Cell Systems

Supplemental Information

Pan-Cancer Analysis of Mutation Hotspots in Protein Domains

Martin L. Miller, Ed Reznik, Nicholas P. Gauthier, Bülent Arman Aksoy, Anil Korkut, Jianjiong Gao, Giovanni Ciriello, Nikolaus Schultz, and Chris Sander

Supplemental Information for:

Pan-Cancer Analysis of Mutation Hotspots in Protein Domains

Martin L. Miller[1,2,*], Ed Reznik[1], Nicholas P. Gauthier[1], Bülent Arman Aksoy[1], Anil Korkut[1], Jianjiong Gao[1], Giovanni Ciriello[1], Nikolaus Schultz[1], and Chris Sander[1,*].

[1] Computational Biology Program, Memorial Sloan-Kettering Cancer Center, 1275 York Avenue, New York, New York 10065, USA.
[2] Cancer Research UK Cambridge Institute, University of Cambridge, Li Ka Shing Centre, Robinson Way, Cambridge CB2 0RE, UK.
* Correspondence: martin.miller@cruk.cam.ac.uk (M.L.M.); chris@sanderlab.org (C.S.);

Extended Experimental Procedures

Mutation data and data preprocessing

We used the TCGA level 3 variant data (MAF file format) in the cBioPortal[2,3] which were retrieved from the Broad Institutes "Firehose" pipeline for processing of raw TCGA data. Thus, we used high level (processed) data for this study and relied on the variant and protein isoform calls from the Broad Firehose. To filter out mutations in low expressed genes, which has been shown to be associated with mutation biases[5], mRNA sequencing data in the form of normalized RSEM values were obtained from the same data portal. Within each tumor type, we determined the mean RSEM value for each gene and mutations in genes with a mean RSEM value of less than 10 were excluded from the analysis. Samples with extreme genomic instability, or so-called ultra-mutated samples, are generally thought to have many non-functional passenger mutations and could therefore bias mutation hotspot analysis, for example in cases where passenger mutations are tallied across genes in a domain family. Thus, we filtered out ultra-mutated samples by disregarding samples with more than 2,000 non-silent mutations. The TCGA tumor types analyzed were: Acute myeloid leukemia (LAML), Adrenocortical carcinoma (ACC), Bladder urothelial carcinoma (BLCA), Brain lower grade glioma (LGG), Breast invasive carcinoma (BRCA), Cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC), Colorectal adenocarcinoma (COADREAD), Glioblastoma multiforme (GBM), Head and neck squamous cell carcinoma (HNSC), Kidney chromophobe (KICH), Kidney renal clear cell carcinoma (KIRC), Kidney renal papillary cell carcinoma (KIRP), Liver hepatocellular carcinoma (LIHC), Lung adenocarcinoma (LUAD), Lung squamous cell carcinoma (LUSC), Ovarian serous cystadenocarcinoma (OV), Prostate adenocarcinoma (PRAD), Skin cutaneous melanoma (SKCM), Stomach adenocarcinoma (STAD), Thyroid carcinoma (THCA), Uterine carcinosarcoma (UCS), Uterine corpus endometrial carcinoma (UCEC).

Pfam domains and mapping mutations to protein domains

The Pfam-A data base of domains in the human proteome (version 26) as well as all human protein sequences were downloaded from the Pfam ftp server (pfam26.9606.tsv, ftp://ftp.ebi.ac.uk/pub/databases/Pfam). To include only high confidence domain calls, domains with an expectancy value (e-value) larger than $1e^{-5}$ were excluded. Mapping entries between MAF files and Pfam domains was performed using Uniprot accession numbers using the MAF ONCOTATOR.UNIPROT.ACCESSION.BEST.EFFECT field. In cases where the MAF entries did not have Uniprot accession numbers, the biomart webservice (http://www.ensembl.org/biomart/) was used to map between HGNC gene symbols and Uniprot accession numbers. The protein domain coordinates from the Pfam-A database were then matched to the MAF entries to determine if the mutations fell within or outside the boundaries of the protein domains using the MAF ONCOTATOR.PROTEIN.CHANGE.BEST.EFFECT field. MAF entries for which the mutated protein position and amino acid identity did not match with the corresponding amino acid identity in the protein sequences were excluded from the analysis. Furthermore, we excluded MAF entries where the mutated protein position was larger than length of the protein sequence.

Identification of domains with enriched mutation burden

For each domain we tallied the number of missense mutations falling (1) within the domain boundary, and compared it to (2) outside of the boundaries of all other domains in the gene, effectively excluding other domains than the domain in question. To assess if the mutation burden of the

domain was larger than would be expected by chance, we implemented a permutation test. The permutation test compared the observed mutation burden of the domain to the distribution of burdens generated by randomly distributing mutations across genes containing the domain. To generate this distribution, we repeated the following process for each permutation i:

1. For each gene g in the domain family, count the total number of observed mutations in the gene (both within and outside of the domain). Define this quantity to be n_g.
2. For each gene g, randomly redistribute n_g mutations across the gene, allowing for multiple mutations to fall at the same amino acid residue.
3. Count the total number of mutations which fall within the domain boundaries across all genes. Define this quantity to be m_i, the mutation burden of the domain in permutation i.

To calculate a p-value for the observed mutation burden of the domain, we compared the true mutation burden m_d derived from the data to the distribution of m_i. The p-value was defined to be the proportion of permutations with mutation burden greater than or equal to the observed mutation burden.

Note that by treating each gene separately and summing over the outcome of randomly distributed mutations in each gene, we are able to account for gene-to-gene variation in mutation rate (*e.g.* variation associated with replication timing[5] as well as differences in gene length and the proportion of each gene occupied by domains).

Domains with less than 25 mutations across all cancer types were excluded in the permutation analysis to avoid spurious results due to low mutation counts. Furthermore, to ensure proper random redistribution of mutations across genes and their domains, we omitted domains where the fraction of amino acids assigned as domains was larger than 75% of the all amino acids in the domain-containing proteins.

Domain mutation enrichment score

To calculate an enrichment score of mutations in the domain (e_d), we compared the observed domain mutation burden (m_d) to the expected domain mutation burden (m_e). We calculated m_e based on the total number of mutations observed (n_g) and the fraction of amino acids assigned as domains compared to total length of all genes in the domain family (f_d):

$$e_d = \frac{m_d}{m_e}, m_e = n_g \times f_d \tag{1}$$

Multiple sequence alignment of protein domains

The domain amino acid sequences were obtained as sub-strings from the protein sequences and aligned across domain-containing genes using the MathWorks multialign package with BLOSUM80 as scoring matrix and default parameters. For aligning domains present in only two genes, the Needleman-Wunsch algorithm was by applied using the MathWorks nwalign package with default parameters. After alignment of domains, missense mutations were tallied across analogous residues of domain-containing genes using the coordinates of the multiple sequence alignment. Residues with alignment gaps in more than 75% of the sequences were excluded from the domain hotspot analysis.

Identification of mutation hotspots within domain alignments

To identify putative hotspots for mutations within domains, we used as a null model the case of mutations falling with equal likelihood at all sites within a domain. Following multiple sequence alignment of all genes within a domain family, we tallied the number of observed mutations within the domain. We assumed that, for a particular residue to be called a putative hotspot, more mutations must fall on that residue than would be expected by chance if mutations were randomly distributed throughout the body of the domain. Assuming that each mutation falls at a random site along the domain body, the frequency of mutations at any particular residue follows a binomial distribution:

$$P(m = k) = \binom{n}{k} p^k (1-p)^{n-k} \tag{2}$$

where n is the total number of mutations in the domain, k is the number of mutations falling at a particular residue, and p is the probability of any individual mutation falling at a particular residue, and $P(m = k)$ is precisely the probability of observing k mutations at a single residue, assuming that n mutations were observed across the entire domain. Because our null model assumes an equal likelihood of mutations at any residue, $p = \frac{1}{L}$, where L is the length of the domain.
Thus, to assign a probability to the observation of k mutations falling at a particular site by change (*i.e.* a p-value), we calculate the probability of at least k mutations falling at a particular site from our null model

$$P(m \geq k) = \sum_{i=k}^{n} \binom{n}{k} p^k (1-p)^{n-k} \tag{3}$$

To correct for multiple hypothesis testing, p-values for all considered hotspots (aligned domain residues with more than two mutations) were adjusted using the Bonferroni correction method.

Entropy calculations

To assess how uniformly the mutations in a specific domain are spread across the genes containing such domain, we rely on the notion of Shannons information entropy. The information entropy S of a discrete probability distribution $P(x)$ is defined to be

$$S = -\sum_{i=1}^{n} P(x_i) \ln P(x_i) \tag{4}$$

where $P(x_i)$ is the probability of the i^{th} value of x. The entropy is maximal when $P(x)$ is uniform, *i.e.* each value of x is equally probable ($S_{max} = \ln n$), and minimal when $P(x)$ is equal to 1 for a single value of x ($S_{min} = 0$). In order to facilitate the comparison of entropy values for vectors of different dimension (*e.g.* domain families with different numbers of constituent genes), we use a normalized entropy measure $\bar{S}$ defined as

$$\bar{S} = \frac{-\sum_{i=1}^{n} P(x_i) \ln P(x_i)}{\ln n} \tag{5}$$

where n is the dimension of the vector x.

Supplementary Figures

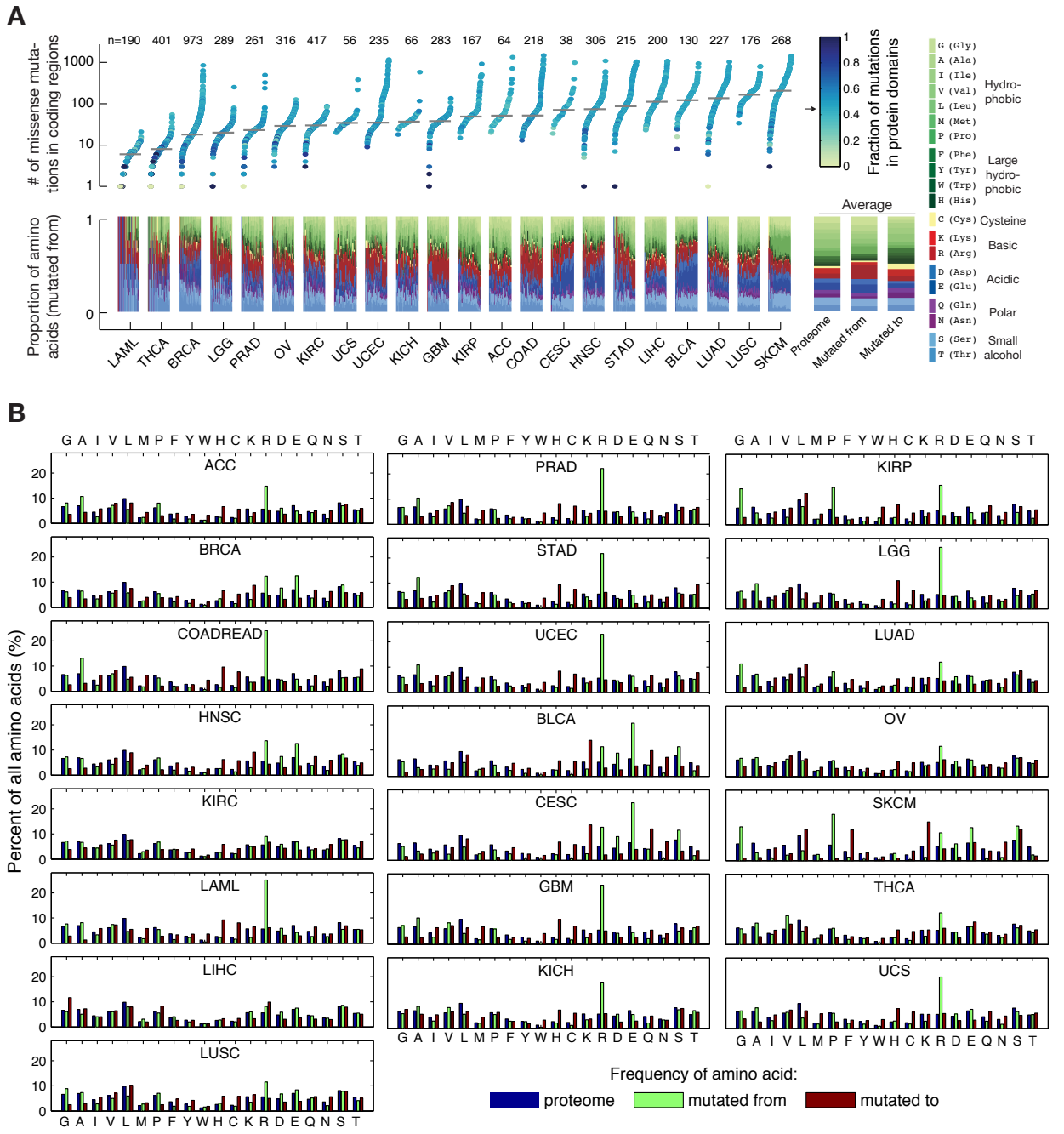

Figure S1: **Related to Figure 1. Mutation frequencies across cancer types and the relative proportion of mutated amino acid types**. (**A**) Within each cancer type the individual samples are ordered by the number of missense mutations in the proteome, and the median number of mutations is indicated (grey line). The color code represents the fraction of mutations that map to protein domains, and the arrow indicates the proportion of the proteome assigned as domains (0.454). The lower panel shows the relative proportion of amino acids altered by missense mutations in coding regions in each same (mutated from). The average proportions are displayed on the right where the first bar is the background frequency of amino acids in the proteome, the second bar is the average of all samples (mutated from), and the third bar is the resulting amino acid change (mutated to). Samples with more than 2000 missense mutations were excluded from the analysis. (**B**) The frequency of amino acids in the human proteome (blue) is compared to the frequency of amino acid before mutation (green) and after mutation (dark red) in the mutated position. Note that some amino acid types are disproportionally altered due to mutation biases in specific cancers[1,5], such as C→G transversions in bladder cancer (BLCA) that disproportionally alter the acidic amino acids aspartic acid (D) and glutamic acid (E), while C→T transitions in melanoma (SKCM) preferentially affect glycine (G) and proline (P). Moreover, arginine (R) is the most mutated amino acid across cancer types due to prevalent C→T transition at CP dinucleotides, which are present in four of six of arginine's codons.

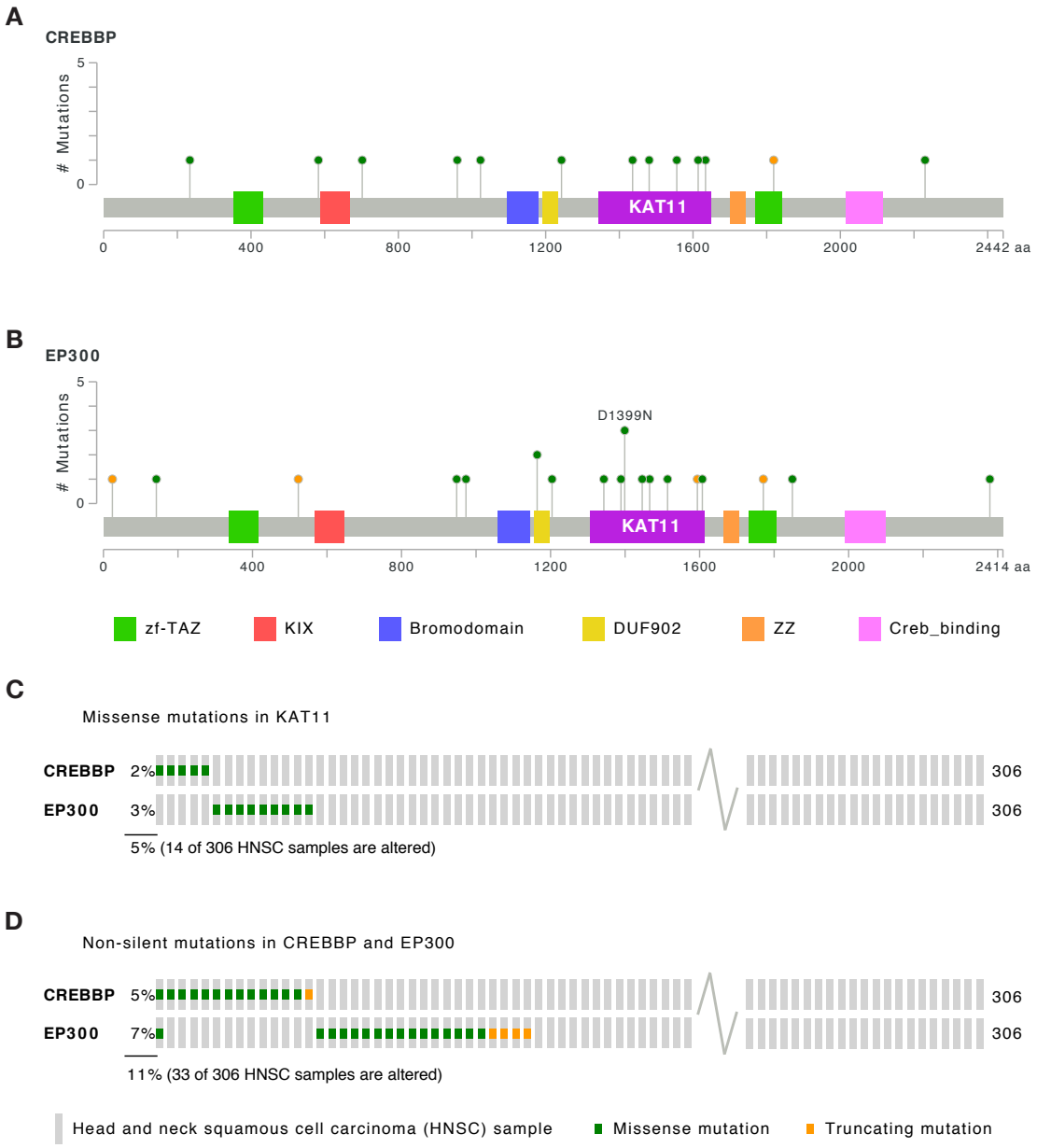

Figure S2: **Related to Figure 2. Mutations in the KAT11 domain of *CREBBP* and *EP300* in head and neck squamous cell carcinoma (HNSC).** Mutation counts across the gene span of *CREBBP* (**A**) and *EP300* (**B**) in HNSC samples. Note that the mutations tend to cluster to the lysine acetylase domain KAT11 in both genes. "Oncoprints" of mutations in individual HNSC samples in the KAT11 domain of *CREBBP* and *EP300* (**C**) and in the full length of the two genes (**D**) are provided to visualize the mutations in the context of each sample.

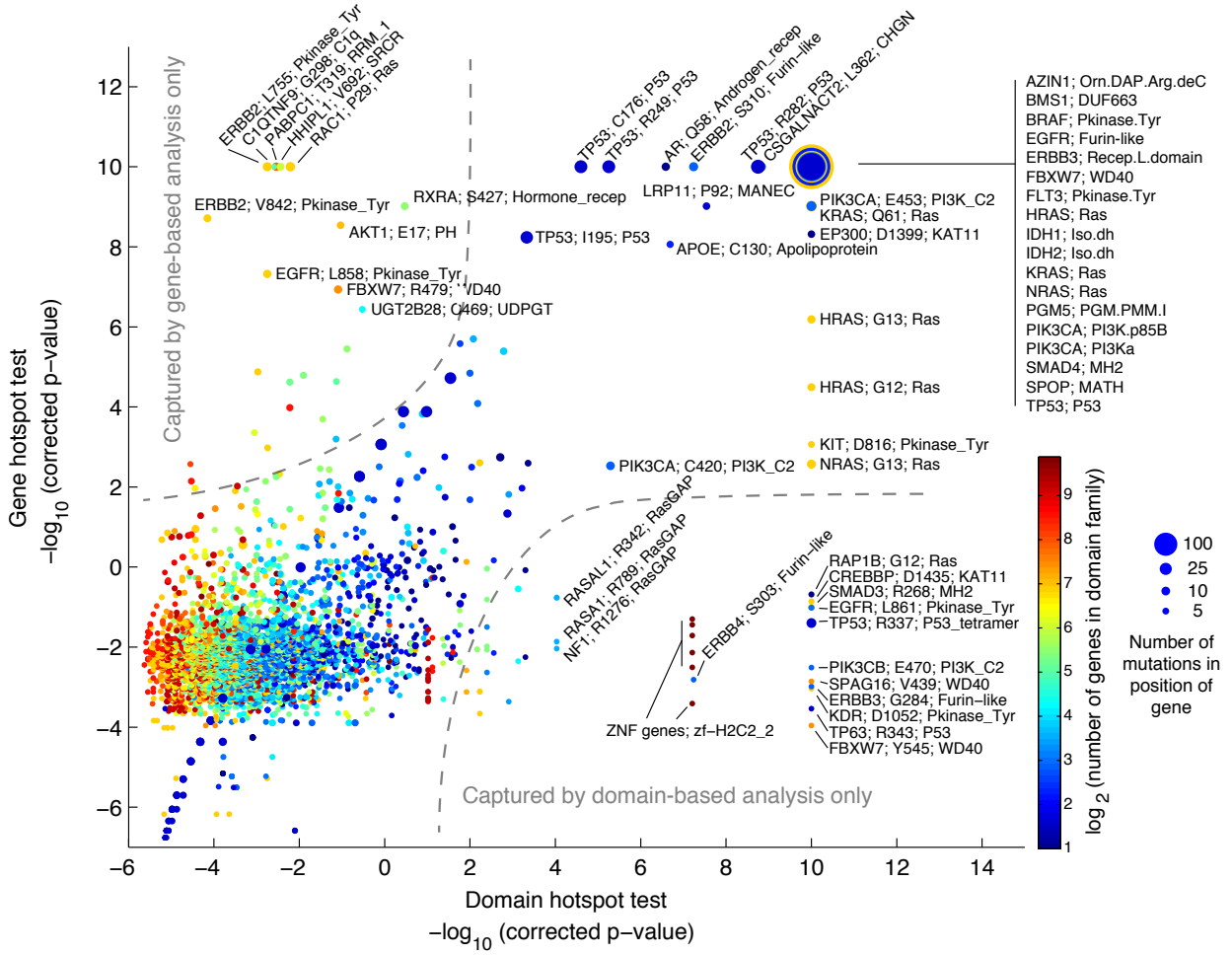

Figure S3: **Related to Figure 3. Systematic analysis of mutation hotspots identified through a domain-centric and a gene-centric approach.** The estimated significance level (Bonferroni corrected) of the identified domain mutation hotspots is plotted against the significance level of hotspots identified through a gene-based approach using similar binomial statistics. Hotspots are named by the gene followed by the mutated site and the domain name. The size of the dots reflects the number of mutations at each site and the dots are color coded by the number of domain-containing genes in the genome. Note that the two approaches are complementary: there are numerous mutations in genes that would not have been significantly associated with hotspots without taking domain sequence similarity into account (68 cases, lower right corner, see **Table S2** for details), and vice versa, there are multiple mutation sites that are not captured by the domain-based approach but only the gene-based approach (upper left corner), particularly for large domain families with high mutation loads which diminishes statistical power for domain hotspot detection.

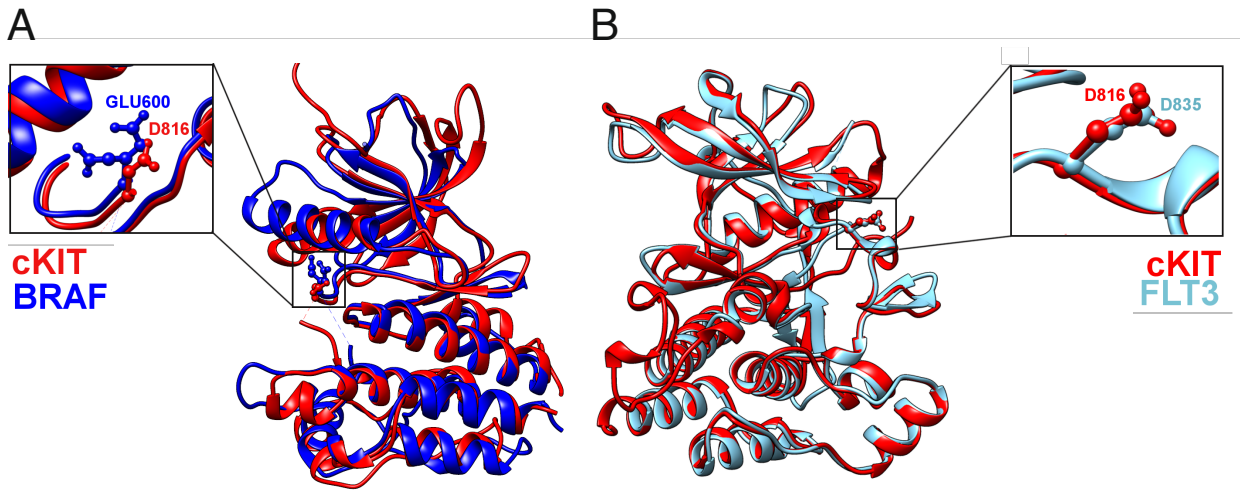

Figure S4: **Related to Table 2. Structural alignment of tyrosine kinases superimposes conserved hotspot residues.** (**A**) The structures of cKIT and BRAF kinase domains are aligned (BRAF PDB ID: 4MNF, cKIT PDB ID: 1PKG). The activation loop is in the active conformation in both structures and the activation loop hotspot residues superpose the hotspots. Note that the mutated (V600E) form of the BRAF structure is shown while the cKIT and FLT3 structure are shown wildtype form. (**B**) The structures of cKIT and FLT3 kinase domains are aligned (FLT3 PDB ID: 1RJB). The activation loop is in an autoinhibited state for both proteins and the hotspot residues superpose perfectly. In each case, we aligned the protein structures to minimize the global root mean square deviation (RMSD) between the two structures and no bias was introduced to superpose the hotspot residues. The structural alignment and analysis was performed with the MatchMaker algorithm implemented in UCSF chimera[6].

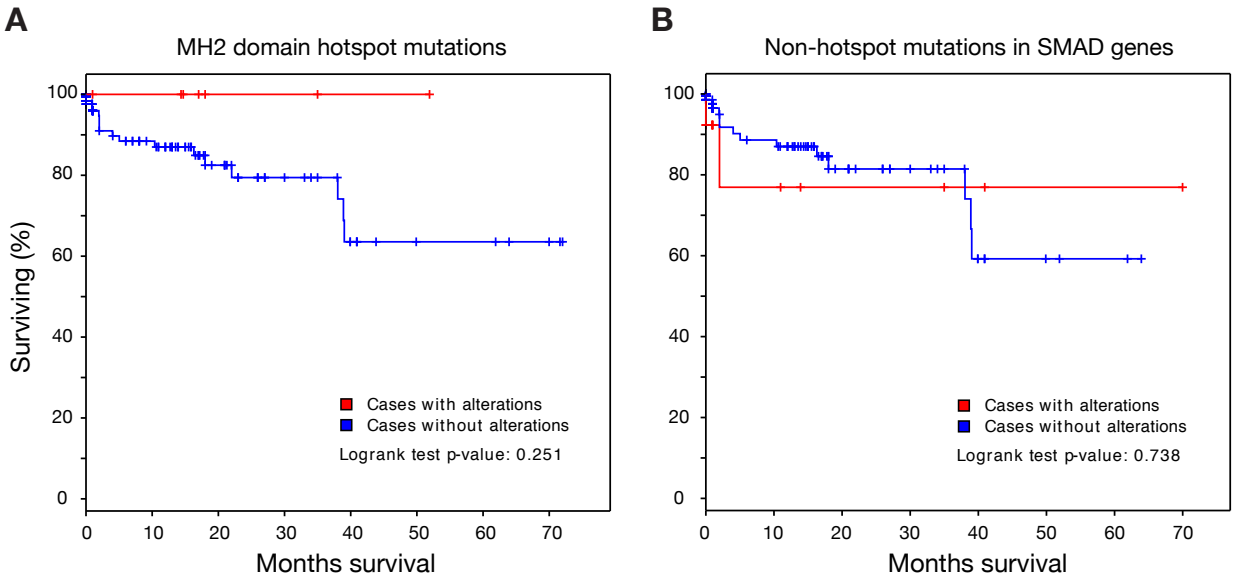

Figure S5: **Related to Figure 4B. Hotspot mutations in the MH2 domain of *SMAD* genes in colorectal adenocarcinoma.** (**A**) Kaplan-meier survival curves of colorectal adenocarcinoma patients with (red) and without (blue) MH2 domain hotspot mutations. Note that in addition to the major hotspot identified at position 46 of the domain alignment, positions 39, 40, and 57 were also considered hotspots in colorectal cancer for this analysis (SMAD1: MUT = R319; SMAD2: MUT = R321 MUT = R321 MUT = D304 MUT = P305; SMAD3: MUT = R268 MUT = D262; SMAD4: MUT = R361 MUT = D355 MUT = P356) (**B**) Similar analysis for mutations found in MH2 domain containing genes (SMAD1-9) except MH2 domain hotspot mutations ("non-hotspot mutations") compared to all other samples. All data were obtained from the cBioPortal for cancer genomics data[2,3].

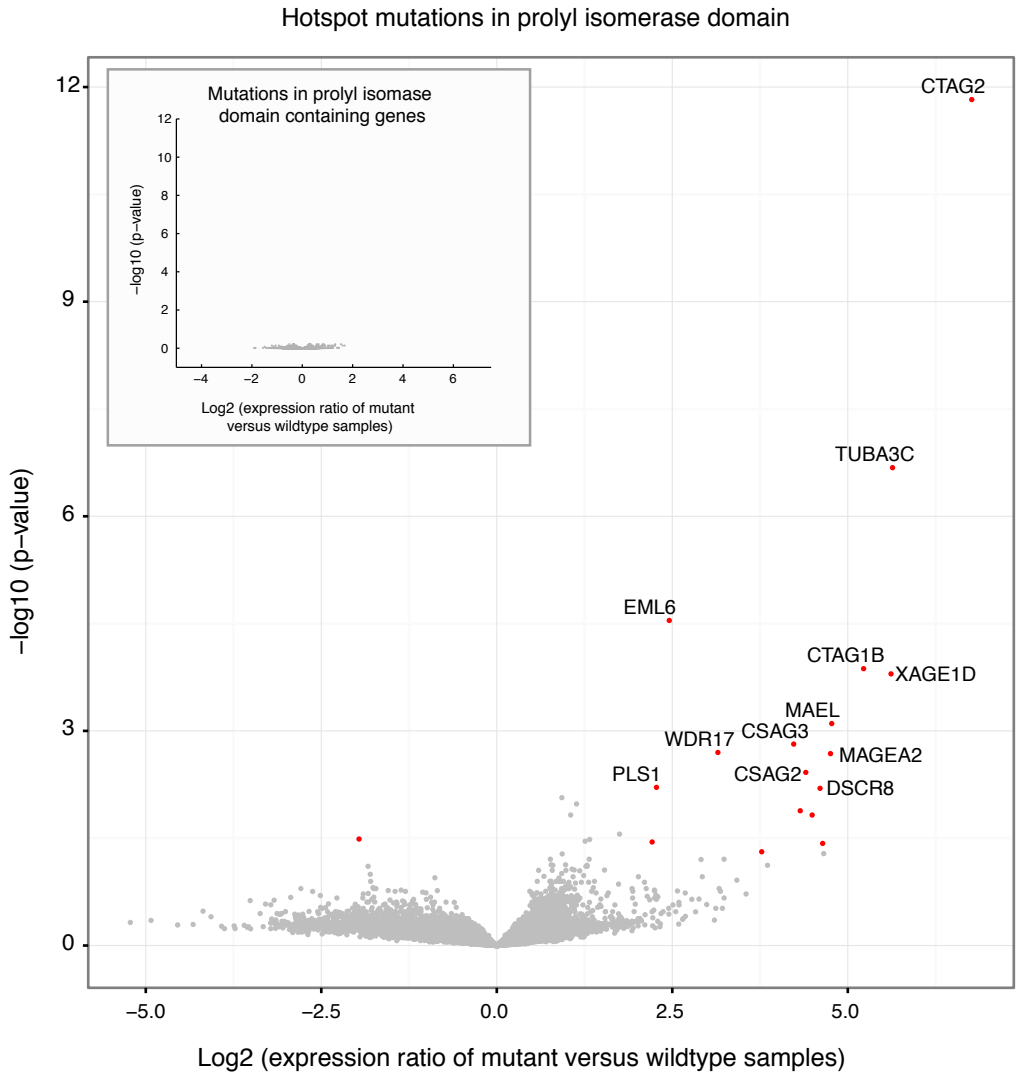

Figure S6: **Related to Figure 5A. Upregulation of genes in melanoma with prolyl isomerase hotspot mutations.** Gene expression difference (ratio of mRNA RSEM values) between melanoma samples containing hotspot mutations in the prolyl isomerase domain (Pro_isomerase) and melanoma samples without hotspot mutations (wildtype) is plotted as a function of the statistical significance (moderated t-statistics adjusted for multiple testing by the Benjamini and Hochberg method). Seven out of 257 melanoma samples had hotspot mutations with mutations in *PPIAL4G* (four R37C mutations), *PPIG* (two R41C mutations), and *PPIA* (one R37C mutation). Highlighted genes (red) are more than 4-fold differentially regulated at the level of FDR < 0.05. Inset shows a similar analysis except mutant samples were considered as all samples with missense mutations in genes containing the domain (excluding samples with hotspot mutations). Note that only hotspot mutations have a perturbing effect on gene expression in melanoma samples although the functional consequence remains unclear. Genes with RSEM values below 32 were excluded to avoid analyzing very low and/or non-expressed genes. All data were obtained from the cBioPortal for cancer genomics data[2,3].

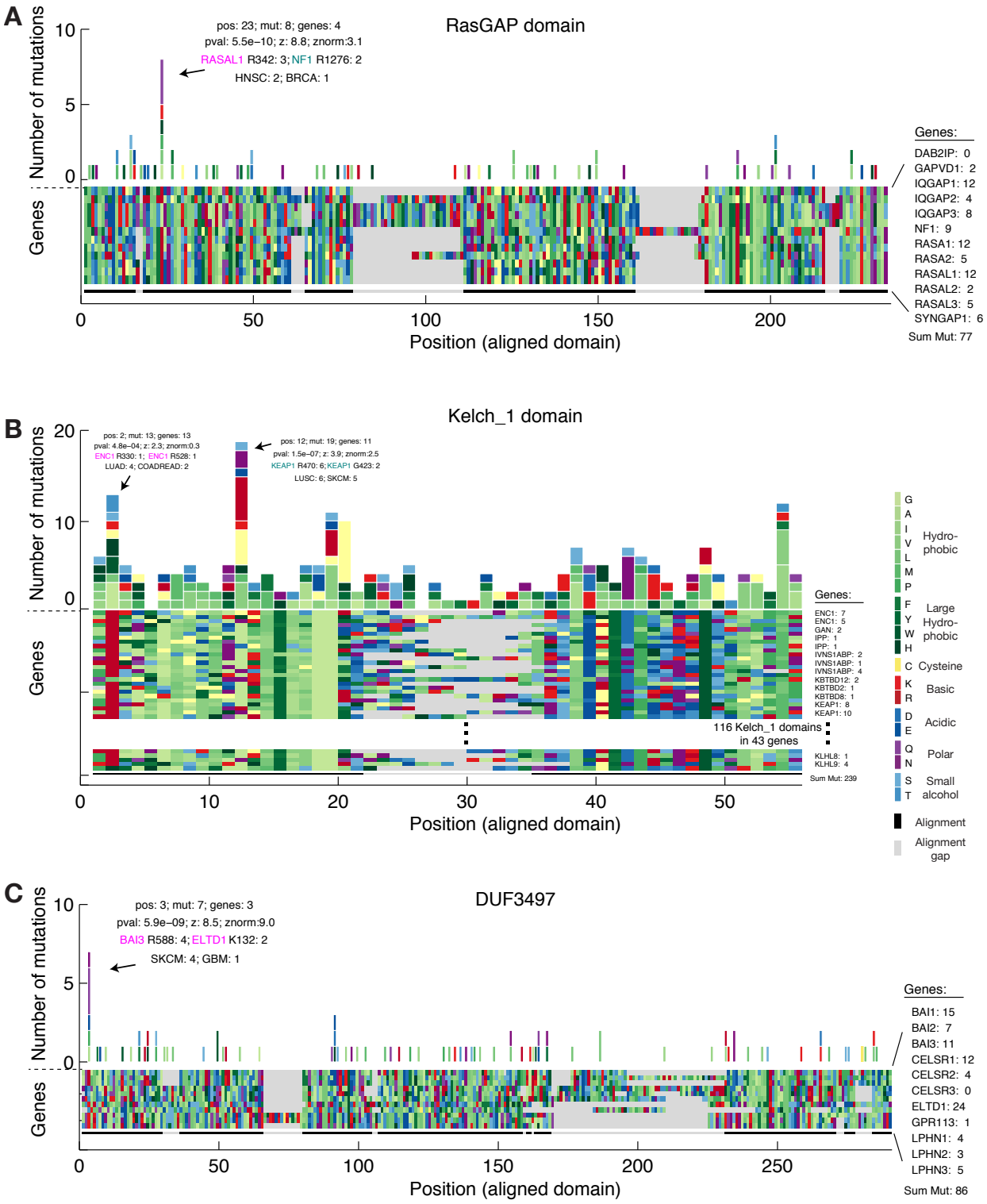

Figure S7: **Related to Table 2. Identification of putative hotspots in RasGAP, Kelch_1, and DUF3497.** (**A**) The sequence alignment of the RasGAP domain is represented as a block of rectangles where the aligned amino acids are color coded by their biochemical properties. Alignment gaps are indicated by gray rectangles. Using the resulting alignment coordinates, missense mutations are tallied across domain-containing genes and color coded according to the generated amino acid types. A significant hotspot is indicated at alignment position 23 with associated p-value and number of mutations in top mutated genes and cancer types. In this hotspot, the mutated genes are *RASAL1* (R342G/H/P), *RASA1* (R789L/Q), *NF1* (R1276Q), and *RASA2* (R397K). (**B**) Similar plot but for the Kelch_1 domain. Of note, some genes contain several repeats of the the Kelch motif. (**C**) Similar plot but for the DUF3497 domain (Domain of Unknown Function). Closer inspection of mutations and genes in the domains can be found a `http://mutationaligner.org/`.

Supplementary Tables

Table S1: **Related to Table 2. Significant mutation hotspots identified in protein domains.** The detected domain hotspots are listed by their Pfam domain identifiers, the number of genes in the domain family, the position of the hotspot in the domain alignment, the Bonferroni-corrected p-values, the entropy score ($\bar{S}$), the number of mutations in the hotspot, the genes with the most mutations in the hotspot, and the cancer type with most mutations in the hotspot. The genes written in italic font if they were reported to be significantly mutated or bold font if they were not in a recent pan-cancer study[4]. The list is sorted by p-value followed by entropy score.

Domain	**Genes** (#)	**Position**	**pValue** (-log$_{10}$)	$\bar{S}$	**Mut** (#)	**Top Gene 1** (# mut, gene, site)	**Top Gene 2** (# mut, gene, site)	**Top Cancer** (# mut, cancer)
KAT11	2	94	10	0.88	10	7 *EP300* D1399	3 *CREBBP* D1435	3 BLCA
Orn_DAP_Arg_deC	3	86	10	0.28	11	10 **AZIN1** S367	1 **ODC1** R369	10 LIHC
Ras	124	17	10	0.27	56	31 *KRAS* G13	11 *NRAS* G13	12 COADREAD
MH2	8	46	10	0.25	14	11 *SMAD4* R361	3 **SMAD3** R268	8 COADREAD
Furin-like	7	119	10	0.20	30	27 *EGFR* A289	2 *ERBB3* G284	23 GBM
PI3K_C2	7	130	10	0.19	17	15 *PIK3CA* E453	2 **PIK3CB** E470	7 BRCA
Ras	124	88	10	0.18	189	142 *NRAS* Q61	22 *HRAS* Q61	78 SKCM
Iso_dh	5	128	10	0.14	292	274 *IDH1* R132	18 *IDH2* R172	232 LGG
WD40	170	16	10	0.13	37	31 *FBXW7* R465	2 *FBXW7* Y545	12 COADREAD
Ras	124	16	10	0.12	224	192 *KRAS* G12	17 *NRAS* G12	74 COADREAD
Pkinase_Tyr	120	291	10	0.09	415	382 *BRAF* V600	14 *FLT3* D835	235 THCA
Recep_L_domain	14	52	10	0.08	17	16 *ERBB3* V104	1 *ERBB2* I101	5 COADREAD
P53	3	181	10	0.07	141	139 *TP53* R273	2 **TP63** R343	44 LGG
PI3Ka	8	27	10	0.02	164	163 *PIK3CA* E545	1 **PIK3CB** E552	66 BRCA
P53	3	156	10	0	99	99 *TP53* R248		14 OV
PI3Ka	8	24	10	0	93	93 *PIK3CA* E542		43 BRCA
P53	3	81	10	0	77	77 *TP53* R175		18 BRCA
P53	3	128	10	0	41	41 *TP53* Y220		8 BRCA
P53	3	85	10	0	39	39 *TP53* H179		7 HNSC
P53	3	101	10	0	39	39 *TP53* H193		9 BRCA
P53	3	153	10	0	36	36 *TP53* G245		9 OV
PI3Ka	8	28	10	0	30	30 *PIK3CA* Q546		11 BRCA
PGM_PMM_I	4	88	10	0	19	19 **PGM5** I98		17 STAD
Iso_dh	5	96	10	0	18	18 *IDH2* R140		17 LAML
PI3K_p85B	3	58	10	0	18	18 *PIK3CA* R88		9 UCEC
P53_tetramer	3	20	10	0	14	14 *TP53* R337		2 HNSC
MATH	8	107	10	0	13	13 *SPOP* F133		13 PRAD
DUF663	2	67	10	0	12	12 **BMS1** E878		10 KIRP
CHGN	8	398	8.83	0	11	11 **CSGALNACT2** L362		5 KIRP
P53	3	190	8.75	0	34	34 *TP53* R282		7 HNSC
MANEC	3	9	7.54	0	7	7 **LRP11** P92		7 ACC
Furin-like	7	136	7.24	0.22	13	11 *ERBB2* S310	2 **ERBB4** S303	4 STAD
zf-H2C2_2	940	7	7.21	0.62	79	2 **ZNF208** R613	2 **ZNF286A** R430	27 UCEC
Apolipoprotein	5	52	6.69	0	6	6 **APOE** C130		6 ACC
Androgen_recep	2	276	6.59	0	9	9 **AR** Q58		2 ACC
PI3K_C2	7	94	5.29	0	10	10 *PIK3CA* C420		4 UCEC
P53	3	157	5.25	0.14	29	28 *TP53* R249	1 **TP63** R319	8 LIHC
P53	3	82	4.6	0	28	28 *TP53* C176		6 OV
RasGAP	12	23	4.03	0.53	8	3 **RASAL1** R342	2 *NF1* R1276	2 HNSC
ELFV_dehydrog	2	206	3.36	0	5	5 **GLUD2** L468		5 KIRP
P53	3	103	3.33	0	26	26 *TP53* I195		9 OV

Domain (Table S1 Continued)	**Genes** (#)	**Position**	**pValue** (-log$_{10}$)	$\bar{S}$	**Mut** (#)	**Top Gene 1** (# mut, gene, site)	**Top Gene 2** (# mut, gene, site)	**Top Cancer** (# mut, cancer)
BicD	2	570	3.3	0.97	5	3 **BICD2** R635	2 **BICD1** R633	2 SKCM
Sox_C_TAD	4	208	3.22	0	4	4 *SOX17* S403		4 UCEC
Pro_isomerase	19	31	3.13	0.48	9	4 **PPIAL4G** R37	2 **PPIG** R41	7 SKCM
DUF3497	11	3	2.91	0.40	7	4 **BAI3** R588	2 **ELTD1** K132	4 SKCM
PI3K_p85B	3	78	2.88	0	9	9 *PIK3CA* R108		4 UCEC
MT	15	36	2.79	0.14	8	7 **DNAH5** D3236	1 **DNAH11** H3139	8 SKCM
VHL	2	32	2.71	0	9	9 *VHL* L89		9 KIRC
Choline_transpo	5	186	2.56	0.42	5	3 **SLC44A1** R437	2 **SLC44A4** R496	1 BRCA
bZIP_2	9	21	2.4	0.61	6	2 **HLF** R243	2 **NFIL3** R91	2 UCEC
Fork_head	42	88	2.4	0.46	11	3 **FOXP1** R514	2 **FOXJ1** R170	4 COADREAD
Kelch_1	116	12	2.23	0.45	19	6 *KEAP1* R470	2 *KEAP1* G423	6 LUSC
GTP_EFTU_D3	8	105	2.18	0	6	6 **EEF1A1** T432		5 LIHC
Annexin	44	32	2.17	0.54	10	2 **ANXA1** R303	2 **ANXA1** R72	2 BRCA
Acyl-CoA_dh_1	12	79	2.08	0	6	6 **ACADS** R330		2 KIRC
ATP-synt_ab_C	5	42	2.04	0.31	5	4 *ATP5B* K459	1 **ATP6V1B2** K457	2 COADREAD
Cu_amine_oxidN2	3	62	2.03	0.95	4	2 **AOC2** I123	1 **ABP1** V100	1 COADREAD
Sec23_trunk	6	188	2	0.28	5	4 **SEC23B** R313	1 **SEC24C** N665	2 SKCM
Glyco_hydro_18	7	350	2	0	6	6 **CHIT1** A359		5 LIHC
CUB	142	45	1.92	0.57	22	3 **CSMD3** R100	2 **CSMD1** D1487	10 SKCM
RunxI	3	45	1.85	0	3	3 **RUNX2** P466		2 KIRC
PI3K_p85B	3	8	1.77	0.34	8	7 *PIK3CA* R38	1 **PIK3CB** R48	5 UCEC
Perilipin	6	341	1.77	0	5	5 **PLIN5** R306		5 ACC
RRM_6	41	2	1.77	0.58	10	2 **GRSF1** R153	1 **ESRP1** R329	3 SKCM
K_tetra	49	110	1.76	0.52	10	2 **KCTD7** R112	2 **KCTD9** R153	2 BRCA
MYT1	4	39	1.73	0.75	4	2 **MYT1L** R659	1 **MYT1** R598	1 COADREAD
UNC45-central	2	3	1.66	0	3	3 **UNC45A** R289		1 KIRP
DIE2_ALG10	2	389	1.66	0.81	4	3 **ALG10** R416	1 **ALG10B** R416	4 SKCM
BAAT_C	5	108	1.65	0.43	4	2 **ACOT4** P310	2 **BAAT** P312	4 SKCM
Pkinase	367	279	1.6	0.58	33	2 **IKBKE** R134	2 **MAP3K4** R1462	6 SKCM
BCL_N	3	51	1.6	0.58	3	2 **BCL7A** T52	1 **BCL7B** T52	2 COADREAD
Xylo_C	2	142	1.59	0	4	4 **XYLT1** R754		1 BRCA
Tmemb_161AB	2	98	1.56	1.00	4	2 **TMEM161A** R98	2 **TMEM161B** H99	2 GBM
P53	3	146	1.54	0.16	23	22 *TP53* C238	1 **TP63** C308	4 BRCA
Creb_binding	2	99	1.5	0.81	4	3 *CREBBP* R2104	1 *EP300* R2088	2 PRAD
FerI	5	50	1.47	0.43	4	2 **FER1L6** T215	2 **OTOF** T388	2 BRCA
MT-A70	3	115	1.42	0	4	4 **METTL14** R298		4 UCEC
NT-C2	4	39	1.41	0	4	4 **EHBP1L1** R50		2 UCEC
Myosin_head	39	23	1.39	0.59	10	2 **MYO18A** R428	1 **MYH4** R109	3 LUAD
ADAM_spacer1	23	112	1.38	0.46	9	4 **ADAMTS2** G824	2 **ADAMTS12** Q800	3 SKCM
Cadherin_2	58	44	1.37	0.62	15	2 **PCDHA7** R62	2 **PCDHGA9** R62	3 COADREAD
CAMSAP_CH	2	55	1.36	0	3	3 **ASPM** T1172		1 HNSC

Table S2. Supplemental spreadsheet. Related to Table 2 and Figure S3. Systematic analysis of mutation hotspots identified through a domain- or a gene-centric approach. Listed are potential mutation hotspots identified in domain-containing genes either through a domain-based (pVal Domain) or a gene-based (pVal Gene) approach using binomial statistics and correcting for multiple hypothesis testing (Bonferroni). The number of mutations in the domain hotspot (Mut Domain) as well as the number of mutations in the gene of interest (Mut Gene) are listed. Only genes with more than two mutations at the same site are considered. Table S2 can be found as a supplemental spreadsheet.

References

[1] Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A. J. R., Behjati, S., Biankin, A. V., Bignell, G. R., Bolli, N., Borg, A., Børresen-Dale, A.-L., Boyault, S., Burkhardt, B., Butler, A. P., Caldas, C., Davies, H. R., Desmedt, C., Eils, R., Eyfjörd, J. E., Foekens, J. A., Greaves, M., Hosoda, F., Hutter, B., Ilicic, T., Imbeaud, S., Imielinsk, M., Jäger, N., Jones, D. T. W., Jones, D., Knappskog, S., Kool, M., Lakhani, S. R., López-Otín, C., Martin, S., Munshi, N. C., Nakamura, H., Northcott, P. A., Pajic, M., Papaemmanuil, E., Paradiso, A., Pearson, J. V., Puente, X. S., Raine, K., Ramakrishna, M., Richardson, A. L., Richter, J., Rosenstiel, P., Schlesner, M., Schumacher, T. N., Span, P. N., Teague, J. W., Totoki, Y., Tutt, A. N. J., Valdés-Mas, R., van Buuren, M. M., van t Veer, L., Vincent-Salomon, A., Waddell, N., Yates, L. R., Zucman-Rossi, J., Andrew Futreal, P., McDermott, U., Lichter, P., Meyerson, M., Grimmond, S. M., Siebert, R., Campo, E., Shibata, T., Pfister, S. M., Campbell, P. J. and Stratton, M. R. (2013). Signatures of mutational processes in human cancer. Nature *500*, 415–421.

[2] Cerami, E., Gao, J., Dogrusoz, U., Gross, B. E., Sumer, S. O., Aksoy, B. A., Jacobsen, A., Byrne, C. J., Heuer, M. L., Larsson, E., Antipin, Y., Reva, B., Goldberg, A. P., Sander, C. and Schultz, N. (2012). The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. Cancer Discovery *2*, 401–404.

[3] Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O., Sun, Y., Jacobsen, A., Sinha, R., Larsson, E., Cerami, E., Sander, C. and Schultz, N. (2013). Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal. Science signaling *6*, pl1–pl1.

[4] Lawrence, M. S., Stojanov, P., Mermel, C. H., Robinson, J. T., Garraway, L. A., Golub, T. R., Meyerson, M., Gabriel, S. B., Lander, E. S. and Getz, G. (2014a). Discovery and saturation analysis of cancer genes across 21 tumour types. Nature *505*, 495–501.

[5] Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V., Cibulskis, K., Sivachenko, A., Carter, S. L., Stewart, C., Mermel, C. H., Roberts, S. A., Kiezun, A., Hammerman, P. S., McKenna, A., Drier, Y., Zou, L., Ramos, A. H., Pugh, T. J., Stransky, N., Helman, E., Kim, J., Sougnez, C., Ambrogio, L., Nickerson, E., Shefler, E., Cortes, M. L., Auclair, D., Saksena, G., Voet, D., Noble, M., DiCara, D., Lin, P., Lichtenstein, L., Heiman, D. I., Fennell, T., Imielinski, M., Hernandez, B., Hodis, E., Baca, S., Dulak, A. M., Lohr, J., Landau, D.-A., Wu, C. J., Melendez-Zajgla, J., Hidalgo-Miranda, A., Koren, A., McCarroll, S. A., Mora, J., Lee, R. S., Crompton, B., Onofrio, R., Parkin, M., Winckler, W., Ardlie, K., Gabriel, S. B., Roberts, C. W. M., Biegel, J. A., Stegmaier, K., Bass, A. J., Garraway, L. A., Meyerson, M., Golub, T. R., Gordenin, D. A., Sunyaev, S., Lander, E. S. and Getz, G. (2014b). Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature *499*, 214–218.

[6] Meng, E. C., Pettersen, E. F., Couch, G. S., Huang, C. C. and Ferrin, T. E. (2006). Tools for integrated sequence-structure analysis with UCSF Chimera. BMC Bioinformatics *7*, 339.