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 widely studied and many kinase inhibitors have been developed into FDA-approved therapeutics. A substantial fraction of the human kinome is nonetheless understudied. In this perspective, members of the NIH Understudied Kinome Consortium mine publicly available databases to assess the functionality of these understudied kinases as well as their potential to be therapeutic targets for drug discovery campaigns. We start with a re-analysis of the kinome as a whole and describe criteria for creating an inclusive set of 710 kinase domains as well as a curated set of 557 protein kinase like (PKL) domains. We define an understudied ('dark') kinome by quantifying the public knowledge on each kinase with a PKL domain using an automatic reading machine. We find a substantial number are essential in the Cancer Dependency Map and differentially expressed or mutated in disease databases such as The Cancer Genome Atlas. Based on this and other data, it seems likely that the dark kinome contains biologically important genes, a subset of which may be viable drug targets.

INTRODUCTION

Protein phosphorylation is widespread in eukaryotic cells (Cohen, 2002) and mediates many critical events in cell fate determination, cell cycle control and signal transduction (Hunter, 1995). The 3D structures (protein folds) (Knighton et al., 1991a) and catalytic activities (Adams, 2001) of eukaryotic protein kinases (ePKs), of which ~500 are found in humans (Manning et al., 2002), 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 fold of ePKs is thought to have arisen in prokaryotes (Lai et al., 2016) and evolved to include tyrosine kinases in metazoans (Darnell, 1997; Wijk and Snel, 2020), resulting in a diverse set of enzymes (Hanks and Hunter, 1995; Hanks et al., 1988) that are often linked - in a single protein - to other catalytic domains, such as SH2 and SH3 peptide-binding domains, and other functional protein 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 (Kanev et al., 2019).

With an accessible binding pocket and demonstrated involvement in many disease pathways, protein kinases are attractive drug targets (Bhullar et al., 2018). Protein kinase inhibitors, and the few activators that have been identified (e.g. AMPK activation by salicylate and A-769662 (Hawley et al., 2012)), are diverse in mechanism and structure. These 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 (Jones, 2018) 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 (FAUVEL and Yasri, 2014). Kinase inhibitors have been intensively studied in human clinical trials and over 50 have been developed into FDA-approved drugs (Kanev et al., 2019).

Despite the general importance of kinases in cellular physiology as well as their druggability and frequent mutation in disease, a substantial subset of the kinome has been relatively little studied. This has given rise to a project within the NIH's *Illuminating the Druggable Genome* Program (IDG) (Finan et al., 2017), to investigate the understudied "dark kinome" and determine its role in human biology and disease. 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 (Huang et al., 2018). Defining the dark kinome necessarily involves a working definition of the full kinome and a survey of the current state of knowledge.

Establishing the membersip of the human kinome is more subtle than might be expected. The first human "kinome" with 514 members, defined as proteins homologous to enzymes shown experimentally to have peptide-directed phosphotransferase activity, was put forward in a groundbreaking 2002 paper by Manning et al. (Manning et al., 2002). This list has subsequently been updated via the KinHub Web resource (Eid et al., 2017a) to include 522 proteins. The kinase domain of protein Kinase A (PKA), a hetero-oligomer of a regulatory and a catalytic subunit, was the first to be crystalized and is often regarded as the prototype of the ePK fold (Knighton et al., 1991a, 1991b). This fold involves two distinct lobes with an ATP-binding catalytic cleft lying between the lobes which is characterized, at the level of primary sequence, by 12 recurrent elements with a total of ~30 highly conserved residues (Lai et al., 2016). Of 514 proteins in the kinome from Manning et al, 478 have an ePK fold.

Kinases have diverged in multiple ways to generate protein folds distinct in sequence and structure from PKA including the eukaryotic like fold, the atypical fold and unrelated folds. The eukaryotic like kinases (eLKs) are similar to ePKs in that they retain significant sequence similarity to the N-terminal region of ePKs but differ in the substrate binding lobe. TP53RK, a regulator of p53 (TP53) (Abe et al., 2001) is an example of a serine/threonine protein kinase with an eLK fold. Kinases with an atypical fold (aPKs) are distinct from ePKs and eLKs in that they have weak sequence similarity to ePKs, but nevertheless adopt an ePK-like three dimensional structure. aPKs include some well-studied protein kinases such as the DNA damage sensing ATM and ATR kinases (Abraham, 2001). The aPK and eLK folds are not limited to protein kinases. The lipid kinase PI3K, one of the most heavily mutated genes in breast cancer (Mukohara, 2015), also adopts an aPK fold (Kanev et al., 2019). Similarly, the choline kinase CHKA, a key player in dysregulated choline metabolism in cancer and a chemotherapy target, adopts the eLK fold (Glunde et al., 2011; K et al., 2016). Over 200 additional proteins are annotated as “kinase” in UniProt. The structures of these kinases are unrelated to the protein kinase fold and they are therefore termed uPKs (unrelated to Protein Kinases). Enzymes with phosphotransferase activity in the uPK family include hexokinases that phosphorylate sugars, but also protein kinases with bromodomains (e.g. BRD2, BRD3 and BRD4) as well as STK19, which displays peptide-directed phosphotransferase activity (Yin et al., 2019). Multiple uPK proteins, including those with bromodomains, bind to small molecule kinase inhibitors (Ciceri et al., 2014) making it useful, from the perspective of drug discovery, to study kinase-like proteins at the same time as kinases themselves.

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 structurally and functionally related lipid kinases and well as many of the protein families relevant to drug discovery. It would also fail to account for a lack of functional data for a substantial number of proteins, potentially excluding from consideration kinases that are physiologically or catalytically active but have not yet been tested in biochemical assays. This has resulted in definitions that rely on sequence alignment and structural data to identify closely related folds (Ciceri et al., 2014); in this definition, uPKs having kinase activity as well as bromodomains that potently bind and are inactivated by kinase inhibitors are often excluded. Moreover, as Hidden Markov Models (HMMs) and other ways of recognizing kinase homology have become more sophisticated, additional proteins have been added to the kinase tree (Briedis et al., 2008)

Kinases directed against molecules other than proteins or peptides regulate signal transduction and other eukaryotic regulatory pathways by phosphorylating second messengers and metabolites (Verheijen et al., 2011). These pathways often intersect with regulatory cascades controlled by protein kinases (Mosca et al., 2012) and some metabolic kinases have been demonstrated to also have activity against peptide or protein substrates (Lu and Hunter, 2018), challenging the conventional notion that small molecule and peptide-directed kinases are distinct families of enzymes. One example is the pyruvate kinase PKM2, which in addition to its well-known function in generating pyruvate from phosphoenolpyruvate, can also phosphorylates histone H3 at T11, thereby activating transcription downstream of EGFR-signaling (Yang et al., 2012).

These data suggest that it would be valuable to define the kinome along multiple axes, based on fold or sequence homology, ability to bind small molecule kinase inhibitors and extent of functional analysis. An expansive list is most likely 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 (Posy et al., 2011)) or chemo-proteomics in which competitive binding to ATP-like ligands on beads (so-called kinobeads (Klaeger et al., 2017) or multiplexed inhibitor beads - MIBs (Cousins et al., 2018)) is assayed using mass spectrometry. In contrast, screens for kinases that phosphorylate a specific protein sequence would logically focus on enzymes known, or likely to have peptide-directed phosphotransferase activity. Discovery programs might logically be directed at the understudied kinases expressed, and potentially functional, in normal cellular physiology and in disease.

In this perspective we generate new lists for membership in the full kinome based on the published literature and a variety of inclusion and exclusion criteria. We also re-compute membership in the understudied dark kinome using the automatic network knowledge assembling machine INDRA (Gyori et al., 2017). We consolidate available data on dark kinase activity and function with the goal of determining which understudied kinase merit additional attention. Functional evidence in this context is typically indirect, such as data from TCGA (The Cancer Genome Atlas, (Weinstein et al., 2013) on the frequency with which a kinase is mutated in a particular type of cancer. In aggregate, the evidence strongly suggests that the understudied kinome is likely to contain multiple 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 (https://darkkinome.org/).

RESULTS

The composition of the human kinome

An initial list of human kinases was obtained from Manning et al. (Manning et al., 2002) (referred to below as 'Manning') and a second from Eid et al.(Eid et al., 2017a) (via the Kinhub Web resource); a list of dark kinases according to IDG was obtained from the NIH solicitation (updated in January 2018) and a fourth list of all 684 proteins tagged as "kinases" was obtained from UniProt; this list includes protein kinases, lipid kinases and other small molecule kinases. 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 (HMM models) and structural analysis (Kannan et al., 2007) to subdivide domains into three 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), (Kannan et al., 2007). PKLs were further subdivided into eukaryotic protein kinases (ePKs, discussed in the introduction), eukaryotic like kinases (eLKs) and kinases with an atypical fold (aPKs) as previously described (Kannan and Neuwald, 2005; Kannan et al., 2007). 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 (Kannan et al., 2007)). aPKs, on the other hand, display no significant sequence similarity to ePKs and eLKs, but nevertheless adopt the canonical protein kinase fold. Most aPKs 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**). Unfortunately, the nomenclature used in making these distinctions 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 kinase group" (AKGs). These domains are usually depicted alongside the familiar

Coral kinase dendogram (Metz et al., 2018) and include protein kinases such as ATM and ATR as well as bromo-domains and TRIM proteins (see below).

As noted previously (Garrett et al., 2011; Manning et al., 2002), 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 (Jassal et al., 2020) 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 and not proteins (Baumlova et al., 2014). Thus, even after extensive computational analysis, some manual curation of the kinome is necessary. We have 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. In many cases, it is unknown if proteins in the extended list have experimentally-validated phospho-transfer activity, and if so whether it is directed against peptides, small molecules or both (Lu and Hunter, 2018). 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 (Reymond et al., 2001) that regulate and are regulated by kinases (Ozato et al., 2008), but have no 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 utility of the extended kinome to drug discovery involving kinase inhibitors can be demonstrated by re-analysis of a large-scale chemo-proteomic dataset collected using multiplexed inhibitor beads (Klaeger et al., 2017). Overall, we found that 48 domains in the extended kinome list and not in the curated list bound to kinobeads and eight 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. Because non-protein kinases participate in metabolic pathways and signaling networks, an expansive list including these non-protein kinases facilitates a systemic analysis of the mechanisms of action (MoA) of kinase inhibitors. Thus, extended and curated kinomes and their different sublists are useful in different settings.

Identifying understudied kinases

The original IDG dark kinome list was assembled using a bibliometric tool, TIN-X (Cannon et al., 2017), 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 (Huang et al., 2018). We therefore took a complementary approach based on the computational tool INDRA (the Integrated Network and Dynamical Reasoning Assembler), (Gyori et al., 2017; Todorov et al., 2019). INDRA uses multiple NLP systems (McDonald et al., 2016; Valenzuela-Escarcega et al., 2017) to extract computable statements about biological mechanism from PubMed abstracts and full text articles in PubMedCentral. Unlike other bibliometric tools, INDRA aggregates data from multiple pathway databases (such as BioGrid (Stark et al., 2006) and PathwayCommons (Cerami et al., 2011)) and specialized resources such as the LINCS compound database and the Target Affinity Spectrum from the Small Molecule Suite database (Moret et al., 2019).

INDRA differs from simpler bibliometric tools because it is able to homogenize and disambiguate biological entities from different sources and maximize the extraction of mechanistic information. This is particularly important when a protein has multiple names, or its name changes over time. For example, the dark kinase PEAK3 was originally known as C19orf35 and was recently found to be a biologically active pseudokinase (Lopez et al., 2019). INDRA consolidates biological statements using both PEAK3 and C19orf35 as identifiers and represents them with the now-standard HGNC name PEAK3. Whenever the information is available, INDRA statements are detailed with respect to molecular mechanism and they are linked to the underlying knowledge support (the database reference or citation). 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 and human-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). Each INDRA statement is therefore a unique biochemical mechanism of action rather than a paper count. 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 can be visualized via the NDEx service (Pratt et al., 2015). 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 only 78 INDRA statements, but a high TIN-X novelty score of ~808. The reason for this inconsistency is still under investigation but in general, such errors reflect the difficulty of linking common names for genes and proteins to their unique identifiers 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. For example, it correctly associates 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* (Moret et al., 2019), 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**). We also mined PHAROS (Nguyen et al., 2017), which classifies drug targets based on whether or not they are bound by an FDA-approved drug (Tclin) or a 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 a 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., 2019) 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 only by 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 and TIN-X 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 list to define the "dark kinome". When the distribution of dark kinases is viewed using the standard Coral kinase dendrogram (Metz et al., 2018), an 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. a branch with four TSSK and another with three STK32 kinases; dashed red outline). In yet other cases, a well-studied kinase is closely related to a dark kinases, SIK1 and SIK1B or WEE1 and WEE2 for example, but it is unknown whether such pairs of isozymes are similar or redundant functionally.

Evidence for dark kinase expression and function

To consolidate existing data on the expression and possible functions of understudied kinases, we analyzed RNAseq data for 1019 cell lines in the Cancer Cell Line Encyclopedia (CCLE) (Nusinow et al., 2020), proteomic data for 375 cell lines in the CCLE (Nusinow et al., 2020) and loss of function data in the Cancer Dependency Map (DepMap). The DepMap was generated using lentivius-based RNAi or CRISPR/Cas9 libraries in pooled screens to identify genes that are essential in one or more of a ~1000 cell lines (Tsherniak et al., 2017).

Based on RNASeq data, non-dark and dark kinases were observed to vary substantially in abundance across 1019 CCLE cell lines. Using the common threshold of RPKM ≥1 (Reads Per Kilobase of transcript, per Million mapped reads) (Kryuchkova-Mostacci and Robinson-Rechavi, 2017) evidence of expression was found in at least one cell 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 lipid kinases PI4KA and PIP4K2C all had maximum expression levels similar to that of the abundant and well-studied LCK tyrosine kinase. Overall, however, dark kinases had significantly lower maximum mRNA expression levels than well studied kinases by multiple criteria (2.1 vs 5.8 RPKM median expression level, p-value=4.6x10$_{-8}$; 36 vs 71 RPKM maximum expression level, p-value=2.2x10$_{-16}$ by Wilcoxon rank sum test). In CCLE proteomic data we observed that 367 kinases from the curated kinome were detected at the level of at least one peptide per protein; 110 of these were dark kinases. The difference between proteomic and mRNA data overall is likely to reflect the lower sensitivity of shotgun proteomics, but some kinases might also be subjected to translational regulation. 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**)**. We conclude that a substantial number of dark kinases are expressed in human cells lines and a subset are 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 (not detectable by shotgun proteomics) and genes can have important functions without necessarily resulting in growth defects assayable by DepMap methodology.

Dark kinases in disease

To study the possible roles of dark kinases in pathophysiology, we mined associations between diseases and either gene mutations or changes in expression. We examined The Cancer Genome Atlas (TCGA), Accelerating Medicines Partnership - Alzheimer's Disease (AMP-AD) and a microarray dataset on changes in gene expression associated with chronic obstructive pulmonary disease (COPD;

(Rogers et al., 2019). COPD progressively impairs a patients' ability to breathe and is the third leading cause of death in the US. In TCGA, 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 (Garraway and Lander, 2013) and we then 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; thus, observed mutations may affect functions other than kinase activity. We performed differential expression and mutation frequency 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 two-sided test with H_0 = light and dark kinases have similar aberrations p=0.86) **(Figure 5)**. For example, in colorectal adenocarcinoma, the dark kinase MAPK4 is one of the three most highly downregulated kinases whereas LMTK3, NEK5 and STK31 represent four of the seven most high upregulated kinases (**Figure S3A**). This is consistent with a report that overexpression of STK31 can inhibit the differentiation of colorectal cancer cells. (Fok et al., 2012).

By mining 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 (DLBCL), the dark kinase ITPKB, which has been reported to phosphorylate inositol 1,4,5-trisphosphate and regulate B cell survival (Schurmans et al., 2011), is more frequently mutated than KDR (~13% vs. 8% of patient samples, **Figure S3B**). Overexpression of KDR is known to promote angiogenesis and correlate with poor patient survival (Gratzinger et al., 2010; Holmes et al., 2007; Jørgensen et al., 2009). More recently a case study reported that patients carrying an ITPKB C873F substitution mutation had Richter's syndrome (which is characterized by sudden development of B cell chronic lymphocytic leukemia into a faster-growing and more aggressive DLBCL) and clones harboring the ITPKB C873F mutation exhibited higher growth rates (Landau et al., 2017). 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. Our analysis 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, making them good candidates for future testing as oncogenes or tumor suppressors.

To explore the roles of dark kinases in other diseases, we analyzed data from the AMP-AD program Target Discovery and Preclinical Validation (Hodes and Buckholtz, 2016)). 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 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 COPD, based on a study by Rogers et al (Rogers et al., 2019) of five COPD microarray datasets from Gene Expression Omnibus (GEO) and two COPD datasets from ArrayExpress that aimed 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 patients (adjusted p-value < 0.05). We analyzed these data and found that the dark kinase PIP4K2C, which is potentially immune regulating (Shim et al., 2016), was significantly downregulated in individuals with COPD (adjusted p-value = 0.048). Additionally, CDC42BPB, nominally involved in cytoskeleton organization and cell migration (Tan et al., 2008, 2011), was significantly upregulated (adjusted p-value = 0.026) (**Figure 5D**). In total, five dark kinases versus fifteen non-dark kinases were differentially expressed in COPD patients. As additional data on gene expression and mutation become available for other diseases, it will be possible further expand the list of dark kinases potentially implicated in human health.

A dark kinase network regulating the cell cycle

Inspection of INDRA networks revealed that dark kinases, like well-studied kinases, function in networks of interacting kinases. One illustrative example involves control of the central regulator of cell cycle progression, CDK1, by the dark kinases PKMYT1, WEE2, BRSK1 and NIM1K **(Figure 6)**. Although the homologues of some of these kinases have been well studied in fission and budding yeast, not as much is known about the human kinases (Wu and Russell, 1993). WEE2, whose expression is described to be oocyte-specific (Sang et al., 2018) (but can is also be detected in seven CCLE lines, six from lung cancer and one from large intestine) is likely to be similar in function to the well-studied and widely-expressed homologue WEE1,which phosphorylates CDK1 on the negative regulatory site T15 (Sang et al., 2018) whereas PKMYT1 phosphorylates CDK1 on the Y14 site to complete the inhibition of CDK1 (Liu et al., 1997; Mueller et al., 1995). These modifications are removed by the CDC25 phosphatase, which promotes cell cycle progress from G2 into M phase (Santamaría et al., 2007). PKMYT1 and WEE1 are essential in nearly all cells, according to DepMap (DepMap, 2019) (although WEE2 is not). Upstream of WEE1, the dark kinases BRSK1 (127 INDRA statements) and NIM1K (28 INDRA statements) and the better-studied BRSK2 (176 INDRA statements) function to regulate WEE1. Neither PKMYT1, BRSK1 nor NIM1K have selective small molecule inhibitors described in the public literature (Asquith et al., 2020); several WEE1 inhibitors are in clinical development however (Matheson et al., 2016), and these molecules are likely to inhibit WEE2 as well. It is remarkable that enzymes so closely associated with the essential cell cycle regulator CDK1, remain relatively understudied in humans (Wu and Russell, 1993). 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 small molecules by using the *selectivity score* (Moret et al., 2019) to mine public data for evidence of known binding and known not-binding. We identified 13 dark kinases that may be 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). For example, the anti-cancer drug sunitinib 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, which are 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)(Bender et al., 2006; Nguyen et al., 2013). 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 by BAFP 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 (Fabian et al., 2005), 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 (Burley et al., 2008) and its successors but have not been subsequently discussed in the published literature. The availability of these resources and data are summarized in Supplementary Table 1.

DISCUSSION

In this perspective we revisit the criteria used to define membership in the human kinome. This is not a trivial task because no single functional, structural or historical definition exists. We have therefore assembled a table of protein domains that can be used to generate more or less expansive sets of kinases in a data-driven manner based on criteria such as function, sequence homology, known or predicted structure etc, (see **Supplementary Table 1**). The table can also be used to generate lists of classical protein kinases, lipid kinases, nucleotide kinases etc. Creation of the master table involved consolidating multiple overlapping kinome lists, which often use different rules. Thus, not all assignments are unambiguous. We also note that the biochemical activities assigned to many kinase remain provisional, particularly in the case of understudied kinases. For example, it is increasingly clear that some kinases can phosphorylate both small molecules and proteins. The NIH IDG group intends to maintain an updated version of **Supplementary Table 1** at its website (https://darkkinome.org/).

The human kinome includes ~50 so-called "pseudokinases" that, based on sequence alignment, 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 (Kwon et al., 2019). 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 (Sliwkowski et al., 1994). ERBB3 over-expression also promotes resistance to therapeutic ERBB2 inhibitors in breast cancer (Garrett et al., 2011). Some proteins commonly annotated as pseudokinases have been found to 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 (Eswaran et al., 2009; Villa et al., 2009). H3 phosphorylation by Haspin changes chromatin structure and mitotic outcome and is therefore physiologically important (Dai et al., 2005). The existence of biologically active pseudokinases, some of which may actually have phosphotransferase activity, is but one way in which equating kinases with a specific fold, sequence similarity, or enzymatic function is inadequate.

Based on our work, the most useful definitions of the human kinome are likely to be an expansive 710 domain "extended kinome" that broadly encompasses related sets of folds, sequences and biological functions. This list will be relevant to machine learning, chemoproteomics, small molecule profiling and genomic studies in which an expansive view of the kinome is advantageous. As one example of such a list we generated a set of 557 "curated kinase" domains that is most similar in spirit to the original definition of the kinome generated by Manning (Manning et al., 2002) nearly two decades ago. This list is most useful in the study of kinases as a family of genes with related biochemistry and cellular function. The computational analysis in this perspective focuses on this curated kinase set.

The amount public knowledge on specific kinases spans at least four orders of magnitude when measured by the number of unique causal and mechanistic statements that can be extracted using INDRA, a text mining and knowledge assembly software. Unsurprisingly, famous drug targets 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: CLK3 and WNK3, for example, both have crystal structures and available as recombinant protein on Reaction Biology and DiscoverX kinase panel. In general, we find that INDRA correlates well with more conventional bibliometric measures, (Huang et al., 2018)(Cannon et al., 2017) and also with whether or not a domain has been successfully targeted with selective or clinical grade small molecules. Following the lead of the NIH IDG program, we have defined the understudied 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.

In the course of preparing this perspective we repeatedly ran into the challenge of correctly associating kinases described in the literature with their canonical (HGNC or UniProt) names. Many kinases have multiple names, often several common ones such as MEK1 or MAPKK1 as well as a standardized one such as MAP2K1 (HGNC:6840). In many cases, the name space changes over time (e.g. PEAK3 instead of C19orf35 (Lopez et al., 2019)). Humans find it arduous to make these associations across a vast literature and, even after extensive human training, tools based on state of the art NLP such as INDRA cannot correctly ground all named entities. We have also observed that many of the regulatory sites on kinases are mis-numbered, either because residue number changed over time or confusion over isoforms and even species (Bachman et al., 2019). This leads to the problem of "unknown knowns" – namely facts that have been established (or data that have been collected) but are no longer findable by the community. One of the tasks of the IDG group is therefore to identify such sources of "lost" information and correctly associate them with systematic knowledge repositories. In the specific case of the dark kinome described here, we welcome information on data we might have missed on specific kinases.

We find that at least 175 of the least studied kinases as defined above, are expressed in CCLE cell lines (the largest available cell line panel analyzed to date) when measured by protein or by mRNA expression (and in many case, by both). We also find that half of all dark kinases are essential in two of more of the 625 cell lines annotated in the DepMap (Tsherniak et al., 2017) and 10 are essential in at least 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 others are differentially expressed in disease databases for Alzheimer's Disease or COPD. These largely indirect findings suggests that a substantial subset of dark kinases are functional in normal physiology and in disease. 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 also be valuable as therapeutic targets.

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(Eswaran et al., 2009; Manning et al., 2002; Talevich et al., 2011). 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(Huo et al., 2017) and the presence/absence of key structural features distinguishing ePKs, eLKs and aPKs, as described previously(Kannan and Neuwald, 2005; Kannan et al., 2007). 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(Yin et al., 2019).

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(Gyori et al., 2017). 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*(Moret et al., 2019), 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 ≤ 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 ≤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(Dempster et al., 2019). 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(Cerami et al., 2012; Gao et al., 2013). 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 two-sided 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([CSL STYLE ERROR: reference with no printed form.]). In summary, these counts were derived from raw reads using the STAR aligner(Dobin et al., 2013) 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(Mostafavi et al., 2018) 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(Love et al., 2014). 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(Zhu et al., 2019).

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/8233

175. Data was preprocessed as described in Rogers et al(Rogers et al., 2019). 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.(Eid et al., 2017b) and accessible via http://kinhub.org/kinases.html; Manning (red) refers to the gene list prepared by Manning et al. in 2002(Manning et al., 2002); Uniprot kinase activity (green) refers to a list of genes annotated as having kinase activity in the Uniprot database(2019) (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)(Manning et al., 2002) 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(Cannon et al., 2017) (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(Metz et al., 2018). 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(Barretina et al., 2012). 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.(Kryuchkova-Mostacci and Robinson-Rechavi, 2017) **(B)** Number of cell lines for which the Dependency Map(Tsherniak et al., 2017) score indicates essentiality (using the recommended *Post-Ceres*(Meyers et al., 2017) 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.(Cerami et al., 2012) 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(Klaeger et al., 2017).

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).

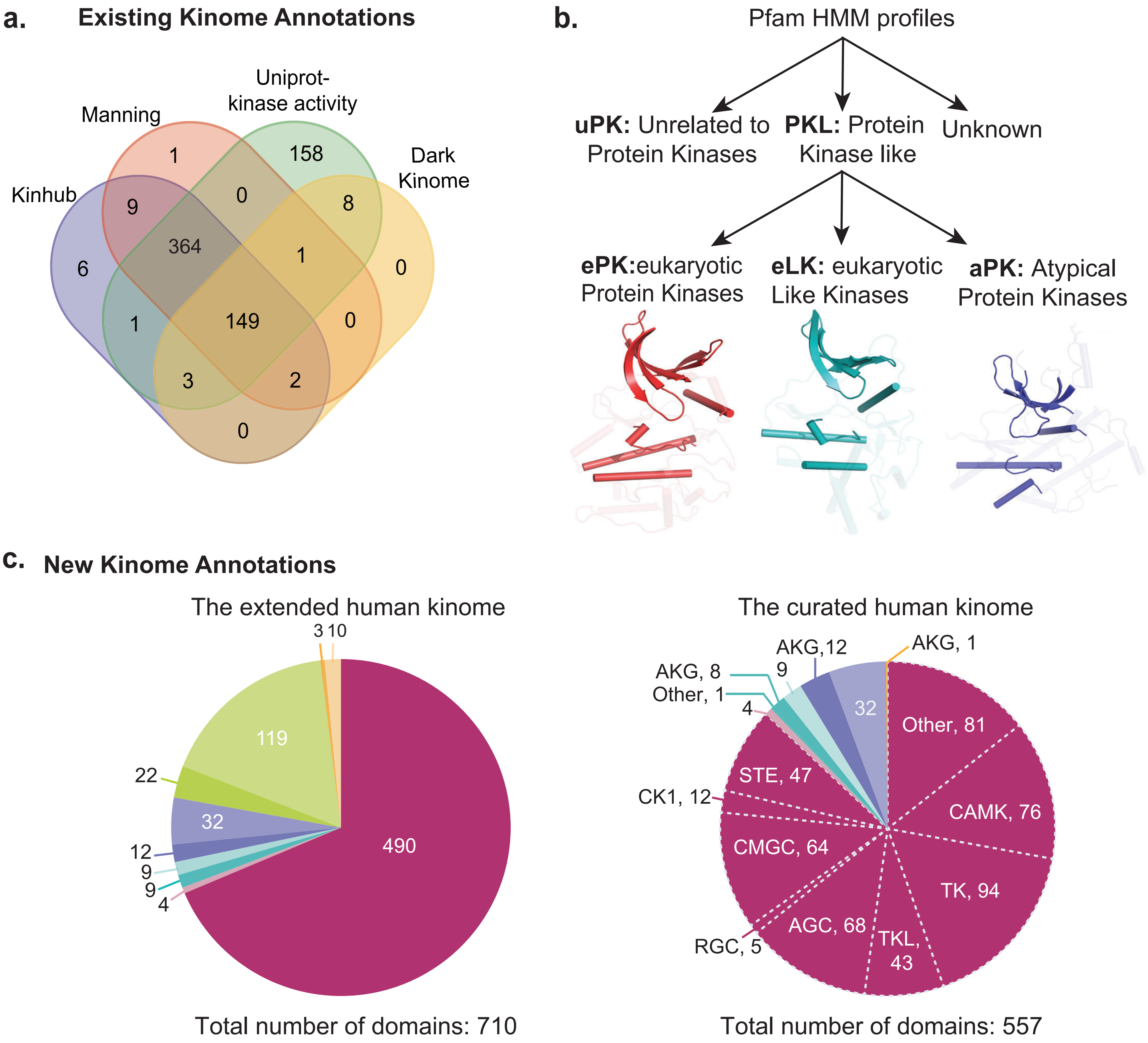
Supplementary Table 3 – Clinical compounds targeting dark kinases

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

Automatic citation updates are disabled. To see the bibliography, click Refresh in the Zotero tab.

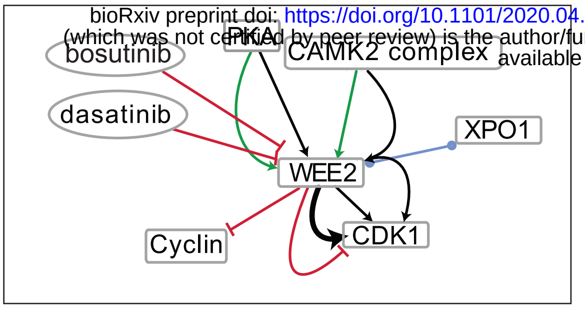
FIGURE 1



	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 PMID: 21454751 PMID: 19550110

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

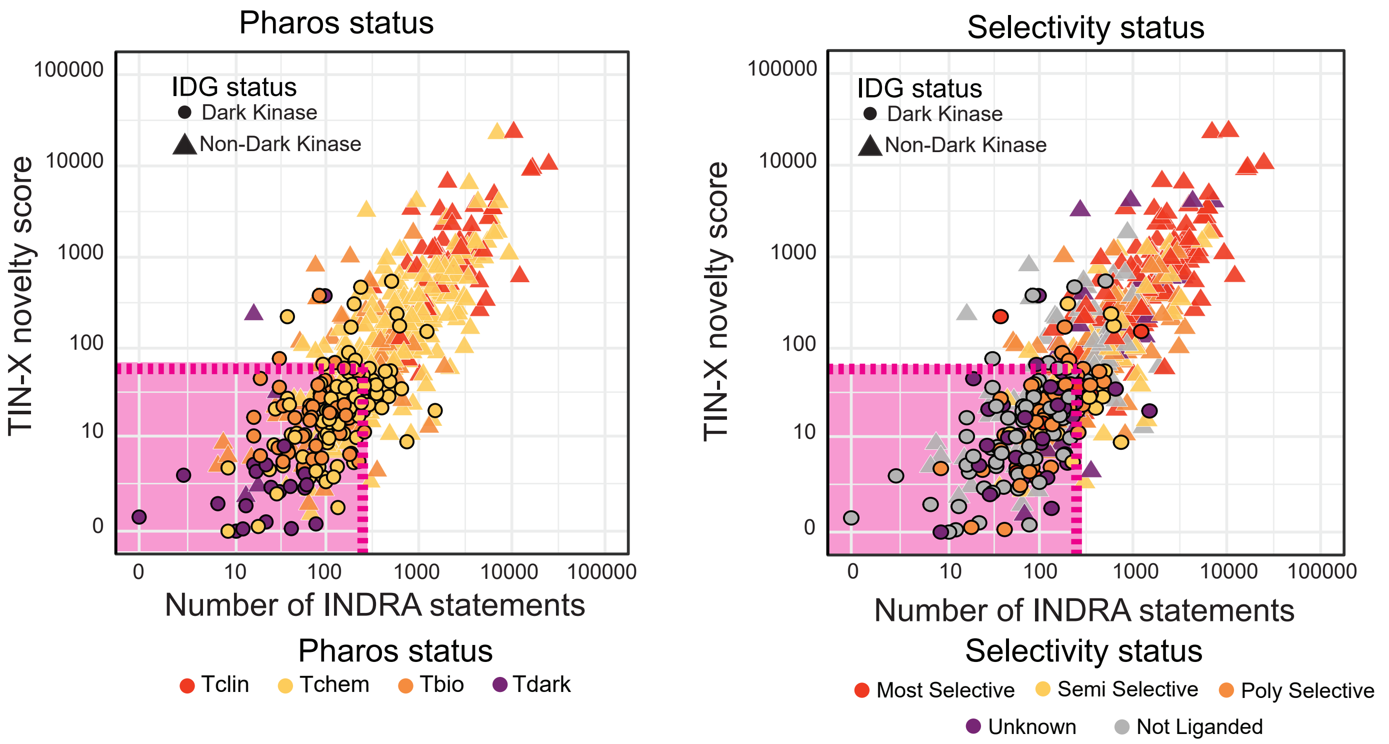


FIGURE 3

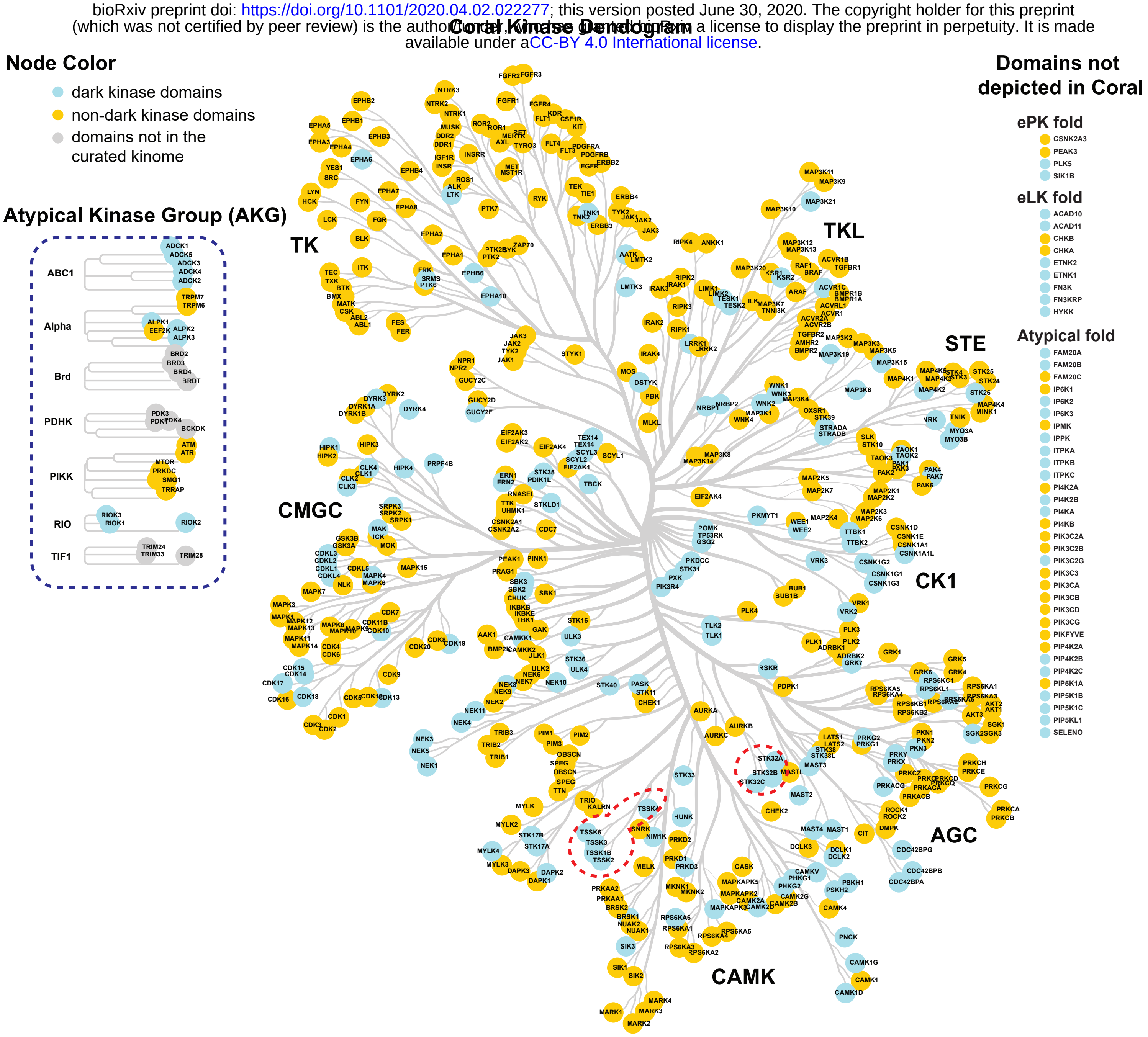

Coral Kinase Dendrogram
Node Color
dark kinase domains
non-dark kinase domains
domains not in the curated kinome
Atypical Kinase Group (AKG)
ABC1
ADCK1
ADCK5
ADCK3
ADCK4
ADCK2
TRPM7
TRPM6
Alpha
ALPK1
EEF2K
ALPK2
ALPK3
Brd
BRD2
BRD3
BRD4
BRDT
PDHK
PDK3
PDK1
PDK4
BCKDK
PIKK
ATM
ATR
MTOR
PRKDC
SMG1
TRRAP
RIO
RIOK3
RIOK1
RIOK2
TIF1
TRIM24
TRIM33
TRIM28
TK
TKL
STE
CK1
AGC
CAMK
CMGC
EPHB2
EPHA5
EPHB1
EPHA3
EPHA4
EPHB3
EPHA6
EPHB4
YES1
SRC
LYN
HCK
FYN
EPHA7
EPHA8
LCK
FGR
BLK
EPHA2
ITK
TEC
TXK
BTK
BMX
MATK
CSK
ABL2
ABL1
FES
FER
FRK
SRMS
PTK6
EPHA1
EPHB6
EPHA10
NTRK3
NTRK2
NTRK1
MUSK
DDR2
DDR1
IGF1R
INSR
INSRR
ROS1
ALK
LTK
ROR2
ROR1
MERTK
AXL
TYRO3
RET
MET
MST1R
PTK7
RYK
FGFR2
FGFR3
FGFR1
FGFR4
FLT1
KDR
CSF1R
KIT
FLT4
FLT3
PDGFRA
PDGFRB
EGFR
ERBB2
ERBB4
ERBB3
TEK
TIE1
TNK2
TNK1
TYK2
JAK1
JAK2
JAK3
PTK2B
PTK2
SYK
ZAP70
STYK1
NPR1
NPR2
GUCY2C
GUCY2D
GUCY2F
MAP3K11
MAP3K9
MAP3K21
MAP3K10
MAP3K12
MAP3K13
RIPK4
ANKK1
AATK
LMTK2
LMTK3
IRAK3
IRAK1
RIPK2
RIPK3
IRAK2
RIPK1
IRAK4
LRRK1
LRRK2
MOS
DSTYK
PBK
MLKL
MAP3K20
KSR1
KSR2
RAF1
BRAF
ARAF
ACVR1B
TGFBR1
ACVR1C
BMPR1B
BMPR1A
ACVRL1
ACVR1
ACVR2A
ACVR2B
TGFBR2
AMHR2
BMPR2
LIMK1
LIMK2
TESK1
TESK2
ILK
MAP3K7
TNNI3K
MAP3K2
MAP3K3
MAP3K5
MAP3K15
MAP3K19
MAP3K6
MAP4K5
MAP4K1
STK4
MAP4K3
STK3
STK25
STK24
MAP4K2
STK26
MAP4K4
MINK1
NRK
TNIK
MYO3A
MYO3B
WNK1
WNK3
WNK2
MAP3K4
MAP3K1
WNK4
NRBP1
NRBP2
OXSR1
STK39
STRADA
STRADB
SLK
STK10
TAOK1
TAOK2
TAOK3
PAK1
PAK2
PAK3
PAK4
PAK7
PAK6
MAP2K5
MAP2K7
MAP2K1
MAP2K2
MAP2K3
MAP2K6
MAP2K4
MAP3K8
MAP3K14
EIF2AK4
PKMYT1
WEE1
WEE2
TTBK1
TTBK2
CSNK1D
CSNK1E
CSNK1A1
CSNK1A1L
CSNK1G2
CSNK1G1
CSNK1G3
VRK3
VRK1
VRK2
POMK
TP53RK
GSG2
PKDCC
STK31
PXK
PIK3R4
BUB1
BUB1B
PLK4
TLK2
TLK1
PLK3
PLK1
PLK2
ADRBK1
ADRBK2
GRK7
GRK1
GRK5
GRK6
GRK4
RSKR
PDPK1
RPS6KA5
RPS6KA4
RPS6KC1
RPS6KL1
RPS6KA1
RPS6KA3
RPS6KB1
RPS6KA6
RPS6KB2
AKT2
AKT1
AKT3
SGK1
SGK3
SGK2
PKN1
PKN2
PKN3
PRKG2
PRKG1
PRKY
PRKX
PRKCH
PRKCE
PRKCZ
PRKCI
PRKCD
PRKCQ
PRKACA
PRKACB
PRKACG
PRKCG
PRKCA
PRKCB
LATS1
LATS2
STK38
STK38L
MAST3
MAST2
MAST4
MAST1
STK32A
STK32B
STK32C
MASTL
CHEK2
ROCK1
ROCK2
DMPK
CIT
CDC42BPG
CDC42BPB
CDC42BPA
AURKA
AURKB
AURKC
STK33
HUNK
CHEK1
PASK
STK11
TSSK4
TSSK6
TSSK3
TSSK1B
TSSK2
SNRK
NIM1K
PRKD2
PRKD1
PRKD3
MELK
DCLK3
DCLK1
DCLK2
CASK
CAMKV
PHKG1
PHKG2
PSKH1
PSKH2
MAPKAPK5
MAPKAPK2
MAPKAPK3
CAMK2A
CAMK2D
CAMK2G
CAMK2B
CAMK4
MKNK1
MKNK2
RPS6KA6
RPS6KA1
RPS6KA4
RPS6KA5
RPS6KA3
RPS6KA2
PNCK
CAMK1G
CAMK1
CAMK1D
PRKAA2
PRKAA1
BRSK2
BRSK1
NUAK2
NUAK1
SIK3
SIK1
SIK2
MARK4
MARK1
MARK3
MARK2
TRIO
KALRN
OBSCN
SPEG
TTN
PIM1
PIM3
PIM2
TRIB3
TRIB2
TRIB1
MYLK
MYLK2
STK17B
STK17A
MYLK4
MYLK3
DAPK3
DAPK2
DAPK1
NEK3
NEK5
NEK1
NEK4
NEK11
NEK2
NEK8
NEK9
NEK6
NEK7
NEK10
STK40
STK36
ULK4
ULK2
ULK1
ULK3
STK16
AAK1
BMP2K
CAMKK1
CAMKK2
GAK
PEAK1
PINK1
PRAG1
SBK3
SBK2
CHUK
SBK1
IKBKB
IKBKE
TBK1
ERN1
ERN2
STK35
PDIK1L
TBCK
RNASEL
TTK
UHMK1
STKLD1
CSNK2A1
CSNK2A2
CDC7
EIF2AK3
EIF2AK2
EIF2AK4
TEX14
SCYL3
SCYL2
SCYL1
EIF2AK1
DYRK2
DYRK3
DYRK1A
DYRK1B
DYRK4
HIPK1
HIPK2
HIPK3
CLK4
CLK1
CLK2
CLK3
HIPK4
PRPF4B
SRPK3
SRPK2
SRPK1
MAK
ICK
MOK
GSK3B
GSK3A
CDKL3
CDKL2
CDKL1
CDKL4
CDKL5
NLK
MAPK4
MAPK6
MAPK15
MAPK7
MAPK3
MAPK1
MAPK12
MAPK13
MAPK11
MAPK14
MAPK8
MAPK10
MAPK9
CDK7
CDK11B
CDK11A
CDK10
CDK4
CDK6
CDK20
CDK8
CDK19
CDK15
CDK14
CDK17
CDK16
CDK18
CDK9
CDK5
CDK12
CDK13
CDK1
CDK3
CDK2
Domains not depicted in Coral
ePK fold
CSNK2A3
PEAK3
PLK5
SIK1B
eLK fold
ACAD10
ACAD11
CHKB
CHKA
ETNK2
ETNK1
FN3K
FN3KRP
HYKK
Atypical fold
FAM20A
FAM20B
FAM20C
IP6K1
IP6K2
IP6K3
IPMK
IPPK
ITPKA
ITPKB
ITPKC
PI4K2A
PI4K2B
PI4KA
PI4KB
PIK3C2A
PIK3C2B
PIK3C2G
PIK3C3
PIK3CA
PIK3CB
PIK3CD
PIK3CG
PIKFYVE
PIP4K2A
PIP4K2B
PIP4K2C
PIP5K1A
PIP5K1B
PIP5K1C
PIP5KL1
SELENO

FIGURE 4

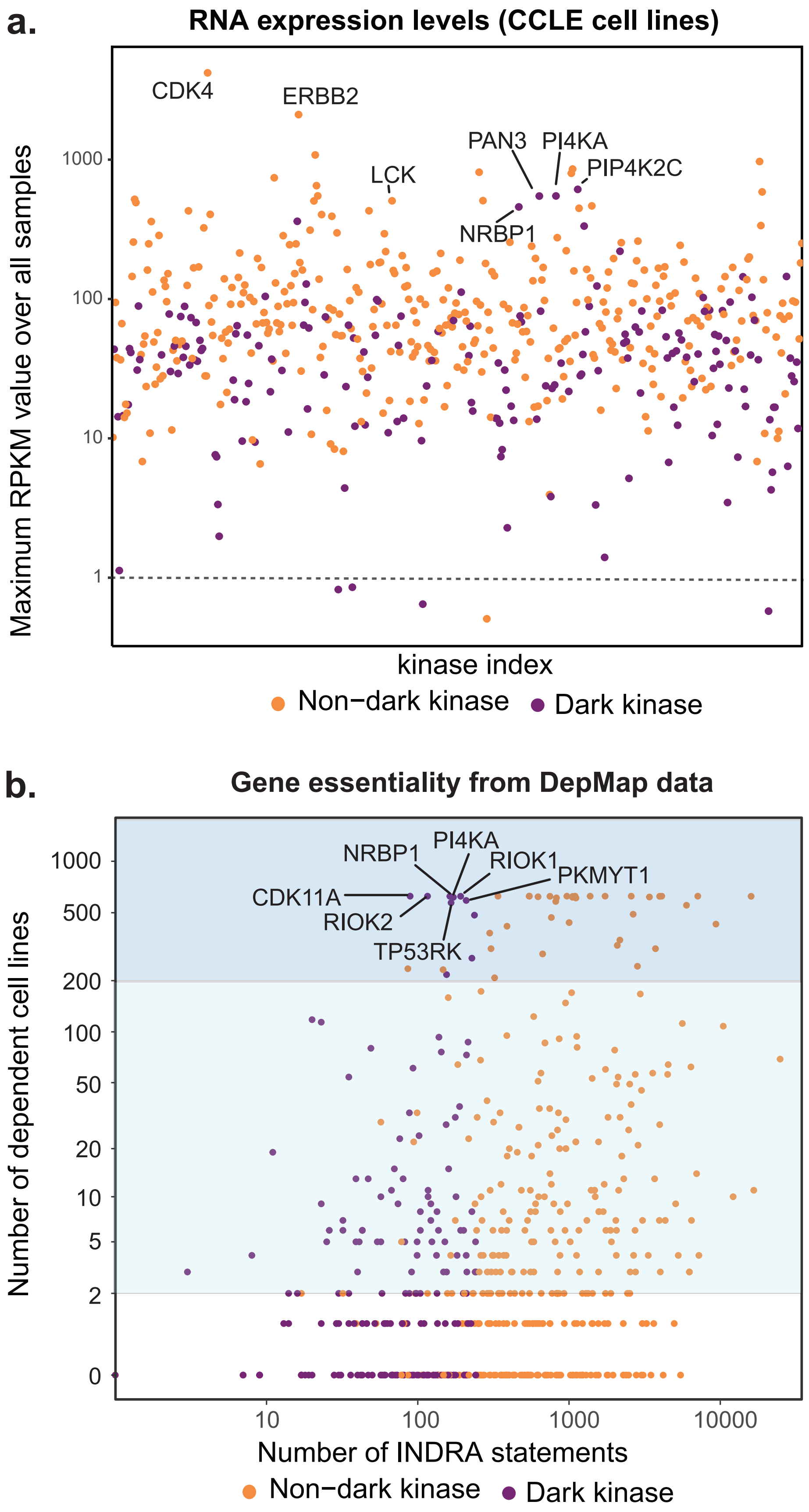

a.
RNA expression levels (CCLE cell lines)
CDK4
ERBB2
PAN3
PI4KA
PIP4K2C
LCK
NRBP1
1000
100
10
1
Maximum RPKM value over all samples
kinase index
● Non−dark kinase
● Dark kinase
b.
Gene essentiality from DepMap data
NRBP1
PI4KA
RIOK1
PKMYT1
CDK11A
RIOK2
TP53RK
1000
500
200
100
50
20
10
5
2
0
Number of dependent cell lines
10
100
1000
10000
Number of INDRA statements
● Non−dark kinase
● Dark kinase

FIGURE 5

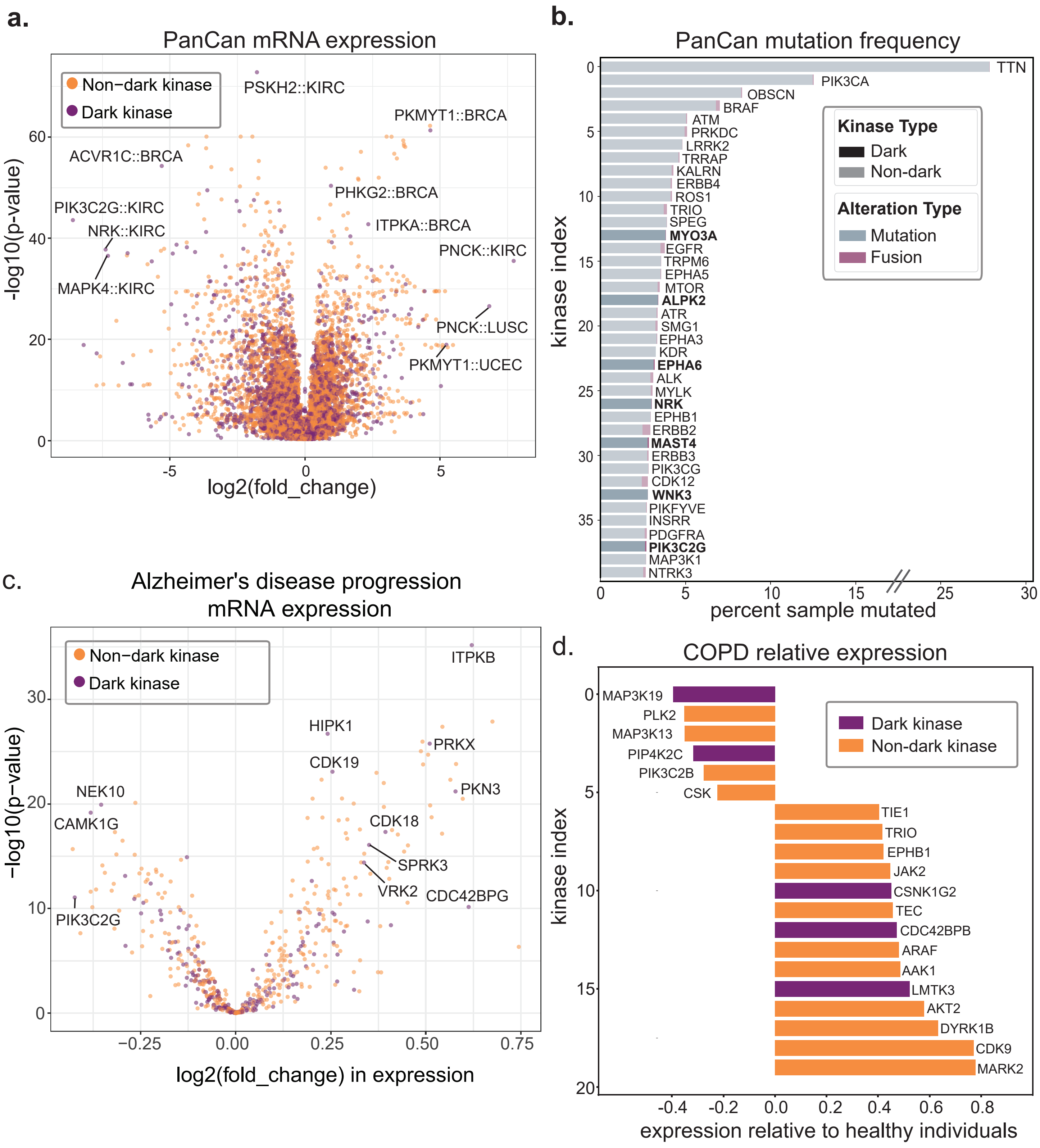
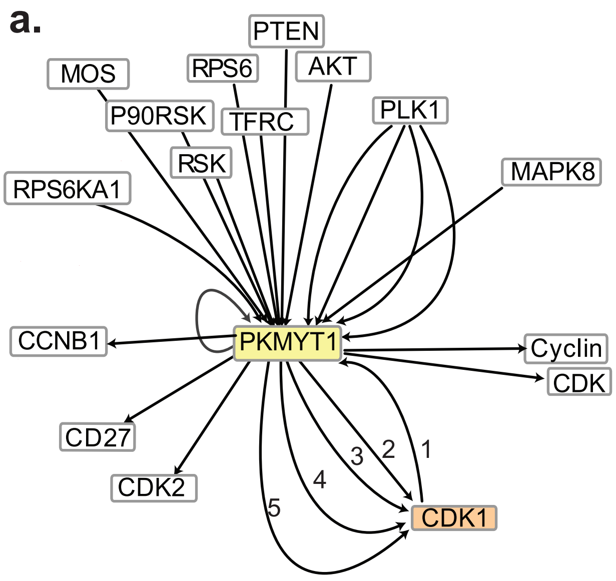
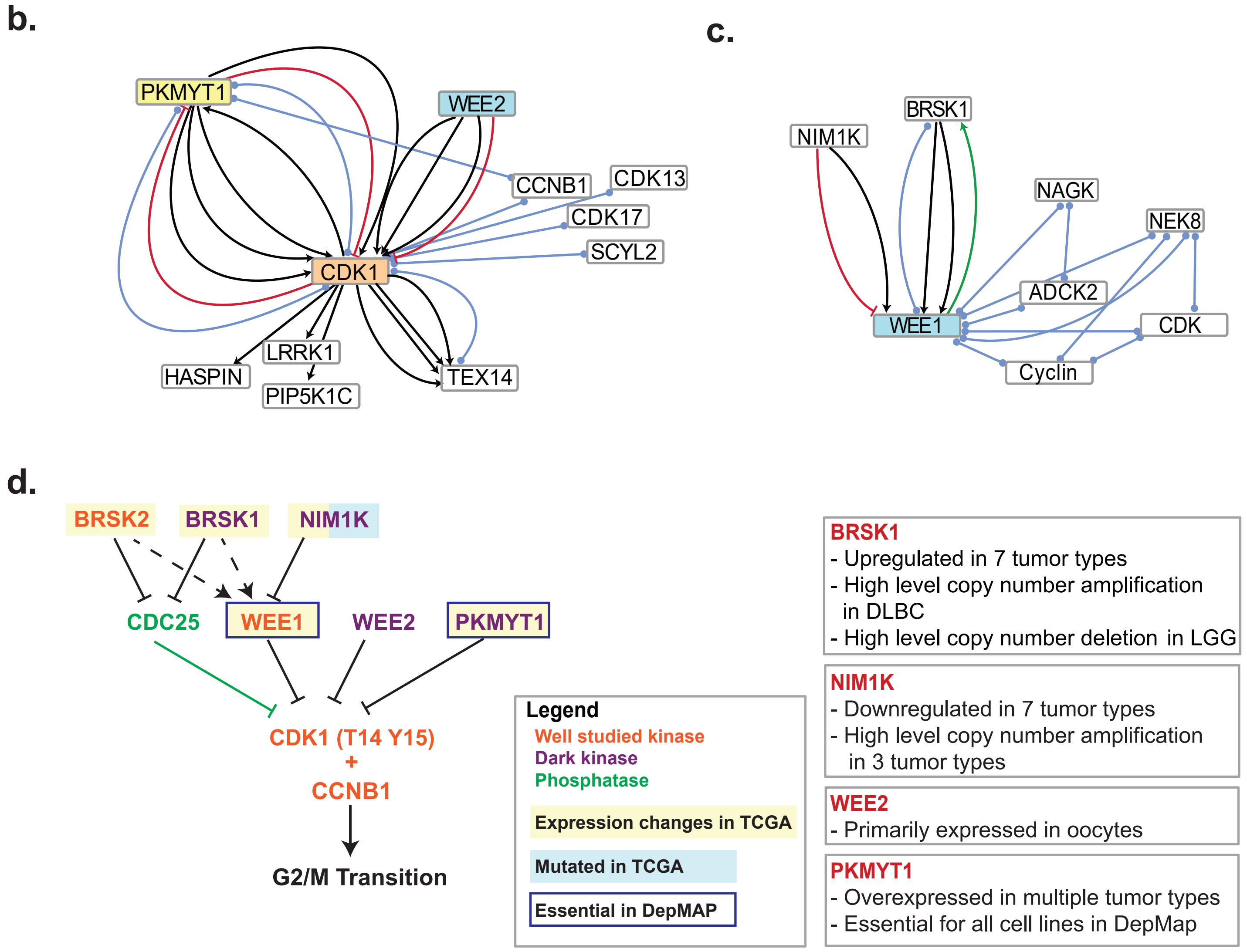

a.
PanCan mRNA expression
Non−dark kinase
Dark kinase
PSKH2::KIRC
PKMYT1::BRCA
ACVR1C::BRCA
PHKG2::BRCA
PIK3C2G::KIRC
ITPKA::BRCA
NRK::KIRC
PNCK::KIRC
MAPK4::KIRC
PNCK::LUSC
PKMYT1::UCEC
-log10(p-value)
60
40
20
0
-5
0
5
log2(fold_change)
b.
PanCan mutation frequency
TTN
PIK3CA
OBSCN
BRAF
ATM
PRKDC
LRRK2
TRRAP
KALRN
ERBB4
ROS1
TRIO
SPEG
MYO3A
EGFR
TRPM6
EPHA5
MTOR
ALPK2
ATR
SMG1
EPHA3
KDR
EPHA6
ALK
MYLK
NRK
EPHB1
ERBB2
MAST4
ERBB3
PIK3CG
CDK12
WNK3
PIKFYVE
INSRR
PDGFRA
PIK3C2G
MAP3K1
NTRK3
Kinase Type
Dark
Non-dark
Alteration Type
Mutation
Fusion
kinase index
0
5
10
15
20
25
30
35
0
5
10
15
25
30
percent sample mutated
c.
Alzheimer's disease progression
mRNA expression
Non−dark kinase
Dark kinase
ITPKB
HIPK1
PRKX
CDK19
PKN3
NEK10
CAMK1G
CDK18
SPRK3
VRK2
CDC42BPG
PIK3C2G
−log10(p−value)
30
20
10
0
−0.25
0.00
0.25
0.50
0.75
log2(fold_change) in expression
d.
COPD relative expression
MAP3K19
PLK2
MAP3K13
PIP4K2C
PIK3C2B
CSK
TIE1
TRIO
EPHB1
JAK2
CSNK1G2
TEC
CDC42BPB
ARAF
AAK1
LMTK3
AKT2
DYRK1B
CDK9
MARK2
Dark kinase
Non-dark kinase
kinase index
0
5
10
15
20
-0.4
-0.2
0.0
0.2
0.4
0.6
0.8
expression relative to healthy individuals

FIGURE 6



	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	...	...	...	...



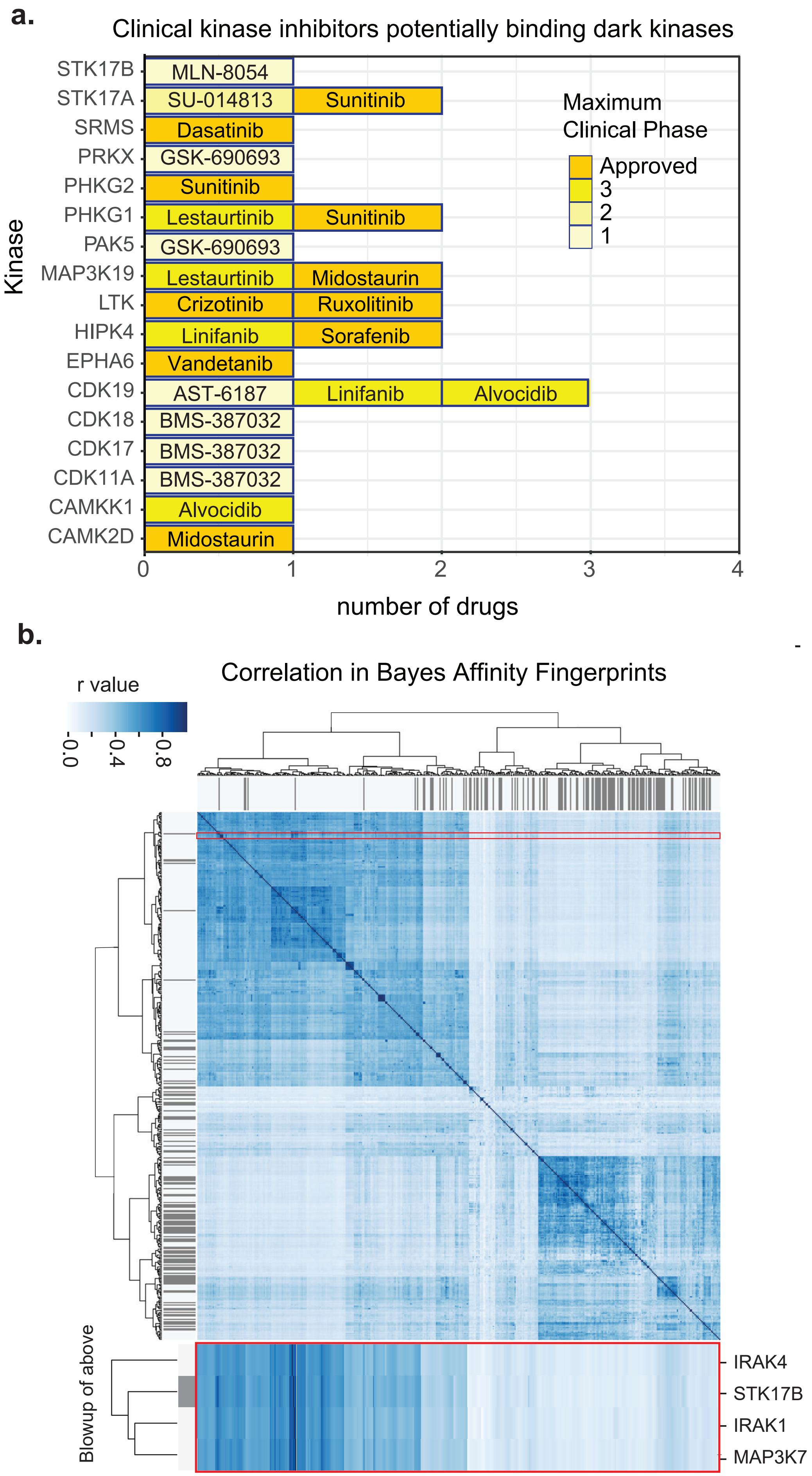
BRSK1
- Upregulated in 7 tumor types
- High level copy number amplification in DLBC
- High level copy number deletion in LGG

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.
Clinical kinase inhibitors potentially binding dark kinases
Kinase
STK17B
MLN-8054
STK17A
SU-014813
Sunitinib
SRMS
Dasatinib
PRKX
GSK-690693
PHKG2
Sunitinib
PHKG1
Lestaurtinib
Sunitinib
PAK5
GSK-690693
MAP3K19
Lestaurtinib
Midostaurin
LTK
Crizotinib
Ruxolitinib
HIPK4
Linifanib
Sorafenib
EPHA6
Vandetanib
CDK19
AST-6187
Linifanib
Alvocidib
CDK18
BMS-387032
CDK17
BMS-387032
CDK11A
BMS-387032
CAMKK1
Alvocidib
CAMK2D
Midostaurin
0
1
2
3
4
number of drugs
Maximum
Clinical Phase
Approved
3
2
1
b.
Correlation in Bayes Affinity Fingerprints
r value
0.0
0.4
0.8
Blowup of above
IRAK4
STK17B
IRAK1
MAP3K7