Redefining the specificity of phosphoinositide-binding by human PH domain-containing proteins

Nilmani Singh[1†], Adriana Reyes-Ordoñez[1†], Michael A. Compagnone[1], Jesus F. Moreno Castillo[1], Benjamin J. Leslie[2], Taekjip Ha[2,3,4,5], Jie Chen[1*]

[1]Department of Cell & Developmental Biology, University of Illinois at Urbana-Champaign, Urbana, IL 61801; [2]Department of Biophysics and Biophysical Chemistry, Johns Hopkins University School of Medicine, Baltimore, MD 21205; [3]Department of Biophysics, Johns Hopkins University, Baltimore, MD 21218; [4]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD 21205; [5]Howard Hughes Medical Institute, Baltimore, MD 21205, USA

[†]These authors contributed equally to this work.

[*]Correspondence: jiechen@illinois.edu.

ABSTRACT

Pleckstrin homology (PH) domains are presumed to bind phosphoinositides (PIPs), but specific interaction with and regulation by PIPs for most PH domain-containing proteins are unclear. Here we employed a single-molecule pulldown assay to study interactions of lipid vesicles with full-length proteins in mammalian whole cell lysates. We examined 67 human PH domain-containing proteins for interaction with PIPs, and found 51% of them to have affinity for PIPs with various specificity, the majority of which had not been reported before. Further investigation of ARHGEF3 revealed structural requirements and functional relevance of its newly discovered PI(4,5)P_2 binding. Based on the assay data of the 67 proteins, we generated a recursive-learning algorithm, which identified amino acid determinants of PIP binding throughout the entire PH domain. This recursive functional classification predicted that 46% of 250 human PH domains bind PIPs with some specificity. Collectively, our findings reveal unexpected lipid-binding specificity of PH domain-containing proteins.

INTRODUCTION

Specific phospholipid-protein interactions are critical to the regulation of many signal transduction processes and cellular functions. These interactions typically involve lipid-binding domains recognized by specific lipid species and/or physical properties of the membrane such as charge or curvature (1-4). The largest family of putative lipid-binding domains (LBDs) is the pleckstrin homology (PH) domain, with over 250 members encoded by the human genome (Pfam database). Originally defined by its presence in the protein Pleckstrin (5,6), this domain of 100-120 amino acids has an invariable structure of seven-stranded β-sandwich lined by a C-terminal α-helix (7). PH domains are found in many different types of proteins that are involved in regulating diverse signaling pathways and functions. Some of the earliest characterized PH domains, including those in Pleckstrin, RasGAP, and GRK2, were found to have affinity for PI(4,5)P_2 (8). The PH domains in PLCδ and AKT1 bind with high selectivity to PI(4,5)P_2 and PI(3,4,5)P_3 (PIP_3), respectively (9,10). Indeed, the interaction with PIP is so specific for those two PH domains that they have been commonly used as sensors to detect PI(4,5)P_2 and PIP_3 in cells (11,12).

Despite those early examples of PH-PIP interactions, to date only a modest number of PH domain-containing proteins have been demonstrated to bind PIPs with specificity. A comprehensive analysis of *S. cerevisiae* PH domains revealed that only one of them had specific affinity for a particular PIP and the rest of them displayed little affinity or selectivity for PIPs (13). Another study examined a large number of mouse PH domains and found 20% of them to translocate to the plasma membrane (PM) in response to PIP_3 production in cells, but most of those PIP_3-responsive PH domains did not show specific binding to PIP_3 in lipid binding assays (14). Thus, it has been proposed that the specific PIP recognition by PLCδ-PH and AKT1-PH

may be the exception rather than the rule for PH domains and that only a small percentage of all PH domains bind PIPs with high affinity and specificity (4). Protein partners have been identified for PH domains, which may cooperate with lipid-PH interaction or operate independently to regulate or mediate PH domain functions (1,15).

It is important to note that our knowledge to date of the lipid binding properties of PH domain-containing proteins is largely derived from studies of isolated PH domains rather than full-length proteins, at least partly owing to the hurdle of purifying full-length proteins for lipid binding assays. Intra-molecular interaction is a universal mechanism for determining protein structure and activity, and this crucial determinant would be eliminated when a PH domain is taken out of the context of the protein. To circumvent the laborious process of purifying proteins and to preserve post-translational modifications of the proteins, we have developed a total internal reflection fluorescence (TIRF) microscopy-based single-molecule pulldown (SiMPull) assay to study protein-lipid interactions with whole cell lysates and lipid vesicles (16), herein referred to as lipid-SiMPull or SiMPull. The sensitivity and specificity of this assay have been demonstrated through highly selective pulldown of various LBDs, as well as full-length AKT1, from cell lysates by vesicles containing PIPs known to interact with those proteins (16). We have now applied the lipid-SiMPull assay to interrogate 67 human PH domain proteins for their binding to vesicles of various compositions. Our results suggest that PIP recognition by PH domain proteins is more prevalent than previously believed. We have also used the assay data to generate a recursive-learning algorithm, which reveals PH domain sequence determinants for PIP binding and predicts PIP binding for the entire family of human PH domains.

RESULTS

Lipid-SiMPull assay for human PH domain proteins

In order to investigate binding of PIPs by PH domain proteins using the lipid-SiMPull assay (16), we created cDNA constructs and transfected them in human embryonic kidney (HEK) 293 cells to express 67 full-length human PH domain proteins with EGFP fused at their C-termini. Those 67 proteins were randomly collected based on availability of intact cDNA clones from the hORFeome V5.1 library. Each cDNA in the final plasmid was sequence-validated in its entirety, and expression of the full-length fusion protein in HEK293 cells was confirmed by western analysis (see examples in Supplementary Figure 1a). Concentrations of EGFP-fusion proteins in the lysates were determined by fluorescence intensity and calculated using a standard emission (ex488/em520) curve generated with purified recombinant EGFP (Supplementary Figure 1b). Small unilamellar vesicles (SUVs) were made with 60 mol % phosphatidylcholine (PC), 15 mol % phosphatidylserine (PS), 20 mol % cholesterol, and 5 mol % of one of six PIPs: PI(3)P, PI(4)P, PI(5)P, PI(3,4)P_2, PI(4,5)P_2, and PI(3,4,5)P_3. The only PIP omitted in our assays was PI(3,5)P_2 due to scarce evidence in the literature for its specific recognition by proteins (17) and the lack of a reliable positive control. Vesicles containing 5% phosphatidic acid (PA), a different type of signaling lipid than PIPs, were also included in our assays. As a negative control, vesicles were made to include an additional 5 mol % PC in place of PIP (referred to as PC vesicles or control hereafter). All vesicles also had 0.01 mol % biotinylated phosphatidylethanol (PE) to facilitate immobilization on imaging slides.

SiMPull assays were performed with the SUVs immobilized on slides and lysates containing 5 nM EGFP-fusion protein flowed into the slide chamber followed by TIRF imaging (Figure 1a). Each of the 67 proteins was assayed against each of the eight types of vesicles. The

SUVs also contained a lipophilic red dye, DiD, to allow visualization of the immobilized vesicles on slides and confirmation of consistent vesicle density across assays. Pulldown of EGFP proteins was quantified by counting green fluorescent spots. Each preparation of SUV was validated by binding to a positive control in SiMPull, and only the experiments with validated SUVs were further processed. Each vesicle-protein pair was assayed using validated vesicles with at least three independent lysates. Based on characterization of many positive and negative controls with the SiMPull assay, we arrived at 100 EGFP spots per area of 1600 μm^2 (after background subtraction) as the threshold for binding and the data were interpreted as a binary outcome – binding or no binding. As shown in Supplementary Table 1, the assay results from three independent experiments were remarkably consistent for the vast majority of the 67 proteins.

New PH domain protein-PIP interactions are revealed

The assay results for all proteins are summarized in Table 1, and the raw data can be found in Supplementary Table 1. Of the 67 proteins examined, 4 proteins displayed promiscuous binding to lipids: CERT1, NET1 and SPATA13 bound all types of SUVs including PC, whereas PLEKHO1 bound most PIPs as well as PA. Thirty proteins were found to bind PIPs with various selectivity, and none of them bound PA. Among the 33 proteins that did not bind lipids in our initial assays, 16 had been previously reported to bind PIPs as either PH domain alone or full-length protein (Supplemental Table 2). We considered the possibility that some of those proteins might require more than 5% PIP in the vesicle for effective binding, since PIPs in the cell membrane can cluster and their asymmetric distribution may lead to drastically higher local concentrations (18). Thus, we performed SiMPull assays with vesicles containing 20% PIP_3 or

$PI(4,5)P_2$ for 11 proteins that had been reported to bind either PIP but not found to bind in our assays (AFAP1L1, ARHGAP12, DOK1, FERMT2, GRB14, GRK2, OSBPL8, PLCD4, PLEK, SBF2, SKAP2). Four of the proteins (ARHGAP12, PLCD4, SBF2, SKAP2) indeed bound 20% PIP_3 or $PI(4,5)P_2$. Hence, in total 34 of the 67 proteins (51%) were found to bind PIPs with some specificity (Table 1, Figure 1b). The positive binding data are also illustrated in a binary table in Figure 1c. Representative SiMPull assay images and quantification are shown in Figure 1d for 6 proteins, each displaying specificity for a different PIP, as well as for AKT1 as a positive control.

Of the 34 proteins found to bind PIPs, our assay results confirmed reported PIP binding by 8 proteins (Group I in Figure 1b & 1c), whereas 26 proteins (Group II) had been either not previously reported to bind any lipid or reported to display different lipid binding profiles than those found in our assays. Of the 29 proteins found not to bind PIP in our assays, 12 had been reported to bind lipids (Group III) and 17 not reported (Group IV). It is possible that those 12 represent false negatives in our assay. Previously we showed that the assay detected LBD-lipid interactions that were reported to have K_d's up to high nanomolar/sub-micromolar (16). To estimate the threshold of detection under our assay conditions, we performed SiMPull with the PX domain of p40Phox, the WT and R60A mutant of which bind PI(3)P with K_d of 5 μM and 17.5 μM, respectively (19). As shown in Supplementary Figure 2, specific binding to PI(3)P was detected for both domains. While binding to the WT p40Phox-PX was robust, binding to the R60A mutant was just above the cut-off of 100 EGFP spots. Therefore, our assay under current conditions has a detection threshold of affinity in the 10-20 μM range. Overall, the specificity of PIP recognition by PH domain proteins revealed by our assay was markedly distinct from that in the current literature.

Phosphatidylserine is not responsible for the novel PH domain protein-PIP interactions

Phospholipids may cooperate in the membrane to determine specific interaction with proteins (2,3). The presence of sphingolipids and PS in the membrane has been found to contribute to enhanced affinity and/or specificity for PIPs by yeast PH domains (20). Since the SUVs in our assays contained 15 mol% PS (based on mammalian cell membrane composition), we wondered whether the newly discovered PIP-PH protein interactions were dependent on PS. To address this possibility, we re-assayed 23 of the new PIP-binding proteins (Group II) with SUVs containing only PC and a specific PIP, omitting PS and cholesterol. Remarkably, all 23 proteins bound their respective PIPs without PS in the vesicles (Figure 2a). Representative SiMPull images are presented in Figure 2b for the same proteins shown in Figure 1d. Therefore, the presence of PS does not appear to account for the novel PIP-PH protein interactions emerged from our studies.

Membrane targeting of PIP_3 binding proteins

One potential consequence of PIP binding is targeting of the protein to specific membranes in the cell. PIP_3 is acutely produced by mitogenic stimulation through activation of PI-3-kinase (PI3K) in the PM, resulting in recruitment of proteins to the PM through their binding to PIP_3. AKT1 is an example of such a protein. As a means to validate PIP binding in the cell, we set out to examine PM translocation of all the proteins found to bind PIP_3 but not $PI(4,5)P_2$ in our SiMPull assays (a total of 8). We excluded those that bind both PIP_3 and $PI(4,5)P_2$ because $PI(4,5)P_2$ is present at high concentrations in the PM and could potentially mask any effect of PIP_3. HEK293 cells transiently expressing EGFP fusion proteins were serum starved and then stimulated with insulin for 5 min, with or without treatment by wortmannin, a

PI3K inhibitor. As shown in Figure 3, AKT1-PH as a positive control was found to translocate to PM in response to insulin stimulation and this translocation was abolished by wortmannin treatment. BTK, ITK, FERMT3, SKAP1 and GAB2, as well as PLCD4 which bound 20% PIP_3 in SiMPull assays, exhibited insulin-stimulated and wortmannin-inhibited PM translocation. Another protein that bound 20% PIP_3, SKAP2, was not found at the PM. ARHGEF39, shown to bind PIP_3 and $PI(3,4)P_2$ in our assays, displayed constitutive PM localization, but its appearance in the cytoplasm was also induced by wortmannin treatment. It is possible that ARHGEF39 is targeted to the PM via both PIP_3-dependent and -independent mechanisms, with the latter involving lipid modification or interaction of ARHGEF39 with a membrane protein. Taken together, the observed membrane recruitment is consistent with the protein-PIP_3 interactions we have discovered with the SiMPull assays.

Isolated PH domains do not have the same lipid-binding properties as their full-length counterparts

To compare the behaviors of PH domains to their full-length protein counterparts, we subcloned 14 PH domains from 11 proteins (3 of them have two PH domains) to express as EGFP-fusions in HEK293 cells and performed SiMPull assays with the lysates. With the exception of AFAP1L1, which had been reported to bind PIP_3 but did not bind in SiMPull assays, those full-length proteins all displayed PIP binding in our assays. All the PH domains were first assayed against the negative control vesicles and PA vesicles, and no binding was observed for any. We then proceeded to assay the PH domains with selected PIP vesicles. The results are summarized in Table 2. Surprisingly, only one PH domain displayed the same PIP binding as its full-length counterpart: PH2 of PLEKHA1 for $PI(3,4)P_2$. The PH domains from the

remaining 10 proteins did not behave like their full-length counterparts: 9 PH domains did not bind any lipid, and 3 PH domains bound lipids differently than the full-length proteins.

One might argue that LBDs other than PH could have been responsible for our observed PIP binding by full-length proteins. However, only 2 of those 10 proteins above contain other known LBDs: BAR domain in ACAP1 and FERM domain in FERMT3. A BAR-PH fragment of ACAP1 bound lipid vesicles non-specifically in our assays (data not shown). The FERM domain was not examined in our study but it had been reported to *not* contribute to PIP binding by FERMT3 (21). Although we cannot rule out the possibility of EGFP interfering with PH domain activity in the fusion proteins, our observations are consistent with the notion that a PH domain without the context of its full-length protein may not have the same lipid-binding properties. It is also important to note that given the diverse domain structures of the 34 proteins found to bind PIPs, with PH domain being the only common domain among them, and the scarcity of other lipid-binding domains in those proteins, the PH domains most likely mediate the PIP binding by those full-length proteins.

PH domain of ARHGEF3 is not sufficient but is necessary for specific lipid binding

Next, we tested the possibility of PH domain mediating PIP binding with ARHGEF3 (also named XPLN for exchange factor found in platelets, leukemic, and neuronal tissues), a GEF for RhoA and RhoB (22). ARHGEF3 (Figure 4a) belongs to the Dbl family of 70 human RhoGEFs, the vast majority of which contain a Dbl homology (DH) domain and a PH domain in tandem (23). There had been no report of specific lipid binding by ARHGEF3. In our assay, the full-length ARHGEF3 protein was pulled down by PI(4,5)P_2 but not by any of the other vesicles tested (Figure 4b). This binding was also observed with recombinant EGFP-ARHGEF3 purified

from bacteria (Supplementary Figure 3), suggesting that the ARHGEF3-PI(4,5)P_2 interaction was direct and independent of any post-translational modification. As shown in Figure 4c, ARHGEF3-PH fused with EGFP at either the N- or C-terminus did not bind PI(4,5)P_2 vesicles, neither did fragments of ARHGEF3 missing the N-terminal 125 amino acids (ΔN) or the PH domain (ΔPH). Expression of the various ARHGEF3 proteins was confirmed by western blotting shown in Supplementary Figure 4a. Hence, it appears that the entire full-length ARHGEF3 may be necessary to confer lipid binding.

As another approach to assess whether the PH domain of ARHGEF3 was required for interaction with PI(4,5)P_2, we considered structural information available in the literature. In reported PH domain-PIP headgroup interactions, negatively charged amino acids in the β1-β2 and β3-β4 loops play critical roles (1). Applying the Patch Finder Plus software (24) to the reported ARHGEF3-PH crystal structure (25), we identified a positively charged patch containing K342, R345, K348, R376 and H427 to be potentially involved in the interaction between ARHGEF3 and PI(4,5)P_2 (Figure 4d). These amino acids were mutated to alanine or acidic residues in full-length ARHGEF3 individually and in combination, and the mutant proteins (Supplementary Figure 4a&b) were subjected to SiMPull assays with PI(4,5)P_2 vesicles. Mutation of any one of the positively charged residues did not result in drastic reduction of PI(4,5)P_2 binding, but three double-mutants (K342E/K348E, R345D/R376D, R345D/H427D) and a triple-mutant (R345A/R376A/H427D) each abolished lipid binding by ARHGEF3 (Figure 4e). Taken together, our observations suggest that the PH domain of ARHGEF3 interacts with PI(4,5)P_2 directly but inter-domain interactions in the full-length protein may also be important for the protein-lipid interaction.

PI(4,5)P_2 binding is necessary for ARHGEF3 membrane targeting, activation of RhoA, and induction of stress fiber formation

Because PI(4,5)P_2 is highly abundant in the PM, we wondered if ARHGEF3 binding to PI(4,5)P_2 played a role in the protein's membrane recruitment. Overexpressed EGFP-ARHGEF3 was found throughout the cell including the nucleus, and the nuclear localization is consistent with the presence of a nuclear localization sequence in ARHGEF3. We observed prominent presence of EGFP-ARHGEF3 in the PM of NIH3T3 cells and, importantly, the lipid binding-deficient mutant of ARHGEF3, K342E/K348E, displayed drastically reduced localization to the PM (Figure 5a). This observation is consistent with ARHGEF3 targeting to the PM via its interaction with PI(4,5)P_2.

Next, we asked whether lipid binding is relevant to ARHGEF3's function in a cell. ARHGEF3 has two unrelated activities: as a GEF for RhoA/B (22) and as a GEF-independent inhibitor of mTORC2 phosphorylation of AKT (26). We examined these activities of ARHGEF3 in HEK293 cells. As shown in Figure 5b, overexpression of wild-type ARHGEF3 suppressed AKT phosphorylation as expected, and overexpression of each lipid binding-deficient mutant of ARHGEF3 (K342E/K348E, R345D/R376D, R345D/H427D, or R345A/R376A/H427D) was equally effective in inhibiting AKT, suggesting that lipid binding is not involved in ARHGEF3's function as an inhibitor of mTORC2. This is consistent with the previous observation that the N-terminal 125-amino acid fragment, devoid of the PH domain, is sufficient to inhibit mTORC2 (26). On the other hand, unlike the wild-type protein, the K342E/K348E mutant ARHGEF3 was ineffective in activating cellular RhoA activity upon overexpression, as measured by pulldown of active RhoA using an effector protein (Figure 5c). Hence, lipid binding is apparently necessary for ARHGEF3's GEF activity toward RhoA in the cell.

We also performed in vitro GEF assays with ARHGEF3 and ARHGEF3-K342E/K348E proteins purified from bacteria, and found that the two proteins displayed near identical activities towards RhoA (Figure 5d). We were not able to assess the effect of lipids on the GEF activity because we found addition of lipid vesicles to be incompatible with the GEF assay. Nevertheless, our observation confirmed that the K342E/K348E mutant retained proper folding and enzymatic activity. The simplest explanation for our combined results is that lipid binding and/or membrane targeting is necessary to achieve maximum GEF activity for ARHGEF3 in the cell.

Consistent with its role as a GEF for RhoA, ARHGEF3 has been reported to regulate actin cytoskeleton reorganization and assemble stress fibers in the cell (22). To probe a potential role of ARHGEF3 lipid binding in this process, we overexpressed EGFP-ARHGEF3 in NIH3T3 cells and visualized actin cytoskeleton with phalloidin staining. As shown in Figure 5e, overexpression of wild-type ARHGEF3, but not the K342E/K348E mutant, led to robust stress fiber formation, suggesting that PI(4,5)P_2 binding may be critical for ARHGEF3 regulation of actin cytoskeleton reassembly, a RhoA-dependent process. Taken together, our observations uncover a novel mechanism of ARHGEF3 regulation and provide functional validation of a protein-lipid interaction identified in our lipid-SiMPull assays.

A recursive-learning algorithm for probabilistic sequence comparison based on SiMPull data

Although our survey of 67 proteins covered only 1/4 of the PH domain-containing protein family in the human genome, we reasoned that our results could have predictive power for the rest of the family. Given the high likelihood that the PH domains mediate PIP binding of the full-length proteins, we compared the sequences of the PH domains of the 67 proteins we

studied plus AKT1-PH as a postive control, excluding the 4 proteins that bound lipids non-specifically (a total of 71 PH domains). Based on reported structural studies of the PH domain-PIP headgroup interaction, the loop between the first two β strands (β1-β2) with a basic sequence motif of "$KX_n(K/R)XR$" plays the most prominent role (3). However, we found the vast majority of the 71 PH domains to contain this motif, with no distinction between those that bound PIPs and those that did not (Figure 6a). Therefore, we next considered the possibility that sequence features throughout the entire PH domain are determinants of PIP binding, as suggested by Park et al. for PIP_3-regulated PH domains (14).

To identify sequence signatures of PIP binding, we employed an unbiased probabilistic sequence comparison strategy developed by Park et al., called recursive functional classification (RFC) (14). This strategy relies on accurate sequence alignement of multiple PH domains, which can be a challenge given the low sequence similarity shared by PH domains despite the 3D structure conservation. We employed the PROMALS3D (PROfile Multiple Alignment with predicted Local Structures and 3D constraints) tool, which makes use of available 3D structural information to guide multiple sequence alignments (27), to align the sequences of 250 PH domains from 222 human proteins in the Pfam and UniProt databases, including the 71 PH domains described above. We found a total of ~280 PH domains in the database, but excluded ~30 of them from the alignment due to their unusual sequence features that would have caused very large gaps in the alignment. Alignment of the 250 PH domains yielded 297 amino acid positions; the 71 PH domains from proteins investigated in SiMPull and the other 179 PH domains are listed in Supplementary Table 3 and Supplementary Table 4, respectively. Following the method described by Park et al. (14), we created two probability matrices of 297 (positions) x 20 (amino acids): P^B for the PIP binding group (36 PH domains including AKT1-

PH), and P^{NB} for the non-binding group (35 PH domains). Based on the assumption that each sequence position contributes independently to PIP binding, we created an RFC-matrix, with each element of the matrix calculated as $\log(P^{B}/P^{NB})$. In the RFC-matrix values > 0 represent positive correlation with PIP binding whereas values < 0 represent negative correlation. This RFC-matrix can be used to calculate the predicted PIP binding score (S_{RFC}) for any PH domain, as previously described (14).

Recursive functional classification identifies amino acid determinants of PIP binding throughout the entire PH domain

Using the RFC-matrix created above, we first scored positions in the β1-β2 loop and its flanking regions to calculate S_{RFC}. The resulting scores separated PIP-binding proteins from non-binding proteins without a large margin (Figure 6b). When we expanded the positions to include β3-β4 and β6-β7 loops and their flanking regions, which are believed to also contribute to PH-PIP interactions (1), the separation of the two groups of proteins improved (Figure 6b). However, when all positions of the PH domain were taken into calculation, the PIP-binding and non-binding groups were even more cleanly separated with markedly higher scores in both the positive and negative ranges (Figure 6b). This outcome validates the idea that the entire PH domain may be involved in determining whether a domain (or full-length protein) binds PIP, similar to what has been reported for PIP_3 regulation (14) and organelle PIP binding (20) by PH domains. Interestingly, the relative contributions of various regions of the PH domain to PIP-binding vary drastically among those found to bind PIPs (Figure 6c), suggesting that the structural basis of PH-PIP interactions may be diverse within this family of proteins.

We created a heat map (Figure 7a) to represent the RFC-matrix for the 71 PH domains studied in SiMPull and aligned it to the crystal structure of AKT1-PH (PDB ID: 1UNP). Amino acids with high impact on lipid binding, either positively or negatively, were distributed throughout the PH domain. Following the method developed by Park et al. (14), we determined the overall contribution of each amino acid position to PIP binding (Figure 7b). The 12 positions with the highest contribution to PIP binding are marked on a representative PH domain structure (Figure 7c). While some of those residues are located in the classically defined head group binding pocket, many are not. For instance, we found significant contribution from residues in β6 and β7 sheets, as well as in the alpha helix. Other than those 12, additional positions spanning the entire PH domain are likely to also play important roles in PIP binding (Figure 7b). It should be noted that this is a composite view, and that an individual PH domain would use only a subset of those residues for interaction with PIP.

The RFC algorithm predicts 46% of human PH domain-containing proteins to bind PIP

Next, we applied the RFC algorithm obtained above to analyze the human PH domains not investigated by SiMPull. To establish a cut-off score for positive correlation with PIP binding, we scrambled the sequences of all 250 PH domains and subjected them to scoring by the same algorithm. As shown in Supplementary Figure 5a, the highest S_{RFC} score for the scrambled sequences was 3.6. The mean S_{RFC} for all 250 scrambled PH domains was -4.04 with a standard deviation (SD) of 2.66 (Supplementary Figure 5b). Using mean + 3x SD, we arrived at 3.94 as the threshold S_{RFC} score for a PH domain to be considered likely to bind PIP. Of the 179 PH domains not tested by SiMPull, 80 scored above the threshold (Figure 7d and Supplementary Table 4). This predicts 46% of the PH domains in the human genome to be PIP-

binding proteins, which is in line with our assay results of the randomly selected 67 proteins (51% bound PIP). It is possible that many proteins with positive S_{RFC} values lower than 3.94 may also bind PIPs. Experiments are ongoing to validate the predictions. Additional data will also further improve the predictive power of the RFC algorithm.

DISCUSSION

Our analysis of a random collection of 67 human PH domain-containing proteins using the lipid-SiMPull assay reveals many specific interactions with PIPs that have not been reported before. One half of all the proteins examined in our assays display PIP binding with various specificity, in contrast to the estimate that only 10% of all PH domains bind PIPs with high affinity and specificity (4). As a proof of principle, we have validated the functional relevance of the novel ARHGEF3-PI(4,5)P_2 interaction. Based on the assay results of the 67 proteins, we have created a recursive-learning algorithm and found that amino acid determinants of PIP binding are not limited to the well-established binding pocket but rather are distributed throughout the entire PH domain. Our recursive functional classification predicts that 46% of all the human PH domains likely bind PIPs with some specificity.

Isolated PH domains often did not recapitulate the lipid binding behaviors of their protein counterparts in our assays, highlighting the need to investigate full-length proteins. The SiMPull assay offers the unique advantage of using full-length proteins expressed in the native environment (i.e., mammalian cells) without the need for purification. This environment would likely confer proper post-translational modifications to the expressed proteins. However, due to overexpression of the EGFP-fusion proteins it is unlikely that any endogenous proteins in the lysate would reach stoichiometric levels to contribute directly to the observed recombinant protein binding to lipid vesicles. Hence, the lipid-protein interactions observed in lipid-SiMPull assays are most likely intrinsic properties of the PH domain-containing proteins.

We have determined the threshold of detection under the current lipid-SiMPull assay conditions to be in the K_d range of 10-20 μM, which is comparable or superior to other methods and sufficiently sensitivity to detect physiologically relevant interactions. Although the

possibility of interference by the EGFP tag has not been ruled out for those proteins that did not display any lipid binding in our assays, we do not consider it likely that many true interactions have been missed by our assay. Instead, we would like to suggest that non-physiological binding conditions (such as in the lipid strips assay) and/or isolation of the PH domains out of the context of full-length proteins may have been accountable for at least some of the discrepancies between our observations and those in the literature. Indeed, of those 12 proteins that were reported to bind PIPs but found not to bind in our assays (Group III), 11 were based on isolated PH domains and/or lipid strips assays (references in Supplementary Table 2).

The PIP binding specificity discovered in our study can potentially lead to new understanding of the functions and mechanisms of the PH domain proteins. For instance, previously the PH domain of CNKSR1 (or CNK1) had been found to bind PIPs weakly and non-specifically (28), but in our assays the full-length CNKSR1 protein bound specifically to PI(4)P. Interestingly, CNKSR1 has been reported to promote PI(4,5)P_2 production from PI(4)P by phosphatidylinositol 4-phosphate 5-kinases (PIP5Ks) at the PM where insulin receptor signaling takes place (29), although the mechanism is unclear. Based on our finding, it is conceivable that CNKSR1 is targeted to PI(4)P-enriched membrane in order to activate PIP5K or to recruit PIP5K to its substrate. Another example is EXOC8 (or Exo84), a component of the exocyst, which is an octameric protein complex involved in tethering secretory vesicles to the plasma membrane for fusion during exocytosis (30). A generally accepted view is that Sec3 and Exo70 in the exocyst bind PI(4,5)P_2 and target the entire octameric complex to the membrane. Our observation of PI(4,5)P_2 binding by EXOC8 suggests that EXOC8 may also contribute directly to the regulation of the exocyst by PI(4,5)P_2, and that further investigation could potentially lead to revision of the current model of exocyst assembly.

The vast majority of the 70 members of the Dbl family of RhoGEFs contain a PH domain immediately following the catalytic DH domain (23). Although those PH domains had been speculated to bind phospholipids and subsequently contribute to membrane localization and/or allosteric regulation of the RhoGEFs, very few of them have been reported to have marked affinity for specific PIPs. Thus, the idea of phospholipids regulating RhoGEFs through their PH domains remains controversial (31). Among the proteins we studied with the SiMPull assay, 14 are RhoGEFs belonging to the Dbl family. PIP binding was observed for 8 of them (ARHGEF3, ARHGEF5, ARHGEF9, ARHGEF16, ARHGEF39, MCF2L, NGEF, VAV1), all of which were either never reported to bind lipids or reported to bind lipids with different specificity than what we observed. For instance, the PH domain of ARHGEF3 was reported to bind phospholipids with little selectivity in lipid strips assays (14). In contrast, we have found that full-length ARHGEF3 exclusively binds PI(4,5)P_2 and that the binding is dependent on the PH domain. Furthermore, using a lipid binding-deficient mutant of ARHGEF3 we have established the importance of PI(4,5)P_2 binding for the subcellular localization, activity and function of ARHGEF3. Follow-up studies on the 7 remaining RhoGEFs and investigation of lipid binding by other Dbl family members will likely be illuminating.

The recursive-learning algorithm developed by Park et al. (14) is an unbiased probabilistic sequence comparison strategy that is now facilitated by accurate multiple sequence alignment incorporating 3D structural information (27). The clean separation of PIP binding and nonbinding groups of PH domains by this algorithm attests to the power of the method. The prediction that nearly one half of all human PH domains are likely to bind PIP with some specificity is surprising given the current thinking in the field. Experimental testing of our predictions with full-length yeast proteins is warranted. As with any recursive learning strategies,

the accuracy of prediction relies on the quality and quantity of the input data. We believe that the lipid-SiMPull assay yields reliable binding data for recursive learning, and that the algorithm can be further improved with more assay data in the future. With the availability of more data, it should also become possible to predict binding to a specific PIP.

MATERIALS AND METHODS

Cell Culture

HEK293 cells were maintained in high-glucose DMEM with 10% FBS, 2 mM L-glutamine, and penicillin/streptomycin at 37 ^{o}C in 5% CO_2. NIH3T3 cells were maintained in high-glucose DMEM with 10% newborn calf serum and 4 mM L-glutamine, and penicillin/streptomycin at 37 ^{o}C in 5% CO_2. For SiMPull experiments, HEK293 cells were transfected for 24 hours in 6-well plates using PolyFect® (3 µL/µg DNA) following manufacturer's recommendations. For immunofluorescence experiments, cells were plated in 12-well plate on poly-L-lysine coated coverslips and transfected with Lipofectamine™ 3000 following manufacturer's recommendations. For active RhoA pulldown assays, HEK293 cells were transfected in 10-cm plates with Lipofectamine™ 3000.

Cell Lysis and Western Blotting

For SiMPull assays, the cells were collected after 24-hr transfection in detergent-free buffer (40 mM HEPES, pH8.0, 150 mM NaCl, 10 mM ß-glycerophosphate, 10 mM sodium pyrophostate, 2 mM EDTA, 1x Sigma protease inhibitor cocktail). The cells were lysed by sonication for 3 seconds on ice followed by ultracentrifugation at 90,000xg in a TLA100.3 rotor for 1 hour at 4°C. For SiMPull, EGFP concentration was measured using a standard emission (ex488/em520) curve of pure recombinant EGFP, and each cell lysate was diluted in vesicle buffer (10 mM Tris·HCl, pH 8.0, 150 mM NaCl) to yield 5 nM EGFP. For western blotting, cells were lysed in 1x SDS sample buffer or as described above and mixed at 1:1 with 2x SDS sample buffer, both containing β-mercaptoethanol at a final concentration of 5%, and heated for 5 min at 95 ^{o}C. Proteins were resolved by SDS-PAGE and transferred onto PVDF membrane. The membrane was incubated with primary and secondary antibodies following manufacturers'

recommendations. HRP-conjugated secondary antibody was reacted with West Pico PLUS Chemiluminescent Substrate, and the signal was detected using an Invitrogen iBright digital imager.

cDNA Cloning and Mutagenesis

Plasmids harboring cDNAs for human PH domain-containing proteins from human cDNA library hORFeome V5.1 were cloned as a pool into the pDest-eEGFP-N1 vector using Gateway Cloning, followed by identification of individual clones. Some of the cDNAs, which were not successfully cloned by Gateway cloning, were cloned individually into the pcDNA3-EGFP vector using Gibson assembly. The PH domains were cloned into the pEGFP-C1 vector using Gibson assembly. ARHGEF3 (XPLN) expression plasmids for full-length and truncations were previously reported (26). p40PhoxPX-EGFP was obtained from Addgene (#19010) (32). All point mutants were created using site-directed mutagenesis using QuikChange Lightning Site-Directed Mutagenesis Kit following the manufacturer's protocol. All cDNAs in the final plasmids were sequence-confirmed in their entirety.

Lipid Vesicle Preparation

All lipids were mixed in chloroform (0.166 μmol total) and dried under nitrogen flow. The dried mixture was re-suspended in 100 μL vesicle buffer (10 mM Tris·HCl, pH 8.0, 150 mM NaCl) to a final concentration of 1.66 mM. After 30 min incubation at room temperature, vesicles were formed by water bath sonication (Laboratory Supplies, Hicksville, NY, model G112SPIT, 600 v, 80 kc, and 0.5 A) in 4 cycles of 4 min each. Small unilamellar vesicles were collected as the supernatant after ultracentrifugation at 194,398xg in a TLA100.3 rotor for 1 hr at 25 ^{o}C.

Single-Molecule Pulldown Assay

Quartz slides were prepared as described in Jain et al. and Arauz et al. (16,33). Briefly, the slides were thoroughly cleaned and passivated with PEG doped with 0.1-0.2% biotin-PEG. Neutravidin (200 μg/mL) was incubated in chambers for 10 min, followed by addition of biotinylated lipid-vesicles. Freshly prepared whole-cell lysate (80 μL) was added by flowing into each slide chamber, replacing the vesicle solution. An inverted total internal reflection fluorescence (TIRF) microscope with Olympus 100x, NA1.4 lens and EMCCD camera (Andor iXon Ultra 897) was used to acquire single-molecule data at 10 frames/second. Diode pumped solid state lasers were used to excite EGFP at 488nm (Coherent) and DiD at 638nm (Cobalt). All SiMPull experiments were performed at room temperature. Average number of spots per imaging area (~1600 μm^2) was calculated from 10 or more images.

SiMPull Image Analysis

TIRF image acquisitions were processed in IDL to generate image files, and the number of fluorescence spots were identified with point-spread function. MATLAB® was used to extract data from data files generated in IDL and import into Excel. The background number of EGFP spots was gathered prior to addition of lysate and subtracted from EGFP spots after lysate addition to generate data for each image. At least 10 SiMPull images (1600 μm^2 each) were analyzed to generate the average number of spots per imaging area for each assay. The IDL scripts used to process raw image files are publicly available from Taekjip Ha laboratory (http://ha.med.jhmi.edu/resources/).

Fluorescence Imaging

HEK293 or NIH3T3 cells on poly-L-lysine-coated glass coverslips were fixed in 3.7% paraformaldehyde and permeabilized with 0.1% Triton X-100. For visualizing stress fibers,

Rhodamine phalloidin was diluted 1:3000 and incubated with 3% BSA/PBS at 4 ^{o}C for 60 min, followed by incubation with DAPI (1:2500) for 20 min at room temperature. For digitonin permeabilization experiments, cells on coverslips were treated with 0.03% digitonin (20 mM HEPES, PH 7.5, 110 mM KOAc, 5 mM NaOAc, 2 mM $Mg(OAc)_2$, 1 mM EGTA) for 2.5 min, followed by fixation with 3.7% paraformaldehyde. A personal deconvolution microscope system (DeltaVision, Applied Precision) was used with a 100× NA 1.4 lens to capture fluorescence images. For PM localization experiments, HEK293 cells were transfected and treated as described in Figure 3 legend, fixed in 3.7% paraformaldehyde, and stained with DAPI for 20 min. Fluorescence imaging was performed as described above with a 60x NA 1.4 lens.

Protein Purification

GST-ARHGEF3 wild-type and K342E/K348E proteins were expressed in E. coli from the pGEX-4T-1 vector and purified using glutathione Sepharose beads following the manufacturer's recommendations. EGFP-ARHGEF3 was expressed in E. coli from the pET His6 EGFP TEV vector, and purified using Ni-affinity chromatography.

GTPase RhoA Activity Assays

Cellular RhoA activity was measured as previously described (26). Briefly, HEK293 cells in 10-cm plate were transfected for 24 hrs and then lysed in 50 mM Tris (pH 7.4), 10mM $MgCl_2$, 500 mM NaCl, 1% (vol/vol) Triton X-100, 0.1%SDS, 0.5% sodium deoxycholate, and 1x Protease Inhibitors cocktail. The cell lysates were incubated for 1 hr at 4 ^{o}C with 60 µg of GST-rhotekin beads, followed by washing with 50 mM Tris (pH 7.4), 10 mM $MgCl_2$, 150 mM NaCl, 1% (vol/vol) Triton X-100, and 1x Protease Inhibitors cocktail, and analysis by western blotting with an anti-RhoA antibody. To measure nucleotide exchange activity of RhoA in vitro, the mant-GTP exchange assay was performed following manufacturer's recommendation in 20

μL reaction volume in a 384-well plate. In short, purified RhoA (1 μM) was added to exchange buffer (40 mM Tris-HCl, pH 7.5, 100 mM NaCl, 20 mM $MgCl_2$, 4 μM mant-GTP) and fluorescence (ex360/em440) was recorded for 5 min before adding purified GST-ARHGEF3 or GST-ARHGEF3-K342E/K348E.

RFC Matrix Generation and PIP-binding Prediction

The recursive-learning strategy described by Park et al. (14) was followed, with the details specific to our study described here. The list of human PH domain-containing proteins was obtained from the Pfam database and individual PH domain sequences were acquired from UniProt. After initial alignment of all sequences using PROMALS3D with default parameters, about 30 sequences were removed due to large insertions, and the remaining 250 PH domains were re-aligned. MATLAB was used to create two 297 x 20 probability matrices for PIP-binding (P^B) and non-binding proteins (P^{NB}), respectively. An amino acid was scored in the probability matrix if it was found in at least one PH domain in the group. The RFC matrix was then calculated by taking the log of the ratio of P^B to P^{NB}. In Figure 6b, the value at each amino acid position is the average of sum-square of values in the corresponding position. For scoring a PH domain, the values in the RFC matrix for matching amino acids at all positions were summed.

ACKNOWLEDGEMENTS

We thank Dr. Lisa Stubbs and Mr. Joe Troy for help with the initial identification of PH domain sequences, and Ms. Lucy Yao and Ms. Eleanor Marcet for assistance with western blotting. This work was supported by grants from the National Institute of General Medical Sciences (R01 GM089771 to JC and R35 GM122569 to TH).

AUTHOR CONTRIBUTIONS

Conceptualization and experimental design by N.S., A.R.-O., B.J.L., T.H. and J.C.; experimentation, data collection and analysis by N.S., A.R.-O., M.A.C. and J.F.M.; resources by B.J.L. and T.H.; writing by N.S., A.R.-O., and J.C.; editing by N.S., A.R.-O., T.H. and J.C.; funding acquisition by T.H. and J.C.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. Balla, T. Inositol-lipid binding motifs: signal integrators through protein-lipid and protein-protein interactions. *J Cell Sci* **118**, 2093-104 (2005).
2. Di Paolo, G. & De Camilli, P. Phosphoinositides in cell regulation and membrane dynamics. *Nature* **443**, 651-7 (2006).
3. Moravcevic, K., Oxley, C.L. & Lemmon, M.A. Conditional peripheral membrane proteins: facing up to limited specificity. *Structure* **20**, 15-27 (2012).
4. Lemmon, M.A. Membrane recognition by phospholipid-binding domains. *Nat Rev Mol Cell Biol* **9**, 99-111 (2008).
5. Haslam, R.J., Koide, H.B. & Hemmings, B.A. Pleckstrin domain homology. *Nature* **363**, 309-10 (1993).
6. Mayer, B.J., Ren, R., Clark, K.L. & Baltimore, D. A putative modular domain present in diverse signaling proteins. *Cell* **73**, 629-30 (1993).
7. Yoon, H.S., Hajduk, P.J., Petros, A.M., Olejniczak, E.T., Meadows, R.P. & Fesik, S.W. Solution structure of a pleckstrin-homology domain. *Nature* **369**, 672-5 (1994).
8. Harlan, J.E., Hajduk, P.J., Yoon, H.S. & Fesik, S.W. Pleckstrin homology domains bind to phosphatidylinositol-4,5-bisphosphate. *Nature* **371**, 168-70 (1994).
9. Lemmon, M.A., Ferguson, K.M., O'Brien, R., Sigler, P.B. & Schlessinger, J. Specific and high-affinity binding of inositol phosphates to an isolated pleckstrin homology domain. *Proc Natl Acad Sci U S A* **92**, 10472-6 (1995).
10. James, S.R., Downes, C.P., Gigg, R., Grove, S.J., Holmes, A.B. & Alessi, D.R. Specific binding of the Akt-1 protein kinase to phosphatidylinositol 3,4,5-trisphosphate without subsequent activation. *Biochem J* **315 (Pt 3)**, 709-13 (1996).
11. Watton, S.J. & Downward, J. Akt/PKB localisation and 3′ phosphoinositide generation at sites of epithelial cell–matrix and cell–cell interaction. *Current Biology* **9**, 433-436 (1999).
12. Varnai, P. & Balla, T. Visualization of phosphoinositides that bind pleckstrin homology domains: calcium- and agonist-induced dynamic changes and relationship to myo-[3H]inositol-labeled phosphoinositide pools. *J Cell Biol* **143**, 501-10 (1998).
13. Yu, J.W., Mendrola, J.M., Audhya, A., Singh, S., Keleti, D., DeWald, D.B., Murray, D., Emr, S.D. & Lemmon, M.A. Genome-wide analysis of membrane targeting by S. cerevisiae pleckstrin homology domains. *Mol Cell* **13**, 677-88 (2004).
14. Park, W.S., Heo, W.D., Whalen, J.H., O'Rourke, N.A., Bryan, H.M., Meyer, T. & Teruel, M.N. Comprehensive identification of PIP3-regulated PH domains from C. elegans to H. sapiens by model prediction and live imaging. *Mol Cell* **30**, 381-92 (2008).
15. Lemmon, M.A. Pleckstrin homology domains: not just for phosphoinositides. *Biochem Soc Trans* **32**, 707-11 (2004).
16. Arauz, E., Aggarwal, V., Jain, A., Ha, T. & Chen, J. Single-Molecule Analysis of Lipid-Protein Interactions in Crude Cell Lysates. *Anal Chem* **88**, 4269-76 (PMC4863138) (2016).
17. Hasegawa, J., Strunk, B.S. & Weisman, L.S. PI5P and PI(3,5)P2: Minor, but Essential Phosphoinositides. *Cell Struct Funct* **42**, 49-60 (2017).
18. van den Bogaart, G., Meyenberg, K., Risselada, H.J., Amin, H., Willig, K.I., Hubrich, B.E., Dier, M., Hell, S.W., Grubmuller, H., Diederichsen, U. & Jahn, R. Membrane protein sequestering by ionic protein-lipid interactions. *Nature* **479**, 552-5 (2011).

19. Bravo, J., Karathanassis, D., Pacold, C.M., Pacold, M.E., Ellson, C.D., Anderson, K.E., Butler, P.J.G., Lavenir, I., Perisic, O., Hawkins, P.T., Stephens, L. & Williams, R.L. The Crystal Structure of the PX Domain from p40phox Bound to Phosphatidylinositol 3-Phosphate. *Molecular Cell* **8**, 829-839 (2001).
20. Vonkova, I., Saliba, A.E., Deghou, S., Anand, K., Ceschia, S., Doerks, T., Galih, A., Kugler, K.G., Maeda, K., Rybin, V., van Noort, V., Ellenberg, J., Bork, P. & Gavin, A.C. Lipid Cooperativity as a General Membrane-Recruitment Principle for PH Domains. *Cell Rep* **12**, 1519-30 (2015).
21. Hart, R., Stanley, P., Chakravarty, P. & Hogg, N. The kindlin 3 pleckstrin homology domain has an essential role in lymphocyte function-associated antigen 1 (LFA-1) integrin-mediated B cell adhesion and migration. *J Biol Chem* **288**, 14852-62 (2013).
22. Arthur, W.T., Ellerbroek, S.M., Der, C.J., Burridge, K. & Wennerberg, K. XPLN, a guanine nucleotide exchange factor for RhoA and RhoB, but not RhoC. *J Biol Chem.* **277**, 42964-72. (2002).
23. Cook, D.R., Rossman, K.L. & Der, C.J. Rho guanine nucleotide exchange factors: regulators of Rho GTPase activity in development and disease. *Oncogene* **33**, 4021-35 (2014).
24. Shazman, S., Celniker, G., Haber, O., Glaser, F. & Mandel-Gutfreund, Y. Patch Finder Plus (PFplus): a web server for extracting and displaying positive electrostatic patches on protein surfaces. *Nucleic Acids Res* **35**, W526-30 (2007).
25. Murayama, K., Kato-Murayama, M., Akasaka, R., Terada, T., Yokoyama, S. & Shirouzu, M. Structure of the Rho-specific guanine nucleotide-exchange factor Xpln. *Acta Crystallographica Section F Structural Biology and Crystallization Communications* **68**, 1455-1459 (2012).
26. Khanna, N., Fang, Y., Yoon, M.S. & Chen, J. XPLN is an endogenous inhibitor of mTORC2. *Proc Natl Acad Sci U S A.* **110**, 15979-84. (PMC3791717) (2013).
27. Pei, J., Kim, B.H. & Grishin, N.V. PROMALS3D: a tool for multiple protein sequence and structure alignments. *Nucleic Acids Res* **36**, 2295-300 (2008).
28. Jaffe, A.B., Aspenstrom, P. & Hall, A. Human CNK1 acts as a scaffold protein, linking Rho and Ras signal transduction pathways. *Mol Cell Biol* **24**, 1736-46 (2004).
29. Lim, J., Zhou, M., Veenstra, T.D. & Morrison, D.K. The CNK1 scaffold binds cytohesins and promotes insulin pathway signaling. *Genes Dev* **24**, 1496-506 (2010).
30. Martin, T.F. PI(4,5)P(2)-binding effector proteins for vesicle exocytosis. *Biochim Biophys Acta* **1851**, 785-93 (2015).
31. Rossman, K.L., Der, C.J. & Sondek, J. GEF means go: turning on RHO GTPases with guanine nucleotide-exchange factors. *Nat Rev Mol Cell Biol* **6**, 167-80 (2005).
32. Kanai, F., Liu, H., Field, S.J., Akbary, H., Matsuo, T., Brown, G.E., Cantley, L.C. & Yaffe, M.B. The PX domains of p47phox and p40phox bind to lipid products of PI(3)K. *Nat Cell Biol* **3**, 675-8 (2001).
33. Jain, A., Liu, R., Ramani, B., Arauz, E., Ishitsuka, Y., Ragunathan, K., Park, J., Chen, J., Xiang, Y.K. & Ha, T. Probing cellular protein complexes using single-molecule pull-down. *Nature* **473**, 484-488. (PMC3103084) (2011).

Table 1.

Group	Protein Name	Entrez Gene ID	Lipid Bound in SiMPull	Reported Lipid Binding
I	BTK	695	PIP_3	PIP_3
	FERMT3[a]	83706	$PI(3,4)P_2$, PIP_3	PIP_3
	GAB2[a]	9846	$PI(3,4)P_2$, PIP_3	PIP_3
	ITK	3702	PIP_3	PIP_3
	PLEKHA1	59338	$PI(3,4)P_2$	$PI(3,4)P_2$
	PLEKHF1	79156	PI(3)P	PI(3)P
	SKAP1	8631	PIP_3	PIP_3
	SKAP2[b,c]	8935	PIP_3	$PI(3,4)P_2$, PIP_3
II	ACAP1	9744	PI(4)P, $PI(3,4)P_2$	all PIPs
	ADAP1	11033	$PI(3,4)P_2$, $PI(4,5)P_2$, PIP_3	$PI(3,4)P_2$, PIP_3
	ARHGAP12[b]	94134	None/$PI(4,5)P_2$	$PI(4,5)P_2$, PIP_3
	ARHGAP15	9938	PI(3)P, PI(4)P, $PI(3,4)P_2$, $PI(4,5)P_2$, PIP_3	Not reported
	ARHGAP26	23092	PI(3)P, $PI(3,4)P_2$	$PI(4,5)P_2$
	ARHGEF3	50650	$PI(4,5)P_2$	Not reported
	ARHGEF5	7984	$PI(3,4)P_2$, $PI(4,5)P_2$	PI(3)P, PI(4)P, PI(5)P
	ARHGEF9	23229	PI(5)P	PI(3)P
	ARHGEF16	27237	PI(3)P	Not reported
	ARHGEF39	84904	$PI(3,4)P_2$, PIP_3	Not reported
	CNKSR1	10256	PI(4)P	all PIPs
	DEF6	50619	$PI(3,4)P_2$	PIP_3
	DOK2	9046	PI(3)P	PI(3)P, PI(4)P, PI(5)P
	EXOC8	149371	$PI(4,5)P_2$	Not reported
	GAB1	2549	PI(4)P	PIP_3
	MCF2L	23263	PI(5)P, $PI(3,4)P_2$, $PI(4,5)P_2$, PIP_3	$PI(4,5)P_2$
	NGEF	25791	PI(5)P	Not reported
	OSBPL5	114879	PI(4)P, PI(5)P, $PI(3,4)P_2$, $PI(4,5)P_2$	$PI(3,4)P_2$, $PI(3,5)P_2$, $PI(4,5)P_2$, PIP_3
	OSBPL10	114884	PI(4)P, $PI(3,4)P_2$	PI(3)P, PI(4)P, PI(5)P, PA
	PLCD4[b]	84812	None/PIP_3	$PI(4,5)P_2$, PIP_3
	PLEKHB2	55041	PI(3)P	$PI(3,4)P_2$, $PI(3,4,5)P_3$, PS
	PLEKHJ1	55111	PI(5)P, $PI(3,4)P_2$, $PI(4,5)P_2$, PIP_3	Not reported
	PSD2	84249	$PI(4,5)P_2$	Not reported
	SBF2[b]	81846	None/ $PI(4,5)P_2$, PIP_3	PIP_3
	SWAP70	23075	PI(3)P	PI(3)P, PI(4)P, $PI(3,4)P_2$, PIP_3
	VAV1	7409	PI(5)P, $PI(3,4)P_2$, $PI(4,5)P_2$, PIP_3	$PI(4,5)P_2$, PIP_3

Table 1 (continued)

Group	Protein Name	Entrez Gene ID	Lipid Bound in SiMPull	Reported Lipid Binding
III	AFAP1L1	134265	None	PIP_3
	APPL1	26060	None	PI(3)P, PI(4)P, PI(5)P, $PI(3,4)P_2$, $PI(3,5)P_2$
	APPL2	55198	None	PI(3)P, PI(4)P, PI(5)P, $PI(3,4)P_2$, $PI(3,5)P_2$
	DOK1	1796	None	PI(5)P, $PI(3,4)P_2$, $PI(4,5)P_2$, PIP_3
	FERMT2	10979	None	$PI(3,5)P_2$, PIP_3
	GRB14	2888	None	all PIPs
	GRB7	2886	None	PI(3)P, PI(4)P, PI(5)P, $PI(3,5)P_2$
	GRK2	156	None	$PI(4,5)P_2$
	OSBPL8	114882	None	$PI(3,4)P_2$, $PI(3,5)P_2$, $PI(4,5)P_2$, PIP_3
	PLEK	5341	None	$PI(4,5)P_2$
	PLEKHF2	79666	None	PI(3)P
	PLEKHN1	84069	None	PI4P, Cardiolipin, PS, PA
IV	AFAP1	60312	None	Not reported
	ARHGAP25	9938	None	Not reported
	ARHGEF7	8874	None	Not reported
	BMX	660	None	Not reported
	DOK3	79930	None	Not reported
	DOK4	55715	None	Not reported
	FGD5	152273	None	Not reported
	KIF1B	23095	None	Not reported
	PHETA2	150368	None	Not reported
	PLCG2	5336	None	Not reported
	PLEKHO2	80301	None	Not reported
	PRKD3	23683	None	Not reported
	PSD4	23550	None	Not reported
	RASA1	5921	None	Not reported
	RASGRF1	5923	None	Not reported
	SOS2	6655	None	Not reported
	STAP1	26228	None	Not reported
Non-specific	CERT1	10087	all vesicles tested	PI(4)P, $PI(4,5)P_2$
	NET1	10276	all vesicles tested	Not reported
	PLEKHO1	51177	PI(3)P, PI(5)P, $PI(3,4)P_2$, $PI(4,5)P_2$, PIP_3, PA	all PIPs
	SPATA13	221178	all vesicles tested	Not reported

Table 1. Summary of lipid-SiMPull assay results for 67 human PH domain-containing proteins. See text for experimental details. Three independent experiments were performed for each protein, and the complete data set is included in Supplementary Table 1. For the reported lipid binding, selected references are provided in Supplementary Table 2. Based on our assay results and information from the literature, the 67 proteins are grouped as follows: **Group I**, proteins previously reported to bind PIP and binding confirmed by SiMPull; **Group II**, proteins

found to bind PIP in SiMPull but not previously reported or previously reported to have different binding specificity; **Group III**, proteins previously reported to bind PIP but binding not observed in SiMPull; **Group IV**, proteins not reported to bind PIP and not found to bind PIP in SiMPull. [a] $PI(3,4)P_2$ had not been examined in the published work. [b] Binding in SiMPull was only observed with 20% PIP. [c] Binding to 20% $PI(3,4)P_2$ was not examined.

PH Domain	Lipid binding by PH domain	Lipid binding by full-length protein
ACAP1_PH	PI(3,4)P_2	PI(3,4)P_2, PI(4)P
ADAP1_PH1	None	PI(3,4,5)P_3, PI(4,5)P_2, PI(3,4)P_2
ADAP1_PH2	None	PI(3,4,5)P_3, PI(4,5)P_2, PI(3,4)P_2
AFAP1L1-PH1	None	None
AFAP1L1-PH2	PI(3,4,5)P_3, PI(4,5)P_2	None
ARHGEF3-PH	None	PI(4,5)P_2
ARHGEF5_PH	None	PI(4,5)P_2, PI(3,4)P_2
ARHGEF9_PH	None	PI(5)P
DOK2-PH	PI(3,4,5)P3, PI(3)P, PI(4)P	PI(3)P
FERMT3-PH	None	PI(3,4,5)P_3, PI(3,4)P_2
PLEKHA1-PH1	None	PI(3,4)P_2
PLEKHA1-PH2	PI(3,4)P_2	PI(3,4)P_2
SPATA13-PH	None	Bound all lipid vesicles
VAV1-PH	None	PI(3,4,5)P_3, PI(4,5)P_2, PI(3,4)P_2, PI(5)P

Table 2. SiMPull assay results for selected PH domains. Experimental details and data processing were as described for Table 1 and Supplementary Table 1. At least three independent experiments were performed with similar outcome for each PH domain.

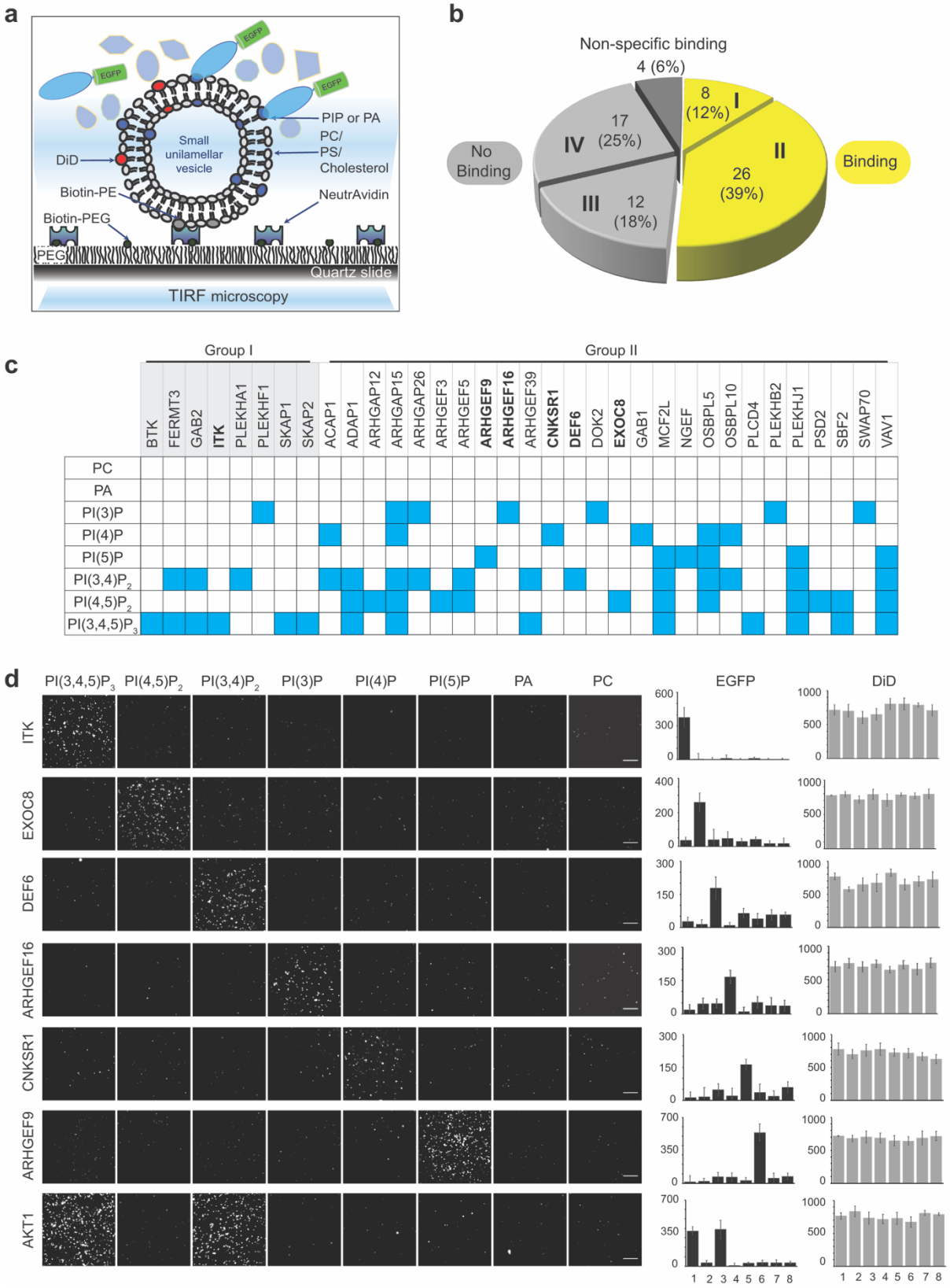

a
EGFP
EGFP
EGFP
EGFP
PIP or PA
PC/
PS/
Cholesterol
Small
unilamellar
vesicle
DiD
Biotin-PE
NeutrAvidin
Biotin-PEG
PEG
Quartz slide
TIRF microscopy
b
Non-specific binding
4 (6%)
8
(12%)
I
II
26
(39%)
Binding
III
12
(18%)
IV
17
(25%)
No
Binding
c
Group I
Group II
BTK
FERMT3
GAB2
ITK
PLEKHA1
PLEKHF1
SKAP1
SKAP2
ACAP1
ADAP1
ARHGAP12
ARHGAP15
ARHGAP26
ARHGEF3
ARHGEF5
ARHGEF9
ARHGEF16
ARHGEF39
CNKSR1
DEF6
DOK2
EXOC8
GAB1
MCF2L
NGEF
OSBPL5
OSBPL10
PLCD4
PLEKHB2
PLEKHJ1
PSD2
SBF2
SWAP70
VAV1
PC
PA
PI(3)P
PI(4)P
PI(5)P
PI(3,4)P2
PI(4,5)P2
PI(3,4,5)P3
d
PI(3,4,5)P3
PI(4,5)P2
PI(3,4)P2
PI(3)P
PI(4)P
PI(5)P
PA
PC
EGFP
DiD
ITK
EXOC8
DEF6
ARHGEF16
CNKSR1
ARHGEF9
AKT1
1 2 3 4 5 6 7 8

Figure 1. Specific PIP binding by full-length human PH domain proteins.

(a) Schematic representation of the lipid-SiMPull assay to detect lipid-protein interactions.
(b) A pie chart summarizing the results of PIP-binding assays for 67 full-length human PH domain proteins against 8 types of vesicles (see text for details). I: reported binding confirmed. II: novel binding. III: binding reported but not found by SiMPull. IV: binding not reported nor found by SiMPull. The grouping corresponds to that in Table 1.
(c) A binary representation of results for the 34 proteins found to bind PIPs in SiMPull. The vesicles are identified by the unique PIP, PA, or PC, but all contained PC, PS, cholesterol, DiD, and biotin-PE. Blue: binding; white: no binding. SiMPull data for the proteins in bold are shown in (d).
(d) Representative TIRF images of seven EGFP-fusions pulled down by vesicles containing 5% PIP or PA, with PC as a negative control. EGFP and DiD spots were quantified from 10 or more image areas to yield the average number of spots per 1600 μm^2 for each assay, shown in the graphs on the right. Each assay was repeated with lysates from at least three independent transfections, and the EGFP data can be found in Supplementary Table 1. The results for 67 proteins are summarized in Table 1. Scale bars: 5 μm.

a

Protein	Lipid
ACAP1	PI(4)P
ARHGAP15	PI(3,4)P_2
ARHGAP26	PI(3,4)P_2
ARHGEF3	PI(4,5)P_2
ARHGEF5	PI(3,4)P_2
ARHGEF9	PI(5)P
ARHGEF16	PI(3)P
ARHGEF39	PI(3,4)P_2
CNKSR1	PI(4)P
DEF6	PI(3,4)P_2
DOK2	PI(3)P
EXOC8	PI(4,5)P_2
GAB1	PI(4)P
MCF2L	PI(4,5)P_2
NGEF	PI(5)P
OSBPL5	PI(4,5)P_2
OSBPL10	PI(3,4)P_2
PLEKHB2	PI(3)P
PLEKHJ1	PI(3,4,5)P_3
PLEKHO1	PI(3,4,5)P_3
PSD2	PI(4,5)P_2
SWAP70	PI(3)P
VAV1	PI(4,5)P_2

b

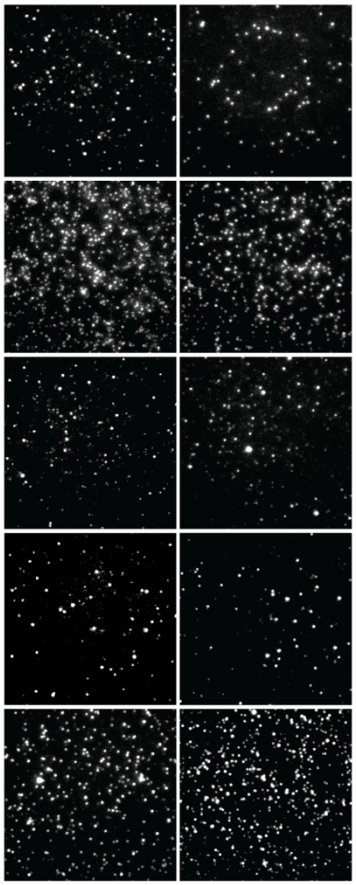


Figure 2. The presence of PS in lipid vesicles does not affect PIP binding by proteins.

(a) SiMPull was performed for 23 proteins against lipid vesicles containing the indicated PIPs, with or without PS and cholesterol. Binding was observed for all with both types of vesicles. For each condition at least two independent experiments were performed with similar results. Data were analyzed as described for Table 1 and Supplementary Table 1.
(b) Representative TIRF images of EGFP pulldown for 5 proteins that are also shown in Figure 1d, each by two types of lipid vesicles. Scale bar: 5 µm.

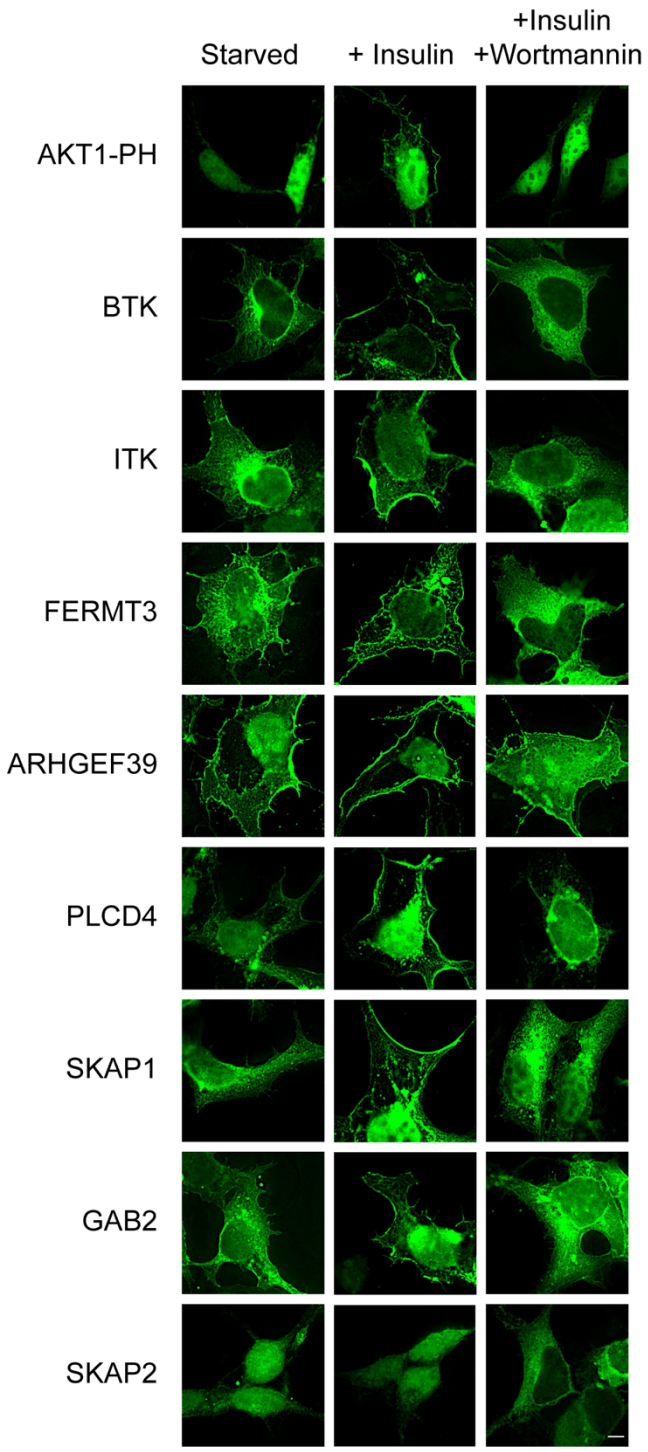


Figure 3. Membrane translocation of PIP_3-binding proteins.

HEK293 cells were transiently transfected with various EGFP-tagged proteins, followed by serum starvation overnight and then stimulation with 200 nM insulin for 5 min, with or without pretreatment by 100 nM wortmannin for 10 min. The cells were fixed for imaging. Representative images are shown. Scale bar: 5 µm.

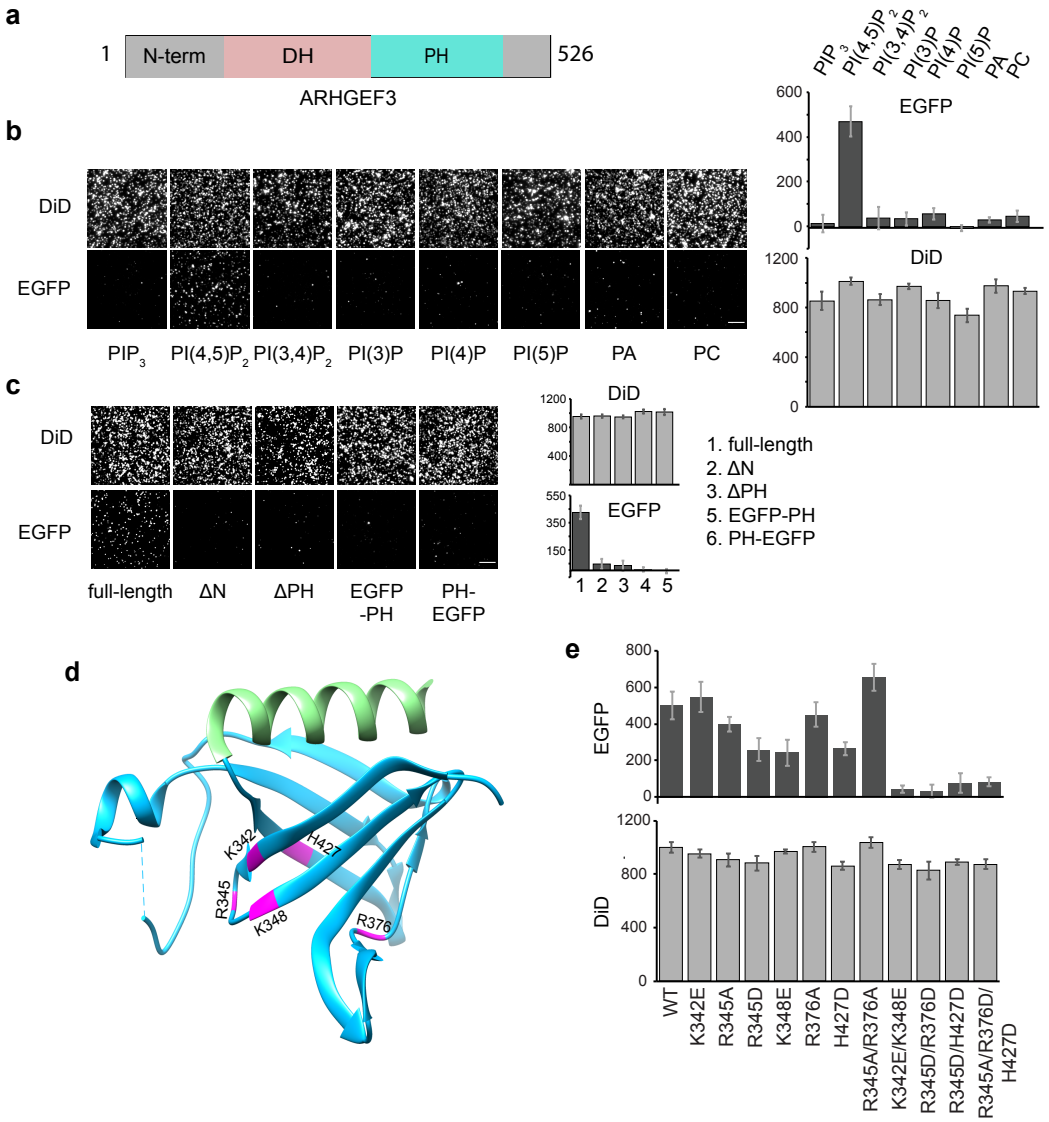


Figure 4. ARHGEF3 binds PI(4,5)P_2 through its PH domain.

(a) Schematic diagram of ARHGEF3 domain structure.
(b) HEK293 cells were transiently transfected with EGFP-ARHGEF3, and the lysates were subjected to SiMPull assays against various types of lipid vesicles as indicated. Representative TIRF images are shown. EGFP and DiD spots were quantified from 10 or more image areas to yield the average number of spots per 1600 μm^2 for each assay, shown in the graphs on the right. Each assay was performed with at least three independent transfections with similar results.
(c) Similar to (b), various EGFP-tagged ARHGEF3 fragments were assayed for binding to vesicles containing 5% PI(4,5)P_2.
(d) Ribbon representation of the structure of ARHGEF3 PH domain (PDB 2Z0Q), with residues predicted to interact with PI(4,5)P_2 highlighted in magenta.
(e) Similar to (b), EGFP-tagged ARHGEF3 mutants were assayed for binding to 5% PI(4,5)P_2 vesicles. Each assay was repeated with lysates from at least three independent transfections. Scale bars: 5 µm.

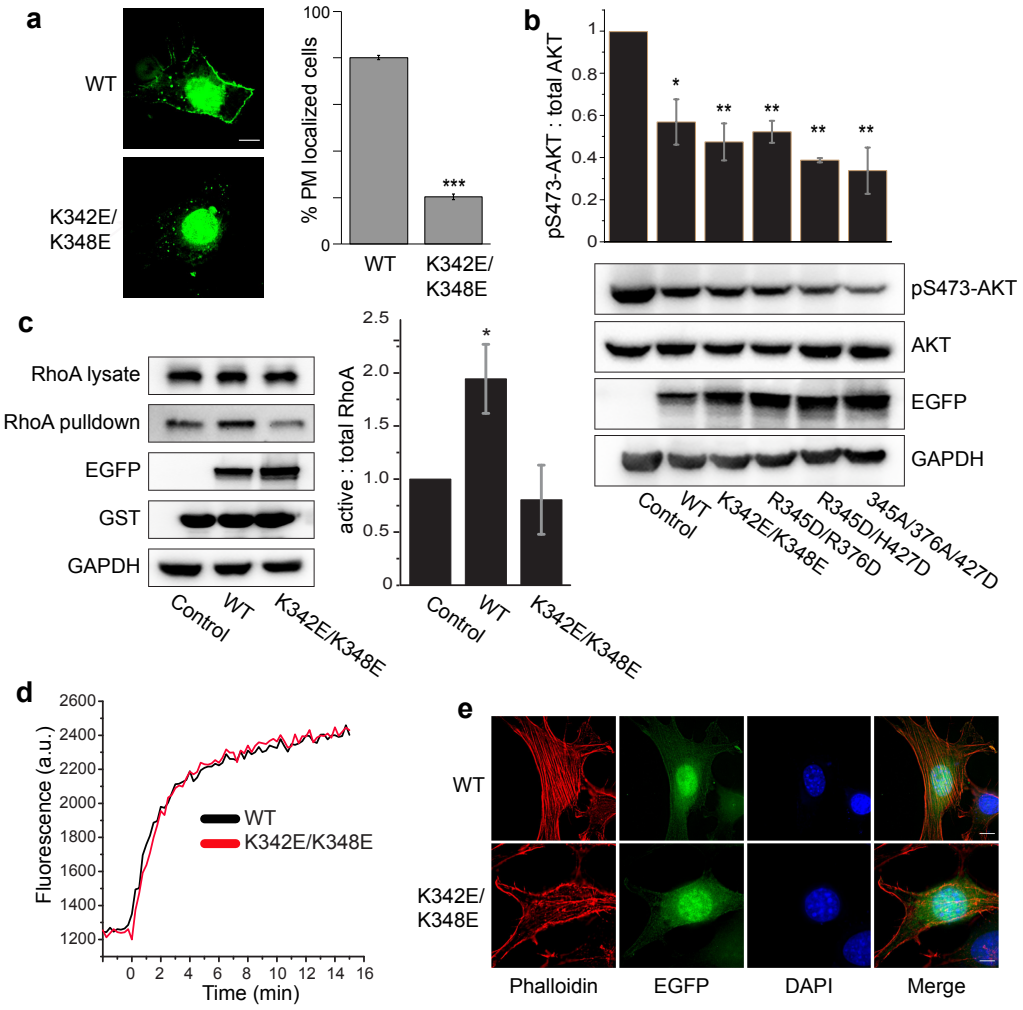


Figure 5. PI(4,5)P_2 binding is necessary for ARHGEF3 membrane targeting, cellular activity and function.

(a) NIH3T3 cells were transfected for 24 hrs with EGFP-tagged wild-type or lipid binding-deficient mutant K342E/K348E of ARHGEF3. The transfected cells were then treated with 0.03% digitonin for 2.5 min to remove cytosolic proteins before fixation. Percentage of transfected cells with PM localization was quantified (n=3).

(b) HEK293 cells were transfected with wild-type or lipid binding-deficient mutants of ARHGEF3, serum-starved overnight, and then stimulated with 10% FBS for 30 min followed by cell lysis and western blotting. Western blot signals were quantified by densitometry to generate ratio of pAKT versus total AKT (n=3).

(c) HEK293 cells were transfected for 24 hrs with empty vector (control), wildtype, or K342E/K348E mutant of ARHGEF3. Cell lysates were subjected to active RhoA pulldown assay using GST-rhotekin beads, and analyzed by western blotting. Western blot signals were quantified by densitometry to generate the relative ratios of RhoA pulled down (active) and RhoA in lysates (total) shown in the graph (n=5).

(d) Purified wild-type and K342E/K348E mutant ARHGEF3 were subjected to in vitro RhoA guanine nucleotide exchange assay. Representative results are shown (n=3).
(e) NIH3T3 cells were transfected for 24 hrs with wild-type or K342E/K348E mutant of ARHGEF3, followed by fixation and phalloidin/DAPI staining. Representative images are shown (n=3).
Error bars represent SEM. Student's t test was performed. $*P < 0.05$; $**P < 0.01$; $***P < 0.001$.
Scale bars: 5 µm.

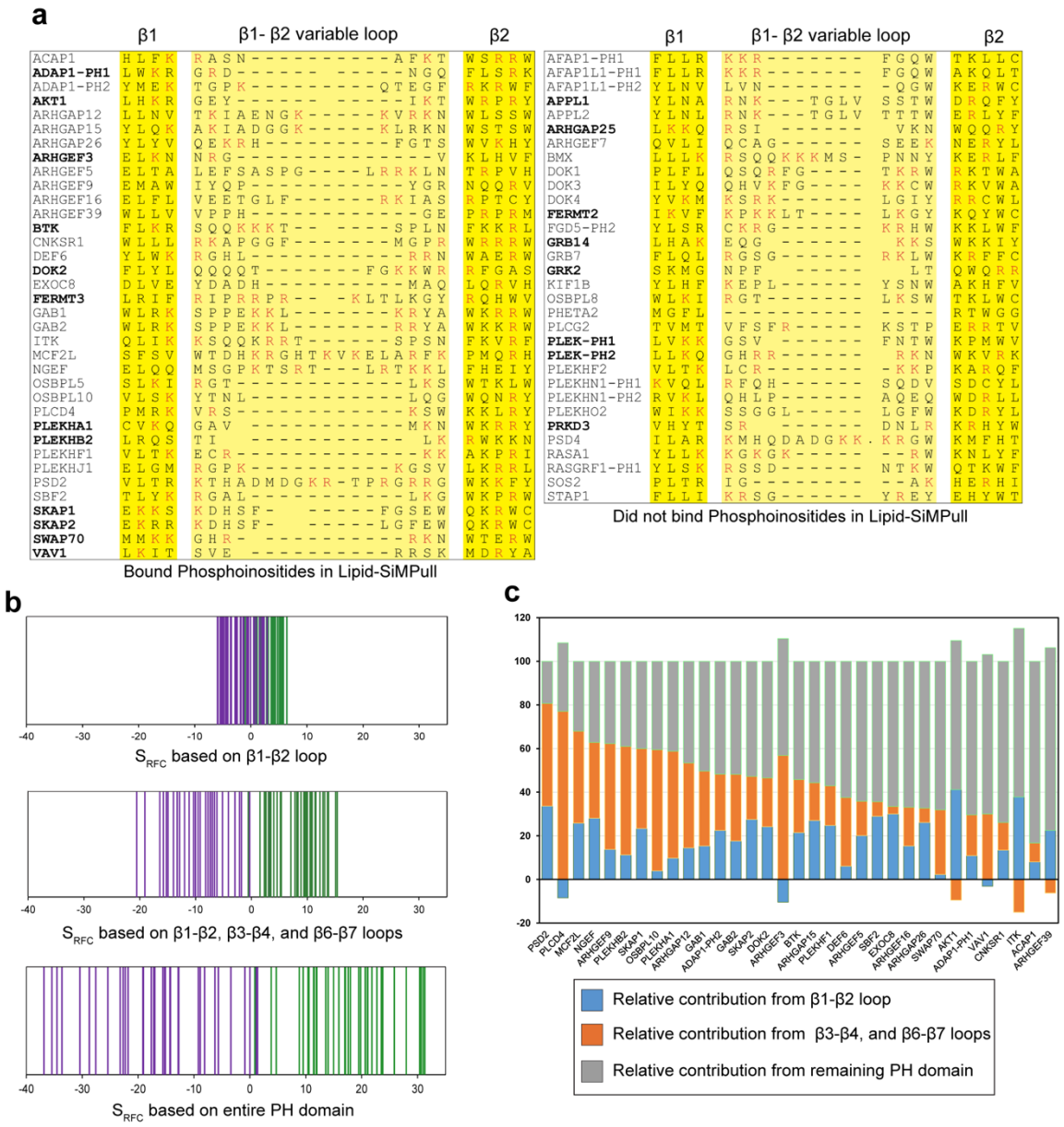


Figure 6. Recursive-learning algorithm distinguishes PIP-binding when taking the entire PH domain into consideration.

(**a**) Sequence alignment of PIP-binding (left) and non-binding (right) PH domains in the β1- β2 region. Bold: PH domains with reported crystal structures. Red: residues conforming to the $KX_n(K/R)XR$ motif. Three PH domains (AFAP1-PH2, FGD5-PH1 and RASGRF1-PH2) were not shown in the alignment due to an unusually short β1 sheet.

(**b**) S_{RFC} scores were calculated for the 71 PH domains in our assays, with amino acids from β1-β2 variable loop, β1-β2, β3-β4, β6-β7 variable loops, or the entire PH domain. Each vertical line represents one PH domain. Green: bound PIP in SiMPull assays; purple: did not bind PIP in SiMPull assays.

(**c**) Relative contribution by β1-β2 loop, β3-β4 and β6-β7 loops, or the rest of the PH domain is plotted based on S_{RFC} for each protein that bound PIP in SiMPull assays.

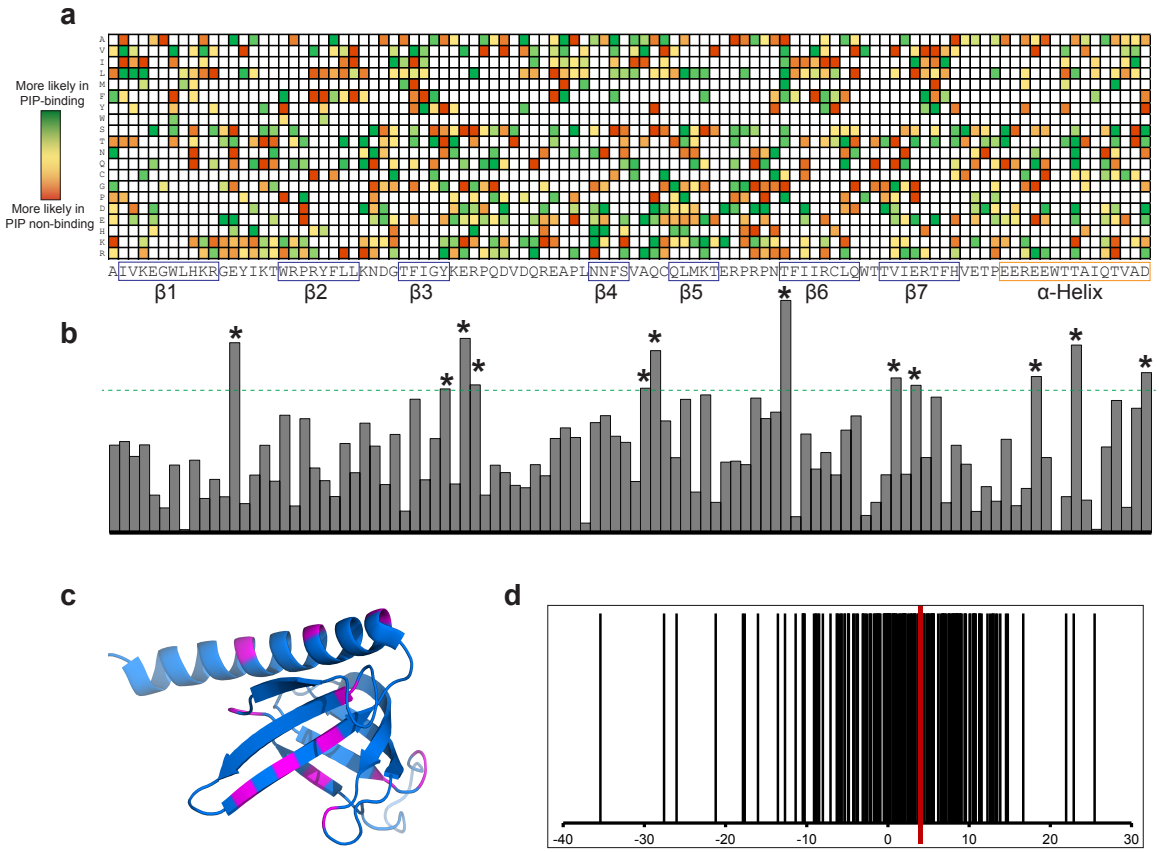
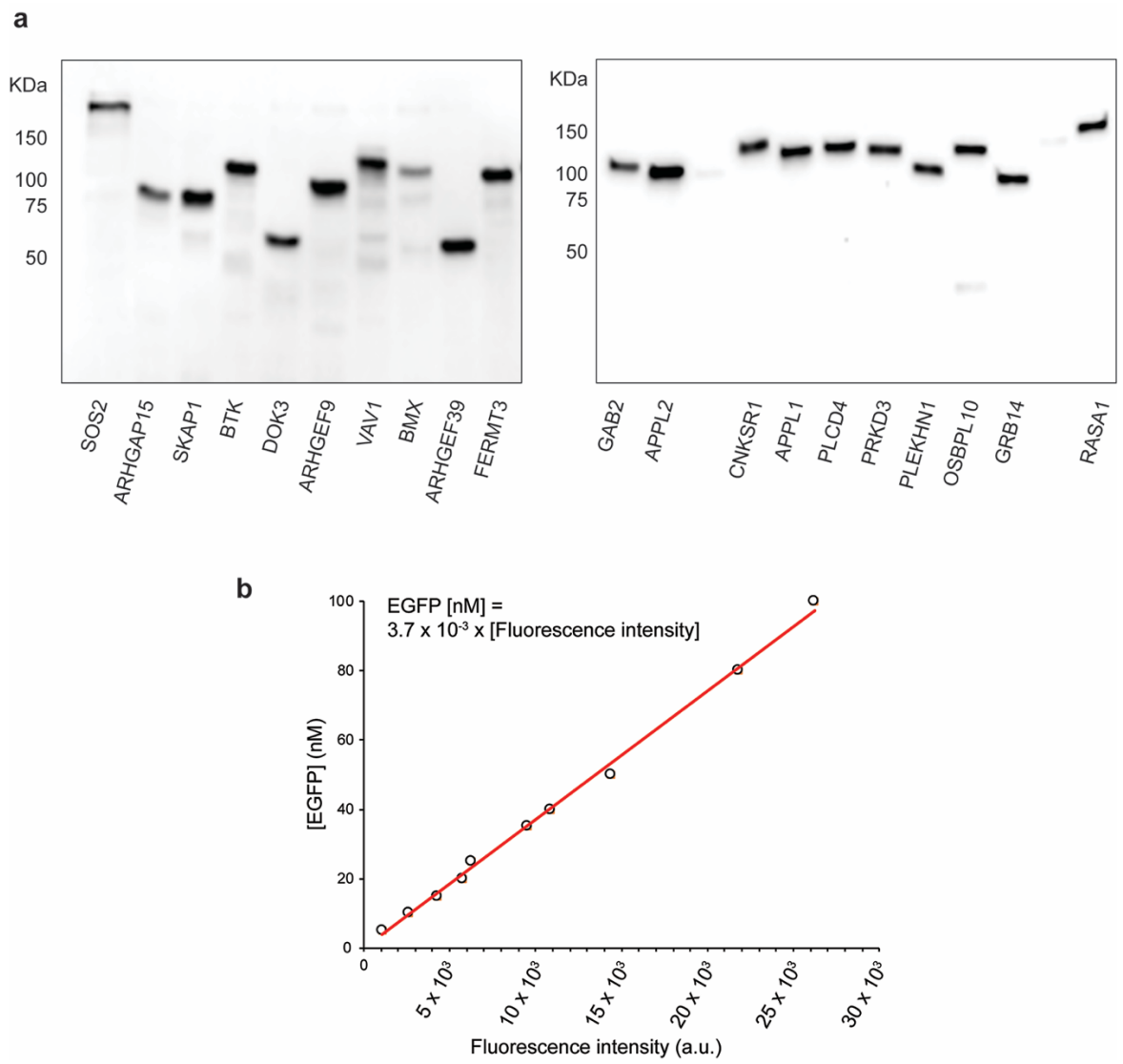


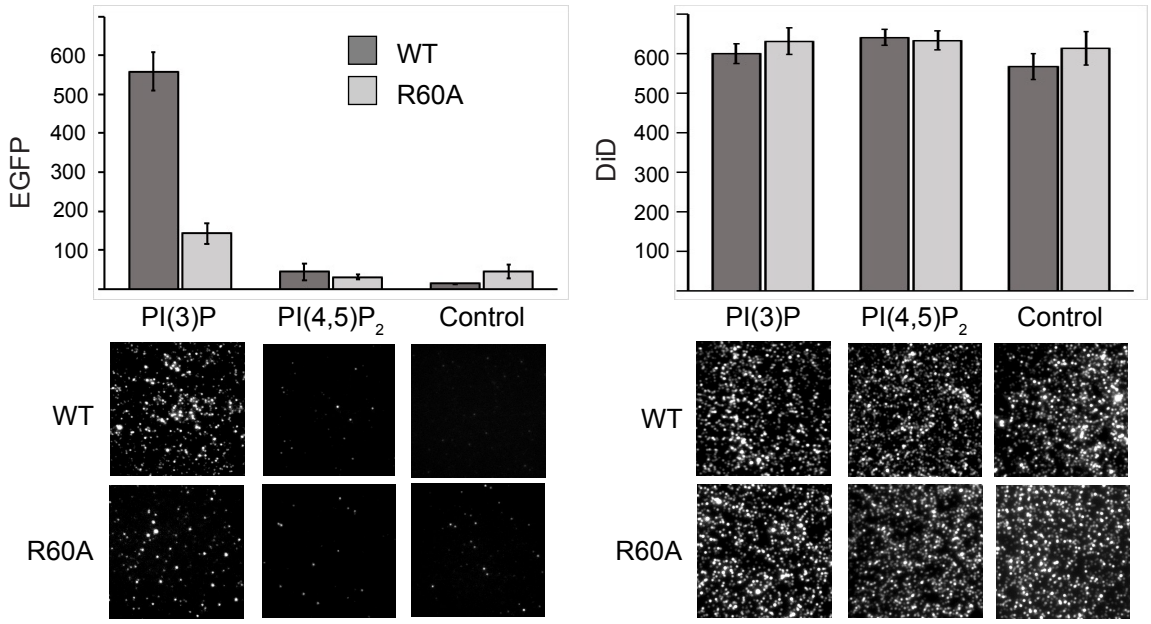
Figure 7. Recursive-learning algorithm identifies amino acid determinants of PIP-binding and predicts binding by human PH domains.

(**a**) A heat map showing the contribution of 20 amino acids at each position for the 71 PH domains. The amino acid positions are aligned to AKT1-PH domain based on its crystal structure (PDB ID: 1UNP).
(**b**) Contribution from different positions of PH domain towards PIP-binding was plotted from RFC-matrix. Twelve positions with the highest scores are marked by stars and aligned to the AKT1-PH sequence in (a).
(**c**) The 12 highest scoring amino acids in predicting PIP-binding are marked (magenta) on a representative PH domain structure (AKT-PH, PDB ID: 1UNP).
(**d**) Prediction of PIP-binding for 179 human PH domains using the RFC-matrix. Each vertical black line represents one PH domain. The red line indicates the cutoff value (3.94) for PIP-binding.



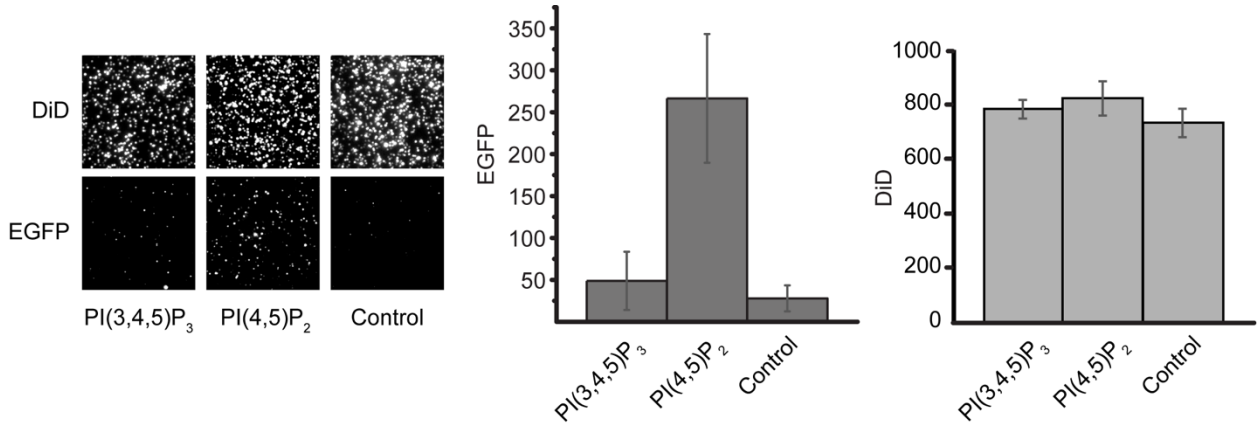
Supplementary Figure 1.

(a) Examples of EGFP-fusion proteins transiently expressed in HEK293 cells. Cell lysates were subjected to western blotting with an anti-GFP antibody.
(b) Fluorescence of pure recombinant EGFP (ex 488 / em 520) was measured at 10 different concentrations on a Spectramax GeminiXPS plate reader (Molecular Devices) to yield a linear standard curve.



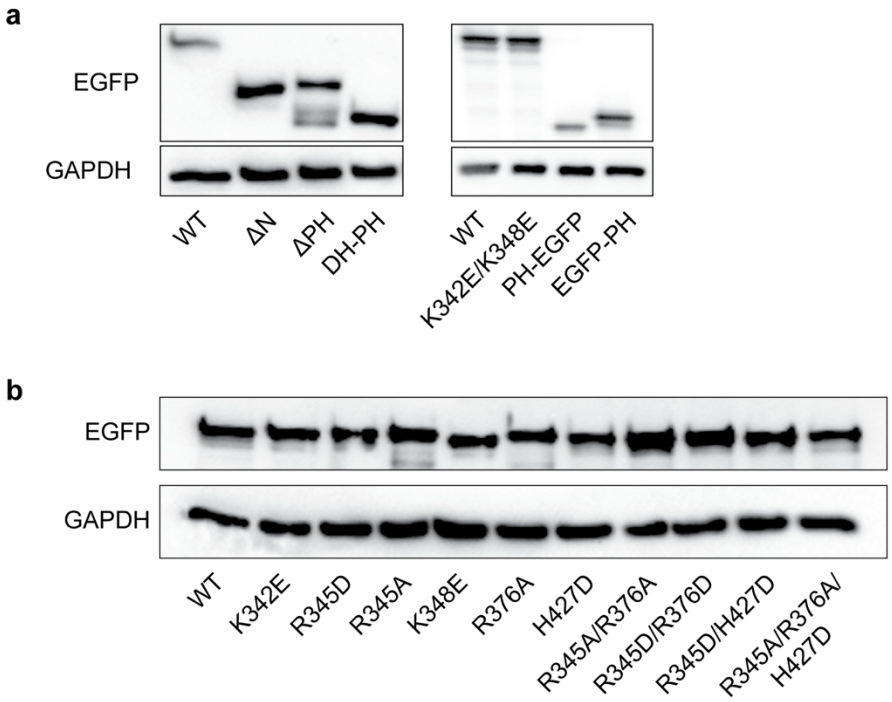
Supplementary Figure 2.

EGFP-p40PhoxPX was expressed in HEK293 cells and cell lysates (5 nM EGFP-fusion) were subjected to SiMPull assay with PI(3)P, PI(4,5)P_2 and PC (control) vesicles. EGFP and DiD (vesicle) spots were quantified as described in Figure 1 legend. The average results of 3 independent experiments are shown with error bars representing SEM.



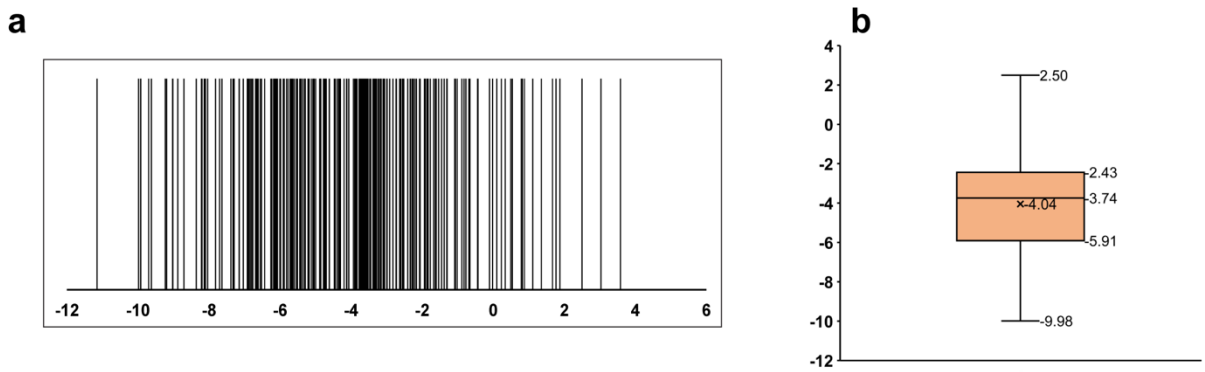
Supplementary Figure 3.

His-tagged EGFP-ARHGEF3 was expressed in E. coli from the pET-His6-EGFP-TEV vector, and purified using Ni-affinity chromatography. SiMPull assays were performed with 5 nM purified 6xHis-EGFP-ARHGEF3 and vesicles as indicated. Representative TIRF images are shown. EGFP (fusion protein) and DiD (vesicle) spots are quantified as described in Figure 1 legend. Control: PC vesicles.



Supplementary Figure 4.

HEK293 cells were transiently transfected to express various truncations **(a)** or point mutants **(b)** of ARHGEF3 fused to EGFP. Cell lysates were analyzed by western blotting with an anti-GFP antibody, and anti-GAPDH blotting served as loading control.



Supplementary Figure 5.

(**a**) S_{RFC} distribution for 250 human PH domains with scrambled sequences. The mean is - 4.04 with standard deviation of 2.66.
(**b**) Box plot for S_{RFC} distribution for the scrambled 250 PH domains. The quartile values with mean and median of the dataset are marked.