RESEARCH BRIEF

Accelerating Discovery of Functional Mutant Alleles in Cancer

Matthew T. Chang[1,2,3], Tripti Shrestha Bhattarai[1,2], Alison M. Schram[4], Craig M. Bielski[5], Mark T.A. Donoghue[5], Philip Jonsson[1,2], Debyani Chakravarty[5], Sarah Phillips[5], Cyriac Kandoth[5], Alexander Penson[1,2], Alexander Gorelick[1,2], Tambudzai Shamu[1,2], Swati Patel[1], Christopher Harris[5], JianJiong Gao[5], Selcuk Onur Sumer[5], Ritika Kundra[5], Pedram Razavi[4], Bob T. Li[4], Dalicia N. Reales[5], Nicholas D. Socci[5,6], Gowtham Jayakumaran[7], Ahmet Zehir[7], Ryma Benayed[7], Maria E. Arcila[7], Sarat Chandarlapaty[1,4], Marc Ladanyi[7], Nikolaus Schultz[2,5], José Baselga[4], Michael F. Berger[5,7], Neal Rosen[4,8], David B. Solit[1,4,5,9], David M. Hyman[4], and Barry S. Taylor[1,2,5]

ABSTRACT Most mutations in cancer are rare, which complicates the identification of therapeutically significant mutations and thus limits the clinical impact of genomic profiling in patients with cancer. Here, we analyzed 24,592 cancers including 10,336 prospectively sequenced patients with advanced disease to identify mutant residues arising more frequently than expected in the absence of selection. We identified 1,165 statistically significant hotspot mutations of which 80% arose in 1 in 1,000 or fewer patients. Of 55 recurrent in-frame indels, we validated that novel *AKT1* duplications induced pathway hyperactivation and conferred AKT inhibitor sensitivity. Cancer genes exhibit different rates of hotspot discovery with increasing sample size, with few approaching saturation. Consequently, 26% of all hotspots in therapeutically actionable oncogenes were novel. Upon matching a subset of affected patients directly to molecularly targeted therapy, we observed radiographic and clinical responses. Population-scale mutant allele discovery illustrates how the identification of driver mutations in cancer is far from complete.

SIGNIFICANCE: Our systematic computational, experimental, and clinical analysis of hotspot mutations in approximately 25,000 human cancers demonstrates that the long right tail of biologically and therapeutically significant mutant alleles is still incompletely characterized. Sharing prospective genomic data will accelerate hotspot identification, thereby expanding the reach of precision oncology in patients with cancer. *Cancer Discov; 8(2); 174–83. ©2017 AACR.*

INTRODUCTION

The rapid adoption of prospective clinical tumor sequencing (1–4) has led to the identification of an increasing number of somatic mutations of unknown significance. Although a small number of mutations are used to guide treatment selection, the vast majority lack biological or clinical validation, limiting the ability of clinicians to use tumor genomic data to guide therapy selection (5, 6). Indeed, such mutations are often presumed to be passenger mutations with no evidence to support

[1]Human Oncology and Pathogenesis Program, Memorial Sloan Kettering Cancer Center, New York, New York. [2]Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, New York. [3]Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, San Francisco, California. [4]Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, New York. [5]Marie-Josée and Henry R. Kravis Center for Molecular Oncology, Memorial Sloan Kettering Cancer Center, New York, New York. [6]Bioinformatics Core, Memorial Sloan Kettering Cancer Center, New York, New York. [7]Department of Pathology, Memorial Sloan Kettering Cancer Center, New York, New York. [8]Molecular Pharmacology Program, Memorial Sloan Kettering Cancer Center, New York, New York. [9]Department of Medicine, Weill Cornell Medical College, Cornell University, New York, New York.

Note: Supplementary data for this article are available at Cancer Discovery Online (http://cancerdiscovery.aacrjournals.org/).

Current address for M.T. Chang: Genentech Inc., South San Francisco, California.

Corresponding Author: B.S. Taylor, Memorial Sloan Kettering Cancer Center, 1275 York Avenue, Box 20, New York, NY 10065. Phone: 646-888-3728; Fax: 646-888-3266; E-mail: taylorb@mskcc.org

doi: 10.1158/2159-8290.CD-17-0321

such a classification. As a consequence, many patients whose tumors harbor mutations of unrecognized clinical significance are not being offered potentially beneficial therapies, and this knowledge gap represents one of the fundamental hurdles to the broader adoption of precision oncology today.

Given the sheer number of mutations of uncertain significance, there is an urgent need to identify and prioritize for biological and clinical study of potentially druggable driver mutations identified within the context of prospective tumor profiling. Unraveling the relationships among different mutant alleles, their comutational patterns and the cell types in which they selectively arise will be critical to defining their function and clinical actionability (7), essential steps to expanding the treatment options for patients with cancer. Historically, however, the incremental laboratory and then clinical validation of novel mutations as sensitizing biomarkers of response or resistance to standard or investigational therapies can take years, preventing current patients from potentially benefiting from such therapies.

Here, we defined driver mutations in the long right tail of somatic mutations in cancer and developed an exploratory framework by which computational weight of evidence alone was utilized in real time to prioritize treatment-refractory patients harboring novel hotspot mutations of uncertain clinical significance for studies of molecularly targeted therapies. In a subset of patients, clinical response rather than laboratory interrogation was employed as the most expedient approach for clinical validation of mutant alleles of unknown function. No example exists to our knowledge where the identification of a novel mutation of likely significance has taken place in the same population used to validate the mutant allele as sensitizing to therapy, a potential acceleration of the clinical validation of variants of unknown significance as putative sensitizing biomarkers.

RESULTS

To identify novel hotspot mutations of biological and potentially therapeutic importance, we analyzed somatic mutational data from 24,592 patients. This cohort consisted of 14,256 retrospectively sequenced predominantly primary untreated human cancers and 10,336 prospectively sequenced patients with active, advanced cancer with recurrent and metastatic disease (43% of specimens were obtained from metastatic sites; see Methods; ref. 8). These samples represent 322 cancer types spanning 32 organ sites, the annotation of which was standardized to conform to an open-source structured disease classification (Supplementary Table S1; http://oncotree.mskcc.org/). We analyzed each of the 32 organ sites independently as well as the full cohort (pan-cancer) to enhance the probability of identifying hotspots that occur rarely in multiple organ types of different mutational burdens and processes (9). To do this, organ type–specific, gene-specific, and context-specific background mutation rates were computed (Supplementary Table S2, see Methods). We also developed a first-of-its-kind computational approach to identify hotspots of candidate oncogenic small in-frame insertions or deletions, which are more challenging to identify than substitutions due to the variability of mutant allele length and position from tumor to tumor.

We identified 1,165 mutational hotspots in 247 genes (1,110 single-codon and 55 indel; median of 2 hotspots per gene; range 1–120; q value < 0.1, false discovery rate of 10%; Supplementary Tables S3 and S4). This analysis recovered nearly all previously identified hotspots (ref. 10; 97%) and identified 840 additional hotspots, reflecting an increased power of detection as well as a more clinically diverse cohort of patients. In total, 5% of these 840 hotspots were identified in cancer types new to this analysis, emphasizing the value of a clinically diverse dataset to power hotspot discovery. The vast majority of hotspots observed here for the first time were due to the large increase in sample size over prior studies, emphasizing how the characterization of the long right tail of the curve of driver mutations in cancer was incomplete. Indeed, the frequency distribution of hotspot-mutant genes had a long right tail (10), the shape of which was independent of the count of unique hotspots in the gene and was different between single-codon and indel hotspots (Fig. 1A and Supplementary Fig. S1). Although the majority of hotspots (n = 596; 51% of total) were statistically significant both pan-cancer and within individual organ types (Fig. 1B), 20% and 29% of hotspots were significant only within an individual organ type or only in the pan-cancer analysis of the full cohort, respectively (Fig. 1C). Many of the mutant alleles identified, both in long-established cancer genes (such as *PIK3CA, MTOR, ERBB2,* and *MAP2K1*) and in genes more recently implicated (such as *CYSLTR2*; ref. 11), were novel, reflecting the greater sensitivity for rare allele discovery with increasing cohort size in even well-characterized cancer genes (Fig. 1B).

Forty-two percent of the patients in this cohort were prospectively sequenced at our institution as part of their clinical care and had advanced and/or previously treated disease. This clinical profile is distinct from that of patients with untreated primary tumors, from which most of the data in the literature are obtained. The inclusion of such patients facilitated the identification of hotspots present almost exclusively in the metastases of treatment-refractory patients. Eleven hotspots were enriched in metastatic disease compared with the primary tumors of a given cancer type (see Methods), nine of which were therapy-associated, arising in specimens after treatment with either antiandrogen, antiestrogen, or tyrosine kinase inhibitor therapies (Supplementary Fig. S2). Notably, some therapy-resistant hotspots were found to arise in treatment-naïve tumors of other cancer types (such as *KIT*^{D820}), suggesting that treatment-associated resistance mutations in one cancer type can arise *de novo* in the absence of therapy as the primary oncogenic driver in another cancer type (12). Other hotspots may reflect new mechanisms of resistance to traditional cytotoxic chemotherapies, such as *TP53*^{N239}, which confers paclitaxel resistance *in vitro* (13) and was the only *TP53* hotspot that arose preferentially in metastatic breast cancers (q value = 0.03), all obtained from tumors that developed resistance to, or rapidly progressed on, taxane-based therapy. Together, these analyses identify a broader range of hotspots than previously recognized and for which biological and clinical study (14, 15) may accelerate clinical translation.

Although substitutions are the most abundant class of mutation in cancer genomes, several recurrent, activating in-frame indels are validated predictive biomarkers of sensitivity to targeted therapies, including indels in exon 19 of *EGFR* in

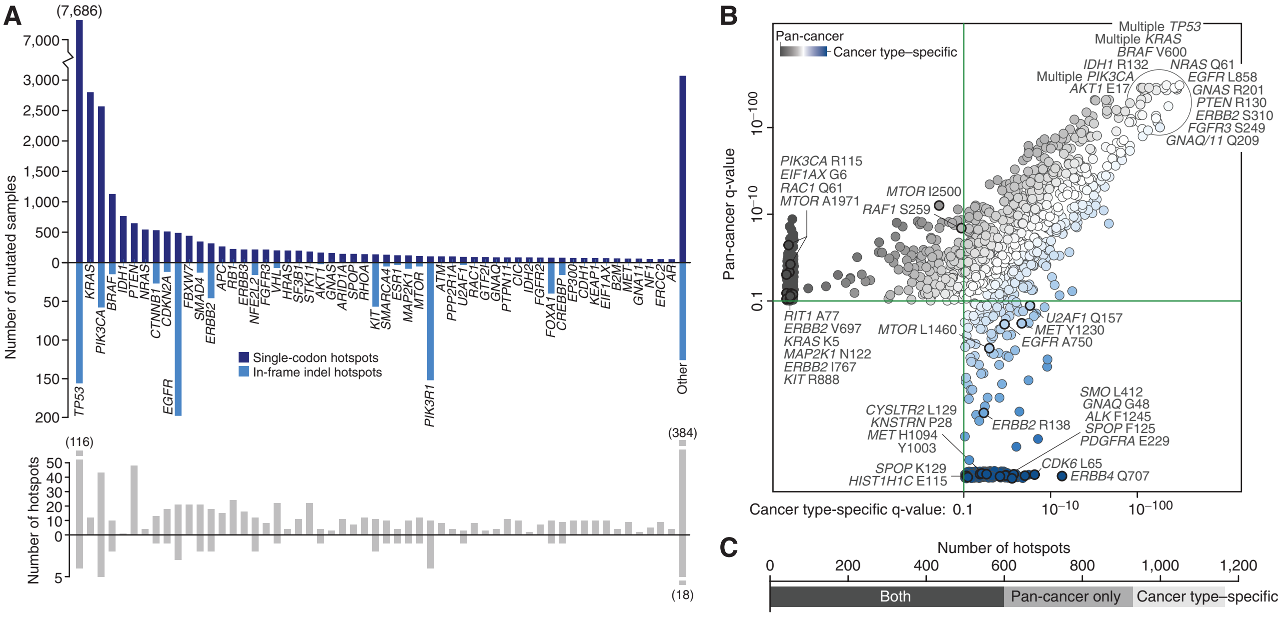

Figure 1. The long tail of mutational hotspots in cancer. **A,** The frequency distribution of genes containing one or more single-codon hotspots (top, dark blue) and in-frame indel hotspots (light blue). At bottom, the count of single-codon and in-frame indel hotspots in the same genes. **B,** Shown is the statistical significance of mutational hotspots inferred from the analysis of the full cohort (pan-cancer, *y* axis) and the most significant individual cancer type (*x* axis). A subset of hotspots are annotated (circled in black) and include mutations significant in both analyses (top right), those significant only when combining all cancer types and data (leftmost), and those significant only within a given cancer type (bottom). **C,** The proportion of hotspots that were significant only in individual organ types, only in the pan-cancer analysis, or both.

lung adenocarcinomas and in exon 11 of *KIT* in gastrointestinal stromal tumors (16, 17). Yet, the unbiased discovery of recurrent oncogenic in-frame indels from population-scale data has not been done. We thus extended our methodology to identify hotspots (clusters) of in-frame indels (see Methods). In total, we identified 55 statistically significant indel hotspots in 36 genes (Supplementary Table S4 and Supplementary Fig. S3). There were 20-fold fewer indel hotspots identified than single-codon hotspots, and although deletions predominated (69%), duplications were enriched in oncogenes (P value < 0.01; Fig. 2A and B). Multiple indel hotspots in *EGFR, ERBB2, KIT,* and *BRAF* were identified (18, 19), as were other novel indel hotspots in *AKT, MTOR, PIK3CA, SRSF2, U2AF1*, and *MYC,* among others (Supplementary Table S4). Given the recently identified clinical activity of AKT inhibitors in AKT-mutant patients (20), we sought to determine whether the previously uncharacterized AKT indels were activating and potentially drug sensitizing. Specifically, we functionally characterized *AKT1* P68_C77dup, one of several paralogous indels found in a hotspot cluster in the pleckstrin homology domain of *AKT1* and *AKT2* (q values = 0.09 and 2×10^{-5}, respectively) proximal to known *AKT1* hotspots (L52 and Q79; q values < 10^{-4}; Fig. 2C and Supplementary Fig. S4). Expression of *AKT1* P68_C77dup in MCF10A cells resulted in a higher level of AKT phosphorylation (T308/S473; Fig. 2D), as well as increased phosphorylation of downstream effectors of AKT such as S6 and PRAS40, as compared with the two most common activating *AKT1* missense mutations, E17K and Q79K. Isogenic cells expressing AKT1 P68_C77dup also demonstrated greater sensitivity to the ATP-competitive pan-AKT kinase inhibitor AZD5363 than did cells expressing AKT1^{E17K} or wild-type AKT1 (Fig. 2E).

Other indels identified here may be associated with therapeutic resistance, such as *ESR1* V422del (q value = 0.045), which arose clonally after failure of antiestrogen therapy in the metastatic site of estrogen receptor–positive breast cancers that otherwise lacked the most common ligand binding domain hotspots E380, L536, Y537, and D538 (q values < 10^{-17}) that are known to confer resistance to estrogen deprivation therapies in breast cancer (ref. 21; Supplementary Fig. S3). Although most indels, like those in *AKT1*, spanned or were physically adjacent to single-codon hotspots in the same genes, indel hotspots in three genes were physically distant (greater than 15 Å) from substitution hotspots in their cognate folded protein. These indels included the aforementioned *ESR1* V422del, the well-characterized FLT3 internal tandem duplication (ITD), and a cluster of indels spanning I99 to I107 in *MAP2K1* (q value = 3.3×10^{-12}), which was 1 of 11 total *MAP2K1* hotspots identified here (Supplementary Tables S3 and S4). Structurally distant indels and single-codon hotspots may imply divergent biological effects, as is the case between ITD and kinase domain mutations in FLT3 (22), but the extent of such differences is unknown.

In BRAF, another effector of aberrant MAPK signaling, V600 mutations induce constitutive kinase activity independent of upstream activation, whereas BRAF^{D594} mutants are kinase dead but cooperatively amplify ERK signaling and tumorigenesis when coexpressed with mutant KRAS (23). Indeed, reflecting this divergent biological function, D594 mutations typically co-occur with activating RAS mutations in patients

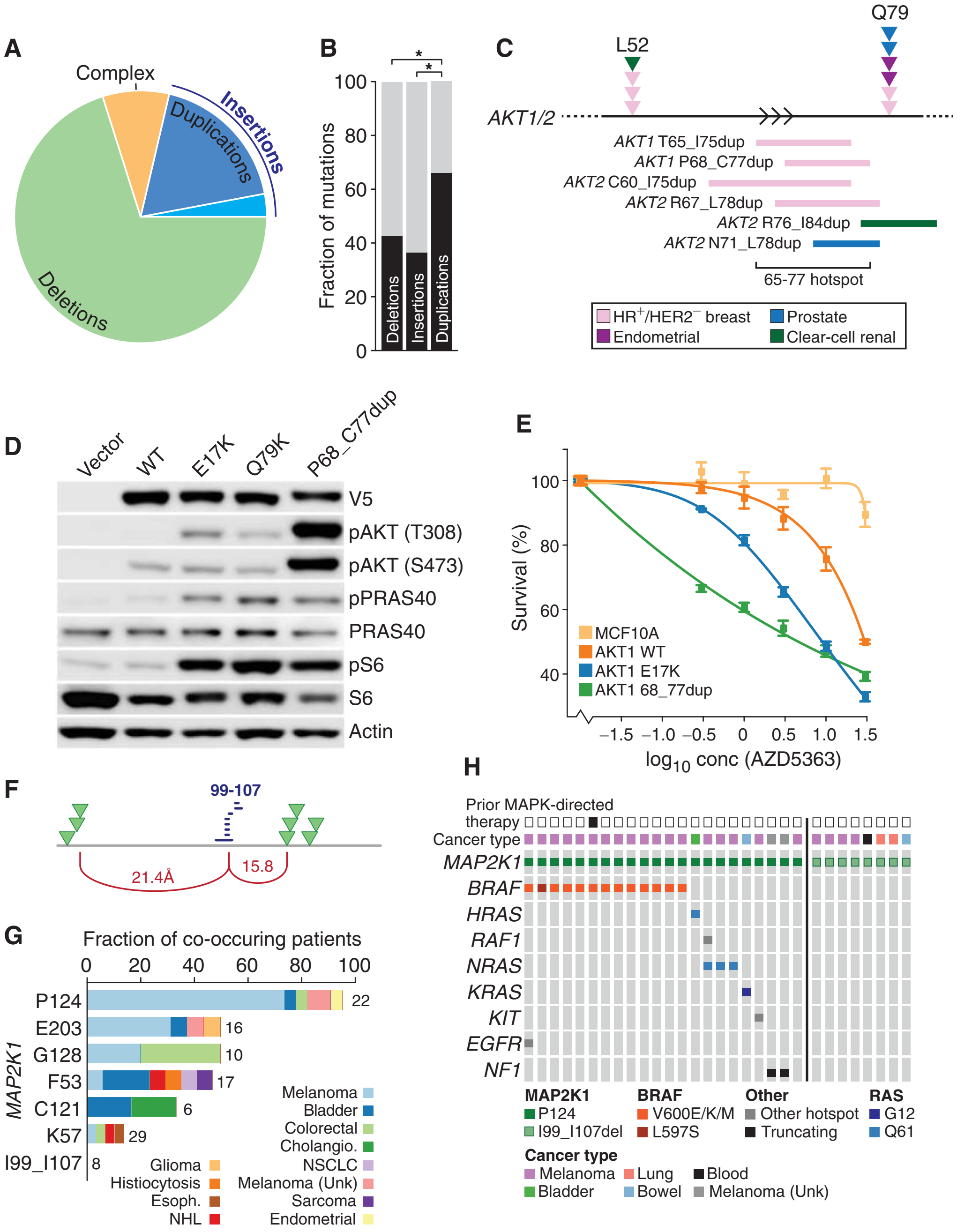

Figure 2. Oncogenic indel hotspots. **A,** The distribution of recurrent indel hotspot types discovered here. **B,** Duplications were significantly more common than either deletions or insertions in oncogenes (*, $P < 0.01$). **C,** The paralogous indels are shown defining the *AKT1* and *AKT2* duplication hotspot. The affected cancer types are similar to those that harbor known activating L52, and Q79, hotspot mutations and include hormone receptor (HR)–positive HER2-negative breast cancers that lack other PI3K pathway alterations. **D,** MCF10A cells stably expressing the indicated *AKT1* mutations are shown, and expression and/or phosphorylation levels were assayed by Western blot indicating that the AKT1 P68_C77dup induces elevated levels of phosphorylated AKT and S6 comparable with or exceeding that of known activating E17K or Q79K hotspots. **E,** Cell survival upon AKT blockade with AZD5363 in *AKT1*-mutant cells indicated that P68-C77dup–mutant cells were most sensitive to AKT inhibition, more so than the canonical E17K hotspot. **F,** Schematic of MAP2K1 from amino acids 60 to 140 indicates the position of single-codon hotspots (green arrows) is distal from the position of the indel hotspot (blue lines are individual indels in affected tumors). Arcing red lines reflect the distance in angstroms between the indels and single-codon hotspots in the protein structure. **G,** The rate of comutation with other MAPK effectors varied by *MAP2K1* hotspot, with P124 mutations always associated with upstream pathway activation and predominantly in melanomas, whereas others (E203, G128, F53, C121, and K57) were only partially comutated, and the *MAP2K1* indel hotspot never arose in tumors with another MAPK driver mutation. **H,** All but one *MAP2K1*^{P124}-mutant tumors possessed another known driver of MAPK signaling, of which most were *BRAF*^{V600E} (59% of total) and these and others were mostly cutaneous melanomas. Conversely, the *MAP2K1* I99_I107 indel hotspot never arose in an otherwise MAPK-altered tumor in a diversity of cancer types.

with melanoma, a pattern missed when all alterations in *BRAF* and *KRAS* are considered together. We therefore sought to determine whether the structurally distant indel and single-codon hotspots in *MAP2K1* may have differing function, as a guide to future functional studies. Using our allele-specific approach, we assessed the statistical significance of co-occurrence among hotspot mutations that arose together in individual tumors more frequently than expected by chance in pairs of genes in MAPK signaling. We identified multiple associations, the most significant of which was between *MAP2K1* and *BRAF* (Supplementary Fig. S5). The pattern and frequency of *MAP2K1* comutational associations varied in an allele-specific manner (Fig. 2G). For example, *MAP2K1*$^{\mathrm{P124}}$ was nearly always comutated with an upstream activating mutation in the MAPK pathway (95%), most often *BRAF*$^{\mathrm{V600E}}$ (55%; Fig. 2H). Conversely, the *MAP2K1* indel hotspot newly identified here arose in a mutually exclusive pattern with other MAPK lesions in affected tumors independent of cancer type. This pattern was not attributable to acquired resistance to MAPK pathway inhibitors, as only one such tumor was sequenced after RAF or MEK inhibitor failure. Overall, these results illustrate how novel computational methodologies can identify previously occult oncogenic in-frame indels of biological and potential therapeutic importance. Moreover, these data indicate that a broader analysis of coincident mutational patterns in multiple pathway effectors can uncover potential allele-specific functional differences that may be missed by gene-level analyses and may condition distinctive signaling biology requiring deeper mechanistic investigation.

Reflecting the long right tail of mutation frequencies in human cancer, 82% of all hotspot mutations were identified in 1 in 1,000 or fewer patients. To assess the future impact of larger cohort size on novel hotspot detection, we performed repeated random downsampling of increasing subsets of the cohort to infer the anticipated rate of future hotspot identification. Principal component analysis of gene-specific rates revealed four distinct classes of genes (saturating, fatiguing, linear, accelerating) that accrued their recurrent mutations, independent of their overall mutational burden, in different patterns with considerable variability from gene to gene (Fig. 3A and Supplementary Fig. S6). We also estimated the number of additional tumors that would need to be sequenced from a cohort of similar cancer type composition to identify the next incremental hotspot in each gene and cluster. One cluster was defined by canonical oncogenes (*IDH1, K/N/HRAS, GNAQ, MYD88*) whose most prevalent hotspots could be identified from the analysis of few samples but, as genes, are approaching saturation and thus additional sequencing is not expected to yield many additional novel hotspots. Indeed, we estimated that an additional 10,000 or more tumors sequenced would be necessary to identify another novel hotspot in many of the genes in this saturating cluster (Fig. 3B). The identification of hotspots in genes in the second cluster initially increased rapidly with increasing cohort size, but their rate is fatiguing yet not saturating, indicating that additional rare alleles will continue to be discovered as additional tumor genomic data become available. Notably, many of the genes in the fatiguing cluster are therapeutically actionable genes such as *BRAF, PIK3CA, ESR1, AKT1,* and *ERBB2,* and thus the identification of novel hotspots in these genes could have immediate clinical implications. The third cluster of genes is still in a linear phase of hotspot identification, and additional sequencing should continue to reveal additional new, albeit uncommon, hotspots in these genes, many of which are therapeutically targetable oncogenes such as *KIT*. The fourth cluster is composed of genes (such as *MET* or *MTOR*) in which even the enormous quantity of sequencing to date has only begun to reveal rare hotspots of potential clinical significance. In this accelerating cluster, fewer than 1,000 additional specimens would be necessary to identify additional hotspots (Fig. 3B). These patterns have important implications for strategies to prioritize and understand emerging mutations and suggest that many additional hotspots could be identified by pooling existing prospectively sequenced tumors that currently reside in siloed institutional or commercial repositories.

The above analysis indicates that we are far from completing the identification of potentially actionable hotspots. As such, patients with functional but rare mutant alleles in targetable cancer genes are not being offered potentially beneficial matched therapies as a result of the unrecognized clinical significance of the mutations identified in their tumor. To determine the scope of such occult actionability, we utilized a curated knowledge base of the oncogenic effects and treatment implications of mutations (http://oncokb.org/; ref. 24) in 18 genes in which one or more mutations are used in current clinical practice to guide routine (FDA-approved or part of established practice guidelines) prescribing of targeted therapy or are being evaluated as investigational biomarkers (see Methods). Of the 196 hotspot mutations identified in these genes, only a minority have been investigated clinically (Fig. 3C), though patterns vary by gene (Supplementary Fig. S7). Fifty hotspots (26%) were newly discovered here, being neither annotated in OncoKB nor identified in further detailed literature review. Because these novel hotspots arise in genes for which targeted therapies are already available, we sought to test the therapeutic hypothesis that these mutations may be similarly sensitizing biomarkers by matching a subset of the affected, prospectively sequenced patients to molecularly targeted therapies. This patient-to-drug matching was performed in the absence of laboratory data confirming that such mutations were activating or sensitizing alleles.

As a proof of this concept, we identified 7 active patients at the time of this study analysis who harbored one of the novel rare hotspots identified here and enrolled them on existing clinical trials where the therapy was targeting the affected oncogene. All 7 patients derived clinical benefit from therapy including 2 patients with a novel *ERBB2*$^{\mathrm{V697}}$ hotspot that were treated with the pan-HER tyrosine kinase inhibitor neratinib (25). One *ERBB2*$^{\mathrm{V697}}$ patient was a patient with heavily pretreated triple-negative breast cancer (Fig. 3D), whereas the other had a cancer of unknown primary involving the head and neck that responded to neratinib monotherapy for 13 months. At the time of progression on therapy, we biopsied and sequenced a cutaneous metastasis, which revealed a clonally related postprogression tumor that lacked any evidence of the *ERBB2*$^{\mathrm{V697}}$ mutation, indicating that loss of the sensitizing mutation was sufficient to confer drug resistance. Four other patients had tumors harboring different novel or previously uncharacterized *PIK3CA* hotspot mutations (P104, T1025, V344, R38), each of which had durable

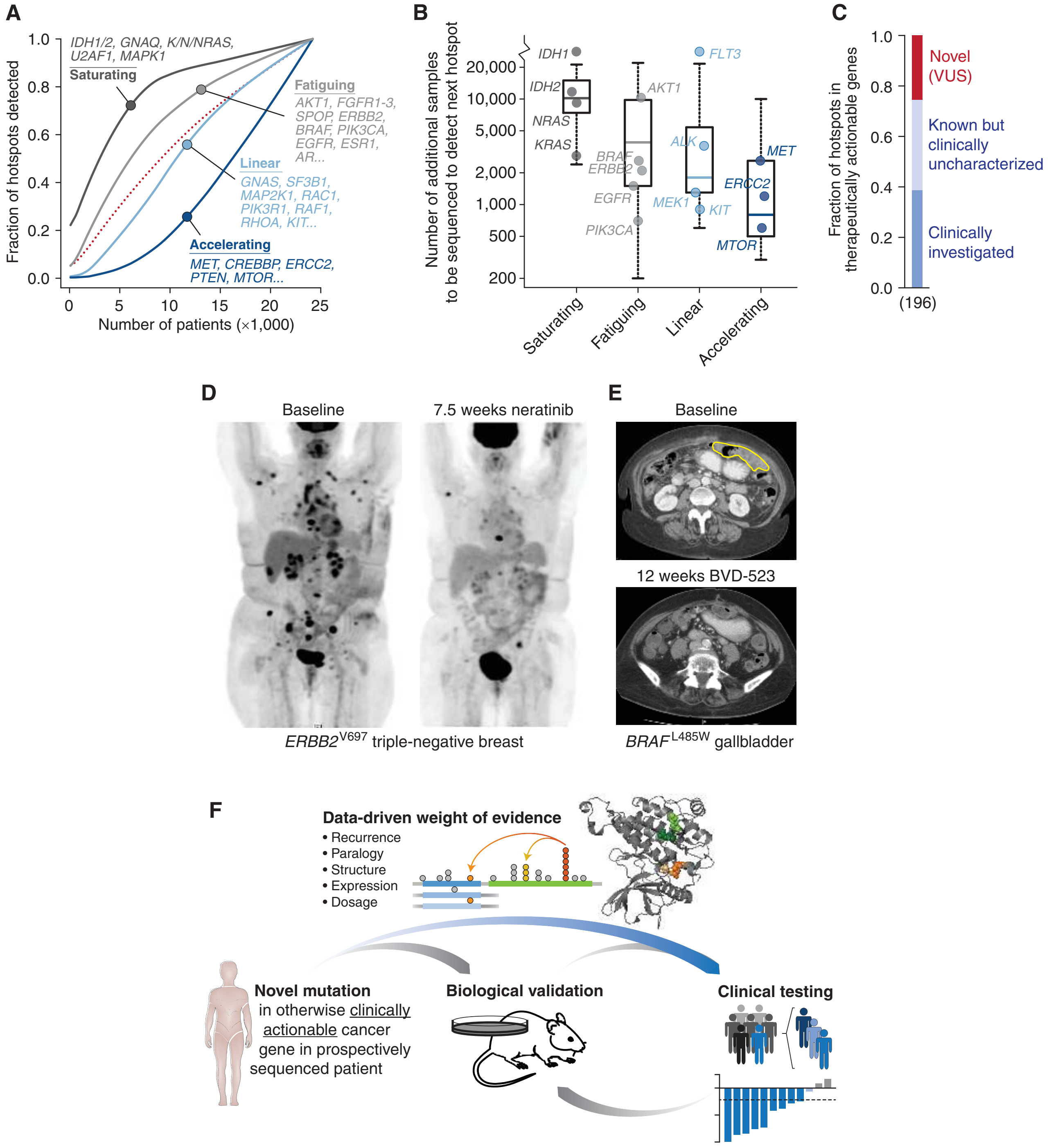

Figure 3. Saturation analysis and the discovery of actionability of mutational hotspots. **A,** Downsampling and clustering analysis revealed four distinct classes of genes with different rates of hotspot acquisition (light and dark gray and light and dark blue) from the number of sequenced samples necessary to identify a given fraction of all hotspots in affected genes. In dark and light gray are genes that either are saturating in their hotspot discovery (dark gray) or were rapidly increasing and now fatiguing (light gray). In light and dark blue are those genes in either their still linear (light blue) and accelerating phases of hotspot discovery (dark blue). **B,** An estimate of the number of additional specimens to be sequenced to identify an additional hotspot in each gene in each of the four aforementioned classes (clinically actionable genes are identified). **C,** Of hotspot mutations identified in 1 of 18 clinically actionable cancer genes (see **B** for genes), the fraction of hotspots used to guide the use of standard-of-care or investigational therapies at present (see Methods) versus those that were identified here but are clinically uncharacterized. **D,** Initial response of a patient with triple-negative breast cancer to neratinib treatment whose tumor harbored a novel *ERBB2*^{V697} hotspot mutation. **E,** A complete response observed in a patient with gallbladder cancer harboring a novel *BRAF*^{L485W} hotspot mutation to the ERK inhibitor BVD-523. **F,** A model by which advanced treatment-refractory patients can be directed to molecularly driven therapies based on computational weight of evidence alone as an efficient means for determining mutant allele function and expanding biomarkers of drug response.

responses to either an isoform-selective *PIK3CA* inhibitor or an mTORC1/2 catalytic inhibitor. The final patient had a gallbladder cancer with a *BRAF*$^{\mathrm{L485W}}$ missense hotspot who achieved a durable complete response to the ERK inhibitor BVD-523 lasting nearly a year (Fig. 3E; ref. 26). Although these 7 patients harboring one of six novel hotspot mutations represent only an exploratory proof of principle, further studies of other hotspots are needed. However, these results indicate that, in some genes, mutation recurrence alone could be used as the initial screen to select otherwise treatment-refractory patients for targeted therapy when biological data do not exist. When affected patients are identified prospectively, such clinical responses to molecularly targeted therapy may be the most efficient way to determine functionality and expand the pool of mutant alleles within a targetable gene that are considered sensitizing biomarkers (Fig. 3F).

DISCUSSION

In sum, we identified 1,165 hotspot mutations across a spectrum of primary and advanced cancers. The rate at which hotspots were identified with increasing cohort size varied widely among genes. In some genes, potentially actionable mutational hotspots were still being identified at a rapid rate with increasing cohort size. As many of the novel hotspots identified here were not previously recognized as functional variants, patients whose tumors harbored such mutations were unlikely to have been offered matched molecularly targeted therapies to which patients with other previously characterized alleles in the same genes have had a profound clinical benefit. The implications of this are especially urgent for those patients with metastatic disease most in need of novel therapeutic approaches.

Pooling prospective genomic data from many sources may quickly achieve the scale needed to saturate the discovery of hotspots in most of the genes targetable with current drugs (27), expanding the reach of precision oncology. To accelerate the identification of novel clinically actionable hotspots, we have deposited all of the data and results at http://cancerhotspots.org for query, visualization, and download to facilitate their dissemination and use by the wider biomedical community. Despite our functional and clinical validation of select novel hotspots, we recognize that hotspot mutations in individual genes may have varying drug sensitivities and potentially allele-specific neomorphic functions. By making all hotspots discovered here available in an easily searchable portal, we aim to catalyze broader functional and clinical validation of individual mutant alleles, results from which we already curate in a knowledge base of oncogenic effects and treatment implications (24). Together, our findings provide a means to prioritize the experimental validation and clinical cross-validation of long-tail driver mutations which will expand the treatment options for molecularly defined populations of patients with cancer.

METHODS

Mutational Data

Retrospective mutational data were obtained from three publicly available sources: (1) The Cancer Genome Atlas (TCGA), (2) the International Cancer Genome Consortium, and (3) independent published sequencing projects (10). The subset of this cohort that was prospectively sequenced consists of 10,945 samples from 10,336 unique patients with advanced cancer whose tumors were profiled as part of their active care between January 2014 and July 2016 at Memorial Sloan Kettering Cancer Center (MSKCC). The consent of these patients, acquisition of specimens, sequencing, analysis, and reporting are described in an accompanying article (8). All such patients provided written and informed consent for sequencing and review of patient medical records for detailed demographic, pathologic, and treatment information (NCT01775072). This study was approved by the MSKCC Institutional Review Board (IRB), and the studies were conducted in accordance with the Declaration of Helsinki, International Ethical Guidelines for Biomedical Research Involving Human Subjects (CIOMS), Belmont Report, or U.S. Common Rule.

Briefly, matched tumor and normal specimens were sequenced (to 500—1,000-fold sequence coverage) with a validated capture-based next-generation sequencing assay called MSK-IMPACT that is New York state–approved for clinical use. This assay captures the coding exons and select introns of oncogenes, tumor suppressor genes, all genes targeted by either approved therapies or those investigational drugs being studied in clinical trials at our Center, and significantly mutated genes reported by large-scale cancer sequencing efforts (Supplementary Table S2). These sequencing data are analyzed as previously described (1) to detect somatic mutations, small insertions and deletions (indels), DNA copy-number alterations, and select translocations using DNA from both frozen and formalin-fixed paraffin-embedded tissues. An IRB protocol facilitates this prospective genomic characterization (IRB #12-245, ClinicalTrials.gov NCT01775072) and enables the return of results to patients. All genomic data generated as part of routine standard-of-care therapy are deposited, along with relevant clinical data, in a HIPAA-compliant manner, in the cBioPortal for Cancer Genomics (28, 29). All somatic nonsynonymous mutations reported were manually reviewed in primary sequencing data as described in ref. 8 and combined with synonymous mutations in the same samples and utilized in this analysis. All mutations in any one of 469 genes that overlap among the retrospective and prospective subsets of the final cohort were uniformly reannotated using vcf2maf ver. 1.6.10 (https://github.com/mskcc/vcf2maf). Variants identified by the Exome Aggregation Consortium (ExAC; ref. 30) as having a minor allele frequency greater than 0.0004 in any subpopulation were excluded as presumed germline unless they were annotated by ClinVar (31) as either pathogenic, a risk factor, or protective.

Single-Codon Hotspot Significance Calculations

The statistical significance of single-codon hotspots was determined in each of 32 separate organ types as well as pan-cancer (full cohort) using an extended version of our previously described method (10). Briefly, statistical significance of every codon was assessed with a truncated binomial probability model in which the expected probability incorporates underlying features of gene-specific rather than genome-wide mutation rates including gene length, gene- and position-specific mutability, and overall mutational burden of the gene (10). This background model is gene-specific and assesses the significance of individual mutant alleles relative to the background of all mutations in the gene in which it emerges rather than across genes. Unlike in our prior study, here we calculated gene- and position-specific mutability on a per-organ type basis to reflect their differences in background mutability and mutational processes. The mutability of each of 32 possible trinucleotides was calculated independently for each organ type as the fraction of mutations affecting the central position of the given trinucleotide *t* across all samples from cancer types belonging to the given organ type (Supplementary Table S1). The mutability of each codon, expected mutability of each gene, and the final binomial probability were calculated as before (10). For 7 of 32 organ types, insufficient whole-exome sequencing data existed to robustly estimate trinucleotide mutability (<50 samples per organ type), so a pan-cancer mutability was calculated as above and utilized. Multiple hypothesis correction for both pan-cancer and organ-specific analyses was performed using the Benjamini and Yekutieli method. Mutational hotspots corresponding to a q value < 0.1 were considered statistically significant (false discovery rate < 10%).

Small In-Frame Insertion/Deletion Significance

We assessed the statistical significance of in-frame small insertions and/or deletions (indels) in a manner similar to single-codon hotspots using the truncated binomial probability model. For this analysis, we excluded frameshift mutations as presumed truncating loss-of-function mutations. As a background model of indel mutability in both normal and disease human genomes is poorly understood, none was utilized here (neither gene- nor position-specific mutability). Also when calculating the expected probability at each site, we allowed the minimum probability to decrease beyond the 20th percentile of all probabilities dataset-wide used for single-codon hotspot detection (10). Due to the allelic variability of indels, in-frame indels were grouped using a maximal common region defined as the contiguous genomic region spanned by overlapping indels. The mutation count for each such region is the sum of all spanning (single bp or more) in-frame indels. Significance was assessed, as with single-codon hotspots, with the binomial model described above. Statistically significant indels that exclusively arose in samples from retrospective data (published or consortial studies) were manually reviewed in aligned sequencing data of representative cases to identify and exclude potential false positives.

Simulating Mutational Acquisition Rates

To assess hotspot acquisition rates within genes, we performed the hotspot analysis on repeated random downsampling of samples in the dataset starting from 100 patients to the final total number of patients in the dataset in 100-sample increments. Only statistically significant hotspots in each downsample were considered if significant in the final analysis. For each gene, we then fit a locally weighted polynomial regression to the distribution of downsamples to estimate the rate of hotspot acquisition for each gene. To infer broader patterns of hotspot acquisition, these fits were then clustered using fuzzy c-means clustering (R package e1071 v1.6-7), and the optimal number of clusters (four) was determined based on reduction of sum of squared error between 1 and 15 clusters.

Mutational Annotation

Hotspots identified here were considered novel if they were absent from the results of prior hotspot studies or, upon detailed literature review, no prior publication described the mutation or its biochemical or biological validation. All mutations were further annotated for their potential prognostic and therapeutic significance utilizing OncoKB, a curated knowledge base of the oncogenic effects and treatment implications of mutations at the individual allele resolution (http://www.oncokb.org/; ref. 24). The potential therapeutic actionability of each mutation (sensitizing to either standard-of-care or investigational therapies) was defined as having one of four levels of evidence based on published clinical or laboratory evidence. Levels are as follows: level 1: genomic alterations that are FDA-approved biomarkers in patients of the indicated cancer type; level 2A: mutations that were deemed to be standard-of-care biomarkers for FDA-approved drugs in the indicated cancer type based on currently accepted practice guidelines such as those issued by the National Comprehensive Cancer Network; level 2B: alterations that are FDA-approved biomarkers in another cancer indication, but not in patients with the affected cancer type; level 3: alterations for which clinical evidence links the biomarker to drug response in patients, but use of the biomarker is not currently a standard of care in any cancer type; and finally level 4: alterations for which compelling preclinical data associate the biomarker with drug response. Only levels 1, 2A, and 3 were utilized for the analyses and results described here.

Enrichment and Clinical Analyses

To test the enrichment of hotspots in either primary or metastatic disease within cancer types, we required that a given hotspot be present in at least 15 samples or 5 metastatic samples in each cancer type. Only samples and cancer types for which we could confirm their primary or metastatic disease status were included in the analysis (TCGA; SU2C prostate, ref. 32; and the prospective MSK-IMPACT series). The significance of enrichment for individual hotspots was assessed on a per-cancer type basis and determined by two-sided Fisher exact test comparing the number of primary samples of a given cancer type that possess the hotspot to metastatic samples of that same type. Both cutaneous melanoma and gliomas were excluded from this analysis due to the high rate of presentation with metastatic disease in the former, and the absence of distant metastasis (local recurrence only) of the latter. Resulting P values were corrected for multiple hypothesis testing with Benjamini and Hochberg method on a per-cancer type basis.

Co-mutational Analysis

To assess the statistical significance of observed comutational frequency, we first constructed a *2-by-j* binary matrix M where each entry m_{ij} referred to the status of the gene i in the sample j and whose value was 1 if sample j had a hotspot alteration in gene i. Co-occurrence analysis was performed for all unique pairwise combinations of genes within a given pathway (whose members were curated from OncoKB; see above). Other than hotspots identified here, for the purposes of this analysis, presumed loss-of-function mutations in tumor suppressor genes in these pathways (*NF1*) were considered altered (nonsense, frameshift insertions or deletions, splice site, nonstop, or translation start site). We generated a null model of random co-occurrence by permuting the observed alterations (10^6 permutations) while preserving the overall frequency of the alterations observed in our cohort. Empirically derived P values were generated as the number of times co-occurrence was observed equal to or more often in this null distribution compared with that of the observed data. Multiple hypothesis correction was performed using Benjamini and Hochberg approach, and significant co-occurrence were those pairwise combinations of genes within pathway of q value < 0.01.

AKT1 Duplication Indel Validation

293-FT cells were obtained from the ATCC and maintained on DMEM supplemented with 10% FBS and 2 mmol/L glutamine. MCF10A cells were similarly acquired from the ATCC (and generously provided by the Solit laboratory), and maintained in DMEM/F-12 base medium containing 5% horse serum and other supplements (20 ng/mL EGF, 0.5 mg/mL hydrocortisone, 100 ng/mL cholera toxin, and 10 mg/mL insulin; complete growth medium). Both cell lines obtained from the ATCC were *Mycoplasma*-free and authenticated by the ATCC using karyotyping, morphology, and PCR-based methods. For experiments, growth factors were withdrawn from the media, and an "assay medium" was used (DMEM/F-12 base medium containing 2% horse serum, hydrocortisone, and cholera toxin). Plasmids, cloning, and stable line generation was performed as follows. *AKT1*-wild-type (WT) and *AKT1*$^{\mathrm{E17K}}$ in pDONR223 vector were provided by the Baselga laboratory. *AKT1* point and indel mutants were generated by site-directed mutagenesis using KAPA HiFi polymerase (KAPA Biosystems) or Q5 mutagenesis kit (New England BioLabs) and verified by Sanger sequencing. *AKT1*-WT and all the other mutants were subsequently subcloned into gateway lentiviral vector pLX302 using LR Clonase II enzyme mix (Invitrogen). Lentiviruses encoding WT or mutant AKT1 were packaged in 293FT cells, and the supernatant media containing viral particles were filtered through 0.45 µm filters and used to infect MCF10A cells. Cells stably expressing the lentiviral constructs were selected with puromycin (2.5 µg/mL).

For Western blot assays, MCF10A cells stably expressing WT and mutant AKT1 were seeded on 6-well plates. After overnight exposure to the assay medium, the cells were lysed, sonicated, and 30 µg protein was loaded onto SDS-PAGE gels, transferred to nitrocellulose membranes, and immunoblotted for pAKT and other downstream molecular targets of AKT pathway activation. Antibodies for pAKT (T308; D25E6), pAKT (S473), pS6RP (S240/244), pPRAS40 (T246), total PRAS40, and total S6 were obtained from Cell Signaling Technology. V5 probe (E10) and actin antibodies were purchased from Santa Cruz Biotechnology. Drug treatment and cell viability assays were performed as follows. AZD5363 was generously provided by AstraZeneca, dissolved in DMSO to yield a 10 mmol/L stock, and diluted in assay medium to achieve the desired concentrations. MCF10A stable lines expressing WT or mutant AKT1 were seeded in 96-well plates, treated with a range of drug concentrations, and cell viability was assessed 72 hours after treatment using the CellTiter-Glo luminescent cell viability assay (Promega).

Data Availability

Both the assembled somatic mutational data and the mutational hotspots identified here have been deposited for visualization, query, and download at http://cancerhotspots.org and in the cBioPortal for Cancer Genomics (http://cbioportal.org/). Levels of clinical evidence for mutational actionability are available at http://oncokb.org/.

Code Availability

The source code for the methods described here is available for download and use in GitHub (https://github.com/taylor-lab).

Disclosure of Potential Conflicts of Interest

B.T. Li is a consultant/advisory board member for ThermoFisher Scientific and Guardant Health. D.B. Solit is a consultant/advisory board member for Pfizer and Loxo Oncology. D.M. Hyman reports receiving commercial research grants from AstraZeneca, Loxo Oncology, and Puma Biotechnology, and is a consultant/advisory board member for AstraZeneca, Atara Biotechnology, Boehringer Ingelheim, Chugai, and CytomX. No potential conflicts of interest were disclosed by the other authors.

One of the Editors-in-Chief is an author on this article. In keeping with the AACR's editorial policy, the peer review of this submission was managed by a senior member of *Cancer Discovery*'s editorial team; a member of the AACR Publications Committee rendered the final decision concerning acceptability.

Authors' Contributions

Conception and design: M.T. Chang, J. Baselga, D.M. Hyman, B.S. Taylor

Development of methodology: M.T. Chang, S. Patel, J. Gao, B.T. Li, B.S. Taylor

Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): T.S. Bhattarai, A.M. Schram, P. Jonsson, T. Shamu, P. Razavi, B.T. Li, D.N. Reales, N.D. Socci, G. Jayakumaran, A. Zehir, R. Benayed, M.E. Arcila, M. Ladanyi, N. Schultz, D.B. Solit

Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): M.T. Chang, T.S. Bhattarai, A.M. Schram, C.M. Bielski, M.T.A. Donoghue, P. Jonsson, D. Chakravarty, S. Phillips, C. Kandoth, A. Penson, A. Gorelick, T. Shamu, C. Harris, B.T. Li, N.D. Socci, G. Jayakumaran, M.E. Arcila, J. Baselga, M.F. Berger, N. Rosen, D.B. Solit, B.S. Taylor

Writing, review, and/or revision of the manuscript: M.T. Chang, A.M. Schram, A. Penson, J. Gao, P. Razavi, B.T. Li, A. Zehir, M.E. Arcila, S. Chandarlapaty, M. Ladanyi, M.F. Berger, N. Rosen, D.B. Solit, D.M. Hyman, B.S. Taylor

Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): S.O. Sumer, R. Kundra, D.N. Reales, D.B. Solit, D.M. Hyman

Study supervision: D.M. Hyman, B.S. Taylor

Acknowledgments

We thank the members of the Marie-Josée and Henry R. Kravis Center for Molecular Oncology for discussions and support. This work was supported by NIH awards P30 CA008748, T32-CA009207 (A.M. Schram), U54 OD020355 (D.B. Solit and B.S. Taylor), and R01 CA207244 (D.M. Hyman and B.S. Taylor); Sontag Foundation (B.S. Taylor); Robertson Foundation (N. Schultz and B.S. Taylor); Prostate Cancer Foundation (N. Schultz and B.S. Taylor); and Cycle for Survival (D.B. Solit and B.S. Taylor).

Received March 24, 2017; revised October 10, 2017; accepted December 4, 2017; published OnlineFirst December 15, 2017.

REFERENCES

1. Cheng DT, Mitchell TN, Zehir A, Shah RH, Benayed R, Syed A, et al. Memorial Sloan Kettering-integrated mutation profiling of actionable cancer targets (MSK-IMPACT): a hybridization capture-based next-generation sequencing clinical assay for solid tumor molecular oncology. J Mol Diagn 2015;17:251–64.
2. Frampton GM, Fichtenholtz A, Otto GA, Wang K, Downing SR, He J, et al. Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. Nat Biotechnol 2013;31:1023–31.
3. Roychowdhury S, Iyer MK, Robinson DR, Lonigro RJ, Wu YM, Cao X, et al. Personalized oncology through integrative high-throughput sequencing: a pilot study. Sci Transl Med 2011;3:111ra21.
4. Van Allen EM, Wagle N, Stojanov P, Perrin DL, Cibulskis K, Marlow S, et al. Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine. Nat Med 2014;20:682–8.
5. Garraway LA, Lander ES. Lessons from the cancer genome. Cell 2013;153:17–37.
6. Stockley TL, Oza AM, Berman HK, Leighl NB, Knox JJ, Shepherd FA, et al. Molecular profiling of advanced solid tumors and patient outcomes with genotype-matched clinical trials: the Princess Margaret IMPACT/COMPACT trial. Genome Med 2016;8:109.
7. Hyman DM, Puzanov I, Subbiah V, Faris JE, Chau I, Blay JY, et al. Vemurafenib in multiple nonmelanoma cancers with BRAF V600 mutations. N Engl J Med 2015;373:726–36.
8. Zehir A, Benayed R, Shah RH, Syed A, Middha S, Kim HR, et al. Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. Nat Med 2017;23:703–13.
9. Alexandrov LB, Stratton MR. Mutational signatures: the patterns of somatic mutations hidden in cancer genomes. Curr Opin Genet Dev 2014;24:52–60.
10. Chang MT, Asthana S, Gao SP, Lee BH, Chapman JS, Kandoth C, et al. Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. Nat Biotechnol 2016;34:155–63.
11. Moore AR, Ceraudo E, Sher JJ, Guan Y, Shoushtari AN, Chang MT, et al. Recurrent activating mutations of G-protein-coupled receptor CYSLTR2 in uveal melanoma. Nat Genet 2016;48:675–80.
12. Sasaki T, Okuda K, Zheng W, Butrynski J, Capelletti M, Wang L, et al. The neuroblastoma-associated F1174L ALK mutation causes resistance to an ALK kinase inhibitor in ALK-translocated cancers. Cancer Res 2010;70:10038–43.
13. Giannakakou P, Poy G, Zhan Z, Knutsen T, Blagosklonny MV, Fojo T. Paclitaxel selects for mutant or pseudo-null p53 in drug resistance

associated with tubulin mutations in human cancer. Oncogene 2000;19:3078–85.

14. Berger AH, Brooks AN, Wu X, Shrestha Y, Chouinard C, Piccioni F, et al. High-throughput phenotyping of lung cancer somatic mutations. Cancer Cell 2016;30:214–28.
15. Kim E, Ilic N, Shrestha Y, Zou L, Kamburov A, Zhu C, et al. Systematic functional interrogation of rare cancer variants identifies oncogenic alleles. Cancer Discov 2016;6:714–26.
16. Demetri GD, von Mehren M, Blanke CD, Van den Abbeele AD, Eisenberg B, Roberts PJ, et al. Efficacy and safety of imatinib mesylate in advanced gastrointestinal stromal tumors. N Engl J Med 2002;347: 472–80.
17. Maemondo M, Inoue A, Kobayashi K, Sugawara S, Oizumi S, Isobe H, et al. Gefitinib or chemotherapy for non-small-cell lung cancer with mutated EGFR. N Engl J Med 2010;362:2380–8.
18. Eisenhardt AE, Olbrich H, Roring M, Janzarik W, Anh TN, Cin H, et al. Functional characterization of a BRAF insertion mutant associated with pilocytic astrocytoma. Int J Cancer 2011;129:2297–303.
19. Foster SA, Whalen DM, Ozen A, Wongchenko MJ, Yin J, Yen I, et al. Activation mechanism of oncogenic deletion mutations in BRAF, EGFR, and HER2. Cancer Cell 2016;29:477–93.
20. Hyman DM, Smyth L, Bedard PL, Oza AM, Dean E, Armstrong A, et al. AZD5363, a catalytical pan-Akt inhibitor, in Akt1 E17K mutation positive advanced solit tumors. Mol Cancer Ther 2015;Abstract B109.
21. Toy W, Shen Y, Won H, Green B, Sakr RA, Will M, et al. ESR1 ligand-binding domain mutations in hormone-resistant breast cancer. Nat Genet 2013;45:1439–45.
22. Grundler R, Miething C, Thiede C, Peschel C, Duyster J. FLT3-ITD and tyrosine kinase domain mutants induce 2 distinct phenotypes in a murine bone marrow transplantation model. Blood 2005;105: 4792–9.
23. Heidorn SJ, Milagre C, Whittaker S, Nourry A, Niculescu-Duvas I, Dhomen N, et al. Kinase-dead BRAF and oncogenic RAS cooperate to drive tumor progression through CRAF. Cell 2010;140:209–21.
24. Chakravarty D, Gao J, Phillips SM, Kundra R, Zhang H, Wang J, et al. OncoKB: a precision oncology knowledge base. JCO Precis Oncol 2017;2017.
25. Hyman DM, Piha-Paul SA, Rodon J, Saura C, Shapiro GI, Quinn DI, et al. Neratinib in HER2 or HER3 mutant solid tumors: SUMMIT, a global, multi-histology, open-label, phase 2 basket study. Cancer Res 2017;77:CT001–CT.
26. Sullivan RJ, Infante JR, Janku F, Wong DJL, Sosman JA, Keedy V, et al. First-in-class ERK1/2 inhibitor ulixertinib (BVD-523) in patients with MAPK mutant advanced solid tumors: results of a phase I dose-escalation and expansion study. Cancer Discov 2018;8:184–95.
27. Grossman RL, Heath AP, Ferretti V, Varmus HE, Lowy DR, Kibbe WA, et al. Toward a shared vision for cancer genomic data. N Engl J Med 2016;375:1109–12.
28. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov 2012;2:401–4.
29. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal 2013;6:pl1.
30. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature 2016;536:285–91.
31. Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, et al. ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res 2016;44:D862–8.
32. Robinson D, Van Allen EM, Wu YM, Schultz N, Lonigro RJ, Mosquera JM, et al. Integrative clinical genomics of advanced prostate cancer. Cell 2015;161:1215–28.