2016

The Simple Chordate Ciona intestinalis Has a Reduced Complement of Genes Associated with Fanconi Anemia

Edward C. Stanley
Paul A. Azzinaro

See next page for additional authors

Creative Commons License

This work is licensed under a Creative Commons Attribution-Noncommercial 3.0 License

Follow this and additional works at: https://digitalcommons.uri.edu/cmb_facpubs

Citation/Publisher Attribution


This Article is brought to you for free and open access by the Cell and Molecular Biology at DigitalCommons@URI. It has been accepted for inclusion in Cell and Molecular Biology Faculty Publications by an authorized administrator of DigitalCommons@URI. For more information, please contact digitalcommons@etal.uri.edu.
Introduction
DNA repair mechanisms are a major way by which organisms avoid mutations that can lead to disease, especially cancer. However, the complexity of DNA repair pathways has hindered progress in fully understanding how they work. We have examined the genome of the simple chordate animal, Ciona intestinalis, which is the closest invertebrate relative of vertebrates, for genes associated with the repair of DNA interstrand cross-links (ICL repair), to see if it might possess a simplified version of this DNA repair mechanism. Fanconi anemia (FA) is clinically characterized by congenital abnormalities, pediatric bone marrow failure, and increased cancer risk during early adulthood. FA is caused by mutations in one of the 24 genes linked in a complex pathway. The FA pathway proteins have been categorized into three distinct groups: group I represents the FA core complex and comprises FANCA, FANCB, FANCC, FANCE, FANCF, FANCG, FANCL, FANCM, and FANCT/UB2T. The FA core complex catalyzes the site-specific monoubiquitination of the FANCD2 and FANCI (group II) proteins.5,7 FANCL is a RING domain containing E3 ubiquitin ligase,8,9 while UBE2T is an E2 ubiquitin-conjugating enzyme.10 FANCM is a large (230 kDa) scaffold protein that possesses DNA binding and ATPase/translocase activities.11,12

Interaction network is extensive and includes numerous other proteins that function in ICL repair, which have not been genetically linked to FA.4

The FA pathway proteins have been categorized into three distinct groups: group I represents the FA core complex and comprises FANCA, FANCB, FANCC, FANCE, FANCF, FANCG, FANCL, FANCM, and FANCT/UB2T. The FA core complex catalyzes the site-specific monoubiquitination of the FANCD2 and FANCI (group II) proteins.5,7 FANCL is a RING domain containing E3 ubiquitin ligase,8,9 while UBE2T is an E2 ubiquitin-conjugating enzyme.10 FANCM is a large (230 kDa) scaffold protein that possesses DNA binding and ATPase/translocase activities.11,12 The functions of the remaining group I proteins remain poorly understood. The group II proteins FANCD2 and FANCI, when monoubiquitinated, facilitate the recruitment of several key DNA repair proteins, including FAN1, FANCP/SLX4, and CtIP, to the ICL.13–18 The group III FA proteins comprise FANCD1/BRCA2, FANCJ/BRIP1, FANCN/PALB2,
FANCO/RAD51C, FANCP/SLX4, FANCQ/ERCC4, FANCR/RAD51, and FANCS/BRCA1 and function downstream of FANCD2 and FANCI monoubiquitination. These proteins function primarily in the homologous recombination (HR) step of ICL repair. For example, FANCD1/BRCA2, FANCN/PALB2, and FANCO/RAD51C regulate the localization and activity of FANCR/RAD51, a well established and key HR protein.19–25 Several of the FA proteins are ubiquitous among the eukaryotes.26 Almost every organism surveyed possesses both of the group II proteins, as well as FANCL, FANCM, and an associated ubiquitin-conjugating (E2) enzyme (Fig. 1). There is no apparent evolutionary pattern associated with the presence or absence of the group I 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 of the chordates, possess at least four of the group I proteins.

C. intestinalis is a tunicate, the group thought to be the closest invertebrate relative of the vertebrates.27 C. intestinalis has a number of characteristics that make it a promising model for human diseases. Its genome is very compact, at only 115 Mb, fully sequenced, most of which has been mapped to chromosomes. The current genebuild on Ensembl has 16,671 coding genes, as compared with 20,313 in humans.28 Homologs of almost all human gene families are represented, but Ciona does not have the duplicate genes created by the genome duplications that occurred in vertebrates.29 There are curated databases with abundant gene expression data,30,31 as well as a proteome database.32 While in many cases Ciona has lost genes reflecting adaptation to its sessile lifestyle,33 it can still be used to model simplified pathways,34–36 as it possesses a simplified version of the vertebrate body plan, most notably as a larva.37

A previous study focusing on zebrafish38 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. A subset of the vertebrate group I proteins do appear to be present in Ciona, according to our study, suggesting that it may possess a minimal FA pathway.

In order to better assess the total complement of FA-associated genes in C. intestinalis, we have analyzed the protein structure, hydrophobicity, and phylogenetic relationships of candidates for each of the FA genes of vertebrates. These analyses indicate that C. intestinalis has both of the group II genes from vertebrates, as expected, but only one-third of the group I and two-thirds of the group III genes. In comparison with other animals, and even the plant Arabidopsis, C. intestinalis appears to have an extremely depauperate FA pathway. These data suggest that C. intestinalis may be a good model organism to study a simplified FA pathway and gain important insight into the poorly understood molecular basis of the developmental defects of FA patients.

Figure 1. Presence/absence of FA gene orthologs in selected eukaryotes, as determined by this study. Filled boxes denote estimated presence of a gene in that taxon. Outside of Ciona and humans, presence/absence was determined only by a Delta-BLAST search of the NCBI database using the human gene as query. The dendrogram at the top of the figure denotes the relationships between organisms.

Materials and Methods

Obtaining sequences. First, a Reciprocal Best BLAST (RBB)39 search on 24 gene products was performed, searching the human genes of the FA pathway (Table 1) 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 step was augmented with a search by the reciprocal smallest distance (RSD) method,\(^{40}\) which in all but three cases returned the same protein as RBB. In these three cases the RSD candidate had a higher percentage of positive matches, so those proteins were the ones listed in Table 1.

BLAT\(^1\) in the JGI genome portal\(^{42}\) as well as OrthoDB\(^{43}\) was used to look for synteny between human and *Ciona* FA genes, but none was detected for any of the candidates.

**Protein information.** Using ClustalX and Clustal\(\Omega\),\(^{44}\) each *Ciona* FA protein sequence was aligned against the human and *Xenopus laevis* sequence. The sequences were imported into Jalview,\(^{45}\) and the most closely aligned regions were isolated. Hydrophobicity plots of each sequence were created using Biopython and code built and modified from Dalke Scientific.\(^{46}\) To determine whether the results were significant, the Pearson coefficients were evaluated for the *Ciona* amino acid (aa) sequence against the human and *Xenopus* sequences (again using Python), a beta distribution derived for each sequence,\(^{47}\) and a comparison of the critical values to a \(P < 0.002\) level of significance was made. As a standard, \(P < 0.05\) level of significance with 24 tests gives about a 30% chance of a false positive (Type I error), so a more thorough bound of significance was required. The Sidak test,\(^{48}\) a familywise error correction method used to reduce type I errors, suggests a \(P\)-value of \(1 - (1 - 0.05)^{24}\), or about 0.0021, where 0.05 is the original level of significance and 24 is the number of comparison tests performed. This assumes that the genes and their products are independent – there does not appear to be any evidence that a mutation in one FA protein leads to the absence of any of the other FA proteins.

Protein structural models (Figs. 2E, F, J, and K and 4F and G) were constructed using Discovery Studio v. 3.1 (BIOVIA), based on pdb files in the RCSB Protein Data Bank, using 50 iterations with loop refinement. The protein motif diagrams were based on the information in Pfam 29.0.\(^{49}\)

**Phylogenetic analysis.** Full protein sequences (see Supplementary Table S1 for accession numbers) were aligned using MAFFT with default settings.\(^{50}\) Poorly aligned regions were excised using TrimAl v. 1.3 using the Gappyout setting on the Phylemon 2.0 web server.\(^{51}\) RAxML v. 8.0.0\(^{52}\) was used to construct a maximum likelihood (ML) tree with bootstrap number determined with the FC bootstrapping criterion and PROTGAMMABLOSUM62 substitution model. User supplied trees with candidate genes rearranged were statistically evaluated using the Shimodaira–Hasegawa (SH) log likelihood test in RAxML.

**Results**

*Ciona* has orthologs of vertebrate FA genes from each functional group. Our analysis revealed that *Ciona* has highly conserved orthologs of genes from each of the three FA protein groups (Fig. 1). Like all the other multicellular organisms examined, *Ciona* has both members of group II: FANCD2 and FANCI. However, only 4 of 9 members of group I and 5 of 8 members of group III were found, as well as only 2 of several “FA associated” proteins. In fact, *Ciona* appears to have as few or fewer members of the FA pathway of any multicellular organism examined, including plants, slime mold, and the primitive metazoan *Nematostella*.

Below we present evidence for or against orthology in *C. intestinalis* of each of the members of the FA pathway. The first analyses described are for those genes that we estimate are present in *Ciona*, organized by the functional group. We then list those that do not have orthologs in *Ciona* according to our methods. The order of the genes in the text is similar to the vertical order in Figure 1.

**Group I orthologs found.** FANCE. FANCE is part of the FA core complex with an unknown function. RBB returns an uncharacterized *C. intestinalis* protein LOC100186252 (XP_002129936). The *Ciona* candidate protein aligns well with the last 250–300 aa of vertebrate FANCE proteins (\(R^2 = 0.202\)), but on the whole, the correlation is only 0.08 (and the region outside the C-terminal registering at only 0.05; Fig. 2A). The *Ciona* candidate is about 400 aa in length, while vertebrate FANCE proteins are all between 550 and 600 aa. 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 shared secondary structure is found in the ELM analysis (data not shown). The ML best tree (Fig. 2B) groups the *Ciona* candidate LOC100186252 (‘C1UP1’) in a sister group to the vertebrate FANCA proteins, more closely related to the plant and fungal candidates. However, if LOC100186252 is forced to group with the vertebrate FANCE proteins (Fig. 2C), the tree is not significantly worse, while moving LOC100186252 more distant from the FANCE clade is statistically worse (Fig. 2D; \(P < 0.01\)), consistent with the orthology of FANCE. In addition, a crystal structure exists for human FANCE,\(^{53}\) allowing us to perform structural homology modeling between the human protein and the inferred *C. intestinalis* protein (Fig. 2E and F). The 3D models indicate that the structure of LOC100186252 is potentially very similar to human FANCE. Taken as a whole, these data provide support for LOC100186252 being the homolog of FANCE in *C. intestinalis*.

**FANCL.** FANCL is an E3 ubiquitin ligase and a component of the FA core complex, which serves to ubiquitinate FANCD2 and FANCN.\(^7\) RBB returns a putative *Ciona* FANCL protein with an E-value of \(2 \times 10^{-24}\) (Table 1). The *Ciona* candidate hydrophobicity plot shows close correspondence to the vertebrate proteins (Fig. 2G). SMART and Pfam primary sequence-based prediction analyses both detect three amino-terminal WD40 repeats and a carboxy-terminal RING domain in *Ciona* fancl (Fig. 2I), similar to that originally described for human FANCL.\(^5,4\) Subsequent structural analyses of *Drosophila* and human FANCL have revealed...
<table>
<thead>
<tr>
<th>GROUP</th>
<th>HUMAN GENE NAME</th>
<th>HUMAN ACC. NO.</th>
<th>BEST Ciona GENE MATCH</th>
<th>Ciona E-VALUE</th>
<th>FANCA</th>
<th>BRCA2</th>
<th>Reciprocal Human ACC. NO.</th>
<th>Reciprocal Human Gene Match</th>
<th>E-VALUE</th>
<th>Best Ciona Gene Match</th>
<th>E-VALUE</th>
<th>Ciona E-VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>i</td>
<td>FANCA</td>
<td>NP_000126</td>
<td></td>
<td></td>
<td>1 X 10^-32</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>FANCC</td>
<td>NP_001018123</td>
<td></td>
<td></td>
<td>6 X 10^-10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>FANCB</td>
<td>XP_009858877</td>
<td></td>
<td></td>
<td>4 X 10^-146</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>i</td>
<td>FANCL</td>
<td>Q8iuf8</td>
<td></td>
<td></td>
<td>1.7 X 10^-308</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>i</td>
<td>FANCO/</td>
<td>AAC39604</td>
<td></td>
<td></td>
<td>2 X 10^-10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>i</td>
<td>FANCP</td>
<td>NP_114432</td>
<td></td>
<td></td>
<td>1 X 10^-32</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>i</td>
<td>FANCP</td>
<td>NP_009381</td>
<td></td>
<td></td>
<td>1 X 10^-32</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>i</td>
<td>FANCP</td>
<td>NP_0028662</td>
<td></td>
<td></td>
<td>2 X 10^-10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>i</td>
<td>FANCP</td>
<td>NP_115820</td>
<td></td>
<td></td>
<td>1 X 10^-32</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>i</td>
<td>FANCP</td>
<td>NP_055273.2</td>
<td></td>
<td></td>
<td>0.0 X 10^-32</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1. BLAST (RBB/RSD) results. Refer to Supplemental Table S1 for additional accession numbers of sequences used in phylogenetic and other analyses.
that FANCL encompasses three distinct domains: an amino-terminal E2-like fold, a central double RWD-like domain, and a carboxy-terminal RING domain. Structural homology modeling of Ciona FANCL, based on the 3.2 Å Drosophila melanogaster FANCL structure (PDB ID 3K1L), indicates the existence of close structural similarity (Fig. 2J and K). In addition, Clustal Omega multiple sequence alignment (MSA) analyses of human, mouse, and Ciona FANCL indicate that K22, a predicted site of autoubiquitination, is conserved in all three species (data not shown). The ML best tree (Fig. 2H) agrees with this finding, showing that the C. intestinalis candidate falls in a clade with the vertebrate FANCL proteins to the exclusion of the second most similar Ciona and human proteins. However, moving the C. intestinalis candidate further from the vertebrate FANCL clade, or as a sister taxon to the vertebrate FANCL genes, does not make for a statistically worse tree (data not shown). This ambiguity indicates that the phylogenetic evidence for orthology is weak. However, based on the

Table 2. Hydrophobicity plot correlations between identities and positive matches.

<table>
<thead>
<tr>
<th>GROUP</th>
<th>NAME</th>
<th>ID%</th>
<th>HYDROPATHY PLOT ID% R²</th>
<th>POS%</th>
<th>HYDROPATHY PLOT POS% R²</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>FANCA</td>
<td>36%</td>
<td>0.052</td>
<td>58%</td>
<td>0.027</td>
</tr>
<tr>
<td></td>
<td>FANCB</td>
<td>37%</td>
<td>0.045</td>
<td>52%</td>
<td>0.026</td>
</tr>
<tr>
<td></td>
<td>FANCC</td>
<td>32%</td>
<td>0.121</td>
<td>46%</td>
<td>0.052</td>
</tr>
<tr>
<td></td>
<td>FANCE</td>
<td>19%</td>
<td>0.082</td>
<td>34%</td>
<td>0.051</td>
</tr>
<tr>
<td></td>
<td>FANCF</td>
<td>47%</td>
<td>0.039</td>
<td>65%</td>
<td>0.021</td>
</tr>
<tr>
<td></td>
<td>FANCN</td>
<td>56%</td>
<td>0.034</td>
<td>74%</td>
<td>0.021</td>
</tr>
<tr>
<td></td>
<td>FANCN</td>
<td>56%</td>
<td>0.037</td>
<td>74%</td>
<td>0.019</td>
</tr>
<tr>
<td></td>
<td>FANCM</td>
<td>52%</td>
<td>0.369</td>
<td>66%</td>
<td>0.321</td>
</tr>
<tr>
<td></td>
<td>FAAP20</td>
<td>18%</td>
<td>0.053</td>
<td>28%</td>
<td>0.027</td>
</tr>
<tr>
<td></td>
<td>FAAP24</td>
<td>26%</td>
<td>0.177</td>
<td>45%</td>
<td>0.098</td>
</tr>
<tr>
<td></td>
<td>FAAP100</td>
<td>35%</td>
<td>0.114</td>
<td>53%</td>
<td>0.053</td>
</tr>
<tr>
<td></td>
<td>UBE2T</td>
<td>34%</td>
<td>0.311</td>
<td>58%</td>
<td>0.222</td>
</tr>
<tr>
<td>II</td>
<td>FANCD2</td>
<td>25%</td>
<td>0.304</td>
<td>44%</td>
<td>0.222</td>
</tr>
<tr>
<td></td>
<td>FANCI</td>
<td>29%</td>
<td>0.331</td>
<td>50%</td>
<td>0.237</td>
</tr>
<tr>
<td>III</td>
<td>FANCD1</td>
<td>31%</td>
<td>0.127</td>
<td>49%</td>
<td>0.046</td>
</tr>
<tr>
<td></td>
<td>FANCC</td>
<td>37%</td>
<td>0.291</td>
<td>53%</td>
<td>0.236</td>
</tr>
<tr>
<td></td>
<td>FANCN</td>
<td>18%</td>
<td>0.099</td>
<td>45%</td>
<td>0.061</td>
</tr>
<tr>
<td></td>
<td>FANCO</td>
<td>41%</td>
<td>0.328</td>
<td>44%</td>
<td>0.345</td>
</tr>
<tr>
<td></td>
<td>FANCP</td>
<td>32%</td>
<td>0.050</td>
<td>49%</td>
<td>0.030</td>
</tr>
<tr>
<td></td>
<td>FANCO/XP</td>
<td>50%</td>
<td>0.566</td>
<td>68%</td>
<td>0.446</td>
</tr>
<tr>
<td></td>
<td>FANC/BRCA1</td>
<td>82%</td>
<td>0.844</td>
<td>92%</td>
<td>0.847</td>
</tr>
<tr>
<td></td>
<td>BRCA1</td>
<td>27%</td>
<td>0.117</td>
<td>46%</td>
<td>0.322</td>
</tr>
<tr>
<td></td>
<td>ERCC1</td>
<td>47%</td>
<td>0.396</td>
<td>65%</td>
<td>0.446</td>
</tr>
<tr>
<td></td>
<td>Fan1</td>
<td>41%</td>
<td>0.447</td>
<td>63%</td>
<td>0.337</td>
</tr>
</tbody>
</table>

Note: ID% refers to the Delta-BLAST result for identical amino acid matches. Pos% refers to Delta-BLAST results for positive amino acid matches, eg, aa from the same functional groups.
Evolutionary Bioinformatics - candidate.

Fig. Blast matches to fanc E fancl protein structures. 12 RBB returns a putative -11,57–61 s loc100186252, showing extreme similarity of overall structures. (i and other G- (d) Table C. intestinalis These data indicate strong support for the orthology of the proteins in a clade with 94% bootstrap support (Fig. 3C). FANCM candidate clusters with the vertebrate FANCM the amino-terminus (Fig. 3A). In the ML tree, the FANCM protein also possesses a degenerate XPF/ERCC4 endonuclease domain that the protein lacks. The human FANCM candidate proteins possess a DEAH-box helicase/DNA-structure analysis shows that both the human and Ciona proteins during ICL repair. remodeling and the chromatin recruitment of the group I core complex and plays a key role in DNA replication fork structure and is the specific one implicated in the monoubiquitination of FANCD2. Patient-derived mutations in the UBE2T gene have recently been discovered in two unrelated patients, leading to a call to denote UBE2T as FANCT.12 The Delta-BLAST search returns Ciona ubc2-17kd as the closest match to human UBE2T. However, the reciprocal BLAST against human proteins returns human UBE2D4 with an E-value of 2 × 10⁻⁷⁷ (Table 1). The RSD method returns Ciona ubc2 J1l with an E-value of 9 × 10⁻⁷⁶. Apparently, these very similar E2 ubiquitin-conjugating enzymes cannot be reliably distinguished by BLAST searches (Fig. 3E). The hydropathy and phylogenetic analyses (Fig. 3D and F) do not help to resolve the exact relationship. In the hydropathy plot, it is apparent that both the Xenopus and Ciona proteins

structural similarities, there is reasonably strong support for the C. intestinalis gene to be a true ortholog of human FANCL.

FANCM. FANCM is also a component of the FA core complex and plays a key role in DNA replication fork remodeling and the chromatin recruitment of the group I proteins during ICL repair.13,57–61 RBB returns a putative Ciona FANCM protein as the closest match. Secondary structure analysis shows that both the human and Ciona candidate proteins possess a DEAH-box helicase/DNA-stimulated ATPase domain (Fig. 3B). The human FANCM protein also possesses a degenerate XPF/ERCC4 endonuclease domain that the Ciona protein lacks.12 The hydrophobicity plot shows high levels of correlation, especially toward the amino-terminus (Fig. 3A). In the ML tree, the Ciona FANCM candidate clusters with the vertebrate FANCM proteins in a clade with 94% bootstrap support (Fig. 3C). These data indicate strong support for the orthology of the C. intestinalis candidate.

FANCT/UBE2T. FANCT/UBE2T is one of the many E2 ubiquitin-conjugating enzymes found in the human proteome and is the specific one implicated in the monoubiquitination of FANCD2 and FANCI.10 In humans, UBE2T interacts with FANCL to ubiquitinate FANCD2. Patient-derived mutations in the UBE2T gene have recently been discovered in two unrelated patients, leading to a call to denote UBE2T as FANCT.12 The Delta-BLAST search returns Ciona ubc2-17kd as the closest match to human UBE2T. However, the reciprocal BLAST against human proteins returns human UBE2D4 with an E-value of 2 × 10⁻⁷⁷ (Table 1). The RSD method returns Ciona ubc2 J1l with an E-value of 9 × 10⁻⁷⁶. Apparently, these very similar E2 ubiquitin-conjugating enzymes cannot be reliably distinguished by BLAST searches (Fig. 3E). The hydropathy and phylogenetic analyses (Fig. 3D and F) do not help to resolve the exact relationship. In the hydropathy plot, it is apparent that both the Xenopus and Ciona proteins

Figure 2. Analysis of FANCE (A–F) and FANCL (G–K) putative homologs in C. intestinalis. (A) Hydrophathy plot of best aligning regions in human, Xenopus, and Ciona putative homologs for FANCE. (B) Best ML tree for alignment of FANCE and putative homologs in C. intestinalis and other eukaryotes. CIU1 (LOC100186252) has 93% bootstrap support for membership in the clade with vertebrate FANCE proteins. (C) Forcing CIU1 into the vertebrate FANCE clade does not result in a statistically worse tree, whereas if the locations of the two best C. intestinalis BLAST matches to FANCE are switched in the ML tree (D), the tree is worse at the P < 0.01 level, giving further support to LOC100186252 as the homolog of FANCE. (E,F) Structural modeling of human FANCE and C. intestinalis LOC100186252, showing extreme similarity of overall structures. (G) Hydrophathy plot of best aligning regions in human, Xenopus, and Ciona putative homologs for FANCL. (H) Best ML tree for alignment of putative FANCL homologs, showing 87% bootstrap support for Cinfcl clustering with vertebrate and other FANCL proteins. (I) Diagrammatic comparison of human and C. intestinalis FANCL inferred protein motifs. (J,K) Modeling of D. melanogaster and C. intestinalis FANCL protein structures.
roughly follow the pattern of human UBE2T, but neither closely matches with the hydropathy of the human protein. Curiously, in the ML phylogenetic analysis, the best tree shows human UBE2T clustering with another Ciona ube2 protein, Ciona ube2D3-like, but not the Ciona ube2–17, or ube2J1l proteins that are the best hits in the RBB and RSD analyses (Fig. 3F). If Ciona ube2J1l is grouped with human UBE2T, the tree is not statistically worse (Fig. 3G), but if Ciona ube2D3I is swapped with Ciona ube2–17, the tree does become significantly worse (Fig. 3H). In short, there are multiple ube2 proteins in Ciona that have such high similarity to the human UBE2T that they alternately appear as putative homologs in different analytic methods. We suggest that it is likely that one of these performs the same E2 ubiquitin conjugation function as UBE2T does in the human FA pathway.

Both group II genes have orthologs in Ciona. FANCD2. FANCD2 is one of the proteins monoubiquitinated by FANCL and FANC7/UBE2T during ICL repair.5,9,10 Both RBB and RSD returned a putative FA 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 \(1.7 \times 10^{-305}\), indicating extremely strong similarity (Table 1). The Ciona fancd2 protein contains 1394 aa, while the most common isoform in humans is 1451 aa long.

When the sequences are aligned and gaps removed, the smoothed hydropobicity plots show multiple similarities (Fig. 4A). 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. The phylogenetic analysis groups the C. intestinalis fancd2 candidate with vertebrate, fly, urchin, and amphioxus putative orthologs, although at low bootstrap support (Fig. 4B). Forcing the C. intestinalis candidate out of the FANCD2 clade makes the tree significantly worse at the \(P < 0.02\) level (Fig. 4C).

In addition, Clustal Omega MSA analyses of human, mouse, and Ciona FANCD2 revealed a strong conservation of the CUE ubiquitin-binding domain,5 the PCNA-interaction motif,6,4 and the site of FANCD2 monoubiquitination K561 (Fig. 4D and E).5 Furthermore, structural homology
modeling of *Ciona* fancd2, based on the 3.4 Å *Mus musculus* Fancd2-Fanci heterodimer structure (PDB ID 3S4W).\(^5\)\(^6\) Reveals a largely favorable structural similarity (Fig. 4F and G). Taken together, we consider that these data provide good support for the presence of a *C. intestinalis* fancd2 gene.

**FANCI.** Like FANCD2, FANCI is monoubiquitinated by FANCL and FANCT/UBE2T during ICL repair. Both RBB and RSD returned a *C. intestinalis* candidate fanci as the closest match to the human FANCI protein, with an E-value of 0. The hydrophobicity plots return an \(R^2\) value of 0.33, but several areas, notably a 150 amino acid stretch toward the carboxy-terminal end of the protein, have much higher correlations (Fig. 4H). Both proteins show multiple globular domains with moderate alignment and no recognizable secondary motifs. Clustal Omega MSA analyses of human, mouse, and *Ciona* FANCI indicate the conservation of K523 and K715, the sites of FANCI monoubiquitination and SUMOylation, respectively (Fig. 4J).\(^6\)\(^7\)\(^6\) In addition, *Ciona* fanci contains multiple conserved SQ/TQ ATM/ATR kinase phosphorylation motifs proximal to the putative
monoubiquitination site (Fig. 4K). In vertebrates, these sites have been demonstrated to be critical for FANCJ regulation and function.\textsuperscript{67,68} On the other hand, the ML phylogenetic analysis is inconclusive with respect to the orthology of the \textit{C. intestinalis} candidate and FANCJ. The best ML tree places the \textit{Ciona} candidate as a sister taxon to a clade of deuterostome plus cnidarian FANCJ proteins (Fig. 4D). However, forcing the \textit{Ciona} candidate into the vertebrate FANCJ clade results in a statistically worse tree, while forcing the \textit{Ciona} candidate to group with the next most similar \textit{Ciona} protein is not significantly different from the best ML tree (data not shown). In spite of the lack of support from the phylogenetic analysis, the sequence motif and structural data strongly suggest that \textit{Ciona} fanc is a true FANCJ ortholog.

**Seven group III orthologs were found.** \textit{FANCJ/BRIP1}. In humans, FANCJ is a 5′–3′ DNA helicase that interacts directly with BRCA1.\textsuperscript{69,70} RBB returns the ERCC2 nucleotide excision repair protein, but RSD returns human FANCJ. There is good alignment between the globular domains in human FANCJ and the \textit{Ciona} candidate, and the hydrophobicity plot shows high correlation (Fig. 5A). The human protein is of a similar size to the \textit{Ciona} protein, and they both possess a DEAH-box helicase domain (Fig. 5B). The ML tree groups \textit{C. intestinalis} fancj in the vertebrate FANCJ clade at 100% bootstrap support, and moving the \textit{C. intestinalis} candidate out of that clade makes the tree significantly worse (Fig. 5C and D). Given these data, the \textit{C. intestinalis} fancj candidate is a clear ortholog of human FANCJ.

**FAN1.** Fanconi-associatedRoute protein 1 is a DNA repair protein known to interact with monoubiquitinated FANCD2\textsuperscript{14} and FANCJ.\textsuperscript{71} The RBB returns a protein annotated as \textit{Ciona} fanc1, with an E-value of $4 \times 10^{-44}$. The fanc1 C-terminal region shows 41% identity and 63% positive matches. The human and \textit{Ciona} proteins align extremely well in the hydropathy plot (Fig. 5E) and both contain a 110 aa VRR nuclease domain (Fig. 5F). The ML tree clusters the \textit{C. intestinalis} candidate with the vertebrate FAN1 proteins (Fig. 5G) and is significantly worse when the \textit{C. intestinalis} protein is taken out of that clade (Fig. 5H; \textit{P} < 0.01). Taken together, the evidence is strongly in favor of \textit{Ciona} fanc1 being a homolog of FAN1.

**FANQ/ERCC4.** The \textit{FANQ} gene product, also known as ERCC4 or XPF, forms a heterodimer with ERCC1 and functions as a DNA repair structure-specific endonuclease.\textsuperscript{72} Both search methods return a \textit{Ciona} xpf as the most closely matching protein, with 50% identity, and 64% positive matches. The hydrophobicity plots show a high correlation, excepting one area corresponding to aa 390–430 in \textit{Ciona} and aa 520–560 in humans (Fig. 6A). Both proteins possess an ERCC4 endonuclease domain of the same size approximately the same distance from the carboxy-terminal end of the protein (Fig. 6C). The ML analysis clusters the \textit{C. intestinalis} xpf in the FANCO clade (Fig. 6B), although moving the \textit{C. intestinalis} protein out of that clade does not make the tree likelihood significantly worse (data not shown). Taken together, we conclude that \textit{C. intestinalis} does have a FANCO ortholog.

**ERCC1.** ERCC1 interacts directly with FANQ/ERCC4. The \textit{Ciona} candidate returned by RBB (XP_009861832) has an extremely similar hydropathy plot as the human and frog ERCC1 proteins, except at the N-terminal-most 50 residues (Fig. 6D), although the \textit{Ciona} candidate appears to lack an intact HhH1 domain present in the human protein (Fig. 6E). The ML analysis groups the \textit{Ciona} candidate within the vertebrate ERCC1 clade (Fig. 6F). Moving the \textit{Ciona} protein outside that clade or grouping it with the next most similar human gene (FAAP24) makes the trees statistically worse at the \textit{P} < 0.02 level (Fig. 6G and H). These data strongly support the orthology of the \textit{Ciona} candidate.

**FANCO/RADS1C.** RADS1C is also required for the maintenance of chromosome stability by functioning in HR repair.\textsuperscript{73} \textit{Ciona} has five potential RADS1 family homologs if the proteins listed as lim15 and xrc2 are included. RBB finds \textit{Ciona} rad51 (XP_002126934) as the closest match to human FANCO. However, if the \textit{Ciona} protein identified as rad51C in GenBank (XP_002130341) is used in the ML analysis with FANCO, \textit{Ciona} rad51C robustly groups with FANCO to the exclusion of \textit{Ciona} rad51 (Fig. 7C). Forcing \textit{Ciona} rad51C out of the FANCO clade results in a statistically worse tree (Fig. 7D, \textit{P} < 0.01). Structurally, the \textit{Ciona} rad51C is more similar to FANCO than the higher BLAST match \textit{Ciona} rad51 (Fig. 7A and B). Based on these analyses, we conclude that \textit{Ciona} does have a FANCO homolog.

**FANCR/RADS1.** In humans, RAD51, recently gaining the name FANCJ, is the major DNA strand exchange protein and is critical for the HR DNA repair process.\textsuperscript{74,75} De novo heterozygous RADS1 mutations have recently been reported in two unrelated individuals with an FA-like syndrome.\textsuperscript{76} RADS1 is known to interact with both FANCS/BRCA1 and FANCD1/BRCA2 in the cellular DNA damage response.\textsuperscript{77} Both search methods return a \textit{Ciona} rad51 as the most likely candidate. The BLASTP matches are 41% identity between human and \textit{Ciona} as well as a 92% level of positive matches, far outstripping any other gene product tested. The \textit{Ciona} product is 338 aa in length, while the human product is 339 aa (Fig. 8B). Both \textit{Ciona} rad51 and the \textit{Ciona} rad51 candidate 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 (Fig. 8A). The ML analysis shows the \textit{Ciona} rad51 candidate grouping with other deuterostome RADS1 proteins (Fig. 8C), while excluding \textit{Ciona} rad51 from that clade results in a statistically worse tree (Fig. 8D, \textit{P} < 0.01). It is highly likely that \textit{Ciona} rad51 is a true ortholog of human FANCR/RADS1.

**FANCS/BRCA1.** The \textit{C. intestinalis} candidate for FANCS by RBB has two BRCT (BRCA1 C-terminal domain) domains at the C-terminus, similar to BRCA1 (Fig. 8F). BRCT domains typically mediate interactions with phosphopeptides. The
The evolutionary bioinformatic analysis of FANJ and FAN1 proteins in C. intestinalis reveals several key points:

1. **Hydropathy Plot**: The hydropathy plot of the C-terminal 500 residues of the C. intestinalis protein, human, and frog proteins show a good degree of similarity. However, the analysis does not predict the C. intestinalis protein from the ANISEED database.

2. **ML Tree**: The ML tree analysis results in a closer relationship to the vertebrate FANCS/BRCA1. Rearranging the trees so that the best BLAST match is moved out of the FANCD1 clade altogether or switching the first and second most similar sequence results in a statistically worse topology.

3. **Protein Homology**: Our analyses found 11 FA or FA-associated proteins in C. intestinalis, which we did not find homologs for in other eukaryotes. Cifancj has 100% bootstrap support for membership in the FANCA, FANCB, FANCC, FANCF, FANCG, FANCD1/BRCA2, FANCN/PALB2, and FANCP/SLX4. We also failed to find the FA-associated proteins FAAP20, FAAP24, and FAAP100.

4. **Prosite Scan**: A Prosite scan indicates that LOC100185089 has two BRC repeats, which may explain the presence of this protein in the tree. However, it may still be the case that this protein in combination with one or more others is fulfilling the function served in humans by BRCA1.

**FA and FA-associated proteins not found in Ciona**: Our analyses found 11 FA or FA-associated proteins present in vertebrates but not in Ciona. These results were based on the four major criteria outlined for each of the predicted FA homologs, as outlined above, namely, BLAST search, structural motif similarity, hydropathy, and phylogenetic (ML) analysis. The FA proteins that we did not find homologs for in Ciona were as follows: FANCA, FANCB, FANCC, FANCF, FANCG, FANCD1/BRCA2, FANCN/PALB2, and FANCP/SLX4. We also failed to find the FA-associated proteins FAAP20, FAAP24, and FAAP100.

For 10 of the 11 cases, RBB and RSD failed to match a Ciona protein sequence with an FA-related protein (Table 1). The exception is FANCD1/BRCA2, for which one match comes up in RBB as an uncharacterized protein. LOC100185089 (Table 1). However, the ML analysis results in another C. intestinalis protein showing a closer relationship to FANCD1/BRCA2. Rearranging the trees so that the best BLAST match is moved out of the FANCD1 clade altogether, or switching the first and second most similar C. intestinalis proteins in the tree, does not result in statistically worse trees, indicating that the evidence for homology of the C. intestinalis proteins is weak (data not shown).

In addition, the hydropathy analysis shows a low correlation ($R^2 = 0.117$, Table 2). A Prosite scan indicates that LOC100185089 has two BRC repeats, which may explain why it comes up in the BLAST search. However, FANCD1 is a much larger protein (3418 aa vs. 724 aa) and has eight BRC repeats. These BRC repeats represent the major sites of interaction between RADS1 and BRCA2. In addition, BRCA2/FANCD1 has an α-helical region, an oligonucleotide/oligosaccharide-binding domain, a TOWER domain, and a second oligonucleotide/oligosaccharide-binding domain.
C. intestinalis LOCl00185089 possesses two BRC repeats only. None of these other domains are present.

There is a possibility that the predicted Ciona protein in the NCBI database is not the full-length sequence. However, we searched a 22 kb region in the Ciona genome, which includes LOC100185089 and flanking regions. No significant similarity to the human sequence outside the region that aligns with LOC100185089 was found, even when the protein sequence not included in LOC100185089 was blasted against the translated Ciona genomic sequence. Thus, we infer that Ciona does not have a complete ortholog of human BRCA2.

For the other 10 of the 11 cases of unlikely homology, the hydropathy $R^2$ statistics are lower numbers than those for the putative homologs, ranging from 0.034 to 0.177 vs. 0.291 to 0.566, respectively (Table 2). Similarly, we did not find good evidence for homology to any C. intestinalis proteins by any of the other three analytical methods used (Table 1, and data not shown). Therefore, we conclude that these 11 FA and FA-associated proteins are missing from C. intestinalis.

Discussion
In this study, we have established that the model marine invertebrate, C. intestinalis, appears to contain all of the necessary functional components to reconstitute a simplified FA pathway (Fig. 9). Of the FA core complex group I proteins, we identified orthologs of FANCL, FANCT/UBE2T, and FANCM, and possibly FANC. FANCL and FANCT/UBE2T are the E3 ubiquitin ligase and E2 ubiquitin conjugase enzymes, respectively, that monoubiquitinate FANCD2 and FANCI. While FANCD2 and FANCI monoubiquitination are largely defective in FA patient cells with mutations in any of the core complex genes (FANCA, B, C, E, F, G, L, and T), several studies have established that FANCL and FANCT/UBE2T, in the presence of an E1 ubiquitin-activating enzyme and DNA, can readily promote FANCD2 and FANCI monoubiquitination in vitro. The roles of the other FA core complex proteins in promoting FANCD2 and FANCI monoubiquitination in vivo remain unknown. The functions provided by these other core complex proteins may be unnecessary in C. intestinalis, or may be provided by other proteins. Interestingly, previous studies have established that the FANCE protein directly interacts with FANCD2, thereby bridging the core ubiquitin ligase machinery and the substrate. C. intestinalis fance may fulfill an analogous function. Similar to human FANCM, C. intestinalis fancm contains an N-terminal DEAH domain-containing Walker A and B motifs typical of an SF2 family
translocate. These proteins are capable of movement along DNA in the absence of helicase activity. FANCM translocate activity is necessary for replication fork stability and ATR-CHK1 checkpoint signaling. The C-terminus of human FANCM contains a degenerate ERCC4 endonuclease domain, which is also the site of binding of its heterodimeric partner FAAP24; yet, this region appears absent in C. intestinalis fancm (Fig. 3B). Since C. intestinalis appears to lack a FAAP24 homolog, it is not surprising that Cifancm lacks the binding site. It has been speculated that the FANCM-FAAP24 heterodimer plays an important DNA-targeting function, and why the formation of a heterodimer might be unnecessary in C. intestinalis is unclear. However, the categorization of FANCM as a true FA gene remains controversial.

The evidence for structural and functional conservation of the FANCD2 and FANCI proteins appears quite strong, with several protein domains and important sites of posttranslational modification being highly conserved (Fig. 4D, E, J, and K). This is consistent with the previous finding indicating considerable depth in their conservation in all eukaryotes. The monoubiquitination of these proteins is a critical step in the activation of the FA pathway and in ICL repair. In the case of FANCD2, monoubiquitination of K561 has been implicated in the recruitment of the FAN1 and FANCP/SLX4 proteins, which participate in, or facilitate, several key nucle-
In the case of Alzheimer’s, transgenic C. intestinalis were produced expressing the human APP gene mutant associated with familial AD. The transgenic protein resulted in the formation of amyloid-β plaques in less than 24 hours in the rapidly developing C. intestinalis larval brain. This result contrasts with a 2–8-month time period for plaques to form in mouse AD models. For FA, study of the pathway in invertebrate model organisms has proven valuable in several cases. For example, the function of FANCJ in maintaining poly(G)/poly(C) tract stability during DNA replication was first shown in the nematode worm C. elegans. It was later demonstrated that human FANCJ has the same helicase function.

It is important to note that of all the three major constellations of FA patient phenotypes, namely, developmental defects, bone marrow failure, and increased cancer risk, the molecular bases of the developmental defects are the most poorly understood. A C. intestinalis model for FA could provide unique insights into these defects. Temporospatial aspects of FA gene expression and developmental consequences of disruption of FA genes using CRISPR/Cas9 or TALEN systems could be highly informative for FA patient developmental defects. Furthermore, another unique benefit to exploring a C. intestinalis model for FA is the prospect of discovering the physiological function(s) of this pathway. While it is well established that FA patient cells

**Figure 9.** (A) A model of the FA pathway in humans. Following exposure to DNA damaging agents or during S-phase of the cell cycle, the FA core complex (group I) proteins catalyze the monoubiquitination of the FANCD2 and FANCI (group II) proteins. Following their monoubiquitination, FANCD2 and FANCI function together with the downstream FA (group III) proteins to repair damaged DNA. Modified from Cybulski and Howlett, 2011. (B) A model of a hypothetical simplified FA pathway in C. intestinalis based on the reduced complement of FA gene homologs found by this study. C. intestinalis possesses the critical E3 ubiquitin ligase (Fanci) and E2 ubiquitin-conjugating enzyme (Fanct) to monoubiquitinate Fancd2 and Fanci, as well as a minimal set of FA group III effector proteins. Proteins shown in gray have lower support for existence in C. intestinalis.

Recombination in metabolism, and DNA repair. Conservation of the FANCD2 K561 and FANCI K523 monoubiquitination sites, as well as several other important sites of posttranslational modification, strongly suggests that this central step is intact in C. intestinalis.

Of the group III proteins, the evidence points to the existence of C. intestinalis orthologs of FANCJ/BRIP1, FANCO/ERCC4, FANCR/RAD51, FANCO/RAD51C, and FANCS/BRCA1. The heterodimeric binding partner of FANCO/ERCC4, ERCC1, is also present, as is the FANCD2-associated nuclease FAN1. Conservation of FANCR/RAD51 and FANCS/BRCA1 is not surprising, given their key roles in multiple cellular processes, including meiotic and mitotic recombination. Targeted disruption of either gene results in early embryonic lethality in mice. However, the absence of FANCD1/BRCA2 is particularly surprising, given its strong conservation among eukaryotes. FANCD1/BRCA2 plays a critical role in regulating FANCR/RAD51 nucleoprotein filament formation and DNA strand exchange. It is also intriguing that C. intestinalis apparently lacks FANCN/PALB2. FANCN/PALB2 interacts directly with FANCD1/BRCA2 and promotes its chromatin localization. Studies of the Ustilago maydis homolog of BRCA2 indicate that BRCA2 promotes RAD51 nucleation at junctions of single-stranded and double-stranded DNA. However, lower eukaryotes such as Saccharomyces cerevisiae and Schizosaccharomyces pombe lack homologs of both FANCS/BRCA1 and FANCD1/BRCA2, indicating that the functions provided by these proteins are unnecessary in certain organisms or may be provided by other proteins.

There is considerable precedent suggestive of the efficacy of studying the FA pathway in C. intestinalis. Study of several human diseases have benefited from the use of invertebrate model organisms. In particular, the genetically tractable invertebrates, such as Drosophila and Caenorhabditis elegans, have been used extensively. Notably, it has recently been shown that even very simple animals, such as sponges and sea anemones, have homologs of many human disease genes. C. intestinalis has only recently emerged as a model system. However, it has already been used to study certain human disease pathways, such as Huntington’s Disease and Alzheimer’s disease (AD). In the case of Alzheimer’s, transgenic C. intestinalis were produced expressing the human APP gene mutant associated with familial AD. The transgenic protein resulted in the formation of amyloid-β plaques in less than 24 hours in the rapidly developing C. intestinalis larval brain. This result contrasts with a 2–8-month time period for plaques to form in mouse AD models. For FA, study of the pathway in invertebrate model organisms has proven valuable in several cases. For example, the function of FANCJ in maintaining poly(G)/poly(C) tract stability during DNA replication was first shown in the nematode worm C. elegans. It was later demonstrated that human FANCJ has the same helicase function.
are hypersensitive to ICL-inducing agents, the relevance of ICLs in the physiological setting is unclear. Recent studies have established an important role for the FA proteins in mitigating endogenously arising aldehyde-mediated DNA damage.\(^{107-109}\) Exploring the pathway in other model systems may lead to a broader understanding of the true function(s) of these key proteins. *C. intestinalis*, as a tunicate, is in the most closely related invertebrate group to the vertebrates.\(^{27,110}\) As such, in spite of being anatomically simpler than a vertebrate, they are genetically more similar than other eukaryotes. However, it is possible that *C. intestinalis* may deploy its FA homologs differently than they function in humans. If this is the case, it may still be relevant to understanding human disease, as it will point to alternative ways of dealing with DNA lesions and may provide information on some of the other defects seen in FA patients.

In summary, our study provides compelling evidence for the existence of a simplified and potentially functional FA pathway in the model chordate *C. intestinalis*. *C. intestinalis* is an excellent model for the study of developmental processes because it is anatomically simple, its gametogenesis and development are well studied, it has a small and well-annotated genome and abundant gene expression data, and good transgenic technology exists.\(^{39,101-111}\) Future studies will seek to determine the patterns and timing of FA gene expression in *C. intestinalis* and the developmental impacts of disruption of the pathway.

**Author Contributions**

Analyzed the data: ECS, PAA, DAV, NGH, SQI. Wrote the first draft of the manuscript: ECS. Contributed to the writing of the manuscript: NGH, SQI. Agreed with manuscript results and conclusions: ECS, PAA, DAV, NGH!SQI. Jointly developed the structure and arguments for the paper: ECS, NGH, SQI. Made critical revisions and approved the final version: ECS, NGH, SQI. All the authors reviewed and approved the final manuscript.

**Supplementary Material**

**Supplemental Table S1.** Accession Numbers for sequences used in analyses.

**REFERENCES**


Stanley et al


