Running head: Novel phosphatase evolution in algae and higher plants

Greg Moorhead
Department of Biological Sciences, University of Calgary, 2500 University Drive NW,
Calgary, AB T2N 1N4, Canada
Email: moorhead@ucalgary.ca Phone: (403) 220-6238

Research category: Genome Analysis
Title:
Evolutionary Radiation Pattern of Novel Protein Phosphatases Revealed by Analysis of Protein Data from the Completely Sequenced Genomes of Humans, Green Algae and Higher Plants

David Kerk¹⁺, George Templeton¹⁺, Greg BG Moorhead¹⁺
1 Department of Biological Sciences, University of Calgary, 2500 University Drive NW, Calgary, AB T2N 1N4, Canada

* Correspondence author
+ These authors contributed equally
Footnotes:

Financial sources: GM is supported by NSERC, GT is supported by NSERC and Alberta Ingenuity, and DK is supported by NSERC and the Alberta Ingenuity Center for Carbohydrate Science.

Corresponding Author: Greg Moorhead, email moorhead@ucalgary.ca
Abstract
In addition to the major PPP, PPM and PTP families, there are novel protein phosphatases, including enzymes with Asp-based catalysis and subfamilies of protein tyrosine phosphatases, whose evolutionary history and representation in plants is poorly characterized. We have searched the protein datasets encoded by the well-finished nuclear genomes of the higher plants Arabidopsis thaliana and Oryza sativa, and the latest draft datasets from the tree Populus trichocarpa and the green algae Chlamydomonas reinhardtii and Ostreococcus tauri, for homologues to several classes of novel protein phosphatases. The Arabidopsis proteins, in combination with previously published data, provide a complete inventory of known types of protein phosphatase in this organism. Phylogenetic analysis of these proteins reveals a pattern of evolution where a diverse set of protein phosphatases was present early in the history of eukaryotes, and the division of plant and animal evolution resulted in two distinct sets of protein phosphatases. The green algae occupy an intermediate position, and show similarity to both plants and animals, depending on the protein. Of specific interest are the lack of cell division phosphatases CDC25 and CDC14, and the seeming adaptation of CDC14 as a protein interaction domain in higher plants. In addition, there is a dramatic increase in proteins containing RNA polymerase C-terminal domain (CTD) phosphatase-like catalytic domains in the higher plants. Expression analysis of Arabidopsis phosphatase genes differentially amplified in plants (specifically the CTD phosphatase-like phosphatases) show patterns of tissue specific expression with a statistically significant number of correlated genes encoding putative signal transduction proteins.
Abbreviations

PPP – serine/threonine specific Phosphoprotein Phosphatase, PPM – Phosphoprotein phosphatase, Mg$^{2+}$ dependent, PTP – Protein Tyrosine Phosphatase, CDC – Cell Division Cycle, CTD – C-terminal Domain of RNA Polymerase II, DSP – Dual Specificity Phosphatase, FCP – TFIIF-associating component of CTD Phosphatase, SCP – Small CTD phosphatase, HAD – Haloacid Dehalogenase, HMM – Hidden Markov Model, NADK - Nicotinamide Adenine Dinucleotide Kinase, NJ – Neighbor Joining, Pars – Maximum Parsimony, ML – Maximum Likelihood, CDK – Cyclin Dependent Kinase, ACR – Arsenate Reductase, LMWPTP – Low Molecular Weight Protein Tyrosine Phosphatase, CPL – RNA polymerase II C-terminal domain phosphatase-like protein phosphatase, BRCT – BRCA-related C-Terminal, ABA – Abscisic Acid

Introduction

The phosphorylation and dephosphorylation of proteins has been found to modify protein function in a multitude of ways (Cohen, 2002). The protein kinase content (kinome) of many eukaryotes and their evolutionary relationships have been studied in depth, revealing both the importance and diversity of these proteins (Manning et al., 2002; Manning et al., 2002; Caenepeel et al., 2004; Champion et al., 2004). With the exception of the few PIKKs (phosphatidyl inositol 3-kinase-like kinases), the protein kinases share a highly conserved catalytic domain. In contrast, the protein phosphatases are more diverse, displaying three different catalytic signatures, and thus can be divided into three broad groups (Moorhead et al., 2007). While many of the phosphatases have been catalogued in the genomes of several organisms (Koh et al., 1997; Kerk et al., 2002; Alonso et al., 2004), this list continues to grow and, in organisms like higher plants, classification schemes are often quite incomplete.

Protein phosphatases were originally identified as enzymes responsible for dephosphorylating serine and threonine residues on enzymes involved in mammalian glycogen metabolism. Purification of these enzymes, cloning and genomics has revealed that this group is composed of two families (PPP and PPM) which represent the major
group of serine and threonine phosphatases in eukaryotes (Table 1; Rayapureddi et al., 2003; Alonso et al., 2004; Gohla et al., 2005; Moorhead et al., 2007). Ten years after the cloning of the first tyrosine kinase, the first tyrosine phosphatase was purified and then cloned. Its catalytic signature \([C(X)\text{R}]\) defined the large protein tyrosine phosphatase (PTP) superfamily (Table I), which now, in addition to the tyrosine specific enzymes, include enzymes that specifically dephosphorylate serine or threonine as well as tyrosine (the dual specificity enzymes, or DSPs), mRNA and phosphoinositides. Based on this catalytic signature, the group has expanded to 107 transcribed genes in humans. Several of these are catalytically inactive, but their gene products function in the cell, and several have been linked to human diseases (Robinson and Dixon, 2006). The third major group of phosphatases was identified most recently and is characterized by a catalytic signature \(\text{DXDXT/V}\) and is referred to here as the Asp-based enzymes (Table I). The phosphatase responsible for dephosphorylation of the C-terminal domain (CTD) of RNA polymerase II (FCP1) was the first enzyme of this group to be identified as a protein phosphatase and the related small C-terminal domain phosphatases (SCPs) have been recognized as part of the group. Several haloacid dehalogenase (HAD) superfamily members, such as Eyes Absent (EyA), which acts as a transcription factor, and chronophin, which controls cofilin phosphorylation, have now been demonstrated to function as protein phosphatases. Like FCP1 and SCP enzymes, HAD superfamily members have a \(\text{DXDXT/V}\) catalytic signature, utilizing a unique aspartate-based catalytic mechanism. The HAD superfamily is potentially very large, but to date only a few members have been demonstrated to display serine or tyrosine phosphatase activity (Moorhead et al., 2007).

A catalogue of enzymes that comprises the PPP, PPM, and some of the PTP family members of Arabidopsis thaliana was presented several years ago (Kerk et al., 2002). Since then both the number of known protein phosphatases and the set of completely sequenced reference genomes has expanded considerably. In this situation a systematic re-visitation of the protein phosphatase repertoire is warranted. We have focused on these new or novel phosphatases by using the eleven phosphatase classes from Homo sapiens to identify homologues from the genomes of A. thaliana, Chlamydomonas reinhardtii, Oryza sativa, Ostreococcus tauri, and Populus trichocarpa. These proteins, along with their
counterparts in humans and other animals, were analyzed to determine their interrelationships. When combined with previous studies from our laboratory, this work defines a complete set of all known varieties of protein phosphatases in Arabidopsis.

Results and Discussion:

Homologues were identified using BLAST searches as well as Hidden Markov Models (HMM) of the catalytic domains. The overall results of this study are summarized in Table II. This lists the number of homologues for each protein phosphatase type which were found in each of the subject organism-specific databases. The structural classes are derived from Table 1 in the recently published study of Moorhead et al. (2007). Results will be discussed in the order of their appearance in this table, followed by an analysis of expression and promoters of a subset of the identified genes. Evidence of expression of all proteins is summarized in Supplemental Table SI (see Methods for details).

i) Protein tyrosine phosphatases

SSU72

Using the sequence of H. sapiens SSU72 (gi:7661832) to search the target protein databases, we found one candidate homologue each in C. reinhardtii (Cre130182), O. tauri (Ot16g02480), Populus (Pop775960), Arabidopsis (At1g73820.1), and O. sativa (Os12g07050.1 and .2). We constructed a multiple sequence alignment, which is presented as Supplement Figure S1. Phylogenetic trees show that both of the algal sequences cluster with the higher plant sequences in two of the three inference methods with high bootstrap support (82.2% Pars; 80.0% ML).

Slingshot

Slingshots, along with chronophin (included in Section iv) dephosphorylate cofilin, and as a result, stimulate the depolymerization of F-actin (Huang et al., 2006). We used the H. sapiens sequences SSH1 (gi:40254884), SSH2 (gi:37674210) and SSH3 (gi:24586675) to search the target protein databases. We found no candidate slingshot homologues in the algae, nor in any of the plant species, indicating that this method of regulation is animal specific.

CDC14
CDC14 belongs to the family of PTPs, and is responsible for control of mitotic exit in organisms of the fungi/metazoan group (Trinkle-Mulcahy and Lamond, 2006). We used the human proteins CDC14A (gi:55976620) and CDC14B (gi:55976216) to search the target protein databases. A single candidate CDC14 homologue was found in C. reinhardtii (Cre112184). Our work shows clearly through several different methods that this sequence has all the structural features expected of a true CDC14. First, analysis with the FFAS03 technique (Rychlewski et al., 2000) shows that there is sequence similarity between Cre112184 and the solved structure of human CDC14B (PDB entry: 1ohc) (Gray et al., 2003) extending for over 300 amino acids, encompassing both the upstream “A” (unique to CDC14) and downstream “B” (PTP/DSP catalytic) domains. The score for this comparison is very high (Z~90 vs. Z~110 for HuCDC14B vs itself; for this technique a Z score of 9.5 or greater is considered significant). Second, the FFAS03 alignment shows high conservation (10/11) of a set of critical residues described in the solved structure of human CDC14B (these include the canonical PTP/DSP catalytic residues [HCX5R] in the “B” domain, plus a number of others unique to CDC14s). Third, a multiple sequence alignment was constructed encompassing the “B” domain of animal CDC14s and, as an outgroup, the protein phosphatase domains of a set of DSPs previously characterized from Arabidopsis (Kerk et al., 2002; Kerk et al., 2006) (presented as Supplement Figure S2). The corresponding phylogenetic tree is presented as Figure 1. It is clear that in this tree Cre112184 is part of a clade with human and X. laevis CDC14s, sharing a common node in all three phylogenetic tree inference methods, with high bootstrap support (100% Neighbor Joining [NJ]; 98.8% Maximum Parsimony [Pars]; 78.2% Maximum Likelihood [ML]).

In the green alga O. tauri there is a domain which is related to CDC14s. It was initially detected with BLAST searches utilizing human CDC14 sequences (E~e-5). It is found as an N-terminal domain in a sequence (Ot15g00870) which is annotated as containing an NADK (Nicotinamide Adenine Dinucleotide Kinase) domain (PF01513). In a reciprocal BLAST search the best hit to this O. tauri domain is human CDC14A (E=0.001), indicating a specific relationship. Using this O. tauri domain sequence as a query, we found similar sequences (BLAST hits E~e-15) in P. trichocarpa (2 sequences; [Pop347211, Pop714589]), Arabidopsis (A1g21640.1) and O. sativa (Os11g08670). Upon
further analysis, it is very clear that these sequences are all related to CDC14. First, when
the FFAS03 technique is applied to each of them, there is similarity to the solved structure
of human CDC14B (PDB entry: 1ohc) along a 300 amino acid region which encompasses
both the “A” and “B” domains. The scores for these comparisons are strong, between Z=20
and Z=50. (Note, however, that these scores are much weaker than those obtained with the
C. reinhardtii sequence Cre112184, presented above). The O. tauri sequence (Z~50),
retains the canonical PTP/DSP catalytic residues (HC₃R), and therefore could be
enzymatically active. It retains some of the hydrophobic pocket residues described in the
solved structure of human CDC14B, but not all of them, and has only two of six acidic
residues in the “acidic groove” region. In the solved structure these acidic residues are
thought to be critical to binding basic residues in the target CDK (Cyclin Dependent
Kinase), while the hydrophobic pocket residues are responsible for maintaining substrate
specificity of CDC14 for proline in the pSer + 1 position (Gray et al., 2003). Therefore,
while structural resemblance is readily apparent, it is doubtful that this domain could
function specifically as a CDC14. The higher plant sequences obtain FFAS03 scores of
between Z=22 and Z=36. They lack nearly all of the set of specific CDC14 residues, and
furthermore have a C to S substitution in the PTP/DSP catalytic loop sequence. They
therefore could not function as catalytic domains. A multiple sequence alignment
encompassing the length of the full “A” and “B” domains of the CDC14-like sequences is
presented as Supplement Figure S3.

When considered in the context of the shorter multiple sequence alignment
(Supplement Figure S2) and the subsequent phylogenetic tree (Figure 1), it is apparent that
sequence Ot15g00870 and the higher plant sequences form a second, distinct clade, sharing
a common node with high bootstrap support in all tree inference methods (99.1% NJ;
98.4% Pars, 82.7% ML). Finally, the C. reinhardtii/animal cluster and the O. tauri/higher
plant cluster are clearly related, compared to the generic Arabidopsis DSPs, sharing a
common node in all three tree inference methods, with high bootstrap support in two of
them (88.6% NJ; 95% Pars; 40.2% ML). It is thus very clear that all these sequences are
“CDC14-like” and evolved from a common ancestor, if not from CDC14 itself. All other
clusters and nodes in this DSP tree are as previous published (Kerk et al., 2006; data not shown).

As an additional note of interest, the higher plant CDC14-like sequences are found as a domain on Nicotinamide Adenine Dinucleotide kinases (NADKs). NADK2, the Arabidopsis protein containing the domain, is a chloroplast localized protein, that has been shown to be a calcium-dependent calmodulin-regulated protein (Turner et al., 2004; Chai et al., 2005). Calmodulin binding takes place on the N-terminus of the protein, which is the location of the CDC14-like domain. The calmodulin binding site was mapped to a 45 amino acid region of the domain, presented in Figure 7 of Turner et al.(2004). When comparing with our multiple sequence alignment, the heart of this binding motif is the mutated PTP consensus site (H[C→S]X₅R). This is reminiscent of "substrate trapping" mutants observed at the altered consensus catalytic motifs of other protein tyrosine phosphatases (Bliska et al., 1992; Milarski et al., 1993; Sun et al., 1993).

**CDC25**

The CDC25 proteins also have a role in control of progression through the cell cycle in fungi and metazoans. While CDC14 controls mitotic exit, CDC25 is involved in the transition from G2 to M phase (Trinkle-Mulcahy and Lamond, 2006). Homo sapiens have three CDC25s (CDC25A [gi:50403734]; CDC25B [gi:21264471]; CDC25C [gi:125625350]). Using these sequences to search the target databases, we found a number of similar sequences, whose protein phosphatase catalytic domains are collected in the multiple sequence alignment presented as Supplement Figure S4. The resulting phylogenetic tree is presented as Figure 2. O. tauri has a candidate CDC25 homologue (Ot02g05470), that is a member of a CDC25 clade encompassing human, animal, and yeast sequences, sharing a common node with high bootstrap support in all three tree inference methods (100% NJ; 97.8% Pars; 75.9% ML). Previous work supports this, as this gene has previously been cloned, and the expressed protein acted as a CDC25 in both yeast complementation and starfish oocyte cell division assays (Khadaroo et al., 2004). They noted the divergence of the N-terminal domain of this sequence (see Figure S5 for an alignment of the O. tauri and H. sapiens sequences), but asserted that there was a conserved 14-3-3 binding site, and several potential phosphorylation sites. Presumably the 14-3-3
binding site reported was 252 - RPLASPP - 258, the closest match to either of the
consensus binding sites (Rxxx[S/T]xP in this case), which, while matching the consensus,
has a proline in both the S-3 and S+1 positions. This has been shown to be unfavourable
(Yaffe et al., 1997), making it unlikely that the O.tauri protein has any 14-3-3 binding
capability. Thus, although this protein clearly can act like a true CDC25 in functional
assays, and we support its classification as a CDC25 based on our sequence analysis, we
anticipate that its regulation in vivo might well differ from that previously described for the
fungal/animal proteins. Our searches also revealed sequences sharing some similarity in C.
reinhardtii (Cre153947, Cre171654, Cre183511, Cre167673), P. trichocarpa (Pop282198),
Arabidopsis (A5g03455.1) and O. sativa (O10g39860, O03g01770). The Arabidopsis
sequence was originally published as a “CDC25” (Landrieu et al., 2004), but more recent
work indicates that it may functionally be an arsenate reductase, indicated by a lack of
arsenate reductase activity in a T-DNA plant line, and in vitro arsenate (V) reductase
activity (Bleeker et al., 2006; Dhankher et al., 2006). The sequences of several known
arsenate reductases from fern (PvACR2), yeast (ScACR2, SpACR2) and the protist
Leishmania major (LmACR2) form a clade with the above candidate algal and higher plant
sequences, with strong bootstrap support in two phylogenetic tree inference methods
(98.8% Pars; 82.7% ML), supporting the categorization of these proteins as arsenate
reductases. Our findings confirm, and extend, with larger sequence sets, the phylogenetic
analyses previously reported (Dhankher et al., 2006; Ellis et al., 2006). The higher plant
and C. reinhardtii sequences also lack any significant N-terminus, which is known to
contain regulatory sites in animal CDC25s, such as phosphorylation and 14-3-3 protein
binding sites. It is apparent that the algal/plant proteins lack the conserved regulatory sites,
as well as the catalytic activity, of the fungal/animal CDC25s.

This viewpoint is supported by a recent article questioning the existence of CDC25
in higher plants, and suggesting that cell cycle control has been reorganized along lines
distinctly different than the fungal/metazoan (presumably ancestral) model (Boudolf et al.,
2006). It seems that with our data, an expansion of this concept is in order. It appears as if
this reorganization of cell cycle control has occurred in higher plants, and that this process
began during the radiation of the green algae. When combined with the above information
on CDC14 in plants, it appears that different algal species are "frozen" at different points in this reorganization, some retaining one mitotic phosphatase or the other, but both lost by the transition to higher plants. Study of these algal species, as well as the higher plants, may give a unique insight into the evolution of these processes.

**LMWPTP**

We used the sequence for the LMWPTP (Low Molecular Weight Protein Tyrosine Phosphatase) from H. sapiens (ACP1 [gi:1709543]) to search the target protein databases. We found one candidate homologue in C. reinhardtii (Cre117512), none in O. tauri, two in P. trichocarpa (Pop821042, Pop594818), one in Arabidopsis (At3g44620.1), and one in O. sativa (Os08g44320.1). The multiple sequence alignment constructed from these sequences is presented as Supplement Figure S6. As seen in the alignment, the proteins are remarkably conserved, indicating an essential, conserved function for the protein in eukaryotes. The lack of a homolog in O. tauri is puzzling, however, and is possibly the result of secondary loss of the protein, as the O. tauri genome is remarkably streamlined (Derelle et al., 2006).

**ii) Asp-based catalysis: FCP-like**

The TFII-interacting RNA Polymerase II C-terminal domain protein phosphatase FCP1 is an essential yeast protein, which acts to dephosphorylate the C-terminal domain of the largest subunit of RNA Pol II (Archambault et al., 1997). This subunit contains an array of repeats of a heptad unit containing serine residues at positions 2 and 5. The transcription initiation and elongation process consists of a variety of mRNA modifying proteins being recruited to the Pol II complex. This appears to be modified by the state of Pol II phosphorylation – it is recruited to the complex in a hypophosphorylated state, phosphorylated during the transcription process, then dephosphorylated to allow termination and recruitment to a new complex (Meinhart et al., 2005; Moorhead et al., 2007). Complex modifications of the “phospho-array” are thus possible, and with it modulation of the transcription process. FCP1 is a metal-binding protein, possessing a DXDXT/V motif which is essential to catalytic activity (Kobor et al., 1999). The isolated phosphatase domain is sufficient for catalytic activity. In plants and algae, we found a large set of proteins sharing a degree of similarity to this prototype sequence. A large
multiple sequence alignment of the protein phosphatase catalytic domain of 99 sequences was constructed (presented as Supplement Figure S7) and the corresponding phylogenetic trees inferred (Figure 3). The Arabidopsis proteins CPL1 (C-terminal domain phosphatase-like protein phosphatase) and CPL2 contain an FCP-like catalytic domain, however they, and their homologues in other plants, are further characterized by the presence of one or two dsRNA binding domains, and are discussed separately from the other members of this family. The trees were composed of several distinct subclusters, which are each presented in turn, based upon the topology of the Neighbor Joining tree. The amino acid residues required for protein phosphatase catalytic activity have been well studied in FCP1 (Hausmann and Shuman, 2003; Hausmann et al., 2004). A set of 11 critical sequence positions have been identified through biochemical analysis, and are indicated in Supplement Figure S7. The majority of sequences in this FCP1-like dataset retain conservation at all these residue positions. However, 44 sequences deviate from the yeast residue pattern in at least one position (see Supplement Figure S7 legend).

SCP (Subcluster A)

SCP proteins are small RNA Pol II C-terminal domain protein phosphatases. In humans there are three proteins, SCP1 (gi: 15278033), SCP2 (gi: 31074179) and SCP3 (gi: 34392247). We found one sequence in Chlamydomonas (Cre149388) and one in Ostreococcus (Ot12g02910) that share similarity with the animal proteins in the phosphatase catalytic domain, and none in the higher plants. Upon multiple sequence alignment and phylogenetic tree inference, these algal sequences form part of a clade with the human and other animal proteins, sharing a common node with varying bootstrap support in the three tree inference methods (99.7% NJ; 44.8% Pars; 34.7% ML). This support is not to our minimum threshold of majority support in two of three methods. To clarify the situation, sequences were added (duplicates with different accession numbers) and removed (more divergent sequences) from the alignment used to generate the trees. In both these situations, the modified alignments, with more or less sequences, met our requirements of support of two of three methods. However, N-terminal (non-catalytic domain) motif analysis shows that the O. tauri sequence does not share motifs found in the animal sequences, and the C. reinhardtii sequence lacks this N-terminal region. Thus,
while not unequivocal, these data support the assignment of these two algal sequences to the SCP cluster.

**Subclusters B-F**

Subcluster B is a mixed group of 19 animal/algal/plant sequences which achieve high to moderate bootstrap support in two tree inference methods (96.7% NJ; 70.1% ML). The sequences in Subcluster B are: Hu6841480 (also known as HSPC129), Xt56605878, Ot14g03430, Cre118402, Pop415530, Pop179057, Pop586980, Pop568466, A5g11860.1, A5g46410.1, O7g10690.3, O07g10690, O05g49120, O1g43870.2, O01g43870, O1g47540.3, O01g47540, O1g47540.4, O1g47540.2. N-terminal motif analysis of the sequences in Subcluster B shows that the Os01g43870 isoforms lack motifs shared by the other sequences. Characterization of the proteins in this group is limited to a very recent study indicating CTD phosphatase activity of the human protein HSPC129 (Qian et al., 2007). There has been moderate radiation of proteins belonging in this subcluster in higher plants, where, discounting splice variants, higher plants have between two and four homologues.

Subcluster C comprises human “Dullard” and its animal homologues (100% NJ; 100% Pars; 98.4% ML). Dullard is a fairly recent discovery in this gene family, and has been implicated in neural tube development, the BMP (bone morphogenetic protein) pathway, and in nuclear membrane biogenesis (Satow et al., 2002; Satow et al., 2006; Kim et al., 2007). N-terminal sequence analysis shows that these sequences share common motifs. The presence of a Dullard homolog in yeast, involved in a conserved pathway (Kim et al., 2007), and the lack of homologs in algae/plants, suggests it arose after the plant/animal evolutionary split.

Subcluster D is a group of six higher plant sequences (Pop545127, A1g29780.1, A1g29770.1, A5g45700.1, O05g11570, Pop235176). This group has high bootstrap support in all three tree inference methods (99.4% NJ; 91.8% Pars; 80.6% ML). N-terminal sequence analysis shows that these sequences share common motifs. Subcluster E contains both algal and higher plant sequences (Cre111940, Pop560916, Pop659512, A3g55960.1, Os01g61640, O05g39070). This group receives high to moderate bootstrap support in two tree inference methods (99.3% NJ; 60.46% ML) and these sequences share common N-
terminal motifs. Subcluster F contains two C. reinhardtii sequences that are not splice variants (Cre187551, Cre142839) and have high to moderate bootstrap support in two of three inference methods (99.7% NJ; 79.4% ML). The proteins of subclusters D-F have yet to be characterized.

Subcluster G
This is a set of eight sequences from animals, algae, and higher plants (TIM50Hu, Cre169672, Ot03g04550, Pop642296, Pop568582, A1g55900.1, O05g43770, O1g55700.1). This group receives moderate bootstrap support in all three tree inference methods (82.0% NJ; 88.7% Pars; 74.7% ML). N-terminal motif analysis shows that most sequences in this cluster share a common signature. TIM50 only weakly shares some elements of this signature, and is clearly the most distantly related sequence in this group. TIM50 (sometimes referred to as TIMM50) is the homolog of the yeast protein of the same name, and is named as translocase of inner mitochondrial membrane 50 kDa. As indicated by its name, TIM50 is involved with the translocation of proteins through the inner membrane into the matrix, as part of the TIM23 complex (mitochondrial protein import reviewed in Neupert and Herrmann, 2007), although a nuclear localized isoform has also been identified in humans (Xu et al., 2005). The Arabidopsis protein was identified in a previous proteomic characterization of mitochondrial import proteins, demonstrating the conservation of the localization of this protein, at the very least (Lister et al., 2004). The human isoform has also been shown to possess phosphatase activity, intriguingly active on phosphorylated serines, threonines and tyrosines (Guo et al., 2004). This subcluster is the only one containing proteins that we can be all but assured are not involved in the dephosphorylation of the CTD (due to their mitochondrial localization), raising questions not only about the substrate specificity of other FCP-like proteins, but also about the specific target of dephosphorylation by these TIM50 homologues.

“FCP Assemblage” (Subcluster H)
This is a large group of sequences (24) from animals, algae and higher plants. The assemblage as a whole receives moderate to low bootstrap support from all three tree inference methods (64.1% NJ; 67.6% Pars; 43.7% ML; note that several sequences are excluded from the subcluster in the maximum parsimony tree, see the Figure 3 legend for
Within it are distinct subclusters formed by the yeast/animal FCP1 group (FCP1_Spomb, FCP1_Hu, Dr49618915, Dr94734487, FCP1_Xle, Xt62858037), the higher plant CPL3 (Pop708815, CPL3_Ath, OsCPL3), and CPL4 (Pop262722, CPL4_Ath, OsCPL4) groups, and associated algal sequences (Cre187332, Cre141879, Ot04g02710, Ot03g04040). Of particular note is a subcluster made up exclusively of Arabidopsis sequences (A5g23470.1, A2g02290.1, A1g20320.1, A1g43600.1, A1g43610.1). There is also a subcluster of closely related sequences from the algae as well as P. trichocarpa (Ot03g00770, Cre149314, Pop560900). Each of these subclusters receives high bootstrap support, but their relative topological interrelationships within the assemblage varies slightly amongst the different tree inference methods.

N-terminal motif analysis of these sequences indicates that some of these sequences share more than simply the same catalytic domain. The higher plant CPL3s have a distinct motif signature (data not shown). Elements of this signature are weakly shared by the two algal sequences: Ot03g00770, and Cre187332. This indicates that these sequences are most closely related to the CPL3s. The higher plant CPL4s also have a distinct motif signature (data not shown). The algal sequence Ot03g04040 shares this motif signature, as does the algal sequence Cre141879 (though with reduced similarity). Finally, the “CPL4” upstream motif signature is shared by the Arabidopsis sequences: A1g20320, A2g02290, and A5g23470. It is likely that all these CPL4-like sequences are related. The animal FCP1s share a common motif signature, which is shared to some extent by yeast FCP1. The CPL4-like sequences share elements of this motif signature with the FCP1s, whereas the CPL3s do not, suggesting a closer relationship of CPL4-like and FCP1 groups. The algal sequence Ot04g02710 shares elements of the FCP1 motif signature, suggesting it is more closely related to the animal and yeast FCP1s. Sequences A1g43600 and A1g43610 have non-existent or short N-termini (respectively) and are likely to be regulated differently from other proteins in this cluster.

Fungal and animal FCP1 proteins have in addition to the protein phosphatase catalytic domain a downstream phospho-protein binding BRCT (BRCA-related C-terminal) domain. The only sequences in our dataset containing this domain are in the “FCP Assemblage”, confirming the relationship of these algal and plant sequences to the FCP1s.
However, while most sequences contain the BRCT domain, some do not. Fifteen of the 24 sequences in the cluster contain it, with the exception being the small Arabidopsis-only sequence cluster (5 sequences), the three algal sequences Ot03g00770, Cre141879, Cre149314 and the P. trichocarpa sequence Pop560900, which is likely a CPL1/2 relative, but was included here because of its ambiguity. Multiple sequence alignment of the BRCT domain sequences show them to be well conserved, and we would therefore expect them to be functional (data not shown). While the Arabidopsis sequences without BRCT domains are likely a result of secondary loss of the domain, the algal sequences could also be an indicator of the original state of the "FCP-like" proteins, before gaining the BRCT domain.

Limited study of the Arabidopsis CPL3 and CPL4 proteins sheds some light on the comparative function of these proteins in plants. Both of these proteins contain a functional BRCT domain, which binds to AtRAP74, a homolog of animal/yeast TFIIF (Bang et al., 2006). Knockout plants for CPL3 display hyperactivation of ABA (abscisic acid) mediated transcription, as well as a general alteration of plant growth and maturation, which can be duplicated with mutations to either the BRCT or catalytic domains (Koiwa et al., 2002). RNAi knockdown of CPL4 also leads to plant growth and maturation defects (Bang et al., 2006).

Subcluster I

This cluster is composed of ten plant and algal sequences (O07g01850, A4g261190.1, Pop290290, A3g29760.1, Pop296466, Pop195150, O3g54870.1, O3g54850.1, Ot05g00010, Ot05g00160; Cre166215 is also included depending on tree). This group receives a range of support from the three tree inference methods (99.5% NJ; 60.4% Pars; 68.1% ML; Cre166215 is included in this subcluster in the maximum parsimony and maximum likelihood trees). The sequences do not seem to have any significant relation outside of the catalytic domain.

Subcluster J

This cluster is composed of eight plant, animal and algal sequences (Cre115803, Ot01g02920, MGC10067Hu [also known as UBLC1P1; ubiquitin like domain containing C-terminal phosphatase 1](Zheng et al., 2005), O01g6450, O1g65450.2, A4g06599.1, Pop274109, Pop830623). This group receives high to moderate support from the three tree
inference methods (100% NJ; 99.0% Pars; 58.2% ML). This group is defined by the presence of a ubiquitin-like domain on the N-terminus of the proteins. This domain is listed in the NCBI conserved domain database, as cd01813, as shared with ubiquitin specific proteases, however all entries appear to be C-terminal domain phosphatase homologues. Despite this, these proteins do have what appears to be a proteasome interacting motif based on the work of Upadhya and Hegde (2003). The human protein in this group, UBLCP1 (listed on the tree and alignment as MGC10067), has been studied and determined to be a functional CTD-phosphatase with a possible preference for ser-5 (Zheng et al., 2005). The combination of the ubiquitin directed proteolysis and RNA Pol II phosphatase activity, in addition to the apparent conservation, make this group of proteins intriguing, and further study of their role in the cell is awaited.

**CPL1, 2**

CPL1 and CPL2 are C-terminal domain phosphatase-like protein phosphatases initially described in Arabidopsis (At4g21670.1 and At5g01270.1, respectively) (Koiwa et al., 2002; Koiwa et al., 2004). As mentioned above, these FCP-like phosphatases are characterized by the presence of double-stranded RNA binding domain(s) on the C-terminus (two in CPL1 and one in CPL2). We used the Arabidopsis sequences to search the target protein databases. We found four candidate homologues in P. trichocarpa (Pop555554, Pop743771, Pop90064, Pop560900), and six candidate homologues in O. sativa (O02g42600, O01g63820, O1g63820.2, O4g44710.1, O4g44710.2, Os38346621; the last being a possible isoform of the previous two). The multiple sequence alignment encompassing these full-length sequences and that of the Arabidopsis proteins is presented in Supplement Figure S8. From an inspection of the C-terminal region of this alignment, it is evident that six of these newly identified sequences have two full predicted RNA binding domains, with a high degree of similarity to CPL1_Ath, and are therefore CPL1s (O02g42600, Os38346621, O4g44710.1, O4g44710.2, Pop555554, Pop743771). In contrast, two of the new sequences (Pop90064 and Pop560900) have a greatly truncated second RNA binding region (very similar to CPL2_Ath) and are therefore CPL2 proteins. The situation with the remaining new sequence is more complex.
The sequence O01g63820 (both isoforms) occupies an intermediate position between the well-defined CPL1 and CPL2 clusters in the phylogenetic trees. There is disagreement between tree inference methods as to whether it is included within the CPL1 cluster (NJ) or the CPL2 cluster (Pars). While the protein contains a well conserved second RNA binding domain, the first domain contains a 12 residue deletion within a normally conserved region, requiring experimental confirmation of function. The sequence has other peculiarities, which might preclude it being a functional (protein) phosphatase. There are several prominent deletions (~405-440, ~550-625, ~635-680, ~700-725, ~735-785, as on the scale in Supplementary Figure S8). In addition, two residues which are known to be critical to the activity of yeast FCP1 (the “DD” at about position 405 of the alignment) are not conserved, although they are also not conserved in several other proteins, including the members of the TIM50 subcluster (subcluster G), despite the demonstration of phosphatase catalytic activity of human TIM50 (Guo et al., 2004). However, on balance, the sequence features are most consistent with classification of this sequence as a CPL1, provided it is shown to have activity. This sequence has a particularly convoluted history, as a highly similar sequence was published as “OsCPL2” when originally identified (Koiwa et al., 2004). Through more recent revisions of both genomic and protein databases, this protein (and the apparent isoform, or duplicate Os01g0857000, whose database entry is still provisional) has come to appear more like a divergent CPL1. N-terminal motif analysis shows that the CPL1s and CPL2s are very uniform, and have a common motif signature, which they do not share with the other FCP1-related sequences.

The Arabidopsis proteins CPL1 and CPL2 have been experimentally characterized to some degree, and their isolated catalytic domains are capable of dephosphorylating Ser5 of the Pol II heptad repeat (Koiwa et al., 2004). Deletion of the C-terminus of the CPL1 protein, containing the dsRNA binding domains, creates a cpl1 phenotype, although the function of the domains is not known (Koiwa et al., 2004).

Overall observations of sequences with an FCP-like domain

Proteins containing an FCP-like catalytic domain can be seen as a microcosm of the evolutionary differences between algae, higher plants and animals on a protein level. Every possible combination of conservation is present, with the important exception of plant and...
animal similarity with algal differences. As mentioned above, this places modern algae directly between plants and animals, making them ideal candidates to study the earliest differences between plants and animals.

iii) Asp-based catalysis: Haloacid dehalogenase (HAD)-like

_EYA (Eyes_Absent)_

These protein phosphatases are part of the HAD family. They have been shown to mediate complex morphogenetic events in animal development (Rebay et al., 2005). We used the H. sapiens sequences EYA1 (gi:26667222), EYA2 (gi:26667240), EYA3 (gi:26667243) and EYA4 (gi:98991760) to search the target protein databases. We found one candidate homologue in P. trichocarpa (Pop356606), one in Arabidopsis (At2g35320.1), and one in O. sativa (Os06g02028.1). The Arabidopsis protein possesses aspartate-based catalytic activity (Rayapureddi et al., 2003), however the function of the plant proteins is currently unknown. We found no candidate homologue in C. reinhardtii or O. tauri. The multiple sequence alignment we constructed of catalytic domains is presented as Supplement Figure S9. In Drosophila melanogaster EYA, the prototype of this group, binding occurs between the protein phosphatase domain and the homeobox transcription factor _sine oculis_. A large N-terminal EYA domain then supplies transactivation functions essential for normal eye development (Pignoni _et al._, 1997). However, the plant homologues we have identified, including the Arabidopsis protein, lack the N-terminal domain of the animal proteins, and thus are unlikely to be directly involved in transcriptional activation. The absence of homologues in algae may indicate that, whatever the mechanism of action, higher plant EYAs may mediate functions similar to their animal counterparts, and thus have been lost in the modern green algae. Importantly, this is the sole example of animals and plants having homologues of a protein that is absent in algae.

Chronophin

Chronophin is a member of the HAD superfamily, involved in the activation of the actin filament regulator coflin (Gohla _et al._, 2005). We used the sequence of human chronophin (gi:10092677) to search the target protein databases. We found three potential homologues in C. reinhardtii (Cre77681, Cre127857, Cre142105), two in O. tauri (Ot08g02300, Ot15g02680), three in P. trichocarpa (Pop55442, Pop696747, Pop671977),
three in Arabidopsis (At5g36790.1, At5g36700.1, At5g44760.1), and two in O. sativa (Os09g08660, Os04g41340). The multiple sequence alignment constructed from the catalytic domain region is presented as Supplement Figure S10. In the phylogenetic trees, sequences Cre127857, Cre142105 and Ot15g02680 cluster together with the animal chronophin sequences with high to moderate bootstrap support (100% NJ; 85.4% Pars; 53.0% ML). The sequences Cre77681 and Ot08g02300 cluster with neither the plant nor the animal sequences in two of the three tree inference methods (Pars, ML). To summarize, higher plants seem to have at least one extra chronophin-like protein, and this trend includes the algae studied. However, the plant and animal chronophins cluster separately with phylogenetic study, and the algae seem to be closer related to the animals in this regard.

iv) Gene Expression

Because of the well studied ability of some FCP1-like protein phosphatases to modify the phosphorylation state of RNA pol II, and thus to alter the dynamics of messenger RNA transcription, Bang et al. (2006) suggested that they might be able to act as regulators of gene expression. To investigate this possibility further, we analyzed the Affimetrix microarray expression data available for probes from this gene set. The results are summarized in Table III. For eight of the fourteen gene probes examined, there proved to be highly correlated gene sets. To further dissect the data, we defined three arbitrary categories of correlated genes: protein kinases/phosphatases, components of the ubiquitination/proteolysis system, and putative transcription factors. Our rationale was that these proteins are capable of post-translational effects that would amplify the significance of potential gene regulatory networks.

The data for correlated gene expression for the FCP1-like C-terminal domain protein phosphatases present an interesting and varied pattern. The number of highly correlated probes varied from 0 to several hundred (Table III). The sets of “top 100” correlated probes for all the FCP1-like driver gene probes contain substantial numbers of potential regulatory protein gene probes. There are between 15 (At2g33540 [255843_at]) and 37 (At1g43600/At1g43610 [262720_s_at]) found in each FCP1-like driver gene correlated probe set. Furthermore, the “balance” of gene probes in the three categories is
quite varied. The statistical significance of the number of probes identified in each category was also determined, as detailed in Methods. Two drivers had a “very highly significant” number of correlated probes (p<1E-07), three had a “highly significant” number (p<1E-04), five had a “statistically significant” number (p<0.01), with the remaining not statistically significant (p>0.01). Finally, it should be pointed out that for each of the genes in the FCP1-like set, there is a single tissue, or a small set of tissues, which display a greatly enhanced level of expression (with the exception of At3g29760, which shows relatively ubiquitous expression). The limited protein data available (for CPL3 and CPL4) lends some support to this, with expression mostly in the roots, when compared to shoots (Bang et al., 2006). This is in contrast to the other more highly conserved genes in this study (e.g. EYAs, SSU72s), where there is more uniformly ubiquitous gene expression (data not shown).

v) Promoter Analysis

The analysis of a group of genes with highly correlated expression may serve to elucidate possible functions and common regulatory mechanisms for expression. One of the first demonstrations of this concept was for yeast and human gene sets, where the statistical measure of co-expression was hierarchical sequence clustering (Eisen et al., 1998). The results clearly established that groups of genes that share common expression patterns, also share common functions. This allows inferences to be made based on previous knowledge of gene function within the set. A similar type of analysis allowed the identification of clusters of circadian-regulated genes in Arabidopsis, and, with analysis of upstream sequences, the identification of the responsible promoter, “Evening element” (Harmer et al., 2000). More recently, another statistical measure of gene co-expression, the Pearson correlation coefficient, has been used to document gene sets enriched in cell wall synthetic enzymes (Jen et al., 2006), and genes responsive to illumination with red light (Manfield et al., 2006). Common promoter motifs were shown to be shared by cold-responsive genes, and other genes in a highly correlated expression set (Jen et al., 2006).

We examined sets of genes whose expression was positively correlated with that of “driver” genes in the FCP1-like gene tree for enrichment of characterized promoter elements (P<10^-3). The results are summarized in Supplement Table S2. In general terms
these might be said to fall into a few major categories (stress response, development/proliferation, and defense). In broad outlines, there are apparent similarities between the promoter elements enriched in the correlated gene sets for CPL3 (At2g33540), CPL2 (At5g01270), and At5g11860 (in FCP-like subcluster 3). Elements associated with abscisic acid (ABA) predominate. Indeed, CPL3 is one of the best studied of the Arabidopsis FCP-like gene set, and based on functional data it has been proposed to be primarily an ABA response gene (Koiwa et al., 2002). It would be logical for other genes in its regulatory network to have similar characteristics. In contrast, CPL1 (At4g21670) has been shown to be a negative modulator of various stress responses, and mutations produce growth and maturation defects distinct from CPL3 (Koiwa et al., 2002). We find that a distinct set of promoter elements is enriched in the correlated gene set for CPL1. This is consistent with a regulatory gene network responding to different conditions than that for CPL3. Gene sets whose expression is correlated with driver probe 257378_s_at (At2g02290 and At5g23470) and driver probe 262720_s_at (At1g43600 and At1g43610) (FCP1-like subcluster 6) contain promoters enriched for the “telo-box” element. This is a motif with similarity to telomeric chromosomal sequences, which is found in promoters of genes upregulated during the cell cycle (Tremousaygue et al., 2003). The correlated gene set for At3g55960 contains promoters enriched for the “W-box” motif. This has been characterized as being essential to the activities of the NPR1 plant defense response induction gene (Yu et al., 2001). Finally, the At3g29760 gene (driver probe 257285) includes the “Evening Element”, which is a promoter element found in circadian regulated genes (Harmer et al., 2000). In no case was a single element found to be common to all members of a gene set. This could be explained by the presence of multiple gene subsets correlated with each driver, or that the uniting promoter element has yet to be discovered.

vi) The complete set of Arabidopsis thaliana protein phosphatases

In combination with previous work on Arabidopsis (Kerk et al., 2002; Schweighofer et al., 2004; Kerk et al., 2006), the results of the present study allow a compilation of the complete inventory of the known various types of protein phosphatase present in this organism. Table I shows a comparison of the number of genes encoding protein phosphatases in various structural classes in Arabidopsis and humans. The human data are
derived from Table 1 in Moorhead et al. (2007). Since the initial inventory of Kerk et al. (2002), there have been the following changes: additions (3 PPP family members including one PP1 and one PP6), 9 PPM members [PP2Cs], and 6 DSPs); deletions (one DSP, one PP2C) (Kim et al., 2002; DeLong, 2006; Kerk, 2007).

Conclusion

As key regulatory enzymes, the presence or absence of any particular protein phosphatase can indicate similarities and differences between species. This analysis for the ‘novel’ phosphatases has indicated several key differences and similarities in the function of algae, higher plants and animals. The essential (for animals) cell cycle control enzymes CDC14 and CDC25 seem to have been lost or co-opted for different use in higher plants, while higher plants have increased their numbers of FCP-like proteins. Other classes, such as the LMWPTP, SSU72 and the ubiquitin-like domain containing CTD phosphatases seem remarkably conserved. These data allow insight into the differences and similarities in the function of plants and animals, and how they originated.

Materials and Methods

Identification of Candidate Protein Phosphatase Homologue Sequences

Representative animal sequences from each structural class were obtained from the published research literature, and used as queries in BLASTP searches (Altschul et al., 1997). Databases searched were: A. thaliana (ftp://ftp.arabidopsis.org/home/tair/Sequences/blast_datasets/TAIR7_blastsets/TAIR7_pep_20070425 {04/25/07}); C. reinhardtii (ftp://ftp.jgi-psf.org/pub/JGI_data/Chlamy/v3.1/Chlre3_1.GeneCatalogProteins.6JUL06. fasta.gz proteins.Chlre3.fasta); O. sativa (ftp://ftp.tigr.org/pub/data/Eukaryotic_Projects/o_sativa/annotation_dbs/pseudomolecules/version_5.0/all.chrs/all.pep TIGR Release 5.0 [all.pep {1/23/2007}]); O. tauri (http://bioinformatics.psb.ugent.be/genomes/Ostreococcus_tauri/ [ostreo_pep.tfa {06/12/06}]); P. trichocarpa (ftp://ftp.jgi-psf.org/pub/JGI_data/Poplar/annotation/v1.1/proteins.Poptr1_1.JamboreeModels.fasta.gz {9/13/06}). Sequences returned from the database with the highest scores and the lowest E values (closest to zero) were examined further. Due to some ambiguity in the CDC14
data, sequence structural similarities were assessed by the “fold compatibility” method of comparison to sequences of solved proteins, at the FFAS03 web site (Rychlewski et al., 2000; http://ffas.ljcrf.edu/ffas-cgi/cgi/ffas.pl) This method returns standardized variable (“Z”) scores - a score of greater than 9 is cited by the authors as being statistically significant. To ensure that no distantly related algal or plant homologues were missed by the initial single query sequence based BLAST search strategy, the same databases were searched again in a recursive fashion using Hidden Markov Models constructed from the validated sequence sets from each structural class (see below for details).

Characterization by Multiple Sequence Alignment

The putative protein phosphatase domains of all the candidate homologue sequences for a particular structural subclass, identified in the database search strategy, were placed together in a multiple sequence alignment. The program Muscle (Edgar, 2004) was used, with default parameters. In the case of the dual-specificity phosphatase (DSP) CDC14, a reference set of catalytic domains from A. thaliana DSP proteins was included in the alignment to test whether potential homologues are more closely related to the specific CDC14s or to the general DSP set. A multiple sequence alignment representing the phosphatase domain of each structural subclass was then further examined for characteristic sequence features cited in the research literature, including patterns of conserved critical residues. In some instances additional multiple sequence alignments were also performed with more extensive regions of the protein sequences (i.e. including the non-phosphatase domains), to examine similarity outside the catalytic domain. In the case of the large, heterogeneous FCP1-like sequence set, the final multiple sequence alignment was constructed from a set of smaller sub-alignments. Each of these was constructed using Muscle, and edited in the sequence display program GeneDoc (Nicholas et al., 1997) to remove poorly aligned regions. This process was guided by evaluation at the T-Coffee web server (Poirot et al., 2004; http://tcoffee.vital-it.ch/cgi-bin/Tcoffee/tcoffee_cgi/index.cgi). Sub-alignments were combined, or sequences combined to alignments, using the Profile-Profile or Sequence-Profile alignment features of ClustalX (Thompson et al., 1997) (default parameter settings). The various multiple sequence alignments were used to generate Hidden Markov Models of the proteins of each structural class using the HMMER package.
(Eddy, 1998) (program commands “hmmbuild”, “hmmcalibrate” and “hmmsearch”. These models were then used to search (threshold E=1) through the plant and algal protein databases, and new hits with were added to the alignments and scrutinized in the same manner as the original BLAST hits. Sequences lacking known critical active site residues were removed, with the exception of the FCP1 like proteins and the potential CDC14 plant homologues (sequences removed because of a lack of active site conservation are listed in the individual Supplemental Figure alignments, and sequences included despite lack of active site residues are listed in the legend of Supplemental Figure S7).

Construction of Phylogenetic Trees

Phylogenetic trees were inferred by the Neighbor Joining functionality of ClustalX (Thompson et al., 1997) (default scoring matrix, “exclude positions with gaps” off, “correct for multiple substitutions” off); Maximum Likelihood, as implemented in TreePuzzle (Schmidt et al., 2002) (“unique topologies”, outgroup specified from the dataset, scoring matrix BLOSUM62, 10 thousand puzzling trees); and Maximum Parsimony, as implemented in PHYLIP (Felsenstein, 1996) (randomized sequenced input order and shuffle, multiple datasets [500], other parameters default). Neighbor Joining topologies were generated as the consensus of 1000 bootstrap alignment replicates; Maximum Likelihood topologies represent the consensus of 10,000 puzzling trees; Maximum Parsimony topologies represent the consensus of 500 bootstrap alignment replicates. Nodes are presented which exceed 50% support in at least two of the three tree inference procedures.

Characterization of non-phosphatase domains

Additional non-phosphatase domains were identified in some instances from citations in the literature, in other cases by searching. RPS-BLAST (default settings [E=10]) (Altschul et al., 1997) was used with the COG (Tatusov et al., 2003), Smart (Letunic et al., 2006), and CDD (Marchler-Bauer et al., 2007) datasets. The HMMER package (default gathering cutoff threshold) (Eddy, 1998) was used to search with the Hidden Markov Models of the Pfam database (Bateman et al., 2004; Pfam_Ls_21 (ftp://selab.janelia.org/pub/Pfam/)). Non-phosphatase sequence regions were characterized by motif analysis with MEME and MAST (Bailey and Elkan, 1995; Bailey and Gribskov, 1998; Bailey et al., 2006). MEME was run with the “zoops” model, default motif length,
and number of motifs set to 10. Motifs identified in MAST were included if they met the
default scoring threshold of p<0.0001.

Determination of Evidence for Candidate Homologue Gene Expression

Candidate homologue protein sequences were examined for evidence of gene
expression by using a variety of data types. Each protein sequence was used as the query in
a TBLASTN search against the appropriate EST database (see below). In addition,
Arabidopsis microarray data were examined (see next section), as well as MPSS data
(Meyers et al., 2004; http://mpss.udel.edu/at; http://mpss.udel.edu/rice/) from A. thaliana
and O. sativa. Sequences were included from A. thaliana and O. sativa only if there was a
strong hit with a database EST from that species. Since EST representation is so much
poorer for the other organisms in this dataset, sequences were included lacking a species-
specific EST hit, if a strong hit were obtained by the query sequence to an EST sequence in
another species within the same genus (for example, a P. trichocarpa query sequence
returning a strong EST hit in another species of Populus). Because of the dearth of EST
data, sequences were also included with no expression data. These candidate homologue
sequences are marked as provisional (grey) in Supplemental Table I.

Mining of Microarray Gene Expression Data

Affimetrix microarray data within the NASC dataset (Craigon et al., 2004) were
analyzed. Probe identities were obtained from input AGI gene numbers at the “Arabidopsis
Coexpression Data Mining Tools” web site (Jen et al., 2006; Manfield et al., 2006;
http://www.arabidopsis.leeds.ac.uk/act/index.php). Analysis of correlated probes was
performed using the “Coexpression Analysis over Available Array Experiments” option.
Tabulated correlation values (Pearson correlation coefficients, "r") were rounded to three
decimal places to save space. Also provided by the web site for each correlated probe is an
accompanying “P value” (the probability of obtaining an “r” value of the stated magnitude
from the microarray database by chance alone) and an “E value” (the number of times an
“r” value of the stated magnitude would be obtained from a random sampling of the
microarray database). A probe whose expression is “highly correlated” with the given
driver probe was arbitrarily defined in a very conservative fashion (to minimize false
positives) as one where P=0 and E=0. The annotations for the top 100 correlated probes
were examined for each “driver” (e.g. input) probe, and classified into three groups: “Protein kinases/protein phosphatases”, “Ubiquitination/Proteolysis System”, and “Transcription Factors”, based upon sequence annotation (criteria for each group presented in next section). Correlated gene sets for each “driver” gene probe are presented as Supplemental Table S3. Spatial patterns of gene expression were examined using tools at the Genevestigator site (Zimmermann et al., 2005; https://www.genevestigator.ethz.ch/)). The “Meta-Profile” option was used to determine sites of maximal gene expression.

**Statistical Determination of “Over-Represented” Gene Probes**

Table III presents results showing the number of gene probes in each of the three functional groups described above which are highly correlated with driver genes in our dataset. To assess the significance of these observations, we used the method described in Jen et al., 2006. The probability of obtaining the stated number (“k” [given in Table III]) of gene probes by chance from a dataset containing “N” total gene probes, with “R” gene probes of the same functional type as the sample “k” is given by the hypergeometric distribution. This is given by the density function:

\[ P(x;N,R,k) = \frac{C(R,x)C(N - R,k - x)}{C(N,k)} \]

where \( C(n,m) \) is the binomial coefficient representing the number of combinations of “m” objects which can be drawn from a population of “n” objects. Obtaining this probability is the equivalent of the Fisher exact test. We performed the calculation using the “HYPGEOMDIST” function of MS Excel. A value of \( p<0.01 \) was deemed to be statistically significant. Values for the parameters “N” and “R” were obtained as detailed below.

**Generation of probe lists for functional protein classes**

Affymetrix gene probe sets were downloaded from the ACT site, and purged of duplicates arising from cross-hybridization. This resulted in a large “22k” probe set containing “N”=21890 probes, and a small “8k” probe set containing “N”=6134 probes. These files were then searched for annotation text features corresponding to three functional protein classes (see below), resulting in probe lists. Each probe set was then purged of duplicates, with the result that non-redundant probe lists were generated for “protein kinases/phosphatases” (“R”= 1084 for the large probe set, “R” = 338 for the small probe set), “ubiquitin/proteolysis proteins” (“R” = 790 for the large probe set), and
“transcription factors” (“R” = 1040 for the large probe set) (the small probe set was only utilized for the first functional class). Text search terms for each functional gene probe class were as follows: “protein kinases/phosphatases” (“protein kinase”, “protein phosphatase”); “ubiquitin/proteolysis proteins” (“ubiquitin-specific protease”, “PF01485” [IBR domain], “PF00646” [F-box domain], “PF00097” [zf-C3HC4 RING finger ubiquitin ligase domain], “PF00443” [UCH ubiquitin carboxy terminal hydrolase], “IPR000626” [ubiquitin], “PF04564” [U box], “TIGR01640” [F-box protein interaction domain], “Ubiquitin-related”, “Ubiquitin-like”, “PS00518” [Zinc finger RING-type signature], “F-box family”; “transcription factors” (“PF01529” [zf-DHHC zinc finger], “IPR001965” [Znf-PHD C4HC3 zinc finger], “PF00642” [zf-CCCH zinc finger], “PF00096” [zf-C2H2 zinc finger], “IPR001487” [Bromo domain], “PF00249” [Myb-like DNA binding domain], “PF00010” [Helix-loop-helix DNA binding domain], “PF00098” [zinc knuckle], “PF00170” [basic-leucine zipper (bZIP) transcription factor], “TATA-binding protein”, “MADS-box”, “homeodomain”, “basic helix-loop-helix”, “transcription factor”).

Promoter Element Analysis

Promoter element architecture was investigated using the Athena web server (O'Connor et al., 2005; http://www.bioinformatics2.wsu.edu/cgi-bin/Athena/cgi/home.pl). Sets of genes correlated with expression of “driver” genes from the current dataset (ACT site, above) were entered into the “Visualization” tool. Promoter elements enriched in the dataset were harvested from the “Enriched TF Sites” panel (p<1E-03). Links following up these promoter elements led to the PLACE (Higo et al., 1999; http://www.dna.affrc.go.jp/PLACE/) or Atcisdb sites (Molina and Grotewold, 2005; http://arabidopsis.med.ohio-state.edu/AtcisDB/).

Acknowledgements

The authors wish to thank Dr. Iain Manfield, Centre for Plant Sciences, University of Leeds, for helpful and stimulating discussion on gene expression, promoter element analysis, and statistical analysis of over-represented gene probes.

Literature Cited
Alonso A, Sasin J, Bottini N, Friedberg I, Friedberg I, Osterman A, Godzik A, Hunter T, Dixon J, Mustelin T (2004) Protein tyrosine phosphatases in the human genome. Cell 117: 699-711

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389-3402

Archambault J, Chambers RS, Kobor MS, Ho Y, Cartier M, Bolotin D, Andrews B, Kane CM, Greenblatt J (1997) An essential component of a C-terminal domain phosphatase that interacts with transcription factor II in Saccharomyces cerevisiae. Proc Natl Acad Sci U S A 94: 14300-14305

Bailey TL, Gribskov M (1998) Combining evidence using p-values: application to sequence homology searches. Bioinformatics 14: 48-54

Bailey TL, Williams N, Misleh C, Li WW (2006) MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic Acids Res 34: W369-373

Bailey TM, Elkan C (1995) Unsupervised Learning of Multiple Motifs in Biopolymers Using Expectation Maximization. Machine Learning 21: 51-80

Bang W, Kim S, Ueda A, Vikram M, Yun D, Bressan RA, Hasegawa PM, Bahk J, Koiva H (2006) Arabidopsis carboxyl-terminal domain phosphatase-like isoforms share common catalytic and interaction domains but have distinct in planta functions. Plant Physiol 142: 586-594

Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, Studholme DJ, Yeats C, Eddy SR (2004) The Pfam protein families database. Nucleic Acids Res 32: D138-141

Bleeker PM, Hakvoort HW, Bliek M, Souer E, Schat H (2006) Enhanced arsenate reduction by a CDC25-like tyrosine phosphatase explains increased phytochelatin accumulation in arsenate-tolerant Holcus lanatus. Plant J 45: 917-929

Bliska JB, Clemens JC, Dixon JE, Falkow S (1992) The Yersinia tyrosine phosphatase: specificity of a bacterial virulence determinant for phosphoproteins in the J774A.1 macrophage. J Exp Med 176: 1625-1630

Boudolf V, Inze D, De Veylder L (2006) What if higher plants lack a CDC25 phosphatase? Trends Plant Sci 11: 474-479

Caenepeel S, Charydczak G, Sudarsanam S, Hunter T, Manning G (2004) The mouse kinome: discovery and comparative genomics of all mouse protein kinases. Proc Natl Acad Sci U S A 101: 11707-11712

Chai MF, Chen QJ, An R, Chen YM, Chen J, Wang XC (2005) NADK2, an Arabidopsis chloroplastic NAD kinase, plays a vital role in both chlorophyll synthesis and chloroplast protection. Plant Mol Biol 59: 553-564

Champion A, Kreis M, Mockaitis K, Picaud A, Henry Y (2004) Arabidopsis kinome: after the casting. Funct Integr Genomics 4: 163-187

Cohen P (2002) The origins of protein phosphorylation. Nat Cell Biol 4: E127-130

Craigon DJ, James N, Okyere J, Higgins J, Jotham J, May S (2004) NASCArrays: a repository for microarray data generated by NASC’s transcriptomics service. Nucleic Acids Res 32: D575-577

DeLong A (2006) Switching the flip: protein phosphatase roles in signaling pathways. Curr Opin Plant Biol 9: 470-477
Derelle E, Ferraz C, Rombauts S, Rouze P, Worden AZ, Robbens S, Partensky F, Degroeve S, Echeynie S, Cooke R, Saeys Y, Wuyts J, Jabbari K, Bowler C, Panaud O, Piegu B, Ball SG, Rai JP, Bouget FY, Piganeau G, De Baets B, Picard A, Delseny M, Demaille J, Van de Peer Y, Moreau H (2006) Genome analysis of the smallest free-living eukaryote Ostreococcus tauri unveils many unique features. Proc Natl Acad Sci U S A 103: 11647-11652

Dhankher OP, Rosen BP, McKinney EC, Meagher RB (2006) Hyperaccumulation of arsenic in the shoots of Arabidopsis silenced for arsenate reductase (ACR2). Proc Natl Acad Sci U S A 103: 5413-5418

Eddy SR (1998) Profile hidden Markov models. Bioinformatics 14: 755-763

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792-1797

Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A 95: 14863-14868

Ellis DR, Gumaelius L, Indriolo E, Pickering IJ, Banks JA, Salt DE (2006) A novel arsenate reductase from the arsenic hyperaccumulating fern Pteris vittata. Plant Physiol 141: 1544-1554

Gohla A, Birkenfeld J, Bokoch GM (2005) Chronophin, a novel HAD-type serine protein phosphatase, regulates coflin-dependent actin dynamics. Nat Cell Biol 7: 21-29

Gray CH, Good VM, Tonks NK, Barford D (2003) The structure of the cell cycle protein Cdc14 reveals a proline-directed protein phosphatase. Embo J 22: 3524-3535

Guo Y, Cheong N, Zhang Z, De Rose R, Deng Y, Farber SA, Fernandes-Alnemri T, Alnemri ES (2004) Tim50, a component of the mitochondrial translocator, regulates mitochondrial integrity and cell death. J Biol Chem 279: 24813-24825

Harmer SL, Hogenesch JB, Straume M, Chang HS, Han B, Zhu T, Wang X, Kreps JA, Kay SA (2000) Orchestrated transcription of key pathways in Arabidopsis by the circadian clock. Science 290: 2110-2113

Hausmann S, Erdjument-Bromage H, Shuman S (2004) Schizosaccharomyces pombe carboxyl-terminal domain (CTD) phosphatase Fcp1: distributive mechanism, minimal CTD substrate, and active site mapping. J Biol Chem 279: 10892-10900

Hausmann S, Shuman S (2003) Defining the active site of Schizosaccharomyces pombe C-terminal domain phosphatase Fcp1. J Biol Chem 278: 13627-13632

Higo K, Ugawa Y, Iwamoto M, Korenaga T (1999) Plant cis-acting regulatory DNA elements (PLACE) database: 1999. Nucleic Acids Res 27: 297-300

Huang TY, DerMardirossian C, Bokoch GM (2006) Cofilin phosphatases and regulation of actin dynamics. Curr Opin Cell Biol 18: 26-31

Jen CH, Manfield IW, Michalopoulos I, Pinney JW, Willats WG, Gilmartin PM, Westhead DR (2006) The Arabidopsis co-expression tool (ACT): a WWW-based tool and database for microarray-based gene expression analysis. Plant J 46: 336-348

Kerk D (2007) Genome-scale discovery and characterization of class-specific protein sequences: an example using the protein phosphatases of Arabidopsis thaliana. Methods Mol Biol 365: 347-370
Kerk D, Bulgrien J, Smith DW, Barsam B, Veretnik S, Gribskov M (2002) The complement of protein phosphatase catalytic subunits encoded in the genome of Arabidopsis. Plant Physiol 129: 908-925

Kerk D, Conley TR, Rodriguez FA, Tran HT, Nimick M, Muench DG, Moorhead GB (2006) A chloroplast-localized dual-specificity protein phosphatase in Arabidopsis contains a phylogenetically dispersed and ancient carbohydrate-binding domain, which binds the polysaccharide starch. Plant J 46: 400-413

Khadaroo B, Robbens S, Ferraz C, Derelle E, Eychenie S, Cooke R, Peaucellier G, Delseny M, Demaille J, Van de Peer Y, Picard A, Moreau H (2004) The first green lineage cdc25 dual-specificity phosphatase. Cell Cycle 3: 513-518

Kim DH, Kang JG, Yang SS, Chung KS, Song PS, Park CM (2002) A phytochrome-associated protein phosphatase 2A modulates light signals in flowering time control in Arabidopsis. Plant Cell 14: 3043-3056

Kim Y, Gentry MS, Harris TE, Wiley SE, Lawrence JC, Jr., Dixon JE (2007) A conserved phosphatase cascade that regulates nuclear membrane biogenesis. Proc Natl Acad Sci U S A 104: 6596-6601

Kobor MS, Archambault J, Lester W, Holstege FC, Gileadi O, Jansma DB, Jennings EG, Kouyoumdjian F, Davidson AR, Young RA, Greenblatt J (1999) An unusual eukaryotic protein phosphatase required for transcription by RNA polymerase II and CTD dephosphorylation in S. cerevisiae. Mol Cell 4: 55-62

Koh CG, Oon SH, Brenner S (1997) Serine/threonine phosphatases of the pufferfish, Fugu rubripes. Gene 198: 223-228

Koiwa H, Barb AW, Xiong L, Li F, McCully MG, Lee BH, Sokolchik I, Zhu J, Gong Z, Reddy M, Sharkhuu A, Manabe Y, Yokoi S, Zhu JK, Bressan RA, Hasegawa PM (2002) C-terminal domain phosphatase-like family members (AtCPLs) differentially regulate Arabidopsis thaliana abiotic stress signaling, growth, and development. Proc Natl Acad Sci U S A 99: 10893-10898

Koiwa H, Hausmann S, Bang WY, Ueda A, Kondo N, Hiraguri A, Fukuhara T, Bahk JD, Yun DJ, Bressan RA, Hasegawa PM, Shuman S (2004) Arabidopsis C-terminal domain phosphatase-like 1 and 2 are essential Ser-5-specific C-terminal domain phosphatases. Proc Natl Acad Sci U S A 101: 14539-14544

Landrieu I, da Costa M, De Veylder L, Dewitte F, Vandepoele K, Hassan S, Wieruszeski JM, Corellou F, Faure JD, Van Montagu M, Inze D, Lippens G (2004) A small CDC25 dual-specificity tyrosine-phosphatase isoform in Arabidopsis thaliana. Proc Natl Acad Sci U S A 101: 13380-13385

Letunic I, Copley RR, Pils B, Pinkert S, Schultz J, Bork P (2006) SMART 5: domains in the context of genomes and networks. Nucleic Acids Res 34: D257-260

Lister R, Chew O, Lee MN, Heazlewood JL, Clifton R, Parker KL, Millar AH, Whelan J (2004) A transcriptomic and proteomic characterization of the Arabidopsis mitochondrial protein import apparatus and its response to mitochondrial dysfunction. Plant Physiol 134: 777-789

Manfield IW, Jen CH, Pinney JW, Michalopoulos I, Bradford JR, Gilmartin PM, Westhead DR (2006) Arabidopsis Co-expression Tool (ACT): web server tools for microarray-based gene expression analysis. Nucleic Acids Res 34: W504-509
Manning G, Plowman GD, Hunter T, Sudarsanam S (2002) Evolution of protein kinase signaling from yeast to man. Trends Biochem Sci 27: 514-520
Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S (2002) The protein kinase complement of the human genome. Science 298: 1912-1934
Marchler-Bauer A, Anderson JB, Derbyshire MK, DeWeese-Scott C, Gonzales NR, Gwadz M, Hao L, He S, Hurwitz DI, Jackson JD, Ke Z, Krylov D, Lanczycki CJ, Liebert CA, Liu C, Lu F, Lu S, Marchler GH, Mullokandov M, Song JS, Thanki N, Yamashita RA, Yin JJ, Zhang D, Bryant SH (2007) CDD: a conserved domain database for interactive domain family analysis. Nucleic Acids Res 35: D237-240
Meinhart A, Kamenski T, Hoeppner S, Baumli S, Cramer P (2005) A structural perspective of CTD function. Genes Dev 19: 1401-1415
Meyers BC, Vu TH, Tej SS, Ghazal H, Matvienko M, Agrawal V, Ning J, Haudenschild CD (2004) Analysis of the transcriptional complexity of Arabidopsis thaliana by massively parallel signature sequencing. Nat Biotechnol 22: 1006-1011
Milarski KL, Zhu G, Pearl CG, McNamara DJ, Dobrusin EM, MacLean D, Thiemeseffer A, Zhang ZY, Sawyer T, Decker SJ, et al. (1993) Sequence specificity in recognition of the epidermal growth factor receptor by protein tyrosine phosphatase 1B. J Biol Chem 268: 23634-23639
Molina C, Grotewold E (2005) Genome wide analysis of Arabidopsis core promoters. BMC Genomics 6: 25
Moorhead GB, Trinkle-Mulcahy L, Ulke-Lemee A (2007) Emerging roles of nuclear protein phosphatases. Nat Rev Mol Cell Biol 8: 234-244
Neupert W, Herrmann JM (2007) Translocation of proteins into mitochondria. Annu Rev Biochem 76: 723-749
Nicholas KB, Nicholas HBJ, Deerfield DWI (1997) GeneDoc: GeneDoc: Analysis and Visualization of Genetic Variation. EMBNEW.NEWS 4
O'Connor TR, Dyreson C, Wyrick JJ (2005) Athena: a resource for rapid visualization and systematic analysis of Arabidopsis promoter sequences. Bioinformatics 21: 4411-4413
Pignoni F, Hu B, Zavitz KH, Xiao J, Garrity PA, Zipursky SL (1997) The eyespecification proteins So and Eya form a complex and regulate multiple steps in Drosophila eye development. Cell 91: 881-891
Poirot O, Suhre K, Abergel C, O'Toole E, Notredame C (2004) 3DCoffee@igs: a web server for combining sequences and structures into a multiple sequence alignment. Nucleic Acids Res 32: W37-40
Qian H, Ji C, Zhao S, Chen J, Jiang M, Zhang Y, Yan M, Zheng D, Sun Y, Xie Y, Mao Y (2007) Expression and characterization of HSPC129, a RNA polymerase II C-terminal domain phosphatase. Mol Cell Biochem 303: 183-188
Rayapureddi JP, Kattamuri C, Steinmetz BD, Frankfort BJ, Ostrin EJ, Mardon G, Hegde RS (2003) Eyes absent represents a class of protein tyrosine phosphatases. Nature 426: 295-298
Rebay I, Silver SJ, Tootle TL (2005) New vision from Eyes absent: transcription factors as enzymes. Trends Genet 21: 163-171
Robinson FL, Dixon JE (2006) Myotubularin phosphatases: policing 3-phosphoinositides. Trends Cell Biol 16: 403-412

Rychlewski L, Jaroszewski L, Li W, Godzik A (2000) Comparison of sequence profiles. Strategies for structural predictions using sequence information. Protein Sci 9: 232-241

Satow R, Chan TC, Asashima M (2002) Molecular cloning and characterization of dullard: a novel gene required for neural development. Biochem Biophys Res Commun 295: 85-91

Satow R, Kurisaki A, Chan TC, Hamazaki TS, Asashima M (2006) Dullard promotes degradation and dephosphorylation of BMP receptors and is required for neural induction. Dev Cell 11: 763-774

Schweighofer A, Hirt H, Meskiene I (2004) Plant PP2C phosphatases: emerging functions in stress signaling. Trends Plant Sci 9: 236-243

Sun H, Charles CH, Lau LF, Tonks NK (1993) MKP-1 (3CH134), an immediate early gene product, is a dual specificity phosphatase that dephosphorylates MAP kinase in vivo. Cell 75: 487-493

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhodrov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA (2003) The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4: 41

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 25: 4876-4882

Tremousaygue D, Garnier L, Bardet C, Dabos P, Herve C, Lescurie B (2003) Internal telomeric repeats and 'TCP domain' protein-binding sites co-operate to regulate gene expression in Arabidopsis thaliana cycling cells. Plant J 33: 957-966

Trinkle-Mulcahy L, Lamond AI (2006) Mitotic phosphatases: no longer silent partners. Curr Opin Cell Biol 18: 623-631

Turner WL, Waller JC, Vanderbeld B, Snedden WA (2004) Cloning and characterization of two NAD kinases from Arabidopsis. identification of a calmodulin binding isofrom. Plant Physiol 135: 1243-1255

Upadhya SC, Hegde AN (2003) A potential proteasome-interacting motif within the ubiquitin-like domain of parkin and other proteins. Trends Biochem Sci 28: 280-283

Xu H, Somers ZB, Robinson ML, 2nd, Hebert MD (2005) Tim50a, a nuclear isoform of the mitochondrial Tim50, interacts with proteins involved in snRNP biogenesis. BMC Cell Biol 6: 29

Yaffe MB, Rittinger K, Volinia S, Caron PR, Aitken A, Leffers H, Gamblin SJ, Smardon SJ, Cantley LC (1997) The structural basis for 14-3-3:phosphopeptide binding specificity. Cell 91: 961-971

Yu D, Chen C, Chen Z (2001) Evidence for an important role of WRKY DNA binding proteins in the regulation of NPR1 gene expression. Plant Cell 13: 1527-1540

Zheng H, Ji C, Gu S, Shi B, Wang J, Xie Y, Mao Y (2005) Cloning and characterization of a novel RNA polymerase II C-terminal domain phosphatase. Biochem Biophys Res Commun 331: 1401-1407
Zimmermann P, Hennig L, Gruissem W (2005) Gene-expression analysis and network discovery using Genevestigator. Trends Plant Sci 10: 407-409

Figure 1: Phylogenetic Tree of CDC14-Like Sequence Relationships

A rectangular cladogram was generated by comparing catalytic domains of CDC14 like proteins (red) with the closest relatives in plants (blue), using the set of A. thaliana DSP proteins as an outgroup (black, from Kerk (2006)). Proteins included are from the following organisms, with the source of the sequences in brackets: A. thaliana (MIPS code without "t"); C. reinhardtii (Crexxxxxx, where xxxxxx is the protein ID from http://plantsp.genomics.purdue.edu/plantsp/data/proteins.Chlre3.fasta); H. sapiens (CDC14A_Hs: NP_003663, CDC14B_Hs:NP_201588); O. sativa (MIPS code); O. tauri (MIPS codes given by https://bioinformatics.psb.ugent.be/gdb/ostreococcus/); P. trichocarpa (Popxxxxxx, where xxxxxxx is the protein ID from the DOE JGI); X. laevis (CDC14A_Xl: NP_001084450, CDC14B_Xl: NP_001084486) and. Multiple sequence alignment construction and phylogenetic tree inference was performed as detailed in Methods. The tree topology shown is that from Neighbor Joining, where 1000 replicates were performed. The CDC14 proteins (red) form a clade [Node “A”: 100% NJ; 98.8% Pars; 78.2% ML] which is distinct from the clade formed by the most closely related plant proteins (blue) [Node “B”: 99.1% NJ; 98.4% Pars, 82.7% ML]. This suggests distinct function, which is discussed in the text. These two groups are related to the exclusion of the set of Arabidopsis DSP proteins [Node “C”: 88.6% NJ; 95% Pars; 40.2% ML]. All other nodes in the tree figure show replicate support from Neighbor Joining only.

Figure 2: Phylogenetic Tree of CDC25-Like Sequence Relationships

A rectangular cladogram was generated by comparing catalytic domains of CDC25 like proteins (red) with the closest relatives in plants and fungi (blue). Proteins included are from the following organisms, with the source of the sequences in brackets: A. thaliana (MIPS code without "t"); C. reinhardtii (Crexxxxxx, where xxxxxx is the protein ID from http://plantsp.genomics.purdue.edu/plantsp/data/proteins.Chlre3.fasta); D. rerio (Drxxxxxxx, where xxxxxxx is the gi); H. sapiens (CDC25A_Hu:NP_001780, CDC25B_Hu:NP_068659, CDC25C_Hu:NP_001781); L. major (LmACR2: genbank AAS73185); O. sativa (MIPS code without "s"); O. tauri (MIPS codes given from https://bioinformatics.psb.ugent.be/gdb/ostreococcus/); P. trichocarpa (Popxxxxx, where xxxxxx is the protein ID from the DOE JGI); P. vittata (PvACR2: genbank ABC26900); S. cerevisiae (ScCDC25: NP_013750, ScACR2: NP_015526); S. pombe (SpCDC25: NP_592947, SpACR2:NP_595247); X. laevis (Xlexxxxxx, where xxxxxxxx is the gi,CDC25A_Xle: NP_001081257). Multiple sequence alignment construction and phylogenetic tree inference was performed as detailed in Methods. The tree topology shown is that from Maximum Likelihood, where 10,000 replicates were performed. The known CDC25 proteins (red) form a clade with the sequence from O. tauri [Node “A”: 100% NJ; 97.8% Pars; 75.9% ML], while the most closely related plant proteins cluster with the arsenate reductases (blue; see text for details) [Node “B”].
Figure 3: Phylogenetic Tree of FCP1-Like Sequence Relationships

A rectangular cladogram was generated by comparison of catalytic domains from FCP/SCP catalytic domain containing proteins from the following species: A. thaliana (MIPS code without "t", with the following exceptions, CPL3_Ath:At2g33540, CPL4_Ath:At5g58003); C. reinhardtii (Crexxxxxx, where xxxxxx is the protein ID from http://plantsp.genomics.purdue.edu/plantsp/data/proteins.Chlre3.fasta); D. rerio (Drxxxxxxx, where xxxxxxx is the gi, except Dullard_Dr:NP_001007310); H. sapiens (SCP1_Hu: NP_067021, SCP2_Hu:NP_005721, SCP3_Hu:NP_001008393, DULLARD_Hu:NP_056158, Hu68a1840:AAF29093, FCP1_Hu:NP_004706, MGC10067Hu:NP_659486 (also known as UBLCP1), TIM50Hu: NP_001001563); O. sativa (MIPS code without "s" with the following exceptions, OsCPL3: Os11g31890, OsCPL4:Os05g32430); O. tauri (MIPS codes given from https://bioinformatics.psb.ugent.be/gdb/ostreococcus/); P. trichocarpa (Popxxxxxx, where xxxxx is the protein ID from the DOE JGI); S. pombe (FCP1_Spombe:NP_594768); X. laevis (FCP1_Xle:NP_001081726); X. tropicalis (Xxxxxxxxxx, where xxxxxxxx is the gi, with the following exceptions from Ensembl, 39992_Xtr:ENSXETP00000039992, 32705_Xtr:ENSXETP00000032705). Multiple sequence alignment construction and phylogenetic tree inference was performed as detailed in Methods. The tree topology shown is that from Neighbor Joining, where 1000 replicates were performed. The proteins segregate into 10 subclusters, which are labeled, colour coded and discussed in the text. The support for each of the labeled nodes is as follows: Node “A” (99.7% NJ; 44.8% Pars; 34.7% ML); Node “B” (96.7% NJ; 48.8% Pars; 70.1% ML); Node “C” (100% NJ; 100% Pars; 98.4% ML); Node “D” (99.4% NJ; 91.8% Pars; 80.6% ML); Node “E” (99.3% NJ; 31.6% Pars; 60.4% ML); Node “F” (99.7% NJ; 31.2% Pars; 79.4% ML); Node “G” (82.0% NJ; 88.8% Pars; 74.7% ML); Node “H” (64.1% NJ; 67.6% Pars; 43.7% ML; sequences Cre187332, Cre149314 and Pop560900 are missing in the maximum parsimony tree); Node “I” (99.5% NJ; 60.4% Pars; 68.1% ML); Node “J” (100% NJ; 99.0% Pars; 58.2% ML).
| Protein Phosphatase Family | Subclass | Arabidopsis Genes | Human Genes |
|----------------------------|----------|-------------------|-------------|
| **Ser/Thr Phosphatase**    |          |                   |             |
| PPP family                 | Total    | 26                | 13          |
|                           | PP1      | 9                 | 3           |
|                           | PP2A     | 5                 | 2           |
|                           | PP2B     | 0                 | 3           |
|                           | PP4      | 2                 | 1           |
|                           | PP5      | 1                 | 1           |
|                           | PP6      | 2                 | 1           |
|                           | PP7      | 1                 | 2           |
|                           | Otherb   | 6                 | N/A         |
|                           | PPM family (PP2C) | 76 | 18 |
| **PTP Superfamily**        |          |                   |             |
| (CX5R)                     |          |                   |             |
| Class I PTP (classic)      | Total    | 1                 | 38          |
|                           | Receptor | 0                 | 21          |
|                           | Non-Receptor | 1 | 17 |
| Class I PTP (DSPs)        | Total    | 22                | 61          |
|                           | MAPKP    | 0                 | 11          |
|                           | Slingshots | 0 | 3 |
|                           | PRLs\(^a\) | 0 | 3 |
|                           | Atypical DSP\(^a\) | 3 | 19 |
|                           | CDC14    | 0                 | 4           |
|                           | PTEN\(^a\) | 3 | 5 |
|                           | Myotubularins\(^a\) | 2 | 16 |
|                           | Other\(^c\) | 14 | N/A |
| Class II PTPs (CDC25)     | None     |                   | 3           |
| Class III PTPs (LMWPTP)   | 1        |                   | 1           |
| SSU72                     | 1        |                   | 1           |
| **Asp-based catalysis**   |          |                   |             |
| (DXDXT/V)                 |          |                   |             |
| FCP-Like                  | 19       |                   | 8           |
| HAD family (Chronophins)  | 3        |                   | 1           |
| HAD family (Eyes-Absent)  | 1        |                   | 4           |
| **Total**                 | 150      |                   | 148         |

Table I: Summary of Protein Phosphatase Gene Types in Arabidopsis and Human

Protein phosphatase genes are summarized from the present study, plus the following references: Kerk et al. (2002), Kim et al. (2002), Schweighofer et al. (2004), Kerk (2007), Moorhead et al. (2007). Multiple protein isoforms are often transcribed from the same gene, but are ignored in
this table, for simplicity. A number of additional sequences have been identified which contain similarity to the Arabidopsis genes, but have been rejected because they lack critical class-specific catalytic residues (see Kerk, 2007). Lipid phosphatases and phosphatases of unknown substrate belonging to the various families are included and denoted with a. Phosphatases of known family but with no direct homologues to characterized mammalian proteins are designated with b (includes At1g03445, At1g07010, At1g08420, At1g18480, At2g27210, At4g03080) or c (includes At1g05000, At2g04550, At2g32960, At2g35680, At3g02800, At3g06110, At3g09100, At3g23610, At3g62010, At4g03960, At5g01290, At5g16480, At5g23720, At5g28210, At5g56610). Since the report in Kerk et al. (2002), excluding the data from the present study, the following changes have occurred to the Arabidopsis protein phosphatase gene set: 3 PPPs have been added (At1g18480 [“Bacterial-like”]; At3g19980 [PP6]; At5g43380 [PP1]); 6 DSPs have been added (At3g09100, At3g19420 [“PTEN homologue”], At3g62010, At4g03960, At5g01290, At5g28210); 9 PP2Cs have been added (At2g30170, At2g46920, At4g03415, At4g11040, At4g16580, At4g32950, At4g33500, At5g26010, At5g66720); one DSP has been deleted (At3g55270 – re-annotated with a greatly shortened DSP domain); one PP2C has been deleted (At1g75010 – determined to be a false positive).
| Protein Phosphatase Family | Sequence Cluster | Human | C. reinhardtii | O. tauri | P. trichocarpa | Arabidopsis | O. sativa | Algae vs Others | Othersa |
|---------------------------|------------------|-------|---------------|----------|---------------|-------------|-----------|----------------|---------|
| **PTP Superfamily (CX_{5}R)** |                  |       |               |          |               |             |           |                |         |
| Class I PTP (SSU72s)      |                  | 1     | 1             | 1        | 1             | 1           | 1         | P-Lb           |         |
| Class I PTP (DSPs) Slingshots |            | 3     | None          | None     | None          | None        | None      | P-L           |         |
| Class I PTP (DSPs) CDC14s |                  | 2     | 1             | None     | None          | None        | None      | Both           |         |
| Class II PTP CDC25s       |                  | 3     | None          | 1        | None          | None        | None      | Both           |         |
| Class III PTP LMWPTP      |                  | 1     | 1             | None     | 2             | 1           | 1         | Same           |         |
| **Asp-based catalysis (DXDX_{5}T/V)** |                  |       |               |          |               |             |           |                |         |
| FCP-Like SCPs             |                  | 3     | 1\(^d\)       | 1\(^d\)  | None          | None        | None      | A-L           |         |
| “FCP Assemblage”          |                  |       |               |          |               |             |           |                |         |
| FCP-Like FCP1-Like Tree Groups A |          | 3     | 5             | 2        | 12            | 7           | 9         | Both          |         |
| FCP-Like FCP1-Like Tree Groups B – G |     |       |               |          |               |             |           |                |         |
| FCP-Like FCP1-Like Tree Groups H |            | 1     | 3             | 3        | 2\(^e\)       | 7           | 2         | P-L           |         |
| FCP-Like FCP1-Like Tree Groups I |            | None  | 1\(^f\)       | 2        | 3             | 2           | 3         | P-L           |         |
| FCP-Like FCP1-Like Tree Groups J |            | 1     | 1             | 1        | 2             | 1           | 1         | A-L           |         |
| FCP-Like Unclassified     |                  | None  | 1\(^f\)       | None     | None          | None        | None      | N/A           |         |
| FCP-Like FCP1-Like: CPL1.2s |                | None  | None          | None     | None          | 4           | 2         | 4              |         |
| HAD Family EYAs           |                  | 4     | None          | None     | 1             | 1           | 1         | Neither        |         |
| HAD Family Chronophins    |                  | 1     | 3             | 2        | 3             | 3           | 2         | A-L\(^j\)     |         |
Table II: Protein Phosphatase Summary Data

Human protein phosphatase sequences were used as queries, and candidate homologue sequences were obtained and analyzed, as detailed in Methods. Protein phosphatase classes are modeled after the scheme used in Moorhead et al. (2007). The number of candidate homologues is given for each species and protein phosphatase class, not including splice variants. Where appropriate, further subdivision is made of classes into sequence cluster subgroups. Details of findings are given in Results.

Note “a”: Similarity of algal sequence number/affinity to plant sequences (“P-L” = Plant-Like); animal sequences (“A-L” = Animal-Like); both plant and animal sequences (“Both”); all organisms equivalent (“Same”); or algae similar to neither plants nor animals (“Neither”)

Note “b”: Algal proteins cluster with the plant homologues

Note “c”:* Algal chronophins are plant-like in that there are multiple proteins, but the sequences which can be assigned an affinity are animal-like.

Note “d”:* Algal SCP-like proteins do not have the same bootstrap support as the animal proteins, see Results for details.

Note “e”:* Protein pop560900 is included with both the FCP proteins and CPL1 and 2 proteins for the phylogenetic trees, but is only counted with CPL1 and 2.

Note “f”:* Cre166215 may be included in either “Unclassified” or “Group I”, depending on the tree inference method chosen (here included with “Group I”)
| Sequence Source | Gene | Affimmetrix Microarray Probe “Driver” | Predominant Site of Expression | Number Highly Correlated Probes | r-value: Best/100 probes | Top 100 Probes: Protein Kinases/Phosphatases | Top 100 Probes: Ubiquitin/Proteolysis System | Top 100 Probes: Transcription Factors |
|----------------|------|-------------------------------------|-------------------------------|--------------------------------|--------------------------|-----------------------------------------------|---------------------------------------------|----------------------------------------|
| FCP1-Like: SubclusterB | At5g46410 | 248901_at | Cork Xylem | 56 | 0.836 0.682 | 7 3 | 16 (p<10^-4) |
| | At5g11860 | 250298_at | Xylem Cork | 8 | 0.750 0.590 | 9 5 | 17 (p<10^-5) |
| | FCP1-Like: SubclusterD | At5g45700 | 248963_at | Pollen | 624 | 0.911 0.831 | 12 3 | 5 |
| | At1g29770 | 255993_at | Pollen Stamen | 42 | 0.969 0.640 | 10 5 | 3 |
| | At1g29780 | 255998_at | Lateral Root Petiole | 0 | 0.404 0.312 | 1 4 | 12 (p<0.01) |
| | FCP1-Like: SubclusterE | At3g55960 | 251773_at | Senescent leaf Endodermis | 0 | 0.620 0.444 | 12 (p<0.01) | 3 7 |
| | FCP1-Like: SubclusterH | At2g02290 | 257378_s_at At5g23470 | Pollen | 296 | 0.952 0.818 | 0 17 (p<10^-7) | 5 |
| | | At1g43600 | 262720_s_at At1g43610 | Stamen | 246 | 0.979 0.797 | 1 30 (p=0) | 6 |
| CPL3_4 FCP1-Like SubclusterH | At2g33540 | 255843_at | Xylem | 0 | 0.474 0.363 | 8 2 | 5 |
| | At5g58000 | 247894_at (CPL4) | Root tip hypocotyl | 0 | 0.699 0.576 | 1 8 | 7 |
| | FCP1-Like SubclusterI | At3g29760 | 257285_at | Mostly ubiquitous | 0 | 0.562 0.441 | 11 (p<0.01) | 0 5 |
| | At4g26190 | 254019_at | Callus Root Tip | 6 | 0.757 0.627 | 2 1 | 14 (p<10^-3) |
| CPL1_2 | At4g21670 | 20554_at (CPL1) | Seed Flower | 277 | 0.978 0.974 | 16 (p<10^-4) | 0 2 |
| | At5g01270 | 251134_at (CPL2) | Senescent leaf Root tip | 0 | 0.567 0.428 | 7 5 | 10 |
Table III: Arabidopsis Microarray Gene Expression Summary Data

Affimetrix microarray gene expression data were explored for various gene probes of the Arabidopsis FCP1-Like gene set, as detailed in Methods. Tissue sites of gene expression were obtained from the Genevestigator web site (https://www.genevestigator.ethz.ch/). Sets of genes whose expression is correlated with the input “driver” gene were obtained from the “Arabidopsis Coexpression Data Mining Tools” site (http://www.arabidopsis.leeds.ac.uk/act/index.php). Note “a”: A “highly correlated probe” is arbitrarily defined as one with P=0 and E=0, where P represents the probability of such a correlation coefficient arising in the entire microarray dataset by chance alone, and E represents the number of times a correlation coefficient of the stated value would arise from the entire microarray dataset by chance alone (see Arabidopsis Coexpression Data Mining Tools [http://www.arabidopsis.leeds.ac.uk/act/index.php] for details) Note “b”: Correlation values are Pearson correlation coefficients, rounded to three decimal places to save space  Note “c”: Number of observed probes in each column were analyzed for statistical significance by calculating the probability of a random sampling result using the hypergeometric distribution (this procedure corresponds to the Fisher exact test) as detailed in Methods. Significant probabilities are indicated, other entries in these columns have non-significant probabilities. Note “d”: This probe came from the Arabidopsis 8K gene array (all other probes came from the 22K gene array).
Figure 3.