PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

# Protein Domain-Level Landscape of Cancer-Type-Specific Somatic Mutations

**Fan Yang[1,2,3], Evangelia Petsalaki[2,3], Thomas Rolland[4,5], David E. Hill[4], Marc Vidal[4], Frederick P. Roth[1,2,3,4,6,7]***

**1** Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada, **2** Donnelly Centre, University of Toronto, Toronto, Ontario, Canada, **3** Lunenfeld-Tanenbaum Research Institute, Mt. Sinai Hospital, Toronto, Ontario, Canada, **4** Center for Cancer Systems Biology (CCSB), Dana-Farber Cancer Institute, Boston, Massachusetts, United States of America, **5** Department of Genetics, Harvard Medical School, Boston, Massachusetts, United States of America, **6** Canadian Institute for Advanced Research, Toronto, Ontario, Canada, **7** Department of Computer Science, University of Toronto, Toronto, Ontario, Canada

* fritz.roth@utoronto.ca

**Citation:** Yang F, Petsalaki E, Rolland T, Hill DE, Vidal M, Roth FP (2015) Protein Domain-Level Landscape of Cancer-Type-Specific Somatic Mutations. PLoS Comput Biol 11(3): e1004147. doi:10.1371/journal.pcbi.1004147

**Editor:** Mona Singh, Princeton University, United States of America

**Received:** August 22, 2014

**Accepted:** January 22, 2015

**Published:** March 20, 2015

**Data Availability Statement:** Data are available from the Catalogue of Somatic Mutations in Cancer) database (version 67) and Pfam protein domain family database (version 27). Domain-domain interactions were collected from Rolland et all, Cell 2014, DOI: http://dx.doi.org/10.1016/j.cell.2014.10.050

**Funding:** This project was supported by an NIH/NHGRI Center of Excellence in Genome Science award (HG004233) and benefitted from NIH/NHGRI award HG001715 (both to MV, DEH, and FPR). FPR was also supported by the Canada Excellence Research Chairs Program, by the Avon Foundation,

## Abstract

Identifying driver mutations and their functional consequences is critical to our understanding of cancer. Towards this goal, and because domains are the functional units of a protein, we explored the protein domain-level landscape of cancer-type-specific somatic mutations. Specifically, we systematically examined tumor genomes from 21 cancer types to identify domains with high mutational density in specific tissues, the positions of mutational hotspots within these domains, and the functional and structural context where possible. While hotspots corresponding to specific gain-of-function mutations are expected for oncoproteins, we found that tumor suppressor proteins also exhibit strong biases toward being mutated in particular domains. Within domains, however, we observed the expected patterns of mutation, with recurrently mutated positions for oncogenes and evenly distributed mutations for tumor suppressors. For example, we identified both known and new endometrial cancer hotspots in the tyrosine kinase domain of the FGFR2 protein, one of which is also a hotspot in breast cancer, and found new two hotspots in the Immunoglobulin I-set domain in colon cancer. Thus, to prioritize cancer mutations for further functional studies aimed at more precise cancer treatments, we have systematically correlated mutations and cancer types at the protein domain level.

## Author Summary

Extensive tumor genome sequencing has provided raw material to understand mutational processes and identify cancer-associated somatic variants. However, fundamental problems remain to: i) separate 'driver' from 'passenger' mutations, ii) further understand the functional mechanisms and consequences of driver mutations, and iii) identify the cancer types in which each driver mutation is relevant. Here we analyze whole-genome and exome tumor sequencing data from the perspective of protein domains—the basic

by an Ontario Research Fund – Research Excellence Award, and by a Canadian Institute for Advanced Research Fellowship.The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

structural and functional units of proteins. Exploring the cancer-type-specific landscape of domain mutations across 21 cancer types, we identify both cancer-type-specific mutated domains and mutational hotspots. Frequently-mutated domains were identified for oncoproteins for which the ‘mutational hotspot’ phenomenon owing to the relative rarity of gain-of-function mutations is well known, and also for tumor suppressor proteins, for which more uniformly distributed loss-of-function driver mutations are expected. A given gene product may be perturbed differently in different cancers. Indeed, we observed systematic shifts between cancer types of the positions at which mutations occur within a given protein. Both known and novel candidate driver mutations were retrieved. Novel cancer gene candidates significantly overlapped with orthogonal systematic cancer screen hits, supporting the power of this approach to identify cancer genes.

## Introduction

Cancer is caused in large part by the accumulation of mutations in oncogenes and tumor suppressor genes. Previous analyses of well-studied cancers, such as colorectal cancer and retinoblastoma, have suggested that as few as three mutations are sufficient for cancer initiation [1–3]. Thousands of cancer genomes have now been sequenced[4, 5], including efforts from The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC) [6, 7]. In recent years, the genetic landscape of mutations has been revealed in several well-studied cancers [8–16]. However, the process of extracting useful knowledge from this vast sequence resource has only begun.

The complexity of cancer genomes represents a challenge to therapy and our basic understanding of the disease and therefore also to therapy. Individual cancers can contain thousands of somatic mutations [17, 18], only a small fraction of which are likely to be driver mutations contributing to tumor initiation or progression [18–22]. Even genes that are well known to cause cancer contain many effectively-neutral passenger mutations [23]. For example, it was reported that 80% of non-synonymous single-base substitutions observed in genes encoding protein kinases are passenger mutations [24]. Most candidate driver gene identification has been done on the basis of observing mutations in a large fraction of tumor samples. However, the list of putative driver mutations generated includes many that are implausible, with a false positive rate that increases with the number of sequenced tumor samples [25, 26].

Determining the effect of mutations on the structure and function of proteins remains challenging [27]. Previous gene-based studies have generally focused on the whole gene or whole protein, but mutations in different protein domains, structural units that often have distinct functions, may have different functional consequences. Thus, gene-level analysis can identify genes that contribute to multiple cancers, but does not map mutations to structural elements. Recently, computational structural studies have explored mutational effects on specific regions of a protein (e.g., the binding site)[28–31]. For example, Joerger and Fersht showed that certain mutations in the p53 protein can determine folding state and affinity of p53 for specific target DNA elements. Also, different p53 mutations affect different protein–protein interaction interfaces dictating either tetramerization of p53 or its interaction with a multitude of other regulatory proteins[30]. Similarly, the effects of mutations in different protein kinase sub-domains have been shown to have different functional impacts[28]. Thus, within a multi-functional gene, different mutations can affect different functions. The structural details of individual mutants can provide the basis for the design of cancer therapeutics[30]. Indeed, a given gene may have different functional roles in different cancers, reflected in shifts in the mutational

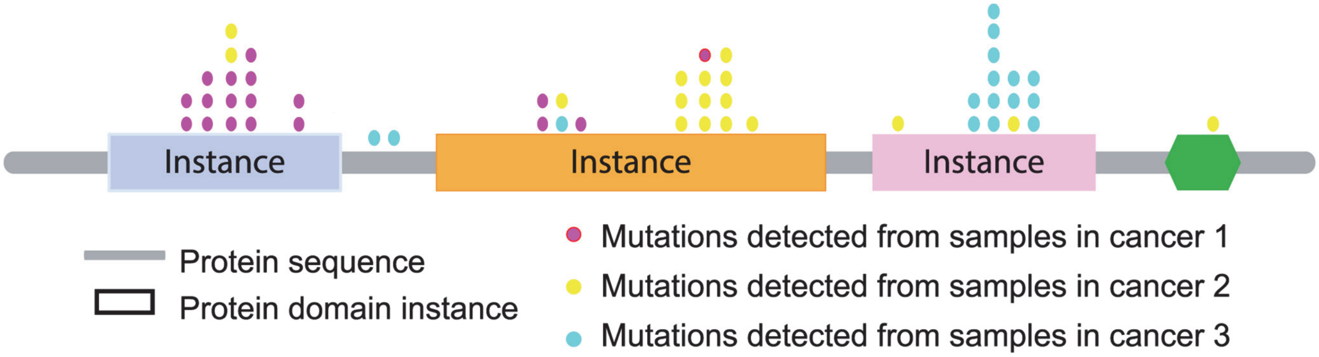
Instance
Instance
Instance
Protein sequence
Protein domain instance
Mutations detected from samples in cancer 1
Mutations detected from samples in cancer 2
Mutations detected from samples in cancer 3

**Fig 1. Mapping mutations detected from different cancers to domain instances.** Rectangles represent protein domain instances in a given gene. Colored dots represent mutations detected in different cancer types.

doi:10.1371/journal.pcbi.1004147.g001

distribution of different cancers. Recently, Nehrt *et al*. examined 100 colon cancer and 522 breast cancer samples to identify specific domain types with heightened mutation rates, succeeding even within genes that have generally lower mutation rates in colon or breast cancer [32, 33]. Mutations occurring within a particular domain are more likely to share structural and functional effects [15]. Two mutations within a given gene may be associated with different human diseases, e.g., potentially by disrupting different protein interactions [34, 35]. Thus, studies that consider mutational positions (e.g., relative to known domains) could be beneficial in elucidating functional effects of mutations.

In this study, a "domain instance" refers to a particular protein domain encoded within a particular gene and a "domain type" refers to a Pfam domain 'pattern' that may correspond to different domain instances encoded by different genes. In other words, a domain instance refers to a specific amino acid subsequence within a given single protein that matches to a given domain type.

To better distinguish this study from previous related studies, such as the domain landscape in colon and breast cancer by Nehrt *et al*, we note that we are systematically analyzing multiple (twenty-one) cancer types. Like Nehrt *et al*., we analyze each Pfam domain type. In addition, however, we specifically analyze each Pfam domain instance. Rather than simply seeking domains with a mutational density that is enriched relative to other genomic regions, we further require that this enrichment is greater in one cancer type than in all other cancer types. This has the advantage of pointing us to interesting differences between cancer types, while also implicitly controlling for region-specific differences in background mutation rate. Thus, we are analyzing the 'domain-centric mutational landscape' by examining the domain-level distribution of missense somatic mutations across multiple cancer types.

We mapped missense somatic mutations to domain instances (Fig. 1. outlines this process) for 21 cancer types (Table 1) and detected 100 cancer-type-specific significantly-mutated domain instances (SMDs) among different cancers. Further examination of these 100 domain instances showed that the proportion of within-domain mutations corresponding to hotspot positions can distinguish oncoproteins from tumor suppressor proteins. We also found that the vast majority of within-domain mutational hotspots shared by multiple cancer types occurred at functional sites. Thus, domain mutational landscape information can be used to prioritize candidate cancer-causing mutations and to elucidate their cancer-type-dependent functional effects.

**Table 1. Prevalence of predicted damaging mutations in domain instances among cancer types.**

| Cancer Type | Mutation Counts | Gene Counts | Domain Families |
|---|---|---|---|
| Neuroblastoma | 166 | 154 | 132 |
| Chondrosarcoma | 68 | 49 | 47 |
| Breast cancer | 4026 | 2568 | 1293 |
| Glioblastoma and medulloblastoma | 1423 | 911 | 559 |
| Cervical cancer | 707 | 570 | 410 |
| Endometrial carcinoma | 12183 | 6105 | 2451 |
| Lymphoma and leukemia | 1443 | 692 | 468 |
| Renal cell carcinoma. | 4774 | 3241 | 1591 |
| Colorectal cancer | 27132 | 9706 | 3426 |
| Liver cancer | 485 | 309 | 235 |
| Lung cancer | 11654 | 5899 | 2401 |
| Meningioma | 89 | 66 | 62 |
| Esophageal adenocarcinoma | 935 | 568 | 366 |
| Ovarian cancer | 4182 | 3035 | 1395 |
| Pancreatic cancer | 703 | 534 | 362 |
| Prostate cancer | 1917 | 1409 | 785 |
| Adenoid cystic carcinomas | 165 | 126 | 107 |
| Melanoma | 1824 | 1352 | 758 |
| Striated muscle | 16 | 13 | 12 |
| Head and neck squamous cell carcinoma | 582 | 445 | 307 |
| Bladder cancer | 2996 | 1289 | 804 |

doi:10.1371/journal.pcbi.1004147.t001

## Results

The results of this study fall into four areas: 1) exploration of cancer-type-specific domain mutation landscapes across 21 cancer types; 2) identification of cancer-type-specific shifts in mutation position within given proteins; 3) comparison of domain-centric mutational patterns between oncoproteins and tumor suppressor proteins; and 4) correlation of mutational hotspots occurring in multiple cancer types to oncogenicity and functional roles.

### Cancer-type-specific domain mutation landscapes across 21 cancer types

Protein domains are generally regarded as the conserved structural and functional units of proteins. We therefore focused on the 237,716 missense somatic mutations, across 21 different human tissues, that fell within protein domain instances. We further focused on the subset of 76,158 mutations that were predicted to compromise the function of the harboring protein, using the IntOGen–mutation platform[36] (Table 1, S1 Table). To avoid observational biases, the above-mentioned mutations were derived only from genome-scale (either whole-genome or exome) sequencing studies (listed in Table 2).

The prevalence of missense somatic mutations can vary from cancer to cancer at the domain level [37, 38]. We found that most domain instances had a mutational density of only one or two missense somatic mutations per megabase in the corresponding DNA sequence (Fig. 2.). For example, a mutated domain of length 209 residues (the average domain instance length) contains an average of one single-amino acid-changing mutation for every 71 patients. Although domains with high mutation rates can be seen for many cancers (Fig. 2.), these mutation rates can be misleading. Given heterogeneity of mutation rates across the genome and

**Table 2. Patient tissue samples from selected cancer genome studies across 21 cancer types.**

| Primary Site | Cancer Type | Sample Counts | References |
|---|---|---|---|
| Autonomic ganglia | **Neuroblastoma** | **134** | [89, 98] |
| Bone | **Chondrosarcoma** | **66** | [85] |
| Breast | **Breast cancer** | **978** | [70, 86, 96, 117] |
| Central nervous system | **Glioblastoma and medulloblastoma** | **525** | [81, 91, 93, 103, 106, 125, 126] |
| Cervix | **Cervical cancer** | **14** | [92] |
| Endometrium | **Endometrial carcinoma** | **261** | [104, 123] |
| Hematopoietic and lymph | **Lymphoma and leukemia** | **415** | [80, 82, 97, 103, 115, 121] |
| Kidney | **Renal cell carcinoma.** | **594** | [104, 113, 116, 119, 120] |
| Colon | **Colorectal cancer** | **762** | [71, 79, 87, 129] |
| Liver | **Liver cancer** | **531** | [84, 101, 111, 114] |
| Lung | **Lung cancer** | **825** | [72, 73, 88, 90, 99, 130] |
| Meninges | **Meningioma** | **39** | [122] |
| Esophagus | **Esophageal adenocarcinoma** | **242** | [120, 131] |
| Ovary | **Ovarian cancer** | **637** | [107, 125] |
| Pancreas | **Pancreatic cancer** | **202** | [83, 109] |
| Prostate | **Prostate cancer** | **423** | [8, 100, 127, 130] |
| Salivary gland | **Adenoid cystic carcinomas** | **60** | [112] |
| Skin | **Melanoma** | **133** | [14, 105] |
| Soft tissue | **Striated muscle** | **13** | [103] |
| Upper aero-digestive tract | **Head and neck squamous cell carcinoma** | **203** | [118, 132] |
| Urinary tract | **Bladder cancer** | **203** | [110] |

doi:10.1371/journal.pcbi.1004147.t002

differences in overall mutation rate for different cancers, domain-instances with the highest mutation density in a given cancer may not be the true drivers of cancer progression[25].

To control for both positional and cancer-type specific differences in mutation rate, we sought domain instances that were highly mutated relative both to the same domain instance in other cancer types and also to other domain instances within the same cancer type (see Materials and Methods). We identified ~100 cancer-type-specific significantly mutated domain instances (SMDs) in 21 cancer types (S2 Table; $P$-value = $10^{-7}$, Fisher's Exact test, False Discovery Rate (FDR) <0.05). The number of cancer-type-specific SMDs in each of the 21 cancer types is listed in Table 3, and in S3 Table in greater detail. With only two exceptions, the smallest number of mutations observed for a domain instance that was declared to be significantly mutated was 6. The exceptions were the Collagen domain instance (with only 2 mutations) within the *COLEC11* gene product in soft tissue cancers, for which only 14 samples were available; and the CCDC14 domain instances (with 3 mutations) encoded by *CCDC14* in cancers of the salivary gland, for which only 60 samples were available (S4 Table).

We found between 3 and 7 SMDs for each cancer type, except for endometrial cancer (with 11 SMDs) as well as hematopoietic and lymphatic cancer (**with 27** SMDs). Of the 94 genes encoding at least one SMD, 40 (42%) had already been implicated in cancer according to the Sanger Cancer Gene Census ('Cancer Census') [39, 40], including well-established cancer-causing genes such as *KRAS, EGFR* and *TP53*. Enrichment for Cancer Census genes was both strong and significant (~12-fold enrichment; $P$-value = $5\times10^{-34}$, Fisher's Exact test), and suggests the remaining 54 genes that are not already known to be cancer drivers represent good candidates. For example, the Syntaphilin protein, encoded by *SNPH* harbors the syntaphilin domain instance, which was significantly mutated in lung cancer. Despite reports that it is

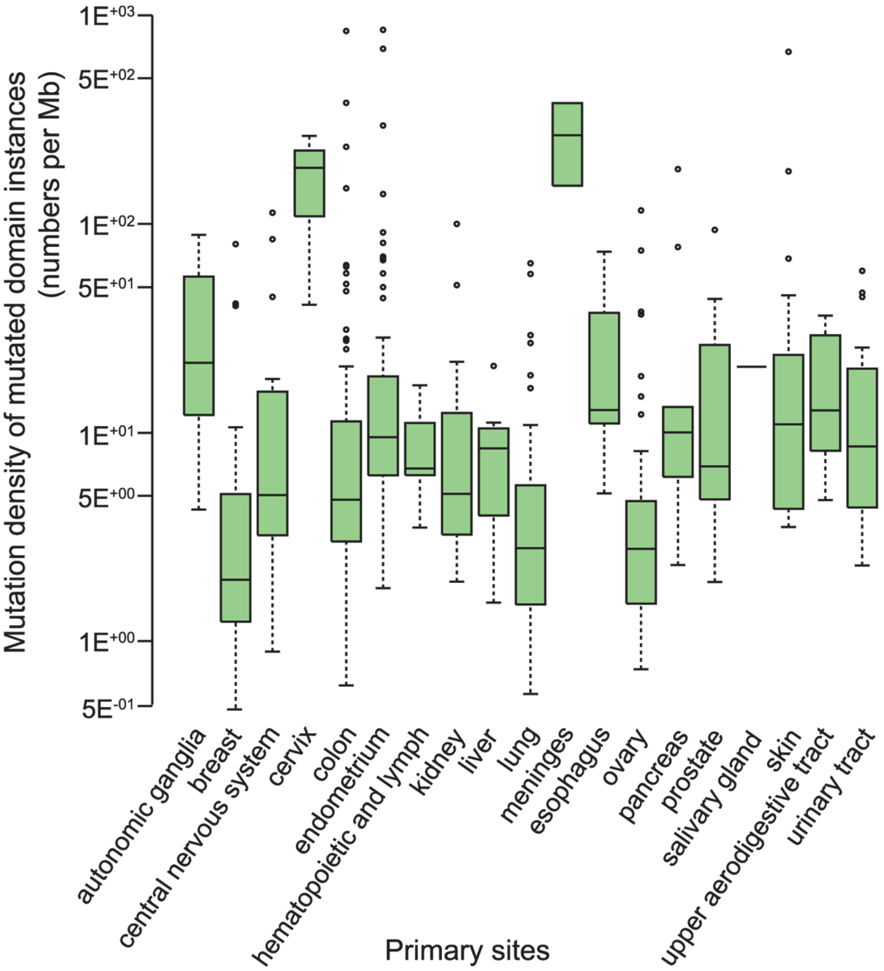

**Fig 2. Mutation Densities for Domain Instances across Cancers.** Box plots display mutation densities for mutated domain instances in different cancers. Outliers are shown as dots. Only predicted- damaging mutations predicted by IntOGen were used for this analysis (S1 Table).

doi:10.1371/journal.pcbi.1004147.g002

brain-specific[41], *SNPH* is expressed in lung according to microarray[42, 43] and RNA-seq studies[44].

We compared the resulting novel cancer gene candidates with cancer gene candidates emerging from a large-scale *in vivo* (mouse) screen via mutagenesis with Sleeping Beauty transposons [45]. Of the 94 genes encoding cancer type-specific SMDs, 24 were found in the Sleeping Beauty dataset ( ~ 3-fold enrichment; *P*-value = $7{\times}10^{-06}$, Fisher's Exact test). Of the subset of 54 candidate genes not already known to be cancer genes, 10 were found in the Sleeping Beauty dataset ( ~ 2-fold enrichment; *P*-value = $5{\times}10^{-3}$, Fisher's Exact test, Table 4).

The distribution of cancer-type-specific SMDs varies across cancer types. Among cancer-type-specific SMDs, most (95%) were only significantly mutated in a single cancer type (Fig. 3.). Five domain instances were found to be significantly mutated in more than one cancer (Table 4): a Ras domain instance of *KRAS*, mutated in lung and pancreatic cancer; a PHD finger domain instance (zf-HC5HC2H) of MLL3, mutated in breast and prostate cancer; a MAD homology 2 (MH2) domain instance of *SMAD4*, mutated in colon and esophageal cancer; a SNF2 family N-terminal (SNF2_N) domain instance of *SMARCA4*, mutated in esophageal cancer and cancer of the central nervous system; and the P53 DNA binding domain instance of

**Table 3. Significantly mutated domain instances and corresponding genes in each cancer type.**

| Cancer Type | Significantly Mutated Domain Instance Counts | Related Gene Counts |
|---|---|---|
| Neuroblastoma | 1 | 1 |
| Chondrosarcoma | 5 | 5 |
| Breast cancer | 4 | 4 |
| Glioblastoma and medulloblastoma | 11 | 9 |
| Cervical cancer | 4 | 4 |
| Colorectal cancer | 3 | 3 |
| Endometrial carcinoma | 11 | 10 |
| Lymphoma and leukemia | 27 | 26 |
| Renal cell carcinoma. | 5 | 5 |
| Liver cancer | 3 | 2 |
| Lung cancer | 9 | 9 |
| Meningioma | 3 | 3 |
| Esophageal adenocarcinoma | 4 | 4 |
| Ovarian cancer | 1 | 1 |
| Pancreatic cancer | 1 | 1 |
| Prostate cancer | 8 | 8 |
| Adenoid cystic carcinomas | 1 | 1 |
| Melanoma | 5 | 5 |
| Striated muscle | 1 | 1 |
| Head and neck squamous cell carcinoma | 3 | 3 |
| Bladder cancer | 1 | 1 |

doi:10.1371/journal.pcbi.1004147.t003

*TP53*, mutated in 8 cancer types. With the exception of *KRAS*, these genes are usually regarded as tumor suppressors. This suggests that, while tumor suppressors may cause different cancers via a common loss of function mechanism, the gain-of-function mechanism of oncogenes is more likely to be tissue-specific.

**Table 4. Genes that encode cancer-type-specific significantly mutated domain instance and overlap with the Sleeping Beauty dataset.**

| Gene Symbol | SignificantlyMutated Domain | Cancer Type | Predicted Category | Reported Category |
|---|---|---|---|---|
| ***CNTN4*** | PF13895.1 | Lymphoma and leukemia | Tumor suppressor | Not determined |
| ***CTNNA3*** | PF01044.14 | Lymphoma and leukemia | Tumor suppressor | Not determined |
| ***EPHA6*** | PF14575.1 | Lymphoma and leukemia | Tumor suppressor | Not determined |
| ***FOXO1*** | PF00250.13 | Lymphoma and leukemia | Tumor suppressor | Tumor suppressor |
| ***GNA13*** | PF00503.15 | Lymphoma and leukemia | Oncogene | Not determined |
| ***MAGI1*** | PF00503.15 | Lymphoma and leukemia | Tumor suppressor | Tumor suppressor |
| ***PCDH11X*** | PF08266.7 | Lymphoma and leukemia | Oncogene | Not determined |
| ***PCDH11X*** | PF00028.12 | Glioblastoma, medulloblastoma | Tumor suppressor | Not determined |
| ***PTPRD*** | PF00102.22 | Prostate cancer | Tumor suppressor | Not determined |
| ***SMAD4*** | PF03166.9 | Colorectal cancer | Tumor suppressor | Tumor suppressor |
| ***SMAD4*** | PF03166.9 | Esophageal adenocarcinoma | Tumor suppressor | Tumor suppressor |
| ***USP25*** | PF00443.24 | Liver cancer | Oncogene | Not determined |

doi:10.1371/journal.pcbi.1004147.t004

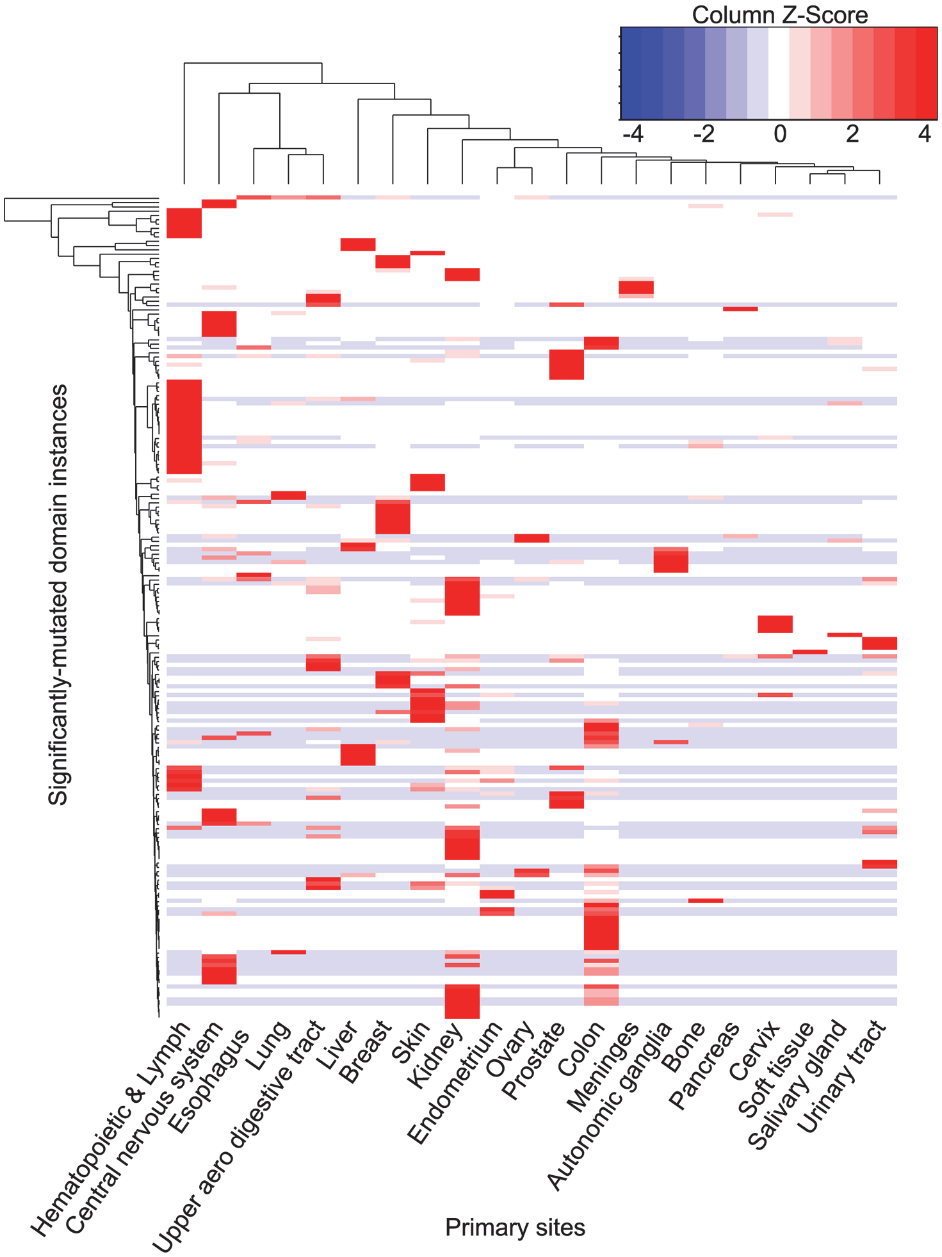

**Fig 3. Clustering of significantly-mutated domain instances across 21 cancer types.** The heatmap reflects the significance of cancer-type-specific mutation density of each domain instance in different cancers. Side bars in the same color indicate domain instances encoded by the same gene, and domain instances belonging to the same domain type.

doi:10.1371/journal.pcbi.1004147.g003

## Cancer-type-specific positioning of mutations within a given gene

Domain instances mutated in a specific cancer type can point to functions that are specifically disrupted in that cancer type. Furthermore, the observation that a given gene product has different domain instances mutated in different cancer types may elucidate how a single gene can play different roles in different cancers. To identify candidate genes with this behavior, we first selected all of the multi-domain genes that contained at least one cancer-type-specific SMD, and examined the cancer-type-specificity of these domains (see Materials and Methods).

Among the 94 genes identified above to contain cancer-type-specific SMDs, 52 genes had multiple domain instances with differing cancer-type-specificity (see S2 Table). These 52 genes were enriched for evidence of involvement in cancer, with 16 being Cancer Census genes (enrichment factor $\sim$ 11.9; $P$-value = $6.7 \times 10^{-13}$, Fisher's Exact test), and 15 being candidate cancer genes according to the Sleeping Beauty screen (enrichment factor $\sim$ 4.5; $P$-value = $1.9 \times 10^{-6}$, Fisher's Exact test).

To illustrate this analysis, we show the distribution of domain mutations in the EGF receptor, encoded by *EGFR*, across five cancers (Fig. 4A.). The EGF receptor is a flexible protein with four distinct domains, including extracellular and transmembrane regions, the intracellular kinase domains, and a long flexible tail (Fig. 4B.). Our analysis recapitulated domain mutation patterns seen in corresponding to previous findings. The extracellular region consists of the furin-like (Furin-Like) domain, the growth factor receptor domain IV (GF_recep_IV) and the L domain (Recep_L_Domain). The Furin-Like and the GF_recep_IV domains were both found to be significantly mutated in cancers of the central nervous system. Mutations in the extracellular region of EGF receptor have been associated with ligand-independent dimerization in cancers of the central nervous system[46], and mutations in the intracellular region of EGF receptor are associated with sensitivity to kinase inhibitors[46].

In lung cancers, mutations were significantly enriched in the tyrosine kinase domain (Pkinase_Tyr). This is a well-known location for oncogenic mutations that hyper-activate downstream pro-survival signaling pathways in lung cancer by causing the auto-phosphorylation of C-terminal residues [51].

The detailed functional consequences of *EGFR* mutations in prostate and colon cancer are still unclear. Differences in the positions of mutations between the extracellular (glioblastoma and prostate cancer) and intracellular regions (colon, lung and ovarian cancer) of the EGF receptor in different cancer types suggest different oncogenic mechanisms and possibly different therapeutic avenues.

Other interesting examples included that of the histone-lysine N-methyltransferase MLL3 protein, for which the PHD finger domain is mutated in breast cancer and prostate cancer, and for which the SET domain is mutated in glioblastoma and medulloblastoma. MLL3 is reported to possess histone methylation activity and is also involved in transcriptional co-activation. Knockdown or deletion of *MLL3* using RNAi or CRISPR is reported to cause acute myeloid leukemia in a mouse model [52].

Domain-associated mutational biases have been reported in several studies focusing on single well-known cancer genes such as the *PI3KCA* gene in colon and breast cancer[32], and the *NOTCH1* gene in leukemia, breast and ovarian cancer [53]. Here, we analyzed the distribution of somatic missense mutations for 14,083 genes across 21 cancer types and identified 52 genes (36 of which are not yet known to be cancer genes) for which different domain instances may contribute to different cancer types.

## Mutational trends of oncoproteins and tumor suppressor proteins

We further analyzed genes with at least one cancer-type-specific SMD. More specifically, we identified a collection of 337 cancer-type-specific mutation hotspots in 68 genes, including some hotspots that appeared in several different types of cancer (Fig. 5., S4 Table). For example, in the EGFR protein, residue p.A289 is a mutational hotspot in central nervous system cancer, p.C231 is a mutational hotspot in prostate cancer (Fig. 5.). Both residues fall in the Furin-like domain of the extra-cellular part of EGFR, but at different domain-domain interaction interfaces.

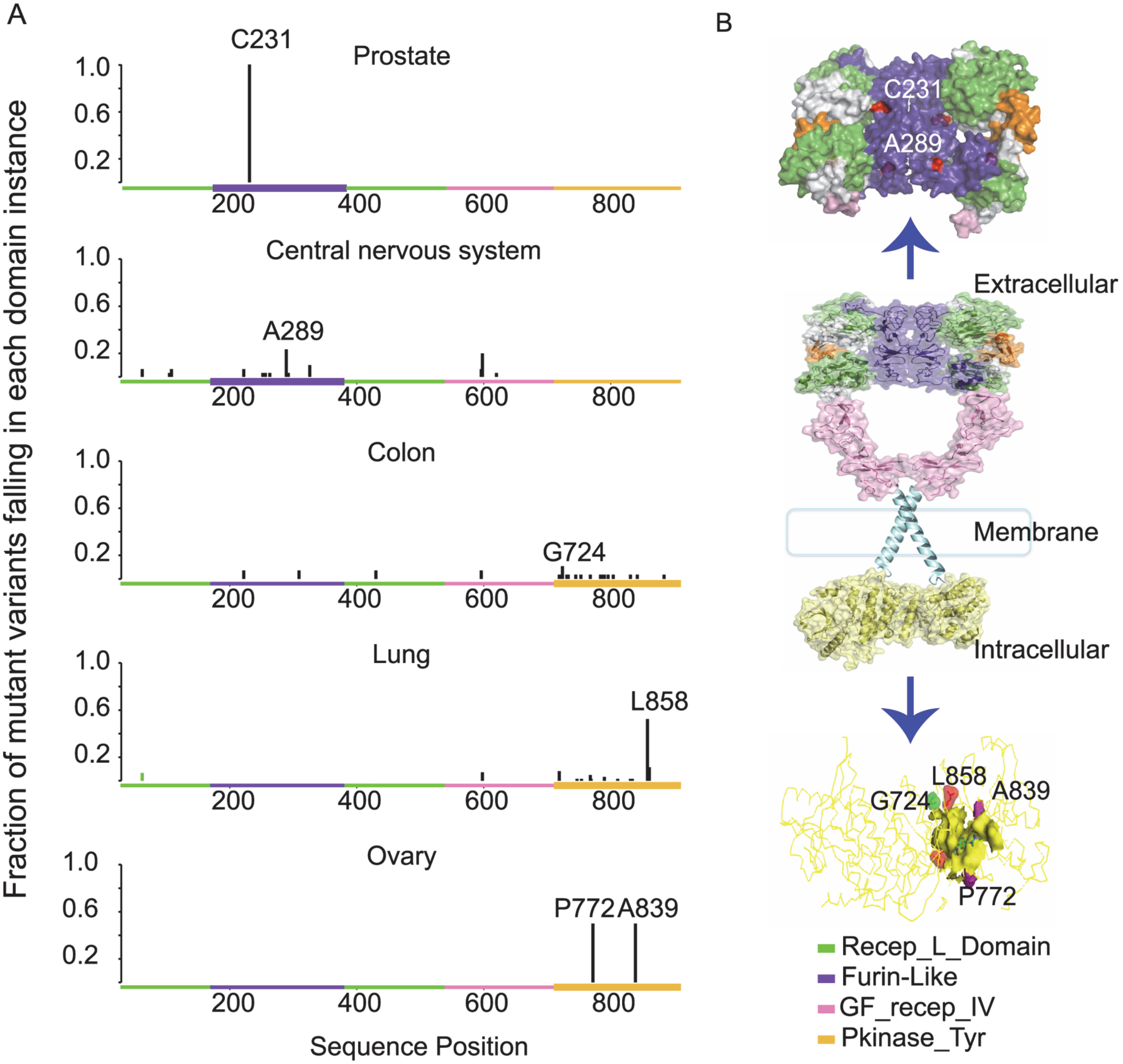

**Fig 4. Mutations in EGFR across 5 different cancers with protein structure context.** (A) The histogram displays the proportions of mutation counts detected at each residue to the total number of mutations that fall in the four different domains encoded by the gene EGFR, in five different cancers. The x-axis indicates the position of mutant residues. Mutations in different domains are shown in different colors. (B) shows the structure of the EGFR protein with epidermal growth factors colored in orange. The arrows point to enlargments of portions of the protein. The tails of the kinase domain are not shown in this structure. The structure visualization was based on Protein Data Bank structure models 1nql, 1ivo, 2jwa, 1m17 and 2gs6[47–50]. Significantly-mutated domain instances (SMDs) were shown as thicker boxes.

doi:10.1371/journal.pcbi.1004147.g004

It has been proposed that oncoproteins tend to be recurrently mutated at the same amino acid residues, while tumor suppressor proteins tend to be mutated throughout their length[15]. Therefore we systematically compared the mutation pattern between tumor suppressor proteins and oncoproteins. In both tumor suppressor proteins and oncoproteins, mutations were enriched in the SMDs, as expected (Fig. 6A., S5 Table). For each cancer-type-specific SMD, to assess whether mutations were recurrent at a few locations as opposed to being evenly spread,

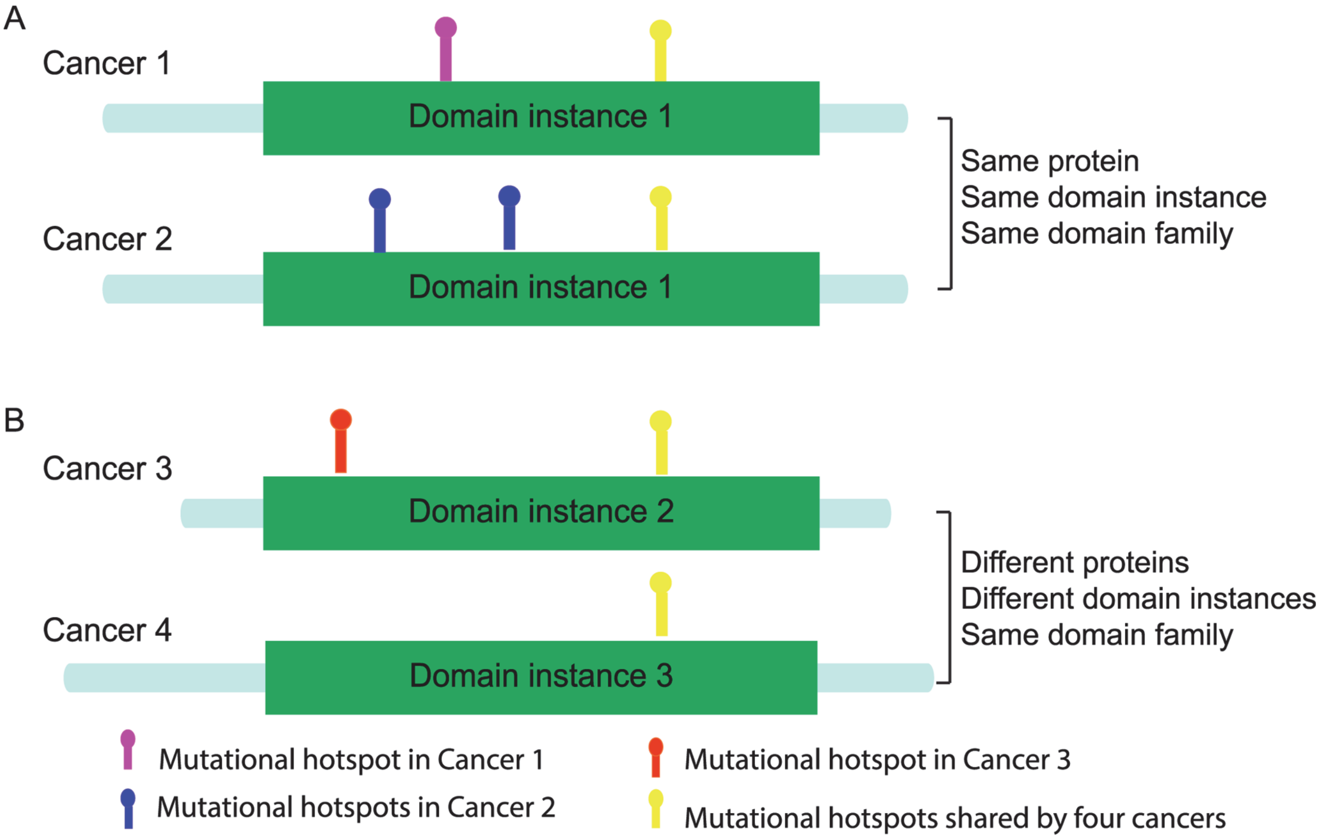

**Fig 5. Cancer-type-specific mutational hotspots and mutational hotspots shared by several cancer types.** A. shows the distribution of mutational hotspots for different cancer types within a given domain instance. B. shows mutational hotspot distribution patterns of different domain instances (encoded by different genes) that each correspond to the same protein domain type. Mutational hotspots are shown as balls and sticks, domain instances are shown as boxes. Mutational hotspots in different colors represent mutations in different cancer types.

doi:10.1371/journal.pcbi.1004147.g005

we compared the ratio of mutational hotspots to the total number of mutated residues within domains. We found this ratio to be significantly higher for oncoproteins than for tumor suppressor proteins (Fig. 6B.; *P* = 0.00026, Mann-Whitney U-test). This is consistent with the known tendency of tumor suppressor proteins to carry loss-of-function mutations that can occur in many places, and that of oncoproteins to harbor more specific gain-of-function mutations[15].

For example, the fibroblast growth factor receptor 2 (FGFR2) is generally regarded as an oncoprotein in breast cancer[15]. Consistent with this view, we found a single hotspot (p. N549) for FGFR2 in breast cancer in the kinase domain, which had not been reported as a hotspot for breast cancer. A previous study of endometrial cancer[54] suggested FGFR2 to be a tumor suppressor protein. Supporting this view, we observed nine evenly-distributed mutated residues in the kinase domain in endometrial cancer, although we also confirm previous observation [54] of the p.N549 hotspot which is more suggestive of an oncoprotein. Four mutational hotspots in the Immunoglobulin I-set domain of FGFR2 were observed in colon cancer, which hints at a tumor suppression role for *FGFR2* in colon cancer (Fig. 7).

We also analyzed the functional properties of the mutational hotspots we observed. Out of the 68 proteins that have at least one mutational hotspot in at least one cancer, we selected 13 proteins for which structures and functional site annotations were available. Of these 13, seven proteins are encoded by oncogenes (*AKT1, BRAF, EGFR, HRAS, KRAS, NRAS*, and *PIK3CA)*

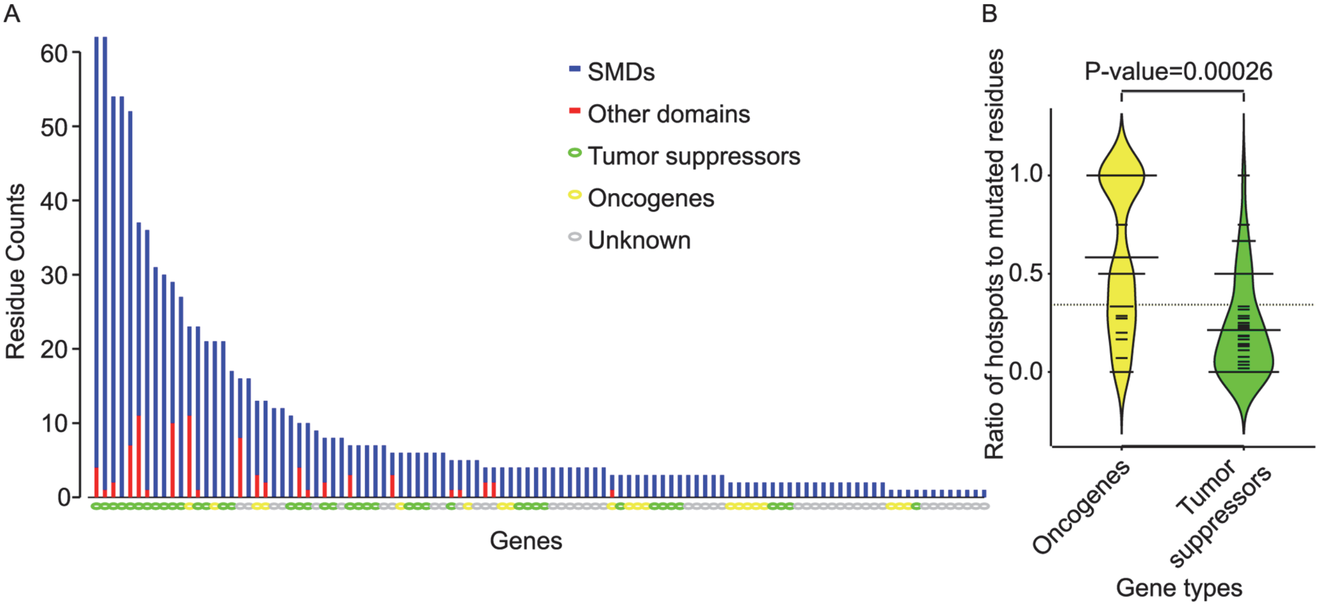

**Fig 6. Distribution of mutated residues within a single gene.** (A) compares the prevalence with which mutations from a specific cancer type fall within significantly mutated domain instances (SMDs) to the prevalence of mutations in other domain instances. Genes with at least one SMD are represented on *x*-axis in descending order by the number of mutated residues. The length of each blue bar shows the number of the mutated residues falling in SMDs for each cancer type, the length of red bars shown the number of mutated residues falling in other domain instances within the same gene. (B) compares the fraction of mutated residues in SMDs that are hotspots in oncogenes (yellow) and tumor suppressors (green).

doi:10.1371/journal.pcbi.1004147.g006

and six are encoded by tumor suppressors (*CDKN2A, FBXW7, PTEN, SMAD4, TP53*, and *VHL*). We mapped all observed mutations to protein structures. For tumor suppressor proteins, we found that most mutational hotspots fell at the interface of domain-domain interactions. We also found that, of 47 mutational hotspots, only 3 (6%) fell at functional sites of tumor suppressor proteins (Table 5). For oncoproteins, of 40 mutational hotspots, 15 (38%) fell at functional sites, including GTP/ATP binding sites and other active sites of enzymes. Functional sites were significantly overrepresented among oncogenic mutational hotspots (Odds ratio = 10.0, $P = 0.0006$, Fisher's Exact Test).

The three mutational hotspots detected at known functional sites of a tumor suppressor protein all fell within the p53 protein. The p.R248 and p.R273 hotspots were within the DNA binding site, and have each been reported as sites of potentially oncogenic mutations in many cancer types, including breast cancer[55]. The hotspot p.R337, found in liver cancer, fell within p53's tetramerization domain, a site of post-translational modification targeted by Protein Arginine N-Methyl Transferase 5 (PRMT5). Methylation of this residue affects the target protein specificity of p53[56, 57]. As shown in Fig. 8., the contact between p.R337 and p.L348, which is a residue in the P53-Tetramerization domain of another chain, may be necessary for tetramerization of the whole protein. The tetramerization of different domains is reported to be essential for the activity of p53[58]. Disruption of tetramerization could have a dominant-negative loss-of-function effect, or a gain- or change-of-function mutation if the un-tetramerized subunits have additional activities. Thus, our analysis points to residue p.R337 being a novel driver mutation in liver cancer.

The different mutational hotspot distribution patterns between oncoproteins and tumor suppressor proteins were generally consistent with the expected gain- and loss-of-function mechanisms of oncogenes and tumor suppressors, respectively[15]. Mutations at functional sites may increase the activation of oncoproteins, while mutations at the inter-chain interfaces

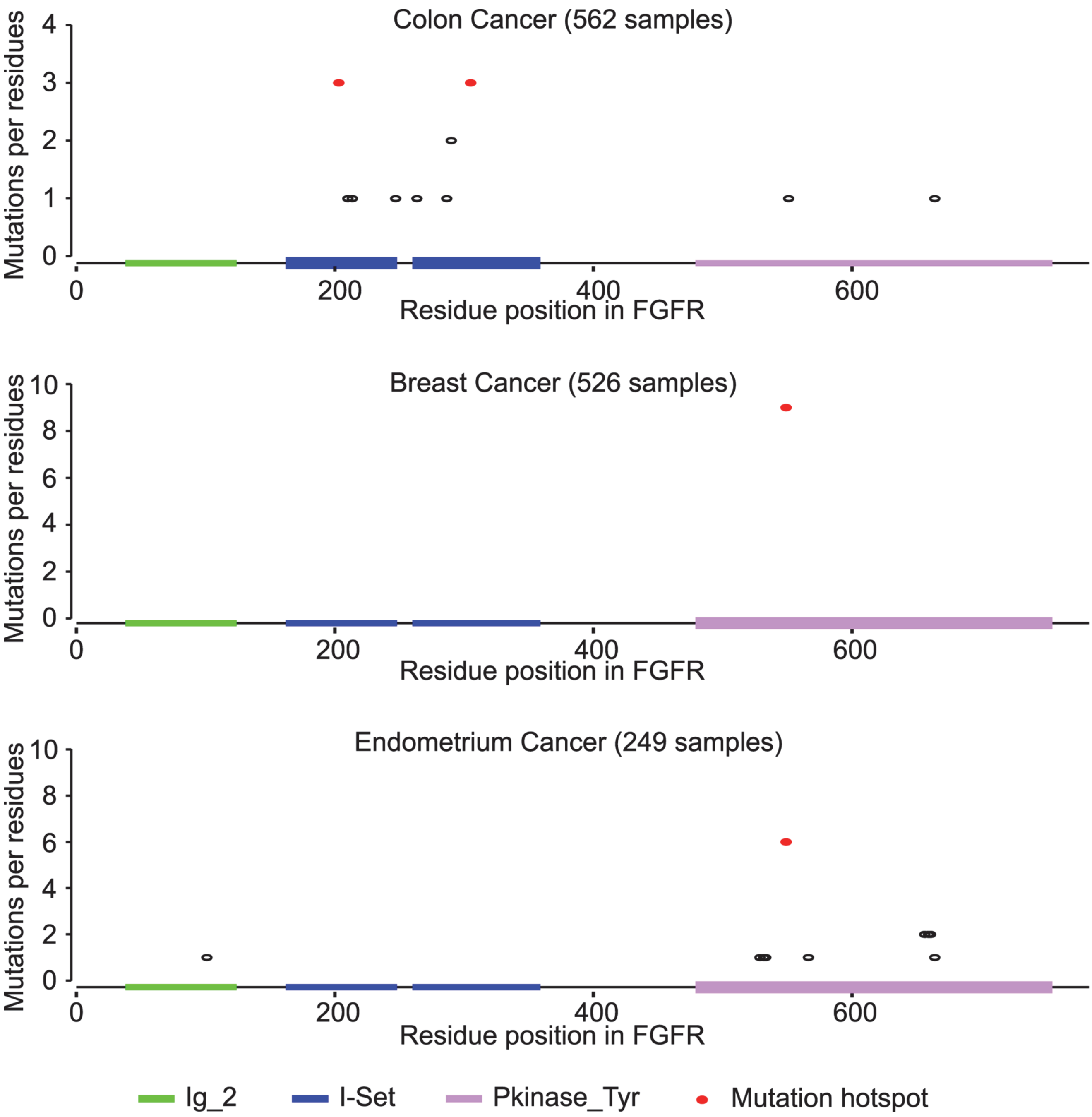

**Fig 7. Distribution of mutated residues in FGFR.** Sequence positions and frequencies of mutated residues in the FGFR protein are shown. Mutational hotspots for each cancer type are displayed as red dots. SMDs are shown as thicker boxes.

doi:10.1371/journal.pcbi.1004147.g007

may destabilize the protein and lead to loss of function in a tumor suppressor. These distinct mutation patterns can help classify newly identified cancer-associated genes for which oncogene or tumor suppressor roles are unknown. We categorized the ten novel cancer candidate genes that overlap with the Sleeping Beauty dataset based on similarity to the hotspot distribution patterns that are characteristic of oncogenes and tumor suppressors (Table 4). Among the ten genes, seven (*CNTN4, CTNNA3, EPHA6, FOXO1, MAGI1, PTPRD* and *SMAD4*) were

**Table 5. Genes that encode more than one cancer-type-specific significantly mutated domain instance.**

| Gene Symbol | Domain Instances per Gene | Significantly Mutated Domain Instance | Primary Site |
|---|---|---|---|
| *TP53* | 2 | P53 | Liver |
| *TP53* | 2 | P53_tetramer | Liver |
| *TP53* | 1 | P53 | Breast |
| *TP53* | 1 | P53 | Central nervous system |
| *TP53* | 1 | P53 | **Hematopoietic and lymph** |
| *TP53* | 1 | P53 | Lung |
| *TP53* | 1 | P53 | Esophagus |
| *TP53* | 1 | P53 | Ovary |
| *TP53* | 1 | P53 | Upper aero-digestive tract |
| *EGFR* | 2 | Furin-like | Central nervous system |
| *EGFR* | 2 | GF_recep_IV | Central nervous system |
| *EGFR* | 1 | Pkinase_Tyr | Lung |
| *KRAS* | 1 | Ras | Lung |
| *KRAS* | 1 | Ras | Pancreas |
| *MLL3* | 1 | zf-HC5HC2H | Breast |
| *MLL3* | 1 | zf-HC5HC2H | Prostate |
| *SMAD4* | 1 | MH2 | Colon |
| *SMAD4* | 1 | MH2 | Esophagus |
| *SMARCA4* | 1 | SNF2_N | Central nervous system |
| *SMARCA4* | 1 | SNF2_N | Esophagus |
| *BCL2* | 2 | BH4 | **Hematopoietic and lymph** |
| *BCL2* | 2 | Bcl-2 | **Hematopoietic and lymph** |
| *DDX3X* | 2 | DEAD | Central nervous system |
| *DDX3X* | 2 | Helicase_C | Central nervous system |
| *PIK3CA* | 2 | PI3Ka | Endometrium |
| *PIK3CA* | 2 | PI3K_p85B | Endometrium |

doi:10.1371/journal.pcbi.1004147.t005

predicted to be tumor suppressors. Using transposon insertion positions, the Sleeping Beauty study [60] had annotated three of these seven genes as loss of function (while not suggesting an annotation for the remaining four). We also reported two potential oncogenes, *USP25* and *GNA13* (Table 4). Finally, we identified one gene, *PCDH11X*, for which the domain mutation patterns suggest an oncogenic role in lymphoma and leukemia but a tumor suppressive role in glioblastoma and medulloblastoma.

## Oncogenic mutational hotspots appearing in multiple cancer types

At the domain level, we noticed that 10 out of 13 oncogenic mutational hotspots shared by at least three cancer types occurred at functional sites (Table 6). This is true not only for domains corresponding to a single gene but also for domain types corresponding to different genes. For example, the Ras domain type (for which instances may be found in multiple genes) was significantly mutated in different cancers (Fig. 9A.). Enrichment of somatic mutations within Ras domains has been reported for different individual genes[61, 62]. Here, we collectively analyzed the domain position-based hotspots for K-RAS, H-RAS, and N-RAS, finding that at least one of the GTP binding site residues p.G12 or p.G13, or the active site residue p.R61 show a relatively high mutation rate in at least five cancer types (Fig. 9B and C.). While each of these three hotspots was known previously for individual genes in individual cancer types, this analysis

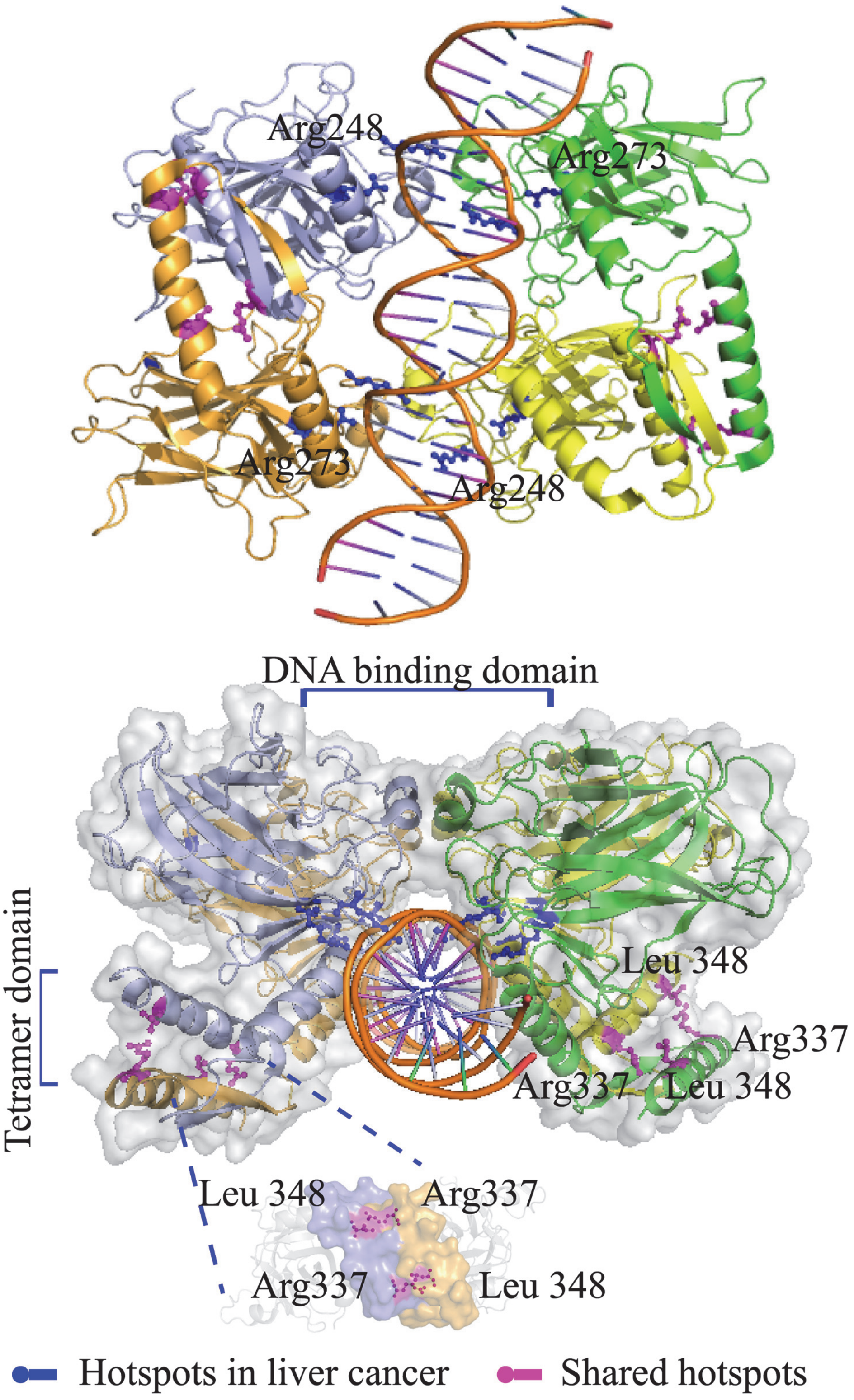

**Fig 8. Structural context of p53 protein (PDB 3q05[59]) mutational hotspots.** Mutational hotspots shared by eight cancers are displayed as blue sticks. Liver-cancer-specific mutational hotspots are displayed as magenta sticks. The p53 protein structure is colored according to amino acid chain.

doi:10.1371/journal.pcbi.1004147.g008

Table 6. Mutational hotspots observed at functional sites.

| Gene Symbol | Category | Mutational Hotspot | Functions |
|---|---|---|---|
| *AKT1* | Oncogene | p.E17 | PH–KD interaction |
| *BRAF* | Oncogene | p.G469 | ATP binding |
| *BRAF* | Oncogene | p.G466 | ATP binding |
| *BRAF* | Oncogene | p.N581 | Enzyme-active |
| *EGFR* | Oncogene | p.G719 | ATP binding |
| *EGFR* | Oncogene | p.G724 | ATP binding |
| *HRAS* | Oncogene | p.G13 | GTP binding |
| *HRAS* | Oncogene | p.Q61 | Enzyme-active |
| *KRAS* | Oncogene | p.G12 | GTP binding |
| *KRAS* | Oncogene | p.G13 | GTP binding |
| *NRAS* | Oncogene | p.G12 | GTP binding |
| *NRAS* | Oncogene | p.Q61 | Enzyme-active |
| *PIK3CA* | Oncogene | p.E545 | Intra-molecular binding |
| *PIK3CA* | Oncogene | p.E542 | Intra-molecular binding |
| *PIK3CA* | Oncogene | p.Q546 | Intra-molecular binding |
| *TP53* | Tumor suppressor | p.R248 | DNA binding |
| *TP53* | Tumor suppressor | p.R273 | DNA binding |
| *TP53* | Tumor suppressor | p.R337 | Post-translational modification |

doi:10.1371/journal.pcbi.1004147.t006

suggests that an increase in statistical power can be gained in the future by grouping protein domain instances of the same domain type.

Beyond the 10 out of 13 oncogenic mutational hotspots occurring at functional sites, there were three oncogenic mutational hotspots shared by at least three cancer types. They are V600 in Serine/threonine-protein kinase B-Raf (encoded by *BRAF*), and R88 and C420 in the phosphatidylinositol-4,5-bisphosphate 3-kinase encoded by *PIK3CA*. We found both C420 and R88 to be positions of mutational hotspots in endometrium, colon and breast cancer. Although the two residues fall within different domains (C420 in C2 domain, and R88 in p85α domain), they both play important roles in maintaining the p110α/ p85α-iSH2 structure [64], and both are at the binding interface of the C2 and p85α domains. Although each of these mutations has been previously studied as a potential driver mutation in each of these three cancer types, this analysis objectively confirms the 'hotspot' status of these mutations.

## Discussion

Major bottlenecks in the systematic study of cancer genomes exist following identification of somatic tumor mutations, including the identification of driver mutations and their functional impacts. By taking advantage of large-scale whole-genome or whole-exome sequencing data and accumulated information about protein structures, we were able to derive and compare the mutational landscapes for 21 cancer types at the domain level. We used a significance test that not only required a domain to have enriched mutational density in a given tumor type relative to other regions in that tumor type, but further required that an enrichment be significantly greater than that observed for all other cancer types taken together. Because region-dependence of mutation rates is similar across tumor types[25], this approach not only identifies cancer-type specific mutational positioning but also implicitly controls for regional differences in mutation rate across the genome.

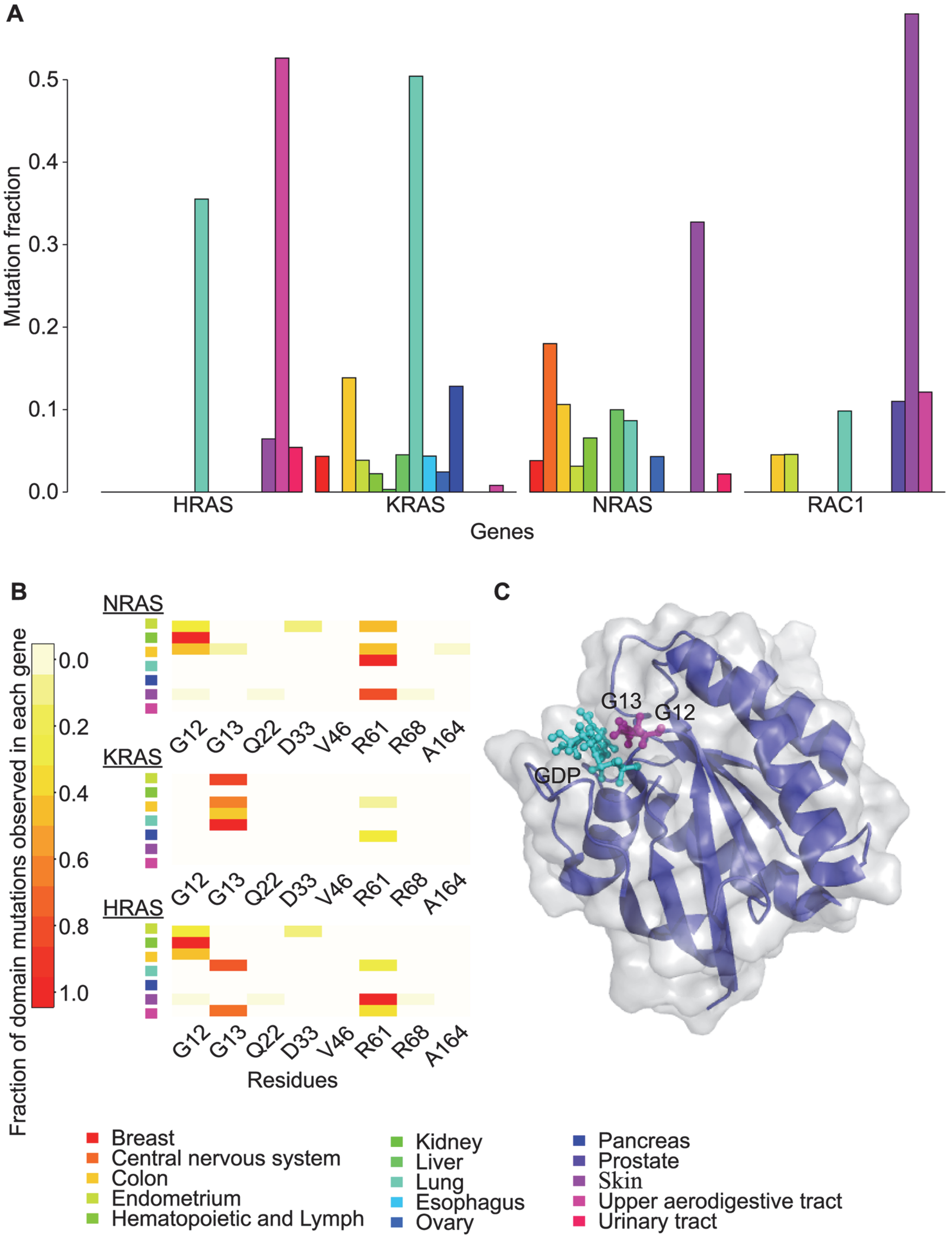

**Fig 9. Mutation distributions of different Ras domain instances and the structure of Ras domain.** (A) bar graph shows Ras domains encoded by different genes have different mutation rates across cancer types. (B) heat map shows fraction of mutations observed at each residue of a given gene in a given cancer. (C) the structure of the Ras domain encoded by the *KRAS* gene (PDB structural model 4lpk[63]). GTP/GDP binding sites are displayed as magenta sticks, GDP binding sites are colored in cyan.

doi:10.1371/journal.pcbi.1004147.g009

We analyzed domain types that are significantly mutated in different cancers, such as the Ras domain type and the Pkinase domain type. We found hotspots that were shared between different domain instances in the same domain type, and which appeared in multiple cancer types. By combining this information with protein structure information, we found that all (10 out of 10) such identified hotspots, where they fell within known oncoproteins, are 'functional hotspots' in the sense that all fell within ligand-binding or active sites. We also found that, in a given cancer type, a functional hotspot corresponding to a given domain type was never mutated in more than one of the domain instances corresponding to that domain type in the same tumor sample. (Sample information is shown in S1 Table.) This suggests that functional hotpots falling within different genes corresponding to a given domain type may contribute to cancer development by a similar and parallel mechanism, and further suggests that only one mutated functional site might be able to increase the activity of those proto-oncogenes and ultimately contribute to cancer initiation. Functional hotspots included oncogenic mutations within proteins that are generally considered to be tumor suppressors, for example p.R248 and p. R273 in *TP53*. Except for the DNA binding sites p.R248 and p.R273 in the p53 DNA binding domain, we did not find mutational hotspots in known tumor suppressors that appeared in more than five cancer types. Providing greater nuance to previous reports that mutations tend to span the entire tumor suppressor gene[15], we found that tumor suppressor mutations detected in a given cancer type tended to be distributed throughout the entirety of a significantly mutated domain instance, and many mutations occurred within core regions important for the stabilization of the protein complex. Mutations detected in different cancers tended to be focused within domain instances, but were distributed across different domain instances of the same gene product.

Mutational positioning information could assist drug design aimed at precisely targeting the region of the protein involved in a particular cancer. In contrast with gene-level studies of mutational frequency, the domain-level view points to particular functional regions, and identifies tendencies of a gene to be mutated in different regions in different cancers. For most genes, only one domain was found to be significantly mutated in a given cancer type. However, we found five genes that each contain two interaction-mediating domain instances that were significantly mutated in the same cancer type. These five interaction-mediating domain pairs are Bcl-2 and BH4 domains encoded by *BCL2*, which play important roles in regulating cell death and survival[65]; DEAD and Helicase_C domains encoded by *DDX3X*, which play important roles in metabolic processes involving RNAs[66]; and PI3Ka and PI3K_p85B domains encoded by *PIK3CA*, which interact with each other to initiate a vast array of signaling events[67]; Furin-like and GF_recep_IV domains encoded by *EGFR*, which are both extracellular domains of receptor tyrosine protein kinases and which interact with each other to regulate the binding of ligands to the receptor[68]; and finally the DNA binding and P53_tetramer (tetramerization) domains encoded by *TP53*. We also identified 117 domain instance pairs that corresponded to interacting proteins[69], for which at least one member of the pair was an SMD. For most interaction-mediating pairs, only one domain instance was significantly mutated (S6 Table). There are only ten cases where both domains of a predicted interaction-mediating domain pair were significantly mutated in the same cancer type (S7 Table). This result raises the possibility that mutations in those domain instances act by disrupting domain-domain interactions. Distinctive mutation landscapes in different cancers could indicate that tumor development mechanisms across different cancer types are dissimilar, although it is also possible that differences in the mutational spectrum between different cancer types alter the probability of mutation in one domain relative to another.

This domain-level study identified known and novel candidate driver mutations and provided clues to the functional effects of tumor-associated somatic mutations. In total, 41 out of the 100 SMDs we identified are encoded by Cancer Census genes (S1 Table). Among the remaining 59 novel candidate driver genes, many domain instances belong to well-known cancer-associated domain types, such as the Pkinase domain type and the WD40 domain type, supporting the idea that this set contains many cancer driver genes that are not yet annotated as such. By comparing the domain-level mutational landscapes of different cancers generated by our study to previously reported gene-level mutation landscapes in small cell lung cancer, melanoma, colon cancer, and breast cancer[14, 70–73], we noticed at least ten cancer-type-specific SMDs that do not correspond to any previously reported highly mutated cancer-associated genes. For each cancer type, we found at least one new potential cancer-associated domain instance, for example, the diacylglycerol kinase domain encoded by *DGKZ* in chondrosarcoma. The DGKZ protein (using the diacylglycerol kinase domain) usually acts as a sentinel and can control p53 function both during normal homeostasis and during stress response [74]. Other examples include the two cadherin domain instances encoded by *PCDH11X* and *PCDH11Y* in glioblastoma. These domain instances are thought to play important roles in cell-cell communication and are essential for a normally-functioning central nervous system [75]. Also, all the eight tumor samples that contained mutations in *PCDH11Y* (on the Y chromosome) were also mutated at a corresponding position in the X-chromosome homolog *PCDH11X*. Another four (female) samples had mutations detected on both alleles of *PCDH11X* [76]. These alleles each contained one of the novel hotspot mutations p.T486 or p.G442 in the cadherin domain, suggesting the potential role for these hotspot mutations as important recessive driver mutations in glioblastoma.

Because Nehrt *et al*.[32] had previously identified significantly mutated domain types for breast and colon cancer, we wished to assess the novelty of the SMDs we found for these cancer types. Of the 23 SMDs we identified for colon cancer, 20 were novel relative to domain types previously identified by Nehrt *et al* (we confirmed three domain types: the PI3K_p85B domain encoded by *PIK3CA*, the MH2 domain encoded by *SMAD4* and the P53 DNA binding domain encoded by *TP53*). Of the 12 SMDs we identified for breast cancer, only three correspond to a certain highly mutated domain type reported in the study by Nehrt *et al* (the PI3K_p85B domain and PI3Ka domain encoded by *PI3KCA*, and the P53 DNA binding domain encoded by *TP53*). We note that, even where an SMD corresponds to a domain type previously found to be significantly mutated, our analysis in this case identifies individual domain instances as significantly mutated, as opposed to domain types for which the mutations may be spread across multiple genes. In summary, 20 out of 23 (87%) of the colon-cancer associated SMDs, and 9 out of 12 (75%) of the breast-cancer-associated SMDs found here are novel relative to Nehrt *et al*.

Our study also differs from Nehrt *et al*.in that we only reported domain instances for which enrichment relative to other regions was significantly greater in one cancer type than in all other cancer types. This procedure controlled both for mutation rates within each cancer type, and for different rates of mutation across cancer types in each domain relative to others. Although a previous study[25] has pointed to the dangers of candidate driver gene identification through mutation frequency analysis, we note that none of the SMDs we identified fell within the 18 genes for which mutational enrichment was reported to be spurious[25]. In addition to correspondence of the discovered SMDs to known cancer-relevant domain families, our set of novel driver gene candidates overlapped significantly with a large-scale screen for cancer genes based on transposon mutagenesis in mouse. Together, these results indicate that we may be far from having a complete catalogue of cancer-associated genes and that domain-level mutation landscape analysis offers an opportunity to identify new driver genes.

We note that the cancer missense somatic mutation data we mined came from 71 unbiased studies, and that data from unbiased studies tends to contain a higher proportion of passenger

**Table 7. Domain position-based mutational hotspots shared by at least three cancers with functional annotations.**

| Domain Types | Corresponding Genes | Cancer Types | Mutational Hotspot | Function |
|---|---|---|---|---|
| **Ras domain** | *KRAS, NRAS, HRAS* | 8 | p.G12 | GTP binding |
| **Ras domain** | *KRAS, NRAS, HRAS* | 5 | p.G13 | GTP binding |
| **Ras domain** | *KRAS, NRAS, HRAS* | 5 | p.Q61 | Enzyme-active |
| **PH domain** | *AKT1* | 5 | p.E17 | Intra-molecular binding |
| **Phosphoinositide 3-kinase family** | *PIK3CA* | 7 | p.E545 | Intra-molecular binding |
| **Phosphoinositide 3-kinase family** | *PIK3CA* | 5 | p.E542 | Intra-molecular binding |
| **P53 DNA binding domain** | *TP53* | 14 | p.R248 | Contact with DNA |
| **P53 DNA binding domain** | *TP53* | 9 | p.G245 | Contact with DNA |
| **P53 DNA binding domain** | *TP53* | 6 | p.R273 | Contact with DNA |
| **P53 DNA binding domain** | *TP53* | 6 | p.R282 | Contact with DNA |

doi:10.1371/journal.pcbi.1004147.t007

mutations compared to data from targeted studies[77]. We therefore chose a relatively conservative significance threshold, necessarily causing us to overlook many candidate driver genes, which might be recovered in the future through larger data sets and use of prior information about cancer relatedness.

## Materials and Methods

To perform the study we first assembled a dataset of somatic mutations. Then, from this dataset we derived a dataset of potentially damaging missense somatic mutations. We analyzed cancer-type-specific SMDs and cancer-type-specific significantly-mutated position-based mutational hotspots. Finally, we analyzed the structural properties of those mutational hotspots.

### Creating the cancer missense somatic mutation dataset

We assembled a total of 237,716 missense somatic mutations in 21 cancer types (Table 7) from 71 whole-genome (WGS) or whole-exome sequencing (WES) studies[8, 14, 16, 24, 70–73, 78–132] included in the COSMIC (Catalogue of Somatic Mutations in Cancer) database (version 67)[127–129]. Most of those studies were conducted by either the International Cancer Genome Consortium (ICGC) [93] or The Cancer Genome Atlas (TCGA) project[136]. The mutations fell within a total of 18,682 genes, corresponding to 22,367 different protein isoforms. Amino acid sequences corresponding to the mutated protein isoforms were also available from the COSMIC database. We used all the protein sequences corresponding to those cancer-associated genes to search against Pfam domain types from the Pfam protein domain family database (version 27) [137], using an *E*-value cutoff of 0.001[138]. A total of 11,633 unique Pfam domain types, encoded by 18,682 mutated genes, were obtained from the Pfam database, considering all transcripts of these genes. Then we mapped the missense somatic mutations to protein domain positions after multiple sequence alignments using HMMER (v3.1b1)[139]. Where a given mutation could be assigned to multiple overlapping domain instances, we mapped the mutation to all of them. Significance of enrichment was calculated separately for each domain instance, so that the results for any given domain instance did not depend on the presence of other overlapping domain instances. We note that the vast majority of all mutations mapped only to a single domain instance (only 6 mutations can be mapped to different domain instances). Finally, among the 11,633 protein domain types, we found 6950 unique Pfam domain types that have at least one missense somatic mutation detected in the studies, all of which had an E-value <0 .0001 (10-fold more stringent than the Pfamrecommended threshold). These 6950 Pfam domain types corresponded to 29,302 unique domain instances. All source code used for extracting missense somatic mutations from COSMIC and mapping them to Pfam domains is provided as supporting material (S1 Protocol).

## Creating the dataset of potentially damaging missense somatic mutations

To predict potential damaging mutations, we used the IntOGen–mutation platform [36, 140], which classified the 237,716 missense somatic mutations into five categories: high, medium, low, unknown and none. We excluded mutations predicted to have no or unknown functional effects from further analyses. This left only 76,158 mutations as potential driver mutations. Those mutations were distributed in 4,509 unique domain types, corresponding to 14,083 genes (Table 1).

## Cancer-type-specific significantly-mutated domain instance analyses

To avoid the possible bias caused by different domain instance lengths and imbalanced sequencing frequency across cancer types, we calculated the cancer-type-specific mutation density as the total number of somatic missense mutations falling in the domain-encoding region of each gene, normalized by the corresponding cumulative domain instance length. We used the Fisher's Exact test to determine whether a certain domain instance is significantly mutated in a given cancer, using the "*stats*" package in R (http://www.r-project.org, [141]). The mutation counts for the R function corresponded to a 2 × 2 contingency table based on whether or not the mutations detected from each cancer type fell (or did not fall) within a given domain instance. We chose a *P*-value threshold ($\alpha = 10^{-7}$) yielding a false discovery rate (FDR) of less than 0.05. We made a heat map representation of the hierarchical clustering of SMDs in different cancers using the "*heatmap.2*" R package based on the −log (*P*-value) of each cancer-type-specific domain instance. We analyzed the tendency of SMDs to co-occur in the same patient sample using Fisher's Exact test ("*stats*" package in R). Also, genes containing at one or more SMDs were regarded as candidate cancer genes in this study. Overlap between our candidate gene set and Cancer Census genes and the Sleeping Beauty gene sets was also analyzed using the Fisher's Exact Test ("*stats*" package in R).

## Cancer-type-specific significantly-mutated position based mutational hotspot analyses

We calculated the mutational hotspots within each domain instance encoded by a single gene based on Fisher's Exact test with a *P*-value cutoff 0.01 (FDR <0.05). False discovery rate analysis was performed using Benjamini & Hochberg FDR[142]. We used the Mann–Whitney U-test to evaluate the significance of difference in distribution patterns of mutation residues between oncoproteins and tumor suppressor proteins. All of these analyses were conducted using the "*stats*" package in R.

## Structural properties for position based mutational hotspots analyses

We downloaded known protein-structure files from the Protein Data Bank[143]. For proteins that had more than one structure file, we chose one, favoring those with larger sequence length and higher crystallographic resolution. For domain-domain interface analysis, mutational hotspots were first mapped onto the available structures by using the Pymol Software (http://www.pymol.org; [144]). The interfacial residues of different domains in different chains were analyzed using Mechismo (http://mechismo.russelllab.org/), ProtInDB (PROTein-protein

INterface residues Data Base; Rafael A Jordan, Feihong Wu, Drena Dobbs and Vasant Honavar. unpublished results), and PDBePISA (Proteins, Interfaces, Surfaces and Assemblies; [145]) servers. We retrieved the functional-site information for those mutational hotspots from the Catalytic Site Atlas[146] and the PhosphoSite[147] databases. We used the odds ratio and Fisher's Exact Test to calculate the tendency of mutational hotspots in oncoproteins to occur at ATP/GTP binding sites or enzyme-active sites, as compared with mutational hotspots in tumor suppressor proteins.

## Supporting Information

**S1 Table. Mutations that are predicted to compromise the function of the harboring protein.**
(XLSX)

**S2 Table. Cancer-type-specific significantly mutated domain instances and corresponding genes in different cancers.** For each cancer type this table lists the significantly mutated domain instances (SMDs], corresponding gene symbols, and number of mutations in each domain instance.
(XLSX)

**S3 Table. Number of cancer types in which each domain instance was significantly mutated.**
(XLSX)

**S4 Table. List of mutational hotspots in different cancer types.**
(XLSX)

**S5 Table. Number of mutational hotspots and number of mutated residues in each domain instance for each cancer type.**
(XLSX)

**S6 Table. List of pairs of significantly mutated domain instances that corresponded to directly-interacting proteins.**
(XLSX)

**S7 Table. Ten pairs of domain instances that are inferred to mediate protein interaction with each other, for which both domains in the pair were found to be significantly mutated in the same cancer type.**
(XLSX)

**S1 Protocol. A 'zipped' file containing source code used for extracting missense somatic mutation information from the COSMIC database, and for mapping mutations to protein domain instances defined by PFAM.**
(ZIP)

## Acknowledgments

We thank Prof. Lincoln Stein and Prof. Frank Sicheri for insightful discussions and advice on this study.

## Author Contributions

Conceived and designed the experiments: FY FPR. Performed the experiments: FY EP FPR. Analyzed the data: FY FPR. Contributed reagents/materials/analysis tools: FY TR DEH MV FPR. Wrote the paper: FY EP TR DEH MV FPR.

## References

1. Luebeck EG, Moolgavkar SH. Multistage carcinogenesis and the incidence of colorectal cancer. Proc Natl Acad Sci U S A. 2002; 99(23):15095–100. Epub 2002/11/05. doi: 10.1073/pnas.222118199 PMID: 12415112; PubMed Central PMCID: PMC137549.
2. Knudson AG Jr. Mutation and cancer: statistical study of retinoblastoma. Proc Natl Acad Sci U S A. 1971; 68(4):820–3. Epub 1971/04/01. PMID: 5279523; PubMed Central PMCID: PMC389051.
3. Beerenwinkel N, Antal T, Dingli D, Traulsen A, Kinzler KW, Velculescu VE, et al. Genetic progression and the waiting time to cancer. PLoS computational biology. 2007; 3(11):e225. Epub 2007/11/14. doi: 10.1371/journal.pcbi.0030225 PMID: 17997597; PubMed Central PMCID: PMC2065895.
4. Wood LD, Parsons DW, Jones S, Lin J, Sjoblom T, Leary RJ, et al. The genomic landscapes of human breast and colorectal cancers. Science signaling. 2007; 318(5853):1108. PMID: 17932254
5. Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, et al. Patterns of somatic mutation in human cancer genomes. Nature. 2007; 446(7132):153–8. PMID: 17344846
6. International Cancer Genome C, Hudson TJ, Anderson W, Artez A, Barker AD, Bell C, et al. International network of cancer genome projects. Nature. 2010; 464(7291):993–8. Epub 2010/04/16. doi: 10.1038/nature08987 PMID: 20393554; PubMed Central PMCID: PMC2902243.
7. Mitra K, Carvunis A-R, Ramesh SK, Ideker T. Integrative approaches for finding modular structure in biological networks. Nature Reviews Genetics. 2013; 14(10):719–32. doi: 10.1038/nrg3552 PMID: 24045689
8. Grasso CS, Wu YM, Robinson DR, Cao X, Dhanasekaran SM, Khan AP, et al. The mutational landscape of lethal castration-resistant prostate cancer. Nature. 2012; 487(7406):239–43. Epub 2012/06/23. doi: 10.1038/nature11125 PMID: 22722839; PubMed Central PMCID: PMC3396711.
9. Stephens PJ, Tarpey PS, Davies H, Van Loo P, Greenman C, Wedge DC, et al. The landscape of cancer genes and mutational processes in breast cancer. Nature. 2012; 486(7403):400–4. Epub 2012/06/23. doi: 10.1038/nature11017 PMID: 22722201; PubMed Central PMCID: PMC3428862.
10. Wood LD, Parsons DW, Jones S, Lin J, Sjoblom T, Leary RJ, et al. The genomic landscapes of human breast and colorectal cancers. Science. 2007; 318(5853):1108–13. Epub 2007/10/13. doi: 10.1126/science.1145720 PMID: 17932254.
11. Stephens PJ, McBride DJ, Lin M-L, Varela I, Pleasance ED, Simpson JT, et al. Complex landscapes of somatic rearrangement in human breast cancer genomes. Nature. 2009; 462(7276):1005–10. doi: 10.1038/nature08645 PMID: 20033038
12. Parsons DW, Li M, Zhang X, Jones S, Leary RJ, Lin JC, et al. The genetic landscape of the childhood cancer medulloblastoma. Science. 2011; 331(6016):435–9. Epub 2010/12/18. doi: 10.1126/science.1198056 PMID: 21163964; PubMed Central PMCID: PMC3110744.
13. Stransky N, Egloff AM, Tward AD, Kostic AD, Cibulskis K, Sivachenko A, et al. The mutational landscape of head and neck squamous cell carcinoma. Science. 2011; 333(6046):1157–60. Epub 2011/07/30. doi: 10.1126/science.1208130 PMID: 21798893; PubMed Central PMCID: PMC3415217.
14. Hodis E, Watson IR, Kryukov GV, Arold ST, Imielinski M, Theurillat JP, et al. A landscape of driver mutations in melanoma. Cell. 2012; 150(2):251–63. Epub 2012/07/24. doi: 10.1016/j.cell.2012.06.024 PMID: 22817889; PubMed Central PMCID: PMC3600117.
15. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA Jr., Kinzler KW. Cancer genome landscapes. Science. 2013; 339(6127):1546–58. Epub 2013/03/30. doi: 10.1126/science.1235122 PMID: 23539594; PubMed Central PMCID: PMC3749880.
16. Watson IR, Takahashi K, Futreal PA, Chin L. Emerging patterns of somatic mutations in cancer. Nature reviews Genetics. 2013; 14(10):703–18. Epub 2013/09/12. doi: 10.1038/nrg3539 PMID: 24022702.
17. Pleasance ED, Cheetham RK, Stephens PJ, McBride DJ, Humphray SJ, Greenman CD, et al. A comprehensive catalogue of somatic mutations from a human cancer genome. Nature. 2010; 463(7278):191–6. Epub 2009/12/18. doi: 10.1038/nature08658 PMID: 20016485; PubMed Central PMCID: PMC3145108.

18. Beroukhim R, Mermel CH, Porter D, Wei G, Raychaudhuri S, Donovan J, et al. The landscape of somatic copy-number alteration across human cancers. Nature. 2010; 463(7283):899–905. Epub 2010/02/19. doi: 10.1038/nature08822 PMID: 20164920; PubMed Central PMCID: PMC2826709.

19. Boyko AR, Williamson SH, Indap AR, Degenhardt JD, Hernandez RD, Lohmueller KE, et al. Assessing the evolutionary impact of amino acid mutations in the human genome. PLoS genetics. 2008; 4(5): e1000083. Epub 2008/06/03. doi: 10.1371/journal.pgen.1000083 PMID: 18516229; PubMed Central PMCID: PMC2377339.

20. Haeno H, Iwasa Y, Michor F. The evolution of two mutations during clonal expansion. Genetics. 2007; 177(4):2209–21. Epub 2007/12/13. doi: 10.1534/genetics.107.078915 PMID: 18073428; PubMed Central PMCID: PMC2219486.

21. Pleasance ED, Stephens PJ, O'Meara S, McBride DJ, Meynert A, Jones D, et al. A small-cell lung cancer genome with complex signatures of tobacco exposure. Nature. 2010; 463(7278):184–90. Epub 2009/12/18. doi: 10.1038/nature08629 PMID: 20016488; PubMed Central PMCID: PMC2880489.

22. McFarland CD, Korolev KS, Kryukov GV, Sunyaev SR, Mirny LA. Impact of deleterious passenger mutations on cancer progression. Proc Natl Acad Sci U S A. 2013; 110(8):2910–5. Epub 2013/02/08. doi: 10.1073/pnas.1213968110 PMID: 23388632; PubMed Central PMCID: PMC3581883.

23. Stratton MR. Exploring the genomes of cancer cells: progress and promise. Science. 2011; 331 (6024):1553–8. Epub 2011/03/26. doi: 10.1126/science.1204040 PMID: 21436442.

24. Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, et al. Patterns of somatic mutation in human cancer genomes. Nature. 2007; 446(7132):153–8. Epub 2007/03/09. doi: 10.1038/nature05610 PMID: 17344846; PubMed Central PMCID: PMC2712719.

25. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013; 499(7457):214–8. doi: 10.1038/nature12213 PMID: 23770567; PubMed Central PMCID: PMC3919509.

26. Bignell GR, Greenman CD, Davies H, Butler AP, Edkins S, Andrews JM, et al. Signatures of mutation and selection in the cancer genome. Nature. 2010; 463(7283):893–8. Epub 2010/02/19. doi: 10.1038/nature08768 PMID: 20164919; PubMed Central PMCID: PMC3145113.

27. Sjoblom T, Jones S, Wood LD, Parsons DW, Lin J, Barber TD, et al. The consensus coding sequences of human breast and colorectal cancers. Science. 2006; 314(5797):268–74. Epub 2006/09/09. doi: 10.1126/science.1133427 PMID: 16959974.

28. Dixit A, Verkhivker GM. Structure-functional prediction and analysis of cancer mutation effects in protein kinases. Computational and mathematical methods in medicine. 2014; 2014:653487. doi: 10.1155/2014/653487 PMID: 24817905; PubMed Central PMCID: PMC4000980.

29. Wan PT, Garnett MJ, Roe SM, Lee S, Niculescu-Duvaz D, Good VM, et al. Mechanism of activation of the RAF-ERK signaling pathway by oncogenic mutations of B-RAF. Cell. 2004; 116(6):855–67. PMID: 15035987.

30. Joerger AC, Fersht AR. Structure-function-rescue: the diverse nature of common p53 cancer mutants. Oncogene. 2007; 26(15):2226–42. doi: 10.1038/sj.onc.1210291 PMID: 17401432.

31. Dixit A, Yi L, Gowthaman R, Torkamani A, Schork NJ, Verkhivker GM. Sequence and structure signatures of cancer mutation hotspots in protein kinases. PloS one. 2009; 4(10):e7485. doi: 10.1371/journal.pone.0007485 PMID: 19834613; PubMed Central PMCID: PMC2759519.

32. Nehrt NL, Peterson TA, Park D, Kann MG. Domain landscapes of somatic mutations in cancer. BMC genomics. 2012; 13 Suppl 4:S9. Epub 2012/07/13. doi: 10.1186/1471-2164-13-S4-S9 PMID: 22759657; PubMed Central PMCID: PMC3394412.

33. Studer RA, Dessailly BH, Orengo CA. Residue mutations and their impact on protein structure and function: detecting beneficial and pathogenic changes. Biochem J. 2013; 449(3):581–94. Epub 2013/01/11. doi: 10.1042/BJ20121221 PMID: 23301657.

34. Zhong Q, Simonis N, Li QR, Charloteaux B, Heuze F, Klitgord N, et al. Edgetic perturbation models of human inherited disorders. Mol Syst Biol. 2009; 5:321. Epub 2009/11/06. doi: 10.1038/msb.2009.80 PMID: 19888216; PubMed Central PMCID: PMC2795474.

35. Sahni N, Yi S, Zhong Q, Jailkhani N, Charloteaux B, Cusick ME, et al. Edgotype: a fundamental link between genotype and phenotype. Current opinion in genetics & development. 2013; 23(6):649–57. doi: 10.1016/j.gde.2013.11.002 PMID: 24287335; PubMed Central PMCID: PMC3902775.

36. Gonzalez-Perez A, Perez-Llamas C, Deu-Pons J, Tamborero D, Schroeder MP, Jene-Sanz A, et al. IntOGen-mutations identifies cancer drivers across tumor types. Nature methods. 2013; 10 (11):1081–2. Epub 2013/09/17. doi: 10.1038/nmeth.2642 PMID: 24037244.

37. Dolle ME, Snyder WK, Gossen JA, Lohman PH, Vijg J. Distinct spectra of somatic mutations accumulated with age in mouse heart and small intestine. Proc Natl Acad Sci U S A. 2000; 97(15):8403–8. PMID: 10900004; PubMed Central PMCID: PMC26960.

38. Jackson AL, Loeb LA. The mutation rate and cancer. Genetics. 1998; 148(4):1483–90. PMID: 9560368; PubMed Central PMCID: PMC1460096.

39. Santarius T, Shipley J, Brewer D, Stratton MR, Cooper CS. A census of amplified and overexpressed human cancer genes. Nat Rev Cancer. 2010; 10(1):59–64. doi: 10.1038/nrc2771 PMID: 20029424.

40. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, et al. A census of human cancer genes. Nat Rev Cancer. 2004; 4(3):177–83. doi: 10.1038/nrc1299 PMID: 14993899; PubMed Central PMCID: PMC2665285.

41. Pruitt KD, Tatusova T, Klimke W, Maglott DR. NCBI Reference Sequences: current status, policy and new initiatives. Nucleic Acids Res. 2009; 37(Database issue):D32–6. doi: 10.1093/nar/gkn721 PMID: 18927115; PubMed Central PMCID: PMC2686572.

42. Wu C, Macleod I, Su AI. BioGPS and MyGene.info: organizing online, gene-centric information. Nucleic Acids Res. 2013; 41(Database issue):D561–5. doi: 10.1093/nar/gks1114 PMID: 23175613; PubMed Central PMCID: PMC3531157.

43. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, et al. A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci U S A. 2004; 101(16):6062–7. doi: 10.1073/pnas.0400782101 PMID: 15075390; PubMed Central PMCID: PMC395923.

44. Hishiki T, Kawamoto S, Morishita S, Okubo K. BodyMap: a human and mouse gene expression database. Nucleic Acids Res. 2000; 28(1):136–8. PMID: 10592203; PubMed Central PMCID: PMC102396.

45. Collier LS, Largaespada DA. Transposons for cancer gene discovery: Sleeping Beauty and beyond. Genome Biol. 2007; 8 Suppl 1:S15. doi: 10.1186/gb-2007-8-s1-s15 PMID: 18047692; PubMed Central PMCID: PMC2106843.

46. Lee JC, Vivanco I, Beroukhim R, Huang JH, Feng WL, DeBiasi RM, et al. Epidermal growth factor receptor activation in glioblastoma through novel missense mutations in the extracellular domain. PLoS medicine. 2006; 3(12):e485. PMID: 17177598

47. Ferguson KM, Berger MB, Mendrola JM, Cho H-S, Leahy DJ, Lemmon MA. EGF activates its receptor by removing interactions that autoinhibit ectodomain dimerization. Molecular cell. 2003; 11(2):507-17.

48. Bocharov EV, Mineev KS, Volynsky PE, Ermolyuk YS, Tkach EN, Sobol AG, et al. Spatial structure of the dimeric transmembrane domain of the growth factor receptor ErbB2 presumably corresponding to the receptor active state. Journal of Biological Chemistry. 2008; 283(11):6950-6.

49. Ogiso H, Ishitani R, Nureki O, Fukai S, Yamanaka M, Kim J-H, et al. Crystal structure of the complex of human epidermal growth factor and receptor extracellular domains. Cell. 2002; 110(6):775-87.

50. Stamos J, Sliwkowski MX, Eigenbrot C. Structure of the epidermal growth factor receptor kinase domain alone and in complex with a 4-anilinoquinazoline inhibitor. Journal of Biological Chemistry. 2002; 277(48):46265-72.

51. Sordella R, Bell DW, Haber DA, Settleman J. Gefitinib-sensitizing EGFR mutations in lung cancer activate anti-apoptotic pathways. Science. 2004; 305(5687):1163–7. Epub 2004/07/31. doi: 10.1126/science.1101637 PMID: 15284455.

52. Chen C, Liu Y, Rappaport Amy R, Kitzing T, Schultz N, Zhao Z, et al. MLL3 Is a Haploinsufficient 7q Tumor Suppressor in Acute Myeloid Leukemia. Cancer Cell. 25(5):652–65. doi: 10.1016/j.ccr.2014.03.016 PMID: 24794707

53. Lobry C, Oh P, Aifantis I. Oncogenic and tumor suppressor functions of Notch in cancer: it's NOTCH what you think. The Journal of experimental medicine. 2011; 208(10):1931–5. Epub 2011/09/29. doi: 10.1084/jem.20111855 PMID: 21948802; PubMed Central PMCID: PMC3182047.

54. Gatius S, Velasco A, Azueta A, Santacana M, Pallares J, Valls J, et al. FGFR2 alterations in endometrial carcinoma. Modern pathology: an official journal of the United States and Canadian Academy of Pathology, Inc. 2011; 24(11):1500–10. Epub 2011/07/05. doi: 10.1038/modpathol.2011.110 PMID: 21725289.

55. Sigal A, Rotter V. Oncogenic mutations of the p53 tumor suppressor: the demons of the guardian of the genome. Cancer research. 2000; 60(24):6788–93. Epub 2001/01/13. PMID: 11156366.

56. Jansson M, Durant ST, Cho EC, Sheahan S, Edelmann M, Kessler B, et al. Arginine methylation regulates the p53 response. Nature cell biology. 2008; 10(12):1431–9. doi: 10.1038/ncb1802 PMID: 19011621.

57. Scoumanne A, Chen X. Protein methylation: a new mechanism of p53 tumor suppressor regulation. Histology and histopathology. 2008; 23(9):1143–9. PMID: 18581285; PubMed Central PMCID: PMC2762123.

58. Chene P. The role of tetramerization in p53 function. Oncogene. 2001; 20(21):2611–7. Epub 2001/06/23. doi: 10.1038/sj.onc.1204373 PMID: 11420672.

59. Petty TJ, Emamzadah S, Costantino L, Petkova I, Stavridi ES, Saven JG, et al. An induced fit mechanism regulates p53 DNA binding kinetics to confer sequence specificity. The EMBO journal. 2011; 30 (11):2167-76.

60. Dupuy AJ, Jenkins NA, Copeland NG. Sleeping beauty: a novel cancer gene discovery tool. Hum Mol Genet. 2006; 15 Spec No 1:R75–9. doi: 10.1093/hmg/ddl061 PMID: 16651372.

61. Prior IA, Lewis PD, Mattos C. A comprehensive survey of Ras mutations in cancer. Cancer research. 2012; 72(10):2457–67. doi: 10.1158/0008-5472.CAN-11-2612 PMID: 22589270; PubMed Central PMCID: PMC3354961.

62. Fernandez-Medarde A, Santos E. Ras in cancer and developmental diseases. Genes & cancer. 2011; 2(3):344–58. Epub 2011/07/23. doi: 10.1177/1947601911411084 PMID: 21779504; PubMed Central PMCID: PMC3128640.

63. Ostrem JM, Peters U, Sos ML, Wells JA, Shokat KM. K-Ras (G12C) inhibitors allosterically control GTP affinity and effector interactions. Nature. 2013; 503(7477):548-51.

64. Hon WC, Berndt A, Williams RL. Regulation of lipid binding underlies the activation mechanism of class IA PI3-kinases. Oncogene. 2012; 31(32):3655–66. doi: 10.1038/onc.2011.532 PMID: 22120714; PubMed Central PMCID: PMC3378484.

65. Kelekar A, Thompson CB. Bcl-2-family proteins: the role of the BH3 domain in apoptosis. Trends in cell biology. 1998; 8(8):324–30. Epub 1998/08/15. PMID: 9704409.

66. Owsianka AM, Patel AH. Hepatitis C virus core protein interacts with a human DEAD box protein DDX3. Virology. 1999; 257(2):330–40. Epub 1999/05/18. doi: 10.1006/viro.1999.9659 PMID: 10329544.

67. Miled N, Yan Y, Hon WC, Perisic O, Zvelebil M, Inbar Y, et al. Mechanism of two classes of cancer mutations in the phosphoinositide 3-kinase catalytic subunit. Science. 2007; 317(5835):239–42. Epub 2007/07/14. doi: 10.1126/science.1135394 PMID: 17626883.

68. Cho HS, Leahy DJ. Structure of the extracellular region of HER3 reveals an interdomain tether. Science. 2002; 297(5585):1330–3. Epub 2002/08/03. doi: 10.1126/science.1074611 PMID: 12154198.

69. Rolland T, Ta An M, Charloteaux B, Pevzner SJ, Zhong Q, Sahni N, et al. A proteome-scale map of the human interactome network. Cell. 2014; 159(5):1212–26. doi: 10.1016/j.cell.2014.10.050 PMID: 25416956.

70. Stephens PJ, Tarpey PS, Davies H, Van Loo P, Greenman C, Wedge DC, et al. The landscape of cancer genes and mutational processes in breast cancer. Nature. 2012; 486(7403):400–4. Epub 2012/06/23. doi: 10.1038/nature11017 PMID: 22722201; PubMed Central PMCID: PMC3428862.

71. Cancer Genome Atlas N. Comprehensive molecular characterization of human colon and rectal cancer. Nature. 2012; 487(7407):330–7. Epub 2012/07/20. doi: 10.1038/nature11252 PMID: 22810696; PubMed Central PMCID: PMC3401966.

72. Kim SC, Jung Y, Park J, Cho S, Seo C, Kim J, et al. A high-dimensional, deep-sequencing study of lung adenocarcinoma in female never-smokers. PLoS One. 2013; 8(2):e55596. Epub 2013/02/14. doi: 10.1371/journal.pone.0055596 PMID: 23405175; PubMed Central PMCID: PMC3566005.

73. Imielinski M, Berger AH, Hammerman PS, Hernandez B, Pugh TJ, Hodis E, et al. Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. Cell. 2012; 150(6):1107–20. Epub 2012/09/18. doi: 10.1016/j.cell.2012.08.029 PMID: 22980975; PubMed Central PMCID: PMC3557932.

74. Tanaka T, Okada M, Hozumi Y, Tachibana K, Kitanaka C, Hamamoto Y, et al. Cytoplasmic localization of DGKzeta exerts a protective effect against p53-mediated cytotoxicity. Journal of cell science. 2013; 126(Pt 13):2785–97. Epub 2013/04/23. doi: 10.1242/jcs.118711 PMID: 23606744.

75. Yoshida K, Sugano S. Identification of a novel protocadherin gene (PCDH11) on the human XY homology region in Xq21.3. Genomics. 1999; 62(3):540–3. doi: 10.1006/geno.1999.6042 PMID: 10644456.

76. Lemaire M, Fremeaux-Bacchi V, Schaefer F, Choi M, Tang WH, Le Quintrec M, et al. Recessive mutations in DGKE cause atypical hemolytic-uremic syndrome. Nat Genet. 2013; 45(5):531–6. Epub 2013/04/02. doi: 10.1038/ng.2590 PMID: 23542698; PubMed Central PMCID: PMC3719402.

77. Tamborero D, Gonzalez-Perez A, Perez-Llamas C, Deu-Pons J, Kandoth C, Reimand J, et al. Comprehensive identification of mutational cancer driver genes across 12 tumor types. Scientific reports. 2013; 3:2650. doi: 10.1038/srep02650 PMID: 24084849; PubMed Central PMCID: PMC3788361.

78. Ciriello G, Miller ML, Aksoy BA, Senbabaoglu Y, Schultz N, Sander C. Emerging landscape of oncogenic signatures across human cancers. Nat Genet. 2013; 45(10):1127–33. Epub 2013/09/28. doi: 10.1038/ng.2762 PMID: 24071851.

79. Zhou D, Yang L, Zheng L, Ge W, Li D, Zhang Y, et al. Exome capture sequencing of adenoma reveals genetic alterations in multiple cellular pathways at the early stage of colorectal tumorigenesis. PLoS One. 2013; 8(1):e53310. Epub 2013/01/10. doi: 10.1371/journal.pone.0053310 PMID: 23301059; PubMed Central PMCID: PMC3534699.

80. Zhang J, Grubor V, Love CL, Banerjee A, Richards KL, Mieczkowski PA, et al. Genetic heterogeneity of diffuse large B-cell lymphoma. Proc Natl Acad Sci U S A. 2013; 110(4):1398–403. Epub 2013/01/08. doi: 10.1073/pnas.1205299110 PMID: 23292937; PubMed Central PMCID: PMC3557051.

81. Yost SE, Pastorino S, Rozenzhak S, Smith EN, Chao YS, Jiang P, et al. High-resolution mutational profiling suggests the genetic validity of glioblastoma patient-derived pre-clinical models. PLoS One. 2013; 8(2):e56185. Epub 2013/02/27. doi: 10.1371/journal.pone.0056185 PMID: 23441165; PubMed Central PMCID: PMC3575368.

82. Yan XJ, Xu J, Gu ZH, Pan CM, Lu G, Shen Y, et al. Exome sequencing identifies somatic mutations of DNA methyltransferase gene DNMT3A in acute monocytic leukemia. Nat Genet. 2011; 43(4):309–15. Epub 2011/03/15. doi: 10.1038/ng.788 PMID: 21399634.

83. Wang L, Tsutsumi S, Kawaguchi T, Nagasaki K, Tatsuno K, Yamamoto S, et al. Whole-exome sequencing of human pancreatic cancers and characterization of genomic instability caused by MLH1 haploinsufficiency and complete deficiency. Genome Res. 2012; 22(2):208–19. Epub 2011/12/14. doi: 10.1101/gr.123109.111 PMID: 22156295; PubMed Central PMCID: PMC3266029.

84. Totoki Y, Tatsuno K, Yamamoto S, Arai Y, Hosoda F, Ishikawa S, et al. High-resolution characterization of a hepatocellular carcinoma genome. Nat Genet. 2011; 43(5):464–9. Epub 2011/04/19. doi: 10.1038/ng.804 PMID: 21499249.

85. Tarpey PS, Behjati S, Cooke SL, Van Loo P, Wedge DC, Pillay N, et al. Frequent mutation of the major cartilage collagen gene COL2A1 in chondrosarcoma. Nat Genet. 2013; 45(8):923–6. Epub 2013/06/19. doi: 10.1038/ng.2668 PMID: 23770606; PubMed Central PMCID: PMC3743157.

86. Shah SP, Roth A, Goya R, Oloumi A, Ha G, Zhao Y, et al. The clonal and mutational evolution spectrum of primary triple-negative breast cancers. Nature. 2012; 486(7403):395–9. Epub 2012/04/13. doi: 10.1038/nature10933 PMID: 22495314; PubMed Central PMCID: PMC3863681.

87. Seshagiri S, Stawiski EW, Durinck S, Modrusan Z, Storm EE, Conboy CB, et al. Recurrent R-spondin fusions in colon cancer. Nature. 2012; 488(7413):660–4. Epub 2012/08/17. doi: 10.1038/nature11282 PMID: 22895193; PubMed Central PMCID: PMC3690621.

88. Seo JS, Ju YS, Lee WC, Shin JY, Lee JK, Bleazard T, et al. The transcriptional landscape and mutational profile of lung adenocarcinoma. Genome Res. 2012; 22(11):2109–19. Epub 2012/09/15. doi: 10.1101/gr.145144.112 PMID: 22975805; PubMed Central PMCID: PMC3483540.

89. Sausen M, Leary RJ, Jones S, Wu J, Reynolds CP, Liu X, et al. Integrated genomic analyses identify ARID1A and ARID1B alterations in the childhood cancer neuroblastoma. Nat Genet. 2013; 45(1):12–7. Epub 2012/12/04. doi: 10.1038/ng.2493 PMID: 23202128; PubMed Central PMCID: PMC3557959.

90. Rudin CM, Durinck S, Stawiski EW, Poirier JT, Modrusan Z, Shames DS, et al. Comprehensive genomic analysis identifies SOX2 as a frequently amplified gene in small-cell lung cancer. Nat Genet. 2012; 44(10):1111–6. Epub 2012/09/04. doi: 10.1038/ng.2405 PMID: 22941189; PubMed Central PMCID: PMC3557461.

91. Robinson G, Parker M, Kranenburg TA, Lu C, Chen X, Ding L, et al. Novel mutations target distinct subgroups of medulloblastoma. Nature. 2012; 488(7409):43–8. Epub 2012/06/23. doi: 10.1038/nature11213 PMID: 22722829; PubMed Central PMCID: PMC3412905.

92. Roberts SA, Lawrence MS, Klimczak LJ, Grimm SA, Fargo D, Stojanov P, et al. An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. Nat Genet. 2013; 45(9):970–6. Epub 2013/07/16. doi: 10.1038/ng.2702 PMID: 23852170; PubMed Central PMCID: PMC3789062.

93. Pugh TJ, Weeraratne SD, Archer TC, Pomeranz Krummel DA, Auclair D, Bochicchio J, et al. Medulloblastoma exome sequencing uncovers subtype-specific somatic mutations. Nature. 2012; 488(7409):106–10. Epub 2012/07/24. doi: 10.1038/nature11329 PMID: 22820256; PubMed Central PMCID: PMC3413789.

94. Pena-Llopis S, Vega-Rubin-de-Celis S, Liao A, Leng N, Pavia-Jimenez A, Wang S, et al. BAP1 loss defines a new class of renal cell carcinoma. Nat Genet. 2012; 44(7):751–9. Epub 2012/06/12. doi: 10.1038/ng.2323 PMID: 22683710; PubMed Central PMCID: PMC3788680.

95. null, Hudson TJ, Anderson W, Artez A, Barker AD, Bell C, et al. International network of cancer genome projects. Nature. 2010; 464(7291):993–8. doi: 10.1038/nature08987 PMID: 20393554

96. Nik-Zainal S, Alexandrov LB, Wedge DC, Van Loo P, Greenman CD, Raine K, et al. Mutational processes molding the genomes of 21 breast cancers. Cell. 2012; 149(5):979–93. Epub 2012/05/23. doi: 10.1016/j.cell.2012.04.024 PMID: 22608084; PubMed Central PMCID: PMC3414841.

97. Morin RD, Mendez-Lago M, Mungall AJ, Goya R, Mungall KL, Corbett RD, et al. Frequent mutation of histone-modifying genes in non-Hodgkin lymphoma. Nature. 2011; 476(7360):298–303. Epub 2011/07/29. doi: 10.1038/nature10351 PMID: 21796119; PubMed Central PMCID: PMC3210554.

98. Molenaar JJ, Koster J, Zwijnenburg DA, van Sluis P, Valentijn LJ, van der Ploeg I, et al. Sequencing of neuroblastoma identifies chromothripsis and defects in neuritogenesis genes. Nature. 2012; 483 (7391):589–93. Epub 2012/03/01. doi: 10.1038/nature10910 PMID: 22367537.

99. Liu J, Lee W, Jiang Z, Chen Z, Jhunjhunwala S, Haverty PM, et al. Genome and transcriptome sequencing of lung cancers reveal diverse mutational and splicing events. Genome Res. 2012; 22 (12):2315–27. Epub 2012/10/04. doi: 10.1101/gr.140988.112 PMID: 23033341; PubMed Central PMCID: PMC3514662.

100. Lindberg J, Mills IG, Klevebring D, Liu W, Neiman M, Xu J, et al. The mitochondrial and autosomal mutation landscapes of prostate cancer. Eur Urol. 2013; 63(4):702–8. Epub 2012/12/26. doi: 10.1016/j.eururo.2012.11.053 PMID: 23265383.

101. Li M, Zhao H, Zhang X, Wood LD, Anders RA, Choti MA, et al. Inactivating mutations of the chromatin remodeling gene ARID2 in hepatocellular carcinoma. Nat Genet. 2011; 43(9):828–9. Epub 2011/08/09. doi: 10.1038/ng.903 PMID: 21822264; PubMed Central PMCID: PMC3163746.

102. Leich E, Weissbach S, Klein HU, Grieb T, Pischimarov J, Stuhmer T, et al. Multiple myeloma is affected by multiple and heterogeneous somatic mutations in adhesion- and receptor tyrosine kinase signaling molecules. Blood Cancer J. 2013; 3:e102. Epub 2013/02/12. doi: 10.1038/bcj.2012.47 PMID: 23396385; PubMed Central PMCID: PMC3584721.

103. Lee RS, Stewart C, Carter SL, Ambrogio L, Cibulskis K, Sougnez C, et al. A remarkably simple genome underlies highly malignant pediatric rhabdoid cancers. J Clin Invest. 2012; 122(8):2983–8. Epub 2012/07/17. doi: 10.1172/JCI64400 PMID: 22797305; PubMed Central PMCID: PMC3408754.

104. Le Gallo M, O'Hara AJ, Rudd ML, Urick ME, Hansen NF, O'Neil NJ, et al. Exome sequencing of serous endometrial tumors identifies recurrent somatic mutations in chromatin-remodeling and ubiquitin ligase complex genes. Nat Genet. 2012; 44(12):1310–5. Epub 2012/10/30. doi: 10.1038/ng.2455 PMID: 23104009; PubMed Central PMCID: PMC3515204.

105. Krauthammer M, Kong Y, Ha BH, Evans P, Bacchiocchi A, McCusker JP, et al. Exome sequencing identifies recurrent somatic RAC1 mutations in melanoma. Nat Genet. 2012; 44(9):1006–14. Epub 2012/07/31. doi: 10.1038/ng.2359 PMID: 22842228; PubMed Central PMCID: PMC3432702.

106. Kannan K, Inagaki A, Silber J, Gorovets D, Zhang J, Kastenhuber ER, et al. Whole-exome sequencing identifies ATRX mutation as a key molecular determinant in lower-grade glioma. Oncotarget. 2012; 3(10):1194–203. Epub 2012/10/30. PMID: 23104868; PubMed Central PMCID: PMC3717947.

107. Jones S, Wang TL, Shih Ie M, Mao TL, Nakayama K, Roden R, et al. Frequent mutations of chromatin remodeling gene ARID1A in ovarian clear cell carcinoma. Science. 2010; 330(6001):228–31. Epub 2010/09/10. doi: 10.1126/science.1196333 PMID: 20826764; PubMed Central PMCID: PMC3076894.

108. Jones DT, Jager N, Kool M, Zichner T, Hutter B, Sultan M, et al. Dissecting the genomic complexity underlying medulloblastoma. Nature. 2012; 488(7409):100–5. Epub 2012/07/27. doi: 10.1038/nature11284 PMID: 22832583; PubMed Central PMCID: PMC3662966.

109. Jiao Y, Shi C, Edil BH, de Wilde RF, Klimstra DS, Maitra A, et al. DAXX/ATRX, MEN1, and mTOR pathway genes are frequently altered in pancreatic neuroendocrine tumors. Science. 2011; 331 (6021):1199–203. Epub 2011/01/22. doi: 10.1126/science.1200609 PMID: 21252315; PubMed Central PMCID: PMC3144496.

110. Iyer G, Hanrahan AJ, Milowsky MI, Al-Ahmadie H, Scott SN, Janakiraman M, et al. Genome sequencing identifies a basis for everolimus sensitivity. Science. 2012; 338(6104):221. Epub 2012/08/28. doi: 10.1126/science.1226344 PMID: 22923433; PubMed Central PMCID: PMC3633467.

111. Huang J, Deng Q, Wang Q, Li KY, Dai JH, Li N, et al. Exome sequencing of hepatitis B virus-associated hepatocellular carcinoma. Nat Genet. 2012; 44(10):1117–21. Epub 2012/08/28. doi: 10.1038/ng.2391 PMID: 22922871.

112. Ho AS, Kannan K, Roy DM, Morris LG, Ganly I, Katabi N, et al. The mutational landscape of adenoid cystic carcinoma. Nat Genet. 2013; 45(7):791–8. Epub 2013/05/21. doi: 10.1038/ng.2643 PMID: 23685749; PubMed Central PMCID: PMC3708595.

113. Guo G, Gui Y, Gao S, Tang A, Hu X, Huang Y, et al. Frequent mutations of genes encoding ubiquitin-mediated proteolysis pathway components in clear cell renal cell carcinoma. Nat Genet. 2012; 44 (1):17–9. Epub 2011/12/06. doi: 10.1038/ng.1014 PMID: 22138691.

114. Guichard C, Amaddeo G, Imbeaud S, Ladeiro Y, Pelletier L, Maad IB, et al. Integrated analysis of somatic mutations and focal copy-number changes identifies key genes and pathways in hepatocellular carcinoma. Nat Genet. 2012; 44(6):694–8. Epub 2012/05/09. doi: 10.1038/ng.2256 PMID: 22561517; PubMed Central PMCID: PMC3819251.

**115.** Green MR, Gentles AJ, Nair RV, Irish JM, Kihira S, Liu CL, et al. Hierarchy in somatic mutations arising during genomic evolution and progression of follicular lymphoma. Blood. 2013; 121(9):1604–11. Epub 2013/01/09. doi: 10.1182/blood-2012-09-457283 PMID: 23297126; PubMed Central PMCID: PMC3587323.

**116.** Gerlinger M, Rowan AJ, Horswell S, Larkin J, Endesfelder D, Gronroos E, et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. N Engl J Med. 2012; 366(10):883–92. Epub 2012/03/09. doi: 10.1056/NEJMoa1113205 PMID: 22397650.

**117.** Galante PA, Parmigiani RB, Zhao Q, Caballero OL, de Souza JE, Navarro FC, et al. Distinct patterns of somatic alterations in a lymphoblastoid and a tumor genome derived from the same individual. Nucleic Acids Res. 2011; 39(14):6056–68. Epub 2011/04/16. doi: 10.1093/nar/gkr221 PMID: 21493686; PubMed Central PMCID: PMC3152357.

**118.** Durinck S, Ho C, Wang NJ, Liao W, Jakkula LR, Collisson EA, et al. Temporal dissection of tumorigenesis in primary cancers. Cancer Discov. 2011; 1(2):137–43. Epub 2011/10/11. doi: 10.1158/2159-8290.CD-11-0028 PMID: 21984974; PubMed Central PMCID: PMC3187561.

**119.** Duns G, Hofstra RM, Sietzema JG, Hollema H, van Duivenbode I, Kuik A, et al. Targeted exome sequencing in clear cell renal cell carcinoma tumors suggests aberrant chromatin regulation as a crucial step in ccRCC development. Hum Mutat. 2012; 33(7):1059–62. Epub 2012/03/31. doi: 10.1002/humu.22090 PMID: 22461374.

**120.** Dulak AM, Stojanov P, Peng S, Lawrence MS, Fox C, Stewart C, et al. Exome and whole-genome sequencing of esophageal adenocarcinoma identifies recurrent driver events and mutational complexity. Nat Genet. 2013; 45(5):478–86. Epub 2013/03/26. doi: 10.1038/ng.2591 PMID: 23525077; PubMed Central PMCID: PMC3678719.

**121.** Ding L, Ley TJ, Larson DE, Miller CA, Koboldt DC, Welch JS, et al. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. Nature. 2012; 481(7382):506–10. Epub 2012/01/13. doi: 10.1038/nature10738 PMID: 22237025; PubMed Central PMCID: PMC3267864.

**122.** Clark VE, Erson-Omay EZ, Serin A, Yin J, Cotney J, Ozduman K, et al. Genomic analysis of non-NF2 meningiomas reveals mutations in TRAF7, KLF4, AKT1, and SMO. Science. 2013; 339(6123):1077–80. Epub 2013/01/26. doi: 10.1126/science.1233009 PMID: 23348505.

**123.** Cancer Genome Atlas Research N, Kandoth C, Schultz N, Cherniack AD, Akbani R, Liu Y, et al. Integrated genomic characterization of endometrial carcinoma. Nature. 2013; 497(7447):67–73. Epub 2013/05/03. doi: 10.1038/nature12113 PMID: 23636398; PubMed Central PMCID: PMC3704730.

**124.** Cancer Genome Atlas Research N. Integrated genomic analyses of ovarian carcinoma. Nature. 2011; 474(7353):609–15. Epub 2011/07/02. doi: 10.1038/nature10166 PMID: 21720365; PubMed Central PMCID: PMC3163504.

**125.** Cancer Genome Atlas Research N. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008; 455(7216):1061–8. Epub 2008/09/06. doi: 10.1038/nature07385 PMID: 18772890; PubMed Central PMCID: PMC2671642.

**126.** Bettegowda C, Agrawal N, Jiao Y, Wang Y, Wood LD, Rodriguez FJ, et al. Exomic sequencing of four rare central nervous system tumor types. Oncotarget. 2013; 4(4):572–83. Epub 2013/04/18. PMID: 23592488; PubMed Central PMCID: PMC3720605.

**127.** Berger MF, Lawrence MS, Demichelis F, Drier Y, Cibulskis K, Sivachenko AY, et al. The genomic complexity of primary human prostate cancer. Nature. 2011; 470(7333):214–20. Epub 2011/02/11. doi: 10.1038/nature09744 PMID: 21307934; PubMed Central PMCID: PMC3075885.

**128.** Bass AJ, Lawrence MS, Brace LE, Ramos AH, Drier Y, Cibulskis K, et al. Genomic sequencing of colorectal adenocarcinomas identifies a recurrent VTI1A-TCF7L2 fusion. Nat Genet. 2011; 43(10):964–8. Epub 2011/09/06. doi: 10.1038/ng.936 PMID: 21892161; PubMed Central PMCID: PMC3802528.

**129.** Barbieri CE, Baca SC, Lawrence MS, Demichelis F, Blattner M, Theurillat JP, et al. Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. Nat Genet. 2012; 44(6):685–9. Epub 2012/05/23. doi: 10.1038/ng.2279 PMID: 22610119; PubMed Central PMCID: PMC3673022.

**130.** Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, et al. Signatures of mutational processes in human cancer. Nature. 2013; 500(7463):415–21. Epub 2013/08/16. doi: 10.1038/nature12477 PMID: 23945592; PubMed Central PMCID: PMC3776390.

**131.** Agrawal N, Jiao Y, Bettegowda C, Hutfless SM, Wang Y, David S, et al. Comparative genomic analysis of esophageal adenocarcinoma and squamous cell carcinoma. Cancer Discov. 2012; 2(10):899–905. Epub 2012/08/11. doi: 10.1158/2159-8290.CD-12-0189 PMID: 22877736; PubMed Central PMCID: PMC3473124.

**132.** Agrawal N, Frederick MJ, Pickering CR, Bettegowda C, Chang K, Li RJ, et al. Exome sequencing of head and neck squamous cell carcinoma reveals inactivating mutations in NOTCH1. Science. 2011;

333(6046):1154–7. Epub 2011/07/30. doi: 10.1126/science.1206923 PMID: 21798897; PubMed Central PMCID: PMC3162986.

133. Bamford S, Dawson E, Forbes S, Clements J, Pettett R, Dogan A, et al. The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. British journal of cancer. 2004; 91(2):355–8. Epub 2004/06/10. doi: 10.1038/sj.bjc.6601894 PMID: 15188009; PubMed Central PMCID: PMC2409828.

134. Forbes SA, Bhamra G, Bamford S, Dawson E, Kok C, Clements J, et al. The Catalogue of Somatic Mutations in Cancer (COSMIC). Current protocols in human genetics / editorial board, Jonathan L Haines [et al]. 2008;Chapter 10:Unit 10 1. Epub 2008/04/23. doi: 10.1002/0471142905.hg1011s57 PMID: 18428421; PubMed Central PMCID: PMC2705836.

135. Forbes SA, Bindal N, Bamford S, Cole C, Kok CY, Beare D, et al. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic acids research. 2011; 39(Database issue):D945–50. Epub 2010/10/19. doi: 10.1093/nar/gkq929 PMID: 20952405; PubMed Central PMCID: PMC3013785.

136. Chin L, Hahn WC, Getz G, Meyerson M. Making sense of cancer genomic data. Genes & development. 2011; 25(6):534–55. doi: 10.1101/gad.2017311 PMID: 21406553; PubMed Central PMCID: PMC3059829.

137. Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, et al. The Pfam protein families database. Nucleic Acids Res. 2010; 38(Database issue):D211–22. doi: 10.1093/nar/gkp985 PMID: 19920124; PubMed Central PMCID: PMC2808889.

138. Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, Lassmann T, et al. Pfam: clans, web tools and services. Nucleic acids research. 2006; 34(Database issue):D247–51. Epub 2005/12/31. doi: 10.1093/nar/gkj149 PMID: 16381856; PubMed Central PMCID: PMC1347511.

139. Eddy SR. Profile hidden Markov models. Bioinformatics. 1998; 14(9):755–63. Epub 1999/01/27. PMID: 9918945.

140. Gundem G, Perez-Llamas C, Jene-Sanz A, Kedzierska A, Islam A, Deu-Pons J, et al. IntOGen: integration and data mining of multidimensional oncogenomic data. Nature methods. 2010; 7(2):92–3. Epub 2010/01/30. doi: 10.1038/nmeth0210–92 PMID: 20111033.

141. Team R. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2011; ISBN: 3-900051-07-0.

142. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B (Methodological). 1995:289–300.

143. Berman H, Henrick K, Nakamura H. Announcing the worldwide Protein Data Bank. Nature structural biology. 2003; 10(12):980. Epub 2003/11/25. doi: 10.1038/nsb1203-980 PMID: 14634627.

144. Schrodinger, LLC. The PyMOL Molecular Graphics System, Version 1.3r1. 2010.

145. Krissinel E, Henrick K. Inference of macromolecular assemblies from crystalline state. J Mol Biol. 2007; 372(3):774–97. Epub 2007/08/08. doi: 10.1016/j.jmb.2007.05.022 PMID: 17681537.

146. Porter CT, Bartlett GJ, Thornton JM. The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. Nucleic acids research. 2004; 32(suppl 1):D129–D33.

147. Hornbeck PV, Chabra I, Kornhauser JM, Skrzypek E, Zhang B. PhosphoSite: A bioinformatics resource dedicated to physiological protein phosphorylation. Proteomics. 2004; 4(6):1551–61. PMID: 15174125