J Mol Evol (2010) 71:444–457
DOI 10.1007/s00239-010-9399-y

Selection and the Cell Cycle: Positive Darwinian Selection in a Well-Known DNA Damage Response Pathway

Mary J. O'Connell

Received: 16 April 2010 / Accepted: 6 October 2010 / Published online: 4 November 2010
© Springer Science+Business Media, LLC 2010

Abstract Cancer is a common occurrence in multi-cellular organisms and is not strictly limited to the elderly in a population. It is therefore possible that individuals with genotypes that protect against early onset cancers have a selective advantage. In this study the patterns of mutation in the proteins of a well-studied DNA damage response pathway have been examined for evidence of adaptive evolutionary change. Using a maximum likelihood framework and the mammalian species phylogeny, together with codon models of evolution, selective pressure variation across the interacting network of proteins has been detected. The presence of signatures of adaptive evolution in BRCA1 and BRCA2 has already been documented but the effect on the entire network of interacting proteins in this damage response pathway has, until now, been unknown. Positive selection is evident throughout the network with a total of 11 proteins out of 15 examined displaying patterns of substitution characteristic of positive selection. It is also shown here that modern human populations display evidence of an ongoing selective sweep in 9 of these DNA damage repair proteins. The results presented here provide the community with new residues that may be relevant to cancer susceptibility while also highlighting those proteins where human and mouse have undergone lineage-specific functional shift. An understanding of this damage response pathway from an evolutionary perspective will undoubtedly contribute to future cancer treatment approaches.

Keywords Cell cycle · Cancer evolution · Cancer selection · Tumor suppressor · Breast cancer · Positive selection · Adaptive evolution · DNA repair · Vertebrate evolution

Introduction

Cancer is not confined to the elderly in a population, it also effects organisms of reproductive age. The presence of neoplasias right across the animal kingdom from molluscs and arthropods to reptiles and mammals, places this as a very ancient disease. Therefore, animals must have developed ways to fight/prevent cancer—at least until after reproductive age. For longer-lived, larger animals and those evolving new morphologies, selection to delay or prevent deaths due to cancer would have been particularly significant. Therefore, it seems reasonable to assume that a selective advantage would exist for those in the population with anticancer adaptations, e.g., reduction of somatic mutation rate (Graham 1992).

Many mechanisms exist to protect the genomic integrity of organisms, with selection on maintaining somatic integrity declining with age. Responses to double strand breaks include cell cycle checkpoints, signal transduction mechanisms, and DNA repair pathways, for review see (O'Driscoll and Jeggo 2006). Efficient workings of DNA damage response pathways are crucial to survival, and so it follows that failure of natural selection in these damage

Electronic supplementary material The online version of this article (doi:10.1007/s00239-010-9399-y) contains supplementary material, which is available to authorized users.

M. J. O'Connell (✉)
Bioinformatics and Molecular Evolution Group,
School of Biotechnology, Dublin City University,
Glasnevin, Dublin 9, Ireland
e-mail: Mary.oconnell@dcu.ie

M. J. O'Connell
Centre for Scientific Computing & Complex Systems Modelling (SCI-SYM), Dublin City University, Glasnevin,
Dublin 9, Ireland

response pathways results in cancer. This poses an important question: Has the maintenance of genomic stability or reduced cancer risk been a target of selection? If this is the case then those species with shorter germ line generation times and higher metabolic rates should show increased selective pressure on tumor suppressor genes, or indeed in pathways responsible for maintaining genomic integrity. Where positive selective pressure acts to improve the fitness of one element of a system, it seems inevitable that a trade-off must occur elsewhere in the system. This is one possible explanation for the close relationship between the data analyzed here and involvement in cancer (da Fonseca et al. 2010). A test of whether selective pressure could be at work to reduce the risk of cancer against a background of rapid developmental changes is presented here.

Fanconi anemia (FA) is an autosomal recessive cancer-susceptibility syndrome (Bagby 2003; D'Andrea and Grompe 2003). It has contributed hugely to our understanding of tumor suppression and DNA damage response mechanisms (Freie et al. 2004; Nakanishi et al. 2005; Offit et al. 2003; Wang et al. 2004). Specifically FA has recently been heralded as a very attractive model for the study of breast cancer (Wang 2007). However, there are notable phenotypic differences between human FA and mouse FA.

Thus far a number of FA proteins (denoted as FANC_) have been identified: FANCA, -B, -C, -D2, -D1(BRCA2), -F, -G, -E, -L, -J, -M, and, -N. These proteins have no sequence similarity to each other or to other known DNA repair genes. The FA proteins interact with well-known DNA damage response proteins, including BRCA-1 and -2, ATM and NBS1 (Hussain et al. 2003; Hussain et al. 2004; Wang et al. 2004; Xia et al. 2007) in a common pathway known as the FA/BRCA pathway, as illustrated in Fig. 1 (for review see (D'Andrea and Grompe 2003)). This FA/BRCA pathway has been shown to be disrupted in a subset of ovarian tumor cell lines (Taniguchi et al. 2003). A subset of FANC proteins, namely: FANCA, -C, -G, -F, -E, -B, -L, and -M, combine to form a functional unit called the FA core complex present in the nucleus of normal cells (Blom et al. 2004; Garcia-Higuera et al. 2001; Hussain et al. 2003). When DNA is subject to damage or mitomycin C, the core complex and BRCA1 co-operate to regulate the monoubiquitylation and thus activation of FANCD2 on lysine 561, thus resulting in mitomycin C resistance (Garcia-Higuera et al. 2001). In response to ioinizing radiation the Ataxia Telangiectasia Mutated (ATM) protein phosphorylates FANCD2 on serine 222, resulting in the establishment on an S phase checkpoint (Taniguchi et al. 2002). Thus, activated FANCD2 is translocated to chromatin and DNA repair foci where it co-localizes with pathway members ATM, BRCA1, BRCA2, NBS1, and RAD51 (D'Andrea and Grompe 2003). What follows is a

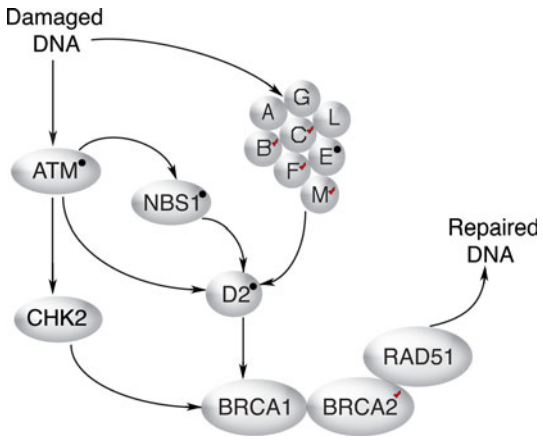

Fig. 1 FA/BRCA pathway of interacting proteins and summary of site-specific results. The *arrows* refer to the order of action of DNA damage response elements in the pathway. Selective pressure analysis results are summarized for those genes displaying site-specific evidence of positive selection by the symbol beside the gene name in the figure. *Ticks* represent those genes with evidence of positive selection solely using site-specific models (M0–M8). *Filled circles* represent those genes with evidence of positive selection under both site (M0–M8) and lineage-specific models (models A–B) of evolution

cascade of events involving RAD51 at the final stage before DNA repair is complete. The remaining FANC proteins are involved in the downstream cascade, see Fig. 1.

Evolutionary studies of oncogenes and tumor suppressors have an application in the identification of functionally essential regions/residues, prioritization of regions for further study, predicting interacting regions, and, predicting those protein residues that are more or less likely to have a negative impact on human health. An understanding of the evolutionary history of DNA damage response pathways has the potential to reveal the molecular basis for species differences in disease severity, and to identify amino acids that have experienced heterogeneous selective pressures. For this reason the BRCA1 protein has been the focus of many independent analyses of evolutionary rate variation (Burk-Herrick et al. 2006; Schmid and Yang 2008). Positive Selection is the term given to describe the retention and spread of advantageous mutations throughout a population. The level of replacement substitution per replacement site (Dn) compared to silent substitution per silent site (Ds) is calculated for each gene. Elevated levels of Dn compared to Ds, or more specifically the ratio of Dn/Ds, known as ω, >1, corresponds to adaptive evolutionary pressures, or selective pressure for change. These are usually indicative of functional shifts in a protein (Davids et al. 2002; Yang 1998), and become fixed more quickly than neutral substitutions in a population due to the advantage they convey to their possessor. It has been demonstrated that the widely used sliding window analysis

of rate variation produces artifactual trends of synonymous and nonsynonymous rate variation (Schmid and Yang 2008). For this reason we have resolved to use the more computationally intense and rigorous framework of the likelihood ratio test (LRT) (Schmid and Yang 2008). In this analysis the entire interacting network of tumor suppressors in the FA/BRCA pathway have been analyzed. There is a direct relationship between mutations in these proteins and increased cancer risk. To determine if there is an ongoing selective sweep in this pathway in modern human populations, the integrated haplotype score (iHS) for each gene for each population was calculated (Voight et al. 2006).

From intensive studies of this pathway over the past 20 years, our knowledge of the underlying molecular reasons for these disorders is rapidly increasing but a number of key issues remain unanswered. The most intriguing of all is what effect, if any, has selection had on the evolution of these pathways and indeed on cancer itself (Crespi and Summers 2005; Crespi and Summers 2006; Leroi et al. 2003; Zimmer 2007; Graham 1992).

Materials and Methods

Data Collection and Alignment Generation

Alignments of orthologous sequences from the BRCA/FA pathway were generated using 1-to-1 orthologues from (i) the Ensembl collection of completed eukaryotic genomes (www.ensembl.org), and (ii), BLASTp searches of GenBank to increase the species set. The genes analyzed here are ATM, BRCA1, BRCA2, CHK2, NBS1, RAD51, FANCA, FANCB, FANCC, FANCD2, FANCE, FANCF, FANCG, FANCL, and FANCM. These 15 genes have single gene orthologues in 6 or more species. All alignments are available in Supplementary File 1. The phylogeny used is the canonical species phylogeny, pruned appropriately for each dataset (Murphy et al. 2001). Information on species names, accession numbers, and alignment lengths are given in the Supplementary Information (SI Table 1). The use of an outgroup sequence that is more distantly related to the human and the mouse than they are to each other permits directionality for the test. All alignments were carried out at the amino acid level using the default settings in ClustalX 1.81 (Thompson et al. 1994), and gap positions were placed into the nucleotide sequences according to where they were found in the protein alignment using in-house software. Unfortunately the level of knowledge of protein interaction domains between these proteins and available 3D structures of (i) suitable quality, and (ii) suitable binding partner, limited the analysis to the sequences rather than the more desirable and biologically realistic tertiary structure analysis (Berglund et al. 2005).

Measuring Selection

A variety of models of codon sequence evolution were employed and using maximum likelihood estimates and the LRT, or Akaike information criterion (AIC) where appropriate, the goodness-of-fit of alternative models to the data were tested (these models are available in the PAML package (Yang 1998; Yang and Nielsen 2002; Yang et al. 2000)). The LRT proceeds by comparing nested models of sequence evolution. This allows us to determine the statistical significance of alternative models to the data, thereby allowing us to determine, for example, whether a model that allows for variations in nonsynonymous to synonymous mutation rates, i.e., Dn/Ds or ω, is a better fit than one that allows only for neutral evolution. These models allow for variable ω ratios among sites, along different branches of a phylogenetic tree or a combination of both. These models imply that there are a variety of classes of sites in a given set of aligned sequences and the LRT provides a method of identifying the model that best describes the evolution of the set of sequences. One model is usually constrained so that $\omega \leq 1$. The more general model allows at least one class of sites to exist where the ω value is dependent on the data. In those cases where the ω value exceeds unity and the resulting increase in the likelihood score is significant, evidence of positive selection in this protein can be inferred. In order to ascertain significance of difference in likelihood score the likelihood statistic is used, $2\Delta l$ with χ^2_v, where v is the number of degrees of freedom and corresponds with the number of free parameters. A brief description of these models now follows, full details of these models have been published previously (O'Connell and McInerney 2005; Yang and Nielsen 2002; Yang et al. 2005).

The simplest model (fewest free parameters) is called M0. In this model it is assumed that there is a single ω value at all sites and across all lineages. This corresponds to the Goldman and Yang model (Goldman and Yang 1994). Model M1 assumes that there are two classes of sites—those with an ω value of zero and those with an ω value of one. Model M2 allows for three classes of sites—class 1 have an ω value of zero, class 2 an ω value of one, and class 3 an ω value that is not fixed to any value and is estimated from the data. Given the relationship between M1 and M2, they can be tested for the significance of the difference of the fit of these two models using an LRT with df = 2. Model M3 allows all ω values to vary freely. There are two variants of this model employed in this analysis. The first is where there are two classes of sites that are free to vary ($k = 2$) and the second is where there are three classes of sites ($k = 3$). M3 ($k = 2$) can be tested for its fit against M0 with df = 2. M3 ($k = 3$) cannot be tested against any of the other models presented here using an

LRT, however, by direct comparison of the likelihood scores under the AIC, its comparison with M3 ($k = 2$) can be interesting if the difference is greater than 1.

A number of models that use discrete approximations to continuous distributions to model variability in ω at different sites has also been applied. The first of these, M7, assumes that variation in ω follows a beta distribution. A total of ten classes of sites are assumed to exist and their ω values are constrained to be between zero and one. The second model, M8, allows the existence of another class of site where the ω value is allowed to be greater than unity. M8 and M7 can be compared with one another using the LRT with df = 2. The null model of M8, i.e., ω is fixed to 1 − M8a, was also applied. M8a and M8 are compared using an LRT with df = 1 using a 50:50 mixture of point mass 0 and χ^2, so the critical χ^2 values are 2.71 at 5% and 5.41 at 1%.

Finally, two models that allow the ω value to vary across sites and across different lineages have been applied to the data. Both human and mouse branches are labeled as foreground for the following reasons: (i) they are present in each alignment and have high quality sequence data, (ii) represent the greatest difference in terms of germ line generation time, metabolic rate and size across the dataset, and (iii), the mouse is the model for the vast majority of human cancers. However, applying a variety of tests such as those outlined here may raise issues for multiple testing, specifically in raising the probability of false rejection to an unacceptable level, particularly if all branches were to be labeled sequentially (Anisimova and Yang 2007). Only the two branches relevant to the hypothesis have been labeled. Estimates of false positive discovery rates for the models described above have varied from 10 to 14% for site-specific models, M0–M8 (Suzuki and Nei 2002) to 20% or even much higher for the first version of the branch-site models, Model A–Model B (Zhang 2004). The branch-site models have since been greatly improved and have been shown to be successful in avoiding high levels of false positives reporting, even in the presence of relaxed selective constraints (Zhang et al. 2005). The most up to date models available have been applied. The first of the lineage-specific models, Model A, is an extension of M1 (referred to as Test 1 in (Zhang et al. 2005)). The null model of Model A (Model A null) where all categories of ω are restricted to between 0 and 1 is also applied to the data. The LRT between Model A null and Model A has df = 2 (referred to as Test 2 in (Zhang et al. 2005)). The second type of lineage-site modeled applied is, Model B, an extension of M3 ($k = 2$). Both of these branch-site models can be compared with their site-specific counterparts using the LRT with df = 2. Individual sites where positive selection is most likely to have occurred are identified using the empirical Bayes approach as described previously (Yang et al. 2005). For all 15 proteins examined in this way, the gene phylogeny and the species phylogeny have been used independently and the results have been compared. The analysis of selective constraints ultimately varied only in a minor way. All gene phylogenies used are shown in Supplementary File 2.

The detection of positively selected residues using the Naïve Empirical Bayes (NEB) or Bayes Empirical Bayes (BEB) methods have been applied here, the preference is for BEB predictions when available as they have been shown to perform better following simulation analysis (Yang et al. 2005). All sites predicted with greater than 50% posterior probability have been given in order to give the complete set of output from codeml. For site directed mutagenesis based on this work, those sites with higher posterior probabilities from BEB predictions should be used.

Integrated Haplotype Score Analysis

The signature of positive directional selection is visible in population data as an unusually long haplotype of low diversity. This is caused by the favored allele increasing in frequency very rapidly in the population while the other genomic regions that do not contain the selected allele tend to have levels of diversity and linkage disequilibrium (LD) that are more typical of the entire genome. To determine if there is an ongoing selective sweep in this pathway in modern human populations, the iHS for each gene for each population was determined (Voight et al. 2006). The iHS has a number of advantages including that it is standardized using the genome wide empirical distributions, and it has an approximately standard normal distribution, i.e., iHS's from different regions can be directly compared (Voight et al. 2006). In comparison to other EHH-based statistics, the iHS outperforms across a broad range of strengths of selective sweep. Using iHS's therefore allows for unusual haplotypes around a SNP to be measured as compared to the genomic background. A large negative iHS indicates that a derived allele has increased in frequency (a selective sweep), while a large positive iHS indicates that the ancestral allele has been segregating. All available SNP data from East Asian (A), Northern and Western European (C), and African Yoruba (Y) were used. The online tool SNP@Evolution and HapMap release II source data was used to carry out the iHS analyses on a gene-by-gene basis (Cheng et al. 2009).

Results

All data used in the analysis are single gene orthologues for the species under analysis. Orthology of these data was

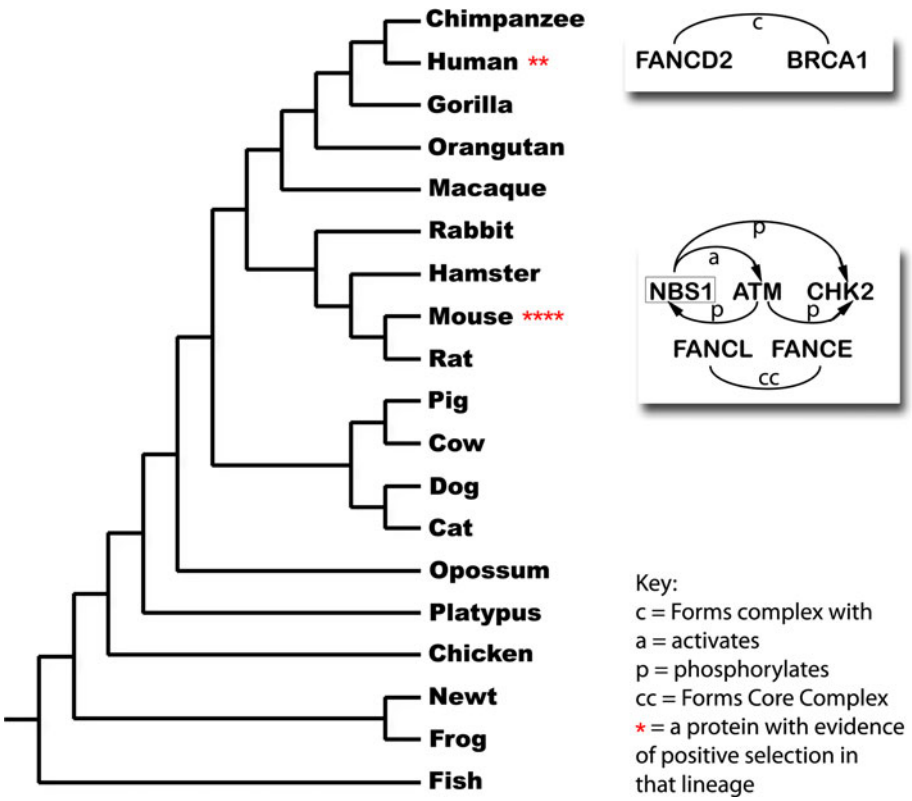

Fig. 2 Species phylogeny used in this analysis and summary of lineage-specific results. For each gene analyzed the available species were used, therefore suitably pruned versions of this species phylogeny were applied. The results of the lineage-specific analysis are shown on the branches for human and mouse. Those genes that show evidence of positive selection are marked with *asterisks* on the lineage where evidence of positive selection was detected. Known interactions between these proteins are shown beside the human and mouse branches, interactions between the proteins are depicted as *arrows*; the direction of the *arrowhead* indicates the direction of phosphorylation and/or activation. NBS1 is boxed because although it shows evidence of lineage specific selective pressure in mouse there is no evidence of positive selection on this protein in this lineage

Table 1 Summary of "best fit" models for FANCD2 following PAML analysis

Model	*P*	*L*	Estimates of parameters	Positively selected sites
Site-specific				
M3:Discrete ($k = 3$)	5	−24847.4083	$p_0 = 0.4018$, $p_1 = 0.4594$, $p_2 = 0.1388$, $\omega_0 = 0.0394$, $\omega_1 = 0.3212$, $\omega_2 = 1.1851$	NEB: 170 > 0.50 38 > 0.95 5 > 0.99
M8: Beta&Omega	4	−24848.4501	$p_0 = 0.9193$, $p = 0.6734$, $q = 2.1633$ ($p_1 = 0.0807$), $\omega = 1.4052$	BEB: 36 > 0.50
Branch Specific				
Model B (Human)	5	−24864.1139	$p_0 = 0.5774$, $p_1 = 0.4045$, ($p_2 = 0.0107$, $p_3 = 0.0075$), *Background*: $\omega_0 = 0.0759$, $\omega_1 = 0.6133$,$\omega_2 = 0.0759$, $\omega_3 = 0.6133$ *Foreground*: $\omega_0 = 0.0759$, $\omega_1 = 0.6133$, $\omega_2 = 999$, $\omega_3 = 999$	NEB: 16 > 0.50 3 > 0.95

The lineage- and site-specific models that fit are shown along with all the likelihood scores and parameter estimates

confirmed using the Ensembl Compara database (www.ensembl.org). Therefore, all phylogenetic gene trees are as in Fig. 2, and are derived from the species phylogeny (Murphy et al. 2001). Individual gene phylogenies were also used in the selection analysis. On comparison of gene phylogeny and species phylogeny results the parameter

Table 2 Summary of PAML analysis on all proteins in the FA/BRCA pathway, the likelihood model only is shown for each protein

Protein	Likelihood Model (Lineage)	*P*	Parameter estimates	Positive selection YES/NO	Number of sites selected
CHK2	Model A (Mouse)	3	$p_0 = 0.7285$, $p_1 = 0.2688$, ($p_2 = 0.0019$, $p_3 = 0.0007$) *Background*: $\omega_0 = 0.0604$, $\omega_1 = 1$, $\omega_2 = 0.0604$, $\omega_3 = 1$ *Foreground*: $\omega_0 = 0.0604$, $\omega_1 = 1$, $\omega_2 = 612$, $\omega_3 = 612$	YES	BEB: 3 > 0.50
FANC_C	M8	4	$p_0 = 0.9388$, $p = 0.8176$, q = 1.2406, ($p_1 = 0.0612$), $\omega = 2.1089$	YES	BEB: 15 > 0.50
FANC_D2	Model B (Human)	5	$p_0 = 0.5774$, $p_1 = 0.4045$, ($p_2 = 0.0107$, $p_3 = 0.0075$), *Background*: $\omega_0 = 0.0759$, $\omega_1 = 0.6133$, $\omega_2 = 0.0759$, $\omega_3 = 0.6133$ *Foreground*: $\omega_0 = 0.0759$, $\omega_1 = 0.6133$, $\omega_2 = 999$, $\omega_3 = 999$	YES	NEB: 16 > 0.50
	M8	4	$p_0 = 0.9193$, $p = 0.6734$, $q = 2.1633$	YES	BEB: 37 > 0.50
FANC_G	M3 ($k = 3$)	4	($p_1 = 0.0807$), $\omega = 1.4052$ $p_0 = 0.2022$, $p_1 = 0.5295$, $p_2 = 0.2683$ $\omega_0 = 0.0392$, $\omega_1 = 0.2899$, $\omega_2 = 0.9916$	NO	None
NBS1	M8	4	$p_0 = 0.9516$, $p = 0.5251$, $q = 1.1043$ ($p_1 = 0.0484$), $\omega = 2.0009$	YES	BEB: 19 > 0.50
	Model B (Mouse)		$p_0 = 0.1907$, $p_1 = 0.2454$, ($p_2 = 0.2466$, $p_3 = 0.3173$) *Background*: $\omega_0 = 0.0575$, $\omega_1 = 0.6100$, $\omega_2 = 0.0575$, $\omega_3 = 0.6100$ *Foreground*: $\omega_0 = 0.0575$, $\omega_1 = 0.6100$, $\omega_2 = 0.0391$, $\omega_3 = 0.0391$	NO	None
RAD51	M3 ($k = 3$)	4	$p_0 = 0.8212$, $p_1 = 0.1695$, $p_2 = 0.0093$ $\omega_0 = 0.0031$, $\omega_1 = 0.0894$, $\omega_2 = 0.4177$	NO	None
BRCA1	Model B (Human)	5	$p_0 = 0.3175$, $p_1 = 0.3176$, ($p_2 = 0.1823$, $p_3 = 0.1825$) *Background*: $\omega_0 = 0.2096$, $\omega_1 = 1.0469$, $\omega_2 = 0.2096$, $\omega_3 = 1.0469$ *Foreground*: $\omega_0 = 0.2096$, $\omega_1 = 1.0469$, $\omega_2 = 5.2485$, $\omega_3 = 5.2485$	YES	NEB: 22 > 0.50
BRCA2	M8	4	$p_0 = 0.9512$, $p = 0.6771$, $q = 0.7185$, ($p_1 = 0.0488$), $\omega = 2.3317$	YES	BEB: 63 > 0.50
ATM	M8	4	$p_0 = 0.9709$, $p = 0.4796$, $q = 2.3781$, ($p_1 = 0.0291$), $\omega = 1.6774$	YES	BEB: 74 > 0.50
	Model B (Mouse)	5	$p_0 = 0.7848$, $p_1 = 0.2047$, ($p_2 = 0.0083$, $p_3 = 0.0022$), *Background*: $\omega_0 = 0.0697$, $\omega_1 = 0.6912$, $\omega_2 = 0.0697$, $\omega_3 = 0.6912$ *Foreground*: $\omega_0 = 0.0697$, $\omega_1 = 0.6912$, $\omega_2 = 9.1272$, $\omega_3 = 9.1272$	YES	NEB: 12 > 0.50
FANC_A	M3 ($k = 3$)	5	$p_0 = 0.1673$, $p_1 = 0.4757$, $p_2 = 0.3571$, $\omega_0 = 0.0139$, $\omega_1 = 0.2287$, $\omega_2 = 0.7885$	NO	None
FANC_B	M3 ($k = 3$)	5	$p_0 = 0.2329$, $p_1 = 0.4928$, $p_2 = 0.2743$, $\omega_0 = 0.1062$, $\omega_1 = 0.5013$, $\omega_2 = 1.0771$	YES	NEB: 191 > 0.50
FANC_F	M3 ($k = 3$)	5	$p_0 = 0.2283$, $p_1 = 0.5472$, $p_2 = 0.2245$, $\omega_0 = 0.0205$, $\omega_1 = 0.2443$, $\omega_2 = 1.0070$	YES	NEB: 68 > 0.50
FANC_M	M8	4	$p_0 = 0.7961$, $p = 0.5964$, $q = 1.5428$, ($p_1 = 0.2039$), $\omega = 1.474$	YES	BEB: 87 > 0.50
FANC_E	M8	4	$p_0 = 0.8312$, $p = 1.0096$, $q = 3.0598$, ($p_1 = 0.1688$), $\omega = 1.5451$	YES	BEB: 65 > 0.50
	Model B (Mouse)	5	$p_0 = 0.6184$, $p_1 = 0.3603$, ($p_2 = 0.0135$, $p_3 = 0.0079$) *Background:* $\omega_0 = 0.1293$, $\omega_1 = 0.9276$, $\omega_2 = 0.1293$, $\omega_3 = 0.9276$ *Foreground*: $\omega_0 = 0.1293$, $\omega_1 = 0.9276$, $\omega_2 = 72.2848$, $\omega_3 = 72.2848$	YES	5 > 0.90 NEB: 4 > 0.50

Table 2 continued

Protein	Likelihood Model (Lineage)	*P*	Parameter estimates	Positive selection YES/NO	Number of sites selected
FANC_L	Model B (Mouse)	5	$p_0 = 0.7055$, $p_1 = 0.2635$, ($p_2 = 0.0226$, $p_3 = 0.0084$) *Background*: $\omega_0 = 0.0715$, $\omega_1 = 0.5054$, $\omega_2 = 0.0715$, $\omega_3 = 0.5054$ *Foreground*: $\omega_0 = 0.0715$, $\omega_1 = 0.5053$, $\omega_2 = 5.9003$, $\omega_3 = 5.9003$	YES	NEB: 3 > 0.50

estimates only varied in a minor way, in most cases the variations occurred after the decimal place. To avoid repetition in this section, the LRTs and AIC tests will be detailed for the analysis of the FANCD2 protein, see Table 1. A summary of the selective pressure analysis on all 15 proteins is given in Table 2, and a detailed set of PAML results for all proteins in the study are available in Supplementary Information (SI Tables 2-16). The notation of the models is described in the materials and methods section. The LRTs performed include a χ^2 test comparing: (1) model M0 with M3 ($k = 2$), (2) M1 with M2, (3) M7 with M8, (4) M8a with M8, (5) M1 with Model A, and (6) Model A null with Model A, (7) M3 ($k = 2$) with Model B, each with two degrees of freedom (df), and (8) a comparison of M3 ($k = 2$) with M3 ($k = 3$). The same LRTs and AICs are performed for all other proteins in this analysis, and the resulting likelihood model calculated for each protein is discussed. Where functional data is available, positively selected residues are examined in detail.

FANCD2 Analysis

Models of sequence evolution were examined where the ω value is allowed to vary from site to site in the alignment, summary shown in Table 1. These models are indicated using the notation M1 to M8 for all proteins analyzed (see Supplementary information Tables 2–16 for full detail).

The comparison of the site-specific models M7 and M8 indicate that there is a statistical improvement in the likelihood score when the LRT is performed, $2\Delta l = 29.44$ with df = 2, ($p \ll 0.05$) with 8% of sites under positive selection, $\omega = 1.405$. BEB predicts 36 sites with $P > 0.50$. The comparison of M8 with its null model, M8a, was 5.7796, this is significant with $p < 0.01$, thus confirming M8 to be a better fit to the data than the alternative models.

The final category of models used is the branch-site models described by Yang (Yang and Nielsen 2002; Zhang et al. 2005). Both the human and mouse branches are labeled as foreground in independent analyses. The LRT results indicate that Model B is significantly better than the discrete models ($2\Delta l = 8.3563$, df = 2, $p < 0.05$) for the mouse branch, and the estimates for ω show no evidence for positive selection in the mouse lineage. However, the likelihood score obtained for FANCD2 using Model B with the human branch treated as foreground is significantly better and indicates positive selection ($2\Delta l = 33.98$, df = 2, $p < 0.01$). Model B indicates that a small amount of the sites ($\sim$2%) are evolving under the influence of positive selection in the human lineage only, while in the remainder of the species these sites are under purifying selection. A total of 58% of the sites in the FANCD2 dataset are under strong purifying selection, with an ω value of 0.07592, and 40% of sites are evolving under slightly less strict purifying selection $\omega = 0.61325$. The inference is that a very small proportion of the FANCD2 protein has been under increased positive selective pressure in the human lineage only. Using NEB estimations, 16 sites are under positive selection, $P > 0.50$, for FANCD2.

The position and function of the 16 sites in the FANCD2 protein was determined. The C terminal 24 residues of FANCD2 isoform 2 are required for its function. From this analysis it is seen that 11 of these 24 essential residues are positively selected, and with a cluster of positively selected sites close to 3 sites which when mutated reduce phosphorylation by ATM. A single positively selected residue lies in the BRCA2 interacting domain. The remaining sites lie in regions without clearly assigned functions. It is of note that none of the 16 residues lie within the FANCE interacting domain.

The same LRTs were carried out for the remaining 14 proteins in the FA/BRCA pathway. What follows is a description of the maximum likelihood model that best fits the data for each member of this pathway, see Table 2 for summary, for detail on all parameter estimates see Supplementary information Tables 2–16. Figure 1 provides a summary of the site-specific results, and Fig. 2 provides a summary of positive selection detected in a lineage-site manner for the two foreground lineages tested (human and mouse). A systematic analysis of all proteins in the pathway reveals links between proteins with evidence of positive selection and their activating, phosphorylating or complex forming partners; this is also depicted in Fig. 2.

CHK2 Analysis

The LRTs for CHK2 show that Model A is the likelihood model with the Mouse lineage treated as foreground ($2\Delta l = 7.04$, df $= 2$, $p < 0.05$). There are 0.3% of the sites in the alignment with $\omega > 1$, indicating positive selection in the mouse lineage while all other lineages are neutral or purifying (see Fig. 2 for illustration of systematic positive selection in mouse between CHK2 and ATM). Using Bayesian estimations (BEB) 3 sites for CHK2 are under positive selection. One of these sites is located in the protein kinase domain, only 5 residues from the ATP-binding site. The other 2 sites are located at the start and end of the protein but as yet have no clearly assigned functions.

ATM Analysis

Under M8, 97% of the sites in ATM are either under purifying selection or neutrally evolving, ω between 0 and 1, and 3% of sites are under positive selective pressure, with an ω estimated at 1.6774. Using Bayesian statistics (BEB), 74 sites are classified as having a greater than 0.5 posterior probability of belonging to the positively selected class of sites. Fourteen of these sites are located in the FAT domain of the ATM protein. These domains are found at the C terminal ends of PIK related kinases and are involved in redox-dependent structural and cellular stability. There are 2 positively selected amino acids in the PI3K/PI4K domain responsible for phosphotransferase activity. Interestingly there are 8 sites under positive selection that are in close proximity (within 4 residues) of disease-implicated sites. The remaining 50 amino acid residues, although part of the functioning protein, have no clear functional assignment in the current literature. The best lineage-specific model for ATM is Model B with mouse treated as foreground. Under this model 1% of the sites in ATM are under positive selection in the mouse lineage alone, with $\omega = 9.1272$. There are 12 sites predicted to be in the positively selected category, two are located in each of the 2 domains: FAT domain and PI3K/PI4K domain. There is 1 site neighboring a residue implicated in ataxia telangiectasia (AT) disease at position 2822.

BRCA1 Analysis

Previous studies of the BRCA1 protein have reported site-specific positive selection under the model M8 (Burk-Herrick et al. 2006). However, the addition of the null model, M8a, to the available suite of models has resulted in the site-specific model M8 being rejected. In the case of the human lineage: it is estimated from Model B that 32% of sites in the BRCA1 protein are under purifying selection with an ω value of 0.2096, 36% have an ω value of 5.2485. 31% of the protein is under positive selection in all lineages, these findings are in agreement with previous studies (Burk-Herrick et al. 2006, Huttley et al. 2000, Pavlicek et al. 2004). 18% of these sites are under positive selection in the human lineage alone with $\omega = 5.2485$, while all background lineages are under strong purifying selective pressure at those sites, $\omega = 0.2096$. In the case of the mouse lineage: the results are consistent with the findings for human, there is a large proportion of the protein under positive selective pressure regardless of lineage with an $\omega = 1.1143$, while in the mouse lineage these positions were under purifying selection with $\omega = 0.2688$. From Bayesian statistics it is estimated that 22 sites in the human lineage alone have a posterior probability (P) of greater than 0.5 of being under positive selection (NEB). Identifying positive selection in this protein in a lineage-specific manner is not surprising, previous studies have reported evidence of positive selection in BRCA1 in the region involved in RAD51 interaction (Wakefield et al. 2005). The results presented here are consistent with previous findings for this protein. There are 6 of these 22 positively selected sites located at or adjacent to sites that are polymorphic in breast cancer, indicating that these regions are likely mutational hotspots.

FANCL Analysis

The results for FANCL show positive selection in the mouse lineage (Model B). There are 4% of the sites that have an ω value of 5.9053 in the mouse lineage while in all other species these positions have ω values of either 0.0715 or 0.5054, i.e., purifying selection. There is currently no information available on the function of these positions.

FANCC Analysis

With model M8 it is estimated that 94% of the sites in FANCC are under purifying selection, with an ω value of 0.0612, and 6% are positively selected with an ω value of 2.1089. From Bayesian estimates (BEB) there are 15 sites identified as being under positive selection with $P > 0.50$. There is little information available for the functional sites in FANCC from the literature and databases. However, one of the sites (residue 555) identified here as positively selected is neighboring a position (residue 554) that has previously been shown to be essential for the functional complementing activity of this Fanconi anemia protein (Gavish et al. 1993). Mutations at this position render the protein inactive and have been reported in Fanconi Anemia patients (Gavish et al. 1993; Strathdee et al. 1992; Verlander et al. 1994).

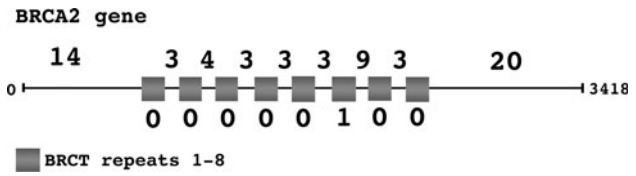

Fig. 3 Diagram summarizing the results for the BRCA2 gene analysis. A linear representation of the BRCA2 protein from position 0 to 3418, the 8 BRCT repeat regions are shown as *shaded blocks* on the protein. *Numbers above* the cartoon depict those sites positively selected in the various regions of the protein excluding the BRCT repeats. The number of sites positively selected in the BRCT repeat regions are given below the image

FANCE Analysis

Under M8, 83% of the sites in FANCE are either under purifying selection or neutrally evolving, ω between 0 and 1, and 17% of sites are under positive selection, with an ω estimated at 1.5451. Using Bayesian statistics (BEB), 5 sites are classified as having a greater than 0.9 posterior probability of belonging to the positively selected class of sites, 3 are in the FANCC interacting domain. The best lineage-specific model for FANCE is Model B with mouse treated as foreground. Under this model >1% of the sites in FANCE have an $\omega = 5.9$.

BRCA2 Analysis

The vast majority of BRCA2 (96%) is either under negative selection or neutrally evolving (model M8), and 4% of the protein is under positive selective pressure for change with an ω value of 2.3728. There are 63 sites selected using Bayesian statistics as having $P > 0.5$ of being under positive selection (BEB). The location of all 63 amino acids on the linear protein sequence has been determined. On closer examination it is evident that these sites form a specific pattern across the BRCA2 protein, see Fig. 3. The amino acid residues are only in those regions at the start and end of the protein and between the 8 BRCT repeats but not within the BRCT repeat domains themselves (with the exception of 1 out of 63 amino acids). Elsewhere in the protein, one residue is located in the FANCD2 interacting domain and is predicted to be under positive selection, this is interesting given that positive selection in the FANCD2 protein has also been predicted in the region that interacts with BRCA2. This pattern may be the result of selective pressure for improved/more rapid recognition of these two interacting partners. A further 7 of the positively selected cohort are within 4 residues of sites that are mutated in breast cancer, again indicating possible mutational hotspots.

NBS1 Analysis

Under M8 95% of the sites in NBS1 have an ω value indicative of purifying or neutral evolution and 5% of sites have an ω value of 2, 19 sites $P > 0.50$ BEB. One of these sites is in close proximity (within 3 residues) of a site implicated in childhood acute lymphoblastic leukemias. There are 3 of the positively selected sites that are each located within 3 residues of phosphoserine sites. Mutation studies have shown that these sites are essential for ATM-dependent phosphorylation. The remaining sites are located throughout the length of the protein sequence.

FANCB Analysis

Under M3 ($k = 3$) there are 27% of the sites with an ω value of 1.0771. NEB estimations predict 191 sites positively selected ($P > 0.50$). There is currently very little information in the literature on this protein and no available data on biochemical or mutagenesis studies.

FANCF Analysis

Under M3 ($k = 3$) the FANCF protein has 22% of sites with an ω value of 1.007. The remaining sites in the alignment have ω values of 0.0205 and 0.2443. Estimating the positions under positive selection the NEB method detects 68 sites under positive selection. There are 4 positively selected positions in close proximity (within 3 residues) of sites that when mutated result in reduction in monoubiquitination of FANCD2.

FANCM Analysis

M8 estimates an ω value of 1.474 for FANCM across 20% of the protein, 87 positions are predicted to be in the positively selected category following BEB analysis. One of these sites is in the helicase ATP-binding domain, and 6 of these sites are in the FAAP24 and EME1 interaction domain. Currently there are very few studies on this protein and so further comment on the functional implications of these sites is not possible.

FANCA, FANCG, and RAD51 results

The results for FANCA are indicative of purifying selection. The best model following LRT calculations is M3 ($k = 3$). Under model M3 ($k = 3$), 17% of the sites are estimated to have an ω value of 0.0139 and are under strong purifying selective pressure, 47% have a value of 0.2287 and 36% have an ω value of 0.7885. Therefore, all categories of sites in the FANCA protein are subject to either purifying selective constraints or are evolving

neutrally. For FANCG, M3 ($k = 3$) predicts no positive selection with ω values ranging from 0.0392 to 0.9916, i.e., ranging from strong purifying selection to neutrally evolving. For RAD51, M3 ($k = 3$) estimates 82% of the sites have an ω value of 0.0031 and 18% of the sites have an ω value of either 0.0894 or 0.4177. Therefore, considering the human and mouse lineages as foreground and all other lineages as background, all categories of sites in the RAD51 protein are subject to selective constraints to remain constant throughout evolution.

Taken together the results of lineage-specific positive selection form an interesting pattern that is suggestive of a systematic selective pressure on the pathway. From Fig. 2 it is seen that the human and mouse lineages display a difference in the levels of positive selection in this pathway, the human has 2 proteins and the mouse 4 proteins with signatures of positive selection unique to these lineages alone. The proteins FANCD2 and BRCA1 combine to form a functional complex and both of these show evidence of positive selection in the human lineage alone. ATM and CHK2 are involved in phosphorylation and activation of one another, as depicted in Fig. 2, both show evidence of positive selection in the mouse lineage alone. The FANCL and FANCE proteins form the core complex and both have evidence of positive selection in the mouse lineage alone.

Analysis of Modern Human Populations

Of the 15 DNA damage response pathway members, there is evidence for an ongoing selective pressure in modern human populations in nine of these genes. In three genes (ATM, BRCA2, FANCC) the evidence spans all three populations: African Yoruba (Y), East Asian (A), and European (C) populations, see Table 3. For CHK2, FANCL, and NBS1 there is evidence that "Y" and "A" populations are undergoing a selective sweep, while BRCA1 displays high iHS for the "Y" and "C" populations, see Fig. 4. In some cases single populations, either "Y", "A", or "C" display characteristics of ongoing selective sweep, these are FANCA for the "Y" population and RAD51 for the "A" population. Figure 4 displays a more detailed analysis of the SNPs for those proteins with significant |iHS|. Analysis of the SNPs in the coding DNA sequences of these proteins identified two replacement substitutions in both ATM (at amino acid position 1420 and 1040) and BRCA1 (amino acid position 1038 and 1613), while BRCA2 has one replacement substitution at amino acid position 2944, and FANCA has one replacement at amino acid position 266. These replacement substitutions are marked by the "*" in Fig. 4 and at present do not have assigned functional consequences. Analysis of the selective pressures on this pathway with the human population is highly indicative of an ongoing selective sweep.

Discussion/Conclusion

For all components of the FA/BRCA pathway, the selective pressures at work across the interacting network of proteins have been determined. In summary, the results show that the protein in the pathway display substitution patterns indicative of positive selection across the *Mammalia* with the exception of 3 proteins (FANCA, FANCG, and

Table 3 Results of analysis of iHS on all proteins across all populations available

	Significant iHS (selective sweep)	Y	C	A	Number of SNPS
ATM	✔	✔	✔	✔	158
CHK2	✔	✔	–	✔	45
BRCA1	✔	✔	✔	–	108
BRCA2	✔	✔	✔	✔	207
FANCA	✔	✔	–	–	12
FANCB	NA	NA	NA	NA	NA
FANCC	✔	✔	✔	✔	212
FANCD2	–	–	–	–	47
FANCE	–	–	–	–	15
FANCF	–	–	–	–	10
FANCG	–	–	–	–	7
FANCL	✔	✔	–	✔	87
FANCM	–	–	–	–	43
NBS1	✔	✔	–	✔	181
RAD51	✔	–	–	✔	37
14	9	8	4	7	1169

The proteins analyzed are listed in alphabetical order. The column titled "significant iHS" gives a summary of whether there is evidence of an ongoing selective sweep (✔) or not (–)

The populations are as per HapMap project: *Y* African Yoruba, *C* Northern and western European, and *A* Asian

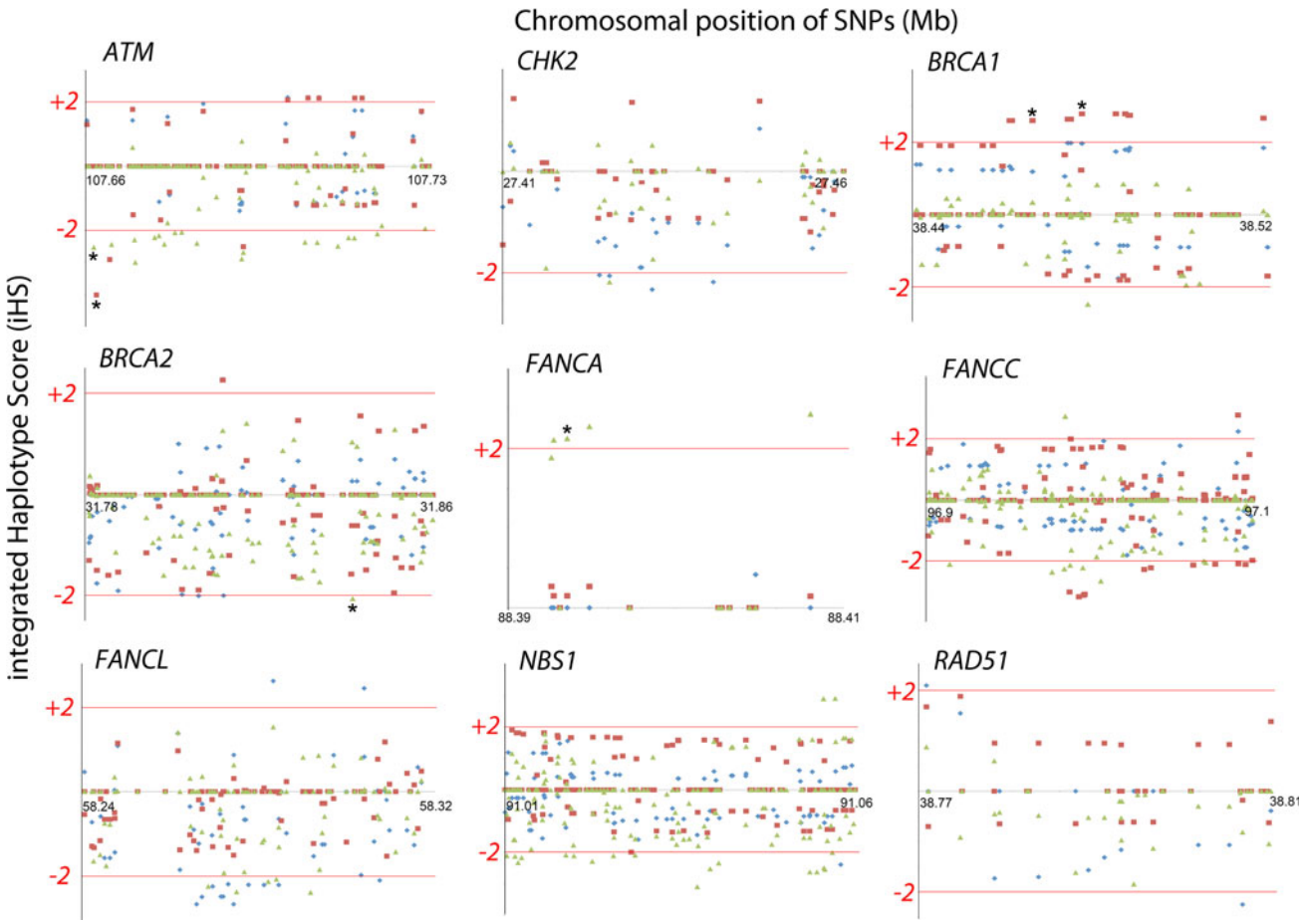

Fig. 4 Summary of integrated haplotype scores for 9 genes in the analysis. The cut off values are ±2 as indicated by the *horizontal lines*. Values of iHS above +2 and below −2 are significant. Each *dot* represents a different SNP. The three populations are as follows: Asian (A) = *diamonds*, European (C) = *squares*, and, African Yoruba (Y) = *triangles*. The *black asterisk* represents those SNPs that cause a nonsynonymous substitution in the coding DNA sequence

RAD51). Evidence for lineage-specific positive selection is present in 6 out of 15 proteins. There are 9 proteins that display patterns indicative of site-specific positive selection (these numbers are inclusive of 3 proteins that have positive selection at both site and lineage level).

Evidence for site-specific positive selection was found in 9 proteins in the pathway, 6 of which have evidence for site-specific positive selection, these are: FANCC, BRCA2, NBS1, FANCB, FANCF, and FANCM. The sites identified as positively selected in these proteins are often in regions of functional importance. For example, from the analysis of the distribution of positively selected sites in BRCA2 it is shown that the BRCT repeats are highly conserved, there is only 1 amino acid from the list of 63 positively selected residues located within the BRCT repeats. Overall this 1 positively selected residue constitutes 0.2% of the overall length of the repeats alone, and therefore is not a significant proportion of sites. The observed pattern of conservation in the BRCA2 repeat regions is not surprising when it is considered that these repeats are known to interact directly with RAD51 which is itself under strong purifying selection. Further biochemical studies of RAD51 mutants would help to elucidate this possibility. An amino acid residue is predicted to be under positive selection in BRCA2 in the region that interacts with FANCD2, and vice versa, illustrating the importance of co-evolution amongst interacting partners in this protein network. Again, as with the BRCA1 results, a significant proportion of the sites identified as positively selected in BRCA2 are direct neighbors of sites mutated in breast cancer. Sites identified in NBS1 as under positive selection are neighboring sites which, when mutated, prevent ATM-dependent phosphorylation. The results for the ATM protein identify sites involved in proto-oncogene interaction, and are the boundaries of regions missing in two types of leukemia. The findings presented here indicate that there has been a selective advantage for those individuals in mammalian evolution with these particular mutations, most likely for reduced cancer risk.

Comparison of evolutionary rate heterogeneity in the human and mouse lineages has revealed a significant difference between these mammals. In the case of the human lineage, these results show that FANCD2 and BRCA1 have been under positive selective pressure in this lineage only while all other species are under purifying selection. The sites under positive selection in FANCD2 are located in a region of the protein involved in phosphorylation, a critical step in the cascade of events in the FA/BRCA damage response pathway and also are found in the region that directly interacts with BRCA2. Mouse shows evidence of lineage-specific positive selection in 4 proteins. Two of these proteins are involved in phosphorylation and activation of each other (ATM and CHK2) while 2 of these form part of the Fanconi anemia core complex (FANCE and FANCL). The two proteins showing evidence of positive selection in the human lineage form a complex together (FANCD2 and BRCA1). Overall this pattern of selection on proteins that interact or heavily rely on one another for function suggests that there is systematic positive selection in these lineages as has previously been described for metabolic pathways (Ardawatia and Liberles 2007). An approach that takes into account the interacting neighborhood of proteins in functional pathways is essential to understanding the evolution of protein function.

Interestingly, while in vivo and in vitro studies of FA mutant mice display sensitivity to DNA crosslinking agents and to ionizing radiation (Whitney et al. 1996), it is of particular interest that the phenotype of the FA mouse is not as severe as humans. The lifespan of the FA mouse is normal, with no reported anemia or increased tumour incidence. Haematopoietic cells of FA mice are more susceptible to apoptosis and anemia, and these mice have low body weight (Whitney et al. 1996). A molecular reason for the discrepancy, between the human disease and its mouse model has not yet been elucidated, but is significant for understanding and modeling of this disease and others (Rangarajan and Weinberg 2003). Overall, the lineage-specific evolution identified here may help to explain why the mouse model for FA phenotype is not as severe as the human.

An important finding of the results presented here is that the level of lineage-specific positive selection detected varies between species, four proteins in mouse compared to two in human. While these numbers are indeed small they are suggestive of a difference between the human and mouse, thus raising an important issue of variation in mutational rate and damage recovery in these species. The metabolic hypothesis states that animals with a greater metabolic rate generate larger amounts of internal mutagens as by-products of metabolism, i.e., reactive oxygen species, and that this causes a greater amount of DNA damage (Martin and Palumbi 1993). The basal metabolic rate per gram of body weight is seven times higher in mouse than in human (Demetrius 2005). It is possible therefore that increased selective pressure may be exerted on species with higher metabolic rates to improve the efficiency of their DNA repair pathways. The rate of metabolic conversion of pro-carcinogen to carcinogen varies hugely between the species with mouse displaying 25-fold higher binding of carcinogen to DNA than human (Demetrius 2005). Genome replication for reproduction also raises the possibility of mutation. The mutation rate for a species is therefore influenced both by the rate of mutation per replication and the number of replicates made per unit time. The germ line generation time hypothesis states that those species with shorter generation times, such as mouse, have faster rates of molecular evolution due to a greater amount of replications per unit time as compared to species with longer generation times such as humans (Bromham et al. 1996). This combined with the fewer number of cell generations it takes to produce a mouse ova compared to a human ova means that mouse has the potential for a much greater number of mutations per generation than human and therefore could be under increased selective pressure for genomic integrity. It has been established that variation in efficiency of error correction between individuals in a population is caused by variations in the coding DNA sequences for DNA repair machinery (Woodruff et al. 1984), therefore these repair processes are visible to—and can be acted upon by—selection. Further comparative analyses using a large number of species and proteins would be necessary to determine with statistical confidence whether the observed trend in this dataset is real.

The results presented here for BRCA1, show that the sites under positive selection are located at or adjacent to sites that are polymorphic in breast cancer. The results for BRCA1 are in strong agreement with previous analyses (Pavlicek et al. 2004; Wakefield et al. 2005). For example, using data from the Breast cancer information core (BIC) database it was predicted that missense mutations conferring the highest predisposition to breast and ovarian cancers are located in the evolutionarily conserved regions, phosphorylated residues and especially in specific protein-binding domains (Pavlicek et al. 2004; Shen and Vadgama 1999; Szabo et al. 2000). Analyses of BRCA1 have shown evidence for positive selective pressures in the region responsible for interacting with RAD51, as presented here, suggesting that BRCA1 evolution is driven by the binding efficiency between these interacting partners for the purpose of DNA repair (Fleming et al. 2003; Wakefield et al. 2005). Studies of BRCA1 in modern human populations have also shown that there is evidence for positive selection against cancer causing mutations even though these cancers are classically late onset and therefore intuitively

may not be classified as visible to selection (Pavard and Metcalf 2007).

There is also evidence for an ongoing selective sweep on this pathway in modern humans as shown by 9/14 proteins displaying unusually long haplotypes of low diversity and significant |iHS|. Previous analyses of the FANCC and RAD51 alleles were strongly indicative of a positive selective sweep in recent human history and the analyses carried out here are in agreement with this earlier study (Wang and Moyzis 2007). Other proteins displaying evidence of ongoing selective sweep from the analysis presented here include: ATM, CHK2, BRCA1, BRCA2, FANCA, FANCL, NBS1, and RAD51. Four of these show evidence of replacement substitutions in the coding DNA sequence (ATM, BRCA1, BRCA2 and FANCA). Using the available data for modern human populations, there is evidence of an ongoing selective sweep in both "Y" and "C" modern human populations for BRCA1. It has been suggested that older alleles present at high frequencies may be linked to a higher mean age of onset and/or lower variance and consequently may be present in the population due to a reduction in negative selection acting upon them (Pavard and Metcalf 2007). On comparison of the "Y" and "C" populations, there is evidence of an ongoing selective sweep occurring independently in both populations for BRCA1 but the allele selected differs. For the "Y" population, there is one SNP with a significantly high negative iHS, i.e., selection on a derived allele, while in the "C" population there are 11 SNPs with high positive iHS, i.e., selection on an ancestral allele.

Not only do these results show that there has been selection acting on this DNA damage response pathway, but also the specific sites under adaptive evolution are neighbors of sites polymorphic in breast cancer and flank regions missing in specific leukemias. This study provides evidence that there is selective pressure on genomic integrity/stability mechanisms in vertebrates, a possible reason for this being reduced cancer risk in a situation where rapid developmental changes have occurred.

Of course as with any biological system there are many functions associated with a given set of proteins, and so positive selection in a cancer gene may be the result of selective pressures caused by infection with a pathogen or other function such as cell apoptosis and an increased risk of cancer may be a negative outcome of this association (da Fonseca et al. 2010). However, with a dataset of 15 proteins, detecting signals of positive selection so widely and in some cases between pairs of interacting proteins—whose sole function is in DNA repair—is strongly suggestive of positive selection acting on this DNA repair mechanism in mammals and also shows ongoing selective pressure on this system in modern human populations.

Acknowledgments I would like to thank Dr. James O. McInerney at NUIM for helpful comments on an early draft of this manuscript and for computation time. ICHEC (Irish Centre for High End Computing) for granting a user license to complete this work, and the OVPR (Office for the Vice President for Research) at DCU for part funding this research through the career start award. I would like to thank Mr Stanley Nolan, Dr James Murphy, Dr Paul McGettigan, and Ms Eithne Smith for preliminary analyses. Dr Mary J. O'Connell is funded by Science Foundation Ireland, RFP award EOB2673.

References

Anisimova M, Yang Z (2007) Multiple hypothesis testing to detect lineages under positive selection that affects only a few sites. Mol Biol Evol 24:1219–1228

Ardawatia H, Liberles DA (2007) A systematic analysis of lineage-specific evolution in metabolic pathways. Gene 387:67–74

Bagby GC Jr (2003) Genetic basis of Fanconi anemia. Curr Opin Hematol 10:68–76

Berglund AC, Wallner B, Elofsson A, Liberles DA (2005) Tertiary windowing to detect positive diversifying selection. J Mol Evol 60:499–504

Blom E, van de Vrugt HJ, de Vries Y, de Winter JP, Arwert F, Joenje H (2004) Multiple TPR motifs characterize the Fanconi anemia FANCG protein. DNA Repair (Amst) 3:77–84

Bromham L, Rambaut A, Harvey PH (1996) Determinants of rate variation in mammalian DNA sequence evolution. J Mol Evol 43:610–621

Burk-Herrick A, Scally M, Amrine-Madsen H, Stanhope MJ, Springer MS (2006) Natural selection and mammalian BRCA1 sequences: elucidating functionally important sites relevant to breast cancer susceptibility in humans. Mamm Genome 17:257–270

Cheng F, Chen W, Richards E, Deng L, Zeng C (2009) SNP@Evolution: a hierarchical database of positive selection on the human genome. BMC Evol Biol 9:221

Crespi B, Summers K (2005) Evolutionary biology of cancer. Trends Ecol Evol 20:545–552

Crespi BJ, Summers K (2006) Positive selection in the evolution of cancer. Biol Rev Camb Philos Soc 81:407–424

D'Andrea AD, Grompe M (2003) The Fanconi anaemia/BRCA pathway. Nat Rev Cancer 3:23–34

da Fonseca RR, Kosiol C, Vinar T, Siepel A, Nielsen R (2010) Positive selection on apoptosis related genes. FEBS Lett 584:469–476

Davids W, Gamieldien J, Liberles DA, Hide W (2002) Positive selection scanning reveals decoupling of enzymatic activities of carbamoyl phosphate synthetase in *Helicobacter pylori*. J Mol Evol 54:458–464

Demetrius L (2005) Of mice and men. When it comes to studying ageing and the means to slow it down, mice are not just small humans. EMBO Rep 6 Spec No: S39–S44

Fleming MA, Potter JD, Ramirez CJ, Ostrander GK, Ostrander EA (2003) Understanding missense mutations in the BRCA1 gene: an evolutionary approach. Proc Natl Acad Sci USA 100: 1151–1156

Freie BW, Ciccone SL, Li X, Plett PA, Orschell CM, Srour EF, Hanenberg H, Schindler D, Lee SH, Clapp DW (2004) A role for the Fanconi anemia C protein in maintaining the DNA damage-induced G2 checkpoint. J Biol Chem 279:50986–50993

Garcia-Higuera I, Taniguchi T, Ganesan S, Meyn MS, Timmers C, Hejna J, Grompe M, D'Andrea AD (2001) Interaction of the Fanconi anemia proteins and BRCA1 in a common pathway. Mol Cell 7:249–262

Gavish H, dos Santos CC, Buchwald M (1993) A Leu554-to-Pro substitution completely abolishes the functional complementing activity of the Fanconi anemia (FACC) protein. Hum Mol Genet 2:123–126

Goldman N, Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol 11:725

Graham J (1992) Cancer Selection: The NEW Theory of Evolution. Aculeus Press, Lexington, Virginia

Hussain S, Witt E, Huber PA, Medhurst AL, Ashworth A, Mathew CG (2003) Direct interaction of the Fanconi anaemia protein FANCG with BRCA2/FANCD1. Hum Mol Genet 12:2503–2510

Hussain S et al (2004) Direct interaction of FANCD2 with BRCA2 in DNA damage response pathways. Hum Mol Genet 13: 1241–1248

Huttley GA, Easteal S, Southey MC, Tesoriero A, Giles GG, McCredie MR, Hopper JL, Venter DJ (2000) Adaptive evolution of the tumour suppressor BRCA1 in humans and chimpanzees. Australian Breast Cancer Family Study. Nat Genet 25:410–413

Leroi AM, Koufopanou V, Burt A (2003) Cancer selection. Nat Rev Cancer 3:226–231

Martin AP, Palumbi SR (1993) Body size, metabolic rate, generation time, and the molecular clock. Proc Natl Acad Sci USA 90:4087–4091

Murphy WJ et al (2001) Resolution of the early placental mammal radiation using Bayesian phylogenetics. Science 294:2348–2351

Nakanishi K, Yang YG, Pierce AJ, Taniguchi T, Digweed M, D'Andrea AD, Wang ZQ, Jasin M (2005) Human Fanconi anemia monoubiquitination pathway promotes homologous DNA repair. Proc Natl Acad Sci USA 102:1110–1115

O'Connell MJ, McInerney JO (2005) Adaptive evolution of the human fatty acid synthase gene: support for the cancer selection and fat utilization hypotheses? Gene 360:151–159

O'Driscoll M, Jeggo PA (2006) The role of double-strand break repair—insights from human genetics. Nat Rev Genet 7:45–54

Offit K et al (2003) Shared genetic susceptibility to breast cancer, brain tumors, and Fanconi anemia. J Natl Cancer Inst 95: 1548–1551

Pavard S, Metcalf CJ (2007) Negative selection on BRCA1 susceptibility alleles sheds light on the population genetics of late-onset diseases and aging theory. PLoS One 2:e1206

Pavlicek A, Noskov VN, Kouprina N, Barrett JC, Jurka J, Larionov V (2004) Evolution of the tumor suppressor BRCA1 locus in primates: implications for cancer predisposition. Hum Mol Genet 13:2737–2751

Rangarajan A, Weinberg RA (2003) Opinion: Comparative biology of mouse versus human cells: modelling human cancer in mice. Nat Rev Cancer 3:952–959

Schmid K, Yang Z (2008) The trouble with sliding windows and the selective pressure in BRCA1. PLoS One 3:e3746

Woodruff RC, Thompson JJJ, Seeger MA, Spivey WE (1984) Variation in spontaneous mutation and repair in natural population lines of *Drosophila melanogaster*. Heredity 53:223–234

Shen D, Vadgama JV (1999) BRCA1 and BRCA2 gene mutation analysis: visit to the Breast Cancer Information Core (BIC). Oncol Res 11:63–69

Strathdee CA, Gavish H, Shannon WR, Buchwald M (1992) Cloning of cDNAs for Fanconi's anaemia by functional complementation. Nature 358:434

Suzuki Y, Nei M (2002) Simulation study of the reliability and robustness of the statistical methods for detecting positive selection at single amino acid sites. Mol Biol Evol 19:1865–1869

Szabo C, Masiello A, Ryan JF, Brody LC (2000) The breast cancer information core: database design, structure, and scope. Hum Mutat 16:123–131

Taniguchi T, Garcia-Higuera I, Xu B, Andreassen PR, Gregory RC, Kim ST, Lane WS, Kastan MB, D'Andrea AD (2002) Convergence of the Fanconi anemia and ataxia telangiectasia signaling pathways. Cell 109:459–472

Taniguchi T, Tischkowitz M, Ameziane N, Hodgson SV, Mathew CG, Joenje H, Mok SC, D'Andrea AD (2003) Disruption of the Fanconi anemia-BRCA pathway in cisplatin-sensitive ovarian tumors. Nat Med 9:568–574

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22: 4673–4680

Verlander PC, Lin JD, Udono MU, Zhang Q, Gibson RA, Mathew CG, Auerbach AD (1994) Mutation analysis of the Fanconi anemia gene FACC. Am J Hum Genet 54:595–601

Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006) A map of recent positive selection in the human genome. PLoS Biol 4:e72

Wakefield MJ, Maxwell P, Huttley GA (2005) Vestige: maximum likelihood phylogenetic footprinting. BMC Bioinformatics 6:130

Wang W (2007) Emergence of a DNA-damage response network consisting of Fanconi anaemia and BRCA proteins. Nat Rev Genet 8:735–748

Wang ET, Moyzis RK (2007) Genetic evidence for ongoing balanced selection at human DNA repair genes ERCC8, FANCC, and RAD51C. Mutat Res 616:165–174

Wang X, Andreassen PR, D'Andrea AD (2004) Functional interaction of monoubiquitinated FANCD2 and BRCA2/FANCD1 in chromatin. Mol Cell Biol 24:5850–5862

Whitney MA et al (1996) Germ cell defects and hematopoietic hypersensitivity to gamma-interferon in mice with a targeted disruption of the Fanconi anemia C gene. Blood 88:49–58

Xia B et al (2007) Fanconi anemia is associated with a defect in the BRCA2 partner PALB2. Nat Genet 39:159–161

Yang Z (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol 15:568–573

Yang Z, Nielsen R (2002) Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol Biol Evol 19:908–917

Yang Z, Nielsen R, Goldman N, Pedersen AM (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155:431–449

Yang Z, Wong WS, Nielsen R (2005) Bayes empirical Bayes inference of amino acid sites under positive selection. Mol Biol Evol 22:1107–1118

Zhang J (2004) Frequent false detection of positive selection by the likelihood method with branch-site models. Mol Biol Evol 21:1332–1339

Zhang J, Nielsen R, Yang Z (2005) Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol 22:2472–2479

Zimmer C (2007) Evolved for cancer? Sci Am 296:68–74, 75A