2015

Uncovering Orthologous Genes of the *Ciona Intestinalis* Fanconi Anemia Pathway

Edward Stanley
University of Rhode Island, ed.c.stanley@gmail.com

UNCOVERING ORTHOLOGOUS GENES of the *Ciona Intestinalis* FANCONI ANEMIA PATHWAY

BY
EDWARD STANLEY

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE

REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

IN

BIOLOGICAL SCIENCES

UNIVERSITY OF RHODE ISLAND

2015

MASTER OF SCIENCE IN BIOLOGICAL SCIENCES

EDWARD STANLEY

APPROVED:

Thesis Committee:

Major Professor Steven Irvine

Marian Goldsmith

Niall Howlett

Nasser H. Zawia
DEAN OF THE GRADUATE SCHOOL

UNIVERSITY OF RHODE ISLAND
2015

ABSTRACT

Fanconi Anemia is a disease caused by any number of mutations in a collection of DNA double-strand repair genes. In silico tests were performed to determine whether any of these genes were conserved in Ciona, which split from humans over 500 million years ago. Among the 22 gene products tested, evidence for 10 orthologs were discovered. Possible orthologs were seen in all three of the major groups of FA proteins.

ACKNOWLEDGEMENTS

I would like to thank Dr. Steve Irvine for taking me on as a student and putting forth the effort necessary to make sure this project was completed. I would also like to acknowledge Drs. Marian Goldsmith and Niall Howlett for adding to their already busy schedules by agreeing to be on my committee. Further thanks are given to Drs. Hany Alashwal, Joanna Norris, Kathleen Peterson and Allison Roberts, Dr. Irvine's lab, the University of Rhode Island Graduate School, and many others at the University of Rhode Island.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Fanconi Anemia is a rare genetic disease characterized by genome instability, an increased risk for leukemia and other types of cancer (D'Andrea, 2003), hypersensitivity to both DNA crosslinking agents (such as mitomycin C) and DNA replication inhibitors (Howlett, 2005), as well as a host of congenital defects (Glanz, 1982). The Fanconi Anemia pathway is extensive, with fifteen known genes causing the disease (Bogliolo, 2013), and several other proteins involved that may or may not manifest themselves as FA-related. The disease is recessive, with the genes (for the most part, all phylogenetically unrelated) in the pathway found throughout the chromosomes, on autosomes, and in one case, the X-chromosome (Meetei, 2004). The pathway itself is involved in a DNA damage response mechanism (Wang, 2007), repairing instances of DNA crosslinkage (Nidernhofer, 2005) as well as double-strand breakage (Kim, 2012).

The proteins involved in the FA pathway are often subdivided into three groups. Group I is the Fanconi Anemia core complex. This group consists of eight known FA causing proteins (A, B, C, E, F, G, L, M) and at least two FA-associated proteins. This complex is required for the ubiquitination of FANCD2 and FANCI (Meetei *et al*, 2003). When at least one of these core complex genes is disabled, downstream ubiquitination does not occur (Garcia-Higueira, 2001). FANCM has been thought to serve as a scaffold for the rest of the core complex as well as for another closely related disease-causing gene, BLM (Bloom's syndrome) (Deans and West, 2009), though its role in causing FA directly is controversial (Meetei, 2005). FANCL serves as an E3 ubiqutin ligase (Meetei *et al*, 2003). The functions of the other Group I proteins are currently unclear. The second group consists of the aforementioned FANCD2 and FANCI proteins, which act on genes

and proteins directly involved in the DNA repair process. The third group contains the remainder of the proteins, assumed to act downstream from FANCD2 and FANCI, as FA patients with mutations in these genes have normal levels of D2 and I ubiquitination (Wang, 2007). These proteins include nucleases and ATPases which function in DNA repair.

Several of the FA proteins are ubiquitous among the eukaryotes (Zhang *et al*, 2009). Almost every organism surveyed possessed both of the group II proteins, as well as FANCL, FANCM and an associated ubiquitin-conjugating (E2) enzyme. There is no apparent evolutionary pattern associated with the presence or absence of the other core complex proteins outside of the vertebrates, as some are found in insects, while others are seen in plants and red algae before seemingly reappearing in *Nematostella* and then again in the vertebrates. Echinoderms, a sister group to chordates, possess at least four of the group I proteins.

Ciona intestinalis is the closest invertebrate relative of the vertebrates. (Delsuc *et al*, 2006). While in many cases *Ciona* has lost genes reflecting adaptation to its sessile niche (Hughes and Friedman, 2005), it can still be used to model simplified pathways (Davidson, 2007 and Philips *et al*, 2003), as it possesses a simplified vertebrate body plan, most notably as a larva (Satoh, 2003). A previous study focusing on zebrafish (Titus, 2006) looked into the *Ciona* FA pathway and was unable to find most of the genes. The genes that were found were concentrated in groups II and III, making it plausible that *Ciona* could at the very least be used as a model for the latter two thirds of the pathway and possibly the minimal pathway for FA function in vertebrates.

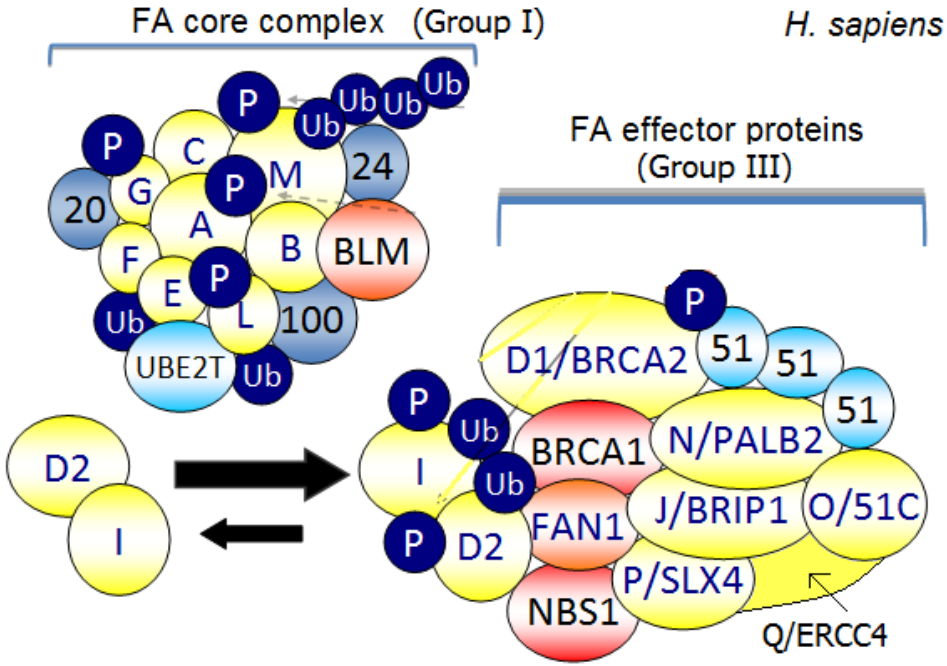


Fig 1 – A possible configuration of the FA pathway in humans. The FA core complex is Group I, the FA effector proteins are group III, and D2/I are group II. Modified from Cybulski and Howlett, 2011

Group	Name	Evidence	Core
I	FancA	Not Known	Yes
I	FancB	Not Known	Yes
I	FancC	Ambiguous	Yes
I	FancE	Ambiguous	Yes
I	FancF	Not Known	Yes
I	FancG	Ambiguous	Yes
I	FancL	Ambiguous	Yes
I	FancM	Ambiguous	Yes
II	FancD2	Yes	No
II	FancI	Ambiguous	No
III	FancD1	Not Known	No
III	FancJ	Yes	No
III	FancN	Not Known	No
III	FancO	Ambiguous	No
III	FancP	Not Known	No
III	FancQ	Ambiguous	No
III	RAD51	Yes	No
III	Fan1	Not Known	Yes
	UBE2T	Ambiguous	Yes
	FAAP20	Not Known	Yes
	FAAP24	Not Known	Yes
	FAAP100	Not Known	Yes

Table 1 – Knowledge of Ciona FA orthologs as of Dec. 2013
Yes – Strong evidence for gene/product in Ciona; Ambiguous – Some evidence of gene/products presence in Ciona; Not Known – no evidence of gene/product in Ciona. Group I is the core complex, Group II consists of FancD2 and FancI, while Group III are the downstream effector proteins.

MATERIALS AND METHODS

Obtaining Sequences: First, a Reciprocal Best BLAST (RBB) (Altschul, 1990) search on twenty gene products was performed, searching the human genes of the FA pathway (Appendix table 2 col. 3) against the *Ciona* proteome, taking the closest match, and then searching the *Ciona* protein back against the human database to see if the same protein was returned as the closest result. This was augmented with a search by the Reciprocal Smallest Distance method (RSD) (Fraser and Hirsh, 2003), which for the most part returned the same protein, but in three cases returned a Fanconi Anemia gene product which had not been found bythe reciprocal best BLAST method. In instances where the two differed, the case with the higher percent of positive matches was used. In all cases this was the FA gene found by the RSD algorithm.

Protein Information: Each *Ciona* and human sequence was run through the European Bioinformatics Institute SAPS (Brendel *et al*, 1992) program for sequence analysis as well as PROSITE (Sigrist *et al*, 2010) and the Eukaryotic Linear Motif (ELM) database (Dinkel *et al*, 2013), which in turn uses SMART (Schultz, Bork, Ponting, 1998) to search for possible overarching structures and motifs. The sequences were also checked in BLAT (Kent, 2002-04, Kent, 2002-06) and the original genes were examined in the JGI genome portal (Nordberg *et al*, 2014) as well as OrthoDB (Zdobnov, 2008) to look for synteny, though none was found for any of the genes.

Using ClustalX and ClustalΩ (Larkin *et al*, 2007), I aligned each *Ciona* FA protein sequence once against the human sequence and *Xenopus* sequence (to serve as a vertebrate comparison), and once against human, *Danio*, *Xenopus*, *Strongylocentrotus purpuratus,* and mouse Fanconi genes, when those sequences were available, and the

closest related human gene by RBB/RSD (Complete listing found in Appendix, Table 3). For the first alignment, I imported the sequences into JALVIEW (Waterhouse *et al*, 2009) and isolated the most closely aligned regions. I then put together hydrophobicity plots of each sequence using biopython (Chapman, 2000) and code built and modified from Dalke scientific (Dalke, 2011). To determine whether the results were significant, I determined the Pearson coefficients for the *Ciona* amino acid sequence against the human and *Xenopus* sequences (again using Python), derived a beta distribution for each sequence (AbouRizk, 1994), and compared the critical values to a $p < 0.02$ level of significance. A standard $p < 0.05$ level of significance with 20 tests gives about a 30% chance of a false positive (Type I error), far too high a level to be acceptable.

Phylogenetic Relationships: The second set of alignments was used as input in PHYLIP (Felsenstein, 1993), where multiple data sets were bootstrapped (5000x), and Tree-Puzzle (default settings) (Schmidt *et al*, 2002), to determine the relation between the proteins. As there is no sequence homology between most FA genes, closer relationships among the same FA gene in different organisms should be seen than ones between different FA genes in the same organism, or between an FA gene and its second best match through RBB/RSD. In particular, *Ciona* FA genes should be especially close to *D. rerio* where available.

RESULTS

Group I:

FANCA: FANCA is part of the FA core complex with an unknown function. Both the Reciprocal Best Blast and Smallest Distance methods return Trafficking Particle Protein Complex subunit 10, with an E-value of $1 \cdot 10^{-122}$ in *Ciona*. The human protein contains five globular domains, while the *Ciona* candidate protein contains only two, with poor alignment. The hydropathy plot shows very little similarity, returning an R^2 value of 0.052. The phylogenetic tree (figure 2) indicates that the *Ciona* candidate protein most closely matches the human TPPC10 rather than any of the Fanconi Anemia group A gene products. However, it may be the case, and the phylogenetic tree would not show this, that two proteins perform the function in *Ciona* that one protein does in vertebrates. That being said, it appears a FANCA orthologue is not present in *Ciona*.

FANCB: FANCB is part of the FA core complex with an unknown function, and is the only known sex-linked FA gene product (Meetei, 2004). Both methods return a dual function lysine demethylase / histidine hydroxylase, but give an E-value of only 0.076, easily the worst measured of any protein. The tree data agrees with this, finding that the *Ciona* candidate protein most closely matches the human demethylase/hydroxylase rather than FANCB. Both *Ciona* and human proteins possess the same number of globular domains, but there is very little alignment between them. The hydrophobicity plot shows similar levels of dissimilarity, returning a correlation of 0.026. It is unlikely that a FANCB orthologue is still present in *Ciona*.

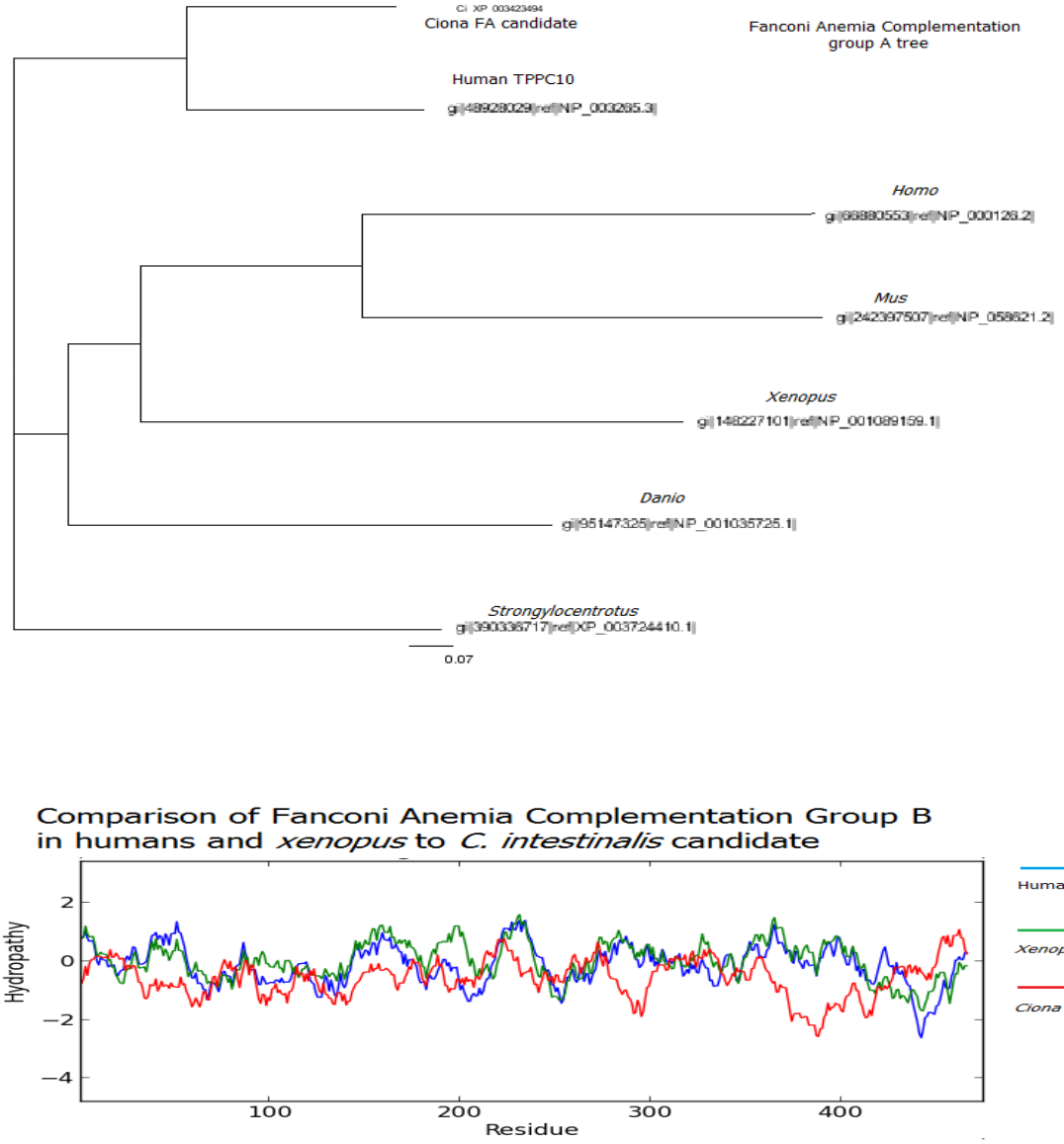


Fig. 2 – Phylogenetic tree data for Fanconi anemia complementation group A, showing that the Ciona candidate gene product is more closely related to a different protein. The base node is a polytomy. ELM and hydropathy plots for FANCA can be found in the supplementary figures.

Fig. 3 – Hydrophobicity plot of FANCB *in humans and* Xenopus *compared to the FANCB candidate protein in C. intestinalis. Ciona's protein shows a very low correlation* (R^2 0.026). *ELM and tree plots can be seen in the supplementary figures.*

FANCC: FANCC is part of the FA core complex with an unknown function. Both methods (RBB & RSD) return the *Ciona* stabilin-2-like protein, at an E-value of less than $1.7 \cdot 10^{-308}$. The human protein returns no notable subdomains in the ELM database, while the *Ciona* candidate protein possesses no less than 15 EGF and EGF-like domains, implying that the *Ciona* candidate is not particularly closely related to the human gene product. The tree data backs this up, indicating the candidate gene is more closely related to human stabilin than any of the Fanconi group C genes. It is unlikely that a FANCC ortholgue exists in *Ciona*.

FANCE: FANCE is part of the FA core complex with an unknown function. The Reciprocal Best BLAST method returns a protein related to human HEXIM-1, while the reciprocal smallest distance method returns the C-terminal domain for FA complementation group E. Phylogenetic data supports this, as the *Ciona* candidate protein groups with the last 250 – 300 amino acids of several FANCE proteins, something not seen when the entirety of the larger FANCE are tested. The *Ciona* candidate is about 400 amino acids, while the other FANCE proteins are all between 550 and 600 aa. Approximately half of the known FancE mutations occur in this C-terminal third of the protein. Moderate alignment is seen between the two globular domains in the *Ciona* candidate and the two C-terminal globular regions in the human protein, though no other secondary structure is seen. For the most part, the last 250 aa of human FANCE match up well with the last 250 aa of the *Ciona* candidate gene ($R^2 = 0.202$), but on the

whole the correlation is only 0.08 (and the region outside the C-terminal registering at only 0.05). ELM, Hydropathy and tree figures for FANCE can be found in the supplementary figures. It is not entirely clear whether FANCE has a full orthologue in Ciona, but a region approximating the C-terminal domain appears to be there.

FANCF: FANCF is part of the FA core complex with an unknown function. Both RBB and RSD return Eukaryotic Elongation Factor 1-like subunit-γ as the best match. The tree data shows human EF-1-like matches the *Ciona* candidate gene product better than any of the Fanconi group F proteins. The *Ciona* candidate possesses two glutathione-transferase domains as well as a region characteristic of an elongation factor, neither of which the human protein contains. FancF does not appear to have an orthologue in *Ciona*.

FANCG / XRCC9: FANCG is part of the FA core complex with an unknown function. Both methods return a Stress-induced Phosphoprotein as the closest match to the *Ciona* candidate gene product, with an E-value of $3 \cdot 10^{-78}$. The tree data confirms that the *Ciona* candidate protein is more closely related to this Phosphoprotein than any of the FA group G genes. Human FANCG possesses four TPR sites, a rather common motif. The *Ciona* candidate protein contains 11 of these domains (as does the human version of this phosphoprotein), which may explain why this protein registers as the closest match to human FancG. With this being the case, it does not appear that FANCG has an orthologue in *Ciona*.

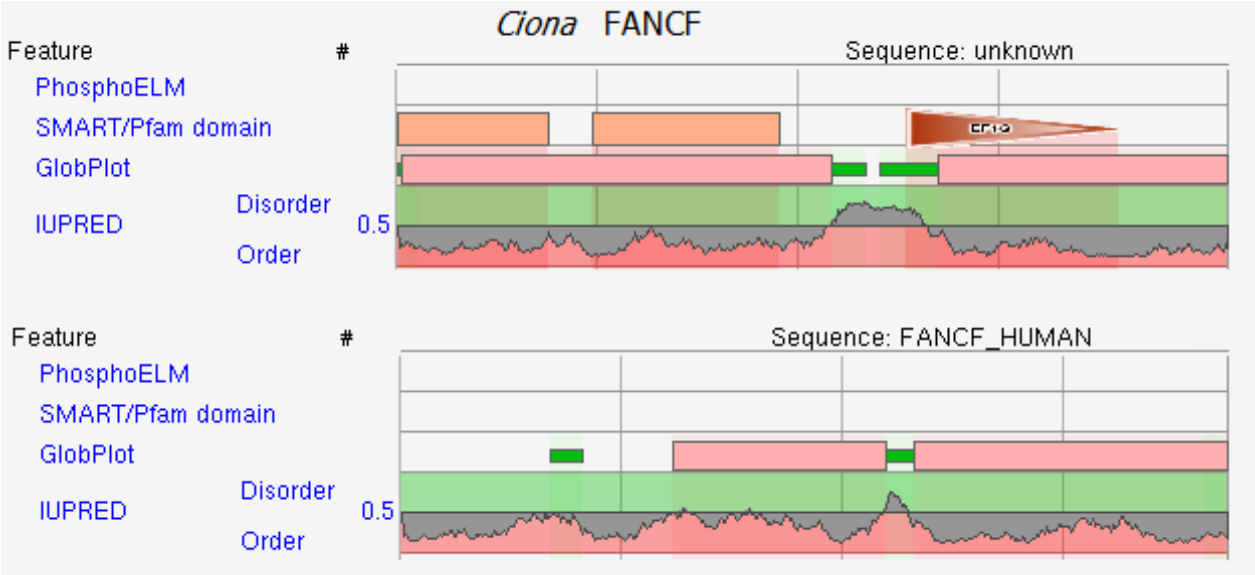
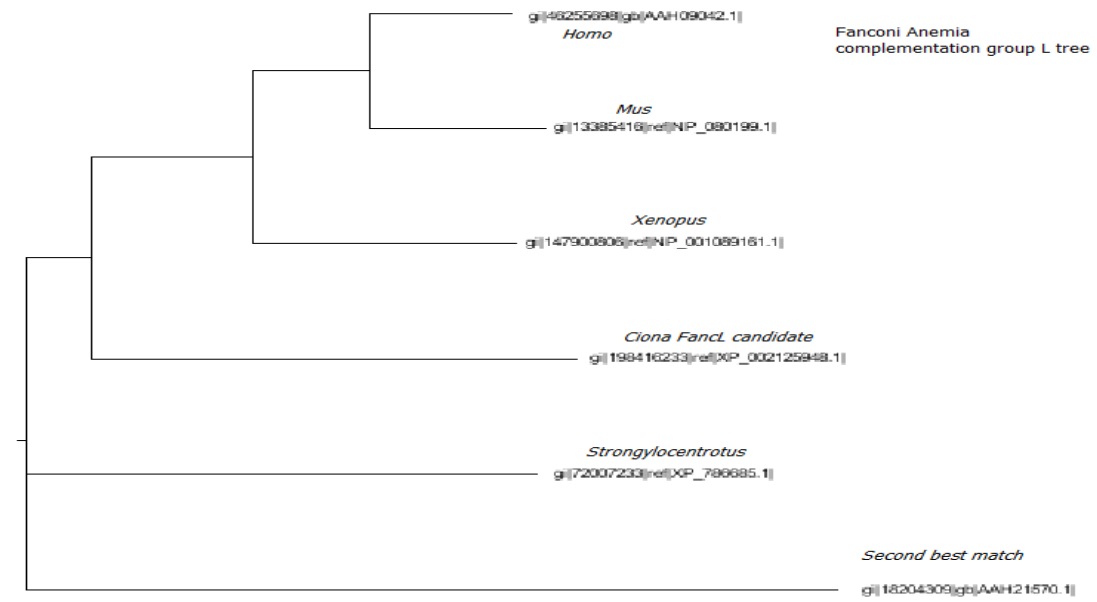


Fig. 4 - ELM comparison for two proteins that do not appear to be related, in this case Human FANCF *and the Ciona FANCF candidate found through reciprocal best BLAST. Note the presence of recognizable secondary structure in Ciona where none is present in the human protein. For Hydropathy plot and tree diagram see supplementary figures.*

FANCL: FANCL is an E3 ubiquitin-ligase and a part of the FA core complex which serves to ubiquitinate FANCD2 and FANCI (Meetei, 2003). Both methods return a putative *Ciona* FANCL protein with a given E-value of $2 \cdot 10^{-74}$. Phylogenetic data agrees with this finding (figure 5), showing that the second best human match for *Ciona*'s FancL candidate is not as closely related to any of the group L proteins as they are to each other. Unlike most of the core complex proteins, FANCL does appear to have an orthologue in *Ciona*.

FANCM: FANCM is a part of the FA core complex known to function as a scaffold for several DNA repair pathways not necessarily involving FA proteins (Deans, 2009), being involved in protein degradation (Ali, Singh, Meetei, 2009) and also serving as a DNA translocase (Gari *et al* 2008). Both methods return a *Ciona* FANCM protein as the closest match. Secondary structure data shows that both the human and *Ciona* candidate proteins possess a DEAD domain as well as a helicase domain in approximately the same positions. The human FANCM also possesses an ERCC endonuclease domain that the *Ciona* protein lacks. The hydrophobicity plot shows high levels of correlation, especially towards the N-terminal end of the protein. This is noteworthy, as most FA causing mutations known in FANCM are far from the N-terminal end. (Auerbach, 2014). It is possible that changes in the N-terminal end of the protein cause it to cease functioning, while mutations towards the C-terminal (including ones that cause FA) are deleterious but not completely loss-of-function inducing. Given the data, *Ciona* is likely to possess a FANCM orthologue.



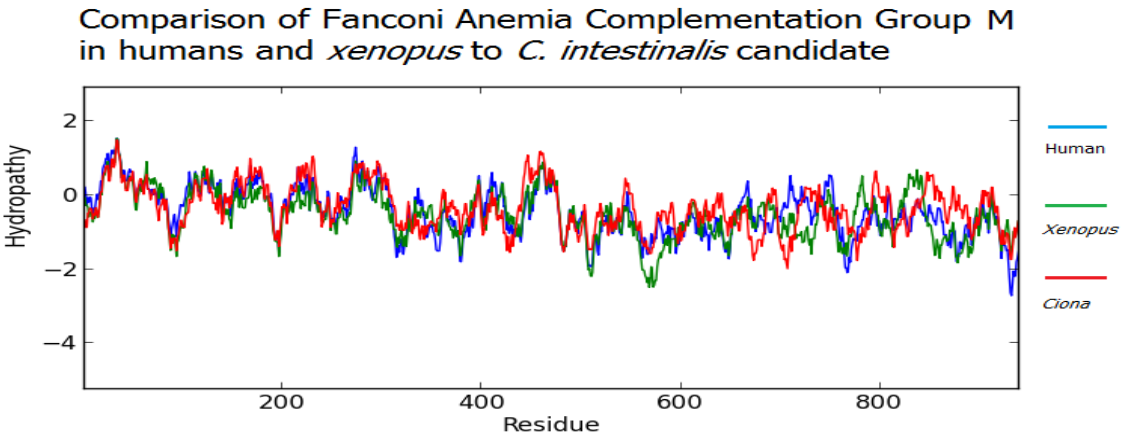


Fig. 5 – Phylogenetic tree for Fanconi Anemia complementation group L, showing that the Ciona candidate is more closely related to the other FANCL gene products than the second best RBB/RSD match.

Fig. 6 – Comparison of hydropathy of FANCM protein in humans and Xenopus *against the aligned regions of the Ciona candidate. The plots show a relatively high correlation (R^2 = 0.37) and the first 500 residues especially so (R^2 = 0.70)*

UBE2T: UBE2T is one of many ubiquitin-conjugating enzymes found in the human proteome, and is the specific one found in the FA pathway (Machida, 2006). The reciprocal best blast method returns UBE2L4 as the closest match with an e-value of $2 \cdot 10^{-24}$, while the reciprocal smallest distance method returns UBE2J1 with an e-value of $9 \cdot 10^{-76}$. In humans, UBE2T interacts with FANCL to ubiquitinate FANCD2. It is entirely possible that instead of using an orthologous protein, C. intestinalis has repurposed a different E2 ubiquitin-conjugating enzyme for the same job.

FAAP20: FAAP20 is a ubiquitin-binding protein known to interact with and stabilize FANCA (Ali, 2012), (Leung, 2012). It is a poor match for any C. intestinalis proteins; both the RBB and RSD methods return a best match of Transcription Elongation Factor SPT6, with an e-value of 0.054. FAAP20 looks unlikely to have a *Ciona* orthologue.

FAAP24: FAAP24 is a protein known to interact with FANCM (Ciccia, 2007) whose mutations do not appear to result in Fanconi Anemia phenotypes. Both methods return DNA polymerase β, with an E-value of $1 \cdot 10\text{-}^{145}$. The tree data indicates the *Ciona* candidate product is more closely related to the DNA polymerase than any of the FAAP24 proteins it was tested against. The human protein possesses a region indicative of the polymerase family, while the *Ciona* candidate has no notable secondary structure. FAAP24 does not appear to have an orthologue in *Ciona*.

FAAP100: FAAP100 is a protein known to interact in the core complex with FANCB and FANCL that works to protect all three from degradation (Ling, 2007). Both methods return an L-fucose kinase as the closest match, though with a poor E-value (0.0002) and the tree data also lends evidence to this. The *Ciona* protein possesses six globular

domains and no other secondary structure motifs. The human FAAP100 contains a signaling peptide region and poor globular domain alignment with *Ciona*. A FAAP100 orthologue in *Ciona* is unlikely.

FAN1: Fanconi-Associated Nuclease 1 is a DNA repair protein known to interact with ubiquitinated FANCD2 (MacKay *et al* 2010) and FANCI (Liu, 2010). The reciprocal best BLAST method returns a protein regulator of cytokinesis, while the reciprocal smallest distance method returns a C-terminal domain of FAN1. The FAN1 C-terminal region returns 41% identity and 63% positive matches. While the original FAN1 protein in humans is 1070 aa in length and the match only 171, they also align extremely well in the hydropathy plot and contain a 110 aa VRR nuclease region. Much like FANCE, a full match between proteins is poor, but some functional FAN1 equivalent appears present.

Group II:

FANCD2: FANCD2 is one of the proteins ubiquitinated by FANCL in order to affect DNA repair proteins downstream (Meetei, 2003). Both the reciprocal best BLAST search and reciprocal smallest distance method returned a probable Fanconi Anemia complementation group D2 protein in *C. intestinalis* as the closest match for this protein in humans, with the BLAST search returning 25% identity, a 44% match on positives and an E-value of less than $1.7 \cdot 10^{-308}$, indicating extremely strong similarity. The *Ciona* FANCD2 protein contains 1394 amino acids, while the most common isoform in humans is 1470 amino acids long.

When the sequences are aligned and gaps removed, the smoothed hydrophobicity plots show multiple similarities. The proteins have highly similar ($R^2 \geq 0.71$) regions at around aligned *Ciona* aa 100 – 125, 240 – 280, 510 – 540, 660 – 760, 1010 – 1045 and

1130 – 1170.Both the human and *Ciona* proteins show five globular domains with moderate alignment, and no recognizable motifs. The presence of a FANCD2 orthologue in *Ciona* is likely.

FANCI: FANCI is the second protein ubiquitinated by FANCL to affect downstream DNA repair proteins (Meetei, 2003). Both methods returned FANCI as the closest matching protein, with an E-value of E-value of less than $1.7 \cdot 10^{-308}$. Phylogenetic data supports this view, as the *Ciona* FANCI candidates appears grouped with the other FANCI proteins rather than the second best RBB/RSD match. The hydrophobicity plots return an R^2 value of 0.33, but several areas, notably an 150 aa stretch towards the C-terminal end of the protein, have much higher correlations. Both proteins show multiple globular domains with moderate alignment and no recognizable secondary motifs. A FANCI orthologue in *Ciona* is likely present.

Group III:

FANCD1 / BRCA2: BRCA2 is among the most well-known of human tumor suppressor genes, functioning in double strand DNA break repair. Both search methods return a BRCA-interacting protein (BCCIP) as the closest match, rather than BRCA2. The sequences show no secondary structure alignment, as the human protein contains multiple globular regions and the *Ciona* candidate only one (shown in supplementary figures). FANCD1 does not appear to have a *Ciona* orthologue, though testing the *Ciona* protein in question against BCCIP could be a follow-up.

FANCJ / BRIP1: In humans, FANCJ is a helicase protein that interacts with BRCA1 (Cantor, 2001), FANCM and Bloom Syndrome protein (Suhasini, 2012). Reciprocal best BLAST returns an ERCC nucleotide excision protein, but reciprocal smallest distance

returns human FA complementation group J. There is good alignment between the globular domains in human FANCJ and the *Ciona* candidate. The human protein is of a similar size to the *Ciona* protein (just over 90 percent as large) and they both possess a DEAD region as well a helicase region in approximately the same positions. A FANCJ orthologue does appear to be present in *Ciona*.

FANCN / PALB2: PALB2 is a protein that interacts with BRCA2 (Nazneen, 2006). Both methods return a WD-repeat containing protein with an E-value of $2 \cdot 10^{-94}$, an identity of 18% and 45% positive matches. Tree data confirms that our *Ciona* FancN candidate is far more closely related to this WD-repeat protein than any Fanconi group N proteins as seen in the supplementary figures. The *Ciona* candidate protein contains 7 of these 40 amino acid WD-repeat regions. The human FancN protein possesses none of these. A FANCN orthologue is unlikely to be in *Ciona*.

FANCO / RAD51C: RAD51C is involved in maintaining chromosome stability as well as playing a part in recombinational repair (Takata, 2001). RBB and RSD both return RAD51 in *Ciona* with an E-value of $1 \cdot 10^{-94}$. The human protein contains two globular domains and the *Ciona* one, and both possess a AAA-ATPase domain. The phylogenetic tree indicates the *Ciona* candidate is more closely related to RAD51 than RAD51C. This will be an interesting protein to investigate further, as though this particular paralog of RAD51 does not appear to be in *Ciona*, it could be that a different member of the same family (of which there are several) provides the same function.

FANCP / SLX4: FANCP is an important coordinator of nucleases (Stoepker, 2011). Both methods of searching return a different Kelch-like protein, Kelch-10 for RBB and Kelch-

20 for RSD. Kelch domains are found in diverse families of proteins. Kelch-10 was found to be involved in mouse fertility (Yatsenko, 2006), while Kelch-20 is involved in regulating HIF-2αf(Higashimura, 2011). The *Ciona* FANCP candidate is 670 amino acids long, while human SLX4 is three times as long. The hydrophobicity plots show very little similarity, only providing a close match at the location of a *Ciona* 50 aa Kelch motif, the most likely reason BLAST shows a relation between the two. The *Ciona* candidate contains five other Kelch motifs in this area as well as a BTB domain 500 amino acids closer to the N-terminus. Human SLX4 only possesses a single BTB domain. Tree data indicates the *Ciona* protein more closely matches these human Kelch proteins rather than any Fanconi group P proteins, and is unlikely to be a FANCP orthologue.

FANCQ / ERCC4 / XPF: FANCQ is a repair endonuclease (Bogliolo, 2013) known to interact with FANCA (Sridharan, 2003). Both search methods return XPF as the most closely matching protein, with 50% identity, and 64% positive matches. The tree data agrees with this, and the hydrophobicity plots show a high correlation, excepting one area corresponding to aa 390 – 430 in *Ciona* and 520 – 560 in humans. As of this writing, no FANCQ mutations relating to this region have been found (Auerbach, 2014). Both proteins possess an endonuclease domain of the same size approximately the same distance from the C-terminal end of the protein. Given this, FANCQ appears to have an orthologue in *Ciona*.

RAD51: In humans, RAD51 interacts with BRCA1 and BRCA2 in a DNA damage response pathway (Chen *et al*, 1999). Both search methods return a *Ciona* RAD51 as the most likely counterpart to the human protein. RAD51 appears to be the most highly

conserved auxiliary protein in the entire FA pathway. The protein possesses 82% identity between human and *Ciona* as well as a 92% level of positive matches, far outstripping any other gene product tested. The *Ciona* product is 338 amino acids, while the human product is 339. The ELM database indicates that both possess a 20 amino-acid helix-hairpin-helix domain starting at about amino acid 60, as well as a 187 aa AAA-ATPase domain ending 33 aa before the C-terminus. The hydrophobicity plots show extreme similarity, returning a Pearson coefficient of 0.92. It is highly likely that this is an orthologue.

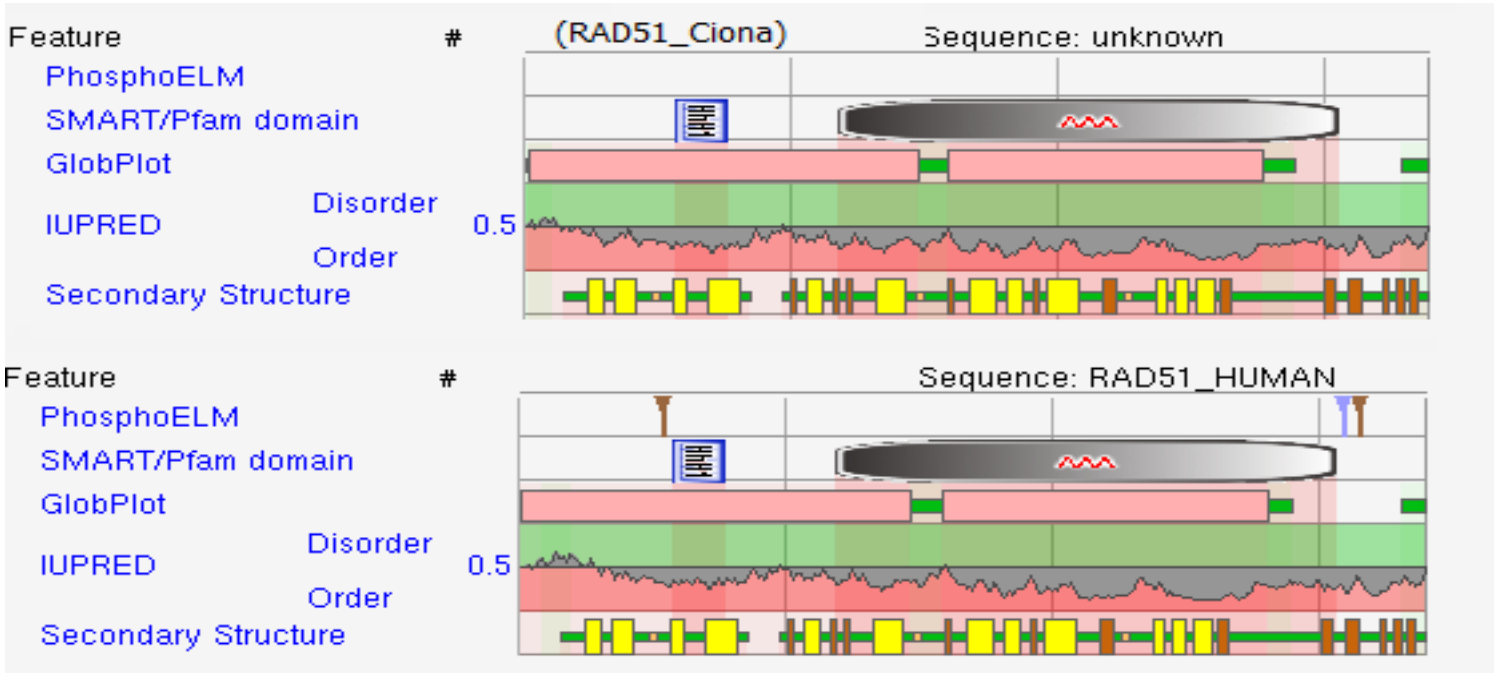


Fig. 7 – ELM information for RAD51 in Ciona (top) and humans (bottom). The Ciona protein is 338 amino acids long and the human protein 339. The placement of the secondary structures is effectively identical. Hydrophobicity and Phylogenetic data can be found in the supplementary figures.

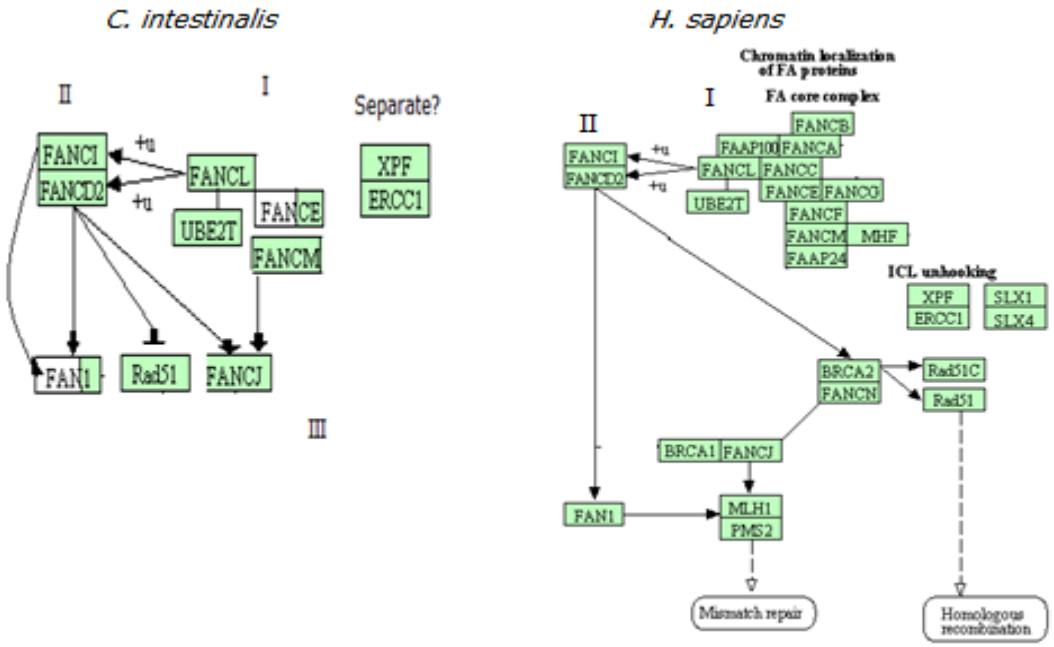


Fig. 8 – Left, A possible FA pathway in C. intestinalis, given the information uncovered in this project. Right – the human pathway, minus an arrow connecting FD2, FANCM and XPF (FANCQ) (both modified from Kanehisa, 2000) I, II and III refer to the FA subgroups.

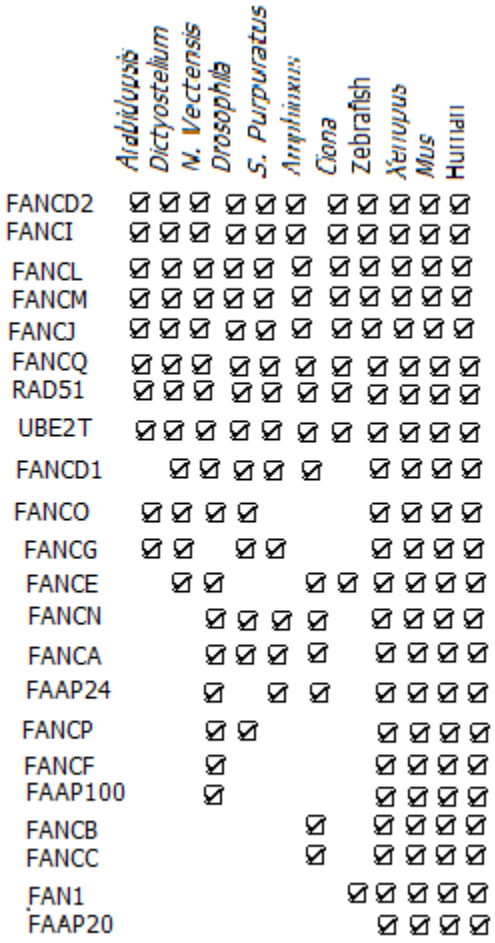


Fig 9. A brief overview of FA pathway genes known in a variety of organisms across several groupings, including plants, amoebae, invertebrates and vertebrates, giving us an idea of plausible, simplified FA pathways.

Ciona intestinalis appears to have both Group II proteins, several group III proteins, and very few group I proteins. In Group I, we have evidence suggesting the presence of FANCL, FANCM, an E2-ubiquitin-conjugating enzyme (though not necessarily human UBE2T) and the carboxy-terminal region of FANCE. In Group III, the evidence points to the existence of FANCJ, FANCQ, RAD51, as well as the carboxy-terminal region of Fan1. The following proteins appear to be absent, or too far derived to be detected with the methods used: FAAP20, FAAP24, FAAP100, A, B, C, D1, F, G, N, O and P.

The evidence for the presence of the proteins thought to be the product of orthologous genes is numerous. The information from both the Reciprocal Best BLAST as well as the Reciprocal Smallest Distance indicate most of these proteins are the best matches for those in humans. The hydrophobicity plots indicate moderate to extremely high conservation, as it is very unlikely for a completely unrelated gene to evolve such high correlation to a gene family by chance. The information provided by the ELM database shows matching secondary structures between human FA proteins and the *Ciona* candidate proteins.

Considering the proteins listed, I expect FANCJ and RAD51 to possess the same roles downstream of FANCD2 and FANCI in *Ciona* that they do in vertebrates. FANCQ may not be involved in the pathway at all, as it is thought to interact with the absent FANCA, but instead possibly maintains its presence in the genome because of its interactions with ERCC1, an endonuclease still present in *Ciona*. Similarly, the ubiquitin-conjugating enzyme most closely matching that found in humans may not be orthologous

but merely a different E2 brought into the pathway from somewhere else to fill the same function.

These data could help in determining the function of the core complex proteins absent in *Ciona* that have no known function. With the exception of FANCB, each of the proteins checked against human and *Ciona* returned a plausible function for a related protein, like human FANCC and *C. intestinalis* stabilin-like. It could be possible that FANCC's function is similar to that of the stabilin-like proteins, which would explain that RBB/RSD result and those like it.

There are three possibilities that could explain the absence of the majority of the core complex proteins. The first is that the FA core complex had not yet come together as a single unit before the divergence of humans and *C. intestinalis*. Proteins of the FA core complex can be seen in organisms as diverse as plants, *Dictyostelium* and arthropods (Zhang, 2009). The second is that many of the DNA repair functions are not necessary for the sessile lifestyle of the tunicate and were not retained as a result - In his paper, Titus (2006) suggested that FANCL may be able to ubiquitinate FANCD2 and FANCI without assistance from the core complex. The third possibility is that the sequences have diverged to an extent that BLAST searches cannot locate them. I feel as this is the least likely scenario, as searches using different block matrices returned the same results as in table 2, with the only difference being the E-values.

With this *in silico* analysis performed, the next step would be to clone these genes and test their expression patterns in living *C. intestinalis*. Once the genes are found to be expressed (and where), one could grow up *Ciona* cell cultures, expose them to DNA-

crosslinking agents and see if and how the genes are upregulated. The effects of these agents on mutated versions of the *Ciona* genes could also be tested.

These results indicate that *Ciona intestinalis* could possibly serve as a model of a minimal FA pathway, and this helps establish a possible lower level at which the pathway can still function. While core complex interactions cannot be done in this organism (and can still be done in humans) *Ciona* is still able to model interactions downstream from the core complex, as most of the known FA effector proteins are still present in the organism.

APPENDICES

Group	Name	Best Match	E-value H→C
I	FancA	Trafficking protein particle complex subunit 10	1×10^{-122}
I	FancB	Lysine demethylase/histidyl hydroxylase MINA	0.076
I	FancC	Stabilin 2-like	$< 1.7 \times 10^{-308}$
I	FancE	Fanconi Anemia Group E protein C-terminal domain	5×10^{-65}
I	FancF	Eukaryotic translation elongation factor 1 γ-like	4×10^{-146}
I	FancG	Stress induced phosphoprotein 1	3×10^{-78}
I	FancL	Ubiquitin ligase Fanconi Anemia group L	2×10^{-74}
I	FancM	FancM protein	5×10^{-156}
II	FancD2	Fanconi Anemia Complementation Group D2	$< 1.7 \times 10^{-308}$
II	FancI	Fanconi Anemia Complementation Group I	$< 1.7 \times 10^{-308}$
III	FancD1	BRCA2 and CDKN1A interacting protein	3×10^{-57}
III	FancJ	Fanconi Anemia group J protein	4×10^{-141}
III	FancN	WD repeat containing protein-5 like	2×10^{-94}
III	FancO	RAD51 homolog 1 isoform 1	6×10^{-90}
III	FancP	Kelch-like protein 20	$< 1.7 \times 10^{-308}$
III	FancQ	DNA repair protein XPF	1×10^{-94}
III	RAD51	RAD51 homolog 1 isoform 1	7×10^{-140}
III	Fan1	Fanconi Associated Nuclease 1 C-terminal domain	9×10^{-41}
	UBE2T	UBE2J1	9×10^{-76}
	FAAP20	Uncharacterized Protein LOC100185826	0.054
	FAAP24	DNA polymerase β	1×10^{-145}
	FAAP100	L-fucose Kinase	0.0002

Table 2 – BLAST results, human to Ciona

Group	Name	Best Match	E-value H→C
I	FancA	Trafficking protein particle complex subunit 10	4×10^{-174}
I	FancB	Lysine demethylase/histidyl hydroxylase MINA	6×10^{-73}
I	FancC	Stabilin 2-like	$< 1.7 \times 10^{-308}$
I	FancE	Fanconi Anemia Group E protein	5×10^{-27}
I	FancF	Eukaryotic translation elongation factor 1 γ	$< 1.7 \times 10^{-308}$
I	FancG	Stress induced phosphoprotein 1	2×10^{-101}
I	FancL	Ubiquitin ligase Fanconi Anemia group L	$< 1.7 \times 10^{-308}$
I	FancM	FancM protein	$< 1.7 \times 10^{-308}$
II	FancD2	Fanconi Anemia Complementation Group D2	$< 1.7 \times 10^{-308}$
II	FancI	Fanconi Anemia Complementation Group I	$< 1.7 \times 10^{-308}$
III	FancD1	BRCA2 and CDKN1A interacting protein	3×10^{-57}
III	FancJ	Fanconi Anemia group J protein	$< 1.7 \times 10^{-308}$
III	FancN	WD repeat containing protein-5 like	7×10^{-116}
III	FancO	RAD51 homolog 1 isoform 1	2×10^{-93}
III	FancP	Kelch-like protein 20	$< 1.7 \times 10^{-308}$
III	FancQ	DNA repair protein XPF	2×10^{-117}
III	RAD51	RAD51 homolog 1 isoform 1	$< 1.7 \times 10^{-308}$
III	Fan1	Fanconi Associated Nuclease 1 C-terminal domain	1×10^{-101}
	UBE2T	UBE2J1	$< 1.7 \times 10^{-308}$
	FAAP20	Serine/Threonine Protein Kinase MRCK	2×10^{-17}
	FAAP24	DNA polymerase β	4×10^{-152}
	FAAP100	L-fucose Kinase	$< 1.7 \times 10^{-308}$

Table 3: BLAST results, Ciona *to human*

Group	Name	Id%	Hydropathy Plot Id% R^2	Pos%	Hydropathy Plot Pos% R^2
I	FancA	36%	0.052	58%	0.027
I	FancB	37%	0.045	52%	0.026
I	FancC	32%	0.121	46%	0.052
I	FancE	19%	0.082	34%	0.051
I	FancF	47%	0.039	65%	0.021
I	FancG	56%	0.034	74%	0.021
I	FancL	36%	0.373	56%	0.199
I	FancM	52%	0.369	66%	0.321
II	FancD2	25%	0.304	44%	0.222
II	FancI	29%	0.331	50%	0.237
III	FancD1	31%	0.127	49%	0.046
III	FancJ	37%	0.291	53%	0.236
III	FancN	18%	0.099	45%	0.061
III	FancO	30%	0.335	44%	0.325
III	FancP	32%	0.050	49%	0.030
III	FancQ	50%	0.566	68%	0.446
III	RAD51	82%	0.844	92%	0.847
III	Fan1	41%	0.447	63%	0.387
	UBE2T	34%	0.311	58%	0.222
	FAAP20	18%	0.053	28%	0.027
	FAAP24	26%	0.177	45%	0.098
	FAAP100	35%	0.114	53%	0.053

Table 4 – Hydrophobicity plot correlations between identities as well as between positive matches.

Table 5 – Main Accession numbers used in this research

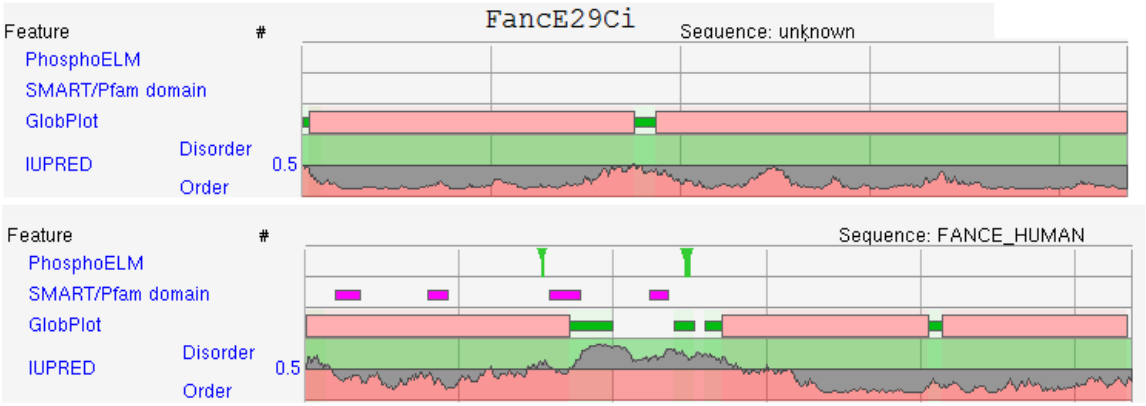
Group	Protein	*Ciona*	Human	*Xenopus*	*Danio*	*Mus*	*S. Purpuratus*
I	A	XP_002123494	NP_000126	NP_001089159	NP_001035725	NP_058621	XP_003724410
I	B	XP_002131324	NP_001018123	NP_001190183	AAY40840	NP_001139553	Did not use
I	C	XP_002122507	NP_000127	NP_001165340	NP_001035727	NP_001036138	Did not use
I	E	XP_002129936	NP_068741	XP_002941449	NP_001035724	NP_001157291	Did not use
I	F	XP_002131081	NP_073562	NP_001086451	NP_001038699	Did not use	Did not use
I	G	XP_002128875	NP_004620	NP_001098745	NP_991202	NP_444311	Did not use
I	L	XP_002125948	AAH09042	NP_001089161	Did not use	NP_080199	XP_786685
I	M	XP_002123226	NP_065988	NP_001171151	NP_001107132	NP_849243	Did not use
II	D2	XP_002130241	NP_149075	NP_001089160	AAI63598	NP_001028416	XP_798439
II	I	XP_002121027	NP_060663	NP_001164682	NP_001070780	NP_666058	XP_001191632
III	D1	XP_002126595	P51587	ABP48763	ABO27625	AAC23702	XP_003727314
III	J	XP_002120239	NP_114432	XP_002933884	Did not use	NP_840094	XP_781875
III	N	XP_002127700	Q86YC2	XP_002932499	XP_001919766	AAH66140	Did not use
III	O	XP_002130341	AAC39604	XP_004911732	AAI64186	Did not use	XP_003725868
III	P	XP_002122519	NP_115820	XP_002932505	XP_003201146	NP_803423	Did not use
III	Q	XP_002125462	Q92889	XP_002935520	NP_956079	AAH26792	XP_003729146
III	RAD51	XP_002126934	CAG38796	NP_001081236	NP_998371	NP_035364	XP_788683
III	FAN1	XP_004227197	NP_055782	XP_002934961	NP_001038546	NP_808561	Did not use
	UBE2T	XP_002128965	NP_054895	NP_001080105	NP_001070763	NP_001265044	Did not use
	FAAP20	XP_002127003	NP_001139782	XP_004915558	Did not use	Did not use	Did not use
	FAAP24	XP_002128462	NP_689479	NP_001088772	Did not use	NP_848578	Did not use
	FAAP100	XP_002122353	NP_079437	XP_002940136	Did not use	NP_082256	Did not use

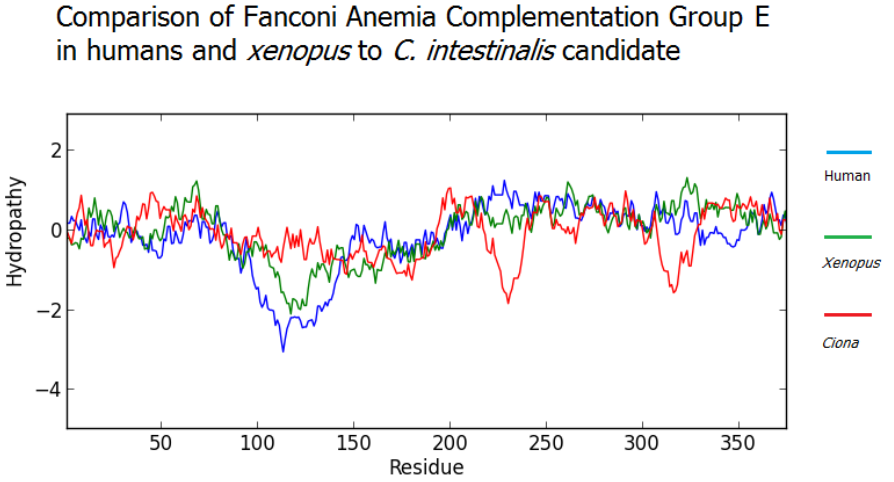
Table 6: Other accession numbers used in this research ("2nd best matches")

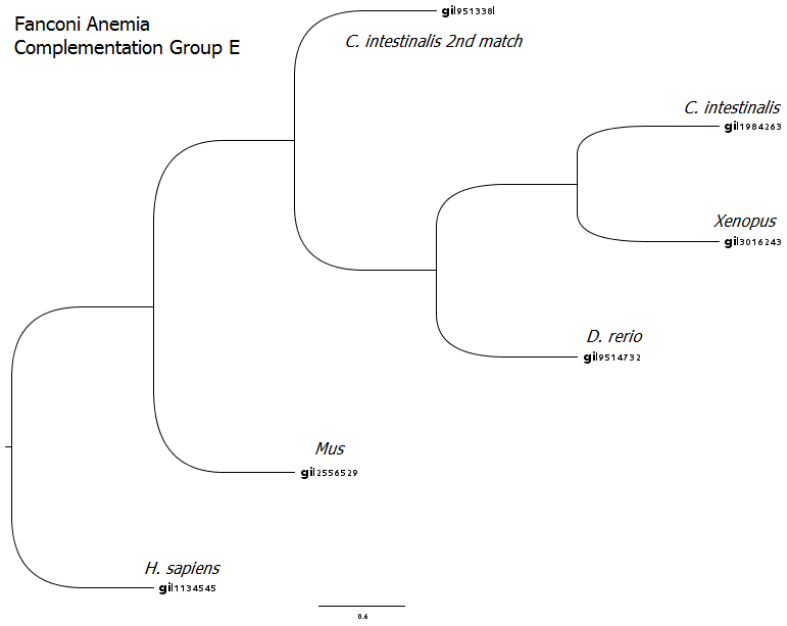
Protein	2nd Closest Matching Gene	Accession #
FANCA	Trafficking Particle Protein Complex Subunit 10	XP_006724110
FANCB	Histidine Hydroxylase MINA	4BXF
FANCC	Stabilin-2	CAC82105
FANCE	CAS / CSE1 homolog	AAC50367
FANCF	Elongation Factor 1-gamma	NP_001395.1
FANCG	Stress-Induced Phosphoprotein 1	NP_001269581
FANCL	Zinc/RING finger protein 3	AAH21570
FANCM	RNA helicase DEAD-box protein	AAG54076
FANCD2	Fanconi Anemia Complementation Group I protein	NP_060663
FANCI	Mitochondrial Poly(A) polymerase	AAH61703
FANCD1	BRCA2 interacting protein	EAW49220
FANCJ	ERCC2 gene product	CAA36463
FANCN	WD-repeat protein 5-like	NP_438172
FANCO	RAD51 homolog 1	CAG38796
FANCP	Kelch-like protein 20	NP_055273
FANCQ	Fanconi Anemia Complementation Group M protein	NP_065988
RAD51	Meiotic Recombination Protein DMC1	BAA10970
FAN1	Unnamed Protein Product (homo sapiens)	BAG60721
UBE2T	UBE2L3-like	XP_002131033
FAAP20	Uncharacterized Protein	XP_002129267
FAAP24	DNA polymerase beta	NP_002681
FAAP100	L-fucose kinase	NP_659496

Supplementary Figures

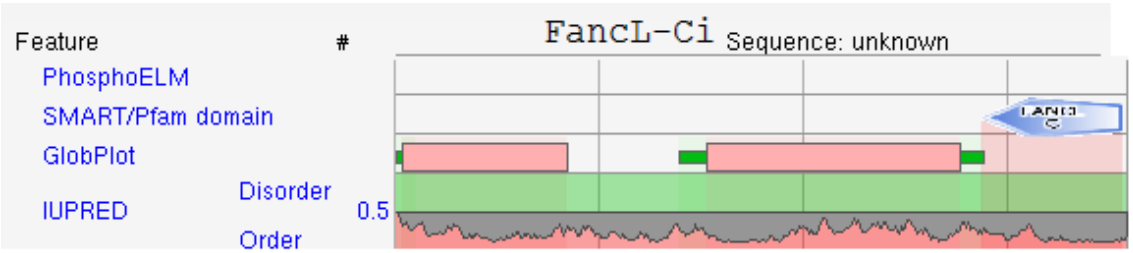
Proteins that appear to be present:
FancE

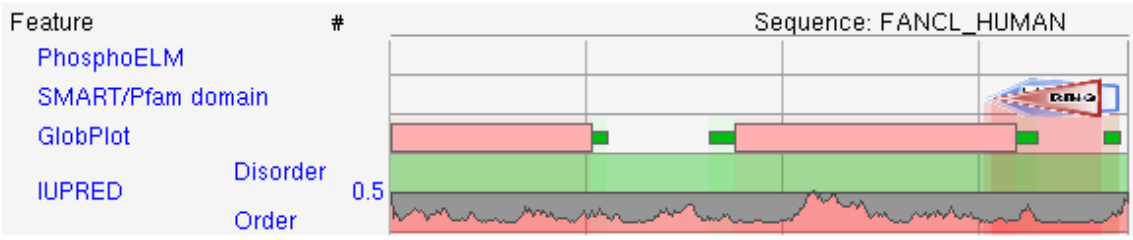


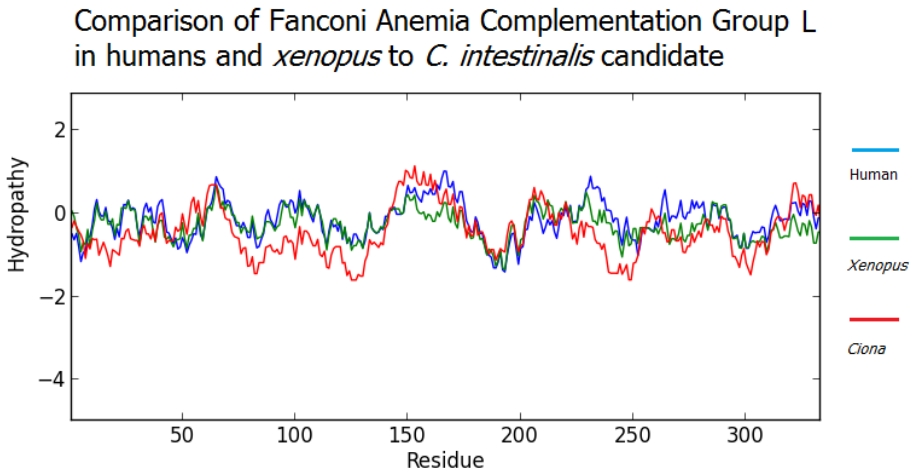




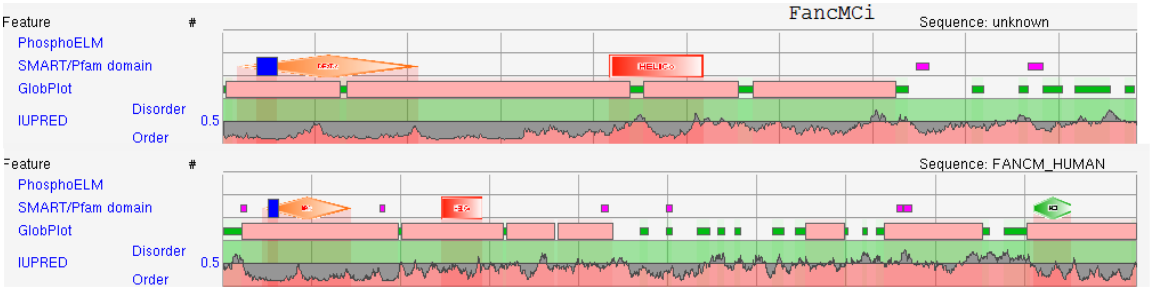
FancL

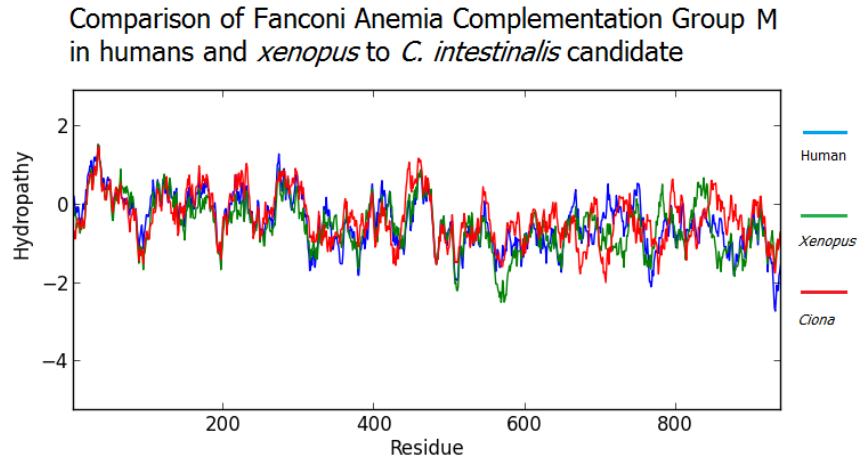

Feature
#
FancL-Ci Sequence: unknown
PhosphoELM
SMART/Pfam domain
GlobPlot
IUPRED
Disorder
0.5
Order

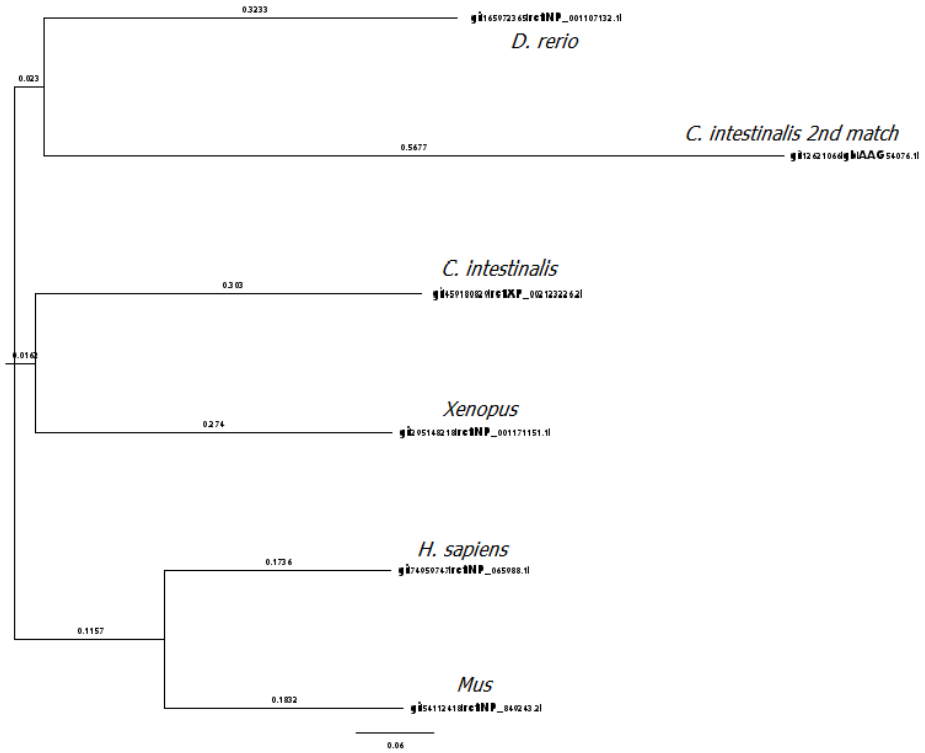

Feature
#
Sequence: FANCL_HUMAN
PhosphoELM
SMART/Pfam domain
GlobPlot
IUPRED
Disorder
0.5
Order


Comparison of Fanconi Anemia Complementation Group L
in humans and xenopus to C. intestinalis candidate
Hydropathy
2
0
−2
−4
50
100
150
200
250
300
Residue
Human
Xenopus
Ciona

FancM

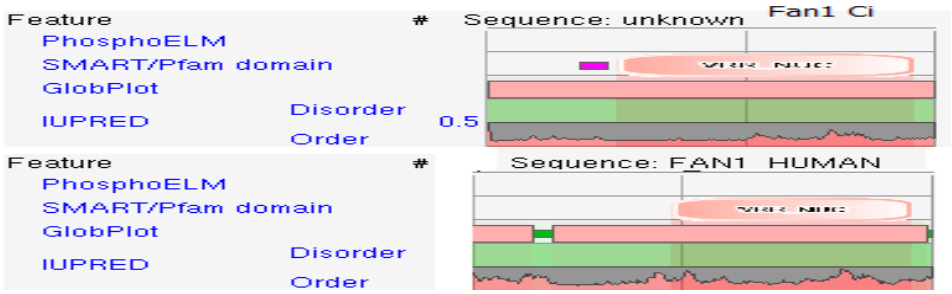

FancMCi
Sequence: unknown
Feature
#
PhosphoELM
SMART/Pfam domain
GlobPlot
IUPRED
Disorder
Order
0.5
HELICc
Sequence: FANCM_HUMAN


Comparison of Fanconi Anemia Complementation Group M
in humans and xenopus to C. intestinalis candidate
Hydropathy
2
0
−2
−4
200
400
600
800
Residue
Human
Xenopus
Ciona

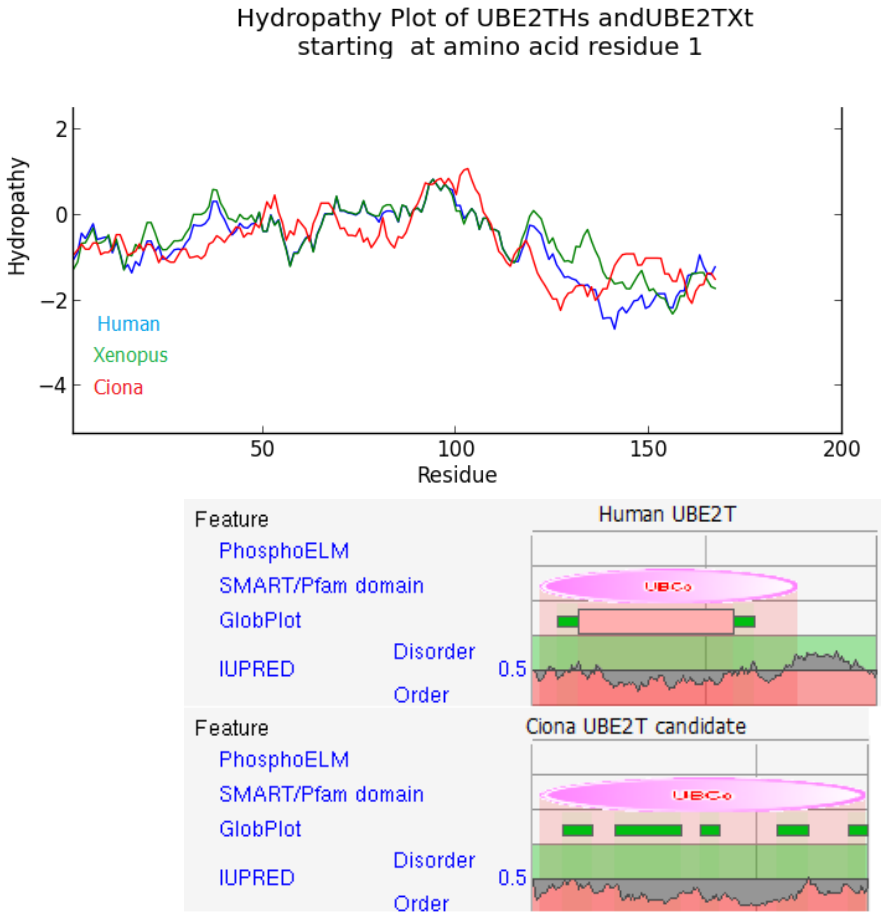

0.3233
D. rerio
0.023
C. intestinalis 2nd match
0.5677
C. intestinalis
0.303
0.0162
Xenopus
0.274
H. sapiens
0.1736
0.1157
Mus
0.1832
0.06

The secondary structure plot for FANCM indicates that both the human and *C. intestinalis* candidate possess a standard helicase as well as a DEAD-like helicase, with the DEAD-like helicase about 200 amino acids upstream in both. The hydropathy plot shows high levels of similarity ($R^2 = 0.321$ for positive matches) and the correlation is even higher towards the N-terminal end ($R^2 = 0.446$ for the first 500 amino acids) The phylogeny data indicates that the *Ciona* candidate is more closely related to vertebrate FANCM than it is the second closest *Ciona* match through RBB/RSD.

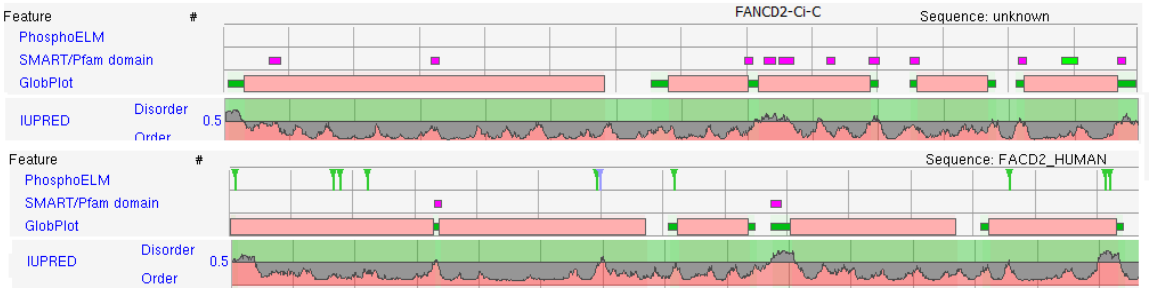
FAN1 (Fanconi-Associated Nuclease 1)

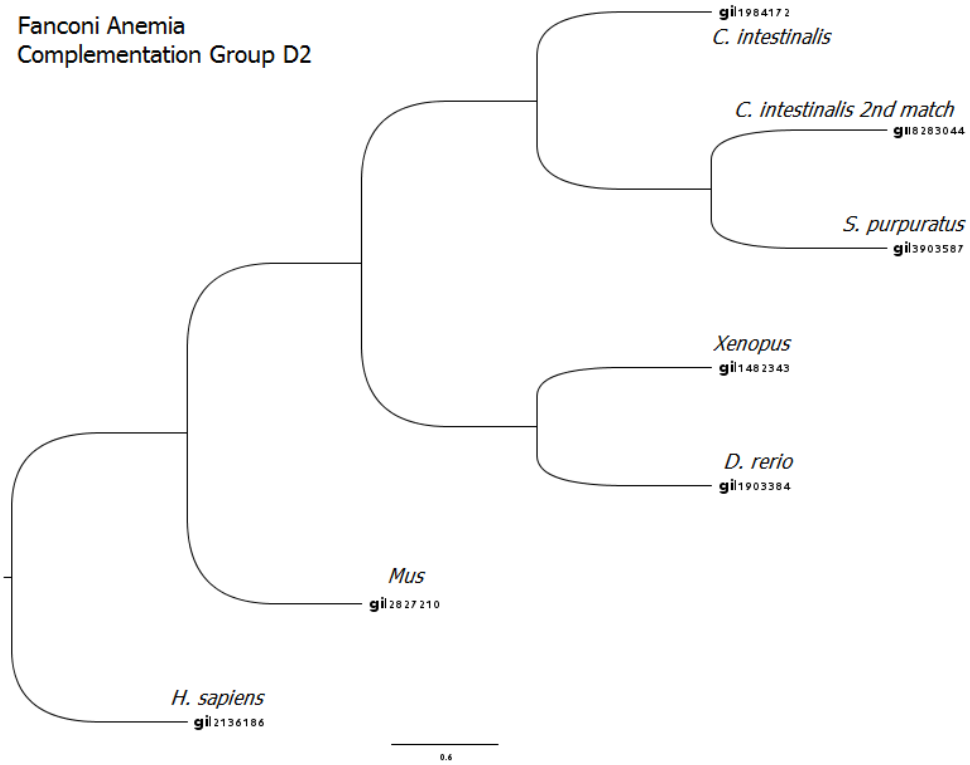


UBE2T (E2 Ubiquitinase):

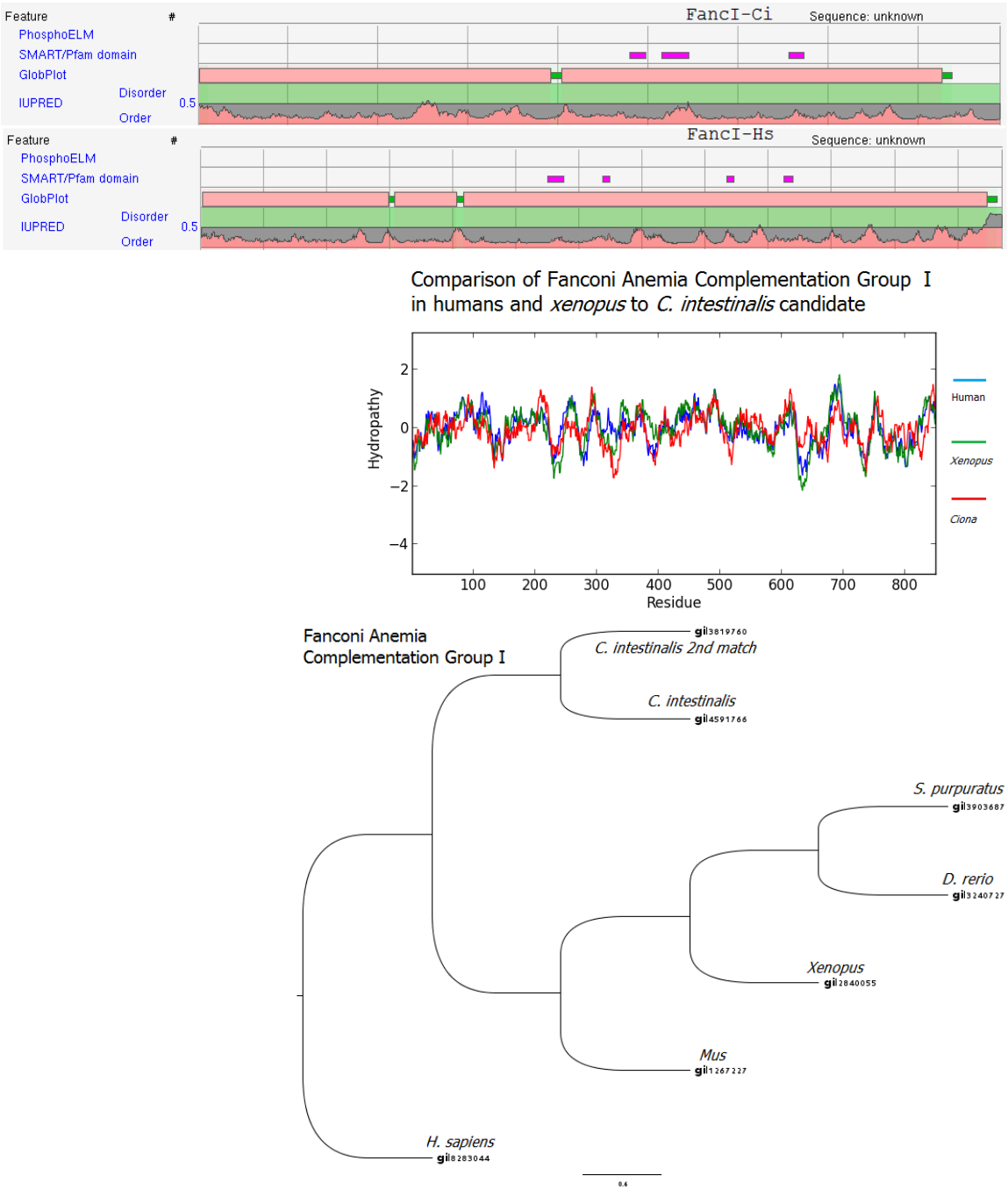


Group II:
FancD2



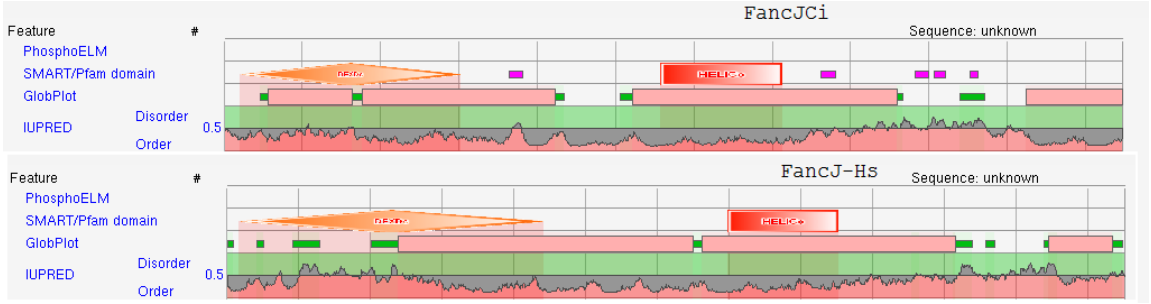


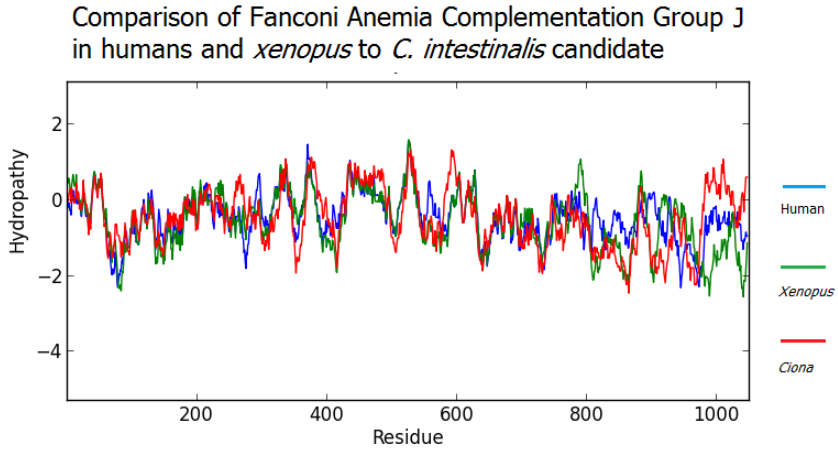
FancI

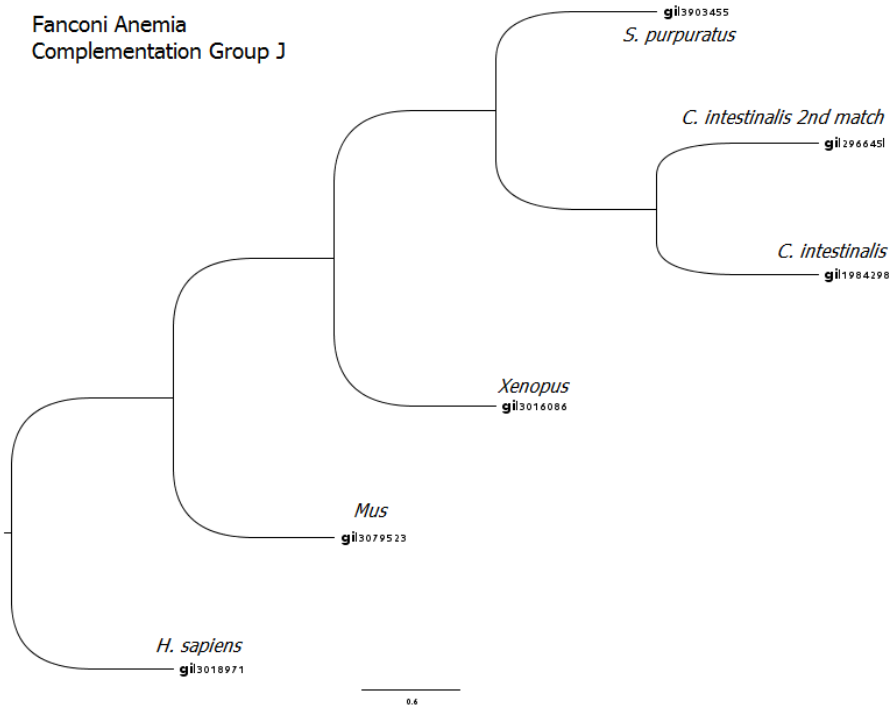


No regions of explicit secondary structure appear on either human or *Ciona* FancI, though that could be a result of regions in the human gene not being in the system. The hydropathy plot shows high levels of similarity, especially at the N-terminal region.

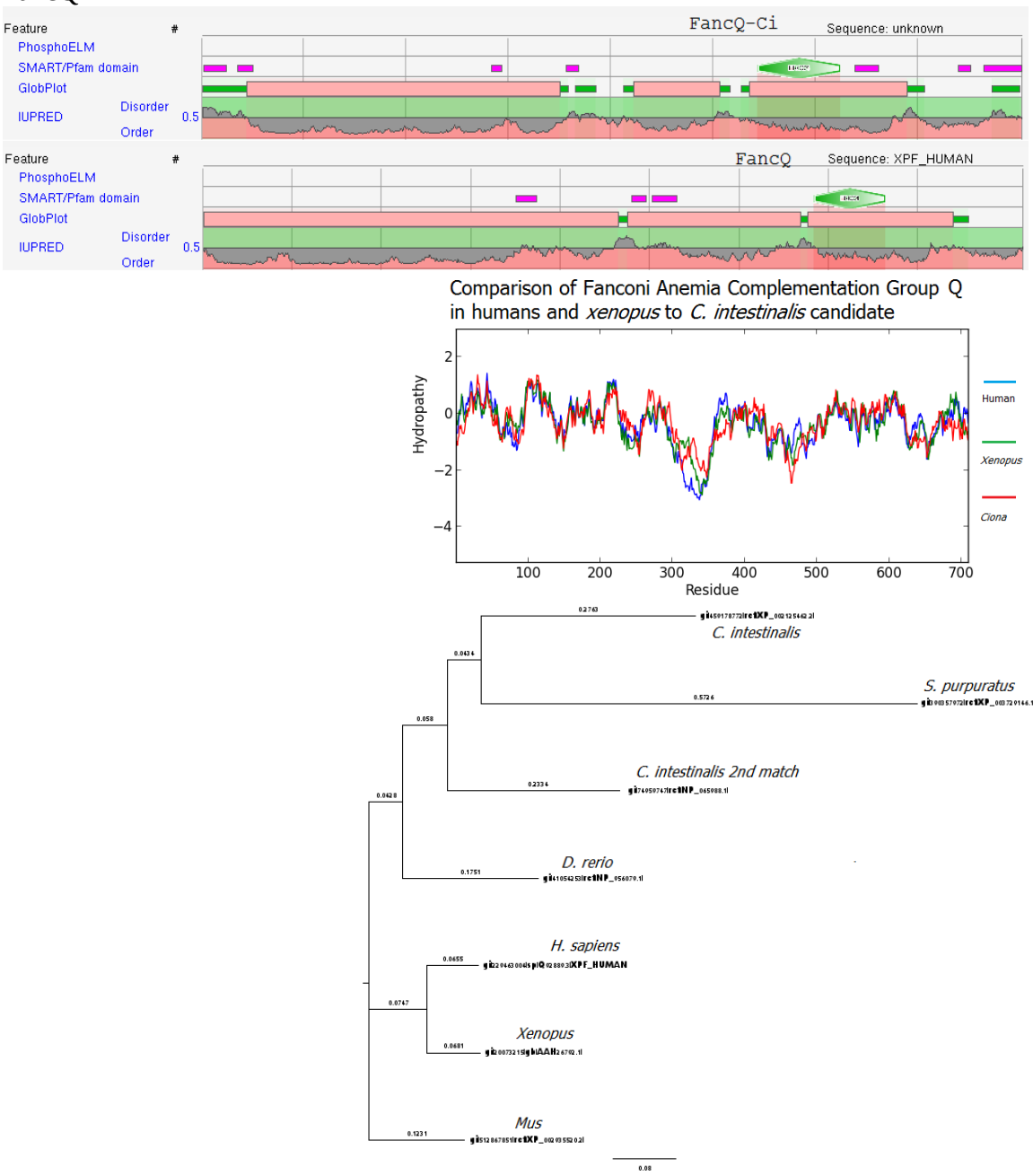
Group III:
FancJ





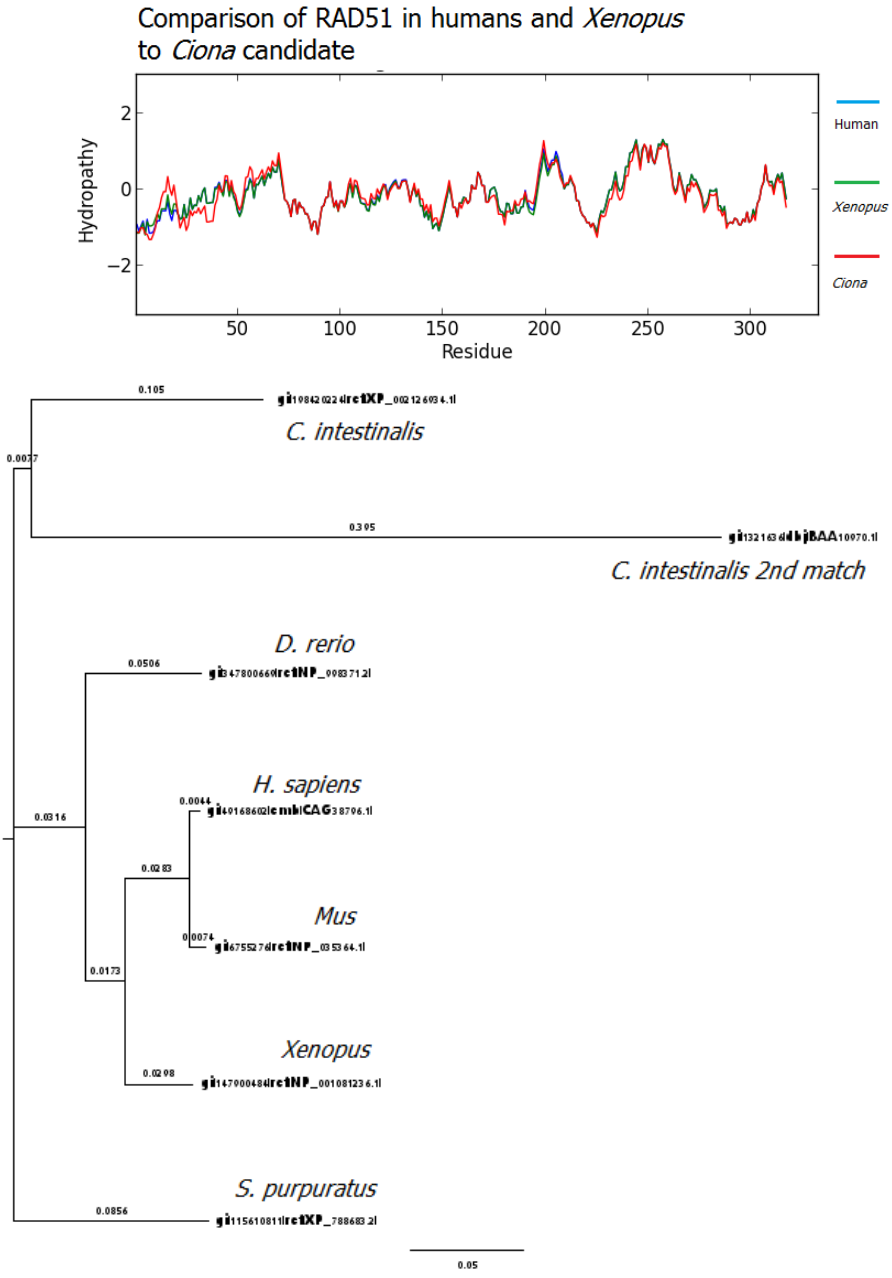


FancQ

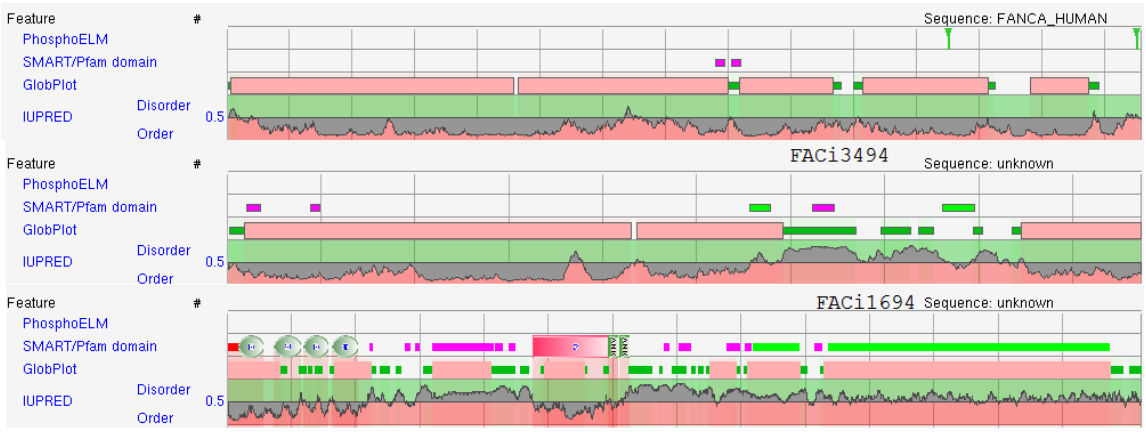


The ELM plot for FANCQ indicates the presence of an ERCC4 nuclease domain approximately 100 amino acids from the C-terminal end of both proteins. The phylogeny data indicates the proteins are less closely related to each other than they are to some of the vertebrate FANCQs.

RAD51:

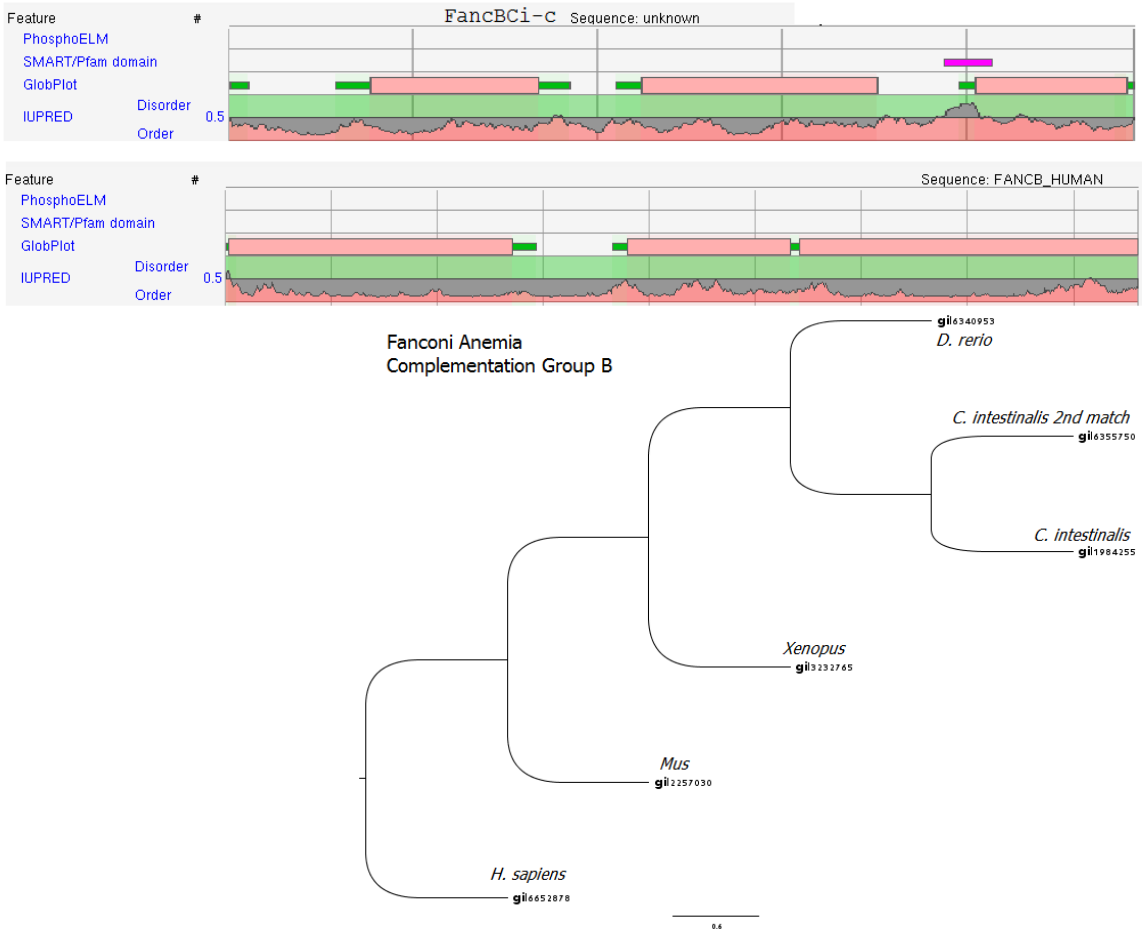

Comparison of RAD51 in humans and *Xenopus* to *Ciona* candidate
Hydropathy
2
0
−2
50
100
150
200
250
300
Residue
Human
Xenopus
Ciona
0.105
C. intestinalis
0.0037
0.395
C. intestinalis 2nd match
D. rerio
0.0506
H. sapiens
0.0044
0.0316
0.0283
Mus
0.0074
0.0173
Xenopus
0.0298
S. purpuratus
0.0856
0.05

Proteins that appear to be absent:
Group I:
FancA



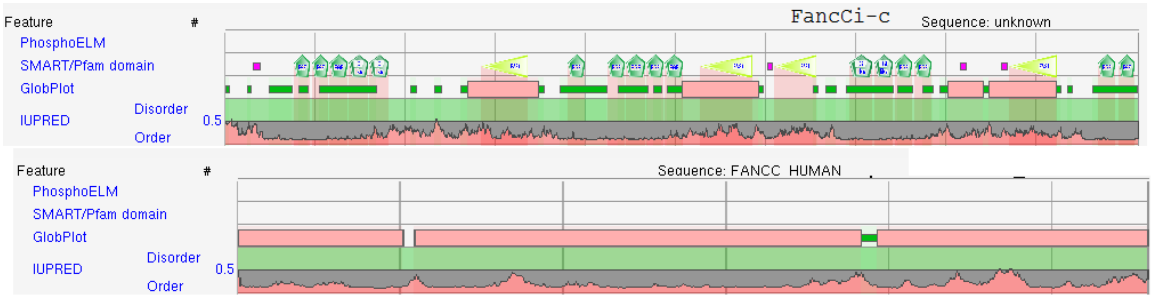
The ELM plots for human FancA protein and the *Ciona* candidate FancA proteins 3494 (best match) and 1694 (second best match) show very little similarity. The best match contains a region of globularity not seen in the human protein, and the protein serving as the second closest match has several protein domains that are not found in the human protein, including a ‘zona pellucida’ region and two ankyrin repeats.

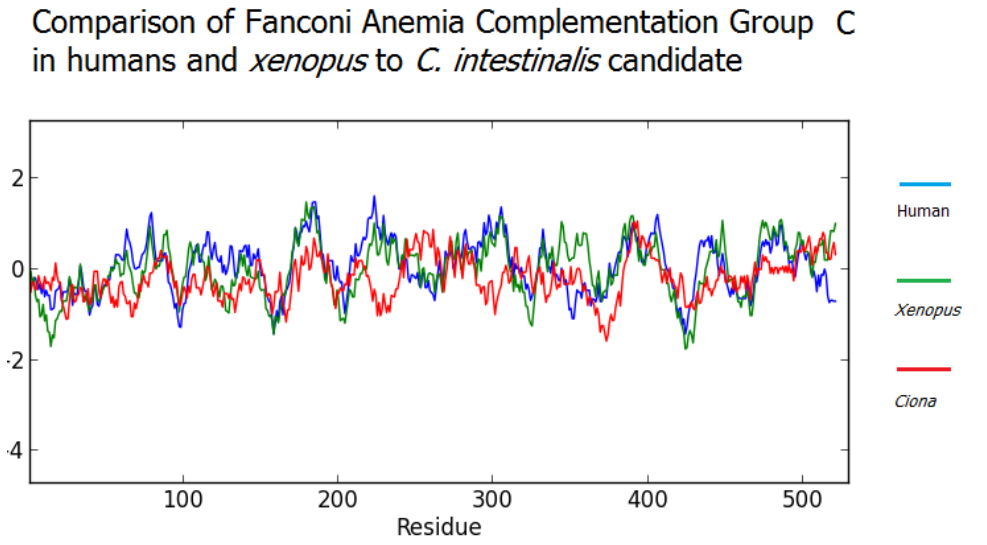
FancB

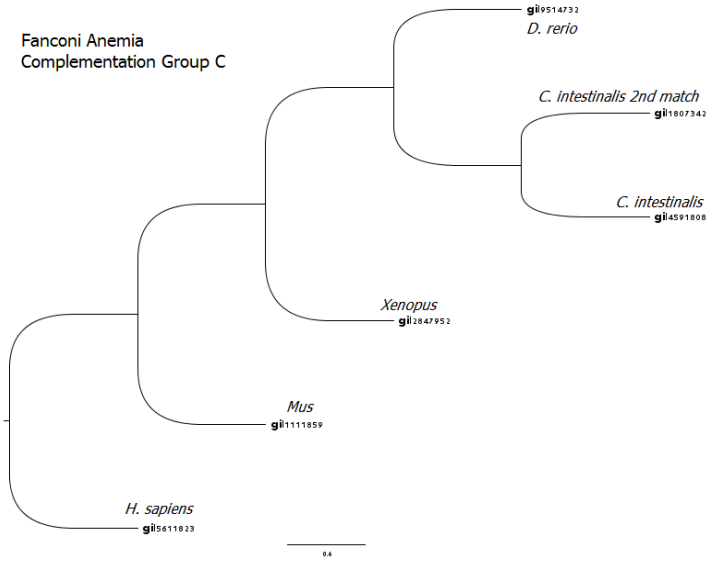


The Fanconi Anemia Group B protein in humans shows a high level of disorder at a point about 75 to 80% of the way through the protein's length where the *Ciona* candidate shows no such region.

FancC





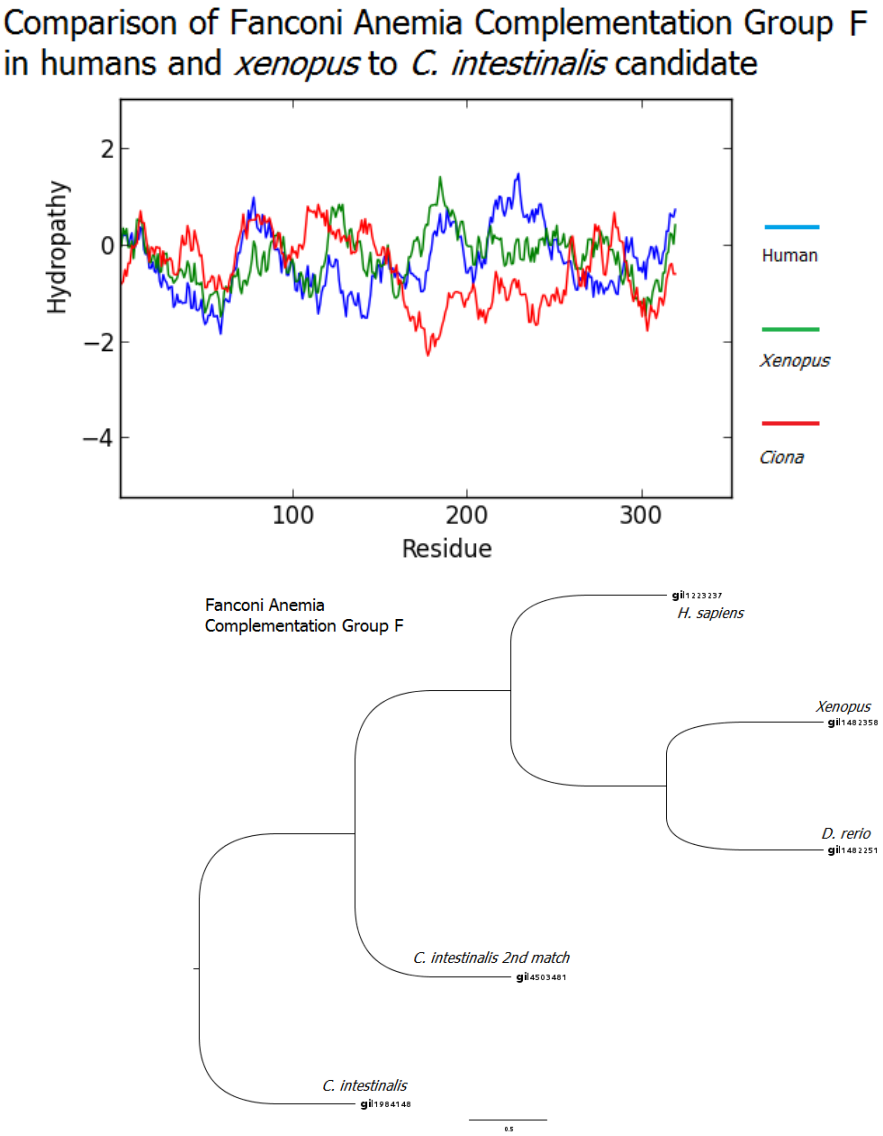


The Fanconi Anemia group C human protein shows none of the secondary structure present in the *Ciona* candidate gene product. The human protein contains three globular domains, while the *Ciona* product possesses four globular domains, 12 EGF or EGF-like domains, and four cell adhesion domains.

The hydrophobicity plot shows two regions of moderate similarity. Amino acids 200 - 325 of the *Ciona* candidate match up with aa 75 - 200 of the human and *Xenopus* proteins. Amino acids 1600 - 1650 closely align to amino acids 350 - 400 of the human protein.

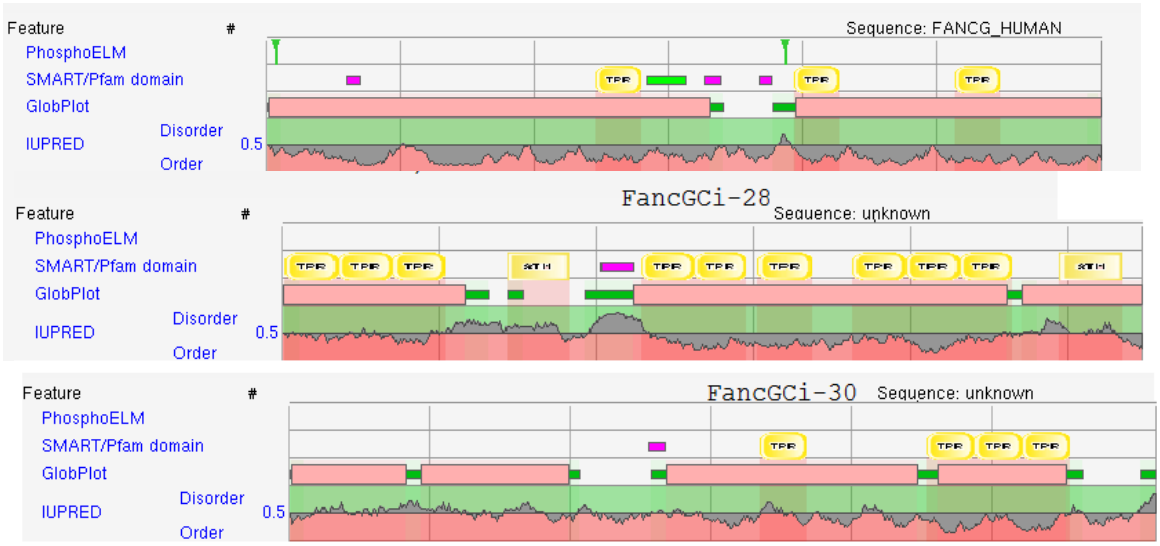
The phylogenetic tree indicates that the *Ciona* candidate gene product is more closely related to the 2nd best BLAST match for *Ciona* than it is to any of the known FancC gene products, giving more evidence that it is not particularly closely related to FancC.

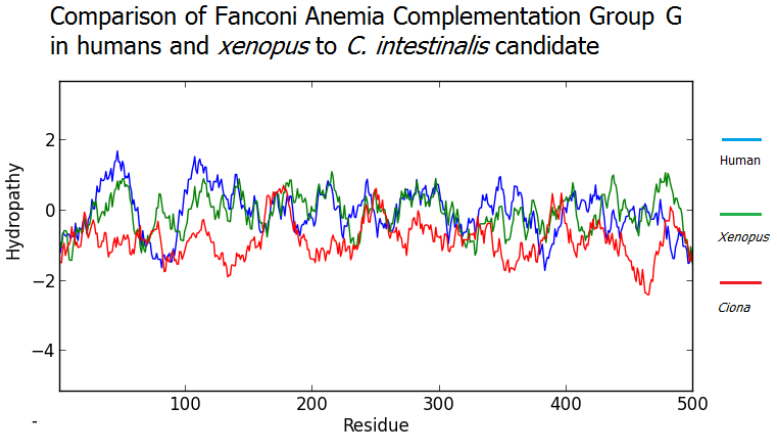
FancF

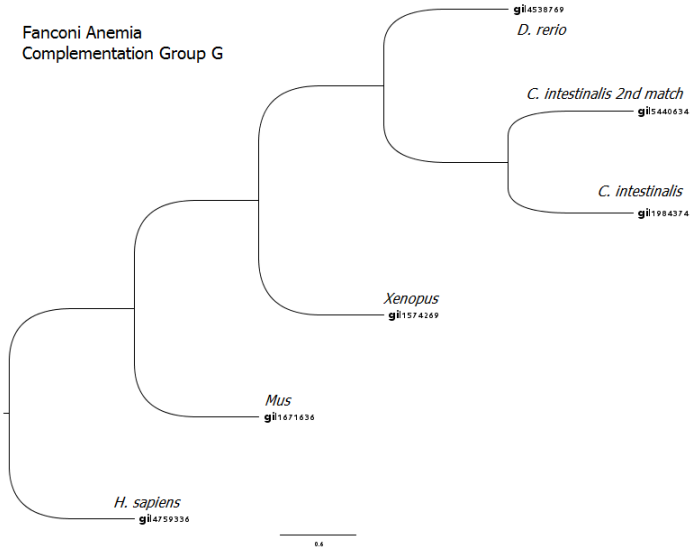


The hydropathy plot of FANCF in humans against FANCF in Xenopus and the Ciona candidate gene shows very little correlation. The phylogenetic tree also shows the Ciona proteins being more closely related to each other than they are to other FANCF proteins.

FancG

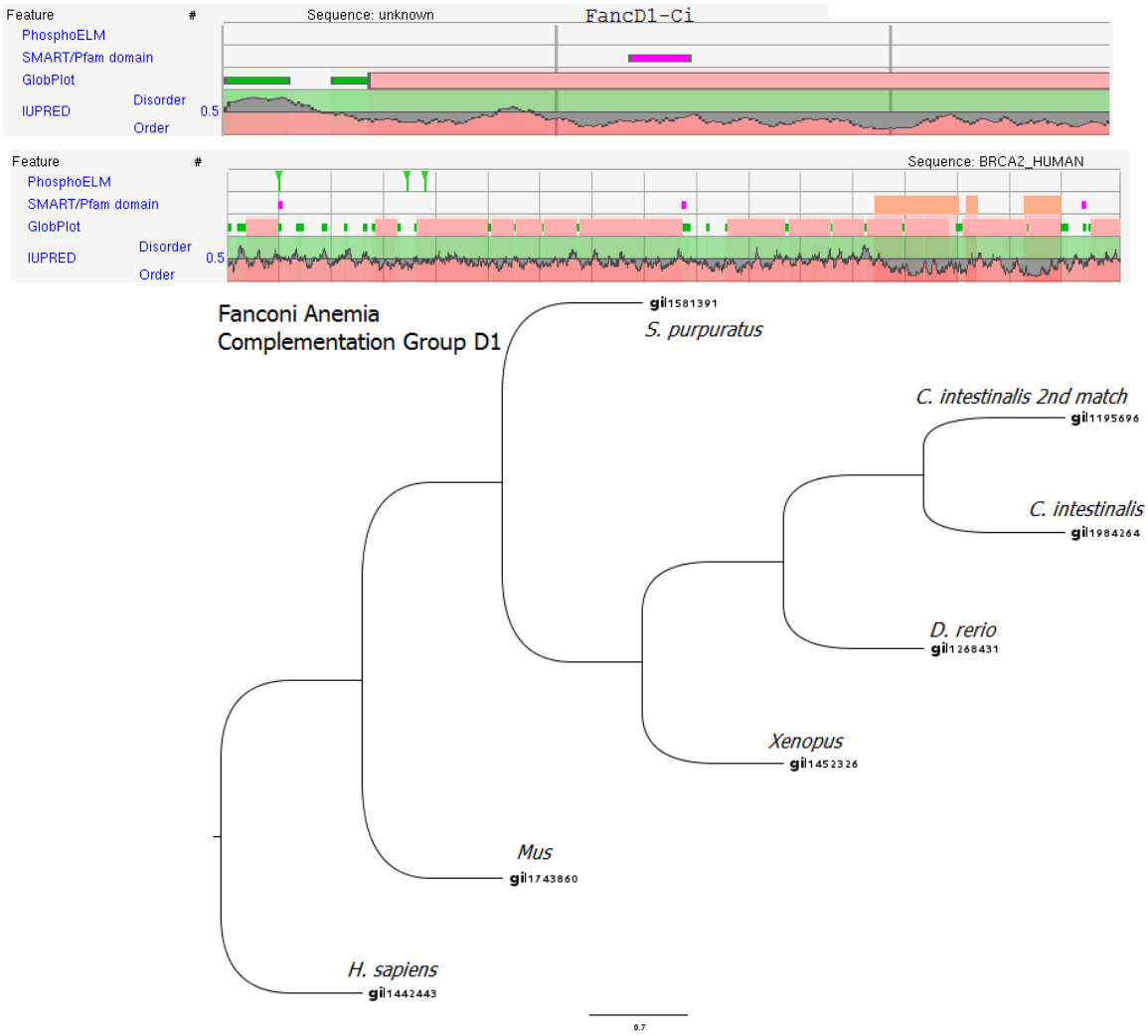

Feature
#
PhosphoELM
SMART/Pfam domain
GlobPlot
IUPRED
Disorder
Order
0.5
Sequence: FANCG_HUMAN
TPR
FancGCi-28
Sequence: unknown
TPR
XTH
FancGCi-30
Sequence: unknown
TPR


Comparison of Fanconi Anemia Complementation Group G
in humans and xenopus to C. intestinalis candidate
Hydropathy
2
0
−2
−4
100
200
300
400
500
Residue
Human
Xenopus
Ciona

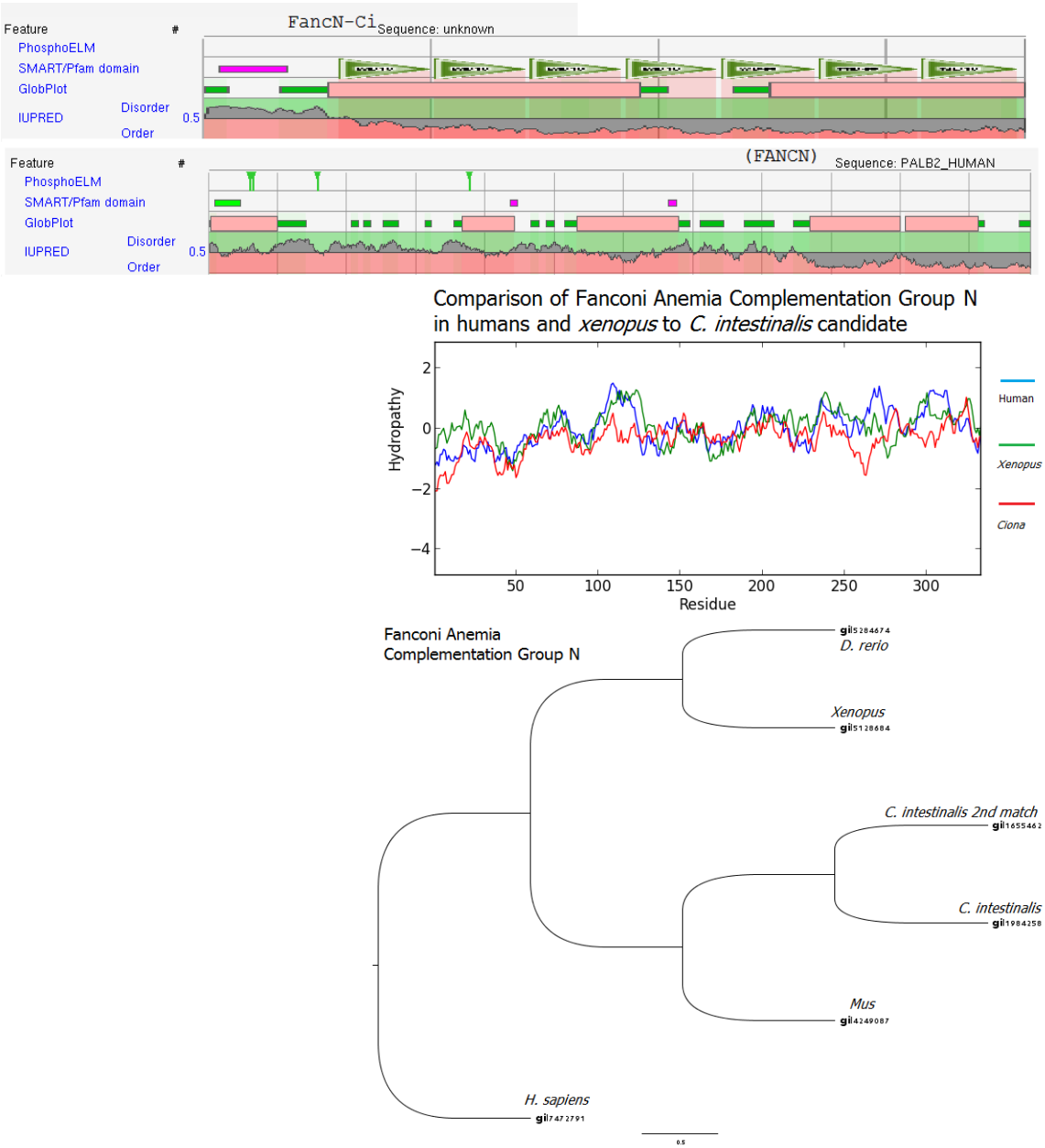

Fanconi Anemia
Complementation Group G
gi4538769
D. rerio
C. intestinalis 2nd match
gi5440634
C. intestinalis
gi1984374
Xenopus
gi1574269
Mus
gi1671636
H. sapiens
gi4759336

The secondary structure plot for FANCG shows the presence of multiple TPR regions, a common scaffolding structure (Blatch, 1999). The phylogenetic data indicates the *Ciona* proteins are more closely related to each other than they are vertebrate Fanconi gene products.

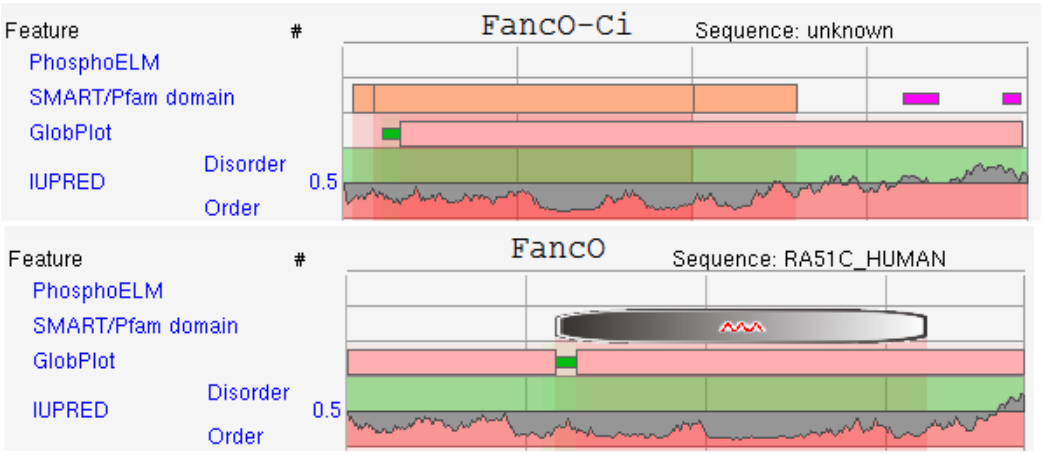
Group III:
FancD1:

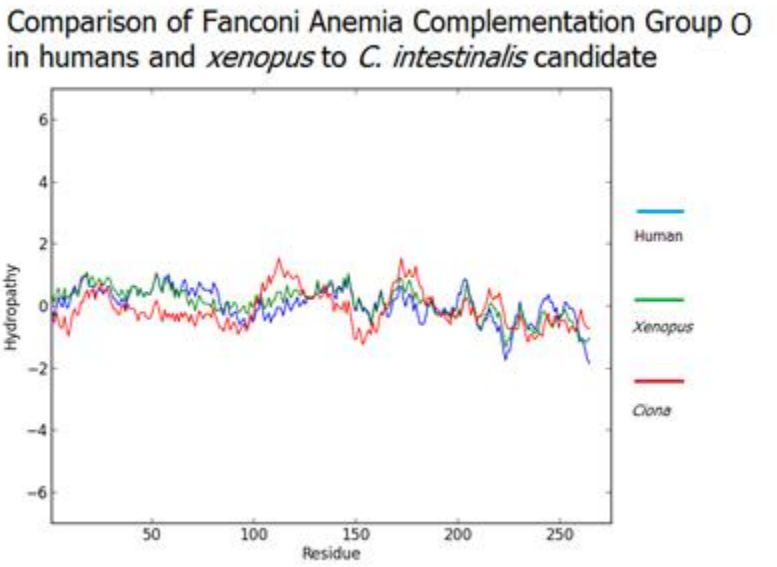


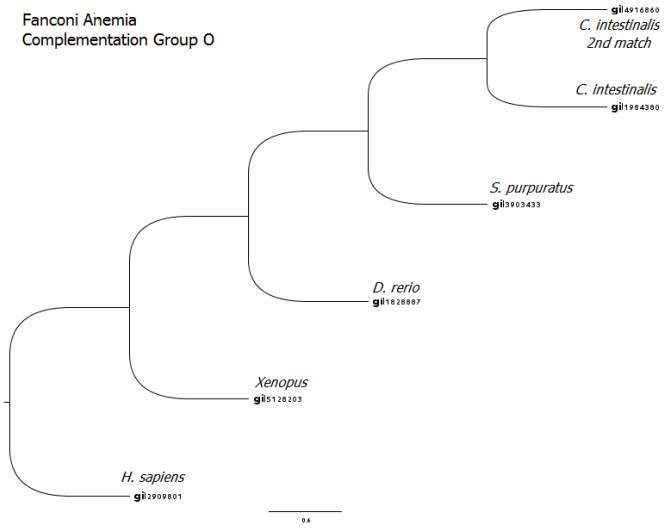
FancN


Feature
#
FancN-Ci
Sequence: unknown
PhosphoELM
SMART/Pfam domain
GlobPlot
IUPRED
Disorder
Order
0.5
(FANCN)
Sequence: PALB2_HUMAN
Comparison of Fanconi Anemia Complementation Group N
in humans and xenopus to C. intestinalis candidate
Hydropathy
Residue
Human
Xenopus
Ciona
Fanconi Anemia
Complementation Group N
D. rerio
Xenopus
C. intestinalis 2nd match
C. intestinalis
Mus
H. sapiens

FancO

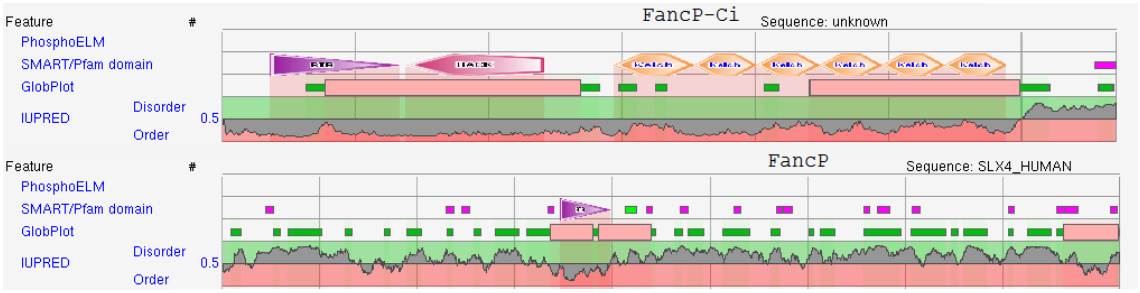

Feature
#
FancO-Ci
Sequence: unknown
PhosphoELM
SMART/Pfam domain
GlobPlot
IUPRED
Disorder
Order
0.5
Feature
#
FancO
Sequence: RA51C_HUMAN
PhosphoELM
SMART/Pfam domain
GlobPlot
IUPRED
Disorder
Order
0.5

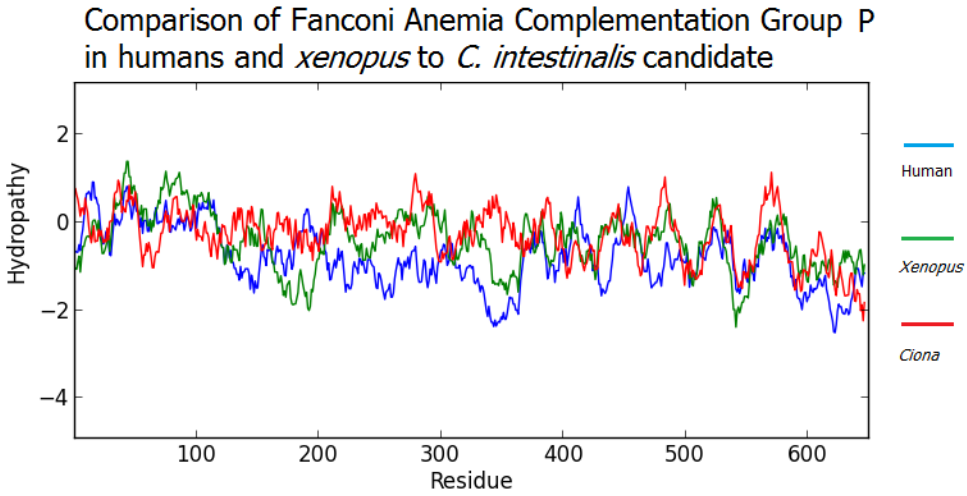

Comparison of Fanconi Anemia Complementation Group O
in humans and xenopus to C. intestinalis candidate
Hydropathy
Residue
Human
Xenopus
Ciona

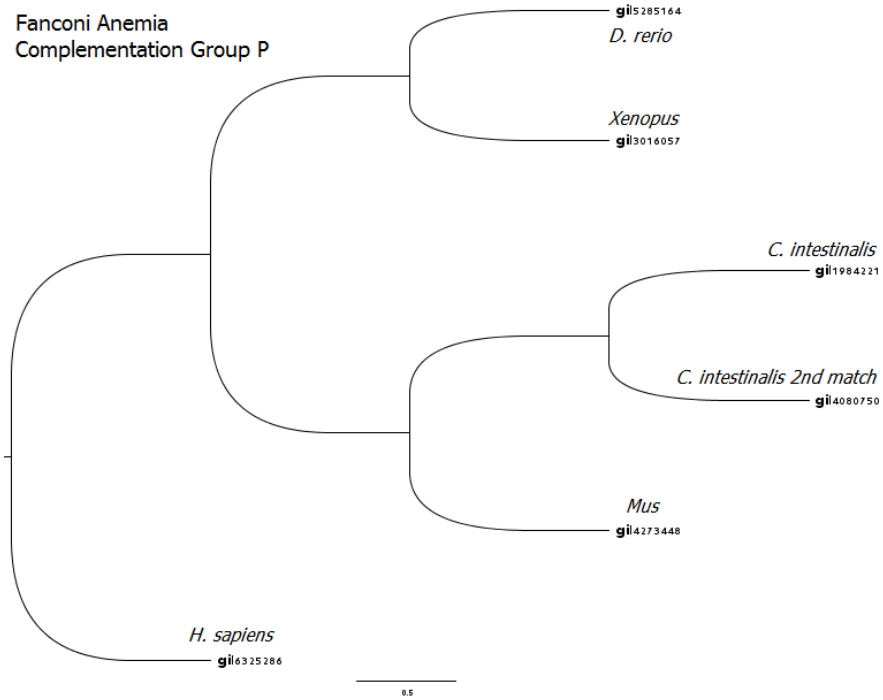

Fanconi Anemia
Complementation Group O
C. intestinalis
2nd match
C. intestinalis
S. purpuratus
D. rerio
Xenopus
H. sapiens

The ELM plot for FANCO indicates the presence of an AAA+ ATPase region in the human protein (which is actually RAD51, a different protein in the pathway) which is entirely absent from the Ciona Fanconi group O candidate.

FancP

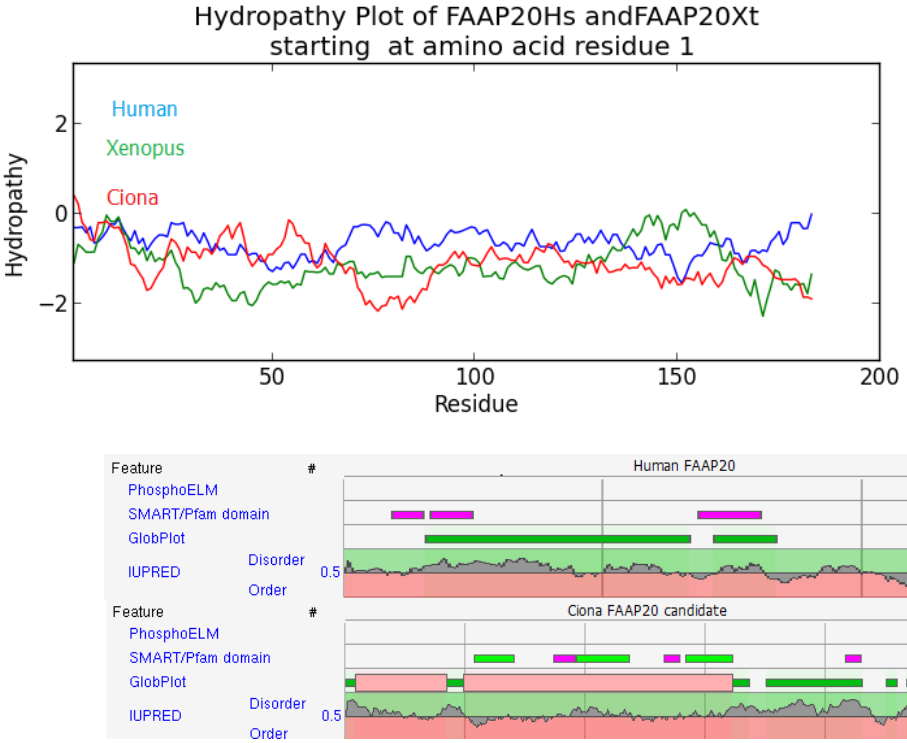

Feature
#
PhosphoELM
SMART/Pfam domain
GlobPlot
IUPRED
Disorder
Order
0.5
FancP-Ci
Sequence: unknown
FancP
Sequence: SLX4_HUMAN


Comparison of Fanconi Anemia Complementation Group P
in humans and xenopus to C. intestinalis candidate
Hydropathy
2
0
−2
−4
100
200
300
400
500
600
Residue
Human
Xenopus
Ciona

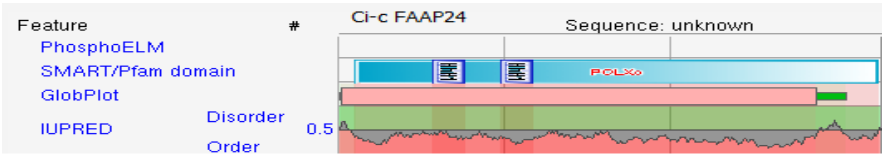

Fanconi Anemia
Complementation Group P
gi5285164
D. rerio
Xenopus
gi3016057
C. intestinalis
gi1984221
C. intestinalis 2nd match
gi4080750
Mus
gi4273448
H. sapiens
gi6325286
0.5

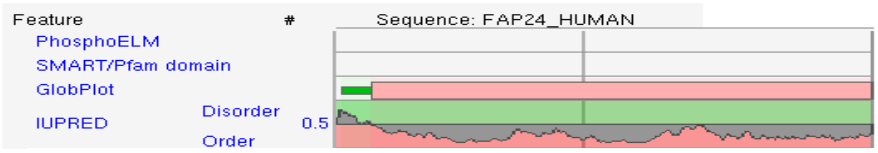
The secondary structure data for FANCP indicates a Bric-a-Brac domain in both the human and *Ciona* candidate genes, which most likely accounts for RBB and RSD giving high similarity scores. The *Ciona* candidate gene product also possesses several Kelch motifs, another structural region. There is a high region of similarity in amino acids around the region of the Kelch motifs in the *Ciona* protein, though there is no indication of these in the human FANCP. The phylogeny data indicates that the *Ciona* candidate gene is more closely related to its second best RBB/RSD match than it is any of the vertebrate FANCPs

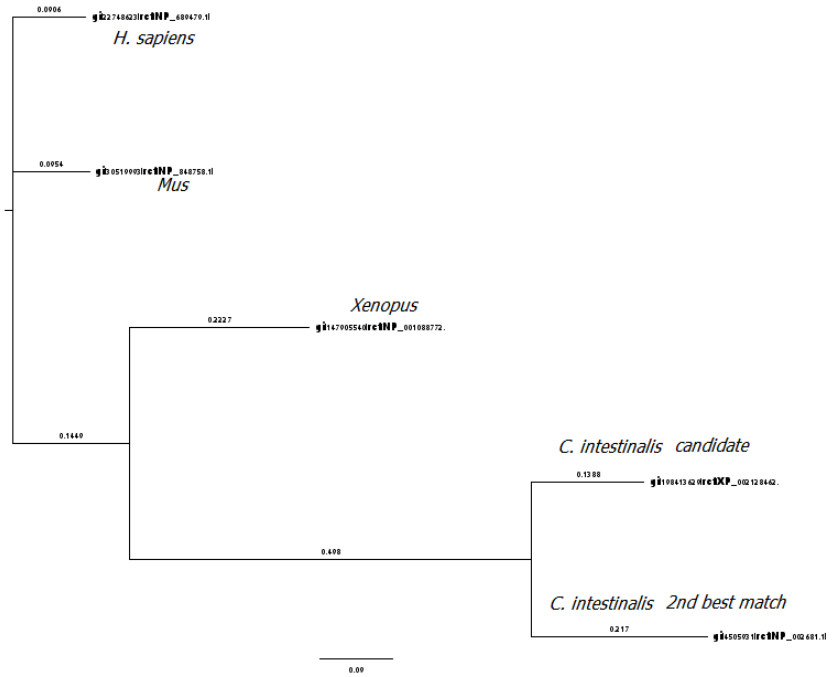
FAAP20:



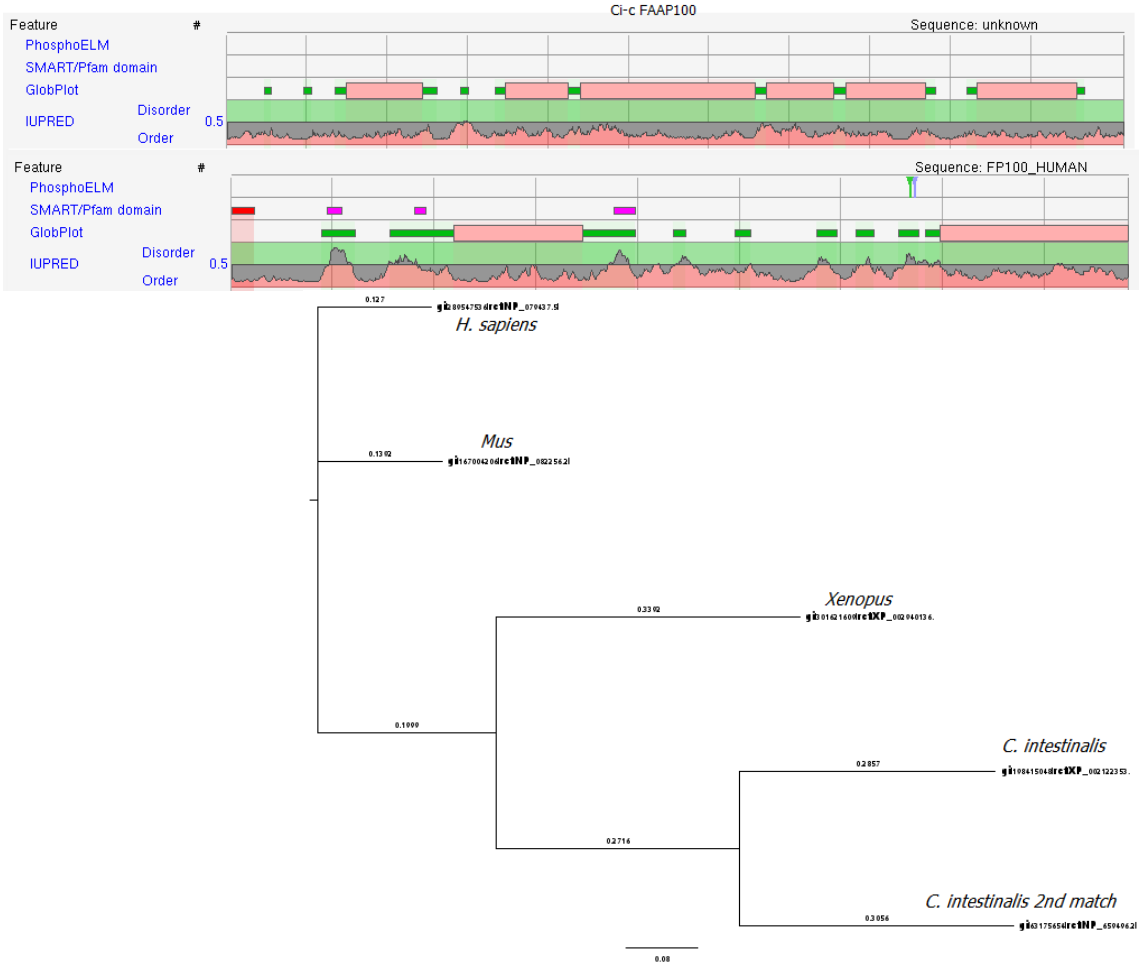
FAAP24:


Feature
#
Ci-c FAAP24
Sequence: unknown
PhosphoELM
SMART/Pfam domain
GlobPlot
IUPRED
Disorder
Order
0.5
POLXc


Feature
#
Sequence: FAP24_HUMAN
PhosphoELM
SMART/Pfam domain
GlobPlot
IUPRED
Disorder
Order
0.5


H. sapiens
Mus
Xenopus
C. intestinalis candidate
C. intestinalis 2nd best match

FAAP100:


Ci-c FAAP100
Feature
#
Sequence: unknown
PhosphoELM
SMART/Pfam domain
GlobPlot
IUPRED
Disorder
Order
0.5
Feature
#
Sequence: FP100_HUMAN
PhosphoELM
SMART/Pfam domain
GlobPlot
IUPRED
Disorder
Order
0.5
0.127
H. sapiens
Mus
0.1392
Xenopus
0.3392
0.1999
C. intestinalis
0.2857
0.2716
C. intestinalis 2nd match
0.3056
0.08

BIBLIOGRAPHY

AbouRizk, Simaan M., Daniel W. Halpin, and James R. Wilson. "Fitting beta distributions based on sample data." *Journal of Construction Engineering and Management* 120.2 (1994): 288-305.

Ali, Abdullah Mahmood, Thiyam Ramsing Singh, and Amom Ruhikanta Meetei. "FANCM–FAAP24 and FANCJ: FA proteins that metabolize DNA." Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis 668.1 (2009): 20-26.

Ali, Abdullah Mahmood, et al. "FAAP20: a novel ubiquitin-binding FA nuclear core-complex protein required for functional integrity of the FA-BRCA DNA repair pathway." *Blood* 119.14 (2012): 3285-3294.

Altschul, Stephen F., *et al*. "Basic local alignment search tool." Journal of molecular biology 215.3 (1990): 403-410.

Auerbach, Arleen, Rockefeller University Fanconi Anemia Mutation Database, <<http://www.rockefeller.edu/fanconi/mutate/>> date of last accession 07/01/14
Blatch, Gregory L., and Michael Lässle. "The tetratricopeptide repeat: a structural motif mediating protein-protein interactions." *Bioessays* 21.11 (1999): 932-939.

Bogliolo, Massimo, *et al*. "Mutations in< i> ERCC4</i>, Encoding the DNA- Repair Endonuclease XPF, Cause Fanconi Anemia." *The American Journal of Human Genetics* 92.5 (2013): 800-806

Brendel, V., Bucher, P., Nourbakhsh, I., Blais- dell, B.E., Karlin, S. (1992) Methods and algorithms for statistical analysis of protein sequences. *Proc. Natl. Acad. Sci*. USA 89: 2002-2006.

Cantor, Sharon B., *et al*. "BACH1, a novel helicase-like protein, interacts directly with BRCA1 and contributes to its DNA repair function." Cell 105.1 (2001): 149-160.

Chapman B, Chang J (August 2000). "Biopython: Python tools for computational biology". ACM SIGBIO Newslett 20 (2): 15–19. doi:10.1145/360262.360268.

Chen, Jun-Jie, *et al*. "BRCA1, BRCA2, and Rad51 operate in a common DNA damage response pathway." Cancer research 59.7 Supplement (1999): 1752s-1756s

Ciccia, Alberto, *et al*. "Identification of FAAP24, a Fanconi anemia core complex protein that interacts with FANCM." Molecular cell 25.3 (2007): 331-343

Cybulski, Kelly E., and Niall G. Howlett. "FANCP/SLX4: a Swiss army knife of DNA interstrand crosslink repair." *Cell Cycle* 10.11 (2011): 1757-1763.

D'Andrea, Alan D., and Markus Grompe. "The Fanconi anaemia/BRCA pathway." Nature Reviews Cancer 3.1 (2003): 23-34.

Dalke Scientific "Hydrophobicity Plots with Matplotlib." (2011), <<http://www.dalkescientific.com/writings/NBN/plotting.html>> date of last accession 07/01/14

Davidson, Brad. "< i> *Ciona* intestinalis</i> as a model for cardiac development." Seminars in cell & developmental biology. Vol. 18. No. 1. Academic Press, 2007.

Deans, Andrew J., and Stephen C. West. "FANCM connects the genome instability disorders Bloom's Syndrome and Fanconi Anemia." *Molecular cell* 36.6 (2009): 943-953.

Delsuc, Frédéric, *et al*. "Tunicates and not cephalochordates are the closest living relatives of vertebrates." *Nature* 439.7079 (2006): 965-968.

Dinkel, Holger, *et al*. "The eukaryotic linear motif resource ELM: 10 years and counting." Nucleic acids research (2013): gkt1047.

Felsenstein, Joseph. "{PHYLIP}: phylogenetic inference package, version 3.5 c." (1993).

Garcia-Higuera, Irene, *et al*. "Interaction of the Fanconi anemia proteins and BRCA1 in a common pathway." Molecular cell 7.2 (2001): 249-262.

Gari, Kerstin, *et al*. "Remodeling of DNA replication structures by the branch point translocase FANCM." *Proceedings of the National Academy of Sciences* 105.42 (2008): 16107-16112.

Glanz, Anthony, and F. Clarke Fraser. "Spectrum of anomalies in Fanconi anaemia." *Journal of medical genetics* 19.6 (1982): 412-416.

Higashimura, Yasuki, *et al*. "Kelch-like 20 up-regulates the expression of hypoxia-inducible factor-2α through hypoxia-and von Hippel–Lindau tumor suppressor protein-independent regulatory mechanisms." *Biochemical and biophysical research communications* 413.2 (2011): 201-205.

Howlett, Niall G., *et al*. "The Fanconi anemia pathway is required for the DNA replication stress response and for the regulation of common fragile site stability." *Human molecular genetics* 14.5 (2005): 693-701.

Hughes, Austin L., and Robert Friedman. "Loss of ancestral genes in the genomic evolution of *Ciona* intestinalis." *Evolution & development* 7.3 (2005): 196-200.

Kanehisa, M. and Goto, S.; KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 28, 27-30 (2000).

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. *Genome Res.* 2002 Jun;12(6):996-1006.

Kent WJ. BLAT - the BLAST-like alignment tool. *Genome Res.* 2002 Apr;12(4):656-64.

Kim, Hyungjin, and Alan D. D'Andrea. "Regulation of DNA cross-link repair by the Fanconi anemia/BRCA pathway." *Genes & development* 26.13 (2012): 1393-1408.

Larkin, Mark A., *et al*. "Clustal W and Clustal X version 2.0." *Bioinformatics* 23.21 (2007): 2947-2948.

Leung, Justin Wai Chung, et al. "Fanconi anemia (FA) binding protein FAAP20 stabilizes FA complementation group A (FANCA) and participates in interstrand cross-link repair." *Proceedings of the National Academy of Sciences* 109.12 (2012): 4491-4496.

Ling, Chen *et al* "FAAP100 is essential for activation of the Fanconi Anemia-associated DNA damage response pathway." TheEMBO journal 26.8 (2007): 2104-2114

Liu, Ting, *et al*. "FAN1 acts with FANCI-FANCD2 to promote DNA interstrand cross-link repair." Science 329.5992 (2010): 693-696.

MacKay, Craig, *et al*. "Identification of KIAA1018/FAN1, a DNA repair nuclease recruited to DNA damage by monoubiquitinated FANCD2." Cell 142.1 (2010): 65-76.

Machida, Yuichi J., et al. "UBE2T is the E2 in the Fanconi anemia pathway and undergoes negative autoregulation." *Molecular cell* 23.4 (2006): 589-596.

Meetei, Amom Ruhikanta, *et al*. "A novel ubiquitin ligase is deficient in Fanconi anemia." Nature genetics 35.2 (2003): 165-170.

Meetei, Amom Ruhikanta, *et al*. "X-linked inheritance of Fanconi anemia complementation group B." *Nature genetics* 36.11 (2004): 1219-1224.

Meetei, Amom Ruhikanta, et al. "A human ortholog of archaeal DNA repair protein Hef is defective in Fanconi anemia complementation group M." *Nature genetics* 37.9 (2005): 958-963.

Niedernhofer, Laura J., Astrid S. Lalai, and Jan HJ Hoeijmakers. "Fanconi anemia (cross) linked to DNA repair." *Cell* 123.7 (2005): 1191-1198.

Nordberg H, Cantor M, Dusheyko S, Hua S, Poliakov A, Shabalov I, Smirnova T, Grigoriev IV, Dubchak I.The genome portal of the Department of Energy Joint Genome Institute: 2014 updates. **Nucleic Acids Res.** 2014,42(1):D26-31.

Philips, Alexandre, *et al*. "Ascidians as a vertebrate-like model organism for physiological studies of Rho GTPase signaling." Biology of the Cell 95.5 (2003): 295-302.

Rahman, Nazneen, *et al*. "PALB2, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene." *Nature genetics* 39.2 (2006): 165-167.

Satoh, Nori, *et al*. *Ciona* intestinalis: an emerging model for whole-genome analyses." Trends in Genetics 19.7 (2003): 376-381.

Schmidt, Heiko A., *et al*. "TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing." Bioinformatics 18.3 (2002): 502-504.

Schultz J, Milpetz F, Bork P, Ponting CP (May 1998). "SMART, a simple modular architecture research tool: identification of signaling domains". *Proc. Natl. Acad. Sci. U.S.A.* **95** (11): 5857–64.

Sigrist, Christian JA, *et al*. "PROSITE, a protein domain database for functional characterization and annotation." *Nucleic acids research* 38.suppl 1 (2010): D161-D166.

Sridharan, Deepa, *et al*. "Nonerythroid αII spectrin is required for recruitment of FANCA and XPF to nuclear foci induced by DNA interstrand cross-links." *Journal of cell science* 116.5 (2003): 823-835.

Stoepker, Chantal, *et al*. "SLX4, a coordinator of structure-specific endonucleases, is mutated in a new Fanconi anemia subtype." Nature genetics 43.2 (2011): 138-141.

Suhasini, Avvaru N., and Robert M. Brosh Jr. "Fanconi anemia and Bloom's syndrome crosstalk through FANCJ–BLM helicase interaction." Trends in Genetics 28.1 (2012): 7-13.

Takata, Minoru, *et al*. "Chromosome instability and defective recombinational repair in knockout mutants of the five Rad51 paralogs." *Molecular and cellular biology* 21.8 (2001): 2858-2866.

Titus, Tom A., *et al*. "The Fanconi anemia gene network is conserved from zebrafish to human." *Gene* 371.2 (2006): 211-223.

Wall, D. P., H. B. Fraser, and A. E. Hirsh. "Detecting putative orthologs." *Bioinformatics* 19.13 (2003): 1710-1711.

Wang, Weidong. "Emergence of a DNA-damage response network consisting of Fanconi Anemia and BRCA proteins *Nature Reviews Genetics* 8.10 (2007): 735 - 748

Waterhouse, A.M., Procter, J.B., Martin, D.M.A, Clamp, M. and Barton, G. J. (2009) "Jalview Version 2 - a multiple sequence alignment editor and analysis workbench" Bioinformatics25 (9) 1189-1191 doi: 10.1093/bioinformatics/btp033

Yatsenko, Alexander N., *et al*. "Non-invasive genetic diagnosis of male infertility using spermatozoal RNA: KLHL10mutations in oligozoospermic patients impair homodimerization." *Human molecular genetics* 15.23 (2006): 3411-3419.

Zdobnov EM (Jan 2008). "OrthoDB: the hierarchical catalog of eukaryotic orthologs". *Nucleic Acids Res.* **36**

Zhang, Xiao-Yin, *et al*. "Xpf and not the Fanconi anaemia proteins or Rev3 accounts for the extreme resistance to cisplatin in Dictyostelium discoideum." *PLoS genetics* 5.9 (2009): e1000645.