A quantitative and multiplexed approach to uncover the fitness landscape of tumor suppression *in vivo*

Zoë N Rogers[1,6], Christopher D McFarland[2,6], Ian P Winters[1,6], Santiago Naranjo[1], Chen-Hua Chuang[1], Dmitri Petrov[2] & Monte M Winslow[1,3–5]

Cancer growth is a multistage, stochastic evolutionary process. While cancer genome sequencing has been instrumental in identifying the genomic alterations that occur in human tumors, the consequences of these alterations on tumor growth remain largely unexplored. Conventional genetically engineered mouse models enable the study of tumor growth *in vivo*, but they are neither readily scalable nor sufficiently quantitative to unravel the magnitude and mode of action of many tumor-suppressor genes. Here, we present a method that integrates tumor barcoding with ultradeep barcode sequencing (Tuba-seq) to interrogate tumor-suppressor function in mouse models of human cancer. Tuba-seq uncovers genotype-dependent distributions of tumor sizes. By combining Tuba-seq with multiplexed CRISPR–Cas9-mediated genome editing, we quantified the effects of 11 tumor-suppressor pathways that are frequently altered in human lung adenocarcinoma. Tuba-seq enables the broad quantification of the function of tumor-suppressor genes with unprecedented resolution, parallelization, and precision.

Genome sequencing has catalogued somatic genomic alterations in human cancers and identified many putative tumor-suppressor genes[1–3]. However, the identification of recurrent genomic alterations does not necessarily reveal their functional importance to cancer growth; the impact of specific alterations remains difficult to glean from cancer genome sequencing data alone[4,5].

The impacts of tumor-suppressor gene losses on neoplastic growth have been investigated using knockdown, knockout, and overexpression studies in cell lines as well as in mouse models. However, the near-optimal growth of cancer cell lines in culture, their widespread genetic and epigenetic changes, and the lack of an autochthonous microenvironment limit the ability of studies on cell lines to provide insight into how tumor-suppressor genes constrain the expansion of tumors *in vivo*. In contrast, genetically engineered mouse models enable the introduction of defined genetic alterations into normal adult cells, which results in the initiation and growth of tumors within their natural *in vivo* setting[6]. While these models have become a mainstay for the analysis of tumor-suppressor gene function, they are neither readily scalable nor sufficiently quantitative.

Recently, CRISPR–Cas9-mediated genome editing in somatic cells has increased the throughput of *in vivo* analyses of gene function in autochthonous cancer models[7–10]. While these systems increase the scale of *in vivo* functional analyses, they continue to rely on relatively crude measurements of tumor growth, which limits their application to the analysis of tumor suppressors with the most dramatic effects.

Molecular barcoding enables precise, multiplexed quantification of evolutionary fitness, selection, and clonal growth[11–17]. We now combine tumor barcoding and high-throughput sequencing (Tuba-seq) with genetically engineered mouse models to quantify tumor growth with unprecedented resolution. Precise quantification of individual tumor sizes uncovered the impact of inactivating different tumor-suppressor genes. Integration of these methods with multiplexed CRISPR–Cas9-mediated genome editing enabled the parallel inactivation and functional quantification of a panel of putative tumor-suppressor genes.

RESULTS

Tuba-seq enables precise and parallel quantification of tumor sizes

Oncogenic KRAS is a key driver of human lung adenocarcinoma, and early stage lung tumors can be modeled using *LoxP-Stop-LoxP Kras*^{G12D} knock-in mice (*Kras*$^{LSL\text{-}G12D/+}$), in which expression of Cre in lung epithelial cells leads to the expression of oncogenic Kras^{G12D} (refs. 18 and 19). *LKB1* and *P53* are frequently mutated tumor-suppressor genes in human lung adenocarcinomas (**Supplementary Fig. 1a**)[20], and *Lkb1* or *p53* deficiency each increase tumor burden in mouse models of oncogenic Kras^{G12D}-driven lung tumors[21,22]. In viral-Cre-induced mouse models of lung cancer, large numbers of tumors can be initiated simultaneously, and individual tumors can be stably tagged by lentiviral-mediated DNA barcoding[23,24]. Therefore, we set out to determine whether high-throughput sequencing of the lentiviral barcode region from bulk-tumor-bearing lungs could quantify the number of neoplastic cells within each uniquely barcoded tumor (**Supplementary Fig. 1b**).

[1]Department of Genetics, Stanford University School of Medicine, Stanford, California, USA. [2]Department of Biology, Stanford University, Stanford, California, USA. [3]Department of Pathology, Stanford University School of Medicine, Stanford, California, USA. [4]Cancer Biology Program, Stanford University School of Medicine, Stanford, California, USA. [5]Stanford Cancer Institute, Stanford University School of Medicine, Stanford, California, USA. [6]These authors contributed equally to this work. Correspondence should be addressed to M.M.W. (mwinslow@stanford.edu).

RECEIVED 4 DECEMBER 2016; ACCEPTED 19 APRIL 2017; PUBLISHED ONLINE 22 MAY 2017 ; DOI:10.1038/NMETH.4297

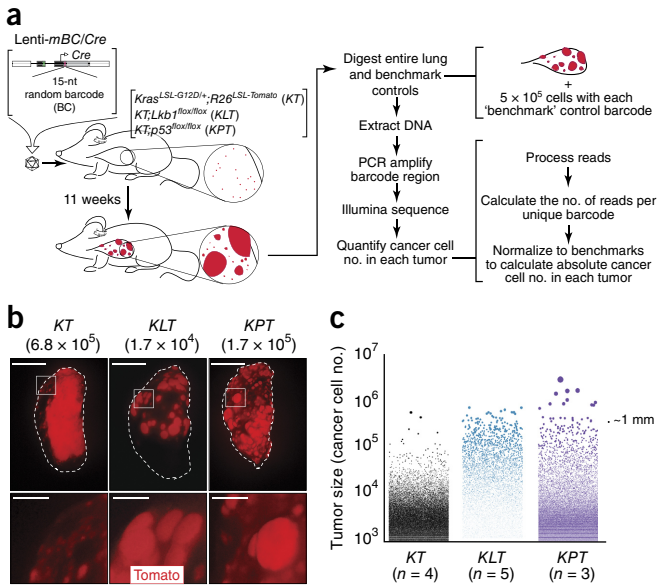

Figure 1 | Tuba-seq combines tumor barcoding with high-throughput sequencing to allow parallel quantification of tumor sizes. (**a**) Schematic of Tuba-seq pipeline to assess lung tumor size distributions. Tumors were initiated in *Kras*$^{LSL\text{-}G12D/+}$;*Rosa26*$^{LSL\text{-}Tomato}$ (*KT*), *KT;Lkb1*$^{flox/flox}$ (*KLT*), and *KT;p53*$^{flox/flox}$ (*KPT*) mice with Lenti-*mBC/Cre* (a pool of lentiviral vectors containing ~10^6 random 15-nt DNA barcodes (BC)). Tumor sizes were calculated via bulk barcode sequencing of tumor-bearing lungs. (**b**) Fluorescence dissecting scope images of lung lobes from *KT*, *KLT*, and *KPT* mice with Lenti-*mBC/Cre*– initiated tumors. Lung lobes are outlined with white dashed lines. The titer of Lenti-*mBC/Cre* is indicated. Scale bars in upper panels, 5 mm. Scale bars in lower panels, 1 mm. (**c**) Tumor size distributions in *KT*, *KLT*, and *KPT* mice (number of mice per group is indicated). Each dot represents a tumor. The area of each dot is proportional to the number of cancer cells in each tumor. A dot corresponding to the approximate number of cancer cells in a 1 mm-diameter spherical tumor is shown to the right of the data for reference.

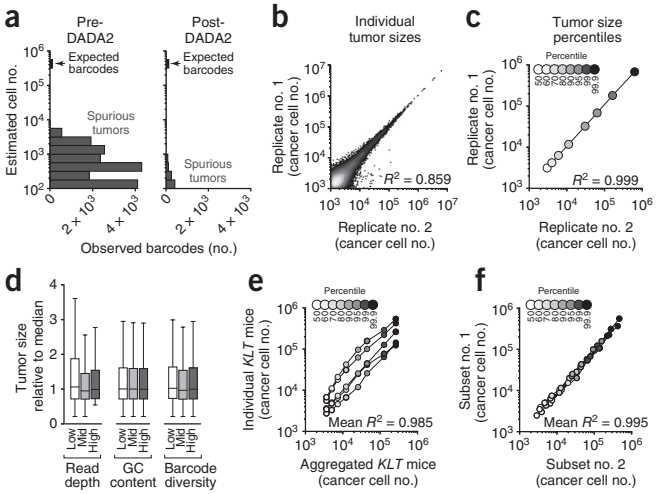

Figure 2 | Tuba-seq precisely and reproducibly quantifies tumor sizes. (**a**) DADA2 eliminates recurrent read errors that can appear as spurious tumors. Cell lines with known barcodes were added to each lung sample (5 × 10^5 cells each). Recurrent read errors derived from these known barcodes generate spurious tumors, which are greatly reduced by DADA2. (**b**) Individual tumor sizes and (**c**) size profiles of tumors at the indicated percentiles of technical replicate sequencing libraries prepared from an individual bulk tumor-bearing lung sample. (**d**) Analysis of the effect of variation in read depth, GC content of the DNA barcodes, and diversity of the barcode library on tumor size calling. Tumors were partitioned into thirds corresponding to high, moderate, and low levels of each technical parameter. Whiskers capped at 1.5 IQR. Boxes depict interquartile range (IQR) with center line at median. (**e**) Size distributions across five *KLT* mice. Sizes of the tumors at the indicated percentiles in individual mice are connected by a line. (**f**) Tumors in each *KLT* mouse were partitioned into two groups (see Online Methods), and the profiles of these groups were compared. Sizes of the tumors at the indicated percentiles in an individual mouse are connected by a line.

To interrogate the growth of oncogenic Kras^{G12D}-driven lung tumors as well as the impact of *Lkb1* and *p53* deficiency on tumor growth, we initiated lung tumors in *Kras*$^{LSL\text{-}G12D/+}$;*Rosa26*$^{LSL\text{-}Tomato}$ (*KT*), *KT;Lkb1*$^{flox/flox}$ (*KLT*), and *KT;p53*$^{flox/flox}$ (*KPT*) mice with a library of lentiviral-Cre vectors containing ~10^6 unique barcodes (Lenti-*mBC/Cre*; **Fig. 1a** and **Supplementary Fig. 1b**). *KT* mice developed widespread hyperplasias and small-tumor masses (**Fig. 1b** and **Supplementary Fig. 1c**). Interestingly, while *KLT* mice had large tumors of relatively uniform size, *KPT* mice had a very diverse range of tumor sizes (**Fig. 1b**).

To quantify the neoplastic cell number in every lesion using high-throughput sequencing, we PCR amplified the integrated lentiviral barcode region from bulk tumor-bearing lung DNA from each mouse and sequenced this to an average depth of >10^7 reads per mouse (**Fig. 1a** and **Supplementary Note**). Our analysis indicated that tumor sizes varied by more than 1,000-fold (**Fig. 1c**). Barcode reads from small lesions could represent unique tumors or be generated from recurrent sequencing errors of similar barcodes from larger tumors. To minimize the occurrence of these spurious tumors, we aggregated reads expected to be derived from the same tumor barcode using an algorithm that generates a statistical model of sequencing errors (DADA2; **Fig. 2** and **Supplementary Fig. 2**)[25]. To enable the conversion of read count to cancer cell number, we added cells with known barcodes to each lung sample at a defined number before tissue homogenization and DNA extraction, and we normalized tumor read counts to 'benchmark' read counts from these cells (**Fig. 1a** and **Supplementary Fig. 3**).

The Tuba-seq pipeline was highly reproducible between technical replicates and was insensitive to typical variation in many technical variables (**Fig. 2b–d**, **Supplementary Fig. 4**, and **Supplementary Note**). Tumor size distributions were also highly reproducible between mice of the same genotype (R^2 > 0.98; **Fig. 2e**, **Supplementary Fig. 4g**, and **Supplementary Note**). Indeed, unsupervised hierarchical clustering of size distributions separated mice according to their genotype, even when tumors were induced with different Lenti-*mBC/Cre* titers (**Supplementary Fig. 4d**). Differences in the spectrum of tumor sizes between mice of the same genotype were far greater than the differences between two fractions of tumors within the same mouse, indicating that the measurement error of Tuba-seq is less than the intrinsic variability between mice (**Fig. 2e**,**f**). Thus, Tuba-seq rapidly and precisely quantifies the number of neoplastic cells within thousands of individual lung lesions in *KT*, *KLT*, and *KPT* mice (**Fig. 1c**, **Supplementary Fig. 4c**, and **Supplementary Note**).

Analysis of tumor sizes uncovers two modes of tumor suppression

To assess the effect of *p53* or *Lkb1* deficiency on tumor growth, we calculated the number of neoplastic cells in the tumors at

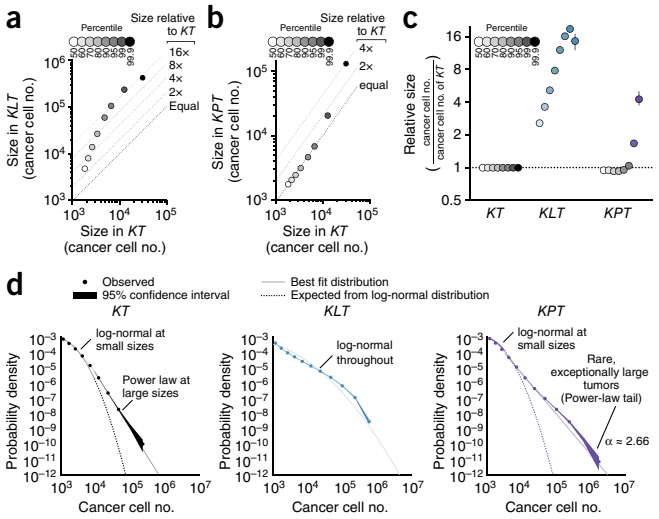

Figure 3 | Massively parallel quantification of tumor sizes enables probability distribution fitting across multiple genotypes. (**a**,**b**) Tumor size at the indicated percentile in *KLT* ($n = 5$) mice (**a**) and *KPT* ($n = 3$) mice (**b**) versus tumor size at the indicated percentile in *KT* mice ($n = 7$). Each percentile was calculated using all tumors from all mice of each genotype 11 weeks after tumor initiation with Lenti-*mBC/Cre*. (**c**) Tumor sizes at the indicated percentiles for each genotype relative to *KT* tumors at the same percentiles. Error bars are 95% confidence intervals obtained via bootstrapping. Percentiles that are significantly different from the corresponding *KT* percentiles are in color. (**d**) Tumor size distributions were most closely fit by log-normal distributions. Tumors in *KLT* mice are best described by a log-normal distribution throughout their entire size spectrum (middle). The tumor size distributions in *KT* mice (left) and *KPT* mice (right) were better explained by combining a log-normal distribution at smaller scales with a power-law distribution at larger scales. Power-law relationships decline linearly on log–log axes, consistent with rare yet very large tumors within the top ~1% of tumors in *KT* mice and ~10% of tumors in *KPT* mice. Note that only tumors in *KPT* mice exceed 10^6 cancer cells after 11 weeks, consistent with *p53* deficiency enabling the generation of the largest tumors in this study.

different percentiles within the distribution. While tumors in *KLT* mice were consistently larger than *KT* tumors, deletion of *p53* allowed only a small fraction of tumors to grow to exceptional sizes (**Figs. 1c** and **3**).

To further investigate the effect of *p53* and *Lkb1* deficiency on tumor growth, we also defined the mathematical distributions that best fit the tumor size distributions in *KT*, *KLT*, and *KPT* mice. *Lkb1*-deficient tumors were log-normally distributed across the full range of the distribution, consistent with exponential tumor growth with normally distributed rates (**Fig. 3d**)[26]. To estimate average tumor size without allowing very large tumors to greatly shift this metric, we calculated the maximum likelihood estimator of the mean number of cancer cells given a log-normal distribution of tumor sizes (LN mean). By this measure *KLT* tumors had, on average, seven-fold more cancer cells than *KT* tumors (**Fig. 3a**,**c**). Despite greater tumor burden and visibly larger tumors in *KPT* mice, *p53* deficiency did not increase the LN mean. Instead, *p53*-deficient tumors were power-law distributed at large sizes, and the elevated tumor burden was driven by rare, exceptionally large tumors (**Fig. 3d** and **Supplementary Note**)[27]. A power-law distribution is consistent with *p53* deficiency allowing tumors to acquire additional rare, yet profoundly tumorigenic, alterations[28–30].

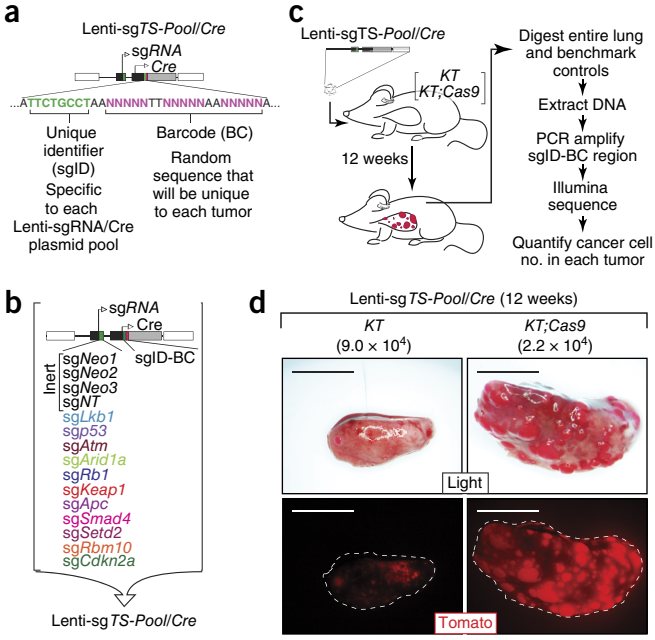

Figure 4 | Rapid quantification of tumor-suppressor phenotypes using Tuba-seq and multiplexed CRISPR–Cas9-mediated gene inactivation. (**a**) Schematic of the Lenti-sg*TS-Pool/Cre* vector. (**b**) Lenti-sg*TS-Pool/Cre* contains 4 vectors with inert sgRNAs and 11 vectors targeting known and candidate tumor-suppressor genes. NT, nontargeting. (**c**) Schematic of multiplexed CRISPR–Cas9-mediated tumor-suppressor inactivation coupled with Tuba-seq to assess the function of each targeted gene on lung tumor growth *in vivo*. (**d**) Bright-field (top) and fluorescence dissecting scope images (bottom) of lung lobes from *KT* and *KT;Cas9* mice 12 weeks after tumor initiation with Lenti-sg*TS-Pool/Cre*. Lung lobes are outlined with dashed white lines in the fluorescence images. Viral titer is indicated. Scale bars, 5 mm.

Multiplexed CRISPR–Cas9-mediated inactivation of tumor-suppressor genes

To simultaneously quantify the tumor-suppressive function of many known and candidate tumor suppressor genes in parallel, we combined Tuba-seq and conventional Cre-based mouse models with multiplexed CRISPR–Cas9-mediated *in vivo* genome editing (**Fig. 4a–c**). Assessing different tumor genotypes within individual mice minimized the effect of mouse-to-mouse variability and maximized the resolution of Tuba-seq (**Supplementary Note**).

Initiation of tumors with lentiviral sg*RNA/Cre* vectors targeting either the tdTomato reporter or *Lkb1* in mice with an $H11^{LSL\text{-}Cas9}$ allele[8] confirmed efficient Cas9-mediated gene inactivation (**Supplementary Fig. 5**). Next, we selected 11 known and putative lung adenocarcinoma tumor-suppressor genes representing diverse pathways and identified the most efficient sgRNA targeting each gene (**Fig. 4b**, **Supplementary Fig. 1a**, and **Supplementary Fig. 6**)[20,31]. To quantify the number of neoplastic cells in each tumor using Tuba-seq, we diversified each Lenti-sg*RNA/Cre* vector with a two-component barcode consisting of a unique 8-nt 'sgID' specific to each sgRNA and a random 15-nt barcode (BC) to uniquely tag each tumor (sgID-BC; **Fig. 4a**,**b** and **Supplementary Fig. 7**).

Parallel quantification of tumor-suppressor function *in vivo*

To quantify the effect of inactivating each gene on lung tumor growth in parallel, we initiated tumors in *KT* and $KT;H11^{LSL\text{-}Cas9}$ (*KT;Cas9*) mice with a pool of the 11 barcoded Lenti-sg*RNA/Cre*

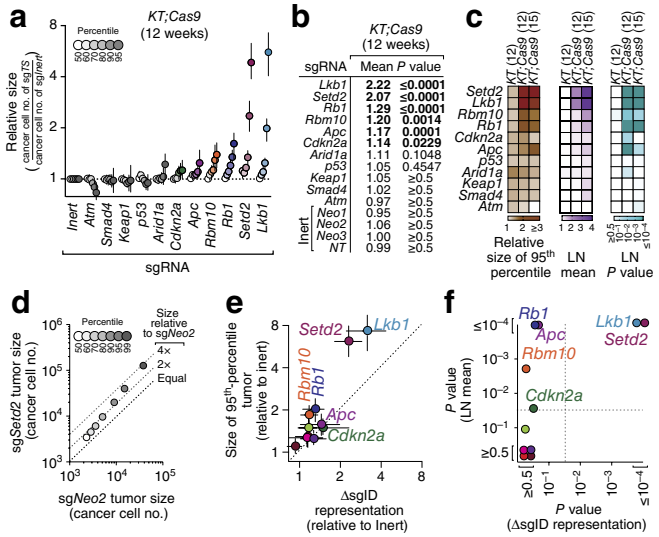

Figure 5 | Tuba-seq uncovers known and novel tumor suppressors with unprecedented resolution. (**a**) Analysis of the relative tumor sizes in *KT;Cas9* mice 12 weeks after tumor initiation with Lenti-sg*TS-Pool/Cre*. Relative size of tumors at the indicated percentiles represents merged data from eight mice, normalized to the average size of sg*Inert* tumors. 95% confidence intervals are shown. Percentiles that are significantly greater than sg*Inert* are in color. Colors correspond to the sgRNA color in **Figure 4b**, and the darker the shade of color the larger the percentile, as shown in the legend in gray scale. (**b**) Estimates of mean tumor size (relative to inert) assuming a log-normal tumor size distribution. Bonferroni-corrected, bootstrapped *P* values are shown. *P* values < 0.05 and their corresponding means are shown in bold. (**c**) Relative size of the 95th-percentile tumors (left), log-normal (LN) mean (middle), and log-normal (LN) *P* value (right) for tumors with each sgRNA in *KT* and *KT;Cas9* mice 12 weeks after tumor initiation, and *KT;Cas9* mice 15 weeks after tumor initiation. Numbers in parentheses denote the number of weeks after tumor initiation that the mice were analyzed. (**d**) Tumor size at the indicated percentile from *KT;Cas9* mice with Lenti-sg*Setd2#1/Cre*-initiated tumors versus Lenti-sg*Neo2/Cre*-initiated tumors ($N = 4$ mice per group). Percentiles were calculated using all tumors from all mice in each group. (**e**,**f**) The relative size of the 95th-percentile tumor and the log-normal statistical significance determined by Tuba-seq were plotted against the average fold change in sgID representation and their associated *P* values, respectively. Error bars in **e** are 95% confidence intervals. Dotted lines in **f** indicate the 0.05 significance threshold. Dot color corresponds to the sgRNA color in **Figure 4b**.

vectors and 4 barcoded Lenti-sg*Inert/Cre* vectors (Lenti-sg*TS-Pool/Cre*; **Fig. 4b**,**c**). 12 weeks after tumor initiation, the number and size of macroscopic tumors was greater in *KT;Cas9* mice, even though they received a lower dose of virus than the *KT* mice (**Fig. 4d** and **Supplementary Fig. 8**). To determine the number of neoplastic cells in each tumor with each sgRNA, we amplified the sgID-BC region from bulk tumor-bearing lung DNA, deep sequenced the product, and applied our Tuba-seq analysis pipeline. For each sgRNA, the number of neoplastic cells in tumors at different percentiles was normalized to tumors from the sg*Inert* distribution (**Fig. 5a**). We also determined the relative LN mean size of tumors containing each of the 11 tumor-suppressor-targeting sgRNAs (**Fig. 5b**).

We analyzed an additional cohort of *KT;Cas9* mice 15 weeks after tumor initiation with Lenti-sg*TS-Pool/Cre*. We confirmed the tumor-suppressive effect of all tumor suppressors identified at 12 weeks post-tumor initiation (**Fig. 5c** and **Supplementary Fig. 8d**,**e**). Importantly, both the LN mean and the relative number of cancer cells in the 95th-percentile tumor were reproducible (**Fig. 5c** and **Supplementary Fig. 8**).

These analyses confirmed the known tumor-suppressive function of Lkb1, Rb1, Cdkn2a, and Apc in Kras^{G12D}-driven lung tumor growth (**Fig. 5a**,**b** and **Supplementary Figs. 6b** and **8**)[7,22,32,33]. Tuba-seq also identified the splicing factor Rbm10 and the methyltransferase Setd2 as suppressors of lung tumor growth. Splicing factors have emerged as potential tumor suppressors in many cancer types, and components of the spliceosome are mutated in 10–15% of human lung adenocarcinomas[2,20,31,34]. *Rbm10* inactivation significantly increased the number of cancer cells in the top 50% of lung tumors and increased the LN mean size (**Fig. 5a**,**b**). Setd2 is the sole histone H3K36me3 methyltransferase and may also affect genomic stability by methylating microtubules[35–37]. *SETD2* is frequently mutated in several major cancer types, including lung adenocarcinoma[2,20,31,33,38]. *Setd2* inactivation dramatically increased tumor size, and these tumors exhibited a log-normal size distribution (**Supplementary Fig. 9**). These data suggest that aberrant pre-mRNA splicing and the absence of Setd2-mediated lysine methylation both have profound protumorigenic effects in lung adenocarcinoma.

To further validate the tumor-suppressive effect of Setd2, we induced tumors in *KT* and *KT;Cas9* mice with lentiviral vectors containing an inert sgRNA (sg*Neo2*) or either of two sgRNAs targeting *Setd2*. *KT;Cas9* mice with tumors initiated with either Lenti-sg*Setd2/Cre* vector developed large adenomas and adenocarcinoma and exhibited greater overall tumor burden than did *KT* mice with tumors initiated with the same virus (**Supplementary Fig. 10**). Analysis of tumor sizes by Tuba-seq confirmed a nearly four-fold increase in the number of neoplastic cells in the largest *Setd2*-deficient tumors relative to control tumors (**Fig. 5d** and **Supplementary Fig. 10**). Importantly, the validation of *Setd2*-mediated tumor suppression by conventional methods required more mice than our initial screen of 11 putative tumor suppressors did; this emphasizes the benefit of multiplexing sgRNAs to increase throughput and decrease costs.

Recapitulation of tumor size distributions within the tumor-suppressor pool

Consistent with the distribution of tumor sizes in *KPT* mice, neither the LN mean nor the analysis of tumors up to the 95th percentile uncovered an effect of targeting *p53* on tumor growth in *KT;Cas9* mice with Lenti-sg*TS-Pool/Cre*-initiated tumors (**Fig. 5**). As anticipated, Lenti-sg*p53/Cre*-initiated tumors exhibited a power-law distribution at larger sizes, and sg*p53* was enriched within the largest tumors in these mice (**Supplementary Fig. 11a**,**b**). The effect of targeting *p53* was greater at the later 15-week time point, consistent with p53's known role in limiting tumor progression (**Supplementary Fig. 11**)[21,29].

In *KT;Cas9* mice with Lenti-sg*TSPool/Cre*-initiated tumors, Lenti-sg*Lkb1/Cre*-initiated tumors exhibited a log-normal distribution of tumor sizes consistent with our data from *KLT* mice (**Figs. 1c** and **2d**; **Supplementary Fig. 9a**). Both *p53*- and *Lkb1*-deficient tumors generated through somatic genome editing have similar size distributions to those of tumors initiated using floxed alleles. Thus, even in this pooled setting, quantification of individual tumor sizes can uncover characteristic distributions of tumor sizes upon tumor suppressor inactivation.

Tuba-seq provides the sensitivity to identify tumor suppressors of small effect

Two-thirds of the tumor suppressors we identified (*Apc*, *Rb1*, *Rbm10*, and *Cdkn2a*) were only identified when we considered the number of neoplastic cells in each barcoded tumor, while they were not identified when we only considered the fold change in sgID representation (**Fig. 5**). In fact, the precision of effect-size estimates, statistical significance, and the detection of tumor suppressors with small effect were all improved using the Tuba-seq pipeline (**Fig. 5e**,**f** and **Supplementary Note**).

As an orthogonal approach to investigate the selection for tumor-suppressor inactivation and to confirm on-target sgRNA-mediated genome editing, we PCR amplified and deep sequenced each sgRNA-targeted region from bulk tumor-bearing lung DNA region from *KT;Cas9* mice with Lenti-*TS-Pool/Cre*-initiated tumors. A relatively high fraction of *Setd2*, *Lkb1*, and *Rb1* alleles had inactivating indels at the targeted sites, which was consistent with on-target sgRNA activity and the expansion of tumors with inactivation of these genes (**Supplementary Figs. 11c–f** and **12**). This analysis also confirmed that all targeted genes contained indels (**Supplementary Fig. 12**). Although all of the genes included in our pool are recurrently mutated in human lung adenocarcinoma (**Supplementary Fig. 1a**)[20,31], *Arid1a*, *Smad4*, *Keap1*, and *Atm* were not identified as tumor suppressors (**Fig. 5**; **Supplementary Figs. 8d**,**e**,**h** and **12a**). That *Atm* deficiency does not increase tumor growth is consistent with results using an Atm^{floxed} allele[39]. We also confirmed the lack of tumor-suppressive function of Smad4 *in vivo* (**Supplementary Fig. 12d**,**e**). For these genes, changes in gene expression or environmental state, additional time, or coincident genomic alterations may be required for inactivation of these pathways to confer a growth advantage in lung cancer cells.

DISCUSSION

While many putative tumor suppressors have been identified from cancer genome sequencing, limited strategies exist to test their function *in vivo* in a rapid, systematic, and quantitative manner (**Supplementary Table 1**). Tuba-seq enables exceptionally precise and detailed quantification of tumor growth *in vivo*. Interestingly, tumors initiated at the same time within the same mouse with the same genomic alterations grew to vastly different sizes after only 12 weeks of growth (**Figs. 1** and **2**). Thus, additional spontaneous alterations, differences in the state of the initial transformed cell, and/or the local microenvironment may impact how rapidly a tumor grows and whether it has the capacity for continued expansion. The growth variability identified by Tuba-seq also revealed properties of gene function. *p53* deficiency generates a tumor size distribution that is power-law distributed for the largest tumors, consistent with a Markov process where very large tumors are generated by additional, rarely acquired driver mutations (**Supplementary Note**)[27]. Conversely, *Lkb1* inactivation increases the size of a majority of lesions, consistent with the role of Lkb1 in restraining proliferation[40]. Interestingly, Setd2 has recently been suggested to methylate tubulin; and *Setd2* deficiency can lead to genomic instability, which would be expected to generate power-law-distributed tumor growth[34]. However, the size distribution of *Setd2*-deficient lung tumors was strictly log normal, which suggests that the main impact of *Setd2* loss on the early stages of tumor growth is the induction of gene-expression programs that generally dysregulate growth (**Supplementary Fig. 9b**,**c**).

Unlike conventional floxed alleles, CRISPR–Cas9-mediated genome editing in the lung only generates homozygous null alleles in approximately half of all tumors (**Supplementary Fig. 5d**). Thus, while the lack of uniform homozygous deletion of targeted genes would reduce the tumor-suppressive signal from bulk measurements, Tuba-seq effectively overcomes this technological limitation by barcoding and analyzing each tumor (**Fig. 5**).

By analyzing a large number of tumor suppressors, our data suggest that early neoplastic cells reside in an evolutionarily nascent state where many tumor-suppressor alterations are adaptive and confer a large growth advantage. In contrast, tumor-suppressor alterations in cancer cell lines often provide little advantage and can even be detrimental[41]. This is consistent with cancer cell lines residing in a much more mature evolutionary state, approaching optimal growth fitness on account of their origin from advanced-stage disease as well as the selection for proliferative ability in culture. Furthermore, the intimate link between tumor suppression and many aspects of the *in vivo* environment underscores the importance of analyzing the effects of tumor-suppressor loss in tumors *in vivo*[42–44].

Notably, the frequency of tumor-suppressor alterations in human cancer does not directly correspond to the magnitude of their tumor-suppressor function. While variation in the mutation rates, inclusive fitness, and genetic context likely contribute to the frequency of mutations in human cancer, our findings highlight the need for rapid and quantitative methods to determine the functional importance of lower frequency putative tumor suppressors, the mutation of which may be profoundly important for individual patients.

Tuba-seq will likely contribute to our understanding of cancer pathogenesis in many other ways. It should permit the investigation of more complex combinations of tumor-suppressor gene loss and facilitate analysis of other aspects of tumor progression. Tuba-seq should be adaptable for studies of other cancer types as well as for genes that normally promote, rather than inhibit, tumor growth[8,10,45,46]. Finally, these applications of Tuba-seq may enable the investigation of genotype-specific therapeutic responses, ultimately leading to more precise and personalized patient treatment.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank P. Chu and R. Ma for technical support; A. Orantes for administrative support; C. Murray, C. Kim-Kiselak, J. Lipsick, B. Callahan, J. Sage, and members of the Petrov and Winslow laboratories for helpful comments; J. Xuhuai and the Stanford Functional Genomics Facility (S10OD018220) for advice and technical assistance; and S. Chan for sequencing expertise. I.P.W. and Z.N.R. were supported by the National Science Foundation Graduate Research Fellowship Program (GRFP). Z.N.R. was also supported by a Stanford Graduate Fellowship. C.D.M. was supported by NIH grant no. E25CA180993. D.P. is the Michelle and Kevin Douglas Professor of Biology. This work was supported by NIH grant nos. R01CA175336 and R21CA194910 (to M.M.W.), R01CA207133 (to D.P. and M.M.W.), and in part by the Stanford Cancer Institute support grant (NIH grant no. P30CA124435).

AUTHOR CONTRIBUTIONS

Z.N.R. tested sgRNA cutting efficiency; generated barcoded vectors; produced lentivirus; and performed mouse analysis, indel analysis, and analysis of single

sgRNA tumor sizes. C.D.M. performed data analysis, including processing sequencing data, designing the tumor-calling procedure, and carrying out all statistical analyses. I.P.W. selected tumor suppressors to investigate, designed sgRNAs, generated Lenti-sg*RNA/Cre* vectors, tested sgRNA cutting efficiency, produced lentivirus, and performed indel analysis. C.-H.C. performed experiments to assess the function of Smad4. D.P. and M.M.W. oversaw the project. C.D.M., Z.N.R., I.P.W., D.P., and M.M.W. wrote the manuscript with comments from all authors.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

1. Lawrence, M.S. *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495–501 (2014).
2. Kandoth, C. *et al.* Mutational landscape and significance across 12 major cancer types. *Nature* **502**, 333–339 (2013).
3. Hahn, W.C. & Weinberg, R.A. Modelling the molecular circuitry of cancer. *Nat. Rev. Cancer* **2**, 331–341 (2002).
4. Lawrence, M.S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).
5. Chin, L., Hahn, W.C., Getz, G. & Meyerson, M. Making sense of cancer genomic data. *Genes Dev.* **25**, 534–555 (2011).
6. Van Dyke, T. & Jacks, T. Cancer modeling in the modern era: progress and challenges. *Cell* **108**, 135–144 (2002).
7. Sánchez-Rivera, F.J. *et al.* Rapid modelling of cooperating genetic events in cancer through somatic genome editing. *Nature* **516**, 428–431 (2014).
8. Chiou, S.H. *et al.* Pancreatic cancer modeling using retrograde viral vector delivery and *in vivo* CRISPR/Cas9-mediated somatic genome editing. *Genes Dev.* **29**, 1576–1585 (2015).
9. Xue, W. *et al.* CRISPR-mediated direct mutation of cancer genes in the mouse liver. *Nature* **514**, 380–384 (2014).
10. Annunziato, S. *et al.* Modeling invasive lobular breast carcinoma by CRISPR/Cas9-mediated somatic genome editing of the mammary gland. *Genes Dev.* **30**, 1470–1480 (2016).
11. Bhang, H.E. *et al.* Studying clonal dynamics in response to cancer therapy using high-complexity barcoding. *Nat. Med.* **21**, 440–448 (2015).
12. Levy, S.F. *et al.* Quantitative evolutionary dynamics using high-resolution lineage tracking. *Nature* **519**, 181–186 (2015).
13. Naik, S.H. *et al.* Diverse and heritable lineage imprinting of early haematopoietic progenitors. *Nature* **496**, 229–232 (2013).
14. Nguyen, L.V. *et al.* Barcoding reveals complex clonal dynamics of *de novo* transformed human mammary cells. *Nature* **528**, 267–271 (2015).
15. Sun, J. *et al.* Clonal dynamics of native haematopoiesis. *Nature* **514**, 322–327 (2014).
16. Venkataram, S. *et al.* Development of a comprehensive genotype-to-fitness map of adaptation-driving mutations in yeast. *Cell* **166**, 1585–1596.e22 (2016).
17. Grüner, B.M. *et al.* An *in vivo* multiplexed small-molecule screening platform. *Nat. Methods* **13**, 883–889 (2016).
18. DuPage, M., Dooley, A.L. & Jacks, T. Conditional mouse lung cancer models using adenoviral or lentiviral delivery of Cre recombinase. *Nat. Protoc.* **4**, 1064–1072 (2009).
19. Jackson, E.L. *et al.* Analysis of lung tumor initiation and progression using conditional expression of oncogenic K-ras. *Genes Dev.* **15**, 3243–3248 (2001).
20. Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543–550 (2014).
21. Jackson, E.L. *et al.* The differential effects of mutant p53 alleles on advanced murine lung cancer. *Cancer Res.* **65**, 10280–10288 (2005).
22. Ji, H. *et al.* LKB1 modulates lung cancer differentiation and metastasis. *Nature* **448**, 807–810 (2007).
23. Caswell, D.R. *et al.* Obligate progression precedes lung adenocarcinoma dissemination. *Cancer Discov.* **4**, 781–789 (2014).
24. Chuang, C.H. *et al.* Molecular definition of a metastatic lung cancer state reveals a targetable CD109–Janus kinase–Stat axis. *Nat. Med.* **23**, 291–300 (2017).
25. Callahan, B.J. *et al.* DADA2: high-resolution sample inference from Illumina amplicon data. *Nat. Methods* **13**, 581–583 (2016).
26. Norton, L. A Gompertzian model of human breast cancer growth. *Cancer Res.* **48**, 7067–7071 (1988).
27. Newman, M.E.J. Power laws, Pareto distributions and Zipf's law. *Contemp. Phys.* **46**, 323–351 (2005).
28. Liu, G. *et al.* Chromosome stability, in the absence of apoptosis, is critical for suppression of tumorigenesis in *Trp53* mutant mice. *Nat. Genet.* **36**, 63–68 (2004).
29. Feldser, D.M. *et al.* Stage-specific sensitivity to p53 restoration during lung cancer progression. *Nature* **468**, 572–575 (2010).
30. Dudgeon, C. *et al.* The evolution of thymic lymphomas in p53 knockout mice. *Genes Dev.* **28**, 2613–2620 (2014).
31. Imielinski, M. *et al.* Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell* **150**, 1107–1120 (2012).
32. Schuster, K. *et al.* Nullifying the *CDKN2AB* locus promotes mutant K-ras lung tumorigenesis. *Mol. Cancer Res.* **12**, 912–923 (2014).
33. Ho, V.M., Schaffer, B.E., Karnezis, A.N., Park, K.S. & Sage, J. The retinoblastoma gene *Rb* and its family member *p130* suppress lung adenocarcinoma induced by oncogenic K-Ras. *Oncogene* **28**, 1393–1399 (2009).
34. Hernández, J. *et al.* Tumor suppressor properties of the splicing regulatory factor RBM10. *RNA Biol.* **13**, 466–472 (2016).
35. Park, I.Y. *et al.* Dual chromatin and cytoskeletal remodeling by SETD2. *Cell* **166**, 950–962 (2016).
36. Yoh, S.M., Lucas, J.S. & Jones, K.A. The Iws1:Spt6:CTD complex controls cotranscriptional mRNA biosynthesis and HYPB/Setd2-mediated histone H3K36 methylation. *Genes Dev.* **22**, 3422–3434 (2008).
37. Edmunds, J.W., Mahadevan, L.C. & Clayton, A.L. Dynamic histone H3 methylation during gene induction: HYPB/Setd2 mediates all H3K36 trimethylation. *EMBO J.* **27**, 406–420 (2008).
38. Fang, D. *et al.* The histone H3.3K36M mutation reprograms the epigenome of chondroblastomas. *Science* **352**, 1344–1348 (2016).
39. Efeyan, A. *et al.* Limited role of murine ATM in oncogene-induced senescence and p53-dependent tumor suppression. *PLoS One* **4**, e5475 (2009).
40. Gan, R.Y. & Li, H.B. Recent progress on liver kinase B1 (LKB1): expression, regulation, downstream signaling and cancer suppressive function. *Int. J. Mol. Sci.* **15**, 16698–16718 (2014).
41. Wang, T., Wei, J.J., Sabatini, D.M. & Lander, E.S. Genetic screens in human cells using the CRISPR-Cas9 system. *Science* **343**, 80–84 (2014).
42. Welford, S.M. & Giaccia, A.J. Hypoxia and senescence: the impact of oxygenation on tumor suppression. *Mol. Cancer Res.* **9**, 538–544 (2011).
43. Jones, R.G. & Thompson, C.B. Tumor suppressors and cell metabolism: a recipe for cancer growth. *Genes Dev.* **23**, 537–548 (2009).
44. Pickup, M.W., Mouw, J.K. & Weaver, V.M. The extracellular matrix modulates the hallmarks of cancer. *EMBO Rep.* **15**, 1243–1253 (2014).
45. Kirsch, D.G. *et al.* A spatially and temporally restricted mouse model of soft tissue sarcoma. *Nat. Med.* **13**, 992–997 (2007).
46. Meuwissen, R. *et al.* Induction of small cell lung cancer by somatic inactivation of both *Trp53* and *Rb1* in a conditional mouse model. *Cancer Cell* **4**, 181–189 (2003).

ONLINE METHODS

Step-by-step protocol. Protocols for plasmid barcoding and library preparation for Tuba-seq are available in the **Supplementary Protocol** and on *Protocol Exchange*[47,48].

Mice and tumor initiation. *Kras*$^{LSL\text{-}G12D}$ (*K*), *Lkb1*^{flox} (*L*), *p53*^{flox} (*P*), *R26*$^{LSL\text{-}Tomato}$ (*T*), and *H11*$^{LSL\text{-}Cas9}$ (*Cas9*) mice have been described[8,19,49–51]. Mice were on a mixed BL6/129 background. Equal numbers of males and females were used for each experiment. The number of mice used in each experiment is specified in the corresponding figure legends and total number of mice used was 50. Lung tumors were initiated by intratracheal administration of viral-Cre vectors to mice as previously described[18]. Tumor burden was assessed by fluorescence microscopy, lung weight, and histology, as indicated. All experiments were performed in accordance with Stanford University Institutional Animal Care and Use Committee guidelines.

Generation of barcoded Lenti-*mBC/Cre* and Lenti-sg*Pool/Cre* vector pools. To enable quantification of the number of cancer cells in individual tumors in parallel using high-throughput sequencing, we diversified lentiviral-Cre vectors with a short barcode sequence that would be unique to each tumor by virtue of stable integration of the lentiviral vector into the initial transduced lung epithelial cell. We generated tumors in a variety of mouse backgrounds with two different pools of barcoded lentiviral vectors. The first was a pool of ~10^6 uniquely barcoded variants of Lenti-PGK-*Cre* (Lenti-millionBC/*Cre*; Lenti-*mBC/Cre*, generated by pooling six barcoded Lenti-U6-sgRNA/PGK-*Cre* vectors), which we used to analyze the number of cancer cells in tumors induced in *Kras*$^{LSL\text{-}G12D/+}$;*R26*$^{LSL\text{-}Tomato}$ (*KT*), *Kras*$^{LSL\text{-}G12D/+}$;*p53*$^{flox/flox}$;*R26*$^{LSL\text{-}Tomato}$ (*KPT*), and *Kras*$^{LSL\text{-}G12D/+}$;*Lkb1*$^{flox/flox}$;*R26*$^{LSL\text{-}Tomato}$ (*KLT*) mice (**Fig. 1**). The second was a pool of 15 barcoded Lenti-U6-sgRNA/PGK-Cre vectors, which we used to assess the tumor-suppressive effect of candidate tumor-suppressor genes in an oncogenic Kras genetic background by infecting *KT;H11*$^{LSL\text{-}Cas9}$ (*KT;Cas9*) and *KT* mice. Our Lenti-sg*Inert/Cre* vectors included three sgRNAs that target the *NeoR* gene within the *Rosa26*$^{LSL\text{-}Tomato}$ allele—these were actively cutting, but functionally inert, negative control sgRNAs.

Design, generation, and screening of sgRNAs. We generated lentiviral vectors carrying Cre as well as an sgRNA targeting each of 11 known and putative lung adenocarcinoma tumor suppressors: sg*Lkb1*, sg*P53*, sg*Apc*, sg*Atm*, sg*Arid1a,* sg*Cdkn2a*, sg*Keap1*, sg*Rb1*, sg*Rbm10*, sg*Setd2*, and sg*Smad4*. We also generated vectors carrying inert guides: sg*Neo1*, sg*Neo2*, sg*Neo3*, sg*NT1*, and sg*NT3*. All possible 20-bp sgRNAs (using an NGG protospacer-adjacent motif (PAM)) targeting each tumor-suppressor gene of interest were identified and scored for predicted on-target cutting efficiency using an available sgRNA design/scoring algorithm[10]. For each tumor-suppressor gene, we selected three unique sgRNAs predicted to be the most likely to produce null alleles; preference was given to sgRNAs with the highest predicted cutting efficiencies as well as to sgRNAs targeting exons conserved in all known splice isoforms (ENSEMBL), closest to splice acceptor or splice donor sites, positioned earliest in the gene-coding region, occurring upstream of annotated functional domains (InterPro; UniProt), and occurring upstream of known human lung adenocarcinoma mutation sites[20,31,52–55]. Lenti-U6-sg*RNA/Cre* vectors containing each sgRNA were generated as previously described[8]. Briefly, Q5 site-directed mutagenesis (NEB E0554S) was used to insert sgRNAs into the parental lentiviral vector containing the U6 promoter as well as PGK-Cre. The cutting efficiency of each sgRNA was determined by transducing LSL–YFP;Cas9[8] cells with each Lenti-sg*RNA/Cre* virus. 48 h after transduction, flow cytometric quantification of YFP-positive cells was used to determine percent transduction. DNA was then extracted from all cells, and the targeted tumor-suppressor-gene locus was amplified by PCR.

PCR amplicons were Sanger sequenced and analyzed using TIDE analysis to quantify percent indel formation[56]. Finally, the indel percent determined by TIDE was divided by the percent transduction of LSL–YFP;Cas9 cells (as determined by flow cytometry) to determine sgRNA cutting efficiency. The most efficient sgRNA targeting each tumor-suppressor gene of interest was used for subsequent experiments. sgRNAs targeting *Tomato* and *Lkb1* have been described[7,8], and we previously validated an sgRNA targeting *p53* (data not shown). Primer sequences used to amplify target indel regions for the top guides used in this study can be found in **Supplementary Table 2**.

Barcode diversification of Lenti-sg*RNA/Cre*. After identifying the best sgRNA targeting each tumor suppressor of interest, we diversified the corresponding Lenti-sg*RNA/Cre* vector with a known 8-nucleotide ID specific to each individual sgRNA (sgID; single underline) and a 15-nucleotide random barcode (BC; double underline) (see **Fig. 4a**). A universal reverse primer (5′ AGCTAGGGATCCGCCGCATAACCAGTG 3′) and barcoded forward primer (5′ AGCTAGTCCGG**NNNNNNNN**AA**NNNNN**TT**NNNNN**AA**NNNNN**ATGCCCAAGAAGAAGAGG AAGGTGTC 3′) were used to PCR amplify a region of the Lenti-PGK-Cre vector that included the 3′ end of the *PGK* promoter and the 5′ end of *Cre*. PCR was performed using PrimeSTAR HS DNA Polymerase (premix) (Clontech, R040A), and PCR products were purified using the Qiagen PCR Purification Kit (28106). The PCR insert was digested with BspEI (NEB, R0540) and BamHI (NEB, R0136) and ligated with the Lenti-sgRNA-Cre vectors cut with XmaI (NEB, R0180) (which produces a BspEI-compatible end) and BamHI.

To generate a large number of uniquely barcoded vectors, we ligated 300 ng of each XmaI, BamHI-digested Lenti-sgRNA-Cre vector with 180 ng of each BspEI, BamHI-digested PCR product using T4 Ligase (NEB, M0202L) and standard protocols (80 µl total reaction volume). Ligations were PCR purified using the Qiagen PCR Purification Kit to remove residual salt. To obtain a pool of the greatest possible number of uniquely barcoded Lenti-sg*RNA/Cre* vectors, 1 µl of purified ligation was transformed into 20 µl of ElectroMAX DH10B cells (Thermo Fisher, 18290015). Cells were electroporated in 0.1 cm GenePulser/MicroPulser Cuvettes (Bio-Rad, 165-2089) in a BD MicroPulser Electroporator (Bio-Rad,165-2100) at 1.9 kV. Cells were then rescued by adding 500 µl media and shaking at 200 r.p.m. for 30 min at 37 °C. For each ligation, bacteria were plated on seven LB-Amp plates (one plate with 1 µl, one plate with 10 µl, and five plates with 100 µl). The following day, colonies were counted on the 1 µl or 10 µl plate to estimate the number of colonies on the 100 µl plates, and this was used as an initial estimation of number of unique barcodes associated with each ID.

10 ml of liquid LB-Amp was added to each plate of bacteria to pool the colonies. Colonies were scraped off of the plates into the liquid, and all plates from each transformation were combined into a flask. Flasks were shaken at 200 r.p.m. for 30 min at 37 °C to mix. DNA was Midi prepped using the Qiagen HiSpeed MidiPrep Kit (12643). DNA concentrations were determined using a Qubit dsDNA HS Kit (Invitrogen, Q32851).

As a quality-control measure, the sgID-BC region from each Lenti-sgRNA-sgID-BC/Cre plasmid pool was PCR amplified with GoTaq Green polymerase (Promega M7123) following the manufacturer's instructions. These PCR products were Sanger sequenced (Stanford PAN facility) to confirm the expected sgID and the presence of a random BC. Since BspEI and XmaI have compatible overhangs but different recognition sites, the Lenti-sgRNA-sgID-BC/Cre vectors generated from successful ligation of the sgID/BC lack an XmaI site. Thus, for pools that had a detectable amount of unbarcoded parental Lenti-sgRNA/Cre plasmid as determined by Sanger sequencing (>5%), we destroyed the parental unbarcoded vector by digesting the pool with XmaI (NEB, 100 µl reaction) using standard methods. These redigested plasmid pools were repurified using the Qiagen PCR Purification Kit, and concentration was redetermined by NanoDrop.

Generation of Lenti-*mBC/Cre* and Lenti-*TS-Pool/Cre*. To obtain a library with approximately 10^6 associated barcodes to use in our initial experiments in mice that lacked the *H11*$^{LSL\text{-}Cas9}$ allele, we pooled six sgID-BC barcoded vectors (*sgLkb1*, *sgp53*, *sgNeo1*, *sgNeo3*, *sgNT1*, and *sgNT3*) to create Lenti-million Barcode/Cre (Lenti-*mBC/Cre*). We then pooled the barcoded Lenti-sgRNA-sgID-BC/Cre vectors (*sgLkb1*, *sgp53*, *sgApc*, *sgAtm*, *sgArid1a*, *sgCdkn2a*, *sgKeap1*, *sgNeo1*, *sgNeo2*, *sgNeo3*, *sgNT1*, *sgRb1*, *sgRbm10*, *sgSetd2*, and *sgSmad4*) to generate Lenti-sg*TS-Pool/Cre*. All plasmids were pooled at equal ratios as determined by Qubit concentration before lentivirus production.

Production, purification, and titering of lentivirus. Lentiviral vectors were produced using polyethylenimine (PEI)-based transfection of 293T cells with the lentiviral vectors, delta8.2 and VSV-G packaging plasmids. Lenti-*mBC/Cre*, Lenti-*sgTS-Pool/Cre*, Lenti-sg*Tomato/Cre*, Lenti-sg*Lkb1/Cre*, Lenti-sg*Setd2#1/Cre*, Lenti-sg*Setd2#2*, Lenti-sg*Neo2/Cre*, and Lenti-sg*Smad4/Cre* were generated for tumor initiation. Sodium butyrate (Sigma Aldrich, B5887) was added at a final concentration of 0.2 mM 8 h after transfection to increase production of viral particles. Virus-containing media were collected 36, 48, and 60 h after transfection, concentrated by ultracentrifugation (25,000 r.p.m. for 1.5–2 h), resuspended overnight in PBS, and frozen at −80 °C. Concentrated lentiviral particles were titered by infecting LSL–YFP cells (a gift from A. Sweet-Cordero, University of California, San Francisco), determining the percent YFP-positive cells by flow cytometry, and comparing the infectious titer with a lentiviral preparation of known titer.

Generation of 'benchmark' cell lines. Three uniquely barcoded Lenti-Cre vectors with the sgID "TTCTGCCT" were used to generate benchmark cell lines that could be spiked into each bulk-tumor-bearing lung sample at a known cell number to enable the calculation of the neoplastic cell number within each tumor. Plasmid DNA from individual bacterial colonies was isolated using the Qiagen QIAprep Spin Miniprep Kit (27106). Clones were Sanger sequenced, lentivirus was produced as described above, and LSL–YFP cells were infected at a very low multiplicity of infection, such that approximately 3% of cells were YFP positive after 48 h. Infected cells were expanded and sorted using a BD Aria II (BD Biosciences). YFP-positive sorted cells were replated and expanded to obtain a large number of cells. After expansion, cells were reanalyzed for percent YFP-positive cells on a BD LSR II analyzer (BD Biosciences). Using this percentage, the number of total cells needed to contain 5×10^5 integrated barcoded lentiviral vectors was calculated for each of the three cell lines, and cells were aliquoted and frozen based on this calculation.

Summary of all mouse infections. Refer to **Supplementary Table 3**.

Isolation of genomic DNA from mouse lungs. For experiments in which barcode sequencing was used to quantify the number of cancer cells in each tumor, the whole lungs from each mouse were homogenized using a Fisher TissueMeiser. 5×10^5 cells from each of the three individually barcoded benchmark cell lines were added before homogenization. Tissue was homogenized in 20 ml lysis buffer (100 mM NaCl, 20 mM Tris, 10 mM EDTA, 0.5% SDS) with 200 µl of 20 mg/ml Proteinase K (Life Technologies, AM2544). Homogenized tissue was incubated at 55 °C overnight. To maintain accurate representation of all tumors, DNA was phenol–chloroform extracted and ethanol precipitated from ~1/10th of the total lung lysate using standard protocols. For lungs weighing less than 0.3 g, DNA was extracted from ~1/5th of the total lung lysate, and for those weighing less than 0.2 g, DNA was extracted from ~3/10th of the total lung lysate to increase DNA yield.

Preparation of sgID-BC libraries for sequencing. Libraries were prepared by amplifying the sgID-BC region from 32 µg of genomic DNA per mouse. The sgID-BC region of the integrated Lenti-sg*RNA-BC/Cre* vectors was PCR amplified using one of 24 primer pairs that contain TruSeq Illumina adapters and a 5′ multiplexing tag (TruSeq i7 index region indicated by underline). This amplification protocol uses a universal forward primer (5′ AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTAC ACGACGCTCTTCCGATCTGCGCACGTCTGCCGCGCTG 3′) and a unique reverse primer (5′ CAAGCAGAAGACGGCATAC GAGAT<u>NNNNNN</u>GTGACTGGACTTCAGACGTGTGCTCTT CCGATCCAGGTTCTTGCGAACCTCAT 3′).

We used a single-step PCR amplification of sgID-BC regions, which we found to be a highly reproducible and quantitative method to determine the number of neoplastic cells in each tumor. We performed eight 100 µl PCR reactions per mouse (4 µg DNA per reaction) using OneTaq 2X Master Mix with Standard buffer (NEB, M0482L) with the following PCR program: (Step 1) 94C 10 min, (Step 2) 94C 30 s, (Step 3) 55C 30 s, (Step 4) 68C 30 s, (Step 5) go back to step 2 (34 x), (Step 6) 68C 7 min, (Step 7) 4C infinity. Pooled PCR products were isolated by gel electrophoresis and gel extracted using the Qiagen MinElute Gel Extraction kit. The concentration of purified PCR products from individual mice was determined by Bioanalyzer (Agilent Technologies) and pooled at equal ratios. Samples were sequenced on an Illumina HiSeq to generate 100 bp single-end reads (ELIM Biopharmaceuticals, Inc.).

doi:10.1038/nmeth.4297

Identifying distinct sgRNAs and tumors via ultradeep sequencing. The unique sgID-BC identifies tumors. These sgID-BCs were detected via next-generation sequencing on an Illumina HiSeq. The size of each tumor, with respect to cell number, was expected to roughly correspond to the abundance of each unique sgID-BC. Because tumor sizes varied by factors larger than the rate of read sequencing errors, distinguishing true tumors from recurrent read errors required careful analysis of the deep-sequencing data.

To this end, tumors and their respective sgRNAs were identified in three steps: (i) abnormal and low-quality reads were discarded from the ultradeep sequencing runs, (ii) unique barcode pileups that we predicted to arise from the same tumor were bundled into groups, and (iii) cell number was estimated from these bundles in the manner that proved most reproducible.

Read preprocessing. Reads contained a two-component DNA barcode (an 8-nucleotide sgID and a 21-nucleotide barcode sequence that contains 15 random nucleotides) that began 49 nucleotides downstream of our forward primer. We discarded unusual reads—specifically, those that lacked the flanking lentiviral sequences, those that contained unexpected barcodes, and those with high error rates. This was accomplished in three steps (**Supplementary Fig. 2a**):

1. We examined the 12 lentiviral nucleotides immediately upstream and downstream of the sgID-BC. These 12 nucleotides were identified using pairs of adjacent 6-mer search strings, such that each 6-mer could tolerate one mismatch. Although we expected these 12 nucleotides to begin at position 37 within the read, we did not require this positioning or leverage this information. A nested 6-mer approach (with two opportunities to identify the lentiviral sequences flanking the sgID-BC) was used to minimize read discarding. For ~7–8% of reads, this 2nd 6-mer match salvaged the read; i.e., the 6-mers immediately flanking the sgID-BC deviated from the reference sequence by more than one nucleotide, yet the 6-mers immediately outside of these inner 6-mer sequences were recognizable and allowed us to salvage the read and identify the barcodes. Salvaging reads is not particularly critical for estimating tumor sizes; however, it is critical for accurate estimation of read error rates, because the nonbarcoded regions of our reads were used to estimate sequencing error rates and, therefore, should not be biased against read errors.
2. We then discarded reads in which the sgID-BC deviated in length by greater than two nucleotides in either direction. Because our first barcode was expected to contain one of the 15 sgIDs, we discarded reads that did not match one of these 15 sequences. One mismatch and one indel were permitted in the matching.
3. We then end trimmed each read such that 18 bp flanked either end of the sgID-BC. We then filtered the trimmed reads according to quality score, retaining those that were predicted to contain no more than two sequencing errors[57]. We also discarded reads with uncalled bases in the second (random) barcode and rectified uncalled bases elsewhere.

In these three stages, 14% of reads were discarded at stage one, ~7% at stage two, and <2% at stage three.

We then examined those reads that failed at each stage. By performing BLAST searches, we determined that those reads discarded at stage one often contained uninformative sequences corresponding to artifacts from either our preparation (Phi X bacteriophage genome and mouse genome) or other samples paired with us on the lane (common plasmid DNAs). In stage two, we found that reads with aberrant barcode lengths often contained large indels or had one or both of their sgID-BCs completely missing. Lastly, very few reads were discarded in stage three on account of the fact that internal regions of the reads exhibited higher quality scores than those of the termini of reads. As a consequence of this trend, it is common practice to end trim reads before discarding those reads predicted to contain more than two sequencing errors[25], as we did.

Clustering of unique read pileups via DADA2. sgID-BC reads were aggregated into sets of identical sequences and counted. The counts of unique DNA barcode pairs do not directly correspond to unique tumors, because large tumors are expected to generate recurrent sequencing errors (**Supplementary Fig. 2b**). We therefore spent considerable effort developing a method to distinguish small tumors from recurrent sequencing errors arising from large tumors. Consider, for example, that a tumor of 10 million cells will produce sequencing-error pileups that mimic a 10,000–100,000-cell tumor, if the error rate is 0.1–1% (a typical rate, given the limitations of PCR amplification and Illumina sequencing machines). DADA2 has previously been used to address this issue in barcoding experiments involving ultradeep sequencing[12]. However, because DADA2 was designed for ultradeep sequencing of full-length Illumina amplicons[25], we had to tailor and calibrate it for our purposes.

In DADA2, the likelihood that barcode pileups will result from a recurrent sequencing error of a larger pileup depends upon (i) the abundance of the larger pileup, (ii) the specific differences in nucleotide sequence between the smaller and larger pileups, and (iii) the average quality scores of the smaller pileup at the variant positions.

Factors i and ii are at first considered heuristically (to maximize computational speed) and then more precisely (when needed) via a Needleman–Wunsch algorithm. DADA2 splits a cluster into two when the probability that a smaller pileup was generated by sequencing errors is less than Ω. This value therefore represents a threshold for splitting larger clusters. When this threshold is large, read pileups are split permissively (many called tumors, perhaps dividing large tumors); and when Ω is small, read pileups are split restrictively (few called tumors, perhaps aggregating distinct small tumors).

The likelihood of sequencing errors was inferred from our ultradeep sequencing data. Phred quality scores provide a theoretical estimate of sequencing error rates; however, these estimates tend to vary from Illumina machine to Illumina machine and do not account for the specifics of our protocol (including, e.g., occasional errors introduced via PCR amplification despite our use of high-fidelity polymerase). Ordinarily, DADA2 will estimate sequencing error rates simultaneously with the unique DNA clusters; however, our lentiviral constructs had nondegenerate regions outside of our sgID-BC region that were used to estimate sequencing error rates directly. Moreover, estimating error rates and barcode clusters jointly is more computationally intensive,

requiring greater than 20,000 central processing unit (CPU) h for clustering our entire data set and exploring the relevant clustering parameters.

A sequencing error model was trained to each Illumina machine by:

1. Generating training pseudoreads by concatenating the 18 nucleotides immediately upstream of our sgID-BC with the 18 nucleotides immediately downstream of the barcodes.
2. Clustering these pseudoreads using a single run of DADA2.
3. Using the error rates estimated from this training run to cluster the sgID-BC using a single run of DADA2.

We used a very low value of $\Omega = 10^{-100}$ to estimate sequencing errors in the training run, as we expected only one cluster of lentiviral sgID-BC-flanking sequences. Altering this value does not affect training results appreciably, but we nonetheless occasionally observed very small derivative clusters from our lentiviral sequence even at this value. These derivative clusters are presumably rare DNA artifacts and never amounted to >2% of our processed reads. We used a very stringent DADA2 run to estimate sequencing errors, because a more permissive threshold might overfit sequencing errors and underestimate sequencing error rates, while the less permissive approach of estimating error rates directly from each read's deviance from expectation (akin to a DADA2 run where $\Omega = 0$) would not accommodate any DNA artifacts in our data and, therefore, would overestimate sequencing error rates.

We trained sequencing error rates on each Illumina machine used in this study (seven in total). Training allowed the probability of every substitution type (A→C, A→T, etc.) to be estimated. The error rates as a function of Phred quality score were determined using LOESS regression of the available data (**Supplementary Fig. 2c**)[25]. In general, error rates were approximately two to three times higher than predicted by the Phred quality scores for transversions (and approximately consistent with expectations for transitions). This elevated error rate is typical[25] and may reflect miscalibration of the machines and/or be due to mutations introduced during PCR.

We then clustered the dual barcodes that passed our preprocessing filters using DADA2. Barcodes were given seven nucleotides of nondegenerate lentiviral flanking regions so that any indels within the barcodes could be identified (without adequate flanking sequences, DNA alignment algorithms sometimes miscall indels as multiple point mutations). During clustering, we also required (i) that clusters deviate from each other by at least two bases (MIN_HAMMING_DISTANCE = 2), (ii) that new clusters only be formed when pileup size exceeded expectations under the error process by at least a factor of two (MIN_FOLD = 2), and (iii) that the Needleman–Wunsch algorithm consider only alignments with at most four net insertions or deletions (BAND_SIZE = 4, VECTORIZED_ALIGNMENT = FALSE). None of these choices affected the results appreciably, but they increased computational performance and offered additional verification that barcodes were aggregated into tumors of reasonable size.

Vetting and calibration of pipeline. We sequenced our first PCR-amplified, multiplexed DNA libraries (from *KT*, *KLT*, and *KPT* tumors) in triplicate to vet and design our tumor-calling approach.

Reproducibility was measured in three ways: (i) by measuring correlation between estimated cell abundances for all barcodes and all mice, (ii) by measuring the variation in the number of lesions called for each sgID in each mouse in our first experiment, and (iii) by measuring the variation in LN mean size for each sgID—a value that should be constant in mice that do not express Cas9. Because the read depth of our triplicate run naturally varied (40.1×10^6, 22.2×10^6, and 34.9×10^6 reads after preprocessing), these three runs were performed on distinct Illumina machines with different sequencing error rates; and, because our initial lentiviral pool contained six different sgIDs with varying levels of barcode diversity, the technical variability in our vetting process approximated the technical variability of later experiments. In our tumor-size analysis pipeline, we found:

1. The mean abundance of our three 'benchmark' DNA barcodes was more reproducible between replicate runs than was the median abundance. Thus, this mean value of benchmark read abundance (corresponding to 500,000 cells) was used to convert read abundance into the absolute cell number of cancer cells in each tumor (**Supplementary Fig. 3**).
2. Ignoring reads with ≥2 errors from the consensus barcode of a cluster improved reproducibility. Typically, ~80–90% of reads in a barcode cluster were exact matches to the consensus barcode; while ~5% of reads were single errors from this read, and ~5–15% of reads deviated at ≥2 errors. These reads, with ≥2 errors, were poorly correlated between replicate runs and hampered our ability to reproducibly estimate absolute cell number/tumor size.
3. The cluster-splitting proclivity of DADA2 was thresholded at $\Omega = 10^{-10}$, and we required that lesions contain ≥500 cells for **Figures 1**–**3** and ≥1,000 cells for **Figures 4**–**5** to maximize reproducibility between replicate runs (**Supplementary Fig. 2d**–**f**). Threshold parameters with high specificity (small Ω, high minimum cell number) called lesion sizes more reproducibly, whereas threshold parameters with high sensitivity (large Ω, low minimum cell number) called lesion quantities more reproducibly. Overprioritizing only one facet of reproducibility would be imprudent. With two thresholds, considering different facets of measurement error, we better balanced these competing priorities.

With this pipeline, we interrogated the diversity of the barcode in our screen in several ways. First, we confirmed that nucleotides in this barcode were evenly distributed among A's, T's, C's, and G's (**Supplementary Fig. 4b**). Second, we found no evidence for an excess of repeated strings (e.g., AAAAA sequences). Third, we calculated the number of random barcodes paired to each sgID in our lentiviral pool. Because of the large number of uniquely barcoded variants of each vector that we generated through our barcode ligation approach (see "Barcode diversification of Lenti-sg*RNA/Cre*"), most barcodes that exist in our lentiviral pool were never detected in any lesions in any of the experiments (because diversity is much higher that total lesion number). Nonetheless, we still inferred the amount of barcode diversity from the observed barcodes.

To infer the barcode diversity of each sgID, we assumed that the probability of observing a barcode in i mice is Poisson distributed; $P(k = i; \lambda) = \lambda^k e^{-\lambda}/k!$, where $\lambda_r = L_r/D_r$ is a ratio of the number of called lesions L_r for each sgID r in our entire data set (a known quantity) divided by the total number of unique barcodes D_r for each sgID (our quantity of interest). By noting that $\lambda_r/(1 - e^{-\lambda r}) = \mu_{\text{nonzero}}$,

doi:10.1038/nmeth.4297

where $\mu_{\text{nonzero}} = \Sigma_{i=1}^{\infty} P(k = i; \lambda_r)$ is simply the mean number of occurrences of each barcode that occurred once or more, we calculated D_r. Across our entire data set, the average probability of the same barcode initiating two distinct tumors in the same mouse was 0.91%.

Good barcode diversity is also demonstrated by the highly reproducible mean size of the six sgIDs in the Lenti-mBC/Cre experiment. If barcode diversity was low, and barcodes overlapped often within a mouse, then the mean sizes of the less diverse sgIDs would increase—as two distinct tumors with the same barcode would be bundled together. However, the mean sizes of tumors containing each sgID vary by <1% within replicate mice, thus refuting the possibility that variation in barcode diversity causes overbundling of tumors. We also assessed our ability to call sgIDs accurately, despite sequencing errors, by processing deep-sequencing runs in two ways—by identifying each read's cognate sgID before clustering based on the raw read sequence or by identifying cognate sgIDs after clustering based on the consensus sequence of the cluster. Using either approach, 99.8% of reads paired to the same cognate sgID, which provided assurance that sgIDs were accurately identified. We opted to employ the latter (after clustering) approach for our final analysis.

By thoroughly developing and vetting our tumor-calling pipeline, we salvaged an extra decade of size resolution (i.e., we could faithfully resolve tumors that were ten-fold smaller than we would have otherwise been able to resolve). Our three DNA benchmarks (added to the lung samples at the very beginning of DNA preparation) (**Supplementary Fig. 3**) offer a glimpse of this resolution. Sequencing errors of the DNA benchmarks are easily identified by the DNA benchmark's unique sgID and known secondary barcodes. While these sequencing errors are usually discarded, we can treat them as ordinary read pileups and observe the properties of potential sequencing errors. Without our calibrated analysis pipeline, the sequencing errors appear as lesions of ~10^3 cells; with our pipeline, these sequencing errors emerge as lesions of ~10^2 cells—below our minimum cell threshold (**Fig. 2a**).

More importantly, our pipeline is robust to technical perturbations. We more intensively profiled reproducibility with two additional technical perturbations in two specific mice from the first experiment. First, a *KLT* 11-week mouse (JB1349) was sequenced at great depth and then randomly downsampled ten-fold to typical read depth (this downsampling was greater than any variability in read depth actually detected throughout our study). Lesion sizes were very highly correlated in this first perturbation (**Supplementary Fig. 4e**,**f**). Additionally, a *KT* 11-week mouse (IW1301) was amplified in two PCR reactions with different multiplexing tags (**Fig. 2b**,**c**). PCR and multiplexing appear to hamper reproducibility more than read depth, although reproducibility is good overall. These mice also display two encouraging reproducibility trends: (i) larger lesions/tumors were most consistent between replicates, and (ii) the overall shapes (histogram) of tumor lesion sizes were better correlated between the replicates than between individual tumors. The excellent reproducibility of size histograms suggests that noise in our tumor size calls is generally unbiased.

Minimizing the influence of GC amplification bias on tumor-size calling. We define each tumor in our study by a size T_{mrb} corresponding to the mouse m that harbored it, the cognate sgRNA r identified by its first barcode, and a unique barcode sequence (consensus of the DADA2 cluster) b. Given the approximately log-normal structure of our data (**Fig. 3d** and **Supplementary Note**, **Fig. 1a** data not shown), we log transformed and normalized sizes such that $\tau_{mrb} = \text{Ln}(T_{mrb}/E_{mr}[T_{mrb}])$. Here $E_{mr}[T_{mrb}] = \Sigma_b T_{mrb}/N_{mr}$ is the expected lesion size for a given mouse m and sgRNA r, and we will use this notation for expectation values. This notation—where aggregated indices are dropped from subscripts—is used throughout. GC biases were subtle; the coefficient of variation (CV) of $E_{mr}[T_{mrb}]$ was 5.0%. This marginal distribution still exhibited a subtle dependence on the GC content of the combined barcode sequence that was best described by a 4$^{\text{th}}$-order least-squares polynomial fit $f_4(b)$ of $E_b[\tau_{mrb}]$ (adjusted $r^2 = 0.994$). The sgIDs were all designed with well-balanced GC content; however, the second barcode comprised random sequences. While the multinomial process of generating barcodes made intermediate levels of GC content most common, some deviation of GC content was observed. Maximal values of $f_4(b)$ arise at intermediate GC content, suggesting that PCR biases amplification toward template DNA of intermediate melting temperature. We subtracted the effects of this GC bias from log-transformed values: $t_{mrb} = \text{Ln}[T_{mrb}] - f_4(b)$. This correction alters tumor sizes by 5% on average.

Analysis of indels at target sites. To confirm CRISPR–Cas9-induced indel formation *in vivo*, the targeted region of each gene of interest was PCR amplified from genomic DNA extracted from bulk-tumor-bearing lung samples using GoTaq Green polymerase (Promega M7123) and primer pairs that yield short amplicons amenable to paired-end sequencing. Primers can be found in **Supplementary Table 4**.

PCR products were either gel extracted or purified directly using the Qiagen MinElute kit. DNA concentration was determined using the Qubit HS assay (Thermo Fisher, Q32851), following the manufacturer's instructions. All 14 purified PCR products were combined in equal proportions for each mouse. TruSeq Illumina sequencing adapters were ligated on to the pooled PCR products with a single multiplexing tag per mouse using SPRIworks (Beckman Coulter, A88267) with standard protocols. Sequencing was performed on the Illumina HiSeq to generate single-end, 150-bp reads (Stanford Functional Genomics Facility).

Custom Python scripts were used to analyze the indel sequencing data. For each of the 14 targeted regions, an 8-mer was selected on either side of the targeted region to generate a 46 bp region. Reads were required to contain both anchors, and no sequencing errors were allowed. The length of each fragment between the two anchors was then determined and compared with the expected length. Indels were categorized according to the number of base pairs inserted or deleted.

The percent of indels for each individual locus in each individual mouse was calculated as follows:

$$\%\text{Indels} = \frac{\text{Total reads} - \text{wildtype reads}}{\text{Total reads}}$$

Then the average percent of indels in the three Neo loci was calculated, and the percent of indels at every other targeted locus was normalized to this value to generate the percent of indels relative to Neo that are plotted in **Supplementary Figure 12a**.

Calculation of *in vitro* cutting efficiency using the Lenti-*TS-Pool/Cre* virus. Cas9-expressing cell lines were infected with

Lenti-*TS-Pool/Cre* virus and harvested after 48 h. gDNA was extracted, and targeted loci were amplified using the above primers (see "Analysis of indels at target sites"). First, all primers were pooled, and 15 rounds of PCR were performed using GoTaq Green polymerase (Promega M7123). These products were then used for subsequent amplification with individual primer pairs as described above. Sequencing libraries were prepared as described above.

Histology, immunohistochemistry, and tumor analysis. Samples were fixed in 4% formalin and paraffin embedded. Immunohistochemistry was performed on 4 µm sections with the ABC Vectastain kits (Vector Laboratories) with antibodies against Tomato (Rockland Immunochemicals, 600-401-379), Smad4 (AbCam, AB40759) and Sox9 (EMD Milipore, AB5535). Sections were developed with DAB and counterstained with haematoxylin. Haematoxylin and eosin staining was performed using standard methods.

Sections from lungs infected with Lenti-sg*Tomato/Cre* were stained for Tomato, and tumors were scored as positive (>95% Tomato-positive cancer cells), Negative (no Tomato-positive cancer cells), or mixed (all other tumors). Tumors were classified and counted from a single section through all lung lobes from four independent mice.

Quantification of tumor area and barcode sequencing of tumors induced with Lenti-sg*Setd2* and Lenti-sg*Neo*. Tumor-bearing lung lobes from mice with Lenti-sg*Setd2#1/Cre,* Lenti-sg*Setd2#2/Cre* or Lenti-sg*Neo2/Cre*-initiated tumors were fixed, embedded in paraffin, sectioned, and stained with haematoxylin and eosin. Percent tumor area was determined using ImageJ.

The distribution of the number of neoplastic cells in individual tumors in *KT;Cas9* mice infected with Lenti-sg*Setd2#1/Cre* and Lenti-sg*Neo2/Cre* was assessed by Illumina sequencing of their respective lentiviral barcodes and subsequent Tuba-seq analysis as described above.

Western blotting for Lkb1 and Cas9. Microdissected Tomato-positive lung tumors from *KT* and *KT;Cas9* mice with Lenti-sg*Lkb1/Cre* initiated tumors were analyzed for Cas9 and Lkb1 protein expression. Samples were lysed in RIPA buffer and boiled with LDS loading dye. Denatured samples were run on a 4–12% Bis–Tris gel (NuPage) and transferred onto a PVDF membrane. Membranes were immunoblotted using primary antibodies against Hsp90 (BD Transduction Laboratories, 610419), Lkb1 (Cell Signaling, 13031P), Cas9 (Novus Biologicals, NBP2-36440), and secondary HRP-conjugated anti-mouse (Santa Cruz Biotechnology, sc-2005) and anti-rabbit (Santa Cruz Biotechnology, sc-2004) antibodies.

Survival analysis of mice with *Cas9*-mediated inactivation of *Smad4*. To investigate tumor suppression by *Smad4*, *KT* and *KT;Cas9* mice were infected intratracheally with 10^5 Lenti-sg*Smad4/Cre*. Mice were sacrificed when they displayed visible signs of distress to assess survival.

Protocols and vectors. Protocols for generation of barcoded vectors and library preparation for Tuba-seq analysis have been uploaded to *Protocol Exchange*[47,48], and the following unbarcoded Lenti-pLL3.3-sgRNA/Cre vectors are available via AddGene: Lenti-sgNT1/Cre (AddGene ID: 66895), Lenti-sgNT3/Cre (AddGene ID: 89654), Lenti-sgNeo1/Cre (AddGene ID: 67594), Lenti-sgNeo2/Cre (AddGene ID: 89652), Lenti-sgNeo3/Cre (AddGene ID: 89653), Lenti-sgSmad4/Cre (AddGene ID: 89651), Lenti-sgSetd2#1/Cre (AddGene ID: 89649), Lenti-sgSetd2#2/Cre (AddGene ID: 89650), Lenti-sgRbm10/Cre (AddGene ID: 89648), Lenti-sgRb1/Cre (AddGene ID: 89647), Lenti-sgp53/Cre (AddGene ID: 89646), Lenti-sgKeap1/Cre (AddGene ID: 89645), Lenti-sgCdkn2a/Cre (AddGene ID: 89644), Lenti-sgAtm/Cre (AddGene ID: 89643), Lenti-sgArid1a/Cre (AddGene ID: 89642), Lenti-sgApc/Cre (AddGene ID: 89641), and Lenti-sgLkb1/Cre (AddGene ID: 66894).

Code availability. User-friendly code has been made available at https://github.com/petrov-lab/tuba-seq.

Data availability statement. Raw sequencing data is publicly available on GEO (GSE98207).

47. Rogers, Z. Barcoding lentiviral Cre vectors for use in experiments involving downstream Tuba-seq analysis. *Protocol Exchange* http://dx.doi.org/10.1038/protex.2017.090.
48. Rogers, Z. Genomic DNA Isolation from Tissue Samples and Library Prep for Tuba-Seq Barcode Analysis. *Protocol Exchange* http://dx.doi.org/10.1038/protex.2017.091.
49. Madisen, L. *et al.* A robust and high-throughput Cre reporting and characterization system for the whole mouse brain. *Nat. Neurosci.* **13**, 133–140 (2010).
50. Jonkers, J. *et al.* Synergistic tumor suppressor activity of BRCA2 and p53 in a conditional mouse model for breast cancer. *Nat. Genet.* **29**, 418–425 (2001).
51. Nakada, D., Saunders, T.L. & Morrison, S.J. Lkb1 regulates cell cycle and energy metabolism in haematopoietic stem cells. *Nature* **468**, 653–658 (2010).
52. Cerami, E. *et al.* The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* **2**, 401–404 (2012).
53. Gao, J. *et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.* **6**, pl1 (2013).
54. Rizvi, N.A. *et al.* Cancer immunology. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. *Science* **348**, 124–128 (2015).
55. Ding, L. *et al.* Somatic mutations affect key pathways in lung adenocarcinoma. *Nature* **455**, 1069–1075 (2008).
56. Brinkman, E.K., Chen, T., Amendola, M. & van Steensel, B. Easy quantitative assessment of genome editing by sequence trace decomposition. *Nucleic Acids Res.* **42**, e168 (2014).
57. Edgar, R.C. & Flyvbjerg, H. Error filtering, pair assembly and error correction for next-generation sequencing reads. *Bioinformatics* **31**, 3476–3482 (2015).

doi:10.1038/nmeth.4297