Published OnlineFirst April 28, 2016; DOI: 10.1158/0008-5472.CAN-15-3190

Integrated Systems and Technologies: *Mathematical Oncology*

Cancer Research

Exome-Scale Discovery of Hotspot Mutation Regions in Human Cancer Using 3D Protein Structure

Collin Tokheim[1], Rohit Bhattacharya[1], Noushin Niknafs[1], Derek M. Gygax[2], Rick Kim[2], Michael Ryan[2], David L. Masica[1], and Rachel Karchin[1,3]

Abstract

The impact of somatic missense mutation on cancer etiology and progression is often difficult to interpret. One common approach for assessing the contribution of missense mutations in carcinogenesis is to identify genes mutated with statistically nonrandom frequencies. Even given the large number of sequenced cancer samples currently available, this approach remains underpowered to detect drivers, particularly in less studied cancer types. Alternative statistical and bioinformatic approaches are needed. One approach to increase power is to focus on localized regions of increased missense mutation density or hotspot regions, rather than a whole gene or protein domain. Detecting missense mutation hotspot regions in three-dimensional (3D) protein structure may also be beneficial because linear sequence alone does not fully describe the biologically relevant organization of codons. Here, we present a novel and statistically rigorous algorithm for detecting missense mutation hotspot regions in 3D protein structures. We analyzed approximately 3×10^5 mutations from The Cancer Genome Atlas (TCGA) and identified 216 tumor-type–specific hotspot regions. In addition to experimentally determined protein structures, we considered high-quality structural models, which increase genomic coverage from approximately 5,000 to more than 15,000 genes. We provide new evidence that 3D mutation analysis has unique advantages. It enables discovery of hotspot regions in many more genes than previously shown and increases sensitivity to hotspot regions in tumor suppressor genes (TSG). Although hotspot regions have long been known to exist in both TSGs and oncogenes, we provide the first report that they have different characteristic properties in the two types of driver genes. We show how cancer researchers can use our results to link 3D protein structure and the biologic functions of missense mutations in cancer, and to generate testable hypotheses about driver mechanisms. Our results are included in a new interactive website for visualizing protein structures with TCGA mutations and associated hotspot regions. Users can submit new sequence data, facilitating the visualization of mutations in a biologically relevant context. *Cancer Res; 76(13); 3719–31. ©2016 AACR.*

Major Findings

We used The Cancer Genome Atlas mutation data and identified 3D clusters of cancer mutations ("hotspot regions") at amino-acid-residue resolution in 91 genes, of which 56 are known cancer-associated genes. The hotspot regions identified by our method are smaller than a protein domain or protein–protein interface and in many cases can be linked precisely with functional features such as binding sites, active sites, and sites of experimentally characterized mutations. The hotspot regions are shown to be biologically relevant to cancer, and we discovered that there are characteristic differences between regions in the two types of driver genes, oncogenes and tumor suppressor genes (TSG). These differences include region size, mutational diversity, evolutionary conservation, and amino acid residue physiochemistry. For the first time, we quantify why the great majority of well-known hotspot regions occur in oncogenes. Because hotspot regions in TSGs are larger, more heterogeneous than those in oncogenes, they are more difficult to detect using protein sequence alone and are likely to be underreported. Our results indicate that protein structure–based 3D mutation clustering increases power to find hotspot regions, particularly in TSGs.

Introduction

Missense mutations are perhaps the most difficult mutation type to interpret in human cancers. Truncating loss-of-function mutations and structural rearrangements generate major changes in the protein product of a gene, but a single missense mutation yields only a small change in protein chemistry. The impact of missense mutation on protein function, cellular behavior, cancer etiology, and progression may be negligible or profound, for reasons that are not yet well understood. Missense mutations are frequent in most cancer types, accounting for approximately 85% of the somatic mutations observed in solid human tumors (1), and the cancer genomics

[1]Department of Biomedical Engineering and Institute for Computational Medicine, Johns Hopkins University, Baltimore, Maryland. [2]In Silico Solutions, Fairfax, Virginia. [3]Department of Oncology, Johns Hopkins University School of Medicine, Baltimore, Maryland.

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

Corresponding Author: Rachel Karchin, Johns Hopkins University School of Medicine, 217 A Hackerman Hall, 3400 N. Charles St., Baltimore, MD 21218. Phone: 410-516-5578; Fax: 410-516-5294; E-mail: karchin@jhu.edu

doi: 10.1158/0008-5472.CAN-15-3190

Published OnlineFirst April 28, 2016; DOI: 10.1158/0008-5472.CAN-15-3190

Quick Guide to Equations and Assumptions

An experimentally determined or theoretically modeled protein structure consists of a set of atoms, each with a unique coordinate in three-dimensional (3D) Euclidean space. Each amino acid residue consists of many atoms and may harbor zero, one, or multiple missense mutations in a cohort of sequenced cancer samples. Two key mathematical concepts in our study are the density of local missense mutations in 3D space, which underlies our statistical measure to define missense mutation hotspot regions, and mutational diversity of a hotspot region. Local missense mutation density

$$D_r^k = \sum_{n \in N_r^k} M_n^k$$

is defined for each amino acid residue r and each protein structure k. It considers the sum of the count of missense mutations that occurred at r and those that occurred at residues proximal to r, that is, in its "neighborhood." Proximity is measured in 3D space and the neighborhood is limited to residues up to 1 nm away from r, where 1 nm was chosen because it is the order of magnitude of an amino acid side chain. The term M_n^k is the missense mutation count for the n^{th} residue neighbor of r. The observed value of D_r^k is compared with simulations of its value under an empirical null distribution, where the total number of missense mutations observed in k remains the same, but they are distributed uniformly in 3D. Residue r has significantly increased D_r^k if its adjusted P value is ≤ 0.01 after multiple testing correction. A 3D hotspot region is a grouping of residues with significantly increased D_r^k that are linked as connected components in a neighbor graph. Our algorithm can find 3D hotspot regions directly on protein complexes, enabling detection of hotspot regions that occur on both sides of a protein–protein interface. It also handles complexes with multiple chains originating from a single gene product (e.g., a homodimer) by running identical simulations simultaneously.

Mutational diversity is computed for each hotspot region s_i^g (where i indexes the region and g indexes the gene) based on the Shannon entropy of the joint probability of a missense mutation occurring at a specific residue r and having a specific mutant amino acid m

$$H\left(R^{s_i^g}, M^{s_i^g}\right) = -\sum_{r \in R^{s_i^g}} \sum_{m \in M_r^{s_i^g}} P\left(R^{s_i^g} = r, M^{s_i^g} = m\right) \log_2 P\left(R^{s_i^g} = r, M^{s_i^g} = m\right)$$

Because the maximum possible Shannon entropy grows with the number of residues in a hotspot region, the score is normalized so hotspot regions of different sizes can be compared.

$$MD\left(s_i^g\right) = \frac{H\left(R^{s_i^g}, M^{s_i^g}\right)}{H_{max}(N, R, A)}$$

N is the number of mutations in the hotspot region, R is the number of residues, and A is the number of possible alternate amino acids per residue. In this work, mutational diversity is found to be significantly different between hotspot regions that occur in oncogenes versus those that occur in tumor suppressor genes.

Major assumptions of the model:

- In the absence of selection for drivers, somatic missense mutations in cancers are equally likely to appear at any amino acid residue position in a protein structure of interest.
- Many driving missense mutations have significantly increased local mutation density.
- Residues with significantly increased mutation density and proximal to each other in three dimensions are likely to be subject to similar selective pressures and can be grouped together into hotspot regions.
- The most parsimonious number of hotspot regions in a protein structure is preferred.
- Carefully filtered theoretical protein structure models are accurate enough to capture local missense mutation densities and groupings of proximal residues with significantly increased densities.

community has prioritized the task of identifying important missense mutations discovered in sequencing studies. Whole-exome sequencing (WES) studies of cancer have created new opportunities to better understand the importance of missense mutations. This enormous collection of data now allows detection of patterns with power that was unheard of a few years ago.

The first approaches to identify cancer drivers from WES mutations looked for significantly mutated genes (SMG), harboring a larger number of somatic mutations than expected by chance (2–5). Metrics to call SMGs now appear to be underpowered given the size of current cohorts in The Cancer Genome Atlas (TCGA) and International Cancer Genome Consortium (ICGC). A recent study suggested that approximately 1,500 cases of endometrial cancer would need to be sequenced to attain 90% power to detect mutations in 90% of genes with a mutation frequency of 2% with the SMG approach (5). The recognition of the limitations of the SMG paradigm has motivated interest in orthogonal analysis techniques to detect mutational patterns associated with drivers (1, 6–9).

Published OnlineFirst April 28, 2016; DOI: 10.1158/0008-5472.CAN-15-3190

Recurrence of somatic missense mutations in cancers at the same amino acid residue position is well known to be a characteristic feature of both oncogenes (OG) and tumor suppressor genes (TSG; ref. 10). The observation that somatic mutations also frequently occur in positions proximal in protein sequence to the most highly recurrent positions has suggested that positional clustering of somatic missense mutations might be used to identify drivers (7). These clusters, known as "hotspots," are regions where somatic missense mutations occur closer together in protein sequence than would be expected by chance. Hotspot regions can be rationalized as areas in a protein under positive selection in the cancer environment; missense mutations occurring in these regions are selected for because they alter protein function in a manner advantageous to the cancer cell. Several algorithms have been developed to identify protein functional domains and genes in which these regions are enriched (8, 11), and to identify specific missense mutations in hotspots (7, 12). These algorithms consider mutations in a coordinate system based on one-dimensional (1D) protein sequence. WES studies of cancer cohorts are increasingly incorporating missense "hotspot" detection as a routine analysis step in the search for new drivers.

Finding missense mutation hotspot regions in 1D is limited by the fact that functional proteins tend to fold into three-dimensional (3D) structures (with the exception of intrinsically disordered regions). Thus, positional clustering done in 1D will likely miss many hotspots that are present in 3D after folding. Gene- and protein domain-level testing may indicate the possibility of a 3D hotspot but cannot identify the specific positions in the hotspot. An algorithm that leverages 3D protein structure information, but still performs clustering in 1D through a dimensionality reduction step, has shown utility in detecting OGs (9). A recent study of an aggregated collection of TCGA cancer mutations from 21 tumor types presented an algorithm to identify cancer genes based on 3D clustering of somatic missense mutations, yielding ten such genes. They reported low correlation between 3D and 1D hotspot regions (13).

Here, we present HotMAPS (Hotspot Missense mutation Areas in Protein Structure), a new, sensitive algorithm and a web-based community resource for high-throughput analysis of cancer missense mutation 3D hotspot regions. HotMAPS finds clusters of amino acid residues with significantly increased local mutation density in 3D protein space, compared with an empirical null distribution. The statistical model is designed to handle higher-order protein complexes and can capture regions that span protein–protein interfaces. We apply HotMAPS to missense mutations from 23 tumor types sequenced by TCGA. By careful use of both experimentally derived protein biologic assemblies in the Protein Data Bank (PDB) and theoretical protein structure models, we substantially increase the number of amino acids that can be mapped into 3D protein space and the number of detectable hotspot regions (13).

HotMAPS systematically delineates 3D hotspot regions on the level of amino acid positions, and we provide a detailed catalog of 216 tumor-type–specific regions. We show how the catalog can be used as a discovery tool so that the links between 3D protein structure and the biologic functions of missense mutations in cancer can be better utilized by the community. The catalog provides comprehensive identification of hotspot regions that overlap with many key biologic features of proteins available in the literature (e.g., residue positions at active sites, small-molecule and metal-binding sites, protein interfaces, positions with published experimental mutagenesis results). This information can potentially provide a researcher with more fine-grained mechanistic understanding of missense mutation cancer relevance than is possible by 1D clustering or domain and gene enrichment approaches. Using the catalog, we were able for the first time to systematically analyze characteristic properties of 3D hotspot regions and differences between 3D hotspot regions in OGs and TSGs.

Materials and Methods

TCGA mutation collection

TCGA mutation annotation format (MAF) file data for 23 tumor types was downloaded from Xena data store (https://genome-cancer.soe.ucsc.edu/proj/site/xena/hub/) using their API.

3D protein structure and theoretical model collection and processing

PDB structures were obtained from the Worldwide Protein Data Bank (10/17/2015). Only structures solved by x-ray crystallography and containing at least one human protein chain were used. Single-domain, theoretical protein structure models constructed on the basis of homology to nonhuman proteins were included to increase coverage over a greater proportion of genes. Theoretical models were obtained from the ModPipe human 2013 dataset (ftp://salilab.org/databases/modbase/projects/genomes/H_sapiens/2013/), built with Modeller 9.11 (14). In addition to criteria required by ModPipe, we filtered the theoretical models to increase the quality of structures used in our assessment based on minimum length, target-template sequence identity, loop content, and radius of gyration (Supplementary Materials and Methods).

Models were assessed by comparing 3D hotspot regions identified by HotMAPS in experimental structures with those identified in theoretical models of the same protein. First, we found all pairs of experimental structures and theoretical models of the same protein, in which there was overlap of the same amino acid residues. The agreement of a structure/model pair was the overlap of their hotspot region–mutated residues. A false-positive error was called when a model had a mutated residue in a hotspot region that was not in a hotspot region for any protein structure that it had been paired with. A false-negative error was called when a structure had a mutated residue in a hotspot region that was not in a hotspot region for any of the models it had been paired with.

HotMAPS algorithm

HotMAPS identifies residue positions with higher local mutation density in each protein structure or model than expected from an empirical null distribution, based on simulations of a discrete uniform distribution. Residues are considered significant for increased local mutation density at FDR threshold of 0.01, after correction for multiple testing (Benjamini Hochberg). Three-dimensional missense mutation hotspots are identified as groupings of significant residues according to the principle of maximum parsimony, based on connected components in a neighbor graph. Construction of the neighbor graph and connected components are illustrated in

Published OnlineFirst April 28, 2016; DOI: 10.1158/0008-5472.CAN-15-3190

Supplementary Fig. S1. HotMAPS is designed to run on both single-chain protein structures and biologic assemblies with multiple chains originating from the same gene. Mathematical details are provided in Supplementary Materials and Methods.

Results

3D missense mutation hotspot regions identified in TCGA whole-exome sequencing

Mutation hotspot regions detectable in 3D. Applying HotMAPS to 19,368 PDB protein structures (PDB bioassemblies in which *in vivo* protein structure is represented) and 46,004 theoretical models, we identified 107 unique 3D mutation hotspot regions (aggregated across tumor types), of which 30 were only detectable by clustering in 3D (Supplementary Table S1). When stratified by tumor type, 216 3D missense mutation hotspot regions were found in 19 of the 23 TCGA tumor types, with none in adrenocortical carcinoma (ACC), kidney renal papillary cell carcinoma (KIRP), liver hepatocellular carcinoma (LIHC), or kidney chromophobe (KICH; Supplementary Table S2). KICH is known to be driven by alterations other than point mutation, such as structural breakpoints in the *TERT* promoter. Among all 23 tumor types, sample and mutation count is lower for these four tumor types ($P = 0.02$ for sample count, $P = 0.04$ for mutation count; Wilcoxon rank-sum test), suggesting that at least for tumor types driven by missense mutations, larger sample size might increase our power to find more 3D regions. Our approach enabled us to consider the 3D protein environment of a much higher fraction of TCGA mutations than has been described previously. We were able to map and analyze approximately 53% of the missense mutations in 23 TCGA tumor types (Supplementary Table S3). Of these missense mutations, approximately 10% could be mapped to PDB protein structures and an additional 42% mapped to theoretical models, in the absence of PDB structure. Using hotspot regions identified in the PDB structures as a control, we estimate that the hotspots called in the models have a false-positive rate of 0.058 and a false-negative rate of 0.138. Therefore, very few hotspot regions found in the models are the result of modeling errors, justifying the increase in mutation coverage obtained.

Genes that harbor 3D mutation hotspot regions. When hotspot regions were stratified by tumor type, 91 genes contained at least one hotspot region in at least one tumor type ($q \leq 0.01$) and 40 of these genes had regions that were only discoverable by consideration in 3D. Of the 91, 19 genes were previously annotated as OGs and 11 genes as TSGs with the 20/20 rule, a ratiometric method based on the proportions of different mutation consequence types observed in a gene (1). Twenty-five of the genes are listed in the Cancer Gene Census (CGC; Table 1; ref. 15). Of the remaining 58 genes, five (*KLF5*, *SMARCA2*, *RASA1*, *TGBFR2*, *KEAP1*) have been identified as candidate TSGs in the literature (16–24), and six (deacetylated *KLF5*, *MAPK1*, *FSIP2*, *RANBP2*, *MTOR*, *CHEK2*) as candidate OGs (16, 25–30). Four of the genes are current or potential drug targets (*SMARCA2*, *HDAC4*, *PARG*, *HLA-A*; refs. 18, 31–33). Two genes (*ERCC2*, *CHEK2*) are involved in hereditary cancer susceptibility when mutated in the germline (1). *GTF2I* is a prognostic biomarker in thymic epithelial tumors (genes with literature support in Table 2, other genes in Supplementary Table S4; ref. 34).

Table 1. Cancer genes with 3D HotMAPS regions identified in TCGA tumor types and in landscapes benchmark or cancer gene census

Gene	Landscapes benchmark	Cancer gene census (CGC)	TCGA tumor type(s)
FGFR3	OG	Dom	BLCA
SF3B1	OG	Dom	BRCA, BLCA
FGFR2	OG	Dom	BRCA, UCEC
KRAS	OG	Dom	CESC, UCS, PAAD, STAD, BLCA, UCEC, LUAD, BRCA
PIK3CA	OG	Dom	ESCA, CESC, UCS, LUSC, GBM, STAD, LGG[a], BLCA, UCEC, PRAD, LUAD, KIRC, BRCA, HNSCC
NFE2L2	OG	Dom	ESCA, HNSCC, BLCA, UCEC, LUSC
IDH1	OG	Dom	GBM, LGG, SKCM
IDH2	OG	Dom	LGG
PTPN11	OG	Dom	LGG
MAP2K1	OG	Dom	LUAD[a], SKCM
GNAS	OG	Dom	PAAD
BRAF	OG	Dom	THCA, GBM, LUAD, SKCM, PRAD[a]
HRAS	OG	Dom	THCA, PCPG, BLCA, HNSCC, LUSC[a]
NRAS	OG	Dom	THCA, SKCM
PPP2R1A	OG	Dom?	UCS, UCEC
SPOP	OG	Rec	PRAD
ERBB2	OG		ESCA[a], BRCA, BLCA
EGFR	OG		GBM, LGG, LUAD
RET	OG		PCPG
PIK3R1	TSG	Rec	BRCA[a], GBM, UCEC[a], LGG[a]
FBXW7	TSG	Rec	CESC[a], UCS, LUSC[a], STAD, BLCA, UCEC, HNSCC
TP53	TSG	Rec	ESCA, UCS, PAAD, LUSC, GBM, STAD, LGG, BLCA, UCEC, PRAD, LUAD, OV, BRCA, HNSCC
CIC	TSG	Rec	LGG
SMARCA4	TSG	Rec	LGG[a]
BCOR	TSG	Rec	UCEC
PTEN	TSG		BRCA, GBM[a], UCEC
CDKN2A	TSG		ESCA[a]
VHL	TSG		KIRC[a]
NOTCH1	TSG		LGG[a]
SMAD4	TSG		STAD[a]
RHOA		Dom	BLCA[a], HNSCC, STAD
RAC1		Dom	HNSCC, SKCM
ERBB3		Dom	STAD

Abbreviations: OG, oncogene; TSG, tumor suppressor gene (landscapes benchmark); Cancer Gene Census; Dom, dominant; Rec, recessive; Dom?, probably dominant; TCGA tumor types, tumor types in which the gene had a significant 3D mutation hotspot region ($q \leq 0.01$). ACC, adrenocortical carcinoma; BLCA, bladder urothelial carcinoma; BRCA, breast invasive carcinoma; CESC, cervical squamous cell carcinoma and endocervical adenocarcinoma; ESCA, esophageal carcinoma; GBM, glioblastoma multiforme; HNSCC, head and neck squamous cell carcinoma; KICH, kidney chromophobe; KIRC, kidney renal clear cell carcinoma; KIRP, kidney renal papillary cell carcinoma; LGG, low-grade glioma; LIHC, liver hepatocellular carcinoma; LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma; OV, ovarian serous cystadenocarcinoma; PAAD, pancreatic adenocarcinoma; PCPG, pheochromocytoma and paraganglioma; PRAD, prostate adenocarcinoma; SKCM, skin cutaneous melanoma; STAD, stomach adenocarcinoma; THCA, thyroid carcinoma; UCEC, uterine corpus endometrial carcinoma; UCS, uterine carcinosarcoma.

[a]At least one 3D hotspot region in the gene/tumor type was not detected with the 1D-only version of the algorithm.

Published OnlineFirst April 28, 2016; DOI: 10.1158/0008-5472.CAN-15-3190

Table 2. Genes with HotMAPS regions identified in TCGA tumor types

Gene	TCGA tumor type(s)	Gene details
AP2B1	ESCA	Involved in FGFR signaling. Knockdown promotes the formation of matrix degrading invadopodia, adhesion structures linked to invasive migration in cancer cells (49).
CAND1	BLCA[a]	Component of many protein complexes involved in proteasome-dependent protein degredation via ubiquitination and neddylation. CAND1 binding to the complexes inactivates ubiquitin ligase activity and may block adaptor and NEDD8 conjugation sites (50). May play a role in PLK4-mediated centriole overduplication and Disrupted in prostate cancer (Korzeniewski 2012).
CHEK2	ESCA, LGG, BLCA, HNSCC, PRAD, LUAD, PCPG, KIRC	Checkpoint kinase involved in DNA damage response signaling. SMG and candidate OG in papillary thyroid carcinoma (PTC) cohort of 296 patients (30). Breast cancer susceptibility gene (inherited germline variants; ref. 1).
CUL1	BLCA	Candidate TSG. SCF complex E3 ubiquitin ligase scaffold protein. Suppressor of centriole multiplication through regulation of PLK4 level (23).
ERCC2	BLCA, LGG[a]	DNA-repair (nucleotide excision repair) protein. Significantly mutated in cisplatin-responders vs. non-responders in cohort of 50 patients with muscle-invasive urothelial carcinoma (MIUC). *ERCC2* mutation status may inform cisplatin-containing regimen usage in MIUC (51). Recurrently mutated in cohort of 17 patients with urothelial bladder cancer (UBC; ref. 52). Xeroderma pigmentosum susceptibility gene (inherited germline variants; ref. 1).
FSIP2	ESCA[a]	Candidate OG. Recurrently amplified in testicular germ cell tumors (TGCTs; ref. 27).
GNA13	BLCA	Significantly mutated in cohort of 55 patients with diffuse large B-cell lymphoma (DLBCL; ref. 53).
GTF2I	UCEC	Highly recurrent missense mutation in Thymic epithelial tumors and associated with increased patient survival (34).
HDAC4	ESCA	Histone deaceytlation enzyme. Drug target. Overexpression shown to promote growth of colon cancer cells via p21 repression. Regulator of colon cell proliferation (54). May regulate cancer cell response to hypoxia via its regulates HIF1A acetylation and stability (55)
HLA-A	BLCA, HNSCC, LGG, PRAD	Immune system. Encodes MHC-Class 1A protein, which presents antigens for T-cell recognition. Somatic mutations previously suggested contributing to tumor immune escape (33).
KEAP1	LUAD[a]	Candidate TSG. Inhibits NRF2 (aka NFE2L2). In cohort of 76 non-small cell lung cancer (NSCLC) patients, *KEAP1* found mutated in two patients with advanced adenocarcinoma and smoking history. *KEAP1* mutation was mutually exclusive of *EGFR*, *KRAS*, *ERBB2*, and *NFE2L2* mutation in the cohort and *KEAP1* mutation status proposed as marker for personalized therapy selection (22). Proposed TSG in lung squamous cell carcinomas (24). Proposed as therapeutic target for thyroid-transcription-factor-1 (*TTF1*)-negative lung adenocarcinoma (LUAD; 56).
KLF5	BLCA[a]	Transcription factor that promotes breast cancer cell proliferation, survival, migration and tumour growth. Upregulates *TNFAIP2*, which interacts with the two small GTPases RAC1 and CDC42, thereby increasing their activities to change actin cytoskeleton and cell morphology (57). Proposed as playing dual role as both TSG when acetylated and OG when deacetylated in prostate cancer (16). Recurrently mutated in mucinous ovarian carcinoma (58).
MAPK1	CESC[a], HNSCC	Kinase involved in cell proliferation, differentiation, transcription regulation, and development; key signaling component of the toll-like receptor pathway. Candidate OG in pancreatic cancer (25), laryngeal squamous cell carcinoma cell lines (26). Significantly mutated in cohort of 91 chronic lymphocytic leukemia CLL patients (59).
MSN	ESCA[a]	Protein homolog of TSG *NF2* (Merlin; ref. 60). Member of the Ezrin-Radixin-Moesin (ERM) protein family. Links membrane and cytoskeleton involved in contact-dependent regulation of EGFR (61). Regulates the motility of oral cancer cells via MT1-MMP and E-cadherin/p120-catenin adhesion complex. Cytoplasmic expression of MSN correlates with nodal metastasis and poor prognosis of oral squamous cell carcinomas (OSCC), may be potential candidate for targeted gene therapy for OSCCs (62).
MTOR	KIRC	Candidate OG. Serine/threonine protein kinase regulates cell growth, proliferation and survival. Frequently activated in human cancer and a major therapeutic target. Randomly selected mutants in HEAT repeats and kinase domain induced transformation in NIH3T3 cells and rapid tumor growth in nude mice (29).
NBPF10	BLCA[a]	Somatic missense mutation reported in prostate cancer cohort of 141 patients (63). In gene family with numerous tandem repeats and pseudogenes, possible read alignment and mutation calling errors.
PARG	GBM, LGG, BLCA, HNSCC, PRAD, LUAD, PCPG, KIRC[a]	Involved in DNA damage repair (with *PARP1*). Cells deficient in these proteins are sensitive to lethal effects of ionizing radiation and alkylating agents. Potential drug target for *BRCA2*-deficient cancers (32).
RANBP2	ESCA	Candidate OG (28). A large multimodular and pleiotropic protein with SUMO E3 ligase function (64). Interacts with MTOR (to regulate cell growth and proliferation via cellular anabolic processes; ref. 65). Hotspot mutation previously found in MSI colorectal cancer. Hotspot suggested as useful for personalized tumor profling and therapy in colorectal cancer (28).
RASA1	HNSCC[a]	Identified as TSG in another squamous cell cancer, cutaneous squamous cell skin cancer (cSCC; ref. 19).
RGPD3	BLCA[a], UCEC, PAAD	Component of ubiqutin E3 ligase complex. Named for similarity to RANBP2.
SIRPB1	HNSCC, PRAD	Ig-like cell-surface receptor. Negatively regulates RTK processes. Related to FGFR signaling.
SMARCA2	BLCA[a]	Actin-dependent regulator of chromatin. Its ATPase domain named as Drug target in SWI/SNF–mutant cancers (e.g., lung, synovial sarcoma, leukemia, and rhabdoid tumors; ref. 66). Proposed TSG, and synthetic lethal target in *SMARCA4* (aka *BRG1*)-deficient cancers (18).
TGFBR2	HNSCC	TSG in HNSCC (20) MSI CRC (21), epithelial transformation and invasive squamous cell carcinoma in the mouse forestomach (46).

NOTE: Genes that are candidate OGs and TSGs, hereditary cancer genes, associated with cancer phenotypes and drug targets.
[a]At least one 3D hotspot region in the gene/tumor type was not detected with the 1D-only version of the algorithm.

Published OnlineFirst April 28, 2016; DOI: 10.1158/0008-5472.CAN-15-3190

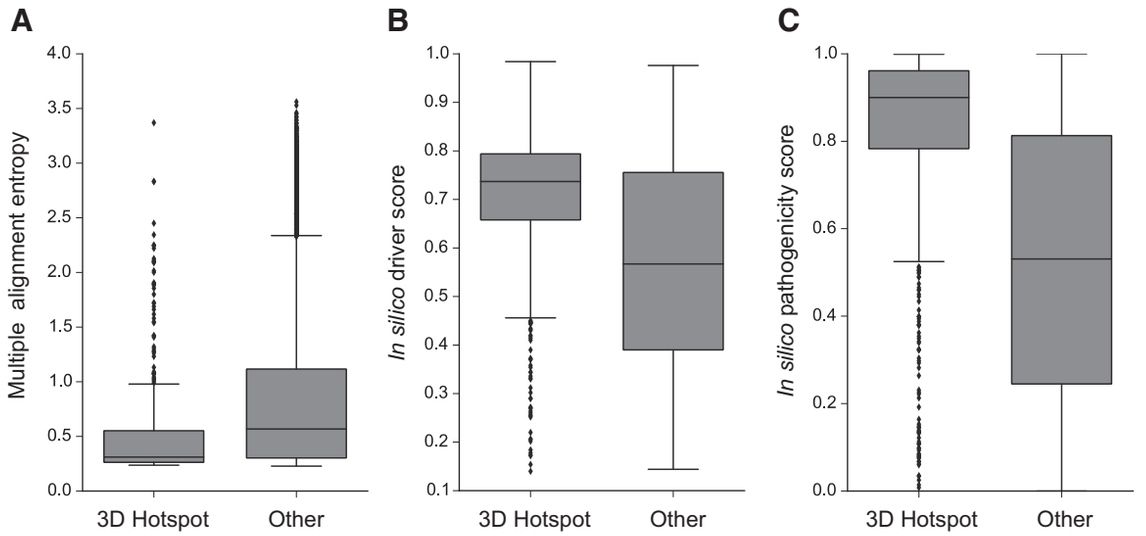

Figure 1.
3D hotspot regions are different from other mutated protein residues. Three distinguishing features of HotMAPS regions. A, HotMAPS-mutated residues are more conserved in vertebrate evolution than mutated residues not in hotspot regions, as shown by lower multiple alignment entropy ($P = 1.2E{-}29$; Mann–Whitney *U* test). Multiple alignment entropy is calculated as the Shannon entropy of protein-translated 46-way vertebrate genome alignments from UCSC Genome Browser, which is lowest for the most conserved residues. B and C, HotMAPS missense mutations have higher *in silico* cancer driver scores from the CHASM algorithm ($P = 5.3E{-}47$; Mann–Whitney *U* test) than those mutations not in hotspot regions (B) and higher *in silico* pathogenicity scores from the VEST algorithm ($P = 7.0E{-}162$; Mann–Whitney *U* test; C). Finally, HotMAPS-mutated residues occur more frequently at protein–protein interfaces ($P = 1.3E{-}11$; one-tailed Fisher exact test; Supplementary Table S8).

3D mutation hotspot regions are important in cancer

3D hotspot regions are enriched in well-known cancer genes. Among the set of genes with available protein structure or models ($n = 15{,}697$), the genes harboring a 3D hotspot region are enriched for OGs and TSGs ($P = 6.1E{-}30$ for OGs and $P = 2.4E{-}13$ for TSGs; one-tailed Fisher exact test). They are also enriched for genes in the CGC list ($P = 1.4E{-}30$; one-tailed Fisher exact test). The subset of these genes harboring only a 3D hotspot region not detectable in 1D is also significantly enriched ($P = 4.3E{-}09$ for OGs, $P = 7.9E{-}12$ for TSGs, $P = 8.0E{-}11$ for CGC genes; one-tailed Fisher exact test). An additional 23 genes that are proposed OGs, TSGs, and/or drug targets or hereditary cancer genes contained at least one 3D hotspot region. This enrichment of known and candidate driver genes supports our claim that many of the regions are biologically relevant and not simply artifacts. While regions were detected in only approximately 18% of established cancer genes, we expect that many of these genes harbor drivers other than missense mutations, some are drivers in tumor types not represented in our study and many lack structural coverage.

Mutations in 3D hotspot regions are different from other somatic mutations in cancers. We examined whether the amino acid residue positions and the missense mutations in the 3D hotspot regions had distinctive features suggestive of a special biologic importance, when compared with the remaining mutations in our study. Four candidate distinguishing features were tested: (i) vertebrate evolutionary conservation; (ii) occurrence at a protein–protein interface, which increases the potential for a missense mutation to disrupt protein–protein interactions; (iii) *in silico* cancer driver scores generated with the CHASM algorithm (6); and (iv) *in silico* pathogenicity scores generated with the VEST algorithm (35), which are predictors of increased missense mutation impact (Fig. 1). In comparison with mutated residues not in 3D hotspot regions, vertebrate evolutionary conservation was higher and protein–protein interface occurrence was higher in the 3D hotspot regions (conservation $P = 2.9E{-}29$, Mann–Whitney *U* test; protein interface $P = 5.2E{-}13$, one-tailed Fisher exact test). *In silico* driver scores and pathogenicity scores were higher for missense mutations in 3D hotspot regions (driver score $P = 3.0E{-}47$, pathogenicity score $P = 3.0E{-}16$; Mann–Whitney U-test) than for the remaining mutations (Fig. 1).

3D hotspot regions are different in OGs and TSGs. The catalog contains 37 regions stratified by tumor type in bonafide TSGs and 77 in bonafide OGs (114 regions in 30 genes), using as a benchmark the classifications of Vogelstein and colleagues (landscapes benchmark; ref. 1). We used these data to explore possible differences between TSG and OG regions at amino acid resolution. We found that in TSGs, 3D hotspot regions were larger than in OGs (region size $P = 9.6E{-}06$; Mann–Whitney *U* test). They were also more mutationally diverse (mutational diversity $P = 2.1E{-}07$; Mann–Whitney *U* test). In addition, OG 3D hotspot regions were more conserved in vertebrate evolution than TSGs and more solvent accessible in protein structure, meaning that they tend to occur at the protein surface (evolution $P = 4.7E{-}07$, solvent accessible $P = 1.5E{-}06$; Mann–Whitney *U* test). TSG hotspot regions harbored increased mutation net change in hydrophobicity ($P = 3.3E{-}07$; Mann–Whitney *U* test) and mutation net change in volume ($P = 2.2E{-}07$; Mann–Whitney *U* test), suggesting that their impact on protein function could be due to decreased stability. The *in silico* missense mutation cancer driver scores

Published OnlineFirst April 28, 2016; DOI: 10.1158/0008-5472.CAN-15-3190

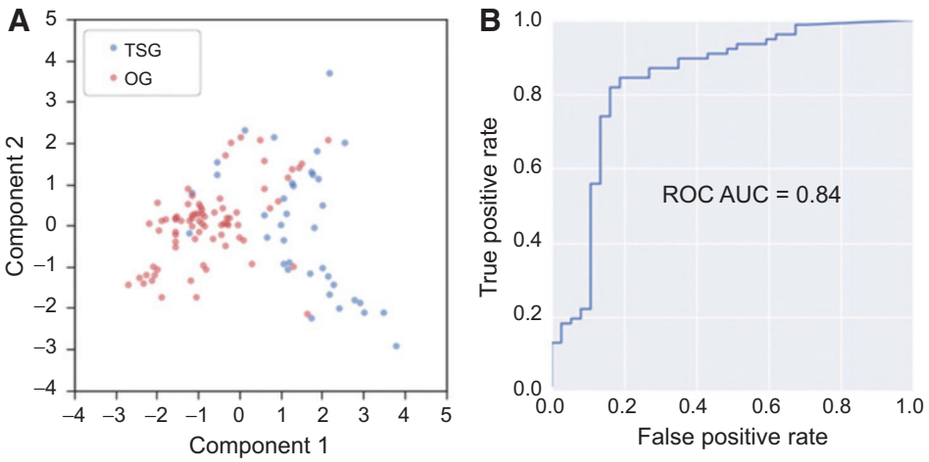

Figure 2.
HotMAPS regions have different characteristic features in OGs and TSGs. A, PCA plot shows a clustering pattern in hotspot regions identified in OGs (red) and TSGs (blue). Each point is a region represented by six numeric features, projected into two dimensions. The features are region size, mutational diversity, vertebrate evolutionary conservation, residue relative solvent accessibility, mutation net change in hydrophobicity, and mutation net change in residue volume. B, OG and TSG HotMAPS regions can be discriminated with machine learning, based on six features. A Gaussian Naive Bayes classifier trained with the landscapes benchmark provides a reasonable separation between the two classes with AUC = 0.84 out of 1.0. Performance of a random classifier is AUC = 0.5. ROC, receiver operating characteristic; AUC, area under the ROC curve.

were higher for OG regions ($P = 0.003$; Mann–Whitney U test). We also tested differences between *in silico* pathogenicity scores and occurrence at protein–protein interfaces between OG and TSG regions, but these were not significant (pathogenicity scores $P = 0.37$, protein interface $P = 0.34$; Mann–Whitney U test).

The fact that these differences between OG and TSG regions were statistically significant suggested that they might have predictive value. Principal components analysis (PCA) of the six significant features indicated some separation (Fig. 2A). Next, we trained a Naive Bayes machine learning classifier to discriminate between OG and TSG hotspot regions, using region size, mutational diversity, vertebrate conservation, residue solvent accessibility, mutation net hydrophobicity change, and residue volume change as features. A rigorous gene-level holdout protocol was used to avoid overfitting (Supplementary Materials and Methods). A Naive Bayes score closer to 1.0 indicates that the hotspot region is likely in an OG while a score closer to 0.0 indicates that it is in a TSG. Area under receiver operating characteristic (ROC) curve or *AUC*, a standard measure of classifier performance, was 0.84 out of 1.0, a result that supports our claim that 3D hotspot regions in OGs and TSGs have distinctive characteristics (Fig. 2B). AUC of a classifier with random performance is 0.5. Performance did not improve when the other features were included in the classifier. Table 3 lists the 30 genes and the median Naive Bayes score, across all regions in each gene. The median values for each predictive feature are also shown.

The ROC performance and PCA plot support our claim that characteristic differences between OG and TSG hotspots can be quantified. However, some hotspot regions remain misclassified, according to their labels in the landscapes benchmark (Discussion).

What is gained by 3D hotspot region detection versus 1D? The larger size and mutational diversity of hotspot regions in TSGs versus OGs suggest that they could be more difficult to detect and perhaps they have been underreported by 1D approaches. OG hotspot regions consisting of recurrent missense mutations at one or two residues can be seen by eye with lollipop plots and are straightforward to detect computationally based on 1D primary sequence. We hypothesized that detection of many TSG hotspot regions might require a 3D algorithm. To maximize the interpretability of this analysis, regions that occurred in multiple tumor types were merged so that each region was represented only once in each gene (Materials and Methods).

For a well-controlled comparison of 3D and 1D hotspot region detection, we applied a 1D version of our method to the protein chain sequences of the same set of PDB protein bioassemblies and theoretical protein structure models to detect nonuniform clustering patterns on primary protein sequence (Supplementary Materials and Methods, Supplementary Tables S5 and S6). Seventy-two percent of hotspot regions identified in 3D were identifiable in 1D.

Next, we compared the number of hotspot regions identified in OGs and TSGs. We considered regions identified in 3D only, in both 3D and 1D, and in 1D only. Using the bona fide OGs and TSGs (Table 1), there were significantly more OG regions that TSG regions identified by the 1D algorithm ($P = 0.03$; one-sided Fisher exact test). The 1D-only version of the algorithm detected 5 OG and 2 TSG regions; 1D further detected an additional 25 OG and 7 TSG regions that were also identified by the 3D algorithm. The 3D algorithm identified an additional 4 OG and 6 TSG regions. To increase our power, we repeated this test again using the bona fide OGs and TSGs plus additional regions in five candidate OGs and TSGs reported in the literature (OGs were *FSIP2*, *MTOR*, *RANBP2*, *CHEK2*, and *MAPK1*; TSGs were *RASA1*, *SMARCA2*, *KEAP1*, *CUL1*, *TGFBR2*; all are listed and cited in Table 2), yielding increased statistical significance ($P = 0.009$, one-sided Fisher exact test). The results suggest that 1D detection methods may be better suited to detecting regions in OGs rather than TSGs.

Published OnlineFirst April 28, 2016; DOI: 10.1158/0008-5472.CAN-15-3190

Table 3. Median scores and feature values for oncogene and tumor suppressor gene hotspot regions

Gene	Landscapes benchmark	Naive Bayes score	Region size	Mutational diversity	Conservation	Change in hydrophobicity	Residue solvent accessibility	Change in volume
FGFR3	OG	1.00	1.00	0.00	0.29	−3.00	0.61	−1.80
KRAS	OG	1.00	2.50	0.58	0.26	4.49	0.73	−1.90
NRAS	OG	1.00	2.50	0.37	0.25	6.24	0.49	−0.98
BCOR	TSG	1.00	1.00	0.00	0.32	−5.40	0.55	0.84
PIK3R1	TSG	1.00	1.00	0.00	0.26	4.20	0.08	−0.70
BRAF	OG	1.00	4.00	0.25	0.25	8.70	0.42	−0.07
HRAS	OG	1.00	2.00	0.62	0.32	4.72	0.43	−1.27
PTPN11	OG	1.00	1.00	0.00	0.41	10.80	0.25	−0.75
FGFR2	OG	1.00	1.50	0.13	0.41	8.91	0.34	−1.81
PPP2R1A	OG	0.99	3.00	0.43	0.29	4.82	0.21	−1.84
PIK3CA	OG	0.99	2.00	0.35	0.26	0.63	0.32	−0.81
MAP2K1	OG	0.97	2.00	0.71	0.25	−1.02	0.29	−0.49
NFE2L2	OG	0.97	3.00	0.69	0.27	−1.49	0.30	−0.34
SF3B1	OG	0.96	3.00	0.30	0.25	0.01	0.23	0.29
ERBB2	OG	0.93	1.50	0.41	0.31	−0.95	0.20	−1.27
RET	OG	0.92	1.00	0.00	0.26	2.20	0.10	1.56
GNAS	OG	0.84	1.00	0.33	0.29	−10.73	0.38	1.10
CIC	TSG	0.79	6.00	0.55	0.41	−8.41	0.39	−0.68
SMAD4	TSG	0.69	5.00	0.92	0.30	−5.15	0.16	−0.34
FBXW7	TSG	0.61	5.00	0.72	0.28	−10.28	0.27	1.96
PTEN	TSG	0.61	6.00	1.00	0.26	−1.33	0.07	0.80
IDH2	OG	0.60	1.00	0.67	0.28	−7.85	0.10	0.49
SMARCA4	TSG	0.11	1.00	0.00	0.30	−13.30	0.08	3.78
NOTCH1	TSG	0.08	1.00	0.00	0.55	−0.80	0.17	0.79
IDH1	OG	0.07	1.00	0.22	0.41	−9.83	0.16	1.09
SPOP	OG	0.06	7.00	0.73	0.25	1.37	0.09	2.46
CDKN2A	TSG	0.05	2.00	0.58	0.60	−4.27	0.35	−0.57
VHL	TSG	0.04	8.00	0.89	0.40	0.92	0.02	0.87
EGFR	OG	0.01	8.00	0.53	0.34	−1.01	0.30	−0.76
TP53	TSG	0.00	30.50	0.81	0.44	−4.11	0.19	−0.09

NOTE: Thirty genes classified as OGs or TSGs in landscapes benchmark. The genes contain a total of 114 tumor-type–specific hotspot regions. Each row in the table shows the median Naive Bayes classification score for the regions in the gene. Scores close to 1.0 predict that a region is in an OG and scores close to 0.0 predict that region is in a TSG. Genes are ranked in decreasing order by the Naive Bayes scores. Ranking generally agrees with the landscapes benchmark. Also shown for each gene is the median value of the six features used to train the Naive Bayes classifier. The features are region size (number of residues), mutational diversity, vertebrate evolutionary conservation (Shannon entropy of alignment position, where lower entropy = higher conservation), mutation net hydrophobicity change, residue solvent accessibility, and mutation net volume change.

A further problem with sequence-based 1D hotspot region detection is that larger regions detectable in 3D may be only partially characterized and/or split into multiple pieces. Figure 3 shows an example of a TSG hotspot region in FBXW7 found in 3D by HotMAPS that has been split into two pieces by the 1D algorithm. In 1D protein sequence, residue 465 is not close enough to residues 502 and 505 to be identified in one hotspot region. On the 3D protein structure of FBXW7 (PDB code 2OVQ), the three residues are spatially close and a single hotspot region is detected.

3D hotspot regions may increase interpretability of driver mechanisms.

Three-dimensional consideration of hotspot regions in protein structure can potentially provide researchers with a rich source of hypothesis generation about driver mechanisms. While gene- or domain-level mutation enrichment analysis can point to potential protein functions, interactions, biologic processes, and pathways important for cancer etiology and progression, more detailed information may be available once a specific set of mutated amino acid residues has been identified as significant.

For many of the 3D hotspot regions found by HotMAPS, the literature contains evidence that they are in direct contact with or proximal to amino acid residues of known functional importance. Figure 4 shows six cancer-associated proteins in which the hotspot region is either overlapping or proximal to important functional sites.

RAC1 hotspot in squamous head and neck cancer. RAC1 is a Rho GTPase important in signaling systems that regulate the organization of actin cytoskeleton and cell motility. The hotspot overlaps the GTP/GDP-binding site and could impact regulation of normal RAC1 cycling between GTP- and GDP-bound states (Fig. 4A). It contains a previously identified recurrent mutation in melanoma (P29S), which dysregulates RAC1 by a fast cycling mechanism (36).

SPOP hotspot in prostate cancer (PRAD). SPOP is the substrate recognition component of a cullin3-based E3 ubiquitin–protein ligase complex, which targets multiple substrates for proteasomal degradation. The hotspot overlaps with a binding groove harboring five residue positions (pink) where mutagenesis has strongly reduced affinity for the substrate (annotated in the UniProtKB).

ERCC2 hotspot in bladder cancer. ERCC2 is an ATP-dependent helicase that is part of the protein complex TFIIH involved in RNA polymerase II transcription and nucleotide excision repair (NER). We identified a hotspot region, proximal to the DEAH box, a highly conserved motif containing residues that interact

Published OnlineFirst April 28, 2016; DOI: 10.1158/0008-5472.CAN-15-3190

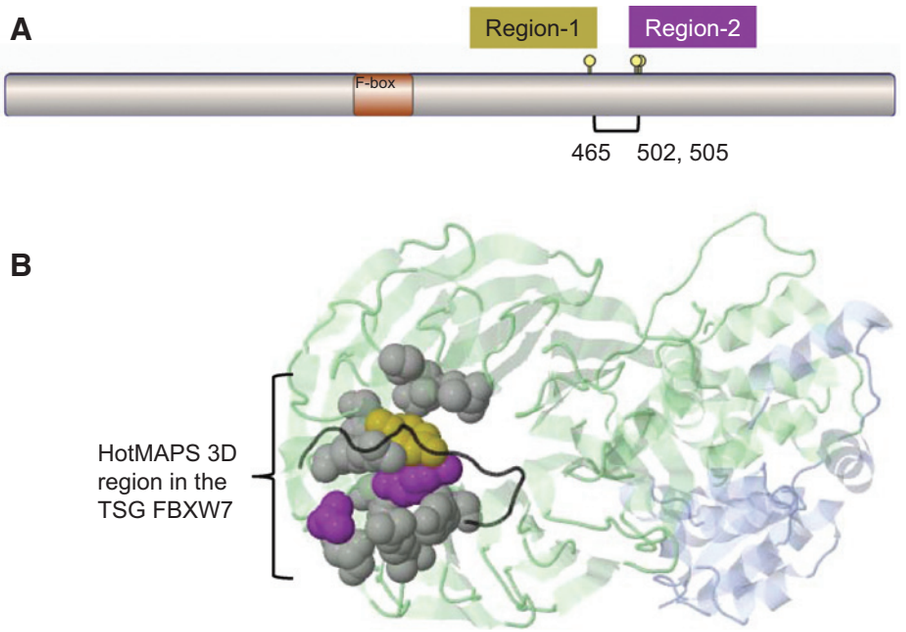

Figure 3.
Comparison of hotspot detection in the TSG FBXW7 in 1D and 3D. A, a simplified 1D version of HotMAPS found two regions in FBXW7. The 3D version of HotMAPS found a single larger region, encompassing both regions. Diagram shows protein sequence of FBXW7, which contains a single F-box functional domain. Region-1, residue 465 (left lollipop); Region-2, residues 502 and 505 (right lollipops). B, HotMAPS identifies a single 3D hotspot region in FBXW7. Structure of SCF^{Fbw7} ubiquitin ligase complex (PDB 2OVQ), containing FBXW7 (green), SKP1 (blue), and CCNE1 fragment (degron peptide; black). Residue coloring: 1D Region-1, gold; 1D Region-2, purple. Residues missed by 1D detection but included in HotMAPS 3D, gray. Although the 1D regions are far in the primary protein sequence, residues 505 and 465 spatially contact at the interface with CCNE1. Protein structure figures were generated by JSMol in MuPIT (http://mupit.icm.jhu.edu/).

with Mg^{2+} and are critical for ATP binding and helicase activity (Fig. 4C). This proximity suggests that the hotspot mutations could disrupt ATPase activity and yield defective NER (37).

PTEN hotspot. PTEN is a phosphatase for both proteins and phosphoinositides, and it removes a phosphate from PIP3, critical for signaling to AKT. The hotspot region identified in endometrial cancer (UCEC) spans two functionally important loops in the protein (P and WPD loops) at the boundary of the active site pocket. Residues in these loops are critical for catalysis (blue dot) and are important for the P-loop's conformation. Mutagenesis of residues in the WPD loop reduces phosphatase activity and increases colony formation in cell culture (38). Pink dots show residues that impact phosphatase activity.

RHOA hotspots. *RHOA* is a small GTPase oncogene, and like *RAC1* is a member of the Ras superfamily (39). We identified hotspot regions in bladder cancer (BLCA), head and neck squamous cell cancer (HNSCC), and stomach adenocarcinoma (STAD). The hotspot regions overlap with the RHOA effector region, a highly conserved motif that is involved in Ras superfamily signaling with downstream effector proteins. The regions are immediately proximal to a magnesium ion, which has been implicated in regulating the kinetics of Rho family GTPases (40).

VHL hotspot (KIRC). VHL is a component of an E3 ubiquitin protein ligase complex, and it ubiquitinates the OG transcription factor HIF1A, targeting it for proteasomal degradation (41). One impact of *VHL* loss of function with failure to ubiquitinate HIF1A is increased protein expression of HIF1A. The hotspot region is proximal to its interaction site with HIF1A and could potentially have an impact on this interaction (Fig. 4F). The TCGA kidney cancer (KIRC) samples were stratified on the basis of their missense mutation status: *VHL* hotspot, non-hotspot, or no missense (WT). HIF1A protein expression was not significantly different between *VHL* non-hotspot and *VHL* WT groups ($P = 0.5$; Mann–Whitney U test), but was significantly higher between *VHL* hotspot and *VHL* WT groups ($P = 0.03$; Mann–Whitney U test). This result is consistent with a special role for *VHL* hotspot missense mutations in regulating HIF1A protein expression. However, increased HIF1A expression in these KIRC samples is likely impacted by additional genetic and other factors. We might see a substantially lower P value if VHL hotspot mutations were the only cause of the observed increase. Also, there are many VHL missense mutations outside of the hotspot region, and it is likely that several of these also have a functional impact. In particular, several of them are at the interface of VHL and the TCEB1 and TCEB2 in the complex and could impact VHL/TCEB binding.

Discussion

Catalog of TCGA 3D missense mutation hotspot regions at amino acid resolution

The large-scale WES and mutation calling efforts of the TCGA have identified hundreds of thousands of somatic missense mutations in human cancers. Although some of these mutations are private, many are shared across multiple patients and multiple tumor types. The biologic and therapeutic relevance of these shared mutations is of great interest to the cancer research community. For example, patients can be stratified for clinical trials and treatment protocols selected on the basis of missense mutation status in several key driver genes, including *BRAF*, *KRAS*, and *EGFR*. A special type of shared missense mutations are those that occur recurrently not only at the same genomic codon, but at neighboring codons in translated protein sequence and more generally, neighboring amino acid residues in protein 3D structure. These clusters of neighboring missense mutations are known as missense

Published OnlineFirst April 28, 2016; DOI: 10.1158/0008-5472.CAN-15-3190

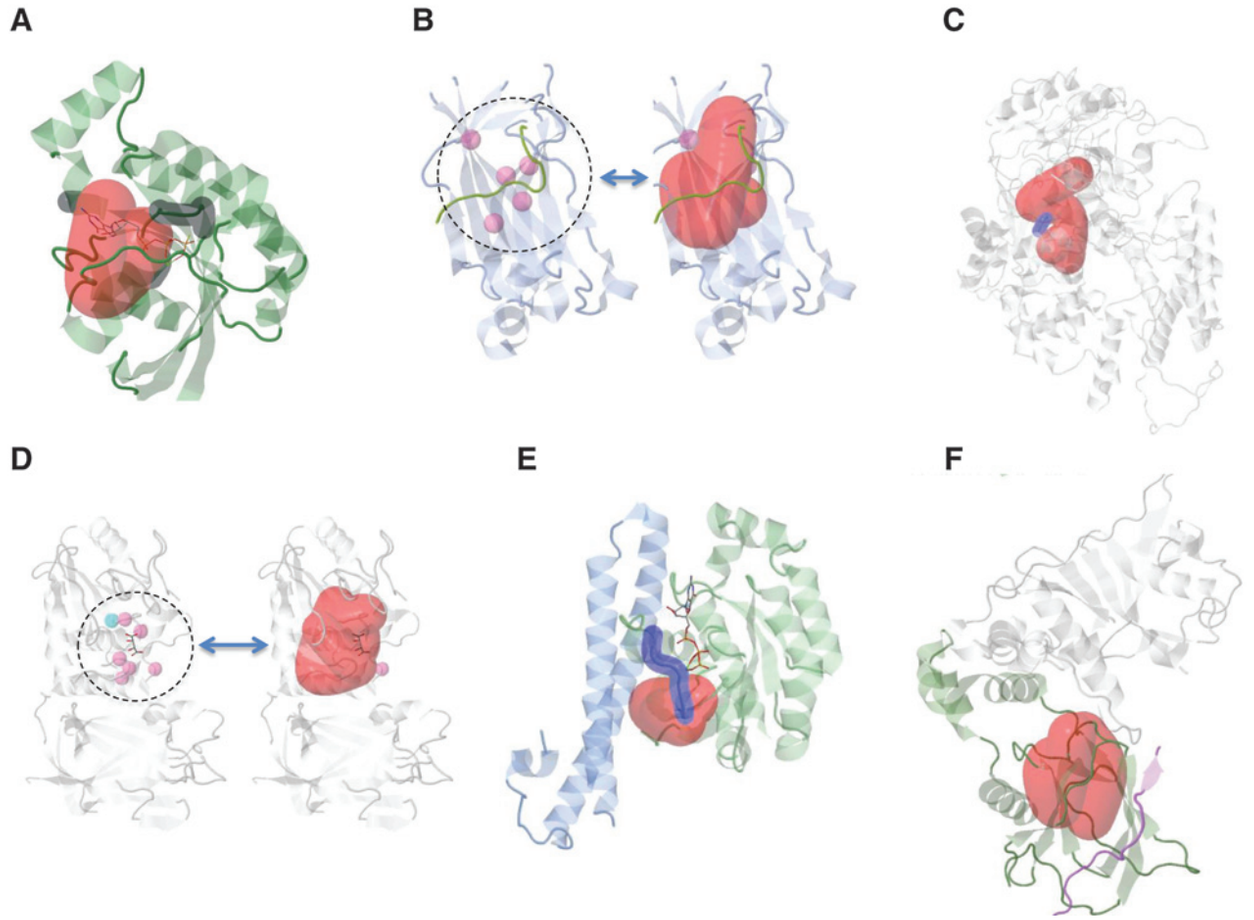

Figure 4.
HotMAPS hotspot regions overlap and are proximal to important functional sites. A, HNSCC hotspot region (red) in RAC1 (green) and GTP/GDP-binding residues (dark gray; PDB 2FJU). B, PRAD hotspot region (red) in SPOP–substrate complex (PDB 3HGH) with SPOP (blue) and H2AFY substrate (green). Left, five residues (pink) that when mutated show strongly reduced affinity for substrate. C, BLCA hotspot region (red) in ERCC2 (gray) shown on theoretical model of ERCC2 helicase ATP-binding domain. The hotspot is proximal to the DEAH box (blue), a highly conserved motif containing residues that interact with $Mg2+$ and are critical for ATP-binding and helicase activity. D, UCEC hotspot region (red) in PTEN (PDB 1D5R) with active site phosphocysteine residue (blue), residues when mutated annotated to reduce phosphatase activity (pink). E, STAD hotspot region (red) in RHOA with a GTP analog bound (sticks; PDB 1CXZ). GTP-binding residues and effector region, dark blue. F, KIRC hotspot region (red) in VHL–TCEB1–TCEB2 complex, bound to HIF1A peptide (PDB 4AJY). Proximity to the interaction site of VHL (green) and HIF1A (blue) suggests possible decreased ubiquitination of HIF1A, resulting in increased protein expression of HIF1A. TCEB1 and TCEB2, gray.

mutation hotspot regions. They have been proposed to have particular relevance to oncogenic processes (12), as the increased frequency of missense mutation at a highly localized region in a protein may be a signature of positive selection (42). Missense hotspot regions may be informative in detecting driver genes (7). A number of groups have developed algorithms to detect enrichment of these regions on the gene- and domain-level (7–9, 11–13), but until now, there have been not been systematically characterized on a large number of protein structures and models at the resolution of individual amino acid residues.

We systematically identify 3D missense hotspot regions using TCGA somatic mutation data from 6,594 samples in 23 tumor types. HotMAPS identified 107 unique, tumor-type–aggregated gene-level regions, and 216 unique tumor-type–specific gene-level regions (Materials and Methods; Supplementary Tables S1 and S2). This catalog enables assessment of how the specific missense mutations in a hotspot contribute to cancer-associated molecular mechanisms.

TCGA 3D hotspot regions have functional importance

We compared features of residues in 3D hotspot regions to other missense mutated residues in the TCGA data. The hotspot regions have characteristic features that support their putative functional importance: high evolutionary conservation, high *in silico*-predicted missense mutation impacts, and increased frequency of occurrence at protein interfaces. Genes containing the 3D hotspot regions appear to be particularly relevant to cancer. Landscapes benchmark TSG and OGs are overrepresented and the list includes many candidate TSGs, OGs, drug targets, and hereditary cancer genes (Tables 1 and 2). For several TSGs and OGs, the regions coincide with enzymatic

Published OnlineFirst April 28, 2016; DOI: 10.1158/0008-5472.CAN-15-3190

active sites, positions that have been shown to alter protein function in experimental mutagenesis assays and sites of interaction with protein and nucleotide interaction partners (Fig. 4; Supplementary Table S7).

TCGA hotspot regions are different in OGs and TSGs

Although recurrent missense mutations have long been known to occur in both OGs and TSGs (10), they have been observed more frequently in OGs. We show that there are systematic differences in hotspot regions found in OGs and TSGs. OG regions are smaller, less mutationally diverse, more evolutionarily conserved, and more solvent accessible than TSG regions. TSG regions are more likely to harbor mutations that may impact protein stability through changes in hydrophobicity or volume. Potential explanations for these differences are that there are more ways to lose the function of a protein than to gain function (43). Loss-of-function tumor suppressor mutations can occur at many residue positions and involve many types of amino acid residue substitutions, while oncogene mutations will occur at a few functionally important positions and involve fewer substitution types.

A consequence of these differences is that TSG regions are harder to detect visually or by 1D clustering approaches than OG regions. Thus, they have been missed by 1D analysis methods. A major contribution of 3D analysis is enabling detection of hotspot regions in TSGs in addition to those in OGs. We suspect that novel hotspot regions in TSGs will continue to be discovered as more samples are sequenced in more tumor types.

We are able to leverage the characteristic differences to distinguish between hotspot regions in TSGs and OGs, with a simple machine learning method, achieving an area under the receiver operating characteristic curve (ROC AUC) of approximately 0.80. However, not all regions are correctly classified by this method. Interestingly, we find that some of these "undistinguishable" genes may act as both TSGs and OGs, depending on context or be atypical of their class. *PIK3R1* has been described as an OG (44) and *SPOP* as a TSG (45), in agreement with our Naive Bayes scores, but not with the landscapes benchmark. The OGs IDH1 and IDH2 both have high net hydrophobicity changes, which are protein destabilizing and characteristic of TSGs. IDH1/IDH2 hotspot mutation may cause a (TSG-like) partial loss of enzymatic function, yielding accumulation of 2 hydoxyglutarate (2HG), a carcinogenic catalytic intermediate (46). *EGFR* has two regions in GBM and LGG, which are scored as TSG-like. In these tumor types, *EGFR* mutation patterns are atypical, because *EGFR* amplification is an early event. This amplification has been linked to increased mutation load in *EGFR* itself, including in aberrant extrachromosomal copies of *EGFR* (43). Supplementary Figure S2 indicates the locations of these misclassified regions on the plot.

HotMAPS has increased sensitivity and coverage than previous 3D hotspot detection algorithms

A disadvantage of working with experimentally derived protein structures is that they are available for a limited number of human proteins (39%). For many of these genes, the structure data are incomplete, so that only a single protein domain or small fragment is represented in PDB. In this work, by careful use of biologic assemblies of PDB structures and also theoretical protein models, we mapped approximately 53% of unique residue positions harboring a TCGA missense mutation into 3D protein space. In a recent study of 21 TCGA tumor types that used a different algorithm and PDB structures only, 11.2% positions were mapped (13). We note that theoretical protein models are well suited for this kind of analysis. HotMAPS considers the center of geometry for each amino acid residue, a metric that is not highly sensitive to atomic-resolution errors common in theoretical protein models (47). The increased sensitivity and coverage of HotMAPS are supported by the number of tumor types in which 3D hotspot regions were detected (19 of 23), the total number of regions detected (107 unique, tumor-type–aggregated gene-level regions and 216 unique tumor-type–specific gene-level regions), and the number of genes in which regions were detected (91). The only previous systematic attempt to find 3D hotspots in TCGA data detected statistically significant regions in 10 genes, based on 21 TCGA tumor types (13), and 9 of these were also detected by HotMAPS.

We hope that some HotMAPS regions found by our algorithm point to novel driver genes; however, functional studies are warranted to find out if they are discoveries or false positives.

An interactive 3D protein viewer where users can submit their own mutations and compare with the HotMAPS catalog (48) is available at http://mupit.icm.jhu.edu/.

HotMAPS software is open source at https://github.com/karchinlab/HotMAPS.

Additional material is available in Supplementary Materials and Methods: detailing mapping from genomic coordinates to protein structures (Supplementary Fig. S3), overview flow chart of HotMAPS (Supplementary Fig. S4), an example of single-residue hotspot region discovery in 1D versus 3D (Supplementary Fig. S5), a stratified analysis of HotMAPS properties by solvent accessibility (Supplementary Fig. S6), and in Supplementary Tables, a list of blacklisted residues (Supplementary Table S9).

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Authors' Contributions

Conception and design: C. Tokheim, D.L. Masica, R. Karchin
Development of methodology: C. Tokheim, R. Bhattacharya, M.C. Ryan
Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): C. Tokheim
Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): C. Tokheim, N. Niknafs, R. Kim, M.C. Ryan, D.L. Masica, R. Karchin
Writing, review, and/or revision of the manuscript: C. Tokheim, R. Bhattacharya, R. Karchin
Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): D.M. Gygax, M.C. Ryan
Study supervision: R. Karchin

Acknowledgments

The authors thank Drs. Bert Vogelstein and Jing Zhu of the UCSC Xena team for valuable discussion on the manuscript.

Grant Support

This work was supported by NIH, National Cancer Institute fellowship F31CA200266 (C. Tokheim) and grant U01CA180956 (R. Karchin).

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received November 30, 2015; revised February 25, 2016; accepted April 1, 2016; published OnlineFirst April 28, 2016.

References

1. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA Jr, Kinzler KW. Cancer genome landscapes. Science 2013;339:1546–58.
2. Wood LD, Parsons DW, Jones S, Lin J, Sjoblom T, Leary RJ, et al. The genomic landscapes of human breast and colorectal cancers. Science 2007; 318:1108–13.
3. Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, et al. Patterns of somatic mutation in human cancer genomes. Nature 2007; 446:153–8.
4. Dees ND, Zhang Q, Kandoth C, Wendl MC, Schierding W, Koboldt DC, et al. MuSiC: identifying mutational significance in cancer genomes. Genome Res 2012;22:1589–98.
5. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 2013;499:214–8.
6. Carter H, Chen S, Isik L, Tyekucheva S, Velculescu VE, Kinzler KW, et al. Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. Cancer Res 2009; 69:6660–67.
7. Gonzalez-Perez A, Perez-Llamas C, Deu-Pons J, Tamborero D, Schroeder MP, Jene-Sanz A, et al. IntOGen-mutations identifies cancer drivers across tumor types. Nat Methods 2013;10:1081–2.
8. Nehrt NL, Peterson TA, Park D, Kann MG. Domain landscapes of somatic mutations in cancer. BMC Genomics 2012;13 Suppl 4:S9.
9. Ryslik GA, Cheng Y, Cheung K-H, Modis Y, Zhao H. Utilizing protein structure to identify non-random somatic mutations. BMC Bioinformatics 2013;14:190.
10. Hollstein M, Sidransky D, Vogelstein B, Harris CC. p53 mutations in human cancers. Science 1991;253:49–53.
11. Tamborero D, Gonzalez-Perez A, Lopez-Bigas N. OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. Bioinformatics 2013;29:2238–44.
12. Ye J, Pavlicek A, Lunney EA, Rejto PA, Teng C-H. Statistical method on nonrandom clustering with application to somatic mutations in cancer. BMC Bioinformatics 2010;11:11.
13. Kamburov A, Lawrence MS, Polak P, Leshchiner I, Lage K, Golub TR, et al. Comprehensive assessment of cancer missense mutation clustering in protein structures. Proc Natl Acad Sci U S A 2015;112:E5486–95.
14. Pieper U, Eswar N, Webb BM, Eramian D, Kelly L, Barkan DT, et al. MODBASE, a database of annotated comparative protein structure models and associated resources. Nucleic Acids Res 2009;37(Database issue): D347–54.
15. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, et al. A census of human cancer genes. Nat Rev Cancer 2004;4:177–83.
16. Atala A. Re: Interruption of KLF5 acetylation converts its function from tumor suppressor to tumor promoter in prostate cancer cells. J Urol 2015;194:1505.
17. Helming KC, Wang X, Roberts CW. Vulnerabilities of mutant SWI/SNF complexes in cancer. Cancer Cell 2014;26:309–17.
18. Hoffman GR, Rahal R, Buxton F, Xiang K, McAllister G, Frias E, et al. Functional epigenetics approach identifies BRM/SMARCA2 as a critical synthetic lethal target in BRG1-deficient cancers. Proc Natl Acad Sci U S A 2014;111:3128–33.
19. Pickering CR, Zhou JH, Lee JJ, Drummond JA, Peng SA, Saade RE, et al. Mutational landscape of aggressive cutaneous squamous cell carcinoma. Clin Cancer Res 2014;20:6582–92.
20. Rothenberg SM, Ellisen LW. The molecular pathogenesis of head and neck squamous cell carcinoma. J Clin Invest 2012;122:1951–7.
21. Biswas S, Trobridge P, Romero-Gallo J, Billheimer D, Myeroff LL, Willson JK, et al. Mutational inactivation of TGFBR2 in microsatellite unstable colon cancer arises from the cooperation of genomic instability and the clonal outgrowth of transforming growth factor beta resistant cells. Genes Chromosomes Cancer 2008;47:95–106.
22. Sasaki H, Suzuki A, Shitara M, Okuda K, Hikosaka Y, Moriyama S, et al. Mutations in lung cancer patients. Oncol Lett 2013;6:719–21.
23. Korzeniewski N, Zheng L, Cuevas R, Parry J, Chatterjee P, Anderton B, et al. Cullin 1 functions as a centrosomal suppressor of centriole multiplication by regulating polo-like kinase 4 protein levels. Cancer Res 2009;69:6668–75.
24. Hast BE, Cloer EW, Goldfarb D, Li H, Siesser PF, Yan F, et al. Cancer-derived mutations in KEAP1 impair NRF2 degradation but not ubiquitination. Cancer Res 2014;74:808–17.
25. Furukawa T, Kanai N, Shiwaku HO, Soga N, Uehara A, Horii A. AURKA is one of the downstream targets of MAPK1/ERK2 in pancreatic cancer. Oncogene 2006;25:4831–9.
26. Kostrzewska-Poczekaj M, Giefing M, Jarmuz M, Brauze D, Pelinska K, Grenman R, et al. Recurrent amplification in the 22q11 region in laryngeal squamous cell carcinoma results in overexpression of the CRKL but not the MAPK1 oncogene. Cancer Biomark 2010;8:11–9.
27. Litchfield K, Summersgill B, Yost S, Sultana R, Labreche K, Dudakia D, et al. Whole-exome sequencing reveals the mutational spectrum of testicular germ cell tumours. Nat Commun 2015;6:5973.
28. Gylfe AE, Kondelin J, Turunen M, Ristolainen H, Katainen R, Pitkanen E, et al. Identification of candidate oncogenes in human colorectal cancers with microsatellite instability. Gastroenterology 2013;145: 540–3e22.
29. Murugan AK, Alzahrani A, Xing M. Mutations in critical domains confer the human mTOR gene strong tumorigenicity. J Biol Chem 2013;288: 6511–21.
30. The Cancer Genome Atlas Research Network. Integrated genomic characterization of papillary thyroid carcinoma. Cell 2014;159:676–90.
31. West AC, Johnstone RW. New and emerging HDAC inhibitors for cancer treatment. J Clin Invest 2014;124:30–9.
32. Fathers C, Drayton RM, Solovieva S, Bryant HE. Inhibition of poly(ADP-ribose) glycohydrolase (PARG) specifically kills BRCA2-deficient tumor cells. Cell Cycle 2012;11:990–7.
33. Shukla SA, Rooney MS, Rajasagi M, Tiao G, Dixon PM, Lawrence MS, et al. Comprehensive analysis of cancer-associated somatic mutations in class I HLA genes. Nat Biotechnol 2015;33:1152–58.
34. Petrini I, Meltzer PS, Kim IK, Lucchi M, Park KS, Fontanini G, et al. A specific missense mutation in GTF2I occurs at high frequency in thymic epithelial tumors. Nat Genet 2014;46:844–9.
35. Douville C, Carter H, Kim R, Niknafs N, Diekhans M, Stenson PD, et al. CRAVAT: cancer-related analysis of variants toolkit. Bioinformatics 2013;29:647–48.
36. Davis MJ, Ha BH, Holman EC, Halaban R, Schlessinger J, Boggon TJ. RAC1P29S is a spontaneously activating cancer-associated GTPase. Proc Natl Acad Sci U S A 2013;110:912–7.
37. Winkler GS, Araujo SJ, Fiedler U, Vermeulen W, Coin F, Egly JM, et al. TFIIH with inactive XPD helicase functions in transcription initiation but is defective in DNA repair. J Biol Chem 2000;275:4258–66.
38. Lee JO, Yang H, Georgescu MM, Di Cristofano A, Maehama T, Shi Y, et al. Crystal structure of the PTEN tumor suppressor: implications for its phosphoinositide phosphatase activity and membrane association. Cell 1999;99:323–34.
39. Rojas AM, Fuentes G, Rausell A, Valencia A. The Ras protein superfamily: evolutionary tree and role of conserved amino acids. J Cell Biol 2012; 196:189–201.
40. Zhang B, Zhang Y, Wang Z, Zheng Y. The role of Mg2+ cofactor in the guanine nucleotide exchange and GTP hydrolysis reactions of Rho family GTP-binding proteins. J Biol Chem 2000;275:25299–307.
41. Gossage L, Eisen T, Maher ER. VHL, the story of a tumour suppressor gene. Nat Rev Cancer 2015;15:55–64.
42. Wagner A. Rapid detection of positive selection in genes and genomes through variation clusters. Genetics 2007;176:2451–63.

43. Nikolaev S, Santoni F, Garieri M, Makrythanasis P, Falconnet E, Guipponi M, et al. Extrachromosomal driver mutations in glioblastoma and low-grade glioma. Nat Commun 2014;5:5690.
44. Philp AJ, Campbell IG, Leet C, Vincan E, Rockman SP, Whitehead RH, et al. The phosphatidylinositol 3′-kinase p85alpha gene is an oncogene in human ovarian and colon tumors. Cancer Res 2001;61:7426–9.
45. Li C, Ao J, Fu J, Lee DF, Xu J, Lonard D, et al. Tumor-suppressor role for the SPOP ubiquitin ligase in signal-dependent proteolysis of the oncogenic co-activator SRC-3/AIB1. Oncogene 2011;30:4350–64.
46. Yang H, Ye D, Guan KL, Xiong Y. IDH1 and IDH2 mutations in tumorigenesis: mechanistic insights and clinical perspectives. Clin Cancer Res 2012;18:5562–71.
47. Baker D, Sali A. Protein structure prediction and structural genomics. Science 2001;294:93–6.
48. Niknafs N, Kim D, Kim R, Diekhans M, Ryan M, Stenson PD, et al. MuPIT interactive: webserver for mapping variant positions to annotated, interactive 3D structures. Hum Genet 2013;132:1235–43.
49. Pignatelli J, Jones MC, LaLonde DP, Turner CE. Beta2-adaptin binds actopaxin and regulates cell spreading, migration and matrix degradation. PLoS One 2012;7:e46228.
50. Bosu DR, Kipreos ET. Cullin-RING ubiquitin ligases: global regulation and activation cycles. Cell Div 2008;3:7.
51. Van Allen EM, Mouw KW, Kim P, Iyer G, Wagle N, Al-Ahmadie H, et al. Somatic ERCC2 mutations correlate with cisplatin sensitivity in muscle-invasive urothelial carcinoma. Cancer Discov 2014;4:1140–53.
52. Balbas-Martinez C, Sagrera A, Carrillo-de-Santa-Pau E, Earl J, Marquez M, Vazquez M, et al. Recurrent inactivation of STAG2 in bladder cancer is not associated with aneuploidy. Nat Genet 2013;45:1464–9.
53. Lohr JG, Stojanov P, Lawrence MS, Auclair D, Chapuy B, Sougnez C, et al. Discovery and prioritization of somatic mutations in diffuse large B-cell lymphoma (DLBCL) by whole-exome sequencing. Proc Natl Acad Sci U S A 2012;109:3879–84.
54. Wilson AJ, Byun DS, Nasser S, Murray LB, Ayyanar K, Arango D, et al. HDAC4 promotes growth of colon cancer cells via repression of p21. Mol Biol Cell 2008;19:4062–75.
55. Geng H, Harvey CT, Pittsenbarger J, Liu Q, Beer TM, Xue C, et al. HDAC4 protein regulates HIF1alpha protein lysine acetylation and cancer cell response to hypoxia. J Biol Chem 2011;286:38095–102.
56. Cardnell RJ, Behrens C, Diao L, Fan Y, Tang X, Tong P, et al. An Integrated Molecular Analysis of Lung Adenocarcinomas Identifies Potential Therapeutic Targets among TTF1-Negative Tumors, Including DNA Repair Proteins and Nrf2. Clin Cancer Res 2015;21:3480–91.
57. Jia L, Zhou Z, Liang H, Wu J, Shi P, Li F, et al. KLF5 promotes breast cancer proliferation, migration and invasion in part by upregulating the transcription of TNFAIP2. Oncogene 2016;35:2040–51.
58. Ryland GL, Hunter SM, Doyle MA, Caramia F, Li J, Rowley SM, et al. Mutational landscape of mucinous ovarian carcinoma and its neoplastic precursors. Genome Med 2015;7:87.
59. Wang L, Lawrence MS, Wan Y, Stojanov P, Sougnez C, Stevenson K, et al. SF3B1 and other novel cancer genes in chronic lymphocytic leukemia. N Engl J Med 2011;365:2497–506.
60. Golovnina K, Blinov A, Akhmametyeva EM, Omelyanchuk LV, Chang LS. Evolution and origin of merlin, the product of the Neurofibromatosis type 2 (NF2) tumor-suppressor gene. BMC Evol Biol 2005;5:69.
61. Chiasson-MacKenzie C, Morris ZS, Baca Q, Morris B, Coker JK, Mirchev R, et al. NF2/Merlin mediates contact-dependent inhibition of EGFR mobility and internalization via cortical actomyosin. J Cell Biol 2015;211:391–405.
62. Li YY, Zhou CX, Gao Y. Moesin regulates the motility of oral cancer cells via MT1-MMP and E-cadherin/p120-catenin adhesion complex. Oral Oncol 2015;51:935–43.
63. Manson-Bahr D, Ball R, Gundem G, Sethia K, Mills R, Rochester M, et al. Mutation detection in formalin-fixed prostate cancer biopsies taken at the time of diagnosis using next-generation DNA sequencing. J Clin Pathol 2015;68:212–7.
64. Zhu T, Wang J, Pei Y, Wang Q, Wu Y, Qiu G, et al. Neddylation controls basal MKK7 kinase activity in breast cancer cells. Oncogene 2015.
65. Kazyken D, Kaz Y, Kiyan V, Zhylkibayev AA, Chen CH, Agarwal NK, et al. The nuclear import of ribosomal proteins is regulated by mTOR. Oncotarget 2014;5:9577–93.
66. Vangamudi B, Paul TA, Shah PK, Kost-Alimova M, Nottebaum L, Shi X, et al. The SMARCA2/4 ATPase domain surpasses the bromodomain as a drug target in SWI/SNF-mutant cancers: insights from cDNA rescue and PFI-3 inhibitor studies. Cancer Res 2015;75:3865–78.

Published OnlineFirst April 28, 2016; DOI: 10.1158/0008-5472.CAN-15-3190

Cancer Research
The Journal of Cancer Research (1916–1930) | The American Journal of Cancer (1931–1940)

AACR American Association for Cancer Research

Exome-Scale Discovery of Hotspot Mutation Regions in Human Cancer Using 3D Protein Structure

Collin Tokheim, Rohit Bhattacharya, Noushin Niknafs, et al.

Cancer Res 2016;76:3719-3731. Published OnlineFirst April 28, 2016.