A hierarchical mixture model for CRISPR pooled screens

Timothy Daley¹,², Xueqiu Lin¹, Yanxia Liu¹, Zhixiang Lin², Lei S. Qi¹,³,⁴,⁶,* and Wing Hung Wong²,⁵,*

¹Department of Bioengineering, Stanford University
²Department of Statistics, Stanford University
³Chemical and Systems Biology, Stanford University
⁴ChEM-H Institute, Stanford University
⁵Department of Biomedical Data Science, Stanford University
⁶To whom correspondence should be addressed to: tdaley@stanford.edu, stanley.qi@stanford.edu, and whwong@stanford.edu

*These authors jointly supervised this work
**Supplementary Figures**

**Figure S1:** DESeq2 normalized log2 fold changes for known essential genes, non-essential genes, and negative control guides in Evers *et al.* 2016 and Rosenbluh *et al.* 2017 in **a** CRISPRko and **b** CRISPRi screens.
Figure S2: ROC curves for our hierarchical normal mixture, MAGeCK RRA, and MAGeCK MLE on data from Evers et al. 2016.
Figure S3: Normal fit to DESeq2 normalized log2 fold change of the negative control guides of a a CRISPRi screen for ricin phenotypes (Gilbert et al. 2014); b a CRISPRi screen for functional lncRNAs in HeLa cells (Liu et al. 2017); and c a CRISPRi screen and d a CRISPRa screen for rigosertib phenotypes (Jost et al. 2017).
Figure S4: Skew-$t$ fit to DESeq2 normalized log2 fold change of the negative control guides of a CRISPRi screen for ricin phenotypes (Gilbert et al. 2014); b a CRISPRi screen for functional lncRNAs in HeLa cells (Liu et al. 2017); and c a CRISPRi screen and d a CRISPRa screen for rigosertib phenotypes (Jost et al. 2017).
Figure S5: Normal fit to DESeq2 normalized log2 fold change of the negative control guides of CRISPRko screens of a genome-wide screen for essential genes in K562 cells from Morgens et al. 2016, b a targeted loss of function screen in HT29 cells from Rosenbluh et al. 2017, c a genome-wide screen for essential genes in CADOES1_BONE cell line from Project Achilles (Aguirre et al. 2016), d a genome-wide screen for essential genes in EWS502_BONE cell line from Project Achilles.
**Figure S6:** Skew-$t$ fit to DESeq2 normalized log2 fold change of the negative control guides of CRISPRko screens of a genome-wide screen for essential genes in K562 cells from Morgens et al. 2016, b a targeted loss of function screen in HT29 cells from Rosenbluh et al. 2017, c a genome-wide screen for essential genes in CADOES1_BONE cell line from Project Achilles (Aguirre et al. 2016), d a genome-wide screen for essential genes in EWS502_BONE cell line from Project Achilles.
Table S1: Bayesian Information Criteria (BIC) of the normal model and the skew-$t$ fit for the negative control guides of the indicated CRISPRko experiments. A smaller BIC is preferred and is in bold for each experiment.
**Figure S7**: DESeq2 normalized log2 fold changes for positive essential genes, simulated negative genes, and negative control guides.
Figure S8: DESeq2 normalized log2 fold changes for the non-control guides, along with the mixture distribution fits.
Figure S9: The estimated local fdr values estimated by CRISPhieRmix and the p-values obtained from MAGeCK RRA and MAGeCK MLE.
**Figure S10**: DESeq2 normalized log2 fold changes for the top 18 genes, clearly identified by CRISPhieRmix, and the other essential genes. A clear separation is observed between the two groups.
Figure S11: Estimated FDR plotted against the empirical FDR for the normal mixture model when applied to the Rosenbluh CRISPRi simulated data.
Figure S12: A comparison of algorithms as a function of sgRNAs per gene in the simulated CRISPRko screen. The area under the receiver operator curves (AUC), true positive rate (TPR) at a threshold of 0.1 global false discovery rate (FDR), empirical FDR at an estimated FDR of 0.1, and running times of all algorithms are shown. The dashed line indicates an FDR of 0.1, or perfect control of the FDR.
Figure S13: CRISPheRmix fit to the $\log_2$ fold changes of the untreated population relative to the initial population.
**Figure S14**: ROC curves for known essential and non-essential genes for a CRISPRi screen from Gilbert et al. 2014.
Figure S15: Histogram of log₂ fold changes of guides corresponding to the 105 essential genes called by both MAGeCK MLE and CRISPhieRmix (left) and the 67 essential genes called only by CRISPhieRmix.
Figure S16: Gaussian kernel smoothed densities of negative control guides and two example genes. AQR was identified by both MAGeCK MLE and CRISPhieRmix, while ALYREF was identified by only CRISPhieRmix. Both genes are first in alphabetical order in their respective classes.
Figure S17: Density plots of ESeq2 calculated log2 fold changes for negative control guides and gene targeting guides.
Figure S18: Evaluation of CRISPhieRmix, MAGeCK RRA, MAGeCK MLE, and the Mann-Whitney test on the 6 base library experiments of Hart et al. 2015 using the reference sets of Hart et al. 2014.
**Figure S19:** Example empirical densities of $\log_2$ fold changes for simulated gene targeting guides and negative control guides with varying guide efficiencies, indicated above each plot. Details on simulation are available in the main text.
Figure S20: The true positive rate (the percentage of true interesting genes correctly identified) by the Mann-Whitney test at an Benjamani-Hochberg corrected FDR of 0.1 (left) and ROC-AUC (right) for simulated data with varying percent of effective sgRNAs and number of sgRNAs per gene.