nature chemical biology

ARTICLE

PUBLISHED ONLINE: 12 JUNE 2017 | DOI: 10.1038/NCHEMBIO.2405

Genetically encoding phosphotyrosine and its nonhydrolyzable analog in bacteria

Xiaozhou Luo[1,2], Guangsen Fu[3], Rongsheng E Wang[2,5], Xueyong Zhu[4], Claudio Zambaldo[1,2], Renhe Liu[3], Tao Liu[1,2], Xiaoxuan Lyu[3], Jintang Du[3], Weimin Xuan[1,2], Anzhi Yao[1,2], Sean A Reed[1,2], Mingchao Kang[1,2], Yuhan Zhang[3], Hui Guo[3], Chunhui Huang[1,2], Peng-Yu Yang[1,2], Ian A Wilson[2,4], Peter G Schultz[1–3]* & Feng Wang[3]*

Tyrosine phosphorylation is a common protein post-translational modification that plays a critical role in signal transduction and the regulation of many cellular processes. Using a propeptide strategy to increase cellular uptake of *O*-phosphotyrosine (pTyr) and its nonhydrolyzable analog 4-phosphomethyl-L-phenylalanine (Pmp), we identified an orthogonal aminoacyl-tRNA synthetase–tRNA pair that allows site-specific incorporation of both pTyr and Pmp into recombinant proteins in response to the amber stop codon in *Escherichia coli* in good yields. The X-ray structure of the synthetase reveals a reconfigured substrate-binding site, formed by nonconservative mutations and substantial local structural perturbations. We demonstrate the utility of this method by introducing Pmp into a putative phosphorylation site and determining the affinities of the individual variants for the substrate 3BP2. In summary, this work provides a useful recombinant tool to dissect the biological functions of tyrosine phosphorylation at specific sites in the proteome.

Tyrosine phosphorylation is a key mechanism by which cells control a wide variety of biological processes[1–5]. However, the dynamic interconversion of phosphorylated and dephosphorylated isoforms of a protein, and phosphorylation at multiple sites, often complicates the study of the roles of individual phosphorylation events[6,7]. Access to site-specifically phosphorylated proteins has greatly facilitated mechanistic investigations of this post-translational modification[8–10]. Unfortunately, there is a lack of robust methods for generating individual phosphoprotein isoforms[11]. Native or engineered kinases lack the selectivity to modify specific tyrosine residues, and such modifications are subject to relatively nonselective removal by phosphatases in the cell. Semisynthetic methods allow the site-specific introduction of pTyr residues into proteins; however, the application of this approach to large proteins is limited by synthetic complexity and by difficulties in protein refolding after chemical synthesis. An *in vitro* translation system has also been developed, using a chemically aminoacylated suppressor tRNA, but affords relatively low yields of the phosphoproteins[12]. Moreover, these approaches do not allow site-specific tyrosine phosphorylation in living cells; therefore, they lack the potential to study biological functions of this important post-translational modification *in vivo*.

Several nonhydrolyzable phosphotyrosine analogs that are readily bioavailable in *Escherichia coli* have been genetically incorporated into proteins using cognate orthogonal amber suppressor tRNA–aminoacyl-tRNA synthetase (aaRS) pairs, including sulfotyrosine (sTyr) and 4-(carboxymethyl)phenylalanine (CMF)[13–21]. Furthermore, the nonhydrolyzable pTyr mimetic Pmp (**Fig. 1a**) has been selectively introduced into peptides to study the function of individual phosphorylation sites in SHP-1 and SHP-2 (refs. 22–24). The ability to recombinantly incorporate Pmp at specific sites of proteins would simplify the study of native pTyr sites in the proteomes of living cells[2]. However, attempts to genetically encode Pmp remain challenging because of poor cellular uptake of this amino acid.

Here, we report the use of a propeptide strategy to increase the cytoplasmic concentrations of pTyr and Pmp allowing the identification of an aaRS that selectively incorporates pTyr or Pmp at desired sites in recombinant proteins in *E. coli* in response to an amber nonsense codon. The molecular basis of synthetase specificity was determined by X-ray crystallography and revealed a substantial reconfiguration of the active site. In addition, we demonstrated the utility of this method by the incorporation of Pmp into a putative phosphorylation site and determined the affinities of the individual variants for their substrate.

RESULTS

Dipeptides increase cellular availability of pTyr and Pmp

Many negatively charged molecules, including pTyr (**1**) and Pmp (**2**) (**Fig. 1a**), are impermeable to the *E. coli* cell wall, resulting in poor cellular bioavailability[25]. On the other hand, it has been shown in both *E. coli*[26,27] and *Caenorhabditis elegans*[28] that poorly bioavailable amino acids can be transported into cells in dipeptide form with an ATP-binding cassette transporter. In *E. coli*, the dipeptide transporter DppA recognizes all 20 canonical amino acids as the N-terminal amino acid, with little restriction toward the C-terminal residue[29,30]. Following uptake, the dipeptide is hydrolyzed by intracellular nonspecific peptidases to afford the free amino acids. To determine whether this strategy could be used to transport pTyr or Pmp into bacteria, we synthesized two dipeptides (**Supplementary Note**), H_2N-lysine-phosphotyrosine-COOH (Lys-pTyr, **3**) and H_2N-lysine-4-phosphomethyl-L-phenylalanine-COOH (Lys-Pmp, **4**), that should be highly soluble and good substrates for DppA (**Fig. 1a**).

The metabolically stable Lys-Pmp was used for bacterial uptake studies, as analysis is not complicated by the activity of endogenous phosphatases. Lys-Pmp and Pmp were incubated with *E. coli* DH10B cultures during exponential growth, and, after repeated washing and subsequent cell lysis, the cytosolic concentrations of Pmp in

[1]Department of Chemistry, The Scripps Research Institute, La Jolla, California, USA. [2]Skaggs Institute for Chemical Biology, The Scripps Research Institute, La Jolla, California, USA. [3]California Institute for Biomedical Research (Calibr), La Jolla, California, USA. [4]Department of Integrative Structural and Computational Biology, The Scripps Research Institute, La Jolla, California, USA. [5]Present address: Department of Chemistry, Temple University, Philadelphia, Pennsylvania, USA. *e-mail: fwang@calibr.org or schultz@scripps.edu

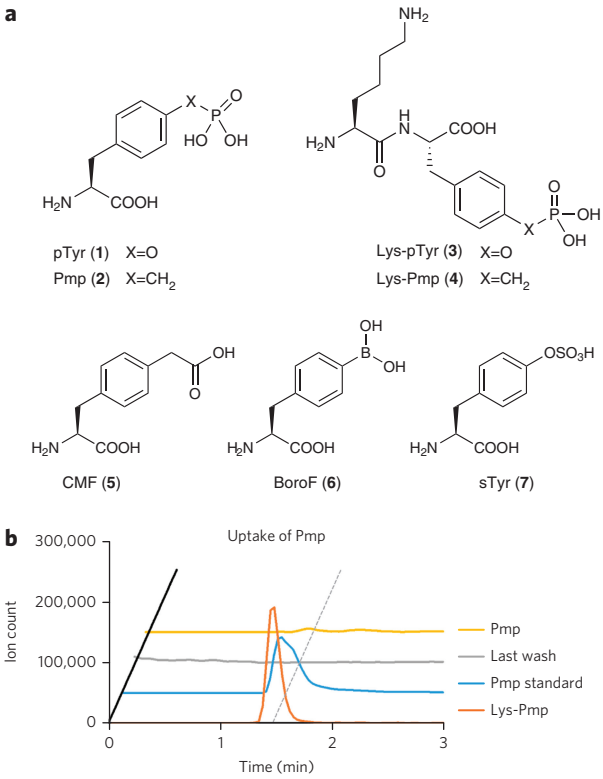

Figure 1 | Propeptide strategy to increase the intracellular concentrations of pTyr and Pmp. (**a**) Structures of ncAAs used in this study. (**b**) The extracted-ion chromatogram for uptake of Lys-Pmp. 5 mM Lys-Pmp (orange) or Pmp (yellow) was added during incubation. Blue, 50 μM Pmp was added to the cell lysate as standard; gray, buffer from last wash before cell lysis.

each sample were determined by LC–MS analysis (the buffer used for the last wash of each sample was also analyzed to exclude the possibility of contamination). Pmp (50 μM) was added to the lysate from cells as a standard for the LC–MS analysis. No signal corresponding to Pmp in cells treated with 5 mM Pmp was observed in the extracted-ion chromatogram, indicating poor uptake of Pmp by cells (**Fig. 1b**). By contrast, the lysate of cells treated with 5 mM Lys-Pmp showed a signal with the same mass and retention time as that of Pmp standard (**Fig. 1b**). Importantly, the peak was absent in the last wash sample standard, indicating that the observed peak is not due to contamination by extracellular material. The area under the curve (AUC) for lysate from cells supplemented with 5 mM Lys-Pmp (AUC = 29,551) was comparable to that of 50 μM Pmp standard (AUC = 27,751). These results confirmed the successful uptake of Lys-Pmp into *E. coli* cytoplasm and encouraged us to pursue the genetic incorporation of pTyr and Pmp into proteins.

Genetic incorporation of pTyr and Pmp in *E. coli*

We next screened three *Methanocaldococcus jannaschii* (*Mj*) tyrosyl-tRNA synthetase (*Mj*TyrRS) mutants for their ability to aminoacylate pTyr and Pmp. These synthetases were originally evolved to incorporate noncanonical amino acids (ncAAs) with negatively charged side chains, including CMF (**5**)[18], *p*-boronophenylalanine (BoroF) (**6**)[31] and sTyr (**7**)[13] (**Fig. 1a**). To assess the suppression efficiency of these aaRSs with Lys-pTyr and Lys-Pmp, a plasmid encoding a myoglobin (Myo) mutant with a permissive amber codon at Lys99 (K99TAG) was transformed into *E. coli* DH10B along with the aaRS–$tRNA_{CUA}$ pair encoded on a pEVOL vector[32]. Only cells transformed with the synthetase for CMF (CMFRS; Y32S, L65A, F108K, Q109H, D158G and L162K) afforded full-length protein after nickel–nitrilotriacetic acid (Ni-NTA) purification in the presence of the ncAAs. Yields were 61 and 60 mg L^{-1} in Terrific Broth (TB) in the presence of 2 mM Lys-pTyr and Lys-Pmp, respectively (**Fig. 2a**); only background tyrosine incorporation into protein (1.4 mg L^{-1}) was observed in the presence of CMFRS and 2 mM pTyr or 2mM Pmp, or in the absence of ncAA (**Fig. 2a**). Mutant proteins were analyzed by SDS–PAGE (**Supplementary Results**, **Supplementary Fig. 1**) and ESI–MS to confirm the incorporation of pTyr or Pmp. Both Myo-K99Pmp and Myo-K99pTyr migrated as a single band at ~18 kDa on an SDS–PAGE gel. However, only Myo-K99Pmp showed the expected mass that would be consistent with its amino acid sequence (calculated and observed, 18,466 Da) (**Fig. 2b**); no background incorporation was observed. In the case of Myo-K99pTyr, only a signal with a mass of 18,386 Da, corresponding to the dephosphorylated product Myo-K99Tyr (calculated, 18,388 Da) (**Supplementary Fig. 2a**), was identified, presumably because of post-translational dephosphorylation of pTyr by endogenous phosphatases in *E. coli*. Indeed, when 1 mM sodium orthovanadate (a relatively nonspecific phosphatase inhibitor) was added during protein expression, a mass signal corresponding to Myo-K99pTyr (calculated, 18,468 Da; observed, 18,465 Da) was observed in addition to the signal corresponding to the dephosphorylated Myo-K99Tyr (calculated, 18,388 Da; observed, 18,387 Da). The apparent ratio of Myo-K99pTyr to Myo-K99Tyr was 1:10 (**Supplementary Fig. 2b**). Finally, site-specific incorporation of Pmp was further confirmed by GluC protease digestion and LC–MS/MS analysis of the Pmp-containing peptide (H_2N-LKPLAQSHATKHPmpIPIKYLE-COOH; **Fig. 2c**). Collectively, these data demonstrate that CMFRS can be readily used to charge pTyr or Pmp onto *Mj*-$tRNA_{CUA}^{Tyr}$, and the propeptide strategy with Lys-pTyr or Lys-Pmp generates sufficient quantities of pTyr or Pmp in the cytosol to allow efficient incorporation of these ncAAs into myoglobin.

Analysis of recombinant Pmp-containing proteins

Next, we used this methodology to generate site-specifically phosphorylated proteins and to then evaluate the effects of these modifications on protein–protein interactions. Phosphotyrosine sites Tyr30 and Tyr52 have been discovered in Abelson murine leukemia viral oncogene homolog 1 (ABL1) SH3 domain. Whereas Tyr52 can be phosphorylated by commonly used kinases, no such kinase has been identified to selectively phosphorylate Tyr30 (ref. 33). Native chemical ligation (NCL) methods have been used to synthesize this domain with pTyr at positions 30 and 52; however, the low efficiency and the complexity of synthesis and purification steps afforded very small quantities of protein. To recombinantly express ABL1 variants containing Pmp, we inserted the CMFRS gene into the pUltra vector, which was reported to exhibit a higher suppression activity[34] (**Supplementary Table 1**). The genes for the ABL1 SH3 domain with an amber codon at position 30 or 52 were synthesized and inserted into the pET-T5 vector, where their expression was controlled by a T5 promoter. After cotransformation of DH10B *E. coli* cells with both plasmids, Abl1-SH3 variants were expressed in liquid TB medium supplemented with 2 mM Lys-Pmp. Full-length Abl1-Pmp30 and Abl1-Pmp52 were purified using Ni-NTA chromatography in yields of 0.75 and 1.0 mg L^{-1}, respectively (**Fig. 3a**). For comparison, wild-type Abl1 (Abl1-WT) was similarly expressed in the absence of ncAA with a purified yield of 7.0 mg L^{-1}. The resulting Abl1 mutants were characterized by SDS–PAGE analysis, where they migrated as a single band. LC–QTOF–MS analysis showed the correct masses for Abl1-WT (calculated and observed, 8,129 Da) as well as for both Abl1-Pmp30 variants (calculated and observed, 8,207 Da) and Abl-Pmp52 (calculated and observed, 8,207 Da) (**Supplementary Fig. 3**). Finally, the site-specific incorporation of Pmp at position 52 was further confirmed by trypsin digestion and LC–MS/MS analysis of the Pmp-containing peptides (H_2N-NGQGWVPSNPmpITPVNSVDHHHHHH-COOH) (**Supplementary Fig. 4**).

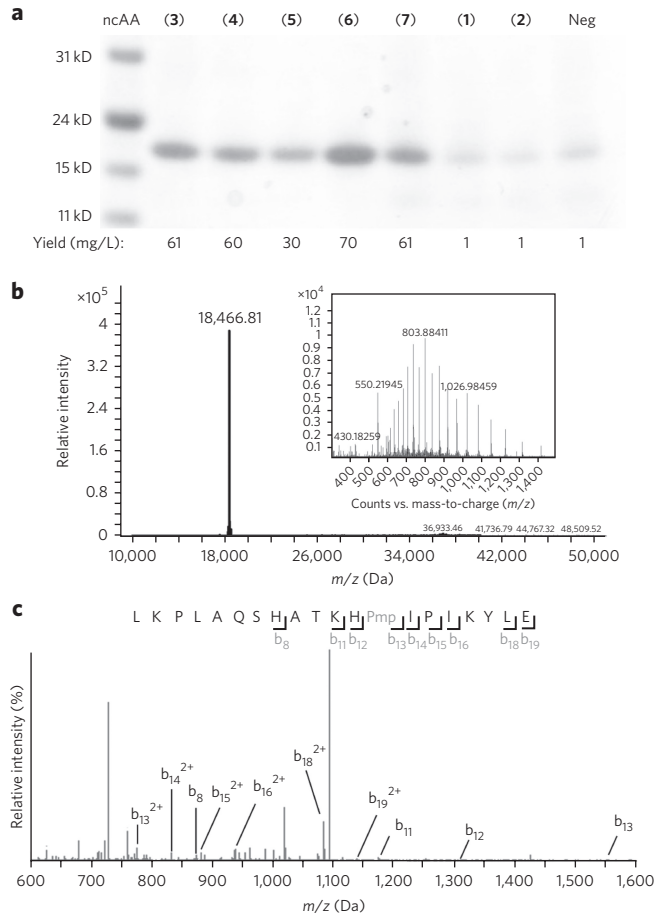

Figure 2 | Incorporation and characterization of variant myoglobins containing different ncAAs. (**a**) SDS–PAGE analysis of Myo-K99TAG expressed with CMFRS in the presence or absence (neg) of various ncAAs or dipeptides and stained with Coomassie blue (full gel image in **Supplementary Fig. 1**). (**b**) Deconvoluted mass spectrum of Myo-K99Pmp; inset, mass spectrum before deconvolution. (**c**) MS/MS fragmentation of GluC-digested peptides derived from Myo-K99Pmp. The spectrum confirmed that Pmp had been incorporated at the desired site.

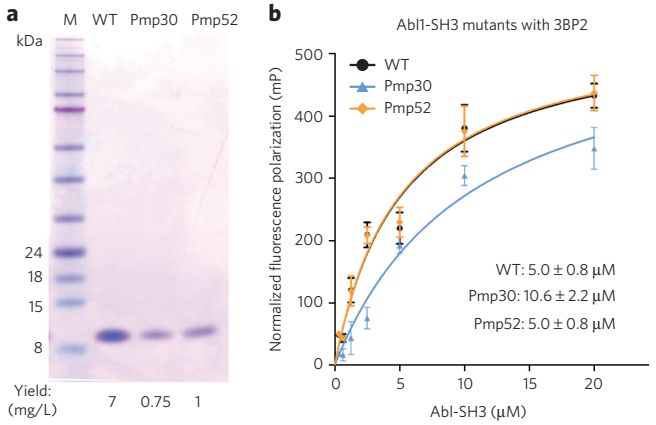

Figure 3 | Expression and characterization of human Abl1 and its Pmp-containing variants in *E. coli*. (**a**) SDS–PAGE analysis of Ni-NTA-purified human Abl1 mutants. Abl1-WT (lane 2) was expressed in the absence of ncAA. Abl1-Pmp30 (lane 3) and Abl1-Pmp52 (lane 4) were expressed in the presence of 2 mM Lys-Pmp. (**b**) Binding affinities of Abl1-WT (black), Abl1-Pmp30 (blue), and Abl1-Pmp52 (orange) to their native substrate 3BP2 peptide were determined by fluorescence polarization. Assays were performed in triplicate and error bars represent the s.d. ± mean. The numbering of tyrosine residues reflects their position in the SH3 domain, as found in the literature.

We next determined the binding affinities of the mutants for the 3BP2 peptide, a native ligand from the adaptor protein 3BP2 (ref. 35), using a reported fluorescence polarization assay[33]. Abl1-Pmp52 exhibited an equilibrium dissociation constant (K_d) of 5.0 ± 0.8 μM, which is identical to that of Abl1-WT; Abl1-Pmp30 exhibited a somewhat reduced binding affinity (K_d = 10.6 ± 2.2 μM) (**Fig. 3b**). These K_d values are in agreement with previous measurements obtained using chemically synthesized peptides containing pTyr (13 μM for Abl1-pTyr30 and 4 μM for Abl1-pTyr52)[33]. These experiments demonstrate that Pmp can be incorporated in response to the TAG codon and can act as a pTyr mimic to probe the roles of protein phosphorylation.

Molecular basis for ncAA incorporation

CMFRS exhibits a unique polyspecificity toward CMF, pTyr and Pmp, which are all ncAAs with negatively charged side chains[36]. This prompted us to investigate the molecular basis for this specificity by X-ray crystallography; the structure of the CMFRS was solved to a resolution of 3.0 Å, with an R_{cryst} of 0.20 and R_{free} of 0.27 (**Fig. 4a**; **Supplementary Table 2**). Although crystals were also obtained in the presence of Pmp, no electron density of the substrate could be observed. The overall structures of CMFRS (Y32S, L65A, F108K, Q109H, D158G and L162K) are aligned well with the wild-type *Mj*TyrRS apo structure (PDB entry 1U7D), with an r.m.s. deviation of 0.6 Å in Cα. However, substantial conformational changes were observed between CMFRS and the wild-type *Mj*TyrRS apo synthetase for helix α8, with an r.m.s. deviation of 3.3 Å in Cα for residues 144–164. The D158G mutation acts as a helix breaker for α8 and alters residues N157–Y162 from a helical structure to a more flexible loop conformation (**Fig. 4b**; **Supplementary Fig. 5**). The B factors for Cα of residues 157–162 range from 66 to 105 Å^2, in contrast to the 52 Å^2 observed for the surrounding residues (**Supplementary Fig. 6**). The glycine at this position has a ϕ of 55.5° and a ψ of 82.0°, angles that are disfavored for any other canonical amino acids. In the *Mj*TyrRS–tyrosine co-crystal structure, the helical structure of N157–L162 is part of the substrate-binding pocket, and D158 forms a hydrogen bond with the hydroxyl group of the substrate tyrosine. It is likely that the D158G mutation and the rearrangement of N157–L162 open up the pocket to accommodate the phosphate and carboxyl groups of the ncAAs. The Y32S and L65A mutations on β2 and β3 likely create additional space on the opposite side of the substrate-binding pocket. The positively charged F108K, Q109H and L162K mutations may form salt bridges with the side chains of negatively charged amino acid substrates, which may explain the selectivity of CMFRS to this series of ncAAs (**Fig. 4c**).

Based on this structural analysis, we hypothesized that CMFRS could potentially aminoacylate other ncAAs with charged side chains. To test this notion, we evaluated BoroF and sTyr by expressing full-length Myo-K99TAG in the presence of 2 mM ncAA. After expression and purification, Myo-K99 variants containing BoroF, sTyr and CMF were produced in yields of 70, 61, and 30 mg L^{-1}, respectively (**Fig. 2b**). MS analysis further confirmed the incorporation of sTyr (calculated and observed, 18,468 Da), pBoroF (calculated, 18,417 Da; observed, 18,416 Da), and CMF (calculated and observed, 18,431 Da) in the protein (**Supplementary Fig. 7**). These results suggest that CMFRS may be a useful polyspecific aaRS for charged amino acids and may be a valuable starting point for the evolution of aaRSs for other amino acids with charged side chains such as reactive pTyr analogs (for example, phosphonodifluoromethyl phenylalanine), which could be useful tools for activity-based protein profiling[37].

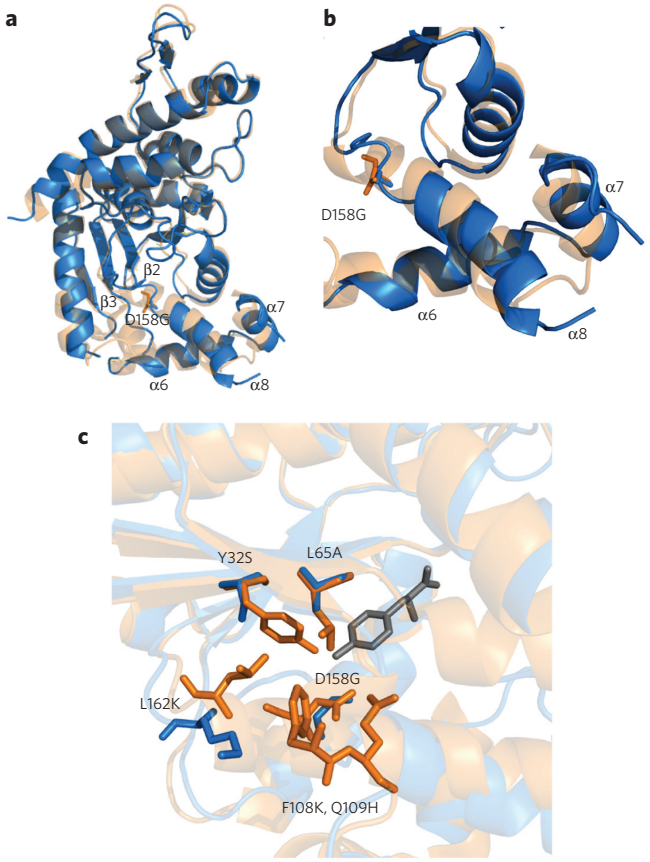

Figure 4 | X-ray structure of CMFRS. (**a**) Superposition of apo wild-type *Mj*TyrRS (orange, PDB: 1U7D) and CMFRS (blue, PDB: 5U36). Secondary structures α6–α8, β2 and β3 are labeled accordingly. (**b**) Comparison of α6–α8 of CMFRS (blue) and apo wild-type *Mj*TyrRS (orange), showing the conformational changes in α8 induced by the D158G mutation. (**c**) Superposition of the active sites of apo wild-type *Mj*TyrRS (orange) and CMFRS (blue); the side chains of mutated residues are shown as sticks. K108 and H109 were disordered and therefore not defined. Tyrosine, the substrate of wild-type *Mj*TyrRS (gray) from the wild-type *Mj*TyrRS–tyrosine complex (PDB: 1J1U), is shown by superposition to illustrate the spatial relationship between the mutant sites and the substrate.

DISCUSSION

We have reported the site-specific incorporation of pTyr and its nonhydrolyzable analog Pmp into recombinantly expressed proteins in *E. coli*. Key to this advance is a propeptide strategy that increases cellular uptake of ncAAs and the use of the polyspecific synthetase CMFRS. It is worth noting that, unlike Pmp or pTyr, other ncAAs with charged side chains (i.e., sTyr, BoroF, and CMF) do not require a propeptide strategy to get into cells. This may be because they may share transporters with other amino acids or small molecules, as demonstrated for other ncAAs[38]. It has also been reported that tRNAs carrying negatively charged amino acids are poor substrates for elongation factor (EF)-Tu. For example, the incorporation of phosphoserine (pSer) was only observed in the presence of an engineered EF-Tu[39]. However, the wild-type EF-Tu in *E. coli* seems adequate for the incorporation of Pmp with good efficiency, although it cannot be ruled out that engineering of EF-Tu may further improve efficiency. Efforts to increase the cellular availability of pTyr by deleting the endogenous phosphatases in *E. coli*[40] gave relatively low yields of protein containing pTyr. Nevertheless, it is very likely that the combination of these strategies could lead to increased yields of variant proteins.

This method enables the generation of homogeneous proteins containing nonhydrolyzable pTyr analogs at different tyrosine phosphorylation sites in a controlled fashion, thereby facilitating studies of the function of natural phosphotyrosine post-translational modifications in a variety of protein sizes. In a proof-of-concept study, Pmp was site-specifically incorporated into Abl1 at different tyrosine sites to afford the nonhydrolyzable analogs of the tyrosine-phosphorylated Abl1. The binding affinities of Abl1-Pmp30 and Abl1-Pmp52 were comparable to those of their pTyr counterparts. However, it should be noted that, in some cases, five-fold lower affinity has been observed for Pmp-containing peptides compared to the corresponding pTyr-containing peptides, likely due to the higher pK_a2 of the phosphonate group (7.1 compared to 5.7 for phosphate) and the loss of a hydrogen bond between the phenolic oxygen atom and the binding partner.

In addition, this strategy allows construction of libraries of phosphorylated peptides or proteins for phage display, which was previously inaccessible by other approaches. Moreover, this method could potentially be expanded to encode a set of pTyr-containing proteins or even a set of pTyr- and pSer-containing proteins when coupled with genetic incorporation of pSer in a desired pattern. The crystal structure of CMFRS sheds light on the molecular basis of the polyspecificity toward ncAAs with charged side chains; this may enable the development of *Mj*TyrRS mutants to incorporate other valuable charged or hydrophilic ncAAs. Finally, we are attempting to adapt a similar strategy for eukaryotic cells to allow the controlled expression of tyrosine-phosphorylated proteins in real time.

Received 17 November 2016; accepted 14 March 2017;
published online 12 June 2017

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

References

1. Manning, G., Whyte, D.B., Martinez, R., Hunter, T. & Sudarsanam, S. The protein kinase complement of the human genome. *Science* **298**, 1912–1934 (2002).
2. Pawson, T. Specificity in signal transduction: from phosphotyrosine-SH2 domain interactions to complex cellular systems. *Cell* **116**, 191–203 (2004).
3. Hunter, T. & Cooper, J.A. Protein-tyrosine kinases. *Annu. Rev. Biochem.* **54**, 897–930 (1985).
4. Alonso, A. *et al.* Protein tyrosine phosphatases in the human genome. *Cell* **117**, 699–711 (2004).
5. Eckhart, W., Hutchinson, M.A. & Hunter, T. An activity phosphorylating tyrosine in polyoma T antigen immunoprecipitates. *Cell* **18**, 925–933 (1979).
6. Tailor, P., Gilman, J., Williams, S., Couture, C. & Mustelin, T. Regulation of the low molecular weight phosphotyrosine phosphatase by phosphorylation at tyrosines 131 and 132. *J. Biol. Chem.* **272**, 5371–5374 (1997).
7. Feng, G.S., Hui, C.C. & Pawson, T. SH2-containing phosphotyrosine phosphatase as a target of protein-tyrosine kinases. *Science* **259**, 1607–1611 (1993).
8. Tarrant, M.K. *et al.* Regulation of CK2 by phosphorylation and O-GlcNAcylation revealed by semisynthesis. *Nat. Chem. Biol.* **8**, 262–269 (2012).
9. Chen, Z. & Cole, P.A. Synthetic approaches to protein phosphorylation. *Curr. Opin. Chem. Biol.* **28**, 115–122 (2015).
10. Serwa, R. *et al.* Chemoselective Staudinger-phosphite reaction of azides for the phosphorylation of proteins. *Angew. Chem. Int. Ed. Engl.* **48**, 8234–8239 (2009).
11. Olsen, J.V. *et al.* Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell* **127**, 635–648 (2006).
12. Arslan, T., Mamaev, S.V., Mamaeva, N.V. & Hecht, S.M. Structurally modified firefly luciferase. Effects of amino acid substitution at position 286. *J. Am. Chem. Soc.* **119**, 10877–10887 (1997).
13. Liu, C.C. & Schultz, P.G. Recombinant expression of selectively sulfated proteins in *Escherichia coli*. *Nat. Biotechnol.* **24**, 1436–1440 (2006).
14. Liu, C.C. *et al.* Protein evolution with an expanded genetic code. *Proc. Natl. Acad. Sci. USA* **105**, 17688–17693 (2008).
15. Liu, C.C., Choe, H., Farzan, M., Smider, V.V. & Schultz, P.G. Mutagenesis and evolution of sulfated antibodies using an expanded genetic code. *Biochemistry* **48**, 8891–8898 (2009).

16. Liu, C.C., Cellitti, S.E., Geierstanger, B.H. & Schultz, P.G. Efficient expression of tyrosine-sulfated proteins in *E. coli* using an expanded genetic code. *Nat. Protoc.* **4**, 1784–1789 (2009).
17. Liu, C.C., Brustad, E., Liu, W. & Schultz, P.G. Crystal structure of a biosynthetic sulfo-hirudin complexed to thrombin. *J. Am. Chem. Soc.* **129**, 10648–10649 (2007).
18. Xie, J., Supekova, L. & Schultz, P.G. A genetically encoded metabolically stable analogue of phosphotyrosine in *Escherichia coli*. *ACS Chem. Biol.* **2**, 474–478 (2007).
19. Rust, H.L. *et al.* Using unnatural amino acid mutagenesis to probe the regulation of PRMT1. *ACS Chem. Biol.* **9**, 649–655 (2014).
20. Guerra-Castellano, A. *et al.* Mimicking tyrosine phosphorylation in human cytochrome c by the evolved tRNA synthetase technique. *Chemistry* **21**, 15004–15012 (2015).
21. Zheng, Y., Lv, X. & Wang, J. A genetically encoded sulfotyrosine for VHR function research. *Protein Cell* **4**, 731–734 (2013).
22. Lu, W., Shen, K. & Cole, P.A. Chemical dissection of the effects of tyrosine phosphorylation of SHP-2. *Biochemistry* **42**, 5461–5468 (2003).
23. Zhang, Z., Shen, K., Lu, W. & Cole, P.A. The role of C-terminal tyrosine phosphorylation in the regulation of SHP-1 explored via expressed protein ligation. *J. Biol. Chem.* **278**, 4668–4674 (2003).
24. Lu, W., Gong, D., Bar-Sagi, D. & Cole, P.A. Site-specific incorporation of a phosphotyrosine mimetic reveals a role for tyrosine phosphorylation of SHP-2 in cell signaling. *Mol. Cell* **8**, 759–769 (2001).
25. Leive, L. A nonspecific increase in permeability in *Escherichia coli* produced by EDTA. *Proc. Natl. Acad. Sci. USA* **53**, 745–750 (1965).
26. Fickel, T.E. & Gilvarg, C. Transport of impermeant substances in *E. coli* by way of oligopeptide permease. *Nat. New Biol.* **241**, 161–163 (1973).
27. Ames, B.N., Ames, G.F., Young, J.D., Tsuchiya, D. & Lecocq, J. Illicit transport: the oligopeptide permease. *Proc. Natl. Acad. Sci. USA* **70**, 456–458 (1973).
28. Parrish, A.R. *et al.* Expanding the genetic code of *Caenorhabditis elegans* using bacterial aminoacyl-tRNA synthetase/tRNA pairs. *ACS Chem. Biol.* **7**, 1292–1302 (2012).
29. Payne, J.W. & Gilvarg, C. The role of the terminal carboxyl group on peptide transport in *Escherichia coli*. *J. Biol. Chem.* **243**, 335–340 (1968).
30. Verkamp, E., Backman, V.M., Björnsson, J.M., Söll, D. & Eggertsson, G. The periplasmic dipeptide permease system transports 5-aminolevulinic acid in *Escherichia coli*. *J. Bacteriol.* **175**, 1452–1456 (1993).
31. Brustad, E. *et al.* A genetically encoded boronate-containing amino acid. *Angew. Chem. Int. Ed. Engl.* **47**, 8220–8223 (2008).
32. Young, T.S., Ahmad, I., Yin, J.A. & Schultz, P.G. An enhanced system for unnatural amino acid mutagenesis in *E. coli*. *J. Mol. Biol.* **395**, 361–374 (2010).
33. Zitterbart, R. & Seitz, O. Parallel chemical protein synthesis on a surface enables the rapid analysis of the phosphoregulation of SH3 domains. *Angew. Chem. Int. Ed. Engl.* **55**, 7252–7256 (2016).
34. Chatterjee, A., Sun, S.B., Furman, J.L., Xiao, H. & Schultz, P.G. A versatile platform for single- and multiple-unnatural amino acid mutagenesis in *Escherichia coli*. *Biochemistry* **52**, 1828–1837 (2013).
35. Viguera, A.R., Arrondo, J.L.R., Musacchio, A., Saraste, M. & Serrano, L. Characterization of the interaction of natural proline-rich peptides with five different SH3 domains. *Biochemistry* **33**, 10925–10933 (1994).
36. Young, D.D. *et al.* An evolved aminoacyl-tRNA synthetase with atypical polysubstrate specificity. *Biochemistry* **50**, 1894–1900 (2011).
37. H. S. Lu, C., Liu, K., Tan, L.P. & Yao, S.Q. Current chemical biology tools for studying protein phosphorylation and dephosphorylation. *Chemistry* **18**, 28–39 (2012).
38. Liu, D.R. & Schultz, P.G. Progress toward the evolution of an organism with an expanded genetic code. *Proc. Natl. Acad. Sci. USA* **96**, 4780–4785 (1999).
39. Park, H.-S. *et al.* Expanding the genetic code of *Escherichia coli* with phosphoserine. *Science* **333**, 1151–1154 (2011).
40. Fan, C., Ip, K. & Söll, D. Expanding the genetic code of *Escherichia coli* with phosphotyrosine. *FEBS Lett.* **590**, 3040–3047 (2016).

Acknowledgments

The authors acknowledge K. Williams for the assistance in manuscript preparation. X-ray diffraction data were collected at the Advanced Photon Source (APS) beamline 23ID-B. Use of the Advanced Photon Source for data collection was supported by the DOE, Basic Energy Sciences, Office of Science, under contract no. DE-AC02- 06CH11357. GM/CA CAT has been funded in whole or in part with federal funds from NCI (grant Y1-CO-1020) and NIGMS (grant Y1-GM-1104). The NIH and DOE funders at the beamlines had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. This work was supported by NIH Grant 5R01 GM062159-14 (to P.G.S). This is manuscript 29424 of The Scripps Research Institute.

Author contributions

X. Luo, P.G.S., and F.W. designed the research. X. Luo, G.F., and R.E.W. performed protein expression, purification, and crystallization. X. Luo, R.E.W., C.Z., R.L., W.X., C.H. and P.-Y.Y. performed chemical synthesis. X. Luo, T.L., J.D., M.K., and Y.Z. performed the cloning and screening of synthetases, expression of target proteins. X.Z. performed X-ray diffraction experiments.; X. Luo, G.F., X.Z., X. Lyu., I.A.W. and F.W. performed crystallographic analysis and data deposition. X. Luo, H.G., and A.Y. performed fluorescence polarization assay. X. Luo, T.L., W.X., P.G.S. and F.W. analyzed the data; and X. Luo, S.A.R., P.G.S. and F.W. wrote the paper.

Competing financial interests

The authors declare no competing financial interests.

Additional information

Any supplementary information, chemical compound information and source data are available in the online version of the paper. Reprints and permissions information is available online at http://www.nature.com/reprints/index.html. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to P.G.S. or F.W.

ONLINE METHODS

Synthesis of dipeptides. Chemical synthesis of dipeptides is described in the **Supplementary Note**.

General information. *E. coli* DH10B (Invitrogen, Carlsbad, CA; Catalog number 18290015) was used for general cloning and propagation. DPBS and Tris, borate and EDTA buffers were obtained from Cellgro. LB medium and LB agar were purchased from Fisher Scientific. IPTG was purchased from Anatrace, and 4–12% (wt/vol) Bis–Tris gels for SDS–PAGE were purchased from Invitrogen. Q5 DNA polymerase kit, Q5 Site-Directed Mutagenesis Kit, restriction enzymes and T4 DNA ligase were obtained from New England BioLabs, and oligonucleotides were purchased from Integrated DNA Technologies (San Diego, CA). Plasmid DNA preparation was carried out with the QIAprep Spin Miniprep Kit (Qiagen). Unless otherwise mentioned, all other chemicals were purchased from Sigma-Aldrich and used without further purification. Protein concentrations were measured using a Coomassie Plus (Bradford) Protein Assay kit from Pierce. All peptides were purchased from Innopep (San Diego, CA)

Agilent 6520 accurate-mass quadrupole-time-of-flight (QTOF) instrument, which was equipped with reverse phase liquid chromatography and an ESI source, was used to carry out the high-resolution MS for all protein samples. Samples in PBS were injected (10 µl) at a concentration of 0.2 mg/ml and separated on a 150 mm reverse phase C8 wide pore column heated to 70 °C to improve peak resolution. Proteins were eluted in a gradient of H_2O + 0.1% formic acid (solvent A) and acetonitrile + 0.1% formic acid (solvent B) using the following method: 5% B for 2 min, 5–60% B for 10 min, 60–80% B for 1 min, followed by a wash (95% B) and re-equilibration (5% A) phase. ESI source settings were 350 °C, 10 l/min, 40 psig nebulizer nitrogen gas, 200 V fragmentor, and 4,500 V capillary. Figures were created by extraction of the total ion count across the entire area of protein elution (roughly 8–10 min) and deconvolution of charge envelopes using Agilent Qualitative Analysis software with BioConfirm.

Amino acid uptake assay. An overnight culture of DH10B *E. coli* in GMML medium was diluted 1:100 in GMML and was grown to saturation (~18 h). 1 ml stocks of ncAAs (25 mM in GMML) were added to 4 ml diluted cultures at an optical density of 600 nm (OD_{600}) 0.8 to make a final concentration of 5 mM. 1 ml of *p*-acetylphenyalanine (pAcF) was used as positive control for the method and 1 ml of GMML as negative control. The mixtures were incubated at 37 °C at 250 r.p.m. for 2 h. The cells were pelleted by centrifugation 13,000 × *g* for 5 min. The media supernatants were saved and the pellets were washed four times with 1 ml ice-cold GMML lacking heavy metals, thiamine and biotin (GMML minus) at 4 °C to minimize leaching out of ncAAs from cells. The last washes were saved and the cell pellets were resuspended in 150 µl lysis buffer (1 mg/ml lysozyme, 5 µg/ml DNAse I and 10 µg/ml RNAse in water) and incubated at 37 °C at 250 r.p.m. for 1 h before being centrifuged at 13,000 × *g* for 20 min at 4 °C. The cell lysates were filtered and subjected to LC–MS analysis together with the saved media and last wash.

The LC–MS analysis was performed on an Agilent 1100 Series LC/MSD instrument. The chromatographic peak corresponding to mass of each ncAAs was extracted and processed using Agilent LC/MSD ChemStation (revision B.03.02).

Myoglobin expression and purification. An overnight culture of DH10B *E. coli* cotransformed with the appropriate pEVOL plasmids (pEVOL-CMF, pEVOL-sTyr and pEVOL-BoroF) and pBAD-Myo(K99TAG) in 2YT medium at 37 °C with chloramphenicol (50 µg/ml) and ampicillin (100 µg/ml) was diluted 1:50 in Terrific Broth (TB) supplemented with chloramphenicol (50 µg/ml) and ampicillin (100 µg/ml) in the presence or absence of ncAAs (the final concentration for sTyr and CMF is 1 mM for Pmp and pTyr and 2 mM for their corresponding dipeptides). The cultures were grown at 37 °C to an OD_{600} of 0.8 when arabinose was added to a final concentration of 0.2% to induce the protein expression. The cultures were grown in 30 °C for 18 h and were pelleted and lysed with BugBuster (Novagen), and the proteins were purified on Ni-NTA spin column (Qiagen) according to manufacturer's protocol.

Construction of pUltra-CMF and pET-T5-ABL1 and its mutants. CMFRS gene without stop codon was amplified from pEVOL-CMFRS[18] with primers PMP001 and PMP002 (**Supplementary Table 1**), and inserted into the pUltra-MjTyrRS vector[34] digested with NotI by Gibson assembly to afford pUltra-CMF. Two additional amino acids, Thr and Glu, were inserted at the Abl1 amino terminus between the first Met and second Asn to increase its stability in *E. coli* during expression while a His_6 tag was added to the C terminus for affinity purification. The gene was synthesized by IDT and amplified with primers PMP003 and PMP004, and was inserted into a pET-T5 backbone amplified with primers PMP005 and PMP006 using Gibson assembly to provide pET-T5-ABL1-wt. Site-directed mutagenesis of this plasmid with primers PMP007/PMP008 and PMP009/PMP010 using Q5 Site-Directed Mutagenesis Kit (NEB) according to manufacturer's protocol resulted in pET-T5-ABL1-Y30TAG and pET-T5-ABL1-Y52TAG, which was used to express Abl1 mutants containing Pmp (**Supplementary Table 1**).

Construction of CMFRS expression plasmid. The CMFRS gene without stop codon was amplified from pEVOL-CMFRS[18] with primers PMP011 and PMP012 (**Supplementary Table 1**), and inserted into the pET-22b(+) (Novagen, catalog number 69744-3) using NdeI and XhoI restriction sites to provide pET-CMFRS with a His6 tag engineered at C terminus of the CMFRS. The plasmid was electroporated into *E. coli* BL21(DE3)-competent cell.

CMFRS expression and purification. The transformed *E. coli* BL21(DE3) was plated on an LB plate containing ampicillin (100µg/ml) and allowed to grow overnight at 37 °C. A single colony was picked to inoculate 50 ml LB medium and cultured overnight. 50 ml bacterial culture was then centrifuged and the bacteria were resuspended and then transferred into 1 L fresh LB medium. All LB media used contained ampicillin (100 µg/ml). The cells were induced with 1 mM IPTG when the culture reached an OD_{600} of 0.6–0.8. After induction, cultivation was continued for 3.5 h at 37 °C. The cells were harvested by centrifugation at 6,000*g* at 4 °C for 15 min. The cell pellets obtained were stored at –20 °C for subsequent use. The purification was completed using a Ni–NTA affinity column (Qiagen) and an ÄKTA purification system (GE Healthcare). The cell pellets were resuspended in 30ml lysis buffer (20 mM Tris–HCl pH 8.0, 500 mM NaCl, 10 mM imidazole) with 0.1 mM PMSF and the cells were lysed by ultrasonication. The lysate was centrifuged at 24,000*g* at 4 °C for 40 min. The supernatant was loaded directly onto a Ni–NTA column, which was pre-equilibrated with lysis buffer. The column was washed with wash buffer (20 mM Tris–HCl pH 8.0, 500 mM NaCl, 20 mM imidazole). CMFRS was eluted with elution buffer (20 mM Tris–HCl pH 8.0, 500 mM NaCl, 250 mM imidazole). The eluate was concentrated by ultrafiltration (Millipore) at 4 °C and buffer changed to 20 mM Tris–HCl, pH 8.0. Then, the CMFRS solution was loaded onto a Mono Q 5/50 GL column (GE Healthcare) for ion exchange purification. As running the gradient of 20 mM Tris–HCl, 1 M NaCl, the column flow rate was 0.5 $mlmin^{-1}$ and all peak fractions were collected. All fractions were identified by SDS–PAGE. CMFRS was eluted at the NaCl concentration of 50–200 mM. The eluant was concentrated and the buffer was exchanged to 20 mM Tris–HCl, pH 8.0, 50 mM NaCl, 10 mM Pmp or sTyr, 5 mM DTT. The final CMFRS concentration was 15 mg/ml, and protein concentrations were determined by Nanodrop 2000c.

Crystallization of CMFRS. The crystallization experiments were conducted using the hanging-drop vapor-diffusion method at 289K. The crystallization conditions were optimized by varying the type and concentration of precipitants (PEGs and glycerol), the pH, the types of additives (Hampton Additive kit) and the protein concentration. As the result of several rounds of optimization, crystals of CMFRS were obtained by mixing equal volumes of protein solution (32 mg/ml) and crystallization solution containing 15–25% PEG 3350, 3–7% glycerol, 10 mM Pmp or sTyr, 5 mM DTT, 50mM sodium cacodylate trihydrate pH 6.0–7.0 and equilibrating over 0.5ml reservoir solution (15–25% PEG 3350, 50mM sodium cacodylate trihydrate pH 6.0–7.0) at 277K for about a week. Prior to data collection, CMFRS crystals were transferred to cryoprotectant solution containing 15% ethylene glycol, 16% PEG 3350, 10 mM Pmp or sTyr, 5 mM DTT, 50mM sodium cacodylate trihydrate

doi:10.1038/nchembio.2405

pH 6.0–7.0 for about 2min; they were then immediately flash cooled in a liquid-nitrogen stream on cryoloops.

Abl1 mutant expression and purification. For wild-type Abl1, an overnight culture of DH10B *E. coli* transformed with pET-T5-ABL1-wt in 2YT media with ampicillin (100 μg/ml) was diluted 100-fold into 200 ml TB media supplemented with ampicillin (100 μg/ml) at 37 °C. For Abl1 containing amber codon, an overnight culture of DH10B *E. coli* cotransformed with pUltra-CMFRS and appropriate pET-T5-ABL1-Y30TAG or -Y52TAG in 2YT media with ampicillin (100 μg/ml) and spectinomycin (50 μg/ml) was diluted 100-fold into 200 ml TB media supplemented with ampicillin (100 μg/ml) and spectinomycin (50 μg/ml). The cells were allowed to grow for 3–5 h when the OD_{600} reached 0.6 and IPTG was added to a final concentration of 1 mM to induce protein expression. The cells were grown for an addition 30 h at 30 °C before harvesting by centrifugation at 6,000*g* for 10 min. The cell pellet was lysed by sonication and the resulting cell lysate was clarified by centrifugation at 15,500*g* for 35 min at 4 °C. Abl1 wild-type and mutants were purified on Ni-NTA resin (Qiagen) following the manufacturer's instructions.

Data collection, structure determination and refinement. Diffraction data for CMFRS were collected at the wavelength of 1.03320Å on beamline 23-IDB APS at 100 K and HKL2000 (ref. 41) (HKL research) was used for the processing, reduction and scaling of the diffraction data.

The data collection statistics of CMFRS crystals are shown in **Supplementary Table 2**. The crystal diffracted to 3.0 Å resolution. The diffraction data were indexed in space group $P2_1$ with unit-cell parameters for the native crystal of a = 58.6 Å, b = 130.6, c = 60.0 Å, α = 90°, β = 117.7° and γ = 90°. Two molecules were found in the asymmetric unit with a solvent content of 57.0%.

The structure was determined by molecular replacement using PDB 4PBR as a model and the program Phaser[42] in CCP4. A preliminary model with about 60% completeness was fitted to the electron density. The rest of the residues were manually modeled using Coot[43]. The refinement was carried out using Refmac[44,45] in the CCP4 suite and Phenix[46].The statistics of data collection and refinement are summarized in **Supplementary Table 2**. All structural pictures were rendered in PyMOL (The PyMOL Molecular Graphics System, Version 1.8 Schrödinger, LLC.). Ramachandran statistics: favored, 94.8%; outliers, 1.4%).

Fluorescence polarization binding assay. The fluorescence polarization assays were performed by using PerkinElmer black 96-well flat-bottom plates and the fluorescence polarization was determined by the SpectraMax M5 (Molecular Devices) plate reader. To each well, FAM-labeled peptide in PBS was added to a final concentration of 5 nM, the recombinant Abl1-wt or mutants were added in two-fold dilution from 20 mM concentration. PBS was used to fill up each well to 100 μl. The plate was incubated for 15 min after which the polarization values were determined in millipolarization units (mP) with excitation wavelength of 485 nm and emission wavelength of 535 nm. The acquired data were analyzed by using Prism (GraphPad Software, Inc., San Diego, CA, USA) and the K_d values were determined by nonlinear fitting.

Data availability. Coordinates and structure factors have been deposited in the Protein Data Bank under accession codes 5U36 for CMFRS. All other data that support the findings of this study are in the published article (and its supplementary information files) or are available from the corresponding author upon reasonable request.

41. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276**, 307–326 (1997).
42. McCoy, A.J. *et al.* Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658–674 (2007).
43. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D Biol. Crystallogr.* **60**, 2126–2132 (2004).
44. Vagin, A.A. *et al.* REFMAC5 dictionary: organization of prior chemical knowledge and guidelines for its use. *Acta Crystallogr. D Biol. Crystallogr.* **60**, 2184–2195 (2004).
45. Murshudov, G.N., Vagin, A.A. & Dodson, E.J. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr. D Biol. Crystallogr.* **53**, 240–255 (1997).
46. Adams, P.D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 213–221 (2010).