Protein Domain Hotspots Reveal Functional Mutations across Genes in Cancer

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] Present Address: 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 and sander@cbio.mskcc.org

ABSTRACT

In cancer genomics, frequent recurrence of mutations in independent tumor samples is a strong indication of functional impact. However, rare functional mutations can escape detection by recurrence analysis for lack of statistical power. We address this problem by extending the notion of recurrence of mutations from single genes to gene families that share homologous protein domains. In addition to lowering the threshold of detection, this sharpens the functional interpretation of the impact of mutations, as protein domains more succinctly embody function than entire genes. Mapping mutations in 22 different tumor types to equivalent positions in multiple sequence alignments of protein domains, we confirm well-known functional mutation hotspots and make two types of discoveries: 1) identification and functional interpretation of uncharacterized rare variants in one gene that are equivalent to well-characterized mutations in canonical cancer genes, such as uncharacterized *ERBB4* (S303F) mutations that are analogous to canonical *ERRB2* (S310F) mutations in the furin-like domain, and 2) detection of previously unknown mutation hotspots with novel functional implications. With the rapid expansion of cancer genomics projects, protein domain hotspot analysis is likely to provide many more leads linking mutations in proteins to the cancer phenotype.

INTRODUCTION

The landscape of somatic mutations in cancer is extraordinarily complex, making it difficult to distinguish oncogenic alterations from passenger mutations. Many approaches use the recurrence of alterations in a single gene across tumor samples to identify potential driver genes. However, the molecular functions of genes are often pleiotropic, and in many cases it may not be a gene as an entity itself, but rather the specific function of a gene or a set of genes that is under selective pressure in cancer. For example, in T-cell acute lymphoblastic leukemia, the transmembrane signaling receptor *NOTCH1* is activated by mutations in the heterodimerization and the PEST domains (Weng *et al.*, 2004), while in squamous cell carcinomas, notch-signaling has a tumor suppressive role and notch receptors (*NOTCH1-4*) are inactivated by mutations in the ligand binding EGF-like domains (Wang *et al.*, 2011). Thus, an alternative approach to assessing the relevance of somatic alterations is to determine the recurrence of mutations in genes involved in similar molecular functions. One powerful method for systematically assessing common biological function of genes is through the analysis of protein domains, which are evolutionarily conserved, structurally related functional units encoded in the protein sequence of genes (Holm & Sander, 1996; Chothia *et al.*, 2003). By coupling the observation of mutations across genes in a domain family together, it may be possible to identify putative functional alterations that confer a selective, functional advantage to cancer cells.

Large cross-institutional projects, such as The Cancer Genome Atlas (TCGA), have recently profiled the major human cancer types genomically, including glioblastoma (McLendon *et al.*, 2008), lung (Hammerman *et al.*, 2012; Ding *et al.*, 2008), ovarian (Bell *et al.*, 2011), breast (Koboldt *et al.*, 2012), endometrial (Getz *et al.*, 2013), kidney (Creighton *et al.*, 2013) and colorectal cancer (Cancer Genome Atlas Network, 2012). Through whole-exome sequencing (WES) of tumor-normal pairs, these and other studies have provided catalogues of somatically mutated genes that are frequently altered and therefore likely associated with disease development. However, despite a collection of mutation data from nearly 5,000 samples encompassing 21 tumor types, the results from a recent pan-cancer study illustrate that by using recurrence of mutations in genes, thousands of samples per tumor type are needed to confidently identify genes that are mutated at low but clinically relevant frequencies (2-5%) (Lawrence *et al.*, 2014a).

Several analytical approaches have been developed to detect genes associated with oncogenesis (Gonzalez-Perez & Lopez-Bigas, 2012; Dees *et al.*, 2012; Lawrence *et al.*, 2014b). One of these widely applied algorithms, MutSigCV, compares the gene-specific mutation burden to a background model using silent mutations in the gene and gene neighborhood to estimate the probability that the gene is significantly mutated (Lawrence *et al.*, 2014b). Importantly, the method also incorporates contextual information, such as genomic parameters that correlate strongly with the mutation background rate (DNA replication timing and the general level of transcriptional activity) (Lawrence *et al.*, 2014b) as well as the tendencies of mutations to cluster to specific sites and to occur at positions that are evolutionarily conserved (Lohr *et al.*, 2012; Lawrence *et al.*, 2014a). Additional approaches have been developed to predict the functional impact of specific amino acid changes. These approaches generally rely on analyzing physico-chemical properties of amino acid substitutions (*e.g.*, changes in size and polarity), structural information (*e.g.*, hydrophobic propensity and surface accessibility), and the evolutionary conservation of the mutated residues across a set of related genes (Reva *et al.*, 2011; Yue *et al.*, 2006; Bromberg *et al.*, 2008; Ng, 2003; Adzhubei *et al.*, 2010). Other approaches analyze mutations across sets of functionally

related genes (*e.g.*, genes in the same signaling pathway) to test for a possible enrichment of mutation events (Cerami *et al.*, 2010; Ciriello *et al.*, 2012; Hofree *et al.*, 2013; Torkamani & Schork, 2009).

Protein domains represent particular sequence variants that have been formed over evolution by duplication and/or recombination (Holm & Sander, 1996; Chothia *et al.*, 2003). Domains often encode structural units associated with specific cellular tasks, and large proteins with multiple domains can have several molecular functions each exerted by a specific domain. The structure-function relationship encoded in domains has been used as a tool for understanding the effect of mutations across functionally related genes. For example, some of the most frequent oncogenic mutations in human cancer affect analogous residues of the activation segment of the kinase domain and cause constitutive activation of several oncogenes, including *FLT3* D835 mutations in acute myeloid leukemia, *KIT* D816 mutations in gastrointestinal stromal tumors, and *BRAF* V600 mutations in melanoma (Dibb *et al.*, 2004; Greenman *et al.*, 2007). In the *SMAD* tumor suppressor genes, mutations in conserved residues of the MAD homology 2 (MH2) domain have analogous effects in *SMAD2* and *SMAD4*, disrupting homo- and hetero-oligomeric interactions critical for SMAD signaling (Shi *et al.*, 1997). Proteome-wide bioinformatics analysis of mutations in domains have been performed to identify domains enriched for alterations (Nehrt *et al.*, 2012; Peterson *et al.*, 2012) as well as to detect significantly mutated domain hotspots through multiple sequence analysis (Peterson *et al.*, 2010; Yue *et al.*, 2010). However, due in part to the scarcity of data available at the time of analysis, these studies did not perform a systematic pan-cancer analysis of mutations in domains and provided limited biological insights.

Here, we performed a systematic and comprehensive analysis of mutations in protein domains using data from more than 5,000 tumor-normal pairs from 22 cancer types profiled by the TCGA consortium and domains from the protein family database Pfam-A (Punta *et al.*, 2011). We confirmed that signaling domains in canonical oncogenes are recurrently altered in cancer and further identified domains that are enriched for mutations contributed by infrequently altered genes not previously associated with cancer. Using multiple sequence analysis, we determined if conserved residues in protein domains were affected by mutations across related genes. This analysis enabled us to identify putative "domain hotspots". For example, we discovered novel hotspots with putative driver mutations in the prolyl isomerase domain and in the DNA-binding forkhead domain. We further exposed rare mutations that associated with well-characterized oncogenic mutations, including the furin-like domain where uncharacterized mutations in *ERBB4* (S303F) are analogous to known oncogenic mutations in the same domain of *ERRB2* (S310F), suggesting similar functional consequences. In several cases, we associated rare mutations in potential cancer genes with therapeutically actionable hotspots in known oncogenes, underlining the potential clinical implications of our findings. We have made all results freely available to the research community through an interactive web resource (`http://www.mutationaligner.org`) that will be continuously updated as data become available from cancer genomics projects.

RESULTS

Mapping somatic mutations to protein domains

To systematically analyze somatic mutations in the context of conserved protein domains, we collected WES data from 5496 tumor-normal pairs of 22 different tumor types profiled by the TCGA consortium. To obtain a uniform data set of mutation calls, annotation of somatic mutations were based on the publicly available data (Oct 2014) from the cBioPortal for cancer genomics data (Cerami *et al.*, 2012; Gao *et al.*, 2013) (**Fig. 1**). After filtering out ultra-mutated samples and mutations in genes with low mRNA expression levels (**Methods**), the data consisted of a total of 727,567 mutations in coding regions with 463,842 missense, 192,518 silent, and 71,207 truncating or small in-frame mutations (**Supplementary Fig. 1**). Focusing on missense mutations, we observed that the relative proportion of amino acids affected by mutations varied considerably between cancer types (**Fig. 2**). These amino acid mutation biases are due to a combination of variations in the codon usage between different amino acids and the variations in the base-pair transitions and transversions observed between different cancer types (Lawrence *et al.*, 2014b; Alexandrov *et al.*, 2013). Because of the high mutation rate of CG dinucleotides across all cancers, arginine (R) is the most frequently altered amino acid despite being the 9^{th} most common amino acid as CG dinucleotides are present in four out of six of arginine's codons (**Supplementary Fig. 2**)

We next mapped the mutations to conserved protein domains obtained from the database of protein domain families, Pfam-A version 26.0 (Punta *et al.*, 2011) (**Fig. 1A**). Overall, 4401 of 4758 unique Pfam domains in the human genome were mutated at least once across all samples. The fraction of missense mutations that map to domains (46.7%, 216,676 of 463,842) was consistent across samples and tumor types and was similar to the proportion of the proteome assigned as conserved domains (45.4%, **Fig. 2**).

Identification of domains with enriched mutation burden

Our first aim was to identify domains that display an increased mutation burden. We defined the domain mutation burden as the total number of missense mutations in a domain, excluding domains only present in only one gene. After tallying mutations across samples, the domain with the highest mutation burden was the protein kinase domain with 7203 mutations in 353 genes (not including genes with tyrosine kinase domains), while the P53 domain present in *TP53*, *TP63*, and *TP73* had the most mutations when normalizing for the domain length and the size of the domain family (**Supplementary Fig. 3**). To systematically investigate if the mutation burden for a given domain was larger than would be expected by chance, we performed a permutation test that takes into account the number of mutations within and outside of the domain, the domain length, and the length and number of genes in the domain family. To specifically compare domain versus non-domain areas, other domains present in the domain-containing gene family were excluded. Assuming that each mutation is an independent event and that all residues of the protein have an equal chance of being mutated, we randomly reassigned all mutations 10^6 times across each gene separately and calculated if the observed domain mutation burden was significantly different from the distribution of burdens observed by chance (**Fig. 1B**).

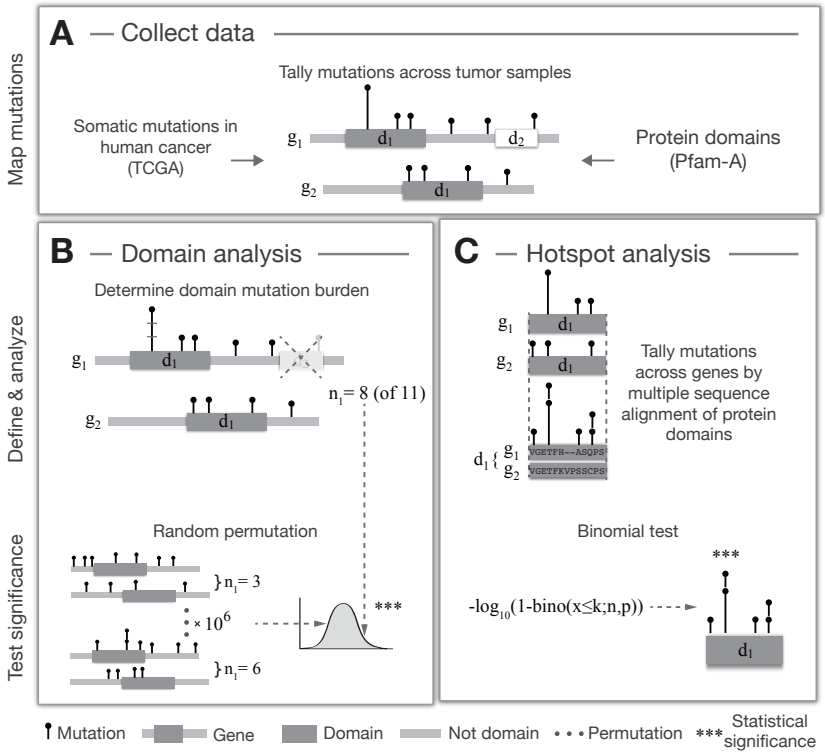

Figure 1: Work flow for analyzing recurrently mutated protein domains in cancer. (**A**) Missense mutation data from recent genomic profiling projects of human cancers (TCGA) are collected and all mutations are tallied across tumor samples and cancer types. Mutations are mapped to protein domains obtained from the Pfam-A database, which contains a manually curated set of highly conserved domain families in the human proteome. Two separate analyses are performed on this data to (**B**) identify domains enriched for missense mutations and (**C**) to detect mutation hotspots in domains through multiple sequence alignment. In the first analysis (**B**), the observed mutation burden (n_1) of a specific domain (d_1) is calculated by counting the total number of mutations in all domain-containing genes ($g_1 \& g_2$). Mutations in other domains (*e.g.*, d_2) are excluded. A permutation test is applied to determine if the observed mutation burden ($n_1 = 8$) is larger than expected by chance. Mutations are randomly shuffled 10^6 times across each gene separately and the observed mutation count is compared to the distribution of randomly estimated mutation counts. In the second analysis (**C**), domains are aligned across related genes by multiple sequence alignment and mutations are tallied at each residue of the alignment. A binomial test is applied to determine if the number of mutations at a specific residue is significantly different than the number of mutations observed at other residues of the alignment.

Using this permutation approach, we identified 14 domains that were significantly enriched for missense mutations within the domain boundaries compared to other areas of the same genes ($p < 0.05$, Bonferroni corrected, **Fig. 3** and **Table 1**). As both the number of gene members per domain (domain family) and the number of mutations per gene varies greatly, we wanted to distinguish between two cases: 1) only a single or a few genes in the domain family contributed to the domain mutation burden, and 2) genes contributed more evenly to the mutations in the domain. We were particularly interested in the latter as mutations in domains contributed by many infrequently mutated genes may represent new functional alterations that would not have been discovered using traditional gene-by-gene approaches. To investigate this, we calculated a

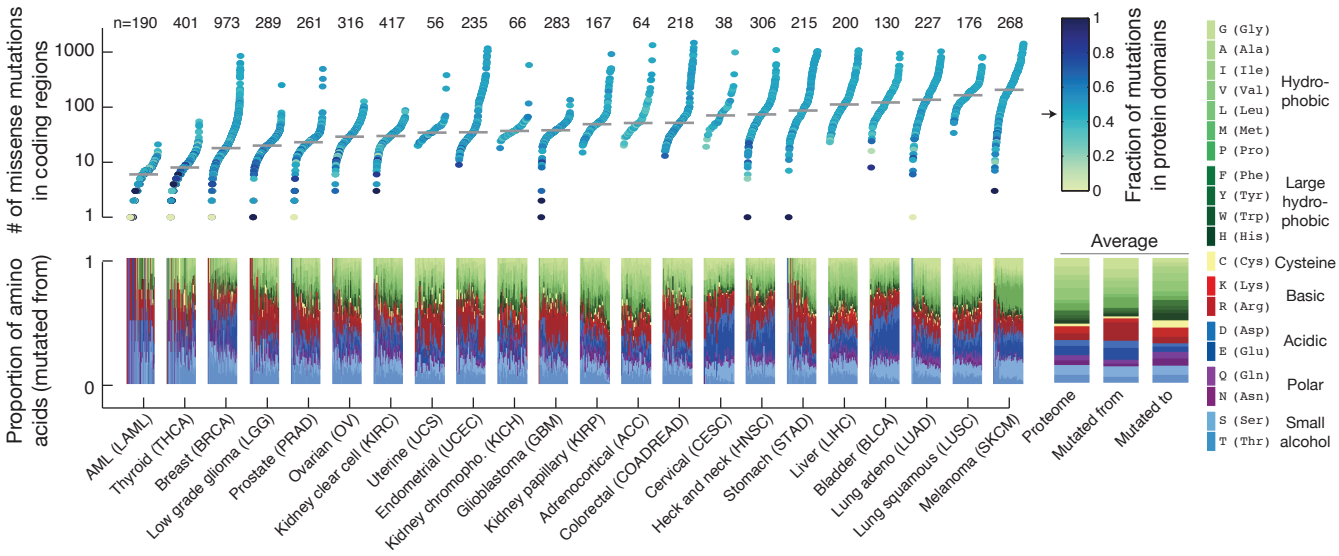

Figure 2: Mutations frequencies across cancer types and the relative proportion of mutated amino acid types. Within each cancer type the individual samples are ordered by the number of missense mutations in the proteome, and the median number of mutations is indicated (grey line). The color code represents 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). The amino acids are color coded by their biochemical properties as indicated. The number of samples in each cancer type is shown at the top. Samples with more than 2000 missense mutations were excluded from the analysis. Note that some amino acid types are disproportionally altered due to mutation biases in specific cancers (Lawrence *et al.*, 2014b; Alexandrov *et al.*, 2013), 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 proline (P).

entropy score ($\bar{S}$) that was normalized to the size of the domain family, so a low score indicates that the mutation burden is unevenly distributed between domain-containing genes and a high score indicates that the mutation burden is distributed evenly among the genes in the domain family (**see Methods**).

As expected, we found that the Von Hippel-Lindau (VHL) and the P53 domains were significantly enriched for mutations and had low entropy scores as they were dominated by mutations in the canonical tumor suppressor genes *VHL* and *TP53*, respectively (**Fig. 3** and **Table 1**, row 6 and 9). On the other end of the spectrum, the KAT11 domain encoding the lysine acetyltransferase (KAT) activity of *CREBBP* and *EP300* was significantly mutated and had a high normalized entropy score with around 30 mutations in each gene (**Table 1**, row 1). *CREBBP* and *EP300* are transcriptional co-activators that regulate gene expression through acetylation of lysine residues of histones and other transcription factors (Liu *et al.*, 2008). In our analysis, head and neck squamous cell carcinoma (HNSC) was the tumor type with most mutations in KAT11 and nearly half fell in the domain (14 of 29) that spans only about 4% of the length of both

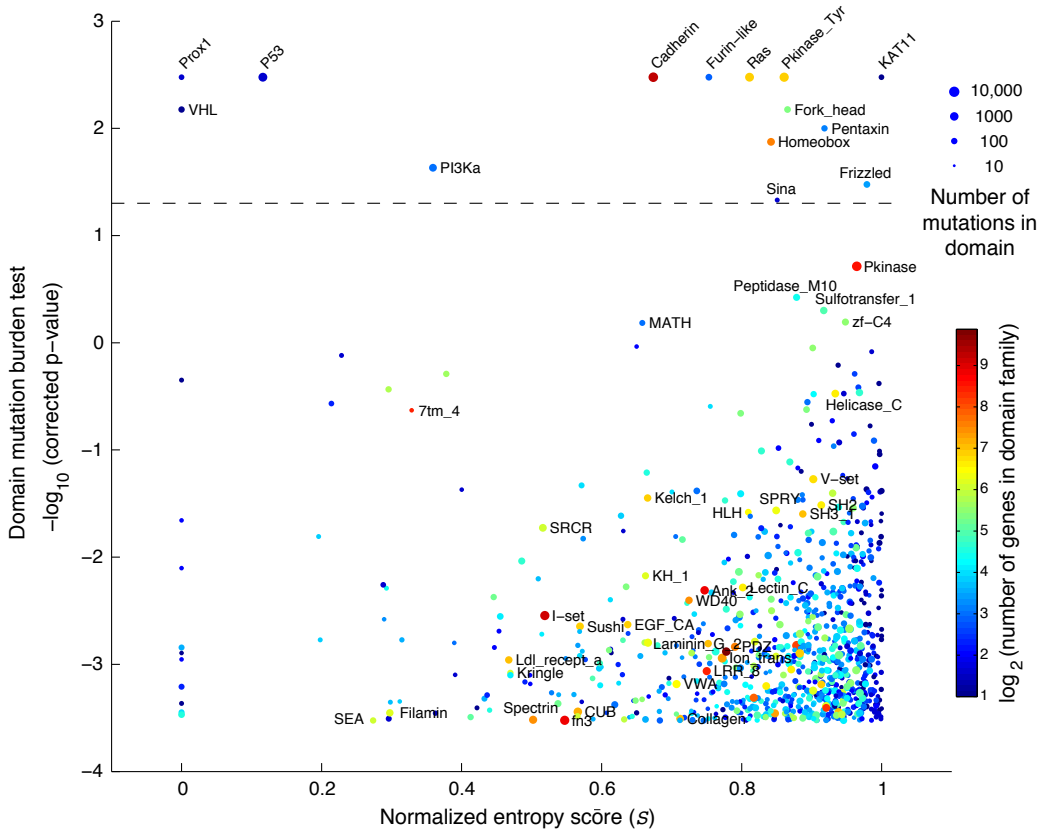

Figure 3: Identification of protein domains enriched for missense mutations. The estimated significance level of the domain mutation burden test is plotted against the domain entropy score ($\bar{S}$). The domain mutation burden test captures the enrichment of mutations within the domain boundaries compared to non-domain areas of the same genes. The entropy score captures the degree to which individual or multiple genes contribute to the mutations in the domain, where low and high scores indicate that the mutation burden is unevenly or evenly distributed between domain-containing genes, respectively. $\bar{S}$ is normalized to the highest possible score so that when $\bar{S} = 1$ all genes are evenly mutated and when $\bar{S} = 0$ then only one gene is mutated. Domains with a significant mutation burden are indicated above the dashed line ($p < 0.05$, Bonferroni corrected). The sizes of the dots reflect the number of mutations in each domain. Domains are color coded by the number of genes in the domain family.

CREBBP and *EP300* (**Supplementary Fig. 4**). Supporting the enrichment of mutations in this domain, inactivating mutations in KAT11 have been associated with oncogenesis in tumor types not part of this analysis, including B-cell lymphoma (Pasqualucci *et al.*, 2012; Cerchietti *et al.*, 2010; Morin *et al.*, 2012) and small-cell lung cancer (Peifer *et al.*, 2012). In small-cell lung cancer, *CREBBP* and *EP300* were reported to be deleted in a mutually exclusive fashion (Peifer *et al.*, 2012), which often indicates that genes are functionally linked (Ciriello *et al.*, 2012). In HNSC, we also found that mutations in *CREBBP* and *EP300* tend to occur in a mutually exclusive pattern (**Supplementary Fig. 4**), indicating a potential function role for both genes in this disease although only one of the two (*EP300*) has previously been linked to HNSC (Lawrence *et al.*, 2014a).

Confirming canonical mutation events in cancer, we found mutations clustering in domains of genes involved in receptor tyrosine kinases (RTKs) signaling, including the tyrosine kinase do-

Row	Domain	Genes (#)	p-val ($-\log_{10}$)	$\bar{S}$	Mutations (#)	e_d	Top Genes (Gene symbol and # of mutations)	Top Cancers (Cancer and # of mutations)
1	KAT11	2	2.48	1	65	1.94	CREBBP 33; EP300 32;	HNSC 14; BLCA 10;
2	Pkinase_Tyr	120	2.48	0.861	2059	1.17	BRAF 427; EGFR 50; ERBB2 43;	SKCM 428; THCA 246;
3	Ras	124	2.48	0.811	1369	1.08	KRAS 269; NRAS 177; HRAS 46;	SKCM 206; COADREAD 166;
4	Furin-like	7	2.48	0.753	147	1.77	EGFR 67; ERBB3 30; ERBB2 23;	GBM 42; LGG 14;
5	Cadherin	614	2.48	0.674	3358	1.06	FAT4 214; FAT3 184; FAT1 116;	SKCM 607; LUAD 474;
6	P53	3	2.48	0.116	1333	1.63	TP53 1301; TP63 23; TP73 9;	OV 182; BRCA 171;
7	Prox1	3	2.48	0	56	1.05	PROX1 56; PROX2 0;	HNSC 13; SKCM 11;
8	Fork_head	42	2.18	0.866	165	1.4	FOXA1 25; FOXA2 10; FOXK2 10;	BRCA 22; SKCM 18;
9	VHL	2	2.18	0	103	1.24	VHL 103; VHLL 0;	KIRC 95; SKCM 2;
10	Pentaxin	9	2	0.918	99	1.42	NPTX2 22; SVEP1 17; NPTXR 16;	SKCM 23; HNSC 13;
11	Homeobox	190	1.87	0.842	375	1.23	ZFHX4 28; NKX3-1 11; ONECUT2 11;	SKCM 48; UCEC 44;
12	PI3Ka	8	1.63	0.359	356	1.21	PIK3CA 296; PIK3CG 20; PIK3CB 13;	BRCA 126; UCEC 44;
13	Frizzled	11	1.48	0.979	164	1.21	FZD10 24; FZD9 20; FZD3 19;	BRCA 18; LUAD 18;
14	Sina	3	1.33	0.851	31	1.43	SIAH2 16; SIAH1 12; SIAH3 3;	UCEC 5; COADREAD 4;

Table 1: Protein domains significantly enriched for mutations. Domains are listed by their Pfam domain identifiers, the number of genes in the domain family, the Bonferroni-corrected p-value, the entropy score ($\bar{S}$), the number of mutations in the domain, the mutation enrichment score (e_d) expressed as the ratio of the observed number of domain mutations to the expected number of domain mutations, the genes with the most domain mutations, and the two cancers with most domain mutations. The genes are color coded based on being reported as significantly mutated (green) or not (magenta) in any cancer type in a recent pan-cancer study (Lawrence *et al.*, 2014a). The list is sorted by p-value followed by entropy score.

main itself (Pkinase_Tyr), the furin-like domain involved in RTK aggregation, and downstream signaling through genes with the ras GTPase domain and the phosphatidylinositol 3-kinase (PI3Ka) domain (**Table 1**, row 2, 3, 4 and 12). These domains have also been reported in other systematic studies of mutations in domains (Yue *et al.*, 2010; Nehrt *et al.*, 2012), consistent with the fact that the RTK signaling pathways are often high-jacked in cancer (Hanahan & Weinberg, 2011). In a similar manner, we identified multiple domains in genes that have previously been associated with cancer, including the DNA-binding forkhead domain in Fox family transcription factors and the frizzled domain in G protein-coupled receptors of the Wnt signaling pathway. Interestingly, these domains have high entropy scores with a substantial amount of mutations contributed by genes not reported as altered in a recent pan-cancer study (Lawrence *et al.*, 2014a) (see color code in **Table 1**, row 8 and 13). Thus, from the perspective of the structure-function relationship encoded in domains, these are candidate cancer driver genes due to the enrichment of mutations in these functional regions.

We also identified several domain families in which most of the genes had no apparent link to cancer. Such domains include the homeobox domain involved in DNA-binding and the cadherin domain involved in cell adhesion (**Table 1**, row 5 and 11). As cell-cell adhesion and DNA-binding are critical cellular processes, it is plausible that domain-contained genes involved in these processes are under positive selective pressure in the cancer environment, although it remains to be tested if mutations in these domains are functionally disruptive and may play a critical role in cancer. Several additional domains were found to be enriched for mutations and may potentially be of interest in a cancer context (**Table 1**).

Protein domain alignment reveals mutation hotspots across related genes

We next aligned each domain using multiple sequence alignment and tallied mutations across analogous residues of domain-containing genes (**Fig. 1C**). The goals of this analysis were to identify new domain hotspots with recurrent mutations across functionally related genes and to associate hotspots in well-established cancer genes with rare events in genes not previously linked to cancer. We used a binomial test to determine if a mutation peak at a specific residue was significantly different from other residues in the domain alignment, and we applied the same entropy analysis to investigate the degree to which individual or multiple genes contributed mutations to each hotspot. In total we identified 82 significant hotspots in 42 different domains (**Supplementary Table 1**).

We recapitulated several well-known hotspots in domains where only one gene was mutated such as the P53 and PI3Ka domains with mutations in *TP53* and *PIKC3A*, respectively (entropy$\simeq$0, **Fig. 4** and **Table 2**, row 13, 14, and 17). We also confirmed several known domain-specific hotspots such as the isocitrate/isopropylmalate dehydrogenase domain (Iso_dh) with homologous mutations in *IDH1* (position R132) and *IDH2* (R172) as well as the ras domain with mutations in *KRAS*, *NRAS*, and *HRAS* at positions G12, G13, and Q61 in the GTP binding region (**Table 2**, row 3, 7, 8, and 10). Furthermore, we found that well-characterized hotspots in *KIT* D816 in acute myeloid leukemia (AML), *FLT3* D835 in AML, and *BRAF* V600 in thyroid carcinoma and melanoma aligned perfectly in the conserved activation segment of the tyrosine kinase domain (**Table 2**, row 11). These mutations are known to cause constitutive kinase activity, which promotes cell proliferation independent of normal growth factor control (Hanahan & Weinberg, 2011; Dibb *et al.*, 2004). We further superimposed the crystal structures of the three proteins and found that the residues overlap in structure space (**Supplementary Fig. 5**), offering support that the alignment approach captures structurally relevant information. Notably, in the same domain hotspot many singleton mutations in lung adenocarcinoma and lung squamous cell carcinoma mapped to the equivalent position in other RTKs, including *EPHA2* V763M, *FGFR1* D647N, *PDGFRA* D842H, and three mutations in *EGFR* L861Q. Although these are rare events in lung cancer, this analysis reveals that they likely affect the same activation loop residue and may be therapeutically actionable in a similar manner as the hotspot mutations in *KIT*, *FLT3*, and *BRAF*. Encouragingly, non-small cell lung cancer patients with *EGFR* L861 mutations have recently shown positive clinical response when treated with *EGFR*-targeted therapy (Wu *et al.*, 2011).

Similar to the previous analysis of entropy in recurrently mutated domains, we were interested in domain hotspots with high entropy scores. Again, the lysine acetylase domain, KAT11, was identified with high entropy for a significant hotspot at position 94 of the domain alignment with mutations in *EP300* at D1399 and *CREBBP* at D1435 (**Table 2**, row 1). These sites are located in the substrate binding loop of KAT11 and mutations in these residues affect the structural conformation of the substrate binding loop (Liu *et al.*, 2008). Recently, both genes have been implicated in other cancers not analyzed here such as small-cell lung cancer (Peifer *et al.*, 2012) and B-cell lymphoma (Pasqualucci *et al.*, 2012; Cerchietti *et al.*, 2010; Morin *et al.*, 2012). Confirming the functional relevance of the identified hotspot, both *EP300* D1399 and *CREBBP* D1435 mutations have been found to reduce lysine acetylase activity *in vitro* (Peifer *et al.*, 2012; Pasqualucci *et al.*, 2012; Liu *et al.*, 2008). We additionally identified a potential hotspot in KAT11 at position 105 with mutations in *CREBBP* (R1446) although this hotspot was not significant when correcting

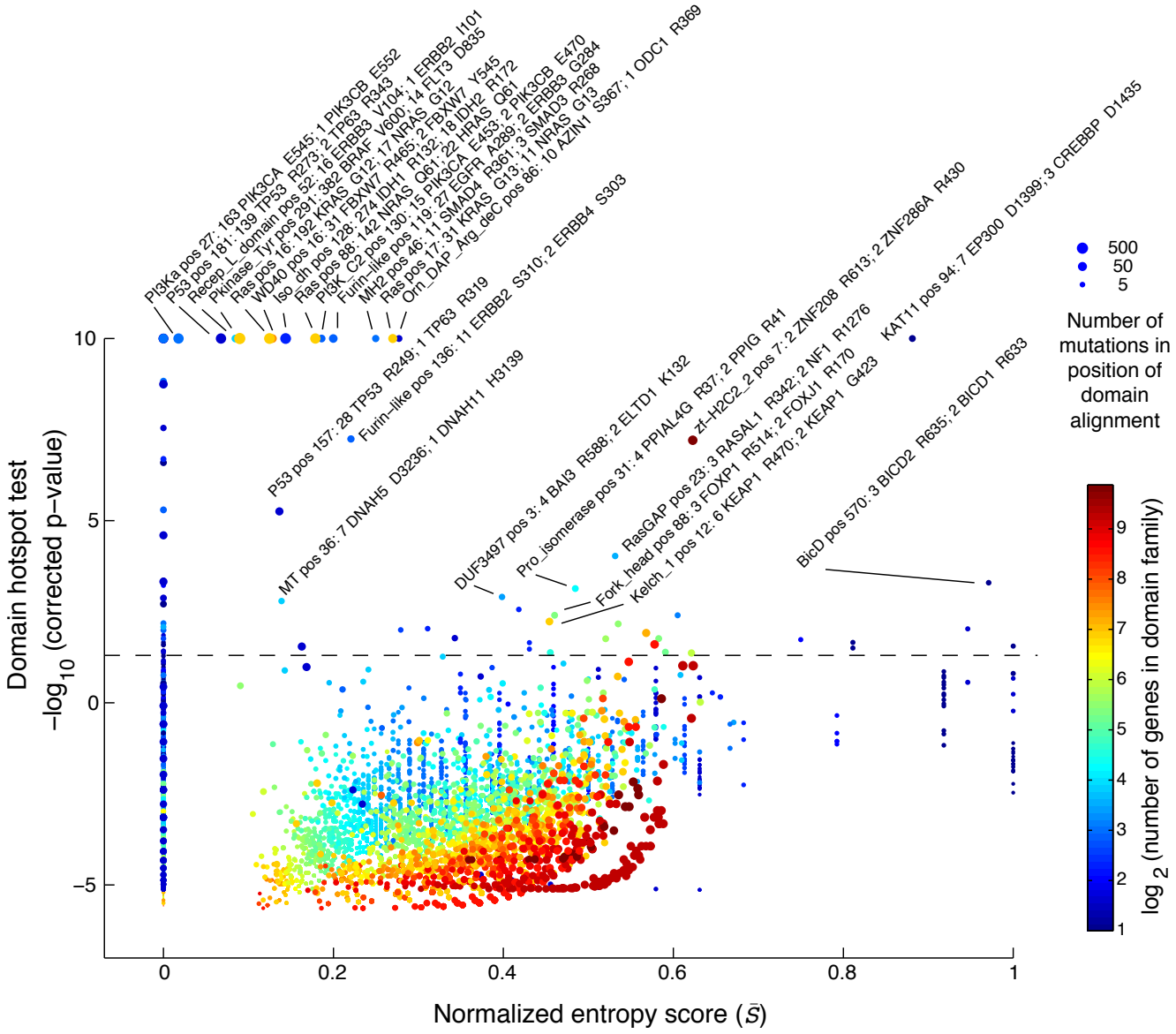

Figure 4: Domain alignment detects mutation hotspots across related genes. The estimated significance level of each mutation hotspot in the domain alignment is plotted against the domain entropy score ($\bar{S}$), which is described elsewhere. Significant hotspots are indicated above the dashed line ($p < 0.05$, Bonferroni corrected). The maximal significance was set to 10 [-$\log_{10}$(p-value)]. Hotspots are named by the Pfam identifiers followed by the position in the domain alignment and the number of mutations in the top two mutated genes. The size of the dots reflects the number of mutations at each residue and the dots are color coded by the number of domain-containing genes in the genome.

for multiple hypothesis testing ($p = 2.6e^{-6}$, corrected $p = 0.59$, **Fig. 5A**). *CREBBP* R1446 is also located within the substrate binding loop (Liu *et al.*, 2008) and R1446 mutations have been found in B-cell neoplasms (Pasqualucci *et al.*, 2012).

Row	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)
1	KAT11	2	94	10	0.88	10	7 EP300 D1399	3 CREBBP D1435	3 BLCA
2	Orn_DAP_Arg_deC	3	86	10	0.28	11	10 AZIN1 S367	1 ODC1 R369	10 LIHC
3	Ras	124	17	10	0.27	56	31 KRAS G13	11 NRAS G13	12 COADREAD
4	MH2	8	46	10	0.25	14	11 SMAD4 R361	3 SMAD3 R268	8 COADREAD
5	Furin-like	7	119	10	0.20	30	27 EGFR A289	2 ERBB3 G284	23 GBM
6	PI3K_C2	7	130	10	0.19	17	15 PIK3CA E453	2 PIK3CB E470	7 BRCA
7	Ras	124	88	10	0.18	189	142 NRAS Q61	22 HRAS Q61	78 SKCM
8	Iso_dh	5	128	10	0.14	292	274 IDH1 R132	18 IDH2 R172	232 LGG
9	WD40	170	16	10	0.13	37	31 FBXW7 R465	2 FBXW7 Y545	12 COADREAD
10	Ras	124	16	10	0.12	224	192 KRAS G12	17 NRAS G12	74 COADREAD
11	Pkinase_Tyr	120	291	10	0.09	415	382 BRAF V600	14 FLT3 D835	235 THCA
12	Recep_L_domain	14	52	10	0.08	17	16 ERBB3 V104	1 ERBB2 I101	5 COADREAD
13	P53	3	181	10	0.07	141	139 TP53 R273	2 TP63 R343	44 LGG
14	PI3Ka	8	27	10	0.02	164	163 PIK3CA E545	1 PIK3CB E552	66 BRCA
15	Furin-like	7	136	7.24	0.22	13	11 ERBB2 S310	2 ERBB4 S303	4 STAD
16	zf-H2C2_2	940	7	7.21	0.62	79	2 ZNF208 R613	2 ZNF286A R430	27 UCEC
17	P53	3	157	5.25	0.14	29	28 TP53 R249	1 TP63 R319	8 LIHC
18	RasGAP	12	23	4.03	0.53	8	3 RASAL1 R342	2 NF1 R1276	2 HNSC
19	BicD	2	570	3.3	0.97	5	3 BICD2 R635	2 BICD1 R633	2 SKCM
20	Pro_isomerase	19	31	3.13	0.48	9	4 PPIAL4G R37	2 PPIG R41	7 SKCM
21	DUF3497	11	3	2.91	0.40	7	4 BAI3 R588	2 ELTD1 K132	4 SKCM
22	MT	15	36	2.79	0.14	8	7 DNAH5 D3236	1 DNAH11 H3139	8 SKCM
23	Choline_transpo	5	186	2.56	0.42	5	3 SLC44A1 R437	2 SLC44A4 R496	1 BRCA
24	bZIP_2	9	21	2.4	0.61	6	2 HLF R243	2 NFIL3 R91	2 UCEC
25	Fork_head	42	88	2.4	0.46	11	3 FOXP1 R514	2 FOXJ1 R170	4 COADREAD

Table 2: Identified mutation hotspots 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 two genes with the most mutations in the hotspot, and the cancer type with most mutations in the hotspot. The genes are color coded based on being reported as significantly mutated (green) or not (magenta) in a recent pan-cancer study (Lawrence *et al.*, 2014a). The list is sorted by p-value followed by entropy score. Hotspots in domains where only one gene was mutated ($S = 0$) were excluded. All significant domain hotspots (82) are provided in (**Supplementary Table 1**).

Associating rare mutations with known oncogenic hotspots

The MAD homology 2 (MH2) domain is found in *SMAD* genes and mediates interaction between SMAD proteins and their interaction partners through recognition of phosphorylated serine residues (Wu *et al.*, 2001). We found the known R361H/C hotspot mutation in *SMAD4* (Shi *et al.*, 1997; Ohtaki *et al.*, 2001) aligned with three R268H/C mutations in *SMAD3* (**Table 2**, row 4). In both proteins these residues are located in the conserved loop/helix region that is directly involved in binding *TGFBR1* (Shi *et al.*, 1997). R361C mutations inactivate the tumor suppressor *SMAD4* (Shi *et al.*, 1997), and recently, R268C mutations in *SMAD3* were also found to repress SMAD3-mediated signaling (Fleming *et al.*, 2013), supporting our association of rare arginine mutations in *SMAD3* with known inactivating mutations in *SMAD4*. The majority of the mutations in the hotspot were from colorectal adenocarcinoma samples (COADREAD) and it is known that *SMAD* genes are recurrently mutated in this disease (Fleming *et al.*, 2013). Interestingly, we found a tendency towards better survival for patients with hotspot mutations in colorectal cancer although more data is needed to confirm this, to our knowledge, unreported observation (**Supplementary Fig. 6**).

We associated several known hotspots in well-characterized cancer genes with rare but poten-

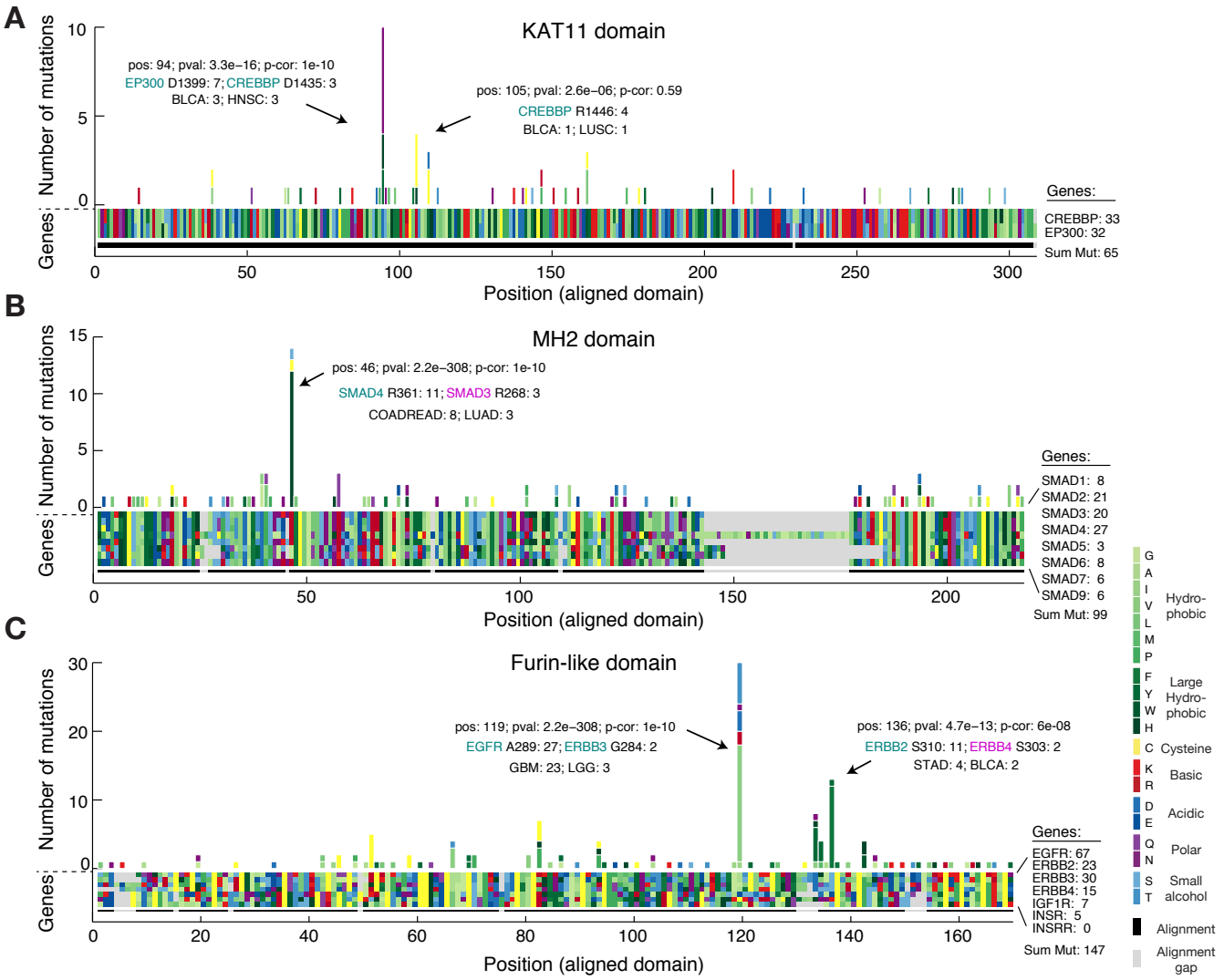

Figure 5: Multiple sequence alignment of domains identifies mutation hotspots and associates rare mutations with known oncogenic hotspots. (**A**) The amino acid sequence alignment of the KAT11 histone acetylate domain in *CREBBP* (position 1342–1648) and *EP300* (position 1306–1612) is represented as a block of two by 308 rectangles. Using the resulting alignment coordinates, missense mutations are tallied across the domains of the two genes. Both amino acids of the alignment (block) and the resulting amino acids due to mutations (histogram) are color coded by their biochemical properties. Alignment gaps are indicated by gray rectangles. Significant hotspots are indicated with position in alignment (pos), p-value (pval), Bonferoni corrected p-value (p-cor), and number of mutations in top mutated genes and cancer types. Similar plots are shown for the MAD homology 2 (MH2) domain involved in SMAD protein-protein interactions (**B**) and the furin-like domain involved in RTK aggregation and signal activation (**C**).

tially functional mutations in genes not frequently mutated in cancer. For example, we found rare mutations in *PIK3CB* at E470 and at E552 in the PI3K_C2 domain and PI3Ka domain, respectively, that associated with known recurrent hotspots in *PIK3CA* (**Table 2**, row 6 and 14). Furthermore, in the cysteine-rich Furin-like domain, which is involved in receptor aggregation and signaling activation of ERBB-family RTKs, we identified several significant hotspots including rare mutations in *ERBB3* (G284R) and *ERBB2* (A293V) that aligned with the known activat-

ing driver mutations in *EGFR* (A289V/T) in glioblastoma (Lee *et al.*, 2006) (**Fig. 5C**). Recently, one of these mutations, *ERBB3* G284R, was found to promote tumorigenesis in mice (Jaiswal *et al.*, 2013), suggesting that the singleton *ERBB2* A293V mutation found in a melanoma sample could represent an infrequent oncogenic event. We also identified a hotspot at position 137 of the alignment with rare S303F mutations in *ERBB4* aligning with S310F/Y mutations in *ERBB2*. Interestingly, in a functional analysis of *ERBB2* mutations in lung cancer cell lines, S310F/Y mutations were found to increase *ERBB2* signaling activity, promote tumorigenesis, and enhance sensitivity to *ERBB2* inhibitors *in vitro* (Greulich *et al.*, 2012). Future work will show if analogous mutations in *ERBB4* (S303F) may have similar effects.

Identification of new hotspots in protein domains

We also identified several additional hotspots in domains with mutations in genes not previously associated with cancer. We detected a hotspot in the prolyl isomerase domain (Pro_isomerase) with nine mutations distributed between *PPIAL4G* (R37C), *PPIG* (R41C), *PPIA* (R37C), *PPIE* (R173C) and *PPIL2* (I308F) (**Fig. 6A**). The Pro_isomerase domain-containing genes catalyze cis-trans isomerization of proline imidic peptide bonds and have been implicated in folding, transport, and assembly of proteins (Göthel & Marahiel, 1999). Seven of the nine mutations found in this hotspot were from melanoma samples, and interestingly we found that in melanoma these mutations correlate with significant upregulation of about a dozen genes including the cancer-testis antigens *CTAG2*, *CTAG1B*, *CSAG2*, and *CSAG3* (**Supplementary Fig. 7**).

The forkhead domain mediates DNA binding of forkhead box (Fox) transcription factors and encodes a conserved "winged helix" structure comprising three α-helices and three β-sheets flanked by one or two "wing"-like loops (Carlsson & Mahlapuu, 2002). In the forkhead domain, we identified a hotspot with 11 mutations distributed between *FOXP1* (R514C/H), *FOXK2* (R307C/H), *FOXK1* (R354W), *FOXJ1* (R170G/L), and *FOXP4* (R516C) in several different cancer types (**Fig. 6B**). The identified hotspot was located in the third α-helix (H3), which exhibits a high degree of sequence homology across Fox proteins and binds to the major groove of DNA targets. Specifically, the arginine residue that we found mutated forms direct hydrogen bonding with DNA in both in *FOXP* and *FOXK* family transcription factors (Wu *et al.*, 2006; Stroud *et al.*, 2006; Chu *et al.*, 2011) (Tsai *et al.*, 2006) (**Fig. 6C, D**). Furthermore, experimental R307A substitution in *FOXK2* abolishes DNA binding (Tsai *et al.*, 2006), suggesting that the identified arginine mutations may play an important role in cancer by inhibiting DNA-binding of *FOXP*, *FOXK*, and related Fox transcription factors.

We identified several other domain hotspots of potential interest such as a hotspot in the ras-GAP GTPase activating domain with mutations in the tumor suppressors *NF1*, *RASA1*, and *RASAL1* (**Supplementary Fig. 8**) and a hotspot in the kelch motif (Kelch_1 domain) with mutations in *KEAP1* and *KLHL4* (**Supplementary Fig. 9**). The potential biological consequence of these mutations remains to be elucidated. Many additional domain hotspots were identified and we make all analysis of hotspots in protein domains available via an interactive web-service at `http://www.mutationaligner.org`.

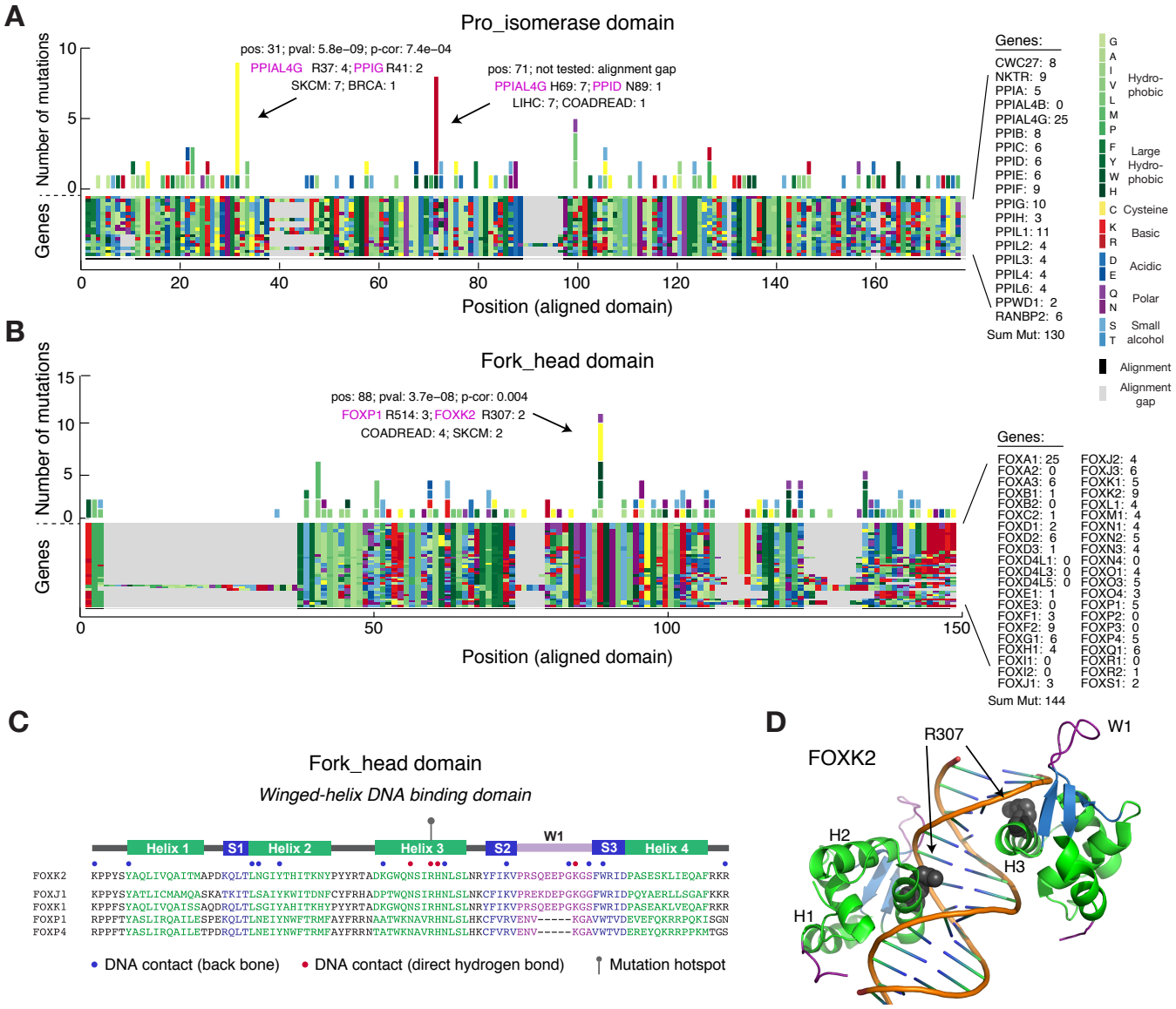

Figure 6: Identification of new hotspots affecting conserved residues of protein domains. Missense mutations are tallied across multiple sequence alignments of genes containing the prolyl isomerase (Pro_isomerase) domain (**A**) and the forkhead domain (**B**). (**C**) Secondary structure of the forkhead domain consisting of four α-helices (H1-4), three β-sheets (S1-3), and one wing-like loop (W1). Sequences are shown for selected Fox transcription factors that had mutations in the identified hotspot in H3. Of note, the selected genes have a fourth α-helix rather than the canonical second wing-like loop found in other Fox genes. (**D**) Ribbon drawing of the crystal structure of two *FOXK2* forkhead domains binding to a 16-bp DNA duplex containing a promoter sequence (pdb ID: 2C6Y) (Tsai *et al.*, 2006). The R307 residue that we identified as mutated in the hotspot is shown with spheres.

DISCUSSION

In this work we used protein domains, rather than individual genes, as the basis for the discovery of cancer-relevant alterations. By coupling the observation of mutations across common members of a domain family together, we identified domains enriched for mutations as well as mutation hotspots within these domains. Many of these domain mutations were contributed by groups of relatively rare mutations which that otherwise have been considered as spurious passengers. We further associated putative function with infrequent mutations as we identified hotspots where rare and well-characterized mutations affected analogous residues in the domain alignments. Based on the structure-function relationship encoded in domains, these rare mutation events may potentially be therapeutically actionable in cases where drugs have been developed to target related genes sharing the mutated domain. Finally, we identified entirely new hotspots in domains with mutations in genes not previously associated with cancer, hereby nominating new potential cancer-related genes for further investigation.

The fundamental assumptions underlying this work are that mutations at analogous sites of a domain family have a common effect and that recurrent mutations in domains are likely associated with cancer. As many protein domains have been functionally characterized, one of the strengths of our approach is that such knowledge can provide mechanistic insight into the potential effect of alterations. For example, we confirmed that canonical signaling domains present in large gene families are enriched for mutations (*e.g.*, tyrosine kinase and ras domains), reflecting the fact that a range of genes involved in mitogenic signaling are often high-jacked in cancer (Hanahan & Weinberg, 2011). We also identified less characterized domains that were significantly altered, such as the homeobox domain of DNA-binding genes and the cadherin domain of cell adhesion genes. While the individual domain-containing genes are not recurrently altered, these new findings suggest that mutations in the DNA-binding and cell adhesion machinery as general phenomena may confer a selective advantage in cancer. Similarly, in the prolyl isomerase domain, we identified a new hotspot in melanoma with arginine to cysteine mutations in *PPIA*, *PPIG*, and *PPIAL4G*. We also identified a new hotspot in the forkhead domain, where the crucial DNA-contacting arginine residue in the third helix of the forkhead-encoded winged-helix structure was mutated in several *FOXP* and *FOXK* family transcription factors. We speculate that this is a novel inactivating oncogenic event, although more research is needed to elucidate this.

Relatively few genes are recurrently mutated in cancer and the majority of somatic mutations are observed in infrequently mutated genes (Stephens *et al.*, 2013; Garraway & Lander, 2013). As above, we can use the biological knowledge associated with protein domains to help interpret the consequences of rare mutations in well-characterized domain families. For example, in the furin-like domain, we associated known oncogenic *ERRB2* S310F/Y mutations with uncharacterized *ERBB4* S303F mutations. In both genes, a small amino acid with an alcohol side chain (S) is mutated to a large hydrophobic amino acid (F or Y) in a conserved region involved in receptor interaction and signal activation. S310F/Y mutations in *ERBB2* are tumorigenic *in vitro* and it has been speculated that S310F/Y mutations promote hydrophobic interactions and receptor dimerization resulting in receptor activation (Greulich *et al.*, 2012). The same report also found that S310F mutations sensitize cell lines to the RTK inhibitors neratinib, afatinib, and lapatinib. We identified *ERBB4* S303F mutations in breast and endometrial cancers and predict that *ERBB4* S303F mutations are gain-of-function mutations that increases sensitivity to small-molecule inhibition and therefore represents a rare but druggable oncogenic event. Additionally, in the MH2

domain we associated known loss-of-function mutations *SMAD4* (R361H/C) in colorectal and lung cancers with rare but potentially functional mutations in *SMAD3* (R268H/C) in the same cancers. These examples illustrate how the association of mutations in infrequently altered genes with known mutations can provide high-confidence predictions and hypotheses for further experimental testing.

By analyzing mutations in a set of functionally related genes, we introduce the risk of detecting false positive hotspots where passenger mutations are aggregated across domain-containing genes. The detection of spurious domain hotspots could be exacerbated by mutation biases that alter amino acids disproportionally (**Fig. 2**). For example, arginine is the most frequently altered amino acid as it has four codons containing CG dinucleotides, which are frequently subject to C→T transitions due to deamination of methylated cytosine to thymine. In several domains, such as the tetramerisation (K_tetra) domain of potassium channel proteins (**Supplementary Table 1**) and the zf-H2C2_2 domain of zinc finger proteins (**Table 2**, row 16), we identify significant hotspots where arginine mutations align across a large set of genes in the domain family. Although the frequency of such hotspots may be driven by the arginine mutation bias, the mutations themselves may nevertheless be functional. Thus, we do not penalize for amino acid mutation biases in the detection of hotspots. However, we do provide relevant information about the potential biases by calculating a normalized hotspot z-score that takes into account the relative amino acid mutation frequency in each cancer type (**Supplementary Table 1** and **Methods**).

Future work will aim to refine the analysis of mutations in domains and expanding the scope of our analysis to other functional elements in genes. Our focus here was on somatic missense mutations, but this requirement may be relaxed to include germ-line mutations or other somatic alterations (*e.g.,* truncating mutations and small in-frame insertions and deletions). Importantly, truncating mutations may not necessarily be localized to the domain regions as they affect the entire gene, which is why we excluded them from our current work. An additional extension of our work would be to implement a sliding window for peak detection of clusters of mutations in domain alignments. However, our own observations suggest that mutation hotspots are largely localized to single residues. Structural information can also be used to analyze the proximity of mutated residues in 3D space. Other types of regulatory protein motifs can be analyzed, including short linear motifs that guide protein phosphorylation by kinases, which has previously been shown to be enriched for mutations in cancer genes (Reimand & Bader, 2013; Reimand *et al.*, 2013). Finally, assessment of the functional impact of mutations using structural information and evolutionary sequence conservation, for example as applied in our mutation assessor method (Reva *et al.*, 2011), can be incorporated to provide additional insight into the potential role of mutations in cancer.

As more data become available, integrative approaches combining mutation evidence across multiple scales such as genes, domains, and signaling pathways will be needed to improve the computational pipelines for variant function prediction. To make the results of our analysis useful and relevant to the community at large, we have made all findings available through an interactive web service (`http://www.mutationaligner.org`) that will be continuously updated as new tumor samples are genomically profiled.

EXPERIMENTAL PROCEDURES

Mutation data and data preprocessing

All TCGA mutation data were obtained in MAF file format from the cBioPortal for cancer genomics data (Cerami *et al.*, 2012; Gao *et al.*, 2013). To filter out mutations in low expressed genes, which has been shown to be associated with mutation biases (Lawrence *et al.*, 2014b), 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. To filter out ultra-mutated cancers, samples with more than 2,000 non-silent mutations were disregarded. 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 (Lawrence *et al.*, 2014b) 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:

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

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 $P(n = k)$ 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

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

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.

Calculating z-scores for mutation counts

To provide an estimate whether the number of mutations at a given mutation hotspot (mutation count) is different from the mutation counts at other residues of the domain alignment, we calculated a z-score for each position of the alignment. Assuming a normal distribution, the z-score expresses the number of standard deviations a given hotspot is above the mean of the distribution of mutations counts observed in the alignment. Additionally, we calculated a normalized z-score that takes into account the biases in amino acid mutation frequencies observed in different cancers by calculating the z-score based on a normalized mutation count ($\hat{c}$):

$$\hat{c} = c \times \frac{f_p}{f_m} \tag{4}$$

where c is the observed mutation count, f_p is the background frequency of a given amino acid in the proteome, and f_m is the observed mutation frequency of a given amino acid in a specific cancer type.

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

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

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

$$\bar{S} = \frac{-\sum_{i=1}^{n} P(x_i) \ln P(x_i)}{\ln n} \tag{6}$$

where n is the dimension of the vector x.

Author contributions

M.L.M. and E.R. designed analysis. N.P.G. designed and developed the web-site. M.L.M., E.R., N.P.G, B.A.A., A.K., J.G., and N.S. analyzed data. M.L.M. conceived and developed the concept. M.L.M. and C.S. managed the project. All authors contributed to discussions and editing of the manuscript.

Acknowledgments

We acknowledge C. Kandoth for technical assistance. C. Carmona-Fontaine and A. Hanrahan for helpful discussions. V.A. Pedicord for helpful comments on the manuscript. This work was funded in part by the National Cancer Institute Cancer Genome Atlas grant (U24 CA143840).

Competing financial interests

The authors declare no competing financial interests.

REFERENCES

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S., & Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. Nature Chemical Biology *7*(4), 248–249.

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., & Stratton, M. R. (2013). Signatures of mutational processes in human cancer. Nature *500*(7463), 415–421.

Bell, D., Berchuck, A., Birrer, M., Chien, J., Cramer, D. W., Dao, F., Dhir, R., DiSaia, P., Gabra, H., Glenn, P., Godwin, A. K., Gross, J., Hartmann, L., Huang, M., Huntsman, D. G., Iacocca, M., Imielinski, M., Kalloger, S., Karlan, B. Y., Levine, D. A., Mills, G. B., Morrison, C., Mutch, D., Olvera, N., Orsulic, S., Park, K., Petrelli, N., Rabeno, B., Rader, J. S., Sikic, B. I., Smith-McCune, K., Sood, A. K., Bowtell, D., Penny, R., Testa, J. R., Chang, K., Dinh, H. H., Drummond, J. A., Fowler, G., Gunaratne, P., Hawes, A. C., Kovar, C. L., Lewis, L. R., Morgan, M. B., Newsham, I. F., Santibanez, J., Reid, J. G., Trevino, L. R., Wu, Y. Q., Wang, M., Muzny, D. M., Wheeler, D. A., Gibbs, R. A., Getz, G., Lawrence, M. S., Cibulskis, K., Sivachenko, A. Y., Sougnez, C., Voet, D., Wilkinson, J., Bloom, T., Ardlie, K., Fennell, T., Baldwin, J., Gabriel, S., Lander, E. S., Ding, L., Fulton, R. S., Koboldt, D. C., McLellan, M. D., Wylie, T., Walker, J., O'Laughlin, M., Dooling, D. J., Fulton, L., Abbott, R., Dees, N. D., Zhang, Q., Kandoth, C., Wendl, M., Schierding, W., Shen, D., Harris, C. C., Schmidt, H., Kalicki, J., Delehaunty, K. D., Fronick, C. C., Demeter, R., Cook, L., Wallis, J. W., Lin, L., Magrini, V. J., Hodges, J. S., Eldred, J. M., Smith, S. M., Pohl, C. S., Vandin, F., Raphael, B. J., Weinstock, G. M., Mardis, E. R., Wilson, R. K., Meyerson, M., Winckler, W., Getz, G., Verhaak, R. G. W., Carter, S. L., Mermel, C. H., Saksena, G., Nguyen, H., Onofrio, R. C., Lawrence, M. S., Hubbard, D., Gupta, S., Crenshaw, A., Ramos, A. H., Ardlie, K., Chin, L., Protopopov, A., Zhang, J., Kim, T. M., Perna, I., Xiao, Y., Zhang, H., Ren, G., Sathiamoorthy, N., Park, R. W., Lee, E., Park, P. J., Kucherlapati, R., Absher, D. M., Waite, L., Sherlock, G., Brooks, J. D., Li, J. Z., Xu, J., Myers, R. M., Laird, P. W., Cope, L., Herman, J. G., Shen, H., Weisenberger, D. J., Noushmehr, H., Pan, F., Triche Jr, T., Berman, B. P., Van Den Berg, D. J., Buckley, J., Baylin, S. B., Spellman, P. T., Purdom, E., Neuvial, P., Bengtsson, H., Jakkula, L. R., Durinck, S., Han, J., Dorton, S., Marr, H., Choi, Y. G., Wang, V., Wang, N. J., Ngai, J., Conboy, J. G., Parvin, B., Feiler, H. S., Speed, T. P., Gray, J. W., Levine, D. A., Socci, N. D., Liang, Y., Taylor, B. S., Schultz, N., Borsu, L., Lash, A. E., Brennan, C., Viale, A., Sander, C., Ladanyi, M., Hoadley, K. A., Meng, S., Du, Y., Shi, Y., Li, L., Turman, Y. J., Zang, D., Helms, E. B., Balu, S., Zhou, X., Wu, J., Topal, M. D., Hayes, D. N., Perou, C. M., Getz, G., Voet, D., Saksena, G., Zhang, J., Zhang, H., Wu, C. J., Shukla, S., Cibulskis, K., Lawrence, M. S., Sivachenko, A., Jing, R., Park, R. W., Liu, Y., Park, P. J., Noble, M., Chin, L., Carter, H., Kim,

D., Karchin, R., Spellman, P. T., Purdom, E., Neuvial, P., Bengtsson, H., Durinck, S., Han, J., Korkola, J. E., Heiser, L. M., Cho, R. J., Hu, Z., Parvin, B., Speed, T. P., Gray, J. W., Schultz, N., Cerami, E., Taylor, B. S., Olshen, A., Reva, B., Antipin, Y., Shen, R., Mankoo, P., Sheridan, R., Ciriello, G., Chang, W. K., Bernanke, J. A., Borsu, L., Levine, D. A., Ladanyi, M., Sander, C., Haussler, D., Benz, C. C., Stuart, J. M., Benz, S. C., Sanborn, J. Z., Vaske, C. J., Zhu, J., Szeto, C., Scott, G. K., Yau, C., Hoadley, K. A., Du, Y., Balu, S., Hayes, D. N., Perou, C. M., Wilkerson, M. D., Zhang, N., Akbani, R., Baggerly, K. A., Yung, W. K., Mills, G. B., Weinstein, J. N., Penny, R., Shelton, T., Grimm, D., Hatfield, M., Morris, S., Yena, P., Rhodes, P., Sherman, M., Paulauskis, J., Millis, S., Kahn, A., Greene, J. M., Sfeir, R., Jensen, M. A., Chen, J., Whitmore, J., Alonso, S., Jordan, J., Chu, A., Zhang, J., Barker, A., Compton, C., Eley, G., Ferguson, M., Fielding, P., Gerhard, D. S., Myles, R., Schaefer, C., Mills Shaw, K. R., Vaught, J., Vockley, J. B., Good, P. J., Guyer, M. S., Ozenberger, B., Peterson, J., & Thomson, E. (2011). Integrated genomic analyses of ovarian carcinoma. Nature *474*(7353), 609–615.

Bromberg, Y., Yachdav, G., & Rost, B. (2008). SNAP predicts effect of mutations on protein function. Bioinformatics (Oxford, England) *24*(20), 2397–2398.

Cancer Genome Atlas Network (2012). Comprehensive molecular characterization of human colon and rectal cancer.. Nature *487*(7407), 330–337.

Carlsson, P. & Mahlapuu, M. (2002). Forkhead Transcription Factors: Key Players in Development and Metabolism. Developmental Biology *250*(1), 1–23.

Cerami, E., Demir, E., Schultz, N., Taylor, B. S., & Sander, C. (2010). Automated Network Analysis Identifies Core Pathways in Glioblastoma. PloS one *5*(2), e8918.

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., & Schultz, N. (2012). The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. Cancer Discovery *2*(5), 401–404.

Cerchietti, L. C., Hatzi, K., Caldas-Lopes, E., Yang, S. N., Figueroa, M. E., Morin, R. D., Hirst, M., Mendez, L., Shaknovich, R., Cole, P. A., Bhalla, K., Gascoyne, R. D., Marra, M., Chiosis, G., & Melnick, A. (2010). BCL6 repression of EP300 in human diffuse large B cell lymphoma cells provides a basis for rational combinatorial therapy. Journal of Clinical Investigation *120*(12), 4569–4582.

Chothia, C., Gough, J., Vogel, C., & Teichmann, S. A. (2003). Evolution of the protein repertoire.. Science (New York, N.Y.) *300*(5626), 1701–1703.

Chu, Y.-P., Chang, C.-H., Shiu, J.-H., Chang, Y.-T., Chen, C.-Y., & Chuang, W.-J. (2011). Solution structure and backbone dynamics of the DNA-binding domain of FOXP1: Insight into its domain swapping and DNA binding. Protein Science *20*(5), 908–924.

Ciriello, G., Cerami, E., Sander, C., & Schultz, N. (2012). Mutual exclusivity analysis identifies oncogenic network modules. Genome Research *22*(2), 398–406.

Creighton, C. J., Morgan, M., Gunaratne, P. H., Wheeler, D. A., Gibbs, R. A., Gordon Robertson, A., Chu, A., Beroukhim, R., Cibulskis, K., Signoretti, S., Vandin Hsin-Ta Wu, F., Raphael, B. J., Verhaak, R. G. W., Tamboli, P., Torres-Garcia, W., Akbani, R., Weinstein, J. N., Reuter,

V., Hsieh, J. J., Rose Brannon, A., Ari Hakimi, A., Jacobsen, A., Ciriello, G., Reva, B., Ricketts, C. J., Marston Linehan, W., Stuart, J. M., Kimryn Rathmell, W., Shen, H., Laird, P. W., Muzny, D., Davis, C., Morgan, M., Xi, L., Chang, K., Kakkar, N., Treviño, L. R., Benton, S., Reid, J. G., Morton, D., Doddapaneni, H., Han, Y., Lewis, L., Dinh, H., Kovar, C., Zhu, Y., Santibanez, J., Wang, M., Hale, W., Kalra, D., Creighton, C. J., Wheeler, D. A., Gibbs, R. A., Getz, G., Cibulskis, K., Lawrence, M. S., Sougnez, C., Carter, S. L., Sivachenko, A., Lichtenstein, L., Stewart, C., Voet, D., Fisher, S., Gabriel, S. B., Lander, E., Beroukhim, R., Schumacher, S. E., Tabak, B., Saksena, G., Onofrio, R. C., Carter, S. L., Cherniack, A. D., Gentry, J., Ardlie, K., Sougnez, C., Getz, G., Gabriel, S. B., Meyerson, M., Gordon Robertson, A., Chu, A., Chun, H.-J. E., Mungall, A. J., Sipahimalani, P., Stoll, D., Ally, A., Balasundaram, M., Butterfield, Y. S. N., Carlsen, R., Carter, C., Chuah, E., Coope, R. J. N., Dhalla, N., Gorski, S., Guin, R., Hirst, C., Hirst, M., Holt, R. A., Lebovitz, C., Lee, D., Li, H. I., Mayo, M., Moore, R. A., Pleasance, E., Plettner, P., Schein, J. E., Shafiei, A., Slobodan, J. R., Tam, A., Thiessen, N., Varhol, R. J., Wye, N., Zhao, Y., Birol, I., Jones, S. J. M., Marra, M. A., Auman, J. T., Tan, D., Jones, C. D., Hoadley, K. A., Mieczkowski, P. A., Mose, L. E., Jefferys, S. R., Topal, M. D., Liquori, C., Turman, Y. J., Shi, Y., Waring, S., Buda, E., Walsh, J., Wu, J., Bodenheimer, T., Hoyle, A. P., Simons, J. V., Soloway, M. G., Balu, S., Parker, J. S., Neil Hayes, D., Perou, C. M., Kucherlapati, R., Park, P., Shen, H., Triche Jr, T., Weisenberger, D. J., Lai, P. H., Bootwalla, M. S., Maglinte, D. T., Mahurkar, S., Berman, B. P., Van Den Berg, D. J., Cope, L., Baylin, S. B., Laird, P. W., Creighton, C. J., Wheeler, D. A., Getz, G., Noble, M. S., DiCara, D., Zhang, H., Cho, J., Heiman, D. I., Gehlenborg, N., Voet, D., Mallard, W., Lin, P., Frazer, S., Stojanov, P., Liu, Y., Zhou, L., Kim, J., Lawrence, M. S., Chin, L., Vandin, F., Wu, H.-T., Raphael, B. J., Benz, C., Yau, C., Reynolds, S. M., Shmulevich, I., Verhaak, R. G. W., Torres-Garcia, W., Vegesna, R., Kim, H., Zhang, W., Cogdell, D., Jonasch, E., Ding, Z., Lu, Y., Akbani, R., Zhang, N., Unruh, A. K., Casasent, T. D., Wakefield, C., Tsavachidou, D., Chin, L., Mills, G. B., Weinstein, J. N., Jacobsen, A., Rose Brannon, A., Ciriello, G., Schultz, N., Ari Hakimi, A., Reva, B., Antipin, Y., Gao, J., Cerami, E., Gross, B., Arman Aksoy, B., Sinha, R., Weinhold, N., Onur Sumer, S., Taylor, B. S., Shen, R., Ostrovnaya, I., Hsieh, J. J., Berger, M. F., Ladanyi, M., Sander, C., Fei, S. S., Stout, A., Spellman, P. T., Rubin, D. L., Liu, T. T., Stuart, J. M., Ng, S., Paull, E. O., Carlin, D., Goldstein, T., Waltman, P., Ellrott, K., Zhu, J., Haussler, D., Gunaratne, P. H., Xiao, W., Shelton, C., Gardner, J., Penny, R., Sherman, M., Mallery, D., Morris, S., Paulauskis, J., & Burnett, K. (2013). Comprehensive molecular characterization of clear cell renal cell carcinoma. Nature *499*(7456), 43–49.

Dees, N. D., Zhang, Q., Kandoth, C., Wendl, M. C., Schierding, W., Koboldt, D. C., Mooney, T. B., Callaway, M. B., Dooling, D., Mardis, E. R., Wilson, R. K., & Ding, L. (2012). MuSiC: Identifying mutational significance in cancer genomes. Genome Research *22*(8), 1589–1598.

Dibb, N. J., Dilworth, S. M., & Mol, C. D. (2004). Switching on kinases: oncogenic activation of BRAF and the PDGFR family.. Nature Reviews Cancer *4*(9), 718–727.

Ding, L., Getz, G., Wheeler, D. A., Mardis, E. R., McLellan, M. D., Cibulskis, K., Sougnez, C., Greulich, H., Muzny, D. M., Morgan, M. B., Fulton, L., Fulton, R. S., Zhang, Q., Wendl, M. C., Lawrence, M. S., Larson, D. E., Chen, K., Dooling, D. J., Sabo, A., Hawes, A. C., Shen, H., Jhangiani, S. N., Lewis, L. R., Hall, O., Zhu, Y., Mathew, T., Ren, Y., Yao, J., Scherer, S. E., Clerc, K., Metcalf, G. A., Ng, B., Milosavljevic, A., Gonzalez-Garay, M. L., Osborne, J. R., Meyer, R., Shi, X., Tang, Y., Koboldt, D. C., Lin, L., Abbott, R., Miner, T. L., Pohl, C., Fewell, G., Haipek, C., Schmidt, H., Dunford-Shore, B. H., Kraja, A., Crosby, S. D., Sawyer,

C. S., Vickery, T., Sander, S., Robinson, J., Winckler, W., Baldwin, J., Chirieac, L. R., Dutt, A., Fennell, T., Hanna, M., Johnson, B. E., Onofrio, R. C., Thomas, R. K., Tonon, G., Weir, B. A., Zhao, X., Ziaugra, L., Zody, M. C., Giordano, T., Orringer, M. B., Roth, J. A., Spitz, M. R., Wistuba, I. I., Ozenberger, B., Good, P. J., Chang, A. C., Beer, D. G., Watson, M. A., Ladanyi, M., Broderick, S., Yoshizawa, A., Travis, W. D., Pao, W., Province, M. A., Weinstock, G. M., Varmus, H. E., Gabriel, S. B., Lander, E. S., Gibbs, R. A., Meyerson, M., & Wilson, R. K. (2008). Somatic mutations affect key pathways in lung adenocarcinoma. Nature *455*(7216), 1069–1075.

Fleming, N. I., Jorissen, R. N., Mouradov, D., Christie, M., Sakthianandeswaren, A., Palmieri, M., Day, F., Li, S., Tsui, C., Lipton, L., Desai, J., Jones, I. T., McLaughlin, S., Ward, R. L., Hawkins, N. J., Ruszkiewicz, A. R., Moore, J., Zhu, H. J., Mariadason, J. M., Burgess, A. W., Busam, D., Zhao, Q., Strausberg, R. L., Gibbs, P., & Sieber, O. M. (2013). SMAD2, SMAD3 and SMAD4 Mutations in Colorectal Cancer. Cancer Research *73*(2), 725–735.

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., & Schultz, N. (2013). Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal. Science signaling *6*(269), pl1–pl1.

Garraway, L. A. & Lander, E. S. (2013). Lessons from the Cancer Genome. Cell *153*(1), 17–37.

Getz, G., Gabriel, S. B., Cibulskis, K., Lander, E., Sivachenko, A., Sougnez, C., Lawrence, M., Kandoth, C., Dooling, D., Fulton, R., Fulton, L., Kalicki-Veizer, J., McLellan, M. D., O'Laughlin, M., Schmidt, H., Wilson, R. K., Ye, K., Ding, L., Mardis, E. R., Ally, A., Balasundaram, M., Birol, I., Butterfield, Y. S. N., Carlsen, R., Carter, C., Chu, A., Chuah, E., Chun, H.-J. E., Dhalla, N., Guin, R., Hirst, C., Holt, R. A., Jones, S. J. M., Lee, D., Li, H. I., Marra, M. A., Mayo, M., Moore, R. A., Mungall, A. J., Plettner, P., Schein, J. E., Sipahimalani, P., Tam, A., Varhol, R. J., Gordon Robertson, A., Cherniack, A. D., Pashtan, I., Saksena, G., Onofrio, R. C., Schumacher, S. E., Tabak, B., Carter, S. L., Hernandez, B., Gentry, J., Salvesen, H. B., Ardlie, K., Getz, G., Winckler, W., Beroukhim, R., Gabriel, S. B., Meyerson, M., Hadjipanayis, A., Lee, S., Mahadeshwar, H. S., Park, P., Protopopov, A., Ren, X., Seth, S., Song, X., Tang, J., Xi, R., Yang, L., Zeng, D., Kucherlapati, R., Chin, L., Zhang, J., Todd Auman, J., Balu, S., Bodenheimer, T., Buda, E., Neil Hayes, D., Hoyle, A. P., Jefferys, S. R., Jones, C. D., Meng, S., Mieczkowski, P. A., Mose, L. E., Parker, J. S., Perou, C. M., Roach, J., Shi, Y., Simons, J. V., Soloway, M. G., Tan, D., Topal, M. D., Waring, S., Wu, J., Hoadley, K. A., Baylin, S. B., Bootwalla, M. S., Lai, P. H., Triche Jr, T. J., Van Den Berg, D. J., Weisenberger, D. J., Laird, P. W., Shen, H., Chin, L., Zhang, J., Getz, G., Cho, J., DiCara, D., Frazer, S., Heiman, D., Jing, R., Lin, P., Mallard, W., Stojanov, P., Voet, D., Zhang, H., Zou, L., Noble, M., Lawrence, M., Reynolds, S. M., Shmulevich, I., Arman Aksoy, B., Antipin, Y., Ciriello, G., Dresdner, G., Gao, J., Gross, B., Jacobsen, A., Ladanyi, M., Reva, B., Sander, C., Sinha, R., Onur Sumer, S., Taylor, B. S., Cerami, E., Weinhold, N., Schultz, N., Shen, R., Benz, S., Goldstein, T., Haussler, D., Ng, S., Szeto, C., Stuart, J., Benz, C. C., Yau, C., Zhang, W., Annala, M., Broom, B. M., Casasent, T. D., Ju, Z., Liang, H., Liu, G., Lu, Y., Unruh, A. K., Wakefield, C., Weinstein, J. N., Zhang, N., Liu, Y., Broaddus, R., Akbani, R., Mills, G. B., Adams, C., Barr, T., Black, A. D., Bowen, J., Deardurff, J., Frick, J., Gastier-Foster, J. M., Grossman, T., Harper, H. A., Hart-Kothari, M., Helsel, C., Hobensack, A., Kuck, H., Kneile, K., Leraas, K. M., Lichtenberg, T. M., McAllister, C., Pyatt, R. E., Ramirez, N. C., Tabler, T. R., Vanhoose, N., White, P., Wise, L., Zmuda, E., Barnabas, N., Berry-Green, C., Blanc, V., Boice, L., Button, M., Farkas,

A., Green, A., MacKenzie, J., Nicholson, D., Kalloger, S. E., Blake Gilks, C., Karlan, B. Y., Lester, J., Orsulic, S., Borowsky, M., Cadungog, M., Czerwinski, C., Huelsenbeck-Dill, L., Iacocca, M., Petrelli, N., Rabeno, B., Witkin, G., Nemirovich-Danchenko, E., Potapova, O., Rotin, D., Berchuck, A., Birrer, M., DiSaia, P., Monovich, L., Curley, E., Gardner, J., Mallery, D., Penny, R., Dowdy, S. C., Winterhoff, B., Dao, L., Gostout, B., Meuter, A., Teoman, A., Dao, F., Olvera, N., Bogomolniy, F., Garg, K., Soslow, R. A., Levine, D. A., Abramov, M., Bartlett, J. M. S., Kodeeswaran, S., Parfitt, J., Moiseenko, F., Clarke, B. A., Goodman, M. T., & Carney, M. (2013). Integrated genomic characterization of endometrial carcinoma. Nature *497*(7447), 67–73.

Gonzalez-Perez, A. & Lopez-Bigas, N. (2012). Functional impact bias reveals cancer drivers. Nucleic Acids Research *40*(21), e169–e169.

Göthel, S. F. & Marahiel, M. A. (1999). Peptidyl-prolyl cis-trans isomerases, a superfamily of ubiquitous folding catalysts.. Cellular and molecular life sciences : CMLS *55*(3), 423–436.

Greenman, C., Stephens, P., Smith, R., Dalgliesh, G. L., Hunter, C., Bignell, G., Davies, H., Teague, J., Butler, A., Stevens, C., Edkins, S., O'Meara, S., Vastrik, I., Schmidt, E. E., Avis, T., Barthorpe, S., Bhamra, G., Buck, G., Choudhury, B., Clements, J., Cole, J., Dicks, E., Forbes, S., Gray, K., Halliday, K., Harrison, R., Hills, K., Hinton, J., Jenkinson, A., Jones, D., Menzies, A., Mironenko, T., Perry, J., Raine, K., Richardson, D., Shepherd, R., Small, A., Tofts, C., Varian, J., Webb, T., West, S., Widaa, S., Yates, A., Cahill, D. P., Louis, D. N., Goldstraw, P., Nicholson, A. G., Brasseur, F., Looijenga, L., Weber, B. L., Chiew, Y.-E., deFazio, A., Greaves, M. F., Green, A. R., Campbell, P., Birney, E., Easton, D. F., Chenevix-Trench, G., Tan, M.-H., Khoo, S. K., Teh, B. T., Yuen, S. T., Leung, S. Y., Wooster, R., Futreal, P. A., & Stratton, M. R. (2007). Patterns of somatic mutation in human cancer genomes. Nature *446*(7132), 153–158.

Greulich, H., Kaplan, B., Mertins, P., Chen, T.-H., Tanaka, K. E., Yun, C.-H., Zhang, X., Lee, S.-H., Cho, J., Ambrogio, L., Liao, R., Imielinski, M., Banerji, S., Berger, A. H., Lawrence, M. S., Zhang, J., Pho, N. H., Walker, S. R., Winckler, W., Getz, G., Frank, D., Hahn, W. C., Eck, M. J., Mani, D. R., Jaffe, J. D., Carr, S. A., Wong, K.-K., & Meyerson, M. (2012). Functional analysis of receptor tyrosine kinase mutations in lung cancer identifies oncogenic extracellular domain mutations of ERBB2.. Proceedings of the National Academy of Sciences *109*(36), 14476–14481.

Hammerman, P. S., Lawrence, M. S., Voet, D., Jing, R., Cibulskis, K., Sivachenko, A., Stojanov, P., McKenna, A., Lander, E. S., Gabriel, S., Getz, G., Sougnez, C., Imielinski, M., Helman, E., Hernandez, B., Pho, N. H., Meyerson, M., Chu, A., Chun, H.-J. E., Mungall, A. J., Pleasance, E., Gordon Robertson, A., Sipahimalani, P., Stoll, D., Balasundaram, M., Birol, I., Butterfield, Y. S. N., Chuah, E., Coope, R. J. N., Corbett, R., Dhalla, N., Guin, R., He, A., Hirst, C., Hirst, M., Holt, R. A., Lee, D., Li, H. I., Mayo, M., Moore, R. A., Mungall, K., Ming Nip, K., Olshen, A., Schein, J. E., Slobodan, J. R., Tam, A., Thiessen, N., Varhol, R., Zeng, T., Zhao, Y., Jones, S. J. M., Marra, M. A., Saksena, G., Cherniack, A. D., Schumacher, S. E., Tabak, B., Carter, S. L., Pho, N. H., Nguyen, H., Onofrio, R. C., Crenshaw, A., Ardlie, K., Beroukhim, R., Winckler, W., Hammerman, P. S., Getz, G., Meyerson, M., Protopopov, A., Zhang, J., Hadjipanayis, A., Lee, S., Xi, R., Yang, L., Ren, X., Zhang, H., Shukla, S., Chen, P.-C., Haseley, P., Lee, E., Chin, L., Park, P. J., Kucherlapati, R., Socci, N. D., Liang, Y., Schultz, N., Borsu, L., Lash, A. E., Viale, A., Sander, C., Ladanyi, M., Todd Auman, J., Hoadley, K. A., Wilkerson, M. D., Shi, Y., Liquori, C., Meng, S., Li, L., Turman, Y. J., Topal, M. D., Tan, D., Waring, S., Buda, E., Walsh,

J., Jones, C. D., Mieczkowski, P. A., Singh, D., Wu, J., Gulabani, A., Dolina, P., Bodenheimer, T., Hoyle, A. P., Simons, J. V., Soloway, M. G., Mose, L. E., Jefferys, S. R., Balu, S., O'Connor, B. D., Prins, J. F., Liu, J., Chiang, D. Y., Neil Hayes, D., Perou, C. M., Cope, L., Danilova, L., Weisenberger, D. J., Maglinte, D. T., Pan, F., Van Den Berg, D. J., Triche Jr, T., Herman, J. G., Baylin, S. B., Laird, P. W., Getz, G., Noble, M., Voet, D., Saksena, G., Gehlenborg, N., DiCara, D., Zhang, J., Zhang, H., Wu, C.-J., Yingchun Liu, S., Lawrence, M. S., Zou, L., Sivachenko, A., Lin, P., Stojanov, P., Jing, R., Cho, J., Nazaire, M.-D., Robinson, J., Thorvaldsdottir, H., Mesirov, J., Park, P. J., Chin, L., Schultz, N., Sinha, R., Ciriello, G., Cerami, E., Gross, B., Jacobsen, A., Gao, J., Arman Aksoy, B., Weinhold, N., Ramirez, R., Taylor, B. S., Antipin, Y., Reva, B., Shen, R., Mo, Q., Seshan, V., Paik, P. K., Ladanyi, M., Sander, C., Akbani, R., Zhang, N., Broom, B. M., Casasent, T., Unruh, A., Wakefield, C., Craig Cason, R., Baggerly, K. A., Weinstein, J. N., Haussler, D., Benz, C. C., Stuart, J. M., Zhu, J., Szeto, C., Scott, G. K., Yau, C., Ng, S., Goldstein, T., Waltman, P., Sokolov, A., Ellrott, K., Collisson, E. A., Zerbino, D., Wilks, C., Ma, S., Craft, B., Wilkerson, M. D., Todd Auman, J., Hoadley, K. A., Du, Y., Cabanski, C., Walter, V., Singh, D., Wu, J., Gulabani, A., Bodenheimer, T., Hoyle, A. P., Simons, J. V., Soloway, M. G., Mose, L. E., Jefferys, S. R., Balu, S., Marron, J. S., Liu, Y., Wang, K., Liu, J., Prins, J. F., Neil Hayes, D., Perou, C. M., Creighton, C. J., Zhang, Y., Travis, W. D., Rekhtman, N., Yi, J., Aubry, M. C., Cheney, R., Dacic, S., Flieder, D., Funkhouser, W., Illei, P., Myers, J., Tsao, M.-S., Penny, R., Mallery, D., Shelton, T., Hatfield, M., Morris, S., Yena, P., Shelton, C., Sherman, M., & Paulauskis, J. (2012). Comprehensive genomic characterization of squamous cell lung cancers. Nature *489*(7417), 519–525.

Hanahan, D. & Weinberg, R. A. (2011). Hallmarks of Cancer: The Next Generation. Cell *144*(5), 646–674.

Hofree, M., Shen, J. P., Carter, H., Gross, A., & Ideker, T. (2013). Network-based stratification of tumor mutations. Nature Chemical Biology *10*(11), 1108–1115.

Holm, L. & Sander, C. (1996). Mapping the protein universe.. Science (New York, N.Y.) *273*(5275), 595–603.

Jaiswal, B. S., Kljavin, N. M., Stawiski, E. W., Chan, E., Parikh, C., Durinck, S., Chaudhuri, S., Pujara, K., Guillory, J., Edgar, K. A., Janakiraman, V., Scholz, R.-P., Bowman, K. K., Lorenzo, M., Li, H., Wu, J., Yuan, W., Peters, B. A., Kan, Z., Stinson, J., Mak, M., Modrusan, Z., Eigenbrot, C., Firestein, R., Stern, H. M., Rajalingam, K., Schaefer, G., Merchant, M. A., Sliwkowski, M. X., de Sauvage, F. J., & Seshagiri, S. (2013). Oncogenic ERBB3 Mutations in Human Cancers. Cancer Cell *23*(5), 603–617.

Koboldt, D. C., Fulton, R. S., McLellan, M. D., Schmidt, H., Kalicki-Veizer, J., McMichael, J. F., Fulton, L. L., Dooling, D. J., Ding, L., Mardis, E. R., Wilson, R. K., Ally, A., Balasundaram, M., Butterfield, Y. S. N., Carlsen, R., Carter, C., Chu, A., Chuah, E., Chun, H.-J. E., Coope, R. J. N., Dhalla, N., Guin, R., Hirst, C., Hirst, M., Holt, R. A., Lee, D., Li, H. I., Mayo, M., Moore, R. A., Mungall, A. J., Pleasance, E., Gordon Robertson, A., Schein, J. E., Shafiei, A., Sipahimalani, P., Slobodan, J. R., Stoll, D., Tam, A., Thiessen, N., Varhol, R. J., Wye, N., Zeng, T., Zhao, Y., Birol, I., Jones, S. J. M., Marra, M. A., Cherniack, A. D., Saksena, G., Onofrio, R. C., Pho, N. H., Carter, S. L., Schumacher, S. E., Tabak, B., Hernandez, B., Gentry, J., Nguyen, H., Crenshaw, A., Ardlie, K., Beroukhim, R., Winckler, W., Getz, G., Gabriel, S. B., Meyerson, M., Chin, L., Park, P. J., Kucherlapati, R., Hoadley, K. A., Todd Auman, J., Fan, C., Turman, Y. J., Shi, Y., Li, L., Topal, M. D., He, X., Chao, H.-H., Prat, A., Silva, G. O.,

Iglesia, M. D., Zhao, W., Usary, J., Berg, J. S., Adams, M., Booker, J., Wu, J., Gulabani, A., Bodenheimer, T., Hoyle, A. P., Simons, J. V., Soloway, M. G., Mose, L. E., Jefferys, S. R., Balu, S., Parker, J. S., Neil Hayes, D., Perou, C. M., Malik, S., Mahurkar, S., Shen, H., Weisenberger, D. J., Triche Jr, T., Lai, P. H., Bootwalla, M. S., Maglinte, D. T., Berman, B. P., Van Den Berg, D. J., Baylin, S. B., Laird, P. W., Creighton, C. J., Donehower, L. A., Getz, G., Noble, M., Voet, D., Saksena, G., Gehlenborg, N., DiCara, D., Zhang, J., Zhang, H., Wu, C.-J., Yingchun Liu, S., Lawrence, M. S., Zou, L., Sivachenko, A., Lin, P., Stojanov, P., Jing, R., Cho, J., Sinha, R., Park, R. W., Nazaire, M.-D., Robinson, J., Thorvaldsdottir, H., Mesirov, J., Park, P. J., Chin, L., Reynolds, S., Kreisberg, R. B., Bernard, B., Bressler, R., Erkkila, T., Lin, J., Thorsson, V., Zhang, W., Shmulevich, I., Ciriello, G., Weinhold, N., Schultz, N., Gao, J., Cerami, E., Gross, B., Jacobsen, A., Sinha, R., Arman Aksoy, B., Antipin, Y., Reva, B., Shen, R., Taylor, B. S., Ladanyi, M., Sander, C., Anur, P., Spellman, P. T., Lu, Y., Liu, W., Verhaak, R. R. G., Mills, G. B., Akbani, R., Zhang, N., Broom, B. M., Casasent, T. D., Wakefield, C., Unruh, A. K., Baggerly, K., Coombes, K., Weinstein, J. N., Haussler, D., Benz, C. C., Stuart, J. M., Benz, S. C., Zhu, J., Szeto, C. C., Scott, G. K., Yau, C., Paull, E. O., Carlin, D., Wong, C., Sokolov, A., Thusberg, J., Mooney, S., Ng, S., Goldstein, T. C., Ellrott, K., Grifford, M., Wilks, C., Ma, S., Craft, B., Yan, C., Hu, Y., Meerzaman, D., Gastier-Foster, J. M., Bowen, J., Ramirez, N. C., Black, A. D., XPATH ERROR unknown variable tname, R. E., White, P., Zmuda, E. J., Frick, J., Lichtenberg, T. M., Brookens, R., George, M. M., Gerken, M. A., Harper, H. A., Leraas, K. M., Wise, L. J., Tabler, T. R., McAllister, C., Barr, T., Hart-Kothari, M., Tarvin, K., Saller, C., Sandusky, G., Mitchell, C., Iacocca, M. V., Brown, J., Rabeno, B., Czerwinski, C., Petrelli, N., Dolzhansky, O., Abramov, M., Voronina, O., Potapova, O., Marks, J. R., Suchorska, W. M., Murawa, D., Kycler, W., Ibbs, M., Korski, K., Spychała, A., Murawa, P., & Brzeziński, J. J. a. (2012). Comprehensive molecular portraits of human breast tumours. Nature *490*(7418), 61–70.

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., & Getz, G. (2014a). Discovery and saturation analysis of cancer genes across 21 tumour types.. Nature *505*(7484), 495–501.

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., & Getz, G. (2014b). Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature *499*(7457), 214–218.

Lee, J. C., Vivanco, I., Beroukhim, R., Huang, J. H. Y., Feng, W. L., DeBiasi, R. M., Yoshimoto, K., King, J. C., Nghiemphu, P., Yuza, Y., Xu, Q., Greulich, H., Thomas, R. K., Paez, J. G., Peck, T. C., Linhart, D. J., Glatt, K. A., Getz, G., Onofrio, R., Ziaugra, L., Levine, R. L., Gabriel, S., Kawaguchi, T., O'Neill, K., Khan, H., Liau, L. M., Nelson, S. F., Rao, P. N., Mischel, P., Pieper, R. O., Cloughesy, T., Leahy, D. J., Sellers, W. R., Sawyers, C. L., Meyerson, M.,

& Mellinghoff, I. K. (2006). Epidermal growth factor receptor activation in glioblastoma through novel missense mutations in the extracellular domain.. PLoS medicine *3*(12), e485.

Liu, X., Wang, L., Zhao, K., Thompson, P. R., Hwang, Y., Marmorstein, R., & Cole, P. A. (2008). The structural basis of protein acetylation by the p300/CBP transcriptional coactivator. Nature *451*(7180), 846–850.

Lohr, J. G., Stojanov, P., Lawrence, M. S., Auclair, D., Chapuy, B., Sougnez, C., Cruz-Gordillo, P., Knoechel, B., Asmann, Y. W., Slager, S. L., Novak, A. J., Dogan, A., Ansell, S. M., Link, B. K., Zou, L., Gould, J., Saksena, G., Stransky, N., Rangel-Escareno, C., Fernandez-Lopez, J. C., Hidalgo-Miranda, A., Melendez-Zajgla, J., Hernández-Lemus, E., Schwarz-Cruz y Celis, A., Imaz-Rosshandler, I., Ojesina, A. I., Jung, J., Pedamallu, C. S., Lander, E. S., Habermann, T. M., Cerhan, J. R., Shipp, M. A., Getz, G., & Golub, T. R. (2012). Discovery and prioritization of somatic mutations in diffuse large B-cell lymphoma (DLBCL) by whole-exome sequencing.. Proceedings of the National Academy of Sciences *109*(10), 3879–3884.

McLendon, R., Friedman, A., Bigner, D., Van Meir, E. G., Brat, D. J., M Mastrogianakis, G., Olson, J. J., Mikkelsen, T., Lehman, N., Aldape, K., Alfred Yung, W. K., Bogler, O., VandenBerg, S., Berger, M., Prados, M., Muzny, D., Morgan, M., Scherer, S., Sabo, A., Nazareth, L., Lewis, L., Hall, O., Zhu, Y., Ren, Y., Alvi, O., Yao, J., Hawes, A., Jhangiani, S., Fowler, G., San Lucas, A., Kovar, C., Cree, A., Dinh, H., Santibanez, J., Joshi, V., Gonzalez-Garay, M. L., Miller, C. A., Milosavljevic, A., Donehower, L., Wheeler, D. A., Gibbs, R. A., Cibulskis, K., Sougnez, C., Fennell, T., Mahan, S., Wilkinson, J., Ziaugra, L., Onofrio, R., Bloom, T., Nicol, R., Ardlie, K., Baldwin, J., Gabriel, S., Lander, E. S., Ding, L., Fulton, R. S., McLellan, M. D., Wallis, J., Larson, D. E., Shi, X., Abbott, R., Fulton, L., Chen, K., Koboldt, D. C., Wendl, M. C., Meyer, R., Tang, Y., Lin, L., Osborne, J. R., Dunford-Shore, B. H., Miner, T. L., Delehaunty, K., Markovic, C., Swift, G., Courtney, W., Pohl, C., Abbott, S., Hawkins, A., Leong, S., Haipek, C., Schmidt, H., Wiechert, M., Vickery, T., Scott, S., Dooling, D. J., Chinwalla, A., Weinstock, G. M., Mardis, E. R., Wilson, R. K., Getz, G., Winckler, W., Verhaak, R. G. W., Lawrence, M. S., O'Kelly, M., Robinson, J., Alexe, G., Beroukhim, R., Carter, S., Chiang, D., Gould, J., Gupta, S., Korn, J., Mermel, C., Mesirov, J., Monti, S., Nguyen, H., Parkin, M., Reich, M., Stransky, N., Weir, B. A., Garraway, L., Golub, T., Meyerson, M., Chin, L., Protopopov, A., Zhang, J., Perna, I., Aronson, S., Sathiamoorthy, N., Ren, G., Yao, J., Wiedemeyer, W. R., Kim, H., Won Kong, S., Xiao, Y., Kohane, I. S., Seidman, J., Park, P. J., Kucherlapati, R., Laird, P. W., Cope, L., Herman, J. G., Weisenberger, D. J., Pan, F., Van Den Berg, D., Van Neste, L., Mi Yi, J., Schuebel, K. E., Baylin, S. B., Absher, D. M., Li, J. Z., Southwick, A., Brady, S., Aggarwal, A., Chung, T., Sherlock, G., Brooks, J. D., Myers, R. M., Spellman, P. T., Purdom, E., Jakkula, L. R., Lapuk, A. V., Marr, H., Dorton, S., Gi Choi, Y., Han, J., Ray, A., Wang, V., Durinck, S., Robinson, M., Wang, N. J., Vranizan, K., Peng, V., Van Name, E., Fontenay, G. V., Ngai, J., Conboy, J. G., Parvin, B., Feiler, H. S., Speed, T. P., Gray, J. W., Brennan, C., Socci, N. D., Olshen, A., Taylor, B. S., Lash, A., Schultz, N., Reva, B., Antipin, Y., Stukalov, A., Gross, B., Cerami, E., Qing Wang, W., Qin, L.-X., Seshan, V. E., Villafania, L., Cavatore, M., Borsu, L., Viale, A., Gerald, W., Sander, C., Ladanyi, M., Perou, C. M., Neil Hayes, D., Topal, M. D., Hoadley, K. A., Qi, Y., Balu, S., Shi, Y., Wu, J., Penny, R., Bittner, M., Shelton, T., Lenkiewicz, E., Morris, S., Beasley, D., Sanders, S., Kahn, A., Sfeir, R., Chen, J., Nassau, D., Feng, L., Hickey, E., Zhang, J., Weinstein, J. N., Barker, A., Gerhard, D. S., Vockley, J., Compton, C., Vaught, J., Fielding, P., Ferguson, M. L., Schaefer, C., Madhavan, S., Buetow, K. H., Collins, F., Good, P., Guyer, M., Ozenberger, B., Peterson, J., & Thomson, E. (2008). Comprehensive

genomic characterization defines human glioblastoma genes and core pathways. Nature *455*(7216), 1061–1068.

Morin, R. D., Mendez-Lago, M., Mungall, A. J., Goya, R., Mungall, K. L., Corbett, R. D., Johnson, N. A., Severson, T. M., Chiu, R., Field, M., Jackman, S., Krzywinski, M., Scott, D. W., Trinh, D. L., Tamura-Wells, J., Li, S., Firme, M. R., Rogic, S., Griffith, M., Chan, S., Yakovenko, O., Meyer, I. M., Zhao, E. Y., Smailus, D., Moksa, M., Chittaranjan, S., Rimsza, L., Brooks-Wilson, A., Spinelli, J. J., Ben-Neriah, S., Meissner, B., Woolcock, B., Boyle, M., McDonald, H., Tam, A., Zhao, Y., Delaney, A., Zeng, T., Tse, K., Butterfield, Y., Birol, I., Holt, R., Schein, J., Horsman, D. E., Moore, R., Jones, S. J. M., Connors, J. M., Hirst, M., Gascoyne, R. D., & Marra, M. A. (2012). Frequent mutation of histone-modifying genes in non-Hodgkin lymphoma. Nature *476*(7360), 298–303.

Nehrt, N. L., Peterson, T. A., Park, D., & Kann, M. G. (2012). Domain landscapes of somatic mutations in cancer. BMC Genomics *13*(Suppl 4), S9.

Ng, P. C. (2003). SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Research *31*(13), 3812–3814.

Ohtaki, N., Yamaguchi, A., Goi, T., Fukaya, T., Takeuchi, K., Katayama, K., Hirose, K., & Urano, T. (2001). Somatic alterations of the DPC4 and Madr2 genes in colorectal cancers and relationship to metastasis.. International journal of oncology *18*(2), 265–270.

Pasqualucci, L., Dominguez-Sola, D., Chiarenza, A., Fabbri, G., Grunn, A., Trifonov, V., Kasper, L. H., Lerach, S., Tang, H., Ma, J., Rossi, D., Chadburn, A., Murty, V. V., Mullighan, C. G., Gaidano, G., Rabadan, R., Brindle, P. K., & Dalla-Favera, R. (2012). Inactivating mutations of acetyltransferase genes in B-cell lymphoma. Nature *471*(7337), 189–195.

Peifer, M., Fernández-Cuesta, L., Sos, M. L., George, J., Seidel, D., Kasper, L. H., Plenker, D., Leenders, F., Sun, R., Zander, T., Menon, R., Koker, M., Dahmen, I., Müller, C., Di Cerbo, V., Schildhaus, H.-U., Altmüller, J., Baessmann, I., Becker, C., de Wilde, B., Vandesompele, J., Böhm, D., Ansén, S., Gabler, F., Wilkening, I., Heynck, S., Heuckmann, J. M., Lu, X., Carter, S. L., Cibulskis, K., Banerji, S., Getz, G., Park, K.-S., Rauh, D., Grütter, C., Fischer, M., Pasqualucci, L., Wright, G., Wainer, Z., Russell, P., Petersen, I., Chen, Y., Stoelben, E., Ludwig, C., Schnabel, P., Hoffmann, H., Muley, T., Brockmann, M., Engel-Riedel, W., Muscarella, L. A., Fazio, V. M., Groen, H., Timens, W., Sietsma, H., Thunnissen, E., Smit, E., Heideman, D. A. M., Snijders, P. J. F., Cappuzzo, F., Ligorio, C., Damiani, S., Field, J., Solberg, S., Brustugun, O. T., Lund-Iversen, M., Sänger, J., Clement, J. H., Soltermann, A., Moch, H., Weder, W., Solomon, B., Soria, J.-C., Validire, P., Besse, B., Brambilla, E., Brambilla, C., Lantuejoul, S., Lorimier, P., Schneider, P. M., Hallek, M., Pao, W., Meyerson, M., Sage, J., Shendure, J., Schneider, R., Büttner, R., Wolf, J., Nürnberg, P., Perner, S., Heukamp, L. C., Brindle, P. K., Haas, S., & Thomas, R. K. (2012). Integrative genome analyses identify key somatic driver mutations of small-cell lung cancer. Nature Genetics *44*(10), 1104–1110.

Peterson, T. A., Adadey, A., Santana-Cruz, I., Sun, Y., Winder, A., & Kann, M. G. (2010). DMDM: domain mapping of disease mutations. Bioinformatics (Oxford, England) *26*(19), 2458–2459.

Peterson, T. A., Nehrt, N. L., Park, D., & Kann, M. G. (2012). Incorporating molecular and functional context into the analysis and prioritization of human variants associated with cancer. Journal of the American Medical Informatics Association *19*(2), 275–283.

Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., Heger, A., Holm, L., Sonnhammer, E. L. L., Eddy, S. R., Bateman, A., & Finn, R. D. (2011). The Pfam protein families database. Nucleic Acids Research *40*(D1), D290–D301.

Reimand, J. & Bader, G. D. (2013). Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers.. Molecular Systems Biology *9*, 637.

Reimand, J., Wagih, O., & Bader, G. D. (2013). The mutational landscape of phosphorylation signaling in cancer. Scientific Reports *3*.

Reva, B., Antipin, Y., & Sander, C. (2011). Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Research *39*(17), e118–e118.

Shi, Y., Hata, A., Lo, R. S., Massagué, J., & Pavletich, N. P. (1997). A structural basis for mutational inactivation of the tumour suppressor Smad4.. Nature *388*(6637), 87–93.

Stephens, P. J., Tarpey, P. S., Davies, H., Loo, P., Greenman, C., Wedge, D. C., Zainal, S., Martin, S., Varela, I., Bignell, G. R., Yates, L. R., Papaemmanuil, E., Beare, D., Butler, A., Cheverton, A., Gamble, J., Hinton, J., Jia, M., Jayakumar, A., Jones, D., Latimer, C., Lau, K., McLaren, S., McBride, D. J., Menzies, A., Mudie, L., Raine, K., Rad, R., Chapman, M. S., Teague, J., Easton, D., Langerød, A., OSBREAC, Lee, M. T. M., Shen, C.-Y., Tee, B. T. K., Huimin, B. W., Broeks, A., Vargas, A. C., Turashvili, G., Martens, J., Fatima, A., Miron, P., Chin, S.-F., Thomas, G., Boyault, S., Mariani, O., Lakhani, S. R., van de Vijver, M., van t Veer, L., Foekens, J., Desmedt, C., Sotiriou, C., Tutt, A., Caldas, C., Reis-Filho, J. S., Aparicio, S. A. J. R., Salomon, A. V., Børresen-Dale, A.-L., Richardson, A., Campbell, P. J., Futreal, P. A., & Stratton, M. R. (2013). The landscape of cancer genes and mutational processes in breast cancer. Nature *486*(7403), 400–404.

Stroud, J. C., Wu, Y., Bates, D. L., Han, A., Nowick, K., Paabo, S., Tong, H., & Chen, L. (2006). Structure of the Forkhead Domain of FOXP2 Bound to DNA. Structure *14*(1), 159–166.

Torkamani, A. & Schork, N. J. (2009). Identification of rare cancer driver mutations by network reconstruction. Genome Research *19*(9), 1570–1578.

Tsai, K. L., Huang, C. Y., Chang, C. H., Sun, Y. J., Chuang, W. J., & Hsiao, C. D. (2006). Crystal Structure of the Human FOXK1a-DNA Complex and Its Implications on the Diverse Binding Specificity of Winged Helix/Forkhead Proteins. The Journal of biological chemistry *281*(25), 17400–17409.

Wang, N. J., Sanborn, Z., Arnett, K. L., Bayston, L. J., Liao, W., Proby, C. M., Leigh, I. M., Collisson, E. A., Gordon, P. B., Jakkula, L., Pennypacker, S., Zou, Y., Sharma, M., North, J. P., Vemula, S. S., Mauro, T. M., Neuhaus, I. M., Leboit, P. E., Hur, J. S., Park, K., Huh, N., Kwok, P.-Y., Arron, S. T., Massion, P. P., Bale, A. E., Haussler, D., Cleaver, J. E., Gray, J. W., Spellman, P. T., South, A. P., Aster, J. C., Blacklow, S. C., & Cho, R. J. (2011). Loss-of-function mutations in Notch receptors in cutaneous and lung squamous cell carcinoma.. Proceedings of the National Academy of Sciences *108*(43), 17761–17766.

Weng, A. P., Ferrando, A. A., Lee, W., Morris, J. P., Silverman, L. B., Sanchez-Irizarry, C., Blacklow, S. C., Look, A. T., & Aster, J. C. (2004). Activating mutations of NOTCH1 in human T cell acute lymphoblastic leukemia.. Science (New York, N.Y.) *306*(5694), 269–271.

Wu, J. W., Hu, M., Chai, J., Seoane, J., Huse, M., Li, C., Rigotti, D. J., Kyin, S., Muir, T. W., Fairman, R., Massagué, J., & Shi, Y. (2001). Crystal structure of a phosphorylated Smad2. Recognition of phosphoserine by the MH2 domain and insights on Smad function in TGF-beta signaling.. Molecular Cell *8*(6), 1277–1289.

Wu, J. Y., Yu, C. J., Chang, Y. C., Yang, C. H., Shih, J. Y., & Yang, P. C. (2011). Effectiveness of Tyrosine Kinase Inhibitors on "Uncommon" Epidermal Growth Factor Receptor Mutations of Unknown Clinical Significance in Non-Small Cell Lung Cancer. Clinical Cancer Research *17*(11), 3812–3821.

Wu, Y., Borde, M., Heissmeyer, V., Feuerer, M., Lapan, A. D., Stroud, J. C., Bates, D. L., Guo, L., Han, A., Ziegler, S. F., Mathis, D., Benoist, C., Chen, L., & Rao, A. (2006). FOXP3 Controls Regulatory T Cell Function through Cooperation with NFAT. Cell *126*(2), 375–387.

Yue, P., Forrest, W. F., Kaminker, J. S., Lohr, S., Zhang, Z., & Cavet, G. (2010). Inferring the functional effects of mutation through clusters of mutations in homologous proteins. Human Mutation *31*(3), 264–271.

Yue, P., Melamud, E., & Moult, J. (2006). SNPs3D: candidate gene and SNP selection for association studies.. BMC Bioinformatics *7*, 166.