Evolution of the phosphatase gene family across nematode worms and flies

Paulina Tsai and Melissa A. Wilson Sayres

Department of Integrative Biology,
College of Letters and Science,
University of California at Berkeley

Keywords: evolution, phosphatase, nematodes, Drosophila, phylogeny

Abstract

Phosphatase genes have been shown to be involved in male meiosis in the nematode worm, Caenorhabditis elegans, and are expressed in the testis in the fruit fly, Drosophila melanogaster. However, the evolution of this multi-gene family among nematodes and flies had not previously been investigated. We conducted a phylogenetic analysis of all genes in the phosphatase gene family across nematodes and flies using sequences from a 6-way alignment of nematode worms and a 15-way alignment of insects, including 12 Drosophila species. We found that: 1) multiple alignments contain spurious alignments that should be filtered for quality control; 2) several gene sequences with incomplete open reading frames are highly conserved, so may actually be functional genes; and, 3) the phosphatase gene family appears to have expanded independently in the common ancestor of nematodes, and again in the common ancestor of flies (but not all insects).

Introduction

Protein Phosphatase Type 1 (PP1) is a major eukaryotic protein serine/threonine phosphatase that regulates a large variety of cellular functions and processes, such as glycogen metabolism, muscle contraction, cell division, and many more, through the dephosphorylation of many substrates (1, 2). Phosphorylation of specific serine, threonine, and/or tyrosine residues controls the expression of about one-third of all eukaryotic proteins (3). Over 400 protein serine/threonine kinases and ~25 protein serine/threonine phosphatase have been identified across eukaryotes (4). Global analyses of proteins and protein domains performed on Caenorhabditis elegans reveal that protein kinases comprise the second largest family of protein domains in nematodes (5). It seems that while the number of protein kinases has steadily increased during eukaryotic evolution, the serine/threonine phosphatases have not increased to the same extent, but the diversity of their interacting polypeptides has increased enormously (3).

PP1 has been shown to be necessary for mitotic progression in Drosophila, and that the other loci cannot supply sufficient activity to complement loss of expression of this gene (6). Further research has shown that PP1 genes are essential for male meiosis in C. elegans (7). Curiously, phosphatase genes are also expressed in the testis in many species, including Drosophila melanogaster (8). Both C. elegans, and D. melanogaster have large phosphatase gene families, but it is not clear whether these genes families expanded in the common ancestor of invertebrates, in the common ancestor of nematodes independent of the common ancestor of flies, or only in the C. elegans and D.melanogaster lineages. We used bioinformatics methods and comparative genomics

Reference species	name	tree code	accession	chr	cdsStart	cdsEnd	homologs
C. elegans	C25A6.1	C25	NM_072031	chrV	5408511	5410055	4
C. elegans	C34D4.2	C34	NM_068724	chrIV	7154706	7155913	1
C. elegans	C47A4.3	C47	NM_070249	chrIV	13741009	13742233	1
C. elegans	F23B12.1	F23	NM_074173	chrV	14429030	14430478	1
C. elegans	F25B3.4	F25	NM_073069	chrV	9565600	9568680	3
C. elegans	F52H3.6	F52	NM_063766	chrII	10026571	10027785	2
C. elegans	gsp-1	gp1	NM_073332	chrV	10682502	10685823	5
C. elegans	gsp-2	gp2	NM_001027445	chrIII	7336549	7338116	4
C. elegans	gsp-3	gp3	NM_059028	chrI	4709356	4710422	2
C. elegans	gsp-4	gp4	NM_058836	chrI	3838390	3839433	1
C. elegans	phophatase	phos	NM_072685	chrV	8084099	8085221	4
C. elegans	pph-1	PpH1	NM_073333	chrV	10702539	10704457	6
C. elegans	T16G12.7	T16	NM_066828	chrIII	10038436	10039983	4
C. elegans	Y71G12B.30	Y71	NM_001026652	chrI	1818989	1822932	4
C. elegans	ZK938.1	ZK9	NM_063716	chrII	9829413	9830618	4
D. melanogaster	flw	flw	NM_167229	chrX	10280544	10301512	15
D. melanogaster	Pp1-13C	P13C	NM_080182	chrX	15253461	15254370	12
D. melanogaster	Pp1-87B	P87B	NM_080198	chr3R	8250501	8251410	12
D. melanogaster	Pp1alpha-96A	P96A	NM_079760	chr3R	20344380	20346216	12
D. melanogaster	PP1Y1	Pp1Y1	AF427493	chrU	8549613	8550146	1
D. melanogaster	PP1Y2	PP1Y2	AF427494	chrYHet	279994	291785	1
D. melanogaster	PpD5	PpD5	NM_079968	chr2R	17929317	17930358	14
D. melanogaster	PpD6	PpD6	NM_080208	chr2L	3109719	3110730	9
D. melanogaster	PpN58A	P58A	NM_058036	chr2R	17768978	17769953	7
D. melanogaster	PpY-55A	P55A	NM_057341	chr2R	13843048	13843993	10

Table 1. Phosphatase gene information.

analyses to take advantage of recent large-scale sequencing efforts, to better understand the evolution of the phosphatase gene family. Specifically we asked when, evolutionarily, the phosphatase gene family expanded. In doing so, we also addressed issues relating to sequencing quality and alignment.

Materials and Methods

Sequences

PP1 DNA sequences (Table 1) were downloaded from UCSC for the 6-way nematode alignment and the 15-way Drosophila alignment (7). We analyzed the currently sequenced nematode genomes in the 6-way nematode alignment (Caenorhabditis elegans: WS190/ce6, C. brenneri: WUGSC 6.0.1/caePb2, C. briggsae: WUGSC 1.0/cb3, C. remanei: WUGSC 15.0.1/caeRem3, C. japonica: WUGSC 3.0.2/caeJap, and Pristionchus pacificus: WUGSC 5.0/priPac1), and the insects available in the 15-way insect alignment including 12 Drosophila species, mosquito, honeybee and red flour beetle (Drosophila melanogaster: BDGP R5/dm3; D. simulans: droSim1; D. sechellia: droSec1; D. yakuba: droYak2; D. erecta: droEre2; D. ananassae: droAna3; D. pseudoobscura: dp4; D. persimilis: droPer1; D. willistoni: droWil1; D. virilis: droVir3; D. mojavensis: droMoj3; D. grimshawi: droGri2; Anopheles gambiae: anoGam1; Apis mellifera: apiMel3; Tribolium castaneum: triCas2). We also downloaded the yeast phosphatase gene sequence (Table 1).

Quality control

All phosphatase sequences from the multiple alignments were analyzed within a species. Spurious alignments, where the same genomic location was mapped to more than one gene, were removed, retaining only the sequence with the highest quality alignment. Sequences were analyzed for potential functionality by assessing open reading frames for frameshift and nonsense mutations. Sequences with premature stop codons, frameshift insertions, or

frameshift deletions were tagged as being potential pseudogenes for downstream analysis.

Alignments
Homologous sequences were aligned using ClustalW2 (9), PRANK (10) and MAFFT (11).

Phylogenetic and evolutionary analysis
Consensus neighbor-joining and parsimony trees with 1000 bootstrap replicates each were constructed using phylip (12). We built maximum likelihood trees and computed branch-specific substitution rates using PhyML (13).

Selection analysis
The alignments were screened for putative pseudogenes and internal stop codons, which were removed before computing the sequences into DataMonkey (www.datamonkey.org) to detect site-specific selection (14).

Results

Sequence quality
We assembled homologous sequences of phosphatase genes from 6 nematodes and 15 insects, and conducted sequence similarity searches across them. We used perl programs to compare the genomic mapping of all sequences within a species, and removed sequences where the genomic location overlapped (suggesting an in silico duplication), retaining the sequence with the most sequence coverage relative to the reference. 46 of the 49 nematode sequences were retained, while 81 of 150 insect sequences were retained.

Alignments
We tested three alignment algorithms, ClustalW2, PRANK and MAFFT, aligning both the nucleotides and amino acids with all methods. Similar to previous analyses (14), we observed that PRANKC (PRANK, aligning the codons) and MAFFT outperformed ClustalW2, but observed similar alignments between PRANKC and MAFFT, except that PRANKC failed multiple times to run for the full dataset of 128 sequences. As such, we proceeded with the analysis using the MAFFT alignments.

Phylogenetic analysis
We tested three different tree-building programs: Neighbor Joining (NJ), Maximum Parsimony (MP), and Maximum Likelihood (ML). All three tree-building methods yielded qualitatively similar results; we present the findings from the ML trees.

Selection analysis
We used Datamonkey (14) to determine the mean dN/dS ratio and positive and negative selection on our alignments. Our results yielded a mean dN/dS ratio of 0.18357 across the tree, which is indicative of purifying selection. There were no positively selected sites detected across the alignment of phosphatase genes, but 337 negatively selected sites (with 0.1 significance) were identified. After applying a Bonferroni correction for multiple testing (one test for each of the 561 codons: 0.05*561 codons = corrected p-value of 8.91266 x 10-5 or less) Datamonkey reported 223 negatively selected sites, suggesting that this gene family evolves under strong purifying selection across lineages.

Discussion

This study showed that the phosphatase gene family appears to have expanded independently in the common ancestor of nematodes, and again in the common ancestor of flies. There was an additional lineage-specific expansion of phosphatase genes in C. elegans. In addition, we determined that great care must be taken when using publicly available multiple alignments to avoid in silico errors that may mislead results.

Sequence quality
During preliminary analyses with the multiple alignments for nematodes and insects, we noticed peculiar clusterings. Upon further investigation, we identified that, in a many species, the same genomic location was mapped to several phosphatase homologs. This led to artificially identifying homologs. To remove these in silico alignment errors, we analyzed all paralogous genes in each species individually, and conducted sequence similarity searches. We compared the genomic mapping of all sequences within a species. We retained the sequences with unique mappings, and, in the case of overlapping segments, we retained only the sequence with the highest coverage relative to the reference. 46 of the 49 nematode sequences were retained, while only 81 of 150 insect sequences were retained. It is noteworthy that nearly all sequences identified as orthologs in the mosquito, honeybee and red flour beetle, are excluded using this quality filter. Although it is possible that all of these orthologs were lost independently in mosquito, honeybee, and red flour beetle, these results suggest that these phosphatase genes likely do not have homologs in the common ancestor of all insects, but that any expansions must have happened after Drosophila diverged from other insects.

The inclusion of nonfunctional genes, called

pseudogenes, might bias the results, so we attempted to computationally define sequences that were likely to be nonfunctional. For the sequences that were uniquely mapped, we defined a set of putative pseudogenes as those sequences that had disrupted open reading frames, including missing start or stop codons, or if the nucleotides were not in multiples of three (frameshift insertions or frameshift deletions). We labeled these sequences with a "q" at the beginning of the file names, for "questionable".

Phylogenetic analysis

Maximum likelihood is a parametric statistical method that uses an explicit model of character evolution, which gives a large number of parameters that can account for differences in the probabilities of particular states, changes, and differences in probabilities of change among the characters. This makes ML more powerful than both NJ and MP, non-parametric statistical methods (15). We used PhyML to compute the Maximum Likelihood tree for each set of alignments. PhyML uses Maximum Likelihood algorithm to estimate phylogenies from alignments of nucleotide or amino acid sequences, which we created in MAFFT. PhyML is good for large data sets, has a large number of substitution models coupled to various options to search for phylogenetic tree topologies, and also incorporates branch-length estimation (13).

Assessing putative pseudogenes

We first analyzed the phylogenetic relationship and branch lengths of all sequences (Figure 1). We expected that the putative pseudogenes might have long branch lengths, relative to orthologs in other species with functional open reading frames. This assumes that the pseudogenization event was old, and so the pseudogenized orthologs have been evolving under relaxed constraints (and so might accumulate more mutations, particularly nonsynonymous mutations) than functional coding orthologs. Curiously, however, the branch length analysis shows that many putative

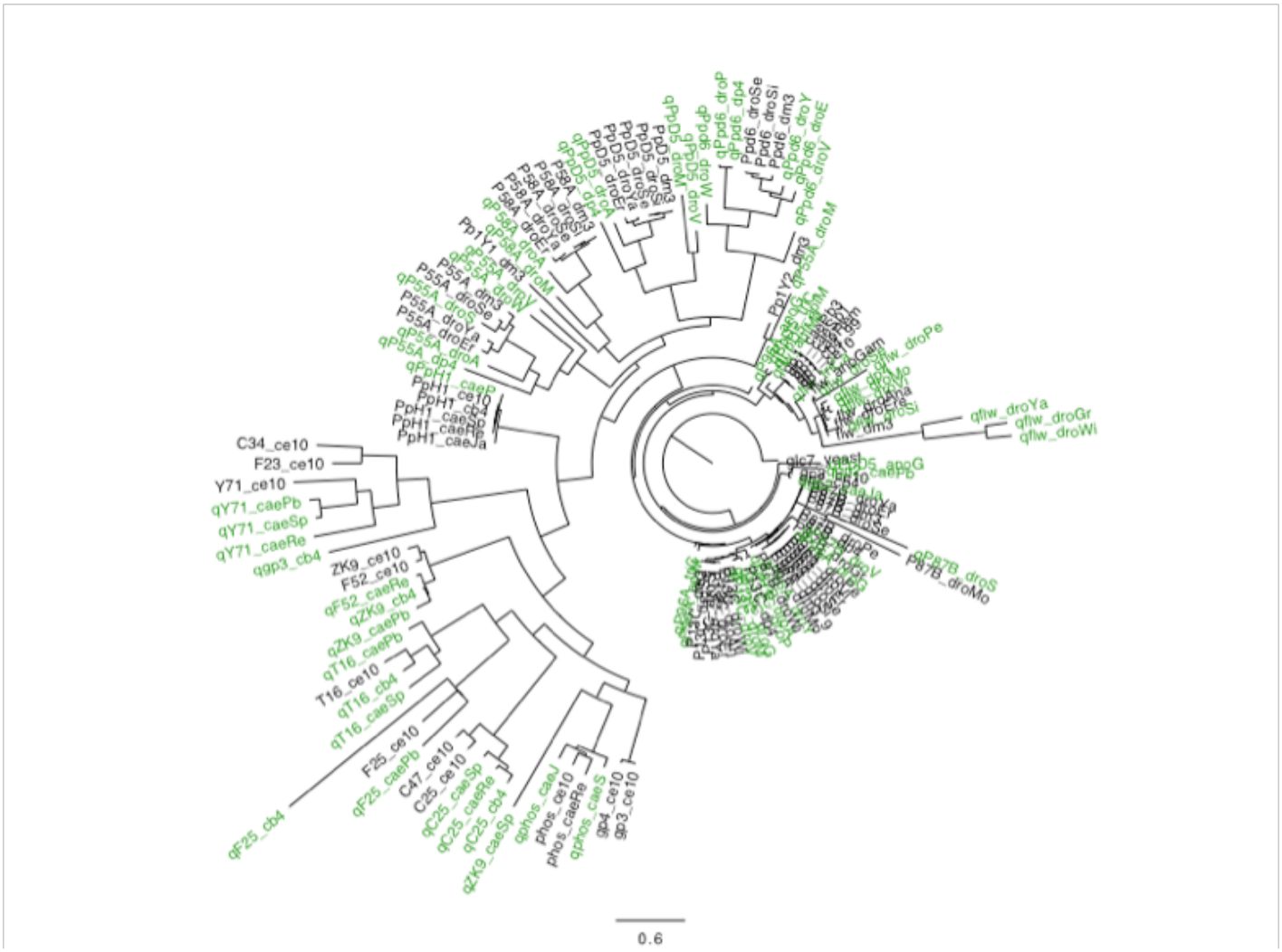

Figure 1

pseudogenes have quite short branch lengths. This would result if the pseudogenization event was very recent, and so there hasn't been enough time for the nonfunctional sequence to accumulate many mutations. Alternatively, because many of these sequences are low coverage, it is possible that many putatitive pseudogenes only have a nonfunctional open reading frame because of a sequencing or assembly error. We do note, however, that some putative pseudogenes do have extremely long branch lengths relative to orthologs with functional open reading frames. These are likely to be nonfunctional sequences.

Figure 1. Maximum likelihood phylogenetic tree comparing all aligned phosphatase homologs across insects, nematodes, and yeast. The aligned MAFFT sequences were run through PhyML. All aligned sequences, including putative pseudogenes, were retained for this analysis for protein alignments. Putative pseudogene sequences, those with an incomplete open reading frame, are highlighted in green (also start with q). For both figures, the species names are abbreviated using the species and genome build (Caenorhabditis elegans: ce6; C. brenneri: caePb2; C. briggsae: cb3; C. remanei: caeRem3; C. japonica: caeJap; Pristionchus pacificus: priPac1; Drosophila melanogaster: dm3; D. simulans: droSim1; D. sechellia: droSec1; D. yakuba: droYak2; D. erecta: droEre2; D. ananassae: droAna3; D. pseudoobscura: dp4; D. persimilis: droPer1; D. willistoni: droWil1; D. virilis: droVir3; D. mojavensis: droMoj3; D. grimshawi: droGri2; Anopheles gambiae: anoGam1; Apis mellifera: apiMel3; and, Tribolium castaneum: triCas2).

Including all sequences, we still observe that most sequences group by orthologous family (Figure 2), and not by paralogs within a species, suggesting that these phosphatase genes expanded prior to the diversification of flies independent of nematodes. However, not all genes are monophyletic by species, so we investigated this further by removing the potentially confounding putatitve pseudogenes.

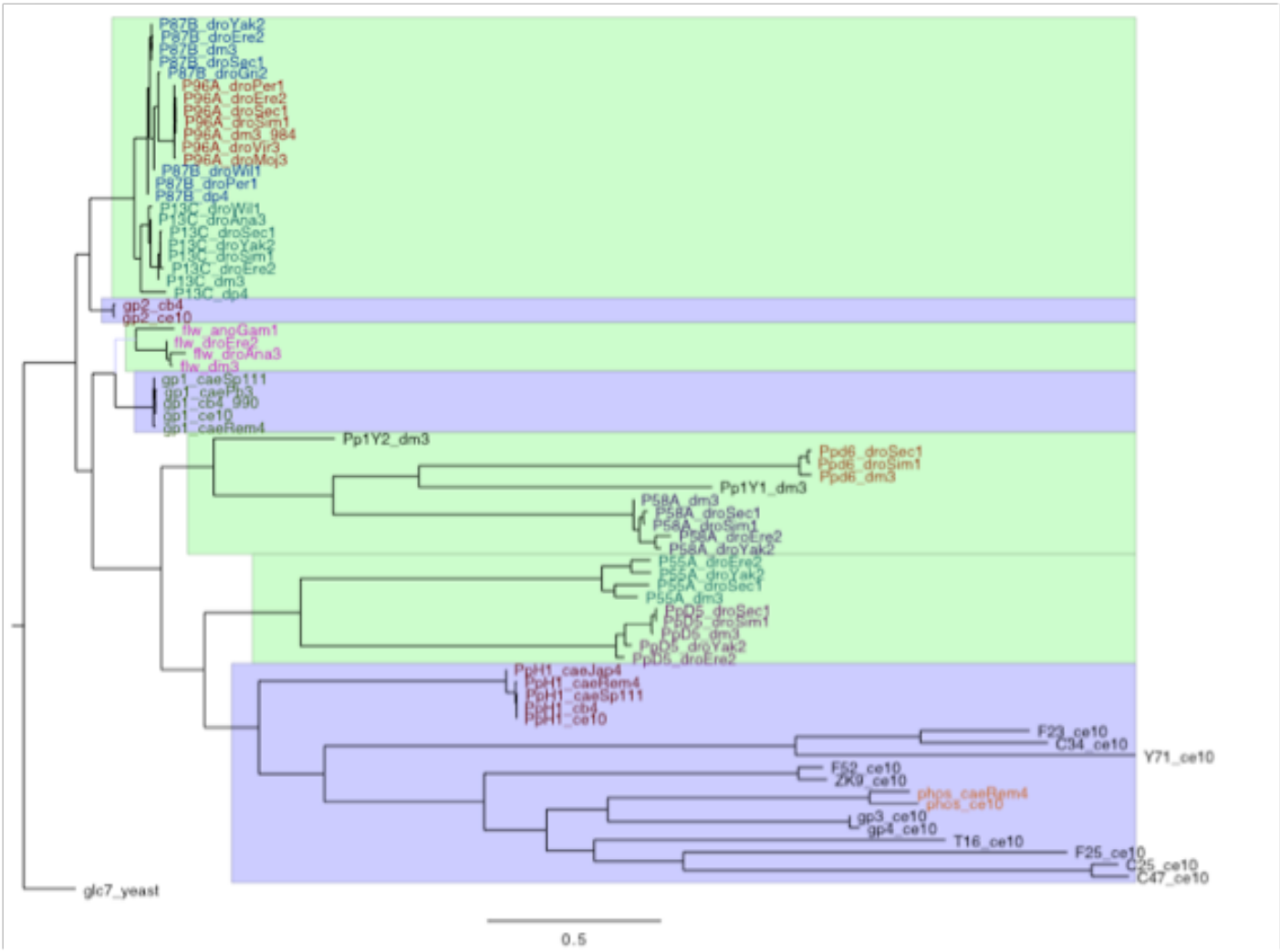

Figure 2

Figure 2. Maximum likelihood phylogenetic tree comparing phosphatase homologs with intact open reading frames across insects, nematodes, and yeast. The aligned MAFFT sequences were run through PhyML. Putative pseudogenes were excluded for this analysis for protein alignments. All gene families are colored in a unique color, and all nematode sequences are highlighted in blue, while insects are highlighted in green.

Evolutionary relationship between genes and species

In the maximum likelihood tree with only sequences with functional open reading frames, we observe a complicated history of the phosphatase gene family. It seems that some phosphatase genes (namely gp1 and gp2) are quite ancestral, whereby the nematode orthologs group closely with a set of homologs in flies. In nematodes, gp1 groups closely with the Drosophila flw genes. Alternatively, the gp2 gene from nematodes groups with an expanded set (P87B, P96A, and P13C) in flies. All of these genes have relatively short branch lengths, suggesting they may be highly conserved. In contrast, the rest of the phylogenetic tree is composed of genes with much longer branch lengths. Notably, there was an expansion of phosphatase genes in C. elegans that has no homologs in any of the flies or in other nematodes, suggesting that, although there was some expansion of these gene families in the common ancestor of nematodes and flies, there has also been a lineage-specific expansion within C. elegans (Figure 2). The C. elegans specific expansion can also be viewed in the tree with putative pseudogene sequences (Figure 1), but is much clearer in the tree without putative pseudogenes (Figure 2).

Selection analysis

223 out of 561 codons (40% of the entire alignment) show evidence of being negatively selected, suggesting that a high proportion of the positions across this gene family are evolutionarily constrained.

Future directions

This study was designed to determine whether or not the phosphatase gene family expanded independently in C. elegans and also in Drosophila flies or if it was more ancestral. To expand upon the results of this project, we can resequence these genes in the non-model organisms, to see if the ORF-disrupting mutations were due to sequencing errors, or are actually ORF-disrupting mutations. We can further investigate lineage-specific selection, and look at the protein domains to detect evidence of neofunctionalizatino or subfunctionalization across this gene family. We can also assess the expression of these genes across the different fly and nematode species, to see if they are sex-specific, or involved in meiosis, as they appear to be in C.elegans. Lastly, we can look at diversity information, and see if there are any signals of selection acting on these genes across or within species.

Conclusion

This study found that 1) multiple alignments may contain spurious alignments, where the same genomic location is mapped multiple times, which is especially a problem for multi-gene families; 2) low quality sequences may result in several ORF-disrupting mutations that may appear to be pseudogenes, but may in fact just be sequencing errors; and 3) the phosphatase gene family appears to have expanded in the common ancestor of nematodes, and again in the common ancestor of flies, with an additional lineage-specific expansion in the nematode C. elegans.

References

1. Cohen, P.T.W., Protein phosphatase 1 – targeted in many directions, Journal of Cell Science, 115, 241-256, 2002.
2. Fong, N.M., Jensen, T.C., Shah, A.S., Parekh, N.N., Saltiel, A.R., and Brady, M.J., Identification of Binding Sites on Protein Targeting to Glycogen for Enzymes of Glycogen Metabolism, Journal of Biological Chemistry, 275, 35034-35039, 2000.
3. Ceulemans, H., and Bollen, M., Functional Diversity of Protein Phosphatase–1, a Cellular Economizer and Reset Button, Physiological Reviews, 84:1, 1-39, 2004.
4. Plowman, G.D., Sudarsanam, S., Bingham, J., Whyte, D., and Hunter, T., The protein kinases of Caenorhabditis elegans: a model for signal transduction in multicellular organisms, PNAS, 96:24, 13603–13610, 1999.
5. Ceulemans, H., Stalmans, W., and Bollen, M., Regulator-driven functional diversification of protein phosphatase-1 in eukaryotic evolution, Bioessays, 24:4, 371–381, 2002.
6. Axton, J.M., Dombradi, V., Cohen, P.W., and Glover, D.M., One of the protein phosphatase 1 isoenzymes in Drosophilia is essential for mitosis, Cell Press, 63:1, 33-46, 1990.
7. Wu, J., Go, A.C., Samson, M., Cintra, T., Mirsoian, S., Wu, T.F., Jow, M.M., Routman, E.J., and Chu, D.S., Sperm Development and Motility are Regulated by PP1 Phosphatases in Caenorhabditis elegans, Genetics, 190:1, 143-157, 2012.
8. Wu, T.F., and Chu, D.S., Sperm Chromatin: Fertile Grounds for Proteomic Discovery of Clinical Tools, Molecular & Cellular Proteomics, 7, 1876-1886, 2008.
9. Larkin M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J., and Higgins D.G., ClustalW and ClustalX version 2, Bioinformatics, 23:21, 2947-2948, 2007.
10. Loytynoja A., Goldman N., An algorithm for progressive multiple alignment of sequences with insertions, PNAS, 102:30, 10557-10562, 2005.
11. Katoh, K., Misawa, K., Kuma, K., Miyata, T., MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform, Nucleic Acids Research, 30:14, 3059-3066, 2002.
12. Edgar, R.C., MUSCLE: multiple sequence alignment with high accuracy and high throughput, Nucleic Acids Research, 32:5, 1792-

1797.
13. Guindon S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O., New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0., Systematic Biology, 59:3, 307-321, 2010.
14. Delport, W., Poon, A.F., Frost, S.D.W., Pond, S.L.K., Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology, Bioinformatics, 26:19, 2455–2457, 2010.
15. Jordan, G., Goldman N., The effects of alignment error and alignment filtering on the sitewise detection of positive selection, Molecular Biology Evolution, 29:4, 1125-1139, 2012.
16. Graur, D., and Li, Wen-Hsiung, Fundamentals of Molecular Evolution: Second Edition, pp. 181-216, 2000.

Acknowledgements

We would like to thank Dr. Diana Chu for preliminary discussions of this topic. Funding for this project was provided by the Miller Institute for Basic Research fellowship to MAWS.