Literature-based automated discovery of tumor suppressor p53 phosphorylation and inhibition by NEK2

Byung-Kwon Choi[a,1], Tajhal Dayaram[b,1], Neha Parikh[b,1,2], Angela D. Wilkins[a,c,1], Meena Nagarajan[d], Ilya B. Novikov[a], Benjamin J. Bachman[a], Sung Yun Jung[e], Peter J. Haas[d], Jacques L. Labrie[d], Curtis R. Pickering[f], Anbu K. Adikesavan[a], Sam Regenbogen[g], Linda Kato[d], Ana Lelescu[d], Christie M. Buchovecky[a], Houyin Zhang[a], Sheng Hua Bao[d], Stephen Boyer[d], Griff Weber[d], Kenneth L. Scott[a], Ying Chen[d], Scott Spangler[d], Lawrence A. Donehower[b,3], and Olivier Lichtarge[a,c,g,3]

[a]Department of Human and Molecular Genetics, Baylor College of Medicine, Houston, TX 77030; [b]Department of Molecular Virology and Microbiology, Baylor College of Medicine, Houston, TX 77030; [c]Computational and Integrative Biomedical Research Center, Baylor College of Medicine, Houston, TX 77030; [d]Watson Health, IBM Almaden Research Center, San Jose, CA 95120; [e]Department of Molecular and Cellular Biology, Baylor College of Medicine, Houston, TX 77030; [f]M.D. Anderson Cancer Center, The University of Texas, Houston, TX 77030; and [g]Department of Pharmacology, Baylor College of Medicine, Houston, TX 77030

Edited by Carol Prives, Columbia University, New York, NY, and approved August 28, 2018 (received for review April 29, 2018)

Scientific progress depends on formulating testable hypotheses informed by the literature. In many domains, however, this model is strained because the number of research papers exceeds human readability. Here, we developed computational assistance to analyze the biomedical literature by reading PubMed abstracts to suggest new hypotheses. The approach was tested experimentally on the tumor suppressor p53 by ranking its most likely kinases, based on all available abstracts. Many of the best-ranked kinases were found to bind and phosphorylate p53 (*P* value = 0.005), suggesting six likely p53 kinases so far. One of these, NEK2, was studied in detail. A known mitosis promoter, NEK2 was shown to phosphorylate p53 at Ser315 in vitro and in vivo and to functionally inhibit p53. These bona fide validations of text-based predictions of p53 phosphorylation, and the discovery of an inhibitory p53 kinase of pharmaceutical interest, suggest that automated reasoning using a large body of literature can generate valuable molecular hypotheses and has the potential to accelerate scientific discovery.

literature text mining | automated hypothesis generation | protein–protein interaction | p53 inhibition | kinase

Developing useful hypotheses depends on understanding and making inferences from prior information. This is a challenge when the scale of the data exceeds human analytical capacity, in which case computational assistance is needed. Algorithms are already highly effective for reasoning and generating solutions when this large-scale information is structured, that is, tabulated in accessible databases (1, 2). When the information is described by words and sentences, however, the generation of hypotheses is more limited (3) and text-mining algorithms tend to focus instead on the retrieval of independent facts (4). In biomedicine, the research literature surpasses 25 million papers. Even restricted domains can include tens of thousands of papers. These numbers highlight a need beyond computational search for new reasoning and discovery applications based on integrative hypothesis generation applied to text (5, 6).

Natural language processing efforts in biomedical literature typically identify the important entities (i.e., proteins, diseases, drugs) and their semantic relationships (7–9). This process relies on curated dictionaries and rules-based approaches to identify and normalize important biological entities (10). A pivotal demonstration of hypothesis generation from the biomedical literature is computer-aided discovery by Swanson linking—that is, if A causes B and B causes C, then A might cause C (11–13)—the original example being between fish oil and Raynaud's disease patients (14). More broadly, mining the literature for proteins, diseases, drugs, and their relationships allows for network-based approaches to identify disease biomarkers (15), repurpose drugs (16), and suggest protein function (17). In recent work (18, 19), we developed an approach to suggest protein interactions by diffusing information over a kinase–kinase network that was built solely from the word context of individual proteins in the abstract in which they appear. To measure the biological gains of the method, we now follow on these limited and retrospective computational studies by testing their predictions prospectively against laboratory experiments.

The tumor suppressor p53 provides an opportune test case. It is the most mutated gene in cancer (20–22), and over 90,000 PubMed studies detail how it responds to genomic stress to coordinate cellular defenses against cancer and other diseases (22). Almost a third of p53 paper abstracts mention kinases—a family

Significance

We adapted natural language processing to the biological literature and demonstrated end-to-end automated knowledge discovery by exploring subtle word connections. General text mining scanned 21 million publication abstracts and selected a reliable 130,000 from which hypothesis generation algorithms predicted kinases not known to phosphorylate p53, but likely to do so. Six of these p53 kinase candidates passed experimental validation. Among them NEK2 was examined in depth and shown to repress p53 and promote cell division. This work demonstrates the possibility of integrating a vast corpora of written knowledge to compute valuable hypotheses that will often test true and fuel discovery.

Author contributions: B.-K.C., T.D., N.P., A.D.W., S.B., G.W., K.L.S., Y.C., S.S., L.A.D., and O.L. designed research; B.-K.C., T.D., N.P., A.D.W., M.N., I.B.N., B.J.B., S.Y.J., P.J.H., J.L.L., C.R.P., A.K.A., S.R., L.K., A.L., C.M.B., H.Z., and S.H.B. performed research; B.-K.C., T.D., N.P., A.D.W., M.N., I.B.N., B.J.B., S.Y.J., P.J.H., J.L.L., C.R.P., A.K.A., S.R., L.K., A.L., C.M.B., H.Z., S.H.B., and S.S. analyzed data; and B.-K.C., T.D., N.P., A.D.W., S.S., L.A.D., and O.L. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1] B.-K.C., T.D., N.P., and A.D.W. contributed equally to this work.

[2] Present address: Department of Diagnostic and Biomedical Sciences, School of Dentistry at Houston, The University of Texas Health Science Center, Houston, TX 77054.

[3] To whom correspondence may be addressed. Email: larryd@bcm.tmc.edu or lichtarge@bcm.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1806643115/-/DCSupplemental.

Published online September 28, 2018.

of evolutionarily related proteins that regulate other proteins through phosphorylation (23) and that are an important source of drug targets (24). The discovery of new kinases that regulate p53 may thus lead to additional therapeutic targets (25). However, the size of this body of literature defies human appraisal and thus limits the scope of current scientific hypotheses (26). By combining predictive algorithms in biology (27) with recent improvements in natural language processing (28, 29), we sought to mine the biological literature and predict biological interactions with support from retrospective computational validation (18, 19). Here, we provide experimental proof that word context information from abstracts alone is sufficient to suggest automated hypotheses that prove correct and lead to the discovery of p53 biological interactions. In-depth molecular studies of one candidate kinase, NEK2, further reveal that this cancer-relevant kinase phosphorylates p53 to negatively regulate p53 functions.

Results

Computational Methods to Identify Kinases That Phosphorylate p53. To identify p53 kinases, a similarity network of human kinases was built from the literature (Fig. 1 and *SI Appendix*). All 21 million PubMed abstracts available in January 2014 were first searched for standard names and synonyms of kinase genes; 240,000 abstracts mentioning human kinases were found. For each kinase entity, the collection of distinct and relevant words contained in abstracts written about that kinase were then counted and summarized as vectors (30). Vector comparisons among all kinases next yielded pairwise distances by taking the cosine similarity. These distances gave rise to a kinase–kinase network in which edges between nodes represent significant word content similarities of abstracts mentioning connected kinases (Fig. 1). Finally, graph information diffusion (17) was performed by globally propagating labels throughout this kinase network from the known p53 kinase nodes to nodes lacking such prior annotation. The result was a ranking of all kinases not previously known to phosphorylate p53. This approach was previously shown to be scalable, but biological evidence for p53 phosphorylation was lacking (18, 19).

Biochemical Screens to Validate Computational Predictions of p53 Kinases. To test prospective predictions of p53 kinases, 26 commercially available kinases were chosen from the top-ranked and the bottom-ranked predictions. These 26 kinases were experimentally tested by two screens. First, an in vitro kinase assay was used to measure whether a purified candidate kinase could phosphorylate purified p53 in the test tube. Of 26 purified candidate kinases tested, 9 kinases exhibited high levels of p53 phosphorylation (Fig. 2*A* and *SI Appendix*, Fig. S1 and Table S1). A second screening assay applied to the 26 candidate kinases was a coimmunoprecipitation assay to detect those kinases that could form a stable protein–protein interaction with p53 in human cells. Twelve immunoprecipitated candidate kinases interacted with p53 and exhibited an intense band on the p53-containing region of the immunoblot after p53 antibody probing (Fig. 2*B* and *SI Appendix*, Fig. S2). These results were corroborated by reciprocal coimmunoprecipitation assays (p53 immunoprecipitation and kinase immunoblotting). PKN1 and NEK2 were confirmed as highly interactive in coimmunoprecipitations (Fig. 2*C*).

Six kinases—NEK2, PLK1, PKN1, PKN2, PAK4, and PAK6—were found to be positive in both the in vitro kinase and coimmunoprecipitation screening assays (Fig. 2*D* and *SI Appendix*, Table S1). The rank distribution of these six kinases was significantly higher than expected by chance (*P* value of 0.0046 by χ^2 test). Likewise, a receiver-operating characteristic (ROC) curve statistical analysis indicated a *P* value of 0.0048 relative to a random distribution of true and false positives (Fig. 2*E*). Based on our screening assays, the top-ranked predictions were therefore significantly enriched for likely p53 kinases.

NEK2 Phosphorylates p53 at Ser315 and Reduces Its Stability. The two screening assays employed above are not definitive in categorizing a p53 kinase. We thus performed extensive experimentation on one of the six candidate kinases positive in both screening assays to more conclusively validate it. One of the highest scoring kinases, NEK2, was chosen for further analysis (*SI Appendix*, Table S1). NEK2 is an important mitosis regulator (31) and may functionally affect p53, a cell-cycle regulator (22). To confirm that NEK2 phosphorylates p53, we immunoblotted in vitro kinase reaction components containing purified p53 with or without purified NEK2 with antibodies specific for p53 phosphoserine 315 and showed a p53 phosphoserine 315-specific band only in the presence of NEK2 (Fig. 3*A*). As a positive control, we also incubated recombinant p53 with purified Aurora kinase A (AURKA), previously demonstrated to phosphorylate p53 at Ser315 (32). Both NEK2 and AURKA showed similar levels of p53 Ser315 phosphorylation in vitro. This in vitro kinase assay was repeated with phospho-specific antibodies to p53 Ser15, Thr18, Ser33, Ser37, Ser46, and Ser392, but these sites were not phosphorylated by NEK2 in vitro (*SI Appendix*, Fig. S3). Mass spectrometry (LC-MS/MS) analyses confirmed p53 Ser315 phosphorylation. Four separate NEK2-p53 in vitro kinase reactions generated robust levels of p53 phosphoserine 315 peptides, whereas p53 incubated without NEK2 did not produce phosphorylated peptides (*SI Appendix*, Fig. S4).

To demonstrate NEK2 phosphorylation of p53 in intact cells, we transfected HCT116 ($p53^{+/+}$) human colorectal cancer cells with a LacZ expression vector, with a wild-type NEK2 vector (WT-NEK2), or with a kinase-dead NEK2 (K37R, KD-NEK2) vector. We immunoprecipitated p53 and performed Western blot analysis on the immunoprecipitates with p53 and phospho-p53 Ser315 antibodies (Fig. 3*B*). p53 Ser315 showed enhanced phosphorylation in cells transfected with WT-NEK2, but not with KD-NEK2.

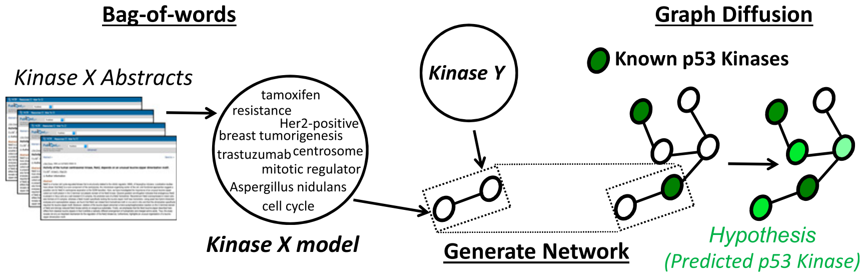

Fig. 1. Computational mining of the scientific literature to build kinase–kinase relationship networks and predict kinase interactions. Model illustrating how gene, protein, biological processes, and word entities are mined from scientific literature abstracts and compared for similarities to build a kinase–kinase network. Graph diffusion is then used to propagate known p53 kinase information through the network to predict undiscovered kinases likely to phosphorylate p53.

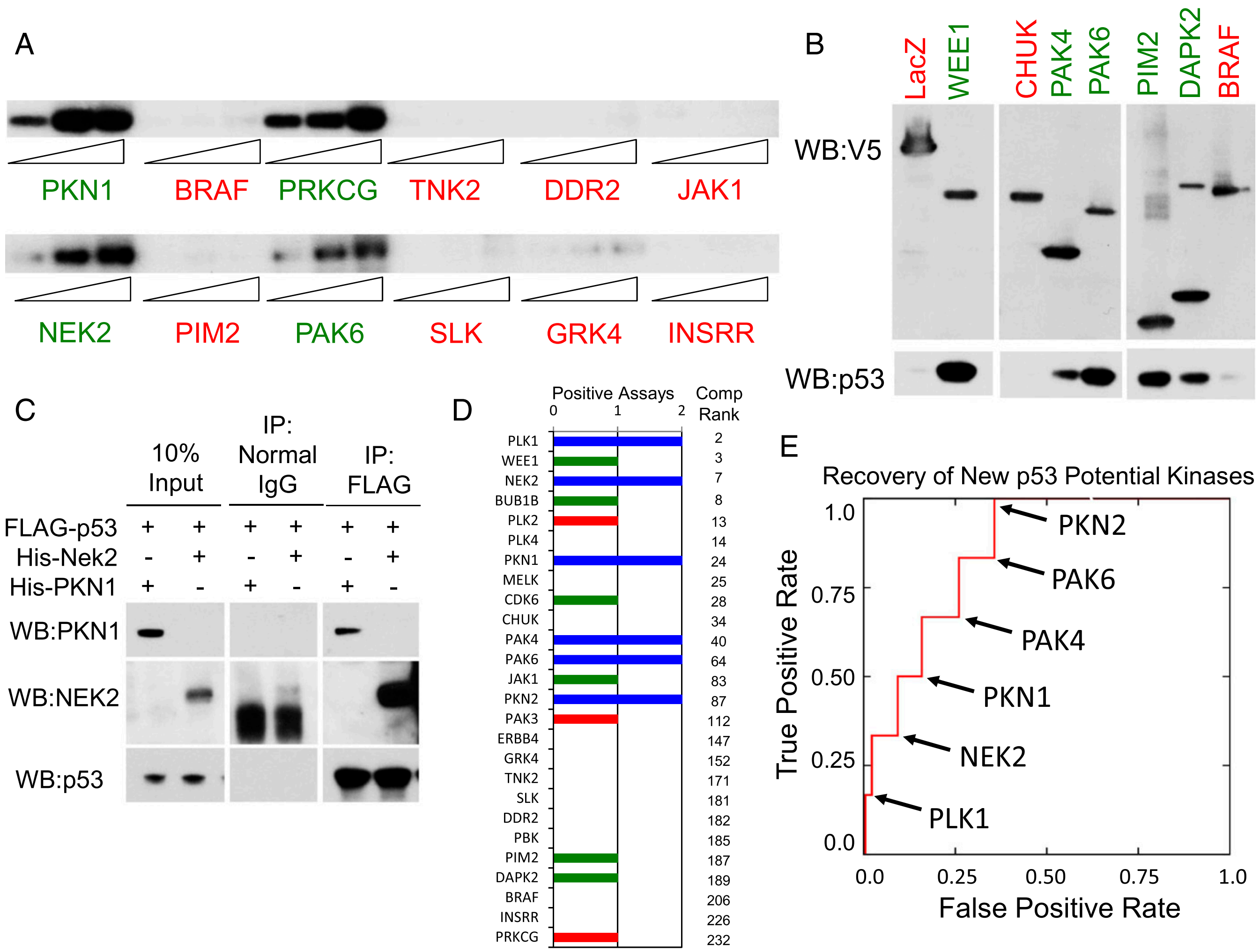

Fig. 2. Screening assays to evaluate computationally predicted kinases for p53 phosphorylation and interaction. (*A*) In vitro kinase assays show robust phosphorylation of p53 by some kinases. In vitro kinase assays were performed using 5, 15, or 50 ng of the indicated kinases and 50 ng of recombinant His-tagged p53 in the presence of ^{32}P-gamma ATP, incubated for 15 min, and resolved by gel electrophoresis. (*B*) Some kinases form protein–protein interactions with p53 in human cells. HEK293 cells were transfected with V5-tagged kinase vectors or a V5-tagged LacZ control. Posttransfection lysates were immunoprecipitated using anti-V5–conjugated agarose, followed by immunoblotting with anti-V5 or anti-p53 antibodies. (*C*) Kinases NEK2 and PKN1 form protein–protein interactions with p53 protein in human cells. NEK2 or PKN1 vectors were cotransfected into HEK293 cells with a Flag-p53 vector. Lysates were immunoprecipitated with an anti-Flag antibody or normal IgG. Immunoblots were assayed for PKN1 and NEK2 interaction with p53 using anti-PKN1, anti-NEK2, and anti-p53 antibodies. (*D*) Experimental validation screen results by in vitro kinase assays and coimmunoprecipitation assays. Of 26 kinases, 6 kinases were positive in both assays (blue bars), 3 kinases were positive only in the in vitro kinase assay (red bars), 6 kinases were positive only in the coimmunoprecipitation assay (green bars), and 11 kinases were negative for both assays (no bars). "Comp Rank" signifies the computational ranks provided by our network algorithms. This distribution was significant by Fisher's exact test ($P = 0.0046$). (*E*) The prospective validation of literature vector models with graph diffusion to predict six potential p53 kinases. ROC curve shows these six targets in their respective ranks ($P = 0.0048$).

To further test whether NEK2 specifically phosphorylates p53 on Ser315, we transfected p53 S315A, S96A, and S96A/S315A point mutant vectors (Ser96 was considered a potential NEK2 phosphorylation site after reviewing mass spectrometry results on in vitro kinase components) into HCT116 (p53$^{-/-}$) cells with or without NEK2 expression vectors. Western blots of transfected cell lysates probed with p53 phosphoserine 315 antibodies showed that, while Ser315 phosphorylation was robust in p53 WT- and p53 S96A-transfected cells and enhanced further with added NEK2, transfection of p53 S315A and S96A/S315A mutants reduced phosphorylation at Ser315 (Fig. 3*C*). We also examined NEK2 inhibition on p53 protein and p53 Ser315 phosphorylation. Three different NEK2 shRNA lentiviral vectors were transduced into HCT116 (p53$^{+/+}$) cells, and p53 Ser315 phosphorylation was reduced relative to cells transduced with scrambled shRNA (Fig. 3*D*).

AURKA, another mitotic kinase that phosphorylates p53 at Ser315, has been shown to destabilize p53 (33, 34). To determine whether NEK2 phosphorylation of p53 affects p53 stability, we cotransfected p53 null HCT116 cells with WT 53 plus GFP vectors together with empty vector, WT p53, kinase-dead mutant NEK2, or WT AURKA vectors. One day posttransfection, transfected cell lysates were subjected to electrophoresis followed by immunoblotting with p53 and GFP antibodies. As shown in Fig. 3*E*, *Left*, both WT-NEK2 and AURKA reduce p53 protein levels in the transfected cells relative to cells transfected with empty vector or kinase-dead NEK2, indicating that both kinases appear to reduce p53 protein stability. Moreover, the NEK2 and AURKA effects on p53 protein levels appear to be dependent on p53 Ser315 phosphorylation, as cells transfected with p53 S315A mutant vector show no reduction in p53 protein levels relative to cells transfected with empty vectors or NEK2 kinase-dead vector (Fig. 3*E*, *Right*).

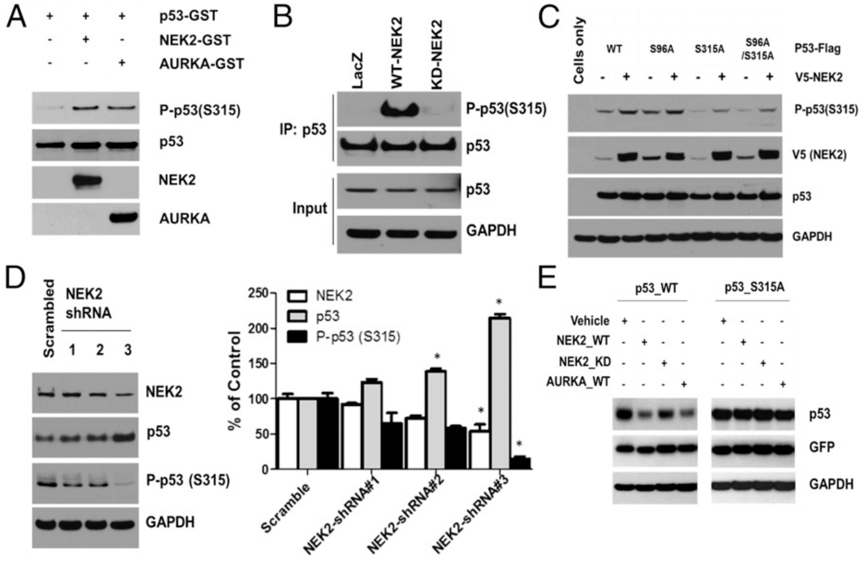

Fig. 3. NEK2 phosphorylates at p53 Serine 315. (*A*) Recombinant NEK2 phosphorylates recombinant p53 at Ser315 in vitro. Purified recombinant p53-GST was incubated without kinase or with recombinant NEK2 kinase or positive control AURKA kinase for 30 min at 30 °C. Reactions were immunoblotted with the indicated antibodies. (*B*) NEK2 overexpression induces enhanced p53 phosphorylation at Ser315 in human cells. HCT116 cells were transfected with lacZ, WT-NEK2, and KD-NEK2 expression vectors, and endogenous p53 was immunoprecipitated from lysates using anti-p53 (DO-1). p53 Ser315 phosphorylation was determined by immunoblot probing with a p53 phosphoserine 315-specific antibody. (*C*) Conversion of p53 Serine 315 to Alanine results in reduced p53 phosphorylation in the presence of overexpressed NEK2. HCT116 p53$^{-/-}$ cells were transfected with WT-p53, mutant p53 (S96A), mutant p53 (S315A), and double mutant p53 (S96A/S315A) with or without WT-NEK2. Cells were lysed and immunoblotted with the indicated antibodies. (*D*) Inhibition of NEK2 expression is associated with reduced phosphorylation of p53 at Ser315. HCT116 cells were transduced with lentivirus expressing nontarget control shRNA and three distinct shRNAs NEK2. Vector-expressing cell lysates were immunoblotted with NEK2, p53 protein, and p53 phosphoserine 315-specific antibodies. (*Right*) Quantitation of the blot data. *$P < 0.05$ ($n = 2$). (*E*) NEK2 reduction of p53 protein levels in human cells is dependent on phosphorylation of p53 Ser315. HCT116 (p53 null) cells were cotransfected with WT p53 plus GFP expression plasmids along with empty vector or WT-p53, kinase-dead NEK2, or WT AURKA expression vectors. Transfected cell lysates were prepared 24 h after transfection and subjected to SDS/PAGE and immunoblotting with the indicated antibodies to the right of each panel.

NEK2 has been shown to regulate mitotic progression through facilitation of centrosome duplication and spindle assembly (32). Given that other kinases that phosphorylate p53 and facilitate cell-cycle progression (AURKA, CDK2) also phosphorylate p53 at Ser315 (33, 34), we examined p53 and NEK2 expression in synchronized cells released from G1 block. The HCT116 (p53$^{+/+}$) cells arrested in G1/S phase showed accumulation of p53 protein. NEK2 protein levels were elevated in S, G2, and G2/M phases of the cell cycle relative to unsynchronized cells, and this was correlated with increased p53 phosphoserine 315 levels (*SI Appendix*, Fig. S5). In contrast, p53 protein levels were moderately reduced in S, G2, and G2/M phase cell populations, consistent with NEK2 suppression of p53 function via p53 Ser315 phosphorylation. Interestingly, total p53 protein levels go back up in M phase while NEK2 levels are reduced.

NEK2 Phosphorylation of p53 Is Correlated with Altered p53 Functions. To assess NEK2 effects on p53 functions, we examined p53-induced transactivation and p53-induced apoptosis. HCT116 (p53 WT) cells were cotransfected with a p53 promoter luciferase reporter construct and vectors expressing CHK1 (a p53 activating kinase), WT-NEK2, or KD-NEK2. Luciferase activity, a proxy for p53 transactivation, was significantly increased with p53 positive regulatory kinase CHK1, while p53 activity was significantly reduced by WT-NEK2 (Fig. 4*A*). Kinase dead NEK2 had no effect on p53 transcription. These results are consistent with NEK2 inhibiting p53 transcriptional activation functions.

To assess NEK2 effects on p53 transcriptional targets, we transfected p53-null Saos-2 cells with vectors expressing LacZ, LacZ plus WT p53, WT-NEK2 plus WT p53, or KD-NEK2 plus p53. Quantitative RT-PCR for the p53 target genes *CDKN1A* (p21$^{\mathrm{CIP1}}$), *GADD45A*, and *FAS* resulted in p53-dependent induction of all three p53 target genes relative to LacZ (Fig. 4*B*) that was significantly attenuated by WT-NEK2. Kinase-dead NEK2 induced little change in p53-induced target expression. Transduction of HCT116 (p53$^{+/+}$) cells with three distinct shRNA lentiviral vectors reduced NEK2 protein expression and increased luciferase activity (Fig. 4*C*). Two of the three NEK2 shRNA-expressing lines that showed the highest knockdown of NEK2 expression exhibited significantly elevated p21$^{\mathrm{CIP1}}$ and FAS mRNA expression (Fig. 4*D*). Thus, NEK2 kinase activity suppresses p53 transactivation function.

We also examined NEK2 effects on p53-induced apoptosis. Apoptosis markers cleaved caspase 3 and cleaved PARP were increased in p53-transfected Saos-2 cells, but cotransfection of p53 plus WT-NEK2 reduced levels of the two apoptosis markers (Fig. 4*E*). Transfection of p53 plus KD-NEK2 did not reduce apoptotic markers. Transfection experiments in which fluorescence-tagged p53 plus or minus WT-NEK2 vectors in Saos-2 cells were examined for p53 immunofluorescence showed cells expressing p53 and undergoing simultaneous condensed nuclear staining (apoptosis markers). Roughly 14% of cells transfected with p53 plus LacZ showed apoptotic phenotypes compared with 2% of cells with NEK2 transfected with p53, indicating that NEK2 suppresses p53-mediated apoptosis (Fig. 4*F*). We also investigated whether NEK2 suppresses p53-mediated apoptosis following DNA damage. Treatment of p53 null Saos-2 cells with 5-Gy–ionizing radiation after transfection with p53 resulted in significantly increased numbers of apoptotic cells relative to control transfected LacZ cells as identified by fluorescence microscopy (*SI Appendix*, Fig. S6). The number of p53-induced apoptotic cells were significantly reduced when p53 was cotransfected with NEK2 expression vectors (*SI Appendix*, Fig. S6). These experiments indicate that NEK2 phosphorylates p53 and functionally inhibits p53.

Discussion

This study tested experimentally multiple text-based predictions of p53 kinases. An algorithm analyzed PubMed abstracts for kinase-relevant information, from which it inferred and ranked protein kinases by their likelihood to target p53. Phosphorylation of p53 activity was enriched among highly ranked kinases, including six that were positive in multiple assays. These six are likely to be p53 kinases, and this possibility was further tested in depth in NEK2. This protein was already known to affect mitosis (32), and we now have shown that it is also a p53 kinase that suppresses one of its cell-cycle–promoting functions.

Like the known p53 mitotic kinase, AURKA (35), NEK2 phosphorylates p53 at Ser315 and reduces p53 stability. While we believe much of the effect of NEK2 on p53 is likely due to

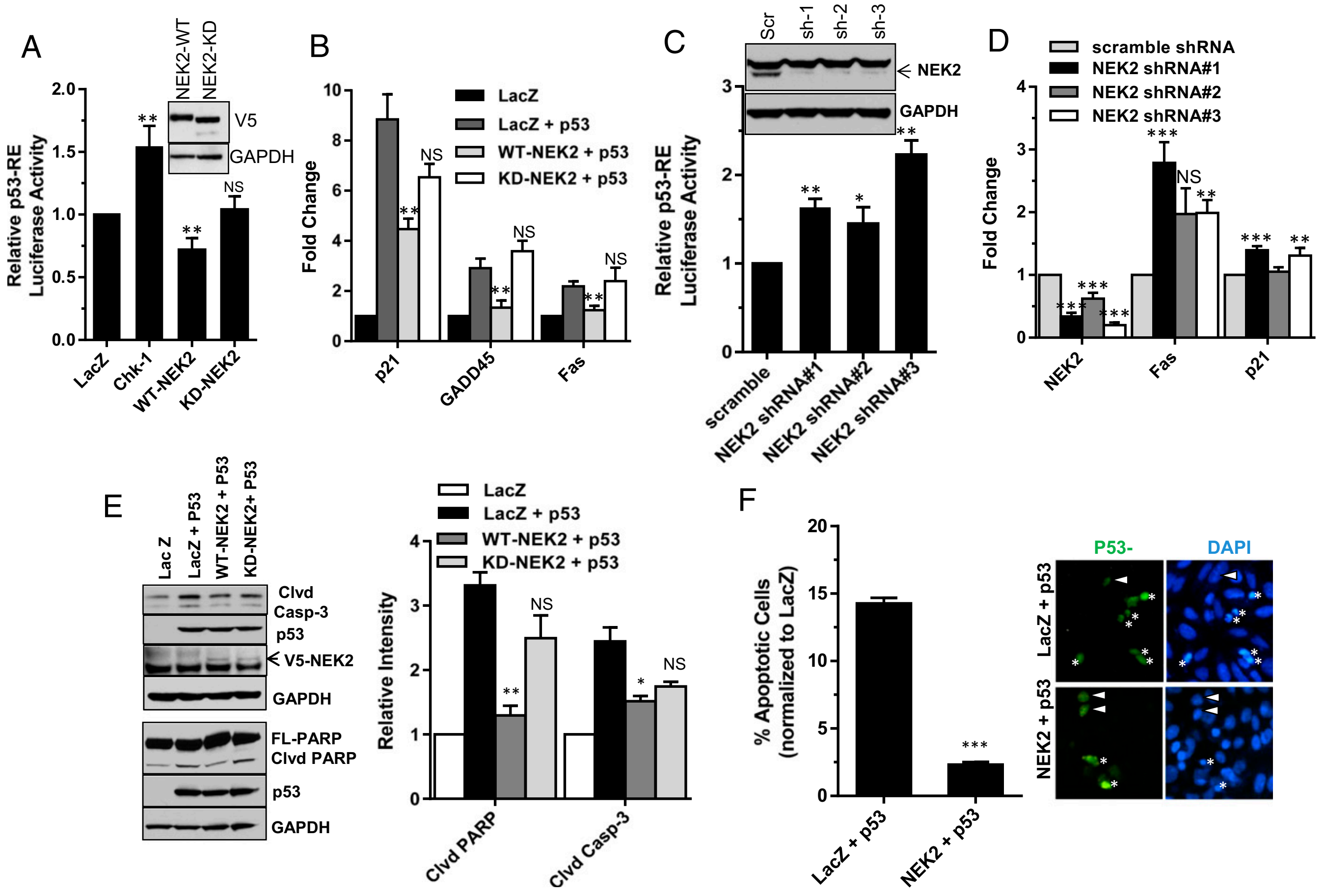

Fig. 4. NEK2 inhibits p53 transcriptional and apoptotic functions. (*A*) NEK2 inhibits p53-mediated transcription in luciferase assays. HCT116 p53$^{+/+}$ cells were transfected with V5-LacZ + pGL3-Luc; V5-LacZ + p53 response element luciferase (p53RE-Luc/PG13-Luc); V5-Chk1 + p53RE-Luc; V5-NEK2 + p53RE-Luc; and V5-KD-NEK2 + p53RE-Luc along with pRL (Renilla luciferase)-TK. Firefly luciferase and Renilla luciferase activities were quantitated, and ratios were normalized to the pGL3 + LacZ condition. The average of three experiments with SEM is plotted. (*B*) NEK2 inhibits transcriptional up-regulation of p53 target genes. Saos-2 cells were transfected with the indicated plasmids. Real-time PCR analysis was performed on p53 target RNAs using primers for p21, GADD45, and FAS. Gene expression was normalized to LacZ-transfected samples ($n = 3$). (*C*) NEK2 inhibition results in enhanced p53 transcriptional activity. HCT116 p53 WT cells stably expressing indicated shRNAs were transfected with pGL3-Luc and p53RE-Luc along with pRL-TK. p53 transcriptional activity was plotted relative to scrambled (Scr) shRNA. The average of three experiments with SEM is plotted. (*Top*) Relative levels of NEK2 (lower band–upper band is a cross-reacting non-NEK2 protein) and GAPDH loading control. (*D*) HCT116 p53 WT cells stably expressing scrambled shRNA, NEK2 shRNA-1, -2, or -3 assessed for NEK2, FAS, and p21 mRNA using qPCR. Average of three experiments with SEM is plotted. (*E*) Markers of p53-induced apoptosis are inhibited by NEK2. Saos-2 cells were transfected with indicated plasmids. Lysate immunoblots were probed with indicated antibodies. (*Right*) Quantitation of the blot data. (*F*) NEK2 inhibits p53-induced apoptosis. Saos-2 cells transfected with indicated vectors were stained posttransfection with p53-FITC antibody and DAPI nuclear stain. Green fluorescent cells were examined for nuclear damage, and percentage of apoptotic cells (hypercondensed DAPI and FITC-stained nuclei) were quantitated relative to total fluorescent transfected cells. Asterisks (*) at the *Right* indicate apoptotic nuclei, and solid triangles indicate nonapoptotic nuclei containing p53. The average percentage of apoptosis from two experiments with SEM was quantitated and is illustrated in the graph. (Magnification: *F*, *Right*, 20×.) *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$; NS, $P \geq 0.05$.

phosphorylation of p53, there may be other direct effects of NEK2 (e.g., protein–protein interactions) on p53 function as well as indirect effects (e.g., phosphorylation of other mitotic p53 kinases) that negatively impact p53 activities in mitosis. We also showed that NEK2 destabilization of p53 is associated with other p53-mediated functions such as transcriptional transactivation and apoptosis induction. As a mitotic kinase, NEK2 has been shown to be overexpressed in many human cancers, and its overexpression is a marker of poor prognosis in several cancer types (36, 37). NEK2 is currently being targeted by small-molecule inhibitors (25, 38). Comparison of the *TP53* mutational status of multiple cancer types from The Cancer Genome Atlas dataset with NEK2 RNA expression status revealed that NEK2 overexpression is strongly correlated with *TP53* mutation (*SI Appendix*, Table S2), consistent with an observed inhibition of NEK2 RNA expression in the presence of stabilized p53 (29). Thus, NEK2 and p53 may be part of a negative feedback autoregulatory loop that becomes dysfunctional when *TP53* is mutated (*SI Appendix*, Fig. S7).

More broadly, this study validates a systematic approach to analyze and reason from unstructured data from the scientific literature to obtain useful scientific hypotheses. This approach has four steps: (*i*) it embeds (39) and compares the word content of the full corpus of biomedical abstracts; (*ii*) this generates a structured network of protein entities; (*iii*) in turn, this network supports semisupervised learning (40); (*iv*) from which predictions of specific types of protein–protein interactions follow. Thus, the integration of text mining with network-based machine learning has led to automated hypothesis generation. Compared with alternative approaches such as network and protein interaction analyses and amino acid sequence alignment against unknown p53 kinases, this literature-based approach considers everything that has been published about each protein, which may be missing in manually curated networks, amino acid sequence information, or experimental interaction values.

While the experimental validations presented above were necessarily focused on a narrow context, p53 kinase activity, the algorithms discussed here should be applicable to other proteins of interest and are not intrinsically limited to predictions of protein–protein interactions. For example, diffusion of time-stamped labels on CTNNB1, GSK3β, and HIST1H3B kinase networks resulted in top-scoring unlabeled nodes that were enriched for relevant kinases discovered after the time stamp (area under the ROC curve = 0.65 for CTNNB1, 0.66 for GSK3β, and 0.68 for HIST1H3B) (*SI Appendix*, *Materials and Methods*, Fig. S8, and Table S3), suggesting that this approach can be used to discover kinases for other proteins that are not as intensively studied as p53. This method has also been used to find prion proteins associated with amyotrophic lateral sclerosis (41). In the future, these algorithms could be expanded to identify hidden connections among many types of biological entities, leading to accelerated discovery in many areas of biological science.

Materials and Methods

Please see *SI Appendix*, *Materials and Methods*, for a detailed description of materials and methods.

To identify p53 kinases, a similarity network of human kinases was built from all PubMed abstracts available in January 2014 that mentioned human kinases. The abstracts for each kinase were summarized as vectors of relevant words (30), and a cosine similarity matrix was produced for the 259 kinases that each had 10 or more abstracts. We converted the matrix into a graph by thresholding the kinases based on their similarity. We then used graph information diffusion (17) to globally propagate labels throughout this kinase network from the known p53 kinase nodes to nodes lacking such prior annotation. This produced a ranking of all kinases not previously known to phosphorylate p53 for functional testing (described in *SI Appendix*).

While it is not possible or practical to release the entire source code for Watson for Drug Discovery, we are happy to make the latest implementation available to any researcher who wants to try out the approach that we describe in this paper. The URL ibm.biz/wdd-trial provides the application for a cloud-based service that scientists and investigators can easily apply to their own problems or use to validate our method against the example described in a paper or any other similar example that they wish to explore. This application is provided free of charge for a period of one month after activation, with extensions possible upon request.

ACKNOWLEDGMENTS. We thank Carl Anderson for a careful reading of the manuscript and Jack Manquen and Danielle Jayanty for technical assistance. This work was supported in part by a gift from the Robert and Janice McNair Medical Foundation to the Computational and Integrative Biomedical Research Center at Baylor College of Medicine; by Defense Advanced Research Projects Agency (DARPA) Contract Grants N66001-14-1-4027 and N66001-15-C-4042; by IBM Research and the IBM Accelerated Discover Lab; and by NIH Grants GM079656 and NSF DBI 1356569 (to O.L.). Imaging was supported by the Integrated Microscopy Core, and mass spectrometry was supported by the Mass Spectrometry Proteomics Core at Baylor College of Medicine. Funding was also provided by the NIH (HD007495, DK56338, and CA125123); the Dan L. Duncan Cancer Center; and the John S. Dunn Gulf Coast Consortium for Chemical Genomics.

1. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. *Nature* 521:436–444.
2. Ghahramani Z (2015) Probabilistic machine learning and artificial intelligence. *Nature* 521:452–459.
3. Blaschke C, Valencia A (2013) The functional genomics network in the evolution of biological text mining over the past decade. *N Biotechnol* 30:278–285.
4. Schneider JH (1971) Selective dissemination and indexing of scientific information. *Science* 173:300–308.
5. Waltz D, Buchanan BG (2009) Computer science. Automating science. *Science* 324:43–44.
6. Gershman SJ, Horvitz EJ, Tenenbaum JB (2015) Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science* 349:273–278.
7. Krallinger M, Valencia A (2005) Text-mining and information-retrieval services for molecular biology. *Genome Biol* 6:224.
8. Rindflesch TC, Fiszman M (2003) The interaction of domain knowledge and linguistic structure in natural language processing: Interpreting hypernymic propositions in biomedical text. *J Biomed Inform* 36:462–477.
9. Rzhetsky A, Foster JG, Foster IT, Evans JA (2015) Choosing experiments to accelerate collective discovery. *Proc Natl Acad Sci USA* 112:14569–14574.
10. Kang N, Singh B, Afzal Z, van Mulligen EM, Kors JA (2013) Using rule-based natural language processing to improve disease normalization in biomedical text. *J Am Med Inform Assoc* 20:876–881.
11. Torvik VI, Smalheiser NR (2007) A quantitative model for linking two disparate sets of articles in MEDLINE. *Bioinformatics* 23:1658–1665.
12. Katukuri JR, Xie Y, Raghavan VV, Gupta A (2012) Hypotheses generation as supervised link discovery with automated class labeling on large-scale biomedical concept networks. *BMC Genomics* 13(Suppl 3):S5.
13. Cameron D, et al. (2013) PREDOSE: A semantic web platform for drug abuse epidemiology using social media. *J Biomed Inform* 46:985–997.
14. Swanson DR (1986) Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspect Biol Med* 30:7–18.
15. Barabási AL, Gulbahce N, Loscalzo J (2011) Network medicine: A network-based approach to human disease. *Nat Rev Genet* 12:56–68.
16. Guney E, Menche J, Vidal M, Barábasi AL (2016) Network-based in silico drug efficacy screening. *Nat Commun* 7:10331.
17. Lisewski AM, et al. (2014) Supergenomic network compression and the discovery of EXP1 as a glutathione transferase inhibited by artesunate. *Cell* 158:916–928.
18. Spangler S, et al. (2014) Automated hypothesis generation based on mining scientific literature. *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Available at https://www-01.ibm.com/common/ssi/cgi-bin/ssialias?htmlfid=HLW03046USEN&. Accessed September 23, 2018.
19. Nagarajan M, et al. (2015) Predicting future scientific discoveries based on a networked analysis of the past literature. *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Available at https://dl.acm.org/citation.cfm?id=2788609. Accessed September 25, 2018.
20. Kandoth C, et al. (2013) Mutational landscape and significance across 12 major cancer types. *Nature* 502:333–339.
21. Soussi T (2014) The TP53 gene network in a postgenomic era. *Hum Mutat* 35:641–642.
22. Vousden KH, Prives C (2009) Blinded by the light: The growing complexity of p53. *Cell* 137:413–431.
23. Brognard J, Hunter T (2011) Protein kinase signaling networks in cancer. *Curr Opin Genet Dev* 21:4–11.
24. Roskoski R, Jr (2015) A historical overview of protein kinases and their targeted small molecule inhibitors. *Pharmacol Res* 100:1–23.
25. Yu X, Narayanan S, Vazquez A, Carpizo DR (2014) Small molecule compounds targeting the p53 pathway: Are we finally making progress? *Apoptosis* 19:1055–1068.
26. Hunter L, Cohen KB (2006) Biomedical language processing: What's beyond PubMed? *Mol Cell* 21:589–594.
27. Radivojac P, et al. (2013) A large-scale evaluation of computational protein function prediction. *Nat Methods* 10:221–227.
28. Mallory EK, Zhang C, Ré C, Altman RB (2016) Large-scale extraction of gene interactions from full-text literature using DeepDive. *Bioinformatics* 32:106–113.
29. Hirschberg J, Manning CD (2015) Advances in natural language processing. *Science* 349:261–266.
30. Baeza-Yates R, Ribeiro-Neto B (1999) *Modern Information Retrieval* (ACM Press, New York), Vol 463.
31. Fry AM, O'Regan L, Sabir SR, Bayliss R (2012) Cell cycle regulation by the NEK family of protein kinases. *J Cell Sci* 125:4423–4433.
32. Marina M, Saavedra HI (2014) Nek2 and Plk4: Prognostic markers, drivers of breast tumorigenesis and drug resistance. *Front Biosci* 19:352–365.
33. Katayama H, et al. (2004) Phosphorylation by aurora kinase A induces Mdm2-mediated destabilization and inhibition of p53. *Nat Genet* 36:55–62.
34. Wang Y, Prives C (1995) Increased and altered DNA binding of human p53 by S and G2/M but not G1 cyclin-dependent kinases. *Nature* 376:88–91.
35. Nabilsi NH, et al. (2013) Local depletion of DNA methylation identifies a repressive p53 regulatory region in the NEK2 promoter. *J Biol Chem* 288:35940–35951.
36. Takahashi Y, et al. (2014) Up-regulation of NEK2 by microRNA-128 methylation is associated with poor prognosis in colorectal cancer. *Ann Surg Oncol* 21:205–212.
37. Zhong X, Guan X, Liu W, Zhang L (2014) Aberrant expression of NEK2 and its clinical significance in non-small cell lung cancer. *Oncol Lett* 8:1470–1476.
38. Hu CM, et al. (2015) Novel small molecules disrupting Hec1/Nek2 interaction ablate tumor progression by triggering Nek2 degradation through a death-trap mechanism. *Oncogene* 34:1220–1230.
39. Charu CA, Chengxiang Z (2012) *Mining Text Data* (Kluwer, Boston).
40. Ben-Hur A, Ong CS, Sonnenburg S, Schölkopf B, Rätsch G (2008) Support vector machines and kernels for computational biology. *PLOS Comput Biol* 4:e1000173.
41. Bakkar N, et al. (2018) Artificial intelligence in neurodegenerative disease research: Use of IBM Watson to identify additional RNA-binding proteins altered in amyotrophic lateral sclerosis. *Acta Neuropathol* 135:227–247.