Title
Multiplexed in vivo homology-directed repair and tumor barcoding enables parallel quantification of Kras variant oncogenicity.

Permalink
https://escholarship.org/uc/item/4hh022tq

Journal
Nature communications, 8(1)

ISSN
2041-1723

Authors
Winters, Ian P
Chiou, Shin-Heng
Paulk, Nicole K
et al.

Publication Date
2017-12-12

DOI
10.1038/s41467-017-01519-y

Peer reviewed
Multiplexed in vivo homology-directed repair and tumor barcoding enables parallel quantification of Kras variant oncogenicity

Ian P. Winters1, Shin-Heng Chiou1, Nicole K. Paulk1,2, Christopher D. McFarland3, Pranav V. Lalgudi1, Rosanna K. Ma1, Leszek Lisowski2,4,5, Andrew J. Connolly6, Dmitri A. Petrov3, Mark A. Kay1,2 & Monte M. Winslow1,6,7,8

Large-scale genomic analyses of human cancers have cataloged somatic point mutations thought to initiate tumor development and sustain cancer growth. However, determining the functional significance of specific alterations remains a major bottleneck in our understanding of the genetic determinants of cancer. Here, we present a platform that integrates multiplexed AAV/Cas9-mediated homology-directed repair (HDR) with DNA barcoding and high-throughput sequencing to simultaneously investigate multiple genomic alterations in de novo cancers in mice. Using this approach, we introduce a barcoded library of non-synonymous mutations into hotspot codons 12 and 13 of Kras in adult somatic cells to initiate tumors in the lung, pancreas, and muscle. High-throughput sequencing of barcoded KrasHDR alleles from bulk lung and pancreas reveals surprising diversity in Kras variant oncogenicity. Rapid, cost-effective, and quantitative approaches to simultaneously investigate the function of precise genomic alterations in vivo will help uncover novel biological and clinically actionable insights into carcinogenesis.
Although somatic mutations in oncogenes typically cluster within a few specific residues, the resulting amino acid changes can produce diverse oncogenic variants that have dramatically different biochemical properties and which correlate with distinct clinical outcomes\textsuperscript{1,2}. The functional consequences of specific oncogenic alterations have been investigated in cell lines and in genetically engineered mouse models. Overexpression of putative oncogenic variants in cell lines can uncover some aspects of their function, but the interpretation of these studies is complicated by the non-physiologic expression of variants, presence of cell line-specific background alterations, and lack of a native in vivo environment\textsuperscript{3}. Conversely, tumors in genetically engineered mice are driven by defined mutations expressed at physiological levels and develop within their natural context. Tumors initiated in these autochthonous mouse models also recapitulate the gene expression programs and histopathological progression of human cancers, including the development of invasive and metastatic disease\textsuperscript{4}. However, genetically engineered mouse models are greatly limited by the time and cost associated with their development and use\textsuperscript{5}.

**KRAS** is the most frequently mutated oncogene in human cancer\textsuperscript{3}. Given the immense importance of oncogenic RAS proteins in human cancer, and the renewed hope for therapeutics targeting this family of proteins, there has been a resurgence of interest in understanding the functional properties of diverse oncogenic variants of RAS proteins\textsuperscript{6}. **KRAS** commonly harbors non-synonymous point mutations in codon 12 or 13 that result in diverse amino acid substitutions. Interestingly, certain point mutations in **KRAS** codon 12 or 13 are identified much more frequently than others, which is thought to be a product of non-uniform mutation rates as well as biological differences between distinct oncogenic **KRAS** variants\textsuperscript{1,7,8}.

Although conventional genetically engineered mice have been used to model oncogenic **KRAS**-driven cancers, very few Kras variants have been studied in autochthonous mouse models. For instance, oncogenic **KRAS**-driven lung cancer has been almost exclusively modeled using knock-in alleles in which Cre-mediated or spontaneous recombination leads to **Kras**\textsuperscript{G12D} expression\textsuperscript{9,10}. However, **KRAS**\textsuperscript{G12D} represents <20\% of KRAS mutations in human lung adenocarcinomas\textsuperscript{9,11}. Nonetheless, data from Kras\textsuperscript{G12D}-based mouse models are often extrapolated to make claims about "oncogenic KRAS" in general, while pre-clinical studies performed in mice with Kras\textsuperscript{G12D}-driven tumors may not predict the response of all oncogenic KRAS-driven human tumors. Our limited understanding of the in vivo oncogenicities of even the most common variants of KRAS underscores the fundamental and urgent need for rapid, quantitative, and precise in vivo methods to model the diverse genetic alterations that occur in human cancers.

**CRISPR/Cas9-mediated somatic genome editing** has recently been used to rapidly investigate the function of tumor suppressor genes in several autochthonous mouse models of cancer\textsuperscript{4,12}. Although the CRISPR/Cas9 system can be used to generate loss-of-function mutations at targeted sites through error-prone non-homologous end-joining, the addition of a homologous DNA template permits the introduction of defined sequences at desired sites through homology-directed repair (HDR). Cas9-mediated HDR has enabled precise genomic editing in cell lines and organoids, and HDR has also been used in mice to introduce reporter genes, correct disease alleles, and introduce targeted point mutations in several adult tissues\textsuperscript{4,12–17}.

By circumventing the need to generate a new genetically engineered mouse line for every oncogenic allele of interest, Cas9-mediated HDR has the potential to provide an enormous decrease in both the time and cost associated with modeling cancers with defined oncogenic driver mutations in vivo. Here, we report a method for efficient de novo cancer modeling in mice using AAV/Cas9-mediated HDR. We demonstrate that this approach enables multiplexed somatic HDR to produce genetically diverse tumors within individual mice. Furthermore, while current autochthonous cancer models rely on semiquantitative techniques to assess tumor burden, we describe an HDR-based DNA barcoding strategy that allows for quantitative tumor analyses using high-throughput sequencing. The integration of AAV/
Cas9-mediated HDR with tumor barcoding and deep sequencing represents a novel platform to rapidly and quantitatively interrogate the function of multiple precise mutations simultaneously in mouse models of diverse human cancer types.

**Results**

**AAV/Cas9-mediated HDR enables alteration of endogenous Kras**

To investigate the oncogenic function of diverse point mutations in vivo, we developed a platform for somatic AAV/Cas9-mediated HDR that incorporates DNA barcoding and high-throughput sequencing in autochthonous mouse models of several cancer types (Fig. 1). We designed, generated, and validated a library of AAV vectors to introduce all possible Kras codon 12 and 13 single-nucleotide non-synonymous point mutations into somatic mouse cells in a multiplexed manner. Each AAV vector pool for Cas9-mediated HDR into the endogenous Kras locus (AAV-Kras<sup>HDR</sup>/sgKras/Cre). Each vector contained an HDR template with 1 of 12 non-synonymous Kras mutations at codons 12 and 13 (or wild-type Kras), silent mutations within the PAM (boxed sequence) and sgRNA homology region (PAM*), and a random 8-nucleotide barcode within the wobble positions of the adjacent codons for stable DNA barcoding of individual tumors. Each vector also contained silent mutations within the sgRNA sequence and associated protospacer adjacent motif (PAM*) to prevent Cas9/sgKras-mediated cleavage of Kras<sup>HDR</sup> alleles (Fig. 2a and Supplementary Fig. 1b, d). To enable the parallel quantification of individual tumors by high-throughput sequencing of the Kras<sup>HDR</sup> allele in DNA from bulk tissue, we diversified the Kras<sup>HDR</sup> template with a random barcode engineered into eight wobble positions of the codons downstream of codons 12 and 13 (Fig. 2a and Supplementary Fig. 1b, c).

The AAV vectors also encoded Cre-recombinase, which enabled tumor initiation in mice containing a Cre-regulated Cas9 allele (H1<sup>LSL</sup>-Cas9<sup>19</sup>) (Fig. 2a). Cre-recombinase also led to labeling of transduced cells via a fluorescent Cre-reporter allele (Rosa26<sup>LSL</sup>-Tomato<sup>20</sup>) and enabled deletion of the tumor suppressor genes p53 or Lkb1 using floxed alleles<sup>21,22</sup>. We packaged the AAV-Kras<sup>HDR</sup>/sgKras/Cre library using an AAV8 capsid that enables high titer production, efficient transduction of mouse lung epithelial cells in vivo (Supplementary Fig. 2), and transduction of a wide range of adult mouse tissues<sup>23</sup>.

We initially transduced Cas9-expressing cells in culture with AAV-Kras<sup>HDR</sup>/sgKras/Cre to determine whether AAV/Cas9-mediated HDR would be an unbiased method to engineer point mutations into the endogenous Kras locus (Fig. 2d). We developed a PCR strategy to specifically amplify Kras<sup>HDR</sup> alleles, thus avoiding amplification of the endogenous Kras locus or any residual episomal AAV-Kras<sup>HDR</sup>/sgKras/Cre (Fig. 2e). High-throughput sequencing of the amplified Kras<sup>HDR</sup> regions from transduced cells confirmed the generation of all point mutant

---

**Fig. 2** Design and validation of an AAV targeting vector library to introduce all Kras codon 12 and 13 single-nucleotide non-synonymous point mutations into somatic mouse cells in a multiplexed manner. **a**, AAV vector pool for Cas9-mediated HDR into the endogenous Kras locus (AAV-Kras<sup>HDR</sup>/sgKras/Cre). Each vector contained an HDR template with 1 of 12 non-synonymous Kras mutations at codons 12 and 13 (or wild-type Kras), silent mutations within the PAM (boxed sequence) and sgRNA homology region (PAM*), and a random 8-nucleotide barcode within the wobble positions of the adjacent codons for stable DNA barcoding of individual tumors. **b**, Representation of each Kras<sup>HDR</sup> allele in the AAV-Kras<sup>HDR</sup>/sgKras/Cre plasmid library. **c**, Diversity of the barcode region in the AAV-Kras<sup>HDR</sup>/sgKras/Cre plasmid library. **d**, Schematic of the experiment to test for HDR bias. A Cas9-expressing cell line was transduced with AAV-Kras<sup>HDR</sup>/sgKras/Cre and Kras<sup>HDR</sup> alleles were sequenced to quantify HDR events. **f**, Schematic of the PCR strategy to specifically amplify Kras<sup>HDR</sup> alleles introduced into the genome via HDR. Forward primer 1 (F1) binds to the sequence containing the 3 PAM* mutations, while reverse primer 1 (R1) binds to the Kras locus, outside the sequence present in the homology arm of the Kras<sup>HDR</sup> template. F2 binds to the Illumina adapter added by F1, R2 binds to a region near exon 2, and R3 binds to the Illumina adapter added in the same reaction by R2. **f**, Frequency of HDR events for each Kras<sup>HDR</sup> allele plotted against the initial frequency of each mutant Kras allele in the AAV-Kras<sup>HDR</sup>/sgKras/Cre plasmid library. High correlation between the initial plasmid library and the representation of mutant Kras alleles following HDR suggests little to no HDR bias.
Kras alleles. The representation of Kras\(^{HDR}\) alleles in these cells was highly correlated with their representation in the plasmid library, suggesting that HDR using our AAV vector was not discernably biased by any single-nucleotide Kras codon 12 or 13 point mutation in the Kras\(^{HDR}\) template (Fig. 2f). Therefore, any differential expansion of tumors harboring specific Kras mutant alleles in vivo could be attributed to differences in the oncogenicity of Kras variants rather than HDR bias.

Somatic HDR initiates mutant Kras-driven lung tumors in vivo. To determine whether HDR in somatic cells could initiate tumors, and to investigate whether Kras variants differ in their ability to drive tumorigenesis, we delivered the AAV-Kras\(^{HDR}\)/sgKras/Cre library intratracheally to the lungs of mice with the H1\(^{LSL-Cas9}\) allele (Fig. 3a and Supplementary Fig. 3a). Three different genotypes of mice were transduced to provide additional insight into whether concurrent inactivation of tumor suppressor genes modulates Kras variant oncogenicity: (1) Rosa26\(^{LSL-Tomato}\), H1\(^{LSL-Cas9}\) (T;H1\(^{LSL-Cas9}\)), p53\(^{lox/lox}\), T;H1\(^{LSL-Cas9}\) (PT;H1\(^{LSL-Cas9}\)), and Lkb1\(^{lox/lox}\);T;H1\(^{LSL-Cas9}\) (LT;H1\(^{LSL-Cas9}\)).

Fig. 3 AAV/Cas9-mediated somatic HDR initiates oncogenic Kras-driven lung tumors. a Schematic of the experiment to introduce point mutations and a DNA barcode into the endogenous Kras locus of lung epithelial cells in Rosa26\(^{LSL-Tomato}\), H1\(^{LSL-Cas9}\) (T;H1\(^{LSL-Cas9}\)), p53\(^{lox/lox}\);T;H1\(^{LSL-Cas9}\) (PT;H1\(^{LSL-Cas9}\)), and Lkb1\(^{lox/lox}\);T;H1\(^{LSL-Cas9}\) (LT;H1\(^{LSL-Cas9}\)) mice by intratracheal administration of AAV-Kras\(^{HDR}\)/sgKras/Cre (8.4 \times 10^{10} vector genomes per mouse). b Representative fluorescence and histology images of Tomato\(^{positive}\) lung tumors in LT;H1\(^{LSL-Cas9}\), PT;H1\(^{LSL-Cas9}\), and T;H1\(^{LSL-Cas9}\) mice transduced with AAV-Kras\(^{HDR}\)/sgKras/Cre. Scale bar = 5 mm. c Quantification of lung tumors in the indicated genotypes of mice transduced with the indicated AAV vectors (with and without sgKras). Each dot represents one mouse. d Representative FACS plot showing Tomato\(^{positive}\) DTCs in the pleural cavity of an LT;H1\(^{LSL-Cas9}\) mouse with AAV-Kras\(^{HDR}\)/sgKras/Cre-initiated lung tumors. Plot shows forward scatter/side scatter (SSC)-gated viable cancer cells (DAPI/CD45/CD31/F4-80/Ter119\(^{negative}\)). e Histology of lymphatic metastases from AAV-Kras\(^{HDR}\)/sgKras/Cre-initiated lung tumors in PT;H1\(^{LSL-Cas9}\) mice. Scale bar = 50 μm. f Number of AAV-Kras\(^{HDR}\)/sgKras/Cre-transduced mice of each genotype that had disseminated tumors cells in their pleural cavity (DTCs > 10) or metastases out of the total number of mice analyzed. g Kras\(^{G12D}/G12D\);LT (KLT) and Kras\(^{G12D}/G12D\);PT (KPT) mice transduced with a 1:10,000 dilution of AAV-Kras\(^{HDR}\)/sgKras/Cre developed approximately half as many tumors as the PT;H1\(^{LSL-Cas9}\) and LT;H1\(^{LSL-Cas9}\) mice transduced with undiluted virus. If all Kras\(^{HDR}\) alleles in the AAV-Kras\(^{HDR}/sgKras/Cre\) library are oncogenic, this suggests that AAV/Cas9-mediated HDR occurs in ~0.02% of transduced cells. Alternatively, if only 20% of the mutant alleles in the AAV-Kras\(^{HDR}/sgKras/Cre\) library are oncogenic, this suggests that HDR occurs in ~0.1% of transduced cells. h Diverse HDR-generated oncogenic Kras alleles in individual lung tumors dissected from LT;H1\(^{LSL-Cas9}\) and PT;H1\(^{LSL-Cas9}\) mice transduced with AAV-Kras\(^{HDR}\)/sgKras/Cre. Number of tumors with each allele is indicated. Alleles that were not identified in any lung tumors are not shown. For the Kras\(^{HDR}\) alleles identified, Bonferroni corrected \(p\) values for likelihood of enrichment relative to WT are shown (Fisher’s exact test generalized for structural zeros; see Methods section).
**Fig. 4** Introduction of mutant Kras variants into pancreatic cells using AAV/Cas9-mediated HDR drives the formation of metastatic PDAC. a Schematic of retrograde pancreatic ductal injection of AAV-Kras HDR/sgKras/Cre (~1.7 x 10^{11} vector genomes per mouse) into PT;H11LSL-Cas9 mice to induce pancreatic cancer. b Histology of pancreatic tumors initiated by retrograde pancreatic ductal injection of AAV-Kras HDR/sgKras/Cre into PT;H11LSL-Cas9 mice. Scale bars = 75 μm. c Histology of metastases in the lymph node (upper panel) and diaphragm (lower panel) in PT;H11LSL-Cas9 mice with PDAC. Scale bars = 50 μm. d Incidence of PDAC, DTCs, or metastases in the indicated genotypes of mice (shown as the number of mice with cancer, DTCs, or metastases out of the total number of mice analyzed), 3–13 months after transduction with either AAV-Kras HDR/sgKras/Cre or AAV-Kras HDR/Cre (~2.9 x 10^{11} vector genomes per mouse). AAV-Kras HDR/sgKras/Cre was administered at the stock concentration (undil.) or at a 1:10 dilution (”1:10”). e HDR-generated oncogenic Kras alleles in individually dissected pancreatic tumor masses. Number of tumors with each allele is indicated. Alleles that were not identified in any pancreatic tumor masses are not shown.

LT;H11^{LSL-Cas9} mice were the first to show signs of tumor development including tachypnea and weight loss between 5 and 7 months after AAV administration. This is consistent with the rapid growth of lung tumors initiated in mice with a Cre-regulated Kras^{G12D} allele and floxed Lkb1 alleles. LT;H11LSL-Cas9 mice had many large primary lung tumors, resulting in high tumor burden (Fig. 3b, c and Supplementary Fig. 3b–d). Histological analysis confirmed the presence of large adenomas and adenocarcinomas in these mice (Fig. 3b and Supplementary Fig. 3b). PT;H11LSL-Cas9 mice also developed numerous large primary lung tumors 6–8 months after viral transduction. Compared to the LT;H11LSL-Cas9 mice, tumors initiated in PT;H11LSL-Cas9 mice had more pronounced nuclear atypia, a characteristic feature of p53 deficiency. Finally, T;H11LSL-Cas9 mice developed smaller, less histologically advanced lesions, even at later time points (>10 months after tumor initiation; Fig. 3b, c and Supplementary Fig. 3b–d). Mice transduced with a 10-fold lower titer of AAV-Kras HDR/sgKras/Cre developed proportionally fewer tumors (Supplementary Fig. 3e). Importantly, delivery of an analogous AAV vector without sgKras (AAV-Kras HDR/Cre) to T, PT, and LT mice did not lead to efficient tumor initiation (Fig. 3c and Supplementary Fig. 4). This indicates that neither p53 nor Lkb1 deficiency, combined with high-level AAV vector transduction, is sufficient to drive lung tumorigenesis. Several LT;H11LSL-Cas9 and PT;H11LSL-Cas9 mice with AAV-Kras HDR/sgKras/Cre-initiated tumors developed invasive primary lung tumors, had disseminated tumor cells (DTCs) in their pleural cavities, and had lymph node metastases (Fig. 3d–f). Thus, AAV-Kras HDR/sgKras/Cre-induced tumors can progress into malignant and metastatic lung cancer.

To approximate the efficiency of somatic HDR in lung epithelial cells in vivo, we compared the number of lung tumors initiated by AAV-Kras HDR/sgKras/Cre in mice containing the H11LSL-Cas9 allele (in which Kras would be activated by HDR) to the number of lung tumors initiated in Kras^{LSL-G12D} mice (in which Kras would be activated by Cre; Fig. 3g). We transduced Kras^{LSL-G12D};PT and Kras^{LSL-G12D};LT mice with a 1:10,000 dilution of AAV-Kras HDR/sgKras/Cre (Fig. 3g). These mice developed approximately half as many tumors as mice in which oncogenic Kras alleles were generated by AAV/Cas9-mediated somatic HDR after transduction of lung epithelial cells with undiluted AAV-Kras HDR/sgKras/Cre (Fig. 3g). This result is consistent with an HDR frequency between 0.02 and 0.1% (Fig. 3g). Importantly, this rate of HDR combined with the high-transduction efficiency of AAV enables the parallel initiation of numerous lung tumors within individual mice (Fig. 3c).

To determine whether tumors initiated using AAV-Kras HDR/sgKras/Cre harbored mutant Kras HDR alleles, we analyzed the Kras locus in cancer cells from large, individual lung tumors from LT;H11LSL-Cas9 and PT;H11LSL-Cas9 mice (Fig. 3c). PCR amplification using primers specific to the Kras HDR allele confirmed the presence of an oncogenic Kras allele with a unique barcode in each tumor (Figs. 2e, 3h). Interestingly, only five Kras variants were identified in ~50 large lung tumors, consistent with differential selection of Kras variants in lung tumorigenesis (Fig. 3h).

By analyzing individual tumors, we were able to carefully assess the Kras HDR allele as well as the second Kras allele in each tumor (Supplementary Fig. 5). Approximately half of the oncogenic Kras HDR alleles resulted from perfect HDR events in which a Kras...
point mutation and a unique barcode were seamlessly recombined into the endogenous \( \text{Kras} \) locus. The remaining \( \text{Kras}^{HDR} \) alleles were seamless from the 5’ end through mutant exon 2, but contained small duplications, insertions, or deletions in intron 2 (Supplementary Fig. 5). Importantly, none of these alterations would be expected to disrupt splicing from mutant exon 2 into exon 3. Additionally, almost all tumors harbored Cas9-induced indels in the second \( \text{Kras} \) allele, which is consistent with the loss of the wild-type \( \text{Kras} \) allele observed in some oncogenic Kras-driven human tumors (Supplementary Fig. 6)\(^2\)\(^8\)\(^9\). While previous studies have documented enhanced \( \text{Kras}^{G12D} \), and \( \text{Kras}^{G61D} \)-driven lung tumor growth following inactivation of the wild-type \( \text{Kras} \) allele in mice\(^1\)\(^0\)\(^3\)\(^1\), our results suggest that wild-type \( \text{Kras} \) may suppress the growth of lung tumors driven by many different oncogenic Kras variants.

**Somatic HDR can initiate pancreatic cancer.** In addition to driving human lung cancer, oncogenic Kras mutations are nearly ubiquitous in human pancreatic ductal adenocarcinoma (PDAC)\(^3\)\(^2\). Furthermore, genomic and clinical data from PDAC patients indicate that KRAS variants differentially impact downstream signaling as well as patient prognoses\(^3\)\(^2\)\(^3\)\(^3\). Expression of \( \text{Kras}^{G12D} \) or \( \text{Kras}^{G12V} \) in pancreatic cells of genetically engineered mice drives widespread formation of pancreatic intra-epithelial neoplasias (PanINs), but rarely leads to PDAC within 1 year\(^3\)\(^2\)\(^3\)\(^4\)\(^5\). However, \( \text{Kras}^{G12D} \) or \( \text{Kras}^{G12V} \) expression in mouse pancreatic cells combined with \( p53 \) mutation—a frequent event in human PDAC—reliably leads to the development of PDAC\(^3\)\(^5\).

Thus, to determine whether AAV/Cas9-mediated somatic HDR could also induce cancer-initiating oncogenic point mutations in pancreatic epithelial cells, we transduced \( PT;H11^{LSL-Cas9} \) mice with AAV-\( \text{Kras}^{HDR}/\text{sgKras}/\text{Cre} \) by retrograde pancreatic ductal injection (Fig. 4a and Supplementary Fig. 7a)\(^1\)^9. These mice developed PanINs as well as PDAC (Fig. 4b and Supplementary Fig. 7b, c). Several mice also developed metastatic PDAC, consistent with the aggressive nature of the human disease (Fig. 4c, d and Supplementary Fig. 7d, e). Sequencing of \( \text{Kras}^{HDR} \) alleles from several individual pancreatic tumor masses uncovered oncogenic \( \text{Kras}^{G12D} \) and \( \text{Kras}^{G12V} \) alleles with unique barcodes (Fig. 4e). Consistent with the requirement for oncogenic Kras to initiate PDAC, transduction of pancreatic cells in \( PT \) mice by retrograde pancreatic ductal injection of our negative control AAV-\( \text{Kras}^{HDR}/\text{Cre} \) vector did not induce any pancreatic tumors (Fig. 4d).

**HDR in somatic skeletal muscle cells induces sarcoma.** The history of the RAS oncoproteins began with their identification as the transforming factors in retroviruses capable of inducing rat sarcoma (hence, RAS)\(^3\)\(^6\). Human sarcomas are a highly heterogeneous cancer type comprised of over 50 histological subtypes that arise from many tissue and cell types\(^3\)\(^7\). Soft tissue sarcomas harbor diverse genomic alterations, including recurrent mutations in the tumor suppressor \( TP53 \) as well as the RAS pathway components \( \text{NF1}, \text{PTEN}, \text{PIK3CA} \)\(^3\)\(^8\)\(^3\)\(^9\). Although Kras is infrequently mutated in sarcomas, KRAS mutations occur in undifferentiated pleomorphic sarcomas as well as in rhabdomyosarcomas, which account for nearly half of pediatric soft tissue sarcomas\(^3\)\(^7\)\(^4\)\(^0\)\(^4\)\(^1\)\(^2\)\(^3\). Importantly, sarcomas have been induced in genetically engineered mouse models through the expression of \( \text{Kras}^{G12D} \) and inactivation of \( p53 \)\(^3\)\(^7\)\(^3\)\(^8\)\(^4\)\(^3\)\(^4\)\(^4\). However, generating mouse models of soft tissue sarcoma remains a challenge given the genetic diversity of the disease, and methods to rapidly model oncogene mutations in sarcomagenesis have not been described\(^3\)\(^7\)\(^4\)\(^4\).

To determine whether AAV/Cas9-mediated somatic HDR could be used to introduce point mutations into Kras and drive sarcoma formation, we performed intramuscular injections of AAV-\( \text{Kras}^{HDR}/\text{sgKras}/\text{Cre} \) into the gastrocnemii of \( PT;H11^{LSL-Cas9} \) mice (Fig. 5a and Supplementary Fig. 8a). These mice developed rapidly growing and invasive skeletal muscle sarcomas that harbored uniquely barcoded oncogenic \( \text{Kras}^{G12D} \), \( \text{Kras}^{G12A} \), or \( \text{Kras}^{G13R} \) alleles (Fig. 5b–e and Supplementary Fig. 8b–d). The successful application of this platform for
modeling tumorigenesis in diverse tissues highlights its broad applicability for multiplexed functional analyses of oncogenic driver mutations in a wide range of cancer types.

**Tumor barcode sequencing reveals Kras variant oncogenicity.** Whereas conventional methods to assess the effects of different genomic alterations in autochthonous cancer models largely rely on manual quantification of tumor number and size, we recently established a simple yet high-throughput approach to determine the genotype and number of neoplastic cells within individual tumors directly from bulk tissue by tumor barcoding and sequencing (Tuba-seq)26. Here, we extended the Tuba-seq platform to enable multiplexed and quantitative measurements of HDR-initiated tumors in vivo by engineering a random barcode into the wobble bases adjacent to an activating point mutation of interest. Since neoplastic cells within AAV-Kras<sup>HDR</sup>/sgKras/Cre-initiated tumors each contain a copy of the same unique DNA barcode, the relative number of neoplastic cells in each tumor can be determined by deep sequencing of the barcode region (Figs. 1d, 6a). Furthermore, by adding a normalization control consisting of a known number of cells with a known barcode to each bulk tissue sample prior to sequencing, the absolute number of neoplastic cells in each tumor can be estimated from relative sequencing read counts26. To determine the Kras genotype and number of neoplastic cells in each tumor in T;H1<sup>LSL-Cas9</sup>, PT;H1<sup>LSL-Cas9</sup>, and LT;H1<sup>LSL-Cas9</sup> mice, we added a normalization control consisting of a known number of cells with a known barcode to each bulk lung sample (Fig. 6a and Supplementary Fig. 9). We then extracted DNA from each sample, PCR-amplified the Kras<sup>HDR</sup> alleles, and deep-sequenced the variant-barcode region (Fig. 6a and Supplementary Fig. 9).
Following high-throughput sequencing, we corrected for recurrent sequencing errors and the possibility of individual tumors having identical barcodes (see Methods section). We then estimated the absolute number of neoplastic cells in each tumor by normalizing tumor barcode-sequencing read counts to the number of reads from the normalization control DNA. This analysis pipeline was exceptionally reproducible with a high degree of concordance in tumor sizes across technical replicates (Supplementary Fig. 10).

High-throughput sequencing of the \( \text{Kras}^{\text{HDR}} \) variant-barcode region uncovered many AAV-\( \text{Kras}^{\text{HDR}} \)/sgKras/Cre-initiated tumor masses in the pancreata of \( \text{PT;H11}^{\text{LSL-Cas9}} \) mice (Supplementary Fig. 11 and Supplementary Data 1). Normalizing tumor number to the initial representation of each \( \text{Kras}^{\text{HDR}} \) allele in the AAV-\( \text{Kras}^{\text{HDR}} \)/sgKras/Cre vector library allowed us to directly compare the in vivo oncogenicity of each \( \text{Kras} \) variant (Fig. 6b–e, Supplementary Fig. 11d, e, and Supplementary Data 2). Across >500 tumors, \( \text{Kras}^{\text{G12D}} \) was the most common variant, consistent with \( \text{KRAS}^{\text{G12D}} \) being the most frequent \( \text{Kras} \) mutation in human lung adenocarcinoma in never-smokers \( ^{45} \). \( \text{Kras}^{\text{G12V}}, \text{Kras}^{\text{G12C}}, \) and \( \text{Kras}^{\text{G12V}} \)—the next most frequent \( \text{Kras} \) variants in lung adenocarcinoma in never-smokers after \( \text{Kras}^{\text{G12D}} \)—were identified as moderate drivers of lung tumorigenesis, but were present in fewer tumors than \( \text{Kras}^{\text{G12D}} \) (Fig. 6b–e and Supplementary Fig. 11a–e). Interestingly, \( \text{Kras}^{\text{G12R}} \) and \( \text{Kras}^{\text{G13R}} \) were also identified as potent oncogenic variants, despite being infrequently mutated in human lung cancer (Fig. 6b–e and Supplementary Fig. 11a–e).

\( \text{Kras} \) expression can be influenced by overall changes in codon usage\(^{46,47} \). As our barcoding approach changed up to eight codons, we investigated whether alterations in \( \text{Kras} \) codon usage influenced tumor size. By calculating the relative change in codon usage introduced by each unique barcode, we found that the degree to which overall codon usage increased or decreased did...
not correlate with tumor size (Supplementary Fig. 12)\textsuperscript{48}. There was also no bias toward increased or decreased codon usage in the tumors identified (Supplementary Fig. 12).

**Kras oncogenicity is robust to tumor suppressor inactivation.** We initiated tumors in LT;H11\textsuperscript{LSL-Cas9}, PT;H11\textsuperscript{LSL-Cas9} and 7;H11\textsuperscript{LSL-Cas9} mice to directly assess whether concurrent tumor suppressor alterations modulate the ability of different Kras variants to initiate and drive tumor growth. In general, Lkb1-deficient tumors from LT;H11\textsuperscript{LSL-Cas9} mice grew much more rapidly than tumors in PT;H11\textsuperscript{LSL-Cas9} and 7;H11\textsuperscript{LSL-Cas9} mice (Supplementary Fig. 3a–d). Surprisingly, no statistically significant pairwise deviations in tumor number were observed for any Kras variant across tumors with different tumor suppressor genotypes. However, by considering the aggregate deviations in Kras variants across genotypes, we observed subtle differences in the spectrum of Kras variants in Lkb1-deficient tumors compared to p53-deficient or otherwise wild-type tumors (Supplementary Fig. 11f). These data are consistent with a model in which the strength of signaling induced by these oncogenic Kras variants in vivo is insufficient to engage the p53 pathway; thus, while p53 functions to constrain tumor progression, it does not appear to limit the initial expansion of tumors with different Kras genotypes\textsuperscript{49}. Additionally, while Lkb1 deficiency increased tumor growth, in general, Lkb1 deficiency did not appear to preferentially synergize with any specific oncogenic Kras variants (Fig. 6c–e and Supplementary Figs. 3d and 11a–e).

**Tuba-seq enables mapping of pancreatic tumors and metastases.** Since our tumor barcoding and sequencing platform (Tuba-seq\textsuperscript{26}) allowed us to characterize many individual tumors in parallel from bulk lungs, we anticipated that we could also use this approach to overcome the challenge of identifying and analyzing individual pancreatic tumor clones in multifocal tumors initiated in vivo\textsuperscript{26}. Since our tumor barcoding and sequencing platform enabled mapping of pancreatic tumors and metastases (Supplementary Fig. 2c) and Supplementary Figs. 3d and 11a, we assessed the functional consequences of specific Kras variants in vivo in a quantitative, scalable, and modular approach for systematic studies of the functional consequences of panels of oncogene mutations in parallel. Using this approach, we rapidly modeled tumors of multiple genotypes in several diverse tissues in vivo. Importantly, this method can be readily integrated with conventional genetically engineered mouse models including those with floxed and reporter alleles. Additionally, as genome engineering strategies continue to facilitate higher-content functional screens in autochthonous cancer models, the lack of correspondingly high-throughput methods for tumor analysis has emerged as a fundamental bottleneck in the functional annotation of cancer genomics data. To hurdle this barrier, we integrated a DNA barcoding strategy into our somatic HDR platform to enable the use of next-generation sequencing to determine the number, size, and genotype of tumors directly from bulk tissue in a quantitative and multiplexed manner (Fig. 1).

Using this HDR platform, we investigated the functional consequences of KRAS point mutations, which are among the most common driving events in human cancer\textsuperscript{25}. Our study addresses the fundamental question of whether tumorigenesis driven by specific KRAS mutations results in different pathological outcomes—indeed, we identify dramatic differences in the oncogenicity of Kras variants in vivo. The elucidation of these quantitative differences in oncogenic potential should facilitate the reclassification and therapeutic stratification of KRAS-driven tumors based on the functional consequences of specific KRAS variants, and contribute to the renewed effort to understand somatic cancer mutations at the level of the nucleotide rather than the gene.

Biochemical differences between KRAS variants as well as the frequency with which certain nucleotide substitutions in KRAS are generated likely both contribute to the non-uniform distribution of KRAS variants in human tumors. In vitro experiments have uncovered biochemical, structural, and cell context-specific differences between KRAS variants that could serve as biological substrates for tumorigenic selection. Notably, 98% of KRAS mutations in human tumors affect active site residues, thus limiting GTPase activity\textsuperscript{1,55}. However, despite measurable differences in both the intrinsic and GAP-mediated GTPase activity of KRAS variants, neither biochemical property correlates with in vivo oncogenicity (Supplementary Fig. 14)\textsuperscript{2}. Oncogenic proclivity was also not well-predicted by affinity for RAF, although variants with relatively low RAF affinities (Kras\textsuperscript{G12D}, Kras\textsuperscript{G12V}, and Kras\textsuperscript{G12R}) tended to be more oncogenic in our models of lung and pancreatic cancer (Supplementary Fig. 14c, i)\textsuperscript{2}. Collectively, these data suggest that the in vivo oncogenicity of KRAS variants may be better described either by an alternate biochemical property, or more likely, through the integration of multiple biochemical inputs.
Though the biochemical mechanisms remain unclear, our data indicate that the primary determinant of mutant KRAS representation in human PDAC is likely of a biological rather than etiological origin; we identified KrasG12D, KrasG12V, and KrasG12R as the most potent drivers of PDAC in vivo, which together account for >90% of all KRAS mutations in human PDAC (Fig. 7b, Supplementary Fig. 15, and Supplementary Data 4). This conclusion is supported by autochthonous zebrafish models of Kras-induced pancreatic tumorigenesis, and suggests that the cells of origin of PDAC may not be subjected to the same levels of mutagenic bias that afflict tissues such as the lung whose mutational burden can be more directly linked to specific external factors59–61.

The forces driving the spectrum of KRAS mutations in human lung cancer appear to be less centralized than in cancers of the pancreas. Differences between the oncogenicity of specific Kras variants in our mouse model compared to their prevalence in human lung cancers indicate that factors other than the intrinsic mutational prevalence in human cancer with careful consideration of the context within which a tumor forms. A simple explanation could be species-specific effects. Although mouse and human KRAS orthologs share 97% (183/188) amino acid identity, the amino acid mismatches are concentrated in the C-terminal hypervariable region (HVR) of KRAS (Supplementary Fig. 17). The RAS HVR regulates the subcellular trafficking and localization, post-translational modification, and effector binding of RAS proteins1,62. Additionally, species-specific regulatory mechanisms could differentially modulate the expression of endogenous KRAS as well as its effectors63,64,65. This is a particularly important consideration since the functional consequences of oncogenic RAS proteins are highly dependent on their expression; too little RAS signaling is non-transforming, while chronic, moderate levels of RAS are highly transforming, and hyperactive RAS signaling can lead to cellular senescence or growth arrest34,49,66,67. This is suggestive of the “Goldilocks Model of Oncogenicity” in which KRAS variants, such as KrasG12R or KrasG13R, conduct either too little or too much RAS signaling in human lung epithelial cells to effectively drive sustained tumorigenesis. Interestingly, while this model may have predicted that p53 deficiency would enable some otherwise “too strong” Kras variants to drive tumor formation, this was in fact not the case in our study. Finally, the identification of increasing numbers of cancer susceptibility loci using both in vivo and in silico approaches indicate that epistatic interactions between KRAS and genetic modifiers represent another intriguing contributor to KRAS mutant oncogenicity61,66.

The high prevalence of KrasG12C in human lung cancer is largely driven by the mutational processes that lead to lung cancer initiation. Indeed, numerous studies have identified mutational signatures in the lung cancer genome that represent extant records of the tumor’s etiological origin56,58,59. By examining the biases associated with tumor-extrinsic mutational signatures59—including the smoking-related signature and the molecular clock-like signature—we estimate that KRASG12C is approximately five times more likely than KRASG12D to occur in the cells of origin of lung adenocarcinomas (Supplementary Fig. 16a and Supplementary Data 5). Furthermore, whereas KRASG12C is the most frequent KRAS mutation in lung adenocarcinomas in current/former-smokers, it is approximately three times less prevalent in never-smokers in which KrasG12D, KrasG12V, and KrasG12A were enriched in lung tumors in mice—as they are in humans—KrasG12C was less frequent than expected given the spectrum of KRAS mutations in the human disease, while KrasG12R and KrasG13R were more frequent (Fig. 6b, Supplementary Fig. 15, and Supplementary Datas 2 and 4).

The high prevalence of KrasG12C in human lung cancer is largely driven by the mutational processes that lead to lung cancer initiation. Indeed, numerous studies have identified mutational signatures in the lung cancer genome that represent extant records of the tumor’s etiological origin56,58,59. By examining the biases associated with tumor-extrinsic mutational signatures59—including the smoking-related signature and the molecular clock-like signature—we estimate that KRASG12C is approximately five times more likely than KRASG12D to occur in the cells of origin of lung adenocarcinomas (Supplementary Fig. 16a and Supplementary Data 5). Furthermore, whereas KRASG12C is the most frequent KRAS mutation in lung adenocarcinomas in current/former-smokers, it is approximately three times less prevalent in never-smokers in which KrasG12D, KrasG12V, and KrasG12A were enriched in lung tumors in mice—as they are in humans—KrasG12C was less frequent than expected given the spectrum of KRAS mutations in the human disease, while KrasG12R and KrasG13R were more frequent (Fig. 6b, Supplementary Fig. 15, and Supplementary Datas 2 and 4).

Neither mutational biases nor biochemical properties appear to fully explain why KrasG12R and KrasG13R were identified as potent drivers of lung tumorigenesis in mice but are rare in human lung adenocarcinomas (Fig. 6b, Supplementary Fig. 15, and Supplementary Datas 2 and 4). Existing structural, biochemical, in vitro, and in vivo data might compel one to believe that these two variants are indeed oncogenic. Arginine substitutions at KRAS residue 12 or 13 prevent formation of transition state complexes with GAPs—an interaction that stimulates the minimal intrinsic GTPase activity of wild-type KRAS by approximately five orders of magnitude.1,60 KrasG12R has a lower GTP-mediated GTPase activity than either KrasG12C or KrasG12D, KrasG12R expression also leads to cellular transformation in culture, and drives de novo tumor formation in the lung and pancreas in various autochthonous systems (Figs. 6b, 7b)57,61. Human cancer genomics data also provides evidence of the oncogenicity of G12R and G13R substitutions in RAS proteins: KRASG12R accounts for >10% of KRAS-driven PDACs (Supplementary Fig. 15c and Supplementary Data 4), while HRA5G13R is the fourth most frequent HRA5 mutation in human cancers and NRASG13R is common in colorectal cancers1,62. Taken together, these data warrant a shift in perspective from the conundrum of why these variants are so oncogenic in our model to the question of why they are not in people.

The answer may lie with one, or perhaps many, of the myriad variables that define the molecular, cellular, and environmental context within which a tumor forms. A simple explanation could be species-specific effects. Although mouse and human KRAS orthologs share 97% (183/188) amino acid identity, the amino acid mismatches are concentrated in the C-terminal hypervariable region (HVR) of KRAS (Supplementary Fig. 17). The RAS HVR regulates the subcellular trafficking and localization, post-translational modification, and effector binding of RAS proteins1,62. Additionally, species-specific regulatory mechanisms could differentially modulate the expression of endogenous KRAS as well as its effectors63,64,65. This is a particularly important consideration since the functional consequences of oncogenic RAS proteins are highly dependent on their expression; too little RAS signaling is non-transforming, while chronic, moderate levels of RAS are highly transforming, and hyperactive RAS signaling can lead to cellular senescence or growth arrest34,49,66,67. This is suggestive of the “Goldilocks Model of Oncogenicity” in which KRAS variants, such as KrasG12R or KrasG13R, conduct either too little or too much RAS signaling in human lung epithelial cells to effectively drive sustained tumorigenesis. Interestingly, while this model may have predicted that p53 deficiency would enable some otherwise “too strong” Kras variants to drive tumor formation, this was in fact not the case in our study. Finally, the identification of increasing numbers of cancer susceptibility loci using both in vivo and in silico approaches indicate that epistatic interactions between KRAS and genetic modifiers represent another intriguing contributor to KRAS mutant oncogenicity61,66.

The key insights of our study are twofold: (1) amino acid substitutions in Kras residues 12 and 13 have quantitatively different and cancer type-dependent abilities to drive tumorigenesis and (2) both the functional consequences of specific Kras mutations, as well as the mechanisms leading to their initiation, play major roles in determining their pervasiveness in human cancer. Multiplexed approaches that enable the genetic dissection of RAS function in vivo represent a valuable complement to ongoing computational, biochemical, and cell culture-based studies of mutant forms of RAS proteins6. Notably, despite mounting evidence that cancers harboring different KRAS mutations can have differential therapeutic responses, the singular genetic inclusion criterion of “confirmed KRAS mutation” is listed for the vast majority of clinical trials in patients with KRAS-driven tumors.

The development and application of novel strategies to more rapidly interrogate the functional consequences of defined cancer mutations will undoubtedly redefine our biological and clinical understanding of many cancers types. We envision that this CRISSPR-based somatic HDR platform will enable an unprecedented understanding of the function of common and rare mutations in oncogenes across diverse cancer types, and will dramatically accelerate both the discovery and pre-clinical validation of therapies for precisely defined genetic subtypes of cancer.

Methods

Design, generation, and screening of sgRNAs targeting Kras. To obtain an sgRNA targeting Kras to enhance HDR in somatic mouse cells, we identified all possible 20 bp sgRNAs (using the consensus Cas9 PAM: NGG) targeting Kras exon 2 and the flanking intronic sequences, and scored them for predicted on-target efficiency using an available sgRNA design/scoring algorithm67. We then empirically determined the cutting efficiency of three sgRNAs targeting Kras (sgKras#1: GCAGCGTTACCTCTATCGTA; sgKras#2: GTAAATCA-GAATCCTTGG; sgKras#3: GACTGAGTATAAACCTTGTTG) (Supplementary Fig. 1a). Briefly, Lenti-U6-sgRNA/Cre vectors were generated for each sgRNA targeting Kras as previously described4. Q5 site-directed mutagenesis (NEB)
was used to insert the sgRNAs into a parental lentiviral vector containing a U6 promoter to drive sgRNA transcription as well as a PGK promoter to drive Cre recombinase. The cutting efficiency of each sgKras was determined via transduction of LSL-YFP:Cas9 cells in culture with each Lenti-sgKras/Cre virus. We isolated YFP<sup>+</sup> cells 48 h post transduction by fluorescence-activated cell sorting (FACS), extracted DNA, PCR-amplified the targeted Kras locus (forward primer: GGCTGTTCCTCGCCGTCCTTTACAAGCGTACG; reverse primer: AAGGCTTCCCTCGTAACTTCCGAGG), and Sanger-sequenced the amplicons (sequencing primer: GCACGGATGGCTGATTTGCAGG). Sequencing traces were analyzed by TIDE (tracking of indels by decomposition) to determine percent indel induction<sup>68</sup>. Since two pools were generated, which generated ~3 × 10<sup>5</sup> bacterial colonies per library. Colonies were scraped into LB-Amp liquid media and expanded for 6 h at 37 °C to increase yield (lowercase) containing multiplexing tags (underlined N<sub>54</sub>) (forward primer: CACACCTAGGTGAGTGTTAAAATATTG; reverse primer: GTAGCTCACTAGTGGTCGCC). Amplicons were digested with AvrII/BsiWI sites were introduced into the TOPO- cloned 2-kras fragment using Q<sup>5</sup> site-directed mutagenesis (NEB) (AvrII forward primer: TGAGTGTAAAATATTAAGTATTGTT; reverse primer: CCTGATGTTGTAACACTGAAATATTTCC), BsiWI forward primer: CGGCGGAGCAAGCAGCTGTAATCAGCTGTA; BsiWI reverse primer: GCAGGACGAGTGAATACACACATAGTCAGA; restriction sites are underlined with mismatching bases in lowercase). The Kras fragment harboring AvrII/BsiWI sites was released from TOPO with Ncol/Xhol and ligated into Ncol/Xhol-digested avv-sgKras/Cre to produce the final AAV-Kras HDR<sup>+</sup> vector.

To generate the control AAV-Kras HDR<sup>+</sup> vector without the sgRNA targeting Kras, PGK-Cre was excised from a TOPO clone with NotI/Xhol and ligated into the NotI/Xhol-digested modified AAV2 plasmid backbone. A BGH polyA sequence and the mouse Kras homology arm were added as described above to produce the control AAV-Kras HDR<sup>+</sup> backbone vector.

To introduce a library activating single point mutations and a DNA barcode into the Kras HDR<sup>+</sup> sequence of the AAV backbones, we synthesized four separate 295 bp Kras fragments with a degenerate “N” base (A, T, C, or G) at each of the first two basepairs of Kras codons 12 or 13 (Integrated DNA Technologies) (Supplementary Fig. 1b). By design, each of the four fragment pools consisted of three mutually non-identical nucleotides at codons 12 or 13 as well as the wild-type Kras sequence to serve as a control. Thus, since each of the four pools contained wild-type fragments, the overall representation of wild-type Kras alleles was expected to be approximately four times higher than each of the mutant Kras alleles. The synthesized fragments also contained silent (synonymous) mutations within the Kras target sequence and the associated PAM* to render the HDR cassette from pLL3.3;U6-sgKrat HDR<sup>+</sup> backbone vector.

A unique barcode AAV-Kras HDR<sup>+</sup> library was generated by targeting the sequence of the AAV backbones, we synthesized four separate pool consisted of three non-synonymous, single nucleotide mutations at codons 12 and 13 as well as random barcode associated with each library (Fig. 1c). The universal bases between Kras codon 12 and the barcode region were used to filter proper MSeq reads. The specific Kras codon 12 or 13 sequence and random barcode associated with each read was then cataloged to determine the representation of the Kras variants and the distribution of barcode nucleotides within each AAV plasmid library.

Availability of AAV-Kras HDR<sup>+</sup>/sgKras/Cre vectors. A uniquely barcoded AAV-Kras HDR<sup>+</sup>/sgKras/Cre plasmid vector for each of the 13 Kras HDR<sup>+</sup> alleles is available from Addgene (#99848-99860).
Analysis of individual lung tumors. Lung tumor-bearing mice displaying symptoms of tumor development were analyzed 4–10 months after viral administration. Lung tumor burden was assessed by lung weight and quantification of macroscopic Tomato<sup>positive</sup> tumors under a fluorescence dissecting scope as indicated (Fig. 3b, c and Supplementary Fig. 3b–e) (a single LTJH1<sup>LSL-Cre</sup> mouse had minimal Tomato<sup>positive</sup> signal that was restricted to a small region of one lung lobe, indicative of improper intratracheal administration of AAV, and was therefore removed from the analyses). The largest individual lung tumors that were visibly multifocal were dissected from bulk lungs under a fluorescence dissecting microscope for sequencing. For the majority of lungs, Tomato<sup>positive</sup> (and DAPI/CD45/CD3/F4/80/Ter119<sup>positive</sup>) cancer cells were purified using FACS instruments (BD Biosciences) within the Stanford FACS facility. Several individually dissected tumor samples were not purified by FACS. Lung lobes from mice of each genotype were also collected for histological analyses.

Analysis of individual pancreatic tumor masses. Pancreatic tumor-bearing mice displayed symptoms of tumor development and were analyzed 3–4 months after viral administration. Since pancreatic tumors generally appeared to be multifocal (Supplementary Fig. 7b), individual regions of the pancreas containing Tomato<sup>positive</sup> tumor masses were dissected and purified for sequencing (a mouse treated with a 1:10 dilution of AAV-Kras/Cre library also developed pancreatic tumor masses and therefore was included in these analyses). Regions of several pancreas were kept for histological analyses (Fig. 4b, c and Supplementary Fig. 7c).

Analysis of individual sarcomas. Sarcoma-bearing mice with obvious tumor development were analyzed 3–7 months after viral administration (Fig. 5 and Supplementary Fig. 8). A region of each sarcoma was sequenced and an adjacent region was saved for histological analysis (Fig. 5b, c and Supplementary Fig. 8c).

Characterization of Kras alleles in individual tumors. Genomic DNA for sequencing was extracted from FACS-purified tumor cells and unsorted tumor samples with a DNeasy Blood and Tissue Extraction kit (Qiagen). To identify Kras point mutations and tumor barcodes, we PCR-amplified and sequenced the Kras<sup>HDR</sup> alleles using two protocols optimized for several variables including annealing temperature, extension time, and primer sequences: (1) forward primer: TCTCTCTTTCTGTTAGTAAAGTTTTTG; reverse primer: CTCATCCACAAAGTGATTCTGA; sequencing primer: GATCTACAC; annealing temperature: 66 °C, extension time: 2 min; (2) forward primer: GCTGTTAGTAAAGTTTTTG; reverse primer: GGTTAGTAAAGTTTTTG; annealing temperature: 64 °C, extension time: 2–3 min. These protocols were used to specifically amplify integrated Kras<sup>HDR</sup> alleles from individual tumors as each incorporated a forward primer overlapping the engineered mutations in the PAM region upstream of codons 12 and 13, and a reverse primer outside the homology arm. Long extension times (2–3 min) were used to enable amplification of all Kras<sup>HDR</sup> alleles, even those containing insertions or duplications in intron 2 of the Kras locus (Supplementary Fig. 5d).

A variety of HindIII-digested oncogenic Kras alleles were identified in 49 individual lung tumors (Fig. 3h), 4 individual pancreatic tumor masses (Fig. 4e), and 5 sarcomas (Fig. 5e). For the Kras<sup>HDR</sup> alleles identified, p values for likelihood of emitting the first round PCR and generic sequencing primers described above) to detect the presence of a Kras<sup>HDR</sup>-containing allele (GTG) with a unique barcode (GGGAAATGCGCCCGTCAGATG) was identified. Genomic DNA from this cell line was used as a normalization control for high-throughput sequencing for all bulk lung samples (Supplementary Fig. 9b).

Bulk lung tissue processing and DNA extraction. Bulk lung samples were dissected from virally-transduced mice and stored at −80 °C prior to processing. To extract genomic DNA for sequencing, samples were thawed and transferred to 50 ml conical tubes. 20 ml of lysis buffer (100 mM NaCl, 20 mM Tris pH 7.6, 10 mM EDTA pH 8.0, 0.5% SDS in H<sub>2</sub>O) plus 200 μl proteinase K (20 mg/ml) were added to each sample. Next, 3 μg (~5 × 10<sup>5</sup> genomes) of normalization control DNA was added to each sample (Fig. 6a and Supplementary Fig. 9b). Samples were then carefully homogenized using a tissue blender, which was cleaned between each generation using clean 10% bleach, 70% ethanol, and 1x PES. Homogenized samples were lysed at 55 °C overnight. Genomic DNA was isolated from tissue lysates by phenol/chloroform extraction followed by ethanol precipitation.

Bulk pancreatic tissue processing and DNA extraction. Pancreatic tumor masses were dissected and digested, and viable neoplastic cells (Tomato<sup>positive</sup> and Cre<sup>positive</sup>) were purified using FACS. No normalization control was added to the pancreatic cancer samples. Genomic DNA was isolated from FACS-isolated neoplastic cells using a DNeasy Blood and Tissue Extraction kit (Qiagen), and then further purified by ethanol precipitation.

Illumina® library preparation and Tuba-seq. To uncover the number and size of tumors harboring each Kras variant in a massively parallel and quantitative manner, we developed a two-round PCR strategy that enabled multiplexed Illumina® sequencing of barcoded template most often occurred within the larger, more distal homology arms, it also occurred at a detectable frequency within very short regions of homology flanked by 5' and 3' mismatches (including the PAM* mutations, a Kras codon 12 or 13 mutation, and up to 11 mismatches within the barcode region).

We initially identified the presence of duplications in Kras<sup>HDR</sup> alleles in some tumors (Supplementary Fig. 5d), we designed PCR primers to specifically amplify duplications of Kras exon 2 that occurred on either side of exon 2 within Kras<sup>HDR</sup> alleles (right-hand duplication—forward primer: TGAACCTAC GATAGAGGTAACG; reverse primer: GTCTCACAGTGCTACAGATG; left-hand duplication—forward primer: TGAGTTGAAATTGATGAAATTTTTGTG; reverse primer: TGGCAGTTAGCTGCTACAGATG; sequencing primer: TGGAGTTGAAA ATATTGAAATTTTTGTG). Each of these duplication-specific PCR protocols used adjacent primer pairs in opposite orientations, ensuring that amplification would only occur if a duplication were present. Duplications of varying lengths were identified, including duplications of the second half of wild-type exon 2 or the entire exon 2 (but lacking critical regions of the splice acceptor) (Supplementary Fig. 5d). Deletions and duplications of regions of intron 2 were also observed. Furthermore, we observed integrations of parts of the AAV transfer vector, including the U6 promoter and viral ITR, into intron 2. Since the size and location of these alterations, none would be expected to change splicing of Kras mutant exon 2 to exon 3, consistent with the requirement of oncogenic Kras to drive tumorigenesis.

Generating a normalization control for Tuba-seq. To obtain a normalization control for tumor barcode sequencing, a single tumor tissue was dissected from an P721L<sup>LSL-Cre</sup> mouse, digested into a single cell suspension, and plated to generate a cell line. After expanding these cells and then extracting genomic DNA, Kras exon 2 was PCR-amplified (forward primer: TCCCTCCTTGGGTGCCGTGTTG; reverse primer: GGGCTGGTGCCGTCCCTATTC). The PCR product was sequenced (using specific and generic sequencing primers described above) to confirm the presence of a Kras<sup>G12V</sup>-containing allele (GTG) with a unique barcode (GGGAAATGCGCCCGTCAGATG) that was identified. Genomic DNA from this cell line was used as a normalization control for high-throughput sequencing for all bulk lung samples (Supplementary Fig. 9b).

Illumina® library preparation and Tuba-seq. To uncover the number and size of tumors harboring each Kras variant in a massively parallel and quantitative manner, we developed a two-round PCR strategy that enabled multiplexed Illumina® sequencing of barcoded template most often occurred within the larger, more distal homology arms, it also occurred at a detectable frequency within very short regions of homology flanked by 5’ and 3’ mismatches (including the PAM* mutations, a Kras codon 12 or 13 mutation, and up to 11 mismatches within the barcode region).

We initially identified the presence of duplications in Kras<sup>HDR</sup> alleles in some tumors (Supplementary Fig. 5d), we designed PCR primers to specifically amplify duplications of Kras exon 2 that occurred on either side of exon 2 within Kras<sup>HDR</sup> alleles (right-hand duplication—forward primer: TGAACCTAC GATAGAGGTAACG; reverse primer: GTCTCACAGTGCTACAGATG; left-hand duplication—forward primer: TGAGTTGAAATTGATGAAATTTTTGTG; reverse primer: TGGCAGTTAGCTGCTACAGATG; sequencing primer: TGGAGTTGAAA ATATTGAAATTTTTGTG). Each of these duplication-specific PCR protocols used adjacent primer pairs in opposite orientations, ensuring that amplification would only occur if a duplication were present. Duplications of varying lengths were identified, including duplications of the second half of wild-type exon 2 or the entire exon 2 (but lacking critical regions of the splice acceptor) (Supplementary Fig. 5d). Deletions and duplications of regions of intron 2 were also observed. Furthermore, we observed integrations of parts of the AAV transfer vector, including the U6 promoter and viral ITR, into intron 2. Since the size and location of these alterations, none would be expected to change splicing of Kras mutant exon 2 to exon 3, consistent with the requirement of oncogenic Kras to drive tumorigenesis.
Importantly, since the characterization of KrasHDR alleles in individual tumors uncovered some variability in HDR (e.g., diverse indels in Kras intron 2; Supplementary Fig. 5), only four cycles (for lung samples, six cycles for pancreas samples) were performed in the second round of PCR to minimize the potential for bias during the amplification of products of variable length. Furthermore, a high-efficiency polymerase (Q5® hot start high-fidelity polymerase, NEB; 64 °C annealing temperature) and a long extension time (3:00 min) were used to ensure robust amplification of all KrasHDR alleles. KrasHDR alleles in genomic lung DNA were amplified using between 4 and 40 separate 100-μL PCR reactions and then pooled following amplification to reduce the effects of PCR jackpotting (Supplementary Fig. 9a). Each of these 100-μL PCR reactions contained 4 μg of DNA template to amplify from a large initial pool of KrasHDR alleles. Following the first round of amplification, all replicate PCR reactions were pooled and 100 μL of each sample was cleaned up using a QIAquick PCR purification Kit (Qiagen). Purified first round PCR amplicons were used as template DNA for a 100 μL second round Illumina library PCR (Q5® hot start high-fidelity polymerase, NEB; 72 °C annealing temperature; 35 cycles for lung samples, 40 cycles for pancreas samples). The second round of PCR amplified a 112 bp region entirely within the Kras exon 2 region present in first round PCR amplicons. The second round reverse primer contained a P7 adapter (italized), a reverse complemented 8 bp custom i7 index (“Nx”), and a reverse complemented Illumina sequencing primer sequence (read 2) (underlined) at the 5’ end to enable dual-indexed, paired-end sequencing of Illumina libraries (second round reverse primer #1: CGAGATCCGACCGCGAGATCGGTGACTGGACTTCAGACGTGGTGCCTCTGGCATCGTGGCTATACTCATCCACA). The second round Illumina library PCR was complementary to the P5 Illumina® adapter added to the amplified KrasHDR allele by the forward primer during the first round PCR (second round forward primer: AATGATACGGGACCGCGAGATCGGTGACTGGACTTCAGACGTGGTGCCTCTGGCATCGTGGCTATACTCATCCACA). This primer was used to amplify second round PCR reactions without amplification of contaminating genomic DNA that may have been carried over from the first round PCR reaction. Furthermore, a second reverse primer encoding the P7 adapter sequence was added to the second round PCR reaction at the same concentration as the two other primers (second round reverse primer #2: CAAGATCCGACCGCGAGATCGGTGACTGGACTTCAGACGTGGTGCCTCTGGCATCGTGGCTATACTCATCCACA). This primer binds the reverse complemented P7 adapter sequence added to the KrasHDR amplicons by second round reverse primer #1. Since the second round PCR was performed over 35–40 cycles, the P7 adapter (second round reverse primer #2) was added to limit the amount of non-specific amplification produced by the lengthy second round reverse primer #1.

After the second round of amplification, 100-μL PCR reactions were run on a 2.5% agarose gel and a band of the expected size was excised. PCR amplicons were then purified on a Bioanalyzer® high-sensitivity DNA chip, and a pool containing the expected fragment size was prepared. The size, concentration, and purity (i.e., the lack of unexpectedly sized peaks indicative of nucleic acid or buffer contaminants) of the purified Illumina® libraries was confirmed using a Bioanalyzer (Agilent). Individual Illumina® libraries with unique dual-indices were then pooled together such that libraries originally derived from mice with greater tumor burden were represented at a higher ratio in the final pool than those from mice with lower tumor burden (Supplementary Fig. 9a). A total of 35 independent samples were combined into two libraries without amplification, one to minimize the potential for bias during the amplification of products of variable length; the other to maximize the amount of non-specific amplification produced by the lengthy second round reverse primer #1.

The size of the final Illumina® library pool was then deep-sequenced on an Illumina® HiSeq lane using a multiplexed, 150 bp paired-end Rapid Run sequencing program (Eliin Biopharmaceuticals).

**Estimating the size and number of barcoded tumors.** We developed a publically available pipeline to call tumors from our de-multiplexed Illumina® sequencing data, termed Tuba-seq (https://github.com/petrov-lab/tuba-seq). The pipeline tunnels unique barcode sequences and eliminates recurrent sequencing errors using an algorithm designed to detect patterns of amplifications indicative of bias. We tailored this algorithm to minimize the occurrence of spurious tumor calls, and minimize technical biases (including variation in read depth, variation in Illumina® sequencing machine error rates, and variation in barcode diversity). This pipeline was modified here for the analysis of tumor genotypes and barcodes following AAV/Cas9-mediated somatic KrasHDR-driven tumor models described above. Although our Illumina® sequencing libraries contained a small 112 bp fragment of the KrasHDR alleles, we performed 150 bp paired-end sequencing of these fragments and merged the overlapping forward and reverse reads to reduce the likelihood of Illumina® sequencing errors in Kras codons 12 and 13 and the barcode region of the KrasHDR alleles. Overlapping paired-end reads were merged, quality-filtered, and trimmed using PANDAseq (fragment length: 60 bp; forward trimming primer: ATAGCTGAGTATAACACT; reverse trimming primer: CTCACTCCAGAAGTAGA)^5^.

Even after merging forward and reverse reads to reduce sequencing errors, an average of ~1–2 errors per 1,000 bases was detected, presumably from recurrent Illumina® sequencing errors or from errors introduced during PCR. Given this error rate, we expected that reads from a large, uniquely barcoded tumor containing single nucleotide mismatches would be called as small, spurious tumors of ~1/10,000th the size of the large real tumor. Even without Tuba-seq, this phenomenon was discernible by eye, as we observed small clusters of spurious “tumors” that were ~3–4 orders of magnitude smaller and contained 1 nucleotide deviation relative to the largest tumor in specific mice. Additionally, each KrasHDR variant-barcode pair also possessed recurrent sequencing errors in mutant base in other Kras codons 12 or 13.

To accurately tally tumors, we developed a computational and statistical pipeline for the analysis of tumor barcode sequencing data. We first estimated the residual rate of sequencing/PCR errors from the seven nucleotides upstream of Kras codon 12 and the seven nucleotides downstream of the final barcode base. We then used our model of sequencing errors to cluster unique read pileups (truncated to within seven nucleotides of the barcoded bases) into unique tumors via DADA2. A minimum confidence in unique origin of the clusters of 0.01 (i.e., omega_a = 0.01) was used. A larger threshold increased the number of unique tumors called in a mouse sample. We chose this larger value rather than those used previously5, as our experiment sequencing gave us greater confidence that unique read pileups were truly distinct tumors. For example, we found that this threshold eliminated all unintended read sequences (e.g., reads with inappropriate nucleotides outside of the barcode), and that this threshold called a total number of lesions within each mouse more consistently between biological replicates. These were important considerations since without proper handling of read errors, the number of called tumors can positively correlate with sequencing read depth. Finally, we removed any tumors with DNA sequences that deviated by only 1 nucleotide from a lesion that was 10,000× larger. This affected only 1.5% of tumor calls.

After generating the read pileups and performing the corrections described above, we normalized the number of reads from each tumor to the number of reads from the normalization control that was spliced into each sample prior to DNA extraction from bulk tumor-bearing tissue lysates (Fig. 6a and Supplementary Fig. 9b). This allowed us to generate a reasonable estimate of the number of neoplastic cells in each tumor and allowed us to merge data from mice of the same genetic and treatment. However, this step is severe, we were able to accurately quantify the absolute number of neoplastic cells within each tumor.

First, some KrasHDR alleles in individual tumors harbored insertions or deletions in Kras intron 2, inside the first round PCR primers for Illumina® sequencing. Although the presence of different sized amplicons could generate a PCR bias, we attempted to reduce this by filtering only to four to six cycles in the first round of Illumina® library PCR, using a long extension time (~3 min), and using a fast (20–30 s/kb), high fidelity polymerase (Q5®; NEB). As the final Illumina® library PCR product in the second round of amplification is short and uniform across all samples, PCR amplification should not be biased at this step. Second, given that the Kras HDR region, found in the endogenous Kras locus, is possible that in some tumors this region is genetically amplified (which has been documented in KrasG12D-driven lung tumors initiated in mouse models of lung cancer5,76). Although Kras amplifications do not typically result in very high Kras copy numbers, any amplification of this region would lead to an overestimation of the number of neoplastic cells in tumors with amplified KrasHDR alleles since our conversion from read count to cell number assumes that each cell contains a single copy of the barcoded KrasHDR allele.

Lastly, the normalization control itself was generated from cells from a tumor with a known duplication in Kras intron 2, which produces a larger PCR product in the second round of the Illumina® library preparation than tumors without duplication. Thus, any PCR bias away from the KrasHDR allele in the normalization control would result in a systematic overestimation of the size of tumors.

Sequencing of tumor barcodes from 35 samples on two lanes of an Illumina® HiSeq Rapid Run, combined with our analysis pipeline, enabled the detection of unique barcodes with read counts covering five orders of magnitude. Thus, the resolution of this approach enables detection of both large lesions and small hyperplasias within bulk tissue. However, the ability to detect a large numbers of lesions within bulk tissue increases the probability of barcode collisions: the occurrence of two or more lesions with the same DNA barcode in the same mouse. Large numbers of collisions can overestimate the size of individual tumors because two small “colliding” tumors would be identified as a single, larger tumor. Therefore, we developed a statistical model of barcode collisions to reduce this likelihood and not overly bias the estimated sizes of called tumors.

Our model of barcode collisions accounts for the likelihood p_i of observing each of the 24,573 possible barcodes i for each Kras HDR variant in our study. A majority of the reproducible variation between barcode frequencies in our pool derives from statistically independent variation in the nucleotide frequencies at each wobble base (i.e., each barcode is not equally likely in the pool because there was subtle variation in nucleotide concentrations during synthesis of the barcodes fragments) (Fig. 2c). To estimate the independent frequency f_i of each barcode at every base b in the barcode and use this data to predict barcode likelihoods based on each barcode’s sequence B(b,w) (where B is 1 if barcode / possesses nucleotide n at position b and 0 otherwise) as follows:

\[ p_i = \prod B(b,w) f_i \]

Here, superscript indices represent contravariant vectors and the dot represents matrix multiplication. This model predicts barcode frequencies with only 21
free parameters. Because some residual over-representation of barcodes persisted in the lung samples, we considered the 10% most frequently observed barcodes, after correcting for nucleotide frequencies, from all lung analyses. These most frequently observed barcodes were identified independent of our mouse experiments by Illumina® sequencing (MiSeq) of our pAAV-KrasHDR/sgKras/Cre plasmid pool prior to viral production. After this processing, we then renormalized the number of barcodes in each sample (on average 4.04%). We then assessed the occurrence of barcode collisions within each mouse using a multinomial sampling process. The mean number of collisions $C_i$ for each observed barcode within each mouse is then:

$$C_i(p, N) = \sum_{k=1}^{N} (k-1)(p_i, p_{-i}) N$$

$$= \mu_i - p_k (k - p_i, p_{-i}) N - 1$$

$$= Np_i + (1 - p_i)^N - 1$$

Here, $p_i$ denotes the mean number of barcodes within each mouse, while $N$ denotes the total number of tumors (both unknowns). $N$ is then determined from the observed number of tumors in each mouse $N^{obs}$ using the equation $N^{obs} = N - \sum C_i(p, N)$, which is solved using Brent's method.

This model found that barcode collisions were generally rare in our mouse samples (on average 4.04%). However, the likelihood of collisions can vary by mouse and by Kras variant. For example, the average predicted number of collisions for WT KrasHDR alleles was 5.8% and as high as 12% in one mouse. WT KrasHDR alleles were expected to experience the highest number of collisions since WT Kras vectors were intentionally represented approximately fourfold more than each mouse Kras vector in the initial pAAV-KrasHDR/sgKras/Cre plasmid pool (Fig. 2b). Thus, we divided the size of each lesion by 1 + $c_i$ to minimize the bias that barcode collisions impart on tumor size distributions. Because collisions are rare events, the particular number of collisions within each mouse can differ substantially from $C_i$. Because of this limitation, we believe that this correction minimizes systematic bias to tumor size distributions resulting from barcode collisions that would alter our final estimates of the mean effect of each Kras variant; however, it cannot effectively identify the specific collisions that occurred among specific barcodes in specific mice.

To determine whether Kras variants had quantitatively different abilities to drive tumorigenesis, we elected to focus on lung tumors estimated to contain $> 100,000$ neoplastic cells (i.e., one-fifth the “size” of the normalization control DNA added to each sample, which was derived from ~5 × 10^5 cells). Regression analysis of lung tumors above this cell number from replicate samples (independent sample preparation, sequencing, and processing) demonstrated high correlations (all $R^2$ values were reported above 0.99; Supplementary Fig. 10). We believed that the estimated number of cells in tumors below this cutoff was less reliable; KrasHDR variant-barcode pairs present at low read counts were much more likely to be biased by barcode collisions and variability in both PCR amplification and DNA sequencing. Additionally, we could not rule out the possibility that these variant-barcode pairs were a product of off-target amplification of our AAV-KrasHDR/Cre control vector, which contains a longer KrasHDR template that partially overlaps the reverse primer used for Illumina® library preparation of bulk samples with AAV-KrasHDR/sgKras/Cre-initiated tumors (as opposed to small but real HDR-initiated lung hyperplasias). We anticipate that future technical and computational improvements will further improve the robustness of this pipeline and enable tumor calling at much smaller cell number cutoffs.

Since Kras expression can be influenced by overall changes in codon usage that effect tumorigenesis^65,67, we also determined whether the use of different nucleotides in the barcode region of KrasHDR alleles had a detectable effect on lung tumor size. To do this, we calculated the relative change in codon usage generated by the unique barcode in each lung tumor. The barcode region of each KrasHDR allele contains three anchor bases that are different from wild-type Kras (but universal to all KrasHDR alleles), as well as random nucleotides in the eight positions that make up the barcode. A "codon usage score" for the KrasHDR allele in each lung tumor was calculated as the sum of the mouse codon frequencies from a publicly available mouse codon usage table^87 across the 22 bp barcode region. A codon usage score was also calculated for wild-type Kras and for a "minimal KrasHDR allele" containing the three universal anchor bases, but wild-type bases at all eight positions of the barcode (Supplementary Fig. 12).

Analysis of sequencing data from bulk tumor-bearing lungs. We quantified the relative numbers of tumors harboring each Kras variant by counting tumors above 100,000 cells in six PT/H11 LSL-Cas9, six LT/H11 LSL-Cas9, and three TH11 LSL-Cas9 mice (Supplementary Fig. 9a). Read counts from tumors with each Kras variant were normalized by dividing by the initial representation of each variant in the AAV-KrasHDR/sgKras/Cre plasmid pool (for this analysis, the initial representation of each variant in the plasmid pool was calculated from the total number of reads associated with each Kras variant after removing barcodes above the 98th percentile of barcode abundance; this restriction did not appreciably alter results, and was simply applied to ensure that extremely abundant variant-barcode pairs did not overly impact the overall representation of specific variants). Resulting values were then scaled such that WT Kras and KrasHDR variants had a representation of 1. A small number of WT KrasHDR alleles appeared to arise from tumors above 100,000 cells. These could represent tumors in which an HDR event created the non-oncogenic Kras WT genotype but which nonetheless evolved into a tumor for other reasons, or the WT Kras variant "hitchhiked" with an oncogenic Kras variant by co-incident HDR in the same lung cell followed by expansion driven by the oncogenic variant.

A small number of residual cells from individual tumors dissected from bulk tissue (and analyzed as described above) were usually detectable in our tumor sequencing data. In all analyses of tumor size, these dissected tumors were excluded as we could not infer their true size. However, when analyzing the number of tumors above 100,000 cells in each treated mouse genotype, we included data from individually dissected tumors since dissectible tumors were always among the largest tumors observed within any mouse and, therefore, certainly above the 100,000 cell threshold.

Statistically significant differences in tumor number were determined using Fisher’s exact test. For each variant, two tests were performed, comparing with either the frequency of G12D or WT KrasHDR alleles (Fig. 6b-e). All p values were Bonferroni-corrected for the number of variants investigated and were two-sided. A two-sided “many cells” Pearson’s $\chi^2$ test was used to compare the distribution of tumor numbers across all Kras variants in PT/H11 LSL-Cas9 and LT/H11 LSL-Cas9 mice relative to TH11 LSL-Cas9 mice (Supplementary Fig. 11f).

Analysis of sequencing data from bulk pancreas tumor masses. Following Illumina® sequencing and read processing as described above, we quantified the number of pancreatic tumors with each KrasHDR allele in four primary tumor masses and three metastasis samples from two PT/H11 LSL-Cas9 mice produced with AAV-KrasHDR/sgKras/Cre by retrograde pancreatic ductal injection (Supplementary Fig. 13a). Since we did not add normalization control DNA to these samples, we used a stringent internal read count cutoff of two times the most abundant WT Kras allele in each sample. Combined with the four tumors identified from individual pancreatic analyses, we detected 25 primary tumors containing KrasHDR alleles across three PT/H11 LSL-Cas9 mice. One KrasHDR allele with a specific variant-barcode pair was identified in an individually dissected tumor as well as in the bulk tumor mass from the same mouse and was therefore only counted once. Two KrasHDR alleles with specific variant-barcode pairs were identified since these were determined to be distinguishable across cell lines generated from individual pancreatic tumors with the same specific variant-barcode pairs. Tumor number counts for each KrasHDR allele were then normalized to the initial representation of each allele in the AAV-KrasHDR/sgKras/Cre library. Statistically significant differences in the number of tumors with each KrasHDR allele were determined by a Fisher’s exact test for likelihood of enrichment relative to WT (Fig. 7b). All p values were Bonferroni-corrected for the number of variants investigated and were two-sided. Spatial and phylogenetic relationships between primary tumors and metastases was established through shared KrasHDR variant-barcode pairs (Fig. 7c and Supplementary 13b).

Comparing the oncoenergicity and biochemistry of Kras variants. To identify any relationships between the oncogenicity of Kras mutants observed in vivo and differences in their biochemical properties, we compared our data to the biochemical data for several of the KRAS variants reported in Hunter et al2 (Supplementary Fig. 14h). Specifically, intrinsic and P120GAPl-stimulated GTP hydrolysis rates (Hunter et al2, Supplementary Data 1), as well as relative RAF kinase activities (Hunter et al2, Supplementary Data 2), were obtained for human KRASG12D, KRASG12V, KRASG12R, KRASG12C, and KRASG13D. Lung and pancreatic tumor numbers plotted against this biochemical data are shown in Supplementary Fig. 14.

Analysis of KRAS mutations in human cancer genomics data sets. The prevalence of KRAS mutations in human lung adenocarcinoma, PDAC, and rhabdomyosarcoma was quantified via alignment of a database of all human cancers available in the Catalog of somatic mutations in cancer (COSMIC; release v81) and the American Association for Cancer Research (AACR) Project Genomics Evidence Neoplasia Information Exchange (GENIE; version 1.0.1) databases, as well as from other independent studies, as outlined below (Supplementary Fig. 15 and Supplementary Data 6). A GISTIC mutation data from both targeted and genome-wide screening were utilized (http://www.cancer.sanger.ac.uk/cosmic). The data were obtained, in part, from cbioPortal^79.

To determine the prevalence of KRAS mutations in human lung adenocarcinoma, we compiled data from COSMIC and GENIE as well as several additional studies (Supplementary Data 48,79,80). All samples were screened for duplicates across databases by comparing sample identifier numbers. All remaining samples were pooled to determine both the overall frequencies of somatic mutations in KRAS and the spectrum of specific KRAS mutations observed in lung adenocarcinoma (Supplementary Fig. 15a and Supplementary Data 4).

We also compared the spectrum of KRAS mutations in lung adenocarcinoma in "never-smoker" and "current/former-smoker" patients from various studies (Supplementary Fig. 16b and Supplementary Data 6). Patients whose smoking history was unknown were not included in the analysis. "Never-smokers" included patients that reported smoking fewer than 100 cigarettes in their lifetime as well as the "current/former-smoker" patients described as "former" or "current" smokers in the included studies, as well as individuals described as "smokers."
To determine the prevalence of KRAS mutations in human PDAC, we combined data from COSMIC and GENIE, as well as The Cancer Genome Atlas (TCGA) provisional data set (NIH NCI Genomic Data Commons Release 7.0) and one additional study.56 We screened for duplicate samples by comparing sample identifier numbers. All remaining samples were pooled to estimate the frequency of somatic mutations in KRAS, as well as the spectrum of specific KRAS mutations in patients with PDAC (Supplementary Fig. 1c and Supplementary Data 4).

To examine KRAS mutations in patients with rhabdomyosarcoma, we similarly compiled data from COSMIC and GENIE, as well as from numerous independent studies (Supplementary Data 4). We screened for duplicate samples by comparing sample identifier numbers. Remaining samples were pooled to estimate overall KRAS mutation frequency and the specific spectrum of KRAS mutations observed in patients with rhabdomyosarcoma (Supplementary Fig. 1b and Supplementary Data 4).

Approximating relative KRAS mutation frequencies. We approximated the relative induction rate of mutations in KRAS codons 12 and 13 from data reported in Campbell et al.58 Campbell et al.58 document six mutational signatures in lung adenocarcinomas: a smoking-related signature (S14), a molecular clock-like signature (S15), an ultraviolet (UV)-related signature (S17), a mismatch-repair signature, and two APOBEC-related signatures (S13 and S12). Since we were interested in estimating the relative rate of induction of KRAS codons 12 and 13 mutations prior to tumorigenesis, we focused on the mutational processes that underlie the tumor-essence frequency—the smoking-related signature and the molecular clock-like signature. Campbell et al.58 (Supplementary Data 7) contains the estimated probabilities of every possible nucleotide substitution within each trinucleotide motif (i.e., the substitution plus the adjacent 5’ and 3’ bases) for each of the mutational signatures. We approximate the relative frequency of specific substitutions in KRAS codons 12 and 13, we summed the trinucleotide substitution probabilities associated with the smoking- and molecular clock-like signatures for each of the KRAS mutations. Normalizing these probabilities within the 12 single-nucleotide mutations in KRAS codon 12 or 13 allowed us to approximate their average relative frequencies in the cells of origin of the lung adenocarcinomas analyzed in Campbell et al. (Supplementary Data 6, current study). These relative frequencies were used to normalize the human data in Supplementary Fig. 16a.

Comparison of mouse vs. human KRas. Coding and amino acid sequences for mouse KRas (Kras_GRCm38.p4, CCDS20693.1) and human KRAS (Kras_GRCh38.p7, CCDS6702.1) were downloaded from Ensembl (http://www.ensembl.org; release 89). Amino acid identity was assessed in a pairwise fashion across each of the 188 amino acids of KRAS. Codon usage for each species was calculated using publicly available codon usage tables. The log-2 fold change in the usage of each codon relative to average (i.e., 1/64) was calculated (Supplementary Fig. 17).

Data availability. All raw Illumina® sequencing data have been deposited in the NCBI Sequence Read Archive (SRA) with study number SRP119241 (accession number SRR119241). The Supplementary Data 7 contains the log-2 fold change in the usage of each codon relative to average (i.e., 1/64) was calculated (Supplementary Fig. 17).
64. Berns, A. Kras and Hras
65. Dimauro, T. & David, G. Ras-induced senescence and its physiological
66. Mckay, J. D. et al. Large-scale association analysis identi
63. Chavan, T. S. et al. High-af
58. Campbell, J. D. et al. Distinct patterns of somatic genome alterations in lung
68. Brinkman, E. K., Chen, T., Amendola, M. & Van Steensel, B. Easy quantitative
60. Gremer, L., Gilsbach, B., Reza Ahmadian, M. & Wittinghofer, A. Fluoride
69. Barzel, A. et al. Promoterless gene targeting without nucleases ameliorates
NATURE COMMUNICATIONS
ARTICLE NATURE COMMUNICATIONS | DOI: 10.1038/s41467-017-01519-y
48. Nakamura, Y., Gojobori, T. & Ikemura, T. Codon usage tabulated from
57. Park, J. T. et al. Differential in vivo tumorigenicity of diverse KRAS mutations
50. Bailey, P. et al. Genomic analyses identify molecular subtypes of pancreatic cancer.
45. Yu, H. A. et al. Prognostic impact of KRAS mutation subtypes in 677 patients
531
13987 (2015).
53
777 (2017).
54. Maresch, R. et al. Multiplexed pancreatic genome engineering and cancer induction by transfection-based CRISP/Cas9 delivery in mice. Nat. Commun. https://doi.org/10.1038/ncomms10770 (2016).
55. Weber, J. et al. CRISP/Cas9 somatic multiplex-mutagenesis for high-throughput functional cancer genomics in mice. Proc. Natl Acad. Sci. USA 112, 13982–13987 (2015).
56. Fernández-Medarde, A. & Santos, E. Ras in cancer and developmental diseases. Genes Cancer 2, 344–358 (2011).
57. Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. Nature 500, 415–421 (2013).
58. Park, I. T. et al. Differential in vivo tumorigenicity of diverse KRAS mutations in vertebrate pancreas: a comprehensive survey. Oncogene 34, 2801–2806 (2015).
59. Campbell, J. D. et al. Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. Nat. Genet. 48, 807–816 (2016).
42. Paulsson, V. et al. High-resolution array CGH identifies common mechanisms that drive embryonal rhabdomyosarcoma pathogenesis. Genes Chromosomes Cancer 39, 381–390 (2004).
43. Kirsch, D. G. et al. A spatially and temporally restricted mouse model of soft tissue sarcoma. Nat. Med. 13, 992–997 (2007).
44. Huang, J. et al. Generation and comparison of CRISPR-Cas9 and Cre-mediated genetically engineered mouse models of sarcoma. Nat. Commun. https://doi. org/10.1038/sknp.2017.54 (2017).
45. Yu, H. A. et al. Prognostic impact of KRAS mutation subtypes in 677 patients with metastatic lung adenocarcinomas. J. Thorac. Oncol. 10, 431–437 (2015).
46. Lampson, B. L. et al. Rare codons regulate KRas oncogenesis. Curr. Biol. 23, 70–75 (2013).
47. Pershing, N. L. K. et al. Rare codons capitate Kras-driven de novo tumorigenesis. J. Clin. Invest. 125, 222–233 (2015).
48. Nakamura, Y., Gojobori, T. & Ikemura, T. Codon usage tabulated from international DNA sequence databases: status for the year 2000. Nucleic Acids Res. 28, 292 (2000).
49. Serrano, M., Lin, A. W., Mccurrach, M. E., Beach, D. & Lowe, S. W. Oncogenic ras provokes premature cell senescence associated with accumulation of p53 and p16INK4a. Cell 88, 393–602 (1997).
50. Bailey, P. et al. Genomic analyses identify molecular subtypes of pancreatic cancer. Nature 531, 47–52 (2016).
51. Wheeler, D. A. & Wang, L. From human genome to cancer genome: the first decade. Genome Res. 23, 1054–1062 (2013).
52. Forrest, A. S. et al. COSMIC cancer genes at high-resolution. Nucleic Acids Res. 45, 777–783 (2017).
53. Maresch, R. et al. Multiplexed pancreatic genome engineering and cancer induction by transfection-based CRISP/Cas9 delivery in mice. Nat. Commun. https://doi.org/10.1038/ncomms10770 (2016).
54. Weber, J. et al. CRISP/Cas9 somatic multiplex-mutagenesis for high-throughput functional cancer genomics in mice. Proc. Natl Acad. Sci. USA 112, 13982–13987 (2015).
55. Fernández-Medarde, A. & Santos, E. Ras in cancer and developmental diseases. Genes Cancer 2, 344–358 (2011).
56. Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. Nature 500, 415–421 (2013).
57. Park, I. T. et al. Differential in vivo tumorigenicity of diverse KRAS mutations in vertebrate pancreas: a comprehensive survey. Oncogene 34, 2801–2806 (2015).
58. Campbell, J. D. et al. Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. Nat. Genet. 48, 807–816 (2016).
59. Alexandrov, L. B. et al. Mutational signatures associated with tobacco smoking in human cancer. Science 354, 618–622 (2016).
60. Gremer, L., Gilbsch, B., Reza Ahmadian, M. & Wittinghofer, A. Fluoride complexes of oncogenic Ras mutants to study the Ras-RasGAP interaction. Biol. Chem. 389, 1163–1171 (2008).
61. Fritz, J. M., Dwyer-Nield, L. D., Russell, B. M. & Malkinsson, A. M. Brief report: the KRAS mutational spectra of chemically induced lung tumors in different inbred mice mimics the spectra of KRAS mutations in adenocarcinomas in smokers vs non-smokers. J. Thorac. Oncol. 5, 254–257 (2010).
62. Ogura, T. et al. Clinopathological characteristics and prognostic impact of colorectal cancers with NRAS mutations. Oncol. Rep. 32, 50–56 (2014).
63. Chovan, T. S. et al. High-affinity interaction of the K-Ras4B hypervariable region with the ras active site. Cancer Discov. 4, 781–789 (2014).
64. Barsel, A. et al. Promoterless gene targeting without nucleases ameliorates haemophilia B in mice. Nature 517, 360–364 (2015).
65. Dimaro, T. & David, G. Ras-induced senescence and its physiological relevance in cancer. Curr. Cancer Drug Targets 10, 869–876 (2010).
66. Mckay, J. D. et al. Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. Nat. Genet. 49, 1126–1132 (2017).
67. Doench, J. G. et al. Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. Nat. Biotechnol. 32, 1262–1267 (2014).
68. Brinkman, E. K., Chen, T., Amendola, M. & Van Steensel, B. Easy quantitative assessment of genome editing by sequence trace decomposition. Nucleic Acids Res. https://doi.org/10.1093/nar/gkx936 (2014).
69. Barzel, A. et al. Promoterless gene targeting without nucleases ameliorates haemophilia B in mice. Nature 517, 360–364 (2015).
70. Grimm, D. Production methods for gene transfer vectors based on adenovirus vector serotypes. Methods 28, 146–157 (2002).
71. Caswell, D. R. et al. Obligate progression precedes lung adenocarcinoma dissemination. Cancer Discov. 4, 781–789 (2014).
72. 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).
73. West, L. J. & Hankin, R. K. S. Exact tests for two-way contingency tables with structural zeros. J. Stat. Softw. https://doi.org/10.18637/jss.v0102.i01 (2008).
74. Callahan, B. J. et al. DADA2: high-resolution sample inference from Illumina amplicon data. Nat. Methods 13, 581–583 (2016).
75. Macmillan, A. P., Bartam, A. K., Truszkowski, I. M., Brown, D. G. & Neufeld, J. D. PANDAseq: paired-end assembler for Illumina sequencing data. BMC Bioinformatics https://doi.org/10.1186/1471-2105-13-31 (2012).
76. Feldser, D. M. et al. Stage-specific sensitivity to p53 restoration during lung cancer progression. Nature 468, 572–575 (2010).
77. Forbes, S. A. et al. COSMIC: exploring the world’s knowledge of somatic mutations in human cancer. Nucleic Acids Res. 43, 805–811 (2015).
78. Cerami, E. et al. The cBio cancer genetics portal: an open platform for exploring multidimensional cancer genetics data. Cancer Discov. 2, 401–404 (2012).
79. Gao, J. et al. Integrative analysis of complex cancer genomics and clinical profiles using the cbioPortal. Sci. Signal. https://doi.org/10.1126/scisignal.aai0488 (2013).
80. Yates, A. et al. Ensemble 2016. Nucleic Acids Res. 44, 710–716 (2016).