputdir`, containing gene expression data in quotation marks as shown.

- o **`outputdir`**: Put the name of the directory to place output files, either as an absolute path or a path relative to the VISAGE run directory, in quotation marks as shown. This directory will be created by VISAGE if necessary.

- o **`drugA, drugB`**: Put the name of the drugs being studied in quotation marks as shown. `drugA` must be the drug that has one dose per column in dose-response viability matrices for each cell line. `drugB` is the drug that has one dose per row.

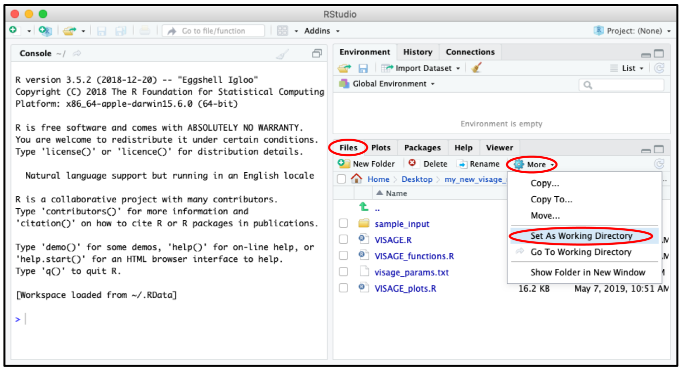
- o **`doseA, doseB`**: List the doses used for `drugA` and `drugB`, respectively, in units of molar. These must be listed in the same order in which the doses are presented in cell line viability data (see above). The first dose of each drug must be zero. Doses should be listed as shown above, as comma separated lists inside the R vector constructor `c()`.

- o **`make_plots`**: If `make_plots` is set to `TRUE` as shown in the image above, VISAGE will make a PDF file for each cell line containing plots (detailed below) regarding drug sensitivity and synergy in that cell line. If it is set to `FALSE` these plots will not be generated.

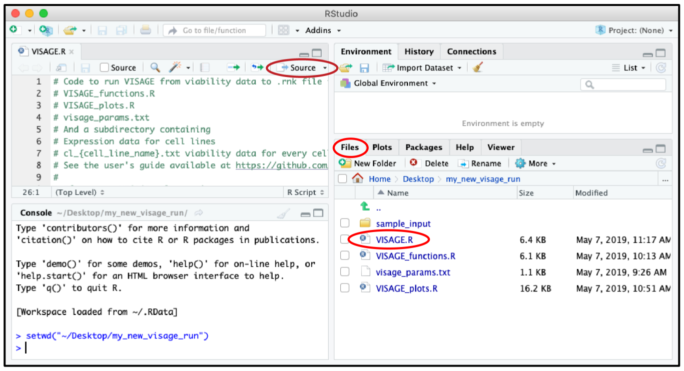
You now should be ready to run VISAGE on your data.

Running VISAGE

1. Open RStudio. Using the "Files" pane on the lower right of the screen, navigate to the directory where you have set up your VISAGE run. From the "More" menu of the file window, select "Set As Working Directory".



```
R version 3.5.2 (2018-12-20) -- "Eggshell Igloo"
Copyright (C) 2018 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin15.6.0 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Workspace loaded from ~/.RData]

> |
```

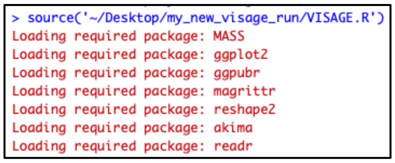


```
# Code to run VISAGE from viability data to .rnk file
# VISAGE_functions.R
# VISAGE_plots.R
# visage_params.txt
# And a subdirectory containing
# Expression data for cell lines
# cl_{cell_line_name}.txt viability data for every cel
# See the user's guide available at https://github.com
#
```

```
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Workspace loaded from ~/.RData]

> setwd("~/Desktop/my_new_visage_run")
> |
```

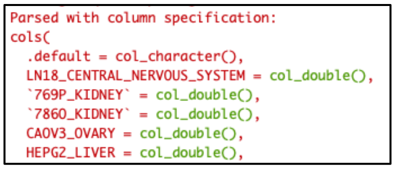
2. In the file pane, click to open VISAGE.R. To run the VISAGE script, click the "Source" button, circled in red at left.

3. You can watch the progress of VISAGE in the console pane at lower left:



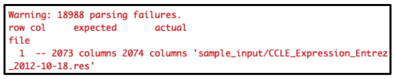
```
> source('~/Desktop/my_new_visage_run/VISAGE.R')
Loading required package: MASS
Loading required package: ggplot2
Loading required package: ggpubr
Loading required package: magrittr
Loading required package: reshape2
Loading required package: akima
Loading required package: readr
```

The first time you run VISAGE, all the necessary R packages will be downloaded and installed if you don't have them. For this reason, an internet connection is necessary the first time you run VISAGE.



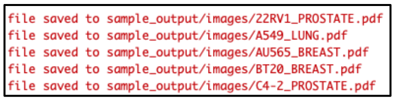
```
Parsed with column specification:
cols(
  .default = col_character(),
  LN18_CENTRAL_NERVOUS_SYSTEM = col_double(),
  `769P_KIDNEY` = col_double(),
  `786O_KIDNEY` = col_double(),
  CAOV3_OVARY = col_double(),
  HEPG2_LIVER = col_double(),
```

Here, VISAGE is reading in gene expression data. Remember that you must include this file manually; it is not provided with VISAGE.



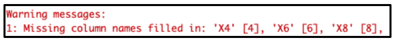
```
Warning: 18988 parsing failures.
row col     expected        actual
file
  1  -- 2073 columns 2074 columns 'sample_input/CCLE_Expression_Entrez
_2012-10-18.res'
```

Because the last column of this file has no heading, VISAGE cannot read it in, but since it is not a cell line's expression data, this is no problem.



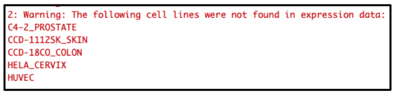
```
file saved to sample_output/images/22RV1_PROSTATE.pdf
file saved to sample_output/images/A549_LUNG.pdf
file saved to sample_output/images/AU565_BREAST.pdf
file saved to sample_output/images/BT20_BREAST.pdf
file saved to sample_output/images/C4-2_PROSTATE.pdf
```

If the `make_plots` flag is set to `TRUE`, you will see a message each time a PDF is generated. If it is set to `FALSE`, these messages will not appear.



```
Warning messages:
1: Missing column names filled in: 'X4' [4], 'X6' [6], 'X8' [8],
```

Since every 2nd column of the gene expression file we used had no heading, VISAGE is warning us that it gave them arbitrary names.

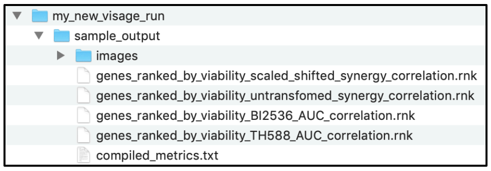


```
2: Warning: The following cell lines were not found in expression data:
C4-2_PROSTATE
CCD-1112SK_SKIN
CCD-18CO_COLON
HELA_CERVIX
HUVEC
```

Finally, we are given a warning about each cell line for which we provided viability data that there was no expression data for. While we get plots about sensitivity and specificity of these lines, they do not contribute to identifying genes or gene sets associated with sensitivity or synergy.

VISAGE Output Files

If run properly, your VISAGE run directory will now contain an output subdirectory (with a name that you provided in `visage_params.txt`) containing files, and, if the `make_plots` flag was set as `TRUE`, a directory of PDFs with sensitivity and synergy plots for each cell line has been created. Each of these is discussed below:

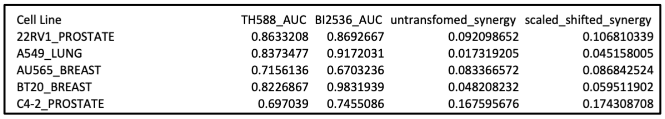


```
▼ my_new_visage_run
    ▼ sample_output
        ▶ images
          genes_ranked_by_viability_scaled_shifted_synergy_correlation.rnk
          genes_ranked_by_viability_untransfomed_synergy_correlation.rnk
          genes_ranked_by_viability_BI2536_AUC_correlation.rnk
          genes_ranked_by_viability_TH588_AUC_correlation.rnk
          compiled_metrics.txt
```

Contents of a VISAGE output directory after a successful run.

- `compiled_metrics.txt`

This file contains drug sensitivity and synergy data calculated from your input viability data:



Cell Line	TH588_AUC	BI2536_AUC	untransfomed_synergy	scaled_shifted_synergy
22RV1_PROSTATE	0.8633208	0.8692667	0.092098652	0.106810339
A549_LUNG	0.8373477	0.9172031	0.017319205	0.045158005
AU565_BREAST	0.7156136	0.6703236	0.083366572	0.086842524
BT20_BREAST	0.8226867	0.9831939	0.048208232	0.059511902
C4-2_PROSTATE	0.697039	0.7455086	0.167595676	0.174308708

First few lines of `compiled_metrics.txt`.

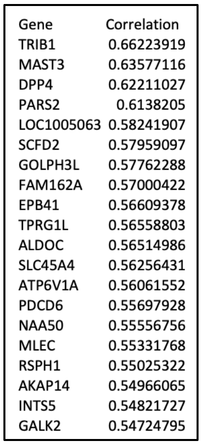
- o `Cell Line`: The names of cell lines come from the top line of each viability input file

- o `{drugA}_AUC`, `{drugB}_AUC`: The fraction of the total possible area under the dose-viability curve for drug A or B alone. A higher value corresponds to lower sensitivity.

- o `untransformed_synergy`: The fraction of the total possible synergy observed in the cell line, calculated as the fractional volume under the expected viability surface minus the fractional volume under the observed viability surface. A higher value corresponds to more synergy

- o `scaled_shifted_synergy`: As `untransformed_synergy`, but before calculation viability data is linearly transformed such that the undrugged control is at 1 and the viability at the dose combination which has minimal viability over the dose-viability matrix is at 0. This is the metric used in the VISAGE manuscript. See the VISAGE manuscript for discussion. A higher value corresponds to more synergy

- `genes_ranked_by_viability_{metric}_correlation.rnk`



Gene	Correlation
TRIB1	0.66223919
MAST3	0.63577116
DPP4	0.62211027
PARS2	0.6138205
LOC1005063	0.58241907
SCFD2	0.57959097
GOLPH3L	0.57762288
FAM162A	0.57000422
EPB41	0.56609378
TPRG1L	0.56558803
ALDOC	0.56514986
SLC45A4	0.56256431
ATP6V1A	0.56061552
PDCD6	0.55697928
NAA50	0.55556756
MLEC	0.55331768
RSPH1	0.55025322
AKAP14	0.54966065
INTS5	0.54821727
GALK2	0.54724795

First lines of a ranked genes file.

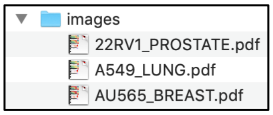
For each of the metrics in compiled_metrics.txt, a file is generated listing the correlation of that metric with gene expression data, across the cell lines for which viability and expression data are both provided. Genes are sorted by correlation from most correlated to most anti-correlated. These .rnk files are suitable for use as inputs to GSEAPreranked, part of GSEA, available at http://software.broadinstitute.org/gsea/.

In the VISAGE manuscript, we ran GSEAPreranked with default parameters on genes ranked by correlation with the scaled & shifted synergy metric, scoring hallmark, KEGG, BioCarta, Reactome, and GO gene sets.

Note: If you open ranked gene files directly with Microsoft Excel, it will incorrectly but automatically rename some genes (e.g., MARCH1, SEPT1) as if they are dates. If you must open these files in Excel, make sure to open it from the File menu in Excel and manually load the column containing gene names as "Text" instead of "General".

VISAGE Output Files (cntd.)

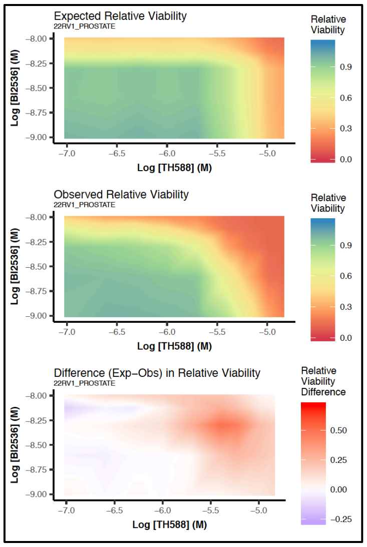
- `genes_ranked_by_viability_{metric}_correlation.rnk`



images
22RV1_PROSTATE.pdf
A549_LUNG.pdf
AU565_BREAST.pdf

PDF files in the `images/` output subdirectory.

If the `make_plots` flag was set to `TRUE` in `visage_params.txt`, your output directory should have a subdirectory called `images/` containing a PDF for each cell line with viability data.



Expected and observed viability surfaces, and the difference between the two.

The first page of each PDF contains 3 plots depicting the expected and observed dose-viability surfaces, and the difference between the two. These surfaces are linearly interpolated from the data computed at the drug doses provided by the user. They use the unscaled viability data, as that is the most intuitive for user examination.

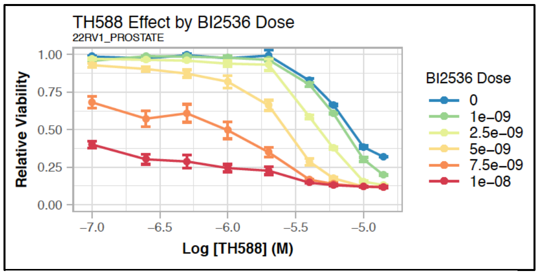
The expected viability at a combination of two drug doses is the product of the viability in the presence of each drug alone (that is, the assumption of Bliss independence).

The observed viability is the data provided in the cell viability input data.

The difference between the two is the degree of unexpected loss of viability caused by the drug combination, where red is greater-than-expected loss of viability, and blue is less-than-expected loss of viability.

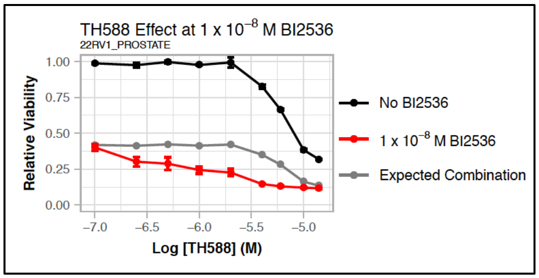
The following pages in each PDF file show "slices" of these dose-viability surfaces, displaying the expected and observed effect of each drug at every dose of the other.

For each drug, the effect of that drug observed at every individual dose of the other drug is shown on a single plot.



Observed effect of TH588 at each dose of BI2536 in 22RV1 prostate cancer cells.

For each drug, several plots show the observed and expected effect of that drug at a single dose of the other drug.



Observed and expected effects of TH588 in the presence or absence of 10 nM BI2536 in 22RV1 prostate cancer cells.