The Journal of Open Source Software

pgMAP: Paired Guide RNA Read Mapping from Dual-Targeting CRISPR Screens

Marissa H. Fujimoto[1], Candace L. Savonen[1], Siobhan O'Brien[1], Phoebe C. R. Parrish[1,2,3], Daniel J. Groso[1,2], James D. Thomas[2,4], Robert K. Bradley[2,3,4], and Alice H. Berger[1,2,3¶]

1 Human Biology Division, Fred Hutchinson Cancer Center, Seattle, WA 98109, USA **2** Computational Biology Program, Public Health Sciences Division, Fred Hutchinson Cancer Center, Seattle, WA 98109, USA **3** Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA **4** Basic Sciences Division, Fred Hutchinson Cancer Center, Seattle, WA 98109, USA ¶ Corresponding author

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor: Open Journals

Reviewers:

- @openjournals

Submitted: 01 January 1970
Published: unpublished

Summary

Genetic interactions occur when two or more genes interact to produce a phenotype that differs from the effect of both genes considered independently. Synthetic lethality is a type of genetic interaction between two genes where the loss of either gene is tolerated individually, but the loss of both genes results in a reduction of cell viability. One possible application of synthetic lethality is cancer therapies. The CRISPR-Cas system allows scalable investigation of synthetic lethality directly in lab-cultured human tumor cells. However, current tools for analyzing dual-targeting CRISPR screens, which is the preferred method to study synthetic lethality at scale, are poorly established compared to existing tools for single-targeting CRISPR screens. We present pgMAP (paired guide RNA MAPper): a command line tool built to read paired guide counts from dual-targeting CRISPR screening sequencing data. By quantifying the relative depletion and enrichment of paired guides between different samples, pgMAP enables the discovery of genetic interactions across many organisms, genetic backgrounds, tissues, and cancer types.

Statement of Need

In recent years, several genome-scale, dual-targeting CRISPR-mediated knockout screening approaches have been developed to map genetic interactions in the human genome (Dede et al., 2020; Gonatopoulos-Pournatzis et al., 2020; Ito et al., 2021; Köferle et al., 2022; Parrish et al., 2021; Tang et al., 2022; Thompson et al., 2021). These methods enable functional profiling of duplicated gene families and expand the range of potentially targetable synthetic lethal interactions in cancer (Dandage & Landry, 2021; Ryan et al., 2023). However, computational methods for genetic interaction mapping from human CRISPR screen data are poorly established, which may impede interpretation of dual-targeting CRISPR screen data and thus prevent identification of actionable synthetic lethal targets.

pgMAP is a command line tool that processes raw sequencing data into tables of counts per genetic perturbation. pgMAP is purpose-built for genetic interaction analysis; rather than utilizing bioinformatics tools designed for general-purpose alignment of DNA sequences, pgMAP exploits the paired structure of dual-targeting CRISPR experiments to efficiently map and count paired guides. Compared to previous approaches, pgMAP produces equivalent results over three times faster while requiring constant memory and disk space relative to the number of reads in the raw sequencing data (Table 1). As a general purpose tool for analyzing

paired knockout CRISPR screen data, pgMAP can be configured with parameters including error tolerances and sequencing strategies to suit a wide berth of experimental designs.

The paired guide counts data from pgMAP can be further analyzed using the gimap R package to perform quality control and statistical analysis of genetic interaction (Savonen et al., 2025). Together pgMAP and gimap provide a complete software suite for the end-to-end data analysis of paired knockout CRISPR screens.

Table 1: Performance comparison of pgMAP and bowtie to count paired guides from a pgPEN screen. Fastqs used from "Discovery of synthetic lethal and tumor suppressor paralog pairs in the human genome" by Parrish Et. Al 2021. Both benchmarks were run using Hyperfine (Peter, 2023) with parameters "–warmup 1 –runs 3" using a Macbook Air M3 with 16 GB of memory.
*Bowtie was run with version 1.3.1 with parameters "-q -v 1 –best –strata –all –sam -p 8" to align the both guide RNAs sequentially. Intermediate file size refers to the sum of the two bam files generated by piping the output from bowtie into samtools (Danecek et al., 2021). A full pipeline built around bowtie would include additional steps for mapping the alignment results into paired guide counts and demultiplexing samples.
**pgMAP was run with the default parameters of gRNA1-error 1, gRNA2-error 1, and barcode-error 1.

Tool	Mean (Minutes)	Min (Minutes)	Max (Minutes)	Peak Memory Usage (GB)	Intermediate File(s) Size (GB)
bowtie pipeline*	72.35	71.69	72.92	0.026	28.64
pgMAP**	22.93	22.22	24.09	0.157	N/A

Description of Algorithm

pgMAP's counting algorithm takes advantage of the structure of paired CRISPR guide libraries. Paired guide libraries are composed of pairs of guide RNA sequences. While the number of unique guide RNA pairs can be in the tens of thousands, there are orders of magnitude fewer unique guide sequences. Therefore, it is possible to directly hash all the guide sequences and efficiently map to them while reading from sequencing data. To support error tolerance, pgMAP hashes every possible error perturbed sequence within an error tolerance of each reference guide sequence. With the suggested error tolerance of a single substitution, pgMAP uses less than 200mb of memory when applied to data generated using the pgPEN CRISPR library, which contains ~33,000 pgRNAs (Parrish et al., 2021).

Compared to an approach using a general purpose aligner like bowtie (Langmead & Salzberg, 2012), there is no need to demultiplex individual samples before aligning. Additionally, no intermediate alignment files are generated by pgMAP and no pipeline of bioinformatics tools is necessary to orchestrate individual steps. pgMAP is implemented with Python and depends only on the Levenshtein Python package for string distance calculations (Necas et al., 2010).

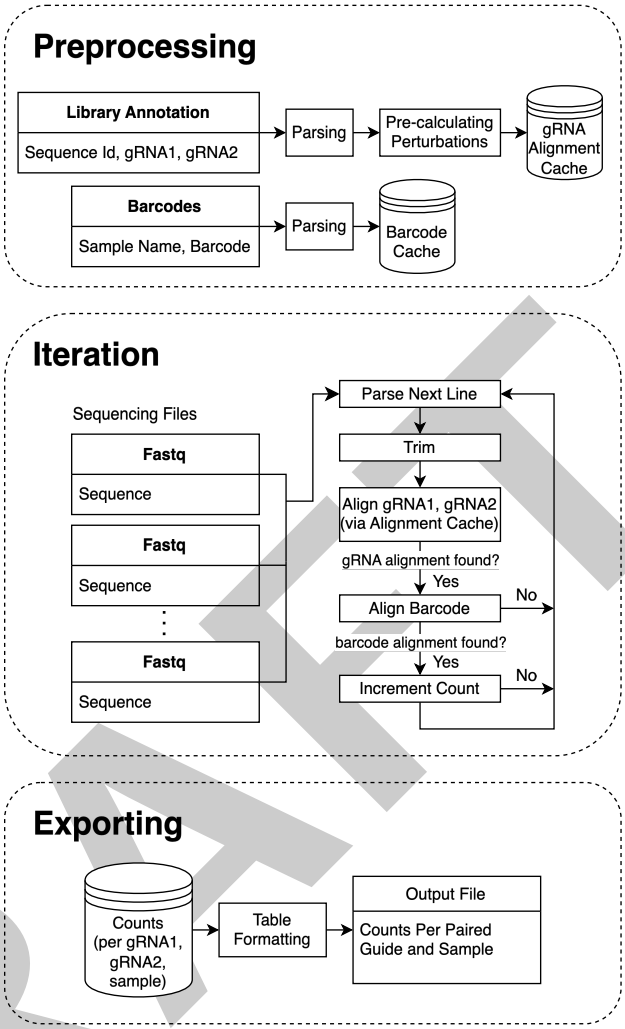

Figure 1: Overview of the pgMAP algorithm. pgMAP first computes and caches every guide RNA (gRNA) in the library annotation along with every tolerated perturbation of each guide. Then the algorithm steps through each sequencing read and trimming and hashes each guide candidate against the reference alignment cache to count guides. Barcode candidates are counted by directly checking each Levenshtein distance to the reference sample barcodes. Finally, the resulting paired guide and sample mappings are exported to a file for subsequent analysis by the user.

Research Using pgMAP

pgMAP is currently being utilized in ongoing research projects in the Berger lab, including follow-up work to previous paired gRNA CRISPR screens (Parrish et al., 2021).

Acknowledgements

M. Fujimoto was supported by the NIH award R25HG012337.

S. O'Brien is a Washington Research Foundation postdoctoral fellow.

This work was supported in part by NCI R01CA262556, the American Cancer Society, and a pilot award from the Fred Hutch Translational Data Science Integrated Research Center.

References

Dandage, R., & Landry, C. R. (2021). Identifying features of genome evolution to exploit cancer vulnerabilities. *Cell Syst.*, *12*(12), 1127–1130. https://doi.org/10.1016/j.cels.2021.08.007

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2). https://doi.org/10.1093/gigascience/giab008

Dede, M., McLaughlin, M., Kim, E., & Hart, T. (2020). Multiplex enCas12a screens detect functional buffering among paralogs otherwise masked in monogenic Cas9 knockout screens. *Genome Biol.*, *21*(1), 262. https://doi.org/10.1186/s13059-020-02173-2

Gonatopoulos-Pournatzis, T., Aregger, M., Brown, K. R., Farhangmehr, S., Braunschweig, U., Ward, H. N., Ha, K. C. H., Weiss, A., Billmann, M., Durbic, T., Myers, C. L., Blencowe, B. J., & Moffat, J. (2020). Genetic interaction mapping and exon-resolution functional genomics with a hybrid Cas9-Cas12a platform. *Nat. Biotechnol.*, *38*(5), 638–648. https://doi.org/10.1038/s41587-020-0437-z

Ito, T., Young, M. J., Li, R., Jain, S., Wernitznig, A., Krill-Burger, J. M., Lemke, C. T., Monducci, D., Rodriguez, D. J., Chang, L., Dutta, S., Pal, D., Paolella, B. R., Rothberg, M. V., Root, D. E., Johannessen, C. M., Parida, L., Getz, G., Vazquez, F., … Sellers, W. R. (2021). Paralog knockout profiling identifies DUSP4 and DUSP6 as a digenic dependence in MAPK pathway-driven cancers. *Nat. Genet.*, *53*(12), 1664–1672. https://doi.org/10.1038/s41588-021-00967-z

Köferle, A., Schlattl, A., Hörmann, A., Thatikonda, V., Popa, A., Spreitzer, F., Ravichandran, M. C., Supper, V., Oberndorfer, S., Puchner, T., Wieshofer, C., Corcokovic, M., Reiser, C., Wöhrle, S., Popow, J., Pearson, M., Martinez, J., Weitzer, S., Mair, B., & Neumüller, R. A. (2022). Interrogation of cancer gene dependencies reveals paralog interactions of autosome and sex chromosome-encoded genes. *Cell Rep.*, *39*(2), 110636. https://doi.org/10.1016/j.celrep.2022.110636

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nat. Methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Necas, D., Ohthamaa, M., Haapala, A., & Bachmann, M. (2010). Levenshtein: Python extension computing string distances and similarities. In *GitHub repository*. GitHub. https://github.com/rapidfuzz/Levenshtein

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 Rep.*, *36*(9), 109597. https://doi.org/10.1016/j.celrep.2021.109597

Peter, D. (2023). *hyperfine* (Version 1.16.1). https://github.com/sharkdp/hyperfine

Ryan, C. J., Mehta, I., Kebabci, N., & Adams, D. J. (2023). Targeting synthetic lethal paralogs in cancer. *Trends Cancer*, *9*(5), 397–409. https://doi.org/10.1016/j.trecan.2023.02.002

Savonen, C., Parrish, P., Isaac, K., Baek, H., Grosso, D., O'Brien, S., Fujimoto, M., & Berger, A. (2025). *Gimap: Calculate genetic interactions for paired CRISPR targets*. The R Foundation. https://doi.org/10.32614/CRAN.package.gimap

Tang, S., Wu, X., Liu, J., Zhang, Q., Wang, X., Shao, S., Gokbag, B., Fan, K., Liu, X., Li, F., Cheng, L., & Li, L. (2022). Generation of dual-gRNA library for combinatorial CRISPR screening of synthetic lethal gene pairs. *STAR Protoc.*, *3*(3), 101556. https://doi.org/10.1016/j.xpro.2022.101556

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. *Nat. Commun.*, *12*(1), 1302. https://doi.org/10.1038/s41467-021-21478-9