# Exploring the understudied human kinome for research and therapeutic opportunities

Nienke Moret[1,2,*], Changchang Liu[1,2,*], Benjamin M. Gyori[2], John A. Bachman[2], Albert Steppi[2], Rahil Taujale[3], Liang-Chin Huang[3], Clemens Hug[2], Matt Berginski[1,4,5], Shawn Gomez[1,4,5], Natarajan Kannan,[1,3] and Peter K. Sorger[1,2,†]

*These authors contributed equally

[†] Corresponding author

[1]The NIH Understudied Kinome Consortium

[2]Laboratory of Systems Pharmacology, Department of Systems Biology, Harvard Program in Therapeutic Science, Harvard Medical School, Boston, Massachusetts 02115, USA

[3] Institute of Bioinformatics, University of Georgia, Athens, GA, 30602 USA

[4] Department of Pharmacology, The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

[5] Joint Department of Biomedical Engineering at the University of North Carolina at Chapel Hill and North Carolina State University, Chapel Hill, NC 27599, USA

**Key Words:** kinase**,** human kinome, kinase inhibitors, drug discovery, cancer, cheminformatics,

[†] Peter Sorger
Warren Alpert 432
200 Longwood Avenue
Harvard Medical School,
Boston MA 02115
peter_sorger@hms.harvard.edu cc: sorger_admin@hms.harvard.edu
617-432-6901

**ORCID Numbers**
Peter K. Sorger 0000-0002-3364-1838
Nienke Moret 0000-0001-6038-6863
Changchang Liu 0000-0003-4594-4577
Ben Gyori 0000-0001-9439-5346
John Bachman 0000-0001-6095-2466
Albert Steppi 0000-0001-5871-6245

## ABSTRACT

The functions of protein kinases have been heavily studied and inhibitors for many human kinases have been developed into FDA-approved therapeutics. A substantial fraction of the human kinome is nonetheless understudied. In this paper, members of the NIH Understudied Kinome Consortium mine public data on "dark" kinases to estimate the likelihood that they are functional. We start with a re-analysis of the human kinome and describe the criteria for creation of an inclusive set of 710 kinase domains and a curated set of 557 protein kinase like (PKL) domains. Nearly all PKLs are expressed in one or more CCLE cell lines and a substantial number are also essential in the Cancer Dependency Map. Dark kinases are frequently differentially expressed or mutated in The Cancer Genome Atlas and other disease databases and investigational and approved kinase inhibitors appear to inhibit them as off-target activities. Thus, it seems likely that the dark human kinome contains multiple biologically important genes, a subset of which may be viable drug targets.

## INTRODUCTION

Protein phosphorylation is widespread in eukaryotic cells[1] and mediates many critical events in cell fate determination, cell cycle control and signal transduction[2]. The structures[3] and catalytic activities[4] of eukaryotic protein kinases (ePKs), of which more than 500 are found in humans[5], have been intensively investigated for many years: to date, structures for over 280 unique domains and ~4,000 co-complexes have been deposited in the PDB database.. The ePK fold is thought to have arisen in procaryotes[6] and evolved to include tyrosine kinases in metazoans[7,8], resulting in a diverse set of enzymes[9,10] that are often linked in a single protein to other catalytic domains and to SH2, SH3 and protein binding domains. In addition, 13 human proteins have two ePK kinase domains. An excellent recent review describes the structural properties of ePKs and the drugs that bind them[11].

The kinase domain of protein Kinase A (PKA), a hetero-oligomer of a regulatory and catalytic subunit, was the first to be crystalized and is often regarded as the prototype of the ePK fold.[3,12] It involves two distinct lobes with an ATP-binding catalytic cleft lying between the lobes. With respect to sequence, ePK are characterized by 12 recurrent elements involving ~30 highly conserved residues. The kinase fold is remarkably adaptable however, and has diverged in multiple ways to generate protein families distinct in sequence and structure from PKA. The eukaryotic like kinases (eLKs) retain significant sequence similarity to the N-terminal region of ePKs but differ in the substrate binding lobe; choline kinase A (CHKA) is a well-studied example of an eLK[13]. Kinases with an atypical fold (aPKs) have weak sequence similarity to ePKs, but nevertheless adopt an ePK like structural fold and include some well-studied kinases such as the DNA damage sensing ATM and ATR enzymes as well as lipid kinases such as PI3K, one of the most heavily mutated genes in breast cancer[14].

In humans, ePKs, eLKs and aPKs are conventionally organized into ten groups based on sequence alignment and structure; this often corresponds to modes of regulation and function. For example, tyrosine kinases represent a distinct branch of the kinome tree that includes 58 human receptor tyrosine kinases[15] (RTKs) that bind extracellular ligands (growth factors) and share an extended regulatory spine that allosterically controls catalytic activity[16]. The AGC group of kinases, in contrast, are regulated by a conserved C-terminal tail flanking the kinase domain[17]. Over 200 additional proteins annotated as “kinase” in UniProt but are unrelated to the protein kinase fold enzymes and therefore termed uPKs (unrelated to Protein Kinases). Enzymes with phosphotransferase activity in the uPKs family include hexokinases that phosphorylate sugars and STK19[18], which displays peptide-directed phosphotransferase activity and also binds protein kinase inhibitors.

The human kinome[5] includes ~50 pseudokinases that lack one or more residues generally required for catalytic activity. These residues include the ATP -binding lysine (K) within the VAIK motif, the catalytic D within the HRD motif and the magnesium binding D within the DFG motif[19].

Many pseudokinases function in signal transduction despite the absence of key catalytic residues. For example, the EGFR family member ERBB3/HER3 is a pseudokinase that, when bound to ERBB2/HER2, forms a high affinity receptor for heregulin growth factors.[20] ERBB3 over-expression also promotes resistance to therapeutic ERBB2 inhibitors in breast cancer[21]. Some proteins commonly annotated as pseudokinases even have phospho-transfer activity. Haspin, for example, is annotated as a pseudokinase in the ProKino database because it lacks a DFG motif in the catalytic domain, but it has been shown to phosphorylate histone H3 using a DYT motif instead;[22,23] H3 phosphorylation by Haspin changes chromatin structure and mitotic outcome and is therefore physiologically important[24].

Protein kinase inhibitors, and the few activators that have been identified (e.g. AMPK activation by salicylate and A-769662[25]), are diverse in mechanism and structure. The molecules include ATP-competitive inhibitors that bind in the enzyme active site and non-competitive "allosteric" inhibitors that bind outside the active site, small molecule PROTAC degraders whose binding to a kinase promotes ubiquitin-dependent degradation[26] and antibodies that target the growth factor or ligand binding sites of receptor kinases or that interfere with a receptor's ability to homo or hetero-oligomerize[27]. Kinase inhibitors have been intensively studied in human clinical trials and over 50 have been developed into FDA-approved drugs[11].

A substantial subset of the kinome has been little studied, despite the general importance of kinases in cellular physiology, their druggability and their frequent mutation in disease. This has given rise to a project within the NIH's *Illuminating the Druggable Genome*[28] (IDG) Program, to investigate the understudied "dark kinome" and determine its role in human biology and disease[29]. IDG has distributed a preliminary list of dark kinases based on estimates of the number of publications describing that kinase and the presence/absence of grant (NIH R01) funding; we and others have started to study the properties of these enzymes[30]. As described in greater detail below, defining the dark kinome necessarily involves a working definition of the full kinome and a survey of the current state of

knowledge. The starting point for this survey is the standard list of kinases put forward in a groundbreaking 2002 paper by Manning et al[5] that found the human kinome to have 514 members; this has subsequently been updated via the KinHub Web resource[31] to include 522 human kinases (although many papers cite a number closer to 520-540).

While protein kinases could in principle be defined strictly as enzymes that catalyze phospho-transfer from ATP onto serine, threonine and tyrosine, such a definition would exclude biologically active pseudokinases and structurally and functionally related lipid kinases. It would also fail to account for a lack of functional data for a substantial number of proteins, potentially excluding kinases that are physiologically or catalytically active. An alternative definition uses sequence alignment and structural data to identify closely related folds, but excludes uPKs having kinase activity as well as bromodomains that are potently bound and inactivated by kinase inhibitors[32]. A less restrictive list is useful for the kinome-wide activity profiling that is a routine part of kinase-focused drug discovery. Profiling typically involves screening compounds against panels of recombinant enzymes (e.g. KINOMEscan)[33] or chemo-proteomics in which competitive binding to ATP-like ligands on beads (so-called kinobeads[34] or multiplexed inhibitor beads - MIBs[35]) is assayed using mass spectrometry. Such screens benefit from a comprehensive list of binding domains for which selectivity can be assayed.

In this perspective we analyze the composition and properties of the dark kinome, with a focus on evidence that understudied kinases are expressed and potentially functional in normal cellular physiology and in disease. As a first step we generate new lists for membership in the full kinome based on a variety of inclusion and exclusion criteria. We also re-compute membership in the dark kinome and consolidate available data on dark kinase activity and function. This evidence is typically indirect, such as data from TCGA (The Cancer Genome Atlas[36]) on the frequency with which a kinase is mutated in particular type of cancer. In aggregate, however, the evidence strongly suggests that the understudied kinome is likely to contain many enzymes worthy of in-depth study, a subset of which may be viable

therapeutic targets. All of the information in this manuscript is available in supplementary materials, and is currently being curated and released via the dark kinome portal.

## RESULTS

### The composition of the human kinome

A list of human kinases was obtained from Manning et al.[5] (referred to below as 'Manning') and a second from Eid et al.[31] (via the Kinhub Web resource); a list of dark kinases according to IDG was obtained from the NIH solicitation[37] (updated in January 2018) and a fourth list of all 684 proteins tagged as "kinases" was obtained from UniProt. These lists are overlapping but not identical (**Figure 1A**). For example, eight IDG dark kinases absent from Manning and Kinhub (CSNK2A3, PIK3C2B, PIK3C2G, PIP4K2C, PI4KA, PIP5K1A, PIP5K1B, and PIP5K1C) are found in the UniProt list. We therefore assembled a superset of 710 domains (the "extended kinome") and used curated alignment profiles and structural analysis[38] to subdivide the domains into the primary categories: "Protein Kinase Like" (PKL), if the kinase domain was similar to known protein kinases in sequence and 3D-structure; "Unrelated to Protein Kinase" (uPK), if the kinase domain was distinct from known protein kinases; and "Unknown" if there was insufficient information to decide (see methods).[38] PKLs were further subdivided into eukaryotic protein kinases (ePKs), eukaryotic like kinases (eLKs) and kinases with an atypical fold (aPKs) as previously described.[38,39] ePKs and eLKs share detectable sequence similarity in the ATP binding lobe and some portions of the substrate binding lobe (up to the conserved F-helix[38]). aPKs, on the other hand, display no significant sequence similarity to ePKs and eLKs, but nevertheless adopt the canonical protein kinase fold. Most aPK lack the canonical F-helix aspartate in the substrate binding lobe, but share structural similarities with ePKs and eLKs in the ATP binding lobe (**Figure 1B)**[38]. Unfortunately, the nomenclature for these families is not consistent across sources. In this perspective aPK refers to a subset of PKLs defined by fold and sequence similarity; this is distinct from

the so-called "atypical protein kinases" (AKGs). These domains are typically depicted off to the side of Coral dendogram[40] and include protein kinases such as ATM and ATR as well as bromo-domains and TRIM proteins (see below).

As noted previously[5,21], structural, sequence-based and functional classifications of kinases are often ambiguous and overlapping. For example, the ATM aPK is known to phosphorylate proteins DYRK2, MDM2, MDM4 and TP53[41] when activated by DNA double-strand breaks and it is also a member of the six-protein family of phosphatidylinositol 3-kinase-related protein kinases (PIKKs). The PIKK family has a protein fold significantly similar to lipid kinases in the PI3K/PI4K family but PI4K2A, for example, modifies phosphatidylinositol lipids not proteins.[42] Thus, even after extensive computational analysis, some manual curation of the kinome is necessary. We have therefore created a sortable table enumerating all of the inclusion and exclusion criteria for individual kinases described in this perspective; it is possible to generate a wide variety sublists from this table based on user-specific criteria (**Supplemental Table S1**).

One drawback of the 710 extended kinome set is that it is substantially larger than the 525-550 domains commonly regarded as comprising the set of human protein kinases. We therefore created a second "curated kinome" comprising 557 domains (544 genes) that includes all 556 PKLs plus the uPK STK19 (**Supplemental Table S2)**; this list omits 15 uPKs found in Manning and 22 found in Kinhub (including multiple TRIM family proteins[43] that regulate and are regulated by kinases[44], but have no known known intrinsic kinase activity). The shorter list also omits bromodomains. The curated 557-domain kinome and the Manning list are compared in **Figure 1C** and **Figure S1A.**

The relevance of the extended kinome to the study of protein kinases and kinase inhibitors is demonstrated in part by re-analysis of a large-scale chemo-proteomic dataset collected using multiplexed inhibitor beads.[34] Overall, 48 domains found in the extended kinome list and not in the curated list, were found to bind to kinobeads and 8 were competed-off in the presence of a kinase

inhibitor, the criterion for activity in this assay (**Figure S1B**). Pyridoxal kinase (PDXK) and adenosine kinase (ADK) were among the enzymes bound by kinase inhibitors, even though these proteins are not conventionally considered when studying kinase inhibitor mechanism of action. We conclude that the extended and curated kinomes are useful in different settings.

**Identifying understudied kinases**

The original IDG dark kinome list was assembled using a bibliometric tool, TIN-X[45], that uses natural language processing (NLP) of PubMed abstracts to assess the "novelty" and " disease importance" of a gene or protein. We have previously found that different ways of performing. bibliometric evaluation yield varying results when applied to the kinome[30]. We therefore took a complementary approach based on the recently developed computational tool INDRA (the Integrated Network and Dynamical Reasoning Assembler)[46,47]. INDRA uses multiple NLP systems to extract computable statements about biological mechanism[48,49] found in PubMed abstracts and full text articles in PubMedCentral. It also aggregates data from multiple pathway databases (such as BioGrid[50] and PathwayCommons[51]) and specialized resources such as the LINCS compound database[52] and the Target Affinity Spectrum from the Small Molecule Suite database[53].

INDRA statements consolidate redundant sources of evidence, link it to the underlying knowledge support (the database reference or citation) and they are frequently detailed with respect to molecular mechanism. For example, the INDRA network for the WEE2 dark kinases (**Figure 2A**) includes statements such as "*Phosphorylation(WEE2(), CDK1())*" and "*Inhibition(WEE2(), CDK1()).*" These machine readable assertions state that WEE2 is active in mediating an inhibitory phosphorylation event on CDK1 (**Figure 2A**). INDRA associates each assertion with its underlying evidence (including database identifiers or specific sentences extracted from text and their PMIDs). INDRA also consolidates overlapping and redundant information: in many cases a single assertion has multiple

pieces of evidence (for example, three PMID citations for the phosphorylation reaction described above). Collections of INDRA Statements can be visualized as networks of mechanisms comprising proteins, small molecules and other biological entities. Thus, INDRA can be used to efficiently explore available information on proteins and protein families.

We generated INDRA networks for all members of the curated kinome and used the number of mechanistic statements as a quantitative measure of knowledge about each kinase; these networks are available via NDEx[54]. We found that prior knowledge about the curated kinome as extracted by INDRA varied by $>10^4$ fold and was correlated with the TIN-X "novelty" score (Pearson's correlation coefficient=0.81). There were some cases in which the two measures were discordant; for example, PI4K2A has 78 INDRA statements, but a high TIN-X novelty score of ~808. The reason for such inconsistency is still under investigation but is likely to reflect the difficulty of linking common names for genes and proteins to their unique identifies in resource such as HGNC (this is known as the process of entity grounding); INDRA has extensive resources to correctly ground entities and resolve ambiguities and can correctly associate MEK kinase with the HGNC name MAP2K1 and not "methyl ethyl ketone.".

To estimate the intensity of drug development for each kinase we used the *Small Molecule Suite*[53], which mines diverse cheminformatic resources to determine which kinases are bound by small molecules in a most-selective (MS) and semi-selective (SS) fashion (**Figure 2C**) as well as PHAROS[55] , which classifies targets based on whether or not they are bound by an FDA-approved drug (Tclin) or tool compound (Tchem). The selectivity levels in the Small Molecule Suite are assigned to target-compound interactions (rather than to compounds *per se*) based on available data on the absolute binding affinity (typically obtained from enzymatic or quantitative protein binding assays), differential "on target" affinity as compared to the "off-target affinity" (typically obtained from a kinase profiling assay), the *p-value* between the distributions for "on" and "off" targets, and "research bias"; the latter

accounts for differences in available binding data (in the absence of bias estimate, a poorly studied compound can appear much more selective than a well-studied one simply because few off-targets have been tested). The MS assertion is assigned to compounds that have an absolute affinity <100 nM, an on-target Kd > 100 times lower than off target Kd, p-value of ≤0.1, and research bias ≤0.2 (see Moret et al.[53] for details). The SS assertion is about 10-fold less strict with regard to absolute and differential affinity (see methods). We found that kinases that were more heavily studied were more likely to have inhibitors classified as Tclin and Tchem in *PHAROS* or MS or SS in *Small Molecule Suite*. However, a substantial number of kinases with high INDRA scores are bound by only relatively non-selective inhibitors and therefore represent opportunities for development of new chemical probes.

The original NIH IDG dark kinase list encompassed approximately one-third of the kinome. Using INDRA scores, we generated a new list of similar scope (schematized by the magenta box in **Figure 2A, 2B**) of the 182 least-studied domains in 181 proteins in the curated kinome, of which 119 were on the original NIH list and 156 in Manning or KinHub (**Figure 2D**). In the analysis that follows we use this recomputed dark kinase list as the "dark kinome". When the distribution of dark kinases is viewed using the standard Coral kinase dendrogram[40], a remarkably even distribution is observed across subfamilies, with the exception that only eight tyrosine kinases are judged to be understudied (**Figure 3**). In many cases light and dark kinases are intermingled on the dendrogram (e.g. the CK1 subgroup) but in some cases an entire sub-branch is dark (e.g. one with 4 TSSK and another with 3 STK32 kinases; dashed red outline).

**Evidence for dark kinase expression and function**

To consolidate existing data on the expression and possible functions of understudied kinases, we analyzed on-line resources including RNAseq data for 1019 cell lines in the Cancer Cell Line Encyclopedia (CCLE)[56], proteomic data for 375 cell lines in the CCLE[56] and loss of function data in the

Cancer Dependency Map (DepMap).[57] DepMap data were generated by using lentivius-based RNAi or CRISPR/Cas9 libraries in pooled screens to identify essential genes across a large panel of cell lines.

Based on RNASeq data, non-dark and dark kinases were observed to vary substantially in abundance across 1019 CCLE cell lines but evidence of expression (using the common threshold of RPKM ≥1 (Reads Per Kilobase of transcript, per Million mapped reads))[58] was obtained in at least one line for 176 of 181 dark kinases (**Figure 4A**). Some dark kinases were as highly expressed as well-studied light kinases: for example, NRBP1 and PAN3 and the PI4KA and PIP4K2C lipid kinases all had maximum expression levels similar to that of the abundant and well-studied LCK tyrosine kinase. Overall, however, dark kinases had lower maximum expression than non-dark kinases by multiple measures (2.1 vs 5.8 RPKM median expression level, p-value=$4.6 \times 10^{-8}$; 36 vs 71 RPKM maximum expression level, p-value=$2.2 \times 10^{-16}$ Wilcoxon rank sum test). In CCLE proteomic data we found that 367 kinases from the curated kinome could be detected with at least one peptide per protein; 110 of these are dark kinases. Analysis of DepMap data showed that 10 dark kinases are essential in at least 1/3 of the 625 cell lines tested to date (**Figure 4B;** dark blue shading**)**, and 88 kinases are essential in at least two lines (light blue shading**)**. Thus, a substantial number of dark kinases are expressed in human cells lines and a subset are known to be required for cell growth. These data are likely to underestimate the breadth of kinase expression and function: proteins can impact cellular physiology when expressed at low levels and genes can have important functions without necessarily resulting in growth defects assayable by DepMap methodology.

**Data on dark kinases in diseases**

To study the roles of dark kinases in pathophysiology, we mined online databases of associations between disease and gene mutations or changes in expression, starting with TCGA, the largest such database. We compared the frequency of mutations in dark and non-dark kinases under the assumption

that the two sets of kinases are characterized by the same ratio of passenger to driver mutations[59] and looked for differential RNA expression relative to matched normal tissue (**Figure 5A)**. In common with most TCGA analysis, mutations and differential expression were scored at the level of genes and not domains and thus, functions other than kinase activity might be affected. We performed these analyses for individual tumor types and for all cancers as a set (the PanCan set). With respect to differential gene expression, we found that dark and light kinases are equally likely to be over or under-expressed in both PanCan data and in data for specific types of cancer (in a Rank-sum test with $H_0$ = light and dark kinases have similar aberrations p=0.15) **(Figure 5)**. For example, in colorectal adenocarcinoma the dark kinases STK31, LMTK3, NEK5 and PKMYT1 represent four of the seven most high upregulated kinases whereas MAPK4 is one of the three most highly downregulated (**Figure S3A**). In PanCan, we also found that five dark kinases were among the 30 most frequently mutated human kinases; for example, the ~3% mutation frequency of the dark MYO3A kinase is similar to that of the oncogenic RTKs EGFR and ERBB4 (but lower than the ~12% mutation frequency for the lipid kinase PIK3CA) (**Figure 4B**). Similarly, in diffuse large B-cell lymphoma, the dark kinase ITPKB is more frequently mutated than KDR/VEGFR2 (~13% vs. 8% of patient samples, **Figure S3B**); over-expression of KDR is known to promote angiogenesis and correlate with poor overall patient survival and poor response to immunotherapy[60–62]. Recurrent mutation, over-expression and under-expression in TCGA data is not evidence of biological significance per se, but systematic analysis of TCGA data has been remarkably successful in identifying genes involved in cancer initiation, progression, and drug resistance[52]. Our analysis therefore shows that dark kinases are nearly as likely to be mutated or differentially expressed in human cancer as their better studied non-dark kinase homologues.

To explore the roles of dark kinases in other diseases, we analyzed data from the AMP-AD program (Accelerating Medicines Partnership - Alzheimer's Disease (AD) Target Discovery and Preclinical Validation)[64]. This large program aims to identify molecular features of AD at different

disease stages. We compared mRNA expression at the earliest stages of AD to those from late-stage disease in age matched samples (**Figure 5C**) and found that the dark kinases ITPKB and PKN3 were among the five most upregulated kinases while NEK10 was substantially downregulated. A similar analysis was performed for Chronic obstructive pulmonary disease (COPD), a common disease that progressively impairs a patients' ability to breathe and is the third leading cause of death in the US[65]. A study by Rogers et al[66] analyzed five COPD microarray datasets from Gene Expression Omnibus (GEO) and two COPD datasets from ArrayExpress to identify genes with significant differential expression in COPD. By comparing the expression of genes in COPD patients to gene expression in healthy individuals, Rogers et al. identified genes significantly up and down regulated in COPD patients (adjusted p-value < 0.05). We analyzed these data and found that the dark kinase PIP4K2C, which is potentially immune regulating,[67] was significantly downregulated in individuals with COPD (adjusted p-value = 0.048). Additionally, CDC42BPB, nominally involved in cytoskeleton organization and cell migration,[68,69] was upregulated in COPD patients (adjusted p-value = 0.026) (**Figure 5D**). In total, 5 dark kinases versus 15 non-dark kinases were expressed differentially in COPD patients. As additional data on gene expression and mutation become available for other pathophysiologies, it will be possible further expand the list of dark kinases potentially implicated in human disease.

**A dark kinase network regulating the cell cycle**

Inspection of INDRA networks revealed that multiple dark kinases are likely to function in networks of interacting kinases. One illustrative example involves regulation of the central regulator of cell cycle progression, CDK1, by the dark kinases PKMYT1, WEE2, BRSK1 and NIM1K **(Figure 6)**. WEE2, whose expression is described to be oocyte-specific[70] (but can is also be detected in seven CCLE lines, six from lung cancer and one from large intestine) as well as its well-studied and widely-expressed homologue WEE1, phosphorylate CDK1 on the negative regulatory site T15[70] whereas PKMYT1

phosphorylates CDK1 on the Y14 site to complete the inhibition of CDK1[71,72]. These modifications are removed by the CDC25 phosphatase, which promotes cell cycle progress from G2 into M phase[73]. PKMYT1 and WEE1 are essential in nearly all cells, according to DepMap[74] (although WEE2 is not). Upstream of WEE1, the dark kinases BRSK1 (127 INDRA statements) and NIM1K (28 INDRA statements) and the well-studied BRSK2 (176 INDRA statements) function to regulate WEE1. Neither PKMYT1, BRSK1 and NIM1K have selective small molecule inhibitors described in the public literature[75]; several WEE1 inhibitors are in clinical development[76], and these are molecules are likely to inhibit WEE2 as well. It is remarkable that enzymes so closely associated with the essential cell cycle regulator CDK1, including several whose homologues have been extensively studied in fission and budding yeast, remain relatively understudied in humans[77]. This is particularly true of PKMYT1 and NIM1K which are frequently upregulated in TCGA data.

**Inhibition of dark kinases by approved drugs**

Kinase inhibitors, including those in clinical development or approved as therapeutic drugs, often bind multiple targets. We therefore asked whether dark kinases are targets of investigational and FDA-approved drugs by using the *selectivity score*[53] to mine public data for evidence of known binding and known not-binding. We identified 13 dark kinases that are inhibited by approved drugs and an additional 12 dark kinases for which MS or SS inhibitors exist among compounds that have entered human trials (although several of these are no longer in active development). The anti-cancer drug sunitinib, for example, is described in the Small Molecule Suite database as binding to the dark kinases STK17A, PHGK1 and PHGK2 with binding constants of 1 nM, 5.5 nM and 5.9 nM respectively (**Figure 7A, Table S3**) as opposed to 30 nM to 1 μM for VEGF receptors (the KDR, FLT1 and FLT4 kinases) and 200 nM for PDGFRA, well established targets for sunitinib. Follow-on biochemical and functional

experiments will be required to determine if dark kinases play a role in the therapeutic mechanisms of these and other approved drugs.

The potential for development of new compounds that inhibit dark kinases based on modification of existing kinase inhibitors can be assessed in part by examining the structures of kinase binding pockets using Bayes Affinity Fingerprints (BAFP)[78,79]. In this cheminformatics approach, each small molecule in a library is computationally decomposed into a series of fragments using a procedure known as fingerprinting. The conditional probability of a compound binding to a specific target (as measured experimentally in profiling or enzymatic assays) given the presence of a chemical fragment is then calculated. Each target is thereby associated with a vector comprising conditional probabilities for binding fragments found in the fingerprints of compounds in the library. Subsequently, the correlation of conditional probability vectors for two proteins is used to evaluate similarity in their binding pockets from the perspective of a chemical probe. BAFP vectors were obtained from a dataset of ~5 million small molecules and 3000 targets for which known binding and non-binding data are available from activity profiling.

We found that the majority of kinase domains fell in two clusters, each of which had multiple dark and non-dark kinases. The close similarity of dark and non-dark kinases in "compound binding space" suggests that many more kinase inhibitors than those described in **Figure 7a** may already bind dark kinases or could be modified to do so (**Figure 7B, Figure S5**). For example, the clustering of IRAK1, IRAK4, STK17B and MAP3K7 by BAFP correlation (highlighted in **Figure 7B**) demonstrates that the STK17B binding pocket is likely very similar to that of IRAK1, IRAK4 and MAP3K7 and that compounds binding these non-dark kinases, such as lestaurtinib and tamatinib may also bind STK17B. Based on this, it may be possible to design new chemical probes with enhanced selectivity for STK17B by starting with the libraries derived from lestaurtinib or tamatinib.

Other useful tools for development of new small molecule probes are commercially available activity assays and experimentally determined NMR or crystallographic protein structures. Of 181 dark kinases 101 can currently be assayed using the popular KINOMEscan platform[80], 91 are available as enzymatic assays (in the Reaction Biology kinase assay panel; www.reactionbiology.com, Malvern, PA), and 74 are found in both. Since the Reaction Biology assay measures phospho-transfer activity onto a peptide substrate, the availability of an assay provides further evidence that at least 91 dark kinases are catalytically active. Searching the Protein Data Bank (PDB) reveals that 53 dark kinases have at least one experimentally determined structure (for at least the kinase domain). Haspin has 18 structures, the highest of all dark kinases, followed by PIP4K2B, CLK3, and CSNK1G3 (14, 10, 10 structures, respectively) (**Supplementary Figure S3, Table S2**). Many of these structures were determined as part of the Protein Structure Initiative[81] and its successors but have not been subsequently discussed in the published literature.

## DISCUSSION

In this perspective we explore the properties of the understudied human kinome. We find that the amount of information in the literature about individual human kinase domains spans at least four orders of magnitude when measured by the number of unique causal and mechanistic statements that can be extracted using INDRA text mining and knowledge assembly software. Not surprisingly, RTKs such as EFGR and cytosolic kinases such as mTOR have high INDRA scores but other kinases are little studied, even ones for which high resolution structures and commercial assays exist. Data from INDRA correlates well with more conventional bibliometric measures,[3045] and also with the degree to which a domain has been successfully targeted with selective or clinical grade small molecules. Following the lead of the NIH IDG program, we define the dark kinome as the least-studied one-third of all kinase

domains. These domains have ~14 fold fewer INDRA statements on average than well-studied kinases and are much less likely to have small molecule ligands.

The goal of the current work is to aggregate knowledge about these domains and determine if dark kinases are likely to be expressed, have detectable phenotypes when knocked out, be mutated, amplified, or deleted in disease and be targeted or potentially targetable using small molecules. INDRA is useful in this regard because it consolidates the available literature evidence, with a focus on any available mechanistic information, in an easy-to parse node-edge graphs available at NDex, (https://home.ndexbio.org/); INDRA statements are also machine readable and can be used to construct large and small-scale computational models[47]. In the largest available cell line panel analyzed to date (the CCLE) we find that 176 dark kinase domains are likely to be expressed as measured by protein or mRNA levels and over half of genes encoding dark kinases are essential in two of more of the 625 cell lines annotated in the DepMap[57]; 10 are essential in two-thirds of DepMap lines. In addition, 27 kinases are among the top ten most mutated kinase in one or more cancer types annotated in TCGA and several kinases are differentially expressed in diseases such as Alzheimer's Disease and COPD. Thus, although available data is largely indirect, it strongly suggests that a substantial subset of dark kinases are functional in normal physiology and disease pathophysiology. This information is of immediate use in studying protein phosphorylation networks and it sets the stage for studies using genetic and chemical tools to understand dark kinase function. Based on available evidence, the possibility exists that some dark kinases may be valuable as therapeutic targets.

Kinases have evolved such that related protein folds can catalyze the phosphorylation of substrates as distinct as peptides and lipids. Conversely, folds with a more distant relationship by sequence and 3D structure can share activity against serine and threonine residues in peptides. Moreover, 50 domains with a kinase fold lack the residues canonically necessary for phospho-transferase activity. These so-called pseudokinases are found in the human kinome and many have well-

established physiological activities. In some cases (Haspin for example) kinases that have historically been classified as pseudokinases have been shown to have catalytic activity against protein substrates. Thus, there is no single definition of the kinome that suffices for all purposes. We have therefore consolidated over a dozen overlapping and often inconsistent criteria to generate two different definitions of the human kinome: an expansive 710 domain "extended kinome" that broadly encompasses related sets of folds, sequences and biological functions. This list of protein domains is likely to be most useful in chemoproteomics, small molecule profiling and genomic studies in which an expansive view of the kinome is advantageous. By sorting this list based on a range of annotated inclusion and exclusion criteria (see **Supplementary Table 1**), it is possible to generate lists with more specific properties (e.g. pseuodokinases, lipid kinases, nucleotide kinases etc). As one example of such a list we also generated a set of 557 "curated kinase" domains that is most similar in spirit to the original definition of the kinome generated by Manning[5] nearly two decades ago. This list is most useful in the study of protein kinases as a family of genes with related biochemistry and cellular fuctions. The computational analysis in this perspective focuses on this curated kinase set.

Several kinase families have regulatory domains whose mutation or over-expression also has the potential to alter function. These include the three kinases in the Protein Kinase A family (cAMP-dependent protein kinase; PRKACA, PRKACB and PRKACC)[82] which function in complex with a family of regulatory subunits, many of which have tissue-specific patterns of gene expression[83]. The cyclin dependent kinases, of which 21 are known – including 8 dark kinases – also function in a complex, in this case with one or more members of a family of at least 20 cyclins[84]. Some cyclins, CCNB3 (cyclin B3, which binds to CDK2) are frequently mutated in cancer and the effects of these mutations must be studied in conjugation with that of the kinases to which they bind (CCNB3 is mutated in 6% of cholangiocarcinomas and 11% of uterine corpus endometrial carcinoma based on TCGA data). CCNB3 is understudied however, with 7 times fewer citations than the classic CDK2 cyclin B1 and with

a similar number of citations as the dark CDK14. Thus, when regulatory subunits are included in the curated kinome, the number of relevant genes is close to 600.

Information available in the public domain and consolidated in this perspective represents the starting point for work being done by the NIH *Illuminating the Druggable Genome* project. The consolidated knowledge base already makes it possible to prioritize a subset of kinases for further investigation. In many cases substantial opportunities appear to exist for biological discovery, development of tool compounds (chemical probes) and analysis of new or existing therapeutics. Caution is nonetheless warranted in the use of this or any other biological knowledge resource because of the extreme difficulty in accurately assembling knowledge about gene function and biological mechanism. Our work on the kinome emphasizes that functional genomics involving literature analysis suffers from a significant problem with "unknown knowns."

For example, PKMYT1 is classified as an understudied kinase and it is without a doubt the least studied member of the WEE kinase family[75], a set of enzymes that regulates cell cycle progression at G2/M. WEE1 has also been targeted recently by investigational anti-cancer therapeutics (e.g. adavosertib)[85]. WEE1 and PKMYT1 are not functionally redundant since both are essential in over 90% of DepMap cell lines and they are known to have different activities against T14 and Y15 of CDK1. The consequences of differential expression of PKMYT1 in cancer cells has only recently been examined, and interest derives in part from evidence that over-expressing PKMYT1 may confer resistance to WEE1 inhibitors.[86] However, automated knowledge assembly (e.g. by INDRA) struggles in consolidating what is known about PKMYT1. Even the best NLP cannot pick up the fact that the PKMYT1 kinase (HGNC:29650) is widely referred to in the literature as "MYT1" kinase and that the HGNC symbol MYT1 (HGNC:7622) refers to an unrelated zinc-finger transcription factor. The literature is replete with these inconsistencies and false cross-references, making accurate grounding and knowledge management by both machines and humans challenging. In the current example, the missed

citations are relevant because many more publications are available citing “MYT1” than “PKMYT1”, even when the citations are manually curated to remove references to the MYT1 transcription factor. A key goal of our consortium is not only to fix these specific problems in kinome annotation but also to advance the state of the art in machine reading so that up-to-date and accurate knowledge can be made widely available about the kinome. We also intend to regularly and automatically mine large-scale database (e.g. AMP and TCGA) to consolidate information from disease genetics and add new experimental data that can be used to prioritize kinases for study in a range of tissues and indications.

## ACKNOWLEDGEMENTS

This work was funded by grant U24-DK116204 and U01-CA239106 as part of the National Institutes of Health Illuminating the Function of the Understudied Druggable Genome Program. The development of INDRA was funded by DARPA grants W911NF-15-1-0544 and W911NF018-1-0124. We thank other members of the Understudied Kinase Consortium including PIs Gary Johnson (UNC), Tim Wilson (UNC), Ben Major (WUSL) and Reid Townsend (WUSL) as well Cat Luria (HMS), Mike East (UNC), Tudor Oprea (UNM) and Jeremy Muhlich (HMS); we thank Tony Hunter (Salk Institute) for reviewing the updated protein kinase list and Yuan Wang and Jeremy Jenkins (Novartis) for providing the Bayesian Affinity Fingerprint vectors. The results on Azheimer’s disease published here are in part based on data obtained from the AMP-AD Knowledge Portal (https://adknowledgeportal.synapse.org/). Study data were provided by the Rush Alzheimer’s Disease Center, Rush University Medical Center, Chicago. Data collection was supported through funding by NIA grants P30AG10161, R01AG15819, R01AG17917, R01AG30146, R01AG36836, U01AG32984, U01AG46152, the Illinois Department of Public Health, and the Translational Genomics Research Institute. Additional data were generated from postmortem brain tissue collected through the Mount Sinai VA Medical Center Brain Bank and were provided by Dr. Eric Schadt from Mount Sinai School of Medicine.

## OUTSIDE INTERESTS

PKS is a member of the SAB or Board of Directors of Applied Biomath and RareCyte Inc and has equity in these companies. In the last five years the Sorger lab has received research funding from Novartis and Merck. Sorger declares that none of these relationships are directly or indirectly related to the content of this manuscript. Other authors declare that they have no outside interests.

## METHODS

### Classification of the "extended kinome" and defining the "curated kinome"

To obtain a list of kinases from UniProt all human proteins annotated to have kinase activity were extracted and filtered based on (i) interaction with ADP/ATP; (ii) presence of a kinase domain; 3) membership in a kinase family (lists of kinase domains and kinase families are available in supplementary material). To identify human kinase sequences that belong to the Protein Kinase Like (PKL) fold, 710 sequences annotated as "kinase" in UniProt were first subjected to a similarity search against well curated ePK profiles to identify and separate out the 8 canonical ePK groups[5,23,87]. eLKs were identified based on detectable sequence similarity with one or more of the ePK sequences. Sequences that share no detectable sequence similarity to ePKs were classified as aPKs. For predicted aPKs, crystal structures of the protein itself or of the closest homolog were inspected manually to check if the kinase domain adopts a canonical ePK fold. Additional support for this classification was obtained by calculating a Hidden Markov Model (HMM)-based distance score between the Pfam domains[88] and the presence/absence of key structural features distinguishing ePKs, eLKs and aPKs, as described previously[38,39]. A subset of sequences that satisfied none of the above criteria i.e. no detectable sequence similarity to ePKs, no clear kinase function and no homologous crystal structures, were grouped into the *unknown protein kinase category* (uPKs). All kinases annotated to have a PKL fold were included in the

curated kinome. STK19 was also included in the curated kinome despite its uPK fold since it is known to be serine/threonine kinase active against peptide substrates[18].

**Curation of INDRA statements and generation of INDRA networks**

INDRA uses natural language processing (NLP) to extract mechanistic information from literature as well as databases and represents them in a standardized format as previously described[47]. In the present study, mechanistic statements for each kinase were obtained from INDRA with the script 'get_kinase_interaction.py'. The number of INDRA statements were counted for each kinase. Regulatory networks were generated by first assembling a mechanistic model for each kinase with the INDRA assembler.cx module and uploading the model to NDex (python scripts to assemble INDRA statements and assemble mechanistic networks are available on the Github repository http://github.com/labsyspharm/dark-kinomes).

**Small molecule selectivity calculations**

The specificity of small molecules was calculated according to the *selectivity score*[53], which uses multiple parameters to assess selectivity: (i) the absolute affinity for the 'on' target; ii) the differential affinity between the 'on' and 'off' targets of each kinase; (iii) the p-value of the difference between the distributions of 'on' and 'off' targets; (iv) the research bias – a score indicating how broadly a compound has been tested for off-targets. The selectivity score was divided in four tiers; Most Selective (MS), Semi Selective (SS), Polyselective (PS) and Unknown (UN). MS levels are defined as an absolute affinity of Kd <100 nM (at least two measurements) ; a differential affinity of 100 (i.e. the affinity of the compound for the 'on' target is 100 times greater than for the 'off' targets), a p-value $\leq$ 0.1 and a research bias <0.2; SS levels are defined as an absolute affinity of Kd<1 µM (at least 4 measurements), a differential affinity of 10, a p-value $\leq$0.1 and research bias <0.2; PS levels are defined as an absolute

affinity Kd< 9000 nM, differential affinity of 1 (e.g. equal affinity for 'on' and 'off' targets) and research bias <0.2; UN levels are defined as an absolute affinity Kd< 9000 nM and differential affinity of 1.

**CCLE analysis**

The data RNA dataset 'CCLE_RNAseq_genes_rpkm_20180929.gct.gz' was downloaded from the CCLE portal (https://portals.broadinstitute.org/ccle/data) and analyzed with the script "analyzing_CCLE_data.r". The maximum expression value over all cell lines was calculated and plotted (**Figure 3A**). Genes were considered 'expressed' if the maximum RPKM was ≥1. The mass spectrometry dataset 'protein_quant_current_normalized.csv' was downloaded from the DepMap portal (https://depmap.org/portal/download/) and analyzed with the script "analyzing_CCLE_data.r". Proteins for which one or more peptides were detected in this dataset were considered to be expressed.

**Determination of Essential Kinases through Dependency Map**

The preprocessed results of genome-wide CRISPR knockout screens were obtained from the DepMap 19Q4 Public data release (https://depmap.org/portal/download/). The results of the screens were processed as described by Dempster et al[89]. For each kinase, cell lines with a CERES score >0.5 were classified as dependent and the number of dependent cell lines for each kinase was then tallied.

**TCGA analysis**

TCGA PanCan gene expression and mutation frequency data was obtained from cBioPortal[90,91]. To identify kinases with abnormal expression in tumors, tumor types with at least 10 paired normal tissue samples were analyzed. For each kinase, the fold change of its median expression in either all tumor tissues (general PanCan analysis) or the individual tumor tissue over its median expression in the paired

healthy tissues was calculated. P-value from Wilcoxon-Mann-Whitney test was calculated based on the distributions of gene expression in tumor and healthy tissues in each tumor type. Adjusted p-values were calculated using the Benjamini-Hochberg procedure. To identify kinases heavily mutated in cancer, the number of patient samples with mutation or gene fusion was counted and normalized to the total number of patient samples (10953 samples).

**AMP-AD analysis**

Preprocessed count matrices of AMP-AD consortium RNA-seq data were downloaded from the AMP-AD Synapse directory[92]. In summary, these counts were derived from raw reads using the STAR aligner[93] and the Gencode v24 human genome annotation. In our analysis, we included all Alzheimer's disease (AD) patients from the Mount Sinai VA Medical Center Brain Bank (MSBB) and the Religious Orders Study and Memory and Aging Project (ROSMAP) study[94] for which RNA-seq data from post-mortem brain was available and their age at death and Braak stage were known. Differential expression analysis was performed using the R package DESeq2[95]. We fitted a generalized linear model to the expression of each gene using the Braak stage as independent variable and adjusted for age at death and study batch effect by including them as covariates. We used the Wald test implemented in DESeq2 to extract differentially expressed genes between early (Braak stages 1 and 2) and late (5 and 6) AD cases. Effect sizes were moderated using the R package apeglm[96].

**COPD differential expression analysis**

Preprocessed dataset combining 5 datasets from GEO and 2 from ArrayExpress was downloaded from https://figshare.com/articles/Meta-analysis_of_Gene_Expression_Microarray_Datasets_in_Chronic_Obstructive_Pulmonary_Disease/8233175. Data was preprocessed as described in Rogers et al[66]. Raw expression data was processed by

generalized least squares (GLS) weighted models to account for heterogeneity between datasets. A Likelihood Ratio Test was used to identify differentially expressed genes. Genes with significant (adjusted p-value <0.5) differential expression in COPD versus healthy individuals and that are within the two-tailed 10% and 90% quantile were identified as genes of interest. Relative expression of these differentially expressed genes was calculated as the effect size of the GLS estimates of the individuals with COPD and healthy individuals.

## FIGURE LEGENDS

**Figure 1 – Composition of the human kinome**.

**(A)** Venn diagram showing the overlap in domains curated as being a kinase depending on the sources. KinHub (purple) refers to a list of kinases described by Eid et al.[97] and accessible via http://kinhub.org/kinases.html; Manning (red) refers to the gene list prepared by Manning et al. in 2002[5]; Uniprot kinase activity (green) refers to a list of genes annotated as having kinase activity in the Uniprot database[98] (see methods and Table S1); Dark Kinome (yellow) refers to a list of 168 understudied kinases as defined by the NIH solicitation for the IDG program and listed in **Supplementary Table S1** . **(B)** Schematic workflow showing how kinases are classified based on kinase three dimensional fold and sequence alignment: PKL – the *protein kinase like* fold most similar to that of the canonical PKA kinase; uPK – folds *unrelated to protein kinases* – but with significant sequence homology and known to encompass kinases against non-protein substrates as well as a limited number of protein kinases. PKLs are further subclassified into eukaryotic Protein Kinases (PKs), eukaryotic Like Kinases (eLKs) and Atypical Kinases (AKs) based on structural properties and sequence alignment. HMM refers to a Hidden Markov Model used to perform classification; see methods for details. **(C)** Pie chart showing the breakdown of 710 domains in the extended human kinome or the 557 domains in the curated kinome as defined in the current work. Subfamilies of kinases (as conventionally defined)[5] are denoted by the

white dotted lines: CAMK – Calcium and Calmodulin regulated kinases; TK – Tyrosine kinase; TKL – Tyrosine kinase like' ACG – named after the initials for kinases within the group PKA, PKC and PKG; CMGC – named after the initials of some members; CK1 – cell kinase group; AKG – atypical protein kinase group. Legend below lists some exemplary kinases from each category. Full details can be found in **Supplementary Table S1.**

**Figure 2 – The composition of the dark kinome.**

**(A)** Illustrative and simplified INDRA network automatically assembled for the WEE2 kinase. The table to the right shows the evidence extracted by INDRA for a single interaction (the bold arrow linking Wee2 and CDK1). An interactive version of this graph and a complete set of evidence can be found on NDex (http://ndexbio.org). **(B, C)** Comparison of number of INDRA statements (X-axis) and TIN-X novelty score[45] (Y-axis) for all domains in the curated human kinome. The number of INDRA statements correlates with TIN-X novelty score a Pearson's correlation coefficient of r = 0.81. The bottom third of domains having the least knowledge according to both INDRA and TIN-X are highlighted in pink and constitutes the dark kinome as defined in this manuscript. In **panel B** the Pharos target designation (solid colors) and IDG status (shape) are shown; in **panel C**, the fill color represents the maximum selectivity of a small molecule compound known to bind to each kinase. See text for details.

**Figure 3 – Dark kinases on the Coral kinase dendogram**

Kinases from the curated kinome are visualized on the Coral kinase dendrogram[40]. The recomputed dark kinome is shown in blue and non-dark kinases are shown in yellow. The atypical kinase group (AGC; denoted by a blue dashed line) as previously defined by Manning and KinHub lies to the right of the dendogram; this set includes multiple genes that are not considered to be members of the curated kinase

family as described in this paper (labelled in gray). The 46 kinases in the curated kinome but not on the Coral dendrogram are listed separately to the right and organized by protein fold. Red dashed lines denote regions of the dendogram in which all kinases are dark.

**Figure 4 – Evidence for dark kinase expression and function.**

**(A)** Maximum expression level (RPKM value) for each gene in the dark kinome list across 1039 cell lines curated in the CCLE database[99]. Dark kinases are colored in purple, non-dark kinases in orange. Dotted line indicates a RPKM threshold of 1, above which genes were designated as "expressed" based on an established metric.[58] **(B)** Number of cell lines for which the Dependency Map[57] score indicates essentiality (using the recommended *Post-Ceres*[100] value of ≤ -0.5). Dark kinases are colored in purple, non-dark kinases in orange; HGNC symbols for genes essential in all cells in the Dependency Map are shown. Blue shading denotes genes essential in one-third or more of cell lines and yellow denotes genes essential in two or more lines.

**Figure 5 – Dark kinases in diseases**

Data on differential expression, mutation, amplification or deletion of genes containing domains from the curated human kinome in disease databases. Dark kinases are colored in purple, non-dark kinases in orange. **(A)** Pan-cancer (PanCan) differential mRNA expression for both the dark and the non-dark kinases based on data in TCGA and accessed via c-BioPortal.[90] No significant difference between the dark and light kinome was observed with respect to the frequency of differential expression relative to matched normal tissue. HGNC symbols and cancer type abbreviations for selected outlier genes and diseases are shown. BRCA - Breast invasive carcinoma; KIRC - Kidney renal clear cell carcinoma; LUSC - Lung squamous cell carcinoma; UCEC - Uterine Corpus Endometrial Carcinoma. **(B)** Mutation frequency for most frequently mutated kinases in PanCan. Dark kinases are shown in solid color; non-

dark kinase in transparent color. Fusion-mutations are shown in magenta. HGNC symbols are displayed next to each bar with bold denoting dark kinases. **(C)** Differential gene expression in early versus late stage Alzheimer's disease. Samples were aged matched prior to calculation of differential expression values. HGNC symbols are shown for outliers displayed. **(D)** Relative gene expression levels in COPD versus healthy individuals. Kinases are sorted by their relative expression. HGNC symbols are displayed next to each bar with purple denoting dark kinases and orange denoting non-dark kinases.

**Figure 6 – A dark kinase network regulating the cell cycle**

**(A)** A partial network (left) and statement table (right) for proteins interacting with PKMYT1 according to the INDRA database. The source 'literature' is denoted by 'L'. **(B)** INDRA network for CDK1 showing interacting dark kinases. Black arrows denote protein modifications; blue lines denote complex formation; red arrows denote inhibition; green arrows denote activation. **(C)** INDRA network for WEE1 showing interacting dark kinases. Color code is the same as in panel B. **(D)** Manually curated signaling network based on known regulatory mechanisms for PKMYT1, CDK1 and WEE1. The network comprises four dark kinases (BRSK1, NIM1K, WEE2 and PKMYT1), three non-dark kinases (BRSK2, WEE1, CDK1), and the protein phosphatase CDC25.

**Figure 7 – Inhibition of dark kinases by clinical grade compounds and approved drugs**

**(A)** Kinase inhibitors in clinical development and FDA-approved small molecule therapeutics targeting dark kinases for which binding is scored as *most selective (*MS) or *semi selective (*SS) based on literature data curated in CHEMBL. Eighteen dark kinases are targeted in total. **(B)** Clustering of dark kinases by Bayes Affinity Fingerprint (BAFP) – a measure of the shape of binding pocket. Dark lines in the margin denote dark kinases. A blowup of BAFP values for four kinases (red box), one of which is dark (STK17B) , is shown below.

## SUPPLEMENTARY FIGURE LEGENDS

### Figure S1 related to Figure 1 - Kinase domains in the standard list

**(A)** Pie chart of kinases in the Manning and Kinhub lists divided into kinase groups as conventionally defined. Letter coding explanation can be found of main figure legend 1C. **(B)** Number of compounds that target dark kinases as determined in a recent large-scale chemoproteomic assay[34].

### Figure S2 related to Figure 2 - INDRA network for WEE2

A partial network (upper panel) and statement table (lower panel) generated by INDRA for the dark kinase WEE2. Table contains full quotes from literature.

### Figure S3 related to Figure 5 - Differential mRNA expression of kinases in selected TCGA datasets

**(A)** Differential mRNA expression for both the dark and the non-dark kinases in colon adenocarcinoma (COAD) based on data in TCGA and accessed via c-BioPortal.82. HGNC symbols and cancer type abbreviations for selected outlier genes. **(B)** Depicted is the alteration frequency in lymphoid neoplasm diffuse large B-cell lymphoma (DLBCL) for dark kinases (darks bars) and non-dark kinases (light bars). Both fusion (magenta) and mutations (black/grey) are indicated.

### Figure S4 related to Figure 6 - Kinase domains with high resolution structures

Number of structures of the kinase domain in Protein Data Bank (PDB) for both non-dark kinases (orange) and dark kinases (purple) sorted in descending order. Top dark kinases with high number of kinase domain structures are labeled.

**Figure S5 related to Figure 6 – Clustering of kinases by binding pocket based on BAFP and mapped to the Coral dendogram**

**(A)** BAFP clusters visualized on the Coral kinase dendrogram. **(B)** The nine labelled BAFP clusters (denoted by color and labelled 1-9) projected on one Coral kinase dendrogram. Dark kinases in each cluster are colored black.

## FOOTNOTES FOR SUPPLEMENTARY TABLES

**Supplementary Table 1 – The extended kinome**

This table describes available information about the 710 kinase domains in the extended kinome. Each domain is annotated with the following pieces of information: gene_id (NCBI gene ID); UniProtEntry (Uniprot ID, Uniprot Entry name and domain index if the kinase contains multiple kinase domains) Entry (the unique and stable short-form Uniprot ID as a number); Entry name (Mnemonic identifier to UniprotKB entry); Gene names (names of genes encoding this protein as obtained from Uniprot), Protein names (full name of the protein provided by UniProt), HGNC ID, HGNC_name (the official gene symbol approved by HGNC), Approved name (the full gene name approved by HGNC), IDG_dark (value of 0 or 1 denoting whether dark in the original NIH list), Kinhub (value of 0 or 1 denoting whether the domain is on the Kinhub list), Manning (value of 0 or 1 denoting whether domain is on the Manning list), Group (membership to one of the ten kinase groups), Family (membership in the kinase families), Uniprot_kinaseactivity (value of 0 or 1 denoting whether domain is on the curated UniProt kinase list), PfamDomain, DomainStart (first residue number of the kinase domain according to UniProt), DomainEnd (last residue number of the kinase domain according to UniProt), ProKinO (value of 0 or 1 denoting wehther the domain is in ProKinO), New_Annotation (further classification of the protein fold as ePK, eLK, Atypical, Unrelated to Protein Kinase or Unknown), Fold (primary classification of the protein fold: protein kinase like – PKL -, unrelated, UPK or unknown),

Pseudokinase? (Yes or No annotation to whether the kinase is a pseudokinase according to ProKinO), Annotation_Score (number to reflect the amount of aggregated information from multiple databases), INDRA_network (URL of the interactive INDRA network on NDEx).

**Supplementary Table 2 – The curated kinome**

A table describing data about the 556 kinase domains in the curated kinome with a PKL fold (plus STK19). Each domain is annotated with all information in Supplementary Table 1 about its identifiers (NCBI gene_id, HGNC identifiers, and UniProt identifiers), inclusion and exclusion criteria based on different kinase lists (Manning, KinHub, kinase group and kinase family according to KinHub, curated UniProt kinase list, NIH dark kinase, ProKinO, pseudokinase), protein fold, and the URL for its INDRA network on NDEx. Each kinase domain also has the following additional annoations: **(i)** amount of existing information (n_indra_statement: number of INDRA statements; TIN-X_Score; tdl (target development level from Pharos); **(ii)** whether the kinase is dark (stat_dark_num: value of 0 or 1 denoting whether a kinase is dark based on number of INDRA statement and TIN-X_Score); **(iii)** PDB structures (PDBID: PDB IDs for any structures of the kinase domain; num_pdb: the total number of pdb structures of the kinase domain); **(iv)** number of MS/SS compounds (num_MSSS_cmpd); **(v)** availability of commercial activity assays (rb_name: the name of the kinase on Reaction Biology (http://www.reactionbiology.com); rb_variants: the phosphorylated form or protein complex available for assay on Reaction Biology; kinomescan_name: the name of the kinase on DiscoverX (https://www.discoverx.com/home); kinomescan_variants: the phosphorylated form or protein complex available for assay on the DiscoverX kinase panel; commercial_assay: value of 0 or 1 denoting whether a Reaction Biology or KinomeScan assay is available); **(vi)** biological relevance and disease implications (num_dep: number of dependent cell lines on DepMap; AMPAD: value of 0 or 1 denoting whether the kinase is differentially expressed in Alzheimer patients; TCGA: value of 0 or 1 denoting

whether the kinase is differentially expressed in any cancer type, among the top 10 most frequently mutated kinases in any cancer type, or among the top 20 most frequently mutated kinases of all cancers; COPD: value of 0 or 1 denoting whether the kinase is differentially expressed in COPD patients).

**Supplementary Table 3 – Clinical compounds targeting dark kinases**

A table with the affinity values of compounds in clinical development (phase 1-3) or approved drugs that have been shown to target dark kinases based on available data in Small Molecule Suite (http://smallmoleculesuite.org). The compounds are annotated with IC50_Q1 (the affinity value per dark kinase), HGNC_symbol (official HGNC symbol of dark kinase), compound_max_phase (the latest stage of clinical development), compound_first_approval (year of first approval if compound is an approved therapeutic), compound_chembl_id (ChEMBL identifier of compound).

## REFERENCES

1. Cohen, P. The origins of protein phosphorylation. *Nat. Cell Biol.* **4**, E127-130 (2002).
2. Hunter, T. Protein kinases and phosphatases: the yin and yang of protein phosphorylation and signaling. *Cell* **80**, 225–236 (1995).
3. Knighton, D. R. *et al.* Crystal structure of the catalytic subunit of cyclic adenosine monophosphate-dependent protein kinase. *Science* **253**, 407–414 (1991).
4. Adams, J. A. Kinetic and catalytic mechanisms of protein kinases. *Chem. Rev.* **101**, 2271–2290 (2001).
5. Manning, G., Whyte, D. B., Martinez, R., Hunter, T. & Sudarsanam, S. The Protein Kinase Complement of the Human Genome. *Science* **298**, 1912–1934 (2002).
6. Lai, S., Safaei, J. & Pelech, S. Evolutionary Ancestry of Eukaryotic Protein Kinases and Choline Kinases. *J. Biol. Chem.* **291**, 5199–5205 (2016).

7. Darnell, J. E. Phosphotyrosine signaling and the single cell:metazoan boundary. *Proc. Natl. Acad. Sci. U. S. A.* **94**, 11767–11769 (1997).

8. Wijk, L. M. van & Snel, B. The first eukaryotic kinome tree illuminates the dynamic history of present-day kinases. *bioRxiv* 2020.01.27.920793 (2020) doi:10.1101/2020.01.27.920793.

9. Hanks, S. K. & Hunter, T. Protein kinases 6. The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification. *FASEB J. Off. Publ. Fed. Am. Soc. Exp. Biol.* **9**, 576–596 (1995).

10. Hanks, S. K., Quinn, A. M. & Hunter, T. The protein kinase family: conserved features and deduced phylogeny of the catalytic domains. *Science* **241**, 42–52 (1988).

11. Kanev, G. K. *et al.* The Landscape of Atypical and Eukaryotic Protein Kinases. *Trends Pharmacol. Sci.* **40**, 818–832 (2019).

12. Knighton, D. R. *et al.* Structure of a peptide inhibitor bound to the catalytic subunit of cyclic adenosine monophosphate-dependent protein kinase. *Science* **253**, 414–420 (1991).

13. Wu, G. & Vance, D. E. Choline kinase and its function. *Biochem. Cell Biol. Biochim. Biol. Cell.* **88**, 559–564 (2010).

14. Mukohara, T. PI3K mutations in breast cancer: prognostic and therapeutic implications. *Breast Cancer Dove Med. Press* **7**, 111–123 (2015).

15. Robinson, D. R., Wu, Y.-M. & Lin, S.-F. The protein tyrosine kinase family of the human genome. *Oncogene* **19**, 5548–5557 (2000).

16. Mohanty, S. *et al.* Hydrophobic Core Variations Provide a Structural Framework for Tyrosine Kinase Evolution and Functional Specialization. *PLoS Genet.* **12**, e1005885 (2016).

17. Kannan, N., Haste, N., Taylor, S. S. & Neuwald, A. F. The hallmark of AGC kinase functional divergence is its C-terminal tail, a cis-acting regulatory module. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 1272–1277 (2007).

18. Yin, C. *et al.* Pharmacological Targeting of STK19 Inhibits Oncogenic NRAS-Driven Melanomagenesis. *Cell* **176**, 1113-1127.e16 (2019).

19. Kwon, A. *et al.* Tracing the origin and evolution of pseudokinases across the tree of life. *Sci. Signal.* **12**, (2019).

20. Sliwkowski, M. X. *et al.* Coexpression of erbB2 and erbB3 proteins reconstitutes a high affinity receptor for heregulin. *J. Biol. Chem.* **269**, 14661–14665 (1994).

21. Garrett, J. T. *et al.* Transcriptional and posttranslational up-regulation of HER3 (ErbB3) compensates for inhibition of the HER2 tyrosine kinase. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 5021–5026 (2011).

22. Villa, F. *et al.* Crystal structure of the catalytic domain of Haspin, an atypical kinase implicated in chromatin organization. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 20204–20209 (2009).

23. Eswaran, J. *et al.* Structure and functional characterization of the atypical human kinase haspin. *Proc. Natl. Acad. Sci.* **106**, 20198–20203 (2009).

24. Dai, J., Sultan, S., Taylor, S. S. & Higgins, J. M. G. The kinase haspin is required for mitotic histone H3 Thr 3 phosphorylation and normal metaphase chromosome alignment. *Genes Dev.* **19**, 472–488 (2005).

25. Hawley, S. A. *et al.* The Ancient Drug Salicylate Directly Activates AMP-Activated Protein Kinase. *Science* **336**, 918–922 (2012).

26. Jones, L. H. Small-Molecule Kinase Downregulators. *Cell Chem. Biol.* **25**, 30–35 (2018).

27. FAUVEL, B. & Yasri, A. Antibodies directed against receptor tyrosine kinases. *mAbs* **6**, 838–851 (2014).

28. Finan, C. *et al.* The druggable genome and support for target identification and validation in drug development. *Sci. Transl. Med.* **9**, (2017).

29. Illuminating the Druggable Genome | NIH Common Fund. https://commonfund.nih.gov/idg.

30. Huang, L.-C. *et al.* Integrative annotation and knowledge discovery of kinase post-translational modifications and cancer-associated mutations through federated protein ontologies and resources. *Sci. Rep.* **8**, 1–16 (2018).

31. Eid, S., Turk, S., Volkamer, A., Rippmann, F. & Fulle, S. KinMap: a web-based tool for interactive navigation through human kinome data. *BMC Bioinformatics* **18**, 16 (2017).

32. Ciceri, P. *et al.* Dual kinase-bromodomain inhibitors for rationally designed polypharmacology. *Nat. Chem. Biol.* **10**, 305–312 (2014).

33. Posy, S. L. *et al.* Trends in kinase selectivity: insights for target class-focused library screening. *J. Med. Chem.* **54**, 54–66 (2011).

34. Klaeger, S. *et al.* The target landscape of clinical kinase drugs. *Science* **358**, eaan4368 (2017).

35. Cousins, E. M. *et al.* Competitive Kinase Enrichment Proteomics Reveals that Abemaciclib Inhibits GSK3β and Activates WNT Signaling. *Mol. Cancer Res. MCR* **16**, 333–344 (2018).

36. Weinstein, J. N. *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).

37. RFA-RM-16-026: Data and Resource Generation Centers for Illuminating the Druggable Genome (U24). https://grants.nih.gov/grants/guide/rfa-files/RFA-RM-16-026.html.

38. Kannan, N., Taylor, S. S., Zhai, Y., Venter, J. C. & Manning, G. Structural and Functional Diversity of the Microbial Kinome. *PLOS Biol.* **5**, e17 (2007).

39. Kannan, N. & Neuwald, A. F. Did protein kinase regulatory mechanisms evolve through elaboration of a simple structural component? *J. Mol. Biol.* **351**, 956–972 (2005).

40. Metz, K. S. *et al.* Coral: Clear and Customizable Visualization of Human Kinome Data. *Cell Syst.* **7**, 347-350.e1 (2018).

41. Jassal, B. *et al.* The reactome pathway knowledgebase. *Nucleic Acids Res.* **48**, D498–D503 (2020).

42. Baumlova, A. *et al.* The crystal structure of the phosphatidylinositol 4-kinase IIα. *EMBO Rep.* **15**, 1085–1092 (2014).

43. Reymond, A. *et al.* The tripartite motif family identifies cell compartments. *EMBO J.* **20**, 2140–2151 (2001).

44. Ozato, K., Shin, D.-M., Chang, T.-H. & Morse, H. C. TRIM family proteins and their emerging roles in innate immunity. *Nat. Rev. Immunol.* **8**, 849–860 (2008).

45. Cannon, D. C. *et al.* TIN-X: target importance and novelty explorer. *Bioinformatics* **33**, 2601–2603 (2017).

46. Todorov, P. V., Gyori, B. M., Bachman, J. A. & Sorger, P. K. INDRA-IPM: interactive pathway modeling using natural language with automated assembly. *Bioinforma. Oxf. Engl.* (2019) doi:10.1093/bioinformatics/btz289.

47. Gyori, B. M. *et al.* From word models to executable models of signaling networks using automated assembly. *Mol. Syst. Biol.* **13**, 954 (2017).

48. McDonald, D. D., Friedman, S. E., Paullada, A., Bobrow, R. & Burstein, M. H. Extending Biology Models with Deep NLP over Scientific Articles. in *AAAI Workshop: Knowledge Extraction from Text* (2016).

49. Valenzuela-Escarcega, M. A. *et al.* Large-scale automated reading with Reach discovers new cancer driving mechanisms. in *Proceedings of the Sixth BioCreative Challenge Evaluation Workshop* 201–203 (2017).

50. Stark, C. *et al.* BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.* **34**, D535-539 (2006).

51. Cerami, E. G. *et al.* Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Res* **39**, D685-90 (2011).

52. Small molecules - HMS LINCS Database - HMS LINCS Project. http://lincs.hms.harvard.edu/db/sm/.

53. Moret, N. *et al.* Cheminformatics Tools for Analyzing and Designing Optimized Small-Molecule Collections and Libraries. *Cell Chem. Biol.* **26**, 765-777.e3 (2019).

54. Pratt, D. *et al.* NDEx, the Network Data Exchange. *Cell Syst.* **1**, 302–305 (2015).

55. Nguyen, D.-T. *et al.* Pharos: Collating protein information to shed light on the druggable genome. *Nucleic Acids Res.* **45**, D995–D1002 (2017).

56. Nusinow, D. P. *et al.* Quantitative Proteomics of the Cancer Cell Line Encyclopedia. *Cell* **180**, 387-402.e16 (2020).

57. Tsherniak, A. *et al.* Defining a Cancer Dependency Map. *Cell* **170**, 564-576.e16 (2017).

58. Kryuchkova-Mostacci, N. & Robinson-Rechavi, M. A benchmark of gene expression tissue-specificity metrics. *Brief. Bioinform.* **18**, 205–214 (2017).

59. Garraway, L. A. & Lander, E. S. Lessons from the cancer genome. *Cell* **153**, 17–37 (2013).

60. Holmes, K., Roberts, O. L., Thomas, A. M. & Cross, M. J. Vascular endothelial growth factor receptor-2: Structure, function, intracellular signalling and therapeutic inhibition. *Cell. Signal.* **19**, 2003–2012 (2007).

61. Gratzinger, D. *et al.* Lymphoma cell VEGFR2 expression detected by immunohistochemistry predicts poor overall survival in diffuse large B cell lymphoma treated with immunochemotherapy (R-CHOP). *Br. J. Haematol.* **148**, 235–244 (2010).

62. Jørgensen, J. M. *et al.* Expression level, tissue distribution pattern, and prognostic impact of vascular endothelial growth factors VEGF and VEGF-C and their receptors Flt-1, KDR, and Flt-4 in different subtypes of non-Hodgkin lymphomas. *Leuk. Lymphoma* **50**, 1647–1660 (2009).

63. Bailey, M. H. *et al.* Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **173**, 371-385.e18 (2018).

64. Hodes, R. J. & Buckholtz, N. Accelerating Medicines Partnership: Alzheimer's Disease (AMP-AD) Knowledge Portal Aids Alzheimer's Drug Discovery through Open Data Sharing. *Expert Opin. Ther. Targets* **20**, 389–391 (2016).

65. COPD | National Heart, Lung, and Blood Institute (NHLBI). https://www.nhlbi.nih.gov/health-topics/copd.

66. Rogers, L. R. K., Verlinde, M. & Mias, G. I. Gene expression microarray public dataset reanalysis in chronic obstructive pulmonary disease. *PLOS ONE* **14**, e0224750 (2019).

67. Shim, H. *et al.* Deletion of the gene Pip4k2c, a novel phosphatidylinositol kinase, results in hyperactivation of the immune system. *Proc. Natl. Acad. Sci.* **113**, 7596–7601 (2016).

68. Tan, I., Lai, J., Yong, J., Li, S. F. Y. & Leung, T. Chelerythrine perturbs lamellar actomyosin filaments by selective inhibition of myotonic dystrophy kinase-related Cdc42-binding kinase. *FEBS Lett.* **585**, 1260–1268 (2011).

69. Tan, I., Yong, J., Dong, J. M., Lim, L. & Leung, T. A tripartite complex containing MRCK modulates lamellar actomyosin retrograde flow. *Cell* **135**, 123–136 (2008).

70. Sang, Q. *et al.* Homozygous Mutations in WEE2 Cause Fertilization Failure and Female Infertility. *Am. J. Hum. Genet.* **102**, 649–657 (2018).

71. Liu, F., Stanton, J. J., Wu, Z. & Piwnica-Worms, H. The human Myt1 kinase preferentially phosphorylates Cdc2 on threonine 14 and localizes to the endoplasmic reticulum and Golgi complex. *Mol. Cell. Biol.* **17**, 571–583 (1997).

72. Mueller, P. R., Coleman, T. R., Kumagai, A. & Dunphy, W. G. Myt1: A Membrane-Associated Inhibitory Kinase That Phosphorylates Cdc2 on Both Threonine-14 and Tyrosine-15. *Science* **270**, 86–90 (1995).

73. Santamaría, D. *et al.* Cdk1 is sufficient to drive the mammalian cell cycle. *Nature* **448**, 811–815 (2007).

74. DepMap, B. DepMap 19Q3 Public. (2019) doi:10.6084/m9.figshare.9201770.v3.

75. Asquith, C. R. M., Laitinen, T. & East, M. P. PKMYT1: a forgotten member of the WEE1 family. *Nat. Rev. Drug Discov.* **19**, 157 (2020).

76. Matheson, C. J., Backos, D. S. & Reigan, P. Targeting WEE1 Kinase in Cancer. *Trends Pharmacol. Sci.* **37**, 872–881 (2016).

77. Wu, L. & Russell, P. Nim1 kinase promotes mitosis by inactivating Wee1 tyrosine kinase. *Nature* **363**, 738–741 (1993).

78. Bender, A. *et al.* "Bayes Affinity Fingerprints" Improve Retrieval Rates in Virtual Screening and Define Orthogonal Bioactivity Space:? When Are Multitarget Drugs a Feasible Concept? *J. Chem. Inf. Model.* **46**, 2445–2456 (2006).

79. Nguyen, H. P. *et al.* Diversity Selection of Compounds Based on 'Protein Affinity Fingerprints' Improves Sampling of Bioactive Chemical Space. *Chem. Biol. Drug Des.* **82**, 252–266 (2013).

80. Fabian, M. A. *et al.* A small molecule–kinase interaction map for clinical kinase inhibitors. *Nat. Biotechnol.* **23**, 329–336 (2005).

81. Burley, S. K., Joachimiak, A., Montelione, G. T. & Wilson, I. A. Contributions to the NIH-NIGMS Protein Structure Initiative from the PSI Production Centers. *Struct. Lond. Engl. 1993* **16**, 5–11 (2008).

82. Taylor, S. S., Zhang, P., Steichen, J. M., Keshwani, M. M. & Kornev, A. P. PKA: lessons learned after twenty years. *Biochim. Biophys. Acta* **1834**, 1271–1278 (2013).

83. Herberg, F. W., Maleszka, A., Eide, T., Vossebein, L. & Tasken, K. Analysis of A-kinase anchoring protein (AKAP) interaction with protein kinase A (PKA) regulatory subunits: PKA isoform specificity in AKAP binding. *J. Mol. Biol.* **298**, 329–339 (2000).

84. Hydbring, P., Malumbres, M. & Sicinski, P. Non-canonical functions of cell cycle cyclins and cyclin-dependent kinases. *Nat. Rev. Mol. Cell Biol.* **17**, 280–292 (2016).

85. Cole, K. A. *et al.* Phase I Clinical Trial of the Wee1 Inhibitor Adavosertib (AZD1775) with Irinotecan in Children with Relapsed Solid Tumors. A COG Phase I Consortium Report (ADVL1312). *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* (2019) doi:10.1158/1078-0432.CCR-19-3470.

86. Lewis, C. W. *et al.* Upregulation of Myt1 Promotes Acquired Resistance of Cancer Cells to Wee1 Inhibition. *Cancer Res.* **79**, 5971–5985 (2019).

87. Talevich, E., Mirza, A. & Kannan, N. Structural and evolutionary divergence of eukaryotic protein kinases in Apicomplexa. *BMC Evol. Biol.* **11**, 321 (2011).

88. Huo, L. *et al.* pHMM-tree: phylogeny of profile hidden Markov models. *Bioinforma. Oxf. Engl.* **33**, 1093–1095 (2017).

89. Dempster, J. M. *et al.* Extracting Biological Insights from the Project Achilles Genome-Scale CRISPR Screens in Cancer Cell Lines. *bioRxiv* 720243 (2019) doi:10.1101/720243.

90. Cerami, E. *et al.* The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. *Cancer Discov.* **2**, (2012).

91. Gao, J. *et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.* **6**, pl1 (2013).

92. rnaSeqReprocessing - syn9702085. https://www.synapse.org/#!Synapse:syn9702085.

93. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinforma. Oxf. Engl.* **29**, 15–21 (2013).

94. Mostafavi, S. *et al.* A molecular network of the aging human brain provides insights into the pathology and cognitive decline of Alzheimer's disease. *Nat. Neurosci.* **21**, 811–819 (2018).

95. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

96. Zhu, A., Ibrahim, J. G. & Love, M. I. Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences. *Bioinforma. Oxf. Engl.* **35**, 2084–2092 (2019).

97. Eid, S., Turk, S., Volkamer, A., Rippmann, F. & Fulle, S. KinMap: a web-based tool for interactive navigation through human kinome data. *BMC Bioinformatics* **18**, 16 (2017).

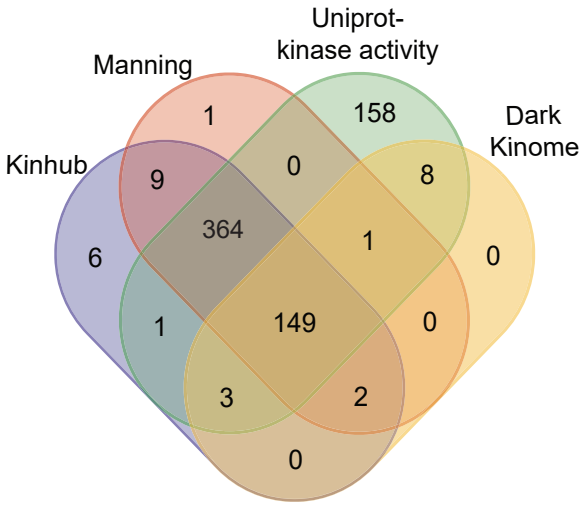
98. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).

99. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012).

100. Meyers, R. M. *et al.* Computational correction of copy-number effect improves specificity of CRISPR-Cas9 essentiality screens in cancer cells. *Nat. Genet.* **49**, 1779–1784 (2017).

# FIGURE 1

**a.** **Existing Kinome Annotations**




**b.** Pfam HMM profiles

**uPK:** Unrelated to Protein Kinases

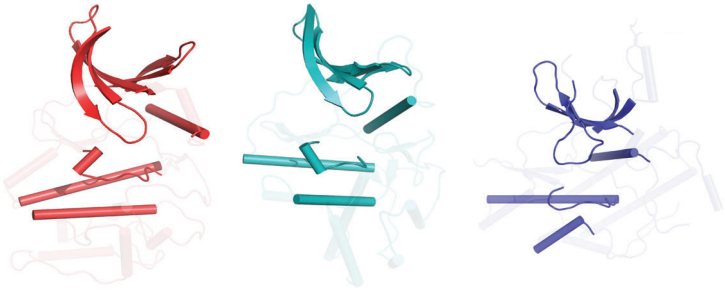
**PKL:** Protein Kinase like

Unknown

**ePK:**eukaryotic Protein Kinases

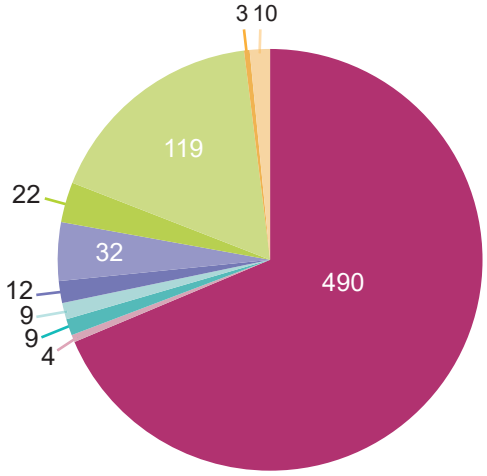
**eLK:** eukaryotic Like Kinases

**aPK:** Atypical Protein Kinases



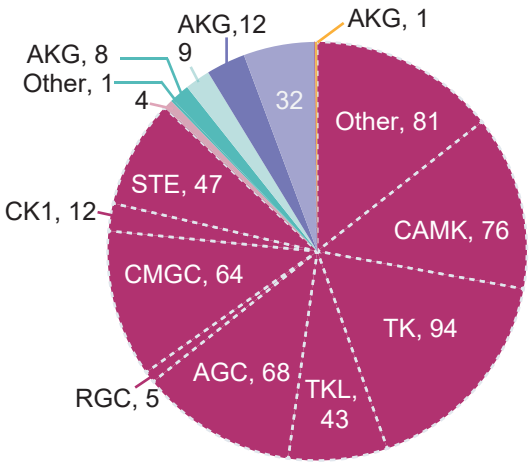
**c.** **New Kinome Annotations**

The extended human kinome




Total number of domains: 710

The curated human kinome



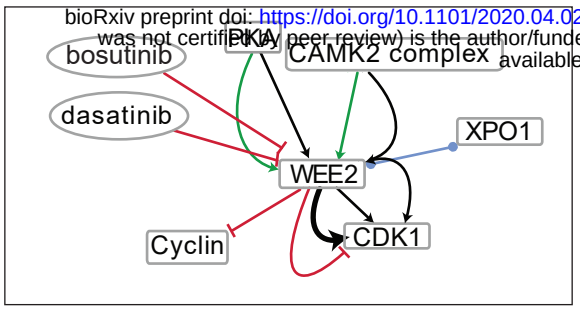

Total number of domains: 557

| | Manning & Kinhub | | Not in Manning & Kinhub | |
|---|---|---|---|---|
| **Fold** | **Color** | **Example Kinases** | **Color** | **Example Kinases** |
| ePK | | PRKACA, KIT, CDK8 | | PEAK3, PLK5, SIK1B |
| eLK | | ADCK1, TP53RK, RIOK1 | | CHKA, ETNK1, HYKK |
| aPK | | ATM, TRRAP, ALPK1 | | PIK3C2A, PI4KA, ITPKA |
| uPK | | BRD2, BCR, TRIM24 | | ADK, PPIP5K1, CKM |
| unknown | | FASTK, HSPB8, STK19 | | FASTKD1, DOLK, NADK2 |

# FIGURE 2



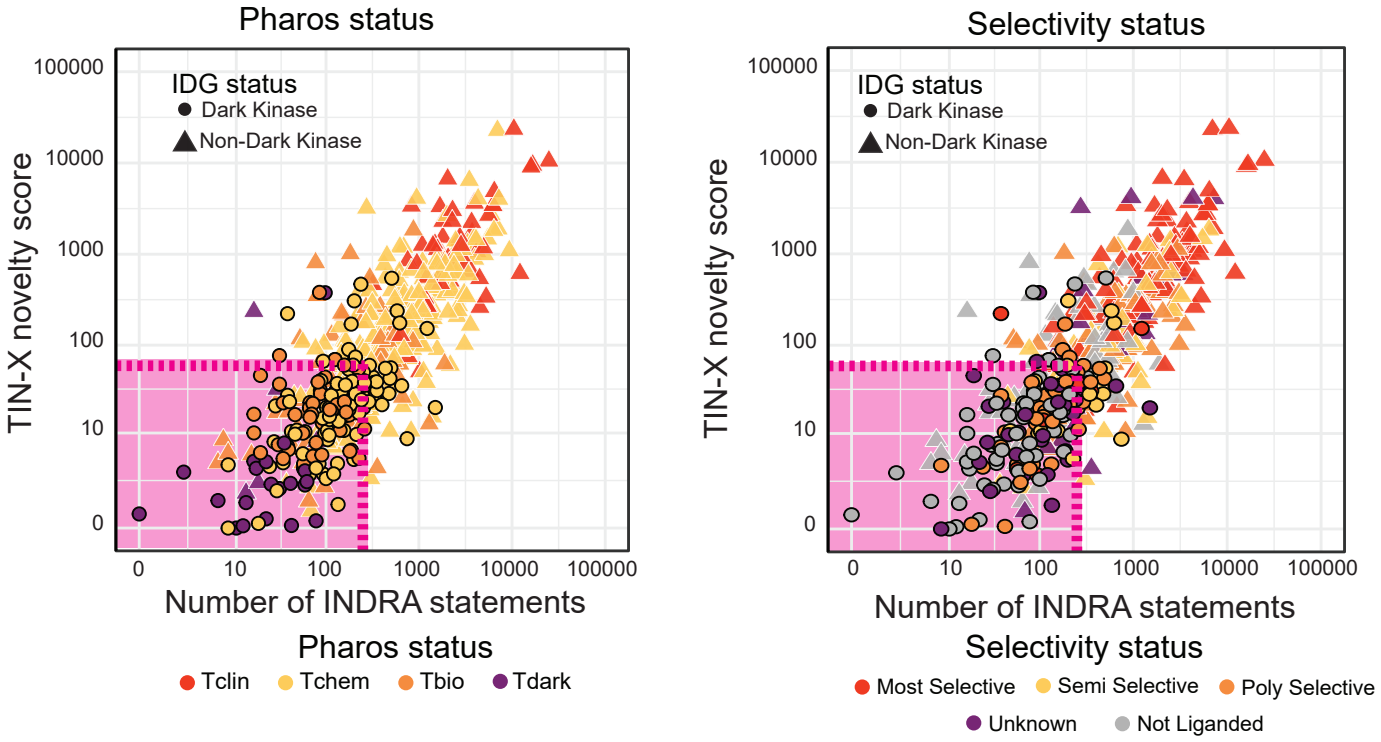
**a.**




**INDRA evidence for a typical interaction**

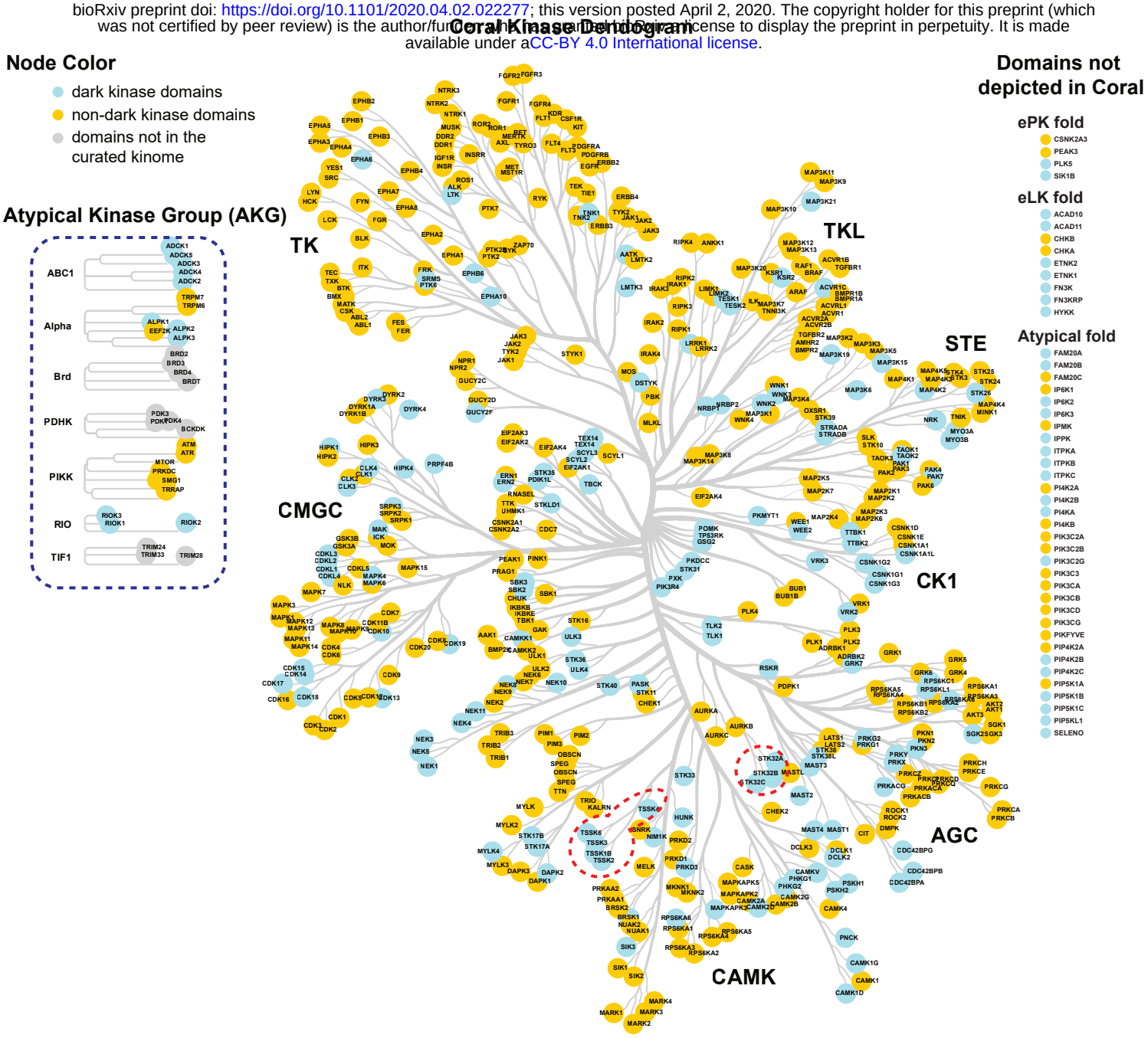
| Type | Sentence extracted | INDRA statement | Source |
|---|---|---|---|
| Modification | "CaMKII regulates resumption of the cell cycle through phosphorylation of Wee1B, which in turn phosphorylates and inhibits Cdc2" | Phosphorylation (WEE2(), CDK1()) | PMID: 23287037<br>PMID: 21454751<br>PMID: 19550110 |

**b.** Knowledge about the kinome extracted by INDRA and TIN-X

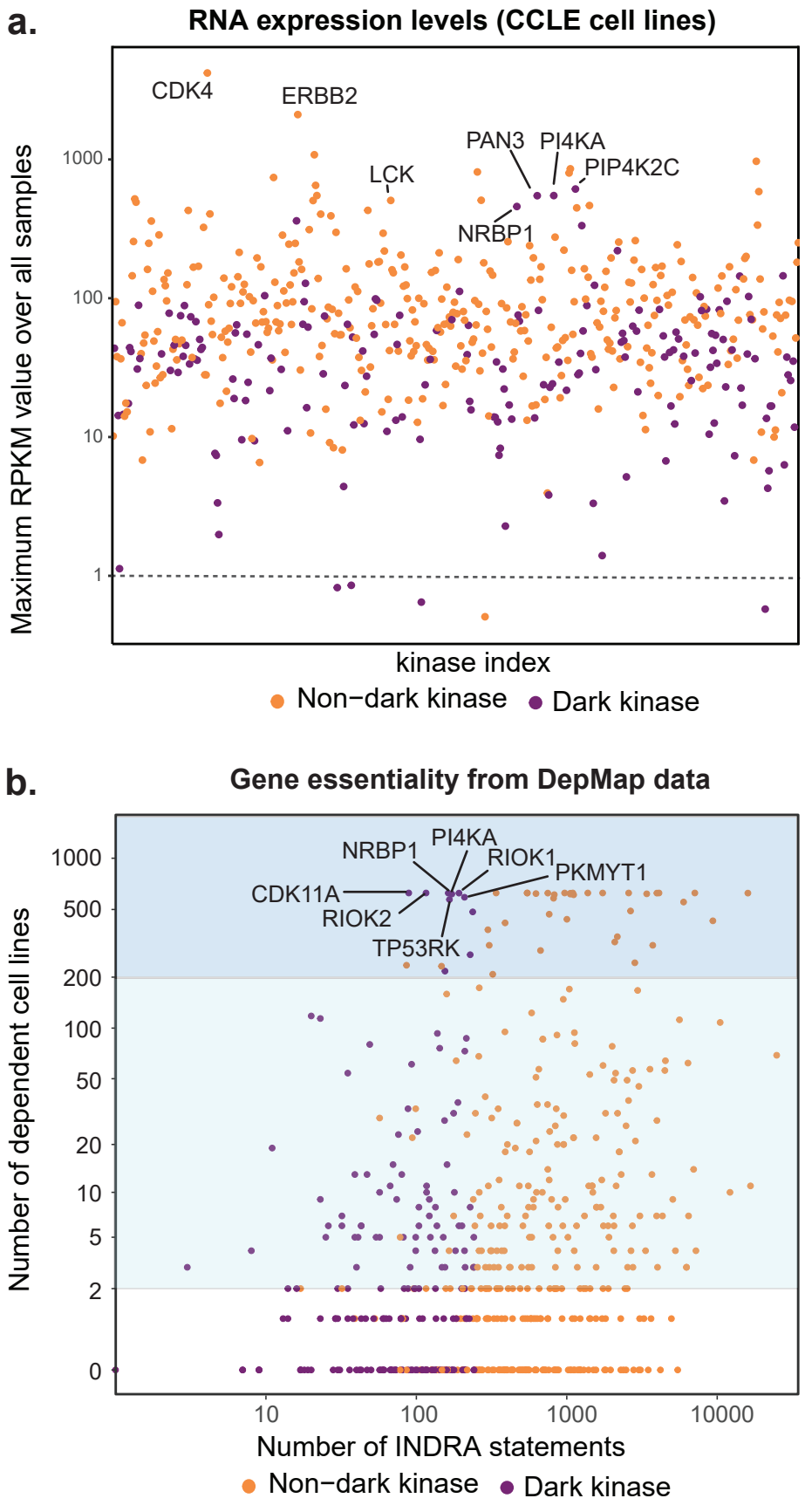


# FIGURE 3

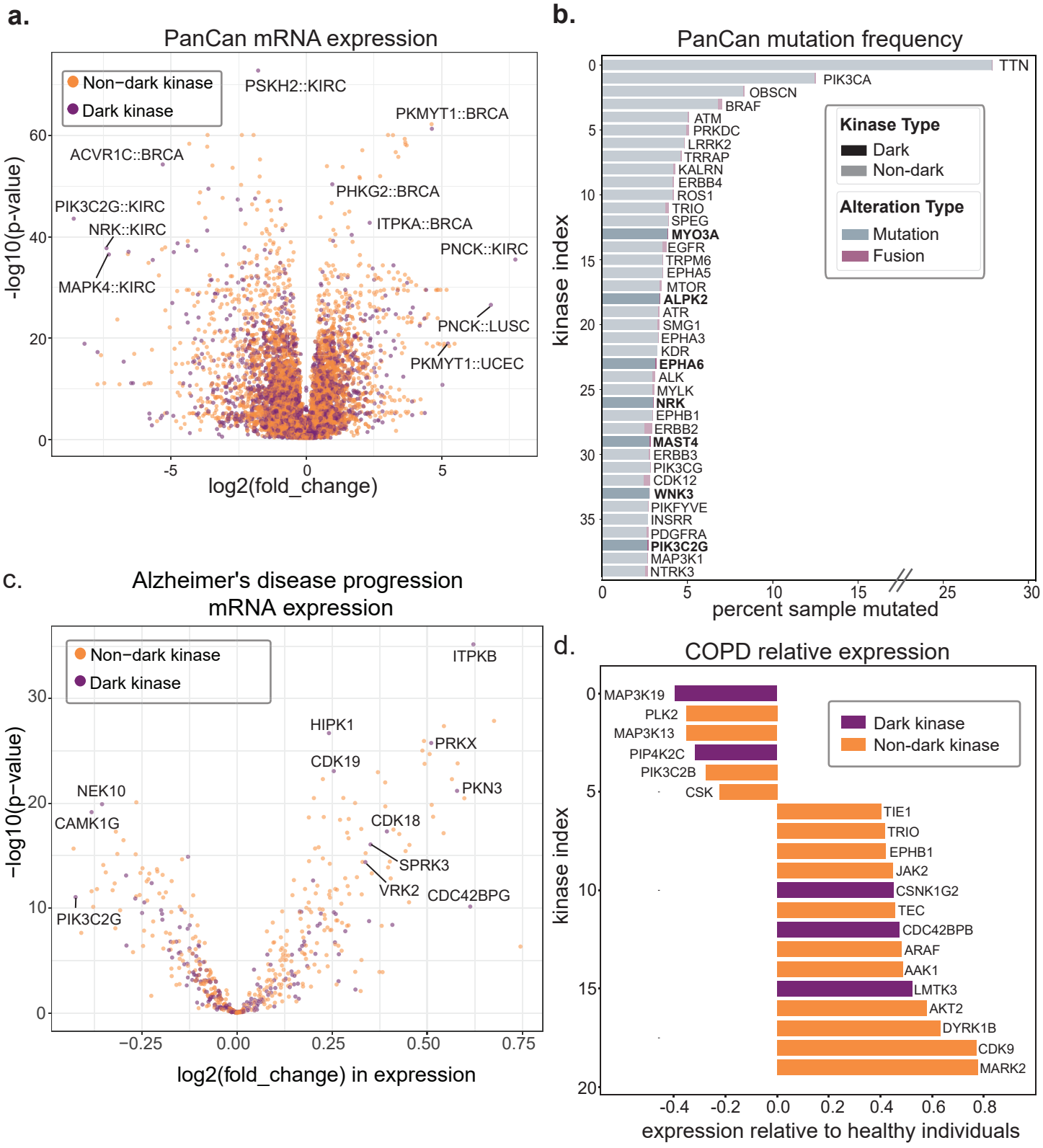




# FIGURE 4

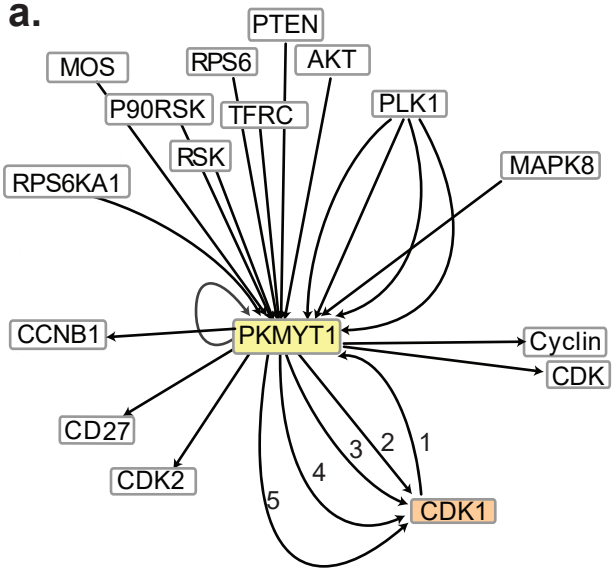


# FIGURE 5



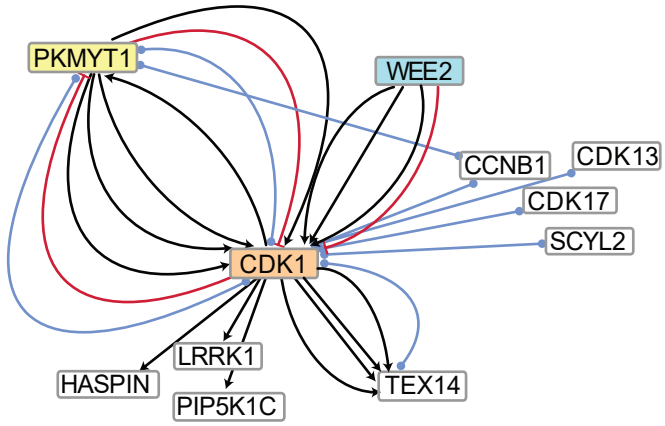
# FIGURE 6

**a.**

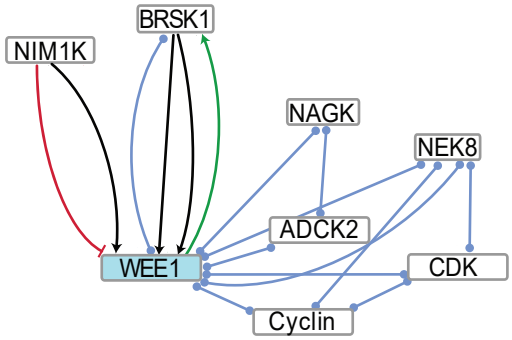



| | interaction type | INDRA statement | interaction description | source type | PMID |
|---|---|---|---|---|---|
| 1 | Modification | Phosphorylation (CDK1(), PKMYT1()) | CDK1 phosphorylates PKMYT1 | L | 21325631 26673326 22726437 |
| 2 | Modification | Phosphorylation (PKMYT1(), CDK1(), Y, 15) | PKMYT1 phosphorylates CDK1 on Tyr 15 | L | 20419782 11202906 22494620 |
| 3 | Modification | Phosphorylation (PKMYT1(), CDK1(), T, 14) | PKMYT1 phosphorylates CDK1 on Thr 14 | L | 20419782 11202906 22494620 |
| 4 | Modification | ... | ... | ... | ... |

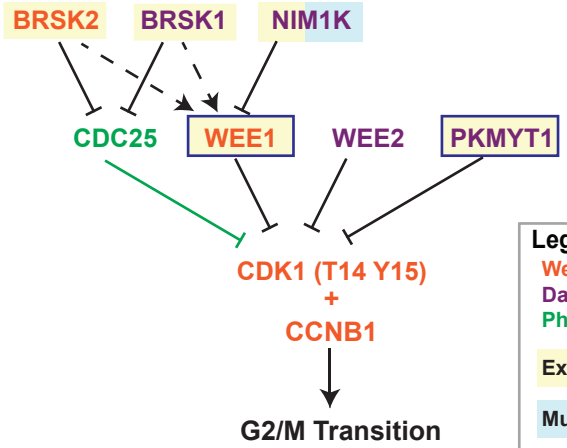
**b.**




**c.**




**d.**




**Legend**

- Well studied kinase
- Dark kinase
- Phosphatase
- Expression changes in TCGA
- Mutated in TCGA
- Essential in DepMAP

**BRSK1**
- Upregulated in 7 tumor types
- High level copy number amplification in DLBC
- High level copy number deletion in LGG

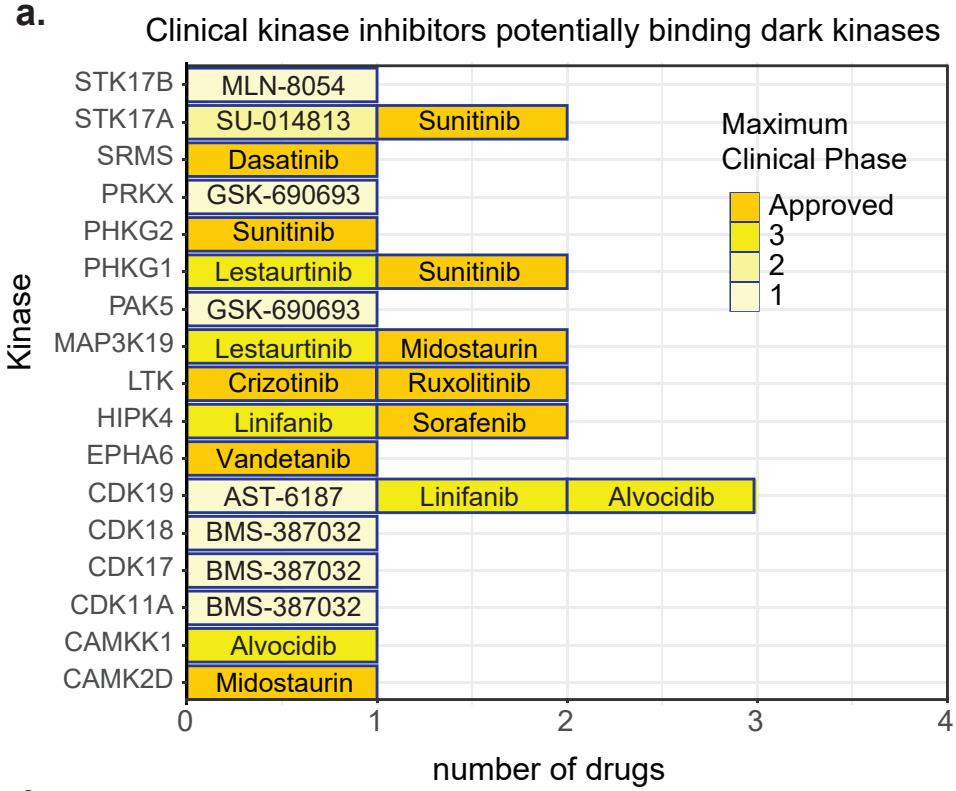
**NIM1K**
- Downregulated in 7 tumor types
- High level copy number amplification in 3 tumor types

**WEE2**
- Primarily expressed in oocytes

**PKMYT1**
- Overexpressed in multiple tumor types
- Essential for all cell lines in DepMap

**FIGURE 7**

**a.**




**b.**