Guide assignment in single-cell CRISPR screens using crispat

Jana M. Braunger[1], Britta Velten[1,2]

[1] Heidelberg University, Centre for Organismal Studies, Heidelberg, Germany
[2] Heidelberg University, Interdisciplinary Center for Scientific Computing, Heidelberg, Germany

Summary

Pooled single cell CRISPR screens have emerged as a powerful tool in functional genomics to probe the effect of genetic interventions at scale. A crucial step in the analysis of the resulting data is the assignment of cells to gRNAs corresponding to a specific genetic intervention. However, this step is challenging due to a lack of systematic benchmarks and accessible software to apply and compare different guide assignment strategies. To address this, we here propose crispat (CRISPR guide assignment tool), a Python package to facilitate the choice of a suitable guide assignment strategy for single cell CRISPR screens.

We demonstrate the package on four single cell CRISPR interference screens at low multiplicity of infection from two studies, where crispat identifies strong differences in the number of assigned cells, downregulation of the target genes and number of discoveries across different guide assignment strategies, highlighting the need for a suitable guide assignment strategy to obtain optimal power in single cell CRISPR screens.

Availability and Implementation

crispat is implemented in python and the source code, installation instructions and tutorials can be found at https://github.com/velten-group/crispat. Code to reproduce all findings in this paper is available at https://github.com/velten-group/crispat_analysis.

Contact

Britta Velten, britta.velten@cos.uni-heidelberg.de

Introduction

Pooled single-cell CRISPR screens are a powerful tool for systematically gaining new insights into the functional consequences of genetic perturbations in high-throughput analyses. In such screens guide RNAs (gRNAs) are introduced into cells expressing Cas proteins and can direct the Cas proteins to a target site in the DNA that can be altered, suppressed or activated[1]. Single-cell sequencing then allows direct read-outs of molecular changes induced in each cell as well as markers linked to the gRNAs that are present in the cell, e.g. using barcodes or direct guide capture[2,3].

To extract biological insights about gene regulatory mechanisms from single-cell CRISPR screens a first step in the computational analysis of the resulting data is guide assignment. Hereby, cells are assigned to specific gRNAs and corresponding genetic perturbations based on gRNA counts in each cell. This step can be challenging due to background contamination and sparsity of single cell measurements and the confidence of the assignment typically varies across cells. Suitable thresholds on the measured gRNA counts per cell have to be found to distinguish background noise from actual guide presence. Currently, a wide range of different guide assignment strategies and thresholds are used[4–13] without guidance on what strategy to choose when. A systematic comparison of multiple strategies is hindered due to the lack of easily accessible software that could facilitate the application and comparison of different guide assignment strategies.

Here, we address this by developing crispat (CRISPR guide Assignment Tool), a well-documented Python package that implements 11 different guide assignment methods and facilitates their comparison. The package includes commonly used methods as well as 5 novel approaches with a focus on low multiplicity of infection (MOI) CRISPR screens, where a majority of transduced cells contains a single gRNA. We compared the performance of the methods in CRISPR interference (CRISPRi) screens with single-cell gene expression read-outs based on the number of uniquely assigned cells, the target gene downregulation in assigned cells, as well as the number of discoveries.

Methods

The starting point for the python package crispat are count matrices of the gRNA measurements in each cell obtained after aligning the sequenced gene and gRNA reads, e.g. using 10X Genomics Cell Ranger[12] or kallisto[14]. Different file formats are supported by crispat, including file formats generated by the widely used Cell Ranger platform or a csv file containing the guide counts for every gRNA and cell. Based on this data, crispat assigns every cell to one or multiple gRNAs, providing users with a choice of 11 different guide assignment methods. These consist of literature-based methods as well as extensions thereof and can be grouped into 4 categories based on the information which is used for the assignment (Fig. 1A). The first group of methods (*independent*) relies only on the individual measurement for each gRNA and cell, by checking whether an entry is higher than a user-defined threshold. While this is a simple strategy, it is widely used in applications with different thresholds[4,7–9]. A second group of methods (*across gRNAs*), considers information

for all gRNAs in a cell, e.g. by using the relative frequency of a gRNA in a cell for the assignment. Similarly to the first group of methods, a suitable threshold needs to be found. A third group of methods (*across cells*) determines a suitable threshold for every individual gRNA by comparing its values across all cells, e.g. using a two-component mixture model to distinguish cells with background counts from transduced cells containing the gRNA. Examples include a Gaussian-Gaussian mixture model employed by 10X Genomics Cell Ranger[12] and a Poisson-Gaussian mixture model proposed by Replogle et al.[5]. A last group of methods (*across cells and gRNAs*) in a similar manner uses the heterogeneity across cells, e.g. to determine suitable thresholds based on mixture models per gRNA, but in addition takes into account the total amount of gRNA counts per cell. An overview of all 11 assignment methods is shown in Supplementary Table 1 and method details can be found in the Supplementary Information and the crispat documentation. All methods relying on mixture models were implemented using pyro[15].

To compare different guide assignment methods using crispat we focused on three criteria: (i) number of cells with single gRNA assigned, (ii) target gene downregulation in the assigned cells, and (iii) number of discoveries defined as differentially expressed genes in cells assigned to a specific perturbation compared to cells assigned to a non-targeting gRNA (control cells) (Fig. 1A). To this end, crispat can be used to compare the number of assigned cells across methods as well as the overall agreement of all assignments between different methods. In addition, the assignments can serve as direct input to differential expression tests, e.g. as demonstrated in the tutorials for scanpy[16] and the R package SCEPTRE[17].

Results

To demonstrate the use of crispat, we applied it to a large-scale single-cell CRISPR interference screen, where 686,464 K562 cells were transduced with gRNAs targeting 2,285 different genes[5]. This enabled us to conduct a systematic comparison of the 11 methods implemented in crispat for guide assignment and evaluate the impact of the method choice on the power of the screen. For methods that require a user-defined threshold (Supp. Table 1) we additionally tested and compared a range of different thresholds (Supp. Fig. 1).

Overall, we observed a high consistency in the guide assignments across methods for cells assigned by multiple methods (Fig. 1B). However, the number of cells with a single gRNA assigned varied strongly from less than 150,000 to more than 600,000 cells (Fig. 1B) with 5,000 to 20,000 cells assigned to non-targeting control gRNAs (control cells) and a median of 100-200 cells per targeting gRNA (Supp. Fig. 2). For every assignment we then quantified differentially expressed genes in cells assigned to a targeting gRNA compared to control cells - also called discoveries in the following - and evaluated the target gene downregulation as well as the number of total and false discoveries for every assignment method (Supplementary Methods). In general, methods with a smaller number of assigned cells tended to have a stronger target gene downregulation (Fig. 1C), indicating a more restrictive assignment of only high-confidence cells. However, methods with a high number of

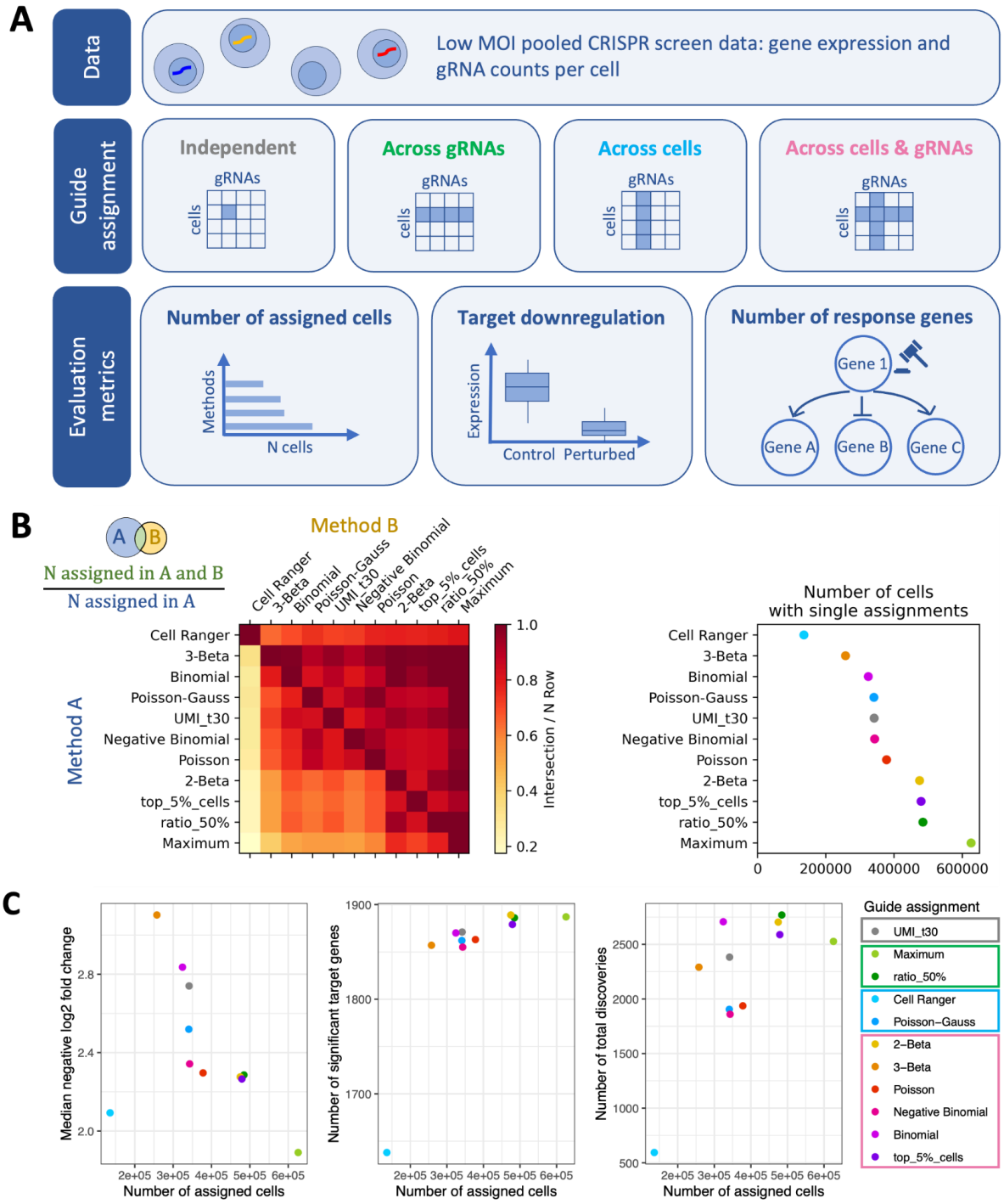

Figure 1 Comparison of guide assignment methods on a CRISPRi screen in K562 cells.
A Overview of the guide assignment comparison using crispat
B Assignment similarity heatmap showing the number of shared gRNA-cell assignments of two methods divided by the number of assignments in method A (left) and number of assigned cells in method A (right). Only cells with a single gRNA assignment are included here.
C Median negative log2 fold change of the target genes, number of significant targets and number of total discoveries (y-axis) vs. the number of assigned cells per method (x-axis). The number of total discoveries was calculated for a subset of 40 selected gRNAs and 5000 response genes (see Supplementary Information for details).

assigned cells nevertheless identified a larger number of target genes as significantly differentially expressed given the larger sample sizes for the testing (Fig. 1C). When inspecting the number of discoveries beyond the target gene we observed a similar positive relationship between the number of assigned cells and the number of discoveries (Fig. 1C). For methods with user-defined thresholds a similar trade-off between sample size and effect size could be observed, with more lenient thresholds resulting in a larger number of assigned cells but a lower median target downregulation in the assigned cells (Supp. Fig. 1). Overall, the number of discoveries varied between 594 and 2,769 depending on the choice of the guide assignment method. While mostly driven by the difference in the number of assigned cells, we also noted differences of more than 800 discoveries between methods with similar cell counts. All methods identified at most two false discoveries for non-targeting control gRNAs (out of 86,880-104,760 tested associations), and seemed to be similarly well calibrated on this data set (Supp. Fig. 3).

A notable outlier in the method comparison was the Cell Ranger assignment based on a Gaussian mixture model, which despite very few cells with a unique guide assignment showed a very low target downregulation in the assigned cells (Fig. 1B,C). As this method is routinely applied when using Cell Ranger for data processing we investigated what is causing the bad performance. We noted that the thresholds inferred by this model were very low for all gRNAs, leading to more than 600,000 cells with at least one assigned gRNA and a majority of cells having more than one gRNA assigned (Supp. Fig. 4A). This is in contrast to the expectations based on the experimental MOI of the screen and typical downstream analyses focus on cells assigned to a single gRNA. When including cells with multiple gRNAs assigned for differential expression testing, the Gaussian mixture model resulted in the highest number of discoveries of all assignment methods (Supp. Fig. 4B) but with an excessive number of false discoveries (n=3,123) (Supp. Fig. 4C). These results indicate that the assignment by the Gaussian mixture model is too lenient, with very low thresholds not being able to discern background contamination. We hypothesized that this problem is especially severe for CRISPR screens, where a high number of different gRNAs results in many cells having zero values for a given gRNA being included in the mixture model. To verify this we re-evaluated the method on semi-synthetic guide counts reducing the number of gRNAs in the Replogle data by aggregating different gRNAs into a smaller set of 500 or 86 gRNA groups, resulting in a performance of the Gaussian mixture model that was comparable to other methods (Supp. Fig. 4D,E). This suggests that the method requires adaptations for CRISPR screens with many gRNAs. We observed that a Gaussian mixture model fitted only on non-zero gRNA count values can remedy this problem to some extent, resulting in a strong increase in the number of significant target genes and total discoveries and a more similar performance to the other mixture models (Supp. Fig. 4F). This is included as an option in crispat.

Another consideration for the choice of a guide assignment strategy is its run time, both for the assignment as well as downstream analyses. For the guide assignment, the Poisson mixture model as well as an adaptation thereof using a negative binomial distribution had the longest run time of up to 170 hours without parallelization (Supp. Fig. 5A). It is important to note here however that the implementation in crispat allows parallelization across gRNAs with a single task run time of around 4.5 minutes. The run time for the downstream analyses

using the SCEPTRE package[17] was mostly driven by the number of assigned cells (Supp. Fig. 5B-D) such that methods achieving the same power using fewer cells would be advantageous in terms of both compute time and memory requirements.

Finally, we investigated to what extent we can derive general recommendations for the guide assignment strategy. For this we first used crispat on a second cell line (RPE1 cells) from the same study[5] and compared the relative performance of the guide assignment methods for the two cell lines. This revealed a very similar ordering of the methods (Supp. Fig. 6-8). However, different technologies and dimensions of the CRISPR screens are likely to influence the suitability and performance of individual methods. To test this, we used crispat on a CRISPRi screen with only 86 gRNAs and compared two different read-out strategies (whole transcriptomics and TAP-seq)[4]. In general, differences across guide assignment methods were less pronounced in these screens (Supp. Fig. 9-12) and methods performing well in the genome-scale screen were not necessarily a good choice for this data, highlighting the need for data-specific comparison of different guide assignment methods.

Conclusion

Guide assignment is a crucial step in single-cell CRISPR screens since downstream analyses rely on the assigned gRNA labels. As shown in this study, the number of assigned cells as well as the number of total discoveries can vary by a factor of 4 depending on the chosen assignment method. Therefore, we created the python package crispat, which facilitates an easy comparison of 11 different assignment methods, and enables users to choose a suitable method based on precision and recall of the different approaches.

Acknowledgements

We thank Timothy Barry for valuable discussions about the SCEPTRE package, as well as Joseph Replogle for providing additional data from their analysis. This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 540147573. The authors acknowledge support by the state of Baden-Württemberg through bwHPC and the German Research Foundation (DFG) through grant INST 35/1597-1 FUGG, as well as the data storage service SDS@hd supported by the Ministry of Science, Research and the Arts Baden-Württemberg (MWK) and the German Research Foundation (DFG) through grant INST 35/1503-1 FUGG.

Conflict of interest

The authors have no competing interests.

Contributions

JB implemented the methods, analyzed the data and generated the figures
JB and BV wrote the manuscript
BV conceived and supervised the project

References

1. Bock, C., Datlinger, P., Chardon, F., Coelho, M.A., Dong, M.B., Lawson, K.A., Lu, T., Maroc, L., Norman, T.M., Song, B., et al. (2022). High-content CRISPR screening. Nat. Rev. Methods Primer *2*, 1–23. 10.1038/s43586-021-00093-4.
2. Schraivogel, D., Steinmetz, L.M., and Parts, L. (2023). Pooled Genome-Scale CRISPR Screens in Single Cells. Annu. Rev. Genet. *57*, annurev-genet-072920-013842. 10.1146/annurev-genet-072920-013842.
3. Cheng, J., Lin, G., Wang, T., Wang, Y., Guo, W., Liao, J., Yang, P., Chen, J., Shao, X., Lu, X., et al. (2023). Massively Parallel CRISPR-Based Genetic Perturbation Screening at Single-Cell Resolution. Adv. Sci. *10*, 2204484. 10.1002/advs.202204484.
4. Schraivogel, D., Gschwind, A.R., Milbank, J.H., Leonce, D.R., Jakob, P., Mathur, L., Korbel, J.O., Merten, C.A., Velten, L., and Steinmetz, L.M. (2020). Targeted Perturb-seq enables genome-scale genetic screens in single cells. Nat. Methods *17*, 629–635. 10.1038/s41592-020-0837-5.
5. Replogle, J.M., Saunders, R.A., Pogson, A.N., Hussmann, J.A., Lenail, A., Guna, A., Mascibroda, L., Wagner, E.J., Adelman, K., Lithwick-Yanai, G., et al. (2022). Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq. Cell *185*, 2559-2575.e28. 10.1016/j.cell.2022.05.013.
6. Datlinger, P., Rendeiro, A.F., Schmidl, C., Krausgruber, T., Traxler, P., Klughammer, J., Schuster, L.C., Kuchler, A., Alpar, D., and Bock, C. (2017). Pooled CRISPR screening with single-cell transcriptome readout. Nat. Methods *14*, 297–301. 10.1038/nmeth.4177.
7. Frangieh, C.J., Melms, J.C., Thakore, P.I., Geiger-Schuller, K.R., Ho, P., Luoma, A.M., Cleary, B., Jerby-Arnon, L., Malu, S., Cuoco, M.S., et al. (2021). Multimodal pooled Perturb-CITE-seq screens in patient models define mechanisms of cancer immune evasion. Nat. Genet. *53*, 332–341. 10.1038/s41588-021-00779-1.
8. Geiger-Schuller, K., Eraslan, B., Kuksenko, O., Dey, K.K., Jagadeesh, K.A., Thakore, P.I., Karayel, O., Yung, A.R., Rajagopalan, A., Meireles, A.M., et al. (2023). Systematically characterizing the roles of E3-ligase family members in inflammatory responses with massively parallel Perturb-seq. Preprint at bioRxiv, 10.1101/2023.01.23.525198 10.1101/2023.01.23.525198.
9. Adamson, B., Norman, T.M., Jost, M., Cho, M.Y., Nuñez, J.K., Chen, Y., Villalta, J.E., Gilbert, L.A., Horlbeck, M.A., Hein, M.Y., et al. (2016). A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response. Cell *167*, 1867-1882.e21. 10.1016/j.cell.2016.11.048.
10. Papalexi, E., Mimitou, E.P., Butler, A.W., Foster, S., Bracken, B., Mauck, W.M., Wessels, H.-H., Hao, Y., Yeung, B.Z., Smibert, P., et al. (2021). Characterizing the molecular regulation of inhibitory immune checkpoints with multimodal single-cell screens. Nat. Genet. *53*, 322–331. 10.1038/s41588-021-00778-2.
11. Jiang, L., Dalgarno, C., Papalexi, E., Mascio, I., Wessels, H.-H., Yun, H., Iremadze, N., Lithwick-Yanai, G., Lipson, D., and Satija, R. (2024). Systematic reconstruction of molecular pathway signatures using scalable single-cell perturbation screens. Preprint at bioRxiv, 10.1101/2024.01.29.576933 10.1101/2024.01.29.576933.
12. Zheng, G.X., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., and Zhu, J. (2017). Massively parallel digital transcriptional profiling of single cells. Nat. Commun. *8*, 14049.
13. Barry, T., Roeder, K., and Katsevich, E. (2024). Exponential family measurement error models for single-cell CRISPR screens. Biostatistics, kxae010. 10.1093/biostatistics/kxae010.
14. Melsted, P., Booeshaghi, A.S., Liu, L., Gao, F., Lu, L., Min, K.H. (Joseph), da Veiga Beltrame, E., Hjörleifsson, K.E., Gehring, J., and Pachter, L. (2021). Modular, efficient and constant-memory single-cell RNA-seq preprocessing. Nat. Biotechnol. *39*, 813–818. 10.1038/s41587-021-00870-2.
15. Bingham, E., Chen, J.P., Jankowiak, M., Obermeyer, F., Pradhan, N., Karaletsos, T.,

Singh, R., Szerlip, P., Horsfall, P., and Goodman, N.D. (2019). Pyro: Deep Universal Probabilistic Programming. J. Mach. Learn. Res. *20*, 1–6.
16. Wolf, F.A., Angerer, P., and Theis, F.J. (2018). SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. *19*, 15. 10.1186/s13059-017-1382-0.
17. Barry, T., Mason, K., Roeder, K., and Katsevich, E. (2023). Robust differential expression testing for single-cell CRISPR screens at low multiplicity of infection. Preprint at bioRxiv, 10.1101/2023.05.15.540875 10.1101/2023.05.15.540875.