RESEARCH ARTICLE

UPDATED A fragmented alignment method detects a putative phosphorylation site and a putative BRC repeat in the *Drosophila melanogaster* BRCA2 protein [v2; ref status: indexed, http://f1000r.es/1wc]

Previously titled: A fragmented alignment method detects a phosphorylation site and a new BRC repeat in the *Drosophila melanogaster* BRCA2 protein, and a new HAT repeat in Utp6 from yeast

Sandeep Chakraborty

Department of Biological Sciences, Tata Institute of Fundamental Research, Mumbai, 400 005, India

v2 **First Published:** 25 Jun 2013, **2**:143 (doi: 10.12688/f1000research.2-143.v1)
Latest Published: 07 Oct 2013, **2**:143 (doi: 10.12688/f1000research.2-143.v2)

Abstract

Mutations in the BRCA2 tumor suppressor protein leave individuals susceptible to breast, ovarian and other cancers. The BRCA2 protein is a critical component of the DNA repair pathways in eukaryotes, and also plays an integral role in fostering genomic variability through meiotic recombination. Although present in many eukaryotes, as a whole the *BRCA2* gene is weakly conserved. Conserved fragments of 30 amino acids (BRC repeats), which mediate interactions with the recombinase RAD51, helped detect orthologs of this protein in other organisms. The carboxy-terminal of the human BRCA2 has been shown to be phosphorylated by checkpoint kinases (Chk1/Chk2) at T3387, which regulate the sequestration of RAD51 on DNA damage. However, apart from three BRC repeats, the *Drosophila melanogaster* gene has not been annotated and associated with other functionally relevant sequence fragments in human BRCA2. In the current work, the carboxy-terminal phosphorylation threonine site (E=9.1e-4) and a new BRC repeat (E=17e-4) in *D. melanogaster* has been identified, using a fragmented alignment methodology (FRAGAL). In a similar study, FRAGAL has also identified a novel half-a- tetratricopeptide (HAT) motif (E=11e-4), a helical repeat motif implicated in various aspects of RNA metabolism, in Utp6 from yeast. The characteristic three aromatic residues with conserved spacing are observed in this new HAT repeat, further strengthening my claim. The reference and target sequences are sliced into overlapping fragments of equal parameterized lengths. All pairs of fragments in the reference and target proteins are aligned, and the gap penalties are adjusted to discourage gaps in the middle of the alignment. The results of the best matches are sorted based on differing criteria to aid the detection of known and putative sequences. The source code for FRAGAL results on these sequences is available at https://github.com/sanchak/FragalCode, while the database can be accessed at www.sanchak.com/fragal.html.

Article Status Summary

Referee Responses

Referees	1	2	3
v1 published 25 Jun 2013	✓ report 2	? report 2	✓ report 1
v2 published 07 Oct 2013 UPDATED		✓	✓ report

1 **Himanshu Sinha**, Tata Institute of Fundamental Research India

2 **Satish Chikkagoudar**, Pacific Northwest National Laboratory USA

3 **Saurabh Sinha**, University of Illinois at Urbana-Champaign USA

Latest Comments

No Comments Yet

Corresponding author: Sandeep Chakraborty (sanchak@gmail.com)

How to cite this article: Chakraborty S (2013) A fragmented alignment method detects a putative phosphorylation site and a putative BRC repeat in the *Drosophila melanogaster* BRCA2 protein [v2; ref status: indexed, http://f1000r.es/1wc] *F1000Research* 2013, **2**:143 (doi: 10.12688/f1000research.2-143.v2)

Grant information: This work was funded by the Tata Institute of Fundamental Research (Department of Atomic Energy), and the Department of Science and Technology (JC Bose Award Grant). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: No competing interests were disclosed.

First Published: 25 Jun 2013, **2**:143 (doi: 10.12688/f1000research.2-143.v1)
First Indexed: 06 Sep 2013, **2**:143 (doi: 10.12688/f1000research.2-143.v1)

UPDATED **Changes from Version 1**

I would like to thank the referees for the insightful observations on my work. I have revised my manuscript in which I have addressed all the specific points raised, and believe that these have improved the manuscript significantly. The main changes incorporated in this version are summarized below:

1. Shortened the title - A fragmented alignment method detects a putative phosphorylation site and a putative BRC repeat in the *Drosophila melanogaster* BRCA2 protein.

2. Added pseudo code of FRAGAL to the methods section.

3. Applied FRAGAL to two new repeats - BIR and TPR. However, I could not detect any new repeat based on FRAGAL results.

4. Added proteins from *Aspergillus nidulans* and *Candida glabrata* which have the HAT repeat in the FRAGAL processing.

5. Used another multiple sequence alignment (MAFFT) to corroborate results obtained from Clustal-W.

6. Table 1 has been simplified based on the suggestion of the referees.

7. Specified the tool used for computing E-values.

Please find my detailed responses underneath each referee report.

See referee reports

Introduction

The breast cancer susceptibility protein BRCA2, first identified in 1995[1], is a critical recombinase regulator[2] that ensures genomic stability through high fidelity repair[3,4] of double stranded breaks (DSB) and prevents stalled replication forks from replicating[5] in the DNA. The primary recombinase in BRCA2 repair of DSB through homologous recombination is the RAD51 protein, belonging to the conserved RecA/RAD51 family[6], that binds to the BRCA2 protein at various segments of ~30 amino acids (BRC repeats)[7,8], and in the C-terminal region in most vertebrates[9,10]. Checkpoint kinases phosphorylate a serine[9] and a threonine[10] at the carboxy-terminal region of BRCA2, thereby regulating its interaction with RAD51. BRCA2 also plays a key role in fostering genomic variability through meiotic recombination[11,12], although a different recombinase (DMC1) is implicated in this pathway in mammalian species[13].

The BRC repeats have helped identify BRCA2 orthologs in various eukaryotic species[14]. Functional characterization of this gene in *Drosophila melanogaster* has demonstrated its interaction with RAD51, and a critical role in mitotic and meiotic DNA repair as well as homologous recombination[11,15]. The copy number of the BRC repeats differs considerably. The BRCA2 homolog in *Ustilago maydis* (a yeast like fungus) has a single BRC repeat[16], the *D. melanogaster* homolog contains only three (known) repeats[14], while there are eight repeats in the human BRCA2 gene[7]. Even among the *Drosophila* genus, the range of BRC repeat numbers is varied - the *D. melanogaster* species has only three repeats, while *D. persimilis* and *D. pseudoobscura* have up to eleven repeats[17]. RAD51 shows varying affinity for the different BRC motifs[18,19]. This difference in repeat numbers in *Drosophila* has raised doubts whether 'this higher repeat number is real or a genome mis-assembly artifact'[20], and also led to speculation on the evolution of these closely related organisms[17,20]. Any such hypothesis would need to be revisited if a new BRC motif were to be identified in *D. melanogaster*.

In the current work, the putative threonine phosphorylation site for checkpoint kinases (Chk1/Chk2) (E=9.1e-4) and a new BRC repeat (E=17e-4) in *D. melanogaster* has been identified, using a fragmented technique for the pairwise alignment of two sequences (FRAGAL). The reference and target sequences are sliced into fragments of equal parameterized length X, sliding along the sequence in intervals of length Y, such that Y is less than X. Thus, the slices have overlaps. An alignment of all pairs of slices in the reference and target proteins is done using the global alignment program 'needle'[21] from the EMBOSS suite[22]. The gap penalties are adjusted to discourage gaps in the middle of the alignment. The results of the best matches are sorted based on differing criteria to aid the detection of known and putative sequences. In order to establish the generic nature of the FRAGAL methodology, the detection of a new half-a-tetratricopeptide (HAT) repeat sequence (E=11e-4) in a nucleolar RNA-associated protein (Utp6) from *Saccharomyces cerevisiae* is also reported. HAT is a helical repeat motif implicated in various aspects of RNA metabolism[23,24]. The characteristic three aromatic residues with a conserved spacing are observed in this new HAT repeat, further strengthening my claim[25].

Existing methods for detecting functional motifs in a given protein sequence have been unable to detect these putative sites. For example, meta servers (http://myhits.isb-sib.ch/cgi-bin/motif_scan, http://www.ebi.ac.uk/Tools/pfa/iprscan/, http://www.genome.jp/tools/motif/) for detecting motifs in a protein have been unable to detect the sites identified using the FRAGAL methodology. These meta servers use one or more motif databases[26–30]. Not all known BRC repeats have a low E-value when aligned with the new BRC repeat. For example, the first BRC repeat in hBRCA2 when aligned to the new dmBRCA2 repeat has an E=0.04, much more than the E=17e-4 observed for the fourth repeat, which is the one I report here. Ideally, if one took all the BRC repeats and did a search in the dmBRCA2 sequence, this new repeat would be reported. Essentially, this is what FRAGAL does - albeit implicitly, by automatically fragmenting the sequence. The same logic applies to the HAT repeat, where the sequences are more varied and thus the choice of the repeat would effect the detection of new motifs.

Spliced alignment techniques have frequently been adopted in the precise identification of eukaryotic gene structures, and in gene assembly. These methods try to solve the exon assembly problem by searching the exon sequence space to find the best fit to known proteins[31,32]. While these methods use graph algorithms to solve the computationally difficult problem of exon chaining, FRAGAL does the converse of finding best matches in known exon chains (i.e. protein sequences).

It is fair to mention that the FRAGAL method is much more computationally intensive than the above mentioned methods. At the same time, FRAGAL makes no assumption of any knowledge of the conserved regions (either the sequence or their position). The choice of the fragment length in FRAGAL depends on the length of repeats that is expected to be present in the protein. Since both repeats (BRC and HAT) discussed in this manuscript are around ~30 amino acid long, I have chosen a fragment length of 50. A larger fragment length might mask the similarity in the core region due to variations in the non-critical regions, whereas a smaller fragment would match irrelevant portions and thus increase false positives.

The significant conservation of the DNA repair and checkpoint pathways in flies and higher organisms[33], the advanced genetic tools available for *Drosophila*, and the viability of the *Drosophila* BRCA2 null mutants in contrast to mammalian mutants[34] establishes *Drosophila* as a model organism for studying these pathways[35]. Significant divergence of key conserved sequences proves to be a serious hurdle for alignment techniques to annotate and associate the conserved sequences in the human BRCA2 to the *Drosophila* BRCA2[36]. Thus, a generic methodology, applicable to distantly evolutionary related proteins like BRCA2 and nucleolar RNA-associated proteins is presented. The methodology has been validated by the identification of two novel functionally relevant sites in the BRCA2 protein from *D. melanogaster*, and a HAT repeat in Utp6 from *S. cerevisiae*.

Materials and methods

The FRAGAL methodology is shown in Supplementary Figure 1.

The sequences are split into fragments of X amino acids, with the starting indices sliding across the sequence length in steps of Y amino acids (SI.A.fasta and SI.B.fasta in Data Files). The score for each match is computed as shown in Equation 1. FRscore is intended to give more weightage to identical residues in the alignment.

$$FRscore = 1/3 * \%onlySimilarity + 2/3 * \%identity; \quad (1)$$

One sorting criteria is to rank the matches based on the best average score, while another takes the cumulative score of a stretch of fragment matches. Stretches of fragments are stitched while ensuring the slices in the sequences are in an increasing order and non-overlapping. The best average criteria will typically select single fragments, while the cumulative scoring criteria will bring forth longer conserved regions.

Algorithm 1: FRAGAL() - A fragmented alignment method

Input: *A*: Query sequence
Input: *B*: Target sequence
Input: *X*: Length of fragments
Input: *Y*: Length of sliding window
Input: *gapopen*: Length of fragments
Input: *gapextend*: Length of fragments
Output: $\phi_{fragments}$: Matching fragments sorted in terms of higher FRscores
begin
  ϕ_{Afrag} = FragementSequenceIntoOverlappingSegments(*A, X, Y*);
  ϕ_{Bfrag} = FragementSequenceIntoOverlappingSegments(*B, X, Y*);
  // Create priority queue, based on FRscore
  $\phi_{fragments} = \varnothing$;
  for *each fragemnt* A_i *in* ϕ_{Afrag} **do**
    for *each fragemnt* B_j *in* ϕ_{Bfrag} **do**
      // See methods section for FRscore
      $FRscore_{ij}$ = RunNeedle(A_i,B_j,*gapopen*, *gapextend*);
      Insert($\phi_{fragments}$,$FRscore_{ij}$);
    end
  end
  return ($\phi_{fragments}$);
end

The threshold for sequence similarity for each fragment is parameterized, and set to 30% in the default mode. A large threshold will exclude more relevant matches, while a smaller threshold might include more false positives. The pairwise alignment for each fragment pair is done by a global alignment program 'needle' from the EMBOSS suite[21,22]. The parameters are set as follows - matrix=BLOSUM62, Gap penalty=25.0 and Extend penalty=0.5. The gap penalty is increased from the default value of 10 to ensure that gaps are discouraged in the middle of the alignment. Single deletions or insertions are rarely expected in conserved fragments. However, once 'needle' has aligned the sequences based on the this penalty, gaps should not have a penalty. It is for this very reason that I have introduced the FRscore as a metric to measure quality of alignment, which creates a weighted score of the %identity and %similarity (Equation 1).

The user is allowed to specify an annotation file for a given protein sequence using the uniprot accession syntax (Supplementary Figure 2). The results from FRAGAL can be filtered based on this annotation, and this provides a easier way to manually inspect and annotate corresponding segments in a query protein sequence.

The FRAGAL package is written in Perl on Ubuntu. Hardware requirements are modest - all results here are from a simple workstation (2GB RAM). The source code for FRAGAL results on these sequences is available at https://github.com/sanchak/Fragal-Code, while the database can be accessed at www.sanchak.com/fragal.html. The multiple sequence alignment was done using ClustalW[37]. PHYML has been used to generate phylogenetic trees from these alignments, which is based on the method of maximum likelihood[38]. The method searches for a tree with the highest probability or likelihood that, given a proposed model of evolution and the hypothesized history, would give rise to the observed data set. The alignment and cladograms images were generated using Seaview[39]. E-values and z-scores have been computed using the Protein Information Resource (http://pir.georgetown.edu/pirwww/search/pairwise.shtml)[40].

BRCA2 sequence fragments and database of the output of FRAGAL for BRCA2 and HAT repeats for different organisms

3 Data Files

http://dx.doi.org/10.6084/m9.figshare.812563

Results

Breast cancer susceptibility protein BRCA2

The *D. melanogaster* gene (CG30169)[41] encodes a 971 amino acid protein (dmBRCA2, Uniprot Accession:Q9W157), and contains three BRC repeat units (conserved sequences of ~30 amino acids that binds to RAD51)[8,14]. In contrast, the human BRCA2 gene product (hBRCA2, Uniprot Accession:P51587) is 3418 amino acids long and contains eight BRC repeats[7]. Further, the hBRCA2 protein is annotated for several sites phosphorylated by checkpoint kinases,

which regulate its interaction with RAD51[9,10]. FRAGAL was run on the dmBRCA2 and hBRCA2 sequences. Table 1 shows the best matches obtained using two different sorting criteria - best average FRscore (see Methods) and best average %similarity - either when the match in hbrca2 is known to be conserved (Table 1A) based on an user defined input file (Supplementary Figure 2) or otherwise (Table 1B).

Detecting the threonine phosphorylation site in the carboxy-terminal region of dmBRCA2. Table 1 shows a significant match (E=9.1e-4, Z-score=100) between fragment 91 of dmBRCA2 to the fragment 337 in hBRCA2, which contains the T3387 that is phosphorylated by the checkpoint kinases Chk1 and Chk2. Z-scores above a value of 8 are considered to be significant[42]. The alignment shows that the T3387 corresponds to the T926 of dmBRCA2 (Figure 1a). The conservation of this region in the *Drosophila* and mammalian species is demonstrated by the multiple sequence alignment of three organisms from each species (Figure 1b). The highly conserved columns in the alignment are highlighted using an asterisk, and can be used to define a Prosite motif ([ST]-E-[ST][ST]-x-[ST]-x(6)-[ED]-x(4)-K-x(4)-[ST]-[ST]-[ST]-x(3)-[DE]-[DE])[27]. Either this motif or FRAGAL alignments failed to detect this site in other species distant from *Drosophila* or mammals (*Ustilago maydis* and *Caenorhabditis elegans*).

Detecting an additional BRC repeat in Drosophila melanogaster. The correct identification of the three BRC repeats in *D. melanogaster* is seen by the significant scores of the FRscore matches of A67-B152 (64), A57-B100 (60) and A75-B152 (60) (Table 1). A significant alignment (E=17e-4, Z-score=95) between A61-B151 (35.8%similarity and 17%identity) (Figure 1c) was also observed. This sequence (634–664:LDTALKRSIESSEEMRSKASKLVVVDTTMR) is now added to the list of sequences previously studied in the *Drosophila* genus[17]. The multiple sequence alignment (obtained using ClustalW[37]) (Figure 2a) and phylogenetic trees (obtained using PHYML[38]) (Figure 2b) shows that this new BRC repeat is more related to *D. willistoni* than other organisms in the *Drosophila* genus. A detailed molecular phylogeny of Drosophilid species has noted that the subgenus *Sophophora* is 'divided into *D. willistoni* and the clade of *D. obscura* and *D. melanogaster* groups', possibly indicating the source of this BRC repeat that has been conserved between *D. willistoni* and *D. melanogaster*[43]. The same inference is drawn when we use a different multiple alignment tool like MAFFT (http://mafft.cbrc.jp/alignment/software/)[44] (Supplementary Figure 3). An iterative methodology, similar to PSI-BLAST (Position-Specific Iterative Basic Local Alignment Search Tool)[45], can be automated to generate comprehensive motifs spanning distant species. The conservation of many key residues in this sequence fragment, as shown by comparing it to the sequence logo of the Prosite BRCA2 profile (PS50138) (Figure 2c) strongly suggests that this is a putative BRC repeat. However, it must be emphasized that such repeats are to be considered putative until verified experimentally[46,47].

Table 1. FRAGAL results for aligning the BRCA2 protein sequences from *Drosophila melanogaster* (A) (Accession:Q9W157) and humans (B) (Accession:P51587): The results are filtered out if the fragment in the hBRCA2 sequence is not marked as conserved (Supplementary Figure 2). This filtering helps in removing already annotated sequences, thus making it easier to observe new sequences. Thus, there are some missing ranks. Multiply index with 10 to get sequence starting position in original sequence. A91 refers to the sequence starting at 910 in A and going till 959, since the fragmenting length is 50. The A91-B337 match corresponds to the phosphorylation site of checkpoint kinases in the carboxy-terminal of BRCA2, while the match A61-B151 (not shown in Table, FRscore=53), corresponds to the new BRC repeat identified in *D. melanogaster*.

Rank	FRscore	Matches
1	73.5	A91-B337
3	64.7	A87-B197
4	64.2	A67-B152
5	64	A87-B194
6	64	A65-B140
9	61.7	A57-B167,A87-B197
10	61.4	A67-B100
11	61.4	A69-B199,A74-B204
12	61.3	A89-B335
14	60.1	A75-B152

Half-a-tetratricopeptide (HAT) motif

HAT is a helical repeat motif implicated in various aspects of RNA metabolism and in protein-protein interactions[23,24]. These repeats are characterized by three aromatic residues with a conserved spacing[25]. A variable number of HAT repeats (9 to 12) are found in different proteins. Figure 3a shows a novel HAT repeat (E=11e-4, Z-score=116) detected in a nucleolar RNA-associated protein (Utp6) from *Saccharomyces cerevisiae* (Uniprot Accession:Q02354) by comparing it to HAT repeats from a human nucleolar RNA-associated protein (Uniprot Accession:Q9NYH9). Q9NYH9 has five annotated HAT repeats (121–153, 156–188, 304–335, 488–520 and 524–557), while Q02354 has three HAT repeats (87–119, 124–156 and 159–191). The new HAT sequence identified in Q02354 (SLIMKKRTDFEHRLNSRGSSINDYIKYINYESN) is from position 30 to 62. It can be seen from the multiple sequence alignment that this sequence has the desired aromatic residues at the proper spacing, a requisite for being considered a HAT repeat (Figure 3a and b). Further, the MSA shows large variation amongst HAT sequences even within the same organism (Figure 3b). Finally, Figure 3b and c shows that certain HAT repeats are more similar to HAT repeats from other organisms than to other HAT repeats in its own sequence. Supplementary Figure 4 shows the alignment and phylogenetic tree when we include more proteins having HAT repeats from organisms closely related to *S. cerevisiae* like *Aspergillus nidulans* and *Candida glabrata*, corroborating the large variation among repeats even within the same organism and that often HAT repeats across organisms show more similarity.

Database for aligning different pairs of BRCA2

A database (www.sanchak.com/fragal.html) which lists the results for the fragmented alignment of various proteins with BRC and HAT repeats sequences has been created. The results have been

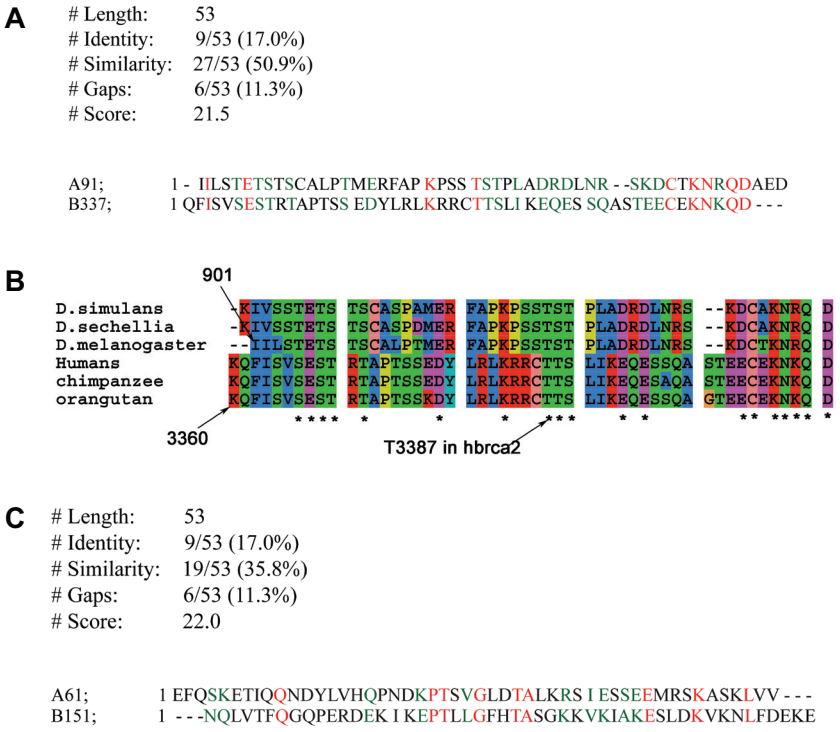

Figure 1. Fragment alignment using 'needle' from the EMBOSS suite[21,22] of previously unknown conserved, and functionally relevant, sequences in dmBRCA2. (red for identity, green for similarity). (**a**) Putative phosphorylation site by checkpoint kinases in the carboxy-terminal of hBRCA2. The threonine that is phosphorylated is highlighted (T3387 in hBRCA2 and T926 in dmBRCA2) (E=0.00091, Zscore=100). (**b**) Conserved sequence in the carboxy-terminal of the BRCA2 protein sequence: Checkpoint kinases Chk1 and Chk2 phosphorylate threonine 3387 in hbrca2, and is seen to be conserved in the mammalian and *Drosophila* species (T926 in *dmBRCA2*). (**c**) Putative BRC repeat identified by the similarity of fragment 61 (634–664:LDTALKRSIESSEEMRSKASKLVVVDTTMR) in *D. melanogaster* to the BRC4 repeat in hBRCA2 (1517–1551) (E=0.0017, Zscore=95) (red for identity, green for similarity).

generated by varying two parameters - length of the fragments and the threshold %similarity value for a significant match in a fragment pair. As mentioned above, the results are presented in two formats - best cumulative score and best average score.

Discussion

Genetic evolution over large time spans often leaves little trace of kinship in different organisms, even when the functional roles of the genes remains conserved. A relevant example is the *BRCA2* gene which, although present in many eukaryotes, is weakly conserved[48]. The BRCA2 protein plays a major role in maintaining genomic stability, fostering genetic variability and also has other cellular functions[2,49]. Individuals with germline mutations in the *BRCA2* gene are at significantly greater risk to a wide range of cancers[50,51]. This is supposed to be primarily due to the instability in chromosome structure and number induced by functional aberrations in BRCA2[52]. Conserved fragments of ~30 amino acids (BRC repeats)[7] that mediates the interaction of BRCA2 with the RAD51 recombinase[53] have been instrumental in identifying BRCA2 orthologs in other species[14,16]. The BRCA2 protein in the *Drosophila* genus assumes significance in this context owing to the advanced tools available for *Drosophila* genetics[35], and has been functionally characterized recently[11,15].

However, weak sequence conservation in this gene has proven to be an impediment in associating experimentally proven functionally relevant gene fragments in humans and *Drosophila*. The variability in the number of BRC repeats even within the *Drosophila* species has provided fodder for further speculation on the evolution of this gene[17,20]. The detection of a new BRC repeat would necessitate the reevaluation of such hypotheses.

Apart from the BRC repeats, RAD51 interacts with BRCA2 in the carboxy-terminal, and this interaction is modulated by checkpoint kinases[9,10]. Since the introduction of BRC repeats in the cell inhibits the formation of RAD51 nucleoprotein filaments[8], a model has been suggested whereby RAD51 binds to both the BRC repeats and the carboxy-terminal in undamaged cells, and DNA damage triggers the release of the carboxy-terminal bound RAD51 via the phosphorylation of a threonine residue[10].

Thus, it is noted that certain functionally significant domains are much more conserved compared to the complete protein[48]. In the current work, a methodology to annotate proteins in such 'twilight' zones[36] by fragmenting and aligning two protein sequences (Figure 1) has been presented. The results are sorted based on differing criteria, and can be directed by a input file in case the sequences have already been annotated. This method helps in quickly honing onto conserved sites through visual inspection (Table 1 and Figure 1). The threonine phosphorylation site (E=9.1e-4) for checkpoint kinases (Chk1/Chk2) (Figure 1) and a new BRC repeat (E=17e-4) using

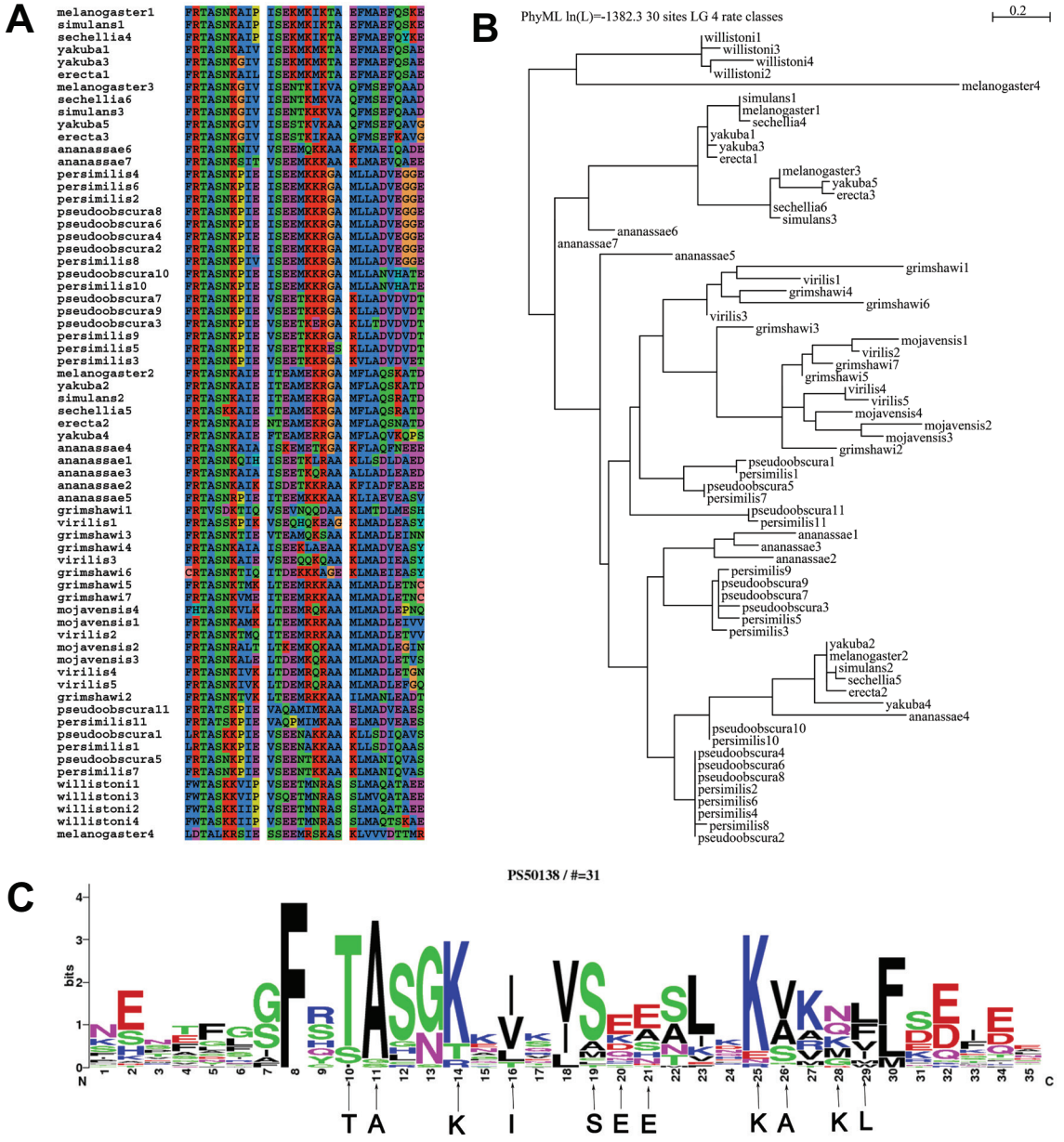

Figure 2. New BRC repeat identified by FRAGAL in *Drosophila melanogaster*. (**a**) The multiple alignment for this new sequence (634–664:LDTALKRSIESSEEMRSKASKLVVVDTTMR) (using ClustalW) highlighted as melanogaster4 and other sequences compared previously in[17]. This putative sequence is more closely related to the sequences in *D. willistoni* than other members of the genus. (**b**) The phylogenetic tree (using PHYML) gives a graphical representation of the relation of the various repeats in the *Drosophila* genus, corroborating the closer relation of the new BRC repeat to *D. willistoni*. (**c**) Alignment of the new BRC repeat to the sequence logo of the Prosite BRCA2 repeat profile PS50138.

FRAGAL (Figure 2) has been identified. Pruning out matches which do not have a corresponding conserved sequence in hBRCA2 helps us to select fragment 61 in dmBRCA2 as a new BRC repeat[7,14], and fragment 91 in dmBRCA2 as the putative threonine site for phosphorylation by checkpoint kinase Chk1 and Chk2[10]. It must be noted that the sites identified remain putative until verified by experimental data, in spite of the low E-values obtained.

The multiple alignments can be used to create (for the carboxy-terminal phosphorylation threonine site) or extend (for the new BRC repeat) Prosite motifs. However, the carboxy-terminal phosphorylation threonine site Prosite motif generated from the multiple alignment of sequences from *Drosophila* and mammals did not result in any matches in other organisms (*Ustilago maydis* and *Caenorhabditis elegans*).

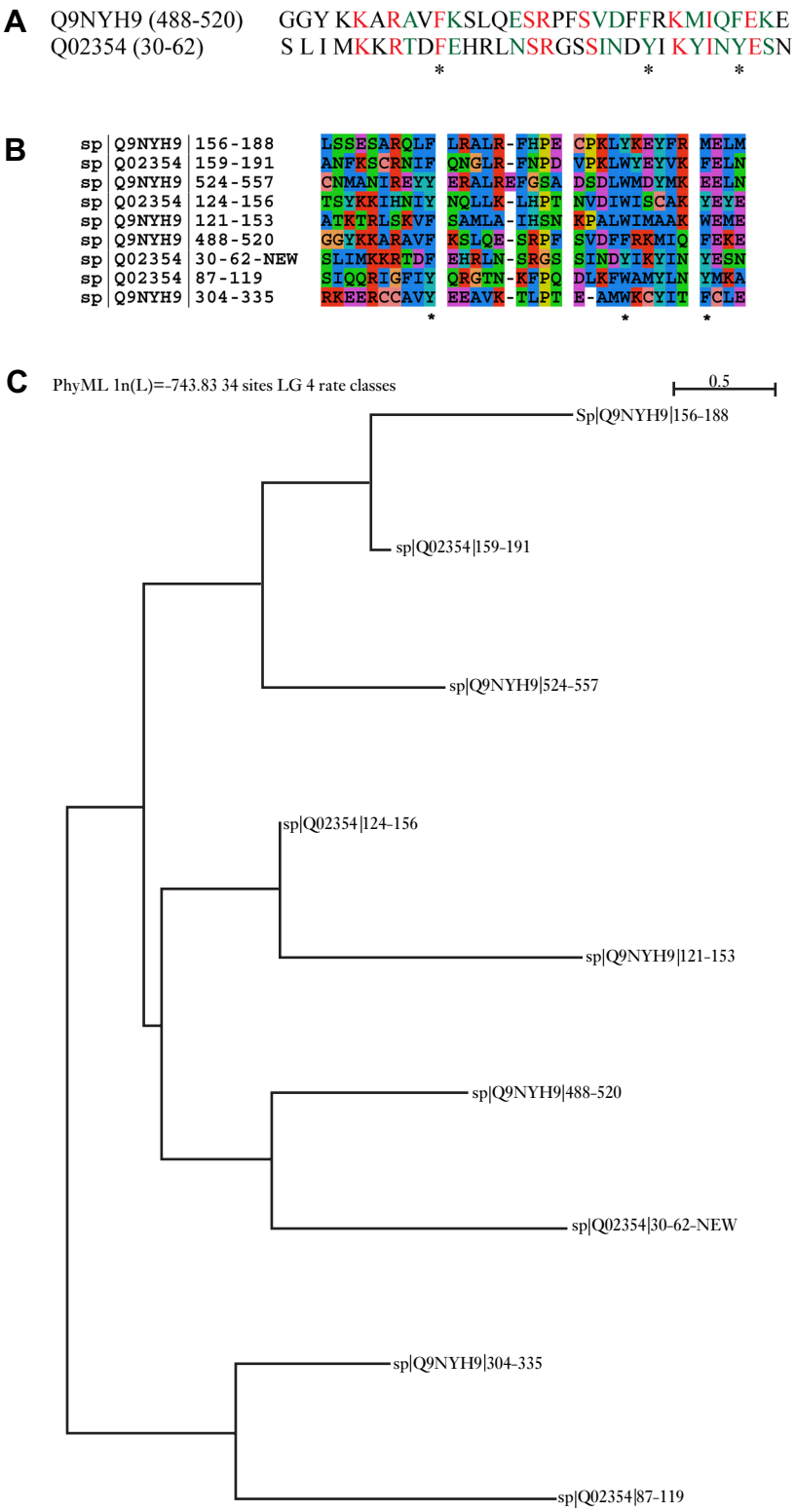

Figure 3. New Half-a-tetratricopeptide (HAT) motif identified by FRAGAL in *Saccharomyces cerevisiae*. (**a**) Pairwise alignment of a previously unannotated HAT motif in *S. cerevisiae* (E=11e-4, Z-score=116) (red for identity, green for similarity). (**b**) The multiple alignment for this new sequence (using ClustalW) with other HAT motifs in *S. cerevisiae* and humans shows large variation amongst HAT sequences even within the same organism. The conserved spacing of the aromatic residues are also highlighted. (**c**) The phylogenetic tree (using PHYML) shows that certain HAT repeats are more similar to HAT repeats from other organisms than to other HAT repeats in its own sequence.

In order to justify this method further, I concentrated on proteins that contain the Half-a-tetratricopeptide (HAT) repeat motifs. The HAT motif is much less ubiquitous than the related tetratricopeptide (TPR) repeat, and has been implicated in various aspects of RNA metabolism[23,24]. HAT motifs are also hypothesized to play a critical role in assembling RNA-processing complexes[25]. A recent study that combined bioinformatics, modeling and mutagenesis studies of the HAT domain used the three tandem HAT motifs in the *Saccharomyces cerevisiae* protein Utp6 to make inferences about the residues that confer structural and/or functional properties to the motif. In the current work, the detection of a new HAT repeat sequence (E=11e-4) in Utp6 from *S. cerevisiae* has been reported. This sequence has the desired aromatic residues at the proper spacing, a requisite for being considered a HAT repeat[25]. The above mentioned study would have gained further by the knowledge of this HAT repeat, a repeat that remained undetected by sequence analysis using other methods. The HAT repeats are much more varied, and thus not suitable for generating motifs (like Prosite[27]). For example, the consensus sequence has been derived from an alignment of 742 HAT motifs from Pfam[30] and had to be manually edited since this alignment included gaps in greater than 90% of the sequences[25]. Moreover, FRAGAL detects that a particular HAT sequence in one protein is more related to HAT sequences from other species that other HAT repeats present in its own sequence. This raises interesting questions about their evolutionary history.

In some of the significant matches in Table 1 the fragment in hBRCA2 is not annotated to be functionally relevant - for example fragments 33 and 87 of dmBRCA2 and fragments 176 and 194 in hBRCA2, respectively. These fragments might suggest an important, yet unknown, functional relevance of that stretch of the human gene as well, since it is conserved across distant species.

An excellent database for *Drosophila* related information is available at http://flybase.org/[54]. A database (www.sanchak.com/fragal.html) for BRCA2 and nucleolar RNA-associated proteins from different organisms, and will be updating this on a regular basis to include more organisms and different repeats has been created. The increasing importance of *Drosophila* as a model system for cancer research[55] in the search for human therapeutics[56–58] can be exploited to the hilt once the conserved mechanism is fully understood. FRAGAL presents the first step by annotating putative conserved sequence fragments in *Drosophila* and humans.

Competing interests
No competing interests were disclosed.

Grant information
This work was funded by the Tata Institute of Fundamental Research (Department of Atomic Energy), and the Department of Science and Technology (JC Bose Award Grant). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements
I gratefully acknowledge Chaitali Khan for introducing me to the lack of annotation of the BRCA2 gene in *Drosophila melanogaster*, and for simulating technical discussions. I am also indebted to B. J. Rao for technical inputs. I would also like to thank Ishita Mehta for helping me in preparing the manuscript.

Supplementary materials

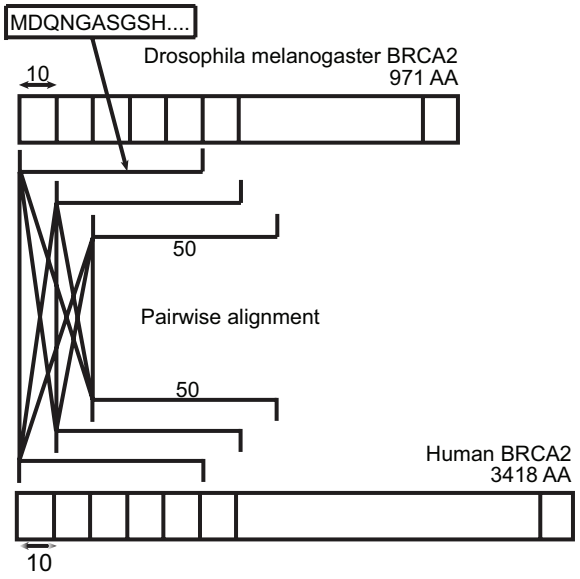

Figure S1. FRAGAL methodology. The BRCA2 protein sequences from *Drosophila melanogaster* and humans are split into fragments of parameterized length (50 in this case), at a parameterized interval (10 in this case). All pairs of fragments are aligned, and the results stitched such that there are no overlap in any given match and the order of the match is not interspersed. The alignment is done using the global alignment program 'needle' from the EMBOSS suite, and the gap penalties are set to 25 to discourage gaps in the middle of the alignment.

```
Repeat  1002 - 1036  35   BRCA2 1
Repeat  1212 - 1246  35   BRCA2 2
Repeat  1421 - 1455  35   BRCA2 3
Repeat  1517 - 1551  35   BRCA2 4
Repeat  1664 - 1698  35   BRCA2 5
Repeat  1837 - 1871  35   BRCA2 6
Repeat  1971 - 2005  35   BRCA2 7
Repeat  2051 - 2085  35   BRCA2 8

Region  1 - 40         40   Interaction with PALB2
Region  639 - 1000    362  Interaction with NPM1
Region  2350 - 2545  196  Interaction with FANCD2


Modified residue  683   1  Phosphoserine Ref.14
Modified residue  755   1  Phosphoserine Ref.14
Modified residue  3291  1  Phosphoserine; by CDK1 and
                           CDK2 Ref.11
Modified residue  3387  1  Phosphothreonine; by CHEK1
                           and CHEK2 Ref.16
```

Figure S2. Annotating the hbrca sequence. The syntax is similar to the one used in the UNIPROT accession site.

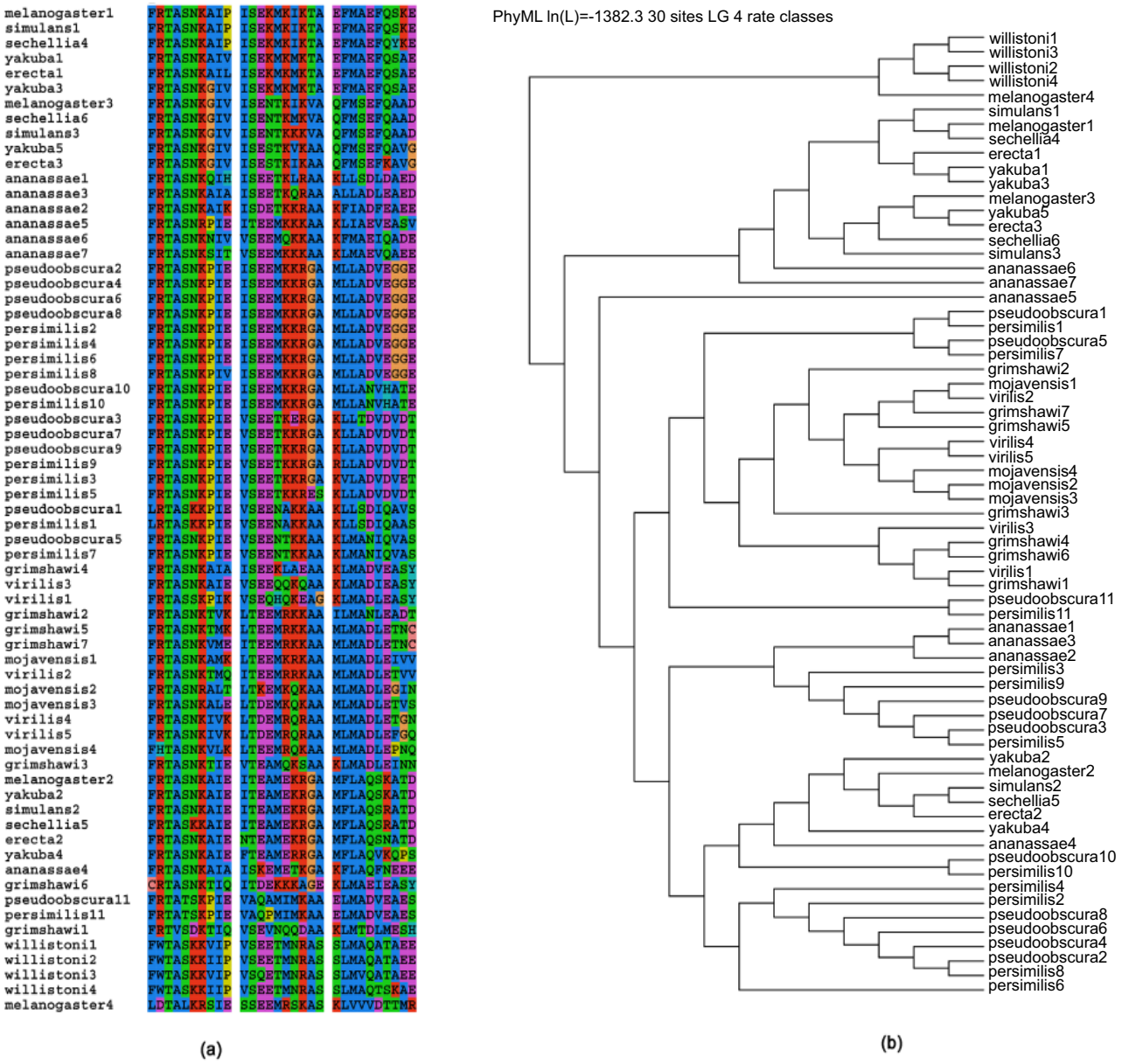

Figure S3. New BRC repeat identifed by FRAGAL in *Drosophila melanogaster* aligned using MAFFT. (**a**) The multiple alignment for this new sequence (634-664:LDTALKRSIESSEEMRSKASKLVVVDTTMR) (using ClustalW) marked as melanogaster4 and other sequences compared previously in[17], using MAFFT for doing multiple sequence alignment. This putative sequence is more closely related to the sequences in *D. willistoni* than other members of the genus, as shown by the alignment done using ClustalW. (**b**) The phylogenetic tree corroborates the closer relation of the new BRC repeat to *D. willistoni*, similar to the ClustalW results.

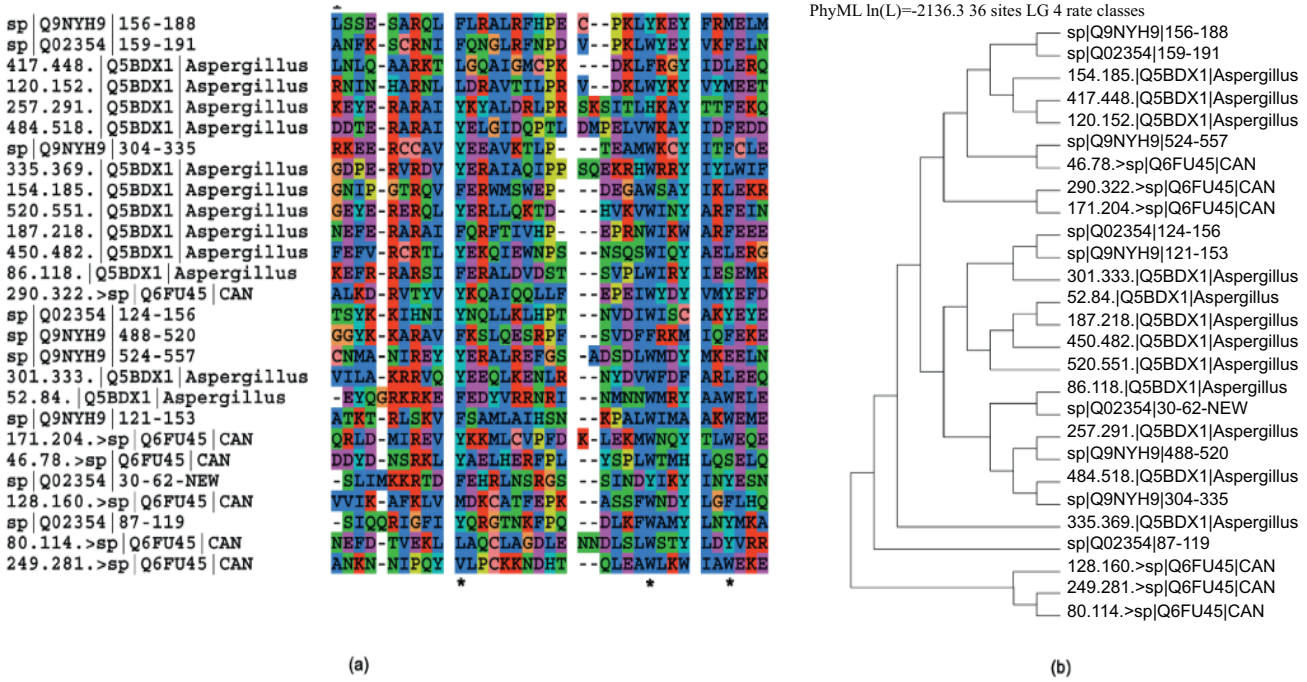

Figure S4. Including more proteins having the HAT motif. (**a**) The multiple alignment for this new sequence (using ClustalW) with other HAT motifs in *S. cerevisiae*, humans, *Aspergillus nidulans* and *Candida glabrata* shows large variation amongst HAT sequences even within the same organism. The conserved spacing of the aromatic residues remains the same. (**b**) Similar to the case when fewer organisms were included, the phylogenetic tree (using PHYML) shows that certain HAT repeats are more similar to HAT repeats from other organisms than to other HAT repeats in its own sequence.

References

1. Wooster R, Bignell G, Lancaster J, *et al.*: **Identification of the breast cancer susceptibility gene BRCA2.** *Nature.* 1995; **378**(6559): 789–792.
PubMed Abstract | Publisher Full Text
2. Thorslund T, West SC: **BRCA2: a universal recombinase regulator.** *Oncogene.* 2007; **26**(56): 7720–7730.
PubMed Abstract | Publisher Full Text
3. Jasin M: **Homologous repair of DNA damage and tumorigenesis: the BRCA connection.** *Oncogene.* 2002; **21**(58): 8981–8993.
PubMed Abstract | Publisher Full Text
4. Liu Y, West SC: **Distinct functions of BRCA1 and BRCA2 in double-strand break repair.** *Breast Cancer Res.* 2002; **4**(1): 9–13.
PubMed Abstract | Publisher Full Text | Free Full Text
5. Schlacher K, Christ N, Siaud N, *et al.*: **Double-strand break repair-independent role for BRCA2 in blocking stalled replication fork degradation by MRE11.** *Cell.* 2011; **145**(4): 529–542.
PubMed Abstract | Publisher Full Text | Free Full Text
6. Holthausen JT, Wyman C, Kanaar R: **Regulation of DNA strand exchange in homologous recombination.** *DNA Repair (Amst).* 2010; **9**(12): 1264–1272.
PubMed Abstract | Publisher Full Text
7. Bork P, Blomberg N, Nilges M: **Internal repeats in the BRCA2 protein sequence.** *Nat Genet.* 1996; **13**(1): 22–23.
PubMed Abstract | Publisher Full Text
8. Pellegrini L, Yu DS, Lo T, *et al.*: **Insights into DNA recombination from the structure of a RAD51-BRCA2 complex.** *Nature.* 2002; **420**(6913): 287–293.
PubMed Abstract | Publisher Full Text
9. Esashi F, Christ N, Gannon J, *et al.*: **CDK-dependent phosphorylation of BRCA2 as a regulatory mechanism for recombinational repair.** *Nature.* 2005; **434**(7033): 598–604.
PubMed Abstract | Publisher Full Text
10. Bahassi EM, Ovesen JL, Riesenberg AL, *et al.*: **The checkpoint kinases Chk1 and Chk2 regulate the functional associations between hBRCA2 and Rad51 in response to DNA damage.** *Oncogene.* 2008; **27**(28): 3977–3985.
PubMed Abstract | Publisher Full Text
11. Klovstad M, Abdu U, Schupbach T: ***Drosophila* brca2 is required for mitotic and meiotic DNA repair and efficient activation of the meiotic recombination checkpoint.** *PLoS Genet.* 2008; **4**(2): e31.
PubMed Abstract | Publisher Full Text | Free Full Text
12. Sharan SK, Pyle A, Coppola V, *et al.*: **BRCA2 deficiency in mice leads to meiotic impairment and infertility.** *Development.* 2004; **131**(1): 131–142.
PubMed Abstract | Publisher Full Text
13. Thorslund T, Esashi F, West SC: **Interactions between human BRCA2 protein and the meiosis-specific recombinase DMC1.** *EMBO J.* 2007; **26**(12): 2915–2922.
PubMed Abstract | Publisher Full Text | Free Full Text
14. Lo T, Pellegrini L, Venkitaraman AR, *et al.*: **Sequence fingerprints in BRCA2 and RAD51: implications for DNA repair and cancer.** *DNA Repair (Amst).* 2003; **2**(9): 1015–1028.
PubMed Abstract | Publisher Full Text
15. Brough R, Wei D, Leulier S, *et al.*: **Functional analysis of *Drosophila melanogaster* BRCA2 in DNA repair.** *DNA Repair (Amst).* 2008; **7**(1): 10–19.
PubMed Abstract | Publisher Full Text
16. Kojic M, Kostrub CF, Buchman AR, *et al.*: **BRCA2 homolog required for proficiency in DNA repair, recombination, and genome stability in Ustilago maydis.** *Mol Cell.* 2002; **10**(3): 683–691.
PubMed Abstract | Publisher Full Text
17. Bennett SM, Noor MA: **Molecular evolution of a *Drosophila* homolog of human BRCA2.** *Genetica.* 2009; **137**(2): 213–219.
PubMed Abstract | Publisher Full Text | Free Full Text
18. Wong AK, Pero R, Ormonde PA, *et al.*: **RAD51 interacts with the evolutionarily conserved BRC motifs in the human breast cancer susceptibility gene brca2.** *J Biol Chem.* 1997; **272**(51): 31941–31944.
PubMed Abstract | Publisher Full Text
19. Chen CF, Chen PL, Zhong Q, *et al.*: **Expression of BRC repeats in breast cancer cells disrupts the BRCA2-Rad51 complex and leads to radiation hypersensitivity and loss of G(2)/M checkpoint control.** *J Biol Chem.* 1999; **274**(46): 32931–32935.
PubMed Abstract | Publisher Full Text
20. Bennett SM, Mercer JM, Noor MA: **Slip-sliding away: serial changes and homoplasy in repeat number in the *Drosophila yakuba* homolog of human cancer susceptibility gene BRCA2.** *PLoS One.* 2010; **5**(6): e11006.
PubMed Abstract | Publisher Full Text | Free Full Text
21. Needleman SB, Wunsch CD: **A general method applicable to the search for similarities in the amino acid sequence of two proteins.** *J Mol Biol.* 1970; **48**(3): 443–453.
PubMed Abstract | Publisher Full Text

22. Rice P, Longden I, Bleasby A: **EMBOSS: the European Molecular Biology Open Software Suite.** *Trends Genet.* 2000; **16**(6): 276–277.
PubMed Abstract | Publisher Full Text

23. Preker PJ, Keller W: **The HAT helix, a repetitive motif implicated in RNA processing.** *Trends Biochem Sci.* 1998; **23**(1): 15–16.
PubMed Abstract | Publisher Full Text

24. Hammani K, Cook WB, Barkan A: **RNA binding and RNA remodeling activities of the half-a- tetratricopeptide (HAT) protein HCF107 underlie its effects on gene expression.** *Proc Natl Acad Sci U S A.* 2012; **109**(15): 5651–5656.
PubMed Abstract | Publisher Full Text | Free Full Text

25. Champion EA, Kundrat L, Regan L, *et al.*: **A structural model for the HAT domain of Utp6 incorporating bioinformatics and genetics.** *Protein Eng Des Sel.* 2009; **22**(7): 431–439.
PubMed Abstract | Publisher Full Text | Free Full Text

26. Pedruzzi I, Rivoire C, Auchincloss AH, *et al.*: **HAMAP in 2013, new developments in the protein family classification and annotation system.** *Nucleic Acids Res.* 2013; **41**(Database issue): D584–589.
PubMed Abstract | Publisher Full Text | Free Full Text

27. Sigrist CJ, Cerutti L, de Castro E, *et al.*: **PROSITE, a protein domain database for functional characterization and annotation.** *Nucleic Acids Res.* 2010; **38**(Database issue): D161–166.
PubMed Abstract | Publisher Full Text | Free Full Text

28. Koua D, Cerutti L, Falquet L, *et al.*: **PeroxiBase: a database with new tools for peroxidase family classification.** *Nucleic Acids Res.* 2009; **37**(Database issue): D261–266.
PubMed Abstract | Publisher Full Text | Free Full Text

29. Attwood TK, Coletta A, Muirhead G, *et al.*: **The PRINTS database: a fine-grained protein sequence annotation and analysis resource--its status in 2012.** *Database (Oxford).* 2012; **2012**: bas019.
PubMed Abstract | Publisher Full Text | Free Full Text

30. Finn RD, Mistry J, Tate J, *et al.*: **The Pfam protein families database.** *Nucleic Acids Res.* 2010; **38**(Database issue): D211–222.
PubMed Abstract | Publisher Full Text | Free Full Text

31. Gelfand MS, Mironov AA, Pevzner PA: **Gene recognition via spliced sequence alignment.** *Proc Natl Acad Sci U S A.* 1996; **93**(17): 9061–9066.
PubMed Abstract | Publisher Full Text | Free Full Text

32. Wu YW, Rho M, Doak TG, *et al.*: **Stitching gene fragments with a network matching algorithm improves gene assembly for metagenomics.** *Bioinformatics.* 2012; **28**(18): i363–i369.
PubMed Abstract | Publisher Full Text | Free Full Text

33. Sekelsky JJ, Brodsky MH, Burtis KC: **DNA repair in *Drosophila:* insights from the *Drosophila* genome sequence.** *J Cell Biol.* 2000; **150**(2): F31–36.
PubMed Abstract | Publisher Full Text | Free Full Text

34. Ludwig T, Chapman DL, Papaioannou VE, *et al.*: **Targeted mutations of breast cancer susceptibility gene homologs in mice: lethal phenotypes of Brca1, Brca2, Brca1/Brca2, Brca1/p53, and Brca2/p53 nullizygous embryos.** *Genes Dev.* 1997; **11**(10): 1226–1241.
PubMed Abstract | Publisher Full Text

35. Bier E: ***Drosophila*, the golden bug, emerges as a tool for human genetics.** *Nat Rev Genet.* 2005; **6**(1): 9–23.
PubMed Abstract | Publisher Full Text

36. Rost B: **Twilight zone of protein sequence alignments.** *Protein Eng.* 1999; **12**(2): 85–94.
PubMed Abstract | Publisher Full Text

37. Larkin MA, Blackshields G, Brown NP, *et al.*: **Clustal W and Clustal X version 2.0.** *Bioinformatics.* 2007; **23**(21): 2947–2948.
PubMed Abstract | Publisher Full Text

38. Guindon S, Lethiec F, Duroux P, *et al.*: **PHYML Online--a web server for fast maximum likelihood-based phylogenetic inference.** *Nucleic Acids Res.* 2005; **33**(Web server issue): W557–559.
PubMed Abstract | Publisher Full Text | Free Full Text

39. Gouy M, Guindon S, Gascuel O: **SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building.** *Mol Biol Evol.* 2010; **27**(2): 221–224.
PubMed Abstract | Publisher Full Text

40. Wu CH, Yeh LS, Huang H, *et al.*: **The Protein Information Resource.** *Nucleic Acids Res.* 2003; **31**(1): 345–347.
PubMed Abstract | Publisher Full Text | Free Full Text

41. Clark AG, Eisen MB, Smith DR, *et al.*: **Evolution of genes and genomes on the *Drosophila* phylogeny.** *Nature.* 2007; **450**(7167): 203–218.
PubMed Abstract | Publisher Full Text

42. Comet JP, Aude JC, Glemet E, *et al.*: **Significance of Z-value statistics of Smith-Waterman scores for protein alignments.** *Comput Chem.* 1999; **23**(3–4): 317–331.
PubMed Abstract | Publisher Full Text

43. Russo CA, Takezaki N, Nei M: **Molecular phylogeny and divergence times of drosophilid species.** *Mol Biol Evol.* 1995; **12**(3): 391–404.
PubMed Abstract

44. Katoh K, Standley DM: **MAFFT multiple sequence alignment software version 7:improvements in performance and usability.** *Mol Biol Evol.* 2013; **30**(4): 772–780.
PubMed Abstract | Publisher Full Text | Free Full Text

45. Altschul SF, Madden TL, Schaffer AA, *et al.*: **Gapped BLAST and PSI- BLAST: a new generation of protein database search programs.** *Nucleic Acids Res.* 1997; **25**(17): 3389–3402.
PubMed Abstract | Publisher Full Text | Free Full Text

46. Shivji MK, Davies OR, Savill JM, *et al.*: **A region of human BRCA2 containing multiple BRC repeats promotes RAD51-mediated strand exchange.** *Nucleic Acids Res.* 2006; **34**(14): 4000–4011.
PubMed Abstract | Publisher Full Text | Free Full Text

47. Shivji MK, Mukund SR, Rajendra E, *et al.*: **The BRC repeats of human BRCA2 differentially regulate RAD51 binding on single- versus double-stranded DNA to stimulate strand exchange.** *Proc Natl Acad Sci U S A.* 2009; **106**(32): 13254–13259.
PubMed Abstract | Publisher Full Text | Free Full Text

48. Warren M, Smith A, Partridge N, *et al.*: **Structural analysis of the chicken BRCA2 gene facilitates identification of functional domains and disease causing mutations.** *Hum Mol Genet.* 2002; **11**(7): 841–851.
PubMed Abstract | Publisher Full Text

49. Ayoub N, Rajendra E, Su X, *et al.*: **The carboxyl terminus of Brca2 links the disassembly of Rad51 complexes to mitotic entry.** *Curr Biol.* 2009; **19**(13): 1075–1085.
PubMed Abstract | Publisher Full Text | Free Full Text

50. Stratton MR, Wooster R: **Hereditary predisposition to breast cancer.** *Curr Opin Genet Dev.* 1996; **6**(1): 93–97.
PubMed Abstract | Publisher Full Text

51. Meyer P, Landgraf K, Hogel B, *et al.*: **BRCA2 mutations and triple-negative breast cancer.** *PLoS One.* 2012; **7**(5): e38361.
PubMed Abstract | Publisher Full Text | Free Full Text

52. Venkitaraman AR: **Linking the cellular functions of BRCA genes to cancer pathogenesis and treatment.** *Annu Rev Pathol.* 2009; **4**: 461–487.
PubMed Abstract | Publisher Full Text

53. Rajendra E, Venkitaraman AR: **Two modules in the BRC repeats of BRCA2 mediate structural and functional interactions with the RAD51 recombinase.** *Nucleic Acids Res.* 2010; **38**(1): 82–96.
PubMed Abstract | Publisher Full Text | Free Full Text

54. McQuilton P, St Pierre SE, Thurmond J, *et al.*: **FlyBase 101--the basics of navigating FlyBase.** *Nucleic Acids Res.* 2012; **40**(Database issue): D706–714.
PubMed Abstract | Publisher Full Text | Free Full Text

55. Vidal M, Cagan RL: ***Drosophila models* for cancer research.** *Curr Opin Genet Dev.* 2006; **16**(1): 10–16.
PubMed Abstract | Publisher Full Text

56. Chelouah S, Monod-Wissler C, Bailly C, *et al.*: **An integrated *Drosophila* model system reveals unique properties for F14512, a novel polyamine-containing anticancer drug that targets topoisomerase II.** *PLoS One.* 2011; **6**: e23597.
PubMed Abstract | Publisher Full Text | Free Full Text

57. Gladstone M, Su TT: **Chemical genetics and drug screening in *Drosophila* cancer models.** *J Genet Genomics.* 2011; **38**(10): 497–504.
PubMed Abstract | Publisher Full Text

58. Das T, Cagan R: ***Drosophila* as a novel therapeutic discovery tool for thyroid cancer.** *Thyroid.* 2010; **20**(7): 689–695.
PubMed Abstract | Publisher Full Text

Current Referee Status:

Referee Responses for Version 2

Saurabh Sinha
Department of Computer Science, University of Illinois at Urbana-Champaign, IL, USA

Approved: 28 November 2013

Referee Report: 28 November 2013
I have read the author's rebuttal carefully, and am now completely satisfied with how he has addressed my previous comments.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Satish Chikkagoudar
Pacific Northwest National Laboratory, Washington, USA

Approved: 22 October 2013

Referee Report: 22 October 2013

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Referee Responses for Version 1

Saurabh Sinha
Department of Computer Science, University of Illinois at Urbana-Champaign, IL, USA

Approved: 06 September 2013

Referee Report: 06 September 2013
In this well-written manuscript, Chakraborty presents a tool for local alignment of two protein sequences that includes a fragment-chaining step. He then uses this tool to identify important putatively functional fragments in two different *Drosophila* proteins by comparison to the respective human ortholog. A database containing results of many more similar applications is also presented and is a nice aspect of

the work.

I have the following specific comments, mostly related to the presentation, that might help the author improve the clarity of the manuscript.

The current title is rather long. The contribution of this work is mainly in the form of the FRAGAL tool, and the title could be trimmed to emphasize only that. The two example applications to finding the phosphorylation site, BRC repeat etc. are not experimentally substantiated biological claims, and may be better off being left out of the title.

Clear discussion should be provided regarding other previous work where pairs of aligned fragments are stitched together (e.g. exon chaining, see Jones & Pevzner 2004, and the chain/net approach to whole genome alignments).

Since the FRScore does not include gap penalties, I am assuming that each pair of fragments is subjected to two distinct similarity-scoring approaches; the gap-based approach when aligning that pair of fragments using 'needle' and the match/mismatch based approach when ranking the aligned pairs. This should be stated clearly, to avoid confusion. Is there a reason why the needle score was not used in place of the FRScore?

It appears that the FR score is the unweighted sum of %similarity and %identity. This should be stated explicitly.

I did not quite understand the formatting of Table 1.

- I think there should be a line separating ‘A’ from ‘B’ (which I think comes after a row with rank 14 in the second column). It took me some time to see that there are two sub-tables being shown here.
- Also, the fact that the same row is used to show different entities was confusing; usually a table is constructed so that a row shows different pieces of information about one entity.
- I assume something like A91-B337 refers to the starting positions of a matching fragment between sequences A and B, and the length of that fragment is not indicated in the row. Is this correct? (On reading further I realize that this interpretation is incorrect, and the numbers in a match are arbitrary indices and not coordinates. This was not clear from the legend.)
- I found that presenting sub-tables ‘A’ and ‘B’ (which I finally realized does not relate to ‘A’ and ‘B’ sequences) leads to more confusion than it helps. If both sub-tables present the same ranked list and the only difference is that ‘A’ filters for “conserved” fragments, then it might be better to show only sub-table ‘B’ and add a column indicating if this is a fragment marked conserved.

Where is the E-value of an FRScore coming from? (I am not familiar with what the '*Protein Information Resource*' provides.) Perhaps this E-value correspond to the global alignment score reported by needle?

The results of Table 1 do not aid ones understanding as to how the fragmented alignment, i.e. stitching together of fragments, helps in this case. As shown, this appears similar in form to a ranked list of matches from a standard local-aligner. Similarly, with respect to Figure 2, the author may wish to discuss why FRAGAL finds the 'melanogaster4' fragment as a BRC repeat where previous annotations (that found three repeats) failed. Was this a matter of previous methods ‘*missing the threshold*’? (This seems unlikely given the strong E-value reported for this.) Similar clarifications for the HAT repeat finding exercise will also be helpful.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

1 Comment

Author Response

Sandeep Chakraborty, Tata Institute of Fundamental Research, India
Posted: 17 Sep 2013

I greatly appreciate the positive comments.

The current title is rather long. The contribution of this work is mainly in the form of the FRAGAL tool, and the title could be trimmed to emphasize only that. The two example applications to finding the phosphorylation site, BRC repeat etc. Are not experimentally substantiated biological claims, and may be better off being left out of the title.

Since this work began in search for unannotated fragments of the dmBRCA2 sequence, and is being followed in the lab actively, I think the reference to an application of FRAGAL is warranted. However, I have removed the reference to the HAT repeat and stated the fact that the phosphorylation site and the BRC repeat is putative in the title.

Clear discussion should be provided regarding other previous work where pairs of aligned fragments are stitched together...

I have discussed the spliced alignment approach to genome assembly in the discussion. However, I have noted that while these methods use graph algorithms to solve the computationally difficult problem of exon chaining, FRAGAL does the converse by finding best matches in known exon chains (i.e. protein sequences).

Since the FRScore does not include gap penalties, I am assuming that each pair of fragments is subjected to two distinct similarity-scoring approaches; the gap-based approach when aligning that pair of fragments using 'needle' and the match/mismatch based approach when ranking the aligned pairs. This should be stated clearly, to avoid confusion. Is there a reason why the needle score was not used in place of the FRScore?

The needle score includes gap penalties, which is something that is not intended for use in FRScore, as you have correctly pointed out. The idea is to direct the alignment to discourage gaps – but once the alignment is done a gap should not have a penalty. It is ‘real' and therefore only the identity or similarity that matters.

It appears that the FR score is the unweighted sum of %similarity and %identity. This should be stated explicitly.

I have empirically assigned more weightage to the %identity based on the fact that we are

searching for repeats, and expect higher conservation. This conservation is magnified a bit more by assigning higher weightage.

I did not quite understand the formatting of Table 1. ... I found that presenting sub-tables A and B (which I finally realized does not relate to A and B sequences) leads to more confusion than it helps.

I apologize for this confusion. The line demarcating subtables A and B was lost in the typesetting - and I missed out on detecting this error. Further, naming the subtables A and B was a poor choice of names, since the sequences were also named A and B. Finally, I agree that the second column was unnecessary, as was two subtables. I have simplified this table.

I assume something like A91-B337 refers to the starting positions of a matching fragment between sequences A and B, and the length of that fragment is not indicated in the row. Is this correct? (On reading further I realize that this interpretation is incorrect, and thenumbers in a match are arbitrary indices and not coordinates. This was not clear from the legend.)

I apologize for this oversight. This is mentioned in the web pages - http://sanchak.com/fragal/ALLRUNS.BRCA2/Caenorhabditiselegans.G5EG86.Homosapiens.P51587.g as `*Multiply index with 10 to get sequence starting position in original sequence'*. Thus A91 refers to the sequence starting at 910 in `A' and going till 959, since the fragmenting length is 50. I have now mentioned this at the beginning of the Results section, and in the legend.

Where is the E-value of an FRScore coming from? ... Perhaps this E-value corresponds to the global alignment score reported by needle?

I have specified the website (http://pir.georgetown.edu/pirwww/search/pairwise.shtm), and cited the paper by Wu C *et al.* (2003) The Protein Information Resource. Nucleic Acids Res 31: 345347.

... with respect to Figure 2, the author may wish to discuss why FRAGAL finds the 'melanogaster4' *fragment as a BRC repeat where previous annotations (that found three repeats) failed. ... Similar clarifications for the HAT repeat finding exercise will also be helpful.*

I could only make an educated guess as to why other tools failed to detect these repeats. I believe that the tools used had a 'sequential' methodology and therefore one match fixed the order of the next searches. Not all known BRC repeats have a low E-value when aligned with the new BRC repeat. For example, the first BRC repeat in hBRCA2 when aligned to the new dmBRCA2 repeat has an E=0.04, much more than the E=17e-4 observed for the fourth repeat (which is the one I report here). Ideally, if one took all the BRC repeats and did a search in the dmBRCA2 sequence, this new repeat would be reported. Essentially, this is what FRAGAL does, albeit implicitly, by automatically fragmenting the sequence. The same logic applies to the HAT repeat, where the sequences are more varied and thus the choice of the repeat would affect the detection of new motifs.

Competing Interests: No competing interests were disclosed.

Satish Chikkagoudar
Pacific Northwest National Laboratory, Washington, USA

Approved with reservations: 22 August 2013

Referee Report: 22 August 2013
The author presents an interesting technique for detecting new BRC repeats. The paper is generally well written, but needs some additional material to bolster its case. The introduction section needs more discussion of the 'state-of-the-art' in the alignment and motif detection area (especially with respect to detecting BRC repeats). The paper's argument can be made stronger by explicitly mentioning the advantages of fragmented alignment over any other recursively applied local alignment method or homology search method. Some discussion of existing methods exists in the discussion/results section, but that needs to be made available in the introduction in order to justify the need for creating a new method An explanation of the choice of parameter values needs to be given. For example, why does FRscore have weights of 1/3 for only Similarity (equation 2)? Also, the reasoning for choosing particular values for gap open and gap extend penalties needs to be mentioned along with whether parameter tuning/search was done to arrive at those values. Otherwise, those numbers seem arbitrary. The author needs to discuss whether he tried any other alignment algorithms apart from Clustal-W. Some other algorithms such as MAFFT, ProbCons or CONTRAlign may yield better results. The author may want to discuss the data in a separate section under methods/materials. A figure describing the FRAGAL pipeline will be useful to visually describe the pipeline/algorithm.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Competing Interests: No competing interests were disclosed.

2 Comments

Author Response

Sandeep Chakraborty, Tata Institute of Fundamental Research, India
Posted: 23 Aug 2013

Dear Dr Chikkagoudar,

I would like to thank you for your insightful suggestions which will help improve the manuscript. I will make the suggested changes, and incorporate them in a new version shortly.

A small clarification - you have asked for a "figure describing the FRAGAL pipeline". There is a supplementary figure S1 doing this. Do you think that this figure is insufficient, or were you suggesting that I move this to the main manuscript?

Best regards,

Sandeep
Competing Interests: No competing interests were disclosed.

Author Response

Sandeep Chakraborty, Tata Institute of Fundamental Research, India
Posted: 17 Sep 2013

I appreciate the positive comments, and hope to have addressed the concerns detailed below.

The introduction section needs more discussion of the 'state-of-the-art' in the alignment and motif detection area (especially with respect to detecting BRC repeats) ... Some discussion of existing methods exists in the discussion/results section, but that needs to be made available in the introduction in order to justify the need for creating a new method.

I have moved this section to the introduction. In response to the comments of another reviewer (Dr Saurabh Sinha), I have also mentioned the possible reasons why existing tools have failed to detect these repeats. Further, I have cited methods that use spliced alignment methods for genome assembly, noting that these methods use graph algorithms to solve the computationally difficult problem of exon chaining. FRAGAL does the converse by finding best matches in known exon chains (i.e. protein sequences).

An explanation of the choice of parameter values needs to be given. For example, why does FRscore have weights of 1/3 for only Similarity (equation 2)?

The extra weightage given to % identity in the score is due to the fact that one expects more sequence conservation in repeats.

...the reasoning for choosing particular values for gap open and gap extend penalties needs to be mentioned along with whether parameter tuning/search was done to arrive at those values.

The gap penalties are set to discourage gaps, but not gap extensions. The gap opening has been set to two values - 10 and 25. Results using both values have been uploaded in the database http://sanchak.com/fragal/BRCA2.html . The results using 25 have been observed to be better. However, a complete statistical analysis of these values is beyond the scope of this work.

The author needs to discuss whether he tried any other alignment algorithms apart from Clustal-W. Some other algorithms such as MAFFT, ProbCons or CONTRAlign may yield better results.

I have used another tool (MAFFT) to generate the multiple sequence alignment. The results from the new alignment tool, mirrors the inference drawn from Clustal-W. This is now a supplementary figure.

A figure describing the FRAGAL pipeline will be useful to visually describe the pipeline/algorithm.

I have added a pseudo code of the FRAGAL program in the main manuscript.
Competing Interests: No competing interests were disclosed.

Himanshu Sinha
Department of Biological Sciences, Tata Institute of Fundamental Research, Mumbai, India

Approved: 12 August 2013

Referee Report: 12 August 2013
A well written paper that proposes a new method, FRAGAL for identifying functional putative motifs within protein sequences which have been hidden from previous analyses. By splitting the sequences into overlapping fragments, this method is able to discover additional motifs. The author has tested his technique on BRC repeats in *Drosophila dm*BRCA2 and HAT repeats in budding yeast Utp6, by comparing them to corresponding human protein sequences. The BRC repeat has been well analysed with comparisons across several *Drosophila* species. However the author does not provide extensive comparison of HAT repeats in *Saccharomyces* species. Since the sequences of several *Saccharomyces* sibling species and closely related fungi such as *Aspergillus*, *Candida*, etc. are known, it would be interesting to see how conserved this new HAT repeat is within the overall conservation of Utp6.

While the author establishes the advantage of FRAGAL technique, it is too early to say that this is a useful generic tool to identify known and novel motifs in protein sequences. I would request the author to run his FRAGAL code on several protein sequences with small motifs to estimate success rates and false discovery rates of his method. A supplementary table should be provided describing several sequences analysed by this method and these rates.

A minor comment, Table 1 should be simplified with the two BRCA2 protein sequences presented in two sub-tables. Please explain why certain ranks are missing in FRscore and %S.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: Although I am affiliated with the same institution as the author, I was not involved at any stage of manuscript preparation and did not collaborate with him at the time the review was written.

2 Comments

Author Response

Sandeep Chakraborty, Tata Institute of Fundamental Research, India
Posted: 23 Aug 2013

Dear Dr Sinha,

I greatly appreciate your comments on my manuscript. I will incorporate your suggested changes and update the manuscript. Your suggestion of including more of such repeats (I plan to do BIR and TPR) is something that will take some computational time and thus the delay.

Best regards,

Sandeep
Competing Interests: No competing interests were disclosed.

Author Response

Sandeep Chakraborty, Tata Institute of Fundamental Research, India
Posted: 17 Sep 2013

I am grateful for the encouraging comments on the work.

...author does not provide extensive comparison of HAT repeats in Saccharomyces *species. ... it would be interesting to see how conserved this new HAT repeat is within the overall conservation of Utp6.*

I have implemented this interesting idea using proteins which have the HAT repeat from *Aspergillus nidulans* and *Candida glabrata*. This is now a Supplementary figure. However, these do not provide any further insights into the evolution of the HAT repeat, and would require sophisticated analyses beyond my expertise.

I would request the author to run his FRAGAL code on several protein sequences with small motifs to estimate success rates and false discovery rates of his method. A supplementary table should be provided describing several sequences analyzed by this method and these rates.

In accordance with this suggestion, I have run FRAGAL on two more motifs (BIR and TPR). However, I failed to detect any new repeats using these two motifs. These are now part of the database - http://sanchak.com/fragal.html.

A minor comment, Table 1 should be simplified with the two BRCA2 protein sequences presented in two sub-tables. Please explain why certain ranks are missing in FRscore...

I have simplified the table considerably based on the comments of another reviewer (please see below). The naming of the sub tables as A and B was confusing given that the query and target sequences were named A and B. Further; the columns for the similarity scoring has been removed. We did not ever use the similarity only score, and this was adding to the confusion. I have now clearly stated the reason for some missing ranks. I apologize for the confusing aspects of this table.
Competing Interests: No competing interests were disclosed.