nature genetics

A multiscale statistical mechanical framework integrates biophysical and genomic data to assemble cancer networks

Mohammed AlQuraishi[1–3], Grigoriy Koytiger[1,3], Anne Jenney[1], Gavin MacBeath[2] & Peter K Sorger[1]

Functional interpretation of genomic variation is critical to understanding human disease, but it remains difficult to predict the effects of specific mutations on protein interaction networks and the phenotypes they regulate. We describe an analytical framework based on multiscale statistical mechanics that integrates genomic and biophysical data to model the human SH2-phosphoprotein network in normal and cancer cells. We apply our approach to data in The Cancer Genome Atlas (TCGA) and test model predictions experimentally. We find that mutations mapping to phosphoproteins often create new interactions but that mutations altering SH2 domains result almost exclusively in loss of interactions. Some of these mutations eliminate all interactions, but many cause more selective loss, thereby rewiring specific edges in highly connected subnetworks. Moreover, idiosyncratic mutations appear to be as functionally consequential as recurrent mutations. By synthesizing genomic, structural and biochemical data, our framework represents a new approach to the interpretation of genetic variation.

TCGA and similar projects have generated extensive data on the mutational landscape of tumors[1]. To understand the functional consequences of these mutations, it is necessary to ascertain how they alter the protein-protein interaction (PPI) networks involved in regulating cellular phenotypes. A wide spectrum of data are available on PPIs, ranging from large-scale binding experiments[2–4] to co-crystal studies. The interpretation of such data is hampered by the absence of an analytical framework for integrating diverse measurements and for modeling the effects of cancer mutations. Such a framework must jointly model the genetic heterogeneity of cancer and the biophysical determinants of PPI specificity at the level of individual protein domains, multidomain proteins and PPI networks. In this report, we describe such a multiscale statistical mechanical (MSM) framework focusing on the subset of PPIs involving protein interaction domains (PIDs) in which SH2 domains bind to phosphotyrosine peptides. Such interactions are essential components of receptor-mediated signaling, and their misregulation is known to have a role in cancer and other diseases[5–8].

The first challenge in modeling PID networks is integrating data from diverse low-throughput and high-throughput assays. Low-throughput methods such as fluorescence polarization spectroscopy provide precise interaction data on a few dozen PIDs and ligands but cannot easily be scaled to the full proteome[2], whereas high-throughput array-based methods provide greater scale but suffer from systematic artifacts and high false positive and false negative rates, resulting in data sets that only partly agree[2]. An additional challenge is modeling the effects of mutations for proteins with multiple binding domains and/or multiple sites of phosphorylation[9], a reality for most signaling proteins (for example, the CRK oncoprotein). Existing methods are either limited to individual domains[10–13] or are insufficiently precise to discern the effects of single-residue changes[14,15].

The MSM framework we have developed combines genomic, binding and structural data and reconciles inconsistencies within and among data sets to generate PID networks for normal and cancer cells. We develop a bottom-up first-principles approach, involving a single mathematical equation based on statistical mechanical ensembles, that models domains, proteins and networks, and we then apply this approach to the analysis of SH2 networks and mutations found in TCGA[16]. We validate newly predicted interactions experimentally and demonstrate the sensitivity of MSM to single-residue mutations that cause subtle changes in binding affinity. Our analysis provides mechanistic insights into an important PID cancer network and validates a computational approach to PID networks that can be applied to other signaling domains and other diseases.

RESULTS

Modeling and data integration

The theory of statistical mechanics relates the energy of a state of a system, such as a particular configuration of bound and free peptides and PIDs, to measurable thermodynamic quantities, such as disassociation constants (**Fig. 1a**); from the energy, it is also possible to compute the probability that a system will assume a particular state. When the function specifying the energy of a state—known as the Hamiltonian—is available, the mathematics of statistical mechanics can be used to directly compute thermodynamic properties (**Fig. 1b**). However, with complex systems such as proteins in solution, the Hamiltonian cannot be readily derived from basic physical principles, such as quantum mechanics. In this work, we recast statistical mechanics

[1]Harvard Medical School Laboratory of Systems Pharmacology, Harvard Medical School, Boston, Massachusetts, USA. [2]Department of Systems Biology, Harvard Medical School, Boston, Massachusetts, USA. [3]These authors contributed equally to this work. Correspondence should be addressed to P.K.S. (peter_sorger@hms.harvard.edu) or M.A. (alquraishi@hms.harvard.edu).

Received 4 June; accepted 9 October; published online 2 November 2014; doi:10.1038/ng.3138

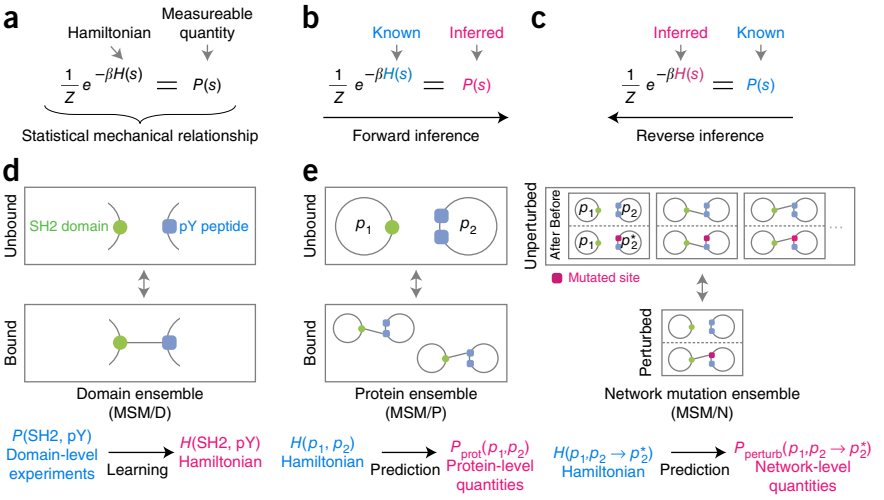

Figure 1 Multiscale statistical mechanical (MSM) framework. (**a**) Statistical mechanics establishes mathematical relationships between the energy of a state *s* of a system, known as the Hamiltonian $H(s)$, and measurable thermodynamic quantities of that state, such as its probability of occurrence $P(s)$. (**b**) In simple physical systems, the Hamiltonian is known, and the mathematics of statistical mechanics can be directly used to infer thermodynamic quantities. (**c**) Experimental data on thermodynamic quantities can be used in the reverse direction to infer the Hamiltonian (more precisely, a pseudo-Hamiltonian) using machine learning techniques. (**d**) In MSM, learning of the Hamiltonian is performed at the single-domain level (MSM/D), by creating ensembles that correspond to bound and unbound SH2-phosphotyrosine (pY) complexes. (**e**) The learned Hamiltonian can be used to make predictions for more complex ensembles. At the whole-protein level (MSM/P), ensembles comprise all physical binding configurations, accounting for the combinatorics of multiple domains and multiple phosphorylation sites. At the network and mutation level (MSM/N), ensembles comprise states that simultaneously represent the behavior of the PPI before and after a mutation is introduced. This selectively captures mutations that result in consequential changes to binding affinity (see **Supplementary Fig. 1** and the **Supplementary Note** for further details).

as a machine learning problem and develop a reverse workflow in which measurements of thermodynamic quantities are used to derive a Hamiltonian for SH2 domain and phosphotyrosine peptide interactions[17] (**Fig. 1c**). Notably, statistical mechanics does not fully constrain the mathematical form of the Hamiltonian or the set of states—known as the ensemble—whose thermodynamic properties are being computed. We exploit this fact by choosing a form for the Hamiltonian that can model arbitrary SH2-peptide interactions, including for mutated domains and unknown phosphotyrosine peptides. This Hamiltonian is defined in terms of interactions between one residue in the peptide and one residue in the SH2 domain, ignoring multi-residue interactions and steric effects between residues in the same protein. Thus, it is more appropriately termed a pseudo-Hamiltonian.

We also exploited freedom in selecting ensembles. When the pseudo-Hamiltonian is learned, an ensemble comprises the bound and unbound states of an SH2 domain and a single phosphotyrosine peptide (**Fig. 1d**). Such domain-level ensembles correspond to readily available thermodynamic measurements, and we can therefore use a wide variety of low- and high-throughput data in inferring the pseudo-Hamiltonian (for a mathematical treatment, see the **Supplementary Note** and **Supplementary Figs. 1** and **2**). Subsequently, we used the pseudo-Hamiltonian to compute difficult-to-obtain thermodynamic quantities for ensembles of multidomain proteins and multiply phosphorylated substrates (**Fig. 1e**). In principle, data on isolated domains, multidomain proteins and multiply phosphorylated substrates can be used in learning. In practice, virtually all experimental data relate to single domain–phosphotyrosine peptide interactions. Thus, ensemble theory allows us to circumvent limitations in available data to predict interactions involving the types of proteins that are actually found in signaling networks.

To perform learning, we combined binding data from four high-throughput data sets[18–21] ('MacBeath', 'Jones', 'Nash' and 'Cesareni') and a diverse set of low-throughput data[21] (**Supplementary Table 1**). The combined data set spanned 111 SH2 domains (out of 122 known; no data were available for 11 domains) and 5,016 phosphotyrosine peptide sequences (the human proteome is estimated to contain ~37,000 phosphotyrosine sites lying in ~12,000 proteins[22]). Data were binarized, yielding a training set of ~20,000 positive and ~400,000 negative interactions. From this training set, we learned the residue-residue energies of the pseudo-Hamiltonian by maximizing the agreement between the predicted and experimental affinities of domains. We also aligned SH2 domains with 25 SH2-peptide co-crystal structures and exploited the resulting spatial information by applying the well-established principle[23] that selectivity is primarily determined by interacting residues in direct physical contact. Mathematically, this principle was imposed by penalizing energy terms in proportion to the distance separating residues, thereby assigning weaker energies to more distant interactions.

Domain model achieves high accuracy and reconciles data sets

We evaluated the performance of the MSM domain model (MSM/D) using a nested cross-validation approach designed to prevent overfitting. Data were randomly divided into three parts: one for fitting residue interaction energies, one for fitting distance-dependent energy penalization and one for model tests (**Supplementary Fig. 3a**). We compared MSM/D to two existing SH2 models (SMALI[10] and PEPINT[11]) and to a general PPI method, PrePPI[14], using a receiver operator characteristic (ROC) curve (**Fig. 2a**). Global and data set–specific performance was evaluated by grouping high-throughput data into four subsets on the basis of their source. We also subdivided data by combining the low-throughput data set with high-confidence interactions (i.e., those confirmed by two (HC2) and three (HC3) data sources). Data for high-confidence interactions were removed from the high-throughput data sets to prevent cross-contamination during training.

We observed that MSM/D substantially outperformed existing methods in the 'high-precision' regime (false positive rate (FPR) = 0.001), yielding a true positive rate (TPR) that was 6 times higher than that for the best existing method across all data and >50 times higher for the HC2 subset (**Fig. 2a**). Integrating the area under the ROC curve (AUC) yields a single performance number, and this too was substantially higher for MSM/D than existing models on all subsets of the data (**Table 1**). Relative performance was particularly high for 'gold standard' subsets, with MSM/D achieving a nearly perfect score of 0.99 AUC on the HC3 data set. The superiority of MSM/D relative to PrePPI does not take into account the latter's ability to model any protein; we include the comparison only to establish a baseline for general PPI methods. In addition, SMALI and PEPINT were trained on data sets only about one-third as large as that used for MSM/D; we therefore retrained MSM/D

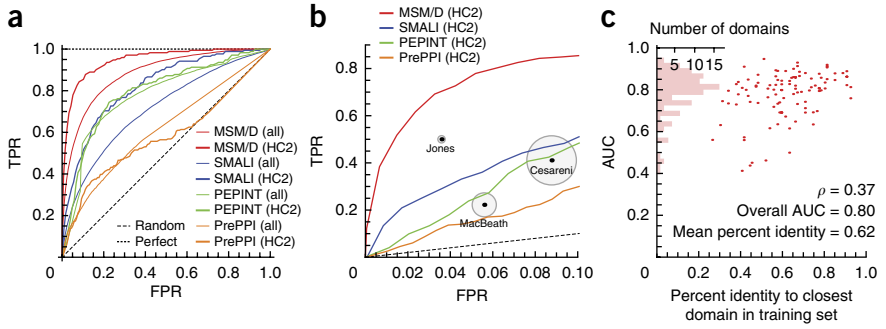

Figure 2 Assessment of domain model (MSM/D) performance. (**a**) ROC curves assessing the performance of MSM/D and other methods (SMALI, PEPINT and PrePPI) in predicting the binding states of SH2-phosphopeptide interactions. ROC curves characterize a model's ability to predict SH2-phosphotyrosine interactions by computing the TPR of predictions as a function of the FPR. A method that makes random guesses will produce a straight line with a slope of 1 (dashed black line), whereas a perfect method produces a constant TPR value of 1 (dotted black line). Tests were performed on the combined data set (all) and a high-confidence subset (HC2). (**b**) A close-up view of **a**, showing the relative performance of high-throughput data sets. Circle size indicates the relative size of the data set. (**c**) The AUCs of MSM/D in predicting held-out SH2 domains are plotted as a function of each domain's sequence identity to its closest homolog in the training set. A histogram of AUC values is overlaid on the *y* axis.

using approximately one-third of the available data (**Supplementary Fig. 3e**) and found that the retrained model attained ~97% of the maximum AUC, which remained substantially superior to the SMALI and PEPINT approaches. We also computed separate ROC and metaparameter sensitivity curves for each cross-validation set (**Supplementary Fig. 3b**,**c**). The curves overlapped almost perfectly, indicating that MSM/D is robust to variation in training data. We conclude that MSM/D is substantially better than available methods at modeling interactions between SH2 domains and phosphotyrosine peptides.

Available data sets for SH2-peptide interaction only partially agree[2,4,24], likely owing to systematic biases and high numbers of false negative and false positive errors. We therefore tested the ability of MSM/D to reconcile disagreement in the experimental record. FPR and TPR were estimated for three high-throughput assays (**Fig. 2b**) through comparison with HC3 data. We found that, at all FPR levels, MSM/D exhibited higher precision and recall than any experimental data set. To test the ability of MSM/D to integrate diverse data, we left out one of the five data sets during training and then tested the model against the excluded data set. MSM/D performance remained high against high-confidence data sets (for example, AUC of 0.938 versus 0.947 on HC2) but dropped when predictions were tested on the high-throughput data set excluded during training (**Table 2**). We interpret this decrease as arising from systematic error in the excluded data that cannot be modeled a priori; however, we cannot exclude impact from the uneven coverage of sequence space. Our results nonetheless show that MSM/D can correct for random and systematic experimental errors to generate a consolidated representation of SH2-phosphotyrosine interactions that is superior to any single data set or simple polling strategies.

To determine the ability of MSM/D to model SH2 domains absent from the training set, we retrained the model on data from which one domain had been excluded and then computed the AUC for the excluded domain; the process was iterated over all SH2 domains. In **Figure 2c**, we plot AUCs as a function of the sequence identity between the excluded domain and its nearest included neighbor. MSM/D modeled excluded SH2 domains with an AUC of ~0.8, even when sequence identity to the nearest neighbor averaged ~62%. In contrast, prediction of unknown domains cannot be performed using PEPINT and SMALI. When we left out ~13 SH2 domains at a time, reducing nearest neighbor sequence identity to ~45% (**Supplementary Fig. 3d**), the AUC dropped to ~0.75, but the model retained substantial predictive capability. We also examined the properties of the 11 SH2 domains for which no experimental data were available and found binding selectivity comparable to that of other SH2 domains but ~50-fold fewer binding partners (at a threshold of $P > 0.85$). Thus, the absence of data for these 11 domains likely reflects the low probability of observing an interaction experimentally[18]. We conclude that MSM/D can model unseen SH2 domains and peptides, including those altered by cancer mutations.

Table 1 Performance breakdown (AUCs)

Model	Overall	MacBeath	Jones	Nash	Cesareni	LT + HC2	HC2	HC3
SMALI	0.713	0.663	0.594	0.709	0.714	0.820	0.821	0.890
PEPINT	0.777	0.627	0.629	0.701	0.793	0.761	0.795	0.899
PrePPI	0.615	0.580	0.576	0.586	0.584	0.708	0.580	0.738
MSM/D	0.882	0.762	0.730	0.769	0.896	0.887	0.947	0.991

LT, low throughput; HC, high confidence.

Protein and mutation models capture multi-site interactions

Interactions between SH2- and phosphotyrosine-containing proteins involve tandem domains and phosphoproteins with up to 25 phosphosites. A majority (82/112) of SH2-containing proteins also contain one or more phosphotyrosine sites. To model such interactions, we constructed statistical mechanical ensembles comprising all physical binding configurations, accounting for the combinatorics of multiple domains and phosphosites (**Fig. 1e** and **Supplementary Fig. 1**), to yield the MSM protein model (MSM/P). Competitive binding between sites is accounted for by the ensemble formulation. By default, MSM/P assumes that all phosphosites are phosphorylated and interacting proteins are equally expressed, but this assumption is unlikely to pertain to actual cells and can be relaxed.

To validate MSM/P experimentally, we focused on GCSAM, a protein previously implicated in B cell lymphoma[25,26] for which a large discrepancy exists between the number of published interacting SH2 partners (two are known: GRB2 (ref. 27) and SYK[28]) and the number predicted by MSM/P (nine more are predicted with affinities comparable to that of GRB2). We coexpressed GCSAM fused to the monomeric red fluorescent protein TagRFP and 1 of 12 SH2 domains tagged with GFP. HEK293T cells were then treated with pervanadate for 5 min to promote tyrosine phosphorylation, lysates were prepared and SH2-containing complexes were immunoprecipitated using anti-GFP beads. Fluorescent imaging of the beads and the supernatant made it possible to normalize the level of bound SH2 domain to the total levels of SH2 and GCSAM expression, resulting in excellent reproducibility between biological replicates (ρ = 0.99; **Fig. 3a**). Using the bead-based assay, we detected binding by all SH2 domains predicted by MSM/P to associate with GCSAM, and the correlation

Table 2 Data set transferability (AUCs)

Data set	Transfer
Overall	0.771
MacBeath	0.671
Jones	0.737
Nash	0.711
Cesareni	0.707
LT + HC2	0.836
HC2	0.938
HC3	0.971

AUCs were computed by withholding an entire data set from the training set and testing the performance of MSM/D exclusively on the excluded data set. LT, HC2 and HC3 were treated as a single data set.

between measured and predicted affinities was high ($\rho = 0.80$; **Fig. 3a**). This is a stringent test of MSM/P as (i) the evaluation of agreement is quantitative and spans a broad range of affinities, (ii) the SH2 domains we tested had diverse primary sequences and (iii) GCSAM contains three phosphotyrosine sites, thereby testing the performance of the protein-level ensemble model.

Next, we modeled the effects of mutations on SH2-phosphotyrosine interactions by constructing ensembles whose states simultaneously represented the behavior of the PPI before and after mutation (**Fig. 1e** and **Supplementary Fig. 1**). The resulting model distinguishes between consequential and non-consequential changes in binding affinity (**Supplementary Fig. 2**) and accounts for the 'protein context' of a mutation, including buffering effects from other phosphosites or domains present in the same protein. To experimentally evaluate the sensitivity of MSM/P to single-amino-acid substitutions, we analyzed binding of the regulatory α subunit of phosphatidylinositol 3-kinase, PIK3R1, to a mutant of the insulin-like growth factor receptor, IGF1R Ala1347Val (COSM12856; encoded by a mutation in squamous cell carcinoma[29]). This represents a stringent test of the model because (i) it does not create a predictable canonical motif such as pYXXM (where pY represents the phosphotyrosine); (ii) the mutation occurs in the complex protein context of IGF1R, which has 11 phosphosites[22]; (iii) we predict a gain-of-function increase in affinity rather than a more common loss-of-function decrease; and (iv) PIK3R1 contains 2 SH2 domains. We coexpressed GFP-tagged IGF1R with a construct for mCherry linked to either the N- or C-terminal PIK3R1 SH2 domain or a construct encoding both domains and performed quantitative coimmunoprecipitation using anti-RFP

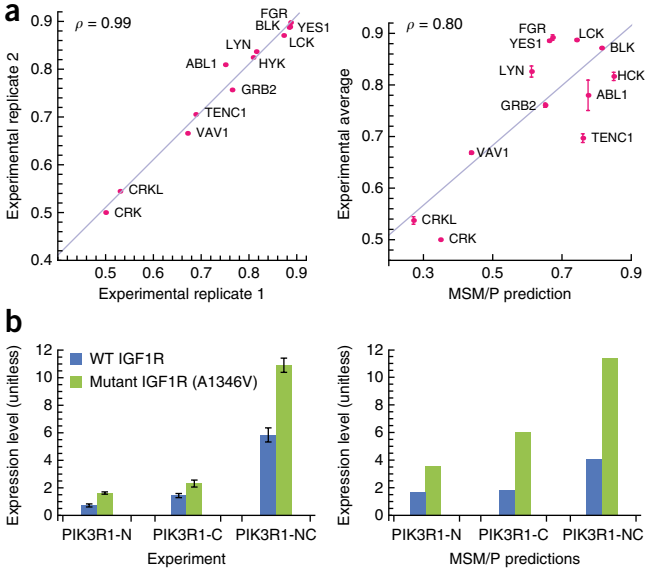

Figure 3 Experimental validation of protein-level interactions for wild-type and mutant proteins. (**a**) Quantitative coimmunoprecipitation signals of GCSAM to partner proteins show excellent experimental reproducibility ($\rho = 0.99$) and a high correlation with MSM/P predictions ($\rho = 0.80$). Error bars represent the standard error of two biological replicates. (**b**) Ala1346Val-mutated IGF1R exhibits higher affinity for the PIK3R1 N-terminal (PIK3R1-N), C-terminal (PIK3R1-C) and tandem N- and C-terminal (PIK3R1-NC) SH2 domains ($P = 6.2 \times 10^{-5}$, $P = 9.0 \times 10^{-3}$ and $P = 5.9 \times 10^{-5}$, respectively, using one-sided *t* tests) as predicted by MSM/P. Error bars represent the standard error of five biological replicates.

beads. As predicted, IGF1R Ala1347Val exhibited stronger binding to the SH2 domains of PIK3R1 than the wild-type protein (one-sided *t* test for five biological replicates of $P = 6.2 \times 10^{-5}$ for the N-terminal domain, $P = 9.0 \times 10^{-3}$ for the C-terminal domain and $P = 5.9 \times 10^{-5}$ for the tandem protein; **Fig. 3b**). The increase in affinity was correctly predicted to stem from stronger binding of both domains to mutant receptor, with the C-terminal domain being the stronger of the two ($P = 0.016$). The effect is potentially biologically relevant, as PIK3R1 and IGF1R are oncoproteins that interact through the adaptor IRS1. We conclude that MSM/P is effective at capturing the effects of mutations on SH2-phosphotyrosine binding affinity. However, the

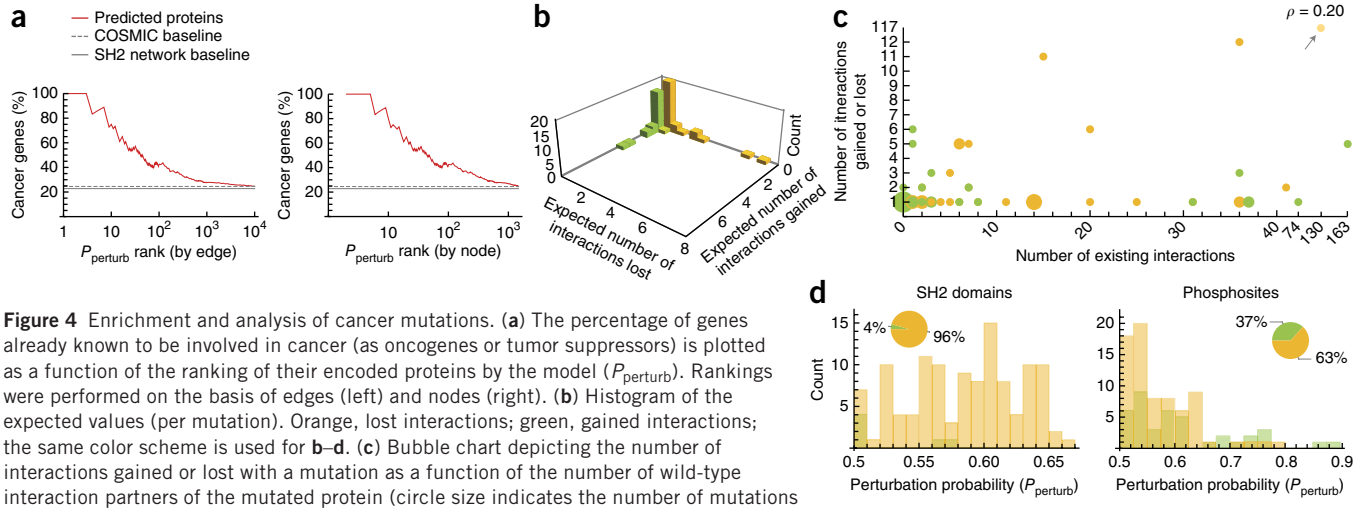

Figure 4 Enrichment and analysis of cancer mutations. (**a**) The percentage of genes already known to be involved in cancer (as oncogenes or tumor suppressors) is plotted as a function of the ranking of their encoded proteins by the model ($P_{perturb}$). Rankings were performed on the basis of edges (left) and nodes (right). (**b**) Histogram of the expected values (per mutation). Orange, lost interactions; green, gained interactions; the same color scheme is used for **b**–**d**. (**c**) Bubble chart depicting the number of interactions gained or lost with a mutation as a function of the number of wild-type interaction partners of the mutated protein (circle size indicates the number of mutations with the same profile). One mutation was removed when calculating correlation (indicated by arrow). (**d**) Distributions of $P_{perturb}$ values for SH2 proteins and phosphoproteins broken down by gain-of-function and loss-of-function mutations.

data also highlight the limitation that MSM/P does not account for non-additive avidity effects in tandem domains and therefore underestimates the affinity of PIK3R1 for wild-type IGF1R.

Cancer network model enriches for causal cancer genes

Cancer cells typically contain many mutations, a subset of which directly promote the transformed phenotype (driver mutations), making it necessary to model the net effect of multiple mutations on a network. We formally treated cancer as a stochastic sampling process in which any given genotype is realized by random draws from a pool of mutations. This simplification ignores the sequential and interdependent accrual of cancer mutations because of insufficient data on these dependences, but it can be relaxed as data become available. Probabilities for mutations can be derived empirically from unbiased whole-genome databanks such as TCGA[16]. We constructed an ensemble over all mutations and associated a perturbed PPI network with each state in this ensemble. The resulting cancer network model (MSM/N) assigned to every potential PPI a quantity, P_{perturb}, defined as the probability that the given PPI would be disrupted or activated by a randomly drawn mutation (**Fig. 1e** and **Supplementary Fig. 1**). P_{perturb} integrates information at the level of domains, proteins and networks to model the impact of mutations affecting multiple proteins and their disease-specific frequencies. P_{perturb} is central to our approach and rigorously captures the concept of a mutation that is causally responsible for a qualitative change in PPI behavior.

We first used MSM/P to reconstruct the human SH2-phosphoprotein network from first principles using primary sequence data on SH2 domains and 2,292 phosphoproteins, about half of which contain multiple phosphosites (**Supplementary Figs. 4** and **5**, and **Supplementary Table 2**). We then obtained all whole-genome tumor sequences from the Catalogue of Somatic Mutations in Cancer (COSMIC)[16], filtered to include mutations mapping to SH2 domains and residues proximate to confirmed sites of tyrosine phosphorylation

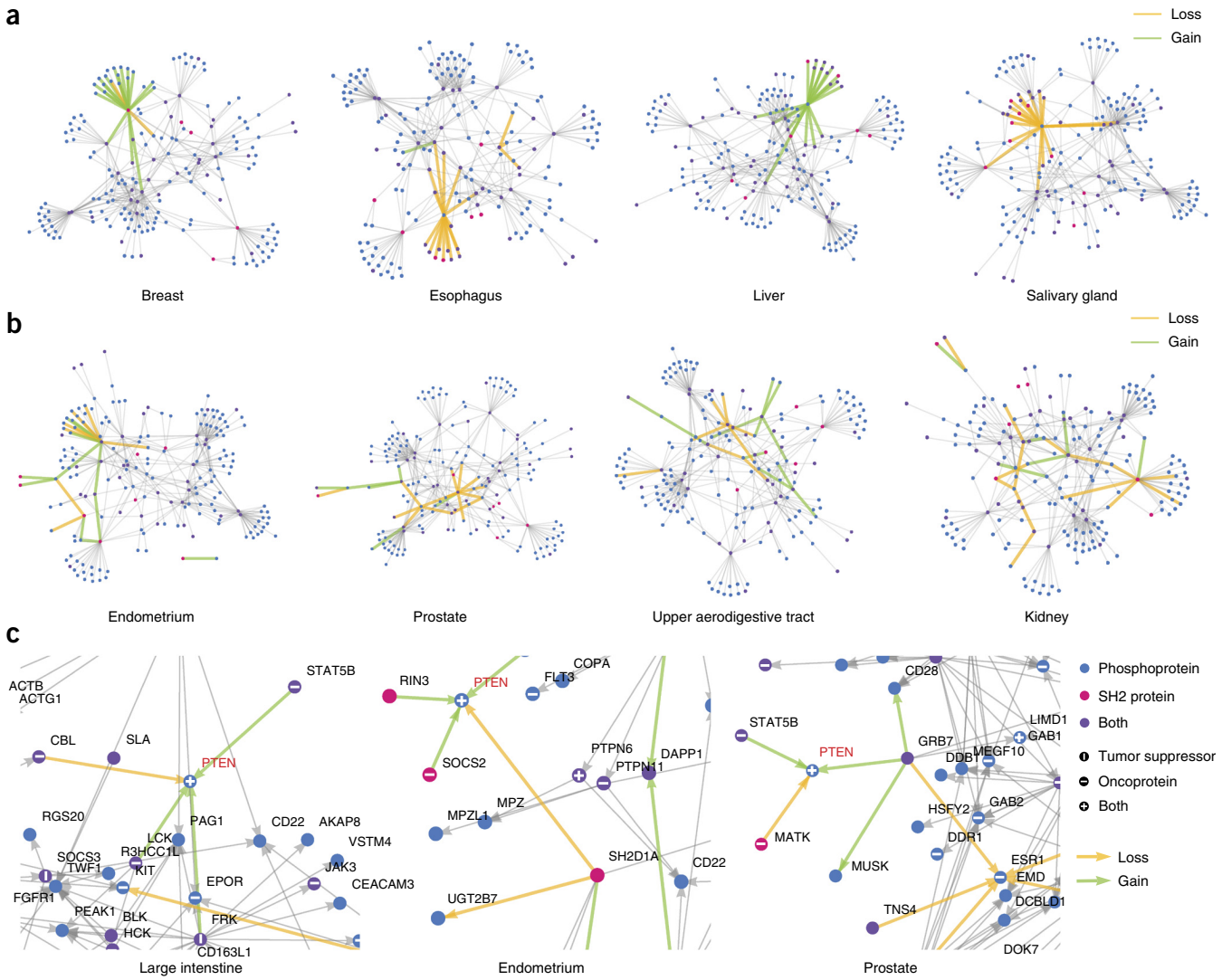

Figure 5 Tissue-specific tumor networks. (**a**) MSM/N predictions of the top 20 interactions gained and lost (green and orange edges, respectively) in 4 tumor networks overlaid on the wild-type SH2 phosphosignaling network (gray edges, each representing an interaction with $P > 0.85$ probability, as in **Supplementary Fig. 4**), showing a bias for the node mode of perturbations. (**b**) Four tumor networks that show a bias for the pathway mode of perturbations. (**c**) Local neighborhoods of the PTEN network in different cancer tissue types. All networks were generated using a spring-electrical embedding in the Mathematica software package.

(i.e., those with two or more sources of experimental support)[22]. This yielded 807 mutations in genes encoding SH2-containing proteins and 4,648 mutations in sequences encoding phosphoproteins across 24 tissue types and 2,206 tumor samples. We pooled all mutations and used MSM/N to derive $P_{perturb}$ values for every PPI (**Supplementary Fig. 6**, **Supplementary Table 3** and **Supplementary Data Sets 1–25**). The resulting tumor network exhibited strong enrichment for cancer-associated genes (**Fig. 4a**). The percentage of SH2-containing proteins or phosphoproteins annotated as oncoproteins or tumor suppressors was ~23% (by TSGene[30] and allOnco) and increased somewhat (to ~25%) when we considered only mutated SH2-containing proteins or phosphoproteins listed in COSMIC. In contrast, cancer gene enrichment increased to 75% for the top 10 interactors and 43% for the top 100 when scored by $P_{perturb}$ (COSMIC and $P_{perturb}$ did not reach parity until ~10,000 PPIs were included). We conclude that COSMIC mutations with high $P_{perturb}$ values are strongly associated with cancer-relevant genes.

Idiosyncratic mutations rewire SH2 signaling networks

Of 5,455 COSMIC mutations that occurred in SH2 proteins or phosphoproteins, 4,254 (78%) were idiosyncratic, occurring in only a single sample. However, we found that idiosyncratic mutations were as likely to rewire PPIs as recurrent mutations. Of 419 recurrent mutations, 23 (5.5%) were predicted to rewire PPIs at >33% probability and 5 (1.2%) were predicted to do so at >50% probability. Of 4,254 idiosyncratic mutations, 262 (6.2%) were predicted to rewire PPIs at >33% probability and 47 (1.1%) were predicted to do so at >50% probability. These results imply that tumor mutations should be analyzed with respect to function rather than frequency alone. MSM is one way to detect potentially functional, non-recurrent mutations.

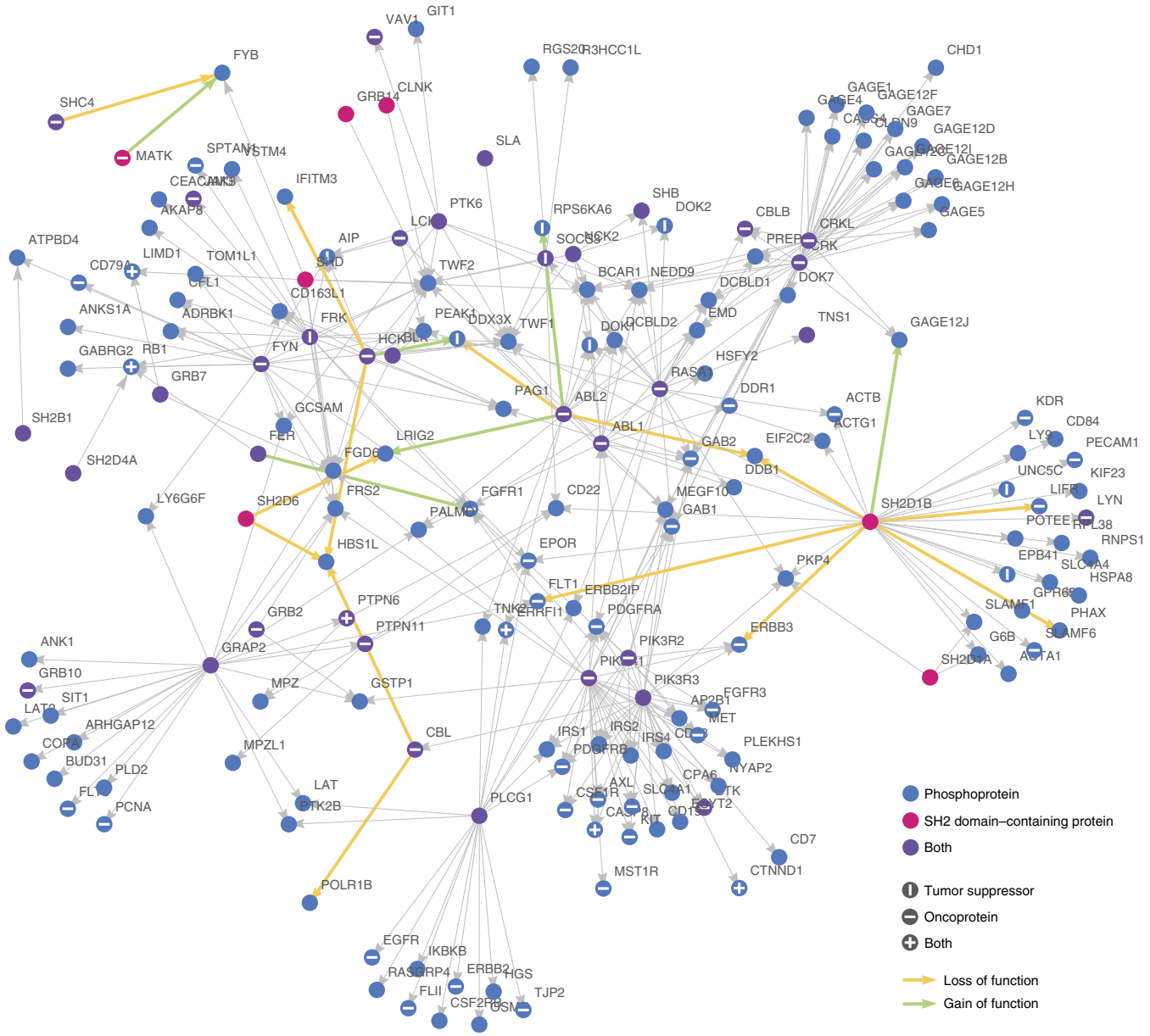

Figure 6 Kidney tumor network. MSM/N predictions of the top 20 perturbed interactions (green and yellow arrows) in kidney cancer overlaid on the wild-type SH2 phosphosignaling network (gray edges, each representing an interaction with $P > 0.85$ probability, as in **Supplementary Fig. 4**). Networks were generated using a spring-electrical embedding in the Mathematica software package.

Most cancer mutations disrupt individual PPIs

Tumorigenic mutations commonly affect enzymatic function by inactivating a tumor suppressor such as PTEN or constitutively activating an oncoprotein such as PI3 kinase[31]. Some mutations may function by selectively rewiring PPIs[6]. We therefore analyzed MSM/N tumor networks to identify mutations that selectively disrupted single high-affinity PPIs while leaving intact higher- and lower-affinity interactions mediated by the same mutated protein. We found that the majority (69%) of strong cancer mutations ($P > 0.5$) targeted a single PPI (**Fig. 4b**). This appeared to hold true irrespective of the number of interactions mediated by the wild-type protein ($\rho = 0.2$ with one outlier interaction removed; **Fig. 4c**). One exception was a mutation mapping to the GRAP2 SH2 domain that changed a tryptophan to a cysteine at a critical residue and was predicted to disrupt 117 of 130 interactions. Mutating a homologous tryptophan in the related GRB2 protein has been shown to similarly abolish its ability to bind phosphotyrosine peptides[32].

We also observed a difference in the predicted effect of cancer mutations on SH2-containing proteins and phosphoproteins (**Fig. 4d**). The vast majority of strong mutations (95%) targeted phosphoproteins and resulted in both gain (37%) and loss (63%) of interactions. Conversely, SH2 mutations led almost universally to loss of interactions (96%). The strength of the perturbational effect also differed. It took, on average, ~100 draws from the pool of phosphoprotein mutations to disrupt an interaction but only ~10 draws from the pool of SH2 mutations. For gain-of-function mutations, it took, on average, ~250 draws for phosphoproteins and ~200 draws for SH2 proteins.

Two modes of network rewiring by cancer mutations

To identify associations between cancer tissue and changes in SH2 networks, we examined the 20 most strongly disrupted PPIs, as ranked by $P_{perturb}$, across 24 cancer tissues of origin. We observed two modes of action: 'node' and 'pathway'. These modes were not mutually exclusive and occurred in combination. In node cases (for example, breast and liver cancers), one or more mutations targeted a single protein and obliterated all its interactions or resulted in a gain-of-function loss of selectivity (**Fig. 5a**). This mode of disruption is analogous to enzymatic mutations and might be therapeutically addressable with drugs that target a single protein. In the pathway mode, multiple mutations disrupted PPIs forming a connected path within a network (**Figs. 5b** and **6**). Random targeting of such connected paths is highly improbable, as each mutation can affect ~250,000 edges, suggesting that selection may be exerted at the level of the signaling pathway. Therapeutic intervention in such cases might require agents that restore pathway-level function. We also observed differential targeting of PPIs involving the same protein across different tissues. For example, the tumor suppressor PTEN was predicted to gain or lose distinct interactions in cancers of the large intestine, endometrium and prostate (**Fig. 5c**), and the affected proteins were pertinent to the tumor type. For instance, in prostate cancer, TNS1, TNS4, BCAR1 and RAC1 regulate cellular motility and invasiveness[33,34] and the estrogen receptor ESR1 enhances proliferation[35].

DISCUSSION

In this report, we attempt to advance the state of the art in functional genomics by developing an analytical framework for reconstructing

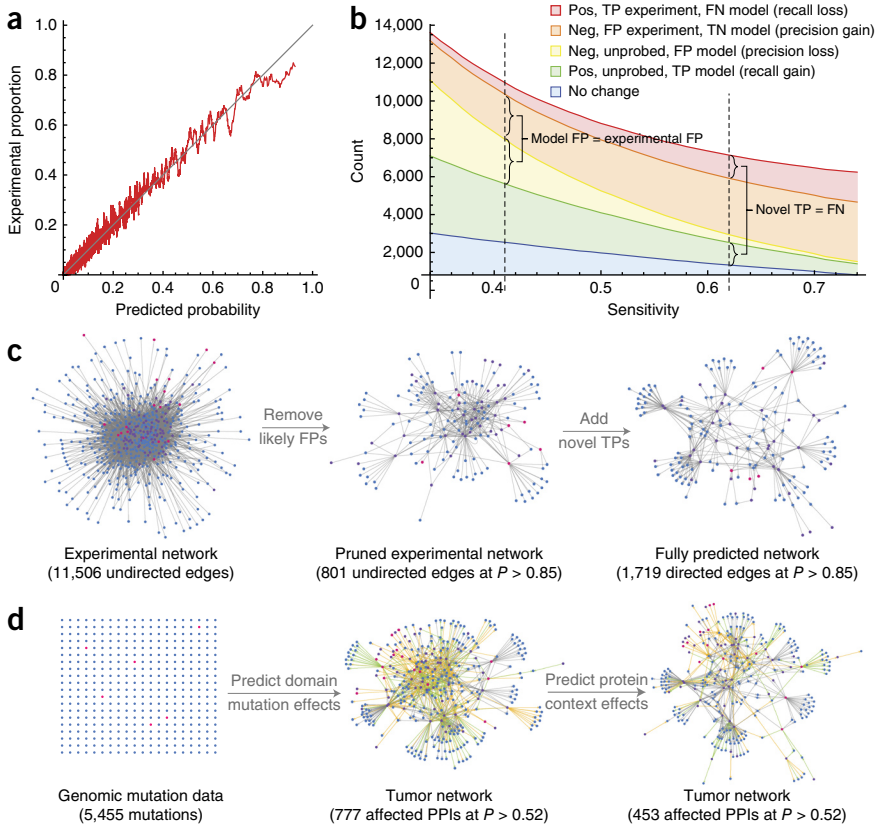

Figure 7 The model enriches high-throughput experiments. (**a**) SH2-phosphotyrosine interactions were rank-ordered by their predicted interaction probability and binned into overlapping windows. The average probability within each bin (*x* axis) is plotted against the proportion of experimental positives in the same bin (*y* axis). We found the agreement to be high, indicating that on a statistical level MSM/D can predict experimental accuracy. (**b**) Expected proportions of various outcomes (TP, true positive; TN, true negative; FP, false positive; FN, false negative) for model and experiment are plotted as a function of model sensitivity. The right dashed vertical line indicates a sensitivity level at which MSM/D is expected to predict as many new interactions (green) as it loses owing to oversensitivity (red). At this threshold, MSM/D is expected to eliminate ~7 times more false positives than it adds (415 model false positives added versus 2,973 experimental false positives eliminated). The left dashed vertical line corresponds to a sensitivity level at which MSM/D is expected to add the same number of false positives (yellow) as it eliminates (orange). At this threshold, MSM/D discovers ~5 times more true positives than it loses (3,091 model true positives added versus 614 experimental true positives lost). Pos, positive; neg, negative. (**c**) Model predictions can be used as quality indicators to enrich high-throughput experiments for true positives by eliminating low-probability interactions. Model predictions can also be used to add novel interactions that have not been experimentally probed. (**d**) Genomic mutation data only provide node-level information (i.e., which gene is mutated). The model converts node-level mutation information into edge-level perturbations and integrates the known or predicted PPI network to model the buffering effects of multi-site proteins.

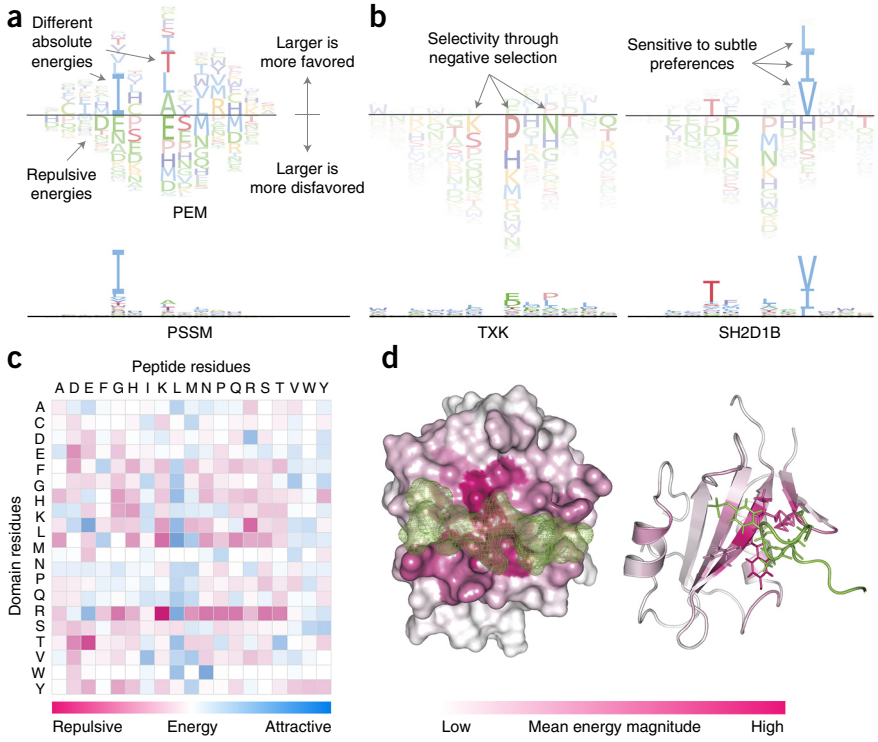

Figure 8 PEMs capture the biophysical basis of SH2 domain specificity. (**a**) PEM representation. Amino acids exhibiting attractive interactions lie above the dividing line, whereas amino acids involving repulsive interactions lie below, with the height of the residue corresponding to the magnitude of the interaction energy. PEMs capture the effects of negative selectivity and differential energy contributions at different residue positions. (**b**) The PEM for the SH2D1B domain shows that a threonine at position –2 (relative to the phosphotyrosine site) contributes less to affinity than a leucine or isoleucine at position +3. In the PSSM, the situation is reversed because the PSSM representation forces each position to contribute equally to the total probability, which causes the dominant valine at position +3 to appear more important than it is in terms of actual energetics. Negative selectivity is also readily evident using PEMs: in the case of the TXK SH2 domain, specificity involves repulsive interactions, specifically proline, asparagine and lysine residues at positions +1, +3, and –1, respectively. These effects on selectivity cannot be discerned from the corresponding PSSM. (**c**) Heat map of pairwise amino acid interaction energies at the SH2-phosphopeptide interface as derived from MSM/D. Instances of strong negative energies (bright pink) correspond to electrostatic repulsion (for example, arginine and lysine), whereas positive energies (bright blue) are electrostatically complementary (for example, arginine and aspartic acid) or involve buried hydrophobic amino acids (for example, leucine and leucine). (**d**) Heat map of the average magnitude of interaction energies per residue position projected onto a structural representative of SH2 domains (white) in complex with phosphopeptide (green) (Protein Data Bank (PDB) 1JU5).

SH2-phosphotyrosine signaling networks in normal and cancer cells via MSM. MSM methodology integrates and reconciles diverse genomic, biochemical and structural data and provides insights into determinants of binding specificity and the consequences of genetic mutations on the basis of biophysical principles (**Fig. 7d**). On the molecular level, we find that the majority of mutations that are consequential for PPIs map to phosphoproteins and are equally likely to result in gain or loss of interactions. Conversely, SH2 domain mutations are mostly loss of function. At the network level, cancer-associated mutations rewire SH2 networks in a bimodal fashion, coordinately rewiring connected subnetworks in one mode and disrupting the total function of individual proteins in the other mode.

To summarize the selectivity of SH2 domains, we developed a new matrix representation: position energy matrices (PEMs) (**Fig. 8a**, **Supplementary Table 4** and **Supplementary Note**). Existing position-specific scoring matrices (PSSMs) describe determinants of binding selectivity by specifying the relative preferences for a base or amino acid at each position in the bound biomolecule. In contrast, PEMs describe per-residue interaction energies using a scale that is universal across SH2 domains and residue positions (**Fig. 8c**). This makes it possible to compare absolute preferences between peptide positions and capture selectivity effects that are obscured by PSSMs (**Fig. 8a**,**b**). The PEM representation also makes clear that residue-specific negative interaction energies (those lying below the horizontal lines in **Fig. 8a**,**b**) have an important role in binding selectivity. By mapping the position-specific MSM/D energies onto the three-dimensional structure of the SH2-phosphotyrosine binding interface (**Fig. 8d**), we find that positive and negative energetic hotspots lie primarily in the peptide-binding pocket, showing that MSM/D captures the physicochemical basis of protein-peptide interaction.

Because it is probabilistic, MSM/D can estimate the proportions of false positives and false negatives in experiments (**Fig. 7a**). An interaction deemed to be an experimental negative but that is assigned a 90% probability by MSM/D has only a 10% probability of being an experimental true negative and a model false positive. Using a sensitivity threshold at which MSM/D is expected to predict as many new interactions (true positives) as it loses (false negatives), MSM/D eliminates ~7 times more false positives than it adds (**Fig. 7b**); at a sensitivity threshold at which MSM/D is expected to add the same number of false positives as it eliminates, MSM/D discovers ~5 times more true positives than it loses. These results suggest a role for MSM in pruning large-scale data as a means to increase quality, sensitivity and concordance across assays (**Fig. 7c**). By analogy, the introduction of Phred quality scores for DNA sequencing was critical in reducing error and increasing throughput in genomics[36]. Data pruning can also be used in an iterative approach involving model training on data and data refinement using a model. More generally, we propose that precise statistical modeling is a superior approach to reconcile irreproducible and discordant data in biomedicine[37] than simple repetition.

The MSM framework is applicable to any PID for which interaction and structural data are available (for example, PTB, SH3 and PDZ domains), including DNA-binding protein domains. Certain PIDs (for example, SH3 domains) will present additional difficulties because they lack the absolute reference frame for peptide alignment provided by phosphotyrosine residues, possibly necessitating threading and structural alignment. Moreover, MSM does not currently take into account levels of protein expression or actual states of tyrosine phosphorylation in cells but can be extended to incorporate these data as they become available (for example, from quantitative

mass spectrometry). The ultimate goal is to model information flow through PID networks under different physiological conditions as a means to understand normal physiology, disease-associated mutations and patient-specific phenotypic responses. Statistical mechanical ensembles such as those described here provide the conceptual framework needed to achieve these aims.

URLs. Website associated with this report, http://lincs.hms.harvard.edu/alquraishi-natgenet-2014/; supporting source code, https://github.com/AlQuraishiLab/sh2-cancer; allOnco Genelists, http://www.bushmanlab.org/links/genelists.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

This work was supported by US National Institutes of Health grants GM68762, GM107618 and GM072872. We used the resources of the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the US Department of Energy under contract DE-AC02-05CH11231.

AUTHOR CONTRIBUTIONS

All authors conceived and designed the study. M.A., G.K. and P.K.S. wrote the manuscript. M.A. developed the mathematical model. G.K. performed the experiments. All authors discussed and interpreted the results.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

1. Lawrence, M.S. *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495–501 (2014).
2. Liu, B.A., Engelmann, B.W. & Nash, P.D. High-throughput analysis of peptide-binding modules. *Proteomics* **12**, 1527–1546 (2012).
3. Ho, Y. *et al.* Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* **415**, 180–183 (2002).
4. Bader, G.D. & Hogue, C.W.V. Analyzing yeast protein-protein interaction data obtained from different sources,. *Nat. Biotechnol.* **20**, 991–997 (2002).
5. Gschwind, A., Fischer, O.M. & Ullrich, A. The discovery of receptor tyrosine kinases: targets for cancer therapy. *Nat. Rev. Cancer* **4**, 361–370 (2004).
6. Zhong, Q. *et al.* Edgetic perturbation models of human inherited disorders. *Mol. Syst. Biol.* **5**, 321 (2009).
7. Ren, J. *et al.* PhosSNP for systematic analysis of genetic polymorphisms that influence protein phosphorylation. *Mol. Cell. Proteomics* **9**, 623–634 (2010).
8. Tamborero, D. *et al.* Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Sci. Rep.* **3**, 2650 (2013).
9. Birge, R.B., Kalodimos, C., Inagaki, F. & Tanaka, S. Crk and CrkL adaptor proteins: networks for physiological and pathological signaling. *Cell Commun. Signal.* **7**, 13 (2009).
10. Li, L. *et al.* Prediction of phosphotyrosine signaling networks using a scoring matrix-assisted ligand identification approach. *Nucleic Acids Res.* **36**, 3263–3273 (2008).
11. Kundu, K., Costa, F., Huber, M., Reth, M. & Backofen, R. Semi-supervised prediction of SH2-peptide interactions from imbalanced high-throughput data. *PLoS ONE* **8**, e62732 (2013).
12. Miller, M.L. *et al.* Linear motif atlas for phosphorylation-dependent signaling. *Sci. Signal.* **1**, ra2 (2008).
13. Wunderlich, Z. & Mirny, L.A. Using genome-wide measurements for computational prediction of SH2-peptide interactions. *Nucleic Acids Res.* **37**, 4629–4641 (2009).
14. Zhang, Q.C. *et al.* Structure-based prediction of protein-protein interactions on a genome-wide scale. *Nature* **490**, 556–560 (2012).
15. Franceschini, A. *et al.* STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.* **41**, D808–D815 (2013).
16. Forbes, S.A. *et al.* COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res.* **39**, D945–D950 (2011).
17. AlQuraishi, M. & McAdams, H.H. Direct inference of protein-DNA interactions using compressed sensing methods. *Proc. Natl. Acad. Sci. USA* **108**, 14819–14824 (2011).
18. Koytiger, G. *et al.* Phosphotyrosine signaling proteins that drive oncogenesis tend to be highly interconnected. *Mol. Cell. Proteomics* **12**, 1204–1213 (2013).
19. Hause, R.J. *et al.* Comprehensive binary interaction mapping of SH2 domains via fluorescence polarization reveals novel functional diversification of ErbB receptors. *PLoS ONE* **7**, e44471 (2012).
20. Liu, B.A. *et al.* SH2 domains recognize contextual peptide sequence information to determine selectivity. *Mol. Cell. Proteomics* **9**, 2391–2404 (2010).
21. Tinti, M. *et al.* The SH2 domain interaction landscape. *Cell Rep.* **3**, 1293–1305 (2013).
22. Hornbeck, P.V. *et al.* 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 (2012).
23. Branden, C. & Tooze, J. *Introduction to Protein Structure* (Garland Science, New York, 1999).
24. von Mering, C. *et al.* Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* **417**, 399–403 (2002).
25. Lossos, I.S., Alizadeh, A.A., Rajapaksa, R., Tibshirani, R. & Levy, R. *HGAL* is a novel interleukin-4–inducible gene that strongly predicts survival in diffuse large B-cell lymphoma. *Blood* **101**, 433–440 (2003).
26. Natkunam, Y. *et al.* Expression of the human germinal center–associated lymphoma (HGAL) protein identifies a subset of classic Hodgkin lymphoma of germinal center derivation and improved survival. *Blood* **109**, 298–305 (2007).
27. Pan, Z. *et al.* Studies of a germinal centre B-cell expressed gene, *GCET2*, suggest its role as a membrane associated adapter protein. *Br. J. Haematol.* **137**, 578–590 (2007).
28. Romero-Camarero, I. *et al.* Germinal centre protein HGAL promotes lymphoid hyperplasia and amyloidosis via BCR-mediated Syk activation. *Nat. Commun.* **4**, 1338 (2013).
29. Davies, H. *et al.* Somatic mutations of the protein kinase gene family in human lung cancer. *Cancer Res.* **65**, 7591–7595 (2005).
30. Zhao, M., Sun, J. & Zhao, Z. TSGene: a web resource for tumor suppressor genes. *Nucleic Acids Res.* **41**, D970–D976 (2013).
31. Watson, I.R., Takahashi, K., Futreal, P.A. & Chin, L. Emerging patterns of somatic mutations in cancer. *Nat. Rev. Genet.* **14**, 703–718 (2013).
32. Marengere, L.E. *et al.* SH2 domain specificity and activity modified by a single residue. *Nature* **369**, 502–505 (1994).
33. Cabodi, S., del Pilar Camacho-Leal, M., Di Stefano, P. & Defilippi, P. Integrin signalling adaptors: not only figurants in the cancer story. *Nat. Rev. Cancer* **10**, 858–870 (2010).
34. Haynie, D.T. Molecular physiology of the tensin brotherhood of integrin adaptor proteins. *Proteins* **82**, 1113–1127 (2014).
35. Ewan, K.B.R. *et al.* Proliferation of estrogen receptor-α–positive mammary epithelial cells is restrained by transforming growth factor-β1 in adult mice. *Am. J. Pathol.* **167**, 409–417 (2005).
36. Ewing, B., Hillier, L., Wendl, M.C. & Green, P. Base-calling of automated sequencer traces using Phred. I. Accuracy assessment. *Genome Res.* **8**, 175–185 (1998).
37. Freedman, L.P. & Inglese, J. The increasing urgency for standards in basic biologic research. *Cancer Res.* **74**, 4024–4029 (2014).

ONLINE METHODS

Quantitative coimmunoprecipitation of GCSAM-interacting proteins. Constructs for N-terminally GFP-tagged SH2 domains and C-terminally TagRFP-tagged GCSAM were cotransfected into HEK293T cells acquired from the American Type Culture Collection (cells were not tested for mycoplasma). After 24 h, cells were treated with freshly prepared pervanadate according to a previously published protocol[27] and subsequently lysed using Cell Signaling Lysis Buffer according to the manufacturer's protocol. Cleared lysate was added to GFP-Trap agarose conjugated beads (ChromoTek, gta-20) and then incubated for 1 h at 4 °C. After centrifugation, 40 µl of supernatant was transferred to a 384-well plate. The beads were subsequently washed twice and also transferred to the same 384-well plate for imaging on the Operetta High-Content Imaging System in both GFP and RFP channels. By normalizing the bead RFP signal by the bead GFP signal and the supernatant RFP signal, a quantitative value was obtained that was linearly related to the association constant of the two species, under the assumption that bead GFP signal primarily reflects the unbound state. To facilitate quantitative comparison, the signals were divided by the signal of the weakest binder, CRK, yielding fold change (f) measurements that were rescaled between 0.5 and 1 using the equation $f/(1 + f)$. This quantity represents the relative occupancy of the bound and unbound states.

Determining the effect of the IGF1R p.Ala1347Val alteration. The mutation encoding p.Ala1347Val was introduced into a plasmid with *IGF1R* using site-directed mutagenesis. IGF1R quantitative coimmunoprecipitation was performed similarly to in GCSAM experiments, except that IGF1R was GFP tagged, SH2 domains were tagged with mCherry and immunoprecipitation was performed with RFP-Trap (ChromoTek, rta-20). Five biological replicates were performed on different days, and the imaging parameters were optimized in each experiment to prevent signal saturation; the data from each biological replicate were therefore rescaled using a constant that related the signal intensities of imaging.

doi:10.1038/ng.3138