UMIBB: A novel nonparametric Bayesian method improves robustness and sensitivity of analysis in pooled CRISPR-Cas9 screens leveraging unique molecular identifiers

1224

Ashley Choi, Samuel Meier, Silvia Fenoglio, Tianshu Feng, Justin Engel, Binzhang Shen, Shangtao Liu, Teng Teng, Tenzing Khendu, Alan Huang, Jannik Andersen, Xuewen Pan, Yi Yu.

ABSTRACT

CRISPR-based functional genomics screens are a powerful tool for identifying synthetically lethal cancer drug targets. Current strategies for analyzing pooled CRISPR screens usually rely on signals from single guide RNAs (sgRNA) with differential relative abundance between two experimental conditions. However, conventional approaches are often susceptible to false positives and false negatives driven by outlier cell clones, since the sgRNA abundance does not account for the heterogenous phenotypes resulting from different editing outcomes of the same sgRNA. To overcome this, we added DNA barcodes to each sgRNA to create unique molecular identifiers (UMIs) for CRISPR libraries and developed a companion analytical platform that enables robust, industry-scale CRISPR screens. Here, we present UMIBB, a novel nonparametric Bayesian approach for analyzing UMI-CRISPR data. The number of UMIs with normalized count depletion or enrichment compared to the control experimental condition for each sgRNA is modeled by a beta-binomial distribution. The gene level statistics is derived by combining z-scores of the sgRNAs level posterior probabilities weighted by the number of UMIs in each sgRNA. This approach minimizes the impact of outlier cell clones on statistics and prioritize genes with consistent count differentials across multiple UMIs in each gene. To assess the power of UMIBB, we benchmarked it on a low coverage (200X) genome-scale negative-selection screen, comparing with results from a high coverage (1000X) screen. These screens were conducted on KRAS mutant cancer cells (A549) treated with trametinib or vehicle control. Despite the high noise level usually observed in lower coverage screens, our method was able to uncover >85% of the validated sensitizer genes for trametinib and achieved the highest sensitivity compared to conventional methods. Furthermore, we applied UMIBB on a genome-scale positive-selection screen and successfully identified novel genes (RAD18 and UBE2K) as key mediators of USP1 dependency in BRCA1/2 mutant cell lines. Our studies demonstrate that UMIBB is highly robust against false positives due to clonal heterogeneity and is more likely to identify true genetic interactions.

INTRODUCTION

Pooled CRISPR-UMI is a more powerful screening platform for assessing gene function than traditional CRISPR screens with only sgRNA abundance data [1]. However, no computational tools for specifically analyzing CRISPR-UMI data have been reported, except applying a modified MAGeCK [2] analysis to model the UMI clonal abundance instead of sgRNA abundance. This approach requires pre-determined parameters, such as the number of outlier clones, which may not be practical in real screens since the number of outlier clones varies across different genes and sgRNAs. To solve this issue, we have developed an industry-scale CRISPR-UMI screening and bioinformatics analysis platform. Here, we have generated a CRISPR-UMI library by combining a genome-wide CRISPR library with 10,000 unique barcodes and present a novel non-parametric Bayesian approach for analyzing pooled CRISPR-UMI screening data.

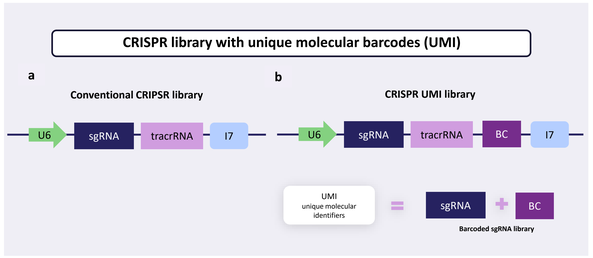


Figure 1: Library design. a) conventional library vector design. b) UMI library vector design. Each sgRNA-barcode pair make unique molecular identifiers (UMIs).

METHODS

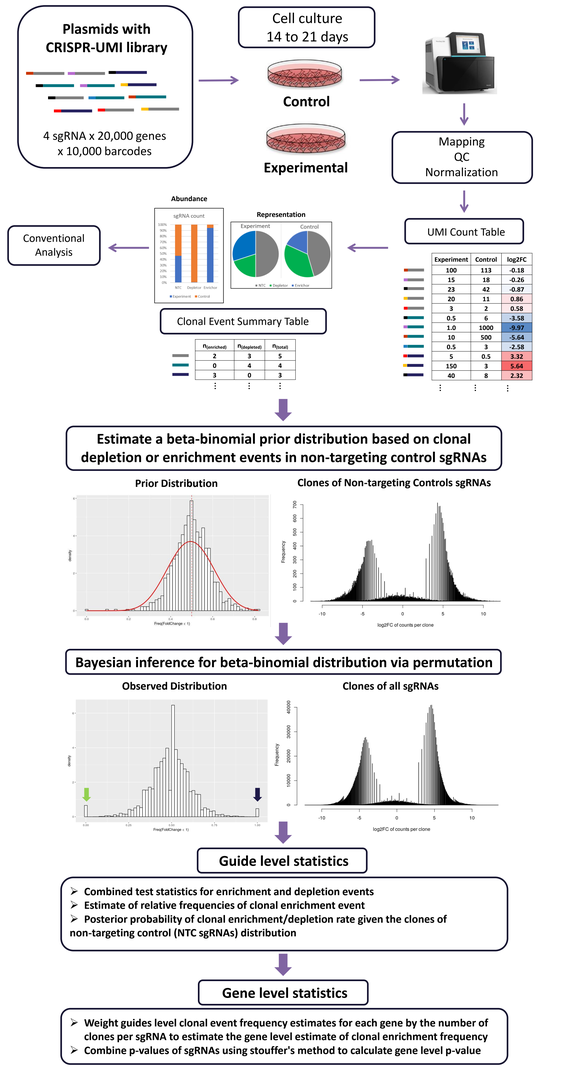
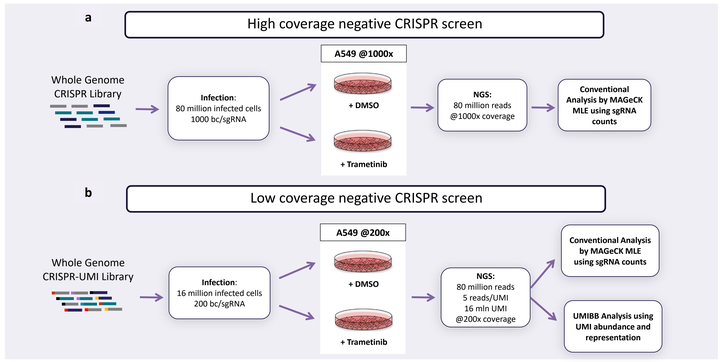


Figure 2: Analytical flow chart of UMIBB test.



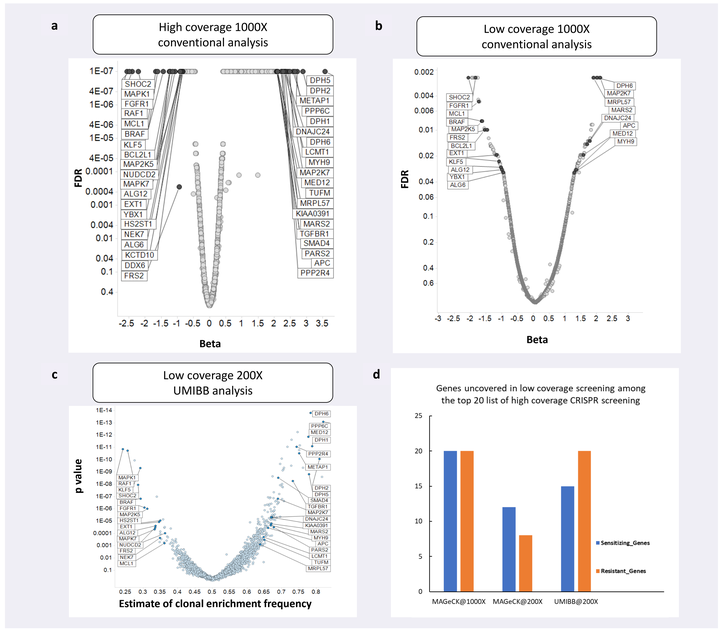


Figure 3: Scheme of CRISPR screenings of A549 treated with 30nM trametinib or DMSO control. a) Conventional analysis on high coverage screening data; b) UMIBB and conventional analysis performed on low coverage screening data.

Figure 4: Comparison of top statistically significant genes in high coverage (1000X) and low coverage (200X) screening analysed by conventional methods (MAGeCK) or UMIBB . a) Top 20 genes (Dark dots) showing strong evidence of sensitizing or resistant effect to trametinib treatment; b) Conventional analysis discovered fewer target genes (Dark dots) with diminishing statistical power in the lower coverage screening; c) UMIBB method uncovers >85% of target genes (Bright blue dots) with highly significant evidence in the low coverage screening; d) Number of target genes discovered using conventional and UMIBB methods in low coverage screening among the top 20 sensitizing or resistant genes identified from the high coverage screening.

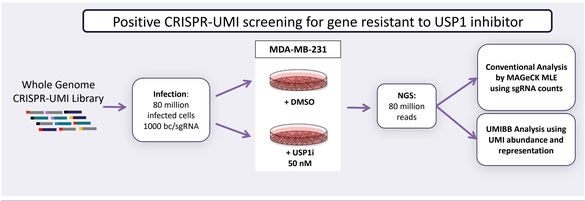


Figure 5: Scheme of CRISPR screenings of MDA-MB-436 (*BRCA1/2* mutant) cell line treated with 50 nM USP1i or DMSO control.

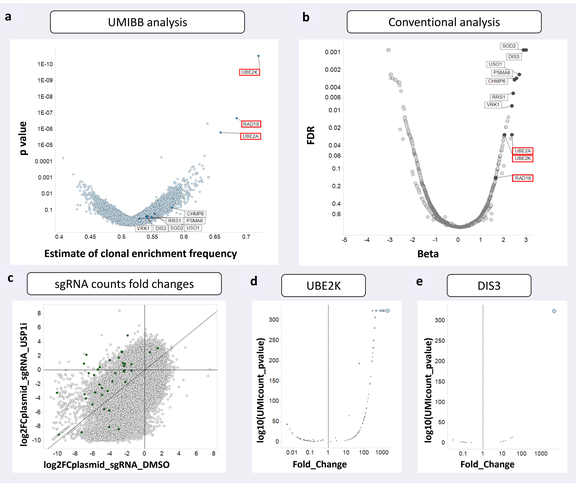


Figure 6: Conventional and UMIBB analysis of USP1 inhibitor screens. Volcano plots of (a) UMIBB analysis shows the true positive rescuer genes (labelled with red squares) for USP1 inhibitor, and false positives (labelled with black squares) which ranked on top by (b) conventional MAGeCK analysis. (c) True or false positive genes are not distinguishable with only sgRNA count fold change data. (d, e) UMI count and fold changes show the difference between a true positive (d) gene with enrichment evidence corroborated by many UMI clones, and a false positive (e) gene whose statistic is only driven by an outlier UMI clone with high count.

CONCLUSION / SUMMARY

- We report here a non-parametric Bayesian method (UMIBB) as an analytical tool for the CRISPR-UMI based target discovery screening and have tested the approach using our whole-genome CRISPR-UMI screening platform
- This UMIBB approach improves screening sensitivity while remaining robust to false positives due to low coverage or unintended noise
- This method may facilitate novel drug target identification with both *in vitro* and *in vivo* CRISPR-UMI screens

REFERENCES

1. Michlits, *et al.* CRISPR-UMI: single-cell lineage tracing of pooled CRISPR–Cas9 screens. *Nat Methods* **14,** 1191–1197 (2017).
2. Li, et al. MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. *Genome Biology* **15,** 554 (2014)