The Journal of Open Source Software

gimap: An R Package for Genetic Interaction Mapping in Dual-Target CRISPR Screens

Candace Savonen[1,2], **Phoebe Parrish**[1], **Kate Isaac**[1], **Daniel Groso**[1], **Marissa Fujimoto**[1], **Siobhan O'Brien**[1], **and Alice Berger**[1]

1 Fred Hutchinson Cancer Center, United States **2** Synthesize Bio, United States

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor:

Submitted: 04 August 2025
Published: unpublished

Summary

The gimap (Genetic Interaction MAPping) R package addresses a fundamental challenge in genomic research: the difficulty of understanding combinatorial interactions among genes. Gene redundancy makes traditional single-gene knockout methods ineffective for identifying therapeutic targets, as backup genes can mask the effects when a single gene is disabled. gimap offers a solution by providing a comprehensive framework for analyzing dual-target CRISPR screening data, where two genes are simultaneously disabled to reveal their backup relationships. This software implements the methods used by Parrish et al. (2021). The package processes raw count data through a multi-step pipeline that includes normalization, calculation of expected and observed CRISPR scores, computation of genetic interaction scores, and statistical analysis to identify significant interactions. Unlike general tools, gimap is specifically tailored for paired guide CRISPR data with built-in quality control reporting and visualization tools. The package makes best practices the default options and is available on GitHub with comprehensive documentation to support the research community in extracting meaningful insights from complex genetic screening experiments.

Statement of Need

When multiple genes have the same function, a common result of evolutionary processes, it becomes challenging to isolate their true functions. This redundancy means that many possible therapeutic targets are missed by traditional methods that disable just one gene at a time (De Kegel & Ryan, 2019; Parrish et al., 2021). A more complementary approach involves disabling two genes simultaneously to reveal these backup relationships (Thompson et al., 2021).

Recent advances in CRISPR technology now allow researchers to knock out gene pairs at once, offering a powerful solution to this problem (https://pubmed.ncbi.nlm.nih.gov/31911676/). Although software solutions exist for single knockout CRISPR, such as MAGeCK, there is no standardized software solution for paired gene CRISPR studies (Li et al., 2014).

The R package, called `gimap` (Genetic Interaction MAPping), was developed specifically for analyzing these dual-target CRISPR experiments. It helps researchers identify important relationships between genes, such as when two genes work together or when disabling both creates a dramatic effect that wouldn't occur by disabling either one alone.

`gimap` is specifically tailored to handle the unique characteristics of paired guide CRISPR data, including the distinction between single-targeting and double-targeting constructs and the need to account for differential double-strand break effects. The package seamlessly integrates with data generated using a specialized pgPEN library but can be adapted for most paired guide CRISPR screening approaches (Parrish et al., 2021).

Implementation

`gimap` addresses this need by providing a comprehensive analytical framework for dual-target CRISPR screening data. The package performs several critical functions: (1) normalization of read count data to account for variable sequencing depth and technical biases, (2) calculation of CRISPR scores that reflect the effect of gene knockouts on cell proliferation, (3) determination of expected CRISPR scores for gene pairs based on single-gene effects, (4) computation of genetic interaction scores that quantify deviations from expected effects, and (5) statistical analysis to identify significant interactions.

Overall design philosophy

In order to ensure usability for the research community we built `gimap` using the following design philosophy.

1. Making best practices as default options and including warning messages for when alternative options are chosen (e.g. if filtering has not been applied).

2. Using elements from familiar packages such as fastqc reports (our `run_qc()` function creates such a report) (*Babraham Bioinformatics - FastQC A Quality Control Tool for High Throughput Sequence Data*, n.d.).
3. Trying to document and inform users of the statistics and decisions that have been made by the software clearly.

gimap data handling

`gimap` implements a multi-step analysis pipeline:

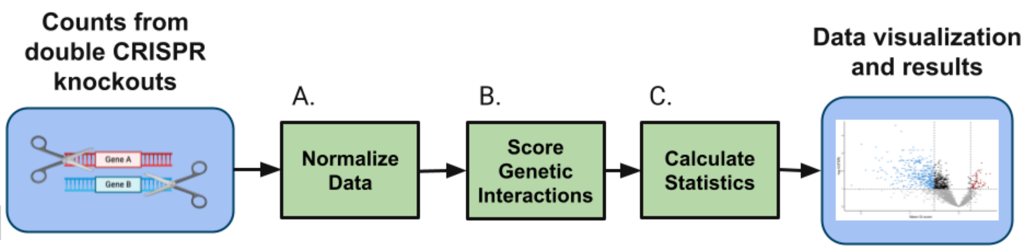

Figure 1: gimap workflow completes 3 main steps. Part A, B, and C of the figure show the major steps of the workflow which are to normalize the data through a multi step process, score genetic interactions based on the expected versus observed CRISPR scores, and finally to calculate statistics to identify statistically significant genetic interactions.

1. **Normalize Data**: Raw count data is transformed into log2 counts per million (CPM) and adjusted by subtracting pre-treatment values to obtain log2 fold changes. These are further normalized based on the distribution of negative (e.g. safe-targeting or non-targeting controls) and positive controls (pgRNAs targeting known essential genes). This scaling normalization is analogous to the normalization methods employed by the Cancer Dependency Map (depmap.org) (Arafeh et al., 2025; DepMap, Broad, 2025).

a. *Log2 Counts Per Million (CPM) Transformation*:

- Let $C_{i,j}$ be the raw count for gene i in sample j
- Let N_j be the total number of counts in sample j

$$L_{i,j} = \log_2\left(\frac{C_{i,j} \times 10^6}{N_j} + 1\right)$$

(The $+1$ is often included to avoid log(0) issues)

b. *Adjustment by Pre-treatment Values*:

- Let $L_{i,j}^{post}$ be the log2 CPM value post-treatment
- Let $L_{i,j}^{pre}$ be the log2 CPM value pre-treatment

$$LFC_{i,j} = L_{i,j}^{post} - L_{i,j}^{pre}$$

c. *Normalization Based on Controls*:

- Let $LFC_{i,j}$ be the log2 fold change calculated above
- Let μ_{neg} and σ_{neg} be the mean and standard deviation of negative controls (safe-targeting or non-targeting)
- Let μ_{pos} and σ_{pos} be the mean and standard deviation of positive controls (pgRNAs targeting essential genes)

$$Z_{i,j} = \frac{LFC_{i,j} - \mu_{neg}}{\mu_{neg} - \mu_{pos}}$$

Or alternatively, using a more complex normalization that accounts for the distributions of both control types:

$$Z_{i,j} = \frac{LFC_{i,j} - \mu_{neg}}{\sigma_{neg}} \times \frac{\sigma_{pos}}{\mu_{neg} - \mu_{pos}}$$

This equation represents the transformation from raw count data to normalized log2 fold changes, calibrated against both negative and positive control distributions.

2. **Score Genetic Interactions**: The goal of this step is to quantify deviations from expected additive effects when two genes are simultaneously targeted, which allows us to identify true genetic interactions beyond what would be predicted from single-gene effects alone.

We model a control distribution for non-interacting genes by assuming that in the absence of genetic interactions, the effect of simultaneously targeting two genes should be additive (i.e., the sum of their individual effects). We further assume that the observed genetic interaction (GI) score is a linear transformation of the expected GI score plus a consistent error term sampled from an approximately normal distribution. This error distribution is shared across all genes and is homoscedastic (constant variance) across expected GI scores, allowing us to use linear modeling approaches to account for systematic biases.

For double-targeting constructs, expected CRISPR scores are calculated as the sum of the corresponding single-targeting scores. For single-targeting constructs, the expected score combines the single-target effect with the mean effect of control constructs. Interaction scores represent the difference between observed and expected CRISPR scores, adjusted using a linear model to account for systematic biases.

For double-targeting constructs: - Let $S_{i,j}^{obs}$ be the observed CRISPR score for a construct targeting genes i and j - Let S_i be the single-targeting score for gene i - Let S_j be the single-targeting score for gene j

The expected score for a double-targeting construct is:

$$S_{i,j}^{exp} = S_i + S_j$$

For single-targeting constructs: - Let S_i^{obs} be the observed CRISPR score for a construct targeting gene i - Let $\mu_{control}$ be the mean effect of control constructs

The expected score for a single-targeting construct is:

$$S_i^{exp} = S_i + \mu_{control}$$

The interaction score calculation, with adjustment for systematic biases: - Let $I_{i,j}$ be the interaction score for genes i and j - Let $S_{i,j}^{obs}$ be the observed score - Let $S_{i,j}^{exp}$ be the expected score - Let β_0 and β_1 be the intercept and slope from a linear regression of observed vs expected scores

The interaction score calculation:

$$I_{i,j} = S_{i,j}^{obs} - (\beta_0 + \beta_1 \cdot S_{i,j}^{exp})$$

Where the genetic interaction score is the difference between the observed score and the linear model prediction based on the expected score, accounting for systematic deviations between observed and expected values.

3. **Calculate Statistics**: T-tests compare the distribution of double-targeting genetic interaction scores for one pair against the background distribution of single-targeting scores, with false discovery rate correction for multiple hypothesis testing.

- S_{double} as the set of double-targeting genetic interaction scores
- S_{single} as the set of single-targeting scores (background distribution)
- μ_{double} as the mean of double-targeting scores
- μ_{single} as the mean of single-targeting scores
- σ_{double} as the standard deviation of double-targeting scores
- σ_{single} as the standard deviation of single-targeting scores
- n_{double} and n_{single} as the sample size of the paired guides

The t-test statistic would be:

$$t = \frac{\mu_{double} - \mu_{single}}{\sqrt{\frac{\sigma^2_{double}}{n_{double}} + \frac{\sigma^2_{single}}{n_{single}}}}$$

For each comparison, we calculate a p-value from this t-statistic.

Then, to account for multiple hypothesis testing, we apply false discovery rate (FDR) correction using Benjamini Hochberg procedure (Benjamini & Hochberg, 1995):

1. Order all p-values: $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(m)}$
2. For a given FDR threshold α (e.g., 0.05), find the largest k such that:

$$p_{(k)} \leq \frac{k}{m} \cdot \alpha$$

3. Reject the null hypothesis for all tests with p-values $\leq p_{(k)}$

The package also provides comprehensive visualization tools including volcano plots to highlight significant genetic interactions and detailed result tables for further analysis.

Use Cases

`gimap` has been successfully used to identify synthetic lethal interactions among paralog genes in cancer cell lines, revealing potential therapeutic targets that single-gene approaches have missed. The package accommodates various experimental designs, including time-course studies and treatment comparisons, offering flexibility for diverse research questions.

Example applications include:
- Identification of genes that provide functional redundancy in critical cellular pathways
- Discovery of context-dependent genetic interactions that emerge under specific conditions or treatments
- Systematic mapping of gene networks based on functional interactions rather than physical associations

Conclusion

`gimap` provides a robust, accessible framework for analyzing paired guide CRISPR screening data and identifying genetic interactions with potential biological and therapeutic significance. By streamlining the computational workflow from raw counts to statistically rigorous interaction scores, `gimap` enables researchers to efficiently extract meaningful insights from complex genetic screening experiments. The package is available on GitHub (https://github.com/Fred-Hutch/gimap) with comprehensive documentation and tutorials to facilitate adoption by the research community.

Acknowledgements

This work is funded by NCI grant R01CA262556 and the Translational Data Science Integrated Research Center of Fred Hutchinson Cancer Center. SO is a Washington Research Foundation postdoctoral fellow.

References

Arafeh, R., Shibue, T., Dempster, J. M., & others. (2025). The present and future of the cancer dependency map. *Nature Reviews Cancer*, *25*, 59–73. https://doi.org/10.1038/s41568-024-00763-x

Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data. (n.d.). Retrieved April 4, 2025, from https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300.

De Kegel, B., & Ryan, C. J. (2019). Paralog buffering contributes to the variable essentiality of genes in cancer cell lines. *PLoS Genetics*, *15*(10), e1008466. https://doi.org/10.1371/journal.pgen.1008466

DepMap, Broad. (2025). *DepMap public 25Q2*. Dataset. https://depmap.org

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

Parrish, P. C. R., Thomas, J. D., Gabel, A. M., Kamlapurkar, S., Bradley, R. K., & Berger, A. H. (2021). Discovery of synthetic lethal and tumor suppressor paralog pairs in the human genome. *Cell Reports*, *36*(9), 109597. https://doi.org/10.1016/j.celrep.2021.109597

Thompson, N. A., Ranzani, M., Weyden, L. van der, Iyer, V., Offord, V., Droop, A., Behan, F., Gonçalves, E., Speak, A., Iorio, F., Hewinson, J., Harle, V., Robertson, H., Anderson, E., Fu, B., Yang, F., Zagnoli-Vieira, G., Chapman, P., Del Castillo Velasco-Herrera, M., … Adams, D. J. (2021). Combinatorial CRISPR screen identifies fitness effects of gene paralogues. *Nature Communications*, *12*(1), 1302. https://doi.org/10.1038/s41467-021-21478-9