UNIVERSITY OF
GLOUCESTERSHIRE

This is a peer-reviewed, final published version of the following document and is licensed under All Rights Reserved license:

Brunette, Gregory J, Jamalruddin, Mohd A, Baldock, Robert A ORCID logoORCID: https://orcid.org/0000-0002-4649-2966, Clark, Nathan L and Bernstein, Kara A (2019) Evolution-based screening enables genome-wide prioritization and discovery of DNA repair genes. Proceedings of the National Academy of Sciences, 116 (39). pp. 19593-19599. doi:10.1073/pnas.1906559116

Official URL: http://dx.doi.org/10.1073/pnas.1906559116
DOI: http://dx.doi.org/10.1073/pnas.1906559116
EPrint URI: https://eprints.glos.ac.uk/id/eprint/9776

Disclaimer

The University of Gloucestershire has obtained warranties from all depositors as to their title in the material deposited and as to their right to deposit such material.

The University of Gloucestershire makes no representation or warranties of commercial utility, title, or fitness for a particular purpose or any other warranty, express or implied in respect of any material deposited.

The University of Gloucestershire makes no representation that the use of the materials will not infringe any patent, copyright, trademark or other property or proprietary rights.

The University of Gloucestershire accepts no liability for any infringement of intellectual property rights in any material deposited but will remove such material from public view pending investigation in the event of an allegation of any such infringement.

PLEASE SCROLL DOWN FOR TEXT.

Evolution-based screening enables genome-wide prioritization and discovery of DNA repair genes

Gregory J. Brunette[a], Mohd A. Jamalruddin[a], Robert A. Baldock[b], Nathan L. Clark[c], and Kara A. Bernstein[a,1]

[a]Department of Microbiology and Molecular Genetics, University of Pittsburgh School of Medicine, Pittsburgh, PA 15213; [b]School of Sport, Health and Social Sciences, Solent University, Southampton SO14 0YN, United Kingdom; and [c]Department of Computational and Systems Biology, University of Pittsburgh School of Medicine, Pittsburgh, PA 15213

Edited by Philip C. Hanawalt, Stanford University, Stanford, CA, and approved August 16, 2019 (received for review April 16, 2019)

DNA repair is critical for genome stability and is maintained through conserved pathways. Traditional genome-wide mammalian screens are both expensive and laborious. However, computational approaches circumvent these limitations and are a powerful tool to identify new DNA repair factors. By analyzing the evolutionary relationships between genes in the major DNA repair pathways, we uncovered functional relationships between individual genes and identified partners. Here we ranked 17,487 mammalian genes for coevolution with 6 distinct DNA repair pathways. Direct comparison to genetic screens for homologous recombination or Fanconi anemia factors indicates that our evolution-based screen is comparable, if not superior, to traditional screening approaches. Demonstrating the utility of our strategy, we identify a role for the DNA damage-induced apoptosis suppressor (*DDIAS*) gene in double-strand break repair based on its coevolution with homologous recombination. DDIAS knockdown results in DNA double-strand breaks, indicated by ATM kinase activation and 53BP1 foci induction. Additionally, DDIAS-depleted cells are deficient for homologous recombination. Our results reveal that evolutionary analysis is a powerful tool to uncover novel factors and functional relationships in DNA repair.

DDIAS | DNA repair | evolution | genome integrity | homologous recombination

DNA repair encompasses a complex network of metabolic and regulatory steps that coordinate with each other to maintain genome stability. Elucidating this complex regulatory network is critical for our understanding of genetic diversity and diseases that arise when DNA repair fidelity is compromised, such as cancer. However, defining the role of a gene in a DNA repair process requires considerable time and experimental effort. Informatics-based approaches bypass the time and resource limitations of experimental screens, and such bioinformatic strategies have correctly inferred relationships between genes through comparative analysis of their gene expression (1) or molecular evolution. Analyses based on molecular evolution work from the premise that functionally related genes often evolve at correlated rates when studied across a variety of species, and this rate correlation has been termed evolutionary rate covariation (ERC) (2). ERC is based on the hypothesis that cofunctioning proteins experience shared changes in selective pressure in different species, which would lead to correlated shifts in amino acid substitution rates between cofunctional proteins. The measured correlation coefficient between 2 proteins' branch-specific rates is referred to as their ERC value. Genes involved in the same biological process often exhibit elevated ERC with each other, comprising a statistically coevolving group (3, 4). Querying additional genes for correlation with these groups has proven effective in uncovering novel roles for genes in previously unrelated processes in model organisms such as worms, flies, and yeast (4–6). More broadly, ERC may be investigated between processes, revealing novel interdependencies at the systems level, and can be expanded to analyze crosstalk between networks (7). Using this approach, we analyzed the evolutionary relationship among genes in specific DNA repair pathways and the rest of the genome to identify relationships and DNA repair factors.

The genome experiences constant damage resulting from both endogenous and environmental sources. Depending on the damage source and cellular context, DNA may incur different lesions, including bulky adducts, single-strand breaks (SSBs), double-strand breaks (DSBs), and interstrand and intrastrand crosslinks (Fig. 1*A*). Consequently, cells have an assortment of DNA repair pathways which recognize specific damage substrates (Fig. 1*A*) (8). For example, DSBs can be resolved by nonhomologous end-joining (NHEJ), which religates the DNA ends (9), or by homologous recombination (HR), which uses the sister chromatid or homolog for templated repair (10). The nucleotide excision repair (NER) pathway processes bulky adducts, such as UV-induced cyclobutane dimers and pyrimidine–pyrimidone (6–4) photoproducts (11). Base excision repair (BER) resolves SSBs and base damage (12). DNA mismatches and replication slippage are remediated by mismatch repair (MMR) (13), and interstrand crosslinks are remediated by the Fanconi anemia (FA) pathway (14). Due to the importance of genomic integrity, these central DNA repair mechanisms are highly conserved (8).

While unrepaired DNA damage can have negative consequences such as genomic instability or cell death, genomic changes also enable genetic diversity. For example, V(D)J recombination enables adaptive immunity to new antigens. Similarly, meiosis, during which a programmed DSB allows the exchange of genetic information between parental homologs, contributes genetic variation

Significance

Genome stability is maintained through conserved DNA repair pathways. By analyzing evolutionary relationships between genes that mediate DNA repair, we show that functional relationships between DNA repair genes are reflected in the variation of their evolutionary rates. Here we ranked mammalian genes for coevolution with 6 DNA repair pathways. Direct comparison to genetic screens for DNA repair factors indicates that our evolution-based screen is comparable, if not superior, to traditional screening approaches. As a proof of principle, we identified a role for the gene *DDIAS* in double-strand break (DSB) repair. DDIAS depletion results in DSB accumulation, increased checkpoint signaling, and defective homologous recombination. Our results reveal that evolutionary analysis is a powerful tool to uncover DNA repair factors.

Author contributions: G.J.B., R.A.B., N.L.C., and K.A.B. designed research; G.J.B., M.A.J., and N.L.C. performed research; N.L.C. contributed new reagents/analytic tools; G.J.B., N.L.C., and K.A.B. analyzed data; and G.J.B., N.L.C., and K.A.B. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

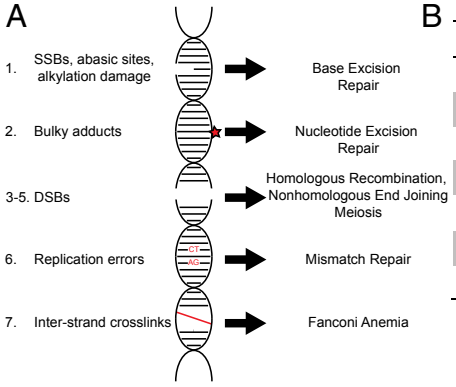
Published under the PNAS license.

[1]To whom correspondence may be addressed. Email: karab@pitt.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1906559116/-/DCSupplemental.

First published September 9, 2019.

Downloaded by guest on June 3, 2021



B

Pathway	*p*	*n*
1. Base Excision Repair	$p = 0.019$	$n = 23$
2. Nucleotide Excision Repair	$p = 0.910$	$n = 30$
3. Homologous Recombination	$p < 0.001$	$n = 42$
4. Meiosis	$p < 0.001$	$n = 58$
5. Nonhomologous End Joining	$p < 0.001$	$n = 21$
6. Mismatch Repair	$p = 0.003$	$n = 7$
7. Fanconi Anemia	$p < 0.001$	$n = 23$

Fig. 1. Genetic constituents of most DNA repair pathways exhibit evolutionary rate covariation. (*A*) Seven mammalian DNA repair pathways are responsible for repairing DNA lesions. In pathway 1, SSBs (indicated by a missing nucleotide), abasic sites, and alkylation damage are repaired by the BER pathway. In pathway 2, bulky adducts (indicated by the red star) are repaired by NER. In pathways 3 to 5, DSBs (indicated by the 2 nicks) are repaired using HR, NHEJ, and meiosis. In pathway 6, replication errors (indicated by the red nucleotide mismatches) are repaired using MMR. In pathway 7, interstrand crosslinks (indicated by the red line) are repaired by the FA pathway. (*B*) The genes mediating the 7 mammalian DNA repair pathways were tested for elevated ERC. Statistical significance was determined by permutation test, where the *P* value is the computed probability of the observed mean ERC or greater from 1,000 equally sized groups of randomly sampled genes. *n* is the number of genes in the indicated group (the gene list is found in *SI Appendix*, Table S1).

in the resulting progeny (15). These mechanisms, as well as the persistence of mutations which evade accurate DNA repair, allow genetic variation and provide the basis for evolution. Therefore, DNA repair processes are central to both the preservation of genome integrity and the emergence of genetic diversity. Through computational methods such as ERC, we can exploit genetic divergence over evolutionary time to identify functional relationships between genes and pathways. Scanning for new genes exhibiting elevated ERC values with a specific pathway has led to the discovery of new genes affecting those functions (4, 7, 16). Because ERC has not yet been used to discover new DNA repair proteins, we demonstrate its potential through comparison to other screening methods.

Here we provide comprehensive evolutionary analysis of the genes involved in the 7 distinct DNA repair pathways including BER, NER, HR, meiosis, NHEJ, MMR, and FA. We investigate individual pathways for elevated ERC signals between their constituent genes and screen the genome for DNA repair factors. We find that genes in unique mammalian DNA repair processes exhibit a coevolutionary signature with each other. Finally, we find a role for the gene C11orf82, DNA damage-induced apoptosis suppressor (*DDIAS*), in DSB repair based on its shared evolutionary signature with HR.

Results

To determine whether genes in different DNA repair pathways evolve at covarying rates, we analyzed genes in 7 DNA repair pathways for ERC. Using published reviews and input from expert colleagues, we compiled lists for the genes comprising each pathway (BER (12), NER (11), HR (10), meiosis (15), NHEJ (9), MMR (13), and FA (14)) and tested each gene group for ERC (Fig. 1*B* and *SI Appendix*, Table S1). Note that these lists are not exhaustive, and we limited the analysis to the central factors indicated in these reviews as well from consultation of experts in each DNA repair pathway. We find that genes in 6 of the distinct DNA repair pathways exhibit significantly elevated mean ERC (Fig. 1*B*; excluding NER), suggesting that the constituents of these DNA repair pathway are subject to similar evolutionary pressures.

Most DNA repair pathways exhibit distinguishable DNA processing steps. Each repair step involves distinct proteins in processing the DNA substrate. For example, HR, which uses a homologous template for repair, can be broken down into 5 distinct steps, including 1) DSB recognition/resection, 2) RAD51 filament mediation, 3) RAD51 filament disruption, 4) joint molecule disruption/DNA heteroduplex extension, and 5) resolution and dissolution (modified from ref. 10 [*SI Appendix*, Fig. S1*A*]). To determine whether genes that function in the same DNA processing step coevolve, we tested the ERC values between proteins that are known to function within each step of HR and observe significantly elevated ERC at each step (*SI Appendix*, Fig. S1*B*). We then analyzed HR genes globally for ERC and performed hierarchical clustering to group genes with similar ERC profiles (Fig. 2). Similar analysis was done for the genes in the remaining DNA repair pathways we studied (BER, NER, meiosis, MMR, FA, and NHEJ; *SI Appendix*, Figs. S2–S7).

After demonstrating that DNA repair pathways exhibit ERC, we implemented this signature to identify genes that may be functioning during DNA repair. To do this, we ranked all genes in the mammalian ERC dataset by their mean ERC with each DNA repair pathway in descending order (Dataset S2). Additionally, we performed genome-wide rankings for ERC with each gene included in our study (Dataset S3; $n = 137$ genes). We divided the ranked lists into 20 bins (5% [874 genes] per bin; Fig. 3*A* and Dataset S2) and first asked if ranking genes by their mean ERC with HR enriches for known HR factors. After plotting the frequency of known HR genes in each bin of the ranked list (*Materials and Methods*), we scanned the distribution of HR gene positions for enriched ranges (Fig. 3*A*). The window of enrichment for elevated HR genes was determined using a scan statistic followed by permutation test as described in ref. 17 (*SI Appendix*, *SI Methods*). We find that the majority of the known HR genes (27 of the 39 HR genes considered) fall in the top 20% of this ranking ($P < 0.001$ [bins 1 to 4]; Fig. 3*A* and Dataset S2). Since the top 20% of ERC-ranked genes was enriched for known HR factors, we performed Gene Ontology (GO) term analysis on this set of genes ($n = 3{,}497$ genes; corresponding to the first 4 bins). We find significant overrepresentation of recombination-related GO terms as well as additional DNA metabolism terms (Table 1; $p_{adj} < 0.05$). Therefore, genes found in the top 20% (3,497 genes; Dataset S2) are enriched for known HR factors and potentially novel HR genes as well as genes that exhibit cofunctionality with HR.

Next, we determined how our evolutionary approach compared to genetic screens for HR factors. Mammalian HR screens routinely use the direct repeat-GFP reporter (drGFP; ref. 18). Here a DSB is induced at a restriction cut site within a nonfunctional GFP copy. HR repair of the DSB using a downstream

Downloaded by guest on June 3, 2021

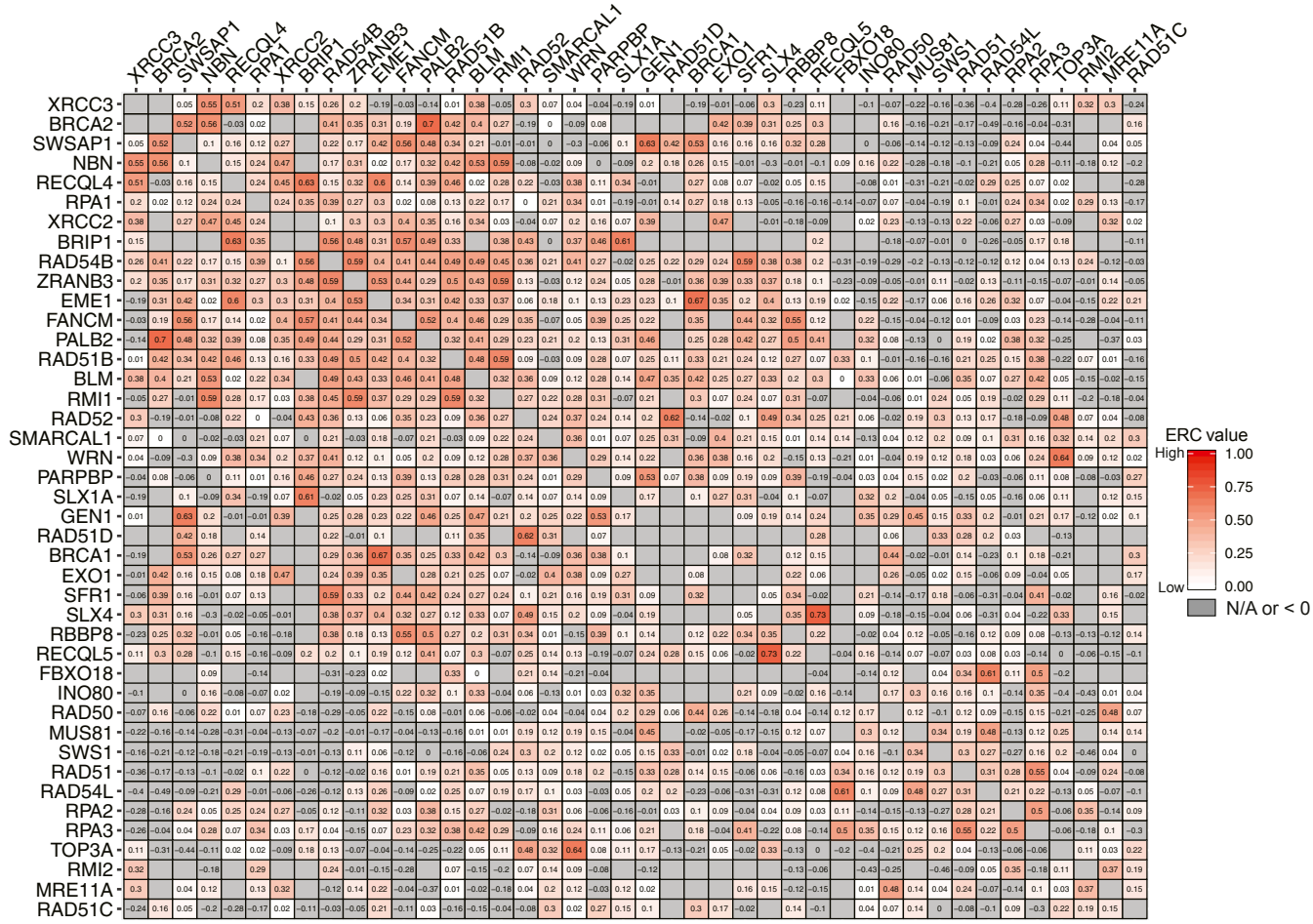


Fig. 2. Evolutionary rate covariation between HR genes. The ERC values between HR gene pairs in each DNA processing step were calculated and plotted using a heat map (ranging from 0 [no covariation] to 1 [positive covariation]). Genes are hierarchically clustered.

GFP template results in GFP expression (18). Paired with a siRNA library against the coding genome (n = 21,040), this assay has been used to identify novel HR factors (19). Analyzing this published study's gene ranking (1,052 genes in 20 bins, ranked by relative drGFP), we find a window of enrichment for HR genes in the top 5% of this ranking ($P < 0.001$ [bin 1]; Fig. 3*B*). We note the top 20% of this drGFP-based ranking contains 17 known HR factors, versus the 27 HR factors we find in the top 20% of the ERC-based ranking (Fig. 3*A*).

We next asked how ERC ranking would compare to other screens for a different DNA repair pathway, FA. Using the same approach described above, we ranked the ERC dataset (n =17,487 genes) by mean ERC with FA (Fig. 3*C* and Dataset S2). We find a window enriched for known FA genes in the top 10%, containing 12 known FA factors ($P < 0.001$ [bins 1 and 2]; Fig. 3*C*). We next compared our ERC analysis to a genome-wide shRNA screen (n = 49,281 shRNAs) to identify genes in the FA pathway (20). In this screen, shRNA treated U2OS cells were screened for mitomycin C (MMC) sensitivity (20). MMC is a DNA crosslinking agent that is routinely used in FA diagnosis. Analyzing this study's shRNA targets, we ranked targets based on MMC sensitivity and divided them into 20 bins (2,464 targets per bin). We find a window enriched for known FA genes in the top 25% of this ranking, which contains 16 of the 20 known FA genes present in both screens ($P < 0.001$ [bins 1 to 4]; Fig. 3*D*). We note that the top 10% of this gene ranking contains 5 known FA factors, versus the 12 FA factors we find in the top 10% of the ERC-based gene ranking (Fig. 3*C*).

Since ranking the genome by ERC with HR and FA enriches for known HR and FA factors, we next analyzed the top 5% of HR-enriched genes for HR factors as a proof of principle. *DDIAS* ranks 63rd out of 17,487 genes in the ERC-based ranking with HR (Dataset S2). In addition to its coevolution with HR, we noted reports that *DDIAS* is overexpressed in colorectal and lung cancer cell lines and tissues, and its depletion results in DNA breaks in a nonsmall cell lung carcinoma cell line (A549) (21). DDIAS also contains an OB-fold domain similar to the ssDNA-binding domain of RPA1 (21). We next analyzed ERC between *DDIAS* and the DNA repair pathways in our study (Fig. 4*A*). Of these pathways, we find that *DDIAS* exhibits significantly elevated ERC with the DSB repair pathways HR and NHEJ and the highest mean ERC with HR (Fig. 4*A*; $P < 0.05$ cutoff). Given both the strong ERC between *DDIAS* and HR genes as well as its links to cancer, we sought to experimentally determine whether *DDIAS* has a role in promoting HR. We first knocked down *DDIAS* by siRNA (Fig. 4*B*).

To determine if DDIAS depletion results in induction of the DSB checkpoint response, we assayed DDIAS-depleted U2OS cells for ATM and ATR activation (Fig. 4*B*). Upon DSB formation, the ATM kinase is phosphorylated at S1981, which is required for ATM stabilization at DSBs (22). In addition to ATM phosphorylation, upon replicative damage, ATR kinase phosphorylates the CHK1 protein at S345. Consistent with increased DSBs, we observe increased ATM phosphorylation (pATM) upon DDIAS knockdown (Fig. 4*B*). We next asked if DDIAS depletion would result in increased CHK1

Downloaded by guest on June 3, 2021

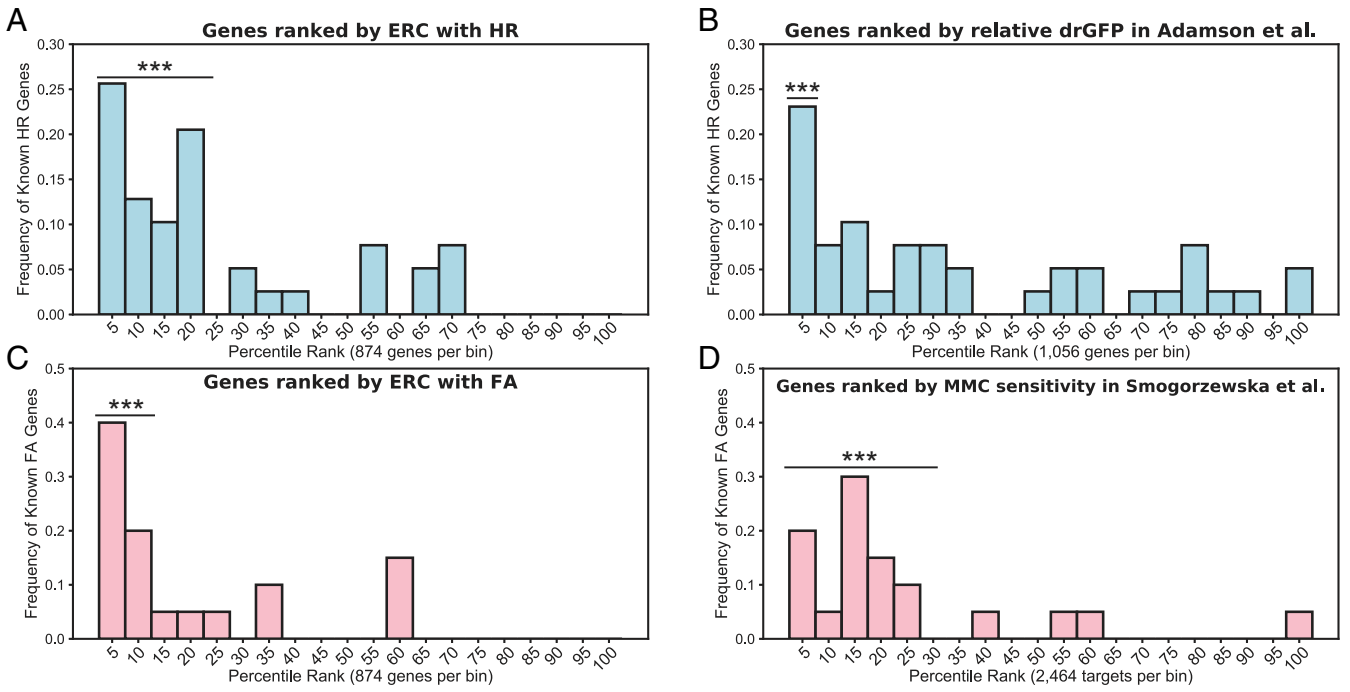


Fig. 3. Genome-wide ranking of 17,487 genes for coevolution with HR and FA enriches for known HR and FA factors when compared to functional screens. (*A*) Genes were ranked by mean ERC with HR. Each bin size contains 874 genes (5% of genes analyzed). The number of HR genes in each bin is plotted relative to the total number of HR genes ($n = 39$) (blue bar). A scan statistic was used to find an enriched window of known HR genes, and significance was determined by permutation test (1,000 nulls; $***P < 0.001$, 0 to 20%). Note that only HR genes present in both screens (the ERC dataset and the screen described in *B*) were included. (*B*) Genes were ranked by relative drGFP in a published siRNA screen ($n = 21,121$) for HR factors (19). Each bin contains 1,056 genes (5% of genes analyzed). The number of HR genes in each bin is plotted relative to the total number of HR genes ($n = 39$) (blue bar). A scan statistic was used to find an enriched window of known HR genes, and significance was determined by permutation test (1,000 nulls; $***P < 0.001$, 0 to 5%). (*C*) Genes were ranked by mean ERC with FA. Each bin size contains 874 genes (5% of genes analyzed). The number of FA genes in each bin is plotted relative to the total number of FA genes ($n = 20$) (pink bar). A scan statistic was used to find an enriched window of known FA genes, and significance was determined by permutation test (1,000 nulls; $***P < 0.001$, 0 to 10%). Note that only FA genes present in both screens (the ERC dataset and the screen described in *D*) were included. (*D*) Genes were ranked by MMC sensitivity in a published shRNA screen ($n = 32,293$) for FA factors (20). Each bin contains 2,464 targets (5% of targets analyzed). The number of FA genes in each bin is plotted relative to the total number of FA genes ($n = 20$) (pink bar). A scan statistic was used to find an enriched window of known FA genes, and significance was determined by permutation test (1,000 nulls; $***P < 0.001$, 0 to 25%).

phosphorylation (pCHK1). In contrast to pATM, we do not observe increased pCHK1 upon DDIAS depletion (Fig. 4*B*). We next analyzed recruitment of the DSB repair protein, 53BP1, into repair foci by fluorescent microscopy upon depletion of DDIAS in U2OS cells. Consistent with a function for DDIAS in DSB repair, we observe an increase in 53BP1 foci upon DDIAS depletion (Fig. 4*C*).

Next, we asked if DDIAS directly impacts HR. We used the sister-chromatid recombination (SCR) reporter, in which a nonfunctional GFP gene is interrupted by an I-*Sce*I restriction cut site (Fig. 4*D*). Upon I-*Sce*I expression and DSB induction, HR repair using an upstream homologous sequence results in functional GFP (Fig. 4*D*; ref. 23). Upon DDIAS depletion using 2 independent siRNAs, we observe a significant decrease in HR, resulting in significantly fewer GFP+ U2OS-SCR cells (Fig. 4*D*; $P < 0.001$). We observe similar results in the HEK293-SCR cell line (*SI Appendix*, Fig. S8; $P < 0.05$). The MRE11 inhibitor, mirin, was used as a positive control for HR inhibition, and an untransfected condition (no I-*Sce*I expressing plasmid) was used as a negative control (Fig. 4*D*). Furthermore, DDIAS knockdown is not accompanied by any gross changes in cell cycle profile (*SI Appendix*, Fig. S9), ruling out the possibility that the HR defect observed is a result of cell cycle arrest. Together, these results indicate that *DDIAS* has a role in mediating an efficient HR response to DSBs.

Discussion

We determined that mammalian DNA repair genes exhibit an evolutionary signature with each other. Apart from NER, we find that genes constituting most major DNA repair pathways exhibit significantly elevated ERC ($P < 0.05$). In addition to finding evidence for coevolution among entire pathways, we also find that the discrete steps within HR coevolve individually. Finally, by ranking the genome based on coevolution with distinct DNA repair pathways, such as HR, we uncovered genes that may be important DNA repair factors such as *DDIAS*. It is interesting to note that DDIAS also contains an OB-fold domain that is similar to the ssDNA binding protein RPA1 (21).

Genome-wide screens are a useful tool in identifying novel members of distinct complexes and pathways. For example, large-scale efforts have documented genome-wide analysis of protein expression (24), cellular localization (25), and genetic and physical interactions (26–28). High-throughput genetic screens using RNAi or sgRNA libraries allow researchers to query the entire coding genome simultaneously (29). These screening methods identify candidate genes for more comprehensive functional analysis. Specifically, gene discovery in DNA repair has benefitted substantially from high-throughput genetic screens in yeast and mammalian systems (19, 20, 27, 28, 30, 31). Although these studies have made important contributions to our understanding of DNA repair, they have limitations including expense, false positives and negatives, and the inability to validate every knockdown by Western blot. Additionally, these studies typically use a single assay, or sensitivity to a single compound, as a readout. This limitation guarantees that genes will be overlooked as not all genes within a given pathway will contribute to a single, specific phenotype. When we compare the

Downloaded by guest on June 3, 2021

Table 1. HR-related GO terms are significantly enriched among genes that exhibit greatest ERC with HR

Padj	Attribute ID	Attribute name	N
<0.001	GO:0000731	DNA synthesis involved in DNA repair	22
<0.001	GO:0000724	DSB repair via HR	36
<0.001	GO:0000725	Recombinational repair	36
<0.001	GO:0006302	DSB repair	63
<0.001	GO:0006310	DNA recombination	63
<0.001	GO:0006281	DNA repair	134
0.001	GO:0006259	DNA metabolic process	191
0.004	GO:0036297	Interstrand cross-link repair	24
0.004	GO:1903046	Meiotic cell cycle process	48
0.004	GO:0006955	Immune response	208
0.006	GO:0006974	Cellular response to DNA damage stimulus	174
0.008	GO:0000732	Strand displacement	16
0.011	GO:0001819	Positive regulation of cytokine production	108
0.017	GO:0032729	Positive regulation of IFN-gamma production	27
0.025	GO:0071897	DNA biosynthetic process	37

GO term enrichment analysis was performed for the 3,497 genes (corresponding to the top 20% in Fig. 3*A*) exhibiting highest mean ERC with HR using the FuncAssociate web tool (llama.mshri.on.ca/funcassociate (38)).

performance of ERC-based gene ranking to genetic screens, we find that ERC is comparable or superior to RNAi-based screens for HR and FA deficiency as measured by drGFP (HR; Fig. 3*B*) and MMC sensitivity (FA; Fig. 3*D*), respectively. Here we used ERC analysis as a gene discovery tool and demonstrate the predictive ability of ERC to infer gene functions such as with the gene *DDIAS*. Like traditional genetic strategies, this computational approach has false negatives and positives. Although we find that 27/39 known HR factors fall in the top 20% of the ERC-based ranking, it is interesting to note the false negatives that result from this ranking. For instance, the RAD51 filament mediator, *RAD51C*, is ranked 11,012 out of 17,487 (63rd percentile). Similarly, *TOP3A* and *MUS81* are ranked 11,585 (66th percentile) and 11,747 (67th percentile), respectively (Fig. 3*A* and Dataset S2). The advantage of this bioinformatic approach is the ability to rank genes without experimental effort and expense.

In systems biology, ERC can be used to analyze evolutionary networks between diverse processes to uncover broader functional relationships. For example, in the context of genetic diseases, elevated ERC inferred related pathogenic mechanisms between supposedly unrelated diseases (7). Functional relationships between distinct cellular processes have clinical importance from the standpoint of synthetic lethality. For instance, MMR deficiency has recently been identified as a robust predictor for response to PD-1 blockade in colorectal tumors (32). Likewise, we find that PDCD1, the gene encoding PD-1, has elevated ERC with several MMR genes (*SI Appendix*, Fig. S10). This finding suggests that ERC analysis could also be used to analyze DNA repair pathways in a broader context and identify new synthetic lethal interactions, possibly of therapeutic use. Overall, evolutionary

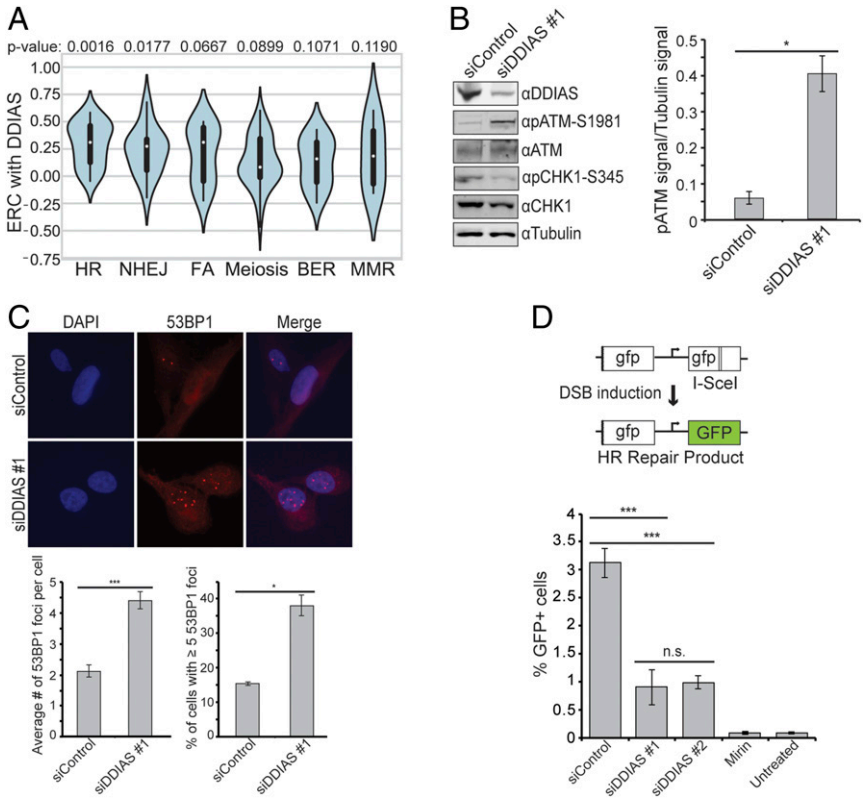


Fig. 4. DDIAS depletion results in defective DSB repair. (*A*) Violin plots show the distribution of ERC values between DDIAS and major DNA repair pathways. Overlaid box plots indicate the quartiles of each distribution, and a white dot represents the median. Permutation *P* values are listed reflecting the significance of DDIAS's ERC elevation with each pathway. (*B*) siDDIAS-U2OS cells were Western blotted for DDIAS, pATM-S1981, ATM, pCHK1-S345, CHK1, and alpha tubulin (*Left*). pATM signal was quantified and normalized to alpha tubulin signal. The mean of 3 experiments is plotted with SEM (*Right*; $*P < 0.05$). (*C*) The 53BP1 foci were quantified in siDDIAS-U2OS cells. The average number of 53BP1 foci (*Bottom Left*) and percentage of cells ≥5 53BP1 foci (*Bottom Right*) were quantified from 2 independent experiments ($n = 200$ cells per condition). Means are plotted with SEM graphed ($*P < 0.05$; $***P < 0.001$). (*D*) DDIAS depletion using 2 independent siRNAs results in reduced HR in U2OS-SCR cells. Schematic of the SCR-GFP reporter system is shown (*Top*). U2OS-SCR-GFP cells were treated with siRNA targeting *DDIAS* (siDDIAS 1 and 2; see *SI Appendix*, Fig. S9, for corresponding Western blot showing DDIAS depletion). Untransfected cells and I-*Sce*I–transfected cells treated with mirin were also measured. Mean %GFP+ is plotted with SE. ($***P < 0.001$.) n.s., not significant.

Downloaded by guest on June 3, 2021

analysis is an important tool to understand the functional relationships between DNA repair genes and pathways and to uncover heretofore unknown components and relationships.

Materials and Methods

Calculation of ERC Values. ERC values were calculated pairwise for 17,487 genes from 33 closely related mammal species as described in Clark and Alani (2, 7). The protein-coding sequences of orthologous genes from 33 mammal species were downloaded from the University of California, Santa Cruz Genome Browser as the 100-way vertebrate alignment (33). The full species set and criteria for species selection are detailed in *SI Appendix, SI Methods*. For each orthologous gene alignment separately, we calculated the branch lengths (amino acid divergence) along the species tree topology using the *aaml* program of the phylogenetic analysis using maximum likelihood (PAML) package (34). The *aaml* program uses a likelihood model for which we chose the Whelan and Goldman empirical substitution matrix. To account for different evolutionary rates across amino acid sites (alignment columns) the model included 3 discrete rate classes and an invariant site class (35). The resulting raw branch lengths for a given gene were then mathematically transformed into relative evolutionary rates (RERs), which represent deviation from the expected amount of divergence. RERs were calculated by normalizing each branch for each gene by the mean length of that branch across all genes. Specifically, this normalization was done by regressing each gene's branch lengths against the genome-wide mean branch lengths and using the residual to represent the amount of change relative to the expectation, wherein positive values represent more change than expected and negative values represent less change (36). The RERs (i.e., the residuals) were thus calculated for all branches, for all genes. ERC between a given gene pair was calculated as the Pearson correlation coefficient between the RER vectors of the 2 genes. ERCs were calculated for all gene pairs, and care was taken to accommodate all gene pairs, despite missing species, by recalculating the RER vectors for every pattern of shared species between all gene pairs.

ERC-Based Gene Ranking. To compare ERC-based rankings to experimental screens, we plotted the number of known HR or FA factors that fall into equally sized bins corresponding to 5% of the total number of genes present in each study. For fair comparison, we only used factors that were present in both the ERC and experimental datasets for ranking. Hence, 39 HR genes and 20 FA genes are considered in the ranking analysis (versus the 42 HR genes and 23 FA genes originally considered in *SI Appendix*, Table S1).

The HR genes considered were BLM, BRCA1, BRCA2, BRIP1, EME1, EXO1, FANCM, INO80, MRE11A, MUS81, NBN, PALB2, PARPBP, RAD50, RAD51, RAD51B, RAD51C, RAD51D, RAD52, RAD54B, RAD54L, RBBP8, RECQL, RECQL4, RECQL5, RMI1, RMI2, RPA1, RPA2, RPA3, SFR1, SLX4, SMARCAL1, SWSAP1, TOP3A, WRN, XRCC2, XRCC3, and ZRANB3.

The FA genes considered were BRCA1, BRCA2, BRIP1, C17orf70, ERCC4, FANCA, FANCB, FANCC, FANCD2, FANCE, FANCF, FANCG, FANCI, FANCL, FANCM, MAD2L2, PALB2, RAD51, RAD51C, and RFWD3.

To account for 1) the different sizes of the screens and 2) the different numbers of genes in unique cellular pathways, we scaled the *x* and *y* axes by the size of the screen and the number of pathway components, respectively. For HR, the *y* axis is defined as

$$\textit{Frequency of known HR genes} = \frac{\textit{number of HR genes in bin}}{\textit{total \# of HR genes}},$$

and the same analysis was performed for FA. Likewise, to directly compare screens of different sizes, we scaled the *x* axis by the size of the respective screen:

$$\textit{Percentile rank} = \frac{\textit{position in ranked list}}{\textit{length of ranked list}}.$$

A detailed description of the scan statistic used to determine the enriched gene window, initially reported in ref. 37, can be found in *SI Appendix, SI Methods*.

For a detailed description of the molecular experiments, please see *SI Appendix, SI Methods*.

ACKNOWLEDGMENTS. This work was supported by the NIH Grants (ES024872 to K.A.B. and HG009299 to N.L.C.), ACS Research Scholar Grant RSG-16-043-01–DMC to K.A.B., and Stand Up to Cancer Innovative Research Grant SU2C-AACR-IRG-02-16. We thank the following experts for their assistance with our gene lists: Drs. Bennett Van Houten, Robert Sobol, Katharina Schlacher, Jeremy Stark, and Judith Yanowitz.

1. K. Natarajan *et al.*, Transcriptional profiling shows that Gcn4p is a master regulator of gene expression during amino acid starvation in yeast. *Mol. Cell. Biol.* **21**, 4347–4368 (2001).
2. N. L. Clark, E. Alani, C. F. Aquadro, Evolutionary rate covariation reveals shared functionality and coexpression of genes. *Genome Res.* **22**, 714–720 (2012).
3. N. L. Clark, E. Alani, C. F. Aquadro, Evolutionary rate covariation in meiotic proteins results from fluctuating evolutionary pressure in yeasts and mammals. *Genetics* **193**, 529–538 (2013).
4. G. D. Findlay *et al.*, Evolutionary rate covariation identifies new members of a protein network required for Drosophila melanogaster female post-mating responses. *PLoS Genet.* **10**, e1004108 (2014).
5. S. Böhm *et al.*, The budding yeast ubiquitin protease Ubp7 is a novel component involved in S phase progression. *J. Biol. Chem.* **291**, 4442–4452 (2016).
6. S. K. Godin *et al.*, Evolutionary and functional analysis of the invariant SWIM domain in the conserved Shu2/SWS1 protein family from Saccharomyces cerevisiae to Homo sapiens. *Genetics* **199**, 1023–1033 (2015).
7. N. Priedigkeit, N. Wolfe, N. L. Clark, Evolutionary signatures amongst disease genes permit novel methods for gene prioritization and construction of informative gene-based networks. *PLoS Genet.* **11**, e1004967 (2015).
8. J. H. Hoeijmakers, Genome maintenance mechanisms for preventing cancer. *Nature* **411**, 366–374 (2001).
9. C. A. Waters, N. T. Strande, D. W. Wyatt, J. M. Pryor, D. A. Ramsden, Nonhomologous end joining: A good solution for bad ends. *DNA Repair (Amst.)* **17**, 39–51 (2014).
10. S. C. Kowalczykowski, An overview of the molecular mechanisms of recombinational DNA repair. *Cold Spring Harb. Perspect. Biol.* **7**, a016410 (2015).
11. J. A. Marteijn, H. Lans, W. Vermeulen, J. H. Hoeijmakers, Understanding nucleotide excision repair and its roles in cancer and ageing. *Nat. Rev. Mol. Cell Biol.* **15**, 465–481 (2014).
12. H. E. Krokan, M. Bjørås, Base excision repair. *Cold Spring Harb. Perspect. Biol.* **5**, a012583 (2013).
13. T. A. Kunkel, D. A. Erie, Eukaryotic mismatch repair in relation to DNA replication. *Annu. Rev. Genet.* **49**, 291–313 (2015).
14. A. D. D'Andrea, Susceptibility pathways in Fanconi's anemia and breast cancer. *N. Engl. J. Med.* **362**, 1909–1919 (2010).
15. F. Baudat, Y. Imai, B. de Massy, Meiotic recombination in mammals: Localization and regulation. *Nat. Rev. Genet.* **14**, 794–806 (2013).
16. A. B. Ziegler *et al.*, The amino acid transporter Jhl-21 coevolves with glutamate receptors, impacts NMJ physiology, and influences locomotor activity in Drosophila larvae. *Sci. Rep.* **6**, 19692 (2016).
17. M. Chang, J. C. Dittmar, R. Rothstein, Long telomeres are preferentially extended during recombination-mediated telomere maintenance. *Nat. Struct. Mol. Biol.* **18**, 451–456 (2011).
18. A. J. Pierce, R. D. Johnson, L. H. Thompson, M. Jasin, XRCC3 promotes homology-directed repair of DNA damage in mammalian cells. *Genes Dev.* **13**, 2633–2638 (1999).
19. B. Adamson, A. Smogorzewska, F. D. Sigoillot, R. W. King, S. J. Elledge, A genome-wide homologous recombination screen identifies the RNA-binding protein RBMX as a component of the DNA-damage response. *Nat. Cell Biol.* **14**, 318–328 (2012).
20. A. Smogorzewska *et al.*, A genetic screen identifies FAN1, a Fanconi anemia-associated nuclease necessary for DNA interstrand crosslink repair. *Mol. Cell* **39**, 36–47 (2010).
21. K. J. Won *et al.*, Human Noxin is an anti-apoptotic protein in response to DNA damage of A549 non-small cell lung carcinoma. *Int. J. Cancer* **134**, 2595–2604 (2014).
22. S. So, A. J. Davis, D. J. Chen, Autophosphorylation at serine 1981 stabilizes ATM at DNA damage sites. *J. Cell Biol.* **187**, 977–990 (2009).
23. N. Puget, M. Knowlton, R. Scully, Molecular analysis of sister chromatid recombination in mammalian cells. *DNA Repair (Amst.)* **4**, 149–161 (2005).
24. D. M. Abd El-Rehim *et al.*, High-throughput protein expression analysis using tissue microarray technology of a large well-characterised series identifies biologically distinct classes of breast cancer confirming recent cDNA expression analyses. *Int. J. Cancer* **116**, 340–350 (2005).
25. P. J. Thul *et al.*, A subcellular map of the human proteome. *Science* **356**, eaal3321 (2017).
26. T. Ito *et al.*, A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. U.S.A.* **98**, 4569–4574 (2001).
27. D. Alvaro, M. Lisby, R. Rothstein, Genome-wide analysis of Rad52 foci reveals diverse mechanisms impacting recombination. *PLoS Genet.* **3**, e228 (2007).
28. E. B. Styles *et al.*, Exploring quantitative yeast Phenomics with single-cell analysis of DNA damage foci. *Cell Syst.* **3**, 264–277.e10 (2016).
29. M. C. Bassik *et al.*, A systematic mammalian genetic interaction map reveals pathways underlying ricin susceptibility. *Cell* **152**, 909–922 (2013).
30. A. Abeyta, M. Castella, C. Jacquemont, T. Taniguchi, NEK8 regulates DNA damage-induced RAD51 foci formation and replication fork protection. *Cell Cycle* **16**, 335–347 (2017).
31. J. P. Svensson, R. C. Fry, E. Wang, L. A. Somoza, L. D. Samson, Identification of novel human damage response proteins targeted through yeast orthology. *PLoS One* **7**, e37368 (2012).

Downloaded by guest on June 3, 2021

32. D. T. Le *et al.*, Mismatch repair deficiency predicts response of solid tumors to PD-1 blockade. *Science* **357**, 409–413 (2017).
33. M. Blanchette *et al.*, Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.* **14**, 708–715 (2004).
34. Z. Yang, PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
35. S. Whelan, N. Goldman, A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.* **18**, 691–699 (2001).
36. T. Sato, Y. Yamanishi, M. Kanehisa, H. Toh, The inference of protein-protein interactions by co-evolutionary analysis is improved by excluding the information about the phylogenetic relationships. *Bioinformatics* **21**, 3482–3489 (2005).
37. V. P. Janeja, A. Vijayalakshmi, "LS3: A linear semantic scan statistic technique for detecting anomalous windows" in *Proceedings of the 2005 ACM Symposium on Applied Computing*, L. M. Liebrock, Ed. (Association for Computing Machinery, New York, 2005), pp. 493–497.
38. G. F. Berriz, O. D. King, B. Bryant, C. Sander, F. P. Roth, Characterizing gene sets with FuncAssociate. *Bioinformatics* **19**, 2502–2504 (2003).

Downloaded by guest on June 3, 2021