Title

On the design of CRISPR-based single cell molecular screens

Authors

Andrew J. Hill[1,*], José L. McFaline-Figueroa[1,*], Lea M. Starita[1], Molly J. Gasperini[1], Kenneth A. Matreyek[1], Jonathan Packer[1], Dana Jackson[1], Jay Shendure[1,2,†], Cole Trapnell[1,†]

[1] Department of Genome Sciences, University of Washington, Seattle, WA, USA
[2] Howard Hughes Medical Institute, Seattle, WA, USA

* These authors contributed equally to this work
[†] These authors contributed equally to this work

Correspondence to coletrap@uw.edu (CT) & shendure@uw.edu (JS)

Abstract

Several groups recently reported coupling CRISPR/Cas9 perturbations and single cell RNA-seq as a potentially powerful approach for forward genetics. Here we demonstrate that vector designs for such screens that rely on *cis* linkage of guides and distally located barcodes suffer from swapping of intended guide-barcode associations at rates approaching 50% due to template switching during lentivirus production, greatly reducing sensitivity. We optimize a published strategy, CROP-seq, that instead uses a Pol II transcribed copy of the sgRNA sequence itself, doubling the rate at which guides are assigned to cells to 94%. We confirm this strategy performs robustly and further explore experimental best practices for CRISPR/Cas9-based single cell molecular screens.

Introduction

Forward genetic screens in cell culture, based on RNA interference or CRISPR/Cas9, enable the functional characterization of thousands of programmed perturbations in a single experiment[1,2]. However, the options for phenotypic assays that are compatible with such screens are often limited to coarse phenotypes such as relative cell growth or survival, and moreover are uninformative with respect to the mechanism by which each positively scoring perturbation mediates its effect.

To circumvent these limitations, several groups recently reported using single cell RNA-seq (scRNA-seq)[3–5] as a readout for CRISPR-Cas9 forward genetic screens, *i.e.* to broadly capture phenotypic effects at the molecular level. A key aspect of these methods is that the guide RNA (sgRNA) present in each single cell is identified together with its transcriptome, either by means of a Pol II transcribed barcode that is linked *in cis* to the sgRNA (CRISP-seq, Perturb-Seq, Mosaic-seq[6–9]) (**Fig. 1a**), or alternatively by capturing the sgRNA sequence itself as part of an overlapping Pol II transcript (CROP-seq[10]) (**Fig. 1b**).

In our own efforts to develop similar methods, we have encountered several important technical challenges not yet fully delineated in the literature. First, we have quantified the impact of template switching during lentiviral packaging on designs that rely on *cis* pairing of each sgRNA with a distally located barcode, and observed swapping of guide-barcode associations at a rate approaching 50%. Second, we have coupled targeted sgRNA amplification[7,8] and the published vector of CROP-seq[10], and shown that our improved protocol is a robust design for scRNA-seq readout of CRISPR-based forward genetic screens. Finally, we have tested an attractive alternative design to CROP-seq, but find it does not result in robust inhibition or editing.

Results

We initially pursued a design similar to recently published methods[6–9] in which each sgRNA was linked to a Pol II transcribed barcode located several kilobases away on the lentiviral construct (**Fig. 1a**). In our vector (pLGB-scKO), the barcode was positioned in the 3' UTR of a blasticidin resistance transgene, such that it could be recovered by scRNA-seq methods that prime off of poly(A) tails (**Fig. S1a-b**). Guides and barcodes were paired during DNA synthesis, which facilitated pooled cloning and lentiviral delivery (**Fig. S1c**).

With this design, we sought to ask how loss-of-function (LoF) of various tumor suppressors altered the transcriptional landscape of immortalized, non-transformed breast epithelial cells[11]. Specifically, we targeted *TP53* and other tumor suppressors in the MCF10A cell line, with or without exposure to doxorubicin, which induces double-strand breaks (DSBs) and a transcriptional response to DNA damage. Cloning and lentiviral packaging was either performed individually for each targeted gene, or in a pooled fashion. In addition to scRNA-seq, we performed targeted amplification to more efficiently recover the sgRNA-linked barcodes present in each cell (**Fig. S1b**; **Fig. S2**).

In a first experiment in which a small number of lentiviral constructs were cloned and packaged separately for each gene ('arrayed lentiviral production'), a substantial proportion of cells in which *TP53* was targeted had a gene expression signature consistent with a failure to activate a cell cycle checkpoint response after DNA damage, in line with *TP53*'s pivotal role in this pathway (*e.g.* lower expression of *CDKN1A* and *TP53I3*; **Fig. S3a**). However, these effects were greatly reduced when we performed a similar experiment with pooled cloning and lentiviral packaging (**Fig. S3b**). Furthermore, markedly fewer genes were differentially expressed across the targets in the pooled experiment than in the arrayed experiment (**Fig. S3c**).

Based on what is known about HIV, we reasoned that template switching during pooled packaging of lentivirus could result in the integration of constructs where the designed sgRNA-barcode pairings are partially randomized. During viral production, lentiviral plasmids are transfected into HEK293T cells at high copy number and transcribed[12]. Lentiviral virions are pseudodiploid, meaning that two viral transcripts are co-packaged within each virion[13]. The reverse transcriptase that performs negative strand synthesis has a remarkably high rate of template switching[14], estimated as roughly 1 event per kilobase (kb)[15]. Template switching would be expected to result in the integration of chimeric products at a rate proportional to the distance between paired sequences, effectively swapping intended sgRNA-barcode associations (**Fig. 1c**). This risk was noted by Adamson et al.[7] and Dixit et al.[8]. It was also altogether avoided by Adamson et al.[7] through arrayed lentiviral production, but pooled lentiviral production was performed in some or all experiments of the other reports[6,8,9]. Although Sack et al.[16] recently

quantified this phenomenon at distances up to 720 bp in vectors designed for bulk selection screens, the implications of template switching at longer distances (e.g., the 2.5 kb+ separation between sgRNAs and barcodes in the pLGB-scKO, CRISP-seq, Perturb-seq, and Mosaic-seq vectors), as well as for scRNA-seq study designs specifically, remain unexplored. Given a large enough distance between the barcode and the sgRNA, *cis* linkage between the two sequences would be lost in ~50% of integration events (an odd number of template switching events) and maintained in ~50% (an even number of template switching events).

To quantify the extent of swapping between two distally located sequences during lentiviral packaging, we cloned BFP and GFP transgenes, which differ by three base pairs, into separate lentiviral vectors (pHAGE-GFP and pHAGE-BFP). We paired each transgene with a unique barcode, separated from the nearest unique bases in BFP/GFP by 2.4 kb (**Fig. 1d**) to approximate the 2.5 kb or greater separation between sgRNAs and barcodes in the pLGB-scKO, CRISP-seq, Perturb-Seq, and Mosaic-seq vectors[6–9]. We then transduced MCF10A cells with lentivirus generated either individually or as a pool of the two plasmids. Finally, we sorted GFP+ or BFP+ fractions of the cells with FACS, and quantified the rate of barcode swapping (**Fig. 1e; Fig. S4**). At this distance, *cis* linkage is lost at the theoretical maximum rate of 50% (**Fig. 1f; Fig. S5**). Our observations are consistent with previous estimates of template switching in HIV[15] and recent studies in lentivirus at distances below 1 kb[16].

In order to simulate the impact of template switching, we obtained data from a pilot experiment of Adamson *et al.*[7] generated using the Peturb-seq vector with arrayed lentiviral production, targeting several transcription factors with CRISPRi. We swapped target labels *in silico* in these data at varying rates, and evaluated the impact on power to detect differential expression. At a 50% swap rate, we observe a 4.8-fold decrease in the number of differentially expressed genes (**Fig. 1g**). This loss in power results from an effective reduction of the useful sample size for each target by twofold and contamination of each target with noise from swapped associations, thus shifting all targets towards the population average of the library.

One of the published strategies for CRISPR-based single cell molecular screens, CROP-seq[10], does not rely on pairing of sgRNAs and barcodes. Instead, the sequence of the integrated sgRNA itself acts as a barcode, as part of an overlapping Pol II transcript. In addition, the sgRNA cassette is copied from the 3' to 5' LTR during positive strand synthesis (**Fig. 1b**). This copy is generated during an intramolecular priming step that does not result in intermolecular swapping at an appreciable rate[17]. A limitation of the CROP-seq method as described is that the sgRNA expressed in each cell is recovered as part of its transcriptome with limited sensitivity (~40-60%)[10]. The roughly half of single cell transcriptomes for which the sgRNA is not identified are discarded. To improve upon this, we modified the CROP-seq protocol to include

133 targeted amplification of the sgRNA region from mRNA libraries that have already been tagged
134 with cellular barcodes, as in our initial pLGB-scKO design (**Fig. 2a; Fig. S6**).
135
136 To evaluate the effectiveness of this approach, we performed a CRISPR-mediated LoF screen of
137 32 tumor suppressors (6 guides per target) and 6 non-targeting control (NTC) guides in MCF10A
138 cells with or without exposure to doxorubicin. Whereas the sgRNA was identified in the shotgun
139 transcriptome of only 42-47% of cells, it was identified in 94% of cells with targeted
140 amplification (**Fig. 2b**). In sharp contrast with our original pooled experiment that also targeted
141 TP53, a tSNE embedding of doxorubicin-exposed cells from this experiment yielded a cluster
142 that is almost entirely composed of cells containing TP53-targeting sgRNAs, highlighting the
143 unique molecular phenotype imparted by TP53 loss when responding to DSBs (**Fig. 2c**).
144 Specifically, the 262 cells in this cluster include 90.5% with TP53-targeting guides, 7.6% with
145 guides targeting other genes, 0% with NTC guides, and 1.9% that were unassigned. In contrast,
146 the remaining 5,617 cells included 3.2% with TP53-targeting guides (presumably cells in which
147 editing failed to occur, or in which editing maintained a functional protein), 84.2% with guides
148 targeting other genes, 7.5% with NTC guides, and 5.2% that were unassigned. Expression levels
149 of the p53 targets *CDKN1A* and *TP53I3*[18,19] were markedly lower in the TP53-targeted cluster
150 (**Fig. 2d**), and 4,277 and 2,186 differentially expressed genes (FDR 5%) were identified relative
151 to cells with NTC guides in the doxorubicin-treated and mock condition, respectively. The clean
152 separation between the TP53-targeted cluster and other cells is presumably a consequence of: (a)
153 the large effect of knocking out TP53; (b) our ability to recover sgRNA labels in a nearly all cells
154 via targeted amplification; (c) the lack of sgRNA/barcode swapping, because of CROP-seq; (d)
155 the high efficiency of CROP-seq expressed sgRNA in mediating nonhomologous end-joining.
156
157 Upon applying dimensionality reduction and clustering to cells in both the mock and doxorubicin
158 treated conditions (**Fig. S7a-b**), we find several other tumor suppressors whose distribution
159 across tSNE clusters is significantly different compared to NTCs (FDR 5%), with more changes
160 observed after doxorubicin exposure (**Fig. S7c-f**). To characterize clusters in which we observe
161 enrichment of particular targets, we tested for target enrichment within clusters and generated
162 average expression profiles for each enriched target-cluster pair. Gene set enrichment analysis of
163 the most highly loaded genes in the principal components of these average expression profiles
164 show many targets to be associated with increased proliferation and a decreased TP53/DNA
165 damage response, with the extent of this effect being largest for TP53 (**Fig. S8**).
166
167 To further assess the impact of lentivirus template switching would have on sensitivity, we
168 permuted target labels at varying rates within our own CROP-seq tumor suppressor screen and
169 find a 2.9-fold reduction in the number of DEGs observed across the targets at a swap rate of
170 50%. Enrichment tests on our tumor suppressor screen with a 50% simulated swap rate
171 substantially decreased the number of knockouts that display a significant change in phenotype

in both mock and doxorubicin treated conditions, recovering just 4/13 (*TP53*, *STK11*, *CHEK1* and *NCOR1*) and 3/14 (*TP53*, *RB1*, and *ARID1B*) targets as enriched in the mock and doxorubicin conditions, respectively. Additionally, swapping simulations on the much larger (50,000 cells) unfolded protein response screen from Adamson *et al.* with arrayed lentiviral production show a 1.9- and 2.8-fold reduction at a simulated swap rate of 50% when using 25,000 and 6,000 cells, respectively (**Fig. S9**). These simulations demonstrate that the effect of barcode-sgRNA pair swapping is dependent on the number of cells captured and also on sequencing depth, number of targets examined and the effect size for those targets.

Although the CROP-seq design is not subject to sgRNA-barcode swapping, it is potentially limited by its placement of the sgRNA cassette in the LTR of the lentiviral vector, as larger intervening sequences may render the LTR non-functional[10]. To enable incorporation of longer cassettes, such as dual sgRNA designs[20–22], we sought to place the sgRNA cassette between the WPRE and LTR. In this design (pHAGE-scKO), a second copy of this cassette would not be generated. However, the guide sequence would still contribute to overlapping Pol II and Pol III transcripts; for the former, it is positioned to ensure its incorporation into the 3' end of the blasticidin resistance gene (**Fig. 2e**).

To evaluate this design, we compared the ability of pHAGE-scKO, CROP-seq, and a standard lentiviral sgRNA expression vector, pKHH030[23], all containing a CRISPRi optimized backbone, to inhibit transcription via CRISPRi. We targeted the promoter of lentivirally-integrated mCherry in both MCF10A and K562 cells, which were then assayed for fluorescence via FACS. Whereas pKHH030 and CROP-seq exhibited efficient inhibition of mCherry, pHAGE-scKO had poor efficacy in both cell lines (**Fig. 2f**). Consistent with this, we observed low editing rates with our new design in MCF10A cells (88% edited with pLGB control vs. 29% edited with pHAGE-scKO). Recent studies have hinted at interference when Pol II and Pol III transcripts overlap[24,25]. We hypothesize that the observed decrease in editing and inhibition efficiency of the pHAGE-scKO design is due to the blasticidin resistance gene (Pol II promoter) inhibiting expression of the Pol III sgRNA. In contrast, CROP-seq likely maintains efficacy because the second integrated copy of the guide expression cassette (copied to the 5' LTR during positive strand synthesis) does not overlap a Pol II transcript.

Discussion

CRISPR-based forward genetic screens that rely on scRNA-seq to phenotype each perturbation have the potential to be extremely powerful. However, as we demonstrate here, there are important technical considerations that must be taken into account. Several published designs, as well as our own initial design (pLGB-scKO), suffer from high rates (50%) of swapped sgRNA-barcode associations, consequent to template switching across the several kilobases

between the sgRNA and barcode during lentiviral production. Importantly, we do not expect that positive conclusions drawn by published studies utilizing such designs in conjunction with pooled lentivirus production[6,8,9] are incorrect. Each of these studies examined a small number of targets and collected large datasets ranging from 25,000 to 100,000 cells per screen with ample sequencing depth, raising their baseline sensitivity. However, given the high cost of single-cell capture and sequencing and the need to expand the number of targets in such screens, our observation of ~50% swapping of sgRNA-barcode associations with pooled lentiviral production using vectors in which the sgRNA and barcode are separated by several Kb, are highly relevant for ongoing and future studies. This loss of power may be overcome in part by filtering cells that appear inconsistent with their assigned knock-out[8], or overcome altogether by performing cloning and lentiviral packaging separately for guides targeting each gene[7]. However, computational filtering of cells has the potential to introduce biases, and itself reduces power by discarding collected data, while performing cloning and lentiviral packaging separately for each sgRNA dramatically limits scalability.

We also explored an alternative design (pHAGE-scKO), that like CROP-seq allows sequencing of the sgRNA sequence directly, in hopes that it would facilitate the use of dual guide RNAs or other designs that require longer cassettes. However, we observe markedly reduced editing/inhibition with this design. It is plausible that methods such as programmed multiplexed guide expression cassettes[20–22,26] could be used in conjunction with CROP-seq due to their reduced length, but it will be important to carefully validate any such designs.

As the community increasingly adopts scRNA-seq as a readout for forward genetic screens, we believe that each of these technical hurdles merit careful consideration. By coupling targeted sgRNA amplification and the published CROP-seq method[10], we doubled the proportion of cells in which guides are assigned to cells, to 94%. The attractive features of this approach include the simplicity of the cloning protocol, its compatibility with lentiviral delivery, the high rate of recovery of sgRNA-cell associations, and no risk of template switching.

Acknowledgements

We thank all members of the Shendure and Trapnell labs for feedback on our manuscript and helpful discussions, particularly Sanjay Srivatsan, Greg Findlay, Aaron McKenna, Riza Daza, Beth Martin, Martin Kircher, Darren Cusanovich, Xiaojie Qiu, and Vijay Ramani. We thank Professors Jesse Bloom and Douglas Fowler for discussions about lentivirus; Dr. Kyuho Han, James Ousey, and Professor Mike Bassik for experimental advice and reagents for CRISPRi experiments. AH thanks Stella the cat for support. This work was supported by the following funding: NIH DP1HG007811 and UM1HG009408 (to JS), DP2HD088158 (to CT), and the W. M. Keck Foundation (to CT and JS). AH and MG are funded by the National Science Foundation

Graduate Research Fellowship. JLM is supported by the NIH Genome Training Grant (5T32HG000035) and the Cardiovascular Research Training Grant (4T32HL007828). CT is partly supported by an Alfred P. Sloan Foundation Research Fellowship. JS is an Investigator of the Howard Hughes Medical Institute.

Data availability.

Data is available on GEO via accession GSE108699 and code along with additional data are available on Github as described below. pHAGE-GFP, pHAGE-BFP, and the CROP-seq vector with the CRISPRi-optimized backbone sequence described in methods are available on Addgene as 106281, 106282, and 106280, respectively (currently pending). The CROP-seq vector with the standard backbone was used as described in Datlinger *et. al* (see methods).

Code Availability

Code for useful scripts and analysis from the paper, more details on data access (as mentioned above), and 10X V2 enrichment PCR details are available at: https://github.com/shendurelab/single-cell-ko-screens.

References

1. Shalem, O., Sanjana, N. E. & Zhang, F. High-throughput functional genomics using CRISPR–Cas9. *Nat. Rev. Genet.* **16,** 299–311 (2015).
2. Mohr, S. E., Smith, J. A., Shamu, C. E., Neumüller, R. A. & Perrimon, N. RNAi screening comes of age: improved techniques and complementary approaches. *Nat. Rev. Mol. Cell Biol.* **15,** 591–600 (2014).
3. Macosko, E. Z. *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161,** 1202–1214 (2015).
4. Klein, A. M. *et al.* Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161,** 1187–1201 (2015).
5. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8,** 14049 (2017).

280 6. Xie, S., Duan, J., Li, B., Zhou, P. & Hon, G. C. Multiplexed Engineering and Analysis of
281 Combinatorial Enhancer Activity in Single Cells. *Mol. Cell* **66,** 285–299.e5 (2017).
282 7. Adamson, B. *et al.* A Multiplexed Single-Cell CRISPR Screening Platform Enables
283 Systematic Dissection of the Unfolded Protein Response. *Cell* **167,** 1867–1882.e21 (2016).
284 8. Dixit, A. *et al.* Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA
285 Profiling of Pooled Genetic Screens. *Cell* **167,** 1853–1866.e17 (2016).
286 9. Jaitin, D. A. *et al.* Dissecting Immune Circuits by Linking CRISPR-Pooled Screens with
287 Single-Cell RNA-Seq. *Cell* **167,** 1883–1896.e15 (2016).
288 10. Datlinger, P. *et al.* Pooled CRISPR screening with single-cell transcriptome readout. *Nat.*
289 *Methods* **14,** 297–301 (2017).
290 11. Debnath, J., Muthuswamy, S. K. & Brugge, J. S. Morphogenesis and oncogenesis of
291 MCF-10A mammary epithelial acini grown in three-dimensional basement membrane
292 cultures. *Methods* **30,** 256–268 (2003).
293 12. Tseng, W. C., Haselton, F. R. & Giorgio, T. D. Transfection by cationic liposomes using
294 simultaneous single cell measurements of plasmid delivery and transgene expression. *J.*
295 *Biol. Chem.* **272,** 25641–25647 (1997).
296 13. Nikolaitchik, O. A. *et al.* Dimeric RNA recognition regulates HIV-1 genome packaging.
297 *PLoS Pathog.* **9,** e1003249 (2013).
298 14. Jetzt, A. E. *et al.* High rate of recombination throughout the human immunodeficiency virus
299 type 1 genome. *J. Virol.* **74,** 1234–1240 (2000).
300 15. Schlub, T. E., Smyth, R. P., Grimm, A. J., Mak, J. & Davenport, M. P. Accurately measuring
301 recombination between closely related HIV-1 genomes. *PLoS Comput. Biol.* **6,** e1000766

302 (2010).

303 16. Sack, L. M., Davoli, T., Xu, Q., Li, M. Z. & Elledge, S. J. Sources of Error in Mammalian
304 Genetic Screens. *G3* **6,** 2781–2790 (2016).

305 17. Yu, H., Jetzt, A. E., Ron, Y., Preston, B. D. & Dougherty, J. P. The nature of human
306 immunodeficiency virus type 1 strand transfers. *J. Biol. Chem.* **273,** 28384–28391 (1998).

307 18. el-Deiry, W. S. *et al.* WAF1, a potential mediator of p53 tumor suppression. *Cell* **75,**
308 817–825 (1993).

309 19. Contente, A., Dittmer, A., Koch, M. C., Roth, J. & Dobbelstein, M. A polymorphic
310 microsatellite that mediates induction of PIG3 by p53. *Nat. Genet.* **30,** 315–320 (2002).

311 20. Gasperini, M. *et al.* CRISPR/Cas9-Mediated Scanning for Regulatory Elements Required
312 for HPRT1 Expression via Thousands of Large, Programmed Genomic Deletions. *Am. J.*
313 *Hum. Genet.* **101,** 192–205 (2017).

314 21. Diao, Y. *et al.* A tiling-deletion-based genetic screen for cis-regulatory element
315 identification in mammalian cells. *Nat. Methods* **14,** 629–635 (2017).

316 22. Vidigal, J. A. & Ventura, A. Rapid and efficient one-step generation of paired gRNA
317 CRISPR-Cas9 libraries. *Nat. Commun.* **6,** 8083 (2015).

318 23. Han, K. *et al.* Synergistic drug combinations for cancer identified in a CRISPR screen for
319 pairwise genetic interactions. *Nat. Biotechnol.* **35,** 463–474 (2017).

320 24. Lukoszek, R., Mueller-Roeber, B. & Ignatova, Z. Interplay between polymerase II- and
321 polymerase III-assisted expression of overlapping genes. *FEBS Lett.* **587,** 3692–3695
322 (2013).

323 25. Yeganeh, M., Praz, V., Cousin, P. & Hernandez, N. Transcriptional interference by RNA

324 polymerase III affects expression of the Polr3e gene. *Genes Dev.* **31,** 413–421 (2017).
325 26. Zhu, S. *et al.* Genome-scale deletion screening of human long non-coding RNAs using a
326 paired-guide RNA CRISPR-Cas9 library. *Nat. Biotechnol.* **34,** 1279–1286 (2016).
327 27. Sanjana, N. E., Shalem, O. & Zhang, F. Improved vectors and genome-wide libraries for
328 CRISPR screening. *Nat. Methods* **11,** 783–784 (2014).
329 28. Chen, B. *et al.* Dynamic imaging of genomic loci in living human cells by an optimized
330 CRISPR/Cas system. *Cell* **155,** 1479–1491 (2013).
331 29. McKenna, A. *et al.* Whole-organism lineage tracing by combinatorial and cumulative
332 genome editing. *Science* **353,** aaf7907 (2016).
333 30. Dixit, A. Correcting Chimeric Crosstalk in Single Cell RNA-seq Experiments. (2016).
334 doi:10.1101/093237
335 31. Qiu, X. *et al.* Single-cell mRNA quantification and differential analysis with Census. *Nat.*
336 *Methods* **14,** 309–315 (2017).

337
Methods

338
Cell Lines and Culture

339 MCF10A immortalized breast epithelial cells were purchased from ATCC and cultured in
340 DMEM/F12 (Invitrogen) supplemented with 10% FBS, 1% penn-strep, 10 ng/mL EGF, 1 µg/mL
341 hydrocortisone, 5 µg/mL insulin and 100 ng/mL cholera toxin.
342

Generating an Inducible Cas9 Expressing MCF10A Cell Line

343 Lentivirus containing a doxycycline inducible and constitutively expressed Cas9 construct was
344 produced by transfecting 293T cells with either pCW-Cas9 (Addgene #50661) or lentiCas9-Blast
345 (Addgene #52962) and plasmids from the ViraPower Lentiviral Expression System (Thermo)
346 according to manufacturer's instructions. 48 hours post transfection, virus containing supernatant
347 was collected and cell debris removed by filtering through a 40 µm syringe filter. MCF10A cells
348 were transduced with filtered supernatant for 48 hours and selected with 1 µg/mL puromycin

349 (pCW-Cas9) or 10 µg/mL blasticidin (lentiCas9-Blast) for 96 hours. For cells expressing a
350 doxycycline inducible Cas9 single cell clones of MCF10A-Cas9 cells were generated by high
351 rate of dilution, individual clones expanded and Cas9 expression of individual clones was
352 confirmed by immunoblotting of cells 96 hours following addition of doxycycline at 1
353 microgram/mL. lentiCas9-Blast cells were maintained as a polyclonal line.
354
355 pCW-Cas9 cells were used for initial arrayed and pooled screens, as well as quantification of
356 editing rates in pHAGE-scKO vector. lentiCas9-Blast cells were used for all CROP-seq
357 experiments.
358

Initial Tagged Transcript Cloning Method

359 Due to our results demonstrating high rates of barcode/sgRNA swapping when using this design,
360 we do not recommend use of this protocol.
361
362 Starting with the standard lentiGuide-puro plasmid (Addgene #52963), this vector was modified
363 to confer blasticidin resistance, a mechanism of selection independent from the pCW-Cas9 (puro
364 resistance) plasmid used to generate MCF10A-Cas9 cells. Puro and its EF-1A promoter were
365 removed via a double digest with NEB SmaI (8 hours at 25 degrees C) and MLU1-HF (8 hours
366 25 degrees C). This product was gel purified using QiaQuick Gel Extraction kit (Qiagen). EF-1A
367 promoter and Blasticidin, each with 20 bp homology on both ends were prepared via PCR from
368 lentiCas9-Blast and gel purified. Fragments were assembled into digested lentiGuide-puro vector
369 using the NEBuilder HiFi DNA Assembly kit with inserts in 2-fold molar excess and
370 transformed into NEB C3040H E. Coli and allowed to incubate overnight at 30 degrees C.
371 Clones were picked from plate, allowed to grow in LB+amp overnight at 30 degrees, and were
372 purified using Qiagen Miniprep kit. Individual clones were validated via Sanger sequencing.
373
374 Lentiguide-blast was linearized using a digest with BsmB1 (Thermo) at 37 degrees for five hours
375 followed by digestion with SalI HF (NEB) overnight and gel purification. Oligos containing both
376 guide sequences and their corresponding barcodes were designed according to the following:
377 tGTGGAAAGGACGAAACACC[G][guide]gttttagagctaGAAAtagcagagacgCGTCTCAgatctccctt
378 tgggccgcctccccgcg[barcode]tcgactttaagaccaatgacttaca
379
380 Where [guide] is a 20 bp guide sequence and [barcode] is an 8 bp barcode sequence uniquely
381 paired to a particular sgRNA. Note that the [G] included prior to the guide is required for
382 expression from a Pol III promoter. Guides that generate an extra BsmB1 restriction site when
383 used in this design were excluded due to incompatibility with our downstream cloning strategy
384 and only barcode sequences that did not generate additional BsmB1 restriction sites were used.
385 RUNX1 was the only target impacted by this filter (4 guides were used instead of 6).

386
387 A library of these oligos was ordered in 96 well format as Ultramers from Integrated DNA
388 Technologies and resuspended at equal molarity. All oligos were resuspended in water, pooled at
389 equimolar concentrations, and amplified using a 50 microliter PCR KAPA HiFi HotStart Ready
390 Mix PCR reaction with 1 ng of input DNA, an annealing temperature of 62 degrees, an extension
391 time of 20 seconds, and all other parameters according to the manufacturer's recommendations.
392 The resulting product was cleaned with a Zymo DNA Clean and Concentrator kit. The purified
393 inserts were assembled into linearized lentiGuide-blast using the NEBuilder HiFi DNA
394 Assembly kit and a molar excess of 1:5 vector to insert ratio. Assembled products were
395 transformed into NEB C3040H E. Coli and grown overnight at 30 degrees in LB+amp. Product
396 was prepared using a plasmid Miniprep kit (Qiagen).
397
398 To prepare the insert for the final reaction, a region spanning from the backbone sequence for the
399 CRISPR sgRNA to a region towards the end of the WPRE element was amplified using the
400 KAPA HiFi Hotstart Master Mix and purified using the Zymo Clean and Concentrator kit. The
401 primers used in this reaction add BsmB1 cut sites that generate complementary ends in the final
402 cloning step following digestion. This amplified fragment was ligated into PGEM-T following
403 the kit protocol and a clone was selected via blue-white screening and validation of individual
404 clones by Sanger sequencing. The validated construct was digested with BsmB1 (Thermo) and
405 gel purified.
406
407 The fragment isolated from PGEM-T was then ligated into this linearized vector using a 3:1
408 molar excess of insert to vector using T4 DNA Ligase (New England Biolabs) and a an overnight
409 incubation at 16 degrees C. Ligation products were transformed into NEB C3040H (stable)
410 competent cells and grown overnight at 30 degrees in LB+amp. Plasmids were recovered using
411 a Plasmid Miniprep kit (Qiagen).
412

pHAGE Vector Cloning

413 The pHAGE_dsRed_IRES_zsGreen vector was modified to contain a multiple cloning site as
414 described in *Quantification of Template Switching in Lentivirus Packaging Using FACS*. The
415 U6-sgRNA cassette containing a 500bp filler removable by Bsmb1 digest was ordered as a
416 gblock (Integrated DNA Technologies). Using the multiple cloning site, the U6-sgRNA cassette
417 was added in the three-prime UTR of the zsGreen/dsRed transgene via Gibson assembly. This
418 vector was further modified to remove the zsGreen/IRES/dsRed cassette and replace the CMV
419 promoter with an EF1a promoter.
420
421 The vector was digested following the protocol outlined in Sanjana and Shalem *et al.* [27]. Oligos
422 corresponding to individual guides with homology for gibson assembly were ordered as standard

423 DNA oligos in 96-well plate format from Integrated DNA Technologies with the following
424 design:
425
426 [GCCTTATTTTAACTTGCTATTTCTAGCTCTAAAAC][GUIDE
427 RC][C][GGTGTTTCGTCCTTTCCACAAGAT]
428
429 Guide RC refers to the reverse complement of the guide sequence. The entire construct may also
430 be reverse complemented, allowing the guide sequence itself to be used rather than the reverse
431 complement. Note that the additional C included here is required for transcription from the Pol
432 III promoter.
433
434 All oligos were resuspended in water, pooled at equimolar concentrations, and amplified using a
435 50 microliter PCR KAPA HiFi HotStart Ready Mix PCR reaction with 1 ng of input DNA, an
436 annealing temperature of 62 degrees, an extension time of 20 seconds, and all other parameters
437 according to the manufacturer's recommendations. The following primers were used for
438 amplification:
439
440 Forward: 5 - GCCTTATTTTAACTTGCTATTTCTAGCT - 3
441 Reverse: 5 - ATCTTGTGGAAAGGACGAAACA - 3
442
443 Reactions were monitored with qPCR and stopped just prior to saturation.
444
445 These reactions were cleaned with a Zymo DNA Clean and Concentrator kit and cloned into the
446 Bsmb1 digested pHAGE vector backbone using the Clontech Infusion HD Cloning Kit.
447 Ligations were performed using 10 fmols of vector and and a 200 fmols of double stranded
448 oligo (1:20 molar ratio of vector to insert). Ligation products were transformed into NEB
449 C3040H (stable) cells according to manufacturer recommendations. Transformations were
450 diluted with 250 µL of LB and spread onto 6 LB-AMP plates and incubated at 30 degrees C for
451 24 hours. Colonies were then scraped into LB, a bacterial pellet was collected and plasmids
452 recovered using a Plasmid Midiprep kit (Qiagen).
453

Quantification of Template Switching in Lentivirus Packaging Using FACS

454 A multiple cloning site was cloned into pHAGE_dsRed_IRES_zsGreen lentiviral vector between
455 the WPRE and 3'LTR. The multiple cloning site was assembled from annealing and extension of
456 WPRE_MCS_insert_W and WPRE_MCS_insert_R:
457
458 WPRE_MCS_insert_W:
459 5- ctttgggccgcctccccgcctgggcgcgccATAACAgctagcTGATGGctcgagcc -3

WPRE_MCS_insert_R:
5- cagctgccttgtaagtcattggtcttaaaggctcgagCCATCAgctagcTGTTATgg -3

The plasmid was amplified by inverse PCR with pHAGE_WPRE_MCS_GIBS_F and R:
pHAGE_WPRE_MCS_GIBS_F
5- TGGctcgagcctttaagaccaatgacttacaaggcagctg -3

pHAGE_WPRE_MCS_GIBS_R
5- ctagcTGTTATggcgcgcccaggcggggaggcggcccaaag -3
The two fragments were cloned by Gibson Assembly. Correct clones of pHAGE_dsRed_IRES_zsGreen_WPRE_MCS were identified by Sanger sequencing and expression of the fluorescent proteins after transfection and lentiviral packaging.

To make the pHAGE EBFP or EGFP_IRES_dsRed_WPRE_MCS, pHAGE_dsRed_IRES_zsGreen_WPRE_MCS was cut with BamHI and ClaI to remove the zsGreen and IRES. The ends were blunted and re-ligated to make pHAGE_dsRed _WPRE_MCS. EGFP or EBFP (amplified with eGFP_gibsF and eGFP_IRES_GibsR) and an IRES (IRES_GibsF, IRES_GibsR) were cloned into the NotI site 5' of the dsRed, by Gibson Assembly. EBFP was ordered as a gBlock (Integrated DNA Technologies) with 3 nucleotide changes from EGFP. Correct clones were identified by sequencing. The dsRed is not expressed in this construct.

eGFP_gibsF:
5- gccatccacgctgttttgacctccatagaagacaccggcATGGTGAGCAAGGGCGAGGAG -3

eGFP_IRES_GibsR:
5- ggatccCTACTTGTACAGCTCGTCCATGCCG -3

IRES_GibsF:
5- ATCACTCTCGGCATGGACGAGCTGTACAAGTAGggatccctcccccccccctaacgttac -3

IRES_GibsR:
5- ctccttgatgacgtcctcggaggaggccatggcggccatgtgtggccatattatcatcgtgtttttcaaagg -3

EBFP
5-
ATGGTGAGCAAGGGCGAGGAGCTGTTCACCGGGGTGGTGCCCATCCTGGTCGAGCT
GGACGGCGACGTAAACGGCCACAAGTTCAGCGTGTCCGGCGAGGGCGAGGGCGATG

499 CCACCTACGGCAAGCTGACCCTGAAGTTCATCTGCACCACCGGCAAGCTGCCCGTGC
500 CCTGGCCCACCCTCGTGACCACCCTGACCCACGGCGTGCAGTGCTTCAGCCGCTACC
501 CCGACCACATGAAGCAGCACGACTTCTTCAAGTCCGCCATGCCCGAAGGCTACGTCC
502 AGGAGCGCACCATCTTCTTCAAGGACGACGGCAACTACAAGACCCGCGCCGAGGTG
503 AAGTTCGAGGGCGACACCCTGGTGAACCGCATCGAGCTGAAGGGCATCGACTTCAA
504 GGAGGACGGCAACATCCTGGGGCACAAGCTGGAGTACAACTTtAACAGCCACAACG
505 TCTATATCATGGCCGACAAGCAGAAGAACGGCATCAAGGTGAACTTCAAGATCCGC
506 CACAACATCGAGGACGGCAGCGTGCAGCTCGCCGACCACTACCAGCAGAACACCCC
507 CATCGGCGACGGCCCCGTGCTGCTGCCCGACAACCACTACCTGAGCACCCAGTCCGC
508 CCTGAGCAAAGACCCCAACGAGAAGCGCGATCACATGGTCCTGCTGGAGTTCGTGA
509 CCGCCGCCGGGATCACTCTCGGCATGGACGAGCTGTACAAG -3
510
511 Fifteen nucleotide barcodes (lenti-barcode and lenti-barcode-r) were then cloned into the
512 multiple cloning site between the WPRE and 3'LTR for both the EBFP and EGFP constructs by
513 Gibson Assembly. Single clones were prepared and the barcode identified by Sanger sequencing.
514
515 lenti-barcode:
516 5-
517 atctccctttgggccgcctccccgcctgggGGATCCAGNNNNNNNNNNNNNNNNtcgagcctttaagaccaatgactt
518 acaagg -3
519
520 lenti-barcode-r:
521 5- CCTTGTAAGTCATTGGTCTTAAAGGCTCGA -3
522
523 Lentivirus was packaged by transfection of barcoded EGFP or EBFP constructs either alone or in
524 an equimolar mix along with helper plasmids (pHDM-Hgpm2, pHDM-Tatlb, pRC-CMVRev1b
525 and pHDM-VSV-G) into HEK293T cells using Lipofectamine 2000 (Invitrogen). Viral
526 supernatant was collected after 48 hours, spun to remove debris, snap frozen in liquid nitrogen
527 and stored at -80˚C. To titer the packaged lentiviruses, they were thawed on ice and added to
528 MCF10A cells with media containing 8 micrograms/ml polybrene, andthe frequency of
529 transduced cells 48 hours post-transduction was determined by flow cytometry.
530
531 To sort blue+ and green+ populations, 400,000 of MCF10A Δ*TP53* cells (Horizon Discovery) in
532 5 ml media plus 8 micrograms/ml polybrene were transduced at a MOI ~0.1, with either of the
533 EGFP or EBFP expressing viruses that had been packaged singly, a mix of the EGFP and EBFP
534 expressing viruses that had been packaged singly or the EGFP and EBFP expressing viruses that
535 had been packaged together. The cells were then cultured for four weeks to avoid residual
536 plasmid contamination following transduction. An equal number of cells transduced with EGFP
537 and EBFP virus were mixed to determine the rate of contamination resulting from FACS error.

538 The mixed cells along with others were sorted for blue+ or green+ populations using a FACS
539 Aria II (Becton Dickinson) that had been compensated for the overlap between the EBFP and
540 EGFP emission spectra. The genomic DNA was harvested from each population using the
541 Qiagen DNeasy kit. The barcodes were amplified from 2-36 ng of genomic DNA in 50 ul Robust
542 polymerase (Kapa) reactions with primers bwds_p5_WPRE_BC_F and
543 bwds_next_WPRE_BC_R.
544
545 bwds_next_WPRE_BC_R:
546 GGCTCGGAGATGTGTATAAGAGACAG
547 5- gaaatcatcgtcctttccttggct -3
548
549 bwds__p5_WPRE_BC_F:
550 5- AATGATACGGCGACCACCGAGAgcgccgatgccttgtaagtcattggtcttaaaggctc -3
551
552 Reactions were removed from the thermocycler just prior to saturation (27-30 cycles). The PCR
553 products were purified with Ampure (Agilent) and P7 index sequences were added by an
554 additional six cycles of PCR. PCR products were purified, quantified, pooled and single-end
555 sequenced on an Illumina Nextseq500 with Read1 primer bwds_WPRE_bc_seqF and standard
556 Illumina i7 primers.
557
558 bwds_WPRE_bc_seqF:
559 5- GCG CCG ATG CCT TGT AAG TCA TTG GTC TTA AAG GCT CGA -3
560

Analysis of FACS Data from pHAGE-GFP and pHAGE-BFP Experiments

561 The background percentage of contaminating barcodes in the BFP/GFP sorted cells from the
562 mixed cells control was first subtracted from the numbers obtained for the pooled virus samples.
563 The fraction of GFP cells, as determined from FACS gating, was fixed and the expected fraction
564 of barcode contamination in the BFP and GFP given this fixed fraction of GFP cells was
565 simulated. Note that the expected contamination of green barcodes in the BFP sorted cells is
566 simply the template switching rate multiplied by the fraction of green cells. The expected rate of
567 contamination of BFP barcodes in the GFP sorted cells is simply the template switching rate
568 multiplied by the BFP cell fraction (1 – GFP cell fraction). The sum of the squared error between
569 the observed and expected values for these to rates of contamination was calculated for a range
570 of different lentivirus swap rates and the minimal value was taken to be the most likely swap rate
571 given the observed data.
572
573 Note that, unlike in a large library of plasmids, in a mix of two plasmids, only half of all
574 chimeric products formed will be detectable as many virions will be homozygous (contain the
575 same construct, and thus chimeric products are identical to the original). To give an analogous

576 example, in a barnyard experiment for a single-cell assay, mouse-mouse or human-human
577 multiplets cannot be detected and thus estimated rates of 'doublets' have to be adjusted
578 accordingly. When the plasmids are equimolar and the swap rate is 50%, for example, one would
579 expect to observe a 75% rate of the intended barcode and a 25% rate of the unintended barcode.
580 This ratio will change according to the molar concentration of the two plasmids. In **Fig. 1f**, we
581 assume that the pool was composed of 61.7% GFP plasmid, corresponding to the fraction of
582 GFP+ cells relative to the total number of GFP+ and BFP+ cells -- 4.59 / (4.59 + 2.85) or 61.7%
583 as explained in **Fig. S4.** This analysis was also performed without fixing the fraction of GFP+
584 cells to the value measured by FACS to ensure results were concordant between the two methods
585 (**Fig. S5**). The minimum sum of squared error over the grid of simulated lentivirus swap rate and
586 fraction of GFP cells were taken to be the most likely set of parameter values.
587

CRISPRi Experiment

588 K562 expressing dCas9-BFP-KRAB (gift of the Bassik lab, Addgene #46911) and MCF10A
589 expressing dCas9-BFP-KRAB (made by transduction with
590 lenti_UCOE_EF1-dCas9-BFP-KRAB, plamid available on Addgene soon; see
591 https://weissmanlab.ucsf.edu/CRISPR/CRISPRiacelllineprimer.pdf) were transduced with
592 lenti-mCherry under control of a CAG-promoter (pCAG_mCherry pKH143, gift of the Bassik
593 lab, unpublished), and sorted such that the resulting population is enriched for mCherry
594 expression.
595

596 A spacer targeting the CAG-promoter was cloned into the KHH030 (Addgene #89358),
597 CROP-seq, and pHAGE sgRNA expression vectors. The CROP and pHAGE were modified by
598 Q5-Site Directed Mutagenesis (New England BioLabs) to use the previously described
599 sgRNA-(F+E)-combined optimized backbone [28]. The CRISPRi mCherry+ K562 and MCF10A
600 cells were transduced with the CAG-targeting sgRNA, and again assayed for mCherry.
601

602 All virus for the CRISPRi experiments were made by the Co-operative Center for Excellence in
603 Hematology Vector Production core. All sorting was performed on a FACS Aria II (Becton
604 Dickinson).
605

606 ***Editing Rate Experiment for pHAGE-scKO***

607 To confirm that our pHAGE-scKO vector exhibited reduced editing efficiency in addition to
608 reduced inhibition efficiency via CRISPRi, we performed editing with a guide to TP53 from our
609 screen (GAGCGCTGCTCAGATAGCGA) in both lentiGuide-Blast and pHAGE-scKO using our
610 pCW-Cas9 MCF10A cells. Cells were passaged for 18 days post-induction of Cas9 expression
611 with dox and gDNA was harvested using Qiagen DNeasy kit and amplified using primers
612 CTAAATGGCTGTGAGAGAGCTCAGCCACACGCAAATTTCCTTCC and

ACTTTATCAATCTCGCTCCAAACCCCCTGCCCTCAACAAGATGT. These were then amplified using KAPA HiFi Hotstart Ready Mix (KAPA) using the following primers to generate final indexed sequencing libraries: AATGATACGGCGACCACCGAGATCTACACacgtaggcCTAAATGGCTGTGAGAGAGCTCAG

CAAGCAGAAGACGGCATACGAGAT[INDEX]gaccgtcggcACTTTATCAATCTCGCTCCAAACC

These reads were then processed using the method described in McKenna and Findlay *et al.* [29] Briefly, low quality bases are trimmed using Trimmomatic, reads are merged using Flash, aligned to the reference of the locus surrounding the guide site using needle, and unique genotypes are quantified. The wild-type genotype fraction was taken to be the proportion of unedited alleles. We did not use UMIs in this experiment. The lack of UMIs may overestimate the editing rate in all samples to some extent due to amplification bias.

KO Experiments

For all screens, each plasmid library was transfected along with plasmids provided with the ViraPower Lentiviral Expression into 293T cells. At 48 and 72 hours post transfection, viral containing supernatant were collected, filtered using a 40 µm steriflip filtration system (EMD Millipore). For arrayed experiments, individual plasmids were transfected and viruses produced as described above. For pHAGE-scKO and arrayed/pooled pLGB-scKO vector experiments, virus concentrated using Peg-it virus concentration solution (SBI). Viral titer of the concentrated lentiviral library was determined by transduction of MCF10A-Cas9 cells for 48 hours at several viral dilutions, splitting cells into replica plates, and subjecting replica plate to blasticidin. Percent control growth was used to assess MOI. MCF10A-Cas9 cells with estimated MOIs of 0.3 were carried forward for further experiments.

For pHAGE-scKO and arrayed/pooled pLGB-scKO vector experiments , media was switched to 1 microgram/mL doxycycline to induce expression of Cas9 in pCW-Cas9 cells. LentiCas9-Blast cells were used for CROP-seq experiments, which do not require induction of Cas9 expression. Editing was allowed take place for 14 days for arrayed and pooled pLGB-scKO and 21 days for pHAGE-scKO and CROP-Seq experiments. Media was changed every 48 hours and cells were cultured every 96 hours. For the first half of editing, cells were cultured in the presence of 5 µg/mL blasticidin and 0.5 µg/mL puromycin to ensure high sgRNA and Cas9 expression.

647

Doxorubicin Treatment

648 After editing, MCF10a cells were seeded in 10 cm plates plates at 1 x 10^6 cells per well, allowed
649 to attach overnight and media replaced with MCF10A media alone (mock) or MCF10A media
650 containing 500 (arrayed and pooled pLGB-scKO experiments) or 100 nM (pHAGE-scKO and
651 CROP-Seq experiments; we ultimately decided that this lower dose was more appropriate and
652 likely to provide more robust signal) doxorubicin prepared from a 500 µM stock of doxorubicin
653 (Sigma) in water. 24 hours after drug exposure untreated and doxorubicin treated cells were
654 harvested by trypsinization, washed with PBS and used for downstream assays.
655

Single-Cell RNA-sequencing

656 Cells were captured using one lane of a 10X Chromium device per sample using 10X V1 Single
657 Cell 3' Solution reagents (10X Genomics). Approximately 4000-7000 cells were captured per
658 lane for each condition. Protocols were performed according to manufacturer recommendations,
659 holding 10-30 ng of full length cDNA out of downstream shearing and library prep steps in
660 order to provide material for barcode enrichment PCR.
661

662 Final libraries were sequenced on NextSeq 500/550. 10X V1 samples were sequenced using the
663 following read configuration on 75 cycle High Output kits:
664 R1: 64, R2: 5, I1: 14, I2: 8
665

666 Our initial arrayed and pooled doxorubicin treated samples using pLGB-scKO were aggregated
667 using cellranger aggregate to normalize the average number of mapped reads per cell. This yields
668 an average of 37,732 reads per cell, 2263 median genes per cell, and a median of 8279 UMIs per
669 cell.
670

671 Our CROP-seq mock sample was sequenced to an average depth of 120,797 raw reads per cell in
672 6598 cells. A median of 4619 genes per cell were detected and a median UMI count of 22,495
673 per cell. Our CROP-seq doxorubicin treated sample was sequenced to an average depth of
674 123,445 raw reads per cell in 6283 cells. A median of 3500 genes per cell were detected and we
675 observed a median UMI count of 15,324 per cell. At this depth the average duplication rate is
676 approximately 78%.
677

Enrichment PCR

678 For all experiments a hemi-nested PCR starting from 5 ng of full length cDNA was used to
679 enrich for the barcodes that assign a target to each cell. All PCR reactions were performed with a
680 P7 reverse primer (as introduced by the 10X Chromium V1 oligo DT RT primer). Importantly,
681 the protocol for the 10X V2 protocol (not used here) would be different -- see

https://github.com/shendurelab/single-cell-ko-screens#enrichment-pcr for more information. For pHAGE-scKO and pLGB-scKO, the first PCR was performed with

5- TCCTGGGATCAAAGCCATAGT -3

and for CROP-Seq with

5- TTTCCCATGATTCCTTCATATTTGC -3

as the forward primer, priming to the blasticidin transcript with no non-templated sequence for pHAGE-scKO and pLGB-scKO, and to part of the U6 promoter in CROP-seq. For pLGB-scKO the second PCR was performed with

5- TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGACGAGTCGGATCTCCCTT -3

for pHAGE-scKO with

5-
TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAACGGACTAGCCTTATTTTAACTTG
-3

and for CROP-Seq with

5- TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGcTTGTGGAAAGGACGAAACAC -3

as the forward primer, priming adjacent to the barcode/guide sequence in each design and adding the standard Nextera R1 primer. Samples were then indexed in a final PCR using standard Nextera P5 index primers of the form:

5- AATGATACGGCGACCACCGAGATCTACAC[8bp Index]TCGTCGGCAGCGTC -3

Each PCR was cleaned with a 1.0X Ampure XP cleanup and one microliter of a 1:5 dilution of the first PCR was carried forward and a 1:25 dilution of the second PCR was carried into the final PCR reaction. PCRs were monitored by qPCR and stopped just prior to reaching saturation to avoid overamplification. The final PCR was run on a Bioanalyzer to confirm expected product size.

Digital Gene Expression Quantification

Sequencing data from each sample was processed using cellranger 1.3.1 to generate sparse matrices of UMI counts for each gene across all cells in the experiment.

Each lane of cells was processed independently using cellranger count, aggregating data from multiple sequencing runs. For the comparison between arrayed and pooled screens, cellranger aggregate was used to downsample data from each screen to an equal average number of mapped reads.

Assigning Cell Genotypes

Barcode enrichment libraries were separately indexed and sequenced as spike-ins alongside the whole transcriptome scRNA-seq libraries. Final UMI and cell barcode assignments were made for each read by processing these samples with cellranger 1.3.1 as was done for the whole transcriptome libraries.

A whitelist of guide or target barcode sequences was constructed using all guides or target barcodes in the library. For each read in the position sorted BAM file output by cellranger 1.3.1, the final cellular barcode and UMI are extracted. If either of these fields is not populated, indicating low sequencing quality for the cell barcode or UMI read, the read is ignored. Using the cDNA read, we attempt to find a perfect match for the sequence immediately preceding the guide or barcode (GTGGAAAGGACGAAACACCG for CROP-seq and CGCCTCCCGCG for pLGB-scKO). If a perfect match is not found, we attempt to locate the sequence in an error-tolerant manner using a striped Smith-Watterman alignment, where alignments must span a length no more than 2bp shorter than the search sequence. If a match or alignment is found, the guide or barcode sequence is extracted. If the extracted sequence does not perfectly match a whitelist sequence, we search for a matching whitelist sequence within an edit distance of half the minimum edit distance between any pair of guides or barcodes in the library (rounded down). If no match is found, the molecule is ultimately discarded. Matches to the whitelist are tracked for each cell.

We also remove likely chimeric sequences using the approach outlined in Dixit [30]. Briefly, within each cell, we first calculate the number of times a given UMI is observed with each observed guide assignment. We then divide these counts by the total instances of the respective UMI across all observed guide assignments within that cell. For UMI-guide assignment combinations where this fraction is less than 20%, we do not count the UMI towards the final observed guide assignment counts. While this has some impact on the raw data, we find the benefits to be modest, in contrast to results reported in Dixit *et al.* [8].

To make a set of final assignments, we take all whitelist sequences with over 10 reads and account for over 7.5% of the whitelist reads assigned to a given cell, where multiple sequences can be assigned to each cell. Whitelist sequences and their corresponding target genes are assigned to each cell. Finally, this set of assignments is merged with the filtered gene expression matrices output by cellranger such that only assignments to the set of high quality cells appear in the final dataset.

Note that when processing CROP-seq data without PCR enrichment, we lowered the requirement for reads supporting a given guide to 3 to account for the decreased coverage of these transcripts.

Estimation of MOI and Capture Rate

The most likely multiplicity of infection and capture rate given the distribution of guide counts per cell were estimated using the generative model described in Dixit *et al.* [8]. Briefly, a log likelihood is calculated using a zero-truncated poisson (represents the multiplicity of infection after selection for cells that harbor a lentiviral construct) convolved with a binomial (represents the incomplete capture of transcripts containing guide sequences from cells). This model is used to to calculate log-likelihood values for a range of MOI values and capture rate values (rate of observing a guide in a cell given that the guide is present). The maximum log-likelihood is taken to be the MOI and capture rate of the experiment.

Removing Low Quality Cells

Despite using the filtered set of cells provided by cellranger to exclude cell barcodes with low UMI counts, we consistently observed a cluster of cells with much lower UMI counts on average than the rest of the dataset when performing dimensionality reduction. To avoid including these cells in downstream analysis, we perform a simple procedure to remove any cluster with low average UMI counts.

First, we perform PCA on the matrix of all cells and genes expressed in at least 50 cells for each condition to reduce to 12 principal components. We then reduce to a two dimensional space using tSNE. Next we perform density peak clustering in the two dimensional space using default parameters. For each cluster, we calculate the average size factor over the cluster as calculated using estimateSizeFactors in monocle2 [31]. We observed that filtering out clusters of cells with an average size factor of -0.85 or lower readily distinguished the low quality cluster of cells. All cells contained in these clusters were removed from downstream analysis. PCA and tSNE were performed using the monocle2 function reduceDimension with default parameters and the tSNE option. Density peak clustering was performed using the monocle2 function clusterCells.

Simulating Loss in Power from Barcode Swapping

Assignments were permuted for a fraction of cells ranging from 0 to 100% and kept fixed for the remaining fraction of cells. The monocle2 function differentialGeneTest was used to test for genes differentially expressed across the target assigned to each cell (only testing genes detectably expressed in at least 50 cells). The number of genes with a qvalue of 0.05 or lower was counted. This was performed over 10 different resamplings for each rate of swapping to obtain a distribution for each swap rate.

The average fold-reduction in DEGs resulting from a 50% swap rate was taken to be the approximate fold change in power resulting from template switching in our original design.

For the simulation performed on our own data, cells with a single target assignment from 100nm doxorubicin treated cells in our CROP-seq experiment were taken as the starting set of cells.

For the simulation on data from Adamson *et al.*, processed data was obtained from GEO (GSE90546). Assignment of cells to targets were used as provided on GEO and only cells that were noted as having high quality assignment of the assigned target and noted as being a single cell were used in downstream analysis. Due to the large number of cells (50,000+) in the UPR experiment from this study and the large number of differential tests required for these simulations, the number of cells assigned to each target was downsampled by 2-fold to reduce runtime. We also performed tests on a dataset further downsampled to approximately 6,000 cells to illustrate the relationship between the initial power of a screen and the impact of simulated target swapping on sensitivity.

tSNE Embedding Demonstrating TP53 Enriched Cluster

Scaling, PCA, and tSNE on PCA results were performed with the monocle2 function reduceDimension. 20 dimensions from PCA were carried into tSNE which was performed to two dimensions using default parameters. All cells (except those excluded in the low size factor cluster) including cells with guides to multiple targets and no assigned target were included in dimensionality reduction for this plot. Percentages of cells with guides to TP53 and ARID1B were calculated including guides that contain guides to multiple targets (all cells with TP53 guides were counted as TP53 cells and all cells without TP53 but with one or more guides to ARID1B were counted as ARID1B cells for the purposes of calculating reported percentages).

822

Enrichment of Tumor Suppressors in Specific Molecular States

823 Only cells containing a guide to a single target were considered in all enrichment testing. A Chi
824 squared test was used to determine whether the distribution of individual sgRNAs and targets in
825 tSNE space was significantly different from non-targeting controls at an FDR cutoff of 5%.
826 Targets which did not pass this test and did not have any individual sgRNA pass the test were
827 excluded from the subsequent enrichment tests. For each sgRNA of the remaining targets, we
828 sought to estimate the functional editing rate (probability of a cell having a true LoF given that it
829 received that sgRNA), but such estimates would be confounded if one accounts for the
830 possibility of edits that cause LoF for the target gene but have incomplete penetrance on the
831 cellular phenotype. Therefore we used an expectation maximization approach to estimate the
832 functional edit rate of each sgRNA relative to the unknown functional edit rate of the most
833 efficient sgRNA for a given target.
834
835 The t-SNE cluster distribution of all cells in which a given sgRNA was detected was modeled as
836 a mixture of the t-SNE cluster distribution of cells with a functional edit for the sgRNA's target
837 gene and the t-SNE cluster distribution of non-targeting controls, where the mixing parameter is
838 the relative functional edit rate for that sgRNA. In the expectation step, the t-SNE cluster
839 distribution of cells with a functional edit for the target is estimated as the weighted average of
840 the empirical t-SNE cluster distributions of each sgRNA for the target, weighted by the current
841 estimates of the relative functional edit rate of the sgRNAs. In the maximization step, the relative
842 functional edit rate of each sgRNA for the target is estimated as that which maximizes the
843 likelihood of the observed t-SNE cluster distribution for cells receiving that sgRNA under the
844 multinomial mixture model.
845
846 After estimating the relative functional edit rate for each sgRNA, a weighted contingency table
847 was constructed where the rows are targets, the columns are t-SNE clusters, and the values are
848 weighted cell counts, where a cell's weight is proportional to the relative functional edit rate for
849 the sgRNA it received. Fractional values were rounded down. Fisher's exact test was applied to
850 this weighted contingency table to test for enrichment of targets amongst t-SNE clusters. Targets
851 were defined as enriched at an FDR of 10%. Chi square and Fisher's exact test were performed
852 using R functions chisq.test and fisher.test, respectively.
853

Principal component and gene set enrichment analysis

854 Pairwise differential gene expression analysis was performed between enriched target cells and
855 non-targeting controls for cells in all significant enriched target-cluster pairs from our enrichment
856 testing. The union of all differentially expressed genes across targets (FDR 5%) was used to
857 perform principal component analysis. Gene set enrichment analysis was performed on genes

that had the top positive and negative loadings for principal component 1 (less than -0.02 or greater than 0.02). Gene set enrichment analysis was performed using the piano R package and the hallmarks gene set from MSigDB. Gene sets were defined as enriched at an FDR cutoff of 1%. PCA was performed using the prcomp function in R, differential expression analysis was done using the monocle2 function differentialGeneTest. The hallmarks gene set collection GMT file was downloaded from the MSiGDB.

866 **Figure Legends**
867
868 **Figure 1** Template switching during lentiviral packaging decreases the sensitivity of designs
869 relying on cis-pairing of sgRNAs and distal barcodes. **A)** Generalized schematic of vectors that
870 rely on *cis* pairing of sgRNAs and barcodes. **B)** Generalized schematic of CROP-seq approach.
871 One copy of the guide is cloned into the 3' LTR of the vector and a second copy of the guide
872 expression cassette is produced in the 5' LTR during lentivirus positive strand synthesis prior to
873 integration. **C)** Schematic of constructs developed to quantify template switching rate at 2.4 kb
874 separation between sequences. Distinguishing bases (3 bp differences) in GFP and BFP are
875 separated from their respective barcodes by 2.4 kb. **D-E)** Cells were transduced with GFP or
876 BFP virus separately or a virus generated from a mix of GFP/BFP produced from individual or
877 combined lentiviral packaging. As an additional control, cells transduced with GFP or BFP only
878 virus were mixed prior to sorting. Cells were sorted on GFP and BFP and the percent GFP and
879 BFP barcodes in each sample is shown as a table. Note that in a mix of two plasmids only
880 approximately half of all chimeric products are detectable due to homozygous virions (see
881 Methods). **F)** Plot of sum of squared errors of observed data vs. expected values at various swap
882 rates assuming a relative proportion of 61.7% GFP+ cells as determined from FACS (see **Fig. S4**
883 for derivation of this percentage and the supplementary methods for a detailed explanation of
884 how the expected values are determined). **G)** Transcription factor pilot screen from Adamson *et*
885 *al.,* used here as a gold standard performed with arrayed lentivirus production, was subjected to
886 simulation of progressively higher fractions of target assignment swapping to mimic the impact
887 of template switching. Number of differentially expressed genes across the target label at FDR of
888 5% is plotted at each swap rate. 0.5 corresponds to the 50% swap rate determined via FACS.
889
890 **Figure 2** CROP-Seq screen of tumor suppressors with high capture rate by PCR enrichment, and
891 assessment of alternate sgRNA placement within a pol II 3'UTR . **A)** Schematic of PCR
892 enrichment of barcoded transcripts from CROP-seq samples. **B)** Determination of the most likely
893 multiplicity of infection and capture rate of barcoded transcripts based on a generative model. **C)**
894 tSNE embedding of a doxorubicin treated sample with colors corresponding to cells with guides
895 to *TP53*, cells that contain non-targeting controls (NTC), cells containing guides to non-*TP53*
896 targets, and cells that are unassigned. **D)** *CDKN1A* and *TP53I3* expression in cells expressing
897 either non-targeting controls or guides to *TP53*. Cells with *TP53* guides are further stratified into
898 cells inside and outside of the *TP53* enriched cluster from panel 2C. **E)** Schematic of pHAGE
899 design with sgRNA placed upstream of the LTR. **F)** CRISPRi knock-down of mCherry in
900 MCF10A and K562 cells not expressing a guide (- control), KHH30 (+ control), CROP-seq, and
901 pHAGE-scKO design. All vectors have been modified to contain a CRISPRi optimized
902 backbone.
903
904

905
906 **Supplementary Figure Legends**
907
908 **Figure S1** Diagram of cloning protocol and barcoded transcript enrichment strategy relying on
909 *cis* pairing of sgRNAs and barcodes (pLGB-scKO). **A)** Schematic of our final vector relying on
910 *cis* pairing of an sgRNA and a distal barcode. **B)** Strategy for PCR enrichment of barcoded
911 transcripts from single-cell RNA-seq data. **C)** Pooled cloning protocol. In 1.1 we start with
912 pLentiguideBlast and digest near the final locations of the sgRNA and paired barcode. In 2.1 an
913 engineered library of oligos containing programmed pairs of sgRNAs and corresponding
914 barcodes are inserted into the digested vector. In 1.2 a portion of pLentiguideBlast is amplified.
915 In 2.2 this fragment is cloned into PGEM-T. Finally, in step 3 vectors resulting from 2.1 and 2.2
916 are digested with Bsmb1 and the insert from 2.2 is ligated into the backbone in 2.1 to produce the
917 final library of sgRNAs and paired barcodes.
918
919 **Figure S2** Barcoded transcript enrichment quality control for arrayed and pooled pLGB-scKO
920 experiments. Each dot represents a barcode sequence observed in a given cell. Plot of reads for a
921 given barcode against the proportion of all barcode reads observed in a given cell for every
922 barcode/cell pair. Red lines indicate the lower-bounds used to distinguish noise from true
923 barcode observations (10 reads and 0.075 proportion within cell). All barcodes observed above
924 the red lines are assigned to their respective cells. Left, doxorubicin treated sample from arrayed
925 experiment. Right, Doxorubicin treated sample from pooled experiment.
926
927 **Figure S3** Comparison of a screen performed with arrayed and pooled lentivirus production
928 using a vector that relies on *cis* pairing of sgRNAs and barcodes. Experiments were performed at
929 different times but under the same conditions. The arrayed experiment was performed as a pilot
930 experiment with 4 targets and observed an overall low rate of cells with detected barcodes. The
931 pooled experiment was performed afterwards with 10 targets and a set of non-targeting controls
932 and we observed a high proportion of cells with detected barcodes and good coverage of the
933 library. To compare these experiments, only the four overlapping targets were considered and the
934 number of cells containing an sgRNA to each target and sequencing depth were matched
935 between samples to control for power differences. **A)** Size-factor normalized *CDKN1A* and
936 *TP53I3* expression across *TP53* and the three other targets in arrayed screen. **B)** *CDKN1A* and
937 *TP53I3* expression across *TP53* and three other targets in pooled screen that overlap with the
938 arrayed screen. **C)** Comparison of the number of differentially expressed genes detected at an
939 FDR of 5% for arrayed across the target label in the arrayed and pooled experiments.
940
941
942
943

Figure S4 Design and sorting of GFP and BFP positive fractions in lentivirus barcode swapping experiment. **A)** Schematic of vectors (pHAGE-GFP and pHAGE-BFP) designed to quantify template switching rate at 2.4 kb using a FACS readout. FACS plots are shown for sorted cells in samples corresponding to **B)** GFP only transduced cells **C)** BFP only transduced cells **D)** GFP and BFP only transduced cells mixed just prior to FACS as a control **E)** cells transduced with BFP and GFP virus that was generated separately but pooled prior to transduction **F)** cells transduced with BFP and GFP virus that was generated from pooled plasmids. The fraction of green plasmids assumed in the determination of lentivirus swap rate from FACS experiments is taken as the fraction of GFP+ cells relative to the total GFP+ and BFP+ cells from this sort (4.59 / (4.59 + 2.85) or 61.7%). This accounts for the fact that plasmids were likely not completely equimolar. The approximate number of total cells cells sorted in each fraction is indicated along the appropriate axes on each plot.

Figure S5 Simulation of concordance between observed and expected data obtained from FACS experiment in **Fig. S4** to quantify template switching rate at 2.4 kb separation between paired sequences. **Fig. 1F** assumed a fraction of 0.617 of GFP plasmid in the original green plasmid / blue plasmid mix as determined from FACS in **Fig. S5**. In this figure, both the fraction of GFP plasmid and lentivirus swap rate are varied to obtain the set of parameters that best fit the collected data. The sum of the squared error between expected and observed values from FACS given each combination of parameters is shown.

Figure S6 Guide transcript enrichment quality control plot for tumor suppressor knock-out screen performed with CROP-seq. Each dot represents a guide sequence observed in a given cell. Plot of reads for a given guide against the proportion of all guide reads observed in a given cell for every barcode/cell pair. Red lines indicate the lower-bounds used to distinguish noise from true guide observations (10 reads and 0.075 proportion within cell). All guide observed above the red lines are assigned to their respective cells. Left, Doxorubicin treated sample from CROP-seq experiment. Right, Mock sample from CROP-seq experiment.

983 **Figure S7** Loss of several targets alter the distribution of mock and doxorubicin exposed cells
984 within tSNE clusters. **A-B)** 3D tSNE embedding and clustering of mock and doxorubicin treated
985 samples, respectively. **C-F)** Chi-squared test qvalues (p values adjusted using the
986 Benjamini-Hochberg method) resulting from testing for differences in the distribution of targets
987 in our screen at both the individual sgRNA **(C and E)** and overall target levels **(D and F).**
988 Comparisons are relative to the distribution of non-targeting controls across tSNE clusters for
989 mock and doxorubicin treated samples, respectively (qvalues were capped to 1e-50 for
990 visualization). Significant differences below a qvalue of 0.05 are colored in red (boundary
991 marked by the grey dashed line).
992
993 **Figure S8** Enriched target-cluster pairs highlight tumor suppressors that share various degrees of
994 a *TP53* deficient signature **A)** Fisher's exact with weights applied to guides according to an
995 expectation maximization procedure were performed for the doxorubicin treated sample to find
996 clusters from Fig. S7 panel B were particular targets were found to be enriched. Cells with
997 target-cluster pairs that showed enrichment were used to generate an aggregate expression profile
998 for every target within genes that are differentially expressed between *TP53* and non targeting
999 controls (NTC). A PCA was performed on these average expression profiles and a distribution of
1000 targets across PC1 is shown colored by the cluster in which they were found to be enriched. **B)**
1001 Gene set enrichments for top positively and negatively loaded (less than -0.02 or greater than
1002 0.02) genes along PC1 (qval < 0.01). **C)** Differential expression tests were performed for cells
1003 within each enriched target-cluster pair, comparing each target to all NTC cells. The proportion
1004 of overlap between these differentially expressed genes and the genes differentially expressed
1005 between *TP53* and NTC is shown.
1006
1007 **Figure S9** Swap rate simulations for our own CROP-seq tumor suppressor screen and the
1008 unfolded protein response screen from Adamson *et al.* Each dataset was subjected to simulation
1009 of progressively higher fractions of target assignment swapping to mimic the impact of template
1010 switching. Number of differentially expressed genes across the target label at FDR of 5% is
1011 plotted at each swap rate. 0.5 corresponds to the 50% swap rate determined via FACS. **A)**
1012 CROP-seq tumor suppressor screen from our study. **B)** Unfolded protein response screen from
1013 Adamson et al. downsampled from ~50,000 to 25,000 cells to make simulations computationally
1014 feasible. **C)** Unfolded protein response screen from Adamson *et al.* downsampled to 6,000 cells
1015 to illustrate how reduced power impacts the observed impact from simulated swapping.
1016

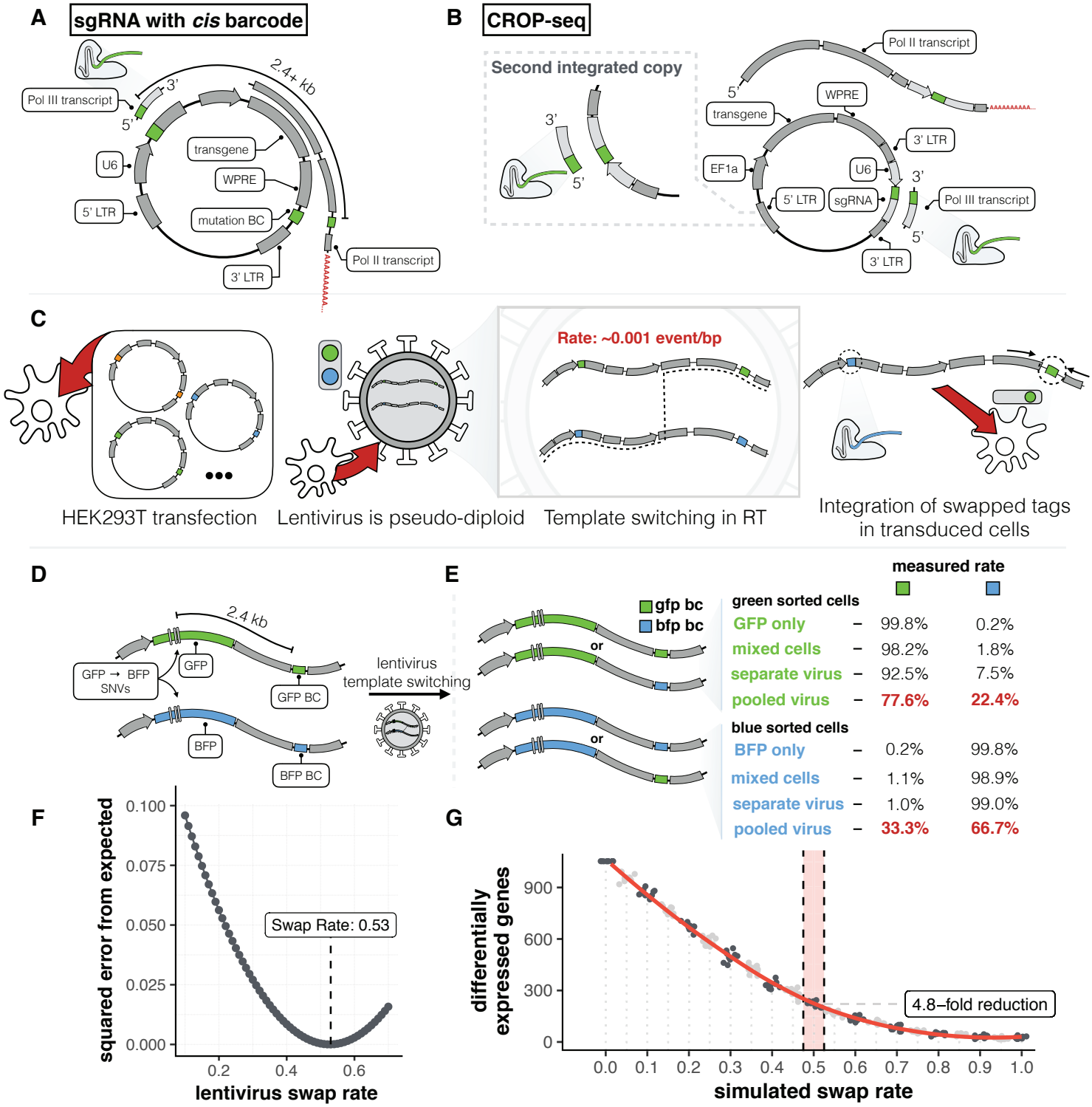

Figure 1 Template switching during lentiviral packaging decreases the sensitivity of designs relying on cis-pairing of sgRNAs and distal barcodes. **A)** Generalized schematic of vectors that rely on *cis* pairing of sgRNAs and barcodes. **B)** Generalized schematic of CROP-seq approach. One copy of the guide is cloned into the 3' LTR of the vector and a second copy of the guide expression cassette is produced in the 5' LTR during lentivirus positive strand synthesis prior to integration. **C)** Schematic of constructs developed to quantify template switching rate at 2.4 kb separation between sequences. Distinguishing bases (3 bp differences) in GFP and BFP are separated from their respective barcodes by 2.4 kb. **D-E)** Cells were transduced with GFP or BFP virus separately or a virus generated from a mix of GFP/BFP produced from individual or combined lentiviral packaging. As an additional control, cells transduced with GFP or BFP only virus were mixed prior to sorting. Cells were sorted on GFP and BFP and the percent GFP and BFP barcodes in each sample is shown as a table. Note that in a mix of two plasmids only approximately half of all chimeric products are detectable. **F)** Plot of sum of squared errors of observed data vs. expected values at various swap rates assuming a relative proportion of 61.7% GFP+ cells as determined from FACS (see **Fig. S4** for derivation of this percentage and the supplementary methods for a detailed explanation of how the expected values are determined). **G)** Transcription factor pilot screen from Adamson *et al.*, used here as a gold standard, was subjected to simulation of progressively higher fractions of target assignment swapping to mimic the impact of template switching. Number of differentially expressed gene across the target label at FDR of 5% is plotted at each swap rate. 0.5 corresponds to the 50% swap rate determined via FACS.

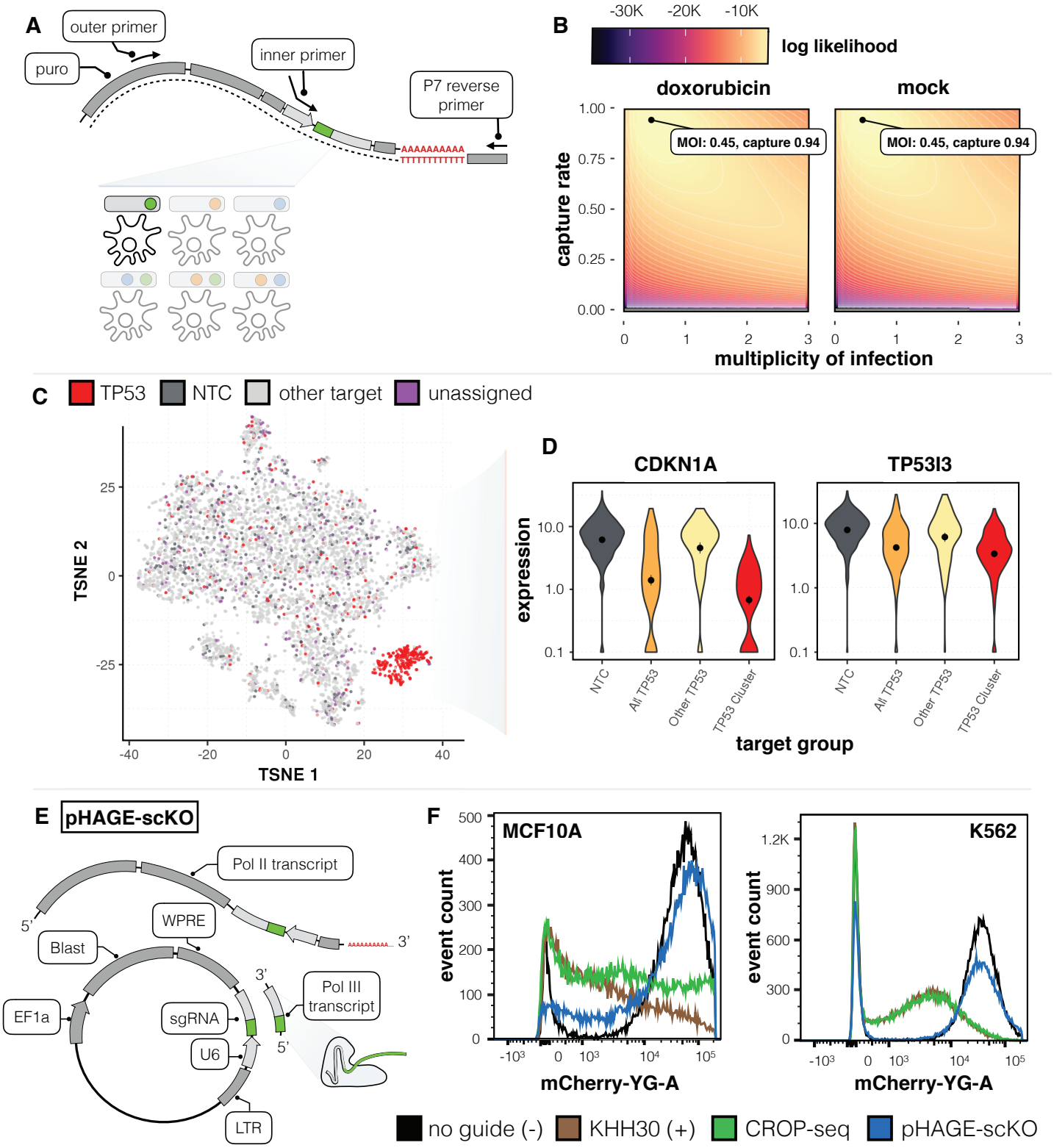

Figure 2 CROP-Seq screen of tumor suppressors with high capture rate by PCR enrichment, and assessment of alternate sgRNA placement within a pol II 3'UTR . **A)** Schematic of PCR enrichment of barcoded transcripts from CROP-seq samples. **B)** Determination of the most likely multiplicity of infection and capture rate of barcoded transcripts based on a generative model. **C)** tSNE embedding of a doxorubicin treated sample with colors corresponding to cells with guides to *TP53*, cells that contain non-targeting controls (NTC), cells containing guides to non-*TP53* targets, and cells that are unassigned. **D)** *CDKN1A* and *TP53I3* expression in cells expressing either non-targeting controls or guides to *TP53*. Cells with *TP53* guides are further stratified into cells inside and outside of the *TP53* enriched cluster from panel 2C. **E)** Schematic of pHAGE design with sgRNA placed upstream of the LTR. **F)** CRISPRi knock-down of mCherry in MCF10A and K562 cells not expressing a guide (- control), KHH30 (+ control), CROP-seq, and pHAGE-scKO design. All vectors contain a CRISPRi optimized backbone in this experiment.

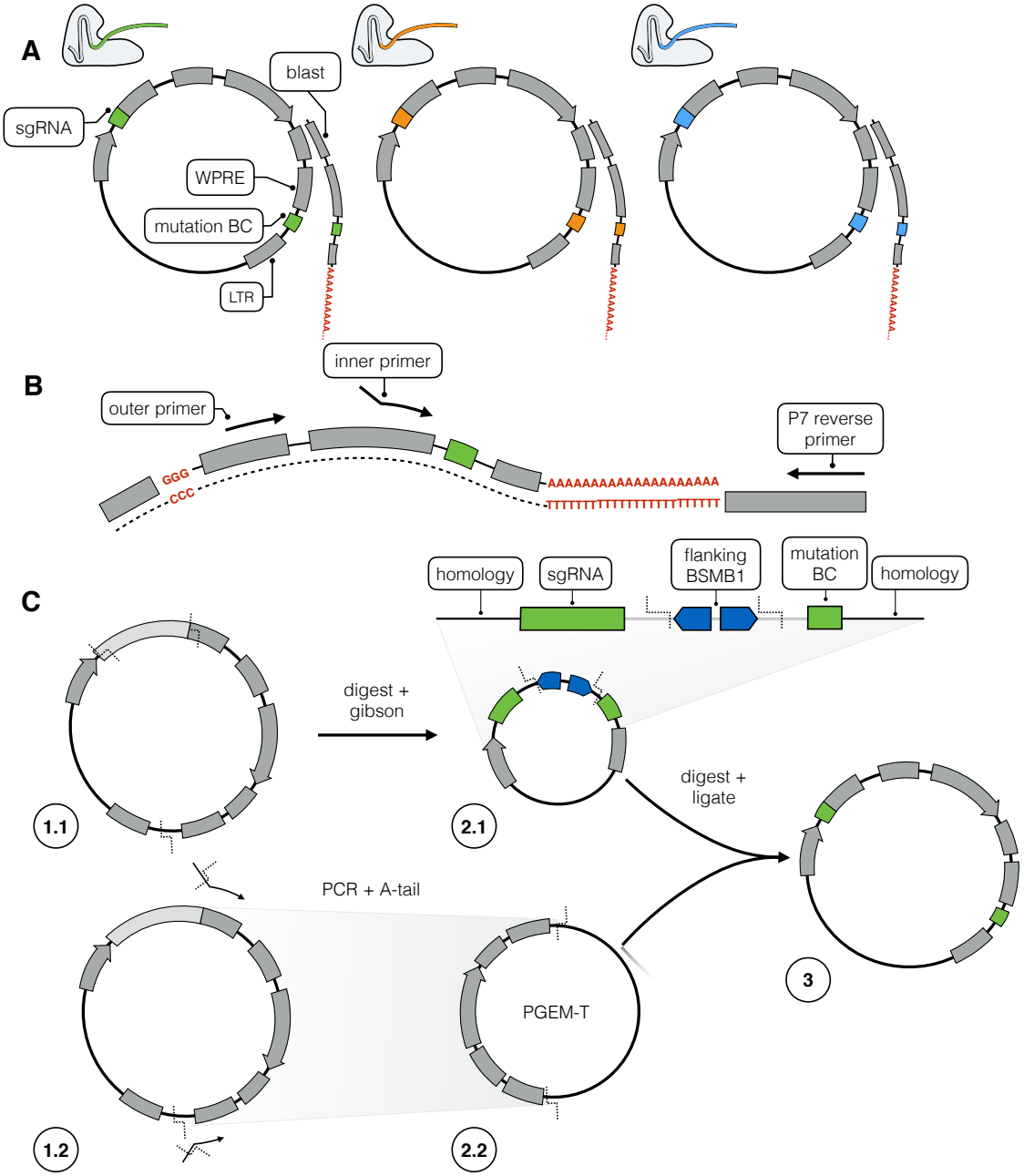

Figure S1 Diagram of cloning protocol and barcoded transcript enrichment strategy relying on *cis* pairing of sgRNAs and barcodes (pLGB-scKO). **A)** Schematic of our final vector relying on *cis* pairing of an sgRNA and a distal barcode. **B)** Strategy for PCR enrichment of barcoded transcripts from single-cell RNA-seq data. **C)** Pooled cloning protocol. In 1.1 we start with pLentiguideBlast and digest near the final locations of the sgRNA and paired barcode. In 2.1 an engineered library of oligos containing programmed pairs of sgRNAs and corresponding barcodes are inserted into the digested vector. In 1.2 a portion of pLentiguideBlast is amplified. In 2.2 this fragment is cloned into PGEM-T. Finally, in step 3 vectors resulting from 2.1 and 2.2 are digested with Bsmb1 and the insert from 2.2 is ligated into the backbone in 2.1 to produce the final library of sgRNAs and paired barcodes.

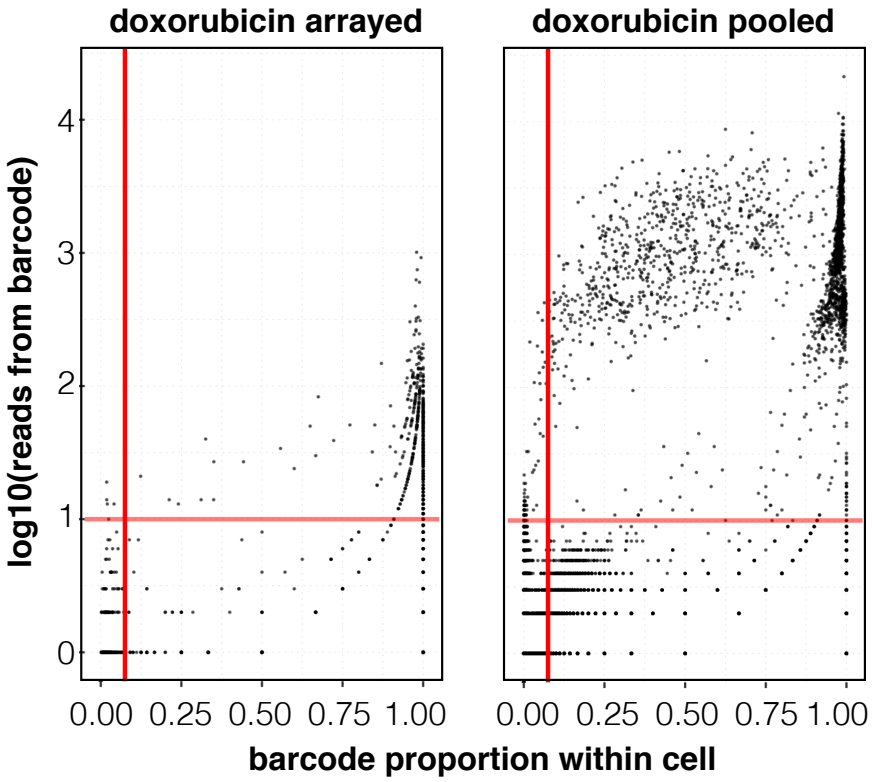

Figure S2 Barcoded transcript enrichment quality control for arrayed and pooled pLGB-scKO experiments. Each dot represents a barcode sequence observed in a given cell. Plot of reads for a given barcode against the proportion of all barcode reads observed in a given cell for every barcode/cell pair. Red lines indicate the lower-bounds used to distinguish noise from true barcode observations (10 reads and 0.075 proportion within cell). All barcodes observed above the red lines are assigned to their respective cells. Left, doxorubicin treated sample from arrayed experiment. Right, Doxorubicin treated sample from pooled experiment.

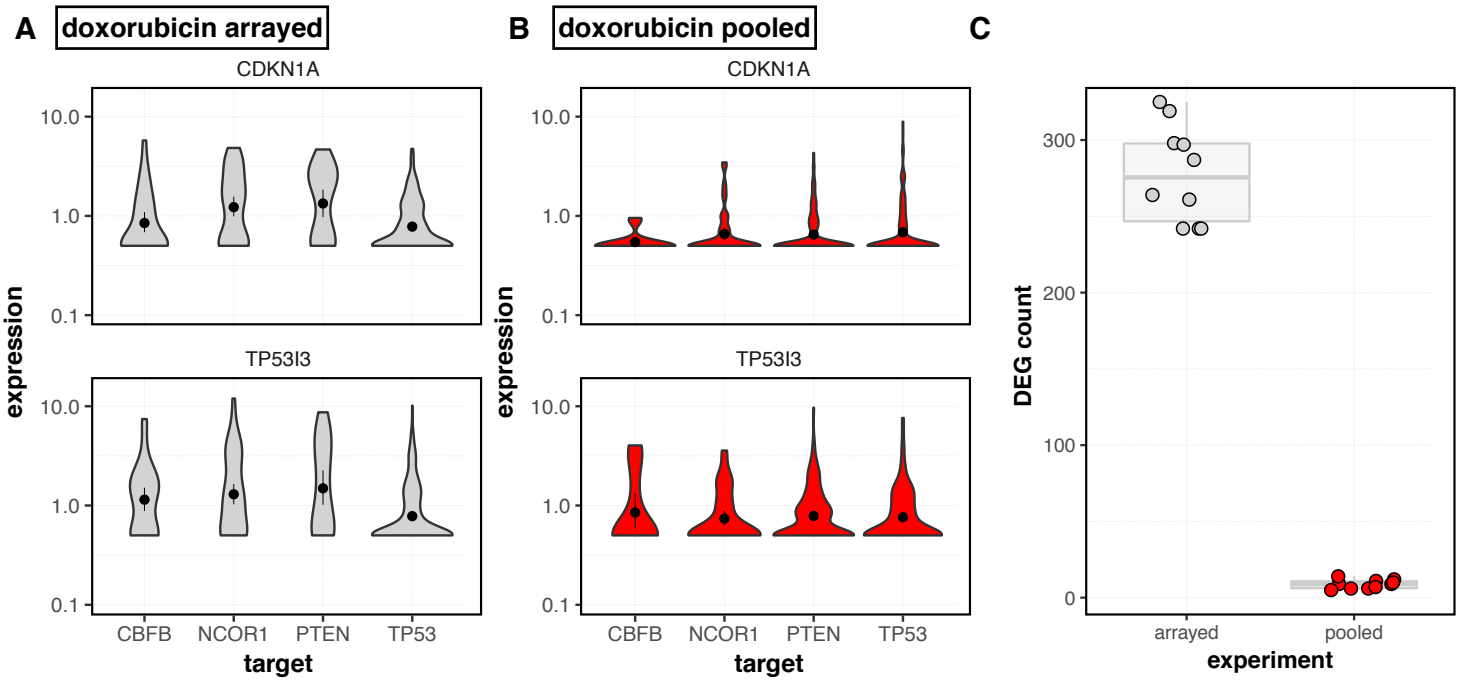

Figure S3 Comparison of a screen performed with arrayed and pooled lentivirus production using a vector that relies on *cis* pairing of sgRNAs and barcodes. Experiments were performed at different times but under the same conditions. The arrayed experiment was performed as a pilot experiment with 4 targets and observed an overall low rate of cells with detected barcodes. The pooled experiment was performed afterwards with 10 targets and a set of non-targeting controls and we observed a high proportion of cells with detected barcodes and good coverage of the library. To compare these experiments, only the four overlapping targets were considered and the number of cells containing an sgRNA to each target and sequencing depth were matched between samples to control for power differences. **A)** Size-factor normalized *CDKN1A* and *TP53I3* expression across *TP53* and the three other targets in arrayed screen. **B)** *CDKN1A* and *TP53I3* expression across *TP53* and three other targets in pooled screen that overlap with the arrayed screen. **C)** Comparison of the number of differentially expressed genes detected at an FDR of 5% for arrayed across the target label in the arrayed and pooled experiments.

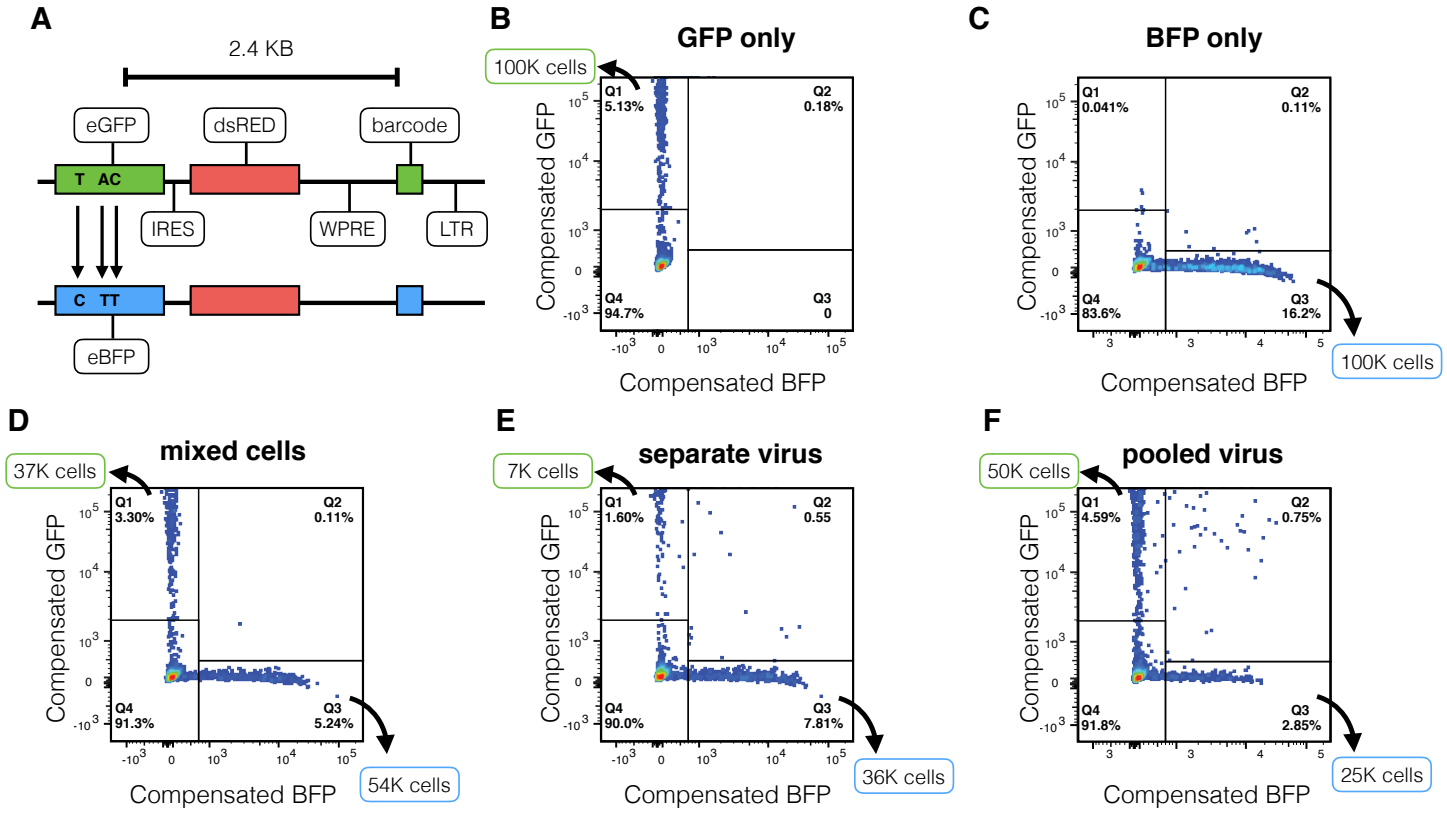

Figure S4 Design and sorting of GFP and BFP positive fractions in lentivirus barcode swapping experiment. **A)** Schematic of vectors (pHAGE-GFP and pHAGE-BFP) designed to quantify template switching rate at 2.4 kb using a FACS readout. FACS plots are shown for sorted cells in samples corresponding to **B)** GFP only transduced cells **C)** BFP only transduced cells **D)** GFP and BFP only transduced cells mixed just prior to FACS as a control **E)** cells transduced with BFP and GFP virus that was generated separately but pooled prior to transduction **F)** cells transduced with BFP and GFP virus that was generated from pooled plasmids. The fraction of green plasmids assumed in the determination of lentivirus swap rate from FACS experiments is taken as the fraction of GFP+ cells relative to the total GFP+ and BFP+ cells from this sort (4.59 / (4.59 + 2.85) or 61.7%). This accounts for the fact that plasmids were likely not completely equimolar. The approximate number of total cells cells sorted in each fraction is indicated along the appropriate axes on each plot.

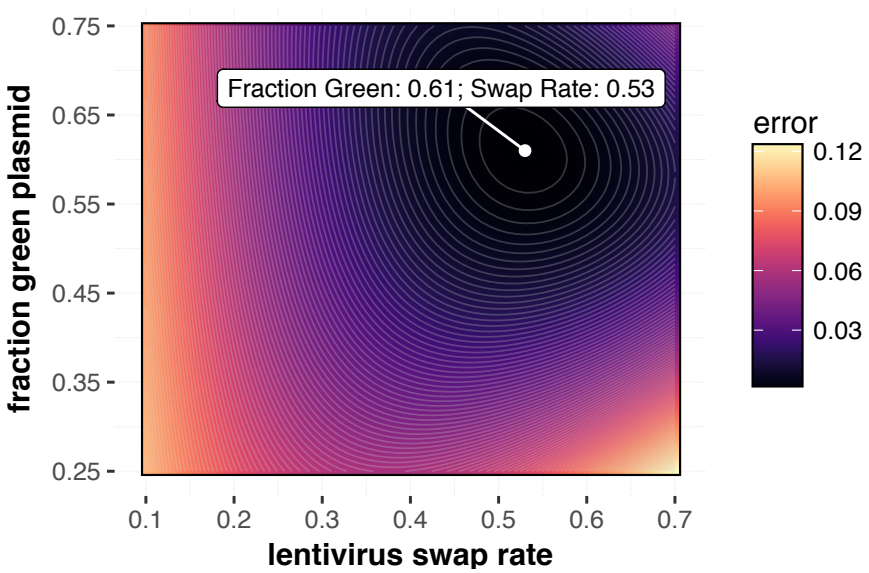

Figure S5 Simulation of concordance between observed and expected data obtained from FACS experiment in **Fig. S4** to quantify template switching rate at 2.4 kb separation between paired sequences. **Fig. 1F** assumed a fraction of 0.617 of GFP plasmid in the original green plasmid / blue plasmid mix as determined from FACS in **Fig. S5**. In this figure, both the fraction of GFP plasmid and lentivirus swap rate are varied to obtain the set of parameters that best fit the collected data. The sum of the squared error between expected and observed values from FACS given each combination of parameters is shown.

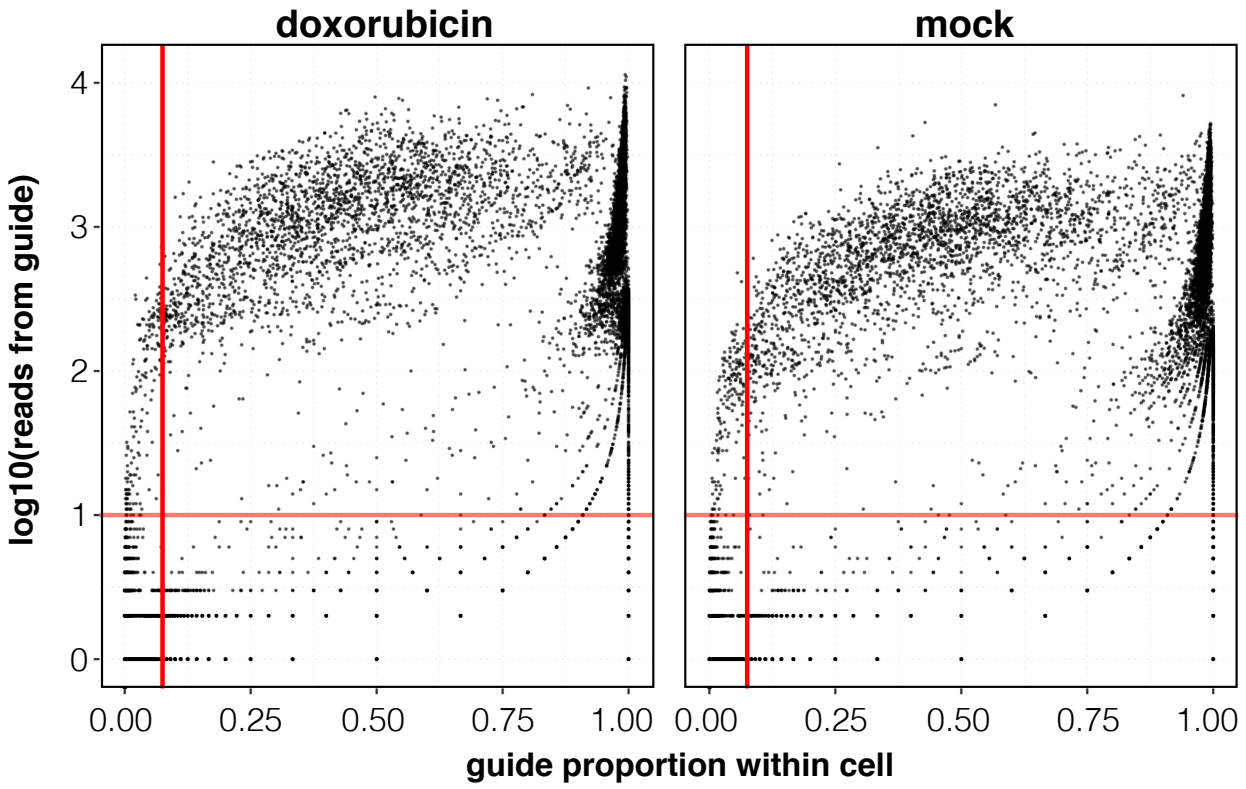

Figure S6 Guide transcript enrichment quality control plot for tumor suppressor knock-out screen performed with CROP-seq. Each dot represents a guide sequence observed in a given cell. Plot of reads for a given guide against the proportion of all guide reads observed in a given cell for every barcode/cell pair. Red lines indicate the lower-bounds used to distinguish noise from true guide observations (10 reads and 0.075 proportion within cell). All guide observed above the red lines are assigned to their respective cells. Left, Doxorubicin treated sample from CROP-seq experiment. Right, Mock sample from CROP-seq experiment.

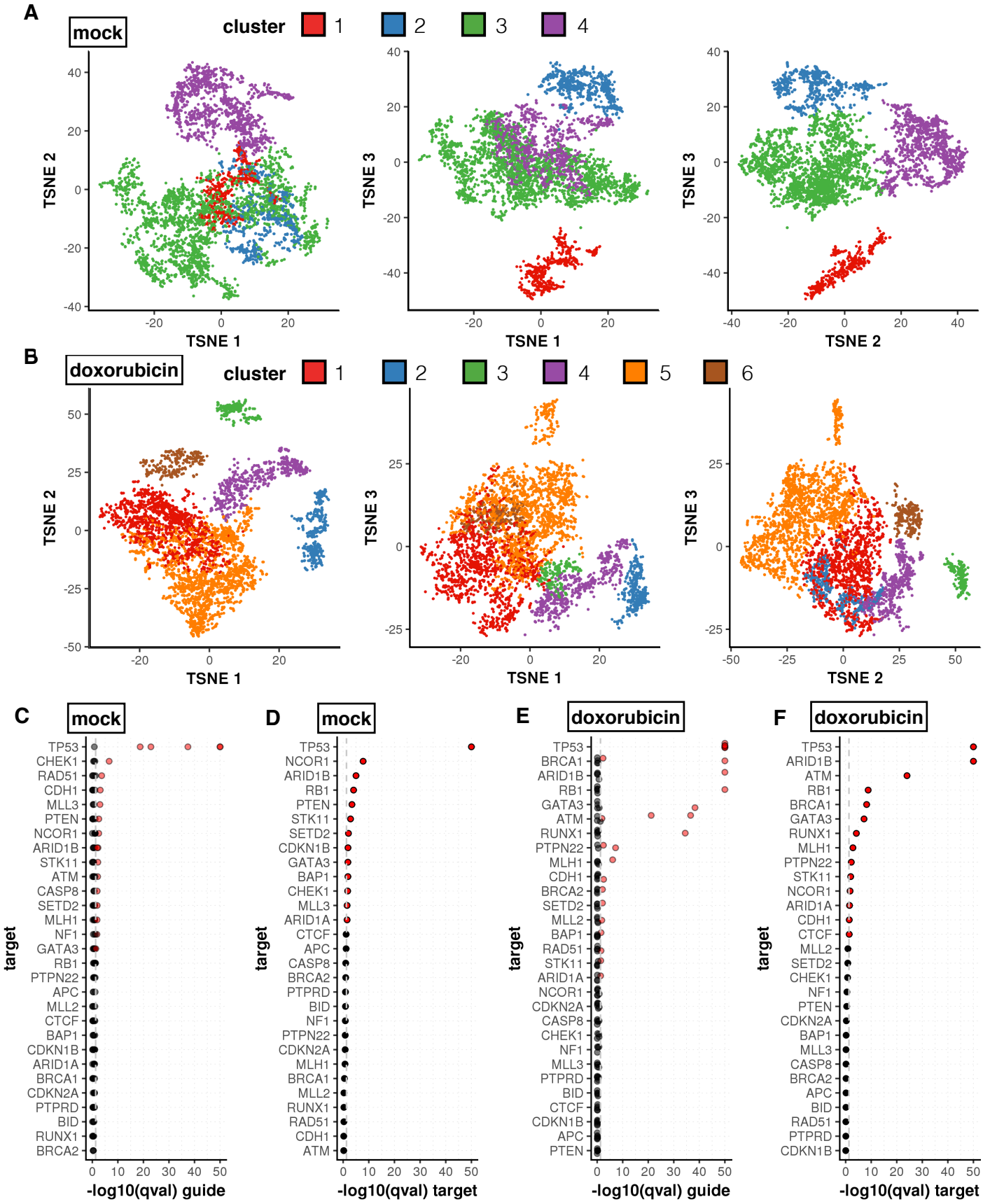

Figure S7 Loss of several targets alter the distribution of mock and doxorubicin exposed cells within tSNE clusters. **A-B)** 3D tSNE embedding and clustering of mock and doxorubicin treated samples, respectively. **C-F)** Chi-squared test qvalues (p values adjusted using the Benjamini-Hochberg method) resulting from testing for differences in the distribution of targets in our screen at both the individual sgRNA **(C and E)** and overall target levels **(D and F).** Comparisons are relative to the distribution of non-targeting controls across tSNE clusters for mock and doxorubicin treated samples, respectively (qvalues were capped to 1e-50 for visualization). Significant differences below a qvalue of 0.05 are colored in red (boundary marked by the grey dashed line).

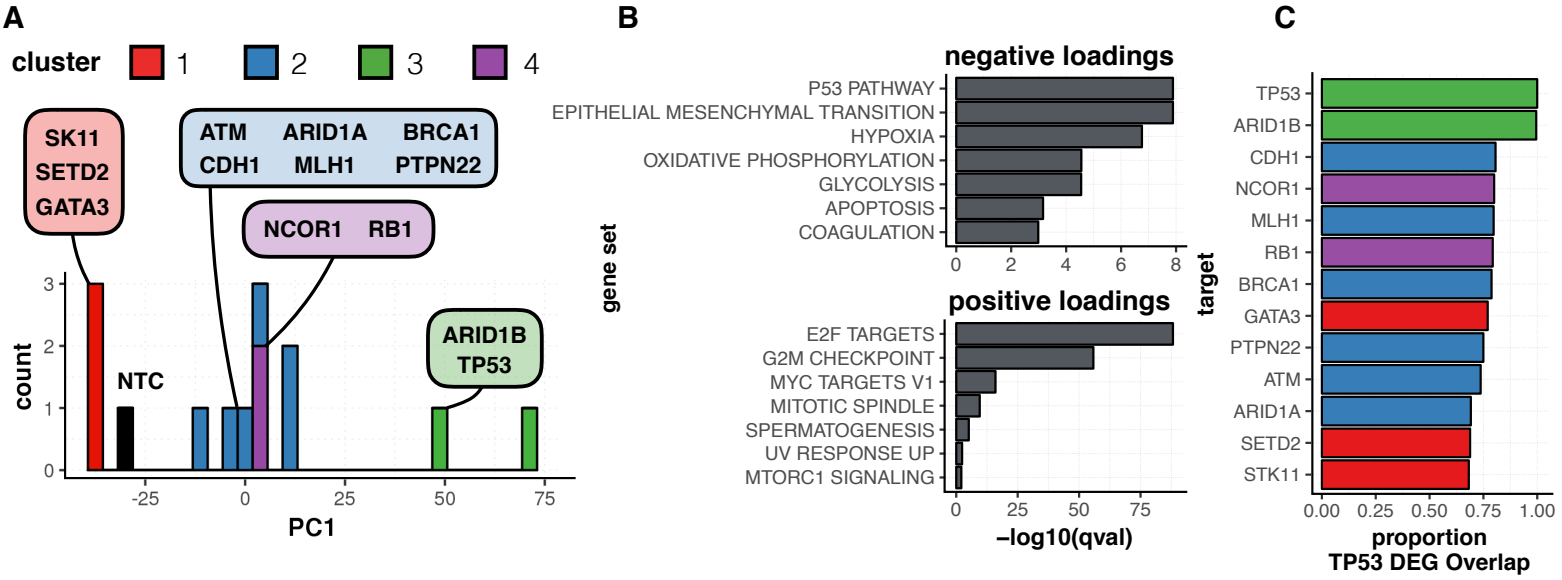

Figure S8 Enriched target-cluster pairs highlight tumor suppressors that share various degrees of a *TP53* deficient signature **A)** Fisher's exact with weights applied to guides according to an expectation maximization procedure were performed for the doxorubicin treated sample to find clusters from Fig. S7 panel B were particular targets were found to be enriched. Cells with target-cluster pairs that showed enrichment were used to generate an aggregate expression profile for every target within genes that are differentially expressed between *TP53* and non targeting controls (NTC). A PCA was performed on these average expression profiles and a distribution of targets across PC1 is shown colored by the cluster in which they were found to be enriched. **B)** Gene set enrichments for top positively and negatively loaded (less than -0.02 or greater than 0.02) genes along PC1 (qval < 0.01). **C)** Differential expression tests were performed for cells within each enriched target-cluster pair, comparing each target to all NTC cells. The proportion of overlap between these differentially expressed genes and the genes differentially expressed between *TP53* and NTC is shown.

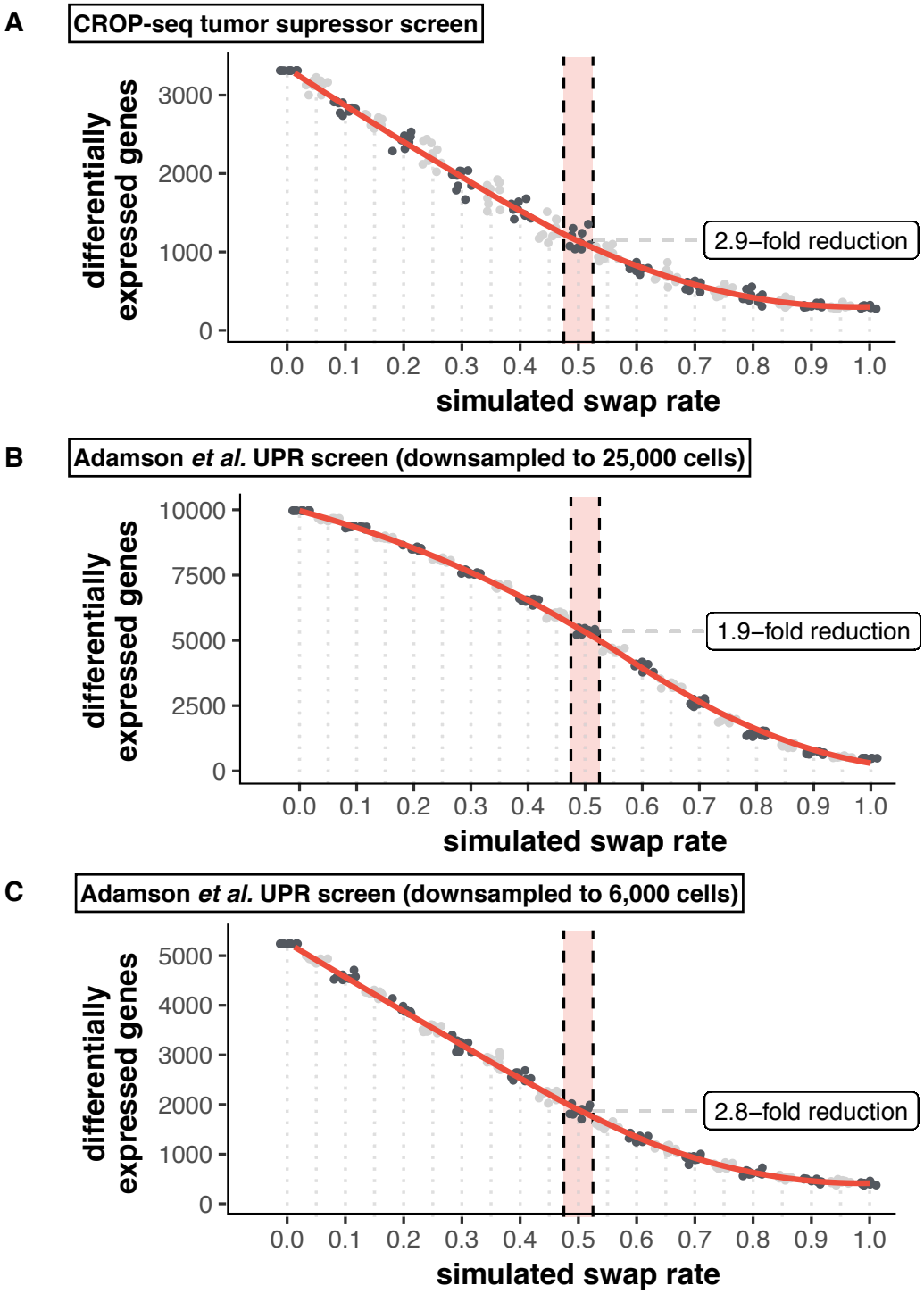

Figure S9 Swap rate simulations for our own CROP-seq tumor suppressor screen and the unfolded protein response screen from Adamson *et al.* Each dataset was subjected to simulation of progressively higher fractions of target assignment swapping to mimic the impact of template switching. Number of differentially expressed genes across the target label at FDR of 5% is plotted at each swap rate. 0.5 corresponds to the 50% swap rate determined via FACS. **A)** CROP-seq tumor suppressor screen from our study. **B)** Unfolded protein response screen from Adamson et al. downsampled from ~50,000 to 25,000 cells to make simulations computationally feasible. **C)** Unfolded protein response screen from Adamson *et al.* downsampled to 6,000 cells to illustrate how reduced power impacts the observed impact from simulated swapping.