Article

Landscape and function of multiple mutations within individual oncogenes

https://doi.org/10.1038/s41586-020-2175-2

Received: 23 July 2019

Accepted: 13 February 2020

Published online: 08 April 2020

Check for updates

Yuki Saito[1,2,10], Junji Koya[1,10], Mitsugu Araki[3], Yasunori Kogure[1], Sumito Shingaki[1], Mariko Tabata[1,4], Marni B. McClure[1], Kota Yoshifuji[1,5], Shigeyuki Matsumoto[6], Yuta Isaka[7], Hiroko Tanaka[8], Takanori Kanai[2], Satoru Miyano[8], Yuichi Shiraishi[9], Yasushi Okuno[3] & Keisuke Kataoka[1]✉

Sporadic reports have described cancer cases in which multiple driver mutations (MMs) occur in the same oncogene[1,2]. However, the overall landscape and relevance of MMs remain elusive. Here we carried out a pan-cancer analysis of 60,954 cancer samples, and identified 14 pan-cancer and 6 cancer-type-specific oncogenes in which MMs occur more frequently than expected: 9% of samples with at least one mutation in these genes harboured MMs. In various oncogenes, MMs are preferentially present in cis and show markedly different mutational patterns compared with single mutations in terms of type (missense mutations versus in-frame indels), position and amino-acid substitution, suggesting a cis-acting effect on mutational selection. MMs show an overrepresentation of functionally weak, infrequent mutations, which confer enhanced oncogenicity in combination. Cells with MMs in the *PIK3CA* and *NOTCH1* genes exhibit stronger dependencies on the mutated genes themselves, enhanced downstream signalling activation and/or greater sensitivity to inhibitory drugs than those with single mutations. Together oncogenic MMs are a relatively common driver event, providing the underlying mechanism for clonal selection of suboptimal mutations that are individually rare but collectively account for a substantial proportion of oncogenic mutations.

The advent of next-generation sequencing has enabled the processing of tens of thousands of tumours of many types for the systematic discovery of driver alterations[3,4]. These efforts have identified thousands of recurrent somatic mutations across cancers, with few highly frequent (major) mutations and a much larger number of rare (minor) mutations. Although the latter account for a substantial portion of the accumulated mutations, even for oncogenes, the vast majority of them are regarded as functionally weak mutations or of uncertain significance; however, how these minor mutations are clonally selected despite limited functionality and what their genetic differences are from major mutations are poorly understood.

Tumour suppressor genes (TSGs) are frequently affected by multiple (typically biallelic) loss-of-function mutations[3,4]. There have also been many cancer cases in which specific secondary mutations—including the *EGFR* T790M and *KIT* V654A mutations—have been acquired in already mutated oncogenes following tyrosine kinase inhibitor (TKI) therapy[5–8]. However, only sporadic reports have investigated MMs arising in the same oncogene during cancer initiation and development, before any therapy[1,2]. Therefore, the frequency, spectrum and genetic features of MMs across oncogenes remain unclear. In addition, whether oncogenic MMs occur as a combination of driver–driver or driver–passenger mutations, what their biological and clinical implications are, and how they differ from single mutations have not been well studied. Here we present a systematic pan-cancer analysis of more than 60,000 samples and delineate the overall landscape and genetic properties of oncogenic MMs. We explore their functionality and influences on molecular machinery and drug sensitivity, as well as potential synergistic interactions (the epistatic effect) between mutations, focusing on *PIK3CA* and *NOTCH1*.

Frequent MMs in a variety of oncogenes

We assembled, annotated and systematically curated a large repository of cancer genome data from 60,954 samples of primary and metastatic cancer across more than 150 cancer types from five cohorts: The Cancer Genome Atlas (TCGA), Therapeutically Applicable Research to Generate Effective Treatments (TARGET), Foundation Medicine (FM), the American Association for Cancer Research (AACR) Project Genomics Evidence Neoplasia Information Exchange (GENIE), and the haematological malignancy (HM) cohort, comprising acute myeloid leukaemia (AML) and non-Hodgkin's lymphoma (NHL) (Extended Data Figs. 1, 2a, b and Supplementary Table 1; see Methods). In particular, to eliminate probable false-positive MMs, we rigorously reviewed certain variant types (Extended Data Fig. 2c–e), namely dinucleotide and trinucleotide

[1]Division of Molecular Oncology, National Cancer Center Research Institute, Tokyo, Japan. [2]Department of Gastroenterology, Keio University School of Medicine, Tokyo, Japan. [3]Department of Clinical System Onco-Informatics, Graduate School of Medicine, Kyoto University, Kyoto, Japan. [4]Department of Urology, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan. [5]Department of Hematology, Graduate School of Medical and Dental Sciences, Tokyo Medical and Dental University, Tokyo, Japan. [6]Medical Sciences Innovation Hub Program, RIKEN Cluster for Science, Technology and Innovation Hub, Yokohama, Japan. [7]Research and Development Group for In Silico Drug Discovery, Center for Cluster Development and Coordination, Foundation for Biomedical Research and Innovation, Kobe, Japan. [8]Laboratory of Sequence Analysis, Human Genome Center, Institute of Medical Science, The University of Tokyo, Tokyo, Japan. [9]Center for Cancer Genomics and Advanced Therapeutics, National Cancer Center, Tokyo, Japan. [10]These authors contributed equally: Yuki Saito, Junji Koya. ✉e-mail: kkataoka-tky@umin.ac.jp

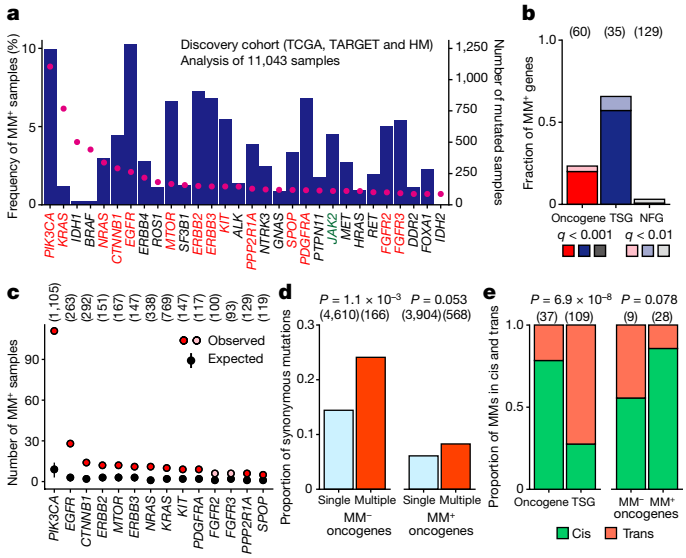


Fig. 1 | Entire landscape of MMs in various oncogenes. a, Number of mutated samples (dots) and frequency of MM$^+$ samples (bars) for 30 oncogenes (with 80 or more mutated samples) in the discovery cohort (n = 11,043). Oncogenes significantly affected by MMs (q < 0.01 and three or more MMs) in pan-cancer and cancer-type-specific analyses are indicated in red and green, respectively. **b**, Fraction of MM$^+$ genes in oncogenes, TSGs and NFGs. Numbers of genes examined are shown in parentheses. **c**, Number of MM$^+$ samples for 14 MM$^+$ oncogenes (with q < 0.01 and three or more MMs). Red (q < 0.001) and pink (q < 0.01) circles show observed values; black circles indicate expected values (median with 95% confidence intervals). Values in parentheses indicate numbers of samples examined. **b**, **c**, One-sided permutation test (n = 10,000) with Benjamini–Hochberg correction. **d**, Proportion of synonymous to total mutations according to MM status (single or multiple) in MM$^-$ and MM$^+$ oncogenes. Values in parentheses indicate numbers of mutations examined. **e**, Proportion of MMs in cis and trans (with distances between mutations of 25 base pairs (bp) or more) by phasing from RNA-seq or WES/WGS in oncogenes and TSGs (left), and in MM$^-$ and MM$^+$ oncogenes (right). Numbers in parentheses indicate numbers of MMs examined. **d**, **e**, Two-sided Fisher's exact test.

variants (DNVs/TNVs)—single genetic events affecting multiple bases, which are usually misannotated as multiple single-nucleotide variants (SNVs) with different amino-acid substitutions, and which frequently affect certain mutational hotspots, such as *KRAS* amino-acid residue G12 and *BRAF* V600; and SNVs around insertions and deletions (indels), which are also often miscalled owing to ambiguous mapping.

First, we focused on whole-exome/whole-genome sequencing (WES/WGS) data of 11,043 primary untreated samples from TCGA, TARGET and HM (together the discovery cohort), and evaluated the frequency of MMs in 60 well-described oncogenes (Extended Data Figs. 1a, 2f and Supplementary Table 2). Although several highly mutated oncogenes, such as *IDH1* and *BRAF*, contained no or few MMs, MMs were frequently observed across a wide variety of oncogenes; 5% or more of the mutated samples carried MMs across nine oncogenes, particularly in *PIK3CA* (10% of samples) and *EGFR* (10%), suggesting that MMs are a relatively common phenomenon that targets various oncogenes (Fig. 1a). MMs in oncogenes were not attributed to hypermutator tumours, and their frequency was consistent across cohorts and samples with varying tumour purity (Extended Data Fig. 3a–c).

MMs in oncogenes may arise as a pair of driver–driver, driver–passenger, or passenger–passenger mutations. To identify genes recurrently affected by driver–driver MMs (q < 0.01), we modified the permutation framework used to detect significantly mutated genes, while accounting for mutational signature, expression and DNA replication time[9] (Extended Data Fig. 3d–f). In contrast with nonfunctional genes (NFGs), of which only 3% were significant, MMs were found frequently in TSGs (66% of the analysed genes). Intriguingly, the significant enrichment of putative driver MMs was observed in 14 (23%) oncogenes (Fig. 1b, c and Supplementary Table 2). These genes ('MM$^+$ oncogenes') included components of the phosphatidylinositol-3-kinase (PI3K) pathway (namely *PIK3CA*, *MTOR* and *PPP2R1A*), members of the RAS family of GTPases (*KRAS* and *NRAS*), and receptor tyrosine kinases (*EGFR*, *ERBB2* and so on) (Extended Data Fig. 3g).

The proportion of synonymous to total mutations indicates the strength and mode of natural selection[10]. Regardless of mutation status (single versus multiple), this proportion was low in TSGs and relatively high in NFGs, suggestive of positive and neutral selection, respectively (Extended Data Fig. 3h). For MM$^-$ oncogenes, the proportion of synonymous mutations was low in samples with single mutations, but significantly increased in MM$^+$ samples (namely those with MMs in the same oncogene), suggesting weakened selective pressure for secondary mutations (that is, enrichment of driver–passenger MMs) in these genes (Fig. 1d and Extended Data Fig. 3i). By contrast, in MM$^+$ oncogenes, the proportion of synonymous mutations remained low (less than 0.1) even in MM$^+$ samples, in which both of the mutations in each MM are considered to undergo positive selection equal to the positive selection on single mutations, suggesting that most MMs arising in these genes are a combination of putative driver–driver mutations.

We investigated the allelic configuration of MMs by phasing from RNA sequencing (RNA-seq) or WES/WGS reads, which revealed that most MMs (78%) in oncogenes were present in cis with concordant allele frequencies, while the majority of MMs (72%) in TSGs were in trans (Fig. 1e, Extended Data Fig. 4a–c and Supplementary Table 3). Although several oncogenes were frequently affected by copy-number amplifications, MMs arose in cis even with concurrent amplification (Extended Data Fig. 4d, e). Notably, the proportion of MMs in cis was particularly high (86%) in MM$^+$ oncogenes, which was validated by Sanger sequencing of complementary DNA (cDNA) and long-read WGS in cell lines (Fig. 1e, Extended Data Fig. 4f–k).

Selection pattern of MMs in *PIK3CA*

The presence of MMs in cis raised the possibility that an initial mutation may influence the clonal selection of subsequent mutations in MM$^+$ oncogenes, as they cause structural changes within the same molecule. Therefore, with the combined data of 40,002 primary samples from all five cohorts, we evaluated the mutational pattern of MMs in these genes (Extended Data Figs. 1a, 2f, 3c and Supplementary Table 4). To minimize the effect of confounding passenger mutations, we concentrated on the established hotspot and/or functionally relevant positions (1,679 positions for 14 MM$^+$ oncogenes) defined by knowledge-based databases and/or computational algorithms (Supplementary Tables 5–7; see Methods). Then, we classified them according to mutation type, position and amino-acid substitution, and compared their fraction of MM$^+$ mutations (namely those found in MM$^+$ samples) (Extended Data Fig. 1b).

We first examined *PIK3CA*, the most common oncogene in various cancers, which contained the largest number of MMs. In *PIK3CA*, MMs were more prevalent in missense mutations than in in-frame indels, and their frequency was quite different by mutational position (Fig. 2a, b and Supplementary Tables 5, 6). More than half of *PIK3CA* missense mutations were located in three major hotspots (E542, E545 and H1047), while almost all other mutations were in 275 minor hotspots (Extended Data Fig. 5a). Regardless of the cohort, MMs occurred more frequently in minor (32%) than major (11%) positions, particularly at E453 and E726, where 50% or more of the mutated cases possessed another mutation (Fig. 2a and Extended Data Fig. 5b). Moreover, even in major positions, less frequent (minor) amino-acid substitutions were more likely to be selected in samples with MMs (16–22%), compared with common (major) amino-acid substitutions (E542K, E545K and H1047R) (Fig. 2c and Supplementary Table 8). Together, these results clearly depict a substantial genetic difference between single and multiple mutations in *PIK3CA*.

Among *PIK3CA* MMs, major–minor combinations—such as E542–E726 and E726–H1047—were significantly enriched (q = 0.0084), followed by

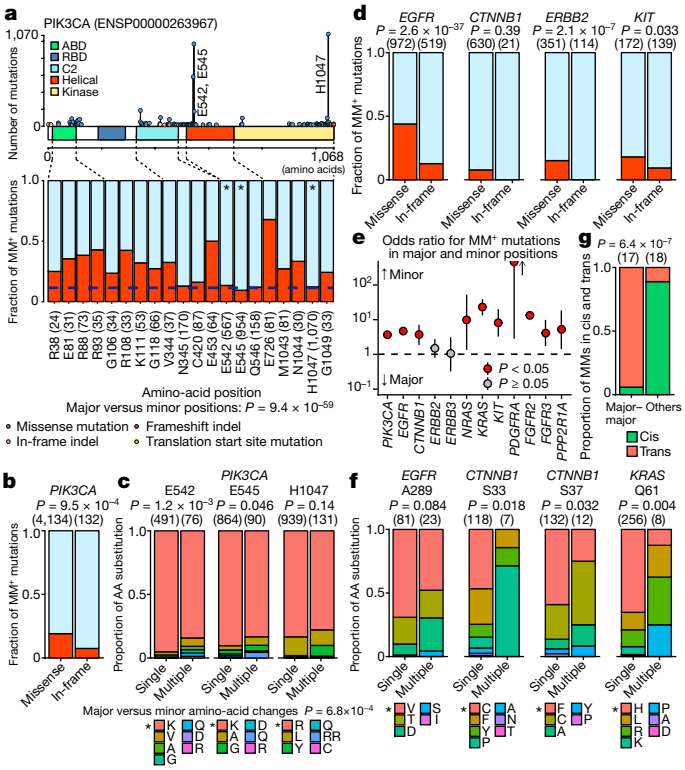


Fig. 2 | Biased selection for MMs across a wide range of oncogenes. a, Distribution of mutations (top) and fraction of MM$^+$ mutations (bottom) for each hotspot/functional position (the top 20 are shown) in *PIK3CA* in recurrently mutated cancer types (defined as those with 20 or more hotspot/functional *PIK3CA* mutations) in primary samples from the total cohort. Asterisks indicate major positions (in which 10% or more of mutations were present in any of the recurrently mutated cancer types). The horizontal blue dotted line represents the mean value of major positions. Examined numbers of each mutation are in parentheses. ENSP00000263967 is the Ensembl database (http://www.ensembl.org/) identification code for PIK3CA protein. **b**, Fraction of MM$^+$ mutations for missense mutations and in-frame indels in *PIK3CA*. **c**, Proportion of each amino-acid (AA) substitution for each major position in *PIK3CA* for samples with single and multiple mutations. Asterisks indicate major amino-acid changes (occurring in more than 33% of samples with single mutations). **d**, Fraction of MM$^+$ mutations for missense mutations and in-frame indels across oncogenes (in which 100 or more mutations were present and 2% or more of them were indels). **e**, Comparison of the fraction of MM$^+$ mutations for major and minor positions across oncogenes (with more than 3 MM$^+$ samples). The *y*-axis shows odds ratios with 95% confidence intervals. See Supplementary Table 4 for sample size. **f**, Proportion of each amino-acid substitution for each major position (in which 100 or more mutations were present and 2% or more of them were from MM$^+$ samples) across oncogenes for samples with single and multiple mutations. Asterisks indicate major amino-acid changes. **b**–**d**, **f**, Values in parentheses indicate numbers of mutations examined. **g**, Proportion of MMs in cis and trans (with distances between mutations of less than 25 bp) by phasing from RNA-seq or WES/WGS in major–major combinations and others. Numbers in parentheses indicate numbers of MMs examined. **a**–**g**, Two-sided Fisher's exact test.

minor–minor combinations ($q < 0.0001$; Extended Data Fig. 5c and Supplementary Table 9). In addition, mutational combinations involving different domains were overrepresented, particularly those involving the adaptor-binding domain (ABD) or C2 domain (such as R88–H1047 and E453–E545) (Extended Data Fig. 5d and Supplementary Table 10). Conversely, major–major combinations were substantially less common than expected, and among them those within the helical domain (E542–E545) were invariably present on different alleles (that is, in trans in the same clone or in different clones) (Extended Data Fig. 5c, e). On the basis of their allele frequencies, most of these mutational combinations were estimated to occur in the same clone (concordant allele frequency), regardless of major or minor hotspots (Extended Data Fig. 5f). In the major–minor combinations showing discordant allele frequencies, mutations at major hotspots tended to have a higher allele frequency than at minor hotspots, implying that the mutations at major hotspots are likely to be earlier events causing intermediate clonal expansions (Extended Data Fig. 5g).

To understand the mechanism that generates MMs, we investigated the mutational signature caused by the cytidine deaminase APOBEC (C-to-G/T at TpCpX trinucleotides). Reflecting the underrepresentation of APOBEC signatures in minor hotspot mutations, we found that samples harbouring *PIK3CA* MMs showed a higher overall APOBEC activity but a lower proportion of APOBEC signature mutations in *PIK3CA* itself, compared with samples harbouring no or single *PIK3CA* mutations. This suggests that a single mutational process does not account for this phenomenon (Extended Data Fig. 5h, i).

Widespread biased selection for MMs

The biased selection of *PIK3CA* MMs led us to assess the mutational spectrum of MMs in other MM$^+$ oncogenes. In MM$^+$ oncogenes, among 12,753 hotspot/functional missense mutations (87% of the total mutations), 79% resided in major hotspot positions (of which 36% showed less frequent (minor) amino-acid substitutions), whereas 21% were in minor positions (Extended Data Fig. 6a and Supplementary Table 4). Similar to the case of *PIK3CA*, missense mutations had, or tended to have, a higher fraction of MMs than did in-frame indels in other examined MM$^+$ oncogenes, most especially in *EGFR* and *ERBB2* (Fig. 2d and Supplementary Tables 5, 6). Conspicuously, missense mutations in minor positions contained a larger fraction of MMs than did those in major positions in various MM$^+$ oncogenes, except for *ERBB2*/*ERBB3* (Fig. 2e and Extended Data Fig. 6b). For instance, in *KRAS* almost no cases with major hotspot mutations (G12, G13 and Q61) harboured MMs, whereas a variable proportion of MMs was observed in minor hotspots. In *CTNNB1*, the fraction of MMs was considerably higher in minor than in major positions, even within the same functional site (β-TrCP-binding site: D32–S37). These findings were strengthened by combinatorial analysis, which showed a greater occurrence of major–minor and/or minor–minor combinations than expected (Extended Data Fig. 6c and Supplementary Table 9). Other genetic characteristics of *PIK3CA* MMs were also observed in these MM$^+$ oncogenes: minor amino-acid substitutions were more common in major hotspots in MM$^+$ oncogenes, including *EGFR* A289, *CTNNB1* S33 and S37, and *KRAS* Q61 (Fig. 2f and Supplementary Table 8); and within MMs in close proximity—in contrast to other combinations mostly present in cis—almost all major–major combinations, such as *NRAS*/*KRAS* G12–G13 pairs, occurred on different alleles (Fig. 2g and Extended Data Fig. 6d). A substantial proportion of these MMs had discordant allele frequencies, and even those with concordant allele frequencies showed lower allele frequencies, probably suggesting that they arise as different subclones (Extended Data Fig. 6e, f). Overall, our findings demonstrate a skewed mutational pattern for MMs that is pervasive across various oncogenes.

Lineage specificity of oncogenic MMs

The frequency of MMs remained consistent among recurrently mutated cancer types in several genes, such as *EGFR* and *KRAS*, whereas it varied extensively across cancer types in other genes, including *PIK3CA* and *CTNNB1*, independent of mutation burden (Fig. 3a and Extended Data Fig. 7a). Particularly, MMs in *PDGFRA* were observed in 12% of the mutated samples in glioma (glioblastoma/low-grade glioma (GBLGG)), but there was no MM$^+$ sample in gastrointestinal stromal tumour (GIST) (Extended Data Fig. 7b). In *PIK3CA*, the distribution of MMs resembled each other across cancer types (Extended Data Fig. 7c). By contrast, the distribution of MMs differed in *EGFR* between non-small-cell lung

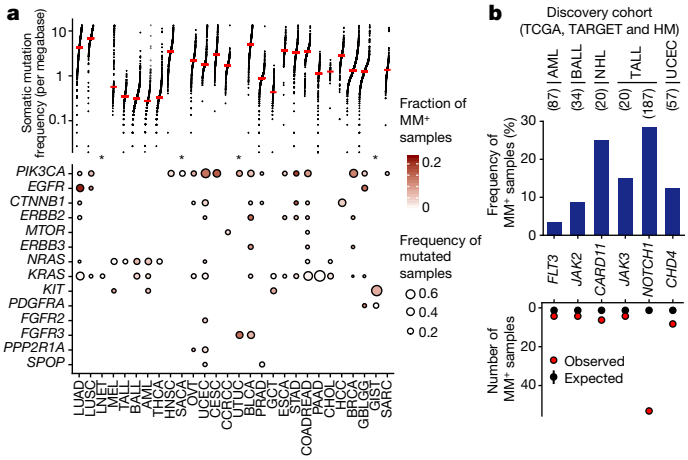


Fig. 3 | Cancer-type specificity and diversity of oncogenic MMs. a, Somatic mutation frequency per megabase calculated from WES/WGS (top), and frequency of mutated samples and fraction of MM$^+$ samples for each MM$^+$ oncogene (bottom), across human cancer types (with more than 150 samples and 20 or more hotspot/functional mutations) in primary samples from the total cohort (n = 40,002). Top, each dot represents a sample and red horizontal lines indicate the median number in each cancer type. Asterisks indicate cancer types without WES/WGS data. BRCA, invasive breast carcinoma; CCRCC, clear cell renal-cell carcinoma; CESC, cervical squamous cell carcinoma; CHOL, cholangiocarcinoma; COADREAD, colorectal adenocarcinoma; ESCA, oesophageal adenocarcinoma; GCT, germ cell tumour; HCC, hepatocellular carcinoma; HNSC, head and neck squamous cell carcinoma; LNET, lung neuroendocrine tumour; LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma; MEL, melanoma; PAAD, pancreatic adenocarcinoma; PRAD, prostate adenocarcinoma; SACA, salivary carcinoma; SARC, soft tissue sarcoma; STAD, stomach adenocarcinoma; THCA, thyroid carcinoma. **b**, Frequency of MM$^+$ samples (top) and number of MM$^+$ samples (bottom) for six cancer-type-specific MM$^+$ oncogenes (with q < 0.01 and three or more MMs) in the discovery cohort. Red and black circles indicate observed and expected (median with 95% confidence intervals) values. One-sided permutation test (n = 10,000) with Benjamini–Hochberg correction. Examined numbers are shown in parentheses.

cancer (NSCLC) and GBLGG: the fraction of MMs in major hotspots and minor positions in the kinase domain was almost equivalent between both tumours, whereas those involving minor positions in the extracellular domain were more common in NSCLC (Extended Data Fig. 7d, e), indicating substantial lineage specificity among oncogenic MMs.

In several genes, MMs were nearly exclusively identified in a certain cancer type or anatomically related ones, including *FGFR2* and *PPP2R1A* mutations in gynaecological cancers (ovarian epithelial tumour (OVT) and/or uterine corpus endometrial carcinoma (UCEC)) and *FGFR3* mutations in urothelial tumours (bladder urothelial carcinoma (BLCA) and upper tract urothelial carcinoma (UTUC)). For *FGFR3* in urothelial tumours, missense mutations within the juxtamembrane region (codons 420–450) were observed only in MM$^+$ samples (Extended Data Fig. 8a). We then performed a permutation test for each cancer type from the discovery cohort with an extended list of 84 oncogenes. This analysis identified six cancer-type-specific MM$^+$ oncogenes that had been overlooked by our pan-cancer analysis—particularly those involving haematological neoplasms, such as *NOTCH1*, *CARD11* and *JAK3* (Fig. 3b, Extended Data Fig. 8b, c and Supplementary Tables 11, 12). Among these, *NOTCH1* mutations in T-cell acute lymphoblastic leukaemia (TALL) showed the highest prevalence of MM$^+$ samples (28%), in contrast to chronic lymphocytic leukaemia (CLL), for which no MM$^+$ samples were observed (Extended Data Fig. 8d, e). In *NOTCH1* MMs, missense mutations or in-frame indels in the HD domain significantly co-occurred with truncating mutations in the PEST domain (q = 0.0050 and q < 0.0001 for HD-N and HD-C, respectively; two-sided simulation test), both of which are gain-of-function mutations[11] (Extended Data Fig. 8f). Consistent with the MM$^+$ oncogenes identified by our pan-cancer analysis, a biased mutational pattern was observed in cancer-type-specific MM$^+$ oncogenes: MMs were more common in missense mutations than in in-frame indels, and in minor than in major positions, in *FLT3* mutations in AML and/or *JAK2* mutations in B-cell acute lymphoblastic leukaemia (BALL) (Extended Data Fig. 8g, h).

We next compared 40,002 primary and 20,952 metastatic cancers, which demonstrated nearly equivalent frequencies of MMs in MM$^+$ oncogenes, except for *EGFR* and *KIT* (Extended Data Fig. 9a, b). After exclusion of previously reported TKI-resistant mutations, including *EGFR* T790M and *KIT* V654A mutations in NSCLC and GIST, respectively[5,6], the frequency of MMs in these genes was comparable between primary and metastatic samples (Extended Data Fig. 9c, d). Also, MMs in metastatic samples were enriched in missense mutations affecting minor hotspot positions, similarly to those in primary samples (Extended Data Fig. 9e–g). These findings suggest that, even when acquired TKI-resistant mutations are excluded, a biased mutational pattern for MMs is observed in metastatic samples.

Functional relevance of MMs

To clarify the biological relevance of oncogenic MMs, we evaluated the functional activity scores for PIK3CA estimated from in vitro and/or in vivo assays of mutant-transduced cell lines[12,13]. Despite their low frequency, in-frame indels had a higher score than missense mutations, suggesting a negative association between functionality and MM fraction (Fig. 4a). Among missense mutations, functional activity scores were inversely correlated with the fraction of MMs, suggesting the selection of functionally weak mutations in MMs (Fig. 4b). In line with this suggestion, Ba/F3 cells transduced with *PIK3CA* major hotspot mutants exhibited increased growth, whereas those transduced with minor hotspot mutants (except R88Q) did not, compared with wild-type-transduced cells in the absence of cytokine (Fig. 4c and Extended Data Fig. 10a). Notably, major–minor double mutants markedly enhanced proliferation compared with single mutants. A major–major double mutant involving different domains (E545K–H1047R) had a similar effect, whereas one within the same domain (E542K–E545K) exhibited no synergistic activity. Moreover, double mutants further augmented tumour growth in vivo compared with single mutants following xenotransplantation of MCF10A cells (Fig. 4d).

These results suggest that individually suboptimal mutations can confer enhanced oncogenic potential in MMs, which might explain their skewed mutational pattern. Notably, analysis of CRISPR–Cas9 loss-of-function screens in Cancer Cell Line Encyclopedia (CCLE) cell lines[14] revealed that most cell lines with *PIK3CA* MMs had the highest relative dependency on this gene within each cancer type (Fig. 4e). Furthermore, cells harbouring *PIK3CA* MMs exhibited stronger *AKT1* dependency and higher sensitivity to PI3K inhibitors than those with no or single *PIK3CA* mutations, pointing to the potential value of MMs as a predictive marker for targeted therapies (Fig. 4f and Extended Data Fig. 10b, c).

In terms of their underlying molecular mechanism, Ba/F3 cells expressing *PIK3CA* major hotspot mutants displayed increased AKT activation (that is, phosphorylation) than those expressing wild-type or minor hotspot mutants, but major–minor double mutants showed much greater AKT activation than single mutants (Fig. 4g and Extended Data Fig. 10d, e). A similar synergistic effect was observed for major–major double mutants affecting different domains, but not for those within the same domain. Moreover, PIK3CA double mutants strongly enhanced mTOR, p70 S6K, PRAS40 and GSK-3β phosphorylation downstream of AKT. These findings were supported by reverse-phase protein array (RPPA) data from CCLE cell lines, with elevated AKT phosphorylation (at S473 and T308) observed in MM$^+$ samples (Fig. 4h and Extended Data Fig. 10f). In addition, TCGA patients showed a greatly increased activation of AKT in MM$^+$ tumours, although a range of activity was evident within each cancer type (Fig. 4i and Extended Data Fig. 10g).

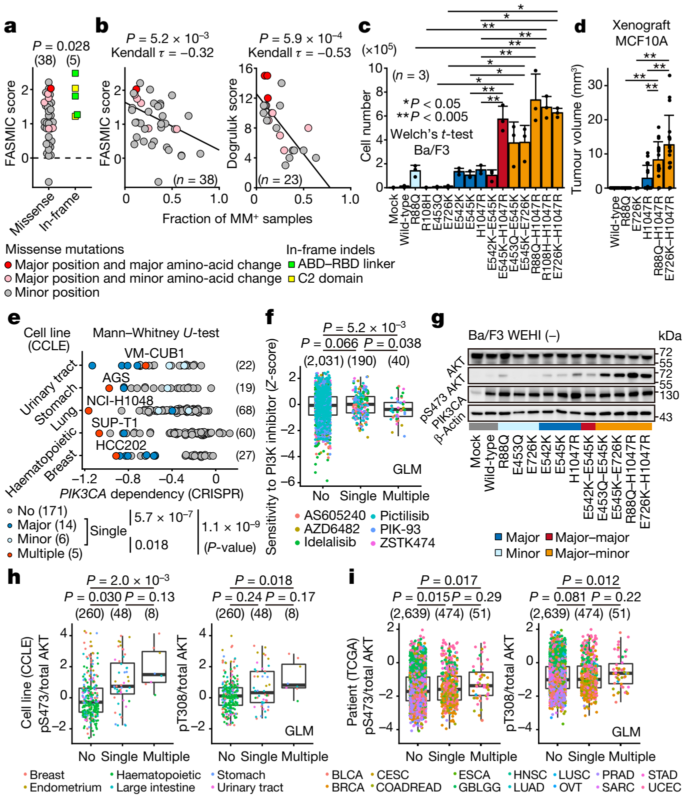


Fig. 4 | *PIK3CA* suboptimal mutations confer enhanced oncogenicity and downstream pathway activation in combination. **a**, Functional activity score for PIK3CA in the Functional Annotation of Somatic Mutations in Cancer database (FASMIC; https://ibl.mdanderson.org/fasmic/) according to mutation type. Two-sided Brunner–Munzel test. **b**, Association between fraction of MM$^+$ samples and PIK3CA functional activity score, in FASMIC[13] (left) and ref. [12] (right). Two-sided Kendall's rank correlation test. **a**, **b**, Each dot represents a mutation (present in five or more samples), shaped and coloured by type, position and amino-acid change. **c**, Proliferation of Ba/F3 cells expressing mock control or wild-type, single-mutant or double-mutant *PIK3CA* without WEHI-3-conditioned medium for three days. Two-sided Welch's *t*-test on log-transformed values. **d**, Xenograft tumour volumes (at eight weeks post-injection) resulting from MCF10A cells expressing mock control or wild-type, single-mutant or double-mutant *PIK3CA*. See Source Data for sample size. Two-sided Welch's *t*-test. **c**, **d**, Data represent means + s.d. **e**, Dependency of 196 CCLE cell lines on *PIK3CA*, with cell lines coloured by MM status, in the DepMap CRISPR–Cas9 knockout data (see Methods). Two-sided Mann–Whitney *U*-test. **f**, Box plots showing sensitivity to six PI3K inhibitors (*Z*-scores) for 417 CCLE cell lines, according to MM status. **e**, **f**, Five cancer types for which there are cell lines harbouring *PIK3CA* MMs were analysed. **g**, Immunoblot analysis of AKT, AKT with phosphorylated S473 (pS473 AKT), PIK3CA and β-actin in Ba/F3 cells expressing mock control or wild-type, single-mutant or double-mutant *PIK3CA* without WEHI-3-conditioned medium. Representative of three independent experiments. See Supplementary Fig. 1 for source images. **h**, **i**, Box plots showing levels of phosphorylated S473 (pS473; left) and T308 (pT308; right) in RPPA data relative to total AKT protein expression for 316 CCLE cell lines (with cancer types for which there are cell lines harbouring *PIK3CA* MMs) (**h**) and in 3,164 TCGA patients (with purity of 50% or more, in recurrently mutated cancer types) (**i**) according to MM status. Each dot represents a sample, coloured by cancer type. **f**, **h**, **i**, Box plots show medians (lines), interquartile ranges (IQR; boxes) and ±1.5 × IQR (whiskers). GLM, generalized linear model (see Methods). Examined numbers are shown in parentheses.

Among genetic alterations involved in the PI3K pathway, *PIK3CA* mutations tended to co-occur with *PTEN* mutations, but be mutually exclusive with *PIK3R1* mutations and *PIK3CA* amplifications in UCEC, which had the highest frequency of the mutated samples (Extended Data Fig. 11a, b). These associations were augmented by *PIK3CA* MMs, suggesting that a second mutation can further enhance the functional interplay among driver alterations of the PI3K pathway.

We then investigated the structural mechanism of MM-mediated PIK3CA (p110α subunit) activation by means of molecular-dynamics simulations of the synergistic mutants R88Q–H1047R, which show the strongest activity in vitro. Examination of the overall structure and various residue–residue contacts confirmed the predicted single-mutant-induced conformational changes. First, the R88–D746 salt bridge between the ABD and kinase domains was rendered unstable by the R88Q mutation, which promoted rotation of the iSH2 domain, contributing to exposure of the kinase domain. Second, the H1047R mutation distorted the orientation of the kinase domain, broadening substrate accessibility[15,16] (Extended Data Fig. 11c, d and Supplementary Table 10). Unexpectedly, although R88Q and H1047R single mutants slightly affected the R38–D743 salt bridge, the R88Q–H1047R double mutant caused its cleavage, leading to further detachment of the interface between the ABD and kinase domains and subsequent kinase exposure (Extended Data Fig. 11e–g). Therefore, a coordinated structural alteration might underlie the enhancement of downstream pathway activation by *PIK3CA* MMs.

We also found MM-induced enhanced functional activity in another gene: CRISPR–Cas9 screens in the CCLE revealed that a cell line with MMs in *NOTCH1* had the highest relative dependency on *NOTCH1*, among cell lines from haematopoietic and lymphoid tissue (Extended Data Fig. 11h). Single missense mutations in the *NOTCH1* HD domain increased NOTCH1 transcriptional activity compared with the wild type; but these mutations in combination with a truncating mutation further augmented its transcriptional activity (Extended Data Fig. 11i). Together, these observations reinforce the idea that MMs in the same oncogene cooperate to potentiate its tumour-promoting activity.

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/s41586-020-2175-2.

1. Kohsaka, S. et al. A method of high-throughput functional evaluation of *EGFR* gene variants of unknown significance in cancer. *Sci. Transl. Med.* **9**, eaan6566 (2017).
2. 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).
3. Garraway, L. A. & Lander, E. S. Lessons from the cancer genome. *Cell* **153**, 17–37 (2013).
4. Vogelstein, B. et al. Cancer genome landscapes. *Science* **339**, 1546–1558 (2013).
5. Heinrich, M. C. et al. Molecular correlates of imatinib resistance in gastrointestinal stromal tumors. *J. Clin. Oncol.* **24**, 4764–4774 (2006).
6. Kobayashi, S. et al. *EGFR* mutation and resistance of non-small-cell lung cancer to gefitinib. *N. Engl. J. Med.* **352**, 786–792 (2005).
7. Pao, W. et al. Acquired resistance of lung adenocarcinomas to gefitinib or erlotinib is associated with a second mutation in the EGFR kinase domain. *PLoS Med.* **2**, e73 (2005).
8. Soh, J. et al. Oncogene mutations, copy number gains and mutant allele specific imbalance (MASI) frequently occur together in tumor cells. *PLoS One* **4**, e7464 (2009).
9. Lawrence, M. S. et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).
10. Martincorena, I. et al. Universal patterns of selection in cancer and somatic tissues. *Cell* **171**, 1029–1041 (2017).
11. Weng, A. P. et al. Activating mutations of *NOTCH1* in human T cell acute lymphoblastic leukemia. *Science* **306**, 269–271 (2004).
12. Dogruluk, T. et al. Identification of variant-specific functions of *PIK3CA* by rapid phenotyping of rare mutations. *Cancer Res.* **75**, 5341–5354 (2015).
13. Ng, P. K. et al. Systematic functional annotation of somatic mutations in cancer. *Cancer Cell* **33**, 450–462 (2018).
14. Ghandi, M. et al. Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature* **569**, 503–508 (2019).
15. Gkeka, P. et al. Investigating the structure and dynamics of the *PIK3CA* wild-type and H1047R oncogenic mutant. *PLOS Comput. Biol.* **10**, e1003895 (2014).
16. Zhang, M., Jang, H. & Nussinov, R. The mechanism of PI3Kα activation at the atomic level. *Chem. Sci.* **10**, 3671–3680 (2019).

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.

Data set preparation

As the discovery data set, we obtained BAM files and mutational data in Mutation Annotation Format (MAF) (compiled by the MC3 Working Group for TCGA) for WES/WGS from TCGA (phs000178.v10.p8)[17–19], TARGET (phs000218.v20.p7)[20–22], and HM (AML (phs001657.v1.p1)[23] and NHL from the University of Iowa/Mayo Clinic Lymphoma Specialized Program of Research Excellence (SPORE) (phs000450.v3.p1)[24]) cohorts from the Genomic Data Commons (GDC) (https://gdc.cancer.gov/), the TARGET Data Matrix (https://ocg.cancer.gov/programs/target/data-matrix/), or the cBioPortal for Cancer Genomics (https://www.cbioportal.org/) (Extended Data Figs. 1a, 2a). As the additional data set, mutation calls for targeted-sequencing data from MSK-IMPACT and the Dana Farber Cancer Institute (DFCI) Oncopanel of the American Association for Cancer Research Project GENIE (Release 5.0; syn7222066)[25] and FM (Release 1.0; phs001179.v1.p1)[26] were downloaded from Synapse (https://www.synapse.org/) and the GDC Data Portal, respectively.

As synonymous mutations were excluded from the original mutational data for the HM AML cohort, mutation calling was performed as described with some modifications[27] (see Supplementary Methods). Genomic coordinates of mutations from alignments to the human reference genome GRCh38 were converted to GRCh37 (hg19) using LiftOver for the FM cohort. Then, the mutation calls were annotated to gene transcripts in Ensembl (release 95), and a single canonical effect per mutation was reported using Variant Effect Predictor (version 95.3) and vcf2maf (version 1.6.16). Germline variants had already been excluded in each cohort before downloading from the above sources. Briefly, matched normal samples were used in TCGA, TARGET, HM NHL and GENIE MSK-IMPACT cohorts. Further filtering for TCGA and HM NHL was performed with a pool of unmatched normal samples. Finally, all cohorts used single-nucleotide polymorphism (SNP) databases, such as Exome Aggregation Consortium (ExAC) and The Genome Aggregate Database, to further remove putative germline variants.

To eliminate probable false-positive MMs, we developed DNVChecker, which detects multiple SNVs occurring in the same codon, evaluates their allelic status using the corresponding BAM files, and changes them into a single alteration if they exist in cis (Extended Data Fig. 2c). We applied this to the discovery cohort and curated 13,482 DNVs and 812 TNVs. As more than 95% of the multiple SNVs in the same codon were DNVs/TNVs in the discovery cohort, all of them were converted to single alterations in the additional cohort, for which BAM files were unavailable. Moreover, we removed missense mutations (mostly SNVs) located within 3 bp of indels, as almost all of them were considered as a single genetic event (Extended Data Fig.2e). On the basis of mutant allele frequencies obtained from MAF files or calculated from BAM files using SAMtools (version 1.4.1) mpileup, mutations with low allele frequencies (less than 0.05), together with non-coding mutations, were removed.

Samples with missing data, derived from xenografts, and duplicated in the same patient were also excluded from the analysis. Hypermutator samples (with 500 or more and 20 or more coding mutations per exome (for the discovery cohort) and per targeted region (for the additional cohort), respectively) were also removed. In the remaining samples of the discovery cohort, 88 samples showed microsatellite instability (defined as having an MSIsensor score of 4 or more[17]) or pathogenic *POLE*/*POLD1* mutations[28], whose effect on MMs was independently evaluated. Only untreated sample data were used in the discovery data set. Although information on prior treatment was not available for the additional data set, we confirmed that the overall frequency and distribution of MMs were consistent across cohorts (Extended Data Figs. 3c, 5a, b).

The total cohort consisted of 40,002 primary and 20,952 metastatic samples representing more than 150 cancer types, which had 1,224,150 (synonymous and nonsynonymous) mutations (Extended Data Figs. 1a, 2b). Among the primary samples, the discovery cohort (n = 11,043) included 924,518 mutations, of which 4,472 mutations arose in 14 MM$^+$ oncogenes, while the additional cohort (n = 28,959) included 15,467 mutations in these oncogenes (Extended Data Fig. 2f). The metastatic samples, consisting mainly of the FM and GENIE cohorts, harboured 149,190 mutations, which contained 11,556 mutations in 14 MM$^+$ oncogenes (Extended Data Fig. 9a). Cancer type was classified on the basis of OncoTree codes (https://www.cbioportal.org/oncotree/), with slight modifications (Supplementary Table 1). Independent sets of AML[29], BALL[30–33] and CLL[34] patients were analysed for *FLT3*, *JAK2* and *NOTCH1*, respectively.

Oncogenes, TSGs and NFGs

To characterize MMs in oncogenes, we compiled a high-confidence list of 60 oncogenes, 35 TSGs and 129 NFGs (Supplementary Table 2). Oncogenes were selected if they were: (i) listed as 'oncogenes' in ref. [35]; or (ii) manually rescued as well described oncogenes (including *JAK2*, *PIK3R2*, *PTPN11*, and *SPOP*[4,18]). Putative TSGs (*CUL3*, *POLE* and *MED12*), target genes of aberrant somatic hypermutation (*BCL2*, *HIST1H1C* and so on) and oncogenes specific to haematological neoplasms (*ABL1*, *FLT3* and so on) were excluded from the pan-cancer oncogene list, although the latter genes were assessed as cancer-type-specific oncogenes. TSGs were selected if they were listed as 'tsg' in PANCAN from the TCGA Pan-Cancer Atlas project[19] and described as 'TSG' in ref. [4]. NFGs included genes frequently affected by passenger hotspot mutations and olfactory genes described in ref. [36] and ref. [9], respectively. Cancer-type-specific oncogenes were selected from those listed as 'oncogene' or 'possible oncogene' in a certain cancer type from the TCGA Pan-Cancer Atlas project and manually curated. Paediatric and haematological cancer-type-specific oncogenes were added from the literature. In total, we compiled an extended list of 84 cancer-type-specific oncogenes, including 38 genes listed in the pan-cancer oncogene list (Supplementary Table 11).

Identification of MM$^+$ oncogenes by permutation

To identify genes significantly affected by driver mutations, a permutation test is widely used, where the expected number of samples with mutations in gene X (the gene of interest) is estimated by permuting all coding (synonymous and nonsynonymous) mutations randomly across the coding region in all samples. Statistical significance is determined by comparing the observed number of samples with nonsynonymous mutations and the expected distribution in gene X. Here, to identify genes significantly affected by putative driver–driver MMs, we modified the permutation framework, such that the expected number of samples with MMs in gene X (that is, with additional gene X mutations) was estimated by permuting all coding mutations other than gene X mutations in samples harbouring gene X mutations (Extended Data Fig. 3d). For the permutation of SNVs, we considered: (i) sequence composition and mutational signature; (ii) expression; and (iii) DNA replication time, all of which are known to affect mutational frequency. Copy-number alterations were not taken into consideration, as they did not affect the frequency of MMs in our analysis (Extended Data Fig. 4d).

To account for sequence composition and mutational signature, permutations were restricted within each of the 64 trinucleotide contexts (2 types of mutated base (C or T) × 2 types of transcriptional strand × 4 types of 5′ base × 4 types of 3′ base). In addition, all coding sequences were divided into 225 bins according to expression (mean expression of 1,156 CCLE cell lines) and DNA replication time[9] (15 bins each), and the synonymous mutation rate was calculated in each bin and used to weight the sampling probability (Extended Data Fig. 3e, f). For example, a C-to-T substitution at ACG was moved to another ACG site of the same strand from the coding region randomly with the weighted

probabilities. Then, the permuted SNVs were classified into nonsynonymous or synonymous mutations, depending on the reading frame. Indels and other mutation types, such as DNVs/TNVs, were separately moved to a random position without weighting. The expected number of samples with additional gene X mutations—including nonsynonymous mutations and indels, but not synonymous mutations—was estimated, and its distribution (10,000 permutations) was compared with the observed number of samples with MMs. The advantage of the permutation test is that gene length and mutation burden (total number of mutations per sample) are considered: samples with high mutation burden have more mutations to be randomized, and genes with long coding sequences have a higher probability to obtain additional mutations. In pan-cancer analysis, 60 oncogenes, 35 TSGs and 129 NFGs were analysed in the discovery cohort, and multiple testing was adjusted for using the Benjamini–Hochberg method. To confirm the results, we estimated the selective pressure for MMs in these genes. Both synonymous and nonsynonymous mutations were collected, and the proportion of synonymous to total mutations was compared between samples with single and multiple mutations. In cancer-type-specific analysis, 200 gene–cancer-type pairs, including 46 oncogenes that had not been analysed in pan-cancer analysis, were evaluated.

Hotspot/functional mutation analysis

We first filtered out loss-of-function mutations, such as splice-site, nonsense, and frameshift mutations, unless they were listed as 'hotspot/functional' mutations (such as *NOTCH1* truncating mutations). Then, we divided the remaining mutations into 'hotspot/functional' and 'non-hotspot/functional' ones by position, and further classified the former into missense mutations and in-frame indels (Extended Data Fig. 1). The positions of hotspot/functional mutations were defined according to four knowledge-based databases (OncoKB[37], PMKB[38], Cancer Genome Interpreter (CGI)[39] and Clinical Interpretations of Variants in Cancers (CIViC)[40]) and four computational algorithms (3D Hotspots[41], Cancer Hotspots[42,43], HotMAPS[44] and CHASMplus[45]) (Supplementary Tables 5–7). For cancer-type-specific mutations, functionally validated mutation positions, such as *JAK2* I682, were also included (Supplementary Tables 13–15). Among 19,154 mutations in 14 MM^+ oncogenes in primary samples from the total cohort (after excluding 482 loss-of-function and 303 synonymous mutations), 14,645 mutations were present in recurrently mutated cancer types (defined as those with 20 or more hotspot/functional mutations in the gene), among which 13,708 mutations (94%) were in 1,679 hotspot/functional positions. These hotspot/functional missense mutational positions were further subdivided into major and minor positions: major ones were defined as those in which 10% or more of mutations were present in any of the recurrently mutated cancer types, and others were classified into minor ones. Amino-acid substitutions in major positions (excluding those frequently affected by major–major combinations in trans) were also subdivided into major and minor changes, which were defined as those present in 33% or more and less than 33% of cases with single mutations, respectively (Supplementary Table 8). Then, the fraction of MM^+ mutations—namely those found in samples harbouring MMs in the same oncogene (namely MM^+ samples)—was compared according to type, position and amino-acid change, in the recurrently mutated cancer types.

Enrichment of mutational combinations

Enrichment of a specific pair of mutations were assessed by Monte-Carlo simulation according to hotspot/functional position and functional domain. For analysis of hotspot/functional position, only combinations consisting of a pair of missense mutations in hotspot/functional positions were analysed. The expected number of combinations was estimated by randomly sampling two mutations (without replacement) on the basis of the mutation frequency in cases with single mutations (with 10,000 iterations). Statistical significance was determined by comparing the observed number of combinations and the expected distribution, and adjusted for multiple testing by the Benjamini–Hochberg method.

Determination of cis and trans phase of MMs

To determine allelic configurations of MMs, we developed CisChecker, which first identifies MMs within the same gene from mutation call data, then extracts sequencing reads encompassing both mutational positions from the corresponding BAM file, and classifies them into those containing both mutant alleles (cis reads), one mutant and one reference allele (trans reads), or both reference alleles (reference reads) (Extended Data Fig. 4a). MMs supported by two or more cis reads and no trans read were considered as cis, and vice versa as trans. Otherwise, the allelic status was assessed by a permutation test, where each mutation was randomly permuted in each position among sequencing reads encompassing both mutational positions. Then, the expected numbers of cis and trans reads were estimated by 10,000 permutations and compared with the observed numbers of cis and trans reads, respectively.

Using the MAF and RNA-seq BAM files, we applied this algorithm to 1,417 MMs (283 and 1,134 MMs in oncogenes and TSGs, respectively) from 1,082 samples in the discovery cohort, among which 220 MMs (16%) were evaluable. Among them, the allelic status for 142 MMs could be assessed by WES/WGS, and almost all of them (more than 99%) were validated. On the basis of this finding, 120 MMs from 101 samples whose RNA-seq data were not available were analysed only by WES/WGS, among which 27 MMs (23%) were evaluable. In the combined data (247 MMs), allelic configuration was compared according to gene category, mutational combination, and distance between mutations. Representative MMs were visualized with the Integrated Genomics Viewer (version 2.4.10). The difference in allele frequency between MMs (combinations of missense mutations only) in the same gene was assessed by Fisher's exact test (with Benjamini–Hochberg correction) for the number of mutant and reference reads, and those with significantly different allele frequencies (with q-values of less than 0.05 and a mutant allele frequency difference of greater than 0.10) were considered as 'discordant', and otherwise as 'concordant'. For *PIK3CA* major–minor combinations showing discordant allele frequencies, we used a binomial test to evaluate whether the proportion of major and minor dominant MMs (based on mutant allele frequency) deviated from the expected distribution in which the order of major and minor mutations was random (that is, the probabilities of major and minor dominant MMs were equal).

Association with clinical and genetic information

Tumour purity inferred by ABSOLUTE (TCGA_mastercalls.abs_tables_JSedit.fixed.txt), In Silico Admixture Removal (ISAR)-corrected Genomic Identification of Significant Targets in Cancer (GISTIC) copy-number data (ISAR_GISTIC.all_thresholded.by_genes.txt) and batch-corrected RPPA data (TCGA-RPPA-pancan-clean.txt) by the TCGA Pan-Cancer Atlas project[19] were obtained from the GDC.

CCLE cell line data

The mutation call data (depmap_19Q1_mutation_calls_v2.csv) for 1,601 cell lines and batch-corrected genome-wide CRISPR–Cas9 knockout screen data (gene_effect_corrected.csv for Public 19Q1) for 558 cell lines, RPPA data (CCLE_RPPA_20181003.csv) for 899 cell lines, and drug-sensitivity data (v17.3_fitted_dose_response.xlsx) for 1,065 cell lines were obtained from the DepMap (https://depmap.org/portal/), CCLE (https://portals.broadinstitute.org/ccle/), and Genomics of Drug Sensitivity in Cancer (GDSC; https://www.cancerrxgene.org/) databases, respectively. Mutations with allele frequencies of 0.05 or more were assessed and cell lines (with less than 5,000 coding mutations) were classified according to *PIK3CA* or *NOTCH1* mutation status. Dependency on *PIK3CA*, *AKT1* and *NOTCH1* was examined by ranking the CERES score (a measure of gene dependency) in each cancer type (Mann–Whitney U-test).

RPPA and drug-sensitivity analysis

For RPPA, levels of AKT with phosphorylated S473 and T308 relative to total AKT protein expression were compared in 14 recurrently mutated cancer types (TCGA) and 6 cancer types for which there are *PIK3CA*-MM-harbouring cell lines (CCLE), using the generalized linear model (GLM) with cancer type as a covariate. For drug-sensitivity analysis, we used at least three *PIK3CA*-MM-harbouring cell lines to test six PI3K inhibitors, namely AS605240 (GDSC drug identification code 224), AZD6482 (156), Idelalisib (238), Pictilisib (1,058), PIK-93 (303) and ZSTK474 (223), by GLM, using drug and cancer type as covariates.

PIK3CA functional activity score

The biological activity of *PIK3CA* hotspot/functional missense mutations and in-frame indels (five or more samples) was evaluated on the basis of two previous functional studies[12,13]. In ref. [12], five different functional assays—including growth-factor- and insulin-free survival, colony formation for MCF10A cells, interleukin-3-less survival for Ba/F3 cells, and tumorigenesis for immortalized mammary epithelial cells—were performed; we scored these results according to their functional impact (no, weak, intermediate and strong phenotype for 0, 1, 2 and 3, respectively) and summed them to yield a total score ranging from 0–15 for each mutant ($n = 23$). For the cell viability data (v1) for Ba/F3 cells obtained from the FASMIC database (https://ibl.mdanderson.org/fasmic/)[13], the cell number of each mutant was averaged in available replicates, normalized to wild-type PIK3CA, and log-transformed ($n = 43$). A functional activity score was calculated as the arithmetic mean at well controlled time points (with wild-type PIK3CA relative to negative control (mCherry and/or Luc) value of 2.5 or less). Correlation between these functional scores and MM fraction was examined using Kendall's rank correlation coefficient.

Cell lines, plasmid constructs and lentiviral transduction

Ba/F3-CL1, WEHI-3 and 293T cell lines were obtained from the RIKEN Cell Bank, the HEC-1 cell line from the JCRB Cell Bank, and the MCF10A, BT-20, NCI-H1048, SUP-T1 and HRT-18 cell lines from ATCC. Lenti-X 293T cells were purchased from TaKaRa. Cell lines were authenticated by the providers using karyotype, isoenzymes, and/or microsatellite profiling. They were cultured according to the providers' instructions, and routinely tested for mycoplasma infection. According to the International Cell Line Authentication Committee (ICLAC) register (https://iclac.org/), cross-contamination has been reported in BT-20 cells, although authentic stocks apparently do exist. We used the BT-20 cell line authenticated by ATCC using short tandem repeat (STR) profiling analysis. This cell line was selected because of the limited number of available cell lines harbouring MMs in *PIK3CA*.

Plasmids containing cDNAs encoding human wild-type (catalogue number 81736), R108H (82875), E453Q (82844), E545K (82881), E726K (82845) and H1047R (82824) *PIK3CA* were obtained from Addgene. R88Q, E542K and E542K–E545K mutants were constructed from the wild type by site-directed polymerase chain reaction (PCR)-based mutagenesis using PrimeSTAR MAX (TaKaRa). The PCR primers are listed in Supplementary Table 16. E545K–H1047R, E453Q–E545K, E545K–E726K, R88Q–H1047R, R108H–H1047R and E726K–H1047R double mutants were constructed from single-mutant vectors using standard restriction-enzyme-mediated cloning. These *PIK3CA* cDNAs were cloned into CSII-EF-RfA-IRES2-Venus (RDB04389; from H. Miyoshi, Riken) using Gateway LR clonase II enzyme mix (Thermo Fisher Scientific). Lenti-X 293T cells were transiently transfected with lentiviral vectors, pMD2.G and psPAX2 (catalogue numbers 12259 and 12260; Addgene) using polyethylenimine 'max' (Polysciences). The viral supernatant was collected 48 h later and used for infection of Ba/F3-CL1 cells with RetroNectin (TaKaRa). After 48 h, the infected cells were harvested, sorted and subjected to culture.

Plasmids containing wild-type human *NOTCH1* cDNA (pFN21AE3300) were obtained from the Kazusa DNA Research Institute. *NOTCH1* L1600P, L1678P, Q2416* (stop codon), L1600P–Q2416* and L1678P–Q2416* mutations were artificially synthesized (FASMAC) and cloned into pcDNA3 (Invitrogen) using standard restriction-enzyme-mediated cloning.

Cell proliferation and viability assay

Ba/F3-CL1 cells were seeded in 12-well plates at 1×10^5 cells per well and incubated for 72 h without WEHI-3 supernatant. Cell number and viability were determined using trypan blue (Nacalai Tesque) staining and cell counter model R1 (Olympus).

In vivo xenograft tumour model

All mouse experiments were approved by the Animal Ethics Committee of the National Cancer Center and strictly adhered to its guidelines. Female BALB/c-nu/nu mice (six weeks old) were obtained from CLEA and maintained under pathogen-free conditions. Mice were subcutaneously injected with 5×10^5 MCF10A cells in 1/1 phosphate-buffered saline (PBS)/matrigel (BD Biosciences), and tumour volumes were measured eight weeks after transplantation. The maximum tumour diameter permitted under the relevant animal protocols is 20 mm, and this limit was not exceeded in any experiment.

Statistical analysis

Statistical analyses were performed with R3.6.0 software (The R Foundation for Statistical Computing). Comparison of categorical and continuous data was performed using two-sided Fisher's exact test and Brunner–Munzel test, respectively, unless otherwise specified. For functional assays, statistical significance was assessed by two-sided Welch's *t*-test, unless otherwise specified.

Reporting summary

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

Data availability

Our findings are supported by data that are available from public online repositories, or data that are publicly available upon request from the data provider. See the 'Data set preparation' section above for details. Long-read WGS data for cell lines have been deposited in the European Genome-phenome Archive (EGA; https://www.ebi.ac.uk/ega/) under accession number EGAS00001003763. Data generated here are available as Source Data files accompanying Fig. 4 and Extended Data Figs. 10, 11.

Code availability

Source codes for CisChecker, DNVChecker and the permutation test are available at https://github.com/nccmo/CisChecker, https://github.com/nccmo/DNVChecker and https://github.com/nccmo/Permutation-test, respectively.

17. Ding, L. et al. Perspective on oncogenic processes at the end of the beginning of cancer genomics. *Cell* **173**, 305–320 (2018).
18. Sanchez-Vega, F. et al. Oncogenic signaling pathways in The Cancer Genome Atlas. *Cell* **173**, 321–337 (2018).
19. Bailey, M. H. et al. Comprehensive characterization of cancer driver genes and mutations. *Cell* **173**, 371–385 (2018).
20. Ma, X. et al. Pan-cancer genome and transcriptome analyses of 1,699 paediatric leukaemias and solid tumours. *Nature* **555**, 371–376 (2018).
21. Mullighan, C. G. et al. Deletion of *IKZF1* and prognosis in acute lymphoblastic leukemia. *N. Engl. J. Med.* **360**, 470–480 (2009).
22. Pugh, T. J. et al. The genetic landscape of high-risk neuroblastoma. *Nat. Genet.* **45**, 279–284 (2013).
23. Tyner, J. W. et al. Functional genomic landscape of acute myeloid leukaemia. *Nature* **562**, 526–531 (2018).

24. Chapuy, B. et al. Molecular subtypes of diffuse large B cell lymphoma are associated with distinct pathogenic mechanisms and outcomes. *Nat. Med*. **24**, 679–690 (2018).
25. AACR Project GENIE Consortium. AACR Project GENIE: powering precision medicine through an international consortium. *Cancer Discov*. **7**, 818–831 (2017).
26. Hartmaier, R. J. et al. High-throughput genomic profiling of adult solid tumors reveals novel insights into cancer pathogenesis. *Cancer Res*. **77**, 2464–2475 (2017).
27. Kataoka, K. et al. Integrated molecular analysis of adult T cell leukemia/lymphoma. *Nat. Genet*. **47**, 1304–1315 (2015).
28. Campbell, B. B. et al. Comprehensive analysis of hypermutation in human cancer. *Cell* **171**, 1042–1056 (2017).
29. Papaemmanuil, E. et al. Genomic classification and prognosis in acute myeloid leukemia. *N. Engl. J. Med*. **374**, 2209–2221 (2016).
30. Gu, Z. et al. *PAX5*-driven subtypes of B-progenitor acute lymphoblastic leukemia. *Nat. Genet*. **51**, 296–307 (2019).
31. Li, J. F. et al. Transcriptional landscape of B cell precursor acute lymphoblastic leukemia based on an international study of 1,223 cases. *Proc. Natl Acad. Sci. USA* **115**, E11711–E11720 (2018).
32. Steeghs, E. M. P. et al. *JAK2* aberrations in childhood B-cell precursor acute lymphoblastic leukemia. *Oncotarget* **8**, 89923–89938 (2017).
33. Forero-Castro, M. et al. Mutations in *TP53* and *JAK2* are independent prognostic biomarkers in B-cell precursor acute lymphoblastic leukaemia. *Br. J. Cancer* **117**, 256–265 (2017).
34. Puente, X. S. et al. Non-coding recurrent mutations in chronic lymphocytic leukaemia. *Nature* **526**, 519–524 (2015).
35. Bielski, C. M. et al. Widespread selection for oncogenic mutant allele imbalance in cancer. *Cancer Cell* **34**, 852–862 (2018).
36. Buisson, R. et al. Passenger hotspot mutations in cancer driven by APOBEC3A and mesoscale genomic features. *Science* **364**, eaaw2872 (2019).
37. Chakravarty, D. et al. OncoKB: a precision oncology knowledge base. JCO Precis. *Oncol*. **1**, 1–16 (2017).
38. Huang, L. et al. The cancer precision medicine knowledge base for structured clinical-grade mutations and interpretations. *J. Am. Med. Inform. Assoc*. **24**, 513–519 (2017).
39. Tamborero, D. et al. Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. *Genome Med*. **10**, 25 (2018).
40. Griffith, M. et al. CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. *Nat. Genet*. **49**, 170–174 (2017).
41. Gao, J. et al. 3D clusters of somatic mutations in cancer reveal numerous rare mutations as functional targets. *Genome Med*. **9**, 4 (2017).
42. Chang, M. T. et al. Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. *Nat. Biotechnol*. **34**, 155–163 (2016).
43. Chang, M. T. et al. Accelerating discovery of functional mutant alleles in cancer. *Cancer Discov*. **8**, 174–183 (2018).
44. Tokheim, C. et al. Exome-scale discovery of hotspot mutation regions in human cancer using 3D protein structure. *Cancer Res*. **76**, 3719–3731 (2016).
45. Tokheim, C. & Karchin, R. CHASMplus reveals the scope of somatic missense mutations driving human cancers. *Cell Syst*. **9**, 9–23 (2019).

Acknowledgements We acknowledge support from the Japan Society for the Promotion of Science (JSPS) KAKENHI (grant numbers 17K19592,18K06594 and 15H05912) and National Cancer Center Research and Development Funds (30-A-1), together with many other funding bodies, organizations and individuals (see Supplementary Note).

Author contributions Y. Saito and K.K. designed the study. Y. Saito, Y.K. and K.K. analysed sequencing data. M.B.M., H.T., T.K., S. Miyano and Y. Shiraishi assisted with the collection and analysis of sequencing data. S.S. and M.T. performed sequencing experiments. J.K. and K.Y. performed immunoblots, capillary-based immunoassays, cell proliferation assays and in vivo xenograft assays. J.K. and Y.K. performed luciferase assays. M.A., S. Matsumoto, Y.I. and Y.O. performed molecular-dynamics simulations. Y. Saito and K.K. generated figures and tables and wrote the manuscript. K.K. led the entire project. All authors participated in discussions and interpretation of the data and results.

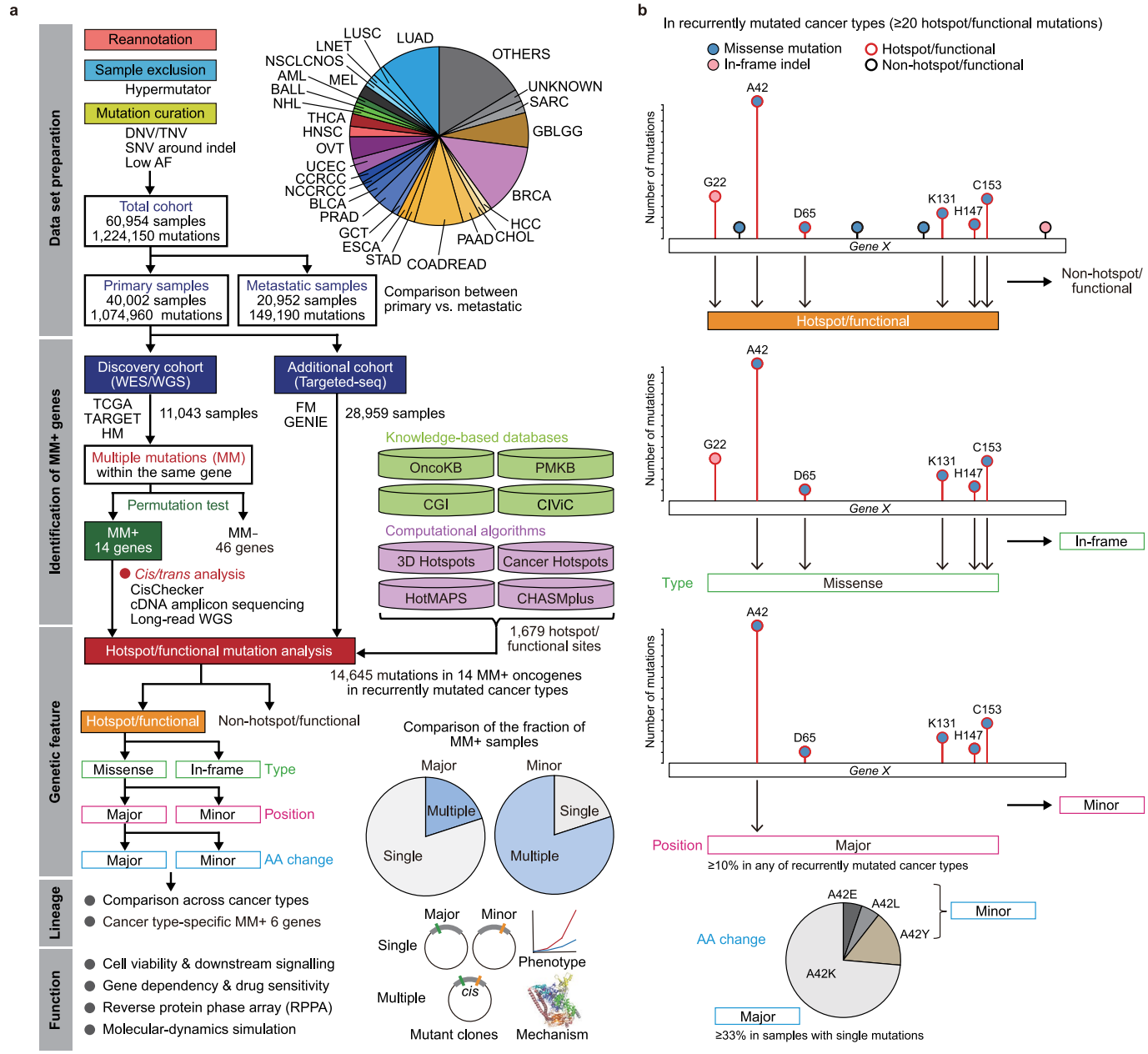
Competing interests The authors declare no competing interests.

Additional information

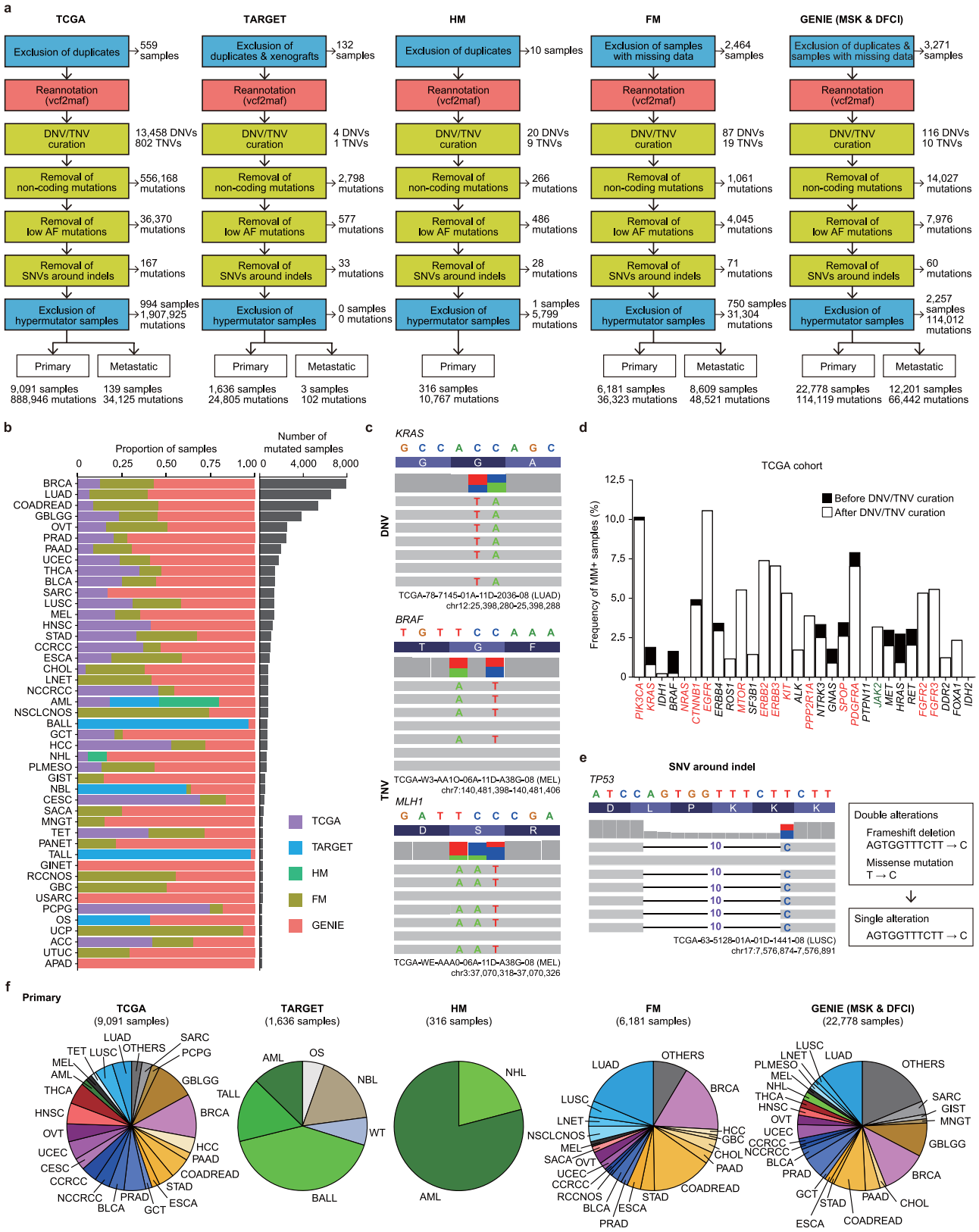
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2175-2.

Correspondence and requests for materials should be addressed to K.K.

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

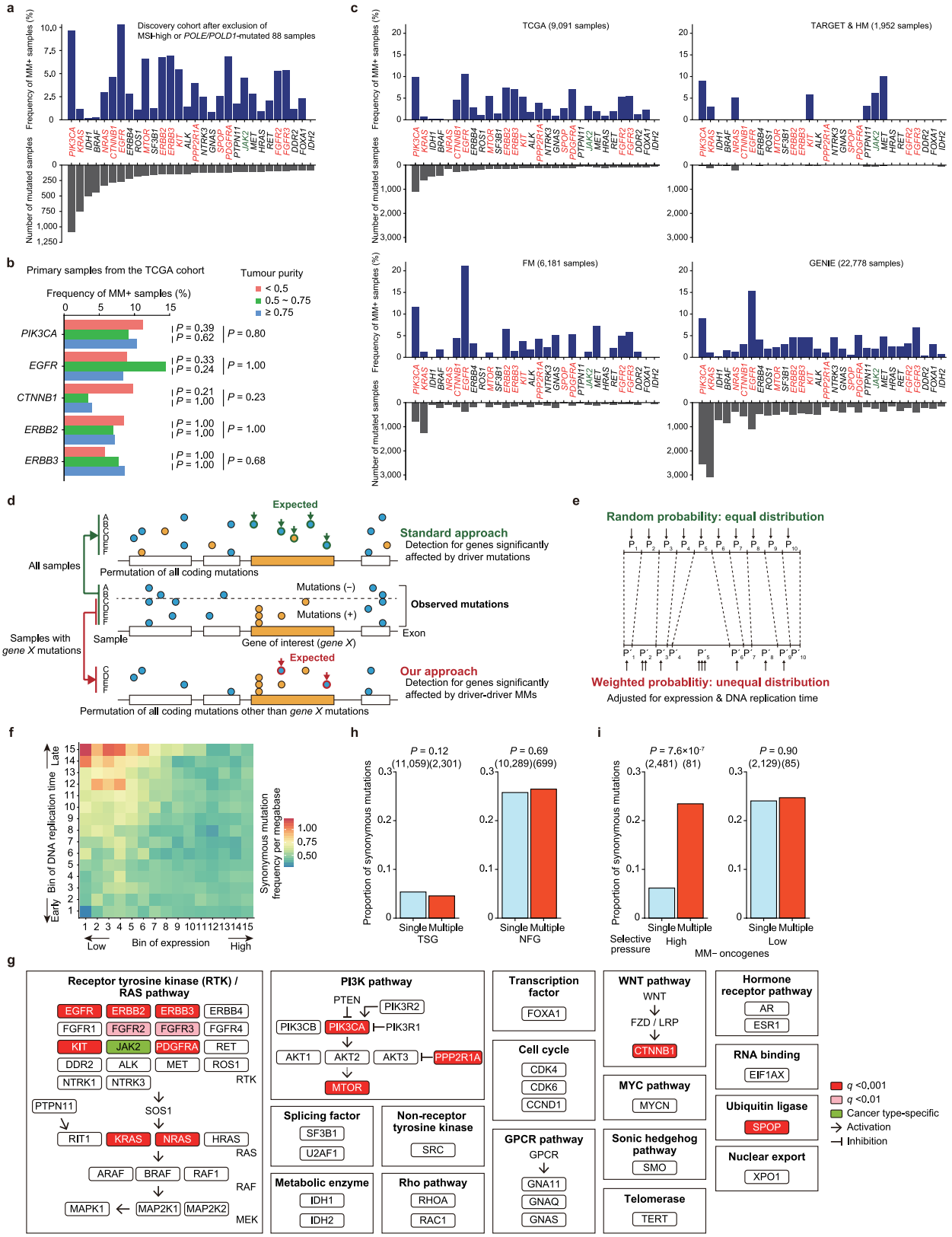


Extended Data Fig. 1 | Workflow showing discovery and assessment of oncogenic MMs. a, Data set preparation, identification of MM$^+$ oncogenes, and analysis of genetic features, lineage specificity and functional role. Distribution of cancer types (Supplementary Table 1) for the total cohort (n = 60,954). Cancer types with less than 1% frequency were combined as 'OTHERS'. AF, allele frequency; NCCRCC, non-clear cell renal-cell carcinoma; NSCLCNOS, NSCLC not otherwise specified. **b**, Classification of hotspot/functional mutations according to mutation type, position and amino-acid change.



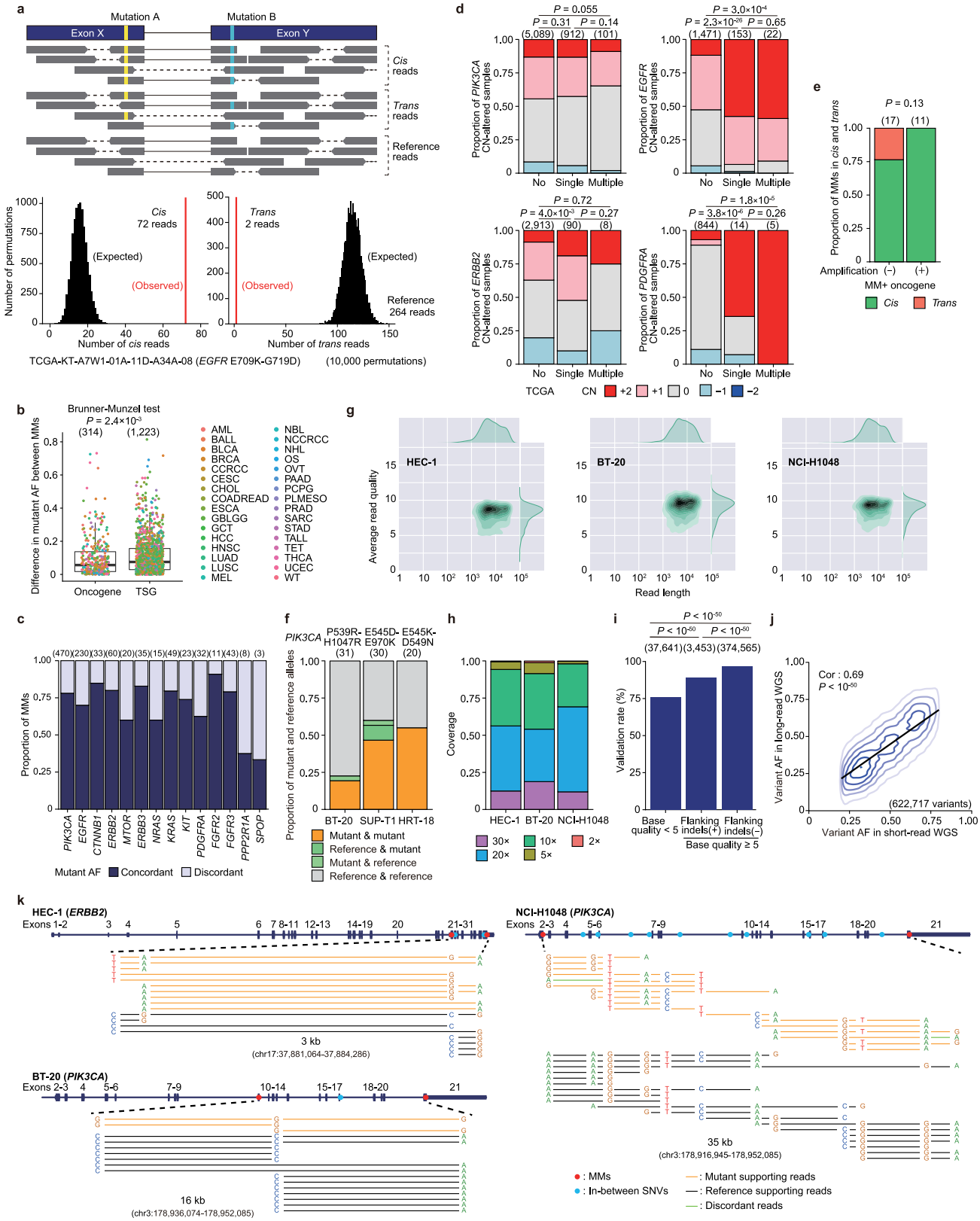
Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | Data set preparation for the TCGA, TARGET, HM, FM and GENIE cohorts. a, Steps involved in reannotation, sample exclusion and mutation curation in each cohort. **b**, Number of mutated samples (right) and proportion of samples in each cohort (left) for 45 cancer types (with 200 or more samples). ACC, adrenocortical carcinoma; APAD, appendiceal adenocarcinoma; GBC, gallbladder cancer; GINET, gastrointestinal neuroendocrine tumours; MNGT, meningothelial tumour; NBL, neuroblastoma; OS, osteosarcoma; PANET, pancreatic neuroendocrine tumour; PCPG, pheochromocytoma and paraganglioma; PLMESO, pleural mesothelioma; RCCNOS, renal-cell carcinoma not otherwise specified; TET, thymic epithelial tumour; UCP, undifferentiated carcinoma of the pancreas, USARC, uterine sarcoma/mesenchymal (see also Supplementary Table 1). **c**, DNVs and TNVs arising in the same codon (considered as a single genomic event) in representative cases. The TCGA identification code for the sample is shown below each chart. **d**, Frequency of MM^+ samples with or without DNV and TNV curation for 30 oncogenes (as in Fig. 1a) in primary samples from TCGA. **e**, Removed SNVs around indels. In this representative case, SNVs (T-to-C) and frameshift deletions (AGTGGTTTCTT-to-C) were independently miscalled, but considered to be a single genomic event. **f**, Distribution of cancer types for primary samples included in the TCGA, TARGET, HM, FM and GENIE cohorts. Cancer types with less than 1% frequency were combined as 'OTHERS' in each cohort. WT, Wilms' tumour.



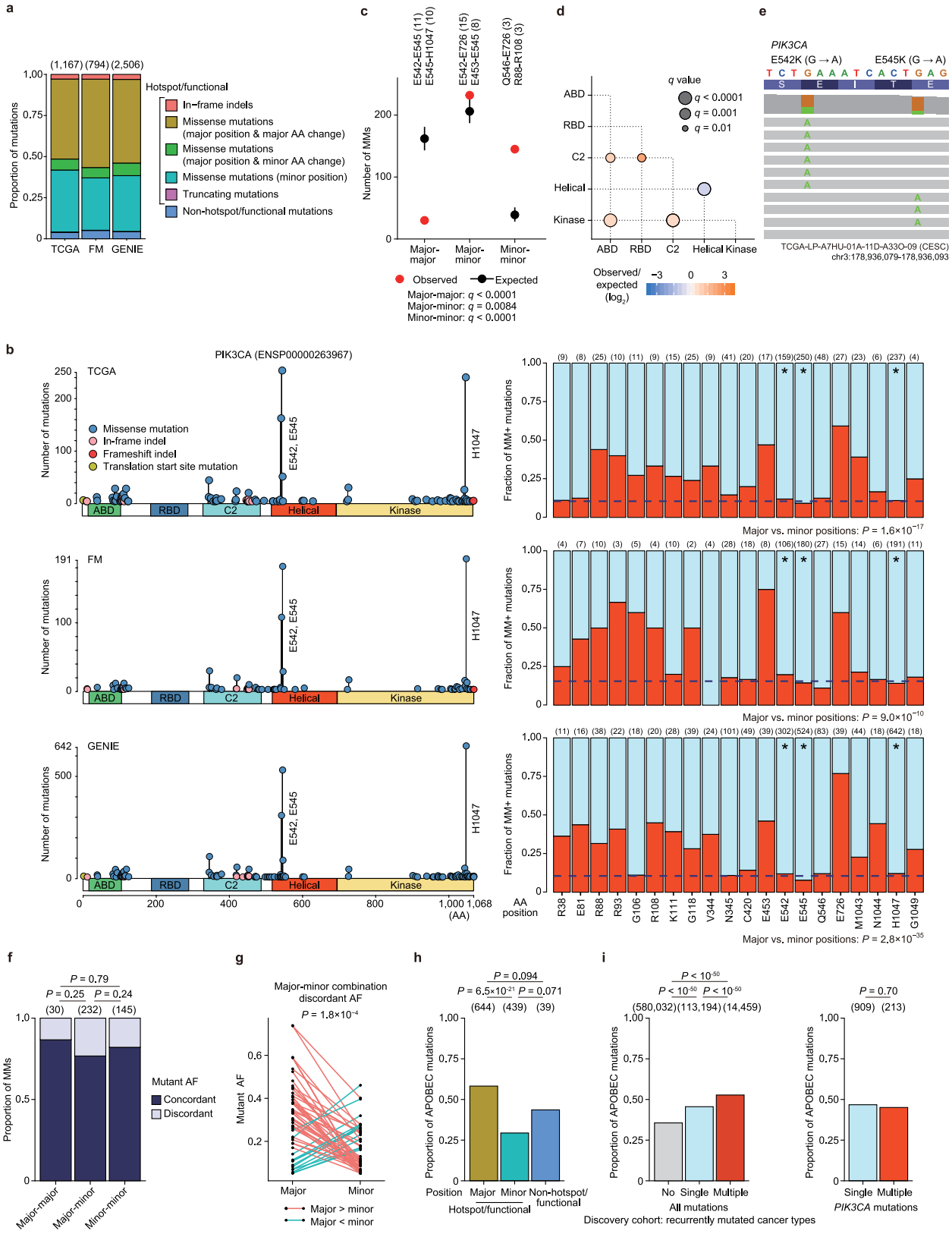
Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | Identification of MM$^+$ oncogenes by a permutation test across cancers. a, Number of mutated samples and frequency of MM$^+$ samples for 30 oncogenes (as in Fig. 1a) after excluding microsatellite instability (MSI)-high or *POLE/POLD1*-mutated samples in the discovery cohort. **b**, Frequency of MM$^+$ samples in *PIK3CA*, *EGFR*, *CTNNB1*, *ERBB2* and *ERBB3* according to tumour purity in primary samples from the TCGA cohort (n = 8,699). **c**, Number of mutated samples and frequency of MM$^+$ samples for 30 oncogenes (as in Fig. 1a) in the TCGA, TARGET, HM, FM and GENIE cohorts. **d**, Representation of the permutation-based framework. In the standard approach, to identify genes significantly affected by driver mutations, the expected number of samples with mutations in gene X (the gene of interest; green) is estimated by permuting all coding mutations randomly across the coding region in all samples (for example, samples A–F). Statistical significance is determined by comparing the observed number of samples with nonsynonymous mutations and the expected distribution in gene X. In our approach, to identify genes significantly affected by putative driver–driver MMs, the expected number of samples with MMs in gene X (red) is estimated by permuting all coding mutations other than gene X mutations in samples harbouring gene X mutations (samples C–F). Statistical significance is determined by comparing the observed number of samples with MMs and the expected distribution in gene X. **e**, In the random-choice model, mutations are moved to another position with equal probability (P), whereas in the weighted-choice model, mutations are moved with unequal probability ($P^{'}$), reflecting expression and DNA replication time. **f**, Synonymous mutation frequency per megabase according to expression and DNA replication time. **g**, Pathways related to 60 oncogenes analysed here. MM$^+$ oncogenes identified in pan-cancer and cancer type-specific analyses are indicated in red (q < 0.001)/pink (q < 0.01) and green, respectively. **h**, Proportion of synonymous to total mutations according to MM status in TSGs and NFGs. **i**, Proportion of synonymous to total mutations according to MM status in MM$^-$ oncogenes under high and low selective pressure (that is, oncogenes in which the proportion of synonymous to total mutations is less than and more than 15% in samples with single mutations, respectively). The proportion of synonymous mutations was substantially increased in MM$^+$ samples, even in MM$^-$ oncogenes under high selective pressure. **b**, **h**, **i**, Two-sided Fisher's exact test. The numbers examined are shown in parentheses.



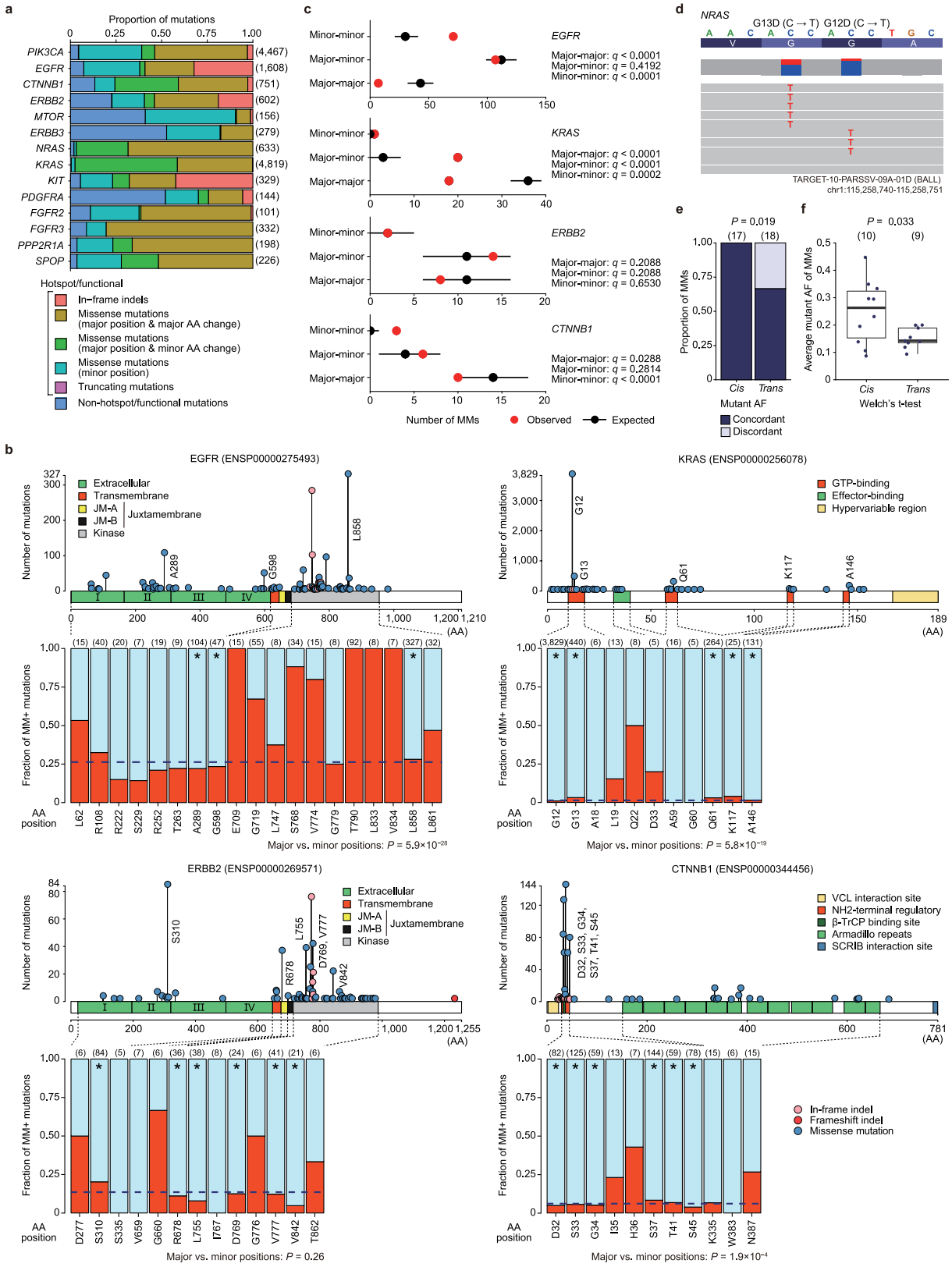
Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | Allelic configuration (cis versus trans) of oncogenic MMs. a, Top, sequencing reads encompassing MMs were classified into those containing both mutant alleles (cis reads), one mutant and one reference allele (trans reads), and both reference alleles (reference reads). Bottom, one-sided permutation test (n = 10,000) for the allelic configuration (cis versus trans) of MMs. In this representative case, the observed numbers of cis (left) and trans (right) reads (red) were significantly higher and lower, respectively, than the expected distribution (black); thus, this example is considered to be cis. **b**, Difference in mutant allele frequency (AF) between MMs across 60 oncogenes and 35 TSGs in the discovery cohort. Each dot represents an MM, coloured by cancer type. Two-sided Brunner–Munzel test. Box plots show medians (lines), interquartile ranges (IQRs; boxes) and ±1.5 × IQRs (whiskers). **c**, Proportion of MMs (combinations of missense mutations only) showing concordant or discordant allele frequencies in MM$^+$ oncogenes in primary samples from the total cohort. **d**, Fraction of *PIK3CA*, *EGFR*, *ERBB2* and *PDGFRA* copy-number (CN) alterations according to MM status in recurrently mutated cancer types (defined as those with 20 or more hotspot/functional mutations) in primary samples from TCGA. **e**, Proportion of MMs in cis and trans (with distances between mutations of 25 bp or more) by phasing from RNA-seq or WES/WGS in MM$^+$ oncogenes with and without concurrent CN amplification of the mutated gene. **f**, Allelic configuration (cis versus trans) assessed by cDNA amplicon sequencing for *PIK3CA* P539R–H1047R, E545D–E970K and E545K–D549N mutations in BT-20, SUP-T1 and HRT-18 cell lines, respectively. Proportions of mutant and reference alleles are shown. **b**–**f**, Examined numbers are shown in parentheses. **g**, Density plot illustrating the distribution of read length and average read quality for each of three long-read WGS samples. **h**, Percentage of bases covered by at least ×2, ×5, ×10, ×20 and ×30 sequencing reads for three long-read WGS samples. **i**, Validation rate of SNV calling from long-read WGS according to base quality and/or flanking indels. Examined read numbers are shown in parentheses. **d**, **e**, **i**, Two-sided Fisher's exact test. **j**, Density plot showing the correlation between variant allele frequencies in short-read and long-read WGS in positions with coverage of at least ×40 and at least ×20, in short-read and long-read WGS, respectively. Two-sided Pearson's correlation test. **k**, Phasing of MMs using long-read WGS reads. Positions of MMs (red) and in-between SNVs (blue) according to their genomic position (top) and long-read WGS reads between them (bottom). Reads supporting both mutant alleles and both reference alleles present in cis are shown in orange and black, respectively; discordant reads are shown in green.



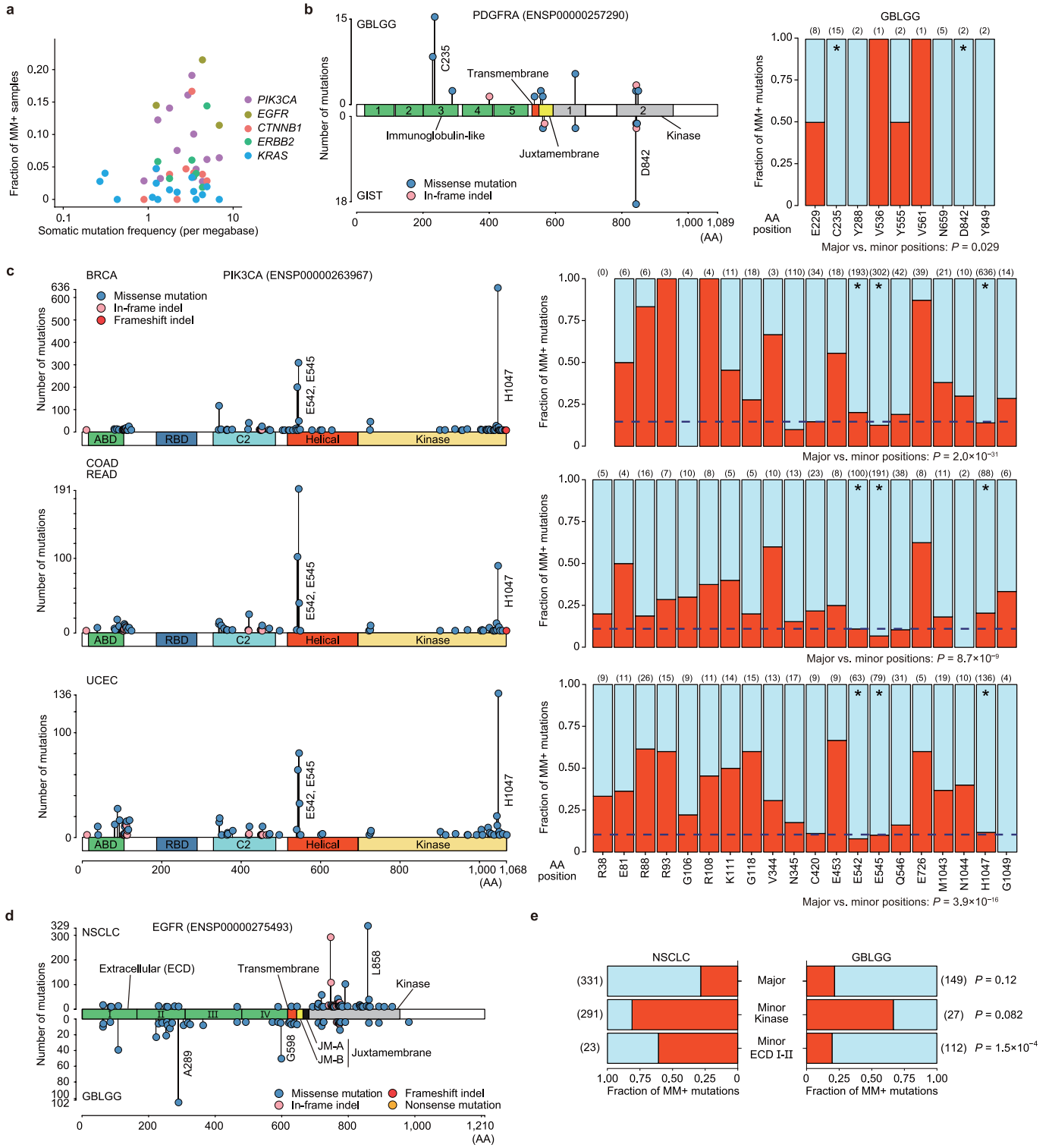
Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | Comparison of *PIK3CA* MMs among cohorts. a, Proportion of *PIK3CA* mutations according to type, position and amino-acid change in recurrently mutated cancer types (defined as those with 20 or more hotspot/functional mutations) in primary samples from the TCGA, FM and GENIE cohorts. **b**, Distribution of mutations and fraction of MM^+ mutations for each hotspot/functional position. Asterisks indicate major positions (in which 10% or more of mutations were present in any of the recurrently mutated cancer types). The horizontal blue dotted lines represent the mean values of major positions. **c**, Number of MMs according to mutational combinations ($n = 407$). Red and black circles indicate, respectively, observed and expected (median with 95% confidence intervals) values. Representative combinations and observed numbers are also shown. **d**, Significant pairwise associations ($q < 0.01$) with observed/expected ratios among functional domains ($n = 471$). Orange and blue colours depict co-occurring (observed number of MMs significantly higher than expected) and mutually exclusive (lower than expected) associations. **c**, **d**, Two-sided simulation test ($n = 10{,}000$) with Benjamini–Hochberg correction. **e**, Major–major combinations on different alleles within close proximity (E542–E545). **f**, Proportion of MMs showing concordant or discordant allele frequencies according to mutational combinations. **g**, Order of major and minor hotspot mutations in major–minor combinations showing discordant allele frequencies ($n = 54$). Two-sided binomial test. **h**, Proportion of mutations with APOBEC signature (C-to-G/T at TpCpX trinucleotides) according to hotspot/functional position in *PIK3CA* in the discovery cohort. **i**, Proportion of mutations with APOBEC signature according to MM status in all coding (left) and *PIK3CA* (right) mutations in the discovery cohort. **b**, **f**, **h**, **i**, Two-sided Fisher's exact test. Examined numbers are shown in parentheses.

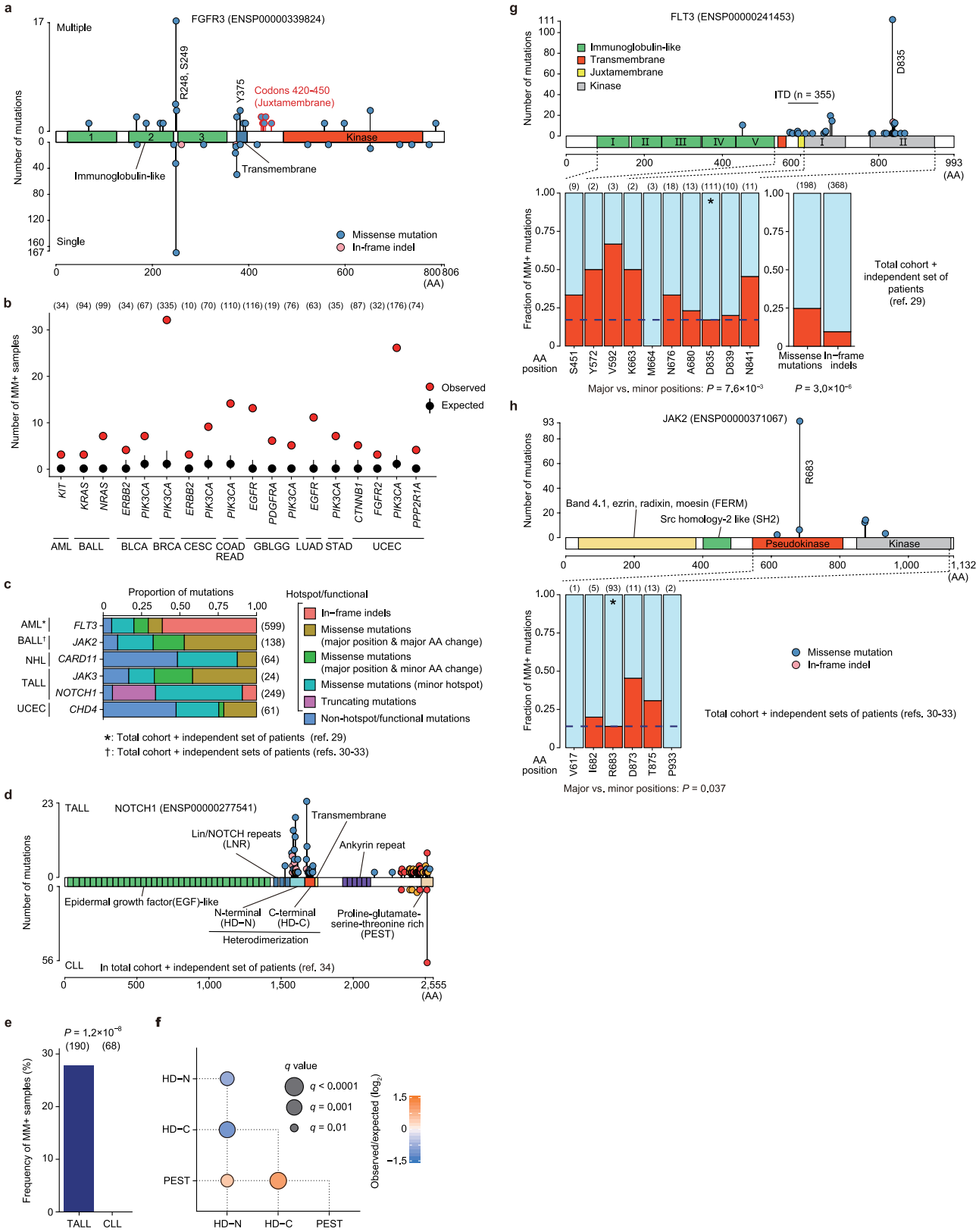


Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | Genetic features of MMs in a variety of oncogenes. **a**, Proportion of mutations according to type, position and amino-acid change across MM^+ oncogenes in recurrently mutated cancer types (defined as those with 20 or more hotspot/functional mutations) in primary samples from the total cohort. **b**, Distribution of mutations and fraction of MM^+ mutations for each hotspot/functional position in *EGFR*, *KRAS*, *ERBB2* and *CTNNB1*. Positions showing five mutations or more (and within the top 20 for *EGFR*) are shown in the bar plots. Asterisks indicate major positions (in which 10% or more of mutations were present in any of the recurrently mutated cancer types). The horizontal blue dotted lines represent the mean values of major positions. **c**, Number of MMs according to mutational combinations in *EGFR* ($n = 185$), *KRAS* ($n = 39$), *ERBB2* ($n = 24$) and *CTNNB1* ($n = 19$). Red and black circles indicate, respectively, observed and expected (median with 95% confidence intervals) values. Two-sided simulation test ($n = 10{,}000$) with Benjamini–Hochberg correction. **d**, Major–major combinations on different alleles within close proximity (*NRAS* G12–G13). **e**, Proportion of MMs showing concordant or discordant allele frequencies (with distance between mutations of less than 25 bp) according to allelic configuration (cis versus trans). **b**, **e**, Two-sided Fisher's exact test. **f**, Average mutant allele frequency of MMs showing concordant allele frequencies present in copy-number-neutral region according to allelic configuration (cis versus trans). Box plots show medians (lines), interquartile ranges (IQRs; boxes) and 1.5 × IQRs (whiskers). Two-sided Welch's *t*-test. Examined numbers are shown in parentheses.

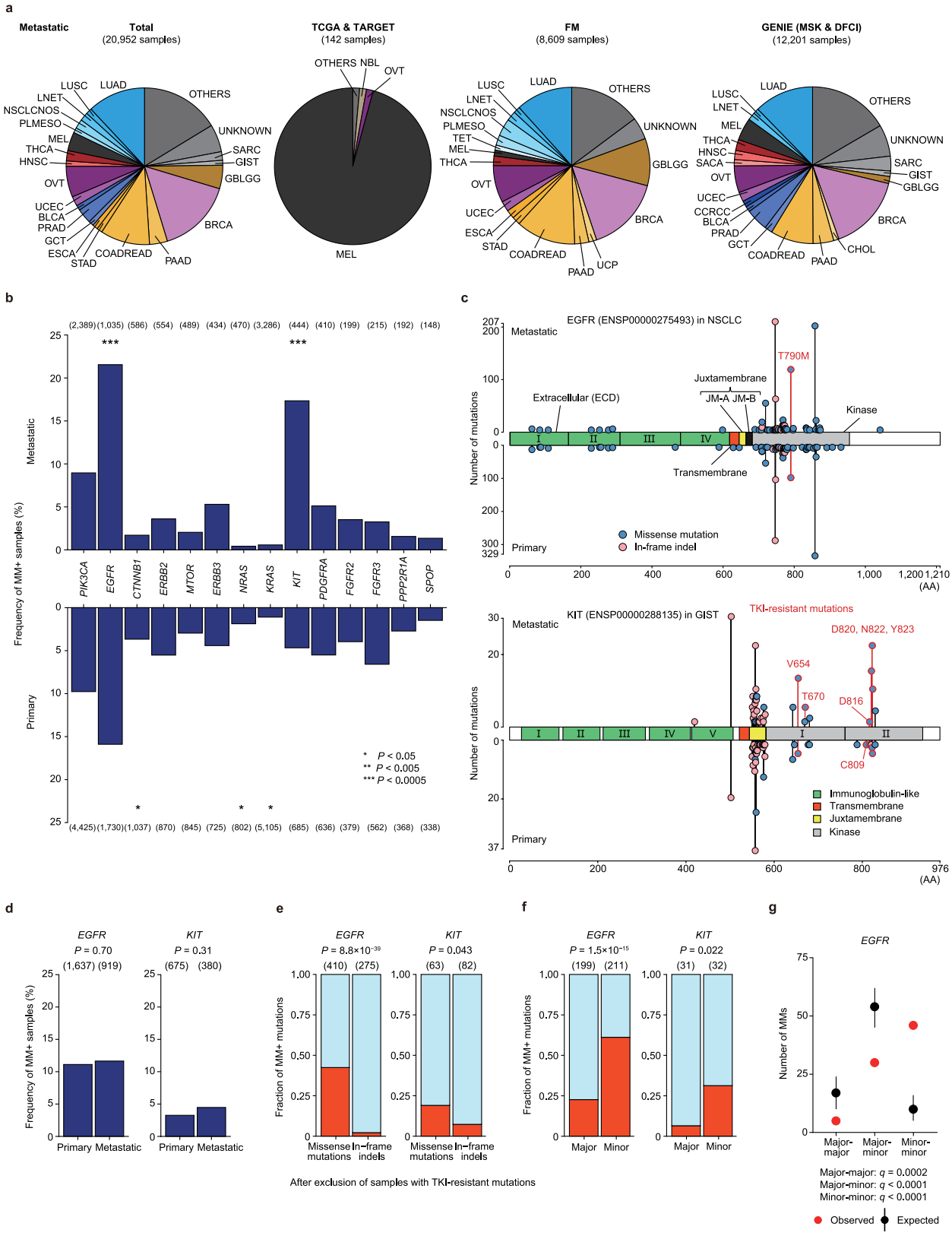


Extended Data Fig. 7 | Similarities and differences of oncogenic MMs across cancer types. a, Correlation between somatic mutation frequency per megabase (median value shown) and fraction of MM$^+$ samples according to MM$^+$ oncogenes in primary samples from the total cohort. Each dot represents a cancer type, coloured by gene. **b**, Distribution of mutations and/or fraction of MM$^+$ mutations for each hotspot/functional position in *PDGFRA* for GBLGG and GIST in primary samples from the total cohort. Asterisks indicate major positions (in which 10% or more of mutations were present in any of the recurrently mutated cancer types (defined as those with 20 or more hotspot/functional mutations)). **c**, Distribution of mutations and fraction of MM$^+$ mutations for each hotspot/functional position in *PIK3CA* for BRCA, COADREAD and UCEC in primary samples from the total cohort. Asterisks indicate major positions. The horizontal blue dotted lines represent the mean values of major positions. **d**, **e**, Distribution of hotspot/functional mutations (**d**) and fraction of MM$^+$ mutations occurring at major positions and minor positions in the ECD I–II and kinase (exon 18–21) domains (**e**) in *EGFR* for NSCLC and GBLGG in primary samples from the total cohort. **b**, **c**, **e**, Two-sided Fisher's exact test. Examined numbers are shown in parentheses.



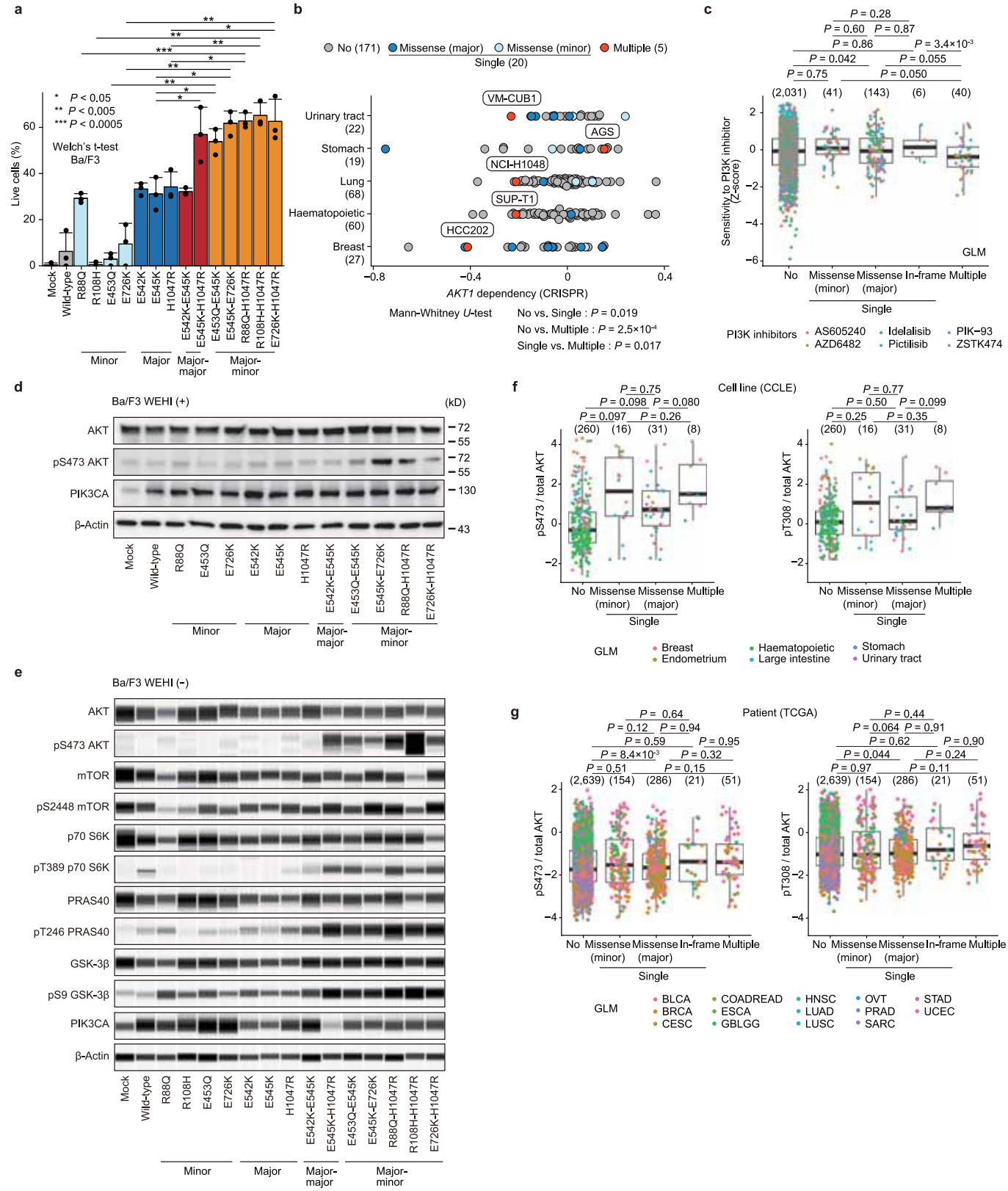
Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | Cancer-type-specific oncogenic MMs. a, Distribution of *FGFR3* mutations in samples with single and multiple mutations in recurrently mutated cancer types (defined as those with 20 or more hotspot/functional mutations) in primary samples from the total cohort. Mutations specific to MM^+ samples are indicated in red. **b**, Cancer-type-specific analysis of MM^+ oncogenes identified by pan-cancer analysis. Shown are numbers of MM^+ samples for MM^+ oncogenes in the corresponding cancer type from the discovery cohort (with $q < 0.01$ and three or more MMs). Red and black circles indicate, respectively, observed and expected (median with 95% confidence intervals) values. One-sided permutation test ($n = 10{,}000$) with Benjamini–Hochberg correction. **c**, Proportion of mutations according to type, position and amino-acid change, in six cancer-type-specific MM^+ oncogenes in the corresponding cancer type. Asterisk, AML, and dagger, BALL were analysed in primary samples from the total cohort and (an) independent set(s) of patients[29–33]. **d**, **e**, Distribution of hotspot/functional mutations (**d**) and frequency of MM^+ samples (**e**) in *NOTCH1* for TALL and CLL in primary samples from the total cohort. An independent cohort[34] was also analysed for CLL. **f**, Significant pairwise associations ($q < 0.01$) with observed/expected ratios among functional domains ($n = 65$). Orange and blue colours depict co-occurring (observed number of MMs significantly higher than expected) and mutually exclusive (lower than expected) associations. Two-sided simulation test ($n = 10{,}000$) with Benjamini–Hochberg correction. **g**, Fraction of MM^+ mutations for missense mutations and in-frame indels (consisting mainly of internal tandem duplications, ITDs) as well as distribution of mutations and fraction of MM^+ mutations for each hotspot/functional position (the top ten are shown) in *FLT3* for AML in primary samples from the total cohort and an independent set of patients. **h**, Distribution of mutations and fraction of MM^+ mutations for each hotspot/functional position in *JAK2* for BALL in primary samples from the total cohort and independent sets of patients. **g**, **h**, Asterisks indicate major positions (in which 10% or more of all mutations were present). **e**, **g**, **h**, Two-sided Fisher's exact test. Examined numbers are shown in parentheses.



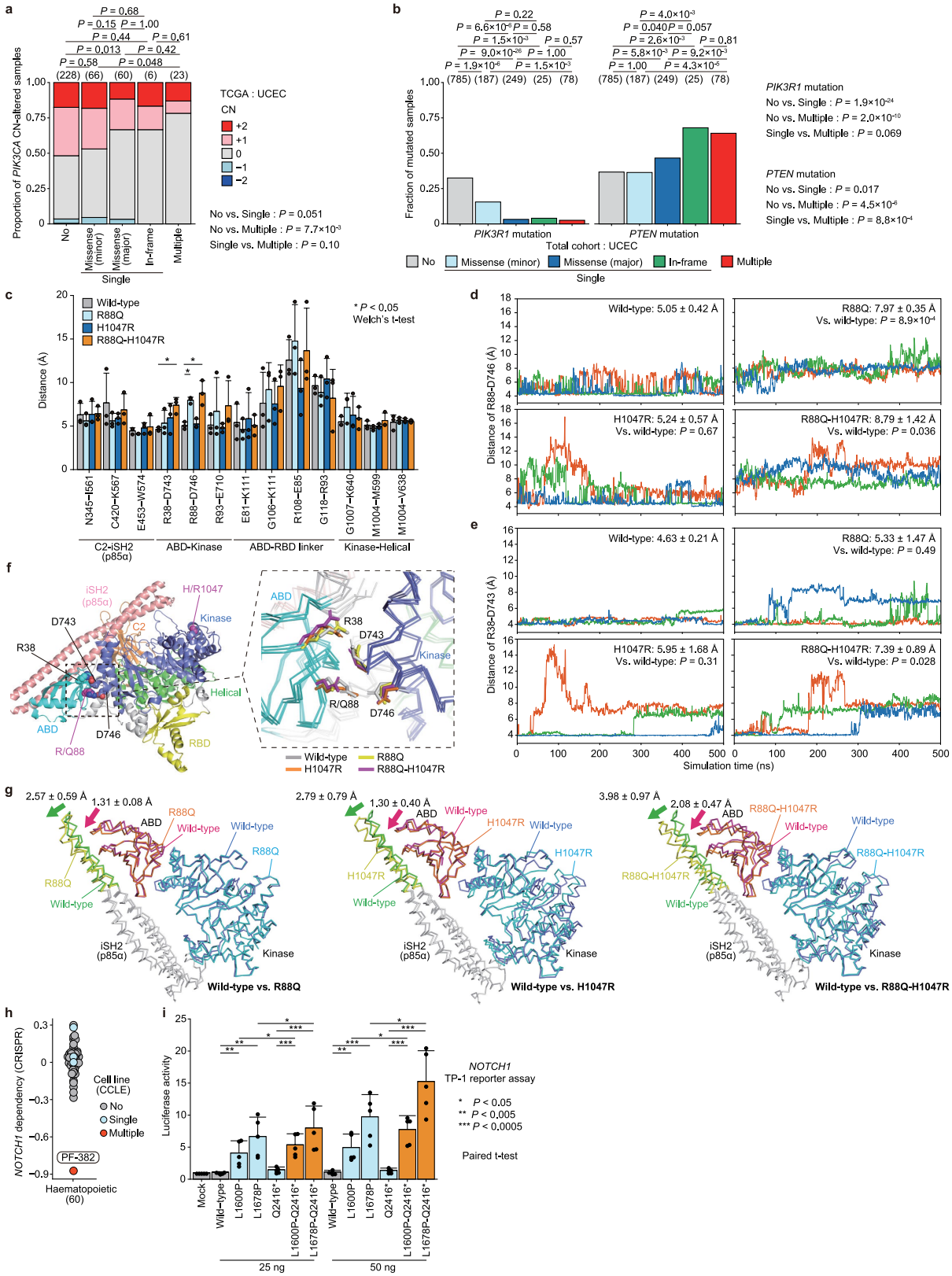
Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | Comparison of oncogenic MMs between primary and metastatic samples across cancers. a, Distribution of cancer types for metastatic samples included in the total, TCGA and TARGET, FM and GENIE cohorts. Cancer types with less than 1% frequency were combined as 'OTHERS' in each cohort. Cancer-type abbreviations are listed in Supplementary Table 1. **b**, Comparison of the frequency of MM^+ samples in primary and metastatic samples across 14 MM^+ oncogenes in the total cohort. **c**, Distribution of *EGFR* and *KIT* hotspot/functional mutations in primary and metastatic samples in NSCLC and GIST, respectively. Acquired TKI-resistant mutations are indicated in red. **d**, Frequency of MM^+ samples for *EGFR* and *KIT* in primary and metastatic samples. **e**, **f**, Fraction of MM^+ mutations for missense mutations and in-frame indels (**e**) and major and minor hotspots (**f**) in *EGFR* and *KIT* in recurrently mutated cancer types in metastatic samples. **b**, **d**–**f**, Two-sided Fisher's exact test. **g**, Number of MMs according to mutational combinations in *EGFR* in recurrently mutated cancer types in metastatic samples ($n = 81$). Red and black circles indicate, respectively, observed and expected (median with 95% confidence intervals) values. Two-sided simulation test ($n = 10{,}000$) with Benjamini–Hochberg correction. **d**–**g**, After exclusion of acquired TKI-resistant mutations, including *EGFR* T790M in NSCLC and *KIT* V654, T670, C809, D816, D820, N822 and Y823 missense mutations in GIST. Examined numbers are shown in parentheses.



Extended Data Fig. 10 | See next page for caption.

Extended Data Fig. 10 | Biological relevance of *PIK3CA* MMs. a, Viability of Ba/F3 cells expressing mock control or wild-type, single-mutant or double-mutant *PIK3CA* without WEHI-3-conditioned medium for three days ($n = 3$). Data represent means + s.d. Two-sided Welch's *t*-test. **b**, Dependency of 196 CCLE cell lines on *AKT1*, with cell lines coloured by *PIK3CA* MM status, in the DepMap CRISPR–Cas9 knockout data (see Methods). Five cancer types for which there are cell lines harbouring *PIK3CA* MMs were analysed by two-sided Mann–Whitney *U*-test. **c**, Box plots showing sensitivity to six PI3K inhibitors (*Z*-scores) for 417 CCLE cell lines, according to MM status. Five cancer types for which there are cell lines harbouring *PIK3CA* MMs were analysed. **d**, Immunoblot analysis of AKT, pS473 AKT, PIK3CA and β-actin in Ba/F3 cells expressing mock control or wild-type, single-mutant or double-mutant *PIK3CA* in the presence of WEHI-3-conditioned medium. Representative of three independent experiments. **e**, Capillary-based immunoassay of AKT, pS473 AKT, mTOR, pS2448 mTOR, p70 S6K, pT389 p70 S6K, PRAS40, pT246 PRAS40, GSK-3β, pS9 GSK-3β, PIK3CA and β-actin in Ba/F3 cells expressing mock control or wild-type, single-mutant or double-mutant *PIK3CA* without WEHI-3-conditioned medium. Representative of two independent experiments. **d**, **e**, See Supplementary Fig. 1 for source images. **f**, **g**, Box plots showing levels of phosphorylated S473 (pS473; left) and T308 (pT308; right) in RPPA data relative to total AKT protein expression for 316 CCLE cell lines (with cancer types for which there are cell lines harbouring *PIK3CA* MMs) (**f**) and in 3,164 TCGA patients (with purity of 50% or more, in recurrently mutated cancer types) (**g**) according to MM status. Each dot represents a sample, coloured by cancer type. **c**, **f**, **g**, Single mutations were classified into missense mutations in major and minor hotspot positions, in-frame indels, and non-hotspot/functional mutations, some of which are not shown owing to small numbers. Box plots show medians (lines), interquartile ranges (IQRs; boxes) and 1.5 × IQRs (whiskers). GLM, generalized linear model (see Methods). Examined numbers are shown in parentheses.



Extended Data Fig. 11 | See next page for caption.

Extended Data Fig. 11 | Molecular mechanism underlying the enhanced functional activity of oncogenic MMs. a, Fraction of *PIK3CA* copy-number (CN) alterations according to *PIK3CA* MM status for 385 UCEC patients in TCGA. **b**, Fraction of *PIK3R1* and *PTEN* mutations according to *PIK3CA* MM status for 1,339 UCEC patients in the total cohort. **a**, **b**, Two-sided Fisher's exact test. Single mutations were classified into missense mutations in major and minor hotspot positions, in-frame indels, and non-hotspot/functional mutations, some of which are not shown owing to small numbers. Examined numbers are shown in parentheses. **c**, The distance between key residues of the p110α/p85α complex[16]. Data represent means + s.d. (250–500 ns). **d**, Distance between R88 and D746 atoms. **e**, Distance between R38 and D743 atoms. **d**, **e**, Data represent means ± s.d. (250–500 ns). **c**–**e**, Two-sided Welch's *t*-test. **d**, **e**, Different colours show independent simulations. **f**, Overall structure of wild-type PIK3CA (left) and a close-up view of the mean structure at the ABD–kinase interface (right). **g**, Mean backbone structures of R88Q, H1047R and R88Q–H1047R mutants, showing the orientation of the ABD, kinase and iSH2 (p85α) domains, superimposed on those of the wild type. Movements of the ABD and iSH2 domains induced by each mutant are indicated by arrows, along with the backbone root mean square deviation (r.m.s.d.) of the ABD (residues 16–105; magenta) and iSH2 (490–540; green) domains from the wild-type structure. **c**–**g**, Molecular-dynamics simulations for wild-type and each mutant PIK3CA (combined results from three independent 500-ns simulations are shown). **h**, Dependency of 60 CCLE cell lines from haematopoietic and lymphoid tissue on *NOTCH1*, with cell lines coloured by MM status, in the DepMap CRISPR–Cas9 knockout data. **i**, Luciferase assays of TP-1 reporter activity in 293T cells transfected with mock control or the indicated amounts of wild-type, single-mutant or double-mutant *NOTCH1* vectors ($n = 5$). Data represent means + s.d. Two-sided paired *t*-test on log-transformed values.

nature research

Corresponding author(s): Keisuke Kataoka

Last updated by author(s): Feb 5, 2020

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

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	ImageQuant LAS 4000 control software version 1.2 Molecular Operating Environment (MOE) program version 2016.8 GROMACS version 2019.1 Compass software version 4.1.0 VICTOR Nivo Control Software version 3.0.2
Data analysis	LiftOver (https://genome.ucsc.edu/cgi-bin/hgLiftOver) Variant Effect Predictor version 95.3 vcf2maf version 1.6.16 SAMtools version 1.4.1 Minimap2 version 2.14 NanoQC version 0.9.0 Integrative Genomics Viewer (IGV) version 2.4.10 Python version 2.7.15 R version 3.6.0 Genomon version 2.6.2 Custom script for the permutation test (https://github.com/nccmo/Permutation-test) CisChecker (https://github.com/nccmo/CisChecker/) DNVChecker (https://github.com/nccmo/DNVChecker) pyMOL version 2.2.0 Gnuplot version 4.6.2

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

The findings of this study are supported by data that are available from public online repositories and data that are publicly available upon request of the data provider. See Methods for detail. Long-read WGS data of cell lines have been deposited in the European Genome-phenome Archive (EGA) under accession EGAS00001003763. Data generated in the current study are available as Source Data files that accompany Fig.4 and Extended Data Figs.10-11.

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

Life sciences study design

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

Sample size	No statistical methods were used to predetermine sample size. Sample size was determined by the availability of sequencing data. We enrolled all samples from public datasets.
Data exclusions	Samples with missing data, derived from xenografts, and duplicated in the same patient were excluded from the analysis. Hypermutator samples [with ≥500 and ≥20 coding mutations per exome (for the discovery cohort) and targeted region (for the additional cohort), respectively] were also removed from the analysis to minimize the effect of confounding passenger mutations. The specific exclusion criteria were not pre-established, although the cutoff values were determined based on previous publications (example: PMID: 29056346). No data were excluded for in vivo and in vitro experiments.
Replication	No experimental replication was performed for sequencing experiments. Molecular dynamics simulations were performed three times with different velocities. For in vivo and in vitro experiments, all attempts at replication were successful with biological replicates performed on separate cohorts of animals/cells.
Randomization	Samples were assigned to each group based on the number of mutations (samples with no mutation, single mutation, or multiple mutations in the gene of interest). Therefore, randomization was not required.
Blinding	Blinding was not relevant to the study, as there was no control and treatment arms involved.

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
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology
☐	☒ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	Rabbit anti-PI3 Kinase p110α (C73F8) (Cell Signaling Technology, #4249, dilution 1:1000 [IB] or 1:50 [CI]) Rabbit anti-β-Actin (13E5) (Cell Signaling Technology, #4970, dilution 1:2000 [IB] or 1:50 [CI]) Rabbit anti-Akt (pan) (C67E7) (Cell Signaling Technology, #4691, dilution 1:1000 [IB] or 1:50 [CI]) Rabbit anti-phospho-Akt (Ser473) (D9E) (Cell Signaling Technology, #4060, dilution 1:2000 [IB] or 1:50 [CI]) Rabbit anti-mTOR (7C10) (Cell Signaling Technology, #2983, dilution 1:50 [CI])

Rabbit anti-phospho-mTOR (Ser2448) (D9C2) (Cell Signaling Technology, #5536, dilution 1:50 [CI])
Rabbit anti-p70 S6 Kinase (Cell Signaling Technology, #9202, dilution 1:50 [CI])
Rabbit anti-phospho-p70 S6 Kinase (Thr389) (Cell Signaling Technology, #9205, dilution 1:50 [CI])
Rabbit anti-PRAS40 (Cell Signaling Technology, #2610, dilution 1:50 [CI])
Rabbit anti-phospho-PRAS40 (Thr246) (C77D7) (Cell Signaling Technology, #2997, dilution 1:50 [CI])
Rabbit anti-GSK-3β (27C10) (Cell Signaling Technology, #9315, dilution 1:50 [CI])
Rabbit anti-Phospho-GSK-3β (Ser9) (5B3) (Cell Signaling Technology, #9323, dilution 1:50 [CI])
IB, immunoblot; CI, capillary-based immunoassay.

Validation

Antibodies with prior manufacturer validation were used. Validation statements at manufacturer's website were as follows:
Rabbit anti-PI3 Kinase p110α (C73F8) (Cell Signaling Technology, #4249): Western blot analysis of extracts from HeLa cells and neonatal mouse brain using PI3 Kinase p110α (C73F8) Rabbit mAb. Currently over 224 citations.
Rabbit anti-β-Actin (13E5) (Cell Signaling Technology, #4970): Western blot analysis of cell extracts from various cell lines (NIH/3T3, HeLa, PAE, A431) using beta-Actin (13E5) Rabbit mAb. Currently over 1,861 citations.
Rabbit anti-Akt (pan) (C67E7) (Cell Signaling Technology, #4691): Western blot analysis of recombinant Akt1, Akt2 and Akt3 proteins, and extracts from various cell lines (HeLa, NIH/3T3, C6, COS), using Akt (pan) (C67E7) Rabbit mAb. Currently over 1,851 citations.
Rabbit anti-phospho-Akt (Ser473) (D9E) (Cell Signaling Technology, #4060): Western blot analysis of extracts from PC-3 cells, untreated or LY294002/wortmannin-treated, and NIH/3T3 cells, serum-starved or PDGF-treated, using Phospho-Akt (Ser473) (D9E) XP® Rabbit mAb (upper) or Akt (pan) (C67E7) Rabbit mAb #4691 (lower). Currently over 3,926 citations.
Rabbit anti-mTOR (7C10) (Cell Signaling Technology, #2983): Western blot analysis of extracts from 293, A431, COS, C6, and C2C12 cells, using mTOR (7C10) Rabbit mAb. Currently over 970 citations.
Rabbit anti-phospho-mTOR (Ser2448) (D9C2) (Cell Signaling Technology, #5536): Western blot analysis of extracts from serum-starved NIH/3T3 cells, untreated or insulin-treated (150 nM, 5 minutes), alone or in combination with λ-phosphatase, using Phospho-mTOR (Ser2448) (D9C2) XP® Rabbit mAb (upper) or mTOR (7C10) Rabbit mAb #2983. Currently over 650 citations.
Rabbit anti-p70 S6 Kinase (Cell Signaling Technology, #9202): Western blot analysis of extracts from HeLa, NIH-3T3, PC12 and COS-7 cells using p70 S6 Kinase Antibody. Currently over 939 citations.
Rabbit anti-phospho-p70 S6 Kinase (Thr389) (Cell Signaling Technology, #9205): Western blot analysis of HeLa, COS, C6 and 3T3 cells, serum-starved overnight, then treated with insulin, l-phosphatase or 20% serum as indicated. Upper panel probed with Phospho-p70 S6 Kinase (Thr389) Antibody #9205; lower panel probed with p70 S6 Kinase Antibody #9202. Currently over 904 citations.
Rabbit anti-PRAS40 (Cell Signaling Technology, #2610): Western blot analysis of extracts from various cell types (MCF-7, 293, HeLa, A204, RD, 3T3, RAW, KNRK, NBTII, and COS7) using PRAS40 Antibody. Currently over 40 citations.
Rabbit anti-phospho-PRAS40 (Thr246) (C77D7) (Cell Signaling Technology, #2997): Western blot analysis of extracts from serum starved H3255, Mkn45 and NIH/3T3 cells, untreated or treated with either Gefitinib (1 µM, 3 hours), Su11274 (1 µM, 3 hours) or insulin (150 nM, 15 minutes), using Phospho-PRAS40 (Thr246) (C77D7) Rabbit mAb (upper) or PRAS40 (D23C7) Rabbit mAb #2691 (lower). Currently over 119 citations.
Rabbit anti-GSK-3β (27C10) (Cell Signaling Technology, #9315): Western blot analysis of extracts from HeLa, NIH/3T3, COS, C6 and 293 cells using GSK-3β (27C10) Rabbit mAb. Currently over 666 citations.
Rabbit anti-Phospho-GSK-3β (Ser9) (5B3) (Cell Signaling Technology, #9323): Western blot analysis of extracts from NIH/3T3 cells, λ-phosphatase- or PDGF-treated, using Phospho-GSK-3β (Ser9) (5B3) Rabbit mAb (upper) or GSK-3β (27C10) Rabbit mAb #9315 (lower). Currently over 294 citations.
See https://www.citeab.com/ for details of relevant citations.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	Ba/F3-CL1, WEHI-3, and 293T cell lines were obtained from the RIKEN Cell Bank, HEC-1 cell line from the JCRB Cell Bank, and MCF10A, BT-20, NCI-H1048, SUP-T1, and HRT-18 cell lines from ATCC. Lenti-X 293T cells were purchased from TaKaRa.
Authentication	All cell lines were authenticated by the providers using karyotype, isoenzmes, and/or microsatellite profiling (short tandem repeat or simple sequence length polymorphism).
Mycoplasma contamination	We confirmed that all cell lines were negative for mycoplasma contamination using MycoAlertTM Mycoplasma Detection Kit (Lonza, LT07-318).
Commonly misidentified lines (See ICLAC register)	According to ICLAR register, cross-contamination was reported in BT-20 in 1976 (Reference PubMed ID: 6451928), although authentic stocks apparently do exist. We used the BT-20 cell line authenticated by ATCC using short tandem repeat (STR) profiling analysis. This cell line was selected because of the limited number of available cell lines harboring multiple mutations in PIK3CA.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	Female BALB/c-nu/nu mice (6 weeks old)
Wild animals	The study did not involve wild animals.
Field-collected samples	The study did not involve samples collected from the field.

Ethics oversight	All mouse experiments were approved by the Animal Ethics Committee of the National Cancer Center and strictly adhered to its guidelines.

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