Molecular Systems Biology 9; Article number 637; doi:10.1038/msb.2012.68
Citation: ***Molecular Systems Biology*** **9:637**
www.molecularsystemsbiology.com

molecular systems biology

Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers

Jüri Reimand* and Gary D Bader*

The Donnelly Centre, University of Toronto, Toronto, Canada
* Corresponding authors. J Reimand or GD Bader, The Donnelly Centre, University of Toronto, 160 College Street, Toronto, Canada M5S 3E1. Tel.: +1 416 978 3935; Fax: +1 416 978 8287; E-mail: Juri.Reimand@utoronto.ca or E-mail: Gary.Bader@utoronto.ca

Received 4.5.12; accepted 6.12.12

Large-scale cancer genome sequencing has uncovered thousands of gene mutations, but distinguishing tumor driver genes from functionally neutral passenger mutations is a major challenge. We analyzed 800 cancer genomes of eight types to find single-nucleotide variants (SNVs) that precisely target phosphorylation machinery, important in cancer development and drug targeting. Assuming that cancer-related biological systems involve unexpectedly frequent mutations, we used novel algorithms to identify genes with significant phosphorylation-associated SNVs (pSNVs), phospho-mutated pathways, kinase networks, drug targets, and clinically correlated signaling modules. We highlight increased survival of patients with *TP53* pSNVs, hierarchically organized cancer kinase modules, a novel pSNV in *EGFR*, and an immune-related network of pSNVs that correlates with prolonged survival in ovarian cancer. Our findings include multiple actionable cancer gene candidates (*FLNB*, *GRM1*, *POU2F1*), protein complexes (HCF1, ASF1), and kinases (PRKCZ). This study demonstrates new ways of interpreting cancer genomes and presents new leads for cancer research.

Molecular Systems Biology **9**: 637; published online 22 January 2013; doi:10.1038/msb.2012.68
Subject Categories: molecular biology of disease; signal transduction
Keywords: cancer drivers; phosphorylation; somatic mutations

Introduction

A major goal of cancer research is to characterize somatic molecular alterations in tumor cells and identify systems driving cancer progression. These aberrations range from small DNA mutations to genomic copy number alterations, changes in chromatin structure and gene expression. To map such changes, international research consortia are collecting thousands of genome-wide molecular profiles of dozens of cancer types (Collins and Barker, 2007; The International Cancer Genome Consortium, 2010). Tumor genome sequencing has revealed a complex landscape of somatic DNA mutations in cancers of multiple types and tissues, including breast, colon, lung, liver, brain, ovary, pancreas, and blood (Wood *et al*, 2007; Cancer Genome Atlas Research Network, 2008, 2011; Ding *et al*, 2008; Jones *et al*, 2008; Parsons *et al*, 2008; Puente *et al*, 2011; Totoki *et al*, 2011). Most identified somatic mutations are thought to be functionally neutral 'passengers', caused by the increased mutation rate in cancer cells, whereas relatively few driver mutations provide selective advantages to tumor cells and are responsible for tumor initiation, maintenance, progression, and metastasis. Discovery of cancer drivers will provide insight into the biology of tumor development, and reveal diagnostic or predictive markers and new avenues of therapy development.

While a number of drivers are well recognized (e.g., *TP53*, *KRAS*, and *EGFR*), due to their frequent mutation rate in tumors and biological characterization, they are not sufficient to explain the phenotypic diversity of cancer and many more drivers likely exist. Of particular interest are driver mutations that occur in multiple cancer types; for instance, the druggable BRAF V600E mutation in melanoma and hairy cell leukemia (Chapman *et al*, 2011; Tiacci *et al*, 2011). Targeted drug development for such mutations may lead to effective multi-cancer therapies. However, most cancer mutations are not highly recurrent, and the observed long tail of infrequently mutated genes indicates the heterogeneous and complex nature of the disease. One likely explanation is that a cancer-related cellular system can be modified in multiple ways to create a neoplastic advantage, and therefore different, rare mutations can have similar phenotypic effects. Comprehensive systems-oriented analysis of integrated cancer data sets may therefore reveal novel, therapeutically relevant cancer driver genes, protein complexes, and pathways.

Tumor development involves a number of pathways with specific protein interactions and post-translational amino-acid modifications (Hanahan and Weinberg, 2011). In particular, protein phosphorylation is central in many hallmark

cancer processes and is often misregulated in the disease. Phosphorylation is a dynamic, reversible post-translational modification involving three major activities. Protein kinases act as writers by adding phosphate groups to serine (S), threonine (T), and tyrosine (Y) residues in substrate proteins, and phosphatases are erasers involved in dephosphorylation. Proteins with phosphorylated residue binding domains (e.g., SH2, PTB, 14-3-3) are readers that mediate context-specific protein interactions (Lim and Pawson, 2010). Phosphorylation can also regulate proteins by causing a conformation change or interfering with interactions with other molecules. Products of known cancer genes are enriched in kinases such as EGFR and SRC, while others such as TP53 and CTNNB1 are regulated by phosphorylation (Morin *et al*, 1997; Chao *et al*, 2006). Phosphorylation is a prime target of drug development, in particular via protein kinases, and a growing class of phosphorylation-related targeted agents such as EGFR inhibitors are now in clinical use (Hynes and Lane, 2005). Many focused studies and large-scale mass spectrometry experiments have mapped thousands of phosphorylation sites in human proteins (Dephoure *et al*, 2008; Nagano *et al*, 2009; Van Hoof *et al*, 2009; Olsen *et al*, 2010). This information helps establish a link between phosphorylation signaling and genomic alterations that can be exploited to identify mutations that affect signaling systems in cancer.

We performed a comprehensive analysis of somatic cancer mutations affecting protein phosphorylation. We hypothesize that statistically unexpected mutation rates in phosphorylation-specific regions within genes and pathways identify these as cancer drivers. We systematically analyzed 800 cancer genomes and thousands of somatic mutations in the context of phosphorylation sites, protein kinase domains, pathways, protein complexes, and kinase–substrate networks. We applied a novel, sensitive statistical model to find genes with unexpected phosphosite and kinase mutations, and carried out pathway and network analyses to identify significantly altered phosphorylation systems. We also detected network modules that significantly correlate with clinical outcome that may be helpful for diagnosis or prognosis.

Results

Global analysis reveals enrichment of cancer mutations in phosphosites and signaling networks

To investigate alterations of phosphorylation signaling in cancer, we integrated multiple data sets of somatic cancer mutations, protein phosphorylation events, kinase–substrate interactions and clinical data (Figure 1A–C). We accumulated 73 872 experimentally determined phosphorylation events in 10 279 phosphoproteins from three public databases (Keshava Prasad *et al*, 2009; Dinkel *et al*, 2011; Hornbeck *et al*, 2012). We collected 10 900 missense single-nucleotide variants (SNVs) from 793 samples of eight cancer types: serous ovarian adenocarcinoma (Cancer Genome Atlas Research Network, 2011), breast and colorectal cancer (Wood *et al*, 2007), pancreatic cancer (Jones *et al*, 2008), lung adenocarcinoma (Ding *et al*, 2008), glioblastoma multiforme (Cancer Genome Atlas Research Network, 2008; Parsons *et al*, 2008), chronic lymphocytic leukemia (Puente *et al*, 2011), and hepatocellular carcinoma (Totoki *et al*, 2011) (Figure 1D). To focus on phosphorylation-associated mutations, we only considered SNVs that directly altered kinase domains or 15-residue 'phosphosites' comprising a central phosphorylated S, T or Y residue and two 7-residue flanking sequences. Overlapping phosphosites were merged into continuous regions, which covered ~7% of the entire non-redundant human proteome. This procedure highlighted 775 distinct genes containing 949 phosphorylation-specific SNVs (pSNVs) in one or more cancer types (Figure 1E), covering 58% (459) of all studied cancer samples. This set includes 650 genes with phosphosite pSNVs, 189 genes with kinase domain pSNVs and 63 genes with both types of pSNVs (Supplementary Tables 1 and 2). We also compiled a kinase-substrate network of 4823 interactions, 379 kinases, and 1879 substrates, involving about 11% (7778) of phosphosites (Supplementary Table 3).

Several global trends confirm the functional significance of phosphosite and kinase pSNVs in cancer. First, phosphosite variants are more frequent than expected, given the genome-wide somatic mutation rate of studied cancer samples (949 observed, 751 ± 27 expected, s.d., $P = 2.5 \times 10^{-12}$, binomial test, Figure 1F). Second, kinase-targeted phosphoproteins are enriched in mutations: 44% (284) of genes with phosphosite-specific mutations are substrates of known kinases, while only 18% (118 ± 10) are expected ($P = 1.0 \times 10^{-54}$, Fisher's exact test). Consequently, 57% (2744) of kinase–substrate interactions are affected by pSNVs (Supplementary Figure 1). Third, pSNVs tend to affect topologically central and highly interacting genes in the kinase–substrate network ($P = 6.5 \times 10^{-25}$ and $P = 7.0 \times 10^{-32}$, Wilcoxon test, Supplementary Figure 2). While this could be due to ascertainment bias in network determination, centrality in our network likely indicates importance in cancer, as known cancer genes also show increased centrality and betweenness ($P = 2.3 \times 10^{-13}$ and $P = 7.3 \times 10^{-6}$, respectively). These results demonstrate the general extent of somatically altered phosphorylation signaling in tumor cells.

Focusing further on phosphosites, we observe 78 direct phosphosite mutations that replace the central S/T/Y residue with a non-phosphorylatable residue (Supplementary Table 4). We also find 37 potentially phosphomimetic mutations introducing negatively charged aspartic and glutamic acid residues that physicochemically resemble residues with constitutive phosphorylation (Supplementary Table 5). Both lists are enriched in known cancer genes, suggesting that pSNVs disable tumor-suppressing phosphorylation and activate signaling that promotes tumorigenesis ($n = 13$, $P = 2.4 \times 10^{-8}$ and $P = 3.3 \times 10^{-12}$, respectively, Fisher's exact test). However, most pSNVs occur in phosphosite-flanking regions (Figure 1F). These may enhance or disrupt sequence recognition and binding affinity of associated reader, writer or eraser proteins, and modify signaling systems while still maintaining the phosphorylation switch. Finally, genes with phosphosite pSNVs are enriched in known drug targets from the DrugBank database ($n = 109$, $P = 2.5 \times 10^{-11}$; Knox *et al* (2011)). Druggable genes with pSNVs cover nearly two-thirds (284) of phospho-mutated cancer samples and 36% of all cancer samples, indicating the potential for therapy development targeting affected genes.

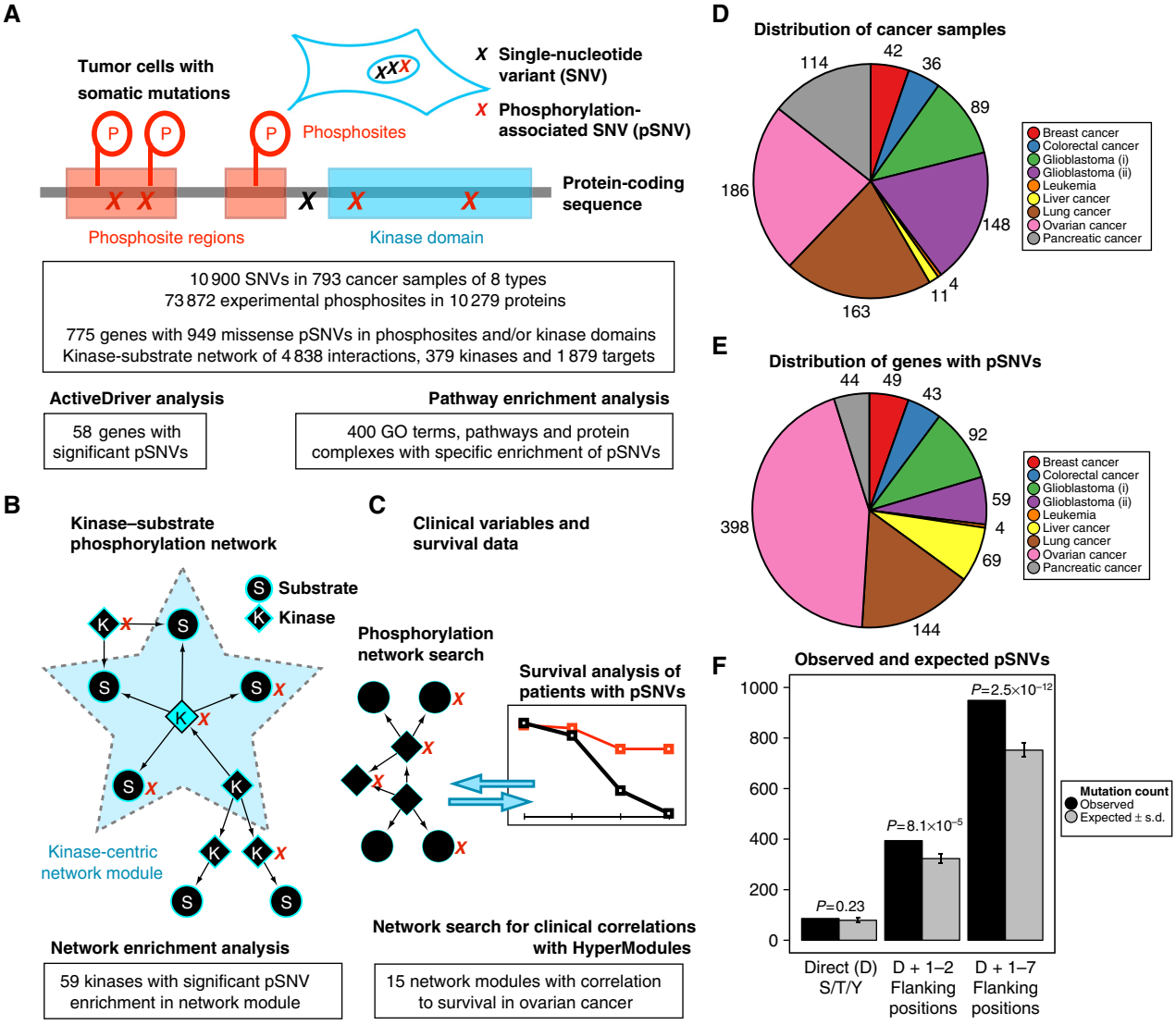

Figure 1 Analysis overview. (**A**) Missense SNVs (crosses) were extracted from cancer genomes and classified as phosphorylation-associated (pSNVs; red crosses) if they affected phosphosites (red P-circles) and their flanking regions (pink rectangles) or kinase domains (blue rectangles). We designed the statistical model ActiveDriver to find cancer genes with significantly enriched or depleted pSNVs. Using pathway enrichment analysis, we identified GO terms, pathways and protein complexes with over-represented pSNVs. (**B**) Phosphorylation network composed of experimentally determined kinase–substrate interactions. To find kinases important in cancer, all kinase-centric signaling modules (light blue star) were tested for statistical enrichment of pSNVs. Each such module comprised a fixed central kinase (blue diamond) and its direct upstream kinases and downstream substrates (black diamonds and circles within the light blue star). (**C**) To find clinically relevant signaling modules, we designed a novel local network search algorithm HyperModules that combines pSNVs, kinase–substrate interactions, and patient survival. (**D**) Distribution of cancer samples across cancer types. Two glioblastoma data sets are shown separately in green (Parsons *et al*, 2008) and purple (Cancer Genome Atlas Research Network, 2008). (**E**) Distribution of genes with pSNVs across cancer types. (**F**) Phosphosites are enriched in somatic cancer mutations in comparison to genome-wide mutation rate averaged across cancer genomes (binomial test, error bars show s.d.).

ActiveDriver—a novel gene-centric method to identify significantly mutated protein sites

We developed a novel statistical method, named ActiveDriver, to identify genes with significant pSNVs (see Materials and methods). This gene-centric method is based on generalized linear regression and tests the following pair of hypotheses for a given gene and its phosphosite region. The null (expected) model states that the phosphosite region follows the same mutation rate as the gene sequence given its structured and disordered regions. The alternative model states that the phosphosite-specific mutation rate is higher or lower than the gene-wide mutation rate. The two models are compared with a likelihood ratio test that refutes the null hypothesis if a distinct mutation rate is required to explain pSNVs observed in the phosphosite region. We use a model-based approach rather than a direct statistical test, as short phosphosites tend to involve low mutation counts and numerous considered sites would reduce statistical power due to multiple testing. To establish the significance of a gene in a cancer type, we multiply *P*-values of all its significantly mutated phosphosites ($P \leqslant 0.05$). A gene is deemed significant if at least one of its

phosphosites displays unexpected mutation rates, while its non-significant sites do not contribute to the final composite *P*-value. Finally, we correct the composite *P*-value for multiple testing with the Benjamini–Hochberg false-discovery rate (FDR) and select genes whose *P*-values exceed the significance of a certain threshold (FDR $P \leqslant 0.05$).

ActiveDriver considers information on pSNV position within a phosphosite, protein structured and unstructured regions and cancer type specific mutation rate. First, the functional impact of each pSNV is determined by the number of adjacent phosphosites and their position relative to the mutation. Direct mutations of S/T/Y are likely to have stronger functional impact than immediate and proximal flanking mutations, and single pSNVs can affect multiple clustered phosphosites. We encode this information in three features of the alternative regression model: for every position *s* in the phosphosite region, we count (i) the number of phosphosites at *s* (zero or one), and the number of phosphosites (ii) within 1–2 residues from *s*, and (iii) within 3–7 residues from *s*. Second, post-translational modifications are known to occur in unstructured (disordered) regions of proteins, and such regions are also believed to evolve more rapidly than structured regions. Therefore, we consider separate mutation rates for disordered and ordered protein regions, and encode this information as confounding factors in the null and alternative regression models. Third, cancer genomes of different types are biologically distinct and involve varying sample sizes (e.g., 4–186 samples, leukemia versus ovarian cancer), baseline mutation rates (e.g., 0.74–11 missense mutations per million amino acids, pancreatic versus lung cancer), and different mutation calling protocols. Cancer-specific mutation rates are therefore compared in independent models, and corrected separately for multiple testing. Such data integration provides greater sensitivity than conventional statistical tests, as we show below.

ActiveDriver identifies known cancer genes with significantly mutated phosphosites

ActiveDriver analysis of pSNVs for nine separate cancer data sets resulted in 44 genes with significantly unexpected numbers of phosphosite mutations (FDR $P \leqslant 0.05$, Figure 2A). We repeated the analysis with a merged collection of SNVs from all cancer types and found 14 additional phospho-mutated genes.

The results are enriched in known cancer genes from the Cancer Gene Census and earlier review papers ($n = 15$, $P = 1.1 \times 10^{-11}$, Fisher's exact test), genes encoding kinase proteins (e.g., *KDR*, *EGFR*, *ABL1*, *FLT4*; $n = 9$, $P = 3.8 \times 10^{-5}$), and transcription factors (e.g., *TP53*, *FOXO3*, *MLL*, *MET*; $n = 15$, $P = 3.0 \times 10^{-4}$). The gene list also includes 13 drug targets ($P = 1.3 \times 10^{-3}$), nine of which are known cancer genes (Knox *et al*, 2011). Ranked pathway enrichment analysis of ActiveDriver results using g:Profiler (Reimand *et al*, 2011) highlights relevant pathways and Gene Ontology (GO) categories, including cell motility, regulation of cell adhesion, regulation of epithelial cell proliferation, and blood vessel development (all FDR $P \leqslant 0.05$, Supplementary Table 6). Thus, ActiveDriver-identified genes are enriched in cancer hallmark processes.

Our predictions include 15 genes with well-established roles in cancer biology. For instance, β-catenin (*CTNNB1*) encodes a transcriptional co-regulator and downstream target of the Wnt pathway involved in organism development (FDR $P = 9.8 \times 10^{-4}$ from ActiveDriver). In the absence of Wnt signaling, CTNNB1 is phosphorylated by GSK3 kinase and continuously degraded. Wnt pathway activation blocks the phosphorylation of CTNNB1, leading to its accumulation in the nucleus and transcriptional activity (van Noort *et al*, 2002). Aberrant activity of CTNNB1 due to altered N-terminal target phosphosites of GSK3 has been observed in colorectal cancer (Morin *et al*, 1997). In our data, all five SNVs in *CTNNB1* are specifically linked to phosphorylation. ActiveDriver highlights the N-terminal region of CTNNB1 that includes phosphosites S33, S37, T41 targeted by multiple kinases (CDK6, CHUK, CSNK1/2A/2B, GSK3A/3B, MAPK8, PRKACA; Figure 2B). These sites are affected by four pSNVs (two in lung and two in liver cancer), including three phosphorylation-disabling mutations (T41A, S37F, S37F) and one potentially phosphomimetic mutation G34E. An additional pSNV in ovarian cancer (G555A) affects three phosphosites that are known targets of AKT1, AKT2 and PRKACA kinases. Consistent with previous observations (Morin *et al*, 1997), we propose that lung and liver cancer pSNVs in *CTNNB1* disrupt its degradation and enable downstream transcriptional programs to benefit cancer progression. In this example, a small number of very specific pSNVs appear in multiple types of cancer and have a statistically significant pattern with a biologically meaningful interpretation. Our strategy is therefore useful for interpreting rare mutations.

About one-third (21) of detected genes have phosphosite-specific mutations in multiple cancer types (Figure 2A), potentially highlighting general cancer driver mechanisms. As integration approaches are less explored in cancer genomics studies, most of our additional findings from the merged data set analysis represent novel candidate cancer genes. Further, IRS1 (insulin receptor substrate 1) and GRM1 (glutamate receptor, metabotropic 1) are known drug targets (Knox *et al*, 2011) and therefore may be directly actionable for therapy development. *KRAS* is the only well-recognized cancer gene in the merged analysis and is listed due to a less-than-expected number of pSNVs (FDR $P = 4.6 \times 10^{-3}$ from ActiveDriver), indicating the role of phosphorylation signaling in its oncogenic activities. These results illustrate the utility of data integration from multiple cancer genomics projects.

ActiveDriver highlights phospho-mutated candidate cancer genes *FLNB*, *GRM1*, *POU2F1*

Our findings also include genes whose cancer-specific roles are less recognized. The highest-ranking candidate cancer gene is filamin-B (*FLNB*, FDR $P = 5.4 \times 10^{-7}$ from ActiveDriver) that has been highlighted in the breast cancer genome project due to frequent mutations (Sjoblom *et al*, 2006). All four SNVs in breast cancer (A1565G, T703K, N663K, R566Q) alter phosphosite-flanking regions, although evidence for phosphorylation events comes from large-scale screens and no targeting kinases are currently known. FLNB is an intracellular signaling protein involved in organization of actin cytoskeleton as well

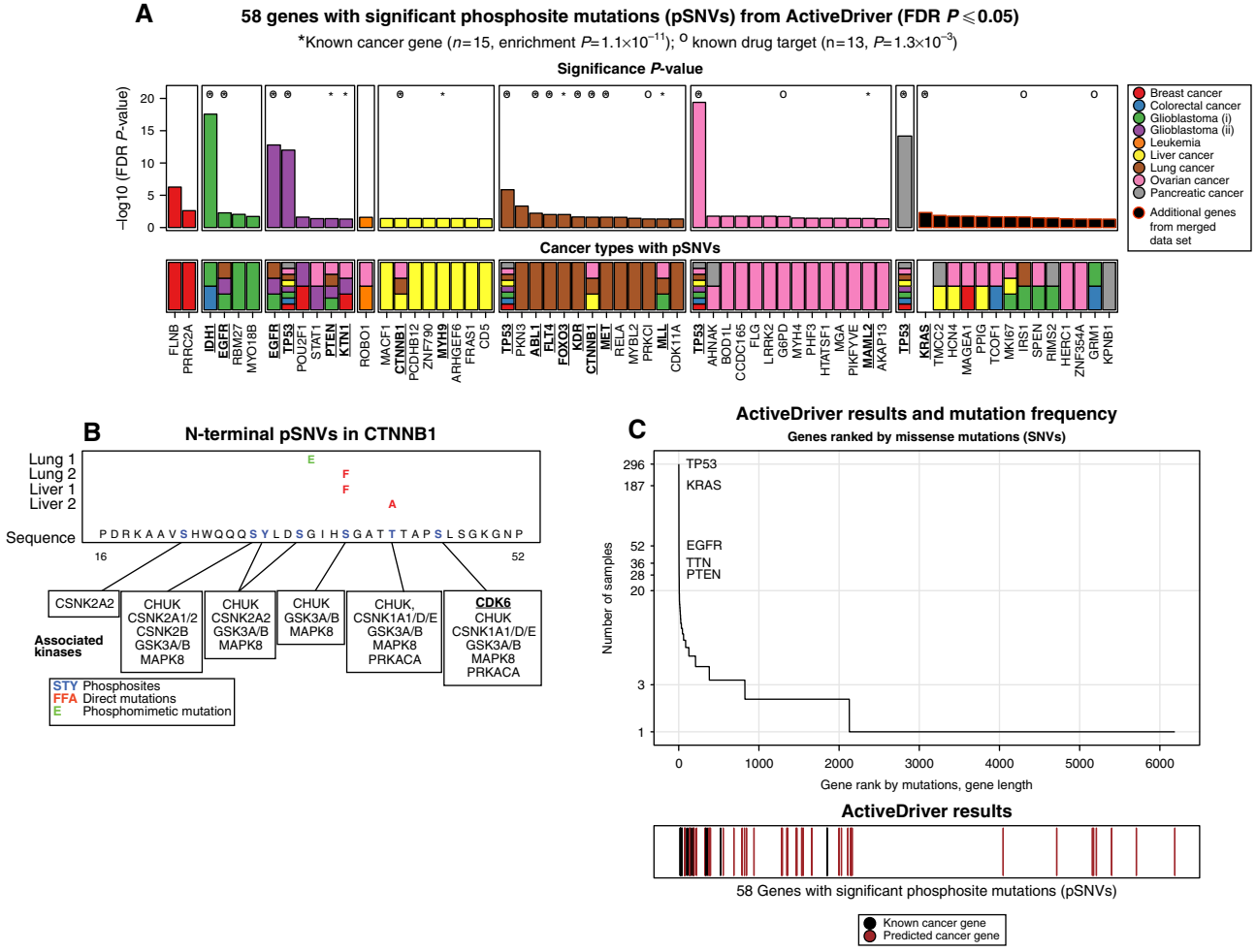

Figure 2 Genes with significant phosphosite mutations (pSNVs). (**A**) ActiveDriver analysis revealed 58 genes with significant mutation rates in phosphosite regions. Top barplot shows gene significance (log10 *P*-value) by cancer type, and bottom color-strip shows cancer types with related pSNVs. Rightmost panel represents 14 additional genes found in a composite analysis of somatic mutations of all cancer types. No pSNVs are known for KRAS, it is listed due to significant depletion of phosphosite mutations. Known cancer genes (*) and drug targets (o) are enriched in the list of discovered genes. (**B**) N-terminal protein sequence region 16–52 of CTNNB1 includes seven phosphosites (blue letters) and is significantly enriched in pSNVs. Four out of five SNVs in *CTTNB1* are found in the region, including three direct mutations (red letters) and one phosphomimetic mutation (green letter). Seven phosphosites (blue letters) are known targets of several kinases, shown in boxes below the sequence. Names of known cancer genes are underlined and printed in bold. (**C**) Comparison of ActiveDriver results with standard mutation frequency-based gene ranking. Top plot shows 6182 genes with at least one missense SNV, ranked in decreasing order by number of missense point mutations across all cancer samples, followed by increasing order by gene length. Bottom plot shows the position of 58 ActiveDriver-predicted genes with significant pSNVs in the global SNV-based ranking. Black lines represent known cancer genes and red lines represent candidate cancer genes.

as skeletal and neuronal development (Lu *et al*, 2007). In particular, knockdown of *FLNB* has been shown to inhibit VEGF-induced angiogenesis (Del Valle-Perez *et al*, 2010), a hallmark of tumor cells. Further, knockdown of filamin genes has been shown to reduce cell migration in cancer cell lines (Baldassarre *et al*, 2009), and germline variants observed in human developmental disorders have been linked to gain-of-function phenotypes manifested in increased F-actin binding (Sawyer *et al*, 2009). We therefore speculate that the observed pSNVs may also act as gain-of-function mutations that enhance angiogenesis, invasion or metastasis of tumor cells.

The G-protein-coupled receptor *GRM1* (glutamate receptor metabotropic 1) is identified in the merged data set analysis of pSNVs due to two flanking pSNVs in glioblastoma (R684C) and colorectal cancer (R696W) (FDR $P = 0.047$ from ActiveDriver). The latter mutation directly flanks a phosphorylation site of protein kinase C α (PRKCA) at T695 that is involved in receptor desensitization in a feedback loop (Francesconi and Duvoisin, 2000). We propose that the pSNV disrupts inhibition of the receptor activity, leading to aberrant activation of tumorigenic processes such as growth, survival, and proliferation via the downstream phosphoinositide 3-kinase (PI3K) pathway. Overexpression of *GRM1* was shown to induce melanoma in mouse models (Pollock *et al*, 2003), and *GRM1* mutations were linked to melanoma in a human genetic association study (Ortiz *et al*, 2007). The family of metabotropic glutamate receptors is also generally enriched in SNVs across all cancer samples ($n = 16$ samples, $P = 3.2 \times 10^{-3}$, Poisson exact test), suggesting the importance of GPCR signaling in tumor biology. *GRM1* is potentially actionable,

as it is the target of several drugs according to Drugbank (e.g., acamprosate, 4-(1-amino-1-carboxy-ethyl)-benzoic acid; Knox *et al*, 2011).

A final example is *POU2F1* (FDR $P = 0.024$ from ActiveDriver), a POU domain transcription factor and a cell-cycle regulator that undergoes phosphorylation-mediated inhibition of DNA-binding activity during mitosis (Segil *et al*, 1991). *POU2F1* is highlighted in our model for two pSNVs in glioblastoma (R296Q) and in breast cancer (S111F). The latter mutation directly modifies a phosphosite targeted by PRKDC kinase, involved in promoting cell survival in response to DNA damage (Schild-Poulter *et al*, 2007).

The detection of many known cancer genes validates our approach and the additional highlighted genes, some known to be druggable, provide novel hypotheses for detailed, functional experiments.

ActiveDriver is complementary to frequency-based methods of global mutation significance

Conventional cancer genomics analysis focuses on frequently mutated cancer genes instead of the long tail of genes with rare mutations. In contrast, our 'gene-centric' model considers each individual gene and detects signaling sites whose mutations are unexpected given the gene-wide mutation rate. We therefore find many genes that harbor infrequent, albeit highly specific mutations that are missed when considering mutation frequency alone (Figure 2C). In particular, our results include nine known cancer genes not found among the top 58 genes ranked by mutation frequency (median rank 319, Supplementary Figure 3).

State of the art global mutation assessment methods such as MutSig compute significance of gene mutations by comparing these with baseline mutation rates estimated from whole genomes or exomes (Banerji *et al*, 2012). To compare ActiveDriver with this general approach, we implemented a simple global strategy similar to MutSig using a standard binomial statistical test. This global strategy highlighted only the four most frequently mutated genes as statistically significant (*TP53*, *KRAS*, *PTEN*, *EGFR*; FDR $P \leqslant 0.05$, Supplementary Figure 4). ActiveDriver identified many more, and also found seven genes with significant pSNVs that have less mutations than expected from the genome average of the corresponding cancer type (*PRKCI*, *PHF3*, *KTN1*, *MLL*, *AKAP13*, *MET*, *CDK11A*). Thus, genes with specific and significant phosphosite mutations would remain unseen in a global analyses.

We further re-implemented ActiveDriver using binomial statistics in place of our disorder-corrected regression model, with all other factors unchanged, to test the effectiveness of our model versus standard methods. This simplified strategy only found five highly mutated genes as statistically significant (*EGFR*, *TP53*, *IDH1*, *KRAS*, *FLNB*; FDR $P \leqslant 0.05$), the first four of which are cancer genes ($P = 2.0 \times 10^{-6}$, Fisher's exact test). As all these were also found with ActiveDriver, we conclude that modeling protein disorder and phosphosite position is important for estimating cancer gene significance. Our method is therefore more sensitive than standard approaches, as it highlights 11 additional cancer genes and many novel candidates with highly specific phosphosite mutations.

Phosphosite mutations of TP53 correlate with extended survival in ovarian cancer and glioblastoma

The tumor suppressor transcription factor *TP53* is the top-ranking gene in our pSNV analysis (167 pSNVs, FDR $P = 7.6 \times 10^{-86}$ from ActiveDriver). It is an active phosphoprotein and a substrate of 43 kinases with 29 phosphosites in nine regions. ActiveDriver revealed a statistically significant mosaic of phosphosites and pSNVs (Figure 3A), including four hotspot regions of phosphosites that are enriched in pSNVs in eight cancer types.

To explore the functional and clinical significance of pSNVs, we studied the survival rates of corresponding glioblastoma and ovarian cancer patients. We found that phosphosite-associated *TP53* mutations significantly correlate to increased survival among ovarian cancer patients (log-rank test $P = 7.2 \times 10^{-3}$, Cox regression $P = 0.025$, Figure 3B). The survival correlation is evident even in comparison to patients with wild-type TP53, suggesting that such pSNVs might be beneficial for tumor suppression. A similar correlation between pSNVs and better prognosis is seen among glioblastoma patients with long-term survival (>1 year), although its statistical significance is low due to small sample sizes (log-rank test $P = 0.066$, Cox regression $P = 0.17$, Figure 3C). In agreement with these data, phosphorylation of T155, S183, S269, T284 by casein and aurora kinases has been associated with TP53 inhibition via post-translational degradation and transcriptional repression (Bech-Otschir *et al*, 2001; Wu *et al*, 2011). The highlighted pSNVs in 118 patients potentially inhibit phosphorylation of these sites, meaning that TP53 will no longer be degraded if mutated. Taken together, our data suggest a double-negative mechanism in which phosphorylation-mediated inhibition of TP53 is potentially disrupted by pSNVs, leading to reduced inhibition of TP53 tumor suppressor function and increased survival.

In contrast to the above mutation hotspots, phosphosites in TP53 termini appear as highly significant mutation deserts (three pSNVs observed, 67 ± 8 expected, $P = 2.8 \times 10^{-30}$ from ActiveDriver). The observed negative selection indicates the importance of these regions as tumorigenic signaling interfaces. The N-terminus of TP53 is the interaction interface of its primary inhibitor MDM2, and TP53 phosphorylation of S15 and S20 inhibits MDM2 binding and leads to stabilization and activation of TP53 (Chehab *et al*, 1999). The absence of mutations in the region suggests that the sequence is required for successful docking of MDM2 and inhibition of apoptosis. This example demonstrates the utility of ActiveDriver in interpreting somatic mutations in known cancer genes.

Pathway analysis of phosphomutated genes reveals cancer hallmarks and predicts novel driver systems

Next, we performed a pathway analysis to find systems of functionally related genes with frequent pSNVs. We assumed that frequent recurrence of cancer mutations in the same biological system is unlikely unless the system is involved in cancer. We searched for pSNV-enriched systems using GO categories (Ashburner *et al*, 2000), Reactome pathways

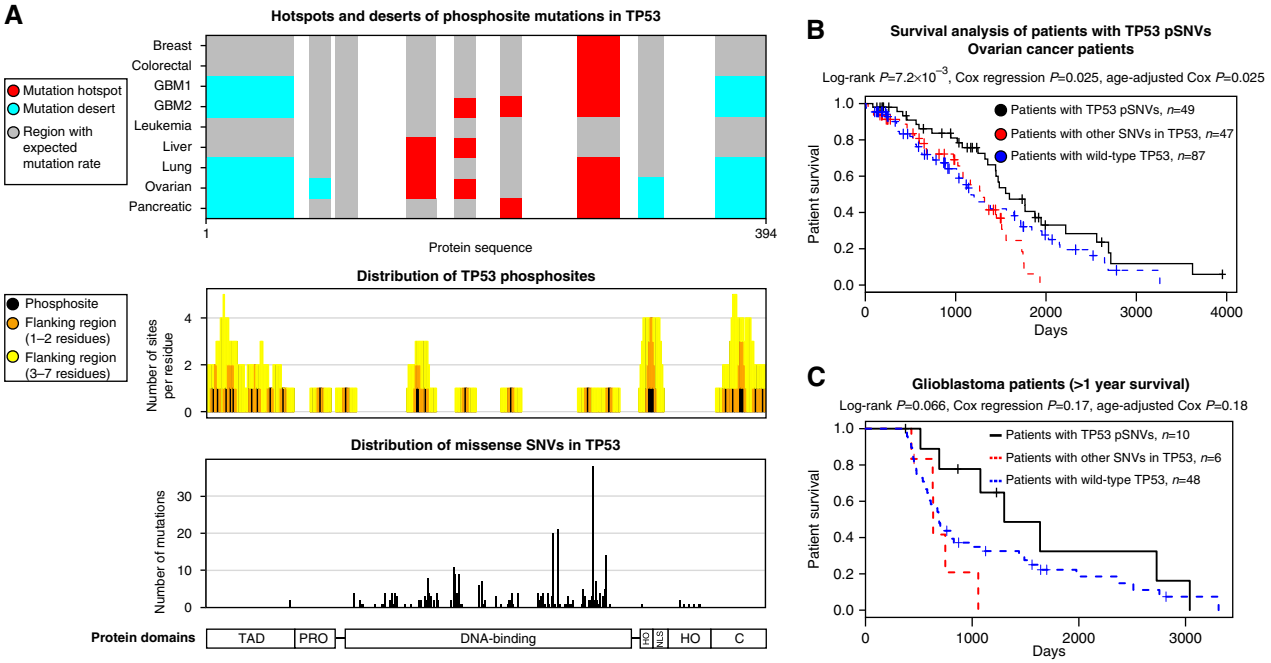

Figure 3 Phosphosite mutations in TP53 correlate with increased patient survival. (**A**) ActiveDriver analysis of *TP53* pSNVs identifies a mosaic of nine phosphosite mutation hotspots (red columns) and deserts (blue columns) across multiple cancer types (top panel). Middle panel shows the protein sequence of TP53 with 29 phosphosites (black bars) and number of flanking phosphosites per residue (yellow and orange). Bottom panel shows number of SNVs per position, with the majority of mutations accumulating to the DNA-binding domain in the central region of the sequence. Protein domains of TP53 are shown below the chart (TAD, transcriptional activation; PRO, proline-rich region; NLS, nuclear localization signal; HO, homo-oligomerization; C, C-terminus). (**B**) Kaplan–Meier analysis of clinical data of ovarian cancer patients shows that TP53 pSNVs correlate with increased survival. Survival of patients with pSNVs (black solid line) is compared to survival of patients with other, non-phosphorylation-associated SNVs (red dashed line) as well as patients with wild-type pSNVs (blue dashed line). (**C**) Long-term survivors of glioblastoma with TP53 pSNVs (black solid line) show a similar correlation with increased survival; however, these observations have borderline statistical significance due to small sample size.

(Matthews *et al*, 2009), and human protein complexes from the CORUM database (Ruepp *et al*, 2010). To reduce bias from extremely highly mutated genes, we excluded *TP53* and *EGFR*, which together cover 39% (181) of the 460 phosphomutated tumor samples.

Our initial pathway analysis revealed nearly 2000 GO terms and pathways with statistically significant enrichment in mutated cancer samples (FDR $P \leqslant 0.05$), many of which match known hallmarks of cancer (Hanahan and Weinberg, 2011) (Supplementary Figure 6). However, such broad themes are also expected in a pathway analysis of standard mutation frequency-ranked gene lists. To identify functional categories specifically affected by phosphorylation mutations, we repeated the pathway enrichment analysis considering all SNVs and subtracted the list of 1590 categories detected in both SNV and pSNV analyses (all FDR $P \leqslant 0.05$). This produced a final set of 400 phospho-specific enriched categories that are not found by analyzing all SNVs (Supplementary Table 7).

The top 50 pSNV-focused GO categories highlight multiple interesting functional themes (Figure 4A). The highest-ranking categories are innate immune response ($n = 52$ samples, FDR $P = 3.7 \times 10^{-4}$, Poisson exact test), cytokines, and specific pathways for Toll signaling ($n = 20$, FDR $P = 1.6 \times 10^{-4}$, Figure 4B) and IκB-NFκB cascade ($n = 36$, FDR $P = 4.9 \times 10^{-6}$), all of which highlight the importance of altered phosphorylation in immune signaling. Avoiding immune destruction and establishing tumor-promoting inflammation are emerging hallmarks of cancer and important therapeutic target systems (Hanahan and Weinberg, 2011). While such pathways are generally expected to be enriched in cancer mutations, the majority of specific pSNVs are infrequent and therefore likely less understood. pSNVs appear in multiple cancer types and affect different components of the same system, suggesting that similar strategies could be applied for drug development. Mutations in some pathways may already be actionable due to existing drugs against known or candidate cancer genes (Figure 4C). For instance, 23 patients associated with the innate immune response category carry mutations in 17 druggable genes such as *HCK*, *SRPK2*, *MAPK9*, *UBC* and *HLA-A*. Thus, pathway analysis of pSNVs may be useful in developing personalized drug treatments based on known drugs.

Our analysis also reveals 21 non-overlapping protein complexes and 17 Reactome pathways with frequent pSNVs (FDR $P \leqslant 0.05$, Figure 4C and D). The identified pathways are often related to cancer hallmarks such as cell-cycle regulation, but not always. For example, the HCF1 complex involves six phospho-mutated genes in ovarian, liver, and brain cancer patients: two transcriptional regulators (*SP1*, *HCFC1*), a histone methyltransferase (*ASH2L*) and three druggable heat shock proteins (*HSP90AA1*, *HSP90AB1*, *HSPA8*). HCF1 selectively modulates chromatin structure and promotes cell proliferation by transcriptional activation and repression (Wysocka *et al*, 2003). Mutations in this complex may therefore lead to aberrant expression of cell-cycle genes and initiation or enhancement of malignant growth.

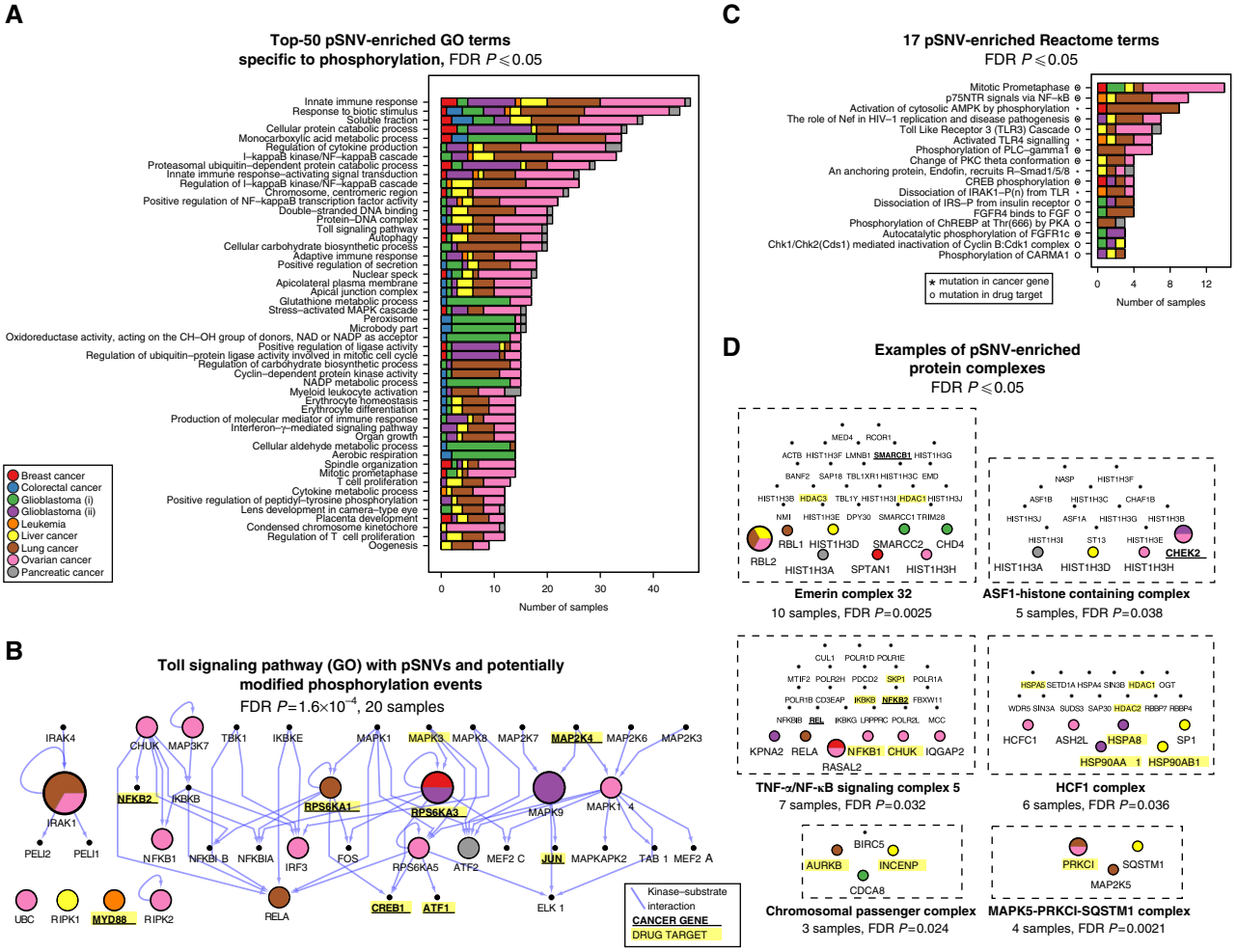

Figure 4 Functional enrichment analysis of pSNVs. (**A**) Top 50 non-redundant, phosphorylation-specific GO categories with statistically significant enrichment of pSNVs (FDR $P \leqslant 0.05$, *TP53* and *EGFR* excluded). Categories are ranked according to number of samples, shown on the *X*-axis. Colors denote different types of cancer. (**B**) Toll signaling pathway is enriched in pSNVs, with genes defined by GO and connections from the kinase–substrate network. Network shows genes with pSNVs (colored circles) and their non-mutated phosphorylation partners (small black circles) in the Toll pathway. Arrows denote kinase–substrate relationships, known cancer genes are labeled using bold and underline, and known drug targets are shown on yellow background. (**C**) The set of 17 non-redundant Reactome pathways with enriched pSNVs, ranked by number of affected samples (*X*-axis). (**D**) Six of 21 CORUM protein complexes with significant pSNV enrichment.

As another example, the ASF1 chaperone complex regulates chromatin assembly during DNA replication (Groth *et al*, 2005). The ASF1 complex is highlighted due to four component genes with five pSNVs: the tumor suppressor kinase *CHEK2* and three H3 histones with pSNVs in tail regions (*HIST1H3A*, *HIST1H3D*, *HIST1H3H*). Interestingly, H3 histone phosphorylation at S10 has dual roles in chromatin condensation and transcriptional activation during different cell-cycle phases (reviewed by Nowak and Corces, 2004), and we see two flanking pSNVs in this site (R9G, R3C). These mutations may also affect other modifications, in particular acetylation and methylation that act as regulatory switches of transcriptionally active chromatin structure. pSNVs in the ASF1 complex could therefore drive cancer using two mechanisms: by altering DNA replication to introduce further mutations in cancer genomes, and by modifying gene expression during cell cycle to increase proliferation.

In summary, functional analysis of pSNVs reveals known and putative cancer driver pathways and complexes, not apparent from standard global analysis of all somatic mutations. All identified complexes, functions and pathways are affected by mutations in multiple cancer types, indicating that our approach is useful for uncovering general mechanisms of tumor biology.

Network analysis of pSNVs reveals hierarchically organized cancer kinases

To identify additional signaling systems important in cancer, we analyzed pSNVs in the kinase–substrate network. The network contains information from many proteomics experiments and is therefore complementary to the known pathways and complexes analyzed above. Multi-kinase signaling systems are enriched in mutations, as 74% of 725 kinase–kinase phosphorylation interactions are altered in cancer ($P = 1.9 \times 10^{-25}$, Fisher's exact test). These are expected to include driver mutations, as alterations of inter-regulator signaling likely affect multiple downstream processes.

To study this further, we constructed kinase-centric network modules containing three types of proteins: a central kinase, its downstream substrates and upstream kinases. The modules for all interacting kinases were studied with pSNV enrichment tests to identify frequently phospho-mutated signaling systems. This analysis revealed 59 kinase-centric modules with surprisingly high pSNV frequency (FDR $P \leqslant 0.05$; Figure 5A; Supplementary Table 8). Many central kinases are associated with cancer, validating our analysis method ($n = 15$, $P = 1.4 \times 10^{-11}$, Fisher's exact test). Further, many kinases are also known drug targets, suggesting that they are potential avenues for therapy development ($n = 23$, $P = 2.7 \times 10^{-10}$).

We grouped the modules according to frequency of pSNV mutations among upstream and downstream interaction partners and revealed a hierarchy of five distinct kinase classes. 'Master kinases' are not highly phospho-mutated themselves, but have numerous pSNVs in downstream substrates (e.g., SRC, AKT1, ABL1). 'Middle manager kinases' include pSNVs in the gene of the central kinase as well as its upstream and downstream interaction partners (STK11, MET, ITK). 'Local kinases' mostly have pSNVs in genes of downstream substrates (CDK6, ERBB2), while 'managed kinases' are characterized by upstream pSNVs (FLT4, BCR). 'Self-employed' signaling systems of KDR and MAPK6 mostly involve mutations in genes of central kinases. Thirteen modules involve pSNVs in feedback loops where the mutated protein is both upstream and downstream of the central kinase.

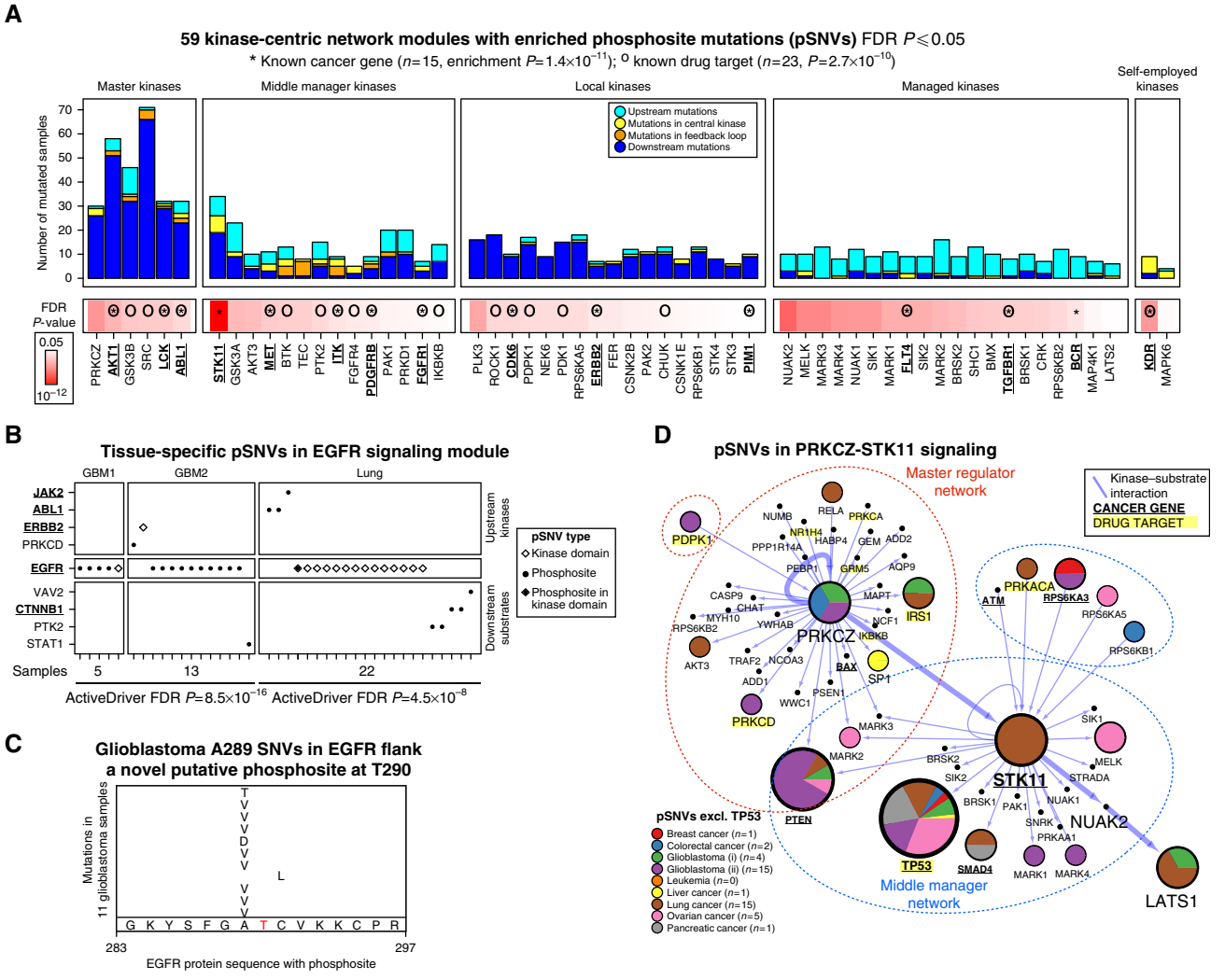

Figure 5 pSNVs in the kinase–substrate network. (**A**) The set of 59 kinase-centric signaling modules with significant pSNV enrichment (FDR $P \leqslant 0.05$), grouped according to their positions in the defined kinase hierarchy. Bars show number of samples with pSNVs in upstream kinases (light blue), downstream substrates (dark blue), and central kinase (yellow). Feedback loop mutations (orange) occur in proteins that are both upstream kinases and downstream substrates of the central kinase. Color-strip under the bars shows statistical significance of pSNV enrichment in each module, asterisks denote known cancer genes, and circles denote known drug targets. (**B**) Tissue-specific phosphosite mutations in EGFR. 14 lung cancer mutations occur in the kinase domain (diamonds), while 14 glioblastoma (GBM) pSNVs associate to phosphosites. (**C**) Eleven glioblastoma pSNVs in EGFR (A289, C291) flank a novel extracellular phosphosite at T290 (shown in red). (**D**) Signaling network of the master kinase PRKCZ and its downstream target STK11 is frequently phospho-mutated and involves several tumor suppressors such as PTEN, TP53, and LATS1. The network involves pSNVs in ~25% of cancer samples. Colored circles denote genes with pSNVs, small black circles denote genes with no pSNVs, and arrows denote kinase–substrate phosphorylation events. Names of known cancer genes are underlined and printed in bold, and drug targets are shown on yellow background.

Analysis of kinase-centric modules provides phosphorylation-associated interpretation of recurrent mutations in well-established cancer genes. For example, the majority (56% = 29) of EGF receptor (*EGFR*) missense point mutations associate with phosphorylation and appear in a tissue-specific, mutually exclusive pattern (Figure 5B). The kinase domain of EGFR is characterized by 15 lung cancer pSNVs (FDR $P = 4.9 \times 10^{-8}$ from ActiveDriver), including the well-studied L858R mutation that affects EGFR autophosphorylation and is used as a clinical marker for therapeutic outcome (Sharma *et al*, 2007). In contrast, 14 SNVs in glioblastoma are associated to EGFR phosphorylation sites, suggesting that the mutations play a role in post-translational activation of this oncogene. Eleven pSNVs alter the residue A289 that directly flanks a novel putative phosphosite T290 (Ruse *et al*, 2008) in the extracellular domain of the protein (FDR $P = 8.5 \times 10^{-16}$ from ActiveDriver, Figure 5C). While extracellular phosphorylation mechanisms are generally poorly understood, a recent study has characterized a novel family of secreted kinases with a role in human disease (Tagliabracci *et al*, 2012). Further study of the role of phosphorylation at this site may help explain the mechanism of highly recurrent glioblastoma mutations.

The master kinase PKC-zeta controls a frequently phospho-mutated tumor suppressor network

Protein kinase C zeta (PRKCZ) is the master kinase with the strongest enrichment of pSNVs among its interaction partners ($n = 30$ samples, FDR $P = 1.2 \times 10^{-6}$, Poisson exact test; Figure 5D). This kinase functions in the PI3K and MAPK pathways and is involved in multiple cellular functions, including cell cycle and proliferation, cell polarity, NFκB signaling and inflammation. While these functions relate to hallmark cancer pathways, the direct role of Protein Kinase C zeta in tumor biology is less established.

Our network analysis suggests that the kinase is involved in tumor-inhibiting signaling systems, as it directly phosphorylates tumor suppressors PTEN and STK11 and may indirectly affect signaling of TP53 and LATS1 via STK11. Altogether, the PRKCZ-STK11 network module involves 44 pSNVs in 19 genes and 7 cancer types. When considering the additional 157 pSNVs in *TP53*, every fourth tumor in our data set involves aberrant signaling in this system. Mutation-ranked functional enrichment analysis of genes in the network reveals categories such as cell cycle (FDR $P = 5.0 \times 10^{-5}$ from g:Profiler), cell differentiation (FDR $P = 1.0 \times 10^{-10}$), and protein phosphorylation (FDR $P = 3.8 \times 10^{-29}$). These functions are consistent with the proposed roles of PRKCZ. In particular, the master kinase network involves a number of frequently mutated downstream kinases that demonstrate the extent of altered signaling in tumor cells. The network is also enriched in specific cancer-related signaling pathways such as WNT (FDR $P = 1.2 \times 10^{-3}$), MAPK (FDR $P = 0.023$), mTOR (FDR $P = 7.0 \times 10^{-15}$), and NGF (FDR $P = 5.5 \times 10^{-9}$). PRKCZ has 11 human paralogs and the significant enrichment of pSNVs and SNVs indicates the importance of PKC signaling in cancer ($n = 15$ pSNVs, $P = 7.6 \times 10^{-13}$ and $n = 27$ SNVs, $P = 1.2 \times 10^{-9}$, respectively). While PRKCZ is not a well-known drug target, multiple inhibitors are available for its immediate upstream kinase PDPK1.

Heuristic network search in the kinase–substrate network identifies signaling modules associated with increased patient survival in ovarian cancer

To explore the clinical relevance of our findings, we studied correlation of pSNVs with patient survival data available in the ovarian cancer genome project (Cancer Genome Atlas Research Network, 2011). First, we repeated the pSNV enrichment analysis for pathways and networks separately for ovarian cancer, and interrogated the resulting 552 GO terms, pathways, and protein complexes for correlations with survival. This revealed seven significant GO categories, however, all are related to survival-associated pSNVs in TP53 (Figure 3B), and no significant correlations were found after removing TP53 from the data set.

To discover modules in the kinase–substrate network that correlate with clinical outcome, we designed a local network search algorithm, called HyperModules, that extends concepts from earlier methods (Chuang *et al*, 2007; Reimand *et al*, 2008; Altmae *et al*, 2011). HyperModules starts from a single mutated 'seed' protein and searches its interaction neighborhood for paths of length two to find other proteins that correlate with a given clinical variable (e.g., survival). Paths are merged into larger modules if they improve the correlation, and the search stops once no further improvements are found. Module correlation with survival is assessed with an age-weighted Cox proportional hazards regression model and a likelihood ratio test. We also compute the significance of our findings by evaluating the distribution of Cox *P*-values expected from the network. Statistical significance of a module is assessed using a permutation test in which the search is repeated in 10 000 random neighborhoods of the seed node. Each random network precisely reflects the topology of the original seed-centered network neighborhood, while node labels (protein names) along with associated pSNV sets are shuffled globally over the entire kinase–substrate network. This strategy is useful for two reasons. First, permutations maintain protein-based survival correlations and break down signals originating from kinase–substrate interactions, therefore highlighting real modules where such interactions are important. Second, the strategy corrects for topological biases like hub proteins with large network neighborhoods, as such neighborhoods produce strong survival correlations even with permuted mutations and are therefore considered less significant. Finally, to test modules originating from a particular seed, we discard modules whose survival significance is similar to *P*-values obtained from random networks (FDR $P \leqslant 0.05$).

We searched for survival correlations across all 164 mutated proteins in the network and found 16 significant signaling modules (FDR $P \leqslant 0.05$, permutation test, Supplementary Table 9). One of our top-ranking modules associates with increased survival and describes mutations in eight patients who were all alive at the end of the ovarian cancer study (Figure 6A). This is highly significant according to the Cox survival regression model used in the search ($P = 4.3 \times 10^{-5}$, Figure 6B). In addition, the permutation test shows that such a survival *P*-value is unlikely to be found in randomly mutated networks ($P = 2.2 \times 10^{-3}$, FDR $P = 0.015$, Figure 6C). The module comprises eight mutually exclusive pSNVs in phosphosites and kinase domains of eight proteins (Figure 6D).

Correlations with other clinical variables further support the module: alive patients and tumor-free patients are enriched in the module (Fisher's exact $P = 6.8 \times 10^{-4}$ and $P = 8.0 \times 10^{-3}$, respectively), while subjects of additional chemotherapy are depleted ($P = 8.4 \times 10^{-3}$, Figure 6E–G).

GO enrichment analysis links the survival module to immune response-activating signal transduction and locomotion, among other terms (FDR $P \leqslant 10^{-4}$ from g:Profiler). The network seed is the hemopoietic cell kinase HCK with a highly specific mutation E410K in its kinase domain that potentially disables the adjacent autophosphorylation site at Y411, required for kinase activity (Porter *et al*, 2000). HCK is involved in immune signaling and cell proliferation in hematopoietic cells and is linked to cancer. High levels of HCK are associated with drug resistance in chronic myeloid leukemia, and its constitutively active isoforms induce solid tumors in mice (Poincloux *et al*, 2009; Pene-Dumitrescu and Smithgall, 2010). These published findings, the likely inactivating mutation in *HCK* and the observed mutation pattern suggest that the pSNVs work to disable HCK signaling for the benefit of patient survival. The positive survival correlation is not surprising, as other cancer mutations with positive prognosis are known, for example *NPM1* mutations in acute myeloid leukemia (Verhaak *et al*, 2005).

While the output of our network search algorithm is not directly applicable for predicting clinical outcome, as it is challenged by infrequent mutations and the small-world property of interaction networks, it is useful as an exploratory tool that helps discover and interpret rare, specific mutations in signaling networks that are significantly correlated with clinical outcome.

Discussion

Many cancer genes are discovered because of high mutation frequency in tumor samples. Our analysis considers more detailed information in the form of SNVs that specifically target experimentally determined phosphosites and kinase–substrate interactions. As a result, we are able to highlight

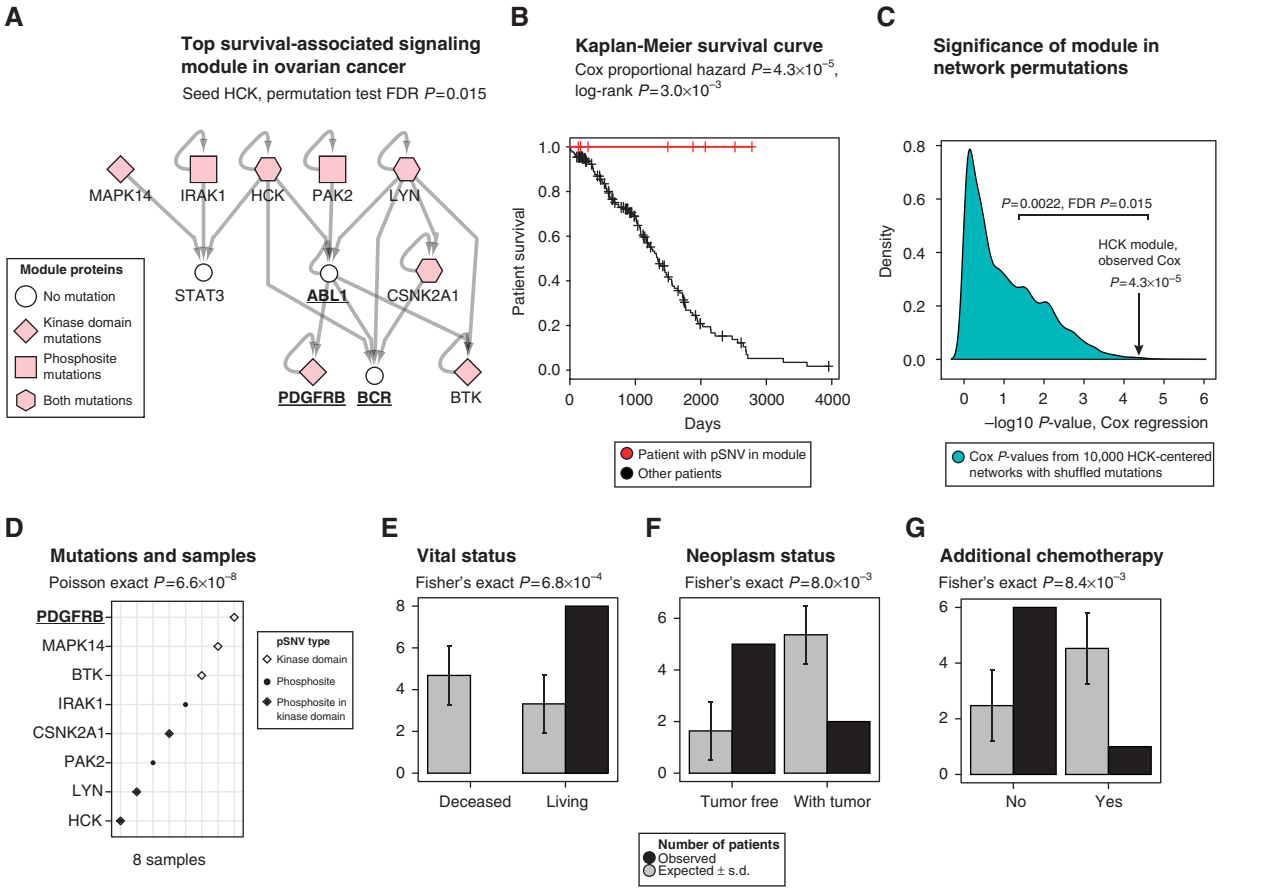

Figure 6 Top survival-associated network module in ovarian cancer. (**A**) One of the top significant survival-associated kinase–substrate signaling modules involves 11 genes and eight pSNVs found in the neighborhood of HCK kinase. (**B**) Kaplan–Meier survival analysis shows a statistically significant difference of module-associated ovarian cancer patients (red line) and other patients (black line), as all patients with mutations in the module were alive at the end of the study. (**C**) Permutation test shows that the observed correlation with increased survival is unlikely to be found in equivalently structured networks with shuffled mutations. (**D**) Survival module involves mutually exclusive pSNVs in eight genes and patients, including the active phosphosite and kinase domain pSNV in the gene of HCK kinase (filled diamond). (**E**–**G**) Additional clinical variables show significant correlation with mutations in the module, including enrichment of alive patients (left) and tumor-free patients (middle) and depletion of subjects of additional chemotherapy (right). Expected values are shown with s.d. from binomial sampling. Names of known cancer genes are underlined and printed in bold.

potential cancer genes that would otherwise remain hidden in the long tail of rare mutations. Pathway and network analysis highlights similar patterns at higher levels of cellular organization. The observed enrichment of known cancer genes in our results validates the analysis, and lends confidence to less-studied genes and systems that we predict to be important in tumor development.

A major challenge in genomics is the functional interpretation of mutations. Functional mutations affect important protein residues, but traditional mutation evaluation methods generally focus on gene and protein specific information, such as amino-acid conservation and DNA sites of alternative splicing (Jordan *et al*, 2010). We can uncover additional information about functional mutations by considering protein sites related to molecular interactions. Protein interaction interfaces are often involved in cell signaling; therefore, significant mutations in these sites are likely to be functional and important in disease. A large class of protein sites in cell signaling are short linear motifs bound by peptide recognition domains (Pawson and Nash, 2003). The writers, readers, and erasers of the phosphorylation machinery recognize phosphorylated motifs, but many other protein domains perform similar functions using other types of sites (Pawson and Nash, 2003). For instance, SH3 domains recognize proline-rich motifs and the histone code has its own set of reader, writer, and eraser domains for post-translational acetylation and methylation. Further, transcription factors recognize regulatory sites at the DNA level. These data are now increasingly available for constructing binding site resolution molecular interaction networks of many human proteins (Chua *et al*, 2004; Dinkel *et al*, 2012; Reimand *et al*, 2012; Wang *et al*, 2012). Such networks allow functional interpretation of mutations with much higher levels of precision than currently possible. For instance, we can identify mutations that alter signaling networks by disrupting or creating binding sites. Our methods are a step in this direction and our future work will include a wider array of signaling domains and their binding motifs.

Our current analysis involves several limitations. First, we only analyzed missense point mutations that are the simplest to interpret, and excluded other mutation types. Considering the short-read DNA sequencing techniques employed by cancer genomics projects, these types of mutations are also likely to be the most accurate. Our methods can be extended to include other mutation types in the future. Second, our analysis treats all phosphosites equally, but phosphorylation is only functional in a cell if the appropriate signaling machinery is properly co-expressed and co-localized. Also, our phosphorylation site data set is derived from a collection of publications, represents a mixture of different cell types, tissues and experimental conditions, and is not specific to the cancer types we study. Cancer-specific phosphoproteomics experiments will hopefully be available in the future to address this limitation. We currently address this limitation by assigning a higher score to genes with phosphorylation sites that are mutated multiple times in different samples and also mutated in multiple sites within the same protein. We thus assume that the highest scoring proteins from our analysis are more likely to be cancer relevant due to the repeated and statistically significant mutation pattern. Finally, mutations in transcriptionally active genes are more likely to be functional than mutations in silent genes, and mutation calls combined with RNA sequencing of corresponding samples therefore reveal an 'active' set of mutations. Mutation filtering will increase the precision of our methods once these data become more widely available. Our models can also consider additional factors that indicate the importance of phosphorylation and other functional sites, such as site conservation (Tan *et al*, 2009), stoichiometry, and number of kinases targeting a site.

Protein phosphorylation machinery, in particular the protein kinase family, is an important drug development target, and several agents such as kinase inhibitors are routinely used in the clinic. Phosphorylation-specific cancer mutations are therefore likely to highlight potential drug targets and may be useful as predictive markers of drug response, or to identify alternate agents in primary-drug-resistant tumors.

We present a wealth of hypotheses for cancer-related discovery, and several novel analysis methods for interpreting cancer genomes. Our approaches will become more powerful as data accumulate from the expanding efforts to decipher cancer genomes.

Materials and methods

Phosphorylation data

Experimentally determined phosphorylated residues and their flanking regions of ± seven residues (referred to as phosphosites) and phosphosite-associated kinases were retrieved from three curated public databases: PhosphoSitePlus (Hornbeck *et al*, 2012), PhosphoELM (Dinkel *et al*, 2011), and Human Protein Reference Database (HPRD) (Keshava Prasad *et al*, 2009) (all downloaded on 30 November 2011). Phosphosites were mapped to high-confidence protein sequences from the Consensus Coding Sequence (CCDS) database (Pruitt *et al*, 2009) (build 20110907, NCBI build 37.3). We mapped phosphopeptides to CCDS sequences using exact sequence matching to avoid discrepancies between protein isoforms, and discarded non-matching peptides. Phosphosites with overlapping flanking sequences were merged into continuous regions. Kinase domains were retrieved from the HPRD database and mapped to protein sequences using a similar procedure. We used HGNC symbols provided by all three phosphosite databases, and retrieved all corresponding isoforms from the CCDS data set after removing 1380 sequences with non-public status and 16 pseudoautosomal genes. We selected the longest isoform of every protein and discarded the remaining isoforms from further analysis.

Somatic mutations in cancer genomes

Somatic mutations from nine cancer genomics projects (Wood *et al*, 2007; Cancer Genome Atlas Research Network, 2008, 2011; Ding *et al*, 2008; Jones *et al*, 2008; Parsons *et al*, 2008; Puente *et al*, 2011; Totoki *et al*, 2011) were downloaded from the International Cancer Genome Consortium (ICGC) data portal (The International Cancer Genome Consortium (2010), version 4–6, downloaded on 30 November 2011). A subset of mutations matching the human genome build 36 were mapped to build 37 with the LiftOver software of the UCSC Genome Browser. We discarded all non-coding mutations, silent mutations, multi-nucleotide substitutions, insertions and deletions from all data sets, and retained only non-synonymous, missense SNVs. Further, all mutations in a given gene and sample were discarded if they contained a mutation that created a premature stop codon or removed the existing stop codon.

Phosphorylation-associated SNVs (pSNVs) comprised SNV modified phosphosites or kinase domains. Phosphosite-associated pSNVs involved mutations that directly modified central, phosphorylated serine (S), threonine (T), and tyrosine (Y) residues (i.e., direct pSNVs),

as well as mutations flanking the central S/T/Y within seven residues. Kinase-associated pSNVs were mapped to domains from HPRD. We compared phosphosite-associated pSNVs with single-nucleotide polymorphisms (SNPs) published by the 1000 Genomes Project Consortium (2010) and found a small set of 12 loci that overlap, representing 1.5% of unique missense pSNVs.

Global analyses of phosphosite mutations

Global enrichment of SNVs in phosphosites was determined using Fisher's exact test, by considering the protein-coding sequence length of all involved cancer samples as the background. The background was corrected for each cancer data set independently, as some studies sequenced all genes and some sequenced a focused set of genes. We considered all genes described in the publication and any additional genes indicated in the ICGC data sets.

Global phosphorylation network properties were evaluated with the non-parametric Wilcoxon test, by comparing the degree distribution of mutated and non-mutated proteins in the kinase–substrate network. To assess protein centrality, we compared the betweenness centrality distribution of mutated and non-mutated proteins in the network with the IGraph R package (Csardi and Nepusz, 2006). An edge in the phosphorylation network was considered mutated if either the kinase or substrate participants had at least one mutation in the kinase domain or any phosphosite. The enrichment of mutated kinase–kinase edges in the phosphorylation network was determined with Fisher's exact test, by taking the collection of all kinase–substrate interactions as the statistical background.

The high-confidence collection of 555 cancer genes was defined as the union of genes described in four review papers (Mitelman, 2000; Hahn and Weinberg, 2002; Futreal *et al*, 2004; Vogelstein and Kinzler, 2004), as collected in the databases of Cancer Genes (Higgins *et al*, 2007), and the Cancer Gene Census (Futreal *et al*, 2004) (downloaded on 15 December 2011). We also compared our list of phosphomutated genes with coding mutations from the Catalog Of Somatic Mutations In Cancer (COSMIC; Forbes *et al* (2010), downloaded on 5 April 2012) and found that 99% were expectedly present as COSMIC contains data from all published cancer genomics projects. The list of 1870 human transcription factors was published by Vaquerizas *et al* (2009). The list of 546 human kinases was compiled using kinase domain information from the HPRD database (Keshava Prasad *et al*, 2009) as well as kinase–substrate interactions from PhosphositePlus, PhosphoELM and HPRD. The set of 1658 genes corresponding to known approved, experimental, and illicit drug targets excluding nutraceuticals were retrieved from the DrugBank database (Knox *et al* (2011), downloaded on 16 October 2012). Paralogs for gene family analysis were retrieved from Ensembl 69. Statistical analyses with these lists were performed with the Fisher's exact test, by considering the final set of unique CCDS genes as the background. Protein disorder predictions were performed with DISOPRED2 (Ward *et al*, 2004), using a local installation of the software and the NCBI BLAST 2.26 NR data bank (downloaded on 21 June 2012).

ActiveDriver analysis of phosphosite mutations

We developed and applied a generalized linear regression model, called ActiveDriver, to find phosphosites whose mutations are unexpected given the background mutation rate. The model assumes that missense mutations in a sequence of a particular gene and cancer type follow the Poisson probability distribution

$$Po(y;\ \lambda) = \frac{\lambda^y \exp(-\lambda)}{y!}$$

where $y \geqslant 0$, $y \in \mathbf{N}$ is the observed number of SNVs, and $\lambda > 0$, $\lambda \in \mathbf{R}$ is the missense mutation rate of the protein sequence of the gene. The rate parameter λ corresponds to the expected number of mutations per residue $E(Y) = \lambda$, as well as its variance $Var(Y) = E(Y^2) - E(Y)^2 = \lambda$, where $Y = y_1, \ldots, y_n$ is the vector of mutation counts per residue in the protein sequence of n residues.

In a data set of n samples, generalized linear regression models express the dependency between a response vector $Y = y_1, \ldots, y_n$ and k predictor vectors $X_1 = x_{11}, \ldots, x_{n1}; \ldots; X_k = x_{1k}, \ldots, x_{nk}$, as

$$g(E(Y \mid X)) = \beta_0 + \sum_{j=1}^{k} \beta_j X_j = X\beta$$

where $\beta = \beta_1 \ldots \beta_k$ is a vector of model coefficients and β_0 is the model intercept, $X = X_1; \ldots; X_k$ is the predictor matrix and $X^T\beta$ is the dot product of the transposed predictor matrix and model coefficients, and $g()$ is a link function for non-linear transformation. Poisson regression expresses the dependency between response and predictors through the Poisson distribution, as

$$E(Y \mid X) = \exp(X^{\mathrm{T}}\beta)$$

where $Y \sim Po(\lambda)$ and the link function $g() = \ln()$ corresponds to the natural logarithm.

The likelihood of a Poisson regression model given data is computed as the product of Poisson probabilities of response values y_i of samples indexed with i, given corresponding values x_{ij} of predictors indexed with j, as

$$L(\beta \mid Y, X) = \prod_{i=1}^{n} \frac{\exp(X_{i*}^{\mathrm{T}}\beta)^{y_i} \exp(-\exp(X_{i*}^{\mathrm{T}}\beta))}{y_i!}$$

where X_{i*} is the vector of k predictor values $x_{i1}, \ldots, x_{ik}$ for a given sample i. The log likelihood of the model is equivalent and more efficient in practice:

$$l(\beta \mid Y, X) = \log(L(\beta \mid Y, X)) = \sum_{i=1}^{n} y_i X_{i*}^{\mathrm{T}}\beta - \exp(X_{i*}^{\mathrm{T}}\beta) - \log(y_i!)$$

Maximum likelihood estimation (MLE) is used to find model coefficients $\hat{\beta}$ that give the best agreement (i.e., smallest absolute log likelihood) between response and predictor values,

$$\hat{l}(\hat{\beta} \mid Y, X) = \arg\max_{\beta} l(\beta \mid Y, X) = \sum_{i=1}^{N} y_i X_{i*}^{\mathrm{T}}\beta - \exp(X_{i*}^{\mathrm{T}}\beta)$$

The factorial term $y_i!$ does not involve coefficients and can be discarded from the estimation. While no analytical solutions exist for the above equation, the Iteratively Reweighted Least Squares algorithm is used to find optimal coefficients β and a corresponding MLE value $\hat{l}$.

In our approach, we used Poisson regression to test whether a specific phosphosite region in a given gene involves a significantly different mutation rate than the gene on average. According to our null hypothesis, mutations across the whole protein sequence of the gene follow the Poisson distribution with the intercept coefficient reflecting background rate β_0 linearly combined with a structure parameter $X_{(s)}$ and corresponding coefficient $\beta_{(s)}$. The null hypothesis is expressed as the following intercept-only model

$$\mathbf{h_0}: E(Y) = \exp(\beta_0 + \beta_{(s)} + X_{(s)})$$

in which $X_{(s)}$ is set to one if the sequence position corresponds to disordered protein sequence, and equals zero if the corresponding region is structured (non-disordered). According to our alternative hypothesis, the mutations in the phosphosite region q are generated by rates λ that are significantly different from the baseline rate λ_0 while considering protein disorder:

$$\mathbf{h_1}: E(Y \mid X) = \exp(\beta_0 + \beta_{(s)} + X_{(s)} + \beta_{(q)} + X_{(q)})$$

In addition, our alternative model may include information about phosphosite density if it significantly improves model fit. The information is expressed in three additional predictor variables $X_{(d)}, X_{(v)}, X_{(w)}$ that are set to zero in protein sequence positions outside the phosphosite region of interest, and otherwise encode relative phosphosite position within the region. Specifically, $x_{i(d)}$ equals one if position i of protein sequence encodes a phosphosite, and zero otherwise. The value $x_{i(v)}$ is set to express the number of nearby phosphosites within a flanking region of $\pm(1 \ldots 2)$ residues around sequence position i. Similarly, the value $x_{i(w)}$ expresses the number of distant phosphosites within a region of $\pm(3 \ldots 7)$ residues. These variables are added to the model one by one using a forward stepwise model selection procedure that evaluates extended models with the

Akaike Information Criterion (AIC):

$$\mathrm{AIC} = -2\times\hat{l}(\mathbf{h_1}) + 2\times v_{\mathbf{h_1}}$$

where $\hat{l}$ is the MLE value and v is the number of degrees of freedom (number of model coefficients). At every step, adjacency-encoding predictors are added to the model one by one, corresponding AIC values are computed, and the predictor that provides the greatest increase in AIC is added to the model. If no additional factors are sufficient for improvement, the original alternative hypothesis is used for significance estimation.

To compare the null $\mathbf{h_0}$ and alternative $\mathbf{h_1}$ models, we use MLE to determine model likelihoods at optimal coefficients, and compute the deviance statistic, as

$$D = -2\times(\hat{l}(\mathbf{h_0}) - \hat{l}(\mathbf{h_1})) = -2\times(\hat{l}(\hat{\beta}_{h_0} \mid Y) - \hat{l}(\hat{\beta}_{h_1} \mid Y, X))$$

A high deviance statistic indicates that the alternative model of phosphosite-specific mutation rates fits the observed mutations considerably better than the null model of gene-wide mutation rate. The deviance statistic approximately follows the χ^2 distribution with $v = v_{\mathbf{h_1}} - v_{\mathbf{h_0}}$ degrees of freedom. The log-likelihood ratio test is used to estimate the statistical significance of deviance

$$P(\mathbf{h_0} \mid X, Y) = P_{\chi^2}(D, v)$$

and the null hypothesis is refuted if the P-value of the likelihood ratio test is $P \leqslant 0.05$.

The alternative hypotheses in ActiveDriver are tested separately for every phosphosite region, gene and cancer type. The cancer type-specific composite P-value for a given gene is computed as a product of significant P-values (only including $P \leqslant 0.05$) for all phosphosite regions. We subsequently correct composite P-values with the Benjamini–Hochberg FDR separately for each cancer type. Genes with no phosphosites or no cancer SNVs are discarded prior to modeling and not included in the multiple testing procedures.

To identify proteins with significant kinase domain mutations, we implemented a version of ActiveDriver that considers a simplified alternative hypothesis for pSNV enrichment detection, as

$$\mathbf{h_1} : E(Y) = \exp(\beta_0 + \beta_{(s)}X_{(s)} + \beta_{(d)}X_{(d)})$$

where $x_{i,(d)}$ is set to one if position i of protein sequence encodes a kinase domain and zero otherwise. All other factors of our method remain the same. We used the simplified model to analyze the subset of kinase proteins with SNVs, and found that only *EGFR* involves a significant enrichment of kinase domain-specific pSNVs in lung cancer samples.

Comparison of ActiveDriver results with alternative approaches

Cancer genes are traditionally identified based on high mutation frequency in tumors. To compare our method to the traditional approach, we first compared the genes ranked by ActiveDriver to all genes with missense mutations ranked according to their mutation frequency (number of missense point mutations) in all studied cancer data sets.

Second, we compared gene mutation rates with global (exome-wide) mutation rates using the binomial statistic. We computed independent gene-based mutation rates for each cancer type using distinct background rates and only genes and cancer samples sequenced in corresponding projects. The two-tailed binomial test was used to assess the number of observed missense mutations in a gene, given the total number of missense mutations and total sequence length (amino-acid positions times number of relevant samples). For a particular cancer type, all sequenced genes were tested with the binomial statistic and corrected for multiple testing (FDR $P \leqslant 0.05$).

Third, we designed a simplified gene-centric binomial statistic to test phosphosite mutation rates in comparison to the Poisson regression model in ActiveDriver. Except for the statistical model as well as protein disorder and phosphosite density integration, all other aspects of the method remained the same. Similarly to the global method, the two-tailed binomial test was used to evaluate number of missense mutations in a particular phosphosite region of a gene of interest, given the number of missense mutations and total protein sequence length. The final P-value for the gene was computed as the product of significant P-values from site-specific tests (only including $P \leqslant 0.05$), followed by multiple testing correction independently for each cancer type.

Pathway enrichment analysis of phosphosite mutations

Pathway enrichment analysis of phosphosite mutations involved testing a given gene list for enrichment of biological process, molecular function and cell component annotations from GO (Ashburner *et al*, 2000), curated human biological pathways from Reactome (Matthews *et al*, 2009), and mammalian protein complexes from the CORUM database (Ruepp *et al*, 2010). We also studied human disease genes from the Human Phenotype Ontology (Robinson and Mundlos, 2010), microRNA target genes from MicroCosm (Griffiths-Jones *et al*, 2008) and transcription factor target genes from TRANSFAC (Matys *et al*, 2006); however, these data sets provided no significant categories after multiple testing correction. Each type of gene set was analyzed and corrected for multiple testing separately. All functional annotations except protein complexes were retrieved from g:Profiler (Reimand *et al*, 2011). We discarded small gene sets with less than three genes and general sets with >1000 genes, and did not use KEGG pathway information due to strong bias towards well-annotated cancer genes (e.g., the 'pathways in cancer' pathway). Ordered functional enrichment analyses were carried out with the g:Profiler software.

Pathway enrichment analysis was conducted across all cancer types simultaneously. The enrichment P-value was computed using a one-sample one-tailed Poisson exact test, using a null hypothesis of uniform background mutation rate across all genes $\lambda_0 = m/n$, where m and n reflect counts of all SNVs and genes, respectively. The P-value of seeing m_p or more mutations in a pathway of n_p genes, given the background gene mutation rate λ_0, is computed as one minus the sum of all less extreme events:

$$p(X \geqslant m_p;\ n_p \mid \lambda_0) = 1 - \sum_{k=0}^{m_p-1} \frac{\lambda_0^k n_p^k}{k!} \exp(-\lambda_0 n_p)$$

Gene sets covering only a single mutated sample, or gene sets with mutations in a single gene were discarded prior to statistical testing. P-values were corrected for multiple testing with Benjamini–Hochberg FDR and filtered for significant results (FDR $P \leqslant 0.05$). We considered the set of genes with kinase domains or phosphosites as the statistical background, after excluding EGFR and TP53 because of their excessive mutation rate in phosphosites (181 samples) and in general (340 samples). The analysis was carried out on the full collection of SNVs to identify gene sets with significant sample coverage, and also with only pSNVs, finally keeping only terms uniquely significant in the pSNV analysis. The non-specific SNV analysis used all unique genes in the CCDS data set as statistical background, while the pSNV analysis involved a more stringent set of genes with kinase domains or phosphosites, both excluding *EGFR* and *TP53*. The analysis focusing on all SNVs additionally excluded *KRAS* from significance tests, since it introduced a functional bias due to SNVs in 187 samples (*KRAS* involves no pSNVs and is highlighted in our model due to pSNV depletion). Gene sets from GO and Reactome were filtered to reduce redundancy, by keeping the most significant term among those with a common ancestor. For significant CORUM protein complexes that covered the exact same samples, only the complex with the strongest P-value was retained in the analysis. A separate sequence of permutation tests was carried out for ovarian cancer samples to identify gene sets with clinical correlation.

Kinase-specific subnetwork analysis

Experimentally determined kinase–phosphosite interactions were retrieved from three public databases PhosphoSitePlus (Hornbeck *et al*, 2012), Phospho ELM (Dinkel *et al*, 2011) and HPRD (Keshava Prasad *et al*, 2009). Kinase-specific subnetworks include (i) a central kinase, (ii) all downstream substrates of the central kinase, and (iii) upstream kinases phosphorylating the central kinase. Cancer samples

were considered to be mutated in subnetworks if they contained at least one mutation in any kinase domain or phosphosite in subnetwork genes. Gene sets covering only a single mutated sample, or gene sets with mutations in a single gene were discarded prior to statistical testing. For evaluating statistical significance of pSNV enrichment, we used a more stringent statistical background of all genes in the kinase–substrate network. We employed the Poisson tests described above, and identified kinases with more than the expected number of pSNV mutations (FDR $P \leqslant 0.05$). As above, direct mutations of *TP53* and *EGFR* where excluded from the analysis and calculation of background mutation rate. Hierarchical classes of kinases were defined as follows. Master regulators are kinases whose modules involved >20 samples with pSNVs (i.e., two-fold median sample coverage over all significant modules) and at least two-thirds of pSNVs occurring downstream of the central kinase. Local kinases, managed kinases and self-employed kinases involved modules with at least two-thirds of pSNVs in downstream targets, upstream targets, and central kinases, respectively. The middle manager class covered the remaining kinases.

Discovery of clinically correlated modules in the kinase–substrate network

Clinical annotation of ovarian cancer samples (Cancer Genome Atlas Research Network, 2011) was retrieved from the Cancer Genome Atlas (TCGA, downloaded on 21 December 2011). We only included the patients that were screened for SNVs as the test set, and excluded patients with missing values in the particular annotation type tested. To test patient survival, we used a regression model involving age, vital status, time of follow-up and mutation status in the given module.

We developed a greedy algorithm to search the kinase–substrate network for signaling modules that cover pSNVs that are maximally informative of a given clinical outcome. We defined statistical objective functions to determine the best steps in the search, comparing (a) the clinical signal in the group of patients defined by a set of pSNVs in a given signaling module and (b) the clinical signal in the remaining patient cohort. For survival analysis, the objective function uses the Cox Proportional Hazards (PH) framework of generalized linear regression, as

$$\lambda(t \mid X) = \lambda_0(t) \exp(X^{\mathrm{T}}\beta)$$

where λ_0 reflects constant positive baseline risk, t corresponds to survival time, X is the matrix of predictor variables and β is the vector of model coefficients. As the null hypothesis, survival is modeled in a univariate regression model with patient age as the confounding predictor. The alternative hypothesis includes an additional indicator predictor that reflects the mutation status of each sample in the given module. The significance of survival difference between the module-related group and the remaining cohort is estimated in the likelihood ratio test by comparing the fit and degrees of freedom of the null model and the alternative model.

The network search starts from a given gene in the kinase–substrate network as the 'seed' and constrains the search space to the neighborhood of distance $d = 2$ from the seed. First, the algorithm extracts all paths of length two, each comprising the seed and two additional proteins. The set of paths is then filtered to reduce the search space, so that only paths linking to additional pSNV mutations are considered. Paths are then iteratively merged into modules containing two or more paths. At each merging iteration, all possible pairwise combinations of modules and/or paths are considered for merging, and the pair with the greatest improvement in clinical correlation according to Cox regression is merged into one module. Merging stops when no such improvement steps are available.

The statistical significance of identified modules in a given seed neighborhood is assessed using 10 000 permutations to evaluate the non-random association between network topology, number of proteins included and corresponding pSNVs. The interaction topology of the neighborhood is retained during permutation, while node labels (protein names) are shuffled globally across the entire kinase–substrate network. This strategy preserves the gene-based correlation with survival, while disrupting the correlation originating from multiple closely interacting proteins. The search algorithm is executed on the set of random neighborhoods with shuffled labels, and the resulting modules and clinical correlations (Cox *P*-values) provide the null model for statistical significance estimation. The significance of each module is computed as the fraction of *P*-values of randomly found modules that have an equivalent or better clinical correlation Cox *P*-value. The *P*-values of all modules originating from a single seed are then filtered with multiple testing correction (FDR $P \leqslant 0.05$). To find all survival-associated modules, we searched the network with every mutated protein as a seed.

To validate survival modules, we also compared related patients and mutations with other types of clinical information using Fisher's exact test, including vital status, neoplasm status, tumor stage, and additional chemotherapy. To form larger sample groups for statistical analysis, we simplified tumor stage classification by removing alphabetical subclasses, resulting in three stages (II, III, IV). Besides age-adjusted Cox regression, modules were validated using alternative significance *P*-values of survival correlation, namely likelihood ratio tests with unadjusted Cox regression and log-rank tests. Clinical correlation tests for ovarian pSNV-enriched GO categories, pathways, protein complexes and individual genes were performed in a similar manner, followed by multiple testing correction and filtering (FDR $P \leqslant 0.05$). Survival correlation of TP53 pSNVs in ovarian cancer was identified using Cox regression and log-rank tests. A separate set of tests was used for the subset of glioblastoma patients with long-term survival (follow-up time >1 year).

ActiveDriver availability

R Source code of ActiveDriver is available at the website www.baderlab.org/Software/ActiveDriver.

Supplementary information

Supplementary information is available at the *Molecular Systems Biology* website (www.nature.com/msb).

Acknowledgements

We thank Shirley Hui, Leopold Parts, Michael D Taylor and Liis Uusküla for critical comments on the manuscript; Ian Clarke, Peter Dirks, Mona Meyer, Jason Moffat, Robert Rottapel, Andrea Uetrecht and all members of the Bader lab for useful discussions; and Ruth Isserlin for providing drug target gene sets. This work was supported by the Canadian Institutes of Health Research grant MOP-84324 and the National Resource for Network Biology (NRNB) under award numbers P41 RR031228 and GM103504. Computational resources for large-scale simulations were provided by the High Performance Computing Centre at the University of Tartu, Estonia.

Author contributions: JR designed the study, analyzed the data, implemented the algorithms, and wrote the first manuscript. GDB designed and supervised the study. Both authors wrote and approved the final manuscript.

Conflict of Interest

The authors declare that they have no conflict of interest.

References

1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. *Nature* **467:** 1061–1073

Altmae S, Reimand J, Hovatta O, Zhang P, Kere J, Laisk T, Saare M, Peters M, Vilo J, Stavreus-Evers A, Salumets A (2012) Research resource: interactome of human embryo implantation: identification of gene expression pathways, regulation, and integrated regulatory networks. *Mol Endocrinol* **26:** 1203–1217

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25:** 25–29

Baldassarre M, Razinia Z, Burande CF, Lamsoul I, Lutz PG, Calderwood DA (2009) Filamins regulate cell spreading and initiation of cell migration. *PLoS ONE* **4:** e7830

Banerji S, Cibulskis K, Rangel-Escareno C, Brown KK, Carter SL, Frederick AM, Lawrence MS, Sivachenko AY, Sougnez C, Zou L, Cortes ML, Fernandez-Lopez JC, Peng S, Ardlie KG, Auclair D, Bautista-Pina V, Duke F, Francis J, Jung J, Maffuz-Aziz A *et al* (2012) Sequence analysis of mutations and translocations across breast cancer subtypes. *Nature* **486:** 405–409

Bech-Otschir D, Kraft R, Huang X, Henklein P, Kapelari B, Pollmann C, Dubiel W (2001) COP9 signalosome-specific phosphorylation targets p53 to degradation by the ubiquitin system. *EMBO J* **20:** 1630–1639

Cancer Genome Atlas Research Network (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455:** 1061–1068

Cancer Genome Atlas Research Network (2011) Integrated genomic analyses of ovarian carcinoma. *Nature* **474:** 609–615

Chao C, Herr D, Chun J, Xu Y (2006) Ser18 and 23 phosphorylation is required for p53-dependent apoptosis and tumor suppression. *EMBO J* **25:** 2615–2622

Chapman PB, Hauschild A, Robert C, Haanen JB, Ascierto P, Larkin J, Dummer R, Garbe C, Testori A, Maio M, Hogg D, Lorigan P, Lebbe C, Jouary T, Schadendorf D, Ribas A, O'Day SJ, Sosman JA, Kirkwood JM, Eggermont AM *et al* (2011) Improved survival with vemurafenib in melanoma with BRAF V600E mutation. *N Engl J Med* **364:** 2507–2516

Chehab NH, Malikzay A, Stavridi ES, Halazonetis TD (1999) Phosphorylation of Ser-20 mediates stabilization of human p53 in response to DNA damage. *Proc Natl Acad Sci USA* **96:** 13777–13782

Chua G, Robinson MD, Morris Q, Hughes TR (2004) Transcriptional networks: reverse-engineering gene regulation on a global scale. *Curr Opin Microbiol* **7:** 638–646

Chuang HY, Lee E, Liu YT, Lee D, Ideker T (2007) Network-based classification of breast cancer metastasis. *Mol Syst Biol* **3:** 140

Collins FS, Barker AD (2007) Mapping the cancer genome. Pinpointing the genes involved in cancer will help chart a new course across the complex landscape of human malignancies. *Sci Am* **296:** 50–57

Csardi G, Nepusz T (2006) The igraph software package for complex network research. *InterJournal* Complex Systems 1695

Del Valle-Perez B, Martinez VG, Lacasa-Salavert C, Figueras A, Shapiro SS, Takafuta T, Casanovas O, Capella G, Ventura F, Vinals F (2010) Filamin B plays a key role in vascular endothelial growth factor-induced endothelial cell motility through its interaction with Rac-1 and Vav-2. *J Biol Chem* **285:** 10748–10760

Dephoure N, Zhou C, Villen J, Beausoleil SA, Bakalarski CE, Elledge SJ, Gygi SP (2008) A quantitative atlas of mitotic phosphorylation. *Proc Natl Acad Sci USA* **105:** 10762–10767

Ding L, Getz G, Wheeler DA, Mardis ER, McLellan MD, Cibulskis K, Sougnez C, Greulich H, Muzny DM, Morgan MB, Fulton L, Fulton RS, Zhang Q, Wendl MC, Lawrence MS, Larson DE, Chen K, Dooling DJ, Sabo A, Hawes AC *et al* (2008) Somatic mutations affect key pathways in lung adenocarcinoma. *Nature* **455:** 1069–1075

Dinkel H, Chica C, Via A, Gould CM, Jensen LJ, Gibson TJ, Diella F (2011) Phospho.ELM: a database of phosphorylation sites—update 2011. *Nucleic Acids Res*, 39: D261–D267

Dinkel H, Michael S, Weatheritt RJ, Davey NE, Van Roey K, Altenberg B, Toedt G, Uyar B, Seiler M, Budd A, Jodicke L, Dammert MA, Schroeter C, Hammer M, Schmidt T, Jehl P, McGuigan C, Dymecka M, Chica C, Luck K *et al* (2012) ELM—the database of eukaryotic linear motifs. *Nucleic Acids Res* **40:** D242–D251

Forbes SA, Tang G, Bindal N, Bamford S, Dawson E, Cole C, Kok CY, Jia M, Ewing R, Menzies A, Teague JW, Stratton MR, Futreal PA (2010) COSMIC (the Catalogue of Somatic Mutations in Cancer): a resource to investigate acquired mutations in human cancer. *Nucleic Acids Res* **38:** D652–D657

Francesconi A, Duvoisin RM (2000) Opposing effects of protein kinase C and protein kinase A on metabotropic glutamate receptor signaling: selective desensitization of the inositol trisphosphate/Ca2+ pathway by phosphorylation of the receptor-G protein-coupling domain. *Proc Natl Acad Sci USA* **97:** 6185–6190

Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR (2004) A census of human cancer genes. *Nat Rev Cancer* **4:** 177–183

Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ (2008) miRBase: tools for microRNA genomics. *Nucleic Acids Res* **36:** D154–D158

Groth A, Ray-Gallet D, Quivy JP, Lukas J, Bartek J, Almouzni G (2005) Human Asf1 regulates the flow of S phase histones during replicational stress. *Mol Cell* **17:** 301–311

Hahn WC, Weinberg RA (2002) Modelling the molecular circuitry of cancer. *Nat Rev Cancer* **2:** 331–341

Hanahan D, Weinberg RA (2011) Hallmarks of cancer: the next generation. *Cell* **144:** 646–674

Higgins ME, Claremont M, Major JE, Sander C, Lash AE (2007) CancerGenes: a gene selection resource for cancer genome projects. *Nucleic Acids Res* **35:** D721–D726

Hornbeck PV, Kornhauser JM, Tkachev S, Zhang B, Skrzypek E, Murray B, Latham V, Sullivan M (2012) PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Res* **40:** D261–D270

Hynes NE, Lane HA (2005) ERBB receptors and cancer: the complexity of targeted inhibitors. *Nat Rev Cancer* **5:** 341–354

Jones S, Zhang X, Parsons DW, Lin JC, Leary RJ, Angenendt P, Mankoo P, Carter H, Kamiyama H, Jimeno A, Hong SM, Fu B, Lin MT, Calhoun ES, Kamiyama M, Walter K, Nikolskaya T, Nikolsky Y, Hartigan J, Smith DR *et al* (2008) Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science* **321:** 1801–1806

Jordan DM, Ramensky VE, Sunyaev SR (2010) Human allelic variation: perspective from protein function, structure, and evolution. *Curr Opin Struct Biol* **20:** 342–350

Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A, Balakrishnan L, Marimuthu A, Banerjee S, Somanathan DS, Sebastian A, Rani S, Ray S, Harrys Kishore CJ, Kanth S, Ahmed M *et al* (2009) Human Protein Reference Database—2009 update. *Nucleic Acids Res* **37:** D767–D772

Knox C, Law V, Jewison T, Liu P, Ly S, Frolkis A, Pon A, Banco K, Mak C, Neveu V, Djoumbou Y, Eisner R, Guo AC, Wishart DS (2011) DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res* **39:** D1035–D1041

Lim WA, Pawson T (2010) Phosphotyrosine signaling: evolving a new cellular communication system. *Cell* **142:** 661–667

Lu J, Lian G, Lenkinski R, De Grand A, Vaid RR, Bryce T, Stasenko M, Boskey A, Walsh C, Sheen V (2007) Filamin B mutations cause chondrocyte defects in skeletal development. *Hum Mol Genet* **16:** 1661–1675

Matthews L, Gopinath G, Gillespie M, Caudy M, Croft D, de Bono B, Garapati P, Hemish J, Hermjakob H, Jassal B, Kanapin A, Lewis S, Mahajan S, May B, Schmidt E, Vastrik I, Wu G, Birney E, Stein L, D'Eustachio P (2009) Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res* **37:** D619–D622

Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K, Voss N, Stegmaier P, Lewicki-Potapov B, Saxel H, Kel AE, Wingender E (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res* **34:** D108–D110

Mitelman F (2000) Recurrent chromosome aberrations in cancer. *Mutat Res* **462:** 247–253

Morin PJ, Sparks AB, Korinek V, Barker N, Clevers H, Vogelstein B, Kinzler KW (1997) Activation of beta-catenin-Tcf signaling in colon cancer by mutations in beta-catenin or APC. *Science* **275:** 1787–1790

Nagano K, Shinkawa T, Mutoh H, Kondoh O, Morimoto S, Inomata N, Ashihara M, Ishii N, Aoki Y, Haramura M (2009) Phosphoproteomic analysis of distinct tumor cell lines in response to nocodazole treatment. *Proteomics* **9:** 2861–2874

Nowak SJ, Corces VG (2004) Phosphorylation of histone H3: a balancing act between chromosome condensation and transcriptional activation. *Trends Genet* **20:** 214–220

Olsen JV, Vermeulen M, Santamaria A, Kumar C, Miller ML, Jensen LJ, Gnad F, Cox J, Jensen TS, Nigg EA, Brunak S, Mann M (2010) Quantitative phosphoproteomics reveals widespread full phosphorylation site occupancy during mitosis. *Sci Signal* **3:** ra3

Ortiz P, Vanaclocha F, Lopez-Bran E, Esquivias JI, Lopez-Estebaranz JL, Martin-Gonzalez M, Arrue I, Garcia-Romero D, Ochoa C, Gonzalez-Perez A, Ruiz A, Real LM (2007) Genetic analysis of the GRM1 gene in human melanoma susceptibility. *Eur J Hum Genet* **15:** 1176–1182

Parsons DW, Jones S, Zhang X, Lin JC, Leary RJ, Angenendt P, Mankoo P, Carter H, Siu IM, Gallia GL, Olivi A, McLendon R, Rasheed BA, Keir S, Nikolskaya T, Nikolsky Y, Busam DA, Tekleab H, Diaz LA, Hartigan J *et al* (2008) An integrated genomic analysis of human glioblastoma multiforme. *Science* **321:** 1807–1812

Pawson T, Nash P (2003) Assembly of cell regulatory systems through protein interaction domains. *Science* **300:** 445–452

Pene-Dumitrescu T, Smithgall TE (2010) Expression of a Src family kinase in chronic myelogenous leukemia cells induces resistance to imatinib in a kinase-dependent manner. *J Biol Chem* **285:** 21446–21457

Poincloux R, Al Saati T, Maridonneau-Parini I, Le Cabec V (2009) The oncogenic activity of the Src family kinase Hck requires the cooperative action of the plasma membrane- and lysosome-associated isoforms. *Eur J Cancer* **45:** 321–327

Pollock PM, Cohen-Solal K, Sood R, Namkoong J, Martino JJ, Koganti A, Zhu H, Robbins C, Makalowska I, Shin SS, Marin Y, Roberts KG, Yudt LM, Chen A, Cheng J, Incao A, Pinkett HW, Graham CL, Dunn K, Crespo-Carbone SM *et al* (2003) Melanoma mouse model implicates metabotropic glutamate signaling in melanocytic neoplasia. *Nat Genet* **34:** 108–112

Porter M, Schindler T, Kuriyan J, Miller WT (2000) Reciprocal regulation of Hck activity by phosphorylation of Tyr(527) and Tyr(416). Effect of introducing a high affinity intramolecular SH2 ligand. *J Biol Chem* **275:** 2721–2726

Pruitt KD, Sharrow J, Harte RA, Wallin C, Diekhans M, Maglott DR, Searle S, Farrell CM, Loveland JE, Ruef BJ, Hart E, Suner MM, Landrum MJ, Aken B, Ayling S, Baertsch R, Fernandez-Banet J, Cherry JL, Curwen V, Dicuccio M *et al* (2009) The consensus coding sequence (CCDS) project: identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res* **19:** 1316–1323

Puente XS, Pinyol M, Quesada V, Conde L, Ordonez GR, Villamor N, Escaramis G, Jares P, Bea S, Gonzalez-Diaz M, Bassaganyas L, Baumann T, Juan M, Lopez-Guerra M, Colomer D, Tubio JM, Lopez C, Navarro A, Tornador C, Aymerich M *et al* (2011) Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. *Nature* **475:** 101–105

Reimand J, Arak T, Vilo J (2011) g:Profiler—a web server for functional interpretation of gene lists (2011 update). *Nucleic Acids Res* **39:** W307–W315

Reimand J, Tooming L, Peterson H, Adler P, Vilo J (2008) GraphWeb: mining heterogeneous biological networks for gene modules with functional significance. *Nucleic Acids Res* **36:** W452–W459

Reimand J, Hui S, Jain S, Law B, Bader GD (2012) Domain-mediated protein interaction prediction: From genome to network. *FEBS Lett* **586:** 2751–2763

Robinson PN, Mundlos S (2010) The human phenotype ontology. *Clin Genet* **77:** 525–534

Ruepp A, Waegele B, Lechner M, Brauner B, Dunger-Kaltenbach I, Fobo G, Frishman G, Montrone C, Mewes HW (2010) CORUM: the comprehensive resource of mammalian protein complexes—2009. *Nucleic Acids Res* **38:** 497–501

Ruse CI, McClatchy DB, Lu B, Cociorva D, Motoyama A, Park SK, Yates JR (2008) Motif-specific sampling of phosphoproteomes. *J Proteome Res* **7:** 2140–2150

Sawyer GM, Clark AR, Robertson SP, Sutherland-Smith AJ (2009) Disease-associated substitutions in the filamin B actin binding domain confer enhanced actin binding affinity in the absence of major structural disturbance: insights from the crystal structures of filamin B actin binding domains. *J Mol Biol* **390:** 1030–1047

Schild-Poulter C, Shih A, Tantin D, Yarymowich NC, Soubeyrand S, Sharp PA, Hache RJ (2007) DNA-PK phosphorylation sites on Oct-1 promote cell survival following DNA damage. *Oncogene* **26:** 3980–3988

Segil N, Roberts SB, Heintz N (1991) Mitotic phosphorylation of the Oct-1 homeodomain and regulation of Oct-1 DNA binding activity. *Science* **254:** 1814–1816

Sharma SV, Bell DW, Settleman J, Haber DA (2007) Epidermal growth factor receptor mutations in lung cancer. *Nat Rev Cancer* **7:** 169–181

Sjoblom T, Jones S, Wood LD, Parsons DW, Lin J, Barber TD, Mandelker D, Leary RJ, Ptak J, Silliman N, Szabo S, Buckhaults P, Farrell C, Meeh P, Markowitz SD, Willis J, Dawson D, Willson JK, Gazdar AF, Hartigan J *et al* (2006) The consensus coding sequences of human breast and colorectal cancers. *Science* **314:** 268–274

Tagliabracci VS, Engel JL, Wen J, Wiley SE, Worby CA, Kinch LN, Xiao J, Grishin NV, Dixon JE (2012) Secreted kinase phosphorylates extracellular proteins that regulate biomineralization. *Science* **336:** 1150–1153

Tan CS, Bodenmiller B, Pasculescu A, Jovanovic M, Hengartner MO, Jorgensen C, Bader GD, Aebersold R, Pawson T, Linding R (2009) Comparative analysis reveals conserved protein phosphorylation networks implicated in multiple diseases. *Sci Signal* **2:** ra39

The International Cancer Genome Consortium (2010) International network of cancer genome projects. *Nature* **464:** 993–998

Tiacci E, Trifonov V, Schiavoni G, Holmes A, Kern W, Martelli MP, Pucciarini A, Bigerna B, Pacini R, Wells VA, Sportoletti P, Pettirossi V, Mannucci R, Elliott O, Liso A, Ambrosetti A, Pulsoni A, Forconi F, Trentin L, Semenzato G *et al* (2011) BRAF mutations in hairy-cell leukemia. *N Engl J Med* **364:** 2305–2315

Totoki Y, Tatsuno K, Yamamoto S, Arai Y, Hosoda F, Ishikawa S, Tsutsumi S, Sonoda K, Totsuka H, Shirakihara T, Sakamoto H, Wang L, Ojima H, Shimada K, Kosuge T, Okusaka T, Kato K, Kusuda J, Yoshida T, Aburatani H *et al* (2011) High-resolution characterization of a hepatocellular carcinoma genome. *Nat Genet* **43:** 464–469

Van Hoof D, Munoz J, Braam SR, Pinkse MW, Linding R, Heck AJ, Mummery CL, Krijgsveld J (2009) Phosphorylation dynamics during early differentiation of human embryonic stem cells. *Cell Stem Cell* **5:** 214–226

van Noort M, Meeldijk J, van der Zee R, Destree O, Clevers H (2002) Wnt signaling controls the phosphorylation status of beta-catenin. *J Biol Chem* **277:** 17901–17905

Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM (2009) A census of human transcription factors: function, expression and evolution. *Nat Rev Genet* **10:** 252–263

Verhaak RG, Goudswaard CS, van Putten W, Bijl MA, Sanders MA, Hugens W, Uitterlinden AG, Erpelinck CA, Delwel R, Lowenberg B, Valk PJ (2005) Mutations in nucleophosmin (NPM1) in acute myeloid leukemia (AML): association with other gene abnormalities and previously established gene expression signatures and their favorable prognostic significance. *Blood* **106:** 3747–3754

Vogelstein B, Kinzler KW (2004) Cancer genes and the pathways they control. *Nat Med* **10:** 789–799

Wang X, Wei X, Thijssen B, Das J, Lipkin SM, Yu H (2012) Three-dimensional reconstruction of protein networks provides insight into human genetic disease. *Nat Biotechnol* **30:** 159–164

Ward JJ, McGuffin LJ, Bryson K, Buxton BF, Jones DT (2004) The DISOPRED server for the prediction of protein disorder. *Bioinformatics* **20:** 2138–2139

Wood LD, Parsons DW, Jones S, Lin J, Sjoblom T, Leary RJ, Shen D, Boca SM, Barber T, Ptak J, Silliman N, Szabo S, Dezso Z, Ustyanksky V, Nikolskaya T, Nikolsky Y, Karchin R, Wilson PA, Kaminker JS, Zhang Z *et al* (2007) The genomic landscapes of human breast and colorectal cancers. *Science* **318:** 1108–1113

Wu L, Ma CA, Zhao Y, Jain A (2011) Aurora B interacts with NIR-p53, leading to p53 phosphorylation in its DNA-binding domain and subsequent functional suppression. *J Biol Chem* **286:** 2236–2244

Wysocka J, Myers MP, Laherty CD, Eisenman RN, Herr W (2003) Human Sin3 deacetylase and trithorax-related Set1/Ash2 histone H3-K4 methyltransferase are tethered together selectively by the cell-proliferation factor HCF-1. *Genes Dev* **17:** 896–911