

UCSF

UC San Francisco Electronic Theses and Dissertations

Title

Systematic identification of driver mutations in cancer

Permalink

<https://escholarship.org/uc/item/0s53q62g>

Author

Chang, Matthew Tsn-Wei

Publication Date

2016

Peer reviewed|Thesis/dissertation

Systematic identification of driver mutations in cancer

by

Matthew Tsn-Wei Chang

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Pharmaceutical Sciences and Pharmacogenomics

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO



Copyright 2016
by
Matthew Tsn-Wei Chang

ACKNOWLEDGEMENTS

Cancer is a disease that affects the people closest to us. First and foremost, I am grateful to have had the opportunity to contribute to the cancer research community during my training at Memorial Sloan Kettering and UC San Francisco.

With that, I want to thank the many people who have generously contributed their time and efforts to the work presented in this thesis:

I want to first thank my mentor, Barry Taylor. I am fortunate that my time as a graduate student has been such a wonderful experience. Barry has taught me how to ask important scientific questions as well as how to creatively approach these questions from multiple perspectives. I am grateful for not only the scientific opportunities but, more importantly, for his guidance and support through uncertainty. I am thankful for his constant faith in me and I am indebted to him for all our fond memories.

The nature of the science described here was very much collaborative and I would like to thank the many colleagues whose data I am presenting. Their direct contributions are credited in the figure legends of this dissertation. I want to also acknowledge the members of the Taylor and Schultz labs for their enthusiasm, support, and friendship that has made New York my second home. Throughout the past 7 years, I have had the fortune to interact with many great mentors (in order of their appearance): Richard Shafer, Leslie Floren, Deanna Kroetz, Igor Mitrovic, Cynthia Watchmaker, Stephen Kahl, Kathy Giacomini, Sook Wah Yee, Saurabh Asthana, Nikolaus Schultz, Cyriac Kandoth, JianJiong Gao, and Debyani Chakravarty. They have taken time out of their busy schedules to provide me invaluable advice and wisdom during the different stages of my training; without them, my achievements would not have been possible.

I also want to thank my thesis committee members: Eric Collison and Joseph Costello. I appreciate their constructive feedback for my thesis project and guidance throughout this process.

Finally, I want to acknowledge my family. My two sisters whose help catalyzed my path into this journey. My wife who serves as the voice of reason in even the most unreasonable circumstances. My mom who embodies that there is no triumph without tribulation. Lastly, my dad who is the reason why I am in cancer research today, whose perseverance I continue to aspire toward, and to whom I dedicate this thesis. Thank you

Systematic identification of driver mutations in cancer

Matthew Tsn-Wei Chang

The advent of next-generation sequencing has accelerated the search for somatic mutations that drive the initiation and progression of human cancers. Much emphasis has been placed on the few mutations that occur at high frequency either within or across cancer types. Yet most mutations in cancer genomes occur infrequently. Nevertheless, in aggregate, these rare mutations play a defining role in as many as one-fourth of all human cancers. Distinguishing which rare mutations, amidst a sea of incidental passenger mutations, drive critical molecular, biological, and clinical phenotypes is a foremost challenge of precision oncology. Here, this dissertation discusses complementary computational approaches to identify putative driver mutations in cancer with a focus on how such mutations can reveal novel biological and clinical insights. The two computational approaches leverage 1) recurrence, one of the best markers of selection, to identify mutations that arise more frequently than expected by chance and 2) protein structures, as orthogonal biological evidence, to credential even private driver mutations through significantly recurrent mutational clusters. These methods are applied to mutational data obtained from both retrospectively sequenced human tumor samples and prospectively sequenced cancer patients who received medical care at Memorial Sloan Kettering. As clinical actionability necessitates first understanding the prevalence and properties of driver mutations across diverse cancer types, exploration of patterns of driver mutation emergence as well as validating novel mutations reveal new insights into the pathogenesis and therapeutic sensitivity of human cancers.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

1.1 Overview	1
1.2 The long tail	2
1.3 Allele-specific approach to mutation discovery	6
1.4 Biological implications of mutational context	12

CHAPTER 2: IDENTIFYING RECURRENT MUTATIONS IN CANCER

2.1 Background	19
2.2 Method	20
2.2.1 Determining significant hotspot mutations	21
2.2.2 Mutational data and pre-processing	24
2.2.3 <i>RAC1</i> functional validation	26
2.3 Results	29
2.3.1 Landscape of hotspot mutations in primary human cancers	30
2.3.2 Unconventional hotspots	40
2.3.3 Lineage diversity and mutant allele-specificity	43
2.3.4 Timing of individual hotspots	46
2.3.5 Population-level hotspots in the long tail	47
2.3.6 Hotspots in transporters and transcriptional regulators	50
2.3.7 Long-tail Ras superfamily hotspots	53
2.4. Discussion	58

CHAPTER 3: IDENTIFYING 3D MUTATIONAL CLUSTERS

3.1 Background	61
3.2 Method	64
3.2.1 Protein 3D structure data collection and pre-processing	64
3.2.2 Determining significantly mutated 3D cluster	65
3.2.3 <i>MAP2K1</i> and <i>RAC1</i> functional validation	66
3.3 Results	68
3.3.1 Classification of mutational clusters in protein structures	70
3.3.2 Rare missense mutations in occult drivers	71
3.3.3 Functional validation of <i>MAP2K1</i> and <i>RAC1</i> mutants	73
3.3.4 Comparison to other 3D hotspot detection algorithms	77
3.4 Discussion	82

CHAPTER 4: BIOMARKER DISCOVERY IN HOTSPOT MUTATIONS

4.1 Background	83
4.2 Method	84
4.2.1 Mutational data and pre-processing	84
4.2.2 Determining pan-cancer and organ-type specific significance	86
4.2.3 Determining in-frame insertions/deletions significance	87
4.2.4 Simulating hotspot identification rates	88

4.2.5 Annotation of biological/clinical significance	88
4.2.6 Enrichment and clinical analyses	89
4.2.7 Co-occurrence analysis	90
4.2.8 <i>AKT1</i> duplication indel validation	91
4.3 Results	92
4.3.1 Identification of lineage-specific hotspots	93
4.3.2 Hotspots enriched in metastatic disease	95
4.3.3 Landscape of recurrent indels in cancer	96
4.3.4 Validation of <i>AKT1</i> duplications as sensitivity biomarkers	100
4.3.5 Co-occurrence of multiple pathway lesions	101
4.3.6 Rate of hotspot identification by gene	104
4.4 Discussion	108
CHAPTER 5: CONCLUSIONS AND FUTURE DIRECTIONS	
5.1 Overview	110
5.2 Lessons from phenotype-to-genotype	110
5.3 Prioritization and validation	112
5.4 Opportunities and challenges	117
REFERENCES	119

List of tables

Chapter 1

Table 1.1 Characteristics of the long right tail	3
--	---

Chapter 2

Table 2.1: Select novel hotspots in known cancer genes	31
Table 2.2: Organ system-specific hotspots	39
Table 2.3: GQ60GK and G60 mutations in Ras genes	54
Table 2.4: Mutational data of 11,119 samples	goo.gl/cZPpzv
Table 2.5: Summary of study cohort	goo.gl/BlfsOr
Table 2.6: Summary of 470 hotspot mutations	goo.gl/9yx6X7
Table 2.7: Summary of presumptive false positives	goo.gl/CjWqVW

Chapter 3

Table 3.1: Select 3D clusters of functional significance	74
Table 3.2: All identified 3D clusters	goo.gl/84JxQr
Table 3.3 Classification of 3D clusters	goo.gl/s0NrBn
Table 3.4: 3D cluster residues per gene by residue class	goo.gl/CC0gKH
Table 3.5: Fraction of samples per gene by residue class	goo.gl/z0rI2j
Table 3.6: Comparison of our method to previous methods	goo.gl/Re7tf5
Table 3.7: Comparison of methods to experimental screen	goo.gl/VbDAHB

Chapter 4

Table 4.1: Summary of 1,165 hotspots	goo.gl/K8p5QS
--------------------------------------	--

List of figures

Chapter 1

Figure 1.1: The long right tail of the frequency distribution in cancer	6
Figure 1.2: Alleles rather than genes	7
Figure 1.3: Context condition rare allele function	13

Chapter 2

Figure 2.1: Hotspot detection components and workflow	28
Figure 2.2: Mutational data and hotspot detection	30
Figure 2.3: Global features of significant hotspots	33
Figure 2.4: RNA expression of tumors with known oncogenic hotspots	34
Figure 2.5: Lineage landscape of hotspot mutation	36
Figure 2.6: Lineage landscape of hotspots in common tumor suppressors	37
Figure 2.7: Squamous cell type-specific hotspots	38
Figure 2.8: Impact of unconventional hotspots	42
Figure 2.9: Lineage diversity and mutant allele specificity	44
Figure 2.10: Candidate GTPase driver mutations in the long tail	49
Figure 2.11: Additional candidate long tail hotspots	51
Figure 2.12: GQ60GK mutations are a single genomic event	55

Chapter 3

Figure 3.1: Mutational 3D cluster analysis method and related resources	63
Figure 3.2: Illustration of the permutation procedure for calculating the statistical significance of 3D clusters	65
Figure 3.3: 3D cluster analysis reveals functional rare mutations	69
Figure 3.4: Examples of mutational 3D clusters in tumor suppressor genes	72
Figure 3.5: Experimental validation of functional impact of mutations in 3D clusters in <i>MAP2K1</i> and <i>RAC1</i>	76
Figure 3.6: Comparison of 3D mutational clusters methods	78

Chapter 4

Figure 4.1: The long tail of mutational hotspots in cancer	94
Figure 4.2: Metastatic enrichment of hotspots identified	96
Figure 4.3: Distance of indel hotspots compared to single-codon hotspots	98
Figure 4.4: Oncogenic indel hotspots	99
Figure 4.5: Treatment history of two <i>ESR1</i> mutant patients	100
Figure 4.6: Distribution of types of indels identified as recurrent	101
Figure 4.7: Co-mutational patterns among hotspots reveals function	102
Figure 4.8: Co-mutational patterns of PI3K pathway hotspots	103
Figure 4.9: Saturation analysis and the discovery of actionability of mutational hotspots	105
Figure 4.10: Clustering the rate of hotspot identification	106
Figure 4.11: Clinical actionability in known and novel hotspots	108

Chapter 5

Figure 5.1 Lifecycle of rare mutation discovery and actionability:	116
--	-----

CHAPTER 1*

INTRODUCTION

1.1 Overview

Cancer can be thought as a “disease of the genome”¹. As normal cells acquire somatic mutations, some contribute to aberrant cell growth that, through successive rounds of clonal expansion², drive the evolution of a tumor³. Few in numbers, these driver mutations co-exist with a much larger number of incidental events, or passenger mutations. Together, these mutations are the result of diverse processes⁴ that shape the somatic mutational profile of tumors and can vary by cell lineage⁵, endogenous or exogenous mutagens⁶, and other factors⁷. Population-scale sequencing of human cancers in the last decade has revealed a new reality of mutational heterogeneity: while a small number of mutations arise frequently, about which we have learned a great deal, most individual mutations arise infrequently. Identifying which rare mutations drive the initiation, progression, and response to therapy of human cancers is a perennial challenge in translational cancer research. Here, this dissertation first outlines the importance of rare driver mutations and how they have guided the treatment of human cancers. The following chapters describe two complementary computational approaches that integrate orthogonal biological information to identify novel, putative driver mutations focusing on how the study of such mutations have revealed facets of cancer biology as well as nominated novel biomarkers for cancer therapy.

¹Chang MT, Taylor BS. On the impact of rare mutations in cancer. *Science*

1.2 The long tail

Most mutations in cancer genomes are rare. Whereas the cancer research community has justifiably focused on understanding the mutations that occur at high frequency in human tumors, but these represent only a small proportion of all known mutations. Consequently, the frequency distribution of somatic mutations in cancer has a “long right tail” of low frequency or private mutations (**Figure 1.1a**). This frequency distribution reflects a profound mutational heterogeneity that appears to operate on several levels and can be defined in many ways (**Figure 1.1b-d** and **Box 1**). However, the shape of the long right tail can vary from gene to gene, even among those that encode components of the same oncogenic signaling pathway (e.g. mutations in the *PIK3CA*, *AKT1*⁸, and *MTOR* genes, all of which activate the PI3K signaling pathway)^{9,10} (**Figure 1.1b**). As pharmacologic inhibitors targeting each of these mutant oncoproteins are being tested in various clinical trials, the identification and characterization of activating mutations in these genes, among many others, is relevant to patient care independent of their frequency. Beyond still uncharacterized rare mutations in well-studied oncogenes, a subset of cancer patients remains so-called “driver negative” as their tumor genomes lack any known driver mutations. For example, 27% of lung adenocarcinomas¹¹, 12% of prostate cancers¹², 6% of cutaneous melanomas¹³, 25% of thyroid cancers¹⁴, and 18% of glioblastoma¹⁵ lack a genetic alteration that is clearly oncogenic. While epigenetic alterations undoubtedly play a role in some of these, the

growth progression of many of these tumors are likely driven by still unrecognized rare or even private mutations.

Table 1.1 Characteristics of the long right tail of rare mutations.

The long right tail can manifest in many ways both within and across genes and cancer types

- Mutant cancer genes within a specific cancer type and across all cancer types
- Individual mutations within a specific cancer type and across all cancer types
- Individual mutations at different positions within a given cancer gene
- Individual mutant amino acids at the same position

Rare mutations in the long right tail can be defined in many, often context-specific ways

- Common mutations overall may arise very rarely in specific lineages
- Rare mutations overall can be much more common in a given, often rarely occurring, cancer type
- Rare mutations can arise in known cancer genes
- Rare mutations in rarely mutated genes at sites paralogous to common mutations in another gene
- Rare mutations in genes with a defined role in cancer, but frequently aberrant by other means

Table 1.1 Characteristics of the long right tail of rare mutations

The existence of large numbers of rare mutations raises important questions that we seek to address in this Review. What defines a rare mutation? Are rare or even “private” mutations (i.e. those found in the tumor of only a single patient) biologically or clinically important? How do we establish a framework for their identification, prioritization, and validation? While we focus here on non-synonymous mutations in protein-coding genes, many of these same questions must be addressed for the vast array of non-coding mutations in tumor genomes as well. Of course, there is no easy definition of what constitutes a rare mutation, and such mutations are often context-dependent (**Table 1.1**). For instance, while *KRAS* G12 is one of the most common somatic mutations in human cancer¹⁶, the presence of which defines certain cancer types (90% of pancreas tumors)^{17,18}, it arises quite rarely in others (<2% of acute myeloid leukemias)¹⁹. Similarly, *IDH1* R132 mutations arise frequently in gliomas,

myeloid leukemias, and cutaneous melanomas but also present rarely in approximately 20 other cancer types in which it still drives aberrant epigenetic phenotypes despite lineage differences^{12,20}. The clinical significance of other rare mutations highlights the value of understanding their biology irrespective of their frequency. For instance, *FLT3* D835²¹⁻²³ is a very rare mutation overall that is more common in a single cancer type (>8% of AMLs)¹⁹, a finding that drove the development of FLT3 kinase inhibitors that are currently under active clinical investigation^{24,25} (ClinicalTrials.gov NCT02335814). Moreover, some rare oncogenic mutations arise in cancer types that are different from the cancer types associated with common mutations in the same gene¹⁶. Likewise, rare mutations also emerge in commonly mutated cancer genes as with *KRAS* L117 or *BRAF* G596 mutations or in oncogenes more commonly amplified²⁶ than mutated, such as with *CDK4* R24 and *MYC* T58^{27,28}. Most challenging of all are rare mutations in rarely mutated genes, events often uncovered by orthogonal approaches for discovering novel regulators of pathways in cancer. The identification of oncogenic *RAB35* as an activator PI3K/Akt signaling revealed very rare gain-of-function mutations resulting in constitutive activation²⁹. Another is *PIK3CB* D1067, a rare oncogenic mutation that so far has arisen in only a single tumor in several cancer types^{16,30,31}, drives aberrant PI3K signaling, and was identified in part by generating preclinical pan-PI3K inhibitor resistance³². These are just a handful of examples, among many more, where the translational significance of a mutation, defined as the possibility of near-term therapeutic intervention, does not always correlate with its frequency in any population.

The identification of oncogenic *RAB35* typifies how rare mutations inform our understanding of the nuance of gene and pathway function. Such rare events may be the key mutational event in a substantial minority of patients, but we cannot currently distinguish these key events from other infrequent passenger mutations based on their recurrence, the most easily measured hallmark of positive selective pressure. These events typify the mutational heterogeneity of human cancers that has led to calls for sequencing many thousands of additional human tumors³³. Yet, recurrence is the differentiating feature of relatively few mutations in human cancers (**Figure 1.1a**). So, while additional molecular profiling will surely credential new recurrent mutations, the number and frequency of rare mutations (both drivers and passengers) will only grow, pushing the long tail out farther but not fundamentally changing the shape of this frequency distribution (**Figure 1.1e**).

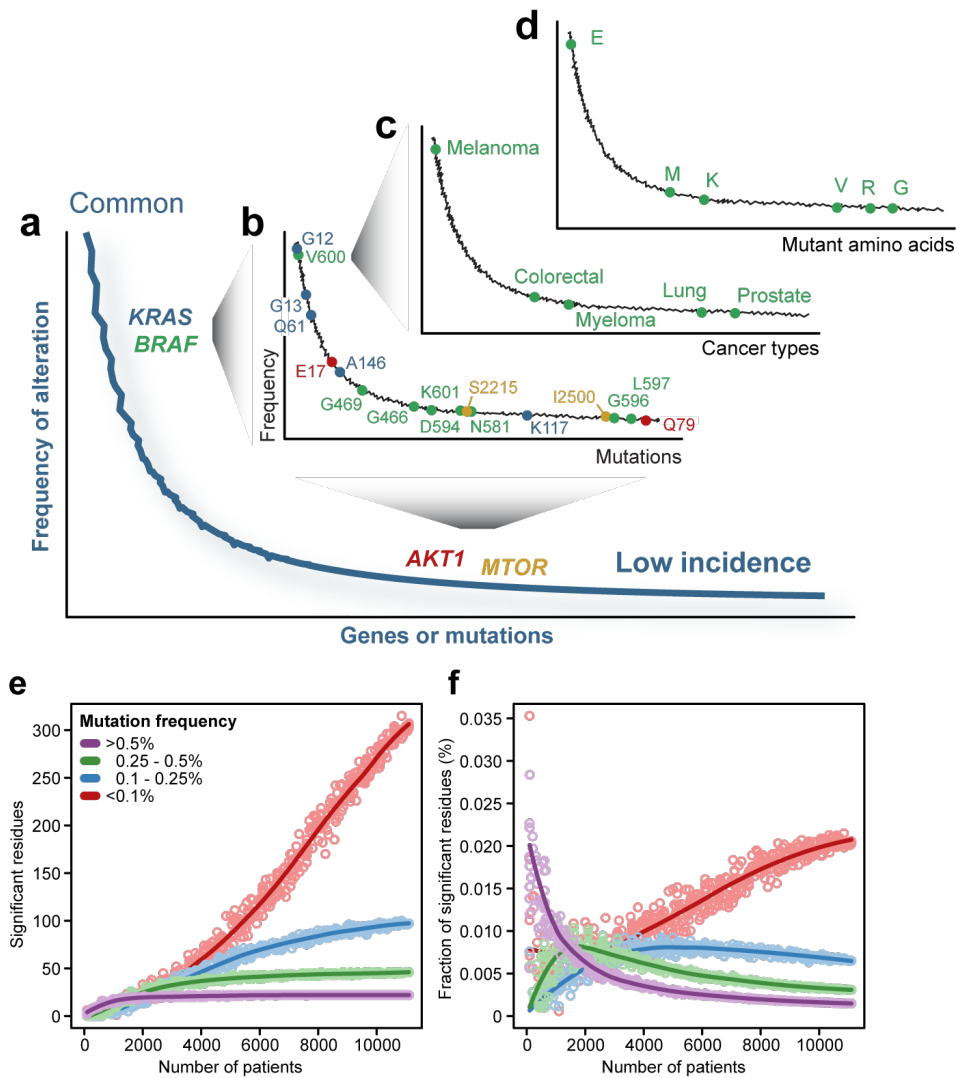


Figure 1.1: The long right tail of the frequency distribution in cancer. The long right tail of the frequency distribution in cancer operates on many levels: **a**) The frequency distribution of somatic mutations in human cancer has a long right tail. Highlighted are the position of four either commonly mutated (*KRAS* and *BRAF*) or uncommonly mutated (*AKT1* and *MTOR*) oncogenes. **b**) Different hotspot mutations in genes [colored as in **(a)**] arise across orders of magnitude difference in frequency despite targeting different effectors of the same pathway. Both the affected cancer types **c**) and the mutant amino acids of a given residue **d**) of an oncogenic mutation (*BRAF* V600) can also form a long tail (e.g. common in melanomas, rare in prostate cancers). **e**) Unlike for common mutations, the rate of rare mutations discovery is still increasing. **f**) Despite the increasing number of rare mutations discoveries **(e)**, the rate of which new recurrent alleles will be discovered relative to the growing number of total mutant residues is rapidly decreasing, arguing that additional sequencing will have diminishing returns over time.

1.3 Allele-specific approach to mutation discovery

Few therapeutic decisions in the era of precision oncology are made at the level of individual genes, but instead at the level of specific mutant alleles. However, to expand the provisioning of therapies targeting mutant oncoproteins, we must first understand the biological and therapeutic significance of all mutant alleles in these genes. Because not all mutations are driver mutations, in even the best characterized and most often therapeutically targeted cancer genes, identifying which mutations are important is an enormous challenge. Moreover, we have only scratched the surface of the potential underlying pleiotropy among the many mutations within a given cancer gene (**Figure 1.2**). To identify allele-specific differences within genes requires both computational and experimental approaches.

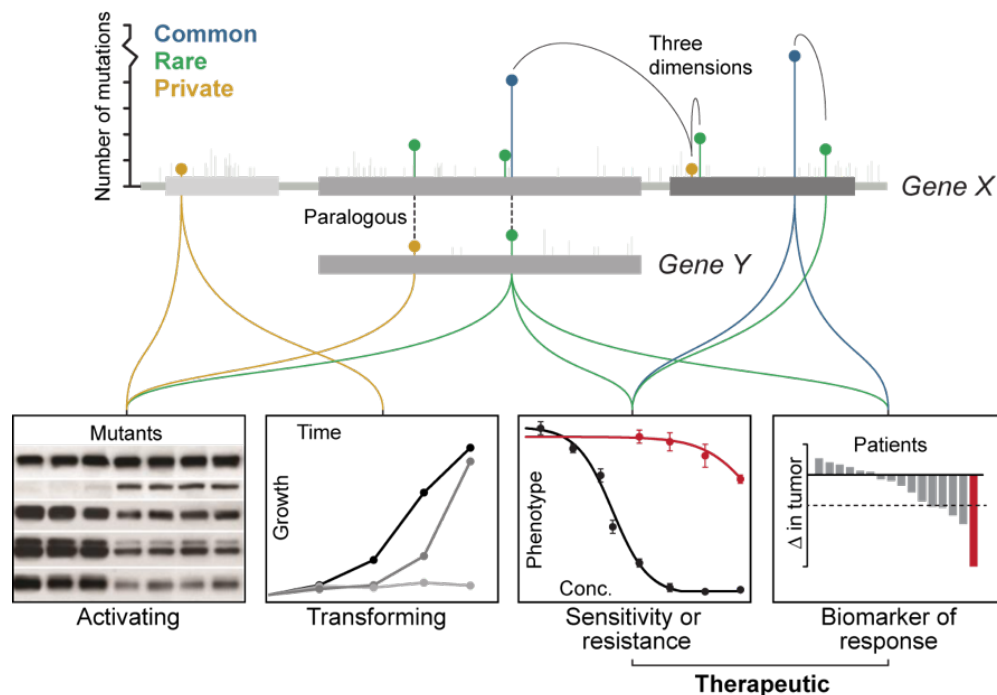


Figure 1.2: Alleles rather than genes: Mutant allele-specific functional complexity confounds the study of genes. Common, rare, or private mutations in a hypothetical gene (X) arise at very different frequencies (as indicated). Some mutations affecting distant residues may cluster in physical proximity when the protein is folded in three-dimensions (arching lines), while other mutations may affect paralogous positions in a closely related family member (gene Y). Each of these mutations may possess different biological or clinical properties (bottom). Mutations may or may not activate signaling (left), transform cells (left middle), sensitize or not to therapy (middle and right) or possess different combinations of these phenotypes. For example, mutations may be sensitizing *in vitro* and a biomarker of response to therapy in patients, but cannot transform cells by themselves (blue), while other mutations may be activating and transforming, but not intrinsically sensitive to drugs (golden).

Distinguishing which among many mutations in each cancer gene are functional is an enormous and often gene-specific challenge that demands orthogonal experimental approaches. This is especially urgent in cancer genes for which a therapeutic agent exists and functional mutations are potentially actionable in patients sequenced in oncology clinics today. Careful analysis of protein structure and mutational patterns can highlight these rare actionable mutations. For instance, *AKT1* and *MAP2K1* are key mediators of PI3K and MAPK signaling pathway activation respectively. While both genes harbor known activating mutations, even the most recurrent mutations in each gene are relatively rare compared to mutations in their upstream effectors, *PIK3CA* and *BRAF*, respectively. Nevertheless, studies show that not all mutations, even in their key domains or adjacent residues, are functional. Mutagenesis experiments in *AKT1* identified mutations only in key residues at the interface of the pleckstrin-homology and kinase domains result in oncogenic activation⁸. Likewise, thus far only *MAP2K1* mutations located within the alpha helix A/N-terminal lobe interface result in deregulated kinase activity³⁴. But, a fundamental gap remains. To realize the promise of clinical sequencing to guide the care of individual cancer patients, we must understand better

the biological and therapeutic significance of any mutant allele in these genes and many others. With the advent of routine clinical sequencing of active cancer patients, the scope and urgency of such studies increases. At Memorial Sloan Kettering Cancer Center, we are performing prospective sequencing to inform the care of our cancer patient. In the over 10,000 patients sequenced to date, beyond the known recurrent mutations in these genes, we have identified mutations affecting 67 and 55 distinct residues in *AKT1* and *MEK1* respectively. Only a minority of these are recurrent and have been experimentally studied (6 and 18 in *AKT1* and *MEK1* respectively). The functional significance of the remaining ~98 mutant residues remains unknown. These are only two examples of many genes with a similar long tail of experimentally uncharacterized mutations for which common mutations in those genes are already biomarkers that guide the use of existing therapies. Because we do not know if these rare mutations are drivers or passengers, it is unclear whether patients would benefit from targeted therapies, thereby limiting the treatment options in advanced cancer patients in greatest need of novel approaches.

While a conservative approach of requiring functional studies prior to clinical testing may be justified in mutant genes emerging now as new drug targets, in other cancer genes in which a subset of mutations is clearly associated with response to specific therapies, a guilty-until-proven-innocent approach is likely warranted. An example of this concept are *EGFR* mutations in lung adenocarcinomas, which predict for a response to *EGFR* tyrosine kinase inhibitors. Rare mutations in the kinase domain of *EGFR* continue to be discovered, but have not necessarily been biochemically validated like

more common *EGFR* alleles. Should patients with rare or even private *EGFR* kinase domain mutations be treated with EGFR inhibitor therapy even before supporting evidence indicates it is similarly a biomarker of response to such therapies? Unlike the aforementioned *AKT1* and *MEK1* mutants where the lack of information prevents clinical action, cases such as rare *KRAS* mutants in colorectal cancers could be presumed to exert similar phenotypes to more recurrent mutations (guilty) until further clinical validation shows otherwise (proven innocent). Other mutations may arise so rarely in an actionable cancer gene as to make their study difficult to justify. If these mutations were present in the tumors of patients who have already failed standard-of-care therapies, might they be presumed a driver and treated accordingly with the corresponding targeted therapy? The absence of a response does not prove the mutation was not a driver, but a response to therapy would justify broader studies and expand therapeutic options for the most advanced patients. The importance of separating such rare functional drivers from neutral alleles extends beyond signaling pathways and targeted inhibitors to other important molecular phenotypes that may guide the provision of newer classes of drugs³⁵⁻³⁷, especially as immunotherapy is being explored in patients with tumors of high somatic mutational load^{38,39}.

Beyond distinguishing functionally significant from neutral mutations in a given cancer gene, the more confounding problem is that subsets of the former may have a different functional impact. The clearest demonstration that not all mutations in each cancer gene have the same function comes from the nearly 40 years of study into *TP53* function. There is an enormous body of literature plumbing the types and patterns of

TP53 mutations in human cancer⁴⁰. Beyond abundant loss-of-function mutations, many *TP53* missense mutations have dominant negative or gain-of-function properties that lead to neomorphic functions and phenotypes^{41,42}. While the extent of these functional differences remains elusive, there is no reason to believe such pleiotropy among individual mutations is specific to *TP53*, as both functional and therapeutic differences have been shown among different mutations in many other well-studied cancer genes including *KRAS*⁴³, *EGFR*⁴⁴, and *PIK3CA*⁴⁵. Translating such functional nuances between different mutations in the same gene into a mechanistic understanding of pathways is now leading to new therapeutic approaches with allele-specific inhibitors^{46,47}.

The distinct biochemical consequences of different mutations in the same gene (**Figure 1.2**) have implications beyond biological nuance and drug development, but have now been shown to affect therapeutic outcomes. For instance, a recent study revealed that *BRAF* mutants with activated kinase activity are all insensitive to ERK-dependent feedback inhibition of Ras because these mutants, as opposed to wildtype Ras, function in a Ras independent manner. However, Ras independence occurs by two different allele-specific mechanisms that confer sensitivity to different types of RAF inhibitors⁴⁸. These rare *BRAF* mutant alleles can be drivers and their mechanism of activation and sensitivity to therapy may be specific to different classes of alleles, a level of complexity that may exist for other targets, but at present is uncharacterized.

Even in genes that lack a common mutation, the study of specific rare mutations can reveal novel mechanisms of pathway activation and specific therapeutic

vulnerabilities. One such example is diverse hypomorphic and loss-of-function mutations in *PIK3R1* that have been documented in several cancer types and result in elevated PI3K signaling⁴⁹. Nevertheless, one rare yet recurrent *PIK3R1* Arg348* truncating mutation is uniquely neomorphic, producing a truncated peptide that results in selective activation of components of the MAPK pathway leading to therapeutic sensitivity to MEK and JNK inhibitors⁵⁰. While uncommon, the differential function of such long tail mutations revealed unique *PIK3R1* mutation-specific biology and a potential biomarker of sensitivity to MAPK and JNK pathway inhibitors. Allele-specific functional differences are also emerging among mutations affecting cell-essential genes such as those involved in the human spliceosome⁵¹⁻⁵³. For instance, S34 and Q157 mutations in *U2AF1*, which together affect ~45-85% of patients with myelodysplastic syndrome⁵³ and arise very rarely in other cancer types¹⁶, are mechanistically distinct, binding distinct sites in the consensus 3' splice site motif of differentially spliced exons⁵⁴. Other such allele-specific differences exist in transcriptional activity and binding affinity among multiple rare mutations in *ESR1* (Y537 and D538) in response to endocrine therapy in breast cancers^{55,56}.

Together, these exemplify the broad spectrum of function in which allele-specific molecular and phenotypic consequences may arise and suggest that many more nuanced functional differences between individual alleles in a cancer gene remain to be discovered. These data argue that a renewed focus on individual alleles and allele-specific effects rather than genes^{5,57-61} can provide unique insight into facets of cancer biology and potential therapeutic vulnerabilities that may, in turn, have the most near-

term therapeutic benefit, especially as more cancer patients have their tumors routinely sequenced to guide their care. Ultimately, precision medicine requires we determine whether alleles identified in potential proto-oncogenes are functional and, if so, characterize their function and their sensitivity to specific inhibitors

1.4 Biological implications of mutational context

The majority of both common and rare mutations arise across cancer types of very different tissues of origin suggesting these mutations may confer a growth advantage across diverse lineages and environments. These mutations may, therefore, confer a growth advantage in diverse cell types and tissue environments. Some of the most intensely studied somatic mutations arise at varying frequencies in nearly all cancer types. If all driver mutations confer a growth advantage, then why do some of them arise at such low frequency? Among many possibilities⁶², the function of a rare driver mutation may be conditioned by the context in which it arises. In this Review, we define “context” as the biological and physiological setting of a given rare mutation arising in patients including the cancer type, cell lineage, amino acid change, the presence or absence of other genomic alterations in the affected tumor, and the selective pressure of ongoing therapy (**Figure 1.3**). Here, we discuss how studying the function of rare mutations in a context-specific manner can enhance our knowledge of basic oncogenic properties and uncover unique context-dependent therapeutic vulnerabilities.

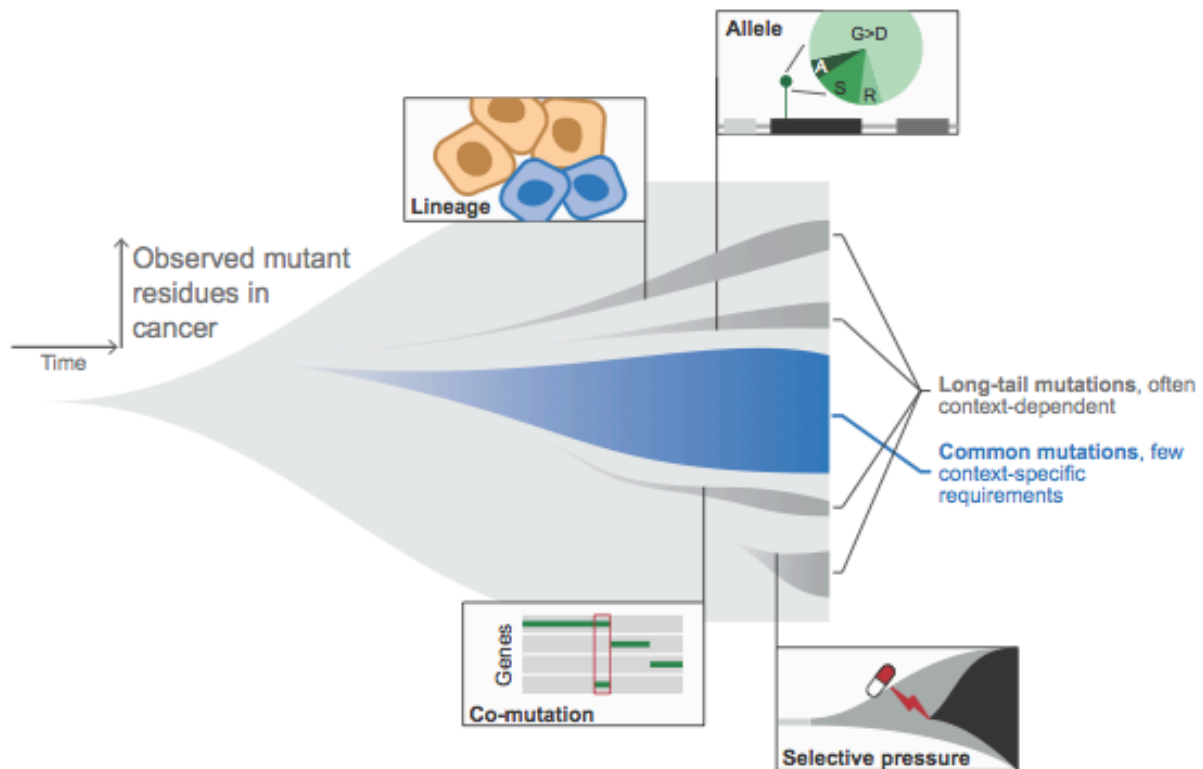


Figure 1.3: Context condition rare allele function: Among all mutations observed in cancer, the vast majority are passenger mutations (light gray). A small minority are driver mutations (dark gray, dark blue) are selected for because they confer gain-of-fitness to the affected cell. Many of these driver mutations are functional across a broad spectrum of biological contexts (e.g. *KRAS* Gly12 mutations; blue, common mutations). Other driver mutations (dark gray) are selected for because they have an aberrant function conditioned by the context in which they arise (see Section 1.3). As these driver mutations have a function that is context-dependent, they are observed less frequently in the long right tail of the frequency distribution of all mutations in cancer.

Lineage. The discovery that many oncogenic hotspot mutations arise in a variety of cancer types, often with low incidence, has excited the community about new models for testing targeted therapies, such as histology-independent “basket” clinical trials in which patients are enrolled based on a defined genetic lesion rather than their tumor’s organ of origin. Such studies may, however, reinforce the importance of lineage as a modifier of mutant allele function⁶³. Receptor tyrosine kinases such as *EGFR* and

ERBB2 are a well-understood class of pharmacological targets with multiple targeted inhibitors of these enzymes approved for the treatment of specific types of cancer. At the population-level, these RTKs have a curious pattern of somatic mutations, with events clustering in either the extracellular or kinase domains as a function of cancer type¹⁶. In the case of *EGFR*, extracellular domain mutations preferentially arise in brain tumors, whereas catalytic domain mutations typically occur in lung adenocarcinomas, only the latter of which result in the active conformation of the mutant protein to which *EGFR* inhibitors bind, explaining in part the differential efficacy of these drugs in these cancer types⁴⁴. While a similar pattern exists among *ERBB2* mutations occurring bladder and breast cancers, it is unknown whether such differences will translate into a similar therapeutic difference, though this hypothesis is being tested clinically (ClinicalTrials.gov NCT01953926). Evidence that lineage is a modifier of function among different mutant alleles within the same gene also exists for two rare mutations in *KRAS*. Both A146 and K117 mutations arise nearly exclusively in colorectal tumors (~3 and <1% respectively), elevate phosphorylated ERK, and can facilitate cellular transformation⁶⁴⁻⁶⁷. However, neither rare mutation can induce pancreatic tumorigenesis *in vivo*⁶⁸ despite the dependence of pancreas tumors on mutant Ras, indicating a functional specificity for colorectal tumors that correlates with their pattern of lineage-specific emergence. A similar pattern of lineage specificity is emerging in the Rho-family GTPase *RAC1* for both recurrent (P29 and A159)^{16,69,70} and rare paralogous mutations, which may possess clinical significance for treatment response⁷¹.

Mutant amino acid. Are different mutant residues at the same codon in the same gene functionally equivalent or distinct? Growing evidence suggests that mutant allele-specific functional differences exist, which adds a dimension of complexity to both common and rare driver mutations. A commonly mutated codon in a given gene may be mutated to one of several amino acids, some of which are quite rare and form a long tail of alleles at that site (**Table 1.1**), but may represent functionally important differences. For instance, while *KRAS* G12 is not a long tail mutation, a diverse set of mutant residues have been observed, many of which are quite rare. Indeed, past studies have shown differences in G12 mutant residue-specific GTPase kinetics and biological potency⁴³. More recently, work suggests that different *KRAS* G12 mutant amino acids vary in their ability to induce pancreatic tumor formation in *vivo*⁶⁸. *KRAS* G12R mutations, a rare allele overall arising preferentially in pancreas cancers rather than other Ras-driven cancer types¹⁶, formed tumors more efficiently than other *KRAS* G12 mutant alleles. Conversely, *KRAS* G12S, another rare allele found predominantly in gastric and not pancreatic cancers¹⁶, failed to form pancreatic lesions in *vivo*. Further biochemical studies are necessary to fully elucidate the consequence of such mutant allele-specific functional differences, the extent of which genome-wide is largely unexplored. Further mechanistic studies are necessary to confirm these findings and determine the extent to which mutant allele-specific phenotypic differences exist across a broad spectrum of alleles and whether these differences have clinical implications.

Co-mutation. The low incidence of rare driver mutations may be driven, in part, by their dependence on a secondary genomic event in order to confer their growth

advantage, as has now been discovered for rare mutations in *BRAF*. While hyperactive *BRAF* mutants result in constitutive ERK signaling and are mutually exclusive with activating Ras mutations⁷²⁻⁷⁴, a recent study described *BRAF* D594 mutations that render the protein catalytically inactive and paradoxically co-occur with activating Ras mutations⁷⁵. These kinase-dead *BRAF* mutants enhanced Ras-mediated transformation through *CRAF* activation. This mechanism by which kinase-dead BRAF potentiates tumorigenesis exemplifies the need to understand how rare co-incident genomic alterations condition aberrant signaling pathways. Such discoveries will provide valuable insight into how patients should be selected for potentially mechanistically distinct targeted therapies, but they may also identify potential synthetic lethal dependences that can be therapeutically exploited.

Selective pressure. Specific selective pressures can uncover rare functional mutations while simultaneously placing them in a specific biochemical, biological, or therapeutic context. The inhibition of specific oncogenic signaling pathways in molecularly defined patients with targeted therapies has proven to be an excellent model of short-term experimental evolution in patients. The first such example was imatinib therapy in BCR-ABL fusion positive CML, which led rapidly to the identification of mechanisms of drug resistance, one of which is the BCR-ABL Thr315Ile gatekeeper mutation⁷⁶ that also arises infrequently in treatment-naïve CML patients⁷⁷. This discovery was a harbinger of what was to come over the next 15 years with the use of various targeted inhibitors of diverse mutant oncoproteins. While similar gatekeeper mutations have been discovered in other targets (most notably *EGFR* T790M⁷⁸, that

also arise infrequently pre-treatment^{79,80}) the landscape of rare mutations mediating resistance has only become more complex with off-target and adaptive mechanisms of resistance emerging. Therapy, therefore, provides an exogenous selective pressure that enriched for a rare functional mutation present in primary cancers as well, the study of which will inform both future drug development and our mechanistic understanding of aberrant target activation⁸¹. Another recent example is the study of a lung cancer patient with an *ALK* rearrangement treated sequentially with three generations of ALK inhibitors (crizotinib, then ceritinib, then lorlatinib), where an initial *ALK* C1156Y mutation led to crizotinib resistance. After third generation lorlatinib treatment, a second *ALK* L1198P mutation arose that led to relapse but paradoxically re-sensitized the tumor to crizotinib therapy⁸². The discovery of such mutations that arise only under the selective pressure of active therapies will likely grow, forming their own right tail, as a greater number of increasingly potent and selective drugs are tested and as cancer genomics refocuses on clinically advanced and post-treatment disease.

To date, studies of hotspot mutations in cancer have been limited to within individual tumor types^{70,83,84} or have focused on individual cancer genes across tumor types⁸⁵. A systematic population-scale, cross-cancer, genome-wide analysis of critical driver mutations has not been performed. As broad-based clinical sequencing has begun to inform the care of individual cancer patients, this would begin to address one of the greatest challenges in the practice of genomically driven cancer medicine: interpreting the biological and clinical significance of mutations in even presumed actionable cancer genes as they arise in oncology clinics. The next two chapters

describe two complementary approaches to identify putative driver mutations. As recurrence is one of the best understood features of positive selection, we first sought to identify mutational hotspots, or mutations that recur more frequently than in the absence of selection. Next, as many oncogenic mutations are thought to alter protein function, rare mutations proximal to known activating mutations are likely to be biologically significant. Therefore, we sought to integrate protein structure to nominate candidate driver mutations in a pan-cancer analysis. Together, these complementary approaches have nominated thousands of potentially novel hotspots in human cancer and uncovered patterns that provide insight into cancer biology.

CHAPTER 2*

IDENTIFYING RECURRENT MUTATIONS IN CANCER

2.1 Background

Among the best-studied therapeutic targets in human cancers are proteins encoded by genes with tumor-specific mutational hotspots, such as *KRAS*, *NRAS*, *BRAF*, *KIT*, and *EGFR*. The acquisition of somatic mutations is one of the major mechanisms responsible for the dysregulation of proliferation, invasion, and apoptosis that is required for oncogenesis. Comprehensive genomic characterization of tumors has produced significant insights into the somatic aberrations that define individual cancer types^{1,61}, broadening our understanding of the dysfunctional molecular pathways that govern tumor initiation, progression, and maintenance. These data have spurred the development of computational algorithms to identify cancer driver genes, defined as those in which molecular abnormalities lead to a fitness advantage for the affected cancer cells.

These computational approaches develop either gene-level statistical models that exploit different mutational patterns^{5,57,58,60} to identify significantly mutated genes or use weight-of-evidence-based methods^{59,61} that are heuristic and ratiometric in approach. Together, these methods focus on identifying cancer genes from a multitude of diverse molecular abnormalities affecting the gene. However, not all genomic alterations in cancer genes are driver alterations. Furthermore, not all driver alterations in a cancer

*Chang MT, Asthana S, Gao SP, Lee BH, Chapman JS, Kandoth C, Gao J, Socci ND, Solit DB, Olshen AB, Schultz N, Taylor BS. Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. *Nature Biotechnology*

gene have the same functional impact, and are therefore likely to have varying clinical significance. The potentially diverse functional effects of different lesions in the same gene are not captured and reported by gene-level models, but are rather assumed to be equivalent. However, emerging data indicate that different hotspot mutations in the same cancer gene can be functionally distinct *in vitro* and *in vivo* and display different clinical phenotypes and drug sensitivity^{44,86-88}. Moreover, it is unknown how widespread such hotspot-specific functional differences may be.

To address this challenge, we develop a computational algorithm to identify driver mutations, rather than driver genes. We assembled and rigorously curated a large repository of cancer genome data consisting of the sequenced tumor exomes and whole genomes of 11,119 human tumors representing 41 tumor types. We developed a biologically aware, statistically principled computational model by combining observed biological phenomena such as nucleotide mutability and varying gene-specific mutation rates into coefficients that we incorporate into binomial statistics. From this, we systematically identify individual recurrent mutations and associate these with related temporal and transcriptional data to investigate lineage-specific variation in mutations, and identify novel hotspots with likely clinical implications.

2.2 Method

For the purposes of this analysis, we first define a driver cancer gene as one in which a molecular abnormality leads to a fitness advantage for the affected cancer cell. This is the broadest definition that encompasses both initiating lesions on which tumor

growth depends as well as lesions arising later in tumor progression that perhaps confer a more modest fitness advantage. We then define a hotspot as an amino acid position in protein-coding gene mutated more frequently than would be expected in the absence of selection. Therefore, all the following mutation types result in the same hotspot: 1) mutations in different nucleotide positions in the same codon of a gene, 2) different nucleotide substitutions at the same site in the same codon that result in different amino acid changes, and 3) mutations where the amino acid substitution is identical but the nucleotide change is different. At present, this analysis is limited to recurrent somatic substitutions, but can be expanded to other classes of somatic alterations such as small insertions and deletions, DNA copy number alterations, and structural rearrangements.

2.2.1 Determining significant mutational hotspots

To determine the statistical significance of individual mutational hotspots, we developed a truncated binomial probability model by incorporating not only underlying features of mutation rates in cancer but also anticipating the gene-specific pattern with which hotspots may arise in different classes of possible cancer genes. In its most general form, if X represents the count of mutations in n samples, the probability of observing k mutations is:

$$\Pr(X = k) = \binom{n}{k} p^k (1 - p)^{n-k} \quad (1)$$

where p is the probability of a mutation in any sample. However, differences exist in the mutability of specific nucleotide contexts in cancer genomes. These vary as a function of the underlying mutational process, potential molecular abnormality in normal DNA maintenance pathways, and possible exposure to exogenous mutagens⁴. Moreover, individual genes have highly variably nucleotide composition and background mutation rates. To address these fundamental characteristics, we integrated a coefficient into a position-specific probability that incorporates both the mutability of the trinucleotide context in which the mutation arose and the trinucleotide composition of the affected gene. For each of the 32 possible trinucleotides, we estimate the mutability of a given trinucleotide t as:

$$m_t = \frac{C_t}{F_t} \quad (2)$$

where C_t is the number of mutations affecting the central position of trinucleotide t across all samples and F_t is the number of occurrences of the trinucleotide t in the coding genome. Too little data existed to compute tumor type- or underlying mutational process-specific mutability. Because a mutated codon in a given gene is comprised of mutations in any one of three trinucleotides that encode that codon, we estimate the mutability of a codon c in gene g as:

$$m_{c,g} = \frac{\sum_{t \in c} m_t n_{t,c}}{n_c} \quad (3)$$

where $n_{t,c}$ is the number of mutations in the central position of trinucleotide t in codon c and n_c is the number of mutations in codon c overall. We estimate the mutability of gene g as $m_g = C_g/(nL_g)$, where C_g is the number of mutations affecting the gene across the n samples and L_g is the length of the gene in amino acids. We then estimated the expected mutability of a given gene g as:

$$m_g = \sum_t \frac{N_{t,g} m_t}{L_g} \quad (4)$$

where $N_{t,g}$ is the number of occurrences of trinucleotide t in gene g . The relative mutability of a codon within a gene is then $r_{c,g} = m_{c,g}/m_g$. This leads to a binomial parameter for hotspot detection of:

$$p_{c,g} = r_{c,g} \mu_g \quad (5)$$

We sought to avoid overestimating the background mutation rate for a gene with several hotspots. This would limit the detection of lower frequency hotspots (*warmspots*) due to the rate of recurrence of one or a few dominant hotspots in the same gene. We therefore developed a truncated form by removing positions in gene g bearing greater than or equal to the 99th percentile of all mutations in the gene. The new background rate is therefore μ'_g , calculated as before where the prime signifies the mutation counts and lengths modified using the above threshold. Then $p'_{c,g} = r_{c,g} \mu'_g$. Finally, in rarely mutated genes where the probability p is exceedingly small (relative to the size of the

cohort N and the length of the protein L), we limited the number of false positive hotspots by allowing $p'_{c,g}$ to get no smaller than the 20th percentile of all p' dataset-wide. Therefore, the final binomial probability is:

$$p''_{c,g} = \max \left\{ \begin{array}{l} p'_{c,g} \\ 20\%ile \text{ of all } p' \end{array} \right. \quad (6)$$

Accordingly, we calculate one-sided p-values for all unique amino acids in every annotated gene per the binomial form given in eq. (1) with probability from eq. (6) and test whether more mutations are observed than would be expected by chance given the pattern of all mutations in the gene; its composition and length; the pattern of its mutability; and the number and type of samples assessed.

To correct for multiple hypotheses, we employed a method for false discovery rate correction that assumes dependence among tests. This correction was performed on the gene level in the following manner. P-values were aggregated per gene on the basis of their codon position. For codons that were not mutated in a given gene and therefore not formally assessed, we padded this with a vector of p-values equal to 1 such that the final set of p-values equaled the amino acid length of the given gene. For all resulting p-values in each gene, they were corrected with the Benjamini and Yekutieli method (implemented in `p.adjust` in the *stats* package in R) and significant hotspots were those sites with q-values < 0.01.

2.2.2 Mutational data, pre-processing, and false-positive filtering

Mutational data were obtained from three publically available sources: 1) The Cancer Genome Atlas (TCGA); 2) the data portal of the International Cancer Genome Consortium (ICGC); 3) various published studies in peer-reviewed journals in which mutational data was made available^{30,31}. Mutation calling algorithms and mutation reporting practices varied from study to study in these curated data, so mutation data review and correction were undertaken where possible. Genomic coordinates of variants from alignments to human reference assembly NCBI36 (hg18) were converted to GRCh37 using LiftOver⁸⁹ with an Ensembl chain file. After standardization to GRCh37, the mutation calls were annotated to gene transcripts in Ensembl release 75 (Gencode release 19), and a single canonical effect per mutation was reported using Variant Effect Predictor (VEP) version 77⁹⁰ and vcf2maf version 1.5. All possible pairs of any two samples with at least 10 somatic mutations were interrogated for sample duplication. For any pair of tumors that shared greater than 80% mutational identity and identical or near-identical clinic-pathological characteristics (upon review of data from the source site/publication), a single tumor in the pair was chosen at random and removed from further analysis as a presumptive duplicate specimen. Furthermore, we excluded small insertions and deletions (indels), despite their presence as true oncogenic hotspots in some genes, due to their greater variability in call quality across datasets. In total, the final dataset included 1,348,424 missense; 524,827 synonymous; 100,866 nonsense; 30,346 splice-site; and 3231 mutations affecting translational start or stop codons. There are also 21,130 oligo-nucleotide variants the majority of which are di-nucleotide mutations along with 71 tri-nucleotide mutations and 13 substitutions

of 4bp or longer. Individual mutations and hotspots of interest (detected as described below) were inspected in individual BAM files from tumor and matched normal specimens of DNA and available RNA sequencing data downloaded from CGHub. When available, expression analyses were based on level-3 RNASeqV2 RSEM normalized gene expression counts from RNA sequencing available via the TCGA Data Coordinating Center. These values were log-transformed and scaled across all samples within each cancer type to facilitate comparisons between cancer types.

Considerable variability exists in the processing and generation of mutational data in individual cohorts by originating centers. To address this variability, we developed several weight-of-evidence based criteria for eliminating presumptive false positives and sequencing artifacts from individual mutation calls as well as from hotspots across the dataset (**Figure 2.1a-b**). Initially, to exclude likely germline variants misattributed as somatic mutations we exclude any mutation identified by both 1000genomes and the NHLBI or those identified only by 1000genomes in two or more samples. We then reasoned that hotspots arising in genes not expressed in a given tumor type are less likely to exert biological impact. We therefore removed from consideration hotspot mutations in genes whose expression was <0.1 transcripts per million (TPM) in 90% or more of the tumors of that type, or for tumors that lacked RNA sequencing data, if more than 95% of all tumors independent of organ of origin had expression of TPM < 0.1 . After determining statistically significant hotspots (described above), hotspots were removed from consideration based on a decision tree model as follows. First, a presumptive true positive (pTPs) list of hotspots was predetermined as coding positions

harboring substitutions in five or more tumor samples (from the August 2013 release of the cBioPortal^{30,31}) in one of 341 key cancer-associated genes sequenced as part of routine CLIA-certified sequencing of matched tumor and normal specimens at Memorial Sloan Kettering Cancer Center⁹¹. Initially, for all samples in which a hotspot was observed and for which the fraction of tumor cells mutated could be calculated from corresponding variant allele frequency and DNA copy number data, we calculated the fraction of tumors in which that site was mutated subclonally (in fewer than 90% of tumor cells). If the fraction of samples in which the hotspot arose subclonally exceeded the maximum such value among pTPs, it was excluded. For remaining sites, we excluded potential hotspots that arose from mutation calling bias from a single source center. We identified cohorts in which subsets of samples were called by different centers and excluded hotspots in which greater than 85% of contributing mutation calls originated from a single mutation-calling center. Next, as local sequence complexity can affect alignment accuracy in various ways based on the read lengths and chemistry of source studies in our dataset, we sought to exclude hotspots on the basis of sequence context. We excluded hotspots where the minimum of Shannon entropy calculated from both 12bp or 24bp of flanking sequence on either the 5' or 3' side of the mutated site was less than the minimum such value among pTPs. We then excluded hotspots that were positioned at either the 5' or 3' end of a mono-, di-, or tri-nucleotide homopolymer runs of 10bp or longer. Remaining hotspots were then excluded if either the sum of their ranked weighted 100 and 24bp alignability (determined by CRG Alignability; UCSC Genome Browser) was less than the minimum value of pTPs or their weighted 24bp

alignability was lower than the 12.5 percentile of all sites. We also excluded any hotspot that while passing these criteria affected a gene that was 1) already rich in presumptive false positives by these criteria (the number of retained hotspots was less than two times the count of hotspots in the gene excluded by one or more of these criteria) or 2) one of 20 well-characterized presumptive “red-herring” cancer genes due to high mutation rates that co-vary with underlying features independent of selection⁴. Finally, we manually inspected the sequencing data contributing to the mutation call for select hotspots in a sampling of affected tumor and matched normal samples. The significant hotspots ($q\text{-value} < 0.01$) that were excluded from consideration on the basis of this model (**Table 2.7**).

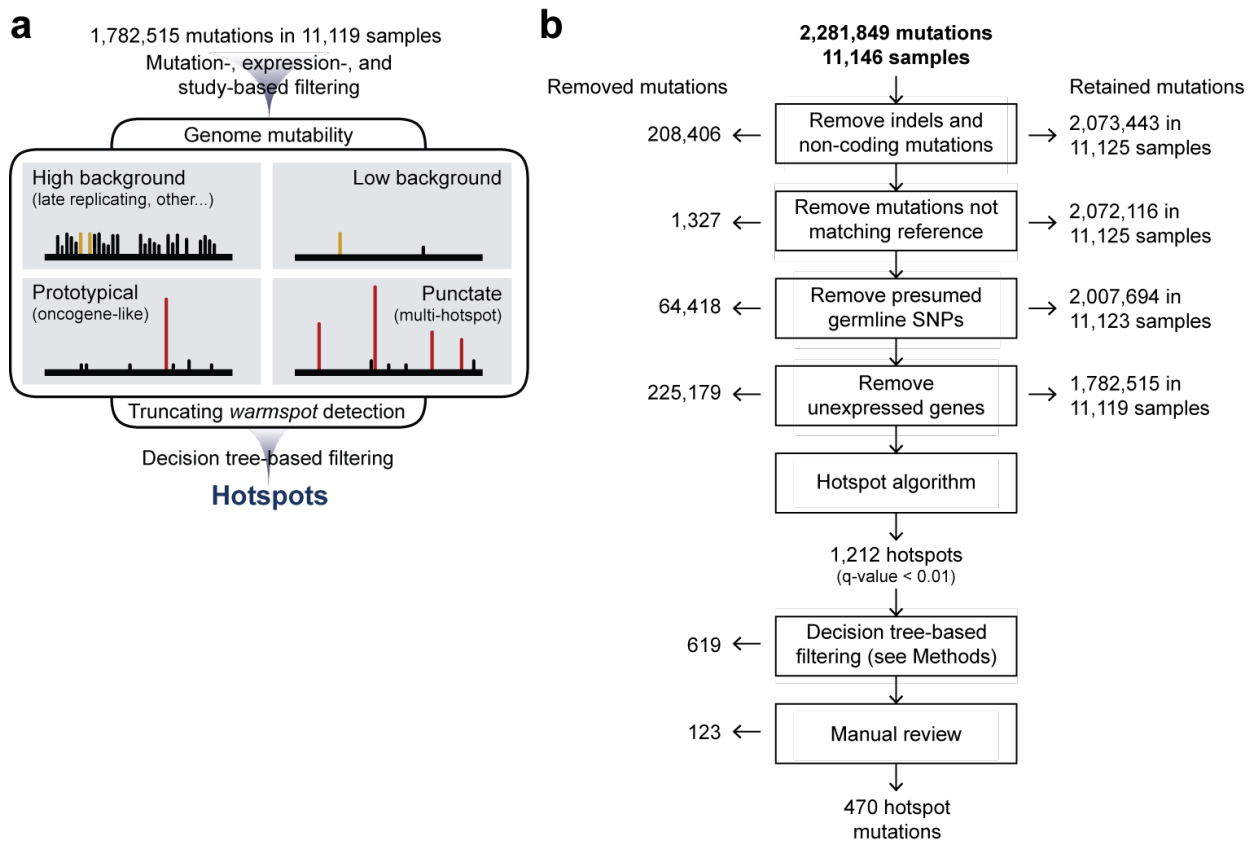


Figure 2.1: Hotspot detection components and workflow

a) Schematic of the hotspot detection methodology employed here is shown. **b)** The steps involved in filtering mutation calls, samples, and genes, as well as generating and curating the final hotspot list.

2.2.3 RAC1 functional validation

DNA coding sequences for wildtype *RAC1* as well as *RAC1*^{P29S}, *RAC1*^{Q61R}, and *RAC1*^{A159V} were generated via site-directed mutagenesis (Genewiz, NJ) to include an N-terminal 3xFLAG epitope tag and were subcloned into a pcDNA3 mammalian expression vector (Life Technologies, NY). The expression constructs were transfected into HEK293T cells using Lipofectamine 2000 (Life Technologies), and cells were harvested after 72 hours. GTP-bound Rac1 (active Rac1) was isolated via immunoprecipitation using recombinant p21-binding domain (PBD) of PAK1 (PAK1-PBD; Active Rac1 Detection Kit, Cat#8815, Cell Signaling, MA), according to the manufacturer's instructions. The Rac1 was detected using kit provided Rac1 primary antibody.

2.3 Results

We collected the mutational data from the sequenced exomes and genomes of 11,119 human tumors in 41 tumor types. These originate from diverse sources including large international consortia and various published studies. This cohort represents a broad range of primary human malignancies with three or more tumor types in each of nine major organ systems (**Figure 2.2a**). The repository consists of 2,007,694 somatic substitutions in protein-coding regions with a median of 57 mutations (25 and 125

mutations; 25th and 75th percentile respectively) per tumor-normal pair with significant variability in mutation rates among and between tumors and types^{4,5}. In total, 19,223 human genes harbor at least one somatic mutation in this dataset.

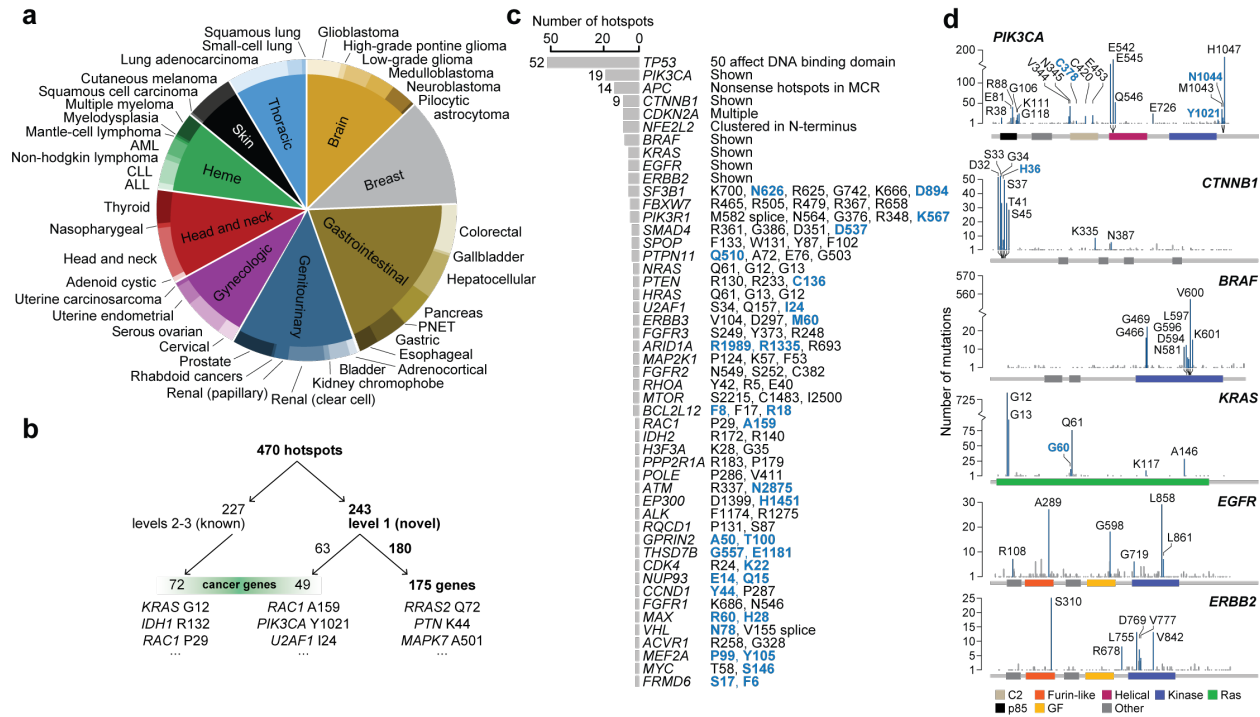


Figure 2.2: Mutational data and hotspot detection

a) The distribution of tumor types included in this analysis. **b)** Breakdown of known and classified novel hotspots and genes. **c)** The number of hotspots in each of 49 genes with two more hotspots detected across the cohort. At right, a summary of hotspots identified. Novel hotspots are bolded blue. **d)** The distribution of mutations and hotspots in six oncogenes refines known patterns and reveals new hotspots.

2.3.1 Landscape of hotspot mutations in primary human cancers

Overall, more than half of all hotspots were determined to be novel (**Figure 2.2b**,

Table 2.1) and 54.8% of all tumors assessed here possessed one or more hotspot mutations.

Pathway/symbol	Codon	q-value	No. of affected tumors	No. of tumor types
Signaling effectors				
<i>KRAS</i>	GQ60	2.28×10^{-6}	11	7
<i>PIK3CA</i>	Y1021	3.18×10^{-6}	9	6
	C378	0.0018	6	5
	N1044	0.0008	6	3
	D1067	0.0068	5	5
<i>PIK3CB</i>	K567	0.0002	5	4
<i>PTEN</i> ^a	C136	2.27×10^{-5}	9	5
<i>RAC1</i>	A159	2.27×10^{-6}	10	5
<i>RRAS2</i>	Q72	8.00×10^{-15}	9	6
<i>GNAQ</i>	T96	7.04×10^{-8}	7	5
<i>ERBB3</i>	M60	0.0083	4	4
<i>MAPK7</i>	A501	9.50×10^{-6}	6	4
<i>PTPN11</i> ^a	Q510	1.84×10^{-6}	7	4
<i>PTN</i>	K44	1.46×10^{-5}	7	4
<i>ARHGAP28</i>	L259	0.0061	5	3
Cell cycle				
<i>CDK4</i> ^b	K22	0.0008	4	2
<i>CCND1</i>	Y44	3.48×10^{-7}	7	2
<i>CDKN2A</i>	E88	4.24×10^{-5}	15	5
	L130	0.007	6	3
Transcription factors				
<i>NFE2L2</i>	E82	1.60×10^{-13}	11	7
	T80	1.96×10^{-10}	9	7
	Q26	9.26×10^{-8}	7	5
	G81	1.34×10^{-9}	10	7
	L30	4.52×10^{-6}	8	5
	G31	0.0001	8	5
	R34	0.0001	13	6
	P99	2.91×10^{-5}	7	6
<i>MEF2A</i>	Y105	0.0061	4	4
	S146	0.0046	6	4
<i>MYC</i>	R60	0.0006	9	6
<i>MAX</i>	H28	0.004	4	1
	I176	0.0001	7	2
<i>FOXA1</i>	I176	0.0001	7	2
Epigenetic modifiers				
<i>ARID1A</i>	R1989	2.45×10^{-8}	17	5
	R1335	0.0062	9	6
<i>ING1</i>	R196	1.06×10^{-6}	11	5
<i>EP300</i>	H1451	0.008	4	4
<i>HIST1H3C</i>	K37	0.0008	5	2
<i>SMARCA4</i>	G1232	0.0006	9	6
DNA damage				
<i>ATM</i>	N2875	4.66×10^{-5}	6	4
RNA splicing				
<i>SF3B1</i>	N626	2.06×10^{-5}	6	4
	D894	0.009	5	4
<i>U2AF1</i>	I24	0.0002	4	4
Wnt pathway				
<i>CTNNB1</i>	H36	0.0001	6	2
Nuclear transport				
<i>NUP93</i>	E14	1.59×10^{-10}	11	6
	Q15	0.0082	4	2
TGF beta signaling				
<i>SMAD2</i>	S464	1.19×10^{-7}	11	5
<i>SMAD4</i>	D351	0.0003	8	6
<i>SMAD4</i>	D537	0.0033	9	3
<i>TGFBR2</i>	R528	0.0013	10	5

Table 2.1: Select new hotspots in cancer genes. A subset of newly identified hotspots is shown

Most affected genes possessed only a single hotspot (**Figure 2.3a**). A subset of genes, however, possessed many hotspots of varying frequency. In total, 49 genes possessed two or more hotspots (**Figure 2.2c**), with many of these also arising in the greatest number of tumor types (**Figure 2.3b**). *TP53* R248 was the most disseminated hotspot, observed in 25 tumor types. Among a subset of even well characterized oncogenes, a pattern of both known and novel hotspots emerge (**Figure 2.2d**). Moreover, the number of observed mutant amino acids at a given hotspot generally increases with its mutational frequency across tumors types (**Figure 2.c**), though 35% (n=164) of hotspots mutate to only a single variant amino acid. In most genes, hotspots bear only a fraction of the total mutational burden across the gene, whereas in a subset of cancer genes, the dominant mutational hotspot constitutes the vast majority of mutations independent of total mutational burden (**Figure 2.2d** and **Figure 2.3d**). Overall, we identified considerable variability in the patterns of mRNA expression of individual hotspots in even canonical oncogenes (**Figure 2.4**), indicating that levels of expression are often not correlated with the biologic significance of known activating mutations.

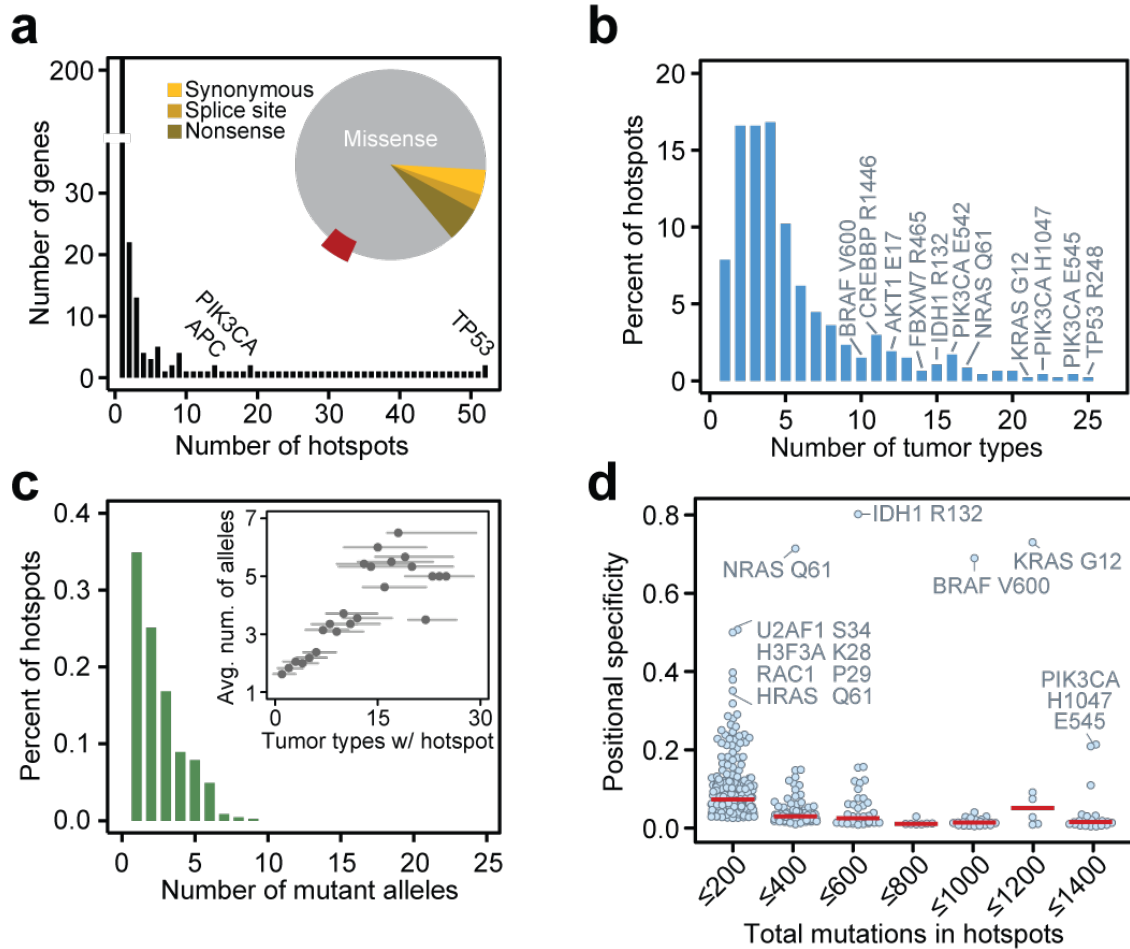


Figure 2.3: Global features of significant hotspots

a) The number of hotspots across the genes identified here (inset: distribution of hotspot type). **b)** The frequency of specific hotspots across the 41 tumor types analyzed here. **c)** The number of mutant alleles distributed among the hotspots detected (inset: number of mutant alleles at a given hotspot increases with the number of tumor types affected, dot is the average number of mutant alleles across the hotspots identified in each of the indicated number of tumor types, bars are the 95% confidence interval). **d)** The fraction of total mutational burden present in the hotspot (positional specificity) of each affected gene.

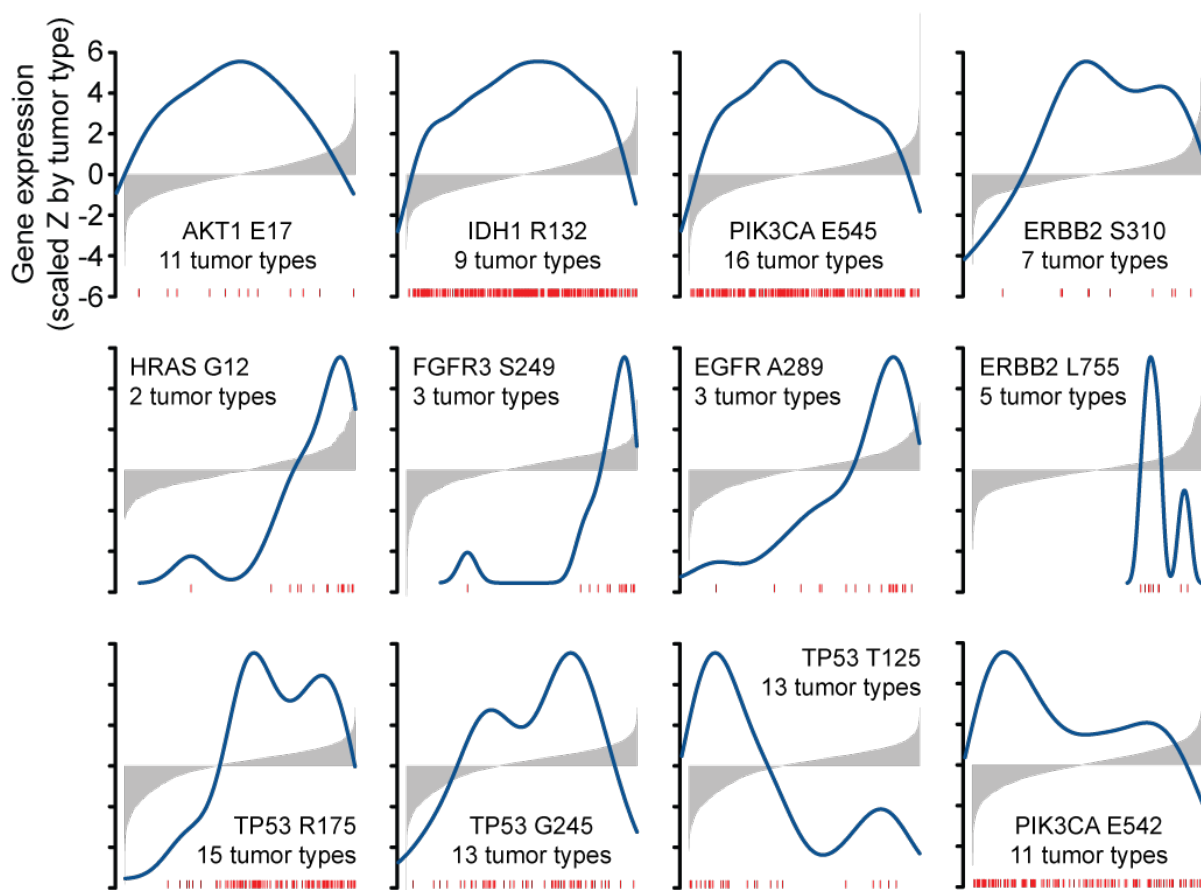


Figure 2.4: RNA expression in tumors with known oncogenic hotspots

The mRNA expression of the indicated gene is shown for all tumors (gray bars) across cancer types in which one or more tumor harbors the indicated hotspot (count of tumor types plotted is indicated, expression is a Z-score of log₂ RSEM normalized count data inferred from level-3 TCGA RNA sequencing data). Tumors harboring the oncogenic hotspot are indicated with red tick marks (x-axis) the density distribution of which is shown in blue. The top row indicates genes with no association between the level of expression and the presence of the hotspot. Middle row are those genes whose expression is elevated in tumors bearing the hotspot. The bottom row indicates genes and hotspots of variable patterns of expression. Multiple hotspots in the same gene with different patterns of expression (*ERBB2* and *PIK3CA*) are shown for reference.

The patterns by which some hotspots emerge support new clinical paradigms for testing targeted agents. Some hotspots that dominate the mutational landscape in one or a few cancer types also arise as uncommon subsets of many others. For instance,

IDH1 R132 is most common in low-grade gliomas, glioblastomas, acute myeloid leukemias (AMLs), and cutaneous melanomas; but it is also present in 1 to 6 tumors in each of 11 additional cancer types. *AKT1* E17K arises in greatest numbers in breast cancer, but also in 1 to 3 tumors of 10 additional cancer types. The distribution of *CREBBP* R1446 mutations is qualitatively different. They were originally identified in relapsed acute lymphoblastic leukemias⁹², but in this cohort of mostly primary disease, we find that they arise in only a small minority (1-3; 0.17-1.7%) of many (11) cancer types. Such patterns reaffirm the value of basket study designs that test mutant-specific inhibitors in early phase clinical trials, where enrollment is based on specific mutations in patients instead of tissue of origin.

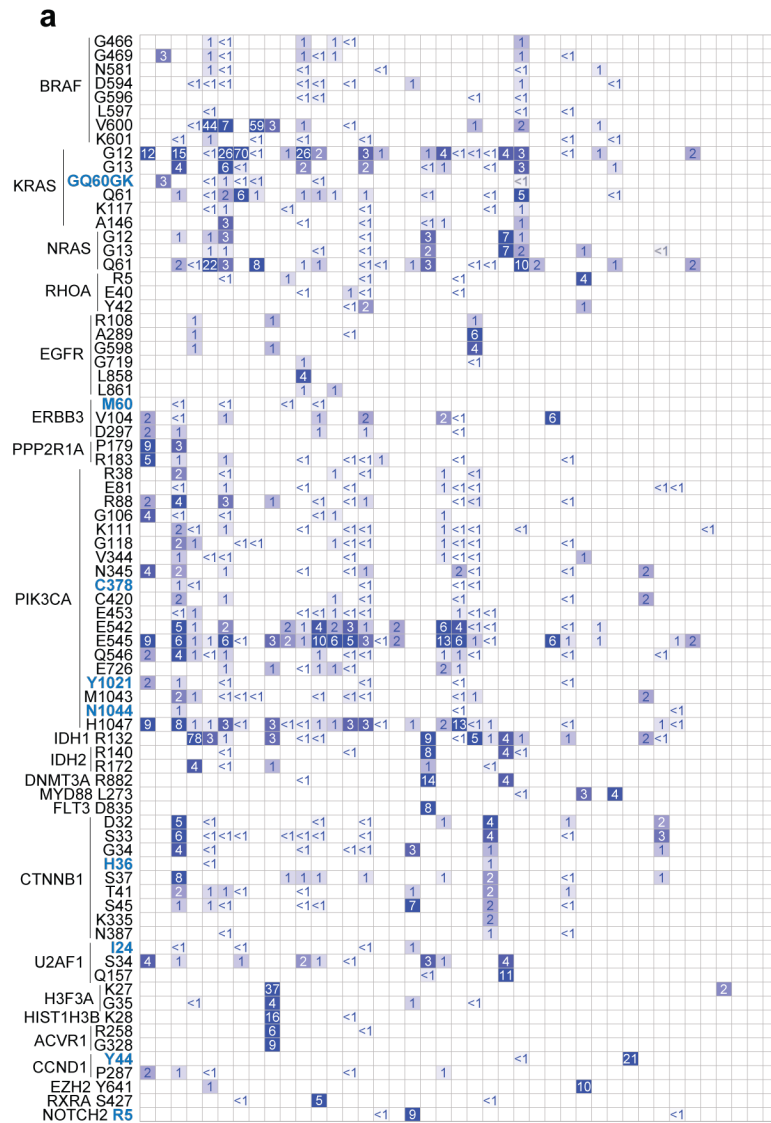
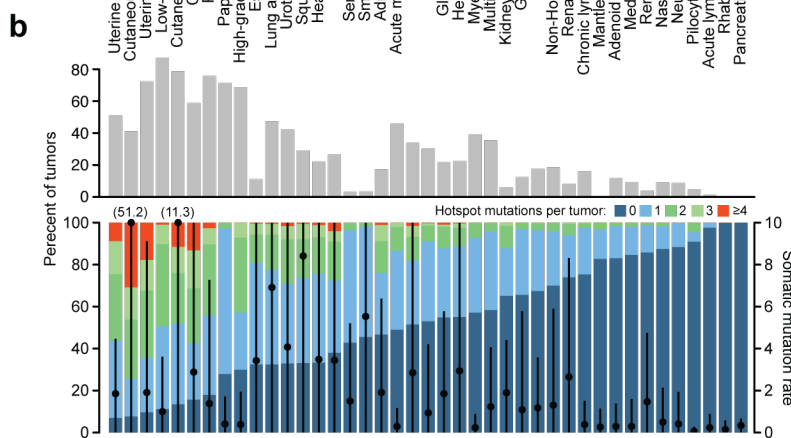


Figure 2.5 Lineage landscape of hotspot mutations.

a) Both common and rare hotspots are largely disseminated across a broad range of malignancies. All hotspots detected in genes with at least one hotspot affecting >5% of tumors of one or more tumor types are shown. Novel hotspots are bolded blue. Genes are grouped broadly by functional similarity, hotspots are ordered by amino acid position, and tumor types (columns, labeled at bottom) are sorted according to the fraction of tumors affected by 1 or more hotspots overall (panel B). The percent of samples altered is represented by colored squares and indicated text. Hotspots in tumor suppressors *TP53*, *PTEN*, *APC*, and *FBXW7* were excluded here (see **Figure 2.6**). **b)** The fraction of tumors of a given type (as indicated) affected by one or more hotspots. Black circles represent the median mutation rate (right axis) in the indicated tumor type (bar is the median absolute deviation). Shown at top is the number of tumors of each type with a hotspot mutation affecting a known or candidate oncogene.



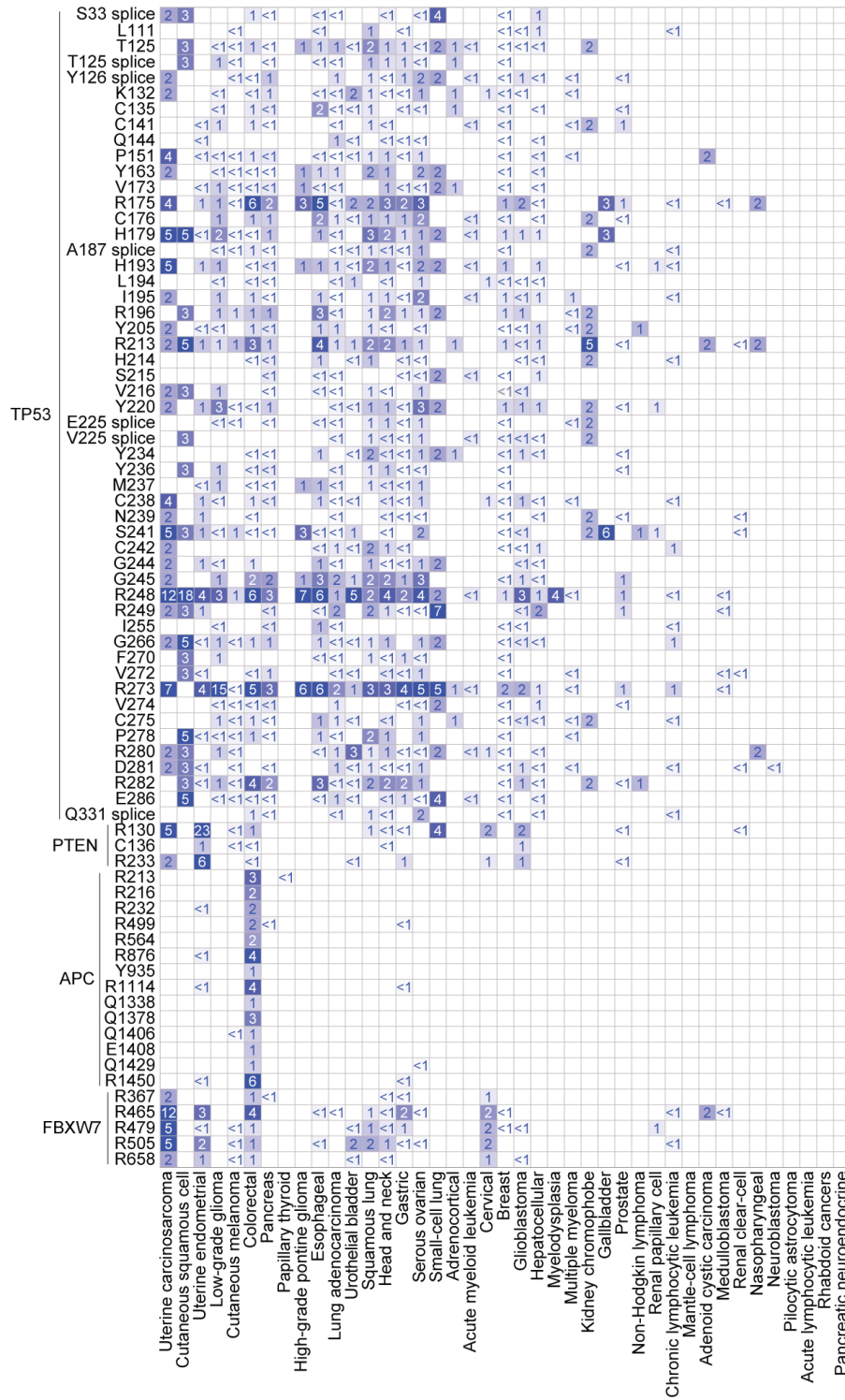


Figure 2.6: Lineage map of hotspots in common tumor suppressors
 As in Figure 2.5, shown here are all hotspots detected in excluded tumor suppressor genes that harbor at least one hotspot affecting >5% of tumors of one or more tumor types are shown. Frequencies are indicated and genes, hotspots, and tumor types are ordered as described Figure 2.5. These included 14 hotspots from R213-1450 of the N-terminal of APC, the mutational cluster region (MCR), affecting between 6 and 37 tumors nearly all of which were colorectal cancers.

A lineage map of all hotspots in genes with at least one common hotspot (**Figure 2.5a** and **Figure 2.6**) indicates most hotspots are defined more by the tissue types rather than the organ systems in which they arise. Of all hotspots, 81% arise in two or more tumor types, suggesting that many hotspot mutations may confer a growth advantage across diverse lineages. Indeed, of hotspots present in multiple tumor types, only 7.6% (n = 36) are confined to a single organ system (**Table 2.2**). Thus, hotspot mutations that arise in a single tumor type may reflect organ-specific growth advantages but they represent only a small minority of all hotspot mutations in cancer. Likewise, a subset of hotspots arises in a cell-type specific manner. Twenty-seven hotspots (5.7%) were more frequently mutated in tumors of a squamous cell lineage (**Figure 2.7a**), the most significant of which were *MAPK1* E322 and *EP300* D1399 (**Figure 2.7b**, q-value = 6×10^{-13} and 1×10^{-11} respectively, X^2) and may potentially confer a squamous cell-type specific growth advantage.

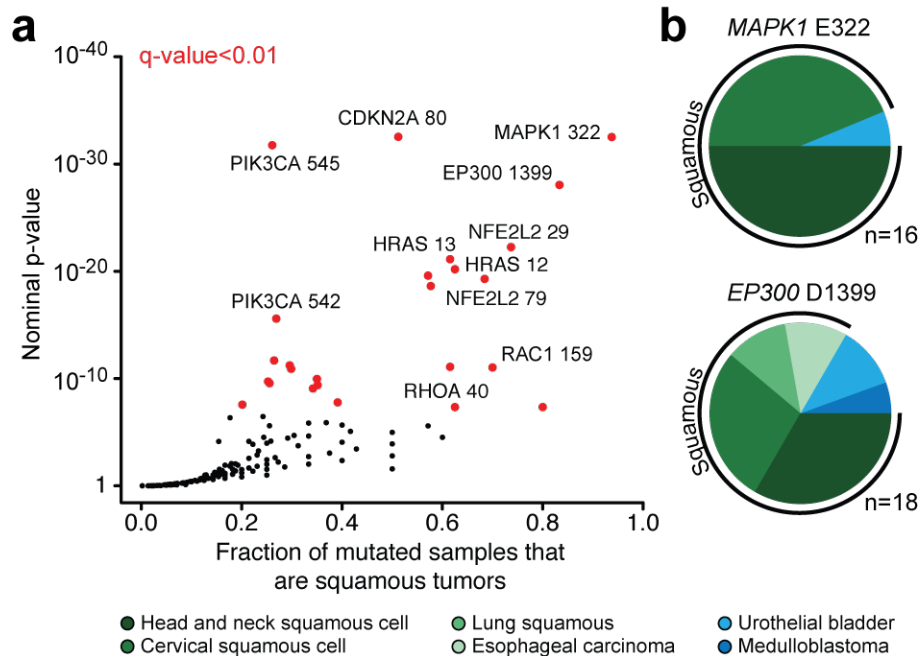


Figure 2.7: Squamous cell type-specific hotspots

a) The enrichment of hotspots in squamous cell tumors (by frequency and significance, as indicated). b) The distribution of tumor types among cases mutated for either *MAPK1* E322 (top) or *EP300* D1399 (bottom).

Symbol	Position	Tumor types ^a	Organ system	No. affected tumors	<i>q</i> -value ^b
<i>H3F3A</i>	K27M	High-grade pontine glioma (26) Pilocytic astrocytoma (1)	Brain	27	2.2×10^{-50}
<i>MYD88</i>	L265P	Chronic lymphocytic leukemia (12) Non-Hodgkin's lymphoma (2) Multiple myeloma (1)	Hematologic	15	1.9×10^{-26}
<i>STK19</i>	D89N	Cutaneous melanoma (13) Squamous cell carcinoma (5)	Skin	18	3.5×10^{-21}
<i>EGFR</i>	G598V/A	Glioblastoma (15) Low-grade glioma (3) High-grade pontine glioma (1)	Brain	19	5×10^{-16}
<i>PPP2R1A</i>	P179R/L	Endometrial (8) Uterine carcinosarcoma (5)	Gynecologic	13	1.6×10^{-12}
<i>FGFR3</i>	Y373C	Urothelial bladder (7) Renal papillary cell carcinoma (1)	Genitourinary	8	2.8×10^{-10}
<i>KNSTRN</i>	S24F	Cutaneous melanoma (11) Squamous cell carcinoma (2)	Skin	13	7.7×10^{-10}
<i>CCND1</i>	Y44D/S/H/F/C/*	Mantel cell lymphoma (6) Multiple myeloma (1)	Hematologic	7	3.5×10^{-7}
<i>CRNKL1</i>	S128F	Cutaneous melanoma (8) Squamous cell carcinoma (2)	Skin	10	4×10^{-7}
<i>EGFR</i>	L861Q	Lung adenocarcinoma (5) Lung squamous cell carcinoma (2)	Thoracic	7	5.4×10^{-7}

Shown are the ten most significant hotspots that arise in multiple tumor types of a single organ system.

Table 2.2: Organ system-specific hotspots. Shown are the 10 most significant hotspots that arise in multiple tumor types of a single organ system.

Overall, the presence, type, and frequency of hotspots by tumor type vary widely (**Figure 2.5b**). In some tumor types, a large proportion of tumors possess one or more hotspot mutations including a significant fraction of tumors with a hotspot in a candidate oncogene (**Figure 2.5b**, top). Conversely, other tumor types never or rarely possess a tumor defined by a hotspot identified here. Some of these differences are certainly attributable to the fact that hotspots are only one of many possible driver genomic

aberrations, including specific gene fusions or focal amplifications and deletions. These other aberrations may define tumors of a given type, but they are not mutually exclusive with hotspots in many cancers. Other differences could not, alone, be explained by the overall mutational burden in these tumor types. For instance, uterine carcinosarcomas and prostate cancers have a similar mutation rate while there is 3-fold greater frequency of hotspot-bearing tumors among the former. Likewise, while papillary thyroid and high-grade pontine gliomas have mutations rates similar to nasopharyngeal tumors and neuroblastomas, the former far more commonly bear hotspot mutations (**Figure 2.5b**).

2.3.2 Unconventional hotspots

In addition to missense mutations, we identified a variety of unconventional hotspot mutations with varied impact. Among these were 13 splice site hotspots. For each of these hotspots, an associated transcript abnormality was identified from RNA sequencing of affected tumors (exon skipping, intron retention, in-frame deletions; **Figure 2.8a**), including two previously characterized in-frame activating mutations (*MET* D1010_splice and *PIK3R1* M582_splice, both exon 14 skipping events). We also identified 70 hotspots in 34 genes for which a nonsense mutation was among a diversity of changes at the affected residue, including 28 hotspots in which only a nonsense mutation was present (**Figure 2.8b**). While nonsense mutations scattered throughout a gene may reflect a pattern of loss-of-function consistent with tumor suppressor activity, a nonsense hotspot would appear to indicate the selection for the selective truncation of specific functional domains. Such events are consistent with the loss of some functions

and the retention of others, as has been observed previously in genes such as *PIK3R1*, *NOTCH1*, and *MET*^{50,93}. These hotspots aside, there was a depletion of nonsense mutations in hotspots in constitutively essential genes (p-value < 10⁻¹⁶, those genes predicted or experimentally verified to be essential across all cell and tissue types and developmental states⁹⁴). Otherwise, the specific impact of nonsense hotspots is generally unknown and belies the disseminated pattern of truncating mutations in likely or proven tumor suppressors (**Figure 2.8c**).

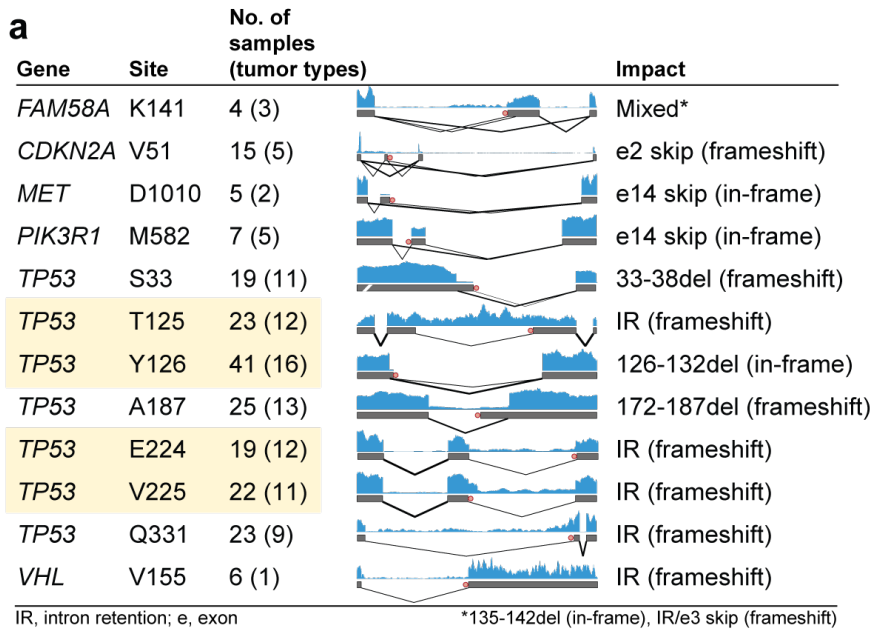
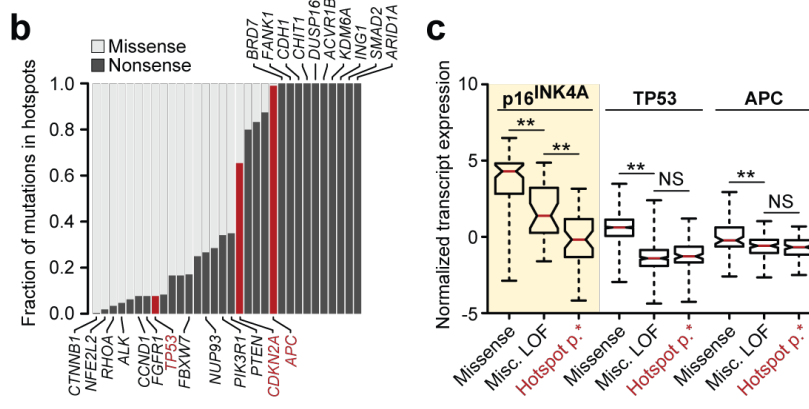


Figure 11: Impact of unconventional hotspots

a) Significant splice site hotspots are shown and have diverse effect on transcript sequence and structure. In blue is the coverage and splicing pattern inferred from RNA sequencing of a representative tumor harboring each hotspot. The impact of each is summarized (rightmost column) and include in-frame and frame-shift events resulting from exon skipping, intron retention, and deletions. Highlighted in yellow are splice site hotspots at opposite ends of the same intron with both similar and dissimilar impact on transcript structure. *SMTNL2* was not



assessable due to little detectable expression in E244 (e3+1)-mutant tumors. **b)** The spectrum of nonsense mutations in hotspots indicate a subset are comprised exclusively of nonsense mutations. **c)** Shown is the impact of nonsense hotspots on transcript expression in *CDKN2A*, *TP53*, and *APC*, three genes affected by the greatest number of nonsense mutations. As expected, the expression (inferred from RNA sequencing of affected cases in TCGA cohorts) of all three genes was significantly decreased between tumors with missense mutations versus those with candidate loss-of-function (LOF) mutations of any kind, including nonsense hotspots. Nevertheless, where no difference in *TP53* or *APC* expression existed between tumors with non-hotspot LOF mutations (labeled Misc. LOF) and those carrying nonsense hotspots, the tumors bearing a nonsense hotspot in *CDKN2A* expressed significantly less transcript levels *p16INK4A* mRNA than did tumors with non-hotspot LOF mutations.

2.3.3 Lineage diversity and mutant allele-specificity

The majority of hotspot mutations arose in diverse tumor types and organ systems, yet widespread differences exist among individual residues and mutant amino acids in hotspots, genes, and tumor types (**Figure 2.9a**). Examining the spectrum of *KRAS* mutations, which includes the most frequently mutated hotspot overall in our study (*KRAS* G12; n=736 mutant tumors, **Figure 2.3d** and **2.5a**), clarified patterns only incidentally observed in the past. We found that gastric cancers were more similar to multiple myeloma in the preponderance of non-G12 mutations compared to endometrial, lung, colorectal, and pancreas tumors (p-value = 5.3×10^{-18}). Only colorectal tumors had *KRAS* A146 mutations whereas pancreas tumors lacked G13 mutations (p-values = 4×10^{-7} and 2.8×10^{-15} respectively). Many of these lineage-specific patterns were present at finer resolution as well. Among *KRAS* G12 mutations, the abundance of G12C mutations are highest in lung adenocarcinomas (p-value = 4×10^{-42}), an event that may be associated with prognostic differences compared with non-G12C *KRAS* mutations⁹⁵⁻⁹⁷. Such mutant amino acid specificity was also apparent in pancreas tumors, where *KRAS* G12R was more common than in any other tumor type (21% versus between 0 and 2.6%; χ^2 p-value = 4.8×10^{-19}). Gastric cancers, on the other hand, had the fewest G12V mutations among all *KRAS* G12-mutant tumor types, but the highest proportion of G12S (p-value = 0.007, **Figure 2.9c**). There is a different balance among hotspots in the other Ras genes. While papillary thyroid cancers nearly exclusively possessed codon Q61 mutations in *HRAS* and *NRAS* (p-value = 4×10^{-7}), there was a higher prevalence of G12 and 13 codon mutations in these genes in AMLs,

colorectal, bladder, and head and neck cancers, which together share few mutational processes in common (p -value = 4×10^{-10} , **Figure 2.9a**).

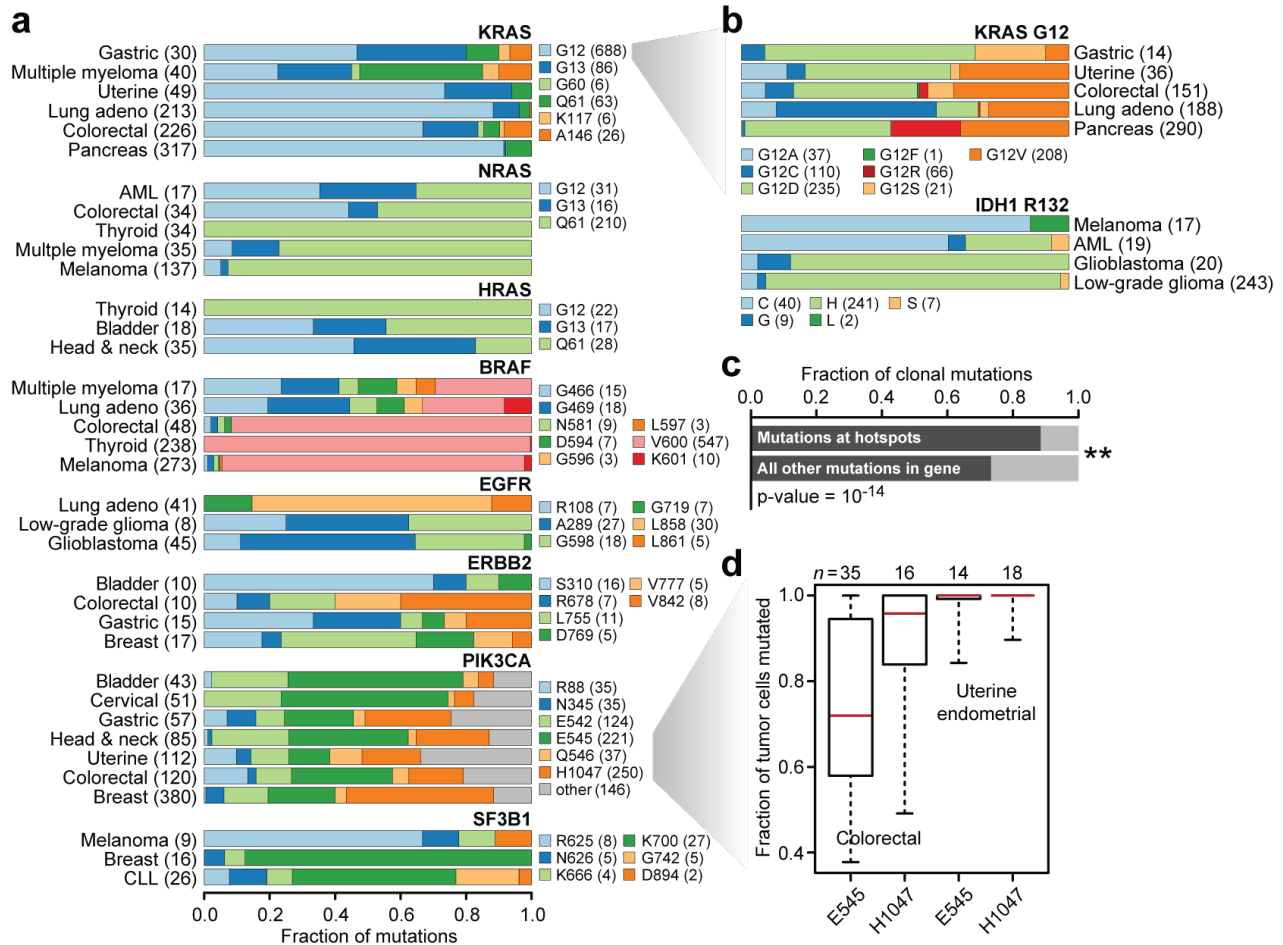


Figure 2.9: Lineage diversity and mutant allele specificity

a) The fraction of cases mutated for each of the most common hotspots in 8 frequently mutated genes in the most commonly mutated lineages indicate substantial lineage diversity and hotspot specificity. **b)** Same as in panel (a), but for *KRAS* G12 and *IDH1* R132 mutations, showing that mutant amino acid specificity exists within individual hotspots across affected tumor types. **c)** The fraction of clonal mutations, those present in 80% or more of the tumor cells of affected samples, was higher among mutations in hotspots versus all other non-recurrent mutations in the same genes (χ^2 , p -value = 1×10^{-14}). **d)** The fraction of tumor cells mutated for *PIK3CA* E545 and H1047 hotspots in affected colorectal and uterine endometrial cancers indicates a pattern of allele-specific subclonality for E545 mutations in colorectal cancer

Similar differences emerged in other driver cancer genes with multiple hotspots. V600E mutations describe nearly all *BRAF* hotspot mutations in melanoma, papillary thyroid, and colorectal carcinomas, whereas multiple myelomas are similar to lung adenocarcinoma in which non-V600E hotspots predominate (p-value = 1.9×10^{-32}). The balance between extracellular and kinase domain mutations in *EGFR* between brain tumors and lung adenocarcinoma (p-value = 3.3×10^{-12}) respectively have been documented previously and affect their biological impact and the efficacy of genotype-directed therapy⁴⁴. *ERBB2* followed a similar pattern, where extracellular domain mutations typified by S310F are far more common than are kinase domain mutations in bladder cancers compared to breast cancers (p-value = 0.006, **Figure 2.9a**). Another notable gene was *PIK3CA*. While bladder and cervical cancers are similar in their distribution of *PIK3CA* hotspot mutations, they vary significantly from breast cancers in the overall balance of helical to kinase domain mutations, possessing far fewer H1047R mutations among *PIK3CA*-mutated cases (p-value = 4.8×10^{-19}). Endometrial and colorectal cancers also have a similar pattern of *PIK3CA* hotspots, but both have a higher prevalence of R88Q mutations than any other tumor type (p-value = 1.3×10^{-11} , **Figure 2.9a**). Such patterns extend beyond essential MAPK or PI3K signaling components, such as with *SF3B1* K700 mutations that predominate in breast cancers and chronic lymphocytic leukemias whereas melanomas more frequently possess R625 mutations (p-value = 0.0001). Finally, mutant amino acid specificity was not limited to hotspots in Ras genes. The *IDH1* R132H hotspot mutation predominated in multiple brain tumor types, but cysteine was the most common *IDH1* R132 mutant amino acid in

melanoma, which is unlikely to be exclusively related to UV light exposure, as this is also true in AMLs that lack a UV-driven etiology (p-value = 3.9×10^{-21}). Together, these results indicate that substantial mutant amino acid specificity exists among hotspot mutations across highly diverse tumor lineages. Two related conclusions may be drawn from these data. First, different hotspots in the same gene may possess in many cases different function, much of which may be lineage-dependent, while not excluding the possibility that some may still arise as a function of differing underlying mutational mechanisms. Second, that perhaps different mutant amino acids within the same hotspot can be functionally different, support for which is growing^{86,88}.

2.3.4 Timing of individual hotspots

We next sought to determine if hotspot mutations, many of which are likely driver mutations and in some cases may serve as the initiating lesion, typically arise earlier than do non-recurrent mutations in the same genes and are therefore more often clonal. Overall, mutations at hotspot residues more often resided in a greater fraction of tumor cells (see Section 2.2) and were therefore earlier arising (presumptive clonal), than were non-hotspot mutations in the same genes (**Figure 2.9c**). So, while prior work has shown that driver genes in lung adenocarcinomas were enriched for clonal mutations⁹⁸, we found that this was true of hotspot mutations across a broad class of cancer genes and tumor types. However, there was considerable variability among hotspots. While colorectal and endometrial cancers have a similar pattern of *PIK3CA* hotspot mutations (**Figure 2.9c**) and share hypermutated subtypes of tumors driven by MSI and *POLE*

exonuclease domain mutations^{35,36}, colorectal tumors were unique in the clonality of the E545 and H1047 mutations. The majority of *PIK3CA* E545 helical domain mutations in colorectal cancers were subclonal, whereas H1047 kinase domain mutations were clonal, a difference that was not apparent in endometrial tumors, in which both are early clonal mutations (**Figure 2.9d**). This may be a function of the pattern of oncogenic co-mutation in these tumors as *PIK3CA* E545, but not H1047, mutations were significantly associated with *KRAS* mutations in these colorectal cancers (χ^2 p-value = 0.0004) and in previous cohorts⁹⁹. Overall, these differences in the molecular timing of specific hotspots augurs potentially important differences in their function in tumor initiation versus progression that requires further study.

2.3.5 Hotspots in the long right tail

Consistent with the so-called *long tail* of the frequency distribution of somatically mutated genes across cancer, we found that 85% of all hotspots identified here were mutated in less than 5% of tumors of all cancer types in which they were found (**Figure 2.10a**). Such findings have led to calls for sequencing up to many thousands of additional specimens from every tumor type³³. However, many hotspots present at low frequency across cancers are not mutated commonly or significantly in even a single cancer type. Indeed, 23% of all hotspots identified here were present in only one or two samples in the tumor types in which it was observed. This included 19 hotspots arising in only one sample of each affected cancer type such as *U2AF1* I24, *MYC* T58, the hyperactivating *MTOR* I2500⁹, *PIK3CB* D1067, *EP300* H1451, and *ERBB3* M60.

Therefore, many driver mutations (rather than genes) may never be found mutated at even the minimal frequencies (2-3%) proposed by previous studies as a goal in each cancer type. Conversely, population-level analysis, rather than by individual cancer type or organ system, allows identification of hotspots that arise as even private mutations in rare malignancies, for which additional broad-scale sequencing is most challenging. While rare, such recurrent alleles are evidence of selection and may be associated with specific phenotypes, such as exceptional responses^{100,101}, de novo resistance to cancer therapy, or reveal specific facets of pathway biology.

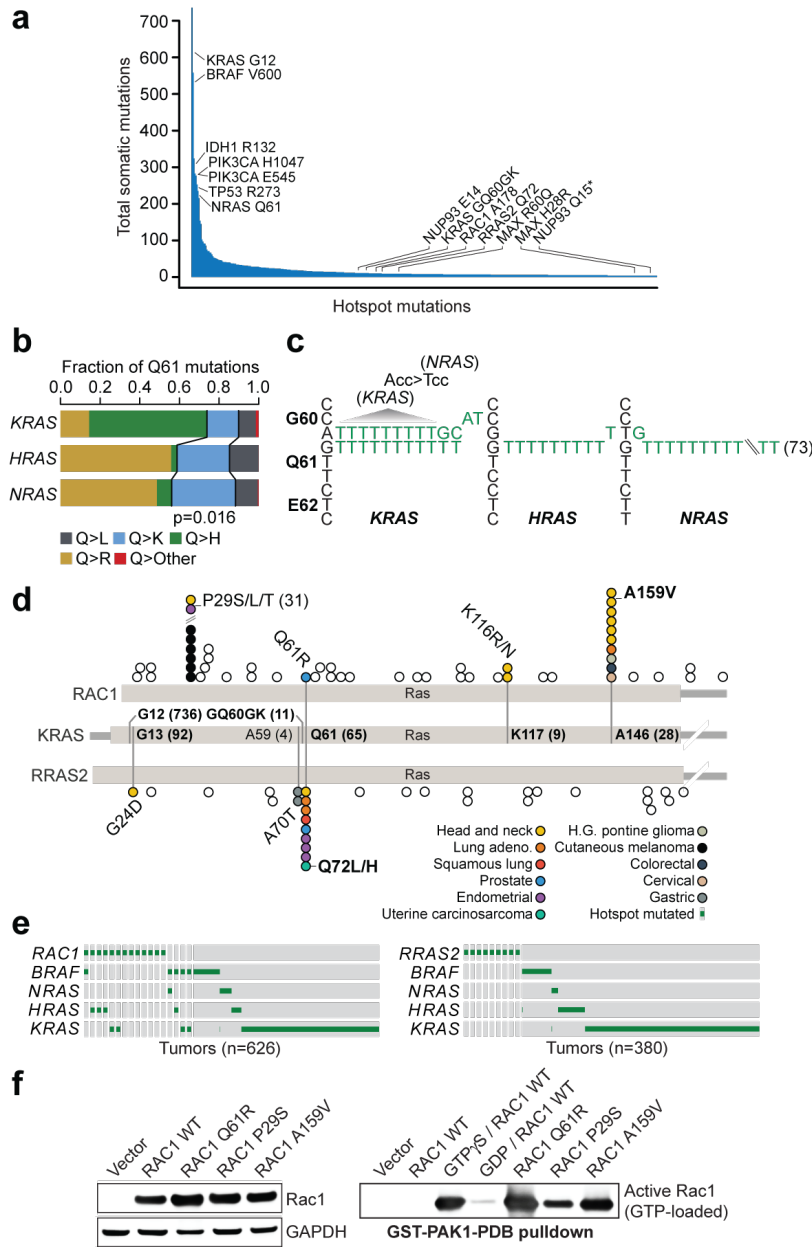


Figure 2.10: Candidate GTPase driver mutations in the long tail

a) The frequency distribution of hotspot mutations in cancer has a long right tail of mutated residues that while recurrent, are not common in any cancer type.

b) There is a statistically significant difference in the pattern of Q61 codon mutations in *KRAS*, *HRAS*, and *NRAS* (χ^2 , p-value = 0.016).

c) The sequence of Gly60-Glu62 of *KRAS*, *HRAS*, and *NRAS* are shown along with mutant alleles from affected cases indicating the GQ60GK dinucleotide mutation was the only source of *KRAS* Q61K mutation, whereas the far more common *HRAS* and *NRAS* Q61K mutations arose almost exclusively from single nucleotide events. The *KRAS* G60G synonymous mutation also creates a G60 codon in sequence (ACC>TCC) identical to wildtype sequence of *NRAS* G60, where Q61 mutations are

the most common. d) *RAC1*, *RRAS2*, and *KRAS* are shown in schematic form indicating the position of novel hotspots *RAC1* A159V and *RRAS2* Q72L/H at paralogous residues in the Ras domain to known activating mutations in *KRAS* (A146 and Q61 respectively). e) The pattern of *RAC1* (left) and *RRAS2* (right) mutations along with those in *BRAF* and Ras genes in affected tumor types. f) *RAC1* activation (GTP-bound *RAC1*) by PAK1 pull-down (right). *RAC1* A159V was associated with significant *RAC1* activation to levels equal to or exceeding the positive control GTP γ S and greater than those of the known oncogenic *RAC1* P29S

* In collaboration with Sizhi Paul Gao

2.3.6 Hotspot mutations in transporters and transcriptional regulators

Among notable long-tail hotspots was E14K in Nucleoporin 93kDa (*NUP93*) (**Figure 2.10a**). This highly expressed essential gene encodes a critical subunit of the nuclear pore complex. This hotspot was present in six breast cancers and one sample each of bladder, head and neck, hepatocellular, lung adenocarcinoma, and papillary thyroid cancers (**Figure 2.11a**, left). Among assessable breast cancers, these appear to arise in HER2-negative luminal tumors and is the fifth most commonly mutated gene in the 1303 breast cancers studied here (after hotspots in *PIK3CA*, *TP53*, *SF3B1* K700E, and *AKT1* E17K) (**Figure 2.11a**, right). Directly adjacent to E14K was a Q15* truncating hotspot, however, affected tumors expressed high levels of both the wildtype and mutant alleles. There was no detectable effect on gene expression of transcripts carrying a mutation predicted to trigger nonsense-mediated decay¹⁰². This is consistent with prior studies of loss-of-function alleles in human genomes¹⁰³, but contrary to the effect of such mutations in other cancer genes such as *TP53*¹⁰⁴ and even *CDKN2A* (**Figure 2.8c**).

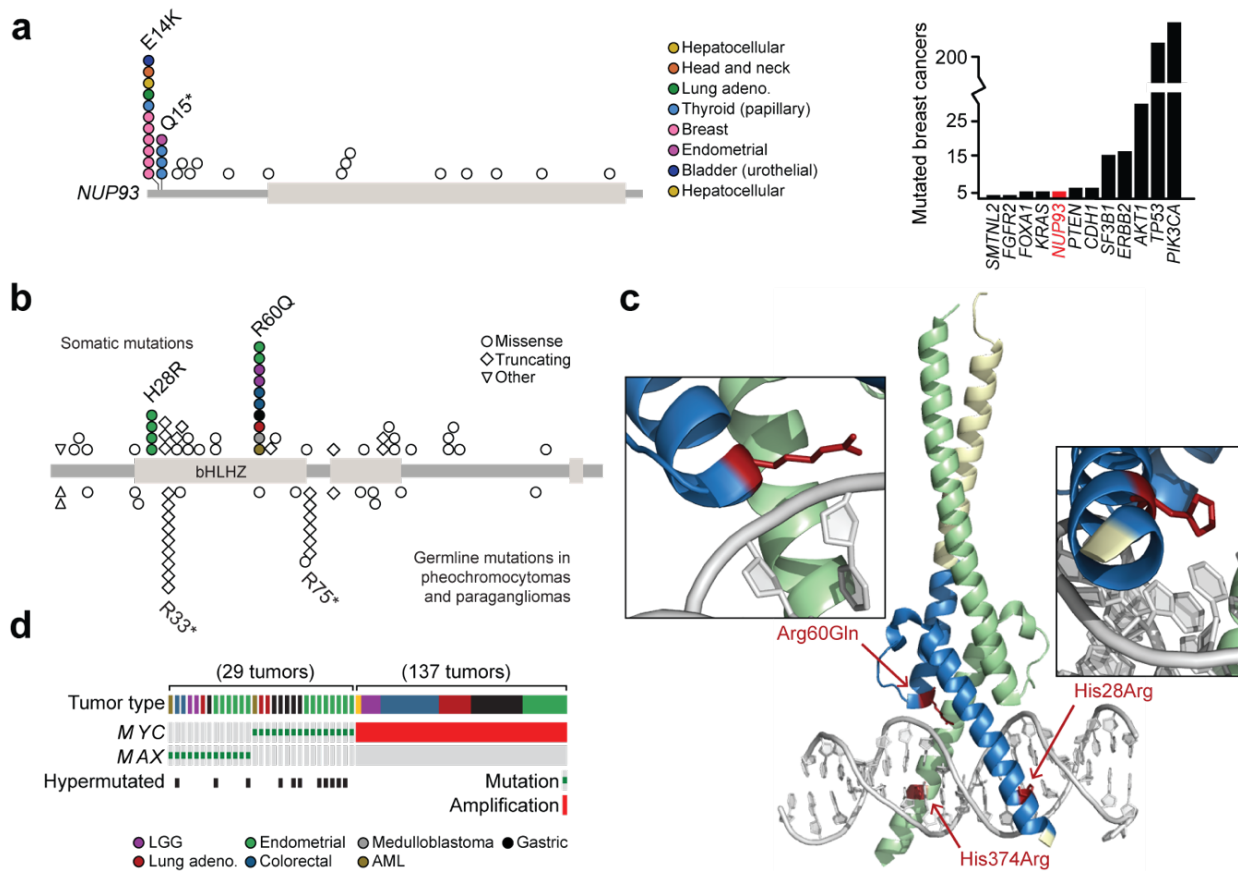


Figure 2.11: Additional candidate long tail hotspots.

a) Two hotspot mutations were detected in the N-terminal of *NUP93* (E14K and Q15*), a constitutively essential gene. The E14K hotspot was recurrently mutated in breast cancer (55% of all E14K mutants), making it among the most common hotspots in breast cancers (right). **b)** Two somatic missense hotspots H28R and R60Q affect the bHLHZ domain of *MAX* in a diversity of tumor types (indicated at bottom of panel **c**). These hotspots are different in type and position to the germline nonsense mutations present in sporadic pheochromocytomas and paragangliomas (bottom). **c)** The MYC:MAX heterodimer bound to DNA in which the DNA binding domain of *MAX* is highlighted (in blue) indicates the position of R60Q and H28R hotspots in highly conserved residues at the 5' and 3' end of the canonical E-box CACGTG motif respectively. A H374R mutation in *MYC* (annotated), also present in a uterine endometrial like *MAX* H28R mutations, is at a site equivalent to *MAX* H28R, extending the affected subset of cases in this tumor type. **d)** *MAX* hotspot mutations are mutually exclusive with mutations and amplification of *MYC* in affected tumor types, irrespective of hypermutation status.

Among other genes with two or more hotspots in the long tail, mutations in the MYC-associated factor X (*MAX*) were notable. *MYC* is an oncogene broadly implicated in the pathogenesis of multiple human cancers. While genomic amplification of *MYC* is common in many tumor types, *MYC* mutations are rare. We identified two *MYC* hotspots in this study (T58 and S146L), in one to three tumors each of head and neck cancers, lung adenocarcinomas, melanomas, lymphomas, neuroblastomas, colorectal cancers. However, *MYC*-mediated transformation through either activation or repression of *MYC* targets is dependent on its heterodimerization with *MAX*¹⁰⁵, which is an integral and constitutively expressed protein. It was notable, therefore, that we identified two *MAX* hotspots mutations (H28R and R60Q) in the helix-loop-helix (bHLHZ) DNA binding domain (**Figure 2.11b**). While recurrent germline *MAX* mutations have been reported in hereditary and sporadic pheochromocytoma and paragangliomas^{106,107}, these were truncating mutations at different residues compared to the somatic missense hotspots detected here (**Figure 2.11b**). The three-dimensional structure of the MYC-MAX heterodimer revealed that the R60 and H28 interact with 5' CA and 3' G of the CACGTG E-box respectively (**Figure 2.11c**), indicating that the mutations target DNA binding of the complex rather than *MYC* dimerization. Notably, all four H28R mutations and 20% of the R60Q mutations arose in endometrial tumors spanning three of the four previously established subtypes, including one *POLE*-ultramutated, three MSI-H hypermutated, and two copy number-low endometrioid-like tumors. Moreover, we also identified in another copy number-low endometrial tumor a *MYC* H374R mutation that is homologous to *MAX* H28R (**Figure 2.11c**). The presence of these mutations in diverse

cancer types and subtypes driven by very different underlying mutational processes indicates they are unlikely passengers due only to the mutational burden of the affected tumors. Finally, whereas the truncating germline mutations in *MAX* imply a tumor suppressor role, we found that *MAX* hotspots mutations were mutually exclusive with *MYC* mutations and genomic amplifications across affected tumor types (**Figure 2.11d**). This suggests that somatic *MAX* hotspots may be gain-of-function. However, due to the complexity of *MYC* function and the functional antagonism of *MAX* heterodimerization with *MAD*¹⁰⁸, functional validation is necessary.

2.3.7 Long-tail Ras superfamily hotspots

Mutations in the Ras family of small GTPases occur widely in human cancers. As expected, these were among the most significant hotspots detected here, affecting 1,335 tumors (12% of all cases). Whereas G12, G13, and Q61 codon hotspots predominate in *KRAS*, *NRAS*, and *HRAS*, albeit at varying frequencies in different tumor types (**Figure 2.5a** and **Figure 2.9a**), we also identified GQ60GK, K117, and A146 hotspots in *KRAS*. Both K117 and A146 are known activating hotspots in the long tail, but we also identified a previously occult GQ60GK dinucleotide substitution (q-value = 2.3×10^{-6}) in 11 tumors. This dinucleotide substitution results in a Q61K mutation accompanied by a G60 synonymous mutation that are present in *cis* (in concomitant RNA sequencing, **Figure 2.12**). Although Q>K mutations at codon 61 can result from 3'G>T single-nucleotide mutations in *KRAS*, 100% of these tumors harbored the dinucleotide substitution, a rare spontaneous event in human genomes. Overall, the

distribution of codon 61 mutations in *KRAS*, *NRAS*, and *HRAS* are very different, with Q>K mutations occurring significantly less frequently in *KRAS* (p-value=0.016; **Figure 2.10b**). GA>TT mutations were the most common dinucleotide substitution producing GQ60GK (**Figure 2.12c**) and converts the ACC codon at *KRAS* G60 to TCC, which is the sequence of the G60 codon in *NRAS*, in which Q61K mutations are far more common and arise nearly exclusively from single-nucleotide mutations. It remains to be determined whether *KRAS* GQ60GK is therefore driven by a pattern of codon usage at the -1 position. Notably, only one tumor had evidence of a non-*KRAS* GQ60GK mutation, an *NRAS*-mutant cutaneous melanoma (**Table 2.3** and **Figure 2.10c**).

Gene	Genomic	cDNA	Protein	Tumor type	Hypermutated	Alternative MAPK driver
<i>NRAS</i>	GT>TG	c.180_181AC>CA	GQ60GK (Q61K)	Cutaneous melanoma	No	--
<i>KRAS</i>	GA>TT	c.180_181TC>AA	GQ60GK (Q61K)	Pancreatic adenocarcinoma	No	--
				Transitional cell bladder	No	--
				Colorectal	MSI-H	--
				Colorectal	MSI-H	--
				Colorectal	MSI-H	Subclonal BRAF R509*
				Papillary thyroid	No	Subclonal BRAF V600E
				Papillary thyroid	No	--
				Multiple myeloma	No	--
				Colorectal	No	--
				Colorectal	No	--
<i>KRAS</i>	GA>TC	c.180_181TC>GA	GQ60GK (Q61K)	Cutaneous melanoma	No	--
	C>T	c.179G>A	G60D	Squamous cell carcinoma	Yes	BRAF S129L
<i>HRAS</i>	C>T	c.178G>A	G60S	Colorectal	No	--

Note: three additional G60 mutations are present in COSMIC (G60D/V/A in a CMML, colorectal metastasis, and papillary thyroid cancer respectively)

Table 2.3: GQ60GK and G60 mutations in Ras genes. Table of GQ60GK and G60 mutations in *K/N/HRAS*. These mutations were found a variety of cancer types both in hypermutated and non-hypermutated samples. Samples with GQ60GK or G60 mutations lack other clonal MAPK pathway drivers.

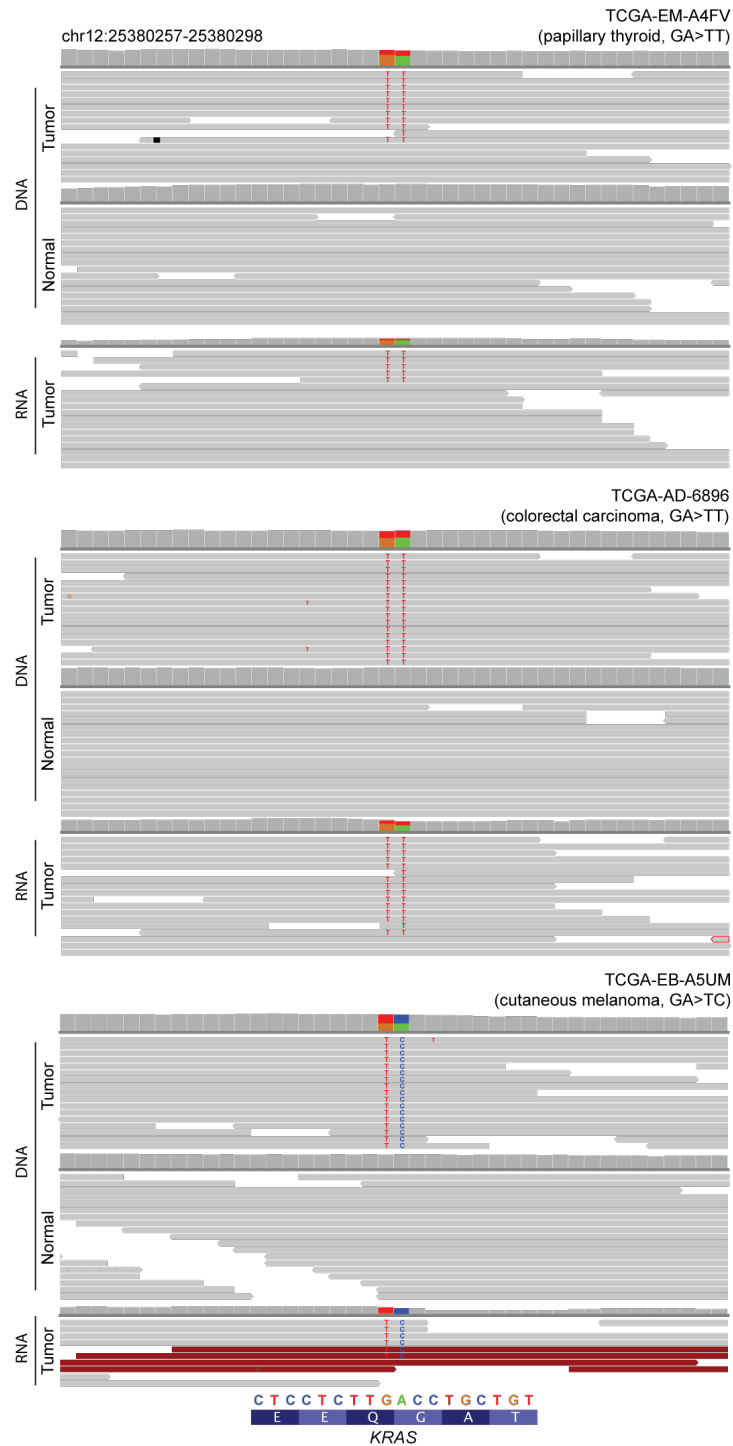


Figure 2.12: GQ60GK mutations are a single genomic event

Shown are aligned reads from whole-exome sequencing of the tumor and matched normal DNA and RNA sequencing of the tumor from representative affected cases indicating that the GQ60GK dinucleotide mutation is a single event expressed in *cis*

We next explored whether *KRAS* GQ60GK may serve as a driver of Ras pathway activity as do conventional *KRAS* hotspots. GQ60GK is indeed present in diverse tumor types that all have well-established Ras-driven subsets (**Table 2.3**). Reasoning that if GQ60GK were a passenger mutation in Ras-driven tumors, alternative MAPK activating mutations may be present in these tumors. Instead, we found that in every GQ60GK-mutant sample where another putative driver of MAPK signaling was present, that lesion was either 1) subclonal, defining a different clone than did GQ60GK; 2) low activity; or 3) a passenger mutation (**Table 2.3**). Also, despite the frequency of GA>TT, there was no evidence that a common underlying mutational process or exogenous mutagen was the source of GQ60GK. There was no evidence of UV light exposure in the clinical histories or nucleotide contexts of most affected cases, only one of which was a cutaneous melanoma. Moreover, GQ60GK arose in both hypermutated (MSI-H colon lacking *BRAF* V600E) and non-hypermutated tumors. Finally, rare G60 missense mutations were evident in *K-* and *HRAS* in this dataset and in the literature (**Table 2.3**)¹⁰⁹. So, while we cannot exclude the possibility that the GQ60GK dinucleotide substitution is simply an alternative mechanism to achieve Q61K, the accompanying *KRAS*-specific G60 synonymous mutation may potentiate a different class of Q61-mutant tumors or cause signaling differences among Q61K-mutant tumors between *K-N-* or *HRAS*. Although further studies will need to explore the molecular properties of *KRAS* GQ60GK, this allele represents the most common dinucleotide substitution spanning two codons in human cancer and a mutation more common than other known hotspots in *KRAS*.

Novel long-tail hotspots were also identified in two other genes that encode members of the Ras superfamily of small GTPases. RAC1, in which we identified two hotspots, is a Rho subfamily member that plays a vital role in various cellular functions. *RAC1* P29S is an oncogenic hotspot in melanomas^{69,70} that we also identified in head and neck and endometrial cancers (**Figure 2.10d**). This mutation can confer resistance to RAF inhibitor treatment *in vitro*⁷¹, and may underlie early resistance in patients¹¹⁰. We also identified a novel *RAC1* A159V hotspot present in 10 tumors (q-value = 2.27×10^{-6} ; **Figure 2.10d**). Notably, *RAC1* A159V is paralogous to *KRAS* A146, a known activating mutation⁶⁵. Whereas activating *KRAS* A146T mutations arise predominantly in colorectal carcinomas, *RAC1* A159V mutations are most common in head and neck cancers and were not present in any melanomas, despite the frequency of *RAC1* P29S in this cancer type. Moreover, similar to P29S mutations, we observed *RAC1* A159V mutations in tumors that are both Ras/Raf wildtype and mutant (**Figure 2.10e**). To determine whether *RAC1* A159V is an activating mutation, we assessed its effect *in vitro*. Active RAC1 is GTP-bound, interacting with PAK1 to activate downstream effectors. Therefore, to quantify RAC1 activation *in vitro*, we utilized a PAK1 pull-down assay. In HEK293T cells expressing *RAC1* A159V, there was significant RAC1 activation to levels equal to or exceeding positive control RAC1 GTPγS cells and greater than even those levels induced by the known *RAC1* P29S oncogenic mutation (**Figure 2.10f**). Moreover, cells expressing *RAC1* Q61R, a mutation we identified in a primary prostate cancer that is paralogous to *KRAS* Q61, also potentially induced RAC1 activation (**Figure 2.10d-f**).

RRAS2 is a Ras-related small GTPase¹¹¹. *RRAS2* is overexpressed or mutated in a small number of cancer cell lines of various origins¹¹²⁻¹¹⁴, and is oncogenic *in vitro* with transforming ability similar to established Ras oncoproteins¹¹⁵. However, it has not been documented as somatically mutated in human tumor specimens. Here, we identified a *RRAS2* Q72 hotspot present in nine tumors (q-value = 8×10^{-15}). Similar to *RAC1* A159V, the *RRAS2* Q72 hotspot is paralogous to *KRAS* Q61 (**Figure 2.10d**). However, unlike *RAC1*, *RRAS2* Q72 does not predominate in any individual tumor type. Also unlike *RAC1*, the *RRAS2* Q72 mutation was present in Ras/Raf wildtype tumors among the affected types (**Figure 2.10e**). This result suggests that *RRAS2* activation may be an alternative avenue for tumors to acquire Ras-like activation as previous studies have shown that *RRAS2* shares many Ras downstream signaling elements including phosphatidylinositol-3 kinase (PI3K)^{116,117}, the Ral GDP dissociation pathway¹¹⁶, and Raf kinases¹¹⁸. Beyond these hotspots, several less common *RAC1* and *RRAS2* mutations affect paralogous residues of highly recurrent alleles in *KRAS* (**Figure 2.10d**), some of which we validated were also activating *in vitro* (**Figure 2.10f**), indicating that the landscape of potentially functional mutations in these genes extends beyond even these less common long-tail hotspots to private mutations as well.

2.4 Discussion

Our work suggests that while a subset of hotspots were prevalent in individual cancer types, most hotspots are present infrequently across many cancer types. This indicates that studies of any individual cancer type may have limited power to identify

novel alleles. We have also begun to detail best practices for the use of diverse public cancer sequencing data in the translational setting. Our approach for hotspot detection incorporates features such as the variable background mutational burden of individual codons and genes, thereby avoiding passenger mutations whose recurrence is due only to their presence in highly mutable amino acids. While the identification of private driver mutations remains challenging, our approach did uncover low-incidence hotspots in highly mutated genes. Though less common, these hotspots are under selection and may confer important clinical phenotypes in cancer patients, such as exceptional responses to cancer therapy^{100,101}.

New mutant alleles in established genes are likely to emerge faster than new cancer genes are identified, extending the long tail of the frequency distribution of somatic mutations. This is especially true as clinical sequencing focuses on profiling advanced and metastatic disease for clinical trial enrollment. Such pre-treated, late-stage cases have been historically under-represented among such population-scale resources, including the one studied here. Moreover, at present there are fewer actionable mutations in cancer than there are cancer genes. Yet the near-term clinical utility of expanding the former is far greater. Our results suggest this will require an understanding of the function of different hotspot mutants in the same gene by lineage, as their function and response to therapy may be mutant amino acid specific. While positive selective pressure may produce the same hotspot mutation, or different variant amino acid changes within the same hotspot residue, it does not imply that they will confer similar selective advantages across lineages. Underlying functional distinctions

may explain the differences observed here in the emergence and frequencies of hotspots across lineages. While this remains speculative or unknown for most hotspots, early evidence suggest that this will be true for even some of the most important alleles in human cancer⁸⁶⁻⁸⁸. Understanding this landscape of distinct molecular function is the necessary translational prerequisite for effective clinical implementation. This focus on mutations rather than genes will spur studies of the biochemical, biological, signaling impact, and drug sensitivity of candidate individual alleles. Collectively, the complementary study of both significantly mutated individual alleles as well as genes will prove indispensable in enabling precision oncology through clinical decision support for patients sequenced at the point of care.

CHAPTER 3*

IDENTIFYING 3D MUTATIONAL CLUSTERS

3.1 Background

Continued sequencing efforts, both within and across a broad spectrum of cancer types, has revealed a complex landscape of somatic mutations in various cancer types¹¹⁹. While the data generated have provided a more complete picture of the genomic aberrations in cancer cells, the interpretation of individual mutations can be difficult. One of the key challenges is distinguishing the few mutations that functionally contribute to oncogenesis (“drivers”) from the many biologically neutral mutations (“passengers”)¹²⁰. While increasing the number of tumor types and samples in this analysis or refining computational methodologies will likely identify additional driver mutations, identifying rare or even private driver mutations in the long tail remains a persistent challenge. Though individually rare, these long-tail mutations are present in a significant fraction of tumors and are likely key molecular events and thus potential drug targets¹. Several methods exist that identify driver genes or mutations in the long tail by incorporating protein-level annotation, such as local positional clustering¹²¹, phosphorylation sites¹²², and paralogous protein domains¹²³.

Recently, three-dimensional (3D) protein structures have also been used to identify driver genes and mutations in cancer and other diseases. For example, Dixit et al.¹²⁴ studied cancer mutations in 3D structures of protein kinases. Wang et al.¹²⁵ generated a

*Gao J[#], Chang MT[#], Johnsen HC, Gao SP, Sylvester BE, Sumer SO, Zhang H, Solit DB, Taylor BS, Schultz N[#], Sander C[#]. 3D clusters of somatic mutations in cancer reveal numerous rare mutations as functional targets. *Genome Medicine*

structurally solved interactome to study genetic diseases. Porta-Pardo et al.¹²⁶ and Engin et al.¹²⁷ used 3D structures to detect protein-protein interaction interfaces that are enriched with cancer mutations. CLUMPS¹²⁸ used 3D clustering of mutations to detect cancer genes and also studied enrichment of mutations in protein-protein interaction interfaces. StructMAN¹²⁹ annotated the amino acid variations of single-nucleotide polymorphisms (SNPs) in the context of 3D structures. SpacePAC¹³⁰, Mutation3D¹³¹, HotMAPS¹³² and Hotspot3D¹³³ used 3D structures to identify mutational clusters in cancer. These efforts have generated interesting sets of candidate functional mutations and illustrate that many rare driver mutations are functionally, and potentially, clinically relevant.

Here, we describe a novel method that identifies mutational 3D clusters, i.e. missense (amino-acid changing) mutations that cluster together in three-dimensional (3D) proximity in protein structures above a random background, with a focus on identifying rare mutations. In this largest 3D cluster analysis of whole exome or genome sequencing data in cancer to date, we analyzed over one million somatic missense mutations in 11,119 human tumors across 32,445 protein structures from 7,390 genes. The analysis identified potential driver mutations, the majority of which are rare mutations (occurring in <0.03% of patients in the dataset), in 3,405 residues clustering in the protein structures of 503 genes (**Figure 3.1**). Many of these 3D clusters were identified in well-characterized cancer genes, such as *KRAS*, *BRAF*, and *TP53*, and include known oncogenic recurrent alleles (e.g., *KRAS* G12D) as well as rare long-tail alleles (e.g., *KRAS* D33E, which has recently been experimentally validated¹³⁴). We

were able to identify new potential driver genes as well as novel candidate driver mutations in clinically actionable cancer genes that were not detected by our mutational single-residue hotspot detection method¹⁶ and previous 3D cluster detection methods¹³¹⁻¹³³. We experimentally tested the activating potential of rare mutations identified in 3D clusters in the *MAP2K1* and *RAC1* proteins, enlarging the number of biologically and potentially clinically significant alleles in these two critical effectors of activated signaling pathways in cancer.

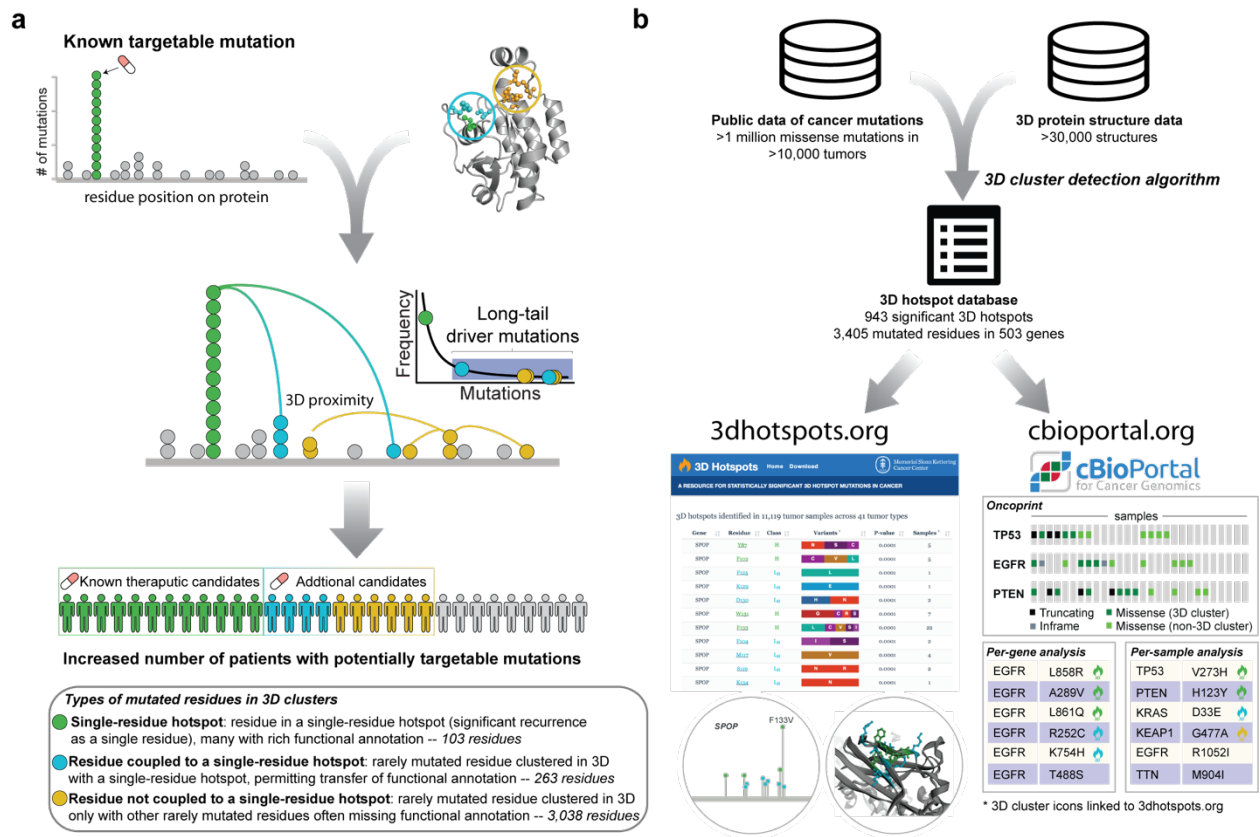


Figure 3.1: 3D method and related resources

a) Process of going beyond single-residue hotspots by considering occurrence in 3D clusters. The colors of different types of mutated residues in 3D clusters are defined in the bottom panel and used throughout the manuscript. **b)** Mutations in 3D clusters can

be explored via the web resource <http://3dhotspots.org>. The results are also made available via a web API service for use by other bioinformatics tools, and mutations viewed in the cBioPortal for Cancer Genomics are annotated if they are part of an identified 3D cluster. The identified 3D clusters are likely to change as the cancer genomics and 3D structure databases grow.

3.2 Method

Mutational data were obtained from publicly available sources including TCGA, ICGC and published studies from literature. Complete description of mutational processing can be found in Section 2.2.1.

3.2.1 Protein 3D structure data collection and pre-processing

Protein structures were downloaded from the RCSB Protein Data Bank (PDB, <http://www.rcsb.org>)¹³⁵. Alignments of protein sequences from UniProt¹³⁶ to PDB were retrieved from MutationAssessor¹³⁷ and SIFTS¹³⁸. Only alignments with sequence identity of 90% or above were included. For each structure chain, a contact map of residues was calculated. Two residues are considered in contact if any pair of their atoms is within 5 angstroms (Å), calculated by BioJava Structure Module¹³⁹. A 3D cluster is defined by a central residue and its contacting neighbor residues (**Figure 3.2a**). All residues are used in turn as centers of clusters. The test of statistical significance (below) is applied separately to each cluster in turn. Clusters are not merged, so each residue can be in more than one cluster, even after filtering for statistical significance of the clusters.

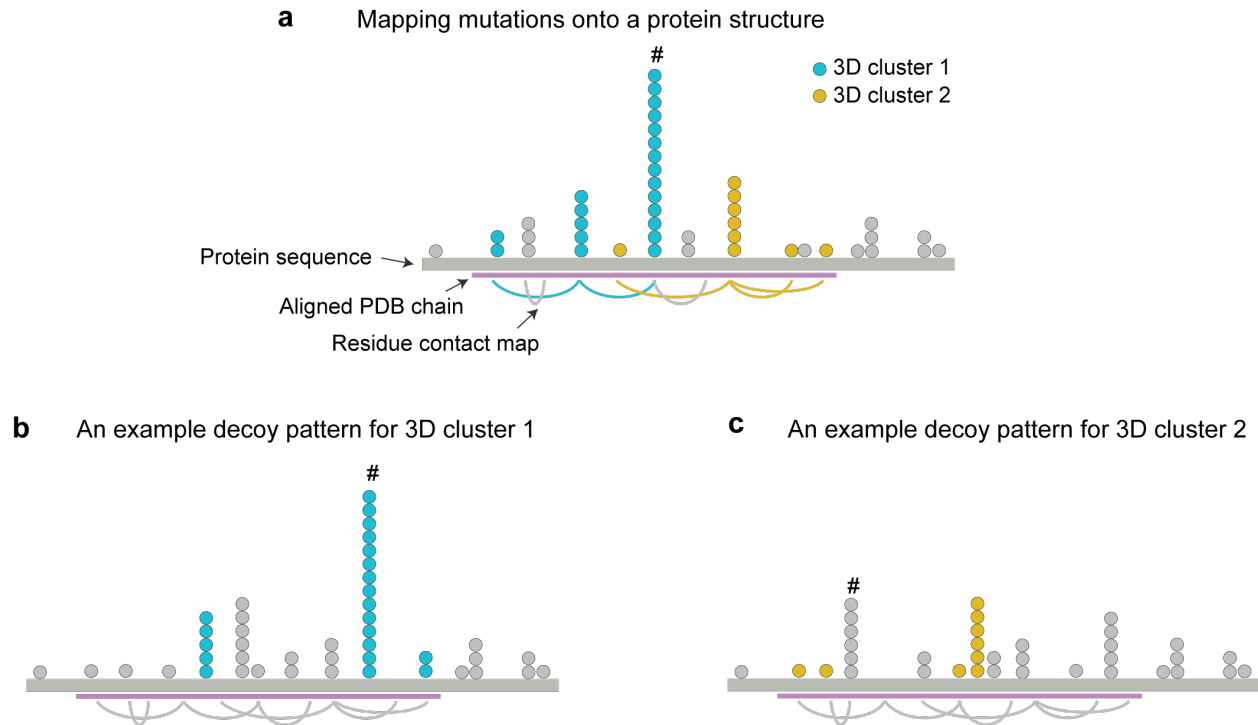


Figure 3.2: Illustration of the permutation procedure for calculating the statistical significance of 3D clusters.

3.2.2 Determining significant mutated 3D clusters

A 3D cluster was identified as significantly mutated, if its member residues were more frequently mutated in the set of samples than expected by chance. Mutations were mapped to the aligned PDB sequences and structures (**Figure 3.2a**) and the total number of mutations across all samples was calculated within each 3D cluster. To determine whether the residues in a 3D cluster in a particular structure were more frequently mutated than expected by chance, a permutation-based test was performed by generating 10^5 decoy mutational patterns on the aligned region of the protein structure. A decoy pattern was generated by randomly shuffling the residue indices (positions in the sequence), with their associated mutation count, on the structure

(**Figure 3.2b-c**). For each decoy mutational pattern, the number of mutations in clusters was calculated as above. For a given 3D cluster, the p -value was calculated as the fraction of decoys for which the number of mutations in the cluster was equal to or larger than the corresponding number (for any cluster position) in the real data. When shuffling the mutations, the mutation count in each residue was maintained, except that we set the maximum number of mutations in one residue in the decoy to the largest number of mutations in the assessed 3D cluster with the intent to ensure detection of less frequently mutated 3D clusters within a gene with one or a few dominant single-residue hotspots (such as *BRAF* V600) **Figure 3.2b-c**). In the rest of the Section, we use the term '3D cluster' as a short alias for 'significantly mutated 3D cluster'.

3.2.3 *MAP2K1* and *RAC1* functional validation

MAP2K1 functional validation. Human embryonic kidney (HEK)-293H cells were maintained in DME-HG medium with 10% fetal bovine serum, supplemented with 2 mM glutamine, and 50 units/ml each of penicillin and streptomycin.

MAP2K1 mutant constructs were generated from the MEK1-GFP plasmid (Addgene, #14746) using the QuikChange II XL Site-Directed Mutagenesis Kit (Stratagene) as recommended. All mutant plasmids were verified by Sanger sequencing. HEK-293H cells were seeded for 70% to 90% confluency at the time of transfection, then transiently transfected with the wild-type or mutant MEK1-GFP plasmid using Lipofectamine® 2000 Transfection Reagent (Invitrogen). Plasmid

transfection levels were standardized according to GFP expression. Cells were collected 24 hours post-transfection.

Cells were lysed in 1% NP-40 buffer with protease and phosphatase inhibitors, then processed for immunoblotting as previously described¹⁴⁰. Rabbit polyclonal antibodies recognizing MEK1/2, phosphorylated ERK1/2 (Thr202/Tyr204), and ERK1/2 were obtained from Cell Signaling. Rabbit monoclonal antibodies recognizing GFP and GAPDH were obtained from Cell Signaling. After incubation with horseradish peroxidase-conjugated secondary antibody, proteins were detected by chemiluminescence (SuperSignal West Dura Chemiluminescent Substrate, Thermo Scientific) and visualized using the Fuji LAS-4000 imager (GE Life Sciences).

HEK-293H cells were transfected with MEK1 wild-type or mutant GFP-tagged plasmid. At 24 hours, cells were treated with 100 nM trametinib and collected after 2 hours. Control cells were treated with dimethyl sulfoxide (DMSO). Cells were lysed for protein and immunoblotted as referenced above.

RAC1 functional validation. Early-passage HEK293T cells, acquired from ATCC and authenticated as mycoplasma free, were cultured at 37°C in 5% CO₂ in Dulbecco's modified Eagle medium (DMEM) supplemented with 10% fetal bovine serum (FBS).

RAC1 mutation validation was performed similar to what was previously described in Section 2.2. DNA coding sequences for mutant RAC1 constructs were generated via site-directed mutagenesis (Genewiz, NJ). All mutant plasmids were verified by Sanger sequencing. RAC1 constructs contained an N-terminal 3xFLAG epitope tag and were subcloned into a pcDNA3 mammalian expression vector (Life Technologies, NY). The

expression constructs were transfected into these cells using Lipofectamine 2000 (Life Technologies).

Cells were harvested 72 hours after transfection. GTP-bound Rac1 (active Rac1) was isolated via immunoprecipitation using recombinant p21-binding domain (PBD) of PAK1 (PAK1-PBD; Active Rac1 Detection Kit, Cat#8815, Cell Signaling, MA), according to the manufacturer's instructions. Total Rac1 was detected using kit-provided Rac1 primary antibody.

3.3 Results

The 1,182,802 somatic missense mutations in our curated dataset of 11,119 human tumors occurred in 1,025,590 residues in 18,100 genes. The vast majority of these residues, 908,009, were mutated only once in the 11,119 samples (**Figure 3.3a**), i.e., most somatic mutations found in cancer are extremely rare. Most of these rare mutations are likely passenger mutations, but some may be unrecognized drivers¹³⁴. Indeed, we found that a small fraction of rare mutations (mutated in three or fewer samples) are members of recurrently mutated clusters in 3D structures (**Figure 3.3a**) and thus probably are functional drivers.

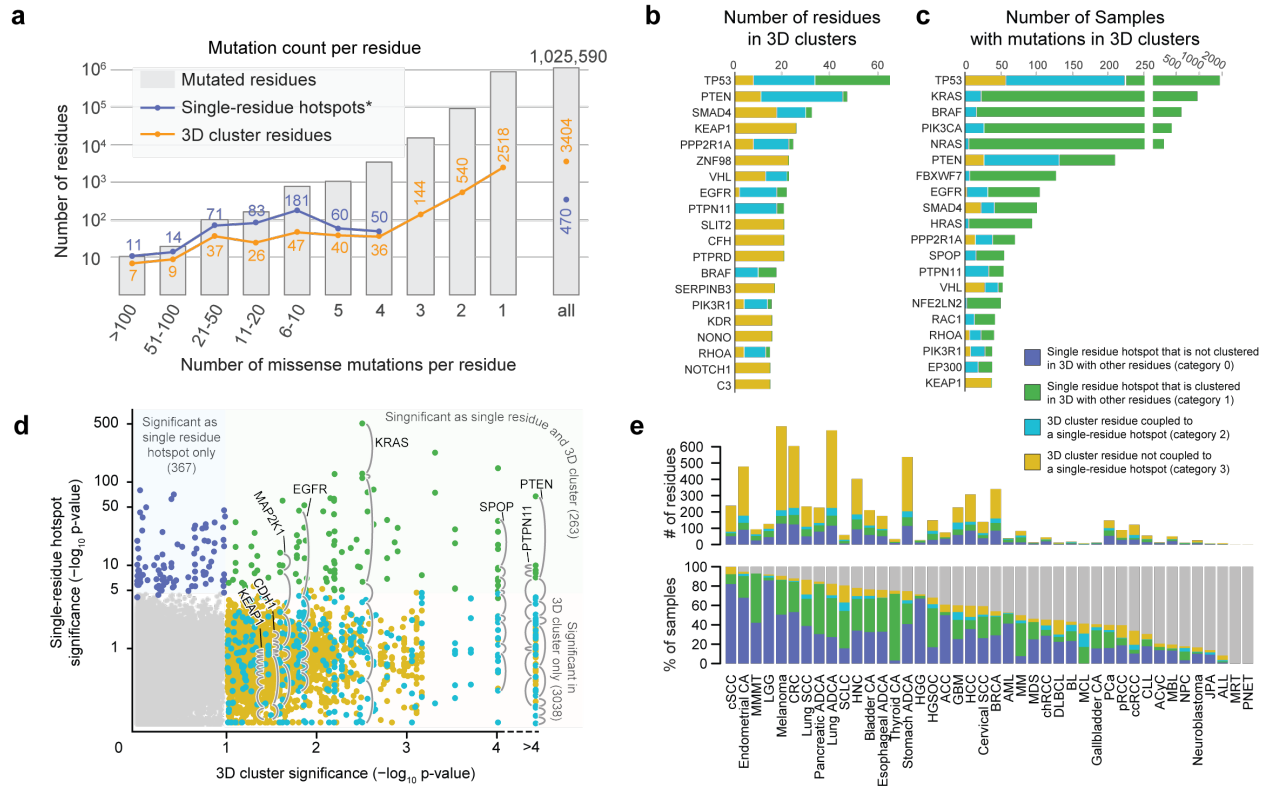


Figure 3.3: 3D cluster analysis reveals numerous potentially functional rare mutations

a) 3D cluster analysis identified a large number of statistically significant, yet rarely mutated residues (mutated 1-3 times in our dataset). The residues were binned by the number of mutations in each residue. The mutation counts for the single-residue hotspots also contain a small fraction of silent, nonsense, and splice-site mutations by Chang et al. 2016. **b)** Genes with the highest number of residues in 3D clusters. **c)** Genes with the highest frequency of tumor samples with mutations clustered in 3D structures across all cancer types. **d)** Per-residue comparison of significance as in single-residue hotspot (vertical axis) and 3D cluster (horizontal axis). Many residues were hotspots as well as parts of 3D clusters (upper right quadrant), but some were detected only as part of 3D clusters (bottom right quadrant). **e)** Number of residues (upper panel) and percentage of samples (bottom panel) with hotspots and 3D clusters per cancer type (see full cancer type names in the Abbreviations section). The category of a sample was assigned based on the lowest category if it had mutations that belonged to different categories.

In total, we identified 943 unique mutational clusters that were statistically significant in 2,382 protein structures (**Table 3.2**). These 3D clusters encompassed 3,404 residues in 503 genes (**Table 3.3**). *TP53* contained the largest number of residues in 3D clusters

(66 residues), followed by *PTEN* (n=48), *SMAD4* (n=33), and *KEAP1* (n=26) (**Figure 3.3b, Table 3.4**). *TP53* mutations in 3D clusters were also the most prevalent across all cancer types (in 1,914 samples, 17%), followed by *KRAS* (8%), *BRAF* (6%), and *PIK3CA* (4%), underscoring the roles of these well-characterized cancer genes in oncogenesis (**Figure 3.3c, Table 3.5**).

3.3.1 Classification of mutational clusters in protein structures

We classified the mutated residues in a 3D cluster into three categories (**Figure 3.1, Figure 3.3d, Table 3.3**) depending on whether the cluster contains single-residue hotspots identified by¹⁶: 1) 103 residues in single-residue hotspots. 2) 263 rarely mutated residues that were clustered in 3D with a single-residue hotspot. 3) 3,038 rarely mutated residues that were clustered in 3D only with other rarely mutated residues. If a residue belongs to category 2 in one cluster and category 3 in another, the residue is assigned category 2. There were 367 hotspots identified by¹⁶ that were not detected in 3D clusters (**Figure 3.3d**), either because they were not part of a significant cluster with other mutated residues or because there was no 3D structure available for the protein or protein region.

Notably, in 5,038 samples (45%) prior frequency-based hotspot analysis failed to identify hotspot driver mutations. By incorporating protein structure data, rare mutations present in 3D clusters were identified in 865 of these samples (17% of the samples without hotspot driver mutations, or 8% of all samples) (**Figure 3.3e**). As an example, 141 (15%) of 961 lung tumors (lung adenocarcinoma, lung squamous cell carcinoma,

and small-cell lung cancer) with no hotspot mutations carried a rare mutation in a mutational 3D cluster. Assuming the diseases of these patients were genetically driven, these 3D cluster mutations were possibly driver events (**Figure 3.3e**).

3.3.2 Rare missense mutations in occult drivers

While tumor suppressor genes are often inactivated by truncating (e.g., nonsense and frameshift) mutations, their function may also be disrupted by missense mutations in critical regions. These missense mutations, unlike hotspot mutations in oncogenes, often are not recurrent at individual positions but instead their recurrence may only be evident in mutational clusters. By using protein structures, we identified potentially inactivating mutational clusters in critical regions of several tumor suppressors including *PTEN*, *CDH1*, and *KEAP1*.

PTEN is one of the most frequently mutated tumor suppressors with mutations occurring in various cancers. In *PTEN*, we identified 15 3D clusters that included 48 residues (2 single-residue hotspots, 46 rarely mutated residues) (**Figure 3.4a**). All these clusters reside in the flanking regions surrounding the phosphatase catalytic core motif (**Figure 3.4a**), a region that is necessary for *PTEN* activity¹⁴¹.

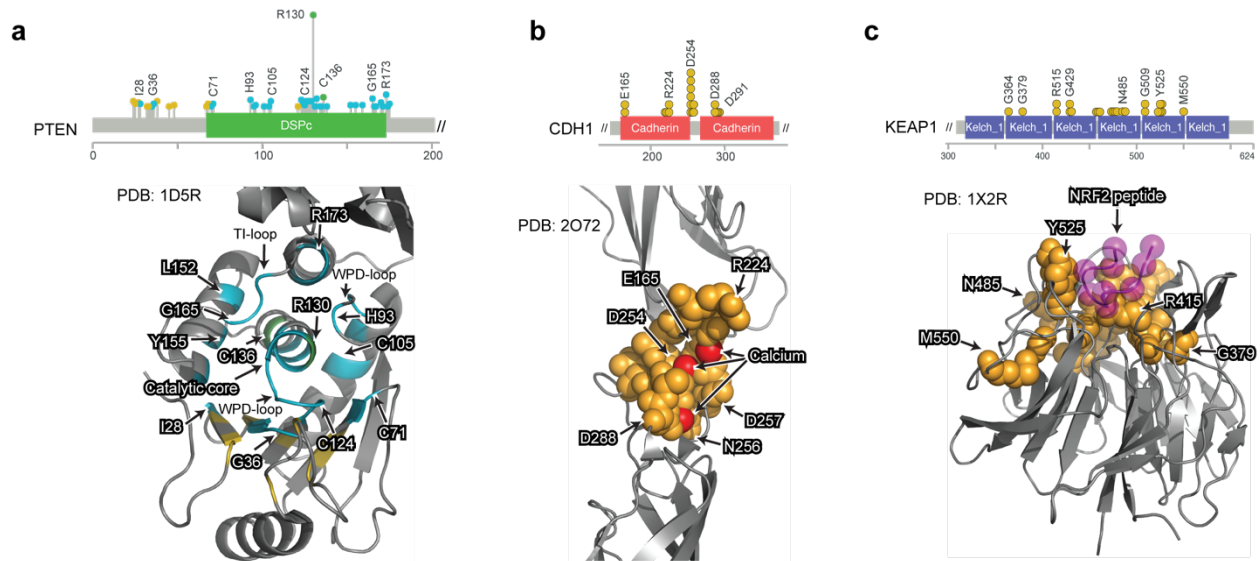


Figure 3.4: Examples of mutational 3D clusters in tumor suppressor genes

a) Residues in 3D clusters in *PTEN* highlighted in the protein sequence (top) and a protein structure (bottom). The 3D cluster residues surround the catalytic site. **b)** Residues in 3D clusters in *CDH1* (E-cadherin) highlighted in the protein sequence (top) and a protein structure (bottom). The 3D cluster mutations are likely to disrupt the critical calcium-binding site (calcium atoms in red). **c)** 3D clusters in *KEAP1* in the protein sequence (top) and a protein structure (bottom). Most of the 3D cluster mutations are in the *NRF2* binding region (*NRF2* peptide in purple).

CDH1 encodes E-cadherin, a transmembrane glycoprotein mainly expressed in epithelial cells. Germline mutations in *CDH1* are associated with an increased risk of gastric and breast cancer¹⁴², and *CDH1* somatic inactivation via epigenetic silencing or truncating mutations is common in both cancer types. We identified 11 3D cluster residues (all rarely mutated residues; mutation frequency 0.01-0.06% individually) in *CDH1* (**Figure 3.4b**). Out of the 19 samples with these 3D cluster mutations, 11 were gastric tumors. Although distant in amino acid position (between the 165th and 291st residues), when mapped in 3D space, these residues surround the junction between the first and second extracellular cadherin domains in the 3D structure (**Figure 3.4b**).

Mutations in these residues are likely to perturb functionally essential calcium-binding sites in the junction region¹⁴³ and hence are likely inactivating and potentially oncogenic.

KEAP1 is a substrate adapter protein for the E3 ubiquitin ligase that targets *NFE2L2* (*NRF2*) for ubiquitination and subsequent degradation. Loss-of-function mutations in key *KEAP1* residues result in accumulation of *NRF2* in the nucleus and contribute to chemoresistance in vitro¹⁴⁴. We identified 26 3D cluster residues (all rarely mutated residues; mutation frequency 0.01-0.03% individually) in *KEAP1* (**Figure 3.4c**). These mutations were localized to the interaction domain of *KEAP1*, suggesting that they likely disrupt *NRF2* binding (**Figure 3.4c**). Notably, out of the 36 samples with these mutations, 18 were lung adenocarcinomas, 6 of which lacked hotspot mutations.

3.3.3 Functional validation of *MAP2K1* and *RAC1* mutants

Identifying mutations in genes for which targeted therapies exist or are being developed, regardless of their individual frequency in the population, is critical for the effective practice of precision oncology. Our analysis identified 3D clusters in several genes for which selective inhibitors are either used as part of standard clinical management or are being actively tested in clinical trials, including *EGFR*, *KIT*, *MTOR*, *PIK3CA*, *MAPK1*, and *FGFR3* (**Table 3.1**). The 3D clusters within these genes contained known activating hotspot mutations as well as rare candidate driver mutations. While the function of most of these rare mutations is unknown, a subset has been functionally characterized in prior studies. For example, *EGFR* T263P has been reported to induce oncogenic EGFR activation¹⁴⁵, and recently, many of the rare

mutations in *MTOR* present within 3D clusters (A1459P, L1460P, Y1463S, T1977R, and V2006I/L) (**Table 3.1**) have been shown to induce increased mTORC1/2 pathway activity⁹.

Gene	PDB_ID:chain	Position (number of mutated samples)	<i>p</i>	Cancer types* (number of mutated samples)
EGFR	1IVO:B	R252(8) F254(1) D256(2) K261(1) T263(2) C264(1) A289(28)	0.016	GBM(30) LGG(8) Stomach ADCA(2) Other(3)
EGFR	2JIU:B	V769(1) R831(2) R832(2) L833(2) L858(30) L861(7) H893(1)	0.025	Lung ADCA(39) Lung SCC(2) CRC(2) Other(2)
KIT	4HVS:A	W557(1) V559(3) V560(1) L576(2)	0.085	Melanoma(6) Stomach ADCA(1)
MTOR	4JT5:B	A1459(1) L1460(2) V1461(1) Y1463(1) K1465(1) M1467(1) R1480(2) C1483(5)	0.035	ccRCC(7) BRCA(1) CRC(1) Other(5)
MTOR	4JSN:A	A1971(3) I1973(2) Y1974(1) T1977(3) M1998(1) V2006(2)	0.047	ccRCC(4) CLL(2) Endometrial CA(2) Other(4)
PIK3CA	2v1y_A	R38(14) E39(5) R88(40) C90(4) R93(15)	0.014	Endometrial CA(27) CRC(19) Other(32)
MAPK1	4FV5:A	E81(2) R135(1) G136(1) D321(3) E322(15)	0.014	Cervical SCC(9) HNC(9) BRCA(1) Other(3)
FGFR3	1RY7:B	R248(9) S249(18) P250(1) D280(2)	0.050	Bladder CA(17) HNC(6) Lung SCC(3) Other(4)

*Full cancer type names are listed in the Abbreviations section

Table 3.1: Select 3D clusters of functional significance. Table of select clinically actionable genes and mutations identified in 3D clusters.

To confirm that the method could identify functional driver alleles that would not have been nominated by previously reported frequency-based methods, we sought to functionally validate several non-recurrent mutations identified by our method in the *MAP2K1* and *RAC1* genes. Components of the MAPK pathway are among the most commonly altered genes in human cancer. 3D cluster analysis revealed 3D clusters in all three RAS proteins (*K/N/H-RAS*), *RAC1*, *BRAF*, *MAP2K1*, and *MAPK1* in a variety of cancer types. MEK1, which is encoded by the *MAP2K1* gene, is a dual specificity kinase that phosphorylates ERK to propagate *MAPK* signaling transduction. Activating mutations in *MAP2K1* have been shown to result in constitutive MAPK pathway activity and to confer resistance to RAF inhibition and MEK inhibitor sensitivity^{146,147}.

We identified a 3D cluster ($p=0.03$) in *MAP2K1* that included 7 mutated residues (R49, A52, F53, Q56, K57, G128, and Y130). Two of these residues (F53 and K57) have been previously identified as single-residue hotspots¹⁶ and shown to induce

constitutive ERK pathway activation³⁴. The other five were infrequently mutated (mutated in 1-3 samples, i.e. mutation frequency of 0.01-0.03%) (**Figure 3.5**). All seven of these mutated residues reside in the shared interface between helix A and the kinase domain (**Figure 3.5b**). As helix A has previously been shown to negatively regulate MEK1 kinase activity by interacting with the kinase domain¹⁴⁸, mutations that disrupt this interaction may result in constitutive ERK pathway activation. We thus experimentally assessed the ability of the mutations in this 3D cluster to induce ERK1/2 phosphorylation in a cellular model. We found that expression of five of the mutated proteins, including G128D, Y130C, as well as the previously characterized F53L, Q56P, and K57N mutations³⁴, induced downstream MAPK signaling as assessed by increased expression of phosphorylated ERK (**Figure 3.5c**). To test whether the Y130C variant protein that is not in a single-residue hotspot but was nominated by 3D cluster analysis, is sensitive to MEK inhibition, we treated HEK293T cells expressing the Y130C mutant, or the Q56P mutant as a positive control, with trametinib, an FDA-approved MEK inhibitor. Trametinib treatment resulted in significant down-regulation of MAPK pathway activity (**Figure 3.5d**). As durable responses to MEK inhibitors have been reported in patients, whose tumors have an activating mutation in MAP2K1¹⁴⁷, this example highlights the potential translational impact of 3D cluster analysis.

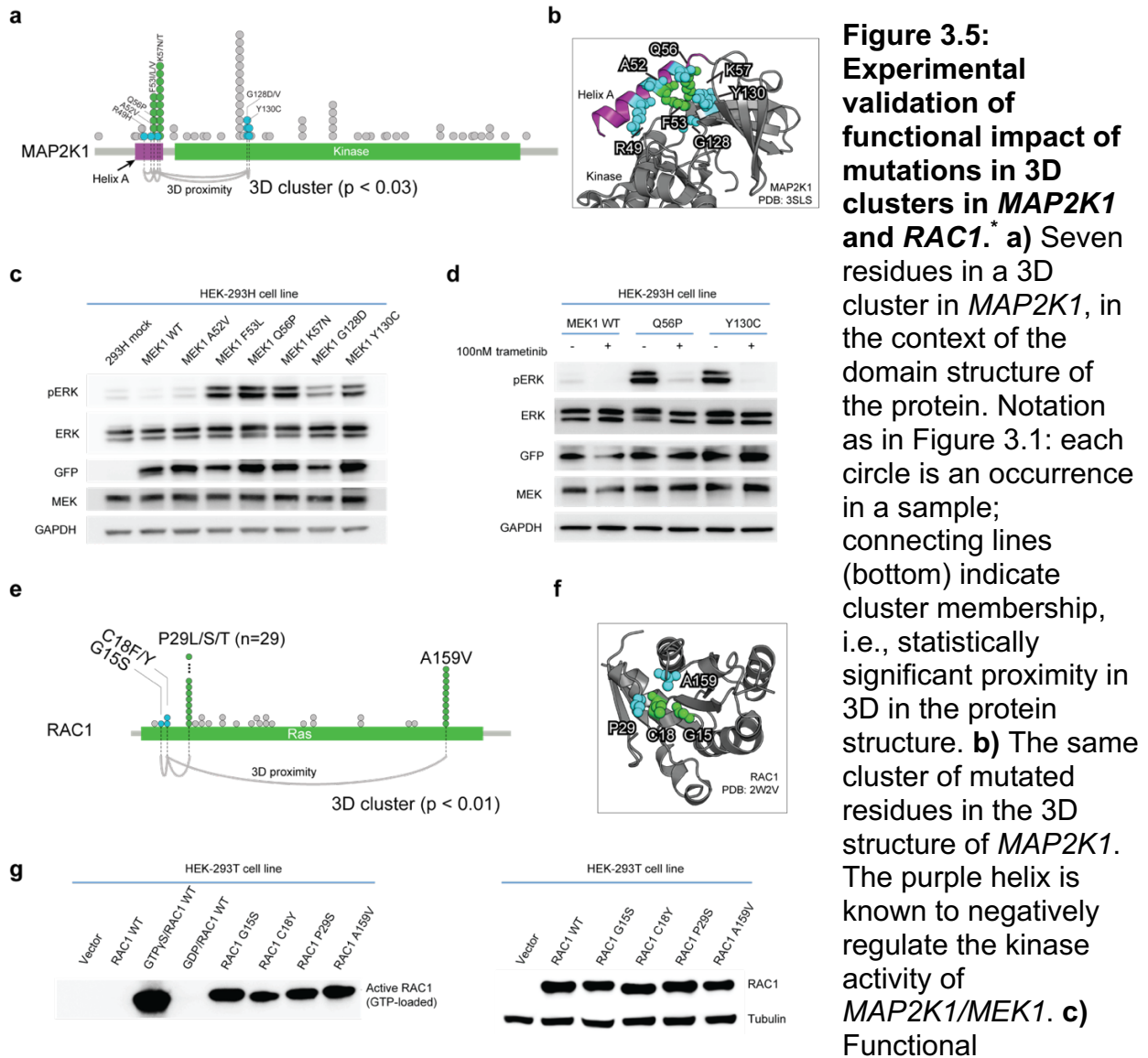


Figure 3.5: Experimental validation of functional impact of mutations in 3D clusters in *MAP2K1* and *RAC1*. **a)** Seven residues in a 3D cluster in *MAP2K1*, in the context of the domain structure of the protein. Notation as in Figure 3.1: each circle is an occurrence in a sample; connecting lines (bottom) indicate cluster membership, i.e., statistically significant proximity in 3D in the protein structure. **b)** The same cluster of mutated residues in the 3D structure of *MAP2K1*. The purple helix is known to negatively regulate the kinase activity of *MAP2K1/MEK1*. **c)** Functional

characterization of *MAP2K1/MEK1* mutants in HEK293H cells. Expression of G128D and Y130C (as well as the previously characterized F53L, Q56P, and K57N) mutants each resulted in increased expression of phosphorylated ERK compared to wild type *MAP2K1* - but not the cluster member A52V. **d)** ERK phosphorylation was inhibited by trametinib in cells expressing the Q56P or Y130C *MAP2K1* mutations in HEK293H cells. **e)** The four residues (two single-residue hotspots: P29 and A159, and two rarely mutated residues: G15 and C18) in the identified 3D cluster in *RAC1* in the linear domain structure of the protein. **f)** The same cluster in the 3D structure of *RAC1*. **g)** Western blot analysis of *RAC1* activation (GTP-bound *RAC1* levels) by PAK1 pull down (left) and of total *RAC1* levels (right) in HEK293T cells. The *RAC1* 3D cluster mutations G15S and C18Y, as well as the previously characterized P29S and A159V, were associated with significant *RAC1* activation, as compared to wild-type *RAC1*.

* In collaboration with Brooke Sylvester, Hannah Johnson, Aphrothiti Hanrahan, and Sizhi Paul Gao

RAC1 is a Rho-family small GTPase that has been recently implicated to confer resistance to RAF inhibition *in vitro* and may underlie early resistance in patients⁷¹. Recently, two oncogenic hotspots in *RAC1* were identified, P29 and A159, both of which activate *RAC1 in vitro*¹⁶. We identified a statistically significant 3D cluster of four residues ($p=0.009$) in *RAC1*, which, in addition to P29 and A159, includes novel rare mutations at amino acids G15 and C18 (mutated in 1 and 2 samples, respectively) (**Figure 3.5e-f**). To confirm that these mutations activate *RAC1*, we utilized a PAK1-pulldown assay to quantify activated *RAC1* expression in cells expressing mutant and wild-type *RAC1* protein. We found that, compared to wild-type *RAC1*, both the G15S and C18Y *RAC1* mutants resulted in elevated active *RAC1* expression (**Figure 3.5g**). These results expand the number of experimentally validated activating alleles in *RAC1*, suggesting that *RAC1* G15 and C18 mutations in this 3D cluster may possess similar biological consequences as the previously characterized *RAC1* hotspot mutations.

3.3.4 Comparison to other 3D hotspot detection algorithms

Previous methods have also detected many mutations that cluster in 3D structures. For example, Mutation3D identified 399 mutated residues in 75 genes¹³¹, HotMAPS identified 398 mutated residues in 91 genes¹³², Hotspot3D identified 14,929 mutated residues in 2,466 genes¹³³, whereas our method identified 3,404 mutated residues in 503 genes (**Table 3.6, Figure 3.6**). Only 15 residues were identified by all four methods, all of which were also previously identified as hotspots¹⁶. 2,908 of the 3,404 mutated residues detected by our method were not identified by any of the other three methods,

including *MAP2K1* Q56 and K57, which we experimentally validated. Comparison to a recent in vivo screening study of rare driver mutations by Kim et al.¹³⁴ also confirmed that the four methods showed different coverage and power to detect rare driver mutations and therefore provide complementary data (**Table 3.7**). The method described here was able to detect the *KRAS* D33E and *SPOP* K134N mutations that were validated by Kim et al.¹³⁴, but the other three methods did not detect these mutations as statistically significant.

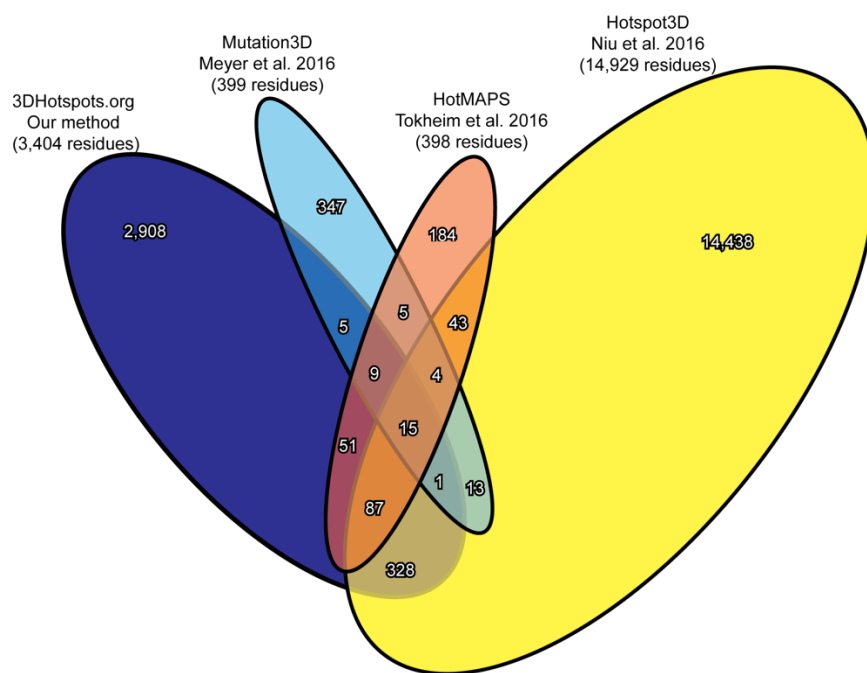


Figure 3.6: Comparison of multiple 3D hotspots approaches. Comparison of mutated residues identified in 3D structure clusters by our method and those by three alternative methods (Mutation3D, HotMAPS, and Hotspot3D)

3.4 Discussion

Tremendous effort has been invested in the discovery of therapeutic agents to suppress oncogenic signaling. These efforts have resulted in several FDA-approved

agents that target a variety of genes and pathways in several different cancer types. For instance, vemurafenib, a selective inhibitor of V600E/K mutant *BRAF* was first approved in metastatic melanoma, a cancer in which approximately 50% of tumors harbor a *BRAF* V600E/K mutation¹⁴⁹. Vemurafenib has since shown activity in a wide spectrum of malignancies that share this actionable mutation⁶³, suggesting that molecular biomarkers can be predictive of drug response across cancer types. However, effective development and use of targeted therapies necessitates identification of “driver” mutations among the far more prevalent passenger mutations in patient genomes. Many of these mutations can be identified by their recurrence in a single position, but others are less common or private to a particular tumor. One property they often share with hotspots and previously functionally characterized mutations is three-dimensional proximity, i.e. rare mutations can be physically close to each other or to a known and common mutation in the same protein, raising the possibility that these mutations are also driver events. To prioritize rare driver mutations for functional or clinical validation, we developed a novel method that identifies significantly mutated regions in 3D protein structures. We applied this method to more than 11,000 tumors analyzed by whole exome or genome sequencing.

Our analysis identified several thousand, mostly novel, candidate functional cancer mutations. While some mutations in the 3D clusters were in single-residue hotspots, which by definition are frequently mutated in cancer, the majority were rare mutations. Functional annotation is often not available or sparse for these rare mutations. On the one hand, rarely mutated residues coupled to a single-residue hotspot often occur in

many well-studied oncogenes (such as *KRAS*, *BRAF*, *EGFR*, *PIK3CA*, *MTOR*, among many others) and in several frequently mutated tumor suppressor genes (such as *TP53* and *PTEN*). It is therefore plausible that the functional impact of such mutations is similar to those in the single-residue hotspot and hence transfer of functional annotation makes sense. On the other hand, the functional annotation of rarely mutated residues, which are not coupled in a 3D cluster to a well-annotated single-residue hotspot, but instead clustered only with other rarely mutated residues, is much less certain.

Fortunately, the placement of the clusters of mutated residues in known 3D structures affords the opportunity for informative mechanistic hypotheses facilitating the design of focused functional studies. For example, we identified a cluster of mutations that likely disrupt critical calcium-binding sites in *CDH1*, a tumor suppressor that mediates cell adhesion. Another example is a cluster of mutations in *KEAP1* that potentially disrupt binding sites with *NRF2*, a key regulator of the cellular oxidative response.

By experimentally validating candidate functional mutations in 3D clusters in *MAP2K1* and *RAC1*, we show that our method readily identifies previously occult rare activating mutations that could not be revealed by positional frequency analyses alone and that a subset of such mutations are potentially biomarkers of sensitivity to targeted inhibitors in individual cancer patients. We showed, for example, that the rare *MAP2K1* G128D and Y130C mutations induce MAPK pathway activation and that such mutations retain sensitivity to MEK inhibitor treatment *in vitro*. While some mutations identified by our analysis were not activating *in vitro*, such as *MAP2K1* mutations of A52, by analyzing mutations in the context of protein structures, we can form hypotheses about

the biochemical reasons for such results: in this case, A52 does not interact strongly with the kinase domain in the wild type 3D structure (**Figure 3.5b**). This example illustrates the potential functional insights resulting from detailed analysis of individual cancer mutations in the context of 3D structures.

A proportion of rare mutations are not only biologically interesting (since they potentially promote tumor initiation or progression), but also clinically important with the advent of genomic-based clinical trial designs (such as the NCI-Molecular Analysis for Therapy Choice (NCI-MATCH) Trial). Forty-five percent of the 11K tumor samples in our dataset lacked a single-residue hotspot driver mutation, and identifying the genetic drivers of these patients is a critical step for choice of therapy, design of clinical trials or drug development. Here, we achieved a partial advance in this direction by identifying potential driver mutations in 17% of the samples without single-residue hotspot driver mutations (8% of all samples). Some of the identified mutations, e.g., those in *MTOR*, *EGFR* and *MAP2K1*, could have immediate translational importance. For example, clinical trials enrolling patients with MAPK pathway mutations, e.g. the NCT01781429 trial, could expand their eligibility criteria beyond single-residue hotspot mutations in the *MAPK* pathway and enroll patients with the *MAP2K1* 3D cluster mutations identified here.

While our approach can identify novel and potentially interesting mutations in cancer genes and in genes previously unknown to be involved in cancer, the method is still limited by the lack of complete protein structure data for many genes. For the 18,100 genes with mutations in our dataset, we were able to align 7,390 of them to one

or more protein structures. However, for many genes, the structures included only individual protein domains, limiting the scope of our analysis. There were only 1,307 genes with a protein structure that covered more than 90% of the protein length, and only 3,183 genes with more than 50% coverage. This limits the ability of our algorithm to detect 3D clusters that were not close in sequence, for example those involved in domain-domain interactions. Fortunately, as protein structure characterization technologies such as cryo-electron microscopy (cryo-EM) advance, more protein structures, and more complete protein structures, are being generated. We can also make use of the remarkable progress in 3D protein structure prediction using evolutionary couplings for proteins that are members of protein families with many known homologous sequences (<http://evfold.org>)¹⁵⁰. We thus plan to periodically include new protein structures in our analysis pipeline, which along with the inclusion of additional sequencing data will allow for the nomination of additional novel 3D clusters. Given the current coverage of human proteins by 3D structural knowledge, one can expect a steady increase in the number of candidate functional mutations identified by methods of this type as more accurate structures of most human proteins become available.

3.4 Discussion

Like any statistical method, the power of our approach is also limited by the number of available tumor samples. For example, a 3D cluster in *AKT1* (R15, E17, W22, and D323), which was not statistically significant ($p=0.11$), included the most frequent

hotspot mutation E17K, which has been evaluated as an indicator of response to AKT-targeted inhibitors in clinical trials. In addition, *in vitro* studies indicate that *AKT1* D323 mutations lead to constitutive activation of AKT¹⁵¹. Fortunately, as more cancer genomic data are generated, additional significant 3D clusters will likely emerge.

We showed that the mutational 3D clusters identified by three alternative methods (Mutation3D¹³¹, HotMAPS¹³², and Hotspot3D¹³³) and our method were largely complementary (**Figure 3.6**). While different mutational and structural datasets used by these four tools may have led to some of the differences observed, methodological differences also likely played a role. For example, HotMAPS predicted as functional the *IDH1* R132 and *IDH2* R172 mutations (both are single-residue hotspots) without linking to other residues in 3D structures, while the other three methods predicted mutations only if a 3D cluster was formed with other mutated residues. Hotspot3D also predicted the *IDH1* R132 and *IDH2* R172 mutations as functional because it utilized long distance interactions in 3D structures, e.g., R172 was detected in a cluster with R140 with a distance of 10Å. Another reason for differences in results may reside in the sensitivity and specificity levels that different methods adopted. Mutation3D and HotMAPS achieved a high specificity and low sensitivity and therefore only predicted about 400 mutated residues in less than 100 genes, most of which were single-residue hotspots. Conversely, Hotspot3D nominated close to 15,000 mutated residues in almost 2,500 genes, which potentially include many false positives. An analysis of *in vivo* screen results by Kim et al.¹⁵² supported this observation: All mutations identified by Mutation3D and most mutations identified by HotMAPS that were included in the screen

were single-residue hotspots, whereas our method and Hotspot3D were able to identify significantly more rare mutations. Finally, the Hotspot3D prediction included many false positives (false detection rate 32% compared to 12% of our method when applied to the Kim et al. data) (**Table 3.7**).

CHAPTER 4*

BIOMARKER DISCOVERY IN HOTSPOT MUTATIONS

4.1 Background

The adoption of prospective clinical sequencing of human tumors has led to the identification of an increasing number of somatic mutations of unknown significance. While a small number of mutations are now used to guide treatment selection, the vast majority of observed mutations lack biological or clinical validation, limiting our ability to use genomic profiling to guide therapy. Even among the small subset of oncogenic mutations targeted by standard-of-care therapies, significant gaps remain in our understanding of what characteristics condition their response⁶³. Other mutations may serve as the basis of extraordinary responses to cancer therapies^{16,100}, expand the number to existing biomarkers that predict therapeutic sensitivity, or elucidate novel biological and molecular phenotypes, but their prospective identification is challenging. Together, this underscores the need to identify and prioritize occult driver mutations in cancer for further biological and clinical study. Unraveling the relationships between mutant alleles, cell types, co-mutational patterns and other factors that determine mutant allele function, and consequently clinical actionability, is essential to expand the treatment options for molecularly defined populations of cancer patients.

* Chang MT, Bhattarai TS, Schram AM, Bielski CM, Donoghue M, Jonsson P, Chakravarty D, Phillips S, Kandoth C, Penson A, Gorelick A, Shamu T, Patel S, Harris C, Gao J, Sumer SO, Gao Y, Yao Z, Kundra R, Razavi P, Reales DN, Socci ND, Jayakumaran G, Zehir A, Chandralapaty S, Ladanyi M, Schultz N, Baselga J, Hyman DM, Solit DB, Berger MF, Rosen N, Taylor BS. Accelerating discovery of functional mutant alleles in cancer. *In preparation*

4.2 Methods

4.2.1 Mutational data and pre-processing

Retrospective mutational data were obtained from three publicly available sources: 1) The Cancer Genome Atlas (TCGA), 2) International Cancer Genome Consortium (ICGC), and 3) independent published sequencing projects. The subset of this cohort that was prospectively sequenced consists of 10,945 samples from 10,336 unique advanced cancer patients and 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 manuscript (Zehir A, et al. submitted). Briefly, matched tumor and normal specimens were sequenced (to 500-1000-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. These sequencing data are analyzed as previously described⁹¹ to detect somatic mutations, small insertions and deletions (indels), DNA copy number alterations (CNAs) and select translocations using DNA from both frozen and formalin fixed-paraffin embedded tissue. 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 is deposited, along with relevant clinical data, in a HIPAA-compliant manner, in the cBioPortal for Cancer Genomics^{30,31}. All somatic nonsynonymous mutations reported were manually reviewed in primary sequencing data as described in Zehir A, et al. submitted 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 re-annotated using vcf2maf ver. 1.6.10 (<https://github.com/mskcc/vcf2maf>). Variants identified by the Exome Aggregation Consortium (ExAC)¹⁵³ as having a minor allele frequency greater than 0.0004 in any subpopulation were excluded as presumed germline unless they were annotated by ClinVar20 as either pathogenic, a risk factor, or protective.

4.2.2 Determining pan-cancer and organ-type specific significance

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 (See Section 2.2.2). Briefly, statistical significance of every codon was assessed with a truncated binomial probability model in which the expected probability incorporates underlying features of mutation rates including gene length, gene- and position-specific mutability, and overall mutational burden of the gene. 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. The mutability of each codon, expected mutability of each gene, and the final binomial probability was calculated as before. 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 were performed using the method Benjamini and Yekutieli method. Mutational hotspots corresponding to a q-value < 0.1 were considered statistically significant (False Discovery Rate < 10%).

4.2.3 Determining in-frame insertions/deletions significance

We assessed the statistical significance of in-frame small insertions and/or deletions (indels) in a manner similar to the identification of single-codon hotspots using the truncated binomial probability model. From 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. 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.

4.2.4 Simulating hotspot identification 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 for between 1 to 15 clusters. To estimate the necessary number of samples needed to sequence to identify an additional hotspot, from the downsampled results from each gene, we fit a regression using a Conway-Maxwell Poisson distribution (R package COMPoissonReg v0.3.5).

4.2.5 Annotation of biological / clinical significance

All mutations were annotated for their potential prognostic and therapeutic significance utilizing OncoKB, a curated knowledgebase of the oncogenic effects and treatment implications of mutations at the individual allele resolution (<http://www.oncokb.org/>). 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: 1) genomic alterations that are FDA-approved biomarkers in patients of the indicated cancer type; 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 (NCCN); 2B) alterations that are FDA-approved biomarkers in another cancer indication, but not in patients with the affected cancer type; 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 4) alterations for which compelling preclinical data associates the biomarker with drug response. Only levels 1, 2A, and 3A were utilized for the analyses and results described here.

4.2.6 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 prostate¹⁵⁴), 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.

4.2.7 Co-mutational analyses

To assess the statistical significance of observed co-mutational frequency, we first construct a 2 -by- j binary matrix M where each entry m_{ij} refers to the status of the gene i in the sample j and whose value is 1 if sample j has 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 (*TSC1/2*, *PTEN*, and *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 to 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.

4.2.8 AKT1 duplication indel validation

293-FT cells were obtained from ATCC and maintained on DMEM supplemented with 10% FBS and 2mM glutamine. MCF10a cells were obtained from the the Solit laboratory and maintained in DMEM/F-12 base medium containing 5% horse serum and other supplements (20ng/ml EGF, 0.5mg/ml hydrocortisone, 100ng/ml cholera toxin and 10mg/ml insulin) (complete growth medium). 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-wildtype (WT) and AKT1-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 sub-cloned 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 was filtered through 0.45µm filters and used to infect MCF10a cells. Cells stably expressing the lentiviral constructs were selected with puromycin (2.5µg/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 30µg protein was loaded onto SDS-PAGE gels, transferred to nitrocellulose membranes, and immunoblotted for p-Akt and other downstream molecular targets of Akt pathway activation. Antibodies for p-Akt (T308) (D25E6), p-Akt (S473), p-S6RP (S240/244), p-GSK-3β (S9), p-4E-BP1 (T37/46), p-ERK1/2 (T202/Y204), and β-actin (8H10D10) were obtained from Cell Signaling Technology. V5 probe (E10) antibody was purchased from Santa Cruz Biotechnology.

For cell growth assays, 50,000 cells per well were seeded in triplicate for each of the MCF10a stable cells expressing WT or mutant AKT1 and maintained in assay medium for the length of the study period. Cells were washed with PBS, trypsinized, resuspended in 1ml of medium, and counted in ViCell-XR to obtain viable cell numbers on days 1, 3, 5 and 7. Drug treatment and cell viability assays were performed as follows. AZD5363 was generously provided by AstraZeneca, dissolved in DMSO to yield a 10mM stock, and diluted in assay medium to achieve the desired concentrations. MCF10A stable lines expressing WT or mutant AKT1 seeded in 96-well plates were treated with a range of drug concentrations, and cell viability was assessed 72 hours post treatment using the Cell Titer Glo assay (Promega).

4.3 Results

We analyzed each of the 32 organ types independently as well as the full cohort (pan-cancer). To ensure sensitivity for detecting hotspots across organ types of often different mutational burdens and processes, we computed organ type-specific gene and position-specific background mutation rates and incorporated gene-specific sample sizes to enable the combination of exome-scale data with prospective clinical sequencing data (Zehir A, et al. submitted). In addition to the identification of substitution hotspots, we also developed a computational approach to identify candidate hotspots of oncogenic small in-frame insertions or deletions (indels, see Methods in Section 4.2.3), which are more challenging than substitutions to identify due to the variability of mutant allele length and position from tumor to tumor. Statistically significant single-codon or indel hotspots were those of q -value < 0.1 (false discovery rate of 10%) either within a given organ type or pan-cancer.

4.3.1 Identification of lineage-specific hotspots

In total, we identified 1,165 mutational hotspots (1,110 single-codon and 55 indel) in 247 genes (median of 2 hotspots per gene, range 1-120) (**Table 4.1**). This analysis recovered nearly all previously identified hotspots¹⁶ (97%) and identified 840 more, reflecting an increased power of detection. The frequency distribution of hotspot-mutant genes across cancer had a long right tail¹⁶, which was independent of the count of unique hotspots in the gene and was different between single-codon and indel hotspots

(Figure 4.1a). While the majority of hotspots (n=596, 51% of total) were significant both pan-cancer and within individual organ types (**Figure 4.1b**), 20 and 29% of hotspots were either significant only within an individual organ type or only in the pan-cancer analysis of the full cohort respectively (**Figure 4.1c**). Some of these mutant alleles arose in genes that did not harbor a significant hotspot in any other facet of this analysis, were new in genes with recently characterized hotspots (such as *CYSLTR2*)¹⁵⁵, or arose in genes with more significant mutant alleles and reflected the impact of cohort size on the sensitivity for rare allele discovery in even well-characterized cancer genes (such as *PIK3CA*, *MTOR*, *ERBB2*, *MAP2K1*) (**Figure 4.1b**).

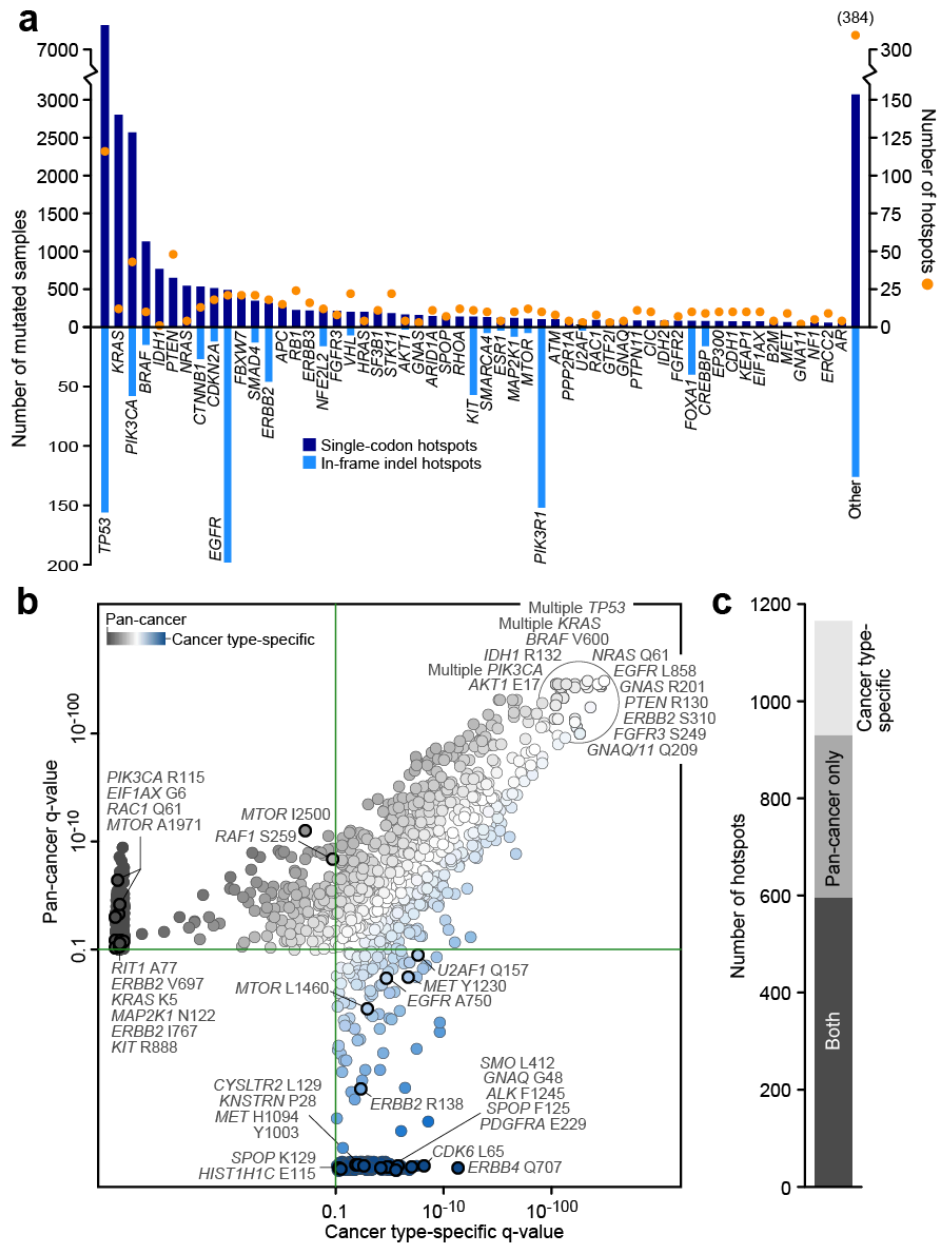


Figure 4.1: The long tail of mutational hotspots in cancer. a) The frequency distribution of genes containing one or more single-codon hotspots (top, dark blue; count of hotspots in orange, right y-axis) and in-frame indel hotspots (light blue). b) Shown is the statistical significance of mutational 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 and include mutations significant in both analyses (upper 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.

4.3.2 Hotspots enriched in metastatic disease

Forty-two percent of the patients in this cohort were prospectively analyzed and had advanced previously treated disease, a clinical profile distinct from the primary untreated data that predominates in the literature. The inclusion of such patients allowed for the identification of hotspots that were present almost exclusively in the metastases of treatment-refractory patients. Eleven hotspots were enriched in metastatic disease compared to the primary tumors of given cancer types (see Methods in Section 4.2.6), nine of which were treatment-associated arising in specimens after treatment with either anti-androgen, anti-estrogen, or tyrosine kinase inhibitor therapies (**Figure 4.2**). Some of the hotspots that mediate drug resistance also arose in treatment-naïve tumors of other cancer types (such as *KIT* D820), suggesting that treatment-associated mutations in one cancer type can provide a selective growth advantage and arise de novo even in the absence of the selective pressure of therapy as a primary driver in other cancer types. Other hotspots may reflect new mechanisms of resistance to systemic therapies, such as *TP53* N239, which was the only *TP53* hotspot that arose preferentially in metastatic breast cancers (q-value = 0.03). As previous work suggests *TP53* N239 mutations confer paclitaxel resistance in vitro¹⁵⁶, it was notable that all five patients with metastatic *TP53* N239-mutant breast cancer had previously received and/or rapidly progressed on taxane-based therapy. Together, these analyses highlight a far broader range of hotspots than previously recognized and for which validation^{134,157} may accelerate clinical translation.

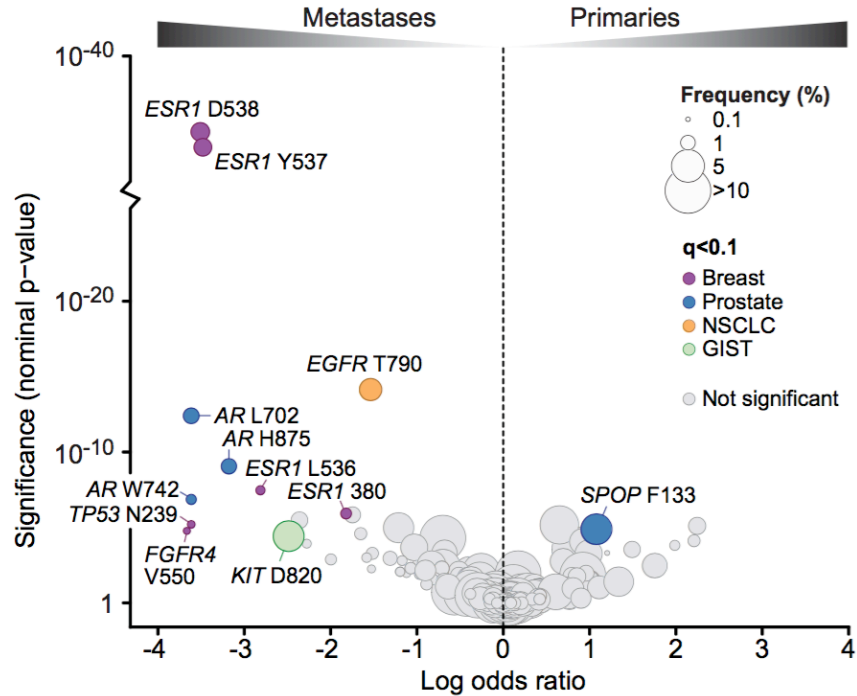


Figure 4.2: Metastatic enrichment of hotspots identified. A subset of mutational hotspots were enriched in metastatic disease compared to primary cancers of a given cancer type. Majority of metastatic-specific hotspots were those associated with targeted and/or hormonal treatment. Y-axis is the nominal p-value each hotspot by the cancer type-specific analysis. Hotspots with q-value < 0.1 were considered statistically significant colored by the affected cancer type.

4.3.3 Landscape of recurrent indels in cancer

While substitution hotspots are the most abundant mutation in cancer genomes, several recurrent constitutively activating in-frame indels in oncogenes are biomarkers for the use of molecular targeted therapies, including indels in exon 19 of *EGFR* in lung adenocarcinomas and in exon 11 of *KIT* in gastrointestinal stromal tumors. Nevertheless, hotspots of activating in-frame indels have never been defined in an unbiased manner before, some of which may similarly sensitize patients to therapy. We, therefore, extended our methodology to identify clusters that represent hotspots of in-

frame indels (see Methods in Section 4.2.6). In total, we identified 55 statistically significant indel hotspots in 36 genes (**Table 4.1**). There were 20-fold fewer indel hotspots identified than single-codon hotspots and the majority of genes harboring at least one indel hotspot also harbored a single-codon hotspot (80%, 31 of 39 genes). In these genes, most indel hotspots span or are physically adjacent to single-codon hotspots (**Figure 4.3**). Three indel hotspots were distal ($>15\text{\AA}$) from single-codon hotspots in the protein structure of the corresponding gene including the well-characterized FLT3 internal tandem duplication (ITD)¹⁵⁸, ESR1 V422del, and a cluster of indels spanning I99 to I107 in MAP2K1 (**Figure 4.4a**). Notably, ESR1 V422del lies approximately 20 angstroms from the ESR1 ligand binding domain (LBD) in which hotspots E380, L536, Y537, and D538 are known to confer resistance to estrogen deprivation therapies⁵⁵ (**Figure 4.2**). Consistent with LBD hotspots, two of the three V422del mutations were clonal within the sampled metastatic site in otherwise *ESR1*-wildtype estrogen receptor-positive breast cancers sequenced after failure of anti-estrogen therapy (**Figure 4.5**). These findings suggest that V422del, unlike the LBD hotspots thought to stabilize ligand-independent confirmation of *ESR1*, may represent a novel mechanism by which mutant *ESR1* confers resistance to anti-estrogen therapy.

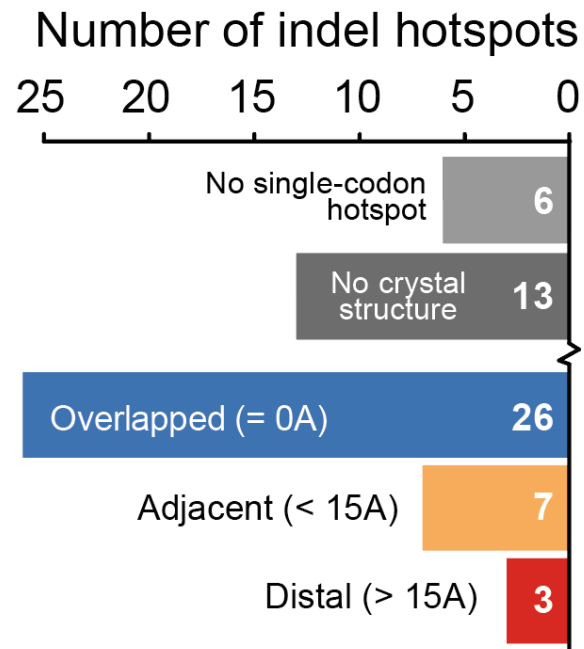


Figure 4.3: Distance of indel hotspots compared to single-codon hotspots. Majority of the 55 significant indel hotspots overlapped directly or were adjacent to other single-codon hotspots. Three indel hotspots were found distal (> 15Å) away from any single-codon hotspot identified in the gene.

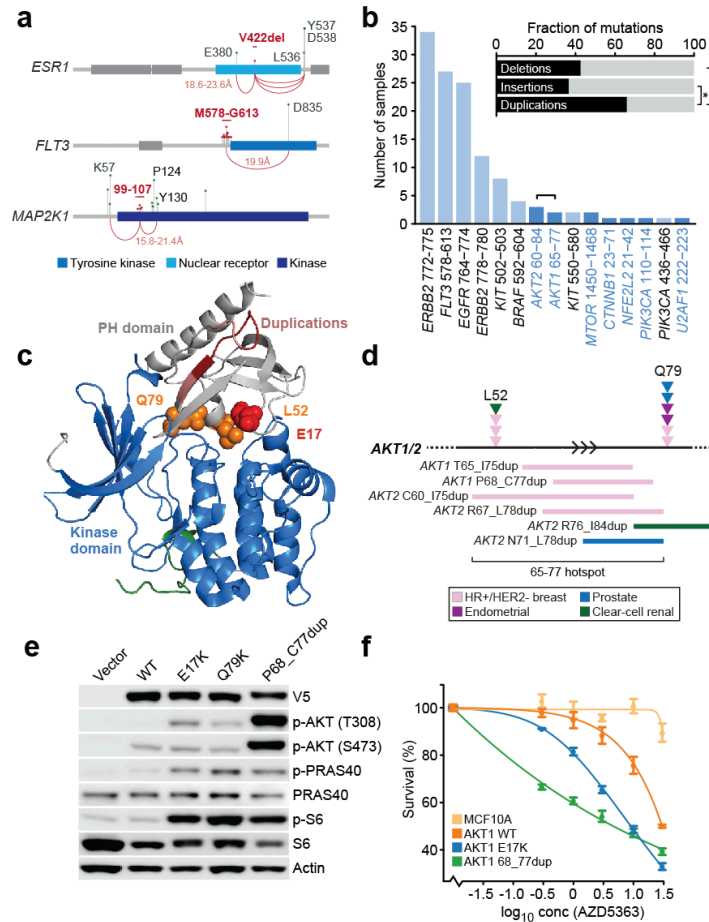


Figure 4.4: Oncogenic indel hotspots* **a)** For the three indel hotspots (red) distal to the position of known single-codon hotspots in the same genes (gray/green), their position and distance is indicated (arcing lines, pink). **b)** The frequency distribution of both previously known (light blue) and novel (dark blue) duplication indels in the study cohort. Inset, duplications were identified in oncogenes significantly more frequently than either deletions or insertions. (asterisk, p -value < 0.01). **c)** The hotspot of duplication indels in *AKT1* are shown in three dimensions (dark red) and lie adjacent to the L52 and Q79 single-codon hotspots in *AKT1* (orange). The E17K hotspot is shown in red for reference. **d)** 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. **e)** MCF10A cells stably expressing the indicated *AKT1* mutations are shown and expression and/or phosphorylation levels were assayed by Western blot indicating the *AKT1* P68_C77dup induces elevated levels of phosphorylated Akt and S6 comparable to or exceeding that of known activating E17K or Q79K hotspots. **f)** 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.

* In collaboration with Tripti Shrestha

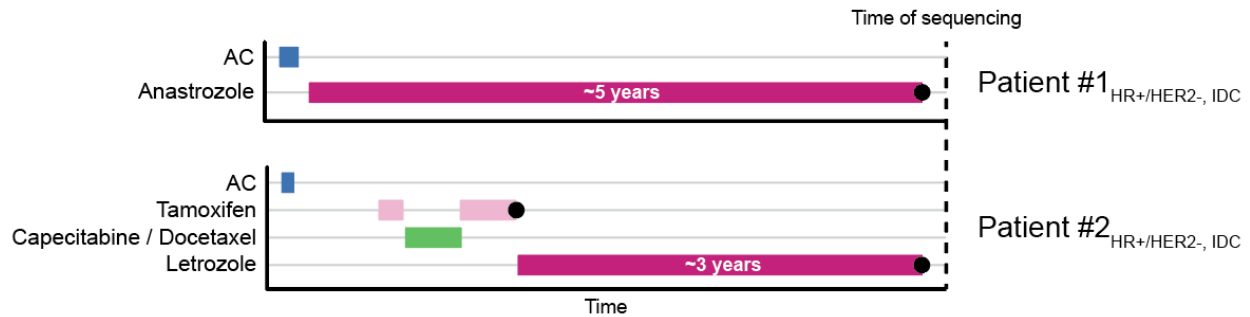


Figure 4.5: Treatment history of two ESR1 V422del mutant HR+/HER2- patients at. Two HR+/HER2- patients who progressed on aromatase inhibitor therapy after 5 and 3 years, respectively.

4.3.4 Validation of AKT1 duplications as sensitivity biomarkers

Overall, while deletions predominate among the indel hotspots (69%, **Figure 4.6**), we found that duplications were enriched in oncogenes (p-value < 0.01, **Figure 4.4b, inset**). The most recurrent of these among previously uncharacterized alleles was paralogous indels in the pleckstrin homology (PH) domain of *AKT1* and *AKT2* (clusters spanning T65-C77 and C60-I84 respectively, q-values = 0.09 and 0.00002) (**Figure 4.4c**). These indels are proximal to two known activating *AKT1* hotspots L52 and Q79 (q-values < 10^{-4}) (**Figure 4.4d**) and arise predominantly in estrogen receptor-positive HER2-negative breast cancers that lack other PI3K alterations. To determine whether indels at this hotspot were activating mutations that sensitized cells to Akt inhibition, we assessed the effects of *AKT1* P68_C77dup and compared it to two known activating hotspots mutations (E17K and Q79). *AKT1* P68_C77dup induced AKT phosphorylation (T308/S473) to levels higher than those achieved by the two known activating single-codon hotspots (**Figure 4.4e**), similar levels of elevated S6 and PRAS40 phosphorylation, and was the most sensitive of tested mutants to AKT inhibition with the selective ATP-competitive pan-AKT kinase inhibitor AZD5363 (**Figure 4.4f**). These

results suggest that novel methodologies to identify previously occult recurrent oncogenic in-frame indels, when coupled with functional validation, can expand the molecular eligibility of genotype-driven trials and ultimately validate as novel predictive biomarkers of inhibitor sensitivity.

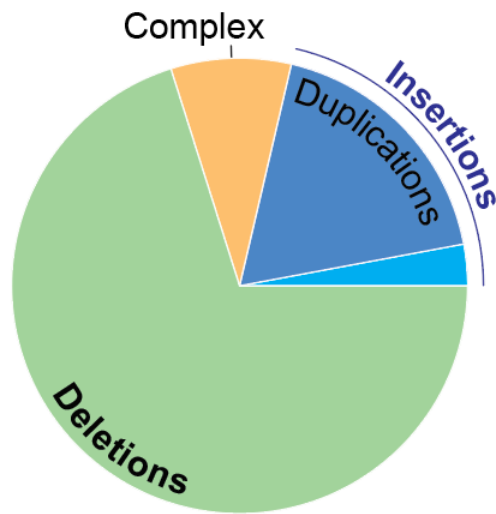


Figure 4.6: Distribution of types of indels identified as recurrent. Majority of indel hotspots were deletions, followed by duplications, complex insertions, and de-novo insertion.

4.3.5 Co-occurrence of multiple pathway hotspots

One limitation of these analyses is that each hotspot is considered individually, without consideration to the other mutations in affected tumors. We hypothesized that analyzing the patterns of co-mutation among hotspots could classify mutations that are independently oncogenic versus cooperative mutant alleles, those that are only functional in a specific mutational context. To formerly test this hypothesis, we assessed co-occurrence among individual hotspots and hotspot-containing genes in the same pathway that arose together in individual tumors more frequently than expected by chance, focusing on MAPK and PI3K pathways as the two most frequently

therapeutically targeted pathways in cancer patients. Overall, the pattern and frequency of co-mutation in these pathways varied widely (**Figure 4.7a** and **Figure 4.8**). The majority of PI3K pathway mutational co-occurrence exists between *PIK3CA* and *PTEN* in endometrial cancers³⁶. The two most significant associations in MAPK signaling involved hotspots in *ERBB3* and *MAP2K1* (q -values $< 10^{-6}$; **Figure 4.7b**). *ERBB3* was most often co-mutated with hotspots in *KRAS* and *ERBB2* in colorectal and breast cancers respectively. While further study of these relationships are necessary, the latter is notable as *ERBB3* preferentially dimerizes with *ERBB2* in vivo suggesting co-mutated heterodimers may further potentiate downstream activation.

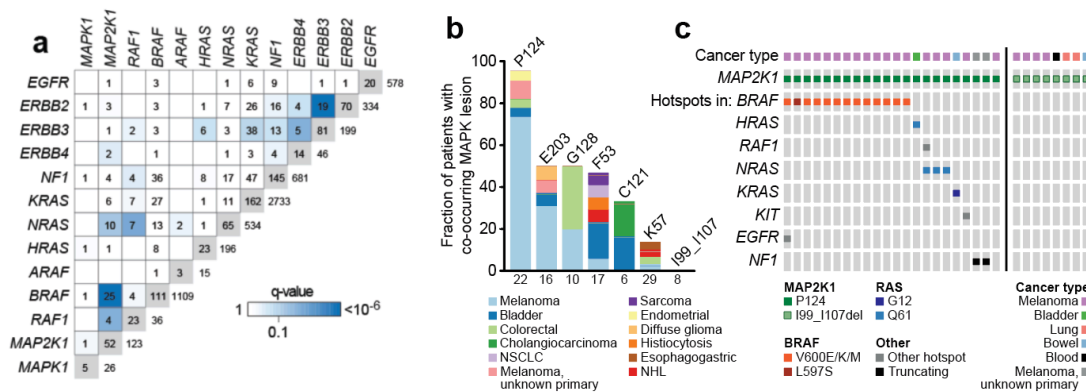


Figure 4.7: Co-mutational patterns among hotspots reveals function. a) The pattern of mutational co-occurrence among hotspots in genes essential to MAPK signaling reveals co-existing hotspot mutations in the same tumors, the most significant of which was *BRAF* and *MAP2K1*. The number in each cell is the count of co-mutated specimens; the numbers at the end of each row is the count of mutated cases for the indicated gene; the numbers in gray are the total co-mutated cases (row and column); and cell shading indicates increasing statistical significance of the association (as indicated in legend). **b)** The rate of *BRAF* and *MAP2K1* co-mutation varied by *MAP2K1* hotspot, with P124 mutations always associated with upstream pathway activation and predominantly in melanomas, while others (E203, G128, F53, C121, and K57) only partially co-mutated while the the *MAP2K1* I99_I107 indel hotspot never arose in tumors with another MAPK driver mutation. **c)** 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.

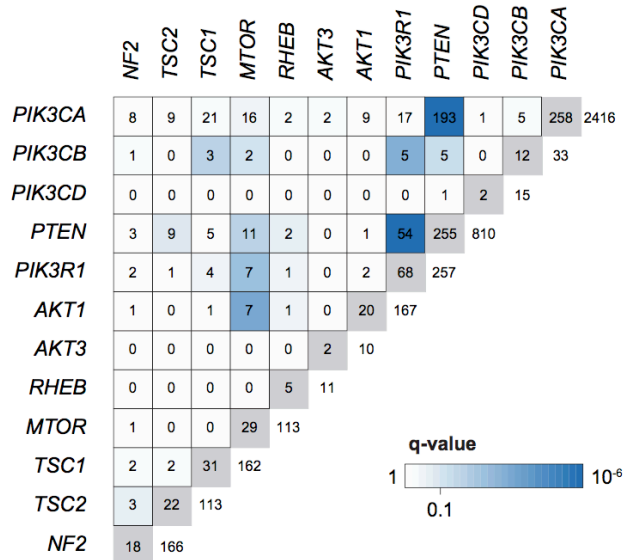


Figure 4.8: Co-mutational analysis of PI3K pathway hotspots. Pattern of co-mutational pattern among hotspots in genes essential to PI3K pathway signaling reveals co-existence of mutations in the same tumors were predominated by PIK3CA-PTEN mutant tumors

In the case of *MAP2K1* (*MEK1*), the pattern and frequency of co-mutation varied in an allele-specific and cancer type-specific manner. *MAP2K1* P124-mutant tumors are nearly always co-mutated with another upstream activating mutation of MAPK signaling (95%), most often with *BRAF* V600E (55%) and largely, but not exclusively, in cutaneous melanoma patients (**Figure 4.7b-c**). Conversely, other *MAP2K1* hotspots like the in-frame indels clustered in the region from I99 to I108 ($q\text{-value} = 3.3 \times 10^{-12}$) arise mutually exclusively with other MAPK lesions in affected tumors independent of cancer type (**Figure 4.7c**). This pattern of co-mutation does not reflect acquired resistance to MAPK pathway inhibitors, as only one such tumor was sequenced after RAF or MEK inhibitor failure. We hypothesized instead that this allele-specific difference in the *MAP2K1* co-mutational pattern suggests that the second mutation in MAPK was required to condition the function of *MAP2K1* P124 but not the *MAP2K1* indels. While

this analysis of co-incident mutations in multiple pathway effectors suggest these lesions can condition distinctive signaling phenotypes, deeper mechanistic investigation is required to validate and fully elucidate the mechanism in which it is achieved.

4.3.6 Rate of hotspot identification by gene

Given the power for rare allele discovery in this cohort (82% of all hotspots were identified in 1 in 1000 or fewer patients), we sought to define the rate with which new hotspots were identified within and across genes as a function of increasing cohort size. We thus performed repeated random downsampling of increasing subsets of the cohort from which we inferred the rate of hotspot identification per gene. Principal component analysis of gene-specific rates revealed four distinct classes of genes that accrue their recurrent mutations, independent of their overall mutational burden, in different patterns and with considerable variability from gene to gene (**Figure 4.9** and **Figure 4.10**). One cluster is defined by canonical oncogenes (*IDH1*, *K/N/HRAS*, *GNAQ*, *MYD88*) whose hotspots can be identified from few samples but, as genes, they are approaching saturation where additional sequencing is not expected to yield many additional currently unrecognized hotspots. 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 additional rare alleles will continue to be discovered including in therapeutically actionable genes in this cluster (such as *BRAF*, *PIK3CA*, *ESR1*, *AKT1*, and *ERBB2*). The third cluster of genes are 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.

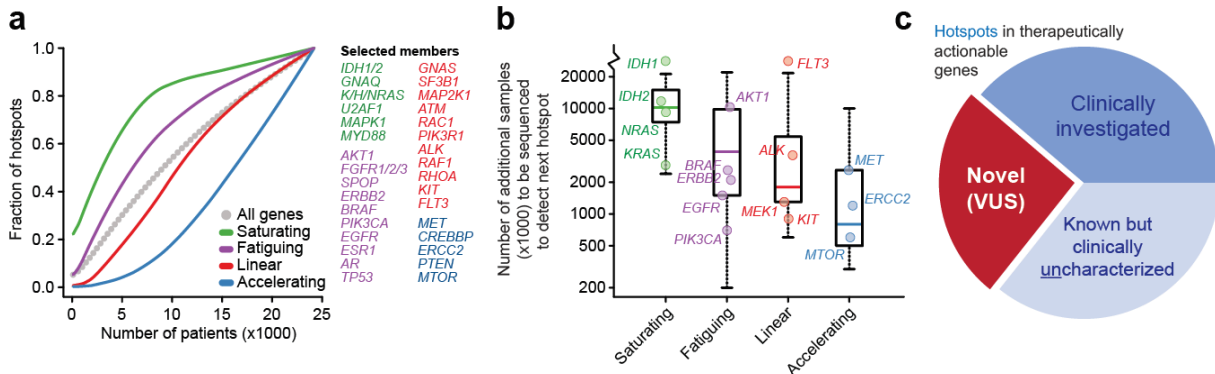


Figure 4.9: Saturation analysis and the discovery of actionability of mutational hotspots. a) Downsampling and clustering analysis revealed four distinct classes of genes with different rates of hotspot acquisition (green, purple, red, and blue) from the number of sequenced samples necessary to identify a given fraction of all hotspots in affected genes. Shown in gray are all genes. In red and purple are genes that are either saturating in their hotspot discovery (green) or were rapidly increasing and now fatiguing (purple). In red and blue are those genes in either their still linear and accelerating phases of hotspot discovery. **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 one of 18 clinically actionable cancer genes (see panel b for genes), the fraction of hotspots used to guide the use of standard-of-care or investigational therapies at present (see Methods in Section 4.2) versus those that were identified here but are clinically uncharacterized.

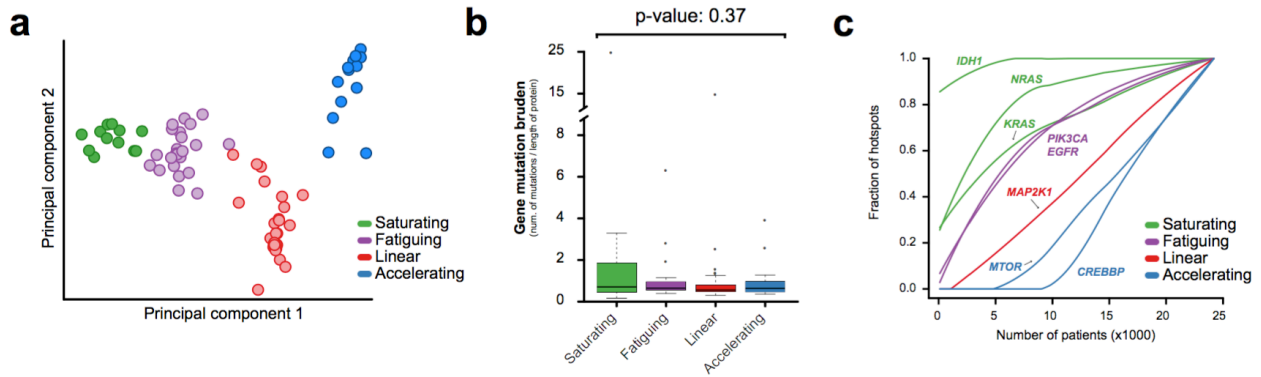


Figure 4.10: Clustering of the rate of hotspot identification. a) Principle component analysis divided genes harboring one or more hotspots into four classes (green, purple, red, and blue). **b)** Overall gene mutational burden was not different between the different clusters and likely not a confounder in the rate in which hotspots were identified between the four clusters. **c)** Variability in the rate of hotspots identified exist both between and within clusters. Highlighting select cluster members colored appropriately to their cluster membership.

As these patterns imply that hotspot discovery in many genes is far from complete, we sought to estimate the number of additional tumors of a cancer type composition similar to the cohort studied here would need to be sequenced to identify the next incremental hotspot in each gene and cluster. We estimate that tumor sequencing data on an additional ten thousand or more patients would be necessary to identify an additional hotspot in many genes in the saturating cluster (**Figure 4.9b**), whereas fewer than 1000 additional specimens would be necessary to identify additional hotspots in genes in the accelerating cluster. The results suggest that many additional hotspots are likely to be identified by pooling the sum of tumors prospectively sequenced across institutions that currently reside in silo repositories that prevent such analyses. Such consortia efforts could accelerate the identification of novel biomarkers for which drugs currently exist and expand treatment options for advanced stage patients.

This analysis indicates that we are far from saturating the identification of potentially actionable hotspots in even therapeutically targetable genes. Thus, affected patients are not being offered potentially beneficial matched therapies. To determine the prevalence of such occult actionability, we utilized a curated knowledgebase of the oncogenic effects and treatment implications of mutations (<http://oncokb.org/>) in 18 genes in which one or more mutations are standard of care (FDA-approved or part of established practice guidelines) or investigational biomarkers used to guide the use of approved or investigational therapies (see Methods in Section 4.2). Of the 196 hotspot mutations identified in these genes, only a minority have been investigated clinically (**Figure 4.10c**), though patterns vary in individual genes (**Figure 4.11**). Fifty hotspots (26%) were newly discovered, being neither annotated in OncoKB nor identified in a detailed literature review. Because these novel hotspots arise in genes for which targeted therapies are available, we sought to test the therapeutic hypothesis that these mutations may be sensitizing biomarkers by matching a subset of the affected, prospectively sequenced patients to molecularly targeted therapies. Two patients with a novel *ERBB2* V697 hotspot, one heavily pre-treated triple negative breast cancer and another cancer of unknown primary of the scalp, were treated with a selective HER tyrosine kinase inhibitor while another patient with a *PIK3CA* P104L-mutant uterine serous carcinoma received a mTORC1/2 catalytic inhibitor, all three of which responded to therapy. Taken together, these results indicate that, in some genes, recurrence alone can be used to select patients for targeted therapy, and that when affected patients are

identified prospectively, such novel mutations can expand selection biomarkers for molecularly targeted therapy.

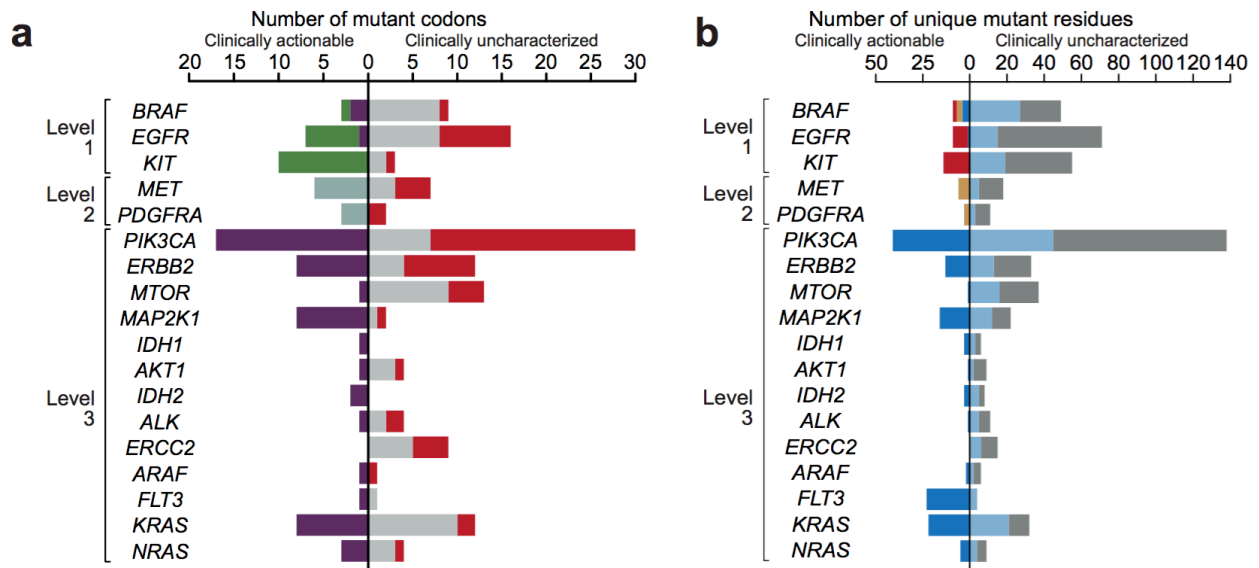


Figure 4.11: a) The number of hotspots per actionable cancer genes is shown and includes (left-facing) those with existing clinical evidence of activity to approved or investigational therapies (green, purple, and teal) or have been biologically studied but clinically characterized (gray) versus the number of hotspots identified here for the first time of the unknown significance (red, right-facing). **b)** Similar to (a), when expanding to distinct mutant residues, the number of clinically actionable and uncharacterized rises but pattern remains consistent.

4.4. Discussion

A central tenet of precision medicine in oncology is to provision therapy that targets the mutant proteins on which the growth and progression of individual human tumors depend. To expand the use of this approach, we identified here 1,165 hotspot mutations that extend the long tail of both common and rare recurrent mutations across a spectrum of 324 detailed tumor types comprised of both primary untreated and advanced post-treatment cancers. We show that the rate at which hotspots are being identified with increasing cohort size varies widely and that, in some genes, novel

potentially actionable mutational hotspots are still being identified at an accelerating rate. Because not all actionable hotspots have been identified yet, existing patients with mutant alleles identified here have largely not been offered matched molecular targeted therapies to which other patients have had a profound clinical benefit. The implications of these findings are especially relevant for patients with advanced, metastatic disease that are most in need of novel therapeutic approaches. Furthermore, our results imply that the development of broadly active targeted therapies that have clinical benefit may necessitate a better understanding of how co-mutations in multiple effectors of a given pathway condition distinct signaling biology and treatment sensitivity. Thus, 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 to expand the reach of precision therapeutic approaches. To accelerate the identification of novel clinically actionable hotspots, we have thus shared the prospective sequencing data described here with the Genomics Evidence Neoplasia Information Exchange (GENIE). Together, our findings provide a means to guide and prioritize experimental validation and clinical cross-validation to expand the treatment options for molecularly defined populations of cancer patients.

CHAPTER 5*

CONCLUSIONS AND FUTURE DIRECTIONS

5.1 Overview

Functionally significant rare mutations are challenging to identify, as these are the very mutations that escape detection by computational tools that use recurrence as a measure to credential driver mutations. A systematic and scalable approach is needed to prioritize and validate these mutations.

5.2 Lessons from phenotype-to-genotype

A conventional genotype-to-phenotype approach first identifies mutations and then seeks to associate them with biological or therapeutic phenotypes such as drug sensitivity or outcome. Such approaches have proven to be an enormous challenge. The mutational complexity and clinical heterogeneity of most cancers make specific phenotypic predictions difficult. Conversely, a phenotype-to-genotype approach can reveal long tail mutations that are the molecular underpinnings of specific complex biological and clinical phenotypes. One of the most effective approaches has been the comprehensive analysis of therapeutic outliers. Here, tumors from individual patients who experienced an unexplained exceptional clinical response to a specific anti-cancer therapy that far exceeds other similarly treated patients undergoes extensive genomic

*Chang MT, Taylor BS. On the impact of rare mutations in cancer. *Science*

sequencing to identify the molecular basis of their therapeutic sensitivity, information that can then be used prospectively in future patients.

Several studies have now shown that rare somatic mutations in proximal activators and inhibitors of mTOR signaling such as *TSC1*, *TSC2*, *MTOR*, and *NF2* can engender pronounced and durable response to mTORC1 inhibitors^{101,159,160}. Other outlier responses have been observed to sorafenib in a lung adenocarcinoma¹⁶¹, erlotinib in a HNSCC¹⁶², and selumetinib¹⁶³ in a low-grade serous ovarian cancer revealing rare sensitizing mutations in *ARAF* (S214C), *MAPK1* (E322K), and *MAP2K1* (in-frame deletion), respectively. Exceptional responses to systemic or combination therapies can also reveal rare mutations that lead to a new mechanistic understanding of a physiologic signaling pathway. For instance, *RAD50* hypomorphism sensitized a tumor to DNA damaging agents by creating a synthetic lethality of simultaneous genetic and pharmacological perturbation of the *ATM* and *ATR* axes of DNA damage response signaling¹⁰⁰.

Not all patients with similar mutations will have as dramatic a clinical response as the index extraordinary responders due in part to many of the context-specific effects discuss above. Assessing clinical responses to a given therapy based on the predicted sensitizing mutation rather than the organ of origin will certainly drive studies to uncover the biological mechanisms that condition variable responses to therapy. The study of exceptional responders has spurred a National Cancer Institute-sponsored initiative, and has accelerated the design of studies to expand and test these clinical hypotheses. A biochemical understanding of the molecular underpinnings, however rare, that

engenders deep and lasting response to anti-cancer therapies will serve as the basis for expanding rare outlier genotypes into a broader panel of biomarkers that bridge the gap between single-patient anecdotes and the identification of broadly applicable biomarkers of drug response.

5.3 Prioritization and validation

Prioritization. Given the number of rare alleles in potentially actionable cancer genes, all with possible context-dependent functional differences, the first step is determining which mutations are highest priority for in-depth characterization. New computational approaches that seek to identify functionally significant mutations look beyond the recurrence of individual mutant alleles, incorporating other complementary information from paralogous residues of proteins within the same family^{16,161}, conservation of affected protein domains¹⁶⁴, protein-protein interactions¹²⁶, or their position in the three dimensional structure of the folded protein¹²⁸. While such algorithmic approaches applied to large datasets can prioritize likely functionally important mutations, they cannot easily prove the converse – that a mutation has no functional effect.

Beyond new computational methods, how can further directed sequencing of patient tumors guide rare allele discovery and prioritization? Rather than undirected sequencing of more primary untreated tumors, which has predominated in the literature to date, the sequencing of underrepresented patient populations such as those with rare cancer types or those with advanced or post-treatment disease may identify important

mutations that arise too rarely in common cancer types to draw the attention of the research community. In rare cancers, rare mutations have been discovered that are obligate events in their pathogenesis including *GNAQ* and *GNA11* (Q209) in uveal melanomas^{165,166}, *H3F3A* (K28) in pontine gliomas^{167,168}, *GNAS* (R201) in intraductal papillary mucinous neoplasms¹⁶⁹, and *PRKD1* (Q710) mutations in salivary gland tumors¹⁷⁰. Even rarer alleles are also emerging now small GTPases such as *RRAS*, *RRAS2*, *RAC1* and *RAC2*¹⁷¹⁻¹⁷³. Likewise, novel and context-rich early-phase basket trials will similarly accelerate the speed of rare mutation discovery. In so doing, these studies are likely to enrich for the discovery of biological and clinically significant rare alleles because all patients are treated with a purpose-built molecularly-driven therapy and share a common genomic alteration. Moreover, integrating pre-treatment and post-progression biopsies and serial collection of circulating tumor DNA into such trials may be as fruitful a strategy for rare mutation detection as mining large-scale cancer genome data from tens of thousands of primary untreated patients. Such clinical studies can, therefore, serve as a far more focused discovery platform as compared to retrospective characterization of otherwise unselected populations of cancer patients.

Finally, an essential adjunct to the aforementioned approaches for rare allele prioritization is the development of high-throughput assays to facilitate broad-scale functional screening of mutations. One example is saturation mutagenesis^{174,175} of all possible mutant alleles in every residue of a given protein^{176,177}. This approach would allow testing of the large number of mutations detected in tumors but also anticipate the discovery of yet unseen mutations. Such assays and related approaches^{134,157} can

parallelize the exploration of a large number of rare alleles in the long right tail, prioritize mutations for in-depth mechanistic characterization, and identify mutations that appear to be biologically silent, facilitating the refinement of the aforementioned computational methods. Nevertheless, one limitation of this approach is that the aberrant function of mutations can be dictated by the context in which they arise and thus oncogenicity may not always be evident in high-throughput assays. Capturing such context-specific biological properties would require screening mutations with these highly parallel approaches across a large number of different cell types and cell growth conditions for multiple phenotypes (eg., growth, signaling, stability, etc.), and combinations thereof. The complexity of such an endeavor may be prohibitive. Therefore, while appropriate as an initial screen, such approaches may be better suited for prioritization rather than the detailed biochemical validation that is necessarily lower throughput.

Validation. Functional biochemistry remains the most powerful tool for investigating the varied and context-specific effects of rare mutations. This approach has been most effective when the physiologic function of the wildtype protein is well established. Discovering the often-subtle differences among driver mutations found in oncoproteins like MTOR, PKC¹⁷⁸, and RAS stems from careful study of the biochemical functions of the wildtype proteins. Moreover, many examples exist of the importance of measuring varied phenotypes (e.g., pathway activation, enzyme activity, protein stability) for proper validation of mutant allele function. For instance, while *NRAS* G12 and Q61 both activate MAPK signaling to a similar degree (shared signaling phenotype), Q61 mutations result in higher rates of melanoma initiation (different phenotype) in mice, a

finding that is consistent with the much higher frequency of Q61 mutations in melanoma patients⁸⁶. Similarly, *BRAF* V600E is sufficient alone for cellular transformation in some cancers¹⁷⁹, but not in others¹⁸⁰, yet is an oncogenic driver of both. But, in which cellular system should such multi-phenotypic validation be performed? With the advent of RNA-guided CRISPR/Cas9 in human cells, genome editing has become an attractive approach for engineering a candidate rare mutation into a cell that does not already possess it. But if context conditions rare allele function, then experiments in engineered cells must control for an enormous number of additional variables, beyond the mutation itself, to recapitulate the ecosystem in which the allele was observed in patients. Instead, an alternative strategy is to remove the mutation from a cell that natively possesses it such as in patient-derived tumor xenografts. Here, candidate driver mutations can be removed either genetically or pharmacologically with selective drugs. While patient-derived models are not without their own limitations, they have the potential to facilitate the study of rare alleles in their native context. Doing so, however, requires that our community heed recent calls to expand the generation of such patient-derived models of human cancer¹⁸¹. Moreover, where prospective clinical sequencing has begun, and rare mutations of interest are now being routinely observed, incorporating into patient consent the ability to perform research biopsies to facilitate the systematic generation of such models is essential, as is sharing them widely¹⁸².

Ultimately, a unified approach that integrates computational and experimental tools with judicious clinical validation will be necessary to identify and validate biologically and therapeutically significant rare driver mutations in real-time (**Figure 5.1**). Such an

accelerated process raises the provocative possibility that patients with rare mutations can be enrolled in clinical trials even before detailed experimental validation, especially when the weight of computational evidence is high and the affected patients have no standard of care treatment options. In such cases, rare mutations can be explored as a predictive biomarker of drug sensitivity in parallel with functional experiments in a co-clinical trial framework. This strategy is especially appealing when coupled to basket studies that assess the therapeutic sensitivity of patients whose disease is defined by 1) a single rare driver mutation that is uncommon in all the cancer types in which it was identified, or 2) one of multiple rare mutations in a given uncommonly mutated cancer gene. Such studies can assess many rare mutations simultaneously while also assessing some of the aforementioned context dependencies such as how cell type conditions function and drug sensitivity.

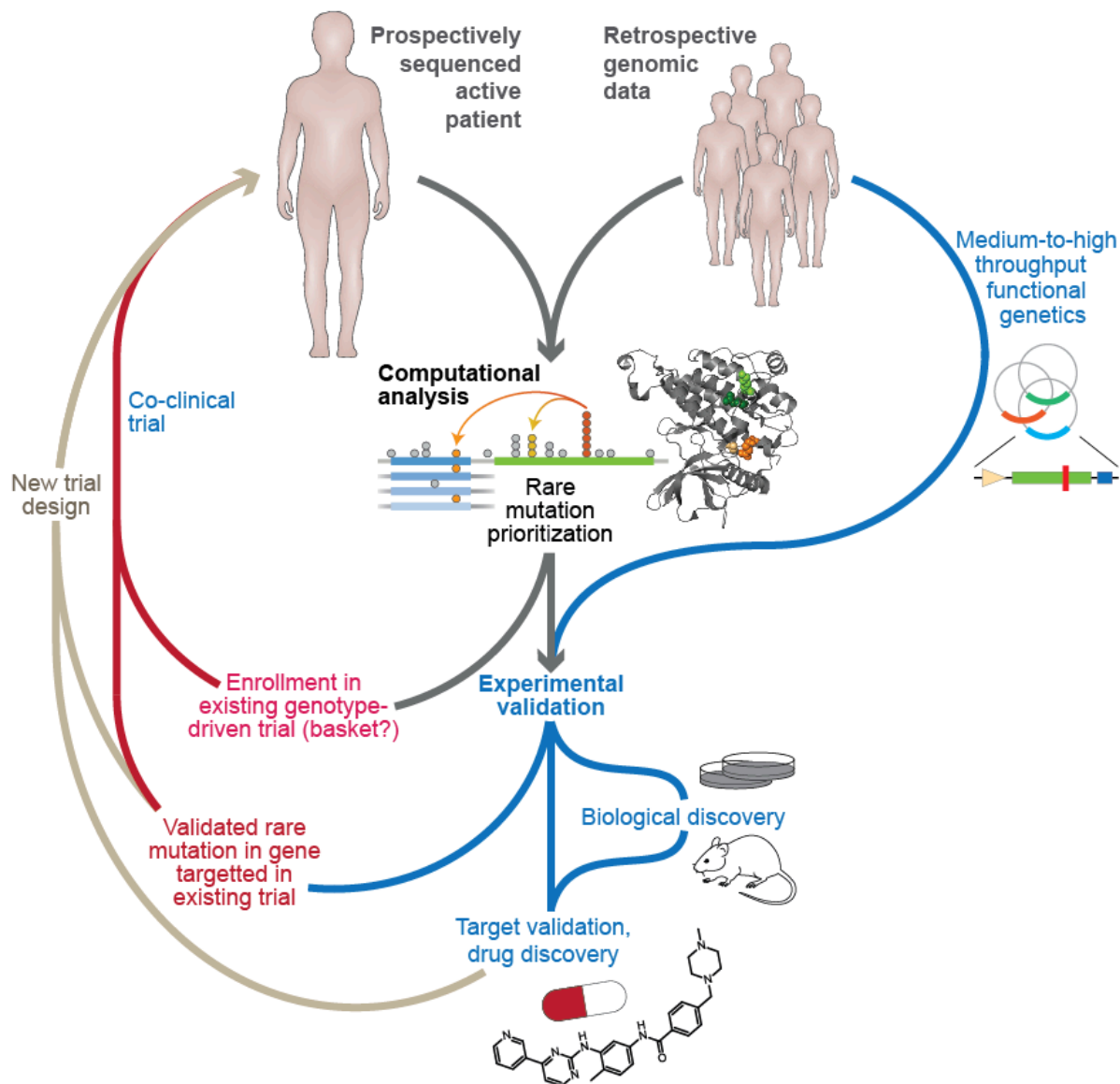


Figure 5.1. Lifecycle of rare mutation discovery and actionability A complementary array of strategies can translate the large number of rare mutations of unknown significance being identified either retrospectively or prospectively in advanced cancer patients (top) into knowledge (bottom). These can include computational analyses (middle) or medium-to-high throughput functional genetic assays (far right) to prioritize mutant alleles for in-depth experimental characterization, which can lead to new biological discoveries and the validate of new drug targets. Results from any of these approaches may also expand the eligibility criteria for enrollment in clinical trials or motivate the design of new trials (left, brown and red).

5.4 Opportunities and challenges

Understanding the biological and therapeutic importance of rare mutations is a prerequisite for the effective practice of precision oncology. A considerable fraction of human cancers, despite broad-based molecular characterization efforts, remain “driver-negative”. An even larger percentage of tumors lack a currently recognizable therapeutically actionable alteration. Conceivably, the growth progression of some of these tumors may be driven by rare or even private mutations that remain to be discovered. Despite these realities, the therapeutic value of rare mutations is often called into question. Do infrequent mutations truly warrant such efforts?

We argue that while rare individually, these mutations in aggregate affect a sizable number of patients across cancer types and may define a truly distinct molecular subtype of disease that contributes to our fundamental understanding of human cancer pathogenesis. Moreover, understanding the properties of rare mutations in effectors of the same pathways affected by common mutations may reveal therapeutic vulnerabilities that extend existing therapies to more patients. Therefore, the near-term therapeutic significance of a rare mutation does not always correlate with its frequency in any patient population. If the study of exceptional responses to anti-cancer therapy has taught us anything, it might be that translating rare mutations into broader therapeutic approaches is possible and can broaden the clinical utility of such therapies. Every rare or even private driver mutation matters when prospectively identified in advanced cancer patients having failed standard-of-care therapies for whom there are few therapeutic options. The clues left behind by the rare mutations characterized to

date will help the field develop better models for distinguishing rare functional mutations from passenger mutations. We have only begun to unravel what conditions the function and therapeutic sensitivity of both common and rare mutations. These efforts in turn may uncover new tumor biology and accelerate clinical translation of biomarkers for mechanism-based cancer treatments.

References:

- 1 Garraway, L. A. & Lander, E. S. Lessons from the cancer genome. *Cell* **153**, 17-37, doi:10.1016/j.cell.2013.03.002 (2013).
- 2 Nowell, P. C. The clonal evolution of tumor cell populations. *Science* **194**, 23-28 (1976).
- 3 Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. *Cell* **100**, 57-70 (2000).
- 4 Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415-421, doi:10.1038/nature12477 (2013).
- 5 Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214-218, doi:10.1038/nature12213 (2013).
- 6 Roberts, S. A. & Gordenin, D. A. Hypermutation in human cancer genomes: footprints and mechanisms. *Nat Rev Cancer* **14**, 786-800, doi:10.1038/nrc3816 (2014).
- 7 Nik-Zainal, S. *et al.* Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979-993, doi:10.1016/j.cell.2012.04.024 (2012).
- 8 Parikh, C. *et al.* Disruption of PH-kinase domain interactions leads to oncogenic activation of AKT in human cancers. *Proc Natl Acad Sci U S A* **109**, 19368-19373, doi:10.1073/pnas.1204384109 (2012).
- 9 Grabiner, B. C. *et al.* A diverse array of cancer-associated MTOR mutations are hyperactivating and can predict rapamycin sensitivity. *Cancer Discov* **4**, 554-563, doi:10.1158/2159-8290.CD-13-0929 (2014).
- 10 Sato, T., Nakashima, A., Guo, L., Coffman, K. & Tamanoi, F. Single amino-acid changes that confer constitutive activation of mTOR are discovered in human cancer. *Oncogene* **29**, 2746-2752, doi:10.1038/onc.2010.28 (2010).
- 11 Cancer Genome Atlas Research, N. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543-550, doi:10.1038/nature13385 (2014).
- 12 Cancer Genome Atlas Research, N. The Molecular Taxonomy of Primary Prostate Cancer. *Cell* **163**, 1011-1025, doi:10.1016/j.cell.2015.10.025 (2015).
- 13 Cancer Genome Atlas, N. Genomic Classification of Cutaneous Melanoma. *Cell* **161**, 1681-1696, doi:10.1016/j.cell.2015.05.044 (2015).
- 14 Cancer Genome Atlas Research, N. Integrated genomic characterization of papillary thyroid carcinoma. *Cell* **159**, 676-690, doi:10.1016/j.cell.2014.09.050 (2014).
- 15 Brennan, C. W. *et al.* The somatic genomic landscape of glioblastoma. *Cell* **155**, 462-477, doi:10.1016/j.cell.2013.09.034 (2013).
- 16 Chang, M. T. *et al.* Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. *Nat Biotechnol* **34**, 155-163, doi:10.1038/nbt.3391 (2016).
- 17 Biankin, A. V. *et al.* Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes. *Nature* **491**, 399-405, doi:10.1038/nature11547 (2012).
- 18 Witkiewicz, A. K. *et al.* Whole-exome sequencing of pancreatic cancer defines genetic diversity and therapeutic targets. *Nat Commun* **6**, 6744, doi:10.1038/ncomms7744 (2015).

- 19 Cancer Genome Atlas Research, N. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N Engl J Med* **368**, 2059-2074, doi:10.1056/NEJMoa1301689 (2013).
- 20 Turcan, S. *et al.* IDH1 mutation is sufficient to establish the glioma hypermethylator phenotype. *Nature* **483**, 479-483, doi:10.1038/nature10866 (2012).
- 21 Yamamoto, Y. *et al.* Activating mutation of D835 within the activation loop of FLT3 in human hematologic malignancies. *Blood* **97**, 2434-2439 (2001).
- 22 Clark, J. J. *et al.* Variable sensitivity of FLT3 activation loop mutations to the small molecule tyrosine kinase inhibitor MLN518. *Blood* **104**, 2867-2872, doi:10.1182/blood-2003-12-4446 (2004).
- 23 Bailey, E. *et al.* FLT3/D835Y mutation knock-in mice display less aggressive disease compared with FLT3/internal tandem duplication (ITD) mice. *Proc Natl Acad Sci U S A* **110**, 21113-21118, doi:10.1073/pnas.1310559110 (2013).
- 24 Li, Z. *et al.* Discovery of AMG 925, a FLT3 and CDK4 dual kinase inhibitor with preferential affinity for the activated state of FLT3. *J Med Chem* **57**, 3430-3449, doi:10.1021/jm500118j (2014).
- 25 Li, C. *et al.* AMG 925 is a dual FLT3/CDK4 inhibitor with the potential to overcome FLT3 inhibitor resistance in acute myeloid leukemia. *Mol Cancer Ther* **14**, 375-383, doi:10.1158/1535-7163.MCT-14-0388 (2015).
- 26 Zack, T. I. *et al.* Pan-cancer patterns of somatic copy number alteration. *Nat Genet* **45**, 1134-1140, doi:10.1038/ng.2760 (2013).
- 27 Yada, M. *et al.* Phosphorylation-dependent degradation of c-Myc is mediated by the F-box protein Fbw7. *EMBO J* **23**, 2116-2125, doi:10.1038/sj.emboj.7600217 (2004).
- 28 Zuo, L. *et al.* Germline mutations in the p16INK4a binding domain of CDK4 in familial melanoma. *Nat Genet* **12**, 97-99, doi:10.1038/ng0196-97 (1996).
- 29 Wheeler, D. B., Zoncu, R., Root, D. E., Sabatini, D. M. & Sawyers, C. L. Identification of an oncogenic RAB protein. *Science* **350**, 211-217, doi:10.1126/science.aaa4903 (2015).
- 30 Cerami, E. *et al.* The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov* **2**, 401-404, doi:10.1158/2159-8290.CD-12-0095 (2012).
- 31 Gao, J. *et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal* **6**, p11, doi:10.1126/scisignal.2004088 (2013).
- 32 Nakanishi, Y. *et al.* Activating Mutations in PIK3CB Confer Resistance to PI3K Inhibition and Define a Novel Oncogenic Role for p110beta. *Cancer Res* **76**, 1193-1203, doi:10.1158/0008-5472.CAN-15-2201 (2016).
- 33 Lawrence, M. S. *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495-501, doi:10.1038/nature12912 (2014).
- 34 Arcila, M. E. *et al.* MAP2K1 (MEK1) Mutations Define a Distinct Subset of Lung Adenocarcinoma Associated with Smoking. *Clin Cancer Res* **21**, 1935-1943, doi:10.1158/1078-0432.CCR-14-2124 (2015).

- 35 Cancer Genome Atlas, N. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330-337, doi:10.1038/nature11252 (2012).
- 36 Cancer Genome Atlas Research, N. *et al.* Integrated genomic characterization of endometrial carcinoma. *Nature* **497**, 67-73, doi:10.1038/nature12113 (2013).
- 37 Shlien, A. *et al.* Combined hereditary and somatic mutations of replication error repair genes result in rapid onset of ultra-hypermuted cancers. *Nat Genet* **47**, 257-262, doi:10.1038/ng.3202 (2015).
- 38 Le, D. T. *et al.* PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. *N Engl J Med* **372**, 2509-2520, doi:10.1056/NEJMoa1500596 (2015).
- 39 Van Allen, E. M. *et al.* Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. *Science* **350**, 207-211, doi:10.1126/science.aad0095 (2015).
- 40 Levine, A. J. & Oren, M. The first 30 years of p53: growing ever more complex. *Nat Rev Cancer* **9**, 749-758, doi:10.1038/nrc2723 (2009).
- 41 Milner, J. & Medcalf, E. A. Cotranslation of activated mutant p53 with wild type drives the wild-type p53 protein into the mutant conformation. *Cell* **65**, 765-774 (1991).
- 42 Zhu, J. *et al.* Gain-of-function p53 mutants co-opt chromatin pathways to drive cancer growth. *Nature* **525**, 206-211, doi:10.1038/nature15251 (2015).
- 43 Donovan, S., Shannon, K. M. & Bollag, G. GTPase activating proteins: critical regulators of intracellular signaling. *Biochim Biophys Acta* **1602**, 23-45 (2002).
- 44 Vivanco, I. *et al.* Differential sensitivity of glioma- versus lung cancer-specific EGFR mutations to EGFR kinase inhibitors. *Cancer Discov* **2**, 458-471, doi:10.1158/2159-8290.CD-11-0284 (2012).
- 45 Dogruluk, T. *et al.* Identification of Variant-Specific Functions of PIK3CA by Rapid Phenotyping of Rare Mutations. *Cancer Res* **75**, 5341-5354, doi:10.1158/0008-5472.CAN-15-1654 (2015).
- 46 Ostrem, J. M., Peters, U., Sos, M. L., Wells, J. A. & Shokat, K. M. K-Ras(G12C) inhibitors allosterically control GTP affinity and effector interactions. *Nature* **503**, 548-551, doi:10.1038/nature12796 (2013).
- 47 Lito, P., Solomon, M., Li, L. S., Hansen, R. & Rosen, N. Allele-specific inhibitors inactivate mutant KRAS G12C by a trapping mechanism. *Science* **351**, 604-608, doi:10.1126/science.aad6204 (2016).
- 48 Yao, Z. *et al.* BRAF Mutants Evade ERK-Dependent Feedback by Different Mechanisms that Determine Their Sensitivity to Pharmacologic Inhibition. *Cancer Cell* **28**, 370-383, doi:10.1016/j.ccell.2015.08.001 (2015).
- 49 Cheung, L. W. *et al.* High frequency of PIK3R1 and PIK3R2 mutations in endometrial cancer elucidates a novel mechanism for regulation of PTEN protein stability. *Cancer Discov* **1**, 170-185, doi:10.1158/2159-8290.CD-11-0039 (2011).
- 50 Cheung, L. W. *et al.* Naturally occurring neomorphic PIK3R1 mutations activate the MAPK pathway, dictating therapeutic response to MAPK pathway inhibitors. *Cancer Cell* **26**, 479-494, doi:10.1016/j.ccell.2014.08.017 (2014).
- 51 Yoshida, K. *et al.* Frequent pathway mutations of splicing machinery in myelodysplasia. *Nature* **478**, 64-69, doi:10.1038/nature10496 (2011).

- 52 Papaemmanuil, E. *et al.* Somatic SF3B1 mutation in myelodysplasia with ring sideroblasts. *N Engl J Med* **365**, 1384-1395, doi:10.1056/NEJMoa1103283 (2011).
- 53 Graubert, T. A. *et al.* Recurrent mutations in the U2AF1 splicing factor in myelodysplastic syndromes. *Nat Genet* **44**, 53-57, doi:10.1038/ng.1031 (2011).
- 54 Ilagan, J. O. *et al.* U2AF1 mutations alter splice site recognition in hematological malignancies. *Genome Res* **25**, 14-26, doi:10.1101/gr.181016.114 (2015).
- 55 Toy, W. *et al.* ESR1 ligand-binding domain mutations in hormone-resistant breast cancer. *Nat Genet* **45**, 1439-1445, doi:10.1038/ng.2822 (2013).
- 56 Fanning, S. W. *et al.* Estrogen receptor alpha somatic mutations Y537S and D538G confer breast cancer endocrine resistance by stabilizing the activating function-2 binding conformation. *Elife* **5**, doi:10.7554/eLife.12792 (2016).
- 57 Gonzalez-Perez, A. *et al.* IntOGen-mutations identifies cancer drivers across tumor types. *Nat Methods* **10**, 1081-1082, doi:10.1038/nmeth.2642 (2013).
- 58 Kandath, C. *et al.* Mutational landscape and significance across 12 major cancer types. *Nature* **502**, 333-339, doi:10.1038/nature12634 (2013).
- 59 Santarius, T., Shipley, J., Brewer, D., Stratton, M. R. & Cooper, C. S. A census of amplified and overexpressed human cancer genes. *Nat Rev Cancer* **10**, 59-64, doi:10.1038/nrc2771 (2010).
- 60 Tamborero, D. *et al.* Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Sci Rep* **3**, 2650, doi:10.1038/srep02650 (2013).
- 61 Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339**, 1546-1558, doi:10.1126/science.1235122 (2013).
- 62 Castro-Giner, F., Ratcliffe, P. & Tomlinson, I. The mini-driver model of polygenic cancer evolution. *Nat Rev Cancer* **15**, 680-685, doi:10.1038/nrc3999 (2015).
- 63 Hyman, D. M. *et al.* Vemurafenib in Multiple Nonmelanoma Cancers with BRAF V600 Mutations. *N Engl J Med* **373**, 726-736, doi:10.1056/NEJMoa1502309 (2015).
- 64 Smith, G. *et al.* Activating K-Ras mutations outwith 'hotspot' codons in sporadic colorectal tumours - implications for personalised cancer medicine. *Br J Cancer* **102**, 693-703, doi:10.1038/sj.bjc.6605534 (2010).
- 65 Janakiraman, M. *et al.* Genomic and biological characterization of exon 4 KRAS mutations in human cancer. *Cancer Res* **70**, 5901-5911, doi:10.1158/0008-5472.CAN-10-0192 (2010).
- 66 Sloan, S. R., Newcomb, E. W. & Pellicer, A. Neutron radiation can activate K-ras via a point mutation in codon 146 and induces a different spectrum of ras mutations than does gamma radiation. *Mol Cell Biol* **10**, 405-408 (1990).
- 67 Tyner, J. W. *et al.* High-throughput sequencing screen reveals novel, transforming RAS mutations in myeloid leukemia patients. *Blood* **113**, 1749-1755, doi:10.1182/blood-2008-04-152157 (2009).
- 68 Park, J. T. *et al.* Differential in vivo tumorigenicity of diverse KRAS mutations in vertebrate pancreas: A comprehensive survey. *Oncogene* **34**, 2801-2806, doi:10.1038/onc.2014.223 (2015).

- 69 Hodis, E. *et al.* A landscape of driver mutations in melanoma. *Cell* **150**, 251-263, doi:10.1016/j.cell.2012.06.024 (2012).
- 70 Krauthammer, M. *et al.* Exome sequencing identifies recurrent somatic RAC1 mutations in melanoma. *Nat Genet* **44**, 1006-1014, doi:10.1038/ng.2359 (2012).
- 71 Watson, I. R. *et al.* The RAC1 P29S hotspot mutation in melanoma confers resistance to pharmacological inhibition of RAF. *Cancer Res* **74**, 4845-4852, doi:10.1158/0008-5472.CAN-14-1232-T (2014).
- 72 Davies, H. *et al.* Mutations of the BRAF gene in human cancer. *Nature* **417**, 949-954, doi:10.1038/nature00766 (2002).
- 73 Brose, M. S. *et al.* BRAF and RAS mutations in human lung cancer and melanoma. *Cancer Res* **62**, 6997-7000 (2002).
- 74 Gorden, A. *et al.* Analysis of BRAF and N-RAS mutations in metastatic melanoma tissues. *Cancer Res* **63**, 3955-3957 (2003).
- 75 Heidorn, S. J. *et al.* Kinase-dead BRAF and oncogenic RAS cooperate to drive tumor progression through CRAF. *Cell* **140**, 209-221, doi:10.1016/j.cell.2009.12.040 (2010).
- 76 Gorre, M. E. *et al.* Clinical resistance to STI-571 cancer therapy caused by BCR-ABL gene mutation or amplification. *Science* **293**, 876-880, doi:10.1126/science.1062538 (2001).
- 77 Willis, S. G. *et al.* High-sensitivity detection of BCR-ABL kinase domain mutations in imatinib-naive patients: correlation with clonal cytogenetic evolution but not response to therapy. *Blood* **106**, 2128-2137, doi:10.1182/blood-2005-03-1036 (2005).
- 78 Yu, H. A. *et al.* Analysis of tumor specimens at the time of acquired resistance to EGFR-TKI therapy in 155 patients with EGFR-mutant lung cancers. *Clin Cancer Res* **19**, 2240-2247, doi:10.1158/1078-0432.CCR-12-2246 (2013).
- 79 Shih, J. Y., Gow, C. H. & Yang, P. C. EGFR mutation conferring primary resistance to gefitinib in non-small-cell lung cancer. *N Engl J Med* **353**, 207-208, doi:10.1056/NEJM200507143530217 (2005).
- 80 Yu, H. A. *et al.* Poor response to erlotinib in patients with tumors containing baseline EGFR T790M mutations found by routine clinical molecular testing. *Ann Oncol* **25**, 423-428, doi:10.1093/annonc/mdt573 (2014).
- 81 Hata, A. N. *et al.* Tumor cells can follow distinct evolutionary paths to become resistant to epidermal growth factor receptor inhibition. *Nat Med* **22**, 262-269, doi:10.1038/nm.4040 (2016).
- 82 Shaw, A. T. *et al.* Resensitization to Crizotinib by the Lorlatinib ALK Resistance Mutation L1198F. *N Engl J Med* **374**, 54-61, doi:10.1056/NEJMoa1508887 (2016).
- 83 Cancer Genome Atlas Research, N. Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature* **507**, 315-322, doi:10.1038/nature12965 (2014).
- 84 Lee, C. S. *et al.* Recurrent point mutations in the kinetochore gene KNSTRN in cutaneous squamous cell carcinoma. *Nat Genet* **46**, 1060-1062, doi:10.1038/ng.3091 (2014).

- 85 Jaiswal, B. S. *et al.* Oncogenic ERBB3 mutations in human cancers. *Cancer Cell* **23**, 603-617, doi:10.1016/j.ccr.2013.04.012 (2013).
- 86 Burd, C. E. *et al.* Mutation-specific RAS oncogenicity explains NRAS codon 61 selection in melanoma. *Cancer Discov* **4**, 1418-1429, doi:10.1158/2159-8290.CD-14-0729 (2014).
- 87 Menzies, A. M. *et al.* Distinguishing clinicopathologic features of patients with V600E and V600K BRAF-mutant metastatic melanoma. *Clin Cancer Res* **18**, 3242-3249, doi:10.1158/1078-0432.CCR-12-0052 (2012).
- 88 Westcott, P. M. *et al.* The mutational landscapes of genetic and chemical models of Kras-driven lung cancer. *Nature* **517**, 489-492, doi:10.1038/nature13898 (2015).
- 89 Karolchik, D. *et al.* The UCSC Genome Browser database: 2014 update. *Nucleic Acids Res* **42**, D764-770, doi:10.1093/nar/gkt1168 (2014).
- 90 McLaren, W. *et al.* Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26**, 2069-2070, doi:10.1093/bioinformatics/btq330 (2010).
- 91 Cheng, D. T. *et al.* Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT): A Hybridization Capture-Based Next-Generation Sequencing Clinical Assay for Solid Tumor Molecular Oncology. *J Mol Diagn* **17**, 251-264, doi:10.1016/j.jmoldx.2014.12.006 (2015).
- 92 Mullighan, C. G. *et al.* CREBBP mutations in relapsed acute lymphoblastic leukaemia. *Nature* **471**, 235-239, doi:10.1038/nature09727 (2011).
- 93 Weng, A. P. *et al.* Activating mutations of NOTCH1 in human T cell acute lymphoblastic leukemia. *Science* **306**, 269-271, doi:10.1126/science.1102160 (2004).
- 94 Hart, T., Brown, K. R., Sircoulomb, F., Rottapel, R. & Moffat, J. Measuring error rates in genomic perturbation screens: gold standards for human functional genomics. *Mol Syst Biol* **10**, 733, doi:10.15252/msb.20145216 (2014).
- 95 Yu, H. A. *et al.* Prognostic impact of KRAS mutation subtypes in 677 patients with metastatic lung adenocarcinomas. *J Thorac Oncol* **10**, 431-437, doi:10.1097/JTO.0000000000000432 (2015).
- 96 Ihle, N. T. *et al.* Effect of KRAS oncogene substitutions on protein behavior: implications for signaling and clinical outcome. *J Natl Cancer Inst* **104**, 228-239, doi:10.1093/jnci/djr523 (2012).
- 97 Garassino, M. C. *et al.* Different types of K-Ras mutations could affect drug sensitivity and tumour behaviour in non-small-cell lung cancer. *Ann Oncol* **22**, 235-237, doi:10.1093/annonc/mdq680 (2011).
- 98 de Bruin, E. C. *et al.* Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. *Science* **346**, 251-256, doi:10.1126/science.1253462 (2014).
- 99 Whitehall, V. L. *et al.* Oncogenic PIK3CA mutations in colorectal cancers and polyps. *Int J Cancer* **131**, 813-820, doi:10.1002/ijc.26440 (2012).

- 100 Al-Ahmadie, H. *et al.* Synthetic lethality in ATM-deficient RAD50-mutant tumors underlies outlier response to cancer therapy. *Cancer Discov* **4**, 1014-1021, doi:10.1158/2159-8290.CD-14-0380 (2014).
- 101 Iyer, G. *et al.* Genome sequencing identifies a basis for everolimus sensitivity. *Science* **338**, 221, doi:10.1126/science.1226344 (2012).
- 102 Nagy, E. & Maquat, L. E. A rule for termination-codon position within intron-containing genes: when nonsense affects RNA abundance. *Trends Biochem Sci* **23**, 198-199 (1998).
- 103 MacArthur, D. G. *et al.* A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**, 823-828, doi:10.1126/science.1215040 (2012).
- 104 Parikh, N. *et al.* Effects of TP53 mutational status on gene expression patterns across 10 human cancer types. *J Pathol* **232**, 522-533, doi:10.1002/path.4321 (2014).
- 105 Amati, B. *et al.* Oncogenic activity of the c-Myc protein requires dimerization with Max. *Cell* **72**, 233-245 (1993).
- 106 Comino-Mendez, I. *et al.* Exome sequencing identifies MAX mutations as a cause of hereditary pheochromocytoma. *Nat Genet* **43**, 663-667, doi:10.1038/ng.861 (2011).
- 107 Burnichon, N. *et al.* MAX mutations cause hereditary and sporadic pheochromocytoma and paraganglioma. *Clin Cancer Res* **18**, 2828-2837, doi:10.1158/1078-0432.CCR-12-0160 (2012).
- 108 Diolaiti, D., McFerrin, L., Carroll, P. A. & Eisenman, R. N. Functional interactions among members of the MAX and MLX transcriptional network during oncogenesis. *Biochim Biophys Acta* **1849**, 484-500, doi:10.1016/j.bbagr.2014.05.016 (2015).
- 109 Forbes, S. A. *et al.* COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res* **43**, D805-811, doi:10.1093/nar/gku1075 (2015).
- 110 Van Allen, E. M. *et al.* The genetic landscape of clinical resistance to RAF inhibition in metastatic melanoma. *Cancer Discov* **4**, 94-109, doi:10.1158/2159-8290.CD-13-0617 (2014).
- 111 Ehrhardt, A., Ehrhardt, G. R., Guo, X. & Schrader, J. W. Ras and relatives--job sharing and networking keep an old family together. *Exp Hematol* **30**, 1089-1106 (2002).
- 112 Barker, K. T. & Crompton, M. R. Ras-related TC21 is activated by mutation in a breast cancer cell line, but infrequently in breast carcinomas in vivo. *Br J Cancer* **78**, 296-300 (1998).
- 113 Clark, G. J., Kinch, M. S., Gilmer, T. M., Burridge, K. & Der, C. J. Overexpression of the Ras-related TC21/R-Ras2 protein may contribute to the development of human breast cancers. *Oncogene* **12**, 169-176 (1996).
- 114 Huang, Y. *et al.* A novel insertional mutation in the TC21 gene activates its transforming activity in a human leiomyosarcoma cell line. *Oncogene* **11**, 1255-1260 (1995).

- 115 Erdogan, M., Pozzi, A., Bhowmick, N., Moses, H. L. & Zent, R. Signaling pathways regulating TC21-induced tumorigenesis. *J Biol Chem* **282**, 27713-27720, doi:10.1074/jbc.M703037200 (2007).
- 116 Rosario, M., Paterson, H. F. & Marshall, C. J. Activation of the Ral and phosphatidylinositol 3' kinase signaling pathways by the ras-related protein TC21. *Mol Cell Biol* **21**, 3750-3762, doi:10.1128/MCB.21.11.3750-3762.2001 (2001).
- 117 Rong, R., He, Q., Liu, Y., Sheikh, M. S. & Huang, Y. TC21 mediates transformation and cell survival via activation of phosphatidylinositol 3-kinase/Akt and NF-kappaB signaling pathway. *Oncogene* **21**, 1062-1070, doi:10.1038/sj.onc.1205154 (2002).
- 118 Rosario, M., Paterson, H. F. & Marshall, C. J. Activation of the Raf/MAP kinase cascade by the Ras-related protein TC21 is required for the TC21-mediated transformation of NIH 3T3 cells. *EMBO J* **18**, 1270-1279, doi:10.1093/emboj/18.5.1270 (1999).
- 119 Cancer Genome Atlas Research, N. *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet* **45**, 1113-1120, doi:10.1038/ng.2764 (2013).
- 120 Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458**, 719-724, doi:10.1038/nature07943 (2009).
- 121 Tamborero, D., Gonzalez-Perez, A. & Lopez-Bigas, N. OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics* **29**, 2238-2244, doi:10.1093/bioinformatics/btt395 (2013).
- 122 Reimand, J. & Bader, G. D. Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. *Mol Syst Biol* **9**, 637, doi:10.1038/msb.2012.68 (2013).
- 123 Miller, M. L. *et al.* Pan-Cancer Analysis of Mutation Hotspots in Protein Domains. *Cell Syst* **1**, 197-209, doi:10.1016/j.cels.2015.08.014 (2015).
- 124 Dixit, A. *et al.* Sequence and structure signatures of cancer mutation hotspots in protein kinases. *PLoS One* **4**, e7485, doi:10.1371/journal.pone.0007485 (2009).
- 125 Wang, X. *et al.* Three-dimensional reconstruction of protein networks provides insight into human genetic disease. *Nat Biotechnol* **30**, 159-164, doi:10.1038/nbt.2106 (2012).
- 126 Porta-Pardo, E., Garcia-Alonso, L., Hrabe, T., Dopazo, J. & Godzik, A. A Pan-Cancer Catalogue of Cancer Driver Protein Interaction Interfaces. *PLoS Comput Biol* **11**, e1004518, doi:10.1371/journal.pcbi.1004518 (2015).
- 127 Engin, H. B., Kreisberg, J. F. & Carter, H. Structure-Based Analysis Reveals Cancer Missense Mutations Target Protein Interaction Interfaces. *PLoS One* **11**, e0152929, doi:10.1371/journal.pone.0152929 (2016).
- 128 Kamburov, A. *et al.* Comprehensive assessment of cancer missense mutation clustering in protein structures. *Proc Natl Acad Sci U S A* **112**, E5486-5495, doi:10.1073/pnas.1516373112 (2015).
- 129 Gress, A., Ramensky, V., Buch, J., Keller, A. & Kalinina, O. V. StructMAN: annotation of single-nucleotide polymorphisms in the structural context. *Nucleic Acids Res* **44**, W463-468, doi:10.1093/nar/gkw364 (2016).

- 130 Ryslik, G. A. *et al.* A spatial simulation approach to account for protein structure when identifying non-random somatic mutations. *BMC Bioinformatics* **15**, 231, doi:10.1186/1471-2105-15-231 (2014).
- 131 Meyer, M. J. *et al.* mutation3D: Cancer Gene Prediction Through Atomic Clustering of Coding Variants in the Structural Proteome. *Hum Mutat* **37**, 447-456, doi:10.1002/humu.22963 (2016).
- 132 Tokheim, C. *et al.* Exome-Scale Discovery of Hotspot Mutation Regions in Human Cancer Using 3D Protein Structure. *Cancer Res* **76**, 3719-3731, doi:10.1158/0008-5472.CAN-15-3190 (2016).
- 133 Niu, B. *et al.* Protein-structure-guided discovery of functional mutations across 19 cancer types. *Nat Genet* **48**, 827-837, doi:10.1038/ng.3586 (2016).
- 134 Kim, E. *et al.* Systematic Functional Interrogation of Rare Cancer Variants Identifies Oncogenic Alleles. *Cancer Discov* **6**, 714-726, doi:10.1158/2159-8290.CD-16-0160 (2016).
- 135 Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res* **28**, 235-242 (2000).
- 136 UniProt, C. UniProt: a hub for protein information. *Nucleic Acids Res* **43**, D204-212, doi:10.1093/nar/gku989 (2015).
- 137 Reva, B., Antipin, Y. & Sander, C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res* **39**, e118, doi:10.1093/nar/gkr407 (2011).
- 138 Velankar, S. *et al.* SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. *Nucleic Acids Res* **41**, D483-489, doi:10.1093/nar/gks1258 (2013).
- 139 Prlic, A. *et al.* BioJava: an open-source framework for bioinformatics in 2012. *Bioinformatics* **28**, 2693-2695, doi:10.1093/bioinformatics/bts494 (2012).
- 140 Pratilas, C. A. *et al.* Genetic predictors of MEK dependence in non-small cell lung cancer. *Cancer Res* **68**, 9375-9383, doi:10.1158/0008-5472.CAN-08-2223 (2008).
- 141 Di Cristofano, A., Pesce, B., Cordon-Cardo, C. & Pandolfi, P. P. Pten is essential for embryonic development and tumour suppression. *Nat Genet* **19**, 348-355, doi:10.1038/1235 (1998).
- 142 Pharoah, P. D., Guilford, P., Caldas, C. & International Gastric Cancer Linkage, C. Incidence of gastric cancer and breast cancer in CDH1 (E-cadherin) mutation carriers from hereditary diffuse gastric cancer families. *Gastroenterology* **121**, 1348-1353 (2001).
- 143 Lee, Y. S. *et al.* Genomic profile analysis of diffuse-type gastric cancers. *Genome Biol* **15**, R55, doi:10.1186/gb-2014-15-4-r55 (2014).
- 144 Jaramillo, M. C. & Zhang, D. D. The emerging role of the Nrf2-Keap1 signaling pathway in cancer. *Genes Dev* **27**, 2179-2191, doi:10.1101/gad.225680.113 (2013).
- 145 Lee, J. C. *et al.* Epidermal growth factor receptor activation in glioblastoma through novel missense mutations in the extracellular domain. *PLoS Med* **3**, e485, doi:10.1371/journal.pmed.0030485 (2006).

- 146 Emery, C. M. *et al.* MEK1 mutations confer resistance to MEK and B-RAF inhibition. *Proc Natl Acad Sci U S A* **106**, 20411-20416, doi:10.1073/pnas.0905833106 (2009).
- 147 Diamond, E. L. *et al.* Diverse and Targetable Kinase Alterations Drive Histiocytic Neoplasms. *Cancer Discov* **6**, 154-165, doi:10.1158/2159-8290.CD-15-0913 (2016).
- 148 Fischmann, T. O. *et al.* Crystal structures of MEK1 binary and ternary complexes with nucleotides and inhibitors. *Biochemistry* **48**, 2661-2674, doi:10.1021/bi801898e (2009).
- 149 Robert, C. *et al.* Improved overall survival in melanoma with combined dabrafenib and trametinib. *N Engl J Med* **372**, 30-39, doi:10.1056/NEJMoa1412690 (2015).
- 150 Goss, D. A. & Grosvenor, T. Rates of childhood myopia progression with bifocals as a function of nearpoint phoria: consistency of three studies. *Optom Vis Sci* **67**, 637-640 (1990).
- 151 Davies, B. R. *et al.* Tumors with AKT1E17K Mutations Are Rational Targets for Single Agent or Combination Therapy with AKT Inhibitors. *Mol Cancer Ther* **14**, 2441-2451, doi:10.1158/1535-7163.MCT-15-0230 (2015).
- 152 Akagi, K. *et al.* Characterization of a novel oncogenic K-ras mutation in colon cancer. *Biochem Biophys Res Commun* **352**, 728-732, doi:10.1016/j.bbrc.2006.11.091 (2007).
- 153 Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285-291, doi:10.1038/nature19057 (2016).
- 154 Robinson, D. *et al.* Integrative clinical genomics of advanced prostate cancer. *Cell* **161**, 1215-1228, doi:10.1016/j.cell.2015.05.001 (2015).
- 155 Moore, A. R. *et al.* Recurrent activating mutations of G-protein-coupled receptor CYSLTR2 in uveal melanoma. *Nat Genet* **48**, 675-680, doi:10.1038/ng.3549 (2016).
- 156 Giannakakou, P. *et al.* Paclitaxel selects for mutant or pseudo-null p53 in drug resistance associated with tubulin mutations in human cancer. *Oncogene* **19**, 3078-3085, doi:10.1038/sj.onc.1203642 (2000).
- 157 Berger, A. H. *et al.* High-throughput Phenotyping of Lung Cancer Somatic Mutations. *Cancer Cell* **30**, 214-228, doi:10.1016/j.ccell.2016.06.022 (2016).
- 158 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* **105**, 4792-4799, doi:10.1182/blood-2004-11-4430 (2005).
- 159 Wagle, N. *et al.* Activating mTOR mutations in a patient with an extraordinary response on a phase I trial of everolimus and pazopanib. *Cancer Discov* **4**, 546-553, doi:10.1158/2159-8290.CD-13-0353 (2014).
- 160 Wagle, N. *et al.* Response and acquired resistance to everolimus in anaplastic thyroid cancer. *N Engl J Med* **371**, 1426-1433, doi:10.1056/NEJMoa1403352 (2014).

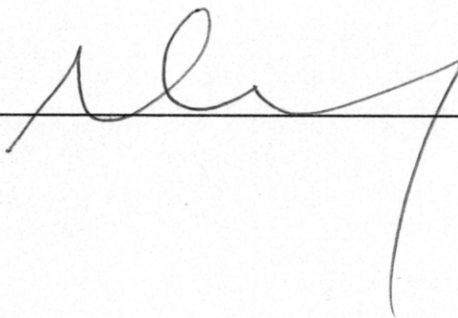
- 161 Imielinski, M. *et al.* Oncogenic and sorafenib-sensitive ARAF mutations in lung adenocarcinoma. *J Clin Invest* **124**, 1582-1586, doi:10.1172/JCI72763 (2014).
- 162 Van Allen, E. M. *et al.* Genomic Correlate of Exceptional Erlotinib Response in Head and Neck Squamous Cell Carcinoma. *JAMA Oncol* **1**, 238-244, doi:10.1001/jamaoncol.2015.34 (2015).
- 163 Grisham, R. N. *et al.* Extreme Outlier Analysis Identifies Occult Mitogen-Activated Protein Kinase Pathway Mutations in Patients With Low-Grade Serous Ovarian Cancer. *J Clin Oncol* **33**, 4099-4105, doi:10.1200/JCO.2015.62.4726 (2015).
- 164 Gauthier, N. P. *et al.* MutationAligner: a resource of recurrent mutation hotspots in protein domains in cancer. *Nucleic Acids Res* **44**, D986-991, doi:10.1093/nar/gkv1132 (2016).
- 165 Van Raamsdonk, C. D. *et al.* Frequent somatic mutations of GNAQ in uveal melanoma and blue naevi. *Nature* **457**, 599-602, doi:10.1038/nature07586 (2009).
- 166 Van Raamsdonk, C. D. *et al.* Mutations in GNA11 in uveal melanoma. *N Engl J Med* **363**, 2191-2199, doi:10.1056/NEJMoa1000584 (2010).
- 167 Schwartzenuber, J. *et al.* Driver mutations in histone H3.3 and chromatin remodelling genes in paediatric glioblastoma. *Nature* **482**, 226-231, doi:10.1038/nature10833 (2012).
- 168 Wu, G. *et al.* Somatic histone H3 alterations in pediatric diffuse intrinsic pontine gliomas and non-brainstem glioblastomas. *Nat Genet* **44**, 251-253, doi:10.1038/ng.1102 (2012).
- 169 Wu, J. *et al.* Recurrent GNAS mutations define an unexpected pathway for pancreatic cyst development. *Sci Transl Med* **3**, 92ra66, doi:10.1126/scitranslmed.3002543 (2011).
- 170 Weinreb, I. *et al.* Hotspot activating PRKD1 somatic mutations in polymorphous low-grade adenocarcinomas of the salivary glands. *Nat Genet* **46**, 1166-1169, doi:10.1038/ng.3096 (2014).
- 171 Caye, A. *et al.* Juvenile myelomonocytic leukemia displays mutations in components of the RAS pathway and the PRC2 network. *Nat Genet* **47**, 1334-1340, doi:10.1038/ng.3420 (2015).
- 172 Nassar, D., Latil, M., Boeckx, B., Lambrechts, D. & Blanpain, C. Genomic landscape of carcinogen-induced and genetically induced mouse skin squamous cell carcinoma. *Nat Med* **21**, 946-954, doi:10.1038/nm.3878 (2015).
- 173 Stieglitz, E. *et al.* The genomic landscape of juvenile myelomonocytic leukemia. *Nat Genet* **47**, 1326-1333, doi:10.1038/ng.3400 (2015).
- 174 Findlay, G. M., Boyle, E. A., Hause, R. J., Klein, J. C. & Shendure, J. Saturation editing of genomic regions by multiplex homology-directed repair. *Nature* **513**, 120-123, doi:10.1038/nature13695 (2014).
- 175 Canver, M. C. *et al.* BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. *Nature* **527**, 192-197, doi:10.1038/nature15521 (2015).
- 176 Brenan, L. *et al.* Phenotypic Characterization of a Comprehensive Set of MAPK1/ERK2 Missense Mutants. *Cell Rep* **17**, 1171-1183, doi:10.1016/j.celrep.2016.09.061 (2016).

- 177 Majithia, A. R. *et al.* Prospective functional classification of all possible missense variants in PPARG. *Nat Genet*, doi:10.1038/ng.3700 (2016).
- 178 Antal, C. E. *et al.* Cancer-associated protein kinase C mutations reveal kinase's role as tumor suppressor. *Cell* **160**, 489-502, doi:10.1016/j.cell.2015.01.001 (2015).
- 179 Kaufman, C. K. *et al.* A zebrafish melanoma model reveals emergence of neural crest identity during melanoma initiation. *Science* **351**, aad2197, doi:10.1126/science.aad2197 (2016).
- 180 Chung, S. S. *et al.* Hematopoietic stem cell origin of BRAFV600E mutations in hairy cell leukemia. *Sci Transl Med* **6**, 238ra271, doi:10.1126/scitranslmed.3008004 (2014).
- 181 Boehm, J. S. & Golub, T. R. An ecosystem of cancer cell line factories to support a cancer dependency map. *Nat Rev Genet* **16**, 373-374, doi:10.1038/nrg3967 (2015).
- 182 Townsend, E. C. *et al.* The Public Repository of Xenografts Enables Discovery and Randomized Phase II-like Trials in Mice. *Cancer Cell* **29**, 574-586, doi:10.1016/j.ccell.2016.03.008 (2016).

Publishing Agreement

It is the policy of the University to encourage the distribution of all theses, dissertations, and manuscripts. Copies of all UCSF theses, dissertations, and manuscripts will be routed to the library via the Graduate Division. The library will make all theses, dissertations, and manuscripts accessible to the public and will preserve these to the best of their abilities, in perpetuity.

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis, dissertation, or manuscript to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.

Author Signature  Date 12/9/2016