Global Analysis of Cdk1 Substrate Phosphorylation Sites Provides Insights into Evolution

Liam J. Holt, *et al.*

Science **325**, 1682 (2009);
DOI: 10.1126/science.1172867

Table 1. IC_{50} values for the inhibition of $PfCRT^{CQR}$-mediated CQ transport by a number of drugs and peptides. $PfCRT^{CQR}$-mediated CQ transport was calculated by subtracting the uptake measured in oocytes expressing $PfCRT^{CQS}$ from that in oocytes expressing $PfCRT^{CQR}$. The data are shown in fig. S7 and Fig. 2F. IC_{50} values were derived by least-squares fit of the equation $Y = Y_{min} + [(Y_{max} - Y_{min})/(1 + ([inhibitor]/IC_{50})^C]$, where Y is $PfCRT^{CQR}$-mediated CQ transport, Y_{min} and Y_{max} are the minimum and maximum values of Y, and C is a constant. All values are mean ± SEM from $n = 3$ or 4 separate experiments, within which measurements were made from 10 oocytes per treatment.

Compound	IC_{50} (μM)
Verapamil	30 ± 3 (n = 4)
Quinine	48 ± 6 (n = 3)
Amantadine	770 ± 63 (n = 3)
YPWF-NH_2 (endomorphin-1)	64 ± 5 (n = 3)
WHWLQL (α_1-mating factor peptide)	99 ± 9 (n = 3)
VVYPWTQR (a hemoglobin peptide)	363 ± 28 (n = 4)
RPPGFSPFR (Bradykinin)	1095 ± 71 (n = 3)

ically effective against both CQS and CQR parasites) had no effect. Amantadine exhibits some antimalarial activity in vitro, particularly against CQR parasites (*16*), and also inhibited transport via $PfCRT^{CQR}$ (table S5).

Several peptides were found to cause a pronounced inhibition of CQ transport via $PfCRT^{CQR}$ (table S5). Most of the peptides that are active against $PfCRT^{CQR}$ have key elements of the CQ-resistance reverser pharmacophore [hydrogen bond acceptor and two hydrophobic aromatic rings (*17*)] (table S6). This pharmacophore can be viewed as defining the basic elements involved in interactions between $PfCRT^{CQR}$ and substrates or inhibitors.

The concentration dependence of inhibition of CQ transport was determined for a number of compounds (Table 1 and fig. S7). YPWF-NH_2 (endomorphin-1; an opioid receptor agonist) was the most effective peptide inhibitor of $PfCRT^{CQR}$-mediated CQ uptake, with an IC_{50} comparable to that of quinine and verapamil. Measurements of [^{3}H]YPWF-NH_2 uptake in oocytes expressing different PfCRT constructs revealed that $PfCRT^{CQR}$, but not $PfCRT^{CQR}$-T76K, $PfCRT^{CQR}$-S163R, or $PfCRT^{CQS}$, mediates the transport of this peptide (fig. S8).

We have demonstrated the transport of CQ via mutant PfCRT, which provides an explanation for the phenomenon of CQ resistance, as well as for the reversal of CQ resistance by reversing agents such as verapamil. The presence of a positive charge (K76 or R163) in the PfCRT substrate-binding site prevents CQH_2^{2+} (or CQH^+) from interacting with the transporter. The K76T mutation removes the positive charge, altering the substrate specificity of PfCRT to allow the transport of the protonated drug. In the parasite, the presence of mutant PfCRT on the digestive vacuole will allow the protonated drug to be transported down its electrochemical gradient, out of the vacuole, and thus away from its site of action (fig. S9). This mechanism is consistent with recent studies implicating $PfCRT^{CQR}$ in the transport of [^{3}H]CQ in CQR parasites (*18*–*20*) and in *Dictyostelium discoideum* transformants expressing PfCRT at endosomal membranes (*21*). It is also consistent with the recent demonstration of a (verapamil-sensitive) CQ-mediated efflux of H^+ from the digestive vacuole of CQR parasites (*22*). The achievement of a robust expression system for PfCRT has the potential to facilitate the rational design of novel CQ-like drugs that bypass the resistance mechanism and/or the design of clinically effective resistance-reversing agents.

References and Notes

1. D. A. Fidock *et al.*, *Mol. Cell* **6**, 861 (2000).
2. A. B. Sidhu, D. Verdier-Pinard, D. A. Fidock, *Science* **298**, 210 (2002).
3. C. D. Fitch, *Proc. Natl. Acad. Sci. U.S.A.* **64**, 1181 (1969).
4. D. J. Krogstad *et al.*, *Science* **238**, 1283 (1987).
5. R. Hayward, K. J. Saliba, K. Kirk, *J. Cell Sci.* **119**, 1016 (2006).
6. N. Klonis *et al.*, *Biochem. J.* **407**, 343 (2007).
7. Y. Kuhn, P. Rohrbach, M. Lanzer, *Cell. Microbiol.* **9**, 1004 (2007).
8. T. N. Bennett *et al.*, *Mol. Biochem. Parasitol.* **133**, 99 (2004).
9. V. Lakshmanan *et al.*, *EMBO J.* **24**, 2294 (2005).
10. R. E. Martin, K. Kirk, *Mol. Biol. Evol.* **21**, 1938 (2004).
11. P. G. Bray *et al.*, *Mol. Microbiol.* **56**, 323 (2005).
12. P. Cortazzo *et al.*, *Biochem. Biophys. Res. Commun.* **293**, 537 (2002).
13. A. A. Komar, T. Lesnik, C. Reiss, *FEBS Lett.* **462**, 387 (1999).
14. Materials and methods are available as supporting material on *Science* Online.
15. Single-letter abbreviations for the amino acid residues are as follows: F, Phe; G, Gly; H, His; K, Lys; L, Leu; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr.
16. D. J. Johnson *et al.*, *Mol. Cell* **15**, 867 (2004).
17. D. A. van Schalkwyk, T. J. Egan, *Drug Resist. Updat.* **9**, 211 (2006).
18. P. G. Bray *et al.*, *Mol. Microbiol.* **62**, 238 (2006).
19. C. P. Sanchez *et al.*, *Biochemistry* **44**, 9862 (2005).
20. C. P. Sanchez *et al.*, *Mol. Microbiol.* **64**, 407 (2007).
21. B. Naude, J. A. Brzostowski, A. R. Kimmel, T. E. Wellems, *J. Biol. Chem.* **280**, 25596 (2005).
22. A. M. Lehane, K. Kirk, *Antimicrob. Agents Chemother.* **52**, 4374 (2008).
23. C. A. Wagner, B. Friedrich, I. Setiawan, F. Lang, S. Bröer, *Cell. Physiol. Biochem.* **10**, 1 (2000).

24. We thank J. Abbey, R. Summers, E. Baker, and R. Slatyer for technical assistance. This work was supported by the Australian National Health and Medical Research Council (NHMRC) (grant 471472) and the Australian Research Council (grant DP0559433). R.E.M. was supported by an NHMRC Australian Biomedical Fellowship (fellowship 520320).

Supporting Online Material

www.sciencemag.org/cgi/content/full/325/5948/1680/DC1
Materials and Methods
Figs. S1 to S9
Tables S1 to S6
References

30 April 2009; accepted 3 August 2009
10.1126/science.1175667

Global Analysis of Cdk1 Substrate Phosphorylation Sites Provides Insights into Evolution

Liam J. Holt,[1]* Brian B. Tuch,[2]* Judit Villén,[3]* Alexander D. Johnson,[2] Steven P. Gygi,[3]† David O. Morgan[1]†

To explore the mechanisms and evolution of cell-cycle control, we analyzed the position and conservation of large numbers of phosphorylation sites for the cyclin-dependent kinase Cdk1 in the budding yeast *Saccharomyces cerevisiae*. We combined specific chemical inhibition of Cdk1 with quantitative mass spectrometry to identify the positions of 547 phosphorylation sites on 308 Cdk1 substrates in vivo. Comparisons of these substrates with orthologs throughout the ascomycete lineage revealed that the position of most phosphorylation sites is not conserved in evolution; instead, clusters of sites shift position in rapidly evolving disordered regions. We propose that the regulation of protein function by phosphorylation often depends on simple nonspecific mechanisms that disrupt or enhance protein-protein interactions. The gain or loss of phosphorylation sites in rapidly evolving regions could facilitate the evolution of kinase-signaling circuits.

Cyclin-dependent kinases (Cdks) drive the major events of the eukaryotic cell-division cycle (*1*). Comprehensive identification and analysis of Cdk substrates would enhance our understanding of cell-cycle control and provide insights into the mechanisms and evolution of regulation by phosphorylation. We therefore developed methods for comprehensive identification of the sites of Cdk1 phosphorylation on large numbers of substrates in vivo. We

used quantitative mass spectrometry to identify sites at which phosphorylation decreased in vivo after specific inhibition of Cdk1 (*2*). We used stable isotope labeling of amino acids in culture (SILAC) in the *cdk1-as1* yeast strain, in which Cdk1 is replaced with a mutant protein engineered to be specifically and rapidly inhibited by the pyrimidine-based inhibitor 1-NM-PP1 (*3*). Cells of a *cdk1-as1; arg4*Δ*; lys1*Δ strain, which require exogenous lysine and arginine to survive, were grown in medium containing lysine and arginine (the "light" culture) or in medium supplied with arginine and lysine labeled with stable heavy isotopes of carbon and nitrogen (^{13}C and ^{15}N) (fig. S1). This "heavy" culture was treated briefly (15 min) with 10 μM 1-NM-PP1 to inactivate Cdk1-as1. The cultures were then mixed together, lysed, and subjected to trypsinization. Phosphopeptides were purified from the peptide mixture and analyzed by means of tandem mass spectrometry (MS/MS). The precise sites of phosphorylation were inferred from the mass signature of peptide ion fragments in MS/MS spectra, and the ratio of heavy to light phosphopeptide in the MS spectra was used to infer relative abundance of all phosphopeptides with and without Cdk1 inhibition. We analyzed three different cell populations: an asynchronous population; a culture arrested in mitosis with the spindle poison nocodazole; and a culture arrested in late mitosis by overexpression of a nondegradable cyclin, Clb2-ΔN (*2*).

We collected 354,560 MS/MS spectra, of which 74,093 were successfully matched to phosphopeptide sequences. In total, we identified 10,656 different phosphorylation sites (database S1), of which 8710 sites on 1957 proteins were assigned a precise position with >95% confidence (database S2). The $\log_2$ heavy/light (H/L) ratios for nonphosphopeptides were tightly distributed around zero (a 1:1 ratio), indicating that global protein abundance was not affected by brief Cdk1 inhibition, whereas the $\log_2$ H/L ratios for phosphopeptides were more broadly distributed (Fig. 1A; see database S2 for a list of H/L ratios). A leftward shift in the H/L ratio of a phosphopeptide indicates that the abundance of that phosphopeptide decreased when Cdk1 was inhibited, as was expected for Cdk1 substrates. Indeed, we observed a leftward shift in peptides phosphorylated at a Cdk1 consensus sequence (S/T*-P, or S/T*-P-x-K/R, where x represents any amino acid and the asterisk indicates the site of phosphorylation), and the phosphopeptides with the lowest H/L ratios ($\log_2$ H/L < –3) were enriched for the Cdk1 consensus site (Fig. 1B), indicating that peptides whose phosphorylation decreased most after Cdk1 inhibition were enriched for direct targets of Cdk1. We therefore used two criteria to define a phosphorylation site as a Cdk1 substrate. First, the phosphorylated serine or threonine must be followed by a proline so as to conform to the minimal Cdk1 consensus sequence. Second, the phosphopeptide must decline in abundance by at least 50% after Cdk1 inhibition (as indicated by $\log_2$ H/L < –1) in one or more of our three experiments. Based on this double filtering, 547 distinct phosphorylation sites were identified on 308 candidate Cdk1 substrates (Fig. 1C and tables S1 and S2, substrate list).

Phosphorylation of Cdk1 consensus sites was observed on 67% (122 of 181) of proteins previously identified as Cdk1 substrates in vitro (*4*). Sixty-six percent (80 of 122) of these proteins contained sites at which phosphorylation decreased ($\log_2$ H/L < –1) after inhibition of Cdk1 (only 45 of 122 are expected if there is no correlation between the experiments in vitro and in vivo; χ^2 test, $P < 10^{-10}$).

A gene ontology analysis of the candidate substrates revealed a strong enrichment for cell cycle–related functional categories (such as GO:0007049, Cell Cycle; hypergeometric $P < 10^{-20}$) (table S3). Substrates are also involved in processes that are not traditionally thought of as being under cell-cycle control, including translation, chromatin remodeling, protein secretion, and nuclear transport (Fig. 2).

To modulate protein function, addition of a phosphate at a specific site can drive a precise

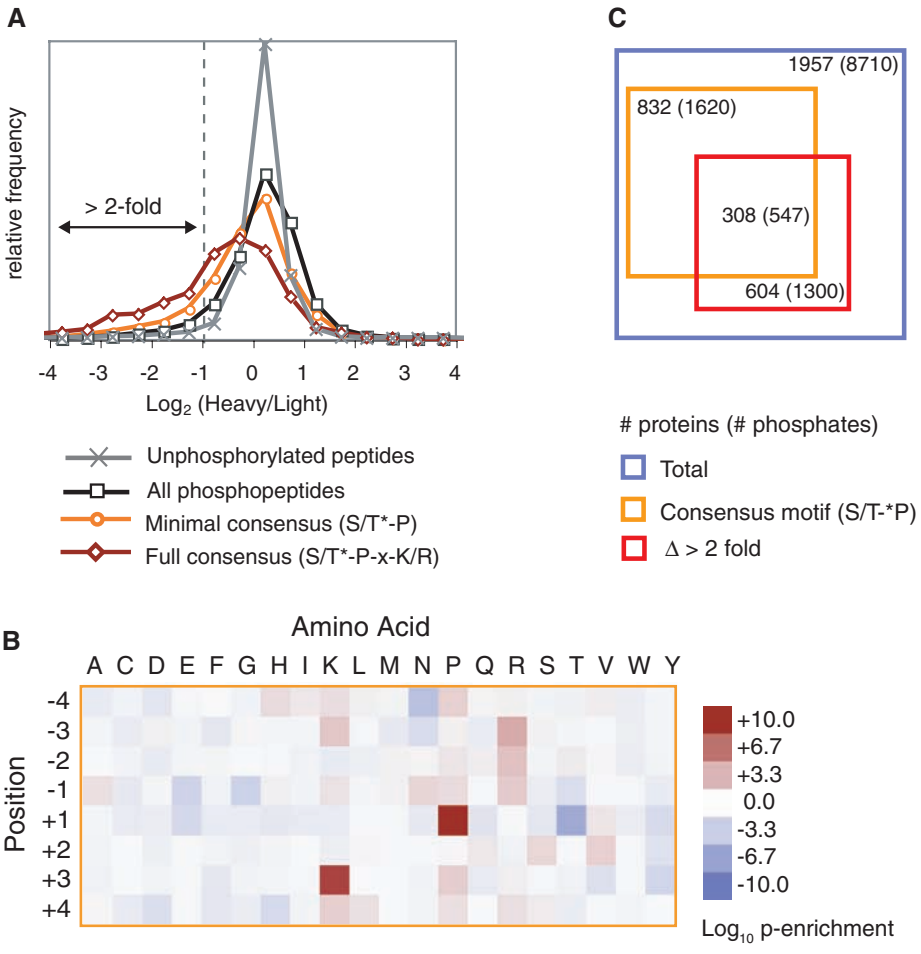

Fig. 1. Large-scale identification of Cdk1 substrates in vivo. (**A**) Distributions of $\log_2$ H/L ratios for unphosphorylated peptides (gray), all phosphopeptides (black), and phosphopeptides containing a minimal (orange) or full (red) Cdk1 consensus phosphorylation motif. (**B**) Log_{10} *P* values (binomial distribution) for the enrichment (red) or depletion (blue) of each amino acid (columns) at each position flanking the phosphorylated serine or threonine (rows) in phosphopeptides that changed greatly in abundance ($\log_2$ H/L < –3) relative to residues flanking serines and threonines proteome-wide. (**C**) Venn diagram representing the number of proteins and unique phosphorylation sites identified in the three experiments (*2*). The blue square indicates the total number of proteins and phosphorylation sites for which H/L ratios could be rigorously determined and for which the precise position of the phosphate could be assigned with 95% confidence. The orange square indicates proteins and phosphopeptides containing a minimal consensus motif, and the red square indicates proteins and phosphopeptides that decreased in abundance by over 50% after Cdk1 inhibition ($\log_2$ H/L < –1). Cdk1 substrates (table S1) were defined by the overlap between the orange and red squares. Squares are scaled to the number of proteins.

[1]Departments of Physiology and Biochemistry and Biophysics, University of California, San Francisco, San Francisco, CA 94158, USA. [2]Department of Microbiology and Immunology, University of California, San Francisco, San Francisco, CA 94158, USA. [3]Department of Cell Biology, Harvard Medical School, Boston, MA 02115, USA.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: steven_gygi@hms.harvard.edu (S.P.G.); david.morgan@ucsf.edu (D.O.M.)

conformational change in a protein loop or domain, thereby altering its activity or its interactions with other proteins (fig. S2A). This mechanism generally relies on coordination of the phosphate by networks of hydrogen bonds and is therefore highly context-dependent and unlikely to arise by a small number of random mutations. Alternatively, the addition of phosphates to a protein surface can directly disrupt interactions with other proteins (*5*, *6*) or can generate new interactions with phosphopeptide-binding modules such as 14-3-3, polo-box, WW, and Src homology 2 domains (fig. S2B) (*7*, *8*). In these cases, the position of the phosphate (or phosphates) is less context-dependent and therefore less constrained, and this form of phosphoregulation is expected to arise more readily through random mutation.

To assess the relative importance of these regulatory mechanisms in Cdk1 function, we analyzed the structural context and conservation of the 547 Cdk1-dependent phosphorylation sites. We found that more than 90% of these sites are predicted to be in loops and disordered regions (Fig. 3A and table S4), which is consistent with previous analyses of phosphorylation sites in general (*9*). Furthermore, we found that many Cdk1 targets have a greater number of phosphates than would be expected by chance ($P < 10^{-145}$; median Mann-Whitney *P* value from comparison of true distribution to 1000 simulations) (Fig. 3B), indicating that Cdk1 substrates tend to be phosphorylated at multiple sites. We also found that Cdk1-dependent phosphorylation sites tend to cluster in the primary amino acid sequence ($P < 10^{-15}$; median Mann-Whitney *P* value from comparison of true distribution to 1000 simulations) (Fig. 3C), suggesting that multiple phosphorylations modulate the same protein surface.

We used the complete genome sequences of 32 fungal species (fig. S3) to examine the evolution of Cdk1 phosphorylation sites. For each Cdk1 substrate, orthologous sequences were identified and aligned (*10*, *11*). A representative short stretch of alignment from the protein Shp1 is illustrated in Fig. 4A. This region of Shp1 contains two experimentally identified phosphorylation sites with different evolutionary dynamics. The precise position of site A, which lies on the edge of a predicted folded domain, has been preserved throughout the lineage. In contrast, the position of site B, which lies in a predicted disordered region, is conserved only in the closely related *sensu stricto Saccharomyces* group. However, Cdk1 consensus sites are found at other positions in this region throughout the lineage. Thus, although phosphorylation in the disordered region appears to be conserved, the precise position of the sites is less constrained.

Hierarchical clustering of all 547 Cdk1 phosphorylation sites showed that relatively few exhibit strong evolutionary conservation of their

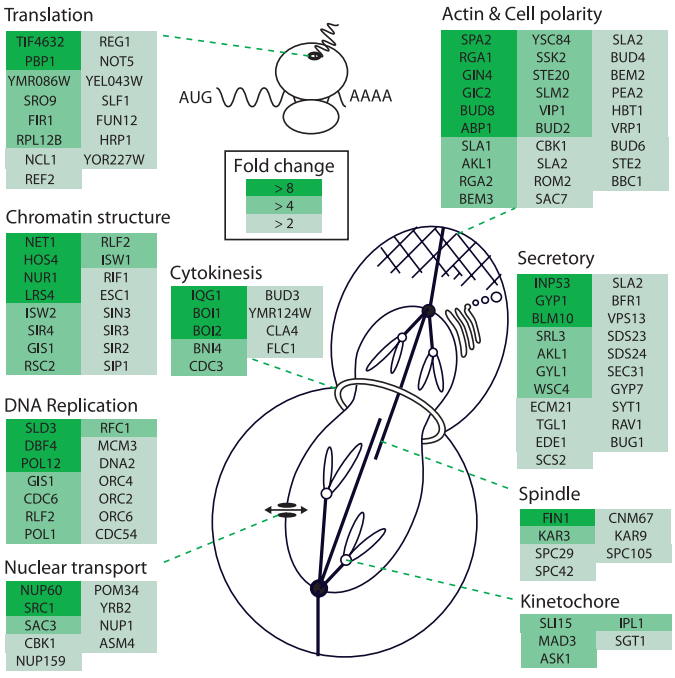

Fig. 2. Selected Cdk1 substrates grouped by cellular process. A subset of proteins phosphorylated in a Cdk1-dependent manner are organized into functional groups. The color of the box surrounding each protein corresponds to the fold-change of the most dynamically regulated phosphorylation site of each protein.

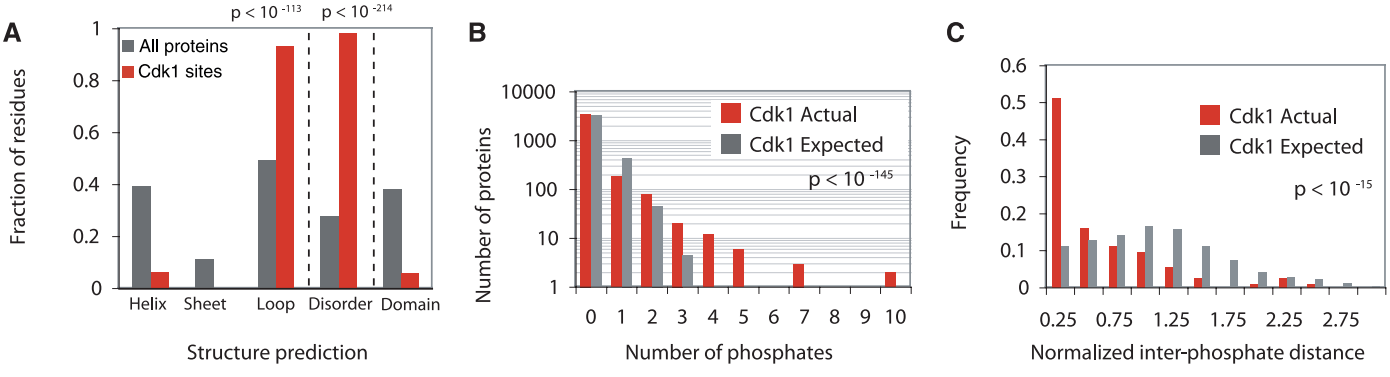

Fig. 3. Structural analysis of Cdk1-dependent phosphorylation sites. (**A**) The predicted structural environment of residues in all proteins in the *S. cerevisiae* genome (gray) and the residues that are phosphorylated in a Cdk1-dependent manner (red). Secondary structure (PsiPred) and disorder (PONDR) prediction algorithms (*2*) were used to predict the structural environment, and precomputed domain predictions were downloaded from the Saccharomyces Genome Database (ftp://ftp.yeastgenome.org/yeast/). All differences are significant at $P < 10^{-69}$ (binomial distribution). See table S4 for details. (**B**) The distribution of Cdk1-dependent phosphorylation sites per protein (red) is compared with the distribution of sites per protein from simulations in which the same number of phosphorylation sites is randomly scattered across a set of mitotic proteins with probability proportional to protein length (gray). To conservatively estimate the number of proteins present in mitosis, we used the set of 3838 proteins that are detectable with Western blotting (*21*). One thousand simulations were performed, and each simulated distribution was compared with the true distribution by calculating the Mann-Whitney *P* value. The "Cdk1 Expected" distribution is the average of the 1000 simulations. (**C**) The distribution of average distances between phosphates within a given protein. The average distances between Cdk1 sites were calculated for all proteins with two or more detected phosphorylation sites (red) and compared with the expected distribution generated by averaging the results of 1000 simulations in which the same number of phosphates was randomly assigned positions within each protein (gray). Because the expected distance between phosphates in a protein depends on the length of the protein, the average distances between phosphates shown here are normalized to protein length. The median Mann-Whitney *P* value from comparison of each of the 1000 simulated distributions with the true distribution is shown.

precise position (Fig. 4B, top, red box, and fig. S4). These phosphorylations might be expected to drive precise conformational changes and might therefore evolve more slowly (fig. S2A). Indeed, this type of substrate is highly enriched for metabolic enzymes (hypergeometric $P = 0.001$ for metabolic enzymes with precise-position age more than 0.5 units greater than enrichment age), which are generally more ancient than other open reading frames (ORFs) (fig. S5) and therefore might have evolved this form of regulation long ago.

A larger number of phosphorylation sites showed a different behavior: The precise position of the phosphorylation was conserved only in very closely related species, but there was a statistically significant enrichment of consensus sites throughout the lineage (Fig. 4B, bottom, blue box, and table S5). This pattern of evolution is consistent with context-independent forms of regulation, as discussed above (fig. S2B).

Precise phosphorylation site positioning might not be required for regulation of a protein by interactions with phosphopeptide-binding domains. We found a highly significant overlap between Cdk1 substrates and the binding partners of the phosphopeptide-binding domain found in 14-3-3 proteins. *S. cerevisiae* has two 14-3-3 proteins, Bmh1 and Bmh2. Ninety-four of 278 Bmh1- or Bmh2-interacting proteins (*12*) were identified as Cdk1 substrates in our studies (hypergeometric $P < 1 \times 10^{-20}$, assuming 3838 total ORFs) (fig. S6A). 14-3-3 proteins typically act as dimers and therefore contain two phosphate-binding sites that bind with higher affinity to multiphosphorylated proteins (*13*). Indeed, substrates that interact with Bmh1 and Bmh2 were more likely to be enriched with multiple Cdk1 consensus sites (Mann-Whitney $P < 10^{-4}$) (fig. S6B). Thus, shifting multisite phosphorylation might act in some cases to create generic interactions with phosphate-binding domains.

Several established Cdk substrates are regulated in multiple species by multisite phosphorylation in rapidly evolving regions (table S6). For example, clusters of Cdk1 phosphorylation sites in components of the prereplicative complex vary in position during evolution but are still likely to confer the same regulation (*14*–*16*). Our work reveals that many Cdk1 substrates are phosphorylated in vivo at rapidly evolving site clusters, which are likely to modify substrate func-

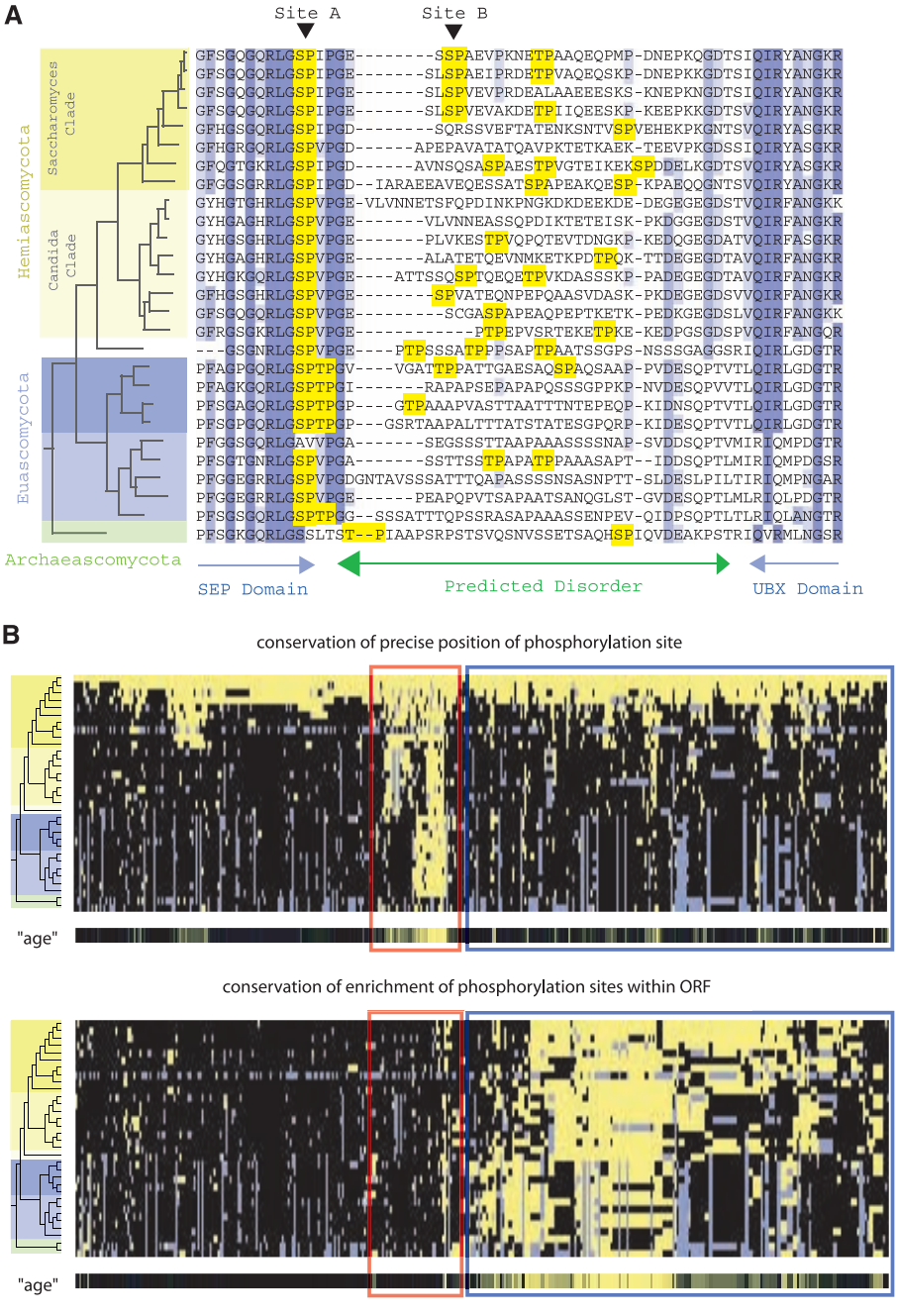

Fig. 4. Evolution of Cdk1-dependent phosphorylation sites. (**A**) Representative multiple sequence alignment of 27 orthologs of *S. cerevisiae* Shp1. The ascomycete phylogeny (fig. S3) is shown to the left of the alignment. Amino acid conservation is indicated by blue boxes, and minimal Cdk1 consensus motifs are highlighted in yellow. Blue arrows indicate predicted domains, and the green arrow indicates a predicted disordered region (PONDR). (**B**) Hierarchical clusters summarize the evolution of all 547 Cdk1 phosphorylation sites: Each row is a different species, and each column is a different phosphorylation site. The phylogeny (32 species) (fig. S3) is represented by a tree at the left. In the top clustergram, yellow indicates that a consensus site (S/T-P) aligns with the phosphorylation site detected in *S. cerevisiae* (top row). Gray indicates that no single ortholog was detected in that species. In the bottom clustergram, yellow indicates that there is an enrichment of Cdk1 consensus sites in the ortholog of the *S. cerevisiae* protein that we identified as a Cdk1 substrate. Enrichment in each ortholog was assessed by assuming that the expected frequency of a consensus motif is equal to the global frequency across all ORFs in the species and then using the Poisson distribution to calculate the probability of observing greater than or equal to the actual number of consensus sites. Enrichment was defined by a P value of less than 0.01 (for example, a typical 400-residue protein is expected to contain 2.8 sites, but must contain eight or more sites to achieve $P < 0.01$) (table S5). Two groups are highlighted within the clustergrams: one with conservation of precise site position (red box) and one with conservation of enrichment of consensus sites (blue box). Beneath each clustergram is a metric termed "age," which summarizes each column as a single conservation score (fig. S4). More intense yellow indicates greater conservation.

tion by simply disrupting or generating protein-protein interactions (fig. S2B).

An important implication of flexibility in phosphorylation site positioning is that combinatorial control by multiple kinases is readily evolved. Indeed, the protein kinase Ime2, a distant relative of Cdk1 that is expressed solely in meiotic cells, phosphorylates a large number of Cdk1 substrates at distinct sites but can still have the same effect as Cdk1 on substrate function (*17*).

The evolution of Cdk1 signaling appears to share features with the evolution of transcriptional regulation (fig. S7). Transcriptional regulators and Cdks both maintain their biochemical specificities (the DNA consensus motif and peptide consensus motif, respectively) over long evolutionary time scales. However, in both cases there is rapid evolution of the intergenic and disordered regions, respectively, that contain these motifs. In transcriptional regulation, DNA sequence motifs can function from many positions relative to the gene being controlled and, because of their short length and sequence degeneracy, can evolve rapidly (*18*–*20*). Similarly, many Cdk1 phosphorylation sites are not tightly constrained within the protein target sequence, and the signals for phosphorylation are short and easily evolved. These features allow cell-cycle control mechanisms to adapt rapidly to developmental challenges and opportunities that arise over time.

References and Notes

1. D. O. Morgan, *The Cell Cycle: Principles of Control* (New Science Press, London, 2007).
2. Materials and methods are available as supporting material on *Science* online.
3. A. C. Bishop *et al.*, *Nature* **407**, 395 (2000).
4. J. A. Ubersax *et al.*, *Nature* **425**, 859 (2003).
5. S. C. Strickfaden *et al.*, *Cell* **128**, 519 (2007).
6. Z. Serber, J. E. Ferrell Jr., *Cell* **128**, 441 (2007).
7. M. B. Yaffe, A. E. Elia, *Curr. Opin. Cell Biol.* **13**, 131 (2001).
8. T. Pawson, G. D. Gish, P. Nash, *Trends Cell Biol.* **11**, 504 (2001).
9. L. M. Iakoucheva *et al.*, *Nucleic Acids Res.* **32**, 1037 (2004).
10. B. B. Tuch, D. J. Galgoczy, A. D. Hernday, H. Li, A. D. Johnson, *PLoS Biol.* **6**, e38 (2008).
11. R. C. Edgar, *Nucleic Acids Res.* **32**, 1792 (2004).
12. K. Kakiuchi *et al.*, *Biochemistry* **46**, 7781 (2007).
13. D. Bridges, G. B. Moorhead, *Sci. STKE* **2005**, re10 (2005).
14. A. M. Moses, J. K. Heriche, R. Durbin, *Genome Biol.* **8**, R23 (2007).
15. A. M. Moses, M. E. Liku, J. J. Li, R. Durbin, *Proc. Natl. Acad. Sci. U.S.A.* **104**, 17713 (2007).
16. L. S. Drury, J. F. Diffley, *Curr. Biol.* **19**, 530 (2009).
17. L. J. Holt, J. E. Hutti, L. C. Cantley, D. O. Morgan, *Mol. Cell* **25**, 689 (2007).
18. G. A. Wray *et al.*, *Mol. Biol. Evol.* **20**, 1377 (2003).
19. S. B. Carroll, J. K. Grenier, S. D. Weatherbee, *From DNA to Diversity: Molecular Genetics and the Evolution of Animal Design* (Blackwell, Malden, MA, 2005).
20. B. B. Tuch, H. Li, A. D. Johnson, *Science* **319**, 1797 (2008).
21. S. Ghaemmaghami *et al.*, *Nature* **425**, 737 (2003).

22. We thank J. Feldman, R. Fletterick, M. Jacobson, H. Li, M. Matyskiela, P. O'Farrell, M. Sullivan, and S. Naylor for helpful comments; A. K. Dunker, E. Garner, C. Oldfield, K. Shimizu, and T. Ishida for disorder prediction algorithms; the Broad Institute, Sanger Center, Génolevures, and the Joint Genome Institute for genome sequence data; and O. Jensen, C. Zhang, and K. Shokat for reagents. This work was supported by grants from NIH (GM50684 to D.O.M., HG3456 to S.P.G., and GM037049 to A.D.J.) and fellowships from NSF (to L.J.H. and B.B.T.).

Supporting Online Material

www.sciencemag.org/cgi/content/full/325/5948/1682/DC1
Materials and Methods
Figs. S1 to S7
Tables S1 to S6
References
Databases S1 and S2

27 February 2009; accepted 15 July 2009
10.1126/science.1172867

Positive Selection of Tyrosine Loss in Metazoan Evolution

Chris Soon Heng Tan,[1,2,3] Adrian Pasculescu,[1] Wendell A. Lim,[4] Tony Pawson,[1,2]* Gary D. Bader,[1,2,3]* Rune Linding[5]*

John Nash showed that within a complex system, individuals are best off if they make the best decision that they can, taking into account the decisions of the other individuals. Here, we investigate whether similar principles influence the evolution of signaling networks in multicellular animals. Specifically, by analyzing a set of metazoan species we observed a striking negative correlation of genomically encoded tyrosine content with biological complexity (as measured by the number of cell types in each organism). We discuss how this observed tyrosine loss correlates with the expansion of tyrosine kinases in the evolution of the metazoan lineage and how it may relate to the optimization of signaling systems in multicellular animals. We propose that this phenomenon illustrates genome-wide adaptive evolution to accommodate beneficial genetic perturbation.

It is a biological paradox that organism complexity shows limited correlation with gene repertoire size (*1*). However, some protein families (*2*) have expanded with organism complexity as measured by number of cell types (*3*), especially those involved in regulation, such as tyrosine kinases in signaling, cell-cell communication, and tissue boundary formation (*4*, *5*). We observed a striking negative correlation of genomically encoded tyrosine content with the number of distinct cell types in metazoan species (Spearman's $\rho = -0.89$, approximate $P = 3.0 \times 10^{-6}$; Pearson's $\rho = -0.89$, approximate $P = 4.0 \times 10^{-6}$) (Fig. 1A). Thus, metazoans with more cell types have proportionally less potential tyrosine phosphosites. Similarly, we observed that the number of tyrosine kinase domains correlates negatively with genomic tyrosine content (Spearman's $\rho = -0.68$, approximate $P = 3.7 \times 10^{-3}$; Pearson's $\rho = -0.81$, approximate $P = 1.3 \times 10^{-4}$) (Fig. 1B). Including dual-specificity mixed-lineage kinases (MLKs) and mitogen-activated protein kinase kinases (MEKs) revealed a similar pattern (fig. S1A).

These observations suggest an evolutionary model in which the acquisition of a tyrosine kinase results in systems-level adaptation to remove deleterious phosphorylation events that cause aberrant cellular behavior and diseases (*4*). Assuming that a cell begins with a single tyrosine kinase, which is subsequently duplicated, it follows that the kinases may functionally diverge, as a result of relaxation in evolutionary constraints, to phosphorylate new substrates. Emerging kinase specificities could be retained if new substrates confer selection advantage. However, it is unlikely that every new phosphorylation event is beneficial. We hypothesize that optimization of newly emerged signaling networks would follow (*6*) through the elimination of detrimental phosphorylation events by tyrosine-removing mutations. Even if many new phosphorylation sites are not deleterious, an organism with minimized noisy signaling systems is likely to have a fitness advantage. This scenario is repeated with the subsequent duplication of tyrosine kinases leading to more tyrosine residues lost (*7*).

Despite several recent systematic phosphoproteomic studies (*8*), many human proteins have no observed phosphotyrosines. Our model suggests that tyrosine loss had occurred predominantly in these proteins in order to minimize tyrosine phosphorylation. To test this hypothesis, we investigated differences in tyrosine loss between these proteins (Non-pTyr) and those that are tyrosine phosphorylated (pTyr). Comparing members of these two groups to their orthologous proteins in *S. cerevisiae* (*7*), which lack conventional tyrosine kinases, enabled us to assess the degree of tyrosine loss that may be triggered by the onset of phosphotyrosine signaling in metazoans.

[1]Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Toronto M5G 1X5, Canada. [2]Department of Molecular Genetics, University of Toronto, Toronto M5S 1A8, Canada. [3]Terrence Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto M5S 3E1, Canada. [4]Howard Hughes Medical Institute and Department of Cellular and Molecular Pharmacology, University of California, San Francisco, San Francisco, CA 94158, USA. [5]Cellular and Molecular Logic Team, Section of Cell and Molecular Biology, The Institute of Cancer Research (ICR), London, SW3 6JB, UK.

*To whom correspondence should be addressed. E-mail: pawson@lunenfeld.ca (T.P.); gary.bader@utoronto.ca (G.D.B.); linding@icr.ac.uk (R.L.)