


Phase and context shape the function of composite oncogenic mutations

<https://doi.org/10.1038/s41586-020-2315-8>

Received: 7 September 2019

Accepted: 6 April 2020

Published online: 27 May 2020

 Check for updates

Alexander N. Gorelick^{1,2}, Francisco J. Sánchez-Rivera³, Yanyan Cai⁴, Craig M. Bielski^{1,2}, Evan Biederstedt², Philip Jonsson⁵, Allison L. Richards⁵, Neil Vasan^{1,6}, Alexander V. Penson^{1,2}, Noah D. Friedman^{1,2}, Yu-Jui Ho³, Timour Baslan³, Chaitanya Bandlamudi⁵, Maurizio Scaltriti⁴, Nikolaus Schultz^{2,5,7}, Scott W. Lowe^{3,8}, Ed Reznik^{2,5,7}✉ & Barry S. Taylor^{1,2,5,7}✉

Cancers develop as a result of driver mutations^{1,2} that lead to clonal outgrowth and the evolution of disease^{3,4}. The discovery and functional characterization of individual driver mutations are central aims of cancer research, and have elucidated myriad phenotypes⁵ and therapeutic vulnerabilities⁶. However, the serial genetic evolution of mutant cancer genes^{7,8} and the allelic context in which they arise is poorly understood in both common and rare cancer genes and tumour types. Here we find that nearly one in four human tumours contains a composite mutation of a cancer-associated gene, defined as two or more nonsynonymous somatic mutations in the same gene and tumour. Composite mutations are enriched in specific genes, have an elevated rate of use of less-common hotspot mutations acquired in a chronology driven in part by oncogenic fitness, and arise in an allelic configuration that reflects context-specific selective pressures. *cis*-acting composite mutations are hypermorphic in some genes in which dosage effects predominate (such as *TERT*), whereas they lead to selection of function in other genes (such as *TP53*). Collectively, composite mutations are driver alterations that arise from context- and allele-specific selective pressures that are dependent in part on gene and mutation function, and which lead to complex—often neomorphic—functions of biological and therapeutic importance.

To study the pattern, prevalence and function of composite mutations (hereafter defined as two or more distinct somatic mutations in the same gene and tumour specimen) in cancer, we analysed the germline blood and matched tumour tissue of 31,359 patients with cancer for whom prospective clinical sequencing was performed to guide treatment decisions for advanced and metastatic disease (Fig. 1a, Extended Data Fig. 1a, Supplementary Table 1).

Selection for composite mutations

In total, 22.7% ($n = 7,874$) of tumours contained at least one composite mutation—which is 56% more frequent than would be expected by chance, when controlling for gene content and mutational burden ($P < 10^{-5}$) (Extended Data Fig. 1b, c, Methods, Supplementary Table 2). Significantly more composite mutations arose than would be expected in cases of modest mutational burden (4–11 mutations per megabase (Mb), about 44% of all tumours, $P < 10^{-5}$) (Fig. 1b, Extended Data Fig. 1d), an enrichment that decreased in tumours of increasing mutational burden. As positive selection cannot be easily distinguished from the predominantly neutral effect of increasing mutational burden, tumours with a high mutational burden were considered to be biologically distinct and were excluded from analysis (Fig. 1c, Methods).

Finally, we also found that known mechanisms of localized hypermutation explain few composite mutations overall (Extended Data Fig. 2).

Composite mutations in tumour-suppressor genes affected a greater proportion of cases than those in oncogenes (14.2% versus 4.8% of all cases; $P < 10^{-308}$, two-sided McNemar's test) (Fig. 1d). Furthermore, 70% of composite mutations in tumour-suppressor genes consisted of one or more truncating variants, compared to only 13% for oncogenes (Fig. 1e); this suggests that biallelic loss drives the enrichment for composite mutations in tumour-suppressor genes. Lineage-specific patterns of driver mutations in individual cancer genes were, in part, reflected in the pattern of composite mutations (Fig. 2a, Extended Data Fig. 3a). This included a higher burden of composite mutations in *PIK3CA* in breast cancers, *APC* in colorectal cancers, *CDK12* in prostate cancers and *EGFR* in both lung cancers and gliomas, among others. By contrast, not all frequently mutated genes—such as *KRAS* in multiple cancers or *VHL* in renal cell carcinomas—had frequent composite mutations, which often evolve serial genetic changes by other means (such as allelic imbalance and/or loss of heterozygosity).

We next sought to determine whether individual cancer genes were enriched or depleted for composite mutations, controlling for the determinants of their background mutation rate⁹ (Methods). In total, 34 genes were significantly enriched for composite mutations ($Q < 0.01$)

¹Human Oncology and Pathogenesis Program, Memorial Sloan Kettering Cancer Center, New York, NY, USA. ²Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, NY, USA. ³Cancer Biology and Genetics Program, Memorial Sloan Kettering Cancer Center, New York, NY, USA. ⁴Department of Pathology, Memorial Sloan Kettering Cancer Center, New York, NY, USA. ⁵Marie-Josée and Henry R. Kravis Center for Molecular Oncology, Memorial Sloan Kettering Cancer Center, New York, NY, USA. ⁶Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, NY, USA. ⁷Weill Cornell Medical College, New York, NY, USA. ⁸Howard Hughes Medical Institute, New York, NY, USA. ✉e-mail: reznike@mskcc.org; taylorb@mskcc.org

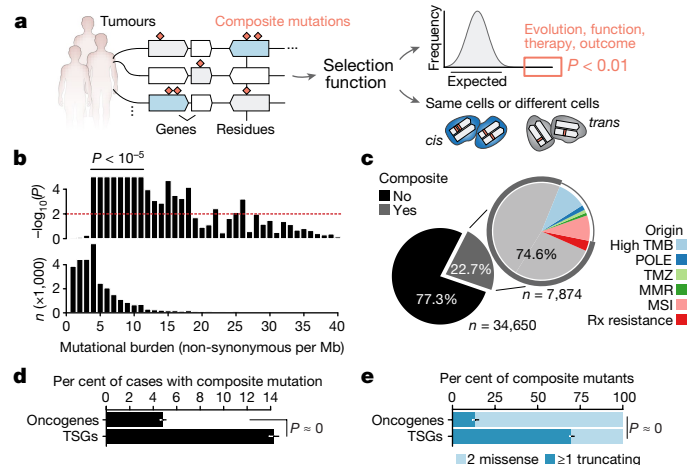


Fig. 1 | Composite mutations in human cancers. **a**, Schematic of the discovery and characterization of composite mutations. **b**, Top, statistically significant enrichment ($P < 10^{-5}$) for composite mutations in tumours of increasing tumour mutational burden. Nominal P based on one-sided permutation tests for enrichment (100,000 permutations) applied independently to the subset of tumours with each indicated tumour mutational burden (bottom, number of cases), $n = 30,505$ biologically independent tumour samples with tumour mutational burden ≤ 40 nonsynonymous exonic mutations per Mb. **c**, Proportion of composite mutations including the fraction ascribed to mutational processes associated with hypermutation. MMR, mismatch repair; MSI, microsatellite instability; POLE, DNA-polymerase- ϵ -associated hypermutation; Rx resistance, acquired resistance to therapy; TMB, tumour mutational burden; TMZ, temozolomide-associated hypermutation; cases excluded from analysis unless otherwise noted. **d**, Percentage of cases with composite mutations by cancer gene function. $P < 10^{-308}$ (numeric limit, two-sided McNemar's test; $n = 29,507$ patients). TSG, tumour-suppressor gene. **e**, Types of composite mutations by cancer-gene function ($P < 10^{-308}$, numeric limit, two-sided Fisher's exact test; $n = 5,954$ composite mutations). Error bars in **d**, **e** are 95% binomial confidence intervals.

(Fig. 2b, Supplementary Table 3), including both tumour-suppressor genes (such as *APC*, *TP53*, *PTEN* and *MAP3K1*) and oncogenes, the most significant of which was *PIK3CA* (9.9% of all mutations in *PIK3CA* were composite, 95% confidence interval 9.0–10.9) (Extended Data Fig. 3b). Other frequently mutated oncogenes were not enriched for composite mutations; these included *IDH1*, which reflects the requirement for heterozygosity in IDH-mutant cells to sustain adequate production of D-2-hydroxyglutarate¹⁰, and *KRAS*, which may reflect selection against further detrimental oncogenic RAS activation^{8,11}. Mutational recurrence alone cannot, therefore, predict whether a cancer gene is enriched for composite mutations.

Consistent with their selection, composite mutants were 2.5-fold more likely than individual mutations to include a hotspot—residues that are mutated in cancer more often than would be expected in the absence of selection^{12,13} ($P < 10^{-308}$, two-sample Z-test for equal proportion) (Fig. 2c). Composite mutations notably lacked the hotspots of greatest positive selection (for example, *KRAS*^{G12} and *BRAF*^{V600}) but were instead prevalent among less common hotspots, which suggests that the selective pressure is greatest for weakly functional alleles. On the basis of differences in their clonality, in 69% of cases the more prevalent hotspot mutation (at the population level) preceded the less prevalent mutation in oncogenes (95% confidence interval 59–78%) (Fig. 2d), consistent with a model in which the less prevalent allele synergizes with a more-potent initial hotspot mutation. Tumour-suppressor genes exhibited no such temporal ordering, which reflects how prevalence is poorly correlated with fitness for predominantly loss-of-function mutations. Together, these data indicate a strong mutant-allele-specific selective pressure for composite mutations that evolve along a chronology driven in part by oncogenic fitness.

Phase and function

The elevated rate of likely driver mutations in composite mutants led us to investigate their allelic configuration. We combined sequencing read support with clonality to phase mutations, and thereby ensured that composite mutations arose in the same tumour cell population. Among evaluable composite mutants, 67% and 19% ($n = 977$ and 275) arose *in cis* (on the same allele) and *in trans* (on different alleles), respectively, and 14% ($n = 210$) were indeterminate. In part, the higher rate of *cis* mutants reflected reduced sensitivity for detecting *trans* mutations from the short-read sequencing used here, an effect we controlled for in subsequent analyses (Methods). Tumour-suppressor genes were substantially more likely to contain composite mutations *in trans*, especially those with two truncating mutations that were consistent with biallelic inactivation (71% *in trans*, $n = 79$ of 111). By contrast, composite-mutant oncogenes with two missense mutations were largely *cis*-acting (91%, $n = 243$ of 268; $P = 3 \times 10^{-33}$, two-sided Fisher's exact test) (Fig. 3a). Composite mutations that involved silent mutations exhibited no such difference in phase among these genes, which suggests that the *cis*-mutant enrichment in oncogenes reflects selective pressure. Notably, although not precluding resistance *in trans*¹⁴, all the secondary resistance mutations that we identified arose *in cis*^{15–17} ($n = 18$; $P = 0.02$, two-sided Fisher's test) (Fig. 3b, Extended Data Fig. 4), which suggests that exogenous selective pressures drive—in part—the phase of composite mutations.

Despite these patterns, extensive variability existed in the phase of composite mutations in individual cancer genes (Fig. 3c). *EGFR*, *TERT* and *PIK3CA* had the highest percentage of *cis* composite mutations among oncogenes (88–97%). Prevalent *cis*-acting composite mutations were observed even among canonical tumour-suppressor genes, comprising 77.1% of all composite mutations in these genes. Here, *TP53* was notable: 43% of all phase-able composite mutations ($n = 70$ of 163) in this gene were *cis*-acting, and enriched in a cluster of residues near the C-terminal end of the DNA binding domain (E287, E285, E271 and R280) (Fig. 3d). Although short-read sequencing technologies restrict phasing to variants within close physical proximity and potentially overestimate the prevalence of *cis* mutations, these data are nevertheless inconsistent with conventional loss of function via biallelic inactivation and may suggest a broader functional effect of composite mutations in *TP53* and other tumour-suppressor genes.

To assess the phenotypic consequence of *cis*-acting composite mutations in the DNA binding domain of *TP53*, we developed an isogenic system for acute reconstitution of *TP53*. As E287D was the most significant mutated residue enriched in composite mutants, we focused on a representative *TP53*^{R280T/E287D} *cis* composite mutant. To model the effect of this composite mutation in the lineage of affected tumours, we transduced *Kras*^{G12D} *Trp53*^{−/−} mouse lung cancer cells with GFP-labelled retroviral constructs that encode complementary (c)DNAs for wild-type *Trp53*, *Trp53*^{R277T}, *Trp53*^{E284D} and *cis* *Trp53*^{R277T/E284D} (orthologous to human wild-type *TP53*, *TP53*^{R280T}, *TP53*^{E287D} and *cis* *TP53*^{R280T/E287D}, respectively), after which we selected GFP-expressing cells and performed RNA sequencing (Fig. 3e, Extended Data Fig. 5a, Supplementary Table 5). *Trp53* mRNA expression was stable and robust, whereas in *Trp53*^{−/−}, *Trp53*^{R277T} and *Trp53*^{R277T/E284D} cells there was a decrease in p21 (encoded by *Cdkn1a*) induction, a surrogate marker of p53 functionality (Extended Data Fig. 5b, c). *Trp53*^{E284D} cells transcriptionally resembled *Trp53*^{+/+} cells, whereas *Trp53*^{R277T} cells resembled *Trp53*^{−/−} cells (Extended Data Fig. 5d). By contrast, *Trp53*^{R277T/E284D} cells had a mixed transcriptional phenotype, bearing a dominant differential expression signature that was equivalent to the one induced in either *Trp53*^{R277T} or *Trp53*^{−/−} cells while retaining a *Trp53*^{E284D}-like downregulation of the AP-1 transcription factor program (Fig. 3f, Extended Data Fig. 5e). These data correlated with human tumour genomics, in which the null-like *TP53*^{R280T} mutation was common but *TP53*^{E287D} mutation was rare and nearly always arose as a composite mutation (Extended Data Fig. 5f).

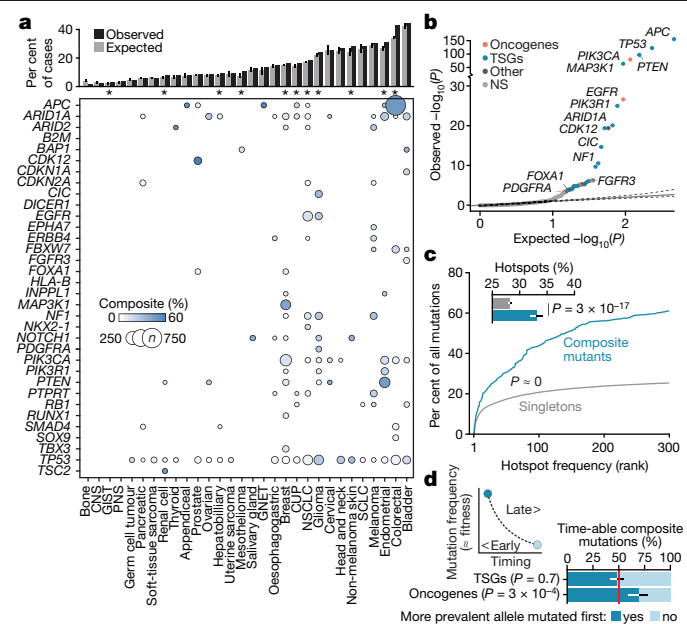


Fig. 2 | Gene- and residue-specific selective pressure for composite mutations. **a**, Prevalence of composite mutations by affected gene and lineage (cancer types of ≥ 100 and ≥ 5 total and composite-mutant cases, $n = 31,563$ samples). Top, percentage of cases with composite mutations, and the expected value on the basis of cohort size and mutational burden. Expected values are the mean percentage of 10,000 random permutations for each lineage. Asterisks denote cancer types with a significantly greater proportion of affected cases than expected by chance (false-discovery-rate (FDR)-adjusted $P < 0.01$). Bars are 95% confidence intervals. CNS, central nervous system; CUP, cancer of unknown primary; GIST, gastrointestinal stromal tumour; GNET, gastrointestinal neuroendocrine tumour; NSCLC, non-small-cell lung cancer; PNS, peripheral nervous system; SCLC, small-cell lung cancer. **b**, The significance of enrichment for composite mutations in cancer genes (FDR-adjusted P values from one-sided binomial test for enrichment, $n = 26,997$; light grey is not significant (NS)). **c**, Hotspot mutation use among composite and singleton mutations by decreasing population-level frequency ($P < 10^{-308}$, numeric limit, two-sided Mann–Whitney U -test, $n = 93,616$ and 2,920 singleton and composite missense mutations, respectively, in 25,037 patients). Inset, the percentage of all missense mutations, comprising composite and singleton mutants that arose at individually significant mutational hotspots. P , two-sided two-sample Z -test for equal proportions, $n = 105,297$ total single-nucleotide variants, error bars are 95% binomial confidence intervals. **d**, Right and left are the proposed and observed temporal order, respectively, of the acquisition of two functional variants in composite mutations in oncogenes (from mutation clonality). Tumour-suppressor genes shown as a negative control. P , two-sided binomial test, error bars in all panels are 95% binomial confidence intervals ($n = 336$ evaluable composite mutations).

A second *cis*-acting composite mutant (*Trp53*^{R277K/E282K}, orthologous to human *TP53*^{R280K/E285K}) also promoted a transcriptional program distinct from its constituent mutations (Extended Data Fig. 5g). Importantly, the *TP53*^{R277T/E284D} mutation was not associated with increased growth in vitro or survival in vivo compared to the individual mutations (Extended Data Fig. 5h, i). Collectively, these data suggest that *cis*-acting *TP53* composite mutations tune mutant p53 transcriptional phenotypes, which leads to a selection of function that is absent from null-like single *TP53* mutations.

Conditionally dependent mutant alleles

The residue-specific transcriptional phenotypes of *TP53*-composite mutants suggest broader allele-specific selection among composite mutations. We therefore identified individual alleles that exhibit an

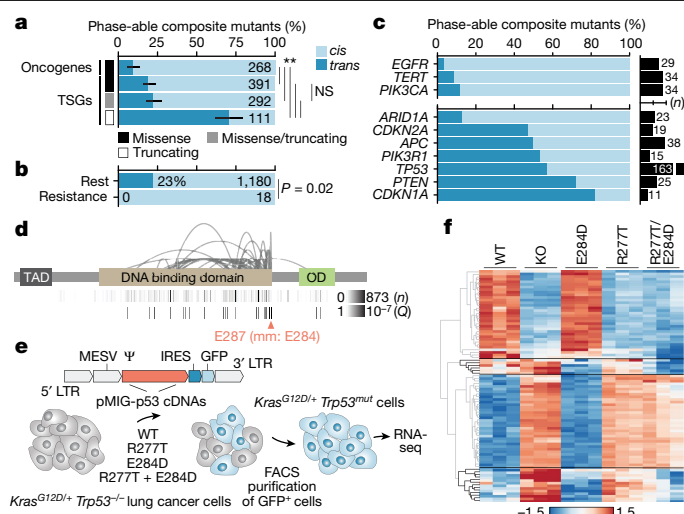


Fig. 3 | *cis*- and *trans*-acting composite mutants. **a–c**, The phase of composite mutations by their type and affected cancer gene (P for asterisked comparisons from left to right are 4×10^{-4} , 2×10^{-5} , 3×10^{-33} , 8×10^{-24} and 3×10^{-19} , and not significant was 0.3, two-sided Fisher's exact test, $n = 1,062$ evaluable composite mutations; error bars are 95% binomial confidence intervals) (**a**); association or not with acquired therapy resistance (P , two-sided Fisher's exact test, $n = 1,198$ evaluable composite mutations) (**b**); and affected individual oncogenes and tumour-suppressor genes (top and bottom, respectively, known and predicted functional mutations in ≥ 10 phase-able tumours; number of cases with phase-able composite mutations as indicated) (**c**). **d**, The pattern of *TP53* composite mutations with arcing lines indicating the position of pairs of mutations in ≥ 2 tumours; height corresponds to recurrence. At the bottom, the number of mutated cases at each individual residue and the Q of significance (FDR-adjusted P value from one-sided binomial test) for each residue as arising in composite are shown. TAD, transactivation domain; OD, oligomerization domain. mm, mouse. **e**, Schematic of the experimental workflow for generating isogenic cells for phenotypic comparison of *TP53* mutations. **f**, Heat map of the top 30 differentially expressed genes between *Trp53*^{R277T}, *Trp53*^{E284D}, and *Trp53*^{R277T/E284D}-mutant cells.

excess of composite mutations (Methods). In total, 86 mutant residues in 24 cancer genes were enriched for arising as composite variants ($Q < 0.01$) (Fig. 4a, Supplementary Table 4). Nearly 70% of these mutations occurred in only 4 genes (*TP53*, *PIK3CA*, *APC* and *EGFR*), with few reaching saturation for discovery at the current cohort size, and 56% also arising as individually significant hotspot mutations¹³ (Fig. 2b, Extended Data Fig. 6). As with *TP53*, several tumour-suppressor genes had mutant-allele-specific enrichment that may suggest selection for something other than conventional loss of function. In *PIK3CA*, mutations that are enriched in composite mutants (in residues E726, E453, K111, R108 and R93) were nearly always *in cis* when phase-able, and often arose through APOBEC-associated mutagenesis (Extended Data Fig. 7). Notably, composite *PIK3CA* mutations drive elevated PI3K activity, downstream signalling, cell proliferation and tumour growth, and may increase sensitivity to PI3K inhibitors¹⁸, confirming that—in addition to introducing passenger mutations—APOBEC and other mutational processes create numerous functional driver mutations.

Multiple significant residues appeared to be conditional alleles—rarely arising without a second *cis* activating mutation (Extended Data Fig. 8a). Among these were *EGFR*-mutant residues (E709, V834 and L833)¹⁹ and the *TERT* promoter mutation 205G>A (Fig. 4a). *TERT* promoter mutations are common in human cancer²⁰ and create novel GABPA binding sites that promote aberrant telomerase activity²¹. The 205G>A mutation was the sixth most common *TERT* promoter mutant, and exclusively arose *in cis* ($n = 13$ of 13) with either the highly prevalent 228G>A or 250G>A hotspots, which—despite their frequency—were never together in composite (Extended Data Fig. 8b). To test whether

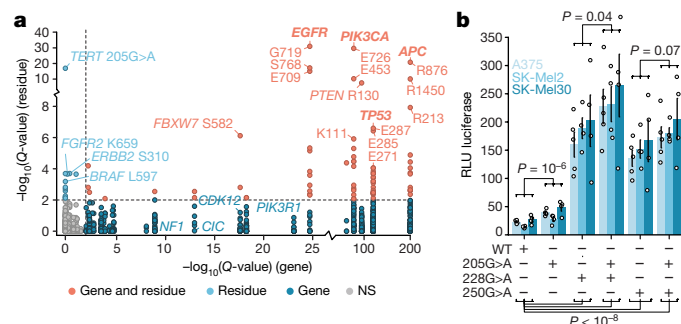


Fig. 4 | Mutant-allele-specific enrichment for composite mutations. a, Enrichment significance of individual mutant residues arising in composite mutations ($n = 1,821$ distinct mutant sites tested; $n = 155,241$ variants overall) compared to significance of composite enrichment among genes (Q for mutant sites is FDR-adjusted one-sided Fisher's exact test; for Q for genes, refer to Fig. 2b). Genes in bold label each of the residues beneath. **b,** The degree of *TERT* expression induced by transient transfection of the indicated mutations individually, or as *cis* composite, in three melanoma cell lines (A375, SK-Mel2 and SK-Mel30). The mean and s.e.m. (error bars) across $n = 4$ or 5 replicates per allele. P , two-way analysis of variance assessing expression as a function of genotype and baseline expression of each cell line (Methods); at the bottom, $P < 10^{-8}$ values from left to right are 3×10^{-9} , 1×10^{-9} , 2×10^{-9} and 2×10^{-11} . RLU, relative luminescence unit.

the 205G>A mutation synergizes with existing promoter mutations to enhance *TERT* expression, we expressed constructs with a luciferase reporter engineered to contain various *TERT* promoter mutations alone or as *cis*-composite mutants in three melanoma cell lines (A375, SK-Mel2 and SK-Mel30). *TERT*^{205G>A} induced modest *TERT* expression compared to wild type, but less than *TERT*^{228G>A} or *TERT*^{250G>A} alone. Consistently, *TERT*^{205G>A} creates a novel motif to which GABPA binds with lower affinity than those motifs created by canonical *TERT* hotspots (Extended Data Fig. 8c). The selective pressure for *TERT*^{205G>A} is therefore probably based on the cooperativity of tandem motifs generated by this mutation, and canonical promoter hotspots bound by GABPA heterotetramer complexes²¹. When expressing *TERT*^{205G>A} as a *cis* composite with either *TERT*^{228G>A} or *TERT*^{250G>A} (thereby modelling the 205G>A-mutant human tumours), *TERT* expression increased relative to either mutation alone (Fig. 4b). These data suggest that 205G>A is hypermorphic, driving modestly elevated *TERT* expression that is weakly selected for and therefore does not arise as an individual hotspot mutation but is instead a conditionally dependent composite allele.

Our results indicate that composite mutations are driver alterations with a selective advantage that appears to be primarily determined by their allelic configuration and context. No single model explains the context-dependent phenotypic consequences of composite mutations. In some cancer genes with dosage-dependent function, *cis*-acting composite mutants are additive and arise predominantly in weakly oncogenic alleles and genes (for example, *PIK3CA*^{22–24}). This suggests an evolutionary model in which the second mutation arises through selection for hypermorphic activity beyond the level sufficient for activation by the first allele. In genes (such as *TP53*) with manifold phenotypic consequences, *cis* mutants seem to drive functional innovation.

With these mutations, the evolutionary advantage consistent with our results is via tuning of the subtle phenotypic differences that are conferred by the asymmetric combination of the output of individual mutations. Mutant cancer genes must ultimately be considered—both biologically and clinically—in their allelic context, with implications for our understanding of cancer gene function, malignant phenotypes and therapy.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-020-2315-8>.

- Vogelstein, B. et al. Cancer genome landscapes. *Science* **339**, 1546–1558 (2013).
- Garraway, L. A. & Lander, E. S. Lessons from the cancer genome. *Cell* **153**, 17–37 (2013).
- Cairns, J. Mutation selection and the natural history of cancer. *Nature* **255**, 197–200 (1975).
- Nowell, P. C. The clonal evolution of tumor cell populations. *Science* **194**, 23–28 (1976).
- Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646–674 (2011).
- Hyman, D. M., Taylor, B. S. & Baselga, J. Implementing genome-driven oncology. *Cell* **168**, 584–599 (2017).
- Knudson, A. G., Jr. Mutation and cancer: statistical study of retinoblastoma. *Proc. Natl Acad. Sci. USA* **68**, 820–823 (1971).
- Bielski, C. M. et al. Widespread selection for oncogenic mutant allele imbalance in cancer. *Cancer Cell* **34**, 852–862.e4 (2018).
- Lawrence, M. S. et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).
- Jin, G. et al. Disruption of wild-type IDH1 suppresses d-2-hydroxyglutarate production in IDH1-mutated gliomas. *Cancer Res.* **73**, 496–501 (2013).
- Mueller, S. et al. Evolutionary routes and KRAS dosage define pancreatic cancer phenotypes. *Nature* **554**, 62–68 (2018).
- Chang, M. T. et al. Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. *Nat. Biotechnol.* **34**, 155–163 (2016).
- Chang, M. T. et al. Accelerating discovery of functional mutant alleles in cancer. *Cancer Discov.* **8**, 174–183 (2018).
- Intlekofer, A. M. et al. Acquired resistance to IDH inhibition through *trans* or *cis* dimer-interface mutations. *Nature* **559**, 125–129 (2018).
- Hidaka, N. et al. Most T790M mutations are present on the same *EGFR* allele as activating mutations in patients with non-small cell lung cancer. *Lung Cancer* **108**, 75–82 (2017).
- Gainor, J. F. et al. Molecular mechanisms of resistance to first- and second-generation ALK inhibitors in ALK-rearranged lung cancer. *Cancer Discov.* **6**, 1118–1133 (2016).
- Kobayashi, S. et al. *EGFR* mutation and resistance of non-small-cell lung cancer to gefitinib. *N. Engl. J. Med.* **352**, 786–792 (2005).
- Vasan, N. et al. Double *PIK3CA* mutations *in cis* increase oncogenicity and sensitivity to PI3Ka inhibitors. *Science* **366**, 714–723 (2019).
- Chen, Z. et al. *EGFR* somatic doublets in lung cancer are frequent and generally arise from a pair of driver mutations uncommonly seen as singlet mutations: one-third of doublets occur at five pairs of amino acids. *Oncogene* **27**, 4336–4343 (2008).
- Huang, F. W. et al. Highly recurrent *TERT* promoter mutations in human melanoma. *Science* **339**, 957–959 (2013).
- Bell, R. J. A. et al. The transcription factor GABP selectively binds and activates the mutant *TERT* promoter in cancer. *Science* **348**, 1036–1039 (2015).
- Berenjeno, I. M. et al. Oncogenic *PIK3CA* induces centrosome amplification and tolerance to genome doubling. *Nat. Commun.* **8**, 1773 (2017).
- Kinross, K. M. et al. An activating *Pik3ca* mutation coupled with *Pten* loss is sufficient to initiate ovarian tumorigenesis in mice. *J. Clin. Invest.* **122**, 553–557 (2012).
- Madsen, R. R. et al. Oncogenic *PIK3CA* promotes cellular stemness in an allele dose-dependent manner. *Proc. Natl Acad. Sci. USA* **116**, 8380–8389 (2019).

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

© The Author(s), under exclusive licence to Springer Nature Limited 2020

Methods

No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment.

Prospective sequencing cohort

Somatic mutation data consisted of 34,650 tumour and matched normal specimens from 31,359 patients with prospectively characterized solid cancers. All patients provided written informed consent and were prospectively sequenced as part of their active care at Memorial Sloan Kettering Cancer Center (MSKCC) between January 2014 and April 2019 as part of an Institutional-Review-Board-approved research protocol (NCT01775072). Details of patient consent, sample acquisition, sequencing and mutational analysis have previously been published^{25,26}. In brief, matched tumour and blood specimens for each patient were sequenced using MSK-IMPACT, a custom hybridization capture-based next-generation sequencing assay. All samples were sequenced with one of three incrementally larger versions of the assay encompassing 341, 410, and 468 cancer-associated genes, respectively. The study cohort consisted of tumours samples with one of 429 distinct subtypes of cancer. For the purposes of grouping histological subtypes into primary cancer diagnosis, we used the OncoTree structured classification of disease (<http://oncotree.mskcc.org>). Histologic subtypes of fewer than 50 tumour samples were aggregated into a miscellaneous category and nonsolid tumour types were excluded from the study cohort (as well as from analyses of The Cancer Genome Atlas (TCGA) data), resulting in a final cohort of 41 distinct types of tumour.

Mutational data and annotation

Somatic nonsynonymous substitutions and small insertions and deletions (indels) were identified with a clinically validated pipeline, as previously described^{26,27}. Each mutation was classified as probably functional if it was previously reported as a mutational hotspot^{12,13}. Truncating variants were considered probably functional if they arose in known tumour-suppressor genes, on the basis of gene function curated by OncoKB²⁸. Finally, any additional somatic mutations that did not satisfy the aforementioned criteria were similarly annotated as probably functional if previously curated via literature mining by OncoKB as oncogenic, probably oncogenic or predicted to be oncogenic²⁸.

For all composite mutants in which one or both mutations were a known therapeutic target or known resistance mutation as defined by OncoKB levels 1 to 4, R1 or R2 alterations (annotation as of April 2019), each mutation was manually reviewed and classified as a likely resistance mutation on the basis of the cancer type of the affected tumour sample, the existence of known resistance mutations to commonly used targeted therapies indicated for the given cancer type and—if available—review of the clinical histories of affected patients. Composite mutations in which one mutation was an established second-site mutation (for example, *EGFR*^{T790M} in non-small cell lung cancer¹⁷ and *AR* mutations in prostate cancer that mediate resistance to anti-androgen therapy) were always classified as resistance mutations. Notably, composite mutations in only 3.4% of cases in this advanced and post-treatment cohort have been associated with therapy resistance, indicating that prior therapy exposure alone cannot explain their prevalence. However, as detailed clinical histories including previous lines of treatment and response phenotypes were not available for all patients, a small number of composite mutations are probably misclassified as non-resistance-associated.

Mutational burden classification

Tumour samples were classified as hypermutated if they contained either MSI MMR deficiency, *POLE*-mediated ultra-mutation, or TMZ-induced hypermutation²⁹. MSI was considered present for any tumour with an MSISensor³⁰ score of greater than or equal to 10, as

previously clinically validated³¹. Tumour samples with *POLE*, MMR and TMZ-induced hypermutation were identified by mutational signature decomposition analysis. In brief, in each tumour specimen with 20 or more substitutions, the proportion of mutations attributable to each of 30 known somatic mutational signatures were calculated on the basis of a basin-hopping algorithm (<https://github.com/mskcc/mutation-signatures>)³². This method uses the distribution of 96 unique trinucleotides generated by 6 possible C- or T-normalized single-nucleotide substitutions (that is, C>A, C>G, C>T, T>A, T>C or T>G) and their 5'- and 3'-adjacent bases to estimate the fraction of mutations attributed to each mutational signature in each specimen. Tumour specimens for which at least 20% of its substitutions were attributed to *POLE* (signatures 10 or 14), TMZ (signature 11), or MMR (signatures 6, 15, 20, 21 or 26) were classified as hypermutated.

To classify tumour specimens with a high mutational burden compared to the majority of cancers of that type, but that otherwise lack one of these known mechanisms of hypermutation, we performed in each individual cancer type of greater than 50 tumour specimens one-dimensional *k*-means clustering of the mutational burden of all tumours (nonsynonymous exonic mutations per Mb). Between 1 and 9 clusters were inferred to best describe the distribution of mutational burden per cancer type. The cluster of lowest mutational burden centred at 20+ mutations per Mb and accounting for <10% of the samples in tumour type established the threshold for high mutational burden, and all tumour specimens in this cluster or those clusters with higher mutational burden were considered to be of high mutational burden.

Composite mutation identification and annotation

For the purposes of this analysis, a composite mutation was the occurrence of two or more somatic mutations to the same gene within a single sequenced tumour specimen. Carriers of pathogenic germline variants with a second somatic mutation were not considered here. We identified composite mutations as arising owing to somatic hypermutation or high mutational burden of unknown aetiology (as defined in 'Mutational burden classification'), or a mechanism of resistance to targeted therapy per the aforementioned annotation ('Mutational data and annotation') in nonhypermutated tumours. Any composite mutation arising in a hypermutated tumour was considered separately, and excluded from primary analyses unless otherwise noted. All composite mutations that did not meet these criteria were analysed further.

Testing of population-, gene- and residue-specific composite mutation enrichment

Multiple somatic mutations will accumulate in a gene in the absence of selection at a rate that correlates with the mutational burden and mutational mechanisms of a given tumour. Using a permutation-based framework, we simulated the burden of composite mutations for a given tumour mutation burden. In brief, the true number of tumour specimens containing a composite mutation was calculated (n^{true}). We assembled an $m \times 2$ matrix, in which m is the total number of nonsynonymous somatic mutations in our cohort. Each row in the matrix identified the sample and the gene in which a particular mutation arose. We constructed a null distribution by randomly permuting the second column of this matrix 100,000 times, thereby preserving the mutation burden of each gene and each tumour specimen. Upon each iteration, the number of tumour specimens containing a composite mutation was reassessed. An empirical *P* value was calculated as the fraction of permutations satisfying $n_i \geq n^{\text{true}}$. We used the same procedure for assessing the enrichment of composite mutations for tumour samples in ranges of specific mutational burdens.

To test for enrichment or depletion for composite mutations within cancer types (in cancer types with more than 50 profiled tumours), we used a modified permutation analysis controlling for the underlying gene-specific tendency for mutated genes within each cancer type to contain a composite. To do so, we defined a mutation event to be

a tumour-sample-mutated gene tuple. A mutation event (s, g) occurs when a tumour sample s was found to contain one or more mutations to a gene, g . Then, we implemented a permutation analysis that shuffles mutations across samples in a manner that preserves (1) gene mutation burden, (2) tumour sample mutation burden and (3) the total number of mutation events that were observed in that cancer type using the `permutswap` function in the R package `vegan`³³. This final constraint enforces that the number of non-zero entries in the mutation event matrix (the binary matrix of patients and genes) remains constant for each permutation. This constraint is particularly relevant in cancer types that have a mutation burden that is dominated by genes that are depleted of composite mutations (for example, *KRAS* in pancreatic cancer and *BRAF* or *KRAS* in thyroid cancer).

We evaluated the enrichment of composite mutations in each gene by modelling composite mutation burden as a function of genomic covariates, testing the likelihood of the observed number of composite mutations (corresponding to the probability of observing this burden of composite mutations by chance) using a binomial test. To parametrize \hat{p} (the background rate of composite mutations in the absence of selection for each gene g), we estimated the expected number of composite mutated samples in a gene n_c from the total number of samples with an observed mutation in the gene n_s , such that $\hat{p}^g = n_c^g / n_s^g$. Dropping the superscript for clarity, n_c was estimated for each gene using negative binomial regression to model the observed number of composite-mutant samples in a gene n_c as a function of the global background rate of composite mutations across all genes, adjusted for multiple covariates per gene, including its replication timing r , coding sequence length l , the per cent of GC content g and the chromatin state of the gene h . Coding sequence length and per cent of GC content were obtained from the Biomart community portal³⁴ for Ensembl human reference genome GRCh37. For the purposes of statistical testing, the noncoding promoter region of *TERT* was added as a distinct unit (gene) for which we computed distinct values of per cent GC content and length for the region targeted by the MSK-IMPACT assay design. Replication timing and chromatin state for each gene were obtained from previous estimates⁹. Additional covariates included the version of the MSK-IMPACT assay in which the gene was introduced i , and the average total DNA copy number of the gene across its mutated samples t . As the composite mutation rate for a gene depends on both the number of composite mutant tumours and the number of samples mutated (that is, the exposure for the count of composite mutants), an offset term was added to the model that represents the log-transformed number of tumour samples containing mutations in the gene of interest. The observed number of composite mutant tumours for a gene was therefore modelled as $n_c \sim \text{NB}(r + l + g + h + i + t + \text{offset}(\log(n_s)))$. Using this model, we predicted the number of composite mutant tumours for each gene arising by chance, \hat{n}_c , calculating the expected fraction of samples with a composite mutation (out of the total number of mutated samples) in each gene \hat{p} . We then used a binomial test to evaluate the null hypothesis that for each gene the observed number of composite mutations arose owing to random chance. Here, we modelled the incidence of composite mutations per gene using a binomial distribution, and calculated the probability of n_s tumour specimens containing composite mutations in n_c tumour specimens by chance given \hat{p} :

$$\Pr(X \geq n_c) = \sum_{i=n_c}^{n_s} \binom{n_s}{i} \hat{p}^i (1 - \hat{p})^{n_s - i}$$

Our parameterization \hat{p} was estimated using nonsynonymous mutations, including those under positive selection in cancer (for example, hotspots), which may reduce overall model sensitivity. We therefore evaluated multiple alternative parameterizations of \hat{p} , including using (1) nonsynonymous mutational data that exclude known hotspot mutations under selection and (2) only synonymous mutations. Neither

alternative parameterization produced a qualitatively distinct result for genes originally detected as significantly enriched, but did increase the overall sensitivity of the test. To ensure appropriate control for potential false-positive findings, we leveraged the parameterization from the complete dataset on nonsynonymous mutational data. Moreover, we observed no difference in the rate of synonymous mutations among genes that were either enriched for composite mutations or not ($P = 0.2$, Mann–Whitney U -test), indicating there was little evidence for the accumulation of variants in the absence of selective pressure.

Finally, all unique individual mutant residues present in five or more nonhypermutated cases, excluding known or likely resistance mutations, were also assessed for the significance of their enrichment for arising as composite mutations. All missense, nonsense, splice-site and translation start-site mutations at a given residue were included, as were unique mutant positions in the promoter of *TERT* and in-frame indels spanning known hotspots of clustered indels¹³. For each residue in a given gene, we assessed whether it arose as part of a composite mutation significantly more often than all other mutant residues in the same gene using a right-sided Fisher's exact test. Mutant residues were considered significant at FDR-adjusted $P < 0.01$ ('Statistical analyses and figures').

Attributing mutations to mutagenic processes

We attributed the individual variants that comprise composite mutations to a mutational origin using 1 of 30 established mutational signatures^{35,36}. Mutational signature decomposition in each tumour was performed as described in 'Mutational burden classification' and a signature was considered present if it accounted for five or more substitutions in the affected specimen (to ensure high-confidence decompositions in targeted sequencing data with comparatively fewer mutations relative to broader-scale sequencing). Multiple signatures of the same aetiology were merged by combining the frequency distribution of trinucleotide contexts (APOBEC signatures 2 and 13; MMR signatures 6, 15, 20, 21 and 26; and smoking-associated signatures 4, 18, 24 and 29). A substitution was attributed to a mutational signature present in a given case if the probability weight of the relevant trinucleotide exceeded 10%. For a substitution attributed to multiple signatures present in an affected tumour, it was attributed to the signature that was most frequently associated with the affected cancer type. To adjust for the nonspecificity of trinucleotide context probabilities for smoking-associated signatures, C>A mutations—regardless of trinucleotide context—were considered smoking-associated in tumours for which mutational signature decomposition identified a smoking signature (in oesophageal squamous and adenocarcinomas; head and neck squamous; hepatobiliary; hepatocellular; lung squamous, adenocarcinoma, and adeno-squamous, oral cavity and renal cell carcinoma)³⁷. Substitutions of a trinucleotide context of insufficient probability in any signature in an affected tumour were considered of ambiguous origin and not attributable, and those mutations that could be attributed to ageing and another signature present in a given tumour were considered nonseparable and classified as being of multiple signatures.

Finally, we also considered several additional mechanisms that can drive site-specific mutation rates as potential sources of composite mutations^{38–40}. First, we estimated the mutation rate within 1 kb up- and downstream of all nucleosome dyads (obtained from <https://bitbucket.org/bbglab/nucleosome-periodicity/src/master/>) mapping to regions sequenced in the MSK-IMPACT panels. Having fit a spline to the mutation rate distribution, we calculated the full-width-half-maximum distances from the dyad and compared the rate of singleton and composite mutations within this region (Extended Data Fig. 2b). We conducted a similar analysis on the potential effect of active coding transcription factor binding sites (TFBSs) on composite mutations. We obtained the positions of active TFBSs in coding regions of the genome via integration with DNase I hypersensitive binding sites in human melanocytes following an established procedure³⁹. The mutation rate within 1 kb

of these active TFBSs were inferred using TCGA cutaneous melanoma samples from the TCGA MC3 dataset to increase the total number of mutations among melanoma samples. We then assessed the proximity of singleton and composite mutations to the elevated mutation rate at TFBS sites as described for nucleosome dyads (Extended. Data Fig. 2).

To investigate the effect of APOBEC3A-mediated mutagenesis, we obtained the position of the optimal stem-loop DNA structure favoured by APOBEC3A from published sources⁴⁰. We investigated the overlap of such optimal sites with those mutant alleles that were enriched for arising as a composite mutation. In total, only 1 of 86 significant residues enriched for arising as a composite mutation was at the position of the optimal APOBEC3A substrate (*ARID1A*^{S226A}). Finally, we compared the rate of composite mutations involving known hotspot mutations as described in 'Mutational data and annotation' with those derived from an orthogonal method optimized to reduce false-positive mutations due to site-specific mutagenesis⁴¹. Controlling for overlapping gene content, there was no difference between the proportion of composite mutations involving hotspot mutations based on the origin of the hotspot mutations (per cent and 95% confidence interval are 9.6 (9.2–10) versus 10 (9.6–10.5), $P = 0.2$, two-sample Z-test), indicating that no excess of false-positive hotspots due to site-specific mutagenesis are driving the results described here.

Phasing composite mutations

The allelic configuration of composite mutations (phase)—*in cis* (arising on the same allele) or *in trans* (arising on different alleles)—was inferred primarily from sequencing read support. In brief, for each pair of somatic mutations in a composite mutant, all reads spanning the relevant loci were re-aligned to the reference genome (hg19) by pairwise sequence alignment using a Needleman–Wunsch algorithm⁴². The numbers of unique reads that supported both alleles being wild type (AB), both alleles being mutant (ab) or a mixture of mutant and wild-type alleles (aB or Ab) were subsequently tabulated. For the purposes of the present study, composite mutations were classified as *in cis* when: (1) three or more spanning reads supported both mutant alleles ($ab \geq 3$) and (2) at least one of these variants was called by two or fewer spanning reads that called the other variant as wild type (that is, $aB \leq 2$ or $Ab \leq 2$, or both). Composite mutations were classified as *in trans* when: (1) each variant was supported by three or more reads that were simultaneously wild type for its partner mutation ($aB \geq 3$ and $Ab \geq 3$), (2) two or fewer reads called both mutant alleles ($AB \leq 2$) and (3) the mutations arose in the same tumour cell population on the basis of their cancer cell fractions (CCFs, see 'Assessing cellular context and molecular timing'). There is an inherent difference in the sensitivity of detection for *cis* and *trans* variants, specifically that *trans* variants must satisfy at least two read-support positive criteria ($aB \geq 3$ and $Ab \geq 3$) and are required to be in the same cell, whereas *cis* variants require only a single positive criterion ($ab \geq 3$) without any constraint of evidence for arising in the same cell. This difference in sensitivity for detection probably explains—to some extent—the increased number of *cis* relative to *trans* composite mutations. To determine the effect of this sensitivity bias, we also phased variants with at least one synonymous mutation. We observed no difference in the rate of synonymous composite mutations in oncogenes versus tumour-suppressor genes (5% versus 7%, $P = 0.2$, Mann–Whitney U-test), in contrast to the significant difference in nonsynonymous composite mutations (14% versus 35%, $P < 10^{-6}$). To control for differences in the sensitivity of detection of *cis* and *trans* mutations, analyses of the effects of allelic configuration on composite mutations compared the relative fraction of *cis* and *trans* mutations between two defined groups (for example, oncogenes versus tumour-suppressor genes).

We additionally inferred the phase of select composite mutants associated with therapeutic resistance mutations in regions of clonal loss of heterozygosity (LOH or copy-neutral LOH). Composite mutants spanned by LOH were assumed to be *in cis* if the spanning locus had a minor copy number of zero and a total copy number of one or more

(LOH via heterozygous loss, copy-neutral LOH or the latter combined with subsequent genomic gains) inferred from the aforementioned purity-corrected integer copy number data from FACETS. These must also have arisen in the same tumour cell population as estimated from CCFs (as described in 'Assessing cellular context and molecular timing') and their observed mutant allele frequencies were approximately equal to the expected mutant allele frequencies for a given copy number state in a *cis* allelic configuration (95% confidence intervals of the observed mutant allele frequency overlap the expected mutant allele frequency of the given copy number configuration, controlling for tumour purity). Composite mutations that did not satisfy any of the aforementioned conditions were not able to be unambiguously phased.

As with other short-read sequencing data, our phasing approach is limited by the requirement that any two mutations arise within sufficient physical proximity in the genome to be spanned by common aligned sequencing reads. Although the higher depth of sequencing coverage in our targeted clinical sequencing platform (about 700-fold median in the tumour samples) does increase the likelihood of sequencing a fragment of tumour DNA encompassing both somatic mutations, and improves the quantification of CCFs by reducing measurement error⁸, this limitation cannot be overcome with short-read sequencing.

Assessing cellular context and molecular timing

We estimated the clonality of all somatic mutations in each affected tumour specimen (the CCF) using the FACETS framework, as described previously⁸. To ensure conservative estimates, all somatic mutations were conservatively assumed to have arisen on the major (more common) allele, thus minimizing the possibility of overestimating the CCF. To determine whether the constituents of a composite mutation arose in the same cell, we defined a criterion based on the confidence intervals of the CCF. Specifically, if the sum of the lower bounds of the 95% confidence intervals for each mutation CCF summed to greater than 1, the two somatic mutations in the same gene and tumour specimen were considered to exist within the same cancer cell population. If either of the two somatic mutations were clonal (the upper bound of the 95% confidence interval overlapped 1), then both mutations were considered to have arisen in the same tumour cell population.

We inferred the chronological order of two somatic mutations in each composite mutation on the basis of their estimated CCFs. Any mutations previously associated with acquired resistance to targeted therapies were excluded, as these will arise after the originating sensitizing lesion and skew results. Only composite mutations determined to arise in the same tumour cell population (based on the sum of CCFs) were considered and required previous evidence establishing both mutations as candidate functional driver mutations individually. The 95% confidence intervals of the CCFs of both mutations were inferred as previously described⁴³. If the lower bound of the 95% confidence interval was greater than the upper bound of the 95% confidence interval for a second variant, then the first mutation was determined to have a greater clonality, and therefore to have arisen first in the tumour. Similarly, if the upper bound of the 95% confidence interval of a mutation was less than the lower bound of the other mutation in the composite, it was considered to have arisen second. If the 95% confidence intervals of CCFs of the two mutations in the composite overlapped, or if there was not sufficient evidence that the two mutations existed in the same cancer cell population in the affected tumour specimen, we considered their chronology to be indeterminate.

TP53 composite mutation analysis and validation studies

For the generation of MSCV-p53-IRES-GFP constructs (pMIG-p53 cDNAs), methods were as follows. Fragments encoding wild-type, single- or composite-mutant *Trp53* (mouse orthologue to human *TP53*) cDNAs were obtained from IDT or SGI-DNA, and cloned into pMIG (Addgene no. 9044) using standard restriction enzyme-based methods. In brief, *Trp53* cDNAs were amplified using primers that add BglII and EcoRI restriction

Article

sites on the 5' and 3' regions, respectively, and subsequently digested and cloned into linearized pMIG backbone containing BglIII and EcoRI cloning overhangs. All constructs were sequence-verified using Sanger sequencing. Primer sequences are available in Supplementary Table 5.

HEK293T (ATCC CRL-3216) cells were obtained from ATCC. Mouse *Kras*^{G12D/+}*Trp53*^{-/-} lung adenocarcinoma cells were provided by the Jacks Laboratory⁴⁴. All cells were maintained in a humidified incubator at 37 °C with 5% CO₂, and grown in DMEM supplemented with 10% FBS and 100 IU/ml penicillin–streptomycin. For virus production, 7.5 million HEK293T cells were plated in 15-cm plates the day before transfection. The following day, cells were transfected with 10 µg pMIG-p53 cDNA (or pMIG-empty as control) and 10 µg of pCL-Eco (Addgene no. 12371) using 50 µl of Lipofectamine 2000 (ThermoFisher). Twenty-four hours later, transfection medium was replaced with fresh DMEM. Two rounds of virus were collected (at 48 and 72 h after transfection), pooled and kept at 4 °C until used for cell transduction. One million *Kras*^{G12D/+}*Trp53*^{-/-} lung adenocarcinoma cells were seeded in 10-cm plates and immediately transduced with retroviral supernatants and 8 µg/ml polybrene. Cells were grown for 48 h before purifying using fluorescence-activated cell sorting (FACS). All transductions were done in triplicate. Following transduction, stable GFP⁺ populations were purified by FACS on a FACSaria (BD Biosciences). One hundred and twenty hours after transduction, total RNA was isolated using the RNeasy Mini Kit (Qiagen) following standard manufacturer protocols.

Purified polyA mRNA was subsequently fragmented, and first- and second-strand cDNA synthesis was performed using standard Illumina mRNA TruSeq library preparation protocols. Double-stranded cDNA was subsequently processed for TruSeq dual-index Illumina library generation. For sequencing, pooled multiplexed libraries were sequenced on NextSeq instrumentation in high-output mode, generating approximately 12 million 76-bp single-end reads per replicate condition. The resulting RNA sequencing data were analysed by first trimming adaptor sequences using Trimmomatic⁴⁵. Sequencing reads were aligned to GRCm38.p5 (mm10) using STAR⁴⁶, and genome-wide transcript quantification was performed using featureCounts⁴⁷. After removing transcripts with fewer than 8 aligned reads (low undetected expression at given library size, $n = 9,848$ transcripts retained), differentially expressed genes were identified using DESeq2, with a cutoff of absolute log₂-transformed fold change ≥ 1 and adjusted $P < 0.01$ between experimental conditions⁴⁸. Mouse genes were mapping to human homologues using gene homologies provided by the Mouse Genome Database project⁴⁹. Principal components analysis was performed with output from DESeq2⁴⁸. For fluorescent competition assays, FACS-purified *Kras*^{G12D/+}*Trp53*^{-/-} lung adenocarcinoma cells stably transduced with either pMIG-empty or pMIG-p53-R277T-E284D were mixed at about 60:40 with untransduced parental cells and cultured in vitro for 10 days. The percentage of GFP⁺ cells was monitored over time using a Guava easyCyte HT flow cytometer (Millipore).

All mouse experiments were approved by the MSKCC Internal Animal Care and Use Committee. No pre-specified sample size was required, and 5 or 10 mice per condition were used. Mice were maintained under specific-pathogen-free conditions, and food and water were provided ad libitum. Mice (Hsd:athymic nude-*Foxn1*^{nu}, abbreviated *Nu/Nu*) were purchased from Envigo (stock no. 069). For experiments involving orthotopic transplantation of *Kras*^{G12D/+}*Trp53*^{-/-} lung adenocarcinoma cells, 100,000 cells stably transduced with either empty vector (pMIG-empty) or *Trp53*-mutant cDNAs (pMIG-p53-R277T, pMIG-p53-E284D or pMIG-p53-R277T-E284D) were resuspended in 200 µl of PBS and injected into the tail vein of 6–8-week-old *Nu/Nu* female mice. These stable cell populations were generated and FACS-purified as described above, and injected at 120 h after transduction.

TERT promoter mutation analysis and validation

TERT promoter mutations present in five or more patients, accounting for multiple samples per patient, were assessed for co-occurrence and

mutual exclusivity among composite mutations via two-sided Fisher's exact test. A pair of somatic mutations with $P < 0.01$ were considered co-occurring (or mutually exclusive) if their log-odds ratio was greater (or less) than zero. To predict the affinity for GABPA to bind TERT promoter mutant alleles, 31-bp DNA sequences for wild-type or mutant TERT centred on each of 205G>A (that is, chromosome 5, 1295205G>A), 228G>A and 250G>A mutations were extracted and generated by editing the appropriate base. The position frequency matrix for GABPA binding profiles in humans was acquired from JASPAR2018⁵⁰ (Matrix identifier MA0062.1), and scores quantifying the predicted affinity of GABPA for each TERT promoter sequence were calculated using TFBSTools⁵¹. Only binding site motifs overlapping the relevant locus in each of the wild-type and mutant sequence were retained. P values quantifying the likelihood of a GABPA binding site in each sequence to arise by chance were calculated using TFMPvalue⁵².

To assess the effect of TERT promoter composite mutations on TERT expression, A375, SK-Mel2 and SK-Mel30 melanoma cell lines were obtained (kindly provided by laboratories of N. Rosen and T. Merg-houb). pGL4.0-TERT wild type, G228A and G250A plasmids were provided by the J. Costello laboratory (Addgene plasmids no. 84924, 84926 and 84925)²¹. pGL4.0-TERT G205A, G205A/G228A and G205A/G250A plasmids were generated using Q5 Site-Directed mutagenesis kit (NEB, E0554S). All plasmids were verified using Sanger sequencing. Thereafter, 1×10^4 cells from A375, SK-Mel2, and SK-Mel30 were seeded into each well of 96-well plates. Cells were transiently transfected with pGL4.0-empty vector (Promega), TERT wild type or mutant plasmids (180 ng per well) along with pGL4.74[hRluc/TK] vector (18 ng per well, Promega) as an internal control using Lipofectamine 3000 (Thermo Fisher). Dual luciferase activity measurement was performed 48 h after transfection using the Dual-Luciferase Reporter Assay System (Promega) following the manufacturer's instructions. The firefly luciferase activity of individual wells was normalized relative to Renilla luciferase activity. Experiments were performed in biological quadruplicates or pentaplicates. To quantify the effect of a specific TERT variant, we compared individual genotypes (for example, TERT^{G205A} to wild type) using linear models of luciferase expression, in which we controlled for the baseline telomerase expression of each cell line—that is, luc - variant + cell line + constant, in which variant is a binary term that encodes the presence or absence of a genotype (relative to the chosen reference), and cell line is a factor introduced to control for the contribution of the baseline expression of each cell line. All cell lines used for either the TERT or TP53 functional validation experiments were authenticated by short-tandem-repeat analysis and confirmed negative for mycoplasma.

Statistical analyses and figures

All statistical analyses were performed using the R statistical programming environment (version 3.5.0). Figures were generated using either base R or the ggplot2 library. Error bars indicate the 95% binomial confidence intervals calculated using the Pearson–Klopper method, unless otherwise noted. Confidence intervals for the down-sampling analysis were calculated using the loess.sd function from the msir library. P values for the difference in proportions were calculated using Fisher's exact test or two-sample Z -tests, unless otherwise noted. P values were corrected for multiple comparisons using the Benjamini–Hochberg method and reported as Q values when applicable.

Reporting summary

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

Data availability

All mutational data from the prospective sequencing cohort are available at http://download.cbioportal.org/composite_mutations_maf.txt.gz. Mutational data from The Cancer Genome Atlas were acquired

from <https://gdc.cancer.gov/about-data/publications/pancanatlas>. RNA sequencing data have been deposited in the Gene Expression Omnibus with accession number GSE136295. All other genomic and clinical data accompany the Article, and are available in the Extended Data and Supplementary Information. All other materials are available upon request from the corresponding authors.

Code availability

Source code for these analyses is available at <https://github.com/taylor-lab/composite-mutations>.

25. Hyman, D. M. et al. Precision medicine at Memorial Sloan Kettering Cancer Center: clinical next-generation sequencing enabling next-generation targeted therapy trials. *Drug Discov. Today* **20**, 1422–1428 (2015).
26. Cheng, D. T. et al. Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT): a hybridization capture-based next-generation sequencing clinical assay for solid tumor molecular oncology. *J. Mol. Diagn.* **17**, 251–264 (2015).
27. Zehir, A. et al. Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. *Nat. Med.* **23**, 703–713 (2017).
28. Chakravarty, D. et al. OncoKB: a precision oncology knowledge base. *JCO Precis. Oncol.* **1**, 1–16 (2017).
29. Campbell, B. B. et al. Comprehensive analysis of hypermutation in human cancer. *Cell* **171**, 1042–1056.e10 (2017).
30. Niu, B. et al. MSIsensor: microsatellite instability detection using paired tumor-normal sequence data. *Bioinformatics* **30**, 1015–1016 (2014).
31. Middha, S. et al. Reliable pan-cancer microsatellite instability assessment by using targeted next-generation sequencing data. *JCO Precis. Oncol.* **1**, 1–17 (2017).
32. Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J. & Stratton, M. R. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep.* **3**, 246–259 (2013).
33. Dixon, P. VEGAN, a package of R functions for community ecology. *J. Veg. Sci.* **14**, 927–930 (2003).
34. Smedley, D. et al. The BioMart community portal: an innovative alternative to large, centralized data repositories. *Nucleic Acids Res.* **43**, W589–W598 (2015).
35. Forbes, S. A. et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res.* **43**, D805–D811 (2015).
36. Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
37. Alexandrov, L. et al. The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
38. Pich, O. et al. Somatic and germline mutation periodicity follow the orientation of the DNA minor groove around nucleosomes. *Cell* **175**, 1074–1087.e18 (2018).
39. Sabarinathan, R., Mularoni, L., Deu-Pons, J., Gonzalez-Perez, A. & López-Bigas, N. Nucleotide excision repair is impaired by binding of transcription factors to DNA. *Nature* **532**, 264–267 (2016).
40. Buisson, R. et al. Passenger hotspot mutations in cancer driven by APOBEC3A and mesoscale genomic features. *Science* **364**, eaaw2872 (2019).
41. Hess, J. M. et al. Passenger hotspot mutations in cancer. *Cancer Cell* **36**, 288–301.e14 (2019).
42. Needleman, S. B. & Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443–453 (1970).
43. McGranahan, N. et al. Clonal status of actionable driver events and the timing of mutational processes in cancer evolution. *Sci. Transl. Med.* **7**, 283ra54 (2015).
44. Dimitrova, N. et al. Stromal expression of miR-143/145 promotes neoangiogenesis in lung cancer development. *Cancer Discov.* **6**, 188–201 (2016).
45. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
46. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
47. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
48. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
49. Bult, C. J., Blake, J. A., Smith, C. L., Kadin, J. A. & Richardson, J. E. Mouse genome database (MGD) 2019. *Nucleic Acids Res.* **47**, D801–D806 (2019).
50. Khan, A. et al. JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res.* **46**, D260–D266 (2018).
51. Tan, G. & Lenhard, B. TFBSTools: an R/bioconductor package for transcription factor binding site analysis. *Bioinformatics* **32**, 1555–1556 (2016).
52. Touzet, H. & Varré, J.-S. Efficient and accurate P-value computation for position weight matrices. *Algorithms Mol. Biol.* **2**, 15 (2007).
53. Supek, F. & Lehner, B. Clustered mutation signatures reveal that error-prone DNA repair targets mutations to active genes. *Cell* **170**, 534–547.e23 (2017).
54. Nik-Zainal, S. et al. Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979–993 (2012).

Acknowledgements We thank the members of the E.R. and B.S.T. laboratories for discussion and support. This work was supported by National Institutes of Health awards P30 CA008748, P01 CA087497 (S.W.L.), U54 OD020355 (S.W.L. and B.S.T.), R01 CA207244 (B.S.T.), R01 CA204749 (B.S.T.), R01 CA245069 (B.S.T.); Brown Performance Group ICI Fund (N.V. and E.R.), Society of MSK (N.V. and E.R.), American Cancer Society, Anna Fuller Fund and the Josie Robertson Foundation (B.S.T.). F.J.S.-R. is an HHMI Hanna Gray Fellow supported in part by an MSKCC Translational Research Oncology Training Fellowship (T32-CA160001). S.W.L. is an investigator of the Howard Hughes Medical Institute.

Author contributions A.N.G., E.R. and B.S.T. conceived the study. C.M.B., E.B., P.J., A.V.P., A.L.R., N.D.F., C.B., N.S., E.R. and B.S.T. assisted with genomic data collection and analytical methodology development. F.J.S.-R., Y.C., N.V., M.S. and S.W.L. designed and performed the experiments. Y.J.H. and T.B. assisted with RNA sequencing. A.N.G., E.R. and B.S.T. wrote the manuscript with input from all authors.

Competing interests N.V. reports advisory board activities for Novartis and consulting activities for Petra Pharmaceuticals. M.S. has received research funding from Puma Biotechnology, Daiichi-Sankio, Immunomedics, Targimmune and Menarini Ricerche; is a cofounder of Medendi.org, and is on the advisory boards of the Bioscience Institute and Menarini Ricerche. S.W.L. is a founder and scientific advisory board member of Oric Pharmaceuticals, Mirimus, Inc. and Blueprint Medicines; and is on the scientific advisory boards of Constellation Pharmaceuticals, Petra Pharmaceuticals and PMV Pharmaceuticals. B.S.T. reports receiving honoraria and research funding from Genentech and Illumina, and advisory board activities for Boehringer Ingelheim and Loxo Oncology, a wholly owned subsidiary of Eli Lilly, Inc. All stated activities were outside of the work described here. The other authors declare no competing interests.

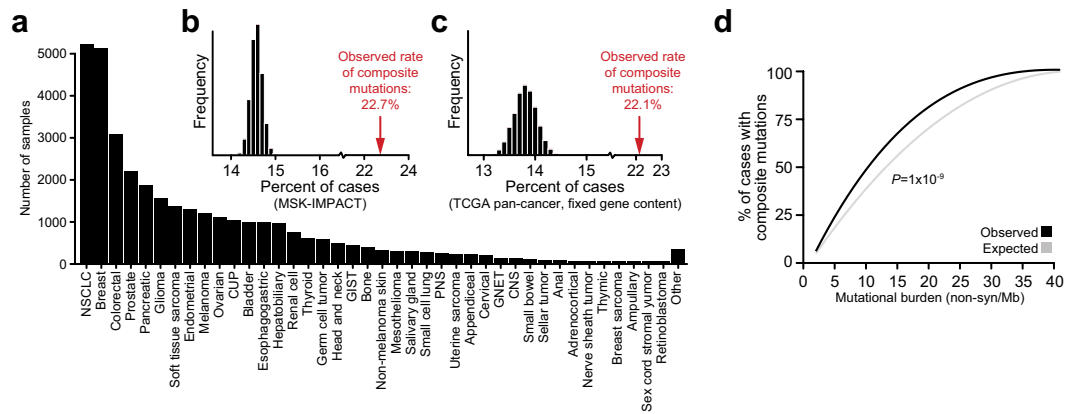
Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-020-2315-8>.

Correspondence and requests for materials should be addressed to E.R. or B.S.T.

Peer review information Nature thanks Moritz Gerstung, Mark Lackner and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

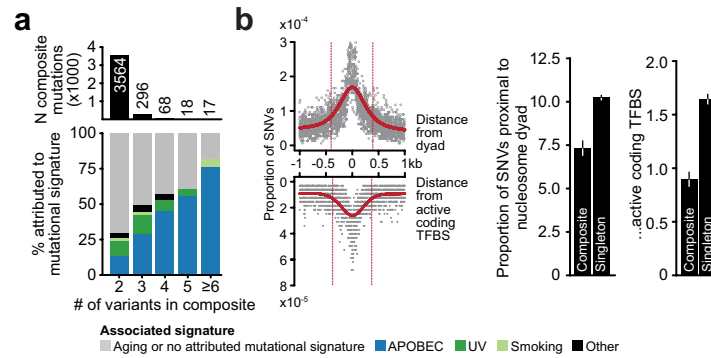
Reprints and permissions information is available at <http://www.nature.com/reprints>.



Extended Data Fig. 1 | Study cohort and rates of composite mutations.

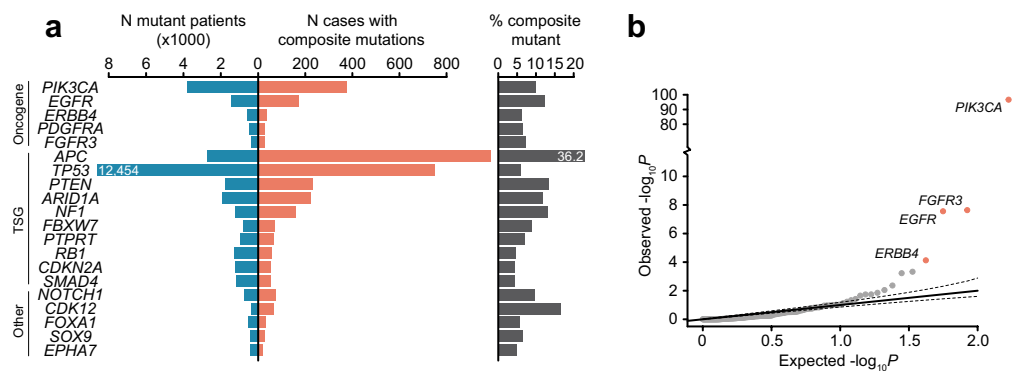
a, Distribution of cancer types in the study cohort. **b**, The rate of composite mutations (22.7% of all tumours) compared to a simulated background rate (black, $P=10^{-5}$ from one-sided permutation test for enrichment with 100,000 random permutation-based simulations (no permutation exceeded observed value)). **c**, The observed rate of composite mutations in the primary untreated cancers of the TCGA cohort ($n=10,908$ solid tumours) when

controlling for gene content for consistency with the targeted sequencing panel of the prospective cohort studied here. The null distribution from sampling (Methods) is shown in black. **d**, The observed and expected rate of composite mutations in tumours of the indicated tumour mutational burden (as in Fig. 1b, $n=30,505$ biologically independent tumour samples with tumour mutational burden ≤ 40 , $P=1 \times 10^{-9}$ from two-sided Wilcoxon signed-rank test).



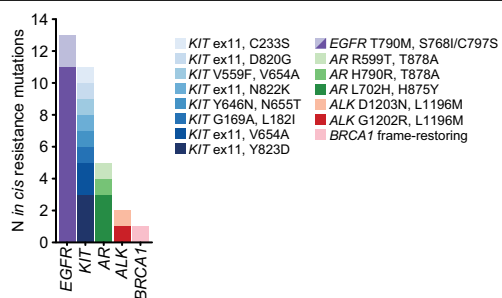
Extended Data Fig. 2 | Sources of local hypermutation. **a**, The number of composite mutations comprising two or more constituent variants (top) and the distribution of likely causative mutational signatures among them (bottom). Composite mutants comprising greater than three mutations were increasingly produced by APOBEC-associated mutagenesis, indicative of localized hypermutation^{53,54}, but accounted for a minority of events cohort-wide. **b**, Left, the somatic mutational data in the study cohort reflect the elevated mutation rates previously observed at both the positions closest to

the nucleosome dyad as well as DNA bound to active transcription-factor binding sites^{38,39}. However, mutations arising in composite events were proportionally less often proximal to such sites (defined here as within the full width at half maximum of the peak of mutation rate (red)) than were singleton mutations (right, $P=10^{-27}$ and 10^{-47} , respectively; two-sided two-sample Z-test, $n=323,883$ single-nucleotide substitutions arising in 471 biologically distinct melanoma samples).

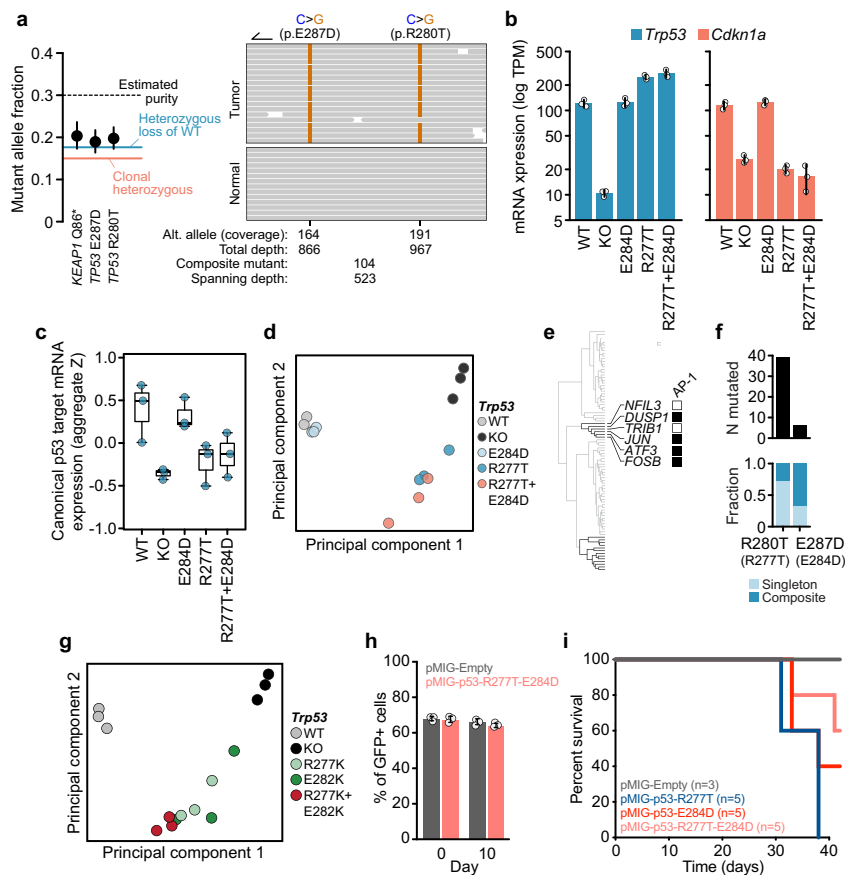


Extended Data Fig. 3 | Number and distribution of composite events across genes. **a**, The number and percentage of cases in the study cohort containing composite mutations in the indicated genes (right) juxtaposed to their overall mutation rate (left). Genes with a significant enrichment of composite mutations are shown ($Q < 0.01$, FDR-adjusted P values from one-sided binomial

test for enrichment, $n = 26,997$ as in Fig. 2b), limited to the top 10 genes by significance in each category of gene function, unless fewer. **b**, The significance of enrichment for composite mutations (n and statistical tests as described in **a** and Fig. 2b) limited to 168 oncogenes.

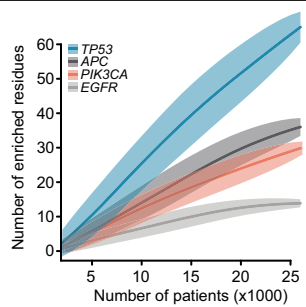


Extended Data Fig. 4 | cis composite secondary-resistance mutations. The cis composite mutations classified as arising in post-treatment specimens due to acquired resistance to one of several molecularly targeted therapies in the study cohort.

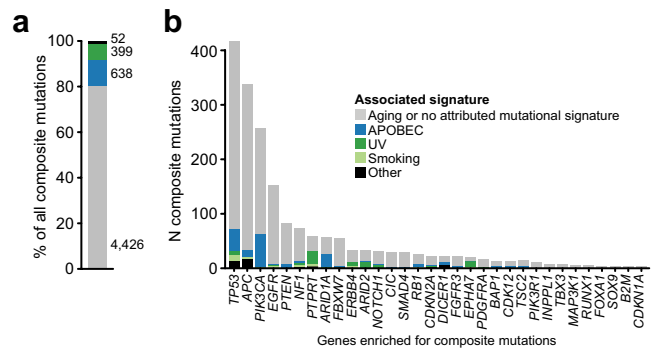


Extended Data Fig. 5 | Phenotypic characterization of *TP53* composite mutants. **a**, *TP53*^{R280T/E287D} mutant lung adenocarcinoma. Left, mutant allele fractions of clonal *TP53* mutations consistent with loss of wild-type *TP53* (error bars, 95% binomial confidence intervals). Expected mutant allele fractions of different copy number states are shown as horizontal lines. Mutant *KEAP1* in the same tumour (with LOH) is shown for reference. Right, spanning reads indicating *cis* mutations. **b**, Right and left, *Trp53* and *Cdkn1a* mRNA expression in *Kras*^{G12D/+}*Trp53*^{Mut} mouse lung cancer cells expressing distinct *Trp53* genotypes. Bars, average of three replicates, error bars are 95% confidence intervals. **c**, The aggregate Z-score per replicate for the mRNA expression of canonical p53-target genes ($n = 3$ replicates per allele; box centre is median, edges are 25% and 75% quartiles, whiskers are minimum and maximum of the most extreme values). **d**, Principal component analysis of the transcriptomes of *Trp53* genotypes ($n = 3$ replicates shown per condition). **e**, Dendrogram as in Fig. 3f, indicating the genes of interest (effectors of the AP-1 transcription

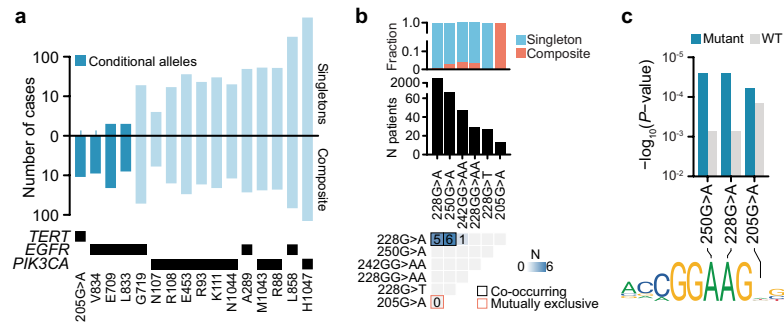
factor network (PID_API_PATHWAY; $Q = 1.4 \times 10^{-7}$ based on computed overlap (using mSigDB) with $n = 5,501$ gene sets from the curated C2 collection)). **f**, The prevalence of *TP53*^{R280T} and *TP53*^{E287D} mutations (top), and the fraction arising as composite mutants (bottom). The corresponding mouse alleles are given in parentheses. **g**, Principal component analysis of the transcriptomes of the *Trp53*^{R277T/E282K} composite mutation genotypes (as in **d**). $n = 3$ replicates per allele. **h**, The percentage of GFP⁺ FACS-purified *Kras*^{G12D/+}*Trp53*^{-/-} lung adenocarcinoma cells stably transduced with pMIG-empty or pMIG-p53-R277T-E284D, and cultured in vitro for 10 days in a 60:40 mixture with untransduced parental cells. Bar indicates mean, error bars are s.d., $n = 3$ independent infections. **i**, Overall survival of immunocompromised mice bearing lung tumours of the indicated *Trp53* genotypes generated by tail vein injection of stably transduced and FACS-purified *Kras*^{G12D/+}*Trp53*^{-/-} lung adenocarcinoma cells ($n = 100,000$ cells).



Extended Data Fig. 6 | Saturation analysis of genes for composite mutation detection. Down-sampling indicates the number of residues identified as enriched for arising in composite mutations in each of four genes ($Q < 0.1$, FDR-adjusted one-sided Fisher's exact tests as in Fig. 4a; $n = 1,000 - 26,997$ patients per down-sample) as a function of the number of tumours sequenced (LOESS fit is shown with 95% confidence interval). Four genes that accounted for the greatest proportion of all enriched residues detected are shown (Fig. 4a). *EGFR* appears to reach saturation for discovery of residues enriched for arising in composite, whereas the other genes have not yet reached saturation for discovery at the current cohort size.



Extended Data Fig. 7 | Mutational signature attribution among composite mutations. **a**, The fraction of all composite mutations identified here in which one or both individual mutations could be unambiguously attributed to an established mutational signature. The majority of composite variants could not be directly attributed to APOBEC, ultraviolet, smoking or other known mutational signatures. **b**, The fraction of composite mutations per gene in which one or both variants could be attributed to an established mutational signature.



Extended Data Fig. 8 | Conditional mutant alleles. **a**, The number of affected cases containing each of the indicated somatic mutations in *TERT*, *EGFR* or *PIK3CA* as either individual mutations (top) or as part of composite mutants (bottom). Conditional mutations were defined as those statistically enriched for arising as part of composite mutations, but seldom as individual hotspot mutations in cancer (predominantly accompanied by a second somatic mutation). **b**, The incidence of *TERT* promoter mutations and the fraction

arising as composite mutations (orange). Bottom, the co-occurrence and mutual exclusivity of composite mutations in the *TERT* promoter (The *P* values for $n = 5$ and 6 co-occurring mutations are 0.002 and 3×10^{-7} , respectively, and for 0 mutually exclusive mutations is 1×10^{-25} ; two-sided Fisher's exact test, $n = 29,507$ patients). **c**, Transcription factor GABPA binding affinity for mutant and wild-type *TERT* promoter sequences at the 228G>A, 250G>A and the conditional 205G>A allele.

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

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

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection	R Statistical Computing environment (v3.5.0)
Data analysis	FACETS (v0.5.6); maf2maf (v1.6.17); MSISensor (v0.2); OncoKB (v1.0.4); R (v3.5.0); R packages: binom (v1.1-1), BiomaRt (v2.36.1), car (v3.0-3), cowplot (v1.0.0), data.table (v1.12.2), DESeq2 (v1.22.2), MASS (v7.3-51.4), Rcpp (v1.0.2), TFBSTools (v1.20.0), TFMPvalue (v0.0.8), vegan (v2.5-6); custom phasing software (https://github.com/taylor-lab/MutationPhaser); mutation signature decomposition (https://github.com/mskcc/mutation-signatures); featureCounts (v1.6.3); Trimmomatic (v0.36); STAR (2.5.3a); Human genome reference (GRCh37); Mouse genome reference (GRCm38.p5, mm10); Source code for these analyses is available at http://github.com/taylor-lab/composite-mutations .

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

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All mutational data from the prospective sequencing cohort is available through the cBioPortal for Cancer Genomics: http://download.cbioportal.org/composite_mutations_maf.txt.gz. Mutational data from The Cancer Genome Atlas was acquired from <https://gdc.cancer.gov/about-data/publications/pancanatlas>. RNA sequencing data were deposited in the GEO with accession number GSE136295. All other genomic and clinical data accompanies the manuscript and is available as Extended Data and Supplementary Information.

Field-specific reporting

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

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

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

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

Sample size	Clinical sequencing data was comprised of 34,650 tumor and matched normal specimens from 31,359 patients prospectively characterized as part of their active care at Memorial Sloan Kettering Cancer Center (MSKCC) between Jan. 2014 and Apr. 2019. Sequencing data from 10,908 primary untreated cancers of The Cancer Genome Atlas cohort were including for comparative frequency analyses (data acquired from https://gdc.cancer.gov/about-data/publications/pancanatlas).
Data exclusions	No exclusion criteria other than including solid tumors were specified for the study population
Replication	Experimental replication was performed as described in the Methods section, which included 3 to 5 replicates per condition, and all attempts at replication were successful.
Randomization	Data were randomized for permutation-based statistical testing as described in the Methods section. No other randomized allocation among groups was performed, and all further allocation was based on stated variables and conditions.
Blinding	Blinding was not applicable for this study/analytical design.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input type="checkbox"/>	<input checked="" type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input type="checkbox"/>	<input checked="" type="checkbox"/> Animals and other organisms
<input type="checkbox"/>	<input checked="" type="checkbox"/> Human research participants
<input type="checkbox"/>	<input checked="" type="checkbox"/> Clinical data

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Eukaryotic cell lines

Policy information about [cell lines](#)

Cell line source(s)	Cell lines utilized here included: A375, Sk-Mel2, Sk-Mel30 (kindly provided by the N. Rosen and T. Merghoub laboratories at MSK), HEK293T (obtained from ATCC, CRL-3216), and murine KP lung adenocarcinoma cells (Kras G12D/+, Trp53-/-; provided by the T. Jacks laboratory, MIT).
Authentication	All cell lines have been authenticated by short tandem repeat analysis.
Mycoplasma contamination	All cell lines were confirmed tested negative for mycoplasma contamination
Commonly misidentified lines (See ICLAC register)	No commonly mis-identified cell lines were utilized.

Animals and other organisms

Policy information about [studies involving animals](#); [ARRIVE guidelines](#) recommended for reporting animal research

Laboratory animals	Mice used in experiments were Hsd:ATHymic Nude-Foxn1nu strain purchased from Envigo (stock #069), 6-8 weeks old, female.
Wild animals	No wild animals were used in this study.

Field-collected samples

No field-collected samples were used in this study.

Ethics oversight

All mouse experiments were approved by the Memorial Sloan-Kettering Cancer Center (MSKCC) Internal Animal Care and Use Committee

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Human research participants

Policy information about [studies involving human research participants](#)

Population characteristics

Age at time of sequencing: median 61.7 years
Male/female: 46.4%/53.6%
Additional details in Supplementary Table 1.

Recruitment

Passive recruitment were for patients who underwent prospective sequencing as part of their active clinical care at Memorial Sloan Kettering Cancer Center (MSKCC) from January 2014 to April 2019. All such patients whose tumor sequencing was performed with a matched normal sample were included and biases include only those related to the demographic composition of the catchment area for cancer patients at MSKCC.

Ethics oversight

MSKCC Institutional Review Board

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about [clinical studies](#)

All manuscripts should comply with the ICMJE [guidelines for publication of clinical research](#) and a completed [CONSORT checklist](#) must be included with all submissions.

Clinical trial registration

NCT01775072

Study protocol

Details available at ClinicalTrials.gov #NCT01775072 or upon request.

Data collection

Locale of data collection: Memorial Sloan Kettering Cancer Center and affiliate sites. Dates of recruitment for prospectively characterized patients utilized here were from January 2014 to April 2019.

Outcomes

Primary and secondary outcome measures not assessed as part of the present study.