In vivo profiling of metastatic double knockouts through CRISPR-Cpf1 screens

Ryan D. Chow1,2,3,4,9, Guangchuan Wang1,2,3,9, Lupeng Ye1,2,3,9, Adan Codina1,2,3,5, Hyunu Ray Kim1,2,3, Li Shen1,2, Matthew B. Dong1,2,3,4,6, Youssef Errami1,2,3 and Sidi Chen1,2,3,4,5,6,7,8*

Systematic investigation of the genetic interactions that influence metastatic potential has been challenging. Here we developed massively parallel CRISPR-Cpf1/Cas12a crRNA array profiling (MCAP), an approach for combinatorial interrogation of double knockouts in vivo. We designed an MCAP library of 11,934 arrays targeting 325 pairwise combinations of genes implicated in metastasis. By assessing the metastatic potential of the double knockouts in mice, we unveiled a quantitative landscape of genetic interactions that drive metastasis.

Metastasis, the major lethal factor of solid tumors, is a complex, multi-step process. A systems-level understanding of the genetic interactions that influence metastatic potential is lacking, as library-throughput screening of DKOs. We focused on genes significantly mutated in a human metastasis cohort (MET500) and the m-malian species has proved challenging. The type V CRISPR system Cpf1 (also known as Cas12a) has empowered simultaneous genome editing at multiple loci. Because Cpf1 does not require a transactivating CRISPR RNA (tracrRNA), multiplexed genome editing can be achieved with a single crRNA array. This characteristic inspired us to develop Cpf1 as a system for interrogating genetic interactions in vivo, with substantial advantages in library design, readout and analysis compared to Cas9-based approaches.

We first established a CRISPR–Cpf1 lentiviral system for characterization of double knockouts in a cancer cell line (KPD). To evaluate the cellular diversity that can be accommodated in vivo, we cloned in a library of random 8-mers (Supplementary Fig. 1). To map the metastatic potential of the MCAP-MET library, we injected the cell pool subcutaneously into nu/nu mice (4 × 106 cells per mouse, × 350 coverage) (n = 10). At this coverage, each BC array is represented by an average of ~23 cells following injection. After 6 weeks we collected the primary tumors (n = 10) and lung lobes (n = 37), and performed crRNA array sequencing. Using the BC array data, we assessed the dynamics of selection in our metastasis model. We chose a 0.001% cutoff by considering the distribution of BC array frequencies in cell samples and quantified the number of ‘clones’ (approximated by BC arrays) per sample, finding clear evidence of progressive selection as the cell pools formed primary tumors and lung metastases (Supplementary Fig. 4). These results were consistent at a frequency cutoff of ≥0.01% (Supplementary Fig. 5). Collectively, the clone-level analyses illustrate the progressive selection pressures on cells as they form primary tumors and metastasize to the lung.

We next considered the data in terms of the 11,934 dual-crRNA arrays (Supplementary Fig. 6 and Supplementary Table 3). Using the 1,326 NTC-NTC arrays as an empirical null distribution, we identified crRNA arrays enriched at a false discovery rate (FDR) < 0.5% in each sample (Supplementary Fig. 7a). We tabulated the percentage of arrays for a given genetic perturbation that were enriched in at least one sample (Supplementary Fig. 7b,c). No single genes had more than 40% of their SKO arrays enriched in lung metastases. In contrast, 62.5% of all arrays targeting the Nf2_Rb1 pair were enriched in at least one lung metastasis, with 56.25% of arrays enriched for Nf2_Trim72 (Supplementary Fig. 7d–f).

We quantitatively determined the metastatic potential of the various perturbations represented in the MCAP-MET library (Supplementary Fig. 8a–c). To identify specific perturbations exhibiting strong selection in vivo, we averaged the crRNA arrays for each gene pair is represented by 16 DKO constructs and each gene is represented by 208 SKO constructs. Additionally, we appended a random 10-mer barcode for clonal analyses. Deep sequencing confirmed complete coverage of the library, and analysis of the 10-mer barcodes revealed the diversity of barcoded crRNA arrays (BC arrays) (n = 774,295) (Supplementary Fig. 3a–d).

We generated lentiviral pools from the MCAP-MET plasmid library and infected Cpf1+ KPD cells (Fig. 1b). Seven and fourteen days after transduction, we sequenced the crRNAs in the cell pool, and found strong concordance with the plasmid library (Supplementary Fig. 3e). In each cell pool we recovered 172,427 ± 2,391 (mean ± s.e.m.) unique BC arrays. To map the metastatic potential of the MCAP-MET library, we injected the cell pool subcutaneously into nu/nu mice (4 × 106 cells per mouse, × 350 coverage) (n = 10). At this coverage, each BC array is represented by an average of ~23 cells following injection. After 6 weeks we collected the primary tumors (n = 10) and lung lobes (n = 37), and performed crRNA array sequencing. Using the BC array data, we assessed the dynamics of selection in our metastasis model. We chose a 0.001% cutoff by considering the distribution of BC array frequencies in cell samples and quantified the number of ‘clones’ (approximated by BC arrays) per sample, finding clear evidence of progressive selection as the cell pools formed primary tumors and lung metastases (Supplementary Fig. 4). These results were consistent at a frequency cutoff of ≥0.01% (Supplementary Fig. 5). Collectively, the clone-level analyses illustrate the progressive selection pressures on cells as they form primary tumors and metastasize to the lung.

We next considered the data in terms of the 11,934 dual-crRNA arrays (Supplementary Fig. 6 and Supplementary Table 3). Using the 1,326 NTC-NTC arrays as an empirical null distribution, we identified crRNA arrays enriched at a false discovery rate (FDR) < 0.5% in each sample (Supplementary Fig. 7a). We tabulated the percentage of arrays for a given genetic perturbation that were enriched in at least one sample (Supplementary Fig. 7b,c). No single genes had more than 40% of their SKO arrays enriched in lung metastases. In contrast, 62.5% of all arrays targeting the Nf2_Rb1 pair were enriched in at least one lung metastasis, with 56.25% of arrays enriched for Nf2_Trim72 (Supplementary Fig. 7d–f).

We quantitatively determined the metastatic potential of the various perturbations represented in the MCAP-MET library (Supplementary Fig. 8a–c). To identify specific perturbations exhibiting strong selection in vivo, we averaged the crRNA arrays

*Corresponding author. E-mail: sidi.chen@yale.edu
for each SKO or DKO condition on a sample-by-sample basis, then aggregated the data by sample type. To pinpoint the perturbations with the strongest selective advantage out of the entire MCAP-MET library, we used all targeting genes/pairs for linear regression modeling. The top gene pairs favored in primary tumors relative to cell pools (outlier test, adjusted P < 0.05) included Nf2_Trim72, Nf2_Chd1, Nf2_Pten, Nf2_Arid1b, Nf2_Kdm6a and Nf2_Rb1 (Fig. 1c). A similar set of gene pairs were enriched in lung metastases compared to cell pools (Supplementary Fig. 8d). Comparing primary tumors to lung metastases, Nf2_Trimg72 and Nf2_Chd1 emerged as the top metastasis-driving mutation pairs (Fig. 1d).

Our analyses suggested that certain gene pairs may be synergistic in promoting metastasis. To identify such mutation combinations, we first identified gene pairs that were significantly more abundant than their respective single gene counterparts (two-sided Wilcoxon rank sum test, adjusted P < 0.05). Because the effects of a mutation combination may simply be additive rather than synergistic, we calculated a synergistic coefficient (SynCo = DKO/NTC – SKO/NTC) for each gene pair (Fig. 2a and Supplementary Table 4). Collectively, we found six DKO s that were significantly more abundant than the corresponding SKOs and with a SynCo > 0: Nf2_Trimg72, Chd1_Nf2, Chd1_Kdm6a, Jak1_Kdm6a, Kdm6a_Pten and Nf2_Pten (Fig. 2b–d and Supplementary Fig. 9). These data were summarized as a library-wide map of the selective advantage of each DKO relative to the corresponding SKOs (Supplementary Fig. 10 and Supplementary Table 5). Some of these synergistic interactions are recapitulated in human cohorts (Supplementary Fig. 11 and Supplementary Table 6).

We then sought to validate the metastatic potential of the strongest gene pair identified in the screen, Nf2_Trimg72. After cloning in five different dual-crRNA arrays with combinations of Rosa26-targeting crRNAs or the top-performing Nf2 and Trimg72 crRNAs from the screen (Rosa26 + Rosa26, Nf2 + Rosa26, Trimg72 + Rosa26, Nf2 + Trimg72 or Trimg72 + Nf2), we assessed mutation efficiency 7 d after lentiviral transduction (n = 5 infection replicates each) and confirmed that array configuration did not influence mutation efficiency (Fig. 3a and Supplementary Fig. 12a). To exclude the possibility that the Nf2_Trimg72 gene pair may have undergone positive selection in vitro before injection, we characterized the 5-ethynyl-2’-deoxyuridine (EdU) incorporation of KPD cells expressing Rosa26 + Rosa26, Nf2 + Rosa26, Trimg72 + Rosa26 or Nf2 + Trimg72 dual-crRNA arrays, and found no significant differences (n = 3 cell replicates each) (Supplementary Fig. 12b–c).

To interrogate the metastatic potential of the Nf2_Trimg72 gene pair, we first performed in vitro Matrigel invasion assays (n = 3 independent experiments) and found that Nf2 + Trimg72 infected cells were more invasive than Rosa26 + Rosa26, Nf2 + Rosa26 or Trimg72 + Rosa26 infected cells (Supplementary Fig. 12d,e). We then proceeded to validate the Nf2_Trimg72 gene pair in vivo, transplanting 1.8×10^6 cells into nu/nu mice (n = 8 mice for each condition). Primary tumors in the Nf2 + Trimg72 group grew significantly larger than Nf2 + Rosa26, Trimg72 + Rosa26 or Rosae + Rosa26 tumors (Fig. 3b). We followed the development of metastasis by luciferase live imaging and, 28 d after the initial transplantation, we harvested the primary tumors and lungs (Supplementary Fig. 13). Mice bearing Nf2 + Trimg72 tumors had significantly more metastatic lung nodules than mice bearing Nf2 + Rosa26, Trimg72 + Rosa26 or Rosa26 + Rosa26 tumors (Fig. 3c). Collectively, these data point to specific mutation combinations with heightened metastatic potential in vivo, and highlight the power of MCAP for high-throughput interrogation of genetic interactions in challenging biological systems.

Several high-throughput double perturbations have been performed in mammalian cells using RNA interference or CRISPR–Cas9 technologies^8–16. However, the dependence of Cas9 on a tracrRNA predicates the need for multiple single-guide RNA (sgRNA) cassettes for combinatorial knockouts, thus complicating library design, cloning, readout and analysis. In comparison, MCAP offers a streamlined approach for double or even higher-order knockout/perturbation screens, with the potential for sequential screens using invertible dual-crRNA arrays. A remaining challenge that limits the broader utility of MCAP is the mutation efficiency of Cpf1, as this necessitates positive selection screens using redundant library designs with several independent constructs representing each perturbation. Of note, progress has been made toward predicting crRNAs that can induce mutations at higher efficiencies^14,15, and Cpf1 itself has been engineered to increase its activity and targeting range^16.
**Fig. 2 | Identification of synergistic mutation combinations.** a. Schematic for calculating the synergy coefficient score (SynCo) and identifying synergistic mutation combinations. For a given gene pair NM, SynCo is defined as DKO_{NM} – SKO_{N} – SKO_{M}. A positive SynCo value indicates that the selective advantage of the gene pair is greater than that of the two individual genes combined. b, c. Scatter plot of –log_{10} adjusted P values for each gene pair (two-sided Wilcoxon rank sum test) (b) and median differential abundance compared to the corresponding single genes (c), in lung metastases (n = 37 from ten mice). Synergistic gene pairs are highlighted in purple. d, Tukey box plots (interquartile range (IQR) boxes with 1.5 × IQR whiskers and notched 95% confidence interval of median) detailing the abundances of Nf2, Trim72 or Nf2, Trim72 arrays in lung metastases (n = 37 from ten mice), with associated two-sided Wilcoxon rank sum P values and SynCo scores noted. Statistics are in reference to Nf2, Trim72 (purple) and colored according to the corresponding SKO conditions (green and orange). Adj., adjusted.

**Fig. 3 | Nf2 and Trim72 mutations jointly promote lung metastasis in vivo.** a. Quantification of T7 endonuclease-1 (T7EI) assays (n = 5 infection replicates each) for Nf2 and Trim72 (mean ± s.e.m.). Nf2 locus: Nf2 + Rosa6 versus Nf2 + Trim72, P = 0.1098; Nf2 + Trim72 versus Trim72 + Nf2, P = 0.6110. Trim72 locus: Trim72 + Rosa6 versus Nf2 + Trim72, P = 0.7450; Nf2 + Trim72 versus Trim72 + Nf2, P = 0.8386. The order of each crRNA within the array is indicated in the array names (that is, Nf2, Nf2). Statistical significance was assessed by two-sided unpaired Welch’s t-test. b. Growth curves of primary tumors derived from cells transduced with Rosa26 + Rosa26, Nf2 + Rosa26, Trim72 + Rosa26 or Nf2 + Trim72 crRNA arrays (mean ± s.e.m.) (n = 8 mice for each condition). Nf2 + Trim72 versus Nf2 + Rosa26, Trim72 + Rosa26 and Rosa26 + Rosa26: P = 0.0396, P = 0.0026 and P = 1.483 × 10^{-5}, respectively. Statistical significance was assessed by two-way analysis of variance (ANOVA). c, Quantification of lung metastases in mice bearing Rosa26 + Rosa26, Nf2 + Rosa26, Trim72 + Rosa26 or Nf2 + Trim72 primary tumors at 28 d post-infection. Data are shown in terms of the number of nodules found in each lung lobe (mean ± s.e.m.) (n = 4–5 lung lobes per mouse, with eight mice for each condition). Nf2 + Trim72 versus Nf2 + Rosa26, Trim72 + Rosa26 and Rosa26 + Rosa26: P = 0.0328, P = 4.263 × 10^{-4} and P = 1.054 × 10^{-4}, respectively. Nf2 + Rosa26 versus Rosa26 + Rosa26 and Trim72 + Rosa26: P = 5.091 × 10^{-5} and P = 8.990 × 10^{-5}, respectively. Trim72 + Rosa26 versus Rosa26 + Rosa26, P = 0.0016. Statistical significance was assessed by two-sided unpaired Welch’s t-test. NS, not significant; *P < 0.05, **P < 0.01, ***P < 0.001.
The MCAP method can be readily applied to different cell types, biological processes and disease models and thus represents a tool for mapping genetic interactions in mammalian species in vivo with unparalleled simplicity and throughput.

Online content
Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41592-019-0371-5.

Received: 18 June 2018; Accepted: 5 March 2019; Published online: 8 April 2019

References
1. Lambert, A. W., Pattabiraman, D. R. & Weinberg, R. A. Cell 168, 670–691 (2017).
2. Schunder, E., Rydzewski, K., Grunow, R. & Heuner, K. Int. J. Med. Microbiol. 303, 51–60 (2013).
3. Zetsche, B. et al. Cell 163, 759–771 (2015).
4. Zetsche, B. et al. Nat. Biotechnol. 35, 1–4 (2017).
5. Chen, S. et al. Cell 160, 1246–1260 (2015).
6. Kamar, M. S. et al. Genes Dev. 23, 2700–2704 (2009).
7. Robinson, D. R. et al. Nature 548, 297–303 (2017).
8. Wong, A. S. L., Choi, G. C. G., Cheng, A. A., Purcell, O. & Lu, T. K. Nat. Biotechnol. 33, 952–961 (2015).
9. Wong, A. S. L. et al. Proc. Natl Acad. Sci. USA 113, 2544–2549 (2016).
10. Rogov, A. et al. Nat. Methods 10, 432–437 (2013).
11. Han, K. et al. Nat. Biotechnol. 35, 463–474 (2017).
12. Shen, J. P. et al. Nat. Methods 14, 573–576 (2017).
13. Horlbeck, M. A. et al. Cell 174, 953–967 (2018).
14. Najm, F. J. et al. Nat. Biotechnol. 36, 179–189 (2018).
15. Erard, N., Knott, S. R. V., & Hannon, G. J. Mol. Cell 67, 348–354 (2017).
16. Boettcher, M. et al. Nat. Biotechnol. 36, 170–178 (2018).
17. Chow, R. D., Kim, H. R. & Chen, S. Nat. Commun. 9, 1903 (2018).
18. Kim, H. K. et al. Nat. Methods 14, 153–159 (2017).
19. Kim, H. K. et al. Nat. Biotechnol. 36, 239–241 (2018).
20. Kleinsteiver, B. P. et al. Nat. Biotechnol. 37, 276–282 (2019).

Acknowledgements
We thank all members of the Chen laboratory, as well as various colleagues in the Department of Genetics, Systems Biology Institute, Immunobiology Program, BBS Program, MSTP Program, Comprehensive Cancer Center and Stem Cell Center at Yale, for their assistance and scientific discussion. We thank the Center for Genome Analysis, Center for Molecular Discovery, Pathology Tissue Services, Histology Services, High Performance Computing Center, West Campus Analytical Chemistry Core and West Campus Imaging Core and Keck Biotechnology Resource Laboratory at Yale, for technical support. S.C. is supported by Yale SIBI/Genetics Startup Fund, Damon Runyon Drey-Fere Award (No. DFS-13-15), Melanoma Research Alliance (No. 412806 and 16-003524), St-Baldrick’s Foundation (No. 426685), Breast Cancer Alliance, Cancer Research Institute, AACR (No. 499395 and 17-20-01-CHEN), The Mary Kay Foundation (No. 017-81), The V Foundation (No. V2017-022), Ludwig Family Foundation, DoD (No. W81XWH-17-1-0235), Sontag Foundation (DSA Award), Chengert Family Foundation and NIH/NCI (No. 1DF2CA238295-01, 1R01CA231112-01, 1U54CA209992-3697, 5P50CA196530-A10805 and 4P50CA121974-A08306). R.D.C. and M.B.D. are supported by the Yale MSTP training grant from NIH (No. T32GM007205). G.W. is supported by CRI Irvington and RJ Anderson Postdoctoral Fellowships. A.C. is supported by a Yale PhD training grant from NIH (No. T32GM007223).

Author contributions
R.D.C. designed the study, performed experiments, developed statistical algorithms and computational pipelines and analyzed the data. G.W. and L.Y. performed molecular, cellular and animal experiments and MCAP readout. A.C. optimized in vivo screens and assisted with validation experiments. M.B.D. assisted with flow cytometry experiments. Y.E. assisted with in vitro assays. S.C. conceived the study, provided conceptual advice, secured funding and supervised the work. R.D.C. and S.C. wrote the manuscript.

Competing interests
The authors have filed a provisional patent related to this work.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41592-019-0371-5.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to S.C.

Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2019
Methods

Additional information can be found in the associated Supplementary Protocol33, as well as the Nature Research Reporting Summary.

Animal work statements and institutional approval. All experimental work involving recombinant DNA was performed under the guidelines of the Yale University Environment, Health and Safety Committee under an approved protocol (No. Chen-rDNA-15-45). All animal work was performed under the guidelines of Yale University Institutional Animal Care and Use Committee with approved protocols (No. Chen-2015-20068 and Chen-2018-20068), and was consistent with the Guide for Care and Use of Laboratory Animals, National Research Council, 1996 (Institutional Animal Welfare Assurance, No. A-3125-01). Six- to eight-week-old mice, both males and females, were used for MCAP screen experiments. For subsequent validation experiments, only female mice were used.

Design of the MCAP-MET library. The top 23 annotated ‘tumor suppressors’ from the human METkooh cohort were compiled, and combined with three top hits from a previous mouse metastasis screen (Nj2, Trim72 and Ube2g3) for a final set of 26 genes. We then analyzed the complete exon sequences of these 26 genes to extract all possible Cpf1 spacers (that is, all 20-mers beginning with the Cpf1 protospacer-adjacent motif (PAM), 5′-TTTV). Each of these 20-mers was then reverse complemented and mapped to the entire mm10 reference genome by Bowtie 1.1.2 (ref. 3), with settings -n 2-118 -p 8 -a -y -best -e 90. After filtering out all alignments that contained mismatches in the final three base pairs (corresponding to the Cpf1 PAM) and disregarding any mismatches in the fourth to last base pair, we obtained the number of genome-wide alignments for each crRNA using all 0, 1 and 2 mismatch (mm) alignments. A total mismatch score (MM score) was calculated for each crRNA using the following formula: MM score = 0 mm × 1,000 + 1 mm × 50 + 2 mm × 1. We also counted the number of consecutive thymidines in each crRNA, and used the following formula: T score = 100 × (max_consecutive_Thymidines) (ref. 5). We then sorted all 20-mc RNAAs corresponding to each target gene by low MM score and high T score. Finally, the top four crRNAs for each gene were chosen. In the event of ties, crRNAs targeting constitutive exons and/or the first exon were prioritized.

Fifty-two NTC crRNAs were randomly selected from a pool of random 20-mers that did not map to the mouse genome with 2 mismatches. In combination with the 104 crRNAs targeting 26 genes, a total of 5,200 DKO, 5,408 SKO and 1,326 NTC-NTC arrays was designed for a total of 11,934 dual-crRNA arrays (MCAP-MET library). With a total pool of 26 genes, the number of possible unique combinations of two different genes is 325. Each of these 325 gene pairs was represented by 16 DKO arrays, whereas each single gene condition was represented by 208 SKO arrays. For SKO crRNA arrays, we placed each gene-targeting crRNA in the first position of the crRNA array and toggled the NTC crRNAs through the second position. For each gene pair, the positioning of the crRNAs representing each of the two genes was determined randomly. For each oligo, we appended a degenerate 10-mer (10× degenerate 8-mer) to each of the predicted crRNA cut sites to generate a library of degenerate 20-mers that did not map to the mouse genome with up to two mismatches.

Cell lines. A non-small-cell lung cancer cell line5 (KPD cell line) was transduced with the 8-mer lentiviral library and selected by puromycin; 4×10^6 KPD-8-mer cells were subcutaneously injected in either Rag1−/− or nsu/nmu mice. Twelve days after transfection, mice were killed and tumors were isolated for genomic preparation and readout.

MCAP in a mouse model of metastasis. Mouse transduction was performed with three infection replicates at high coverage and low multiplicity of infection. Briefly, according to viral titers, MCAP-MET lentiviruses were added to a total of 1×10^6 KPD-Cpf1 cells at a calculated multiplicity of infection of 0.2, and incubated for 24 h before replacing the virus-containing media with 3 μg/ml puromycin containing fresh media to select virus-transduced cells. Approximately 2.5×10^6 cells confer a <2,000-lb body size. MCAP library-transduced cells were cultured under the pressure of 3 μg/ml puromycin for 14 d before injection. MCAP library-transduced KPD-Cpf1 cells were injected subcutaneously into the right flank of nsu/nmu mice at 4×10^6 cells per flank (n=35–100 coverage per transplanted).

Mouse tumor dissector. Mice were killed by carbon dioxide asphyxiation followed by cervical dislocation. Tumors and lungs were manually dissected, fixed in 10% formalin for 24–96 h and transferred into 70% ethanol. Tissues were flash-frozen with liquid nitrogen and ground in a 3-ml frosted polyethylene vial set (No. 2240-PEF) in a 2010 GenoGrinder machine (SPEXSamplePrep).

Homogenized tissues were then used for DNA extraction.

Genomic DNA extraction. Frozen, ground tissue (200–800 mg) was resuspended in 6 ml of NK Lysis Buffer (50 mM Tris, 50 mM EDTA, 1% SDS, pH 8.0) supplemented with 30 μl of 20 mg/ml Proteinase K (Qiagen) in 15 ml-conical tubes, and incubated in a bath at 55 °C overnight. After all the tissues were lysed, 30 μl of 10 mg/ml RNase A (Qiagen) was added, mixed well and incubated at 37 °C for 30 min. Samples were chilled on ice and then 2 ml of pre-chilled 7.5 M ammonium acetate (Sigma) was added to precipitate proteins. The samples were inverted and vortexed for 15–30 s and then centrifuged at 2,400 g for 10 min. The supernatant was carefully decanted into a new 15 ml conical tube, followed by the addition of 6 ml of 100% isopropanol (at a ratio of ~0.7), inverted 30–50 times and centrifuged at 2,400 g for 10 min. At this point, genomic DNA became visible as a small white pellet. After the supernatant had been discarded, 6 ml of freshly prepared 70% ethanol was added, mixed well and then centrifuged at ≥2,000 g for 10 min. The supernatant was discarded by pouring, and remaining residue was removed using a pipette. After air-drying for 10–30 min, DNA was resuspended by the addition of 200–500 μl of nuclease-free H2O. Genomic DNA concentration was measured using a Nanodrop (Thermo Scientific) and normalized to 1,000 ng/ml for the following readout PCR.

MCAP library readout. MCAP library readout was performed using a two-step PCR approach. Briefly, in the first round of PCR, sufficient genomic DNA was used as template to guarantee coverage of the library abundance and representation. Each sample contained 12 μg of genomic DNA, split between six separate PCR reactions. For the first PCR, the sgRNA-containing region was amplified using primers specific to the MCAP library. Then the library was subjected to a second-round PCR, targeting the start position of each biological replicate. The second-round PCR products were then pooled, then 2 μl of well-mixed first-round PCR products were used as template for amplification, with sample-tracking barcode primers having thermocycling conditions set to 98 °C for 1 min, 15 cycles of (98 °C for 1 s, 62 °C for 5 s, 72 °C for 15 s) and 72 °C for 2 min.

Using Plushion Flash High Fidelity Master Mix (Thermo Fisher), the thermocycling parameters for PCR were 98 °C for 2 min, 35 cycles of (98 °C for 1 s, 62 °C for 5 s, 72 °C for 15 s) and 72 °C for 2 min.

Nextera XT library preparation was then performed according to the manufacturer’s protocol with minor modifications. Reads were mapped to the mm10 mouse genome using BWA11, with the settings bwa mem -t 8 -w 20. Indel variants were first processed with Samtools12 with the settings samtools mpileup -b -q -d 10000000000000, then input into VarScan v.2.3.9 (ref. 25) with the settings pileup2indel --min-coverage 2 --min-reads2 2 --min-var-freq 0.00001. Variants occurring within ±7 nt of the predicted crRNA cut sites were trimmed to obtain total mutation frequencies.

Evaluation of in vivo library diversity in the absence of mutagenesis. We synthesized a library of degenerate 8-mers and cloned these into the crRNA expression vector. After lentiviral production, KPD cells were transduced with the 8-mer lentiviral library and selected by puromycin; 4×10^6 KPD-8-mer cells were subcutaneously injected in either Rag1−/− or nsu/nmu mice. Twelve days after transplantation, mice were killed and tumors were isolated for genomic preparation and readout.
E-Gel EX (Life Technologies) using E-Gel Low Range Quantitative DNA Ladder (Thermo Fisher), then the same amounts of each barcoded sample were combined. The pooled PCR products were purified using the QIAquick PCR Purification Kit and further DNeasy Gel Extraction Kit from 2% E-gel EX. The purified pooled library was quantified as above. Diluted libraries with 5–20% PhiX were sequenced with the HiSeq 4000 system (illumina), with 150-bp pair-end read length.

MCAP-MET plasmid library readout and analysis. Raw pair-end fastq read files were first merged to single fastq file by PEAR with the settings -y 8 -g j -v 3. The merged fastq files were then filtered and de-multiplexed using Cutadapt, using two different sets of adaptors for extraction of crRNA array sequences or the 10-mer barcode. For the crRNA array, we used the following settings: cutadapt –discard-untrimmed -g tcgtGTGAAGAGGAGGACCGc, followed by cutadapt –discard-untrimmed -a FGTAGATTTT. The trimmed sequences were then mapped to the MCAP-MET library using Bowtie2.

bowtie2-v 2.3.2. The bowtie settings were: bowtie --end-to-end -v 3 -k 1 -m 1. For the 10-mer barcodes, we used the following cutadapt settings: cutadapt –discard-untrimmed -a aagcgtggcGGATC, followed by cutadapt –disc-untrimmed -g TACTAAGTTGAGATTTTT. The resultantly trimmed sequences were quantified to a reference of all possible 10-mer reads. Sequences that successfully mapped to the MCAP-MET library and contained a valid barcode were tabulated.

Identification of synergistic mutation combinations. We defined the SynCo score for each gene pair with the following formula:

$$\frac{\text{SynCo} = \text{DKO}_{\text{O}_{\text{m}}}-\text{SKO}_{\text{m}}-\text{SKO}_{\text{O}}}{\text{SKO}_{\text{O}}-\text{SKO}_{\text{m}}-\text{SKO}_{\text{O}}},$$

where the DKO value is the median log2 RPM and SKO and SKO are values defined as the median log2 RPM abundance of all corresponding SKO crRNA arrays. We calculated the SynCo of each gene pair within the lung metastasis samples and further assessed whether the DKO abundances were statistically significantly higher than the corresponding SKO abundances, by two-sided Welch’s t-test.

Design of dual-crRNA arrays for validation experiments. Dual-crRNA arrays containing combinations of Rosa26-targeting crRNAs or the best-performing N2 and Trim72 crRNA were designed. The following spacer sequences were used:

- crRosa26:5’- AGGCTATATTCTTCTGCT TCT-3’
- crRosa26:5’- TAGTCCAAGCTCTTGACAG-3’
- crN2: 5’-AAGGCCCTGATCTCCGTTT-3’
- Trim72:5’- TGGCCTGCTGCCTGATCCG-3’

Insertion of NLS–green fluorescent protein (GFP) sequences into the crRNA expression vector. The primary screen experiments were performed using the U6 crRNA expression vector with an EF1α short (EFS) promoter driving expression of puromycin and firefly luciferase. For validation experiments, the coding sequences for NLS-GFP were inserted after puromycin-P2A-luciferase in the crRNA expression vector by Gibson cloning, with a P2A sequence separating the GFP.

Quantification of mutation frequency by T7E1. Seven days after lentiviral transduction and puromycin selection, genomic DNA was extracted from the cells. PCR amplification of the genomic regions flanking the N2 or Trim72 crRNAs was performed using the following primers:

- N2_F: 5’- CCTGAGAAAACACTGTGACCACT-3’
- N2_R: 5’-AAAGCTGTCTGTGGCAGGGTTATTTG-3’
- Trim72_F: 5’-GGAGGAGGGCGGCGAGTGATTTG-3’
- Trim72_R: 5’-GCTGCGCAAGAACGATGTTAATG-3’

Using Phusion Flash High Fidelity Master Mix (Thermo Fisher), the thermocycling parameters for PCR were 98°C for 2 min, 35 cycles of (98°C for 1 s, 60°C for 5 s, 72°C for 2 min) and 72°C for 2 min. The PCR amplicons were then used for T7E1 assays according to the manufacturer’s protocol.

EDU proliferation assay. To assess proliferation, we used the Click-iT Edu Alexa Fluor 647 Flow Cytometry Assay Kit (Thermo Fisher, No. C10419). Cells underwent incubation in culture with 10μM EDU solution for 2 h, followed by fixation, permeabilization and staining. Cells were then analyzed on a BD FACSAria and the data were processed using FlowJo. Statistical significance was assessed by two-sided unpaired Welch’s t-test.

Matrigel invasion assay. For in vitro assessment of invasive potential, unsupplemented DMEM was first mixed with standard Matrigel (Cornin, No. 356234) on ice using pre-chilled pipette tips to a final concentration of 25% Matrigel. After placing 1.5μL Matrigel cell culture inserts with 8-μm pores into a 24-well plate, we added 100μL of the 25% Matrigel to each insert. The inserts were incubated in the cell culture incubator for 1 h to solidify the Matrigel. Cultured cells were then resuspended in unsupplemented DMEM at a concentration of 0.5 × 10⁶ cells ml⁻¹ and 200μL of the cell suspension was gently added on top of the Matrigel layer. Finally, 600μL of 10% FBS DMEM was added to each well, underneath the inserts. Invasive cells were quantified using an inverted microscope 24 h later on the GFP channel. Statistical significance was assessed by two-sided unpaired Welch’s t-test.

Luciferase imaging for tracking metastasis. Mice were anesthetized by isoflurane inhalation and imaged for metastasis using an IVIS machine (PerkinElmer) for 5 min following intraperitoneal injection of fireflyluciferin potassium salt (150mg kg⁻¹ body weight).

Quantification of primary tumors and lung metastases. Mice were anesthetized by isoflurane inhalation, and tumor sizes were quantified
Brief Communication

Nature Methods

every 2–3 d by calipers, using the formula: volume (mm$^3$) = π (6 × x × y × z). Statistical significance was assessed by two-way ANOVA, jointly considering the effect of time and treatment condition. Mice were killed at 28 d post-infection, and lungs were harvested for quantification of lung metastases. Each lung lobe was separately visualized under a dissecting microscope. Lung lobe metastases were quantified on bright-field images with real-time confirmation by GFP expression. Statistical significance was assessed by two-sided unpaired Welch’s t-test.

Genomic comparisons of human primary tumors and metastases. Mutation frequencies from the TCGA PanCancer dataset and the MET500 dataset were filtered for the 26 genes represented in the MCAP-MET library. We determined the statistical significance of Spearman correlation by calculating the t-statistic of the correlation. Identification of gene pairs that were significantly co-mutated was determined by hypergeometric test.

Statistics. All statistical tests are unpaired and two-sided. Details about the statistical tests are described in the corresponding figure legends and Methods subsections.

Blinding statement. Investigators were not blinded for sequencing data analysis, tumor engraftment or organ dissection.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability
MCAP data, sequences of oligos and library design are described in the Methods section and Supplementary Tables. All vectors and libraries have been deposited to Addgene and are available to the academic community. Cell lines and all data supporting this work will be made available to the academic community upon reasonable request to the corresponding author. Genomic sequencing data have been deposited with NCBI SRA (PRJNA515306).

Code availability
Key scripts used to process and analyze the data will be available to the academic community upon reasonable request.

References
21. Chow, R. D., Wang, G., Ye, L. & Chen, S. Protocol Exchange https://doi.org/10.1038/protex.2019.018 (2019).
22. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Genome Biol. 10, R25 (2009).
23. Li, H. & Durbin, R. Bioinformatics 25, 1754–1760 (2009).
24. Li, H. et al. Bioinformatics 25, 2078–2079 (2009).
25. Koboldt, D. C. et al. Genome Res. 22, 568–576 (2012).
26. Zhang, J., Kobert, K., Flouri, T. & Stamatakis, A. Bioinformatics 30, 614–620 (2014).
27. Martin, M. EMBnet J. 17, 10–12 (2011).
Corresponding author(s): Sidi Chen

Statistical parameters
When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

| n/a | Confirmed |
|-----|-----------|
| □   | □         |

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
- Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted
  Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's d, Pearson's r), indicating how they were calculated
- Clearly defined error bars
  State explicitly what error bars represent (e.g. SD, SE, CI)

Software and code
Policy information about availability of computer code

Data collection
No software or code was used for data collection.

Data analysis
Nextera genomic sequencing reads were mapped to the mm10 mouse genome using BWA, with the settings bwa mem -t 8 -w 200. Indel variants were first processed with Samtools with the settings samtools mpileup -B -q 15 -d 10000000000000, then input into VarScan v2.3.9 with the settings pileup2indel --min-coverage 2 --min-reads2 2 --min-var-freq 0.00001. Variants occurring within a ± 7nt window of the predicted crRNA cut sites were summed to obtain total mutation frequencies.

For MCAP plasmid library readout analysis, raw paired-end fastq read files were first merged to single fastq files by PEAR with the settings -y BG -j 8 -v 3. The merged fastq files were then filtered and demultiplexed using Cutadapt, using two different sets of adapters for extraction of crRNA array sequences or the 10mer barcode. For the crRNA array, we used the following settings: cutadapt --discard-untrimmed -g ttctGTGGAAAGGACGAAACCCg, followed by cutadapt --discard-untrimmed -a TGTAGATTTTTTT. The trimmed sequences were then mapped to the MCAP-MET library using Bowtie v1.1.2: bowtie -v 3 -k 1 -m 1. For the 10mer barcodes, we used the following Cutadapt settings: cutadapt --discard-untrimmed -a aagcttgctGGATCCGATATCa, followed by cutadapt --discard-untrimmed -g TACTAAGTGTAGATTTTTTT. The resultant sequences were quantified to a reference of all possible 10mer sequences. Reads that successfully mapped to both the MCAP-MET library and contained a valid barcode were tabulated.

For analyzing MCAP library readout from cells and tumors, PEAR-merged fastq files were filtered and demultiplexed using Cutadapt. To remove extra sequences downstream (i.e. 3' end) of the crRNA array sequences, including the DR and U6 terminator, we used the following settings: cutadapt --discard-untrimmed --e 0.1 -a aagcttgctGGATCCGATATCa -m 80. As the forward PCR primers used to
readout crRNA array representation were designed to have a variety of barcodes to facilitate multiplexed sequencing, we then
demultiplexed these filtered reads with the following settings: cutadapt -g file:fbc.fasta --no-trim, where fbc.fasta contained the 12
possible barcode sequences within the forward primers. Finally, to remove extraneous sequences upstream (i.e. 5’ end) of the crRNA
array spacers, we first used the following settings: cutadapt --discard-untrimmed --e 0.1 -g tcttGTGGAAAGGACGAAACACCg -m 80. Then,
we removed the 5’ DR as follows: cutadapt --discard-untrimmed -e 0.1 -g TAATTTCTACTAAGTGTAGAT -m 80. The filtered fastq reads
were then mapped to the MCAP-MET reference index. To do so, we first generated a Bowtie index of the MCAP-MET library using the
bowtie-build command in Bowtie 1.1.2. Using these bowtie indexes, we mapped the filtered fastq read files using the following settings:
bowtie -n 2 -k 1 -m 1 --best. These settings ensured only single-match reads would be retained for downstream analysis. For data
processing on the level of barcoded-crRNAs, we utilized the same trimmed fastq files as above, but instead used the barcoded-crRNA
plasmid library as the reference index.

Using the resultant mapping output, we quantified the number of reads that had mapped to each crRNA array within the library. We
normalized the number of reads in each sample by converting raw crRNA array counts to reads per million (rpm). The rpm values were
then subject to log-2 transformation for certain analyses. Where applicable, linear regression lines and 95% confidence intervals were
calculated. For comparing cells, primary tumors, and lung metastases, crRNA array abundances were averaged within sample groups and
linear regression was performed using the NTC-NTC arrays as a model for neutral selection. Significant outliers were identified using the
outlierTest function from the car R package, which calculates the studentized residuals of the linear regression and derives the
respective p-values. For gene/gene
code pair analyses, the corresponding SKO and DKO arrays were first averaged together, then aggregated by sample type. Linear regression
was performed using all SKO/DKO genotypes, and outliers were identified as above.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers
upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:
  - Accession codes, unique identifiers, or web links for publicly available datasets
  - A list of figures that have associated raw data
  - A description of any restrictions on data availability

MCAP data, sequences of oligos, and library design are described in the Methods section and Supplementary Tables. All vectors and libraries have been deposited to
Addgene (#123359, #123360, #123361, #123362, and #123363) and are available to the academic community. Cell lines and all data supporting this work will be
available to the academic community upon request. Genomic sequencing data has been deposited to NCBI SRA (PRJNA515306).

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences
☐ Behavioural & social sciences
☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

| Sample size | No specific methods were used to predetermine sample sizes. Sample sizes were chosen based on experience from our prior published work (Chen et al., Cell 2015). |
|-------------|-------------------------------------------------------------------------------|
| Data exclusions | No data were excluded. |
| Replication | All experiments were performed using multiple entirely biologically independent replicates, and replication efforts were successful. |
| Randomization | Animals were randomly allocated to different experimental groups. |
| Blinding | Investigators were not blinded for experiments or analysis. All experimental conditions were carefully controlled and done in parallel, consistent with our established protocols. Multiple independent investigators were involved at all stages of each experiment to ensure accountability and proper experimental practices. |

Reporting for specific materials, systems and methods
Involved in the study
- Unique biological materials
- Antibodies
- Eukaryotic cell lines
- Palaeontology
- Animals and other organisms
- Human research participants

Eukaryotic cell lines
- KPD cells (Chen et al., Genes Dev 2014); HEK293FT cells (ThermoFisher)
- KPD cells were generated by lab personnel (Chen et al., Genes Dev 2014). HEK293FT cells were purchased directly from a reputable manufacturer with rigorous authentication procedures (ThermoFisher).
- KPD and HEK293FT cells tested negative for mycoplasma.
- No commonly misidentified cell lines were used in this study, as defined by the ICLAC register.

Animals and other organisms
- 6-8 week old, both male and female, nu/nu mice or Rag1-/- mice were used for experiments. Only 6-8 week old female nu/nu mice were used in validation experiments.
- The study did not involve the use of wild animals.
- The study did not involve the use of field-collected samples.

Flow Cytometry
- The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).
- The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a ‘group’ is an analysis of identical markers).
- All plots are contour plots with outliers or pseudocolor plots.
- A numerical value for number of cells or percentage (with statistics) is provided.

Methodology
- To assess proliferation, we used the Click-iT EdU Alexa Fluor 647 Flow Cytometry Assay Kit (ThermoFisher, #C10419). We incubated cells in culture with 10 μM for 2 hours, followed by fixation, permeabilization, and staining. Cells were then analyzed on a BD FACS Aria and the data was processed using FlowJo. Statistical significance was assessed by two-sided unpaired Welch’s t-test.
- Gating was done on live single cells (>75% for all samples).
- See Supplementary Figure 12b for gating strategy.
- Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.