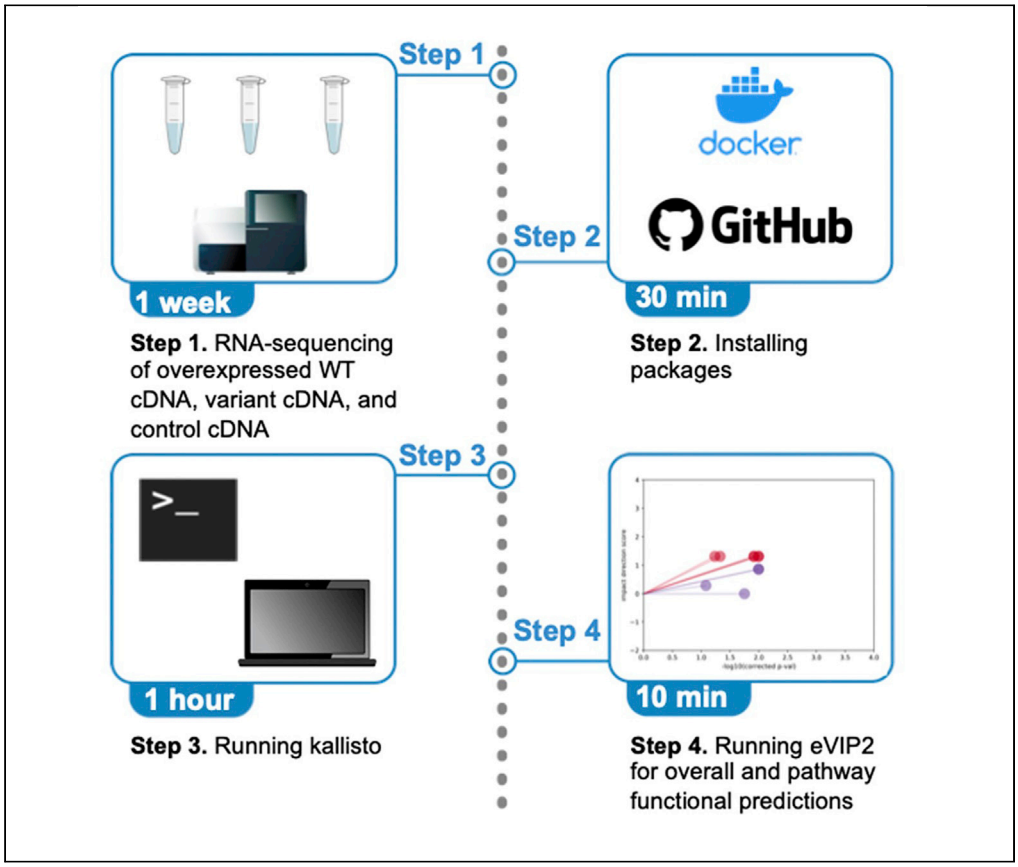
STAR Protocols

CellPress
OPEN ACCESS

Protocol

An expression-based variant impact phenotyping protocol to predict the impact of gene variants in cell lines



We describe a bioinformatics protocol for eVIP2 (expression-based variant impact phenotyping). eVIP2 can predict a gene variant's functional impact by comparing gene expression signatures induced by introduction of wild-type versus mutant cDNAs in cell lines. The predicted functional outcomes of the variants include gain-of-function, loss-of-function, change-of-function, or neutral. eVIP2 improves upon eVIP by being applicable to RNA-seq data and providing pathway-level functional predictions for each mutation. Here, we detail how to run eVIP2 on RNA-seq data from two RNF43 variants.

Publisher's note: Undertaking any experimental protocol requires adherence to local institutional guidelines for laboratory safety and ethics.

Alexis M. Thornton, Manoj Tumu, Angela N. Brooks

anbrooks@ucsc.edu

Highlights

A user-friendly computational tool to predict variant impact

Profiles many variants across multiple genes in a single run

Identification of specific signaling pathways affected by each variant

Static and interactive visualization of results

Thornton et al., STAR Protocols 3, 101651
September 16, 2022 © 2022
https://doi.org/10.1016/j.xpro.2022.101651

STAR Protocols

CellPress
OPEN ACCESS

Protocol

An expression-based variant impact phenotyping protocol to predict the impact of gene variants in cell lines

Alexis M. Thornton,[1,2,3] Manoj Tumu,[1,2] and Angela N. Brooks[1,2,4,*]

[1]Department of Biomolecular Engineering, University of California Santa Cruz, Santa Cruz, CA 95064, USA

[2]UCSC Genomics Institute, University of California Santa Cruz, Santa Cruz, CA 95064, USA

[3]Technical contact: althornt@ucsc.edu

[4]Lead contact

*Correspondence: anbrooks@ucsc.edu
https://doi.org/10.1016/j.xpro.2022.101651

SUMMARY

We describe a bioinformatics protocol for eVIP2 (expression-based variant impact phenotyping). eVIP2 can predict a gene variant's functional impact by comparing gene expression signatures induced by introduction of wild-type versus mutant cDNAs in cell lines. The predicted functional outcomes of the variants include gain-of-function, loss-of-function, change-of-function, or neutral. eVIP2 improves upon eVIP by being applicable to RNA-seq data and providing pathway-level functional predictions for each mutation. Here, we detail how to run eVIP2 on RNA-seq data from two RNF43 variants.

For complete details on the use and execution of this protocol, please refer to Thornton et al. (2021).

BEFORE YOU BEGIN

Experimental validation of cancer-associated gene variants is costly, time-consuming, and requires prior knowledge of a gene's function. Due to these challenges, many cancer variants remain uninvestigated. eVIP2 requires no prior knowledge of the wild-type (WT) gene's function to characterize if a gene variant causes a gain, loss, or change in function, or if it is neutral.

In Berger et al. (2016), eVIP and the L1000 Luminex bead-based gene expression assay were used to characterize 194 somatic mutations in 53 genes identified in primary lung adenocarcinomas (Subramanian et al., 2017; Berger et al., 2016). This work demonstrated the feasibility of high-throughput systematic functional interpretation of variants using gene expression data.

However, because the L1000 assay only measures the abundance of 978 "landmark" genes, we developed eVIP2 for use with RNA-seq data. With RNA-seq data we get a more complete profile of the transcriptome which gives the opportunity for pathway analysis. We used eVIP2 to discover two recurrent frameshift mutations in *RNF43* (G659fs and R117fs) that have different effects on gene function. In the RNF43 G659fs variant, we identified multiple cancer pathways to be impacted, which we validated with reporter assays (Thornton et al., 2021).

To find a variant's impact on pathways, we classify genes into WT-specific and mutant-specific. The WT-specific genes are differentially expressed only in a control condition versus WT and not in a control condition versus mutant. Alternatively, mutation-specific genes are differentially expressed only in a control condition vs mutation and not in a control condition versus WT. These genes represent a

new function caused by the mutant. eVIP2 pathway analysis is performed separately on WT-specific and mutant-specific genes.

The eVIP approach has recently been extended to functionally assess variant impact using single-cell RNA-seq data (sc-eVIP) (Ursu et al., 2022). It has also been applied to features extracted from cell images for cell morphological profiling (cmVIP) (Caicedo et al., 2022). Additionally, the eVIP2 approach can generalize to other data types beyond gene expression. Any generic table can be used as input to get overall functional predictions. A gene variant may cause different effects on different aspects of biology. For example, a mutation may not affect gene expression profiles ("neutral"), but have a strong effect on splicing profiles ("GOF"). Pathway analysis has only been tested on gene expression data and pathway-level functional prediction of other data types is an ongoing area of research.

This protocol describes the steps used for eVIP2 analysis for the two RNF43 frameshift variants described in Thornton et al. (Thornton et al., 2021). Four replicates of *RNF43* WT, *RNF43 R117fs*, *RNF43 G659fs*, and *GFP* (control) cDNAs were overexpressed in HEK293T cells. However, other genes, variants, and cell lines can be used for eVIP2 analysis. We first demonstrate eVIP2 overall and pathway analysis on the Kallisto-processed RNA-seq gene expression of RNF43 (Bray et al., 2016). In part two, we demonstrate using a gene expression table as input for overall eVIP2 predictions. Finally, we demonstrate the use of quantification of alternative splicing events from JuncBASE, to characterize the overall splicing impact of the RNF43 variants (Brooks et al., 2011).

Note: Before beginning, the user should have already performed RNA-seq or L1000 profiling on at least 3 replicates of cell lines upon introduction of (1) the WT version of the gene, (2) the variant, and (3) control (i.e., GFP or RFP) cDNAs. For best results, 4–8 replicates should be performed (Berger et al., 2016; Thornton et al., 2021).

Installing prerequisites

Timing: 10 min

1. Install Docker https://docs.docker.com/get-docker/ (Merkel, 2014).

Note: Users are encouraged to use the provided Docker image which contains all required files, tools, and libraries. However, the prerequisites can be installed separately. The dependencies are listed in https://github.com/BrooksLabUCSC/eVIP2/blob/master/misc/environment.yml.

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
RNA-seq upon overexpression of WT RNF43, RNF43 G659fs, RNF43 R117fs, or GFP in HEK 293 cells	Thornton et al. (2021)	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE141963
GTF file "Homo_sapiens.GRCh38.87.gtf". Any GTF file can be used.	Ensembl (Howe et al., 2021)	http://ftp.ensembl.org/pub/release-87/gtf/homo_sapiens
Hallmark gene set version 6.0 "h.all.v6.0.symbols.gmt". Any gmt file with matching gene IDs can be used.	MSigDB (Liberzon et al., 2011)	https://data.broadinstitute.org/gsea-msigdb/msigdb/release/6.0/
Software and algorithms		
eVIP2	Thornton et al. (2021)	https://github.com/BrooksLabUCSC/eVIP2
Docker	Merkel (2014)	https://docs.docker.com/get-docker/
Kallisto	Bray et al. (2016)	https://pachterlab.github.io/kallisto/download.html

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
juncBASE	(Brooks et al., 2011)	https://github.com/BrooksLabUCSC/juncBASE
Other		
Computer (see materials and equipment for more information)	N/A	N/A

MATERIALS AND EQUIPMENT

This protocol was tested on the following hardware: A 2018 MacBook Pro with 2.7 GHz Quad-Core Intel Core i7 processor and 16 GB of RAM running MacOS Monterey; A Linux server with 160 cores and 1.5 TB of RAM and running CentOS Linux 7; A Lenovo ThinkPad with 2.30 GHz Intel Core i7 processor and 32 GB RAM running Windows 10.

STEP-BY-STEP METHOD DETAILS

eVIP2 characterization of RNF43 variants with Kallisto files as input

Timing: 30 min

Here, we describe how to run eVIP2 on RNA-seq gene expression data from two RNF43 variants. We recommend using Kallisto, but gene quantification from other tools can be used as well, as shown in the sections below. The use of Kallisto is required for eVIP2 pathway analysis.

We prepared a GitHub repository and Docker image containing all the necessary example inputs, reference files, and scripts for the method presented below.

The repository is found on GitHub: https://github.com/BrooksLabUCSC/eVIP2.

The docker image is found at Docker Hub: https://hub.docker.com/r/althornt/evip2_env.

The Docker image contains all files needed to run the following tutorial commands below in the directory named "docker_tutorial_files". Here we describe the files and how they are used.

1. Preparation of RNA-seq data.

Note: We recommend performing general quality control analysis on your data. For information on quality control for raw RNA-seq reads, refer to Conesa et al. (Conesa et al., 2016). Based on our analysis from subsampling read depth, we recommend having at least 20 million reads per replicate.

   a. Run Kallisto on the RNA-seq fastq files with default parameters, or the parameters of your choice. Only default parameters have been tested. Kallisto creates an output directory for each sample, which contains various files, including "abundance.tsv". A directory containing a Kallisto subdirectory for each sample will be used as input to run_eVIP2.py. For this protocol, a folder of Kallisto outputs (`docker_tutorial_files/RNF43_kallisto_outputs`) is provided in the Docker image and can be used as a guide to structure directories for new datasets.

2. Preparation of input files.

To run eVIP2, users need to provide the following required files, which are located within the Docker image and in the GitHub repository:

   a. `-sig_info` This tab delimited file indicates which samples are replicates of which conditions. The sample names listed under distil_id must match the corresponding name of each Kallisto output directory.

distil_id	sig_id	pert_mfc_desc	cell_id	allele
RNF43_WT_4|RNF43_WT_3|RNF43_WT_2|RNF43_WT_1	RNF43_WT	RNF43	293	RNF43_WT
GFP_4|GFP_3|GFP_2|GFP_1	GFP	GFP	293	GFP
RNF43_R117fs_4|RNF43_R117fs_3|RNF43_R117fs_2|RNF43_R117fs_1	RNF43_R117fs	RNF43	293	RNF43_R117fs
RNF43_G659fs_4|RNF43_G659fs_3|RNF43_G659fs_2|RNF43_G659fs_1	RNF43_G659fs	RNF43	293	RNF43_G659fs

b. `-r` The reference file is a tab delimited file that describes which WT to mutant comparisons to do.

wt	mutant
RNF43_WT	RNF43_R117fs
RNF43_WT	RNF43_G659fs

c. `-c` The control file is a tab delimited file to list the controls in the experiment. The control name should match the "sig_id" in the `-sig_info` file. There must be at least one control.

GFP

Note: When running eVIP2 Pathways (by declaring `-eVIPP`) the following additional files are required and are provided on the Docker image:

d. `-gmt` Gene set file in .gmt format. These can be downloaded from MSigDB http://www.gsea-msigdb.org/gsea/msigdb/collections.jsp (Liberzon et al., 2011). Custom gene sets may also be created.

e. `-gtf` A gtf file used to convert transcript counts to gene counts. In this tutorial we use the Ensembl GRCh38 version of the reference genome, but eVIP2 is compatible with any version (Howe et al., 2021). The gtf file used with eVIP2 should be the same as the gtf used as input to Kallisto.

3. Setting up eVIP2 repo and Docker container.

a. Clone the eVIP2 repo to the desired path on your machine:

```
> git clone https://github.com/BrooksLabUCSC/eVIP2.git
```

b. Pull the Docker image.

```
> docker pull althornt/evip2_env
```

c. Verify the Docker image installation.

```
> docker image ls
```

Which should display the `althornt/evip2_env` repository name.

```
REPOSITORY    TAG IMAGE ID    CREATED    SIZE

althornt/evip2_env    latest    8f9af2abd32e    1 h ago    4.9GB
```

d. Run the Docker container:

i. `-v` mounts the locally cloned eVIP2 directory inside of the container.

ii. `-ti` makes the container interactive.

```
> docker run -v /path/to/eVIP2:/eVIP2 -ti althornt/evip2_env
```

4. Running eVIP2 command from kallisto outputs.

a. Now, we demonstrate the first of three independent eVIP2 applications. The following eVIP2 command recreates the overall and pathway analysis of the RNF43 variants presented in Thornton et al. (Thornton et al., 2021). Enter the eVIP2 directory and run eVIP2 with pathway analysis. Table 1 provides an explanation of all parameters of the run_eVIP2.py command:

```
> cd eVIP2

> python2 run_eVIP2.py \

      -input_dir ../docker_tutorial_files/RNF43_kallisto_outputs \

      -out_directory tutorial_files/eVIP2_output_from_kallisto \

      -sig_info tutorial_files/RNF43_sig.info \

      -c tutorial_files/controls.grp \

      -r tutorial_files/comparisons.tsv \

      -gmt tutorial_files/h.all.v6.0.symbols.gmt \

      -gtf ../docker_tutorial_files/Homo_sapiens.GRCh38.87.gtf \

      -num_reps 4 \

      -use_c_pval \

      -eVIPP
```

eVIP2 characterization of RNF43 variants with a gene expression table as input

Timing: 1 min

5. Running eVIP2 command from gene expression table.
   a. For a second eVIP2 application, we demonstrate running eVIP2 using a gene expression table as input as an alternative to using the Kallisto inputs. At this time, pathway analysis has only been tested with Kallisto input; however, you can still run eVIP2 with a generic gene expression table without pathway analysis. To run the eVIP2 pipeline without pathway analysis, the `–input_gene_tpm` or `–input_table` parameters can be used as input with run_eVIP2.py.
   b. Since we are using the same data and experimental setup as in the previous application, we use many of the same input files and parameters files as above. In the following command, we use a gene expression table from the RNF43 experiment to get the overall eVIP2 predictions.

```
python2 run_eVIP2.py \

  -input_gene_tpm tutorial_files/RNF43_gene_exp.tsv \

  ---out_directory tutorial_files/eVIP2_output_from_gene_exp_table \

  ---sig_info tutorial_files/RNF43_sig.info \

  -c tutorial_files/controls.grp \

  -r tutorial_files/comparisons.tsv \

  -num_reps 4 \

  -use_c_pval
```

Table 1. Description of run_eVIP2.py parameters

Parameter	Function
–out_directory	Path to write the eVIP2 output directory
–input_dir	Path to directory containing Kallisto outputs to use as input
–input_table	A generic input table in .tsv format for eVIP2 overall prediction
–input_gene_tpm	A gene expression table in .tsv format for eVIP2 overall prediction

–sig_info	A tsv file with sample information for the following headers: distil_id, sig_id, pert_mfc_desc, cell_id, allele. Each row must list a different group of replicates.
	distil_id = replicate sample names;
	
	sig_id = the replicates condition;
	pert_mfc_desc = the associated WT gene;
	cell_id = name of the cell type used;
	allele = the version of the gene
	
	For example:
	distil_id = RNF43_R117fs_4|RNF43_R117fs_3|RNF43_R117fs_2|RNF43_R117fs_1;
	sig_id = RNF43_R117fs;
	pert_mfc_desc = RNF43;
	cell_id = 293;
	allele = RNF43_R117fs

-c	.grp file containing allele names of control perturbations. If this file is given, a null will be calculated from these
-r	File explicitly indicating which comparisons to do. Assumes the file has a header and it is ignored. The first column is the reference allele and the second column is the test allele. Alleles must match the "allele" column from the –sig_info file
–num_reps	Number of replicates expected for each allele.
–min_tpm	When filtering the gene expression table given with –input_gene_tpm, this value is the minimum TPM value for each gene. If a gene is expressed below this level in all samples, the gene is removed from the table. DEFAULT=1
–conn_thresh	P-value threshold for connectivity vs null. DEFAULT=0.1
–mut_wt_rep_thresh	P-value threshold for comparison of WT and mut robustness. DEFAULT=0.1
–disting_thresh	P-value threshold that tests if mut and wt reps are indistinguishable from each other. DEFAULT=0.1
–mut_wt_rep_rank_diff	The minimum difference in median self replicate correlation between WT and mutation to consider a difference. DEFAULT=0
–use_c_pval	Use the corrected p-values instead of raw p-values
–cond_max_diff_thresh	Threshold for maximum difference between condition correlation medians when determining if a variant is not neutral. DEFAULT = 0.2
–pdf	Create plots in pdf format instead of png.
–xmin	Minimum value on sparkler plot x-axis. DEFAULT = 0
–xmax	Maximum value on sparkler plot x-axis. DEFAULT = 4
–ymin	Minimum value on sparkler plot y-axis. DEFAULT = -3
–ymax	Maximum value on sparkler plot y-axis. DEFAULT = 3
–eVIPP	Perform eVIP2 Pathway analysis. Must also provide –gmt and –gtf
–control	If there are multiple controls in the controls file, designate which to use for DEseq2 when running eVIP2 Pathways
–gtf	GTF reference file used to convert transcript counts to gene counts
–gmt	Gene set file in .gmt format needed when running eVIP2 pathways with –eVIPP
–min_genes	Minimum number of genes per pathway when running eVIP2 Pathways. DEFAULT = 10
–viz_off	Skip creation of heatmaps and scatter plots
–sparkler_off	Skip creation of sparkler plots

eVIP2 RNF43 variants JuncBASE table

Timing: 1 min

6. Running eVIP2 command from JuncBASE table.
   a. So far, we have used gene expression as input to predict variant impact. Now we demonstrate how tables representing other biological measurements can be used as well. Junction Based Analysis of Alternative Splicing Events (JuncBASE) is a tool to identify and quantify alternative splicing in RNA-seq data (Brooks et al., 2011).

b. Here we use a table of quantification of alternative splicing events generated by JuncBASE to see how the RNF43 gene variants impact the splicing profiles using the `run_eVIP2.py` script. We use the `-input_table` parameter to use the JuncBASE table as input.

```
python2 run_eVIP2.py \
  -input_table tutorial_files/RNF43_JuncBASE_PSI_infile.txt \
  -out_directory tutorial_files/eVIP2_output_from_juncBASE \
  -sig_info tutorial_files/RNF43_sig.info \
  -c tutorial_files/controls.grp \
  -r tutorial_files/comparisons.tsv \
  -num_reps 4 \
  -use_c_pval
```

EXPECTED OUTCOMES

The first run_eVIP2.py command above creates four output directories. The "kallisto_files" directory contains the combined, filtered, and log-transformed Kallisto gene count file. The "eVIP_out" directory contains the results of the overall eVIP2 run for each variant. The "deseq2" directory contains the outputs from the DESeq2 runs for each gene variant, which are used for the pathway analysis. The "eVIPP_out" directory contains results of the eVIP2 Pathway analysis run for each variant.

When using tables as input with –input_gene_tpm and –input_table only the eVIP_out directory is created because pathway analysis is not performed. For overall prediction, the "predict.txt" file within eVIP_out is the main result file. Figure 1 describes the eVIP2 decision-tree based algorithm and its corresponding parameters and statistical values.

Table 2 provides explanations for each of the columns in "predict.txt". The most important calculation is the Benjamini-Hochberg false discovery rate corrected impact p-value, which is labeled "wt_mut_rep_vs_wt_mut_conn_c_pval" in the "predict.txt" file. When this p-value is below the cutoff (0.05 or 0.1), the gene variant is considered impactful and is considered neutral when above the cutoff. When using a low number of replicates, a p-value cutoff of 0.1 is suggested. For a more detailed explanation, see Thornton et al. (Thornton et al., 2021).

For eVIP2 pathway analysis, the eVIP approach is run independently on each gene set for all gene variants. Therefore, each gene set has a corresponding predict file and visualizations. The three main visualizations are sparkler plots, impact prediction plots, and scatter plots.

Sparkler plots represent eVIP2 predictions where each variant or pathway is a point. The x-axis represents the Kruskal Wallis "impact test" -log10(adjusted p-value). The y-axis is the "impact direction score", the absolute value of which is equal to the –log10 (adjusted p-value) of a Wilcoxon test directly comparing wild-type and mutant ORF replicate consistency. The sign of the impact direction score is positive if the mutant replicate consistency is greater than WT and negative if the mutant replicate consistency is less than the WT replicate consistency.

Figure 2A shows a sparkler plot where both RNF43 mutations are predicted to be impactful but have opposite trajectories. Figure 2B shows that the mutation-specific pathway results for the RNF43 G659fs variant has multiple change-of-function and gain-of-function pathways. When using the JuncBASE table, we find both RNF43 variants to be neutral. Therefore, we can conclude while both variants impact the gene expression profiles, they do not impact the splicing profiles (Figure 2C).

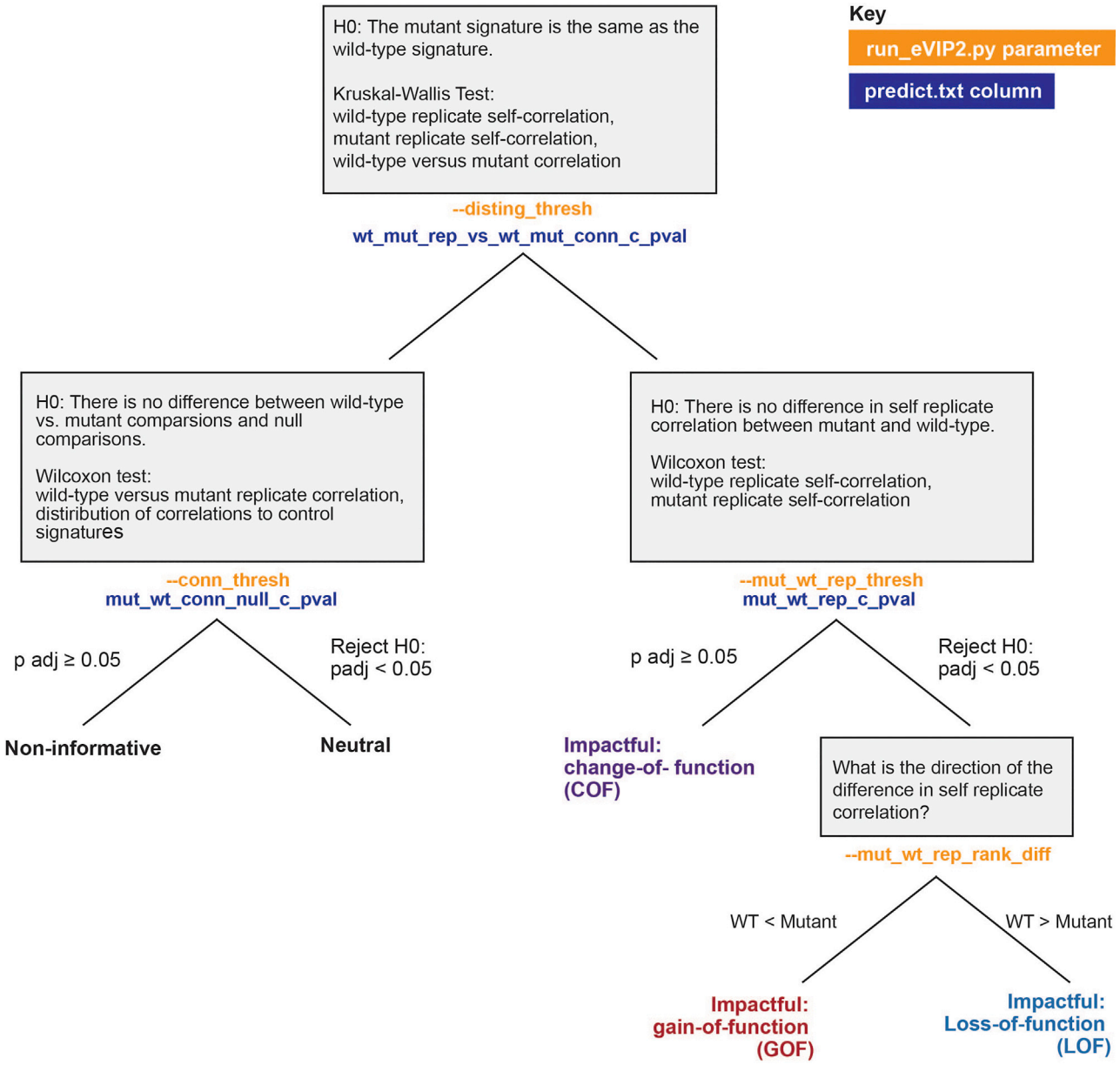


Figure 1. eVIP2 decision-tree algorithm and the associated parameters and names of resulting output values

Schematic of eVIP2 decision-tree based algorithm. Each step of the tree is a statistical test where the p-value cutoff can be set by a parameter (orange) and the test result can be found in the noted column of the predict.txt file (blue).

Interactive sparkler plot HTML files are also created in the directory declared with the –out_directory parameter by using Plotly Dash (Plotly Technologies, Inc, 2015). The interactive sparkler plots are particularly useful when looking at results from many gene variants or pathways, where labels may be overlapping and hard to read.

Impact prediction plots are made for each WT gene and all of its corresponding gene variants. They feature heatmap representations of WT replicate consistency or variant replicate consistency, where the values correspond to Spearman rank correlation. The signature identity (WT vs variant) is represented by heatmaps in the second row. They also feature a dot-plot representation of replicate

Table 2. Explanation of columns in prediction file output (predict.txt)

Column name	Explanation
gene	Name of gene
mut	Name of gene variant
mut_rep	Wild-type replicate self-correlation
wt_rep	Mutant replicate self-correlation
mut_wt_connectivity	Wild-type vs mutant correlation
wt	Name of associated WT gene
cell_line	Name of the cell line used for the variant, which is taken from the sig info file "cell_id" column.
mut_wt_rep_pval	Result of a Wilcoxon test testing the null hypothesis that there is no difference in self replicate correlation between mutant and wildtype. Uses the wt_rep and mut_rep values.
mut_wt_conn_null_pval	Result of a Wilcoxon test testing the null hypothesis there is no difference between the wild-type vs mutant comparisons and the null comparisons.
wt_mut_rep_vs_wt_mut_conn_pval	Result of Kruskal-Wallis test testing the null hypothesis that the mutant signature is the same as the WT signature. Uses the "wt_rep", "mut_rep", "mut_wt_connectivity" correlation values.
kruskal_diff	Difference between the max and minimum between the three values used in the Kruskal-Wallis test ("wt_rep", "mut_rep", "mut_wt_connectivity")
mut_wt_rep_c_pval	mut_wt_rep_pval after Benjamini-Hochberg false discovery rate correction
mut_wt_conn_null_c_pval	mut_wt_conn_null_pval after Benjamini-Hochberg false discovery rate correction
wt_mut_rep_vs_wt_mut_conn_c_pval	wt_mut_rep_vs_wt_mut_conn_pval after Benjamini-Hochberg false discovery rate correction
prediction	Neutral, COF, LOF, GOF, NI

consistency and signature identity measured by Spearman rank correlation . *, adjusted $p < 0.1$. n.s., adjusted $p > 0.1$ (Figure 3).

Lastly, scatter plot matrices are made for each gene variant. The z-scores of gene expression for each replicate of the WT and gene variant are compared to each other. This plot can help visualize comparisons of gene expression profiles among WT replicates, among mutant replicates, and between WT and mutant replicates. For example, for the LOF RNF43 variant, the gene expression profiles of mutant replicates are less self-correlated than the WT replicates (Figure 4).

LIMITATIONS

Due to eVIP2's use of replicate self-correlation to determine the directionality of mutations impact, with only 3 replicates it is likely that most calls will either be change-of-function or neutral. It is unlikely to get any significant loss of function or gain of calls with only 3 replicates. It is recommended that experiments should have 4–8 replicates (Berger et al., 2016; Thornton et al., 2021).

TROUBLESHOOTING

Problem 1

Using eVIP2 without Docker.

Potential solution

If the provided Docker image (which contains all required files, tools, and libraries) cannot be used, the needed libraries can be installed separately. They are listed in https://github.com/BrooksLabUCSC/eVIP2/blob/master/misc/environment.yml and https://github.com/BrooksLabUCSC/eVIP2/blob/master/misc/Dockerfile. Additionally, the code and tutorial files are on the eVIP2 GitHub Repo: https://github.com/BrooksLabUCSC/eVIP2.

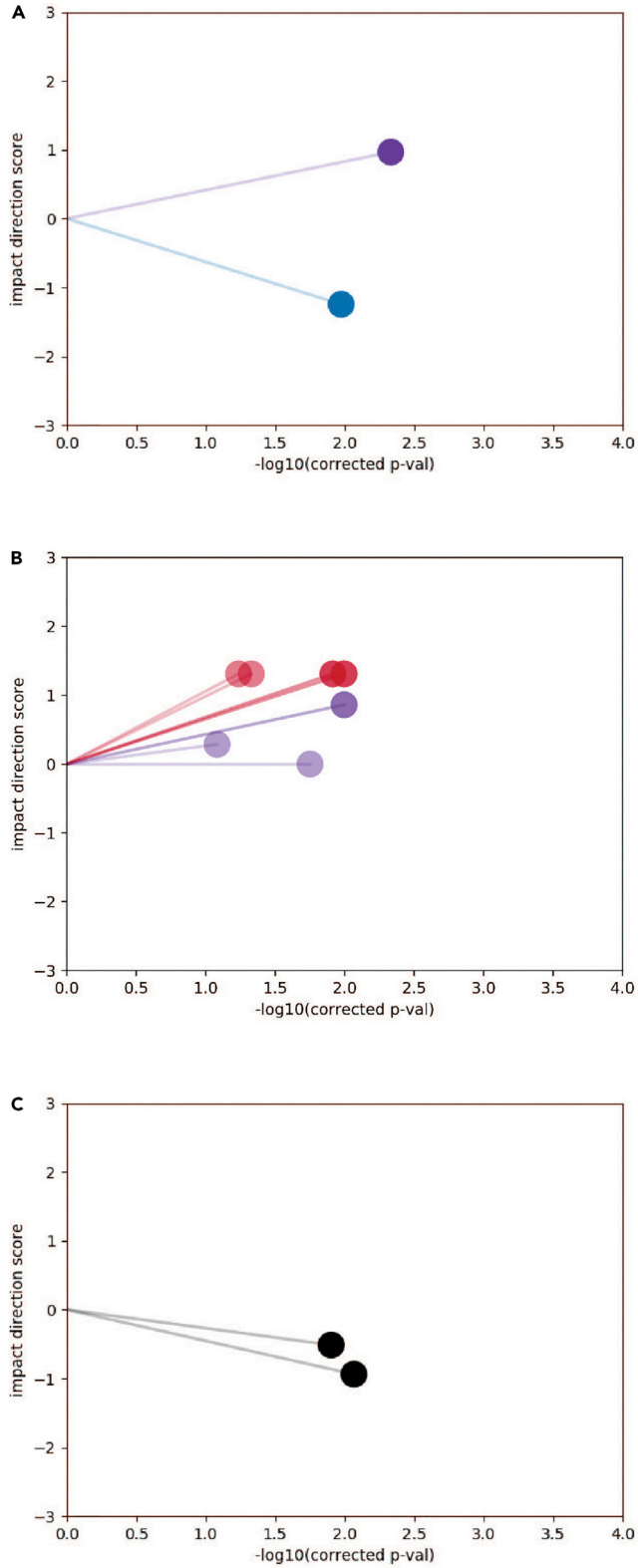


Figure 2. eVIP2 sparkler plot results

(A) Overall impact of RNF43 variants using gene expression. The x axis represents the p value from the Kruskal Wallis "impact test."

(B) RNF43 G659fs mutation-specific pathway impact using gene expression.

(C) Overall impact of RNF43 variants using quantification of alternative splicing events from JuncBASE.

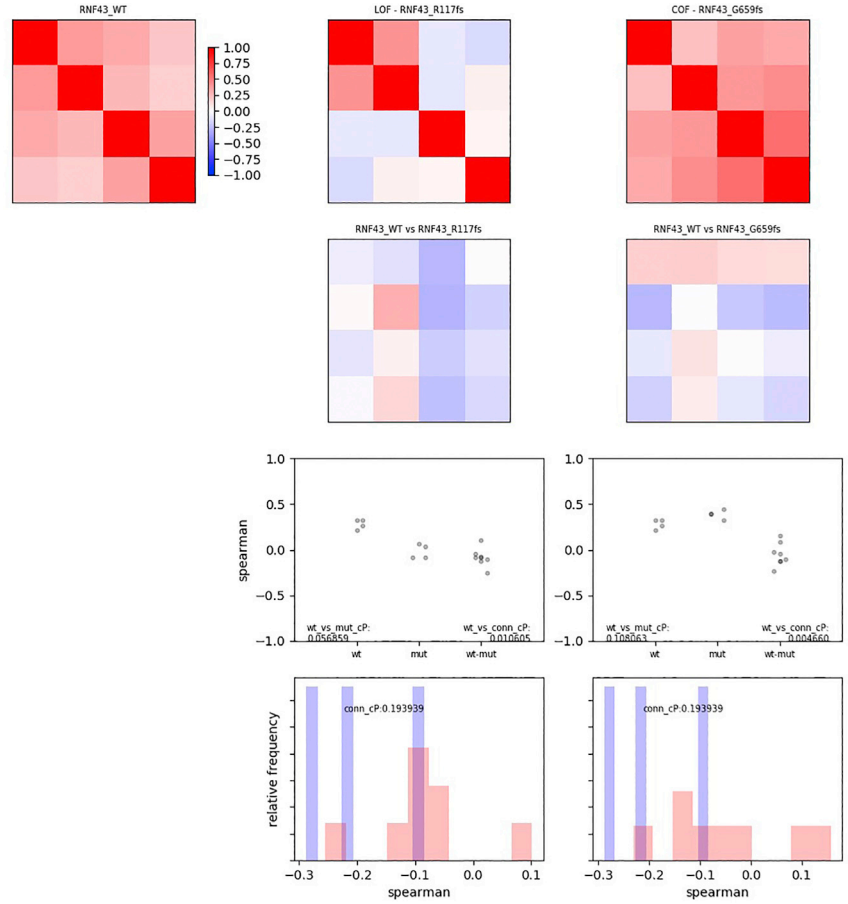


Figure 3. eVIP2 heatmap and impact prediction result

Top, heatmap representations of WT replicate consistency or variant replicate consistency, where the values correspond to Spearman rank correlation. The signature identity (WT vs variant) is represented by heatmaps in the second row. Middle, dot-plot representation of replicate consistency and signature identity measured by Spearman rank correlation. *, adjusted $p < 0.1$. n.s., adjusted $p > 0.1$. Bottom, the distribution of correlations of control signatures in blue and the WT vs mutant in red.

Also, a bash script (tutorial_files/setup.sh in the eVIP2 repo) can be used to download the Kallisto "abundance.tsv" files from GEO and formats them into the original Kallisto output directory structure so they can be used as input to run_eVIP2.py.

Problem 2

Using Singularity.

Potential solution

The eVIP2 docker image can be used with Singularity. We can create a Singularity image file ("evip2_env_latest.sif") by pulling the eVIP docker image. Then, we can run the eVIP2.py commands described above. Unlike the Docker image, the created Singularity image will not contain the Kallisto and gtf tutorial files, which must be downloaded separately.

```
> singularity pull docker://althornt/evip2_env:latest

> git clone <hyperlink refid="https://github.com/BrooksLabUCSC/eVIP2.git">https://gi-
thub.com/BrooksLabUCSC/eVIP2.git
```

```
> cd eVIP2

> singularity exec /path/to/evip2_env_latest.sif python2 run_eVIP2.py \

     -input_dir ../docker_tutorial_files/RNF43_kallisto_outputs \

     -out_directory tutorial_files/eVIP2_output_from_kallisto \

     -sig_info tutorial_files/RNF43_sig.info \

     -c tutorial_files/controls.grp \

     -r tutorial_files/comparisons.tsv \

     -gmt tutorial_files/h.all.v6.0.symbols.gmt \

     -gtf ../docker_tutorial_files/Homo_sapiens.GRCh38.87.gtf \

     -num_reps 4 \

     -use_c_pval \

     -eVIPP
```

Problem 3

Using Windows.

Potential solution

We recommend Windows users use the PowerShell terminal program.

Problem 4

Key errors.

Potential solution

Key errors are likely due to the sample names in the –sig_info file not matching the corresponding names of the kallisto directories. Incorrect formatting of the –sig_info file may also cause issues. Refer to the example files provided in the git repo and check that files are formatted properly.

Problem 5

Sparkler plots are cut off.

Potential solution

Rerun run_eVIP2.py and adjust the axis of the sparkler plot by changing the values for –xmin, –xmax, –ymin, or –ymax.

RESOURCE AVAILABILITY

Lead contact

Further information and requests should be directed to and will be fulfilled by the lead contact, Angela N. Brooks (anbrooks@ucsc.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

The eVIP2 code and tutorial data used in this study are available on GitHub eVIP2: https://github.com/BrooksLabUCSC/eVIP2. The eVIP2 release used in this protocol is: https://doi.org/10.5281/zenodo.6863716. The RNF43 RNA-seq dataset used in this study is available at GEO Accession: GSE141963.

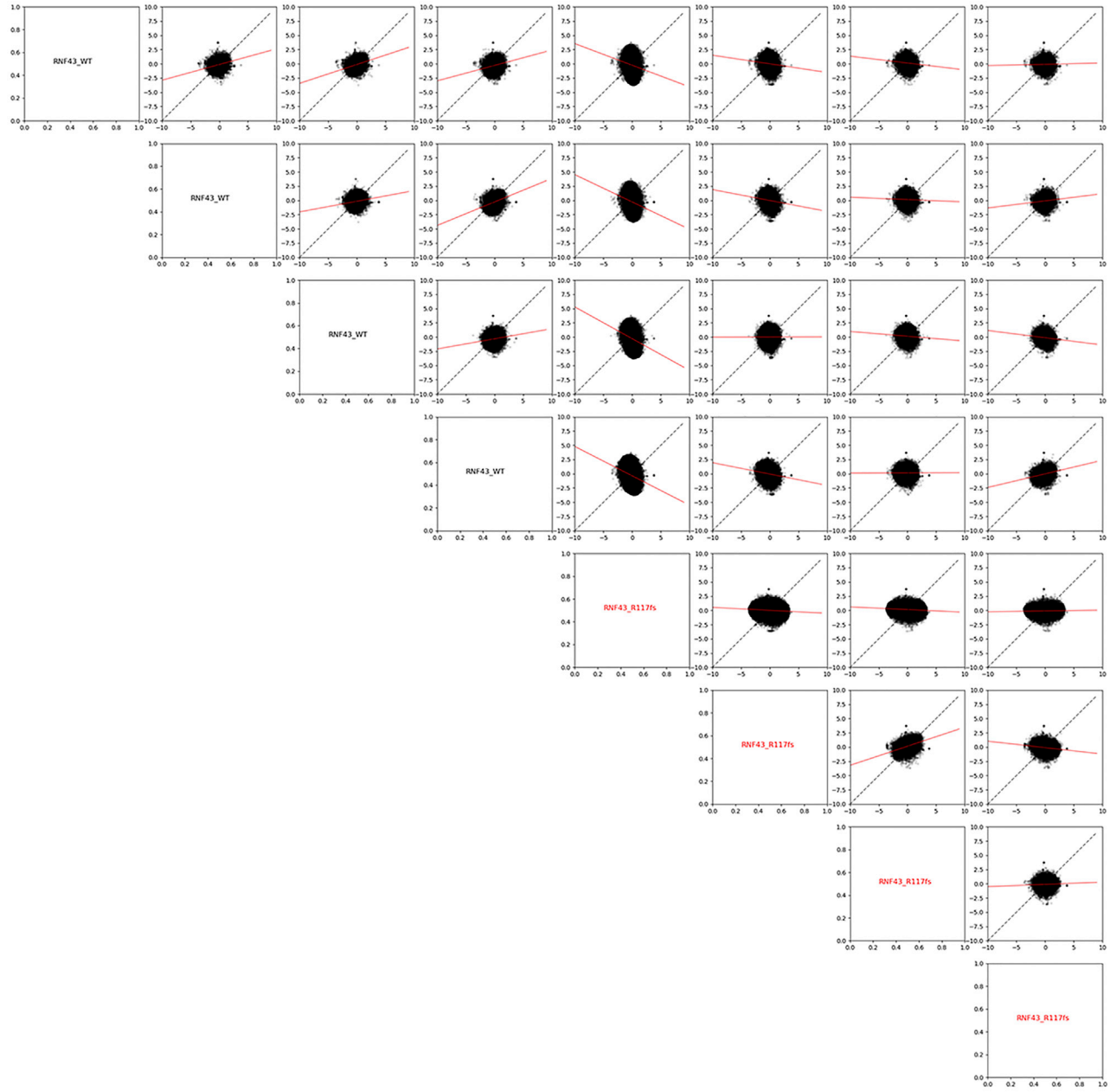


Figure 4. eVIP2 scatter plot results

Scatter plot correlation matrices for the loss-of-function RNF43 R117fs variant.

ACKNOWLEDGMENTS

We would like to thank Roman Reggiardo, April Lo, and Cindy Liang for testing the eVIP2 protocol. A.M.T. was funded through NIH grant 5T32HG008345, the Eugene Cota-Robles Fellowship, and the Bill H. James Foundation scholarship. This work was also partially supported by the Damon Runyon Cancer Research Foundation to A.N.B.

AUTHOR CONTRIBUTIONS

Conceptualization, A.M.T. and A.N.B.; Methodology, A.M.T. and A.N.B.; Software, A.M.T., M.T., and A.N.B.; Validation, A.M.T., M.T., and A.N.B.; Formal Analysis , A.M.T. and A.N.B.;

Writing – Original Draft, A.M.T.; Writing – Review & Editing, A.M.T. and A.N.B.; Visualization, A.M.T., M.T., and A.N.B.; Supervision, A.N.B.; Project Administration, A.N.B.; Funding Acquisition, A.M.T. and A.N.B.

DECLARATION OF INTERESTS

A.N.B. is a consultant for Remix Therapeutics, Inc. All other authors have declared that no competing interests exist.

REFERENCES

Berger, A.H., Brooks, A.N., Wu, X., Shrestha, Y., Chouinard, C., Piccioni, F., Bagul, M., Kamburov, A., Imielinski, M., Hogstrom, L., et al. (2016). High-throughput phenotyping of lung cancer somatic mutations. Cancer Cell *30*, 214–228. https://doi.org/10.1016/j.ccell.2016.06.022.

Bray, N.L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol. *34*, 525–527. https://doi.org/10.1038/nbt.3519.

Brooks, A.N., Yang, L., Duff, M.O., Hansen, K.D., Park, J.W., Dudoit, S., Brenner, S.E., and Graveley, B.R. (2011). Conservation of an RNA regulatory map between Drosophila and mammals. Genome Res. *21*, 193–202. https://doi.org/10.1101/gr.108662.110.

Caicedo, J.C., Arevalo, J., Piccioni, F., Bray, M.-A., Hartland, C.L., Wu, X., Brooks, A.N., Berger, A.H., Boehm, J.S., Carpenter, A.E., and Singh, S. (2022). Cell Painting predicts impact of lung cancer variants. Mol. Biol. Cell *33*, ar49. https://doi.org/10.1091/mbc.E21-11-0538.

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M.W., Gaffney, D.J., Elo, L.L., Zhang, X., and Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. Genome Biol. *17*, 13. https://doi.org/10.1186/s13059-016-0881-8.

Howe, K.L., Achuthan, P., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M.R., Armean, I.M., Azov, A.G., Bennett, R., Bhai, J., et al. (2021). Ensembl 2021. Nucleic Acids Res. *49*, D884–D891. https://doi.org/10.1093/nar/gkaa942.

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., and Mesirov, J.P. (2011). Molecular signatures database (MSigDB) 3.0. Bioinformatics *27*, 1739–1740. https://doi.org/10.1093/bioinformatics/btr260.

Merkel, D. (2014). Docker: lightweight Linux containers for consistent development and deployment. Linux J. *239*, 2.

Plotly Technologies, Inc (2015). Collaborative Data Science (Plotly Technologies Inc.).

Subramanian, A., Narayan, R., Corsello, S.M., Peck, D.D., Natoli, T.E., Lu, X., Gould, J., Davis, J.F., Tubelli, A.A., Asiedu, J.K., et al. (2017). A next generation connectivity map: L1000 platform and the first 1, 000, 000 profiles. Cell *171*, 1437–1452.e17. https://doi.org/10.1016/j.cell.2017.10.049.

Thornton, A.M., Fang, L., Lo, A., McSharry, M., Haan, D., O'Brien, C., Berger, A.H., Giannakis, M., and Brooks, A.N. (2021). eVIP2: expression-based variant impact phenotyping to predict the function of gene variants. PLoS Comput. Biol. *17*, e1009132. https://doi.org/10.1371/journal.pcbi.1009132.

Ursu, O., Neal, J.T., Shea, E., Thakore, P.I., Jerby-Arnon, L., Nguyen, L., Dionne, D., Diaz, C., Bauman, J., Mosaad, M.M., et al. (2022). Massively parallel phenotyping of coding variants in cancer with Perturb-seq. Nat. Biotechnol. *40*, 896–905. https://doi.org/10.1038/s41587-021-01160-7.