Identification and investigation of novel DNA damage repair genes via network analysis

by

Taylor Alexandra Potter

A thesis submitted to the Faculty of Graduate and Postdoctoral Affairs in partial fulfillment of the requirements for the degree of

Master of Science

in

Biology

Carleton University
Ottawa, Ontario

Abstract

While most DNA lesions are repaired faithfully and without genotoxic effect, double-stranded breaks (DSBs) are exceptional. The deleterious potential of a misrepaired DSB represents a severe threat to cellular integrity. Repair machinery defects are frequently observed in tumorigenic and oncogenic cells. General DNA damage repair mechanisms involve homologous recombination (HR), or non-homologous end joining (NHEJ). Ongoing identification of new players in DSB repair leads us to believe there are more undiscovered genes in this pathway. The highly complex and conserved nature of DSB repair across eukaryotic and mammalian cells presents an opportunity for identification of novel genes through a computationally-directed approach. Employing a 'guilt-by-association' model, we analyzed experimental and predicted interaction networks in *S. cerevisiae* to identify previously uncharacterized genes involved in repair. Three novel genes were discovered to influence repair- *GAL7*, *YHI9*, and *YMR130W.* The results of this study implicate all three in the DNA damage response network.

Acknowledgements

Dr. Ashkan Golshani, G-lab, and the third floor crew- The genuine spirit and true kindness of every one of you is something I will carry with me for the rest of my life.

Mom and Dad-

I love you forever.

"It was people who saved me. They still do."[1]

Table of Contents

List of Tables

List of Figures

List of abbreviations

alt-NHEJ	alternative non-homologous end joining
amp^R	ampicillin resistance
c-NHEJ	classical- Non-homologous end joining
DDA	DNA damage array
DDR	DNA damage response
DNA	deoxyribonucleic acid
DSB	double-stranded break
GAL	galactose
GI	genetic interaction
HR	homologous recombination
HU	hydroxyurea
MMEJ	microhomology-mediated end-joining
MMS	methyl methanesulfonate
MS	mass spectrometry
nat^R	nourseothricin resistance
NHEJ	non-homologous end joining
ORF	open reading frame
PIPE	protein- protein interaction prediction engine
PPI	protein- protein interaction
SGA	synthetic genetic array
SDL	synthetic dosage lethality

1. Introduction

1.1 DNA damage

Severe damage to a cell's DNA is inarguably the largest threat to its survival. Each cell possesses a singular set of its genetic code, and from this code the cell must extract information about every process and molecule required for life. Further, this DNA must be faithfully replicated and passed on to every daughter cell. The very blueprint of life is under constant stress in the form of damage, which must be repaired efficiently and accurately if its integrity is to be maintained.

The eukaryotic cell has evolved a set of highly conserved and sophisticated systems to repair DNA damage, both exogenous and endogenous. While mammalian cells do experience endogenously generated breaks (To provide context, a human cell is estimated to experience approximately 100,000 lesions per day from sunlight exposure alone[2].The intricate nature of these systems continues to be a central theme of research due to the direct implications of their failure in the fields of mammalian disease, evolution, and cancer. Of primary research concern are the repair pathways employed following a double stranded break (DSB), an event in which both DNA strands are severed and characteristic of oncogenic cellular activity[3]. Consequences of DSB misrepair are severe and often life-threatening due to their nature.

The DNA damage response pathway is a highly conserved and complex series of functions that primarily involve non- homologous end joining, homologous recombination, and microhomology-mediated end joining[3]. Such a diverse group of potential pathways that exist in the cell for DNA damage response leads to a host of

protein interactions mediating function within the system. A large scale understanding of this pathway will enable the development of targeted drug delivery systems, understanding of the mechanisms involved in the progression of disease, and furthermore provide a deeper understanding of biological processes in the cell.

1.1.1 Double-stranded break repair

In eukaryotic biology, the double-stranded break (DSB) is regarded as the most severe form of DNA damage. The repair process following a DSB event is subsequently high stakes- failure of the cell to respond appropriately can result in sequence mutation, severe chromosomal rearrangement, carcinogenesis, or cell death[4]. DSBs are largely a result of external forces in eukaryotic cells, most commonly ionizing radiation, reactive oxygen species, and wayward nuclease activity[5]. The diversity of both the causes of damage and the pathways involved in repair necessitates a complex, dynamic cellular response, with numerous influential proteins and complexes yet to be characterized.

Classical understanding of the cellular response to DSB repair recognizes two classes of repair pathway- homologous recombination (HR), and non-homologous end joining (NHEJ). HR requires the presence of a donor template to serve as a 'blueprint' for repair of broken ends- by nature it is considered the 'error-free' pathway. In contrast, NHEJ generally involves direct ligation of cohesive broken ends, and is the 'error-prone' pathway. NHEJ dominates as the primary repair mechanism in somatic mammalian cells at all stages of the cell cycle, whereas HR events are limited by template availability[4,6].

1.1.2 Homologous recombination

Repair by HR is understood to be a more sophisticated response than NHEJ- the very nature of HR lends itself to increased accuracy and fidelity at complex break sites[7]. Following DSB induction, various end resection machinery is recruited to the site of the break to prepare the broken ends for recombinational repair. During S and G2 phase, the nuclease Sae2 is believed to interact with the MRX/MRN complex (Mre11, Rad50, Nbs1) to promote end resection[8]. The MRX/MRN complex (budding yeast and mammalian cells, respectively) plays a universal role in DSB detection, recruiting end resection machinery to the broken ends during both HR and NHEJ [9,10]. The complex is conserved from human to yeast, as is the recombinase protein Rad51p. Rad51p binds the newly resected DSB ends, initiating strand invasion of a suitable homologous template[8]. This process is facilitated by the eponymous member of the Rad52 epistasis group of proteins in yeast. Rad52p, a DNA binding protein, is required for Rad51p function in homologous recombination in addition to playing an essential role in DSB repair across the yeast genome. Interestingly, Rad52 has not been observed to perform similar functions in mammalian cells despite extensive sequence homology and evolutionary conservation. Recent work has instead implicated mammalian RAD52 in restarting replication fork collapse induced by oncogenic stress and tumor proliferation through break-induced replication[3].

The template of choice in HR is generally a sister chromatid, close in proximity to the break site- hence the prominence of HR activity in cycle phases or ploidy status where such a template is available. Mammalian cells, due to extensive redundancy of their genome, may also utilize sequence homologies located on the same strand. This process

is a potential contributing factor to the apparent preference towards NHEJ- such a template may be located an unreasonable distance from the site of the lesion and compromise speed of repair.

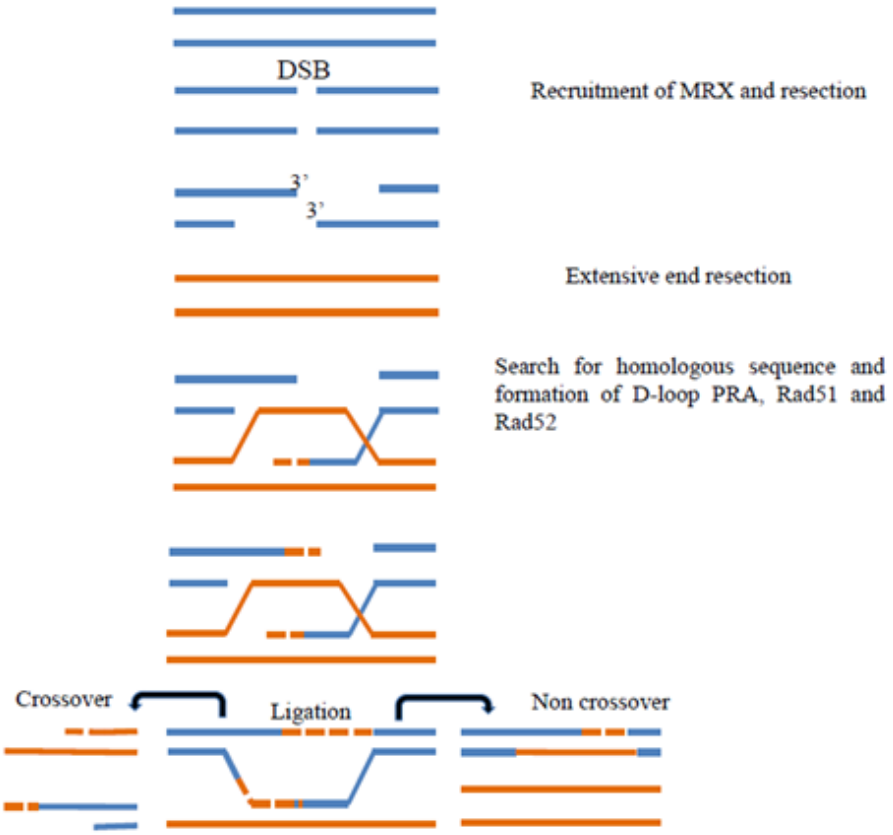

Figure 1: Simple overview of homologous recombination pathway. Adapted from Omidi, 2017[11].

1.1.3 Non-homologous end-joining

The non-homologous end joining (NHEJ) pathway is generally regarded as having greater deleterious potential than HR[12]. While highly efficient, the lack of complex end-processing activity can often lead to a decrease in fidelity of the repair product, dependent on the sequence surrounding the break site or presence/absence of regulatory

proteins[7,12]. Despite these qualities, NHEJ is an essential process across all eukaryotes, capable of repairing an array of broken ends and doing so independently of a repair template[13]. The core machinery involved in classical NHEJ (c-NHEJ) is highly conserved between human and yeast cells, comprising the Ku70/80 heterodimer, the MRX complex, and ligase proteins such as LIG4 in human cells or the complex DNL4 in yeast[8,14]. KU70 and KU80 (Yku70 and Yku80 in yeast) act as first responders, binding to the exposed DNA ends and preventing HR machinery from initiating end resection. Ku70/80 then serves as a docking site, recruiting nuclease complex MRX to process the ends, as well as polymerases (Pol4 in yeast) and ligases or ligase complexes (such as DNL4, comprised of Dnl4, Lif1, and Nej1) to initiate re-ligation[13,14].

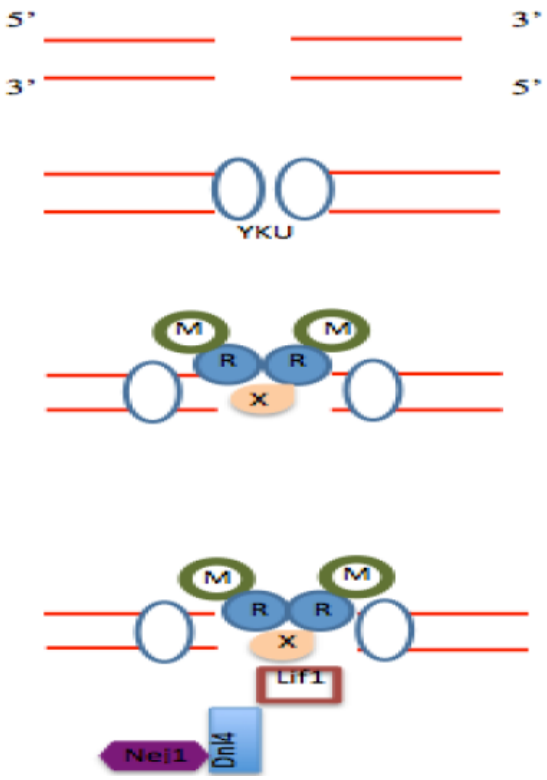

Figure 2: Simplistic overview of the non-homologous end-joining pathway. Adapted from Omidi, 2017[11].

There exist several NHEJ subclasses of repair aside from c-NHEJ. One such subclass involves repair independent of the Ku70/80 heterodimer, as part of alternative NHEJ or microhomology- mediated end joining (alt-NHEJ/ MMEJ). The process is regulated by Mre11, utilizing short homologous sequences (5-25bp) proximal to the break site to repair the lesion[15]. Notably, repair performed through alt-NHEJ frequently leads to mutagenic outcomes and loss of information[12].

1.1.4 Pathway choice in eukaryotes

The mechanisms surrounding cellular commitment to one repair pathway over another are not yet clear. Ploidy, cell cycle stage, and template availability are recognized as contributing factors and have been studied extensively, albeit with varying outcomes and conclusions. It is also possible that the core initiators in each pathway compete in vivo for DSB processing, irreversibly committing the cell to a certain pathway. In budding yeast, HR tends to dominate, exemplified by the observation of diploid cells actively repressing key NHEJ proteins[16]. Mammalian somatic cells, however, rely primarily on NHEJ and by extension its integrity[10].

NHEJ has long been considered inherently error-prone and/or deleterious, a viewpoint which has been challenged in recent years. Estimates of DSB frequency range from 10 to 10^5 per cell, per day (inclusive of endogenous damage)- indicative of the complex and multifaceted nature of damage events[13,17]. The frequency of occurrence and variable complexity of these breaks bring into question the persistence of cellular information through a primary mode of repair that is purportedly genotoxic by nature. It has been

proposed that unlike budding yeast, mammalian HR is burdened by the extensive redundancy of its genome when searching for a repair template, thereby making NHEJ the more attractive fix[18]. Alternatively, NHEJ is a 'fast' pathway, whereas HR is 'slow'. The lag in time between DSB events and repair initiation may very well serve as a deciding factor in pathway commitment.

More recently, evidence for a modern class of DNA-damage genes has emerged- genes whose products may not be directly involved in repair but instead regulate pathway choice and fidelity. On a broad scale, examples of these regulatory elements include the mammalian tumor suppressor BRCA1, which promotes repair fidelity in both HR and c-NHEJ, whilst inhibiting alt-NHEJ activity at chromosomal and episomal breaks, respectively[12]. Tumor suppressor and cell cycle related protein, p53, has been observed to potentially interact with KU80 to prevent carcinogenesis and senescence defects[19]. Indeed, many tumor cells have observed defects in end-joining processes through disruption of regulatory proteins, both independent of or in tandem with defects in core repair machinery.

1.1.5 Cell cycle checkpoints and influences

Regulation of pathway choice and recruitment of corresponding protein complexes in DSB repair is also intricately connected to the cell cycle. There exist three damage checkpoints throughout the mammalian cell cycle- at the G1/S transition, during S phase, and at G2/M[20]. Each checkpoint is associated with a cohort of signaling proteins and kinases that detect damage and subsequently activate the mechanisms necessary for

response. This response is crucial in preventing the progression of compromised cellular information throughout the cycle.

At each checkpoint, the cell must ensure that the integrity of its DNA is upheld before transition by employing an intricate response network. Arrest at G1/S or G2/M tends to be a result of exogenous DSBs, whereas the mid-S phase checkpoint is often initiated due to stalled or collapsed replication forks, slowing progression[21]. The MRX complex again serves a critical role in detection of DSBs and initiation of checkpoint response. Briefly, kinases Mec1 and Tel1 are recruited to the site of the lesion, mediating activation of trigger checkpoint protein Rad9 and kinase Rad53 through phosphorylation of key histone and chromatin modifiers. Further, Rad53 is responsible for performing key modifications of cell cycle effector proteins in order to activate arrest[21].

Cyclin dependent kinases (CDKs) play a major role in this response pathway. In budding yeast, CDK Cdc28 is the master regulator of cell cycle progression and is essential for cellular regulation and division. Recent evidence supports a role in the damage response as well, through interaction with known DNA damage genes such as Yku70 at the beginning of G1[9]. Interactions with specific cyclins at each phase initiates various chromatin and histone modifications, enabling communication with signaling kinases.

Loss or degeneration of cellular information has the potential to impact not only the integrity of DNA, but also gene products and interactions with regulatory substrates. Consideration of these factors is critical and directly relevant to both the understanding of the DNA damage response pathway and the resultant cellular response.

1.2 Systems biology

The complexity of the cell in whole is largely due to the interactive functions between proteins, formation of complexes, and of intricate signaling networks. The wealth of cellular information available to researchers in the post-genomic era has prompted a shift in approach to reflect this- no longer are genes or proteins investigated as isolated entities. Rather, there is a focus on comprehensive data- driven approaches and development of novel techniques designed to probe cellular systems. This is the field of systems biology, the overarching goal of which is to probe the function of uncharacterized genes and proteins towards an understanding of the cell on a global level[22,23].

Whole-genome sequencing provides few insights into the function of genes and their products. The process of relating genotype to phenotype across the cellular landscape remains in progress- *Saccharomyces cerevisiae*, the first eukaryote to have its full genome sequenced decades ago- has still a number of genes that have yet to be assigned function[22]. The last two decades have seen a significant effort towards a complete yeast interactome, the challenge of which perhaps pales in comparison in the context of relating the data to the much larger, highly complex human cell[24]. Bioinformatic approaches have proven an attractive alternative to identifying and characterizing these novel genes and networks, as traditional experimental approaches are time consuming and costly. The production of such massive data collections, representing every facet of cellular life, has undoubtedly revolutionized the landscape of molecular biology.

1.2.1 Functional genomics

The subfield of functional genomics comprises a wide variety of techniques aimed at elucidating gene/ gene product function in a cellular context. The relationship between genetic code and genetic function is evidently non-binary. To date, there are over 20,000 genes encoding a functional protein in the genome of *Homo sapiens*, with more than 40,000 interactions identified out of an estimated potential 400,000[24]. This is projected based on the information gained through study of *Saccharomyces cerevisiae*, possessing a modest 6800 proteins and 44,000 identified interactions[23]. Translating the functional data gained from evolutionarily conserved model organisms is an indispensable tool in achieving full annotation of the human interactome.

Movement towards a complete human interactome requires a characterization of all possible interactions across all cellular components. Protein- protein interactions are core indicators of protein, complex, and pathway function. Cellular functions and processes are realized through interactions, complex formation, and cross-talk between these components. Furthermore, the presumption that a protein or complex fulfills a singular role is losing eminence as emergent data suggests many proteins exhibit a functional hierarchy, or participate in 'moonlighting'[25]. An organism's genome produces proteins with such diverse functions that several different approaches and perturbations are necessary to elucidate the minutiae governing cellular life.

1.2.2 Protein-protein interactions

Proteins recognize function upon interaction with other proteins or macromolecules- in both binary fashion or through formation of complexes. It is therefore desirable to probe not only a protein's function but the larger context of said function. High-throughput methods for systematic functional analyses have been made widely accessible and reproducible following release of the Saccharomyces Gene Deletion Project in 2000. The collection of nearly 6000 mutant ORFs has proven a powerful tool in functional genomics- from study of genetic interaction to chemical genomics screenings and protein-protein interaction assays. The ability to perform high-throughput analyses of gene function in the context of any number of conditions is valuable in both yeast and downstream mammalian investigations. Identification of drug targets, formation of protein complexes, systematic perturbations of mutant cells, and compensation studies are only a small fraction of the work performed in functional genomics over the past two decades. Following several large-scale screens in the early 2000's, functional yeast manipulations and the resultant interaction data skyrocketed. These screens presented functional genomics studies through the yeast-two-hybrid technique (Y2H). Y2H utilizes the structure of the yeast deletion collection to identify protein-protein interactions on a genome-wide scale, with applicability to studies in a number of species- including *Homo sapiens*[25]. There have been two major works published in the field that laid the groundwork for large scale protein-protein analysis study. The first involved completion of two yeast two hybrid screens- resulting in over 900 putative interactions detected[26]. A follow up was performed which provided both a comprehensive yeast two hybrid screen of the *Saccharomyces cerevisiae* genome and invaluable insight into the capabilities and

restrictions of large scale yeast two hybrid screening[22]. These studies were arguably a landmark in the development of modern high throughput yeast two hybrid screening techniques.

To this day, yeast two hybrid screening remains a popular choice for those researching protein- protein interactions and networks. The combination of such a user friendly and direct method for screening combined with a bioinformatic approach is desirable, as seen in a 2015 study where it is claimed "The yeast-two hybrid (Y2H) system is… an important complement to other biochemical approaches."[27]. This study demonstrated the capability of using a bioinformatic and structural approach contiguously with a yeast two hybrid screen to identify specific pathways in the cell. This study emphasized an important consideration in yeast two hybrid screens- two different screens may not provide the same interaction data (replicability is inconsistent), and that protein A as bait may be shown to interact with protein B as prey, but not vice versa. The former limitation of course being addressed using a conjunction of validation techniques post screen, and the latter addressed by performing several tests in each direction.

1.2.3 Genetic interactions

Genetic interactions provide an essential framework for construction of cellular interaction networks, as they govern the relationship between genotype and phenotype[28]. SGA and SDL studies provide a framework with which to discover and interpret genetic interactions. Genes involved in parallel pathways may be identified through their combined effect upon cellular phenotype, where a negative interaction indicates this type of parallel process involvement. A positive interaction may occur where the phenotype is

less severe than is expected and may indicate functional redundancy between genes. Similarity in interaction profiles also enables discovery of novel genes involved in a query pathway.

This model extends to the study of perhaps more dynamic or complex interactions. Gene products or proteins may be further studied through colocalization or expression studies- if protein X and protein Y are conditionally migrating to the same subcellular space, one may be able to infer functional similarity or pathway involvement between said proteins. Similarly, expression profiles of query genes may be generated across a variety of cellular environments and utilized in downstream comparative analyses. Analyses of tertiary or quaternary genetic and protein interactions may enable much greater insight than can be achieved from the study of binary interactions- formation of permanent complexes is central to realization of functional processes.

All functional genomics techniques operate through a hypothesis driven approach, with the end goal being characterization and development of intricate functional networks. Further, these studies can employ well-established base knowledge of 'big players' involved in regulating or initiating cellular processes to discover more nuanced or novel functional profiles. The 'guilt-by-association' philosophy of interaction studies is a central philosophy across multitude of similar techniques- including the Synthetic Genetic Array (SGA) technique and its successor Synthetic Dosage Lethality (SDL), GFP localization studies, and RNA-seq approaches. These dynamic and flexible techniques enable researchers to probe a virtually unlimited range of interactions.

1.3 Computational approaches to functional studies

As the field of computational modelling progresses, genomics too has adapted to the information age, through the implementation of computational models for investigation of protein- protein interactions. Modern, multi-genome databases represent a class of comprehensive tools for systems biology research- a prominent example of which is STRING[1,29]. STRING integrates both experimentally derived data and newfound prediction models into its protein-protein interaction database. The traditional functional genomics workflow has seen remarkable improvements in efficiency and time-cost by utilizing these types of tools, the forefront of which are genome wide protein-protein or protein-DNA interaction centered.

1.3.1 Domain and motif-based protein-protein interaction tools

Domains are protein subunits that are generally conserved across species and confer functional or conformational properties. Identification of the domains, structural or functional, present in a query protein allows putative prediction of function and/or interaction[30]. Domain architecture of a protein is also directly relevant to the goal of transferring functional networks across species, as proteins containing similar domains are believed to behave similarly across conserved homologous genes. Of course, proteins realize these functions upon interaction with other proteins, so the efficacy of domains on prediction of interactions is of special interest. While domain composition can provide general indications of interaction likelihood, their applicability is limited in terms of precision[31]. Transient and permanent interactions are largely dependent on sub- structure and composition of the participating molecules- thus there have been several efforts

towards identifying the precise motifs within these domains responsible for mediating interactions.

Pfam (http://pfam.xfam.org/) and Protein Data Bank (PDB, www.rcsb.org) are two of the most ubiquitous databases of domain interaction and protein sequence information, respectively, from which resources such as DOMINE (protein-domain interactions) and DOMINO (protein- peptide interactions) are able to catalogue large amounts of domain-centered interaction data[32–35]. Aggregation of this data supports a robust tool for identification of novel proteins involved in cellular processes, or discovery of alternative functions for existing proteins.

Many domain-mediated interactions with cellular macromolecules are well characterized, however may not be the most reliable approach to prediction of PPIs due to lack of experimental base data. Resolving domain-domain interactions (DDIs) within the cell experimentally provides little information regarding specific interaction sites or mediating subunits. Despite overarching sequence conservation, short peptides encoded within the domain may exhibit differential affinities for target residues depending on the parent protein[35]. These details may be overlooked in generalized domain models and frequently diminish the capability and sensitivity of the model. Attempts to identify more reliable domain-based interactions have generally utilized known data concerning domain-peptide interactions or resolved 3D structure of the domains themselves, independent of their parent molecule. Web-based app 3D-ID curates and annotates these and similar events to provide context for domain function and affinity across a variety of settings[36]. These catalogues have formed the basis upon which several predictive tools have been built and can be utilized to provide an indication of both interaction type and

likelihood. The conserved properties of domains are also applicable in transferring interaction data across species and identifying homologous proteins. The SMART domain architecture database (http://smart.embl-heidelberg.de) is a catalogue of domain sequences in the context of functional or structural families[37]. The database is distinct from earlier discussed web tools in its approach to non-enzymatic domain classification and the total domain composition of a protein. SMART identifies representative domain composition across family members through multiple sequence alignment, taking into consideration the (often problematic) highly divergent architecture. The alignment data serves as a basis for the analysis protein sequences for domain composition similarity (either similar in structural distribution or with divergent domain arrangement). The result is relevant for elucidation of the domains or short sequences responsible for sub-cellular localization and signaling, among others.

While both domain and motif structures and functionality are conserved across evolutionarily related species, motif-based prediction techniques allow increased precision in terms of interaction location and nature- transient or permanent. Motif-centered tools allow for identification of the short polypeptides responsible for formation of PPIs, whether they are transient interactions or permanent ones, as observed in protein complexes. Protein- Protein Prediction Engine (PIPE) was first presented in 2006 as a linear-motif based PPI prediction tool[31]. The input sequence and interaction data are obtained through protein data repositories UniProt (http://www.uniprot.org) and BioGRID (thebiogrid.org), containing over 20,000 manually curated and reviewed protein sequences[38,39]. High-confidence experimental interaction data is represented by a series of subnetworks of known interacting proteins. The software is designed to identify

short sequences (motifs) present in protein 'X', then scans all proteins in X's network for similar motifs. Positive hits and their neighbours are compiled into a subset of potential interactors for protein X. The potential for X and any other protein in the subset to interact is calculated based on the frequency of occurrence of the motif(s) in known binary interactions.

PIPE presents an attractive alternative to experimental screens with a sensitivity of 23% and false negative rates comparable to traditional methods (e.g. TAP tagging). The software also presents a tangible understanding of issues surrounding human genome complexity- analysis of a yeast protein pair may take seconds, however a human protein pair may require a runtime anywhere between 1 second to 12 hours[40].

1.3.2 Genetic interaction prediction tools

Prediction of genetic interactions (GIs) is a more precarious task than that of protein-protein or protein- DNA interactions. Despite this, the prediction of genetic interactions on a systems biology level remains important, as genetic interactions form the basis of understanding global cellular networks. Mapping of GIs identifies the fundamental relationships between the components of the genetic landscape, from which underlying pathways and processes can be contextualized[25] . Following the millennial boon of large scale network analysis, genetic interaction data, primarily from yeast studies, formed a large portion of the contributing information[24]. The task of scaling these networks from model organisms to the human genome is becoming feasible through integration of predictive computational tools- the use of which is also essential in elucidating novel functions. The practical limitations of experimental network investigation in terms of

coverage, reproducibility, and the sheer time required to exhaust all binary interactions perhaps best exemplifies the necessity of computational prediction.[25]. GI networks serve as the scaffolding onto which alternative functional data can be mapped.

Though scarcer than protein or subunit prediction tools, there have been several successful GI prediction tools developed in recent years. These tools are generally geared towards prediction of synthetic lethal interactions due to the inherent variability and vague nature of the phenotypes/ indicators associated with 'positive' interactions. Building on the value of synthetic lethal interactions, software was introduced in 2009 that utilized known binary interaction data to predict novel lethal phenotypes[41]. Termed the '2-hop' scheme, originally presented by Wong et al[42], predictions are generated from known binary sick or synthetic lethal (SSL) interaction partners and projected onto a third as described:

> 2-hop network motifs capture the relationship between a pair of genes, e.g. A-B, and a third gene, C. In this example, genes A and B share a physical interaction, while genes A and C are synthetic lethal. The 2-hop scheme would suggest that genes B and C might also be synthetic lethal[41].

Incorporating the 2-hop method with random walking of the network resulted in higher confidence and coverage of the query organisms' genetic space, with 10% and 7% false negative rates (*S. cerevisiae and C. elegans*, respectively).

These networks have yet to be transferred or even generated in human successfully- part of the issue appears to stem from the lack of appropriate training data (i.e. 'no interaction'). To solve this, Calzone et al designed a logical framework upon which to

predict genetic interactions in regulatory pathways, validated by comparison to existing experimental data[43]. Their primary model was generated by random walking to ascertain the probability of identifying the phenotype of all possible pairwise interactions, which were then used to apply Boolean values to each node (0 or 1, where 0 was 'loss of function' and 1 was 'gain of function'). The MAPK, cell cycle regulation, and cell fate decision pathways were used to evaluate the performance of the tool to moderate success. This type of approach is of value in terms of attempting to solve the issues of bias and lack of input data in network analysis- by pooling several mathematical approaches and experimental analyses perhaps a complete human interactome is not far away.

1.3.3 Tools for co-expression and co-localization

The dynamics of transcription, localization, interaction, and degradation activity in a variety of cellular contexts are often tightly regulated and reactive. Many proteins are targeted by regulatory bodies for relocation, degradation, interaction, through several types of post-translational modification (PTM). These PTMs modify specific residues or motifs on proteins or substrates in order to regulate cellular activity, and comprise an important part of the genotype-phenotype relationship[44]. The complexity involved in predicting the occurrence, result, and relevance of a modification necessitates a computational approach. Several tools exist to mine sequence data for existence of well-annotated modification pathways and their target sites (e.g. SUMOlyation, phosphorylation, methylation), and to apply this data to in silico predictions[45].

The co-localization of proteins appears to be enriched between interactors[46]. Key proteins involved in cellular response to stress or damage conditions, for example, display notable

changes in subcellular localization upon exposure to damage-inducing drugs or deletion of a major repair gene[47]. Returning to the principle of 'guilt by association', this implicates co- localization prediction as another strong tool in the systems biologists' toolkit. LocTree3 (https://www.rostlab.org/services/loctree3/) uses existing localization annotation to apply sequence- based predictions to homologous proteins[48]. The user is able to access predictions and Gene Ontology (GO) term annotations for input sequences with close homologs and existing annotation as well as sequences with no prior annotation or closely related homologs[49,50].

Functionally related genes are not only co-localized, but often co-expressed. The extensive amount of RNA sequencing (RNA-seq) data generated in recent years has the potential to uncover novel factors and relationships involved in gene function and is frequently exploited towards this goal. A comprehensive tool for prediction of co-expression generated from this type of data may be of significant impact to study of disease and dysregulation in human cells. The difficulty in identifying appropriate training data sets for machine-leaning algorithms, even in lower-complexity genomes (those of model organisms, for example), is likely responsible for the lack of widespread success in developing such a tool. Even with well characterized co-expression data, identifying the regulatory modules responsible for transcription is precarious and renders simplistic/ generalized models unsuitable[51]. There has been success in meta-analyses, utilizing open-chromatin regions, transcription factor motif recognition, and histone modification patterns. Using DNAse-seq data from the ENCODE project (www.encodeproject.org), Natarajan et al were able to identify cell-type specific transcriptional features from open chromatin regions to predict novel expression

regulators and their targets[52,53]. DeepChrome, a deep-learning based approach to predicting expression patterns, identifies combinations of histone modifications responsible for downstream effects on gene regulation and consequently, expression[54]. The web-based Gene Friends tool (http://genefriends.org) employs a more familiar approach of network construction using RNA-seq data[55]. Using expression data for human and mouse, a genome-wide expression network was generated, upon which functional guilt-by-association relationships were identified in co-expressed genes. Further, it has been used to identify functions for novel genes based on co-expression and related transcription factors.

1.3.4 Integration of experimental and predictive data

The relationships between genes and gene products across the genome can be observed through an ever-expanding array of computational analyses- experimentally derived or formed in silico. Integrating the myriad of computational and experimental data across different methods would enable robust identification of novel cellular networks and functional components. Addressing the relevancy of any one type of observations in the context of a multi-omic network is a challenge. In refining the data that best represents a functional module or sub-network, several considerations must be made regarding the occurrence of false negatives and positives, network bias, and drawing inferences from limited or single occurrence data[56]. As an extension of systems biology, network biology integrates these data in a way that enables a researcher to examine the nodes corresponding to the overarching process in question. This is achieved by organizing the networks such that the nodes on each level correspond to a different type of observation, and the edges are representative of the relationship between them- the combined image of

which results in a pseudo-hierarchy of biological networks[57]. One of the current objectives in network biology is to determine the factors involved in the underlying connectedness between network levels- a comprehensive review of existing efforts can be found in Boucher et al, 2013[57]. GeneMANIA (https://GeneMANIA.org/) is a popular tool for integrating network data to predict new genes associated with a user- generated input list, similar to STRING[29,58]. The query genes and their interactors can be weighted based on GO terms related to biological process, molecular function, or cellular component. The tool is optimized for use with lists containing genes that are functionally related, making it an exceptional program for prediction of novel genes in a pathway. Versatility in network biology is a central consideration in applicability- researchers are employing these integrative tools for use across diverse research areas and require relevant presentation of data. Tools that incorporate predictive or in silico generated data provide an unbiased edge when selecting novel candidate genes for study. Unknown or uncharacterized genes are underrepresented in GO term enrichment, text-mining, and homology-based approaches[59]. The usefulness of computational tools in functional genomics is already apparent and is forming the groundwork for solving novel and complex questions across systems biology.

1.4 Focus and objectives

The attraction of computational tools in molecular biology is due largely in part to their potential for robust, efficient, and comprehensive approaches to molecular biology's most complex questions. The ability to predict new or alternative functions for previously uncharacterized components of a genome in a large-scale, systematic approach is powerful. Advanced mathematical models and their potential to generate interaction predictions where pre-existing data is scarce represents an exciting new era in systems biology.

To this end, a focus of this thesis was to investigate the applicability of computational analyses to our research in DNA damage repair. By utilizing predictive tools and existing experimental data, we hypothesized that we could streamline the identification of new players and/or discover genes with alternative function in in DSB repair. Novel genes involved in damage repair as well as regulation of pathway choice and signaling continue to be identified, largely due to the complex and dynamic nature of cellular damage response. Only recently was cNHEJ recognized as a pathway independent from its error-prone cohorts, following decades of experimental analyses. In the last few years, genes with little or no annotated function in DSB repair have been identified as key players in regulation and recruitment of core machinery[60]. Break and repair of DNA is ubiquitous and inevitable, with defects in the latter event imparting obvious consequence upon cellular function, proliferation, and death. DSB repair is frequently observed to be impaired in oncogenic cells, potentiating a modern class of targeted cancer therapies. NHEJ was of pertinence due to its prevalence in human cells and deleterious potential.

In this thesis, our objective was to apply several computational tools, both predictive and non-predictive, to identify genes with potential roles in NHEJ and subset a small selection of candidates for wet lab analysis. Using known major players in the NHEJ pathway as input material, we were especially interested in genes that had no indication of repair activity otherwise, as proof-of-concept. Application of traditional wet lab methods to validate our approach would be performed to assess involvement in end joining activity, sensitivity to genotoxic agents, and genetic interaction with known DNA damage and cell cycle related genes, among several others. The objectives were as follows:

1. Use a variety of data sources, both experimentally generated and computationally predicted, to narrow down a list of candidate genes whose interaction, expression, and localization data indicated involvement in NHEJ.

2. Follow up on candidates with wet-lab analysis to assess the efficacy and potential of our approach.

2. Materials and methods

2.1 Strains and plasmids

Deletion strains were obtained from the yeast gene deletion collection 2.0, in BY4741 background (MATa orfΔ::kanMX4 his3Δ1 leu2Δ0 met15Δ0 ura3Δ0)[61]. Deletion strains in BY4742 (MATα can1Δ::STE2pr-HIS3 lyp11Δ ura31Δ leu21Δ his31Δ met151Δ) background, for SGA analyses, were constructed by PCR mediated gene disruption[62]. Target ORFs were disrupted with a hygromycin B (hyg^R) resistance marker, encoded by HPH1 and flanked by sequences with short homology to the ORF. The plasmid pRS41H, a derivative of pRS40H carrying the hyg^R marker was kindly provided by Smith Lab[63]. Cassettes were transformed into cells by LiOAc/PEG/ssDNA method and successful knockouts were confirmed by colony PCR[64].

For plasmid-based repair assays, a derivative of pRS416 encoding Amp^R bacterial resistance marker and URA3 yeast selection marker was used for 5' overhang repair assays as described previously[65].YCplacIII encoding Amp^R bacterial resistance and LEU2 yeast selection marker was used for blunt end repair assays. pGV-255/LIVE and its derivative pGV-256/DEAD, containing bacterial Amp^R and yeast URA3 selection markers, were used for homologous recombination repair assays.

2.2 Selection of candidate genes

Candidate genes *GAL7*, *YHI9*, and *YMR130W* were identified by first compiling a nonredundant subset of genes with known NHEJ involvement and a set of nonessential DNA damage related genes previously identified as having significant (mutant repair efficiency<50%) impact on repair of DSBs[14]. The set was then used to generate a putative

interaction network through GeneMANIA software[58]. The resultant interactors were compared against a list of nuclear-localized genes/ ORFs with no assigned function in Saccharomyces Genome Database (SGD), and overlap between the lists was used as a primary candidate list[66]. PIPE protein interaction prediction software was utilized to investigate known and predicted interactors, the data from which was used to select final candidate genes[40].

2.3 Plasmid repair assays

Plasmids p416 or YCplacIII, were linearized by restriction digest at their XbaI or SmaI sites, respectively, and linearization was confirmed by agarose gel electrophoresis. Both sites are located within the MCS of their respective vectors, lacking homology to yeast chromosomal DNA. Approximately equal amounts of cells were transformed by LiOAc/PEG method, without single-stranded carrier DNA, and transformed with either intact plasmid or linearized plasmid[64]. Transformed cells were plated on appropriate selective synthetic media, and growth was scored after incubation at 30°C for 3 days. Repair efficiency was calculated as the ratio of colonies formed by linear transformants to colonies formed by intact transformants and related to the wild-type strain ratio. Δ*yku80* and Δ*yku70* were used as positive controls.

2.4 Homologous recombination assay

The assay, adapted from Erdemir et al, 2002, was performed using pGV-255/LIVE and pGV-256/DEAD plasmids[67]. pGV-LIVE encodes a functional lacZ gene, whereas the lacZ coding sequence in pGV-DEAD has been modified to encode a stop codon

interrupted by a unique BglII restriction site. pGV-DEAD was linearized by restriction digest at its BglII site, and linearization was confirmed by gel electrophoresis. Approximately equal amounts of cells were transformed with either pGV-LIVE alone or pGV-DEAD plus a cassette encoding functional lacZ, amplified by PCR from pGV-LIVE, by LiOAc/PEG method. Transformants were plated on synthetic dropout media lacking uracil, and incubated for 3 days at 30°C. The resultant colonies were then replicated onto fresh selective media and incubated for 2-3 days at 30°C. Colonies were then transferred to a nitrocellulose membrane and exposed to liquid nitrogen before incubation in 2% x-gal solution at 30°C for 45min-2h. The ratio of homologous recombination to non-homologous end joining events was scored by dividing the number of blue colonies over the total number of colonies on the same plate. Δ*yku80*, Δ*yku70*, and Δ*rad52* were used as positive controls.

2.5 Validation of automated colony counter

As a combination of manual and automated colony counting was employed in the plasmid repair and homologous recombination assays, it was necessary to ensure no significant differences existed between the counts obtained. The ProtoCOL3 automated colony counter was used for automated counting. Validation of results was performed as described previously using two-tailed paired t-test[68].

2.6 Drug sensitivity tests

Cell cultures were grown to saturation and serial dilutions were spotted onto a series of media containing either 80μM hydroxyurea, 0.05% methyl-methanosulfate, or exposed to 35s UV light. Plates were incubated for 2-3 days at 30°C. Spots were evaluated qualitatively for growth change relative to a no drug control and the wild-type strain.

2.7 Compilation of DNA-damage related deletion array

A deletion mutant array of 360 DNA-damage related genes was manually curated through an exhaustive search of GO term annotations, interactions with previously identified genes involved in DSB repair, and pertinent literature. The array was constructed as a subset of the *Saccharomyces cerevisiae* deletion mutant array, in BY4741 background (MATa orfΔ::kanMX4 his3Δ1 leu2Δ0 met15Δ0 ura3Δ0)[61]. Several key genes involved in NHEJ and HR were included as positive controls. Results were analyzed using SGAtools software[69]. Interactions with a score of +/- 30% relative to the control plate and appearing in 2 or more replicates were selected for GO term enrichment analysis.

2.8 Synthetic genetic array

Synthetic genetic array analyses were performed as described previously[70]. Briefly, double-mutants are generated through a series of mating, sporulation, and selection processes, and the resultant haploid cells are analyzed for growth defects to identify possible genetic interactions. To assess potential genetic interactions between candidate genes and known DNA-damage related genes, deletions of *GAL7, YHI9,* AND *YMR130W*

in BY4742 (MATα) background were crossed to the aforementioned DNA-damage related deletion array, and results were analyzed using SGAtools software[69]. Each experiment was performed 3 times.

3. Results

3.1 Generation of a DNA-damage network

We sought to utilize a computational approach in selecting genes for the study, in contrast to literature-mining or classical selection approaches. Applying the 'guilt-by-association' principle to network biology, we hypothesized that by employing simple analysis of an NHEJ-related network, we might discover novel genes involved in DSB repair. Emphasis was placed on protein-protein interactions, as these have shown to be reliable indicators of function in the past[14].

A broad list of putative candidates related to NHEJ query genes was generated through use of GeneMANIA software and BioGRID interaction database[39,58]. Since our main directive was to identify truly novel genes involved in DNA repair, we crosslinked this first result with a list of genes having no annotated/ assigned function to date- this served an additional role of addressing overrepresentation bias. Genes involved in a prior study performed by our research group were also eliminated[14]. Using PIPE interaction prediction software, we investigated the known and predicted interactors of our candidate genes to finalize selection of *GAL7, YHI9, and YMR130W.*

Table 1: Summary of rationale for candidate gene selection.

Gene Name	*H.sapiens* homolog	GO: process term	Key interactors
GAL7	*GALT*	galactose metabolism	*YKU80, CHK1*
YHI9	undetermined	unknown function	*RAD34, CDC37, COMPASS* complex
YMR130W	undetermined	unknown function	*BRE1, SSB2, DOT1*

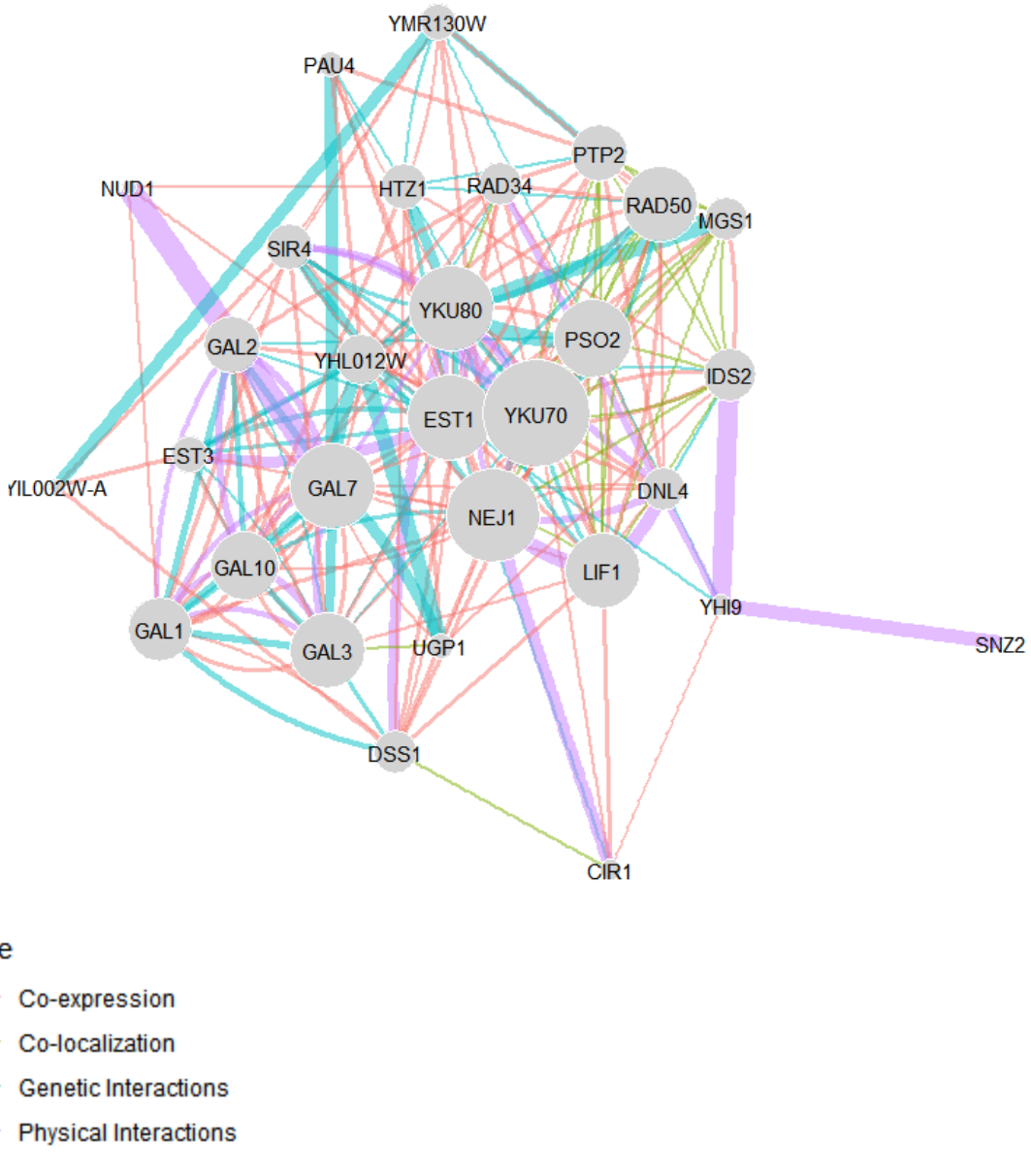

Figure 3: Simplistic overview of the network used to identify candidate genes. Node size is proportional to number of links (degrees). Edge size is representative of weight (network-equal). Datasets and weighting values were obtained from GeneMANIA and BioGRID[39,58].

3.2 Efficiency of repair by NHEJ at varying end structures

Plasmid end-joining assays were employed to assay repair efficiency by NHEJ. Circular and linearized plasmids were transformed in parallel into wild-type, *gal7*Δ, *yhi9*Δ, and *ymr130w*Δ mutant cells. Repair pathway choice is limited to NHEJ due to the location of the cut site- plasmids are linearized at regions with no known homology to the yeast genome.

3.2.1 Repair of cohesive end breaks

We first investigated the repair of short 5' overhang/ cohesive end breaks using p416 and the same plasmid linearized with restriction enzyme. Cells with defects in the NHEJ pathway are expected to display reduced survival compared to the wild-type, assessed by the normalized number of colonies formed from linear transformation divided by colonies formed from circular transformation of the same strain. Key genes such as *YKU80 and/or YKU70* are commonly employed as positive controls due to their known involvement in NHEJ[65,71]. Repair efficiency of 5' overhangs was significantly reduced in Δyhi9 cells compared to the wild type, by approximately 30%. Δyku80, as a reference, showed a reduction to approximately 5%.

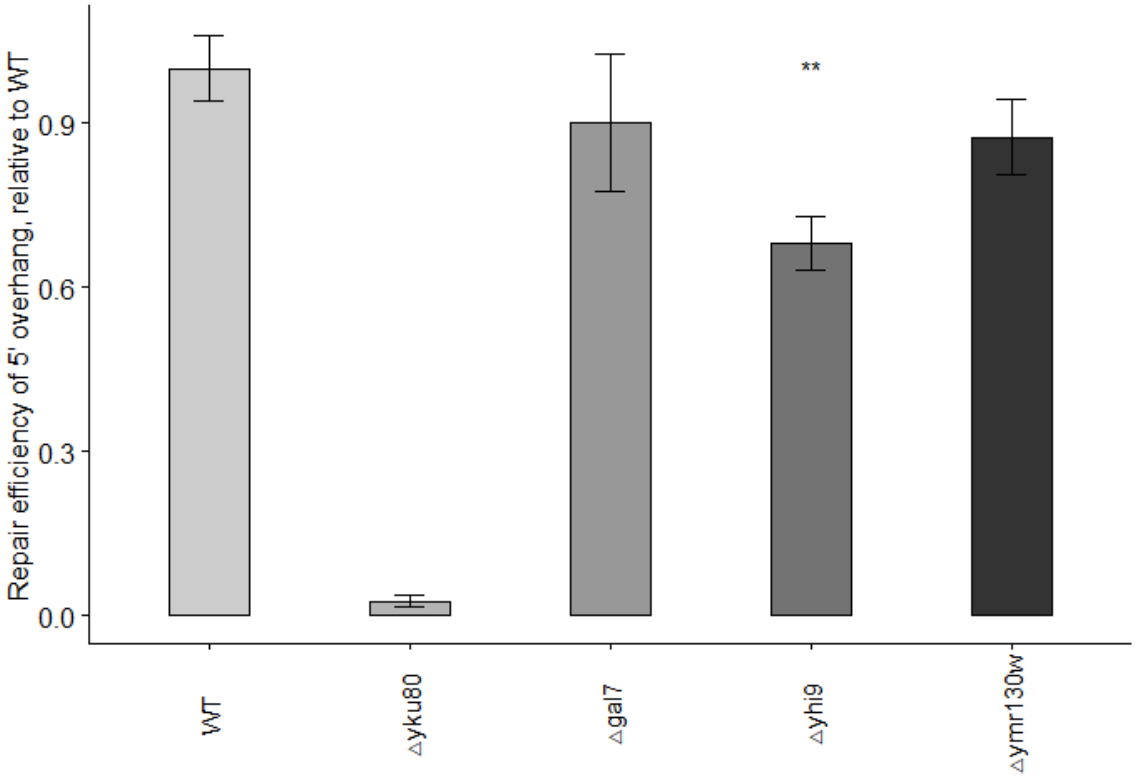

Figure 4: Repair efficiencies of mutant colonies of a 5' overhang break in a plasmid-based religation assay. The assay was performed at least 3 times. **p<0.01, Wilcoxon paired rank-sum test, Benjamini- Hochberg correction method.

3.2.2 Repair of blunt/ non-cohesive end breaks

We also investigated the repair of blunt- end breaks using the same assay described above using intact and linearized YCPlacIII. Normalized repair efficiency was calculated as the ratio of colonies formed after linearized plasmid transformation to those formed after transformation with intact plasmid.

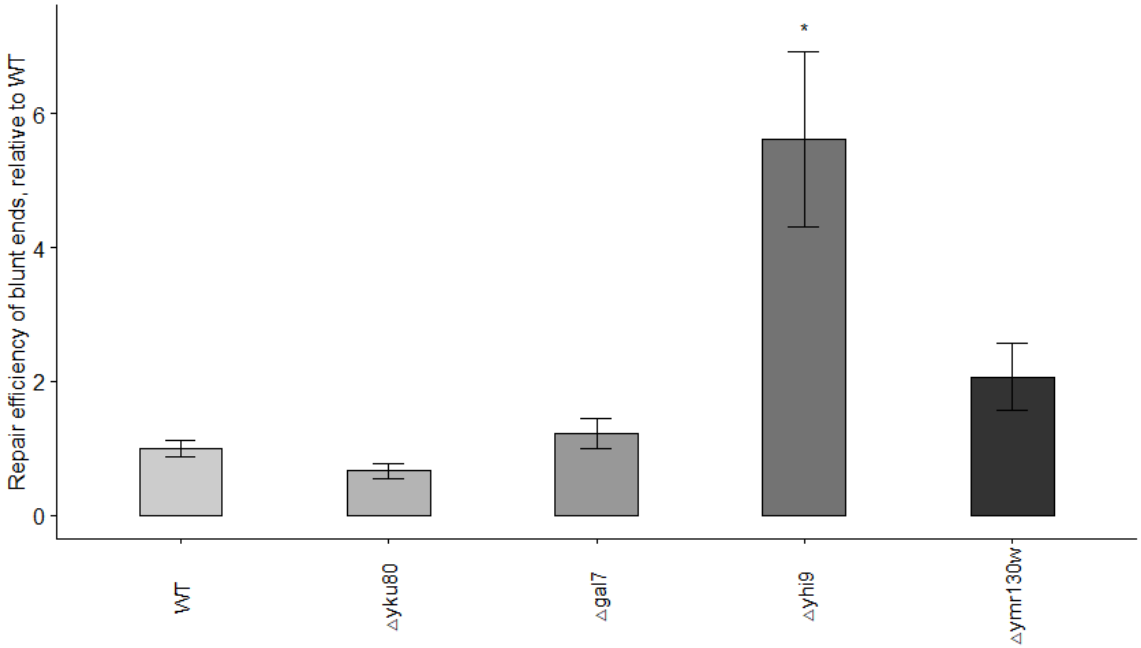

Figure 5: Repair efficiency of blunt-end DSB in a plasmid-based assay. Values represent at least 3 independent trials. *p<0.05, Wilcoxon paired rank-sum test, Benjamini- Hochberg correction method.

Δyhi9 and Δymr130w both showed an increase in NHEJ repair efficiency of blunt end breaks compared to the wild type. Δyhi9 displayed a significant increase of approximately 5-fold over the wild-type, and Δymr130w showed an increase over by wild-type efficiency by 2-fold.

3.3 Effect of gene deletion on repair pathway choice

Many genes involved in NHEJ have also been observed to confer effects on HR when deleted, and vice versa[8,67,72]. The MRX (MRN in *H. sapiens*) complex, for example, has been implicated in both HR and NHEJ[73]. Based on interactions observed with HR related genes from our original network screening, we sought to determine whether our

candidate genes might also play a role in HR. In this assay, mutant and wild type cells are transformed with both a circular plasmid containing a *lacZ* coding region (pGV-255/LIVE), as well as a modified version contained a disrupted *lacZ* frame(pGV-256/DEAD) [67].This disruption is such that the cut site is flanked by residues encoding a STOP codon. The 'live' plasmid is used as a control to assay production of beta-galactosidase, observed as a blue coloured colony in the presence of x-gal substrate. Simple end-joining of the 'dead' plasmid results in no production of beta- galactosidase by the cell, observed as a white coloured colony. The latter is co-transformed with a cassette containing an intact *lacZ* gene, thereby providing the cell an option of repair by either NHEJ or integration of the cassette by HR. The ratio of HR to NHEJ events is assayed by the number of blue colonies over the total number of colonies on the same plate.

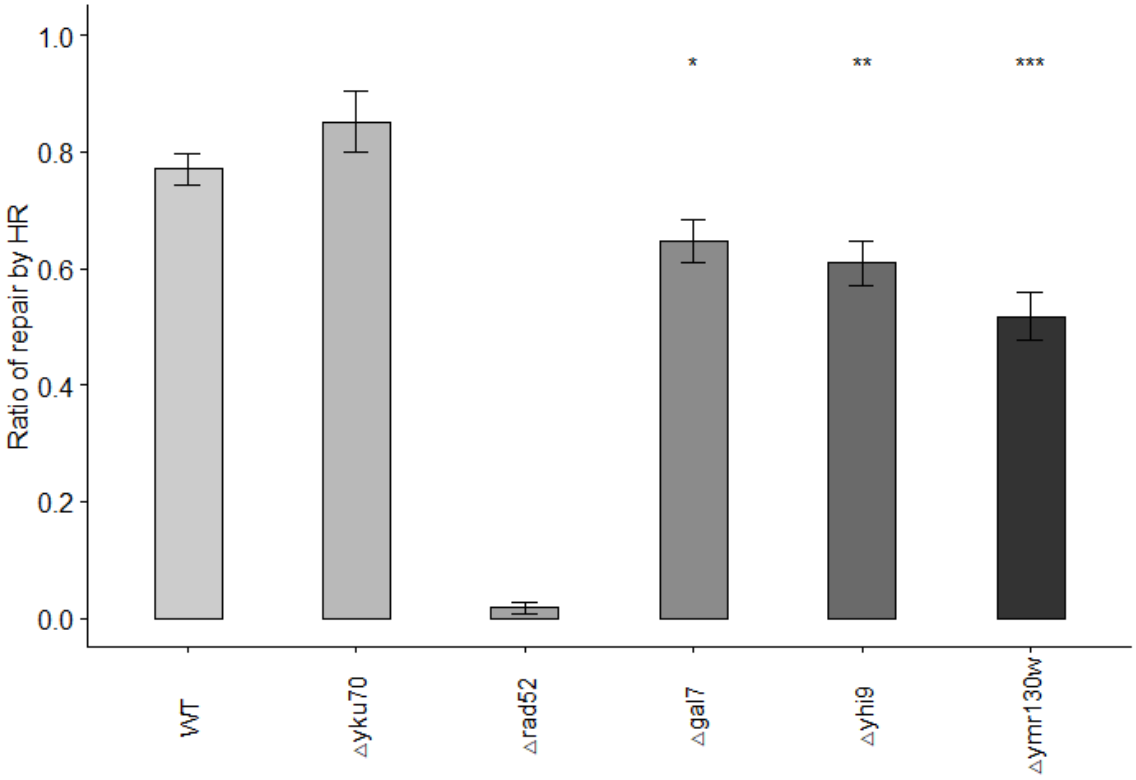

Figure 6: Ratio of colonies expressing functional lacZ product, indicative of repair by HR, to colonies performing repair by NHEJ. Error bars represent S.E.M, where n=4. *p<0.05, **p<0.01,*p<0.001, unpaired t-test with Benjamini- Hochberg correction.**

Δyku70 and Δrad52 mutants were employed as positive controls, with the former modestly increasing or displaying repair by HR equivalent to wild type, and the latter displaying severe HR defects as a reference. All three candidates displayed consistent significant HR defects compared to the wild type, in which HR comprised approximately 80% of total repair events. *gal7*Δ mutants, while having no observed effect on NHEJ in previous end-joining assays, resulted in 65% HR efficiency. *yhi9*Δ and *ymr130w*Δ mutants displayed 60% and 50% total HR events, respectively. These observations indicate involvement in the HR pathway.

3.4 Sensitivity of mutants to DNA-damaging treatment

Known DNA-damaging agents methyl- methanesulfonate (MMS), hydroxyurea (HU), and UV light exposure were used to evaluate the sensitivity of the cell upon deletion of the mutants *gal7Δ, yhi9Δ,* and *ymr130wΔ.* In general, deletion of a gene involved in DNA repair is expected to cause an increase in sensitivity to DNA damaging agents, observable thorough growth or phenotypic defects. HU limits the dNTP pool at replication forks in S phase, slowing cell cycle progression and causing replication fork collapse in checkpoint-impaired cells[74]. UV radiation is the most ubiquitous cause of chromosomal damage, producing pyrimidine dimers and stalling RNA polymerase II at the lesions, which are repaired primarily through nucleotide excision repair[75]. MMS is an alkylating agent thought to stall replication and induce base mispairing events, activating cell cycle checkpoint response.

gal7Δ displayed increased sensitivity compared to the wild type across all 3 treatments, whereas *yhi9Δ* and *ymr130wΔ* cells did not appear to confer substantial changes in sensitivity when compared to wild type cells after HU or UV exposure. *yhi9Δ* and *ymr130wΔ* displayed a perceptible increase in sensitivity to MMS over the wild type, though further analysis is required to ascertain whether these are true phenotypes or a confounding variable of the technical limitations of the assay.

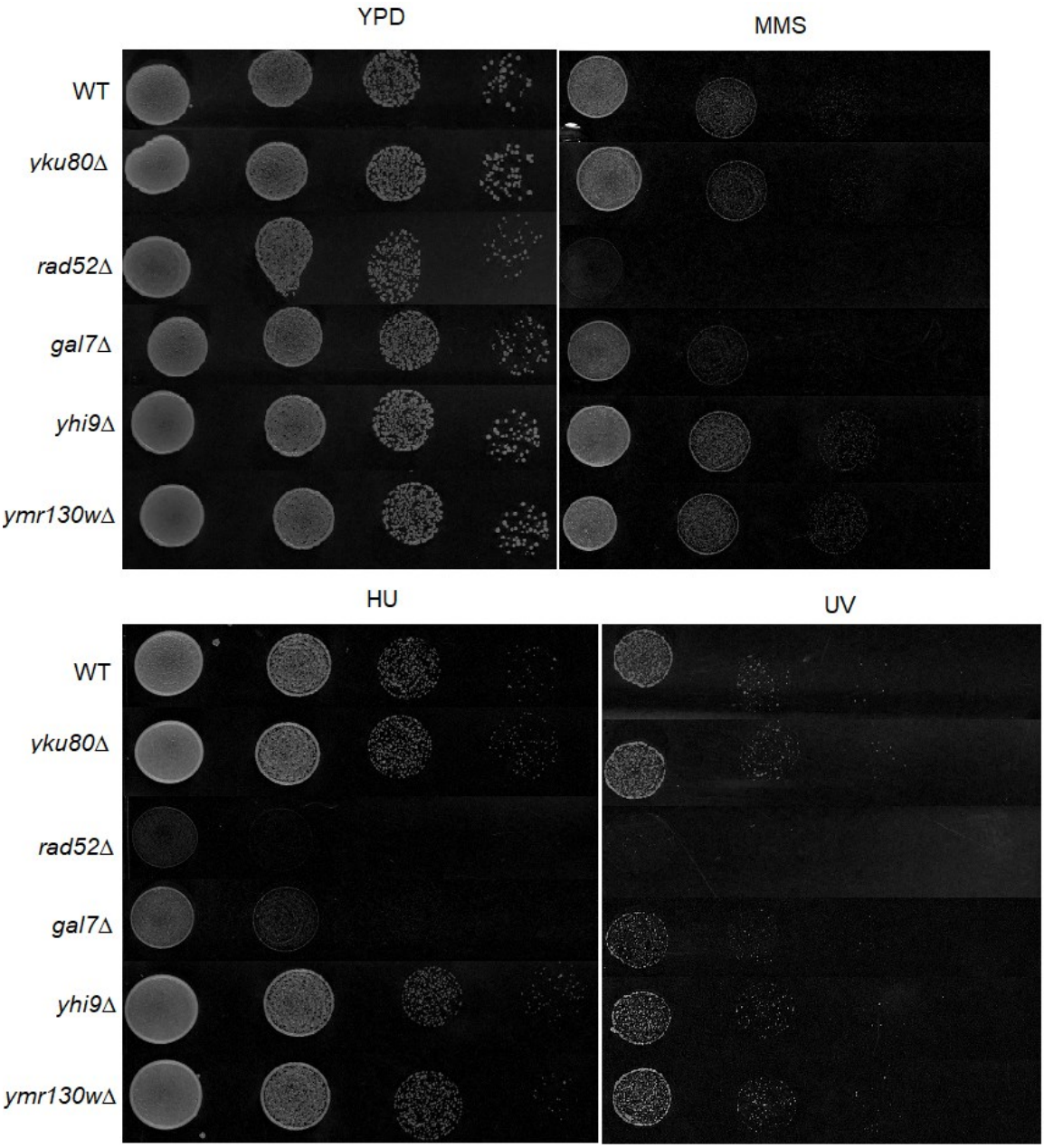

Figure 7: Sensitivity of gene mutants to several DNA damaging drugs was assayed by spot test. Each screen was performed in triplicate, and representative images are shown above, edited for contrast and visibility.

3.5 Genetic interaction with damage-repair associated genes

Synthetic genetic array is a widely- used screening tool for genetic interactions in *S. cerevisiae.* Genetic interactions form the overarching understanding of functional pathways in the cell, and by assessing these interactions we can evaluate functional relationships between genes. A negative interaction is observed by defects in growth or lethality in the cell upon deletion of both genes and indicates that the genes may be involved in parallel functional pathways (i.e. the double deletion renders the cell unable to compensate for deletion of either gene)[70]. A positive interaction is considered to present a phenotype that is less severe than would be expected. Using SGA, we analyzed the candidates for genetic interactions with known DNA repair genes by mating mutant strains to an array of around 370 genes with previously characterized involvement in DNA repair and cell cycle checkpoint regulation. Following a series of selection for double mutants, fitness was assessed using SGAtools software and gene ontology process terms associated with the interactors was obtained using SGD YeastMine[66,69].

The interaction network for *GAL7* included genes involved in the damage checkpoint, transcription regulation in response to damage, homologous recombination repair, nucleotide excision repair, and RNA polymerase II mediation and assembly. *GAL7* showed strong negative interactions with *SIR2, SIR3*, and *SIR4*, members of the Sirtuin family involved in telomeric silencing and cellular aging, defects in which confer gross chromosomal rearrangements. Interaction with DSB-related DNA helicase *SRS2* and chromatin remodeling genes including *ARP8* and recently characterized NHEJ gene *HUR1* were also observed[76].

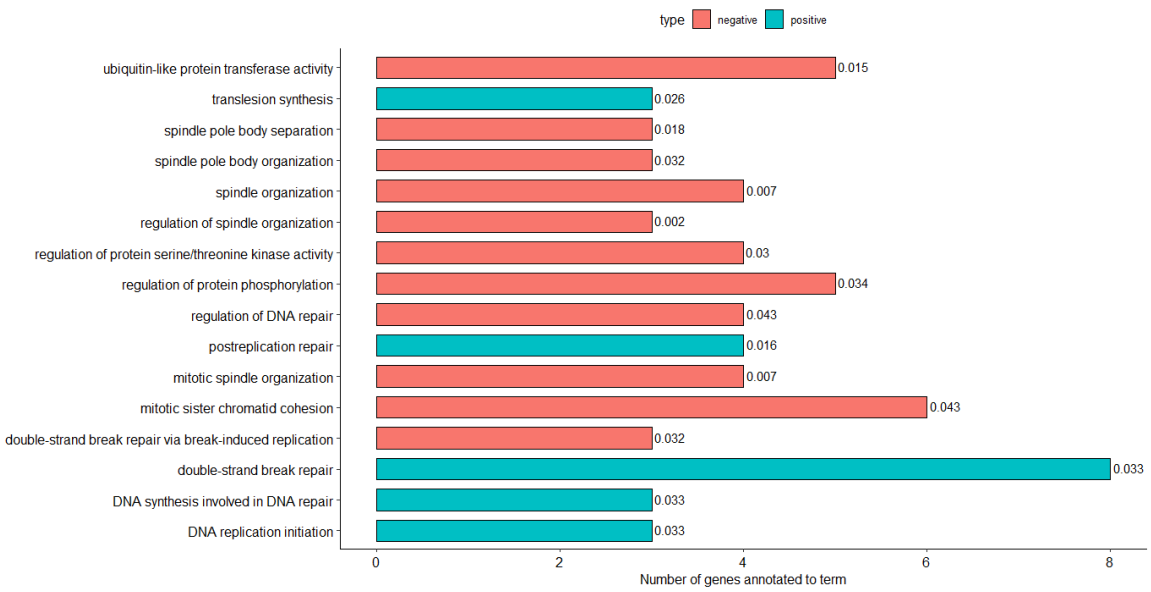

Figure 8: Results of SGA analysis of a *GAL7* mutant strain, presented as GO biological process and molecular function terms that were significantly enriched in the screening. Selection was performed 3 times, and interactions that displayed a growth defect or enhancement of at least 30% in 2 or more trials were used for term analysis. P-values of term enrichment were corrected against enrichment for the plate background.

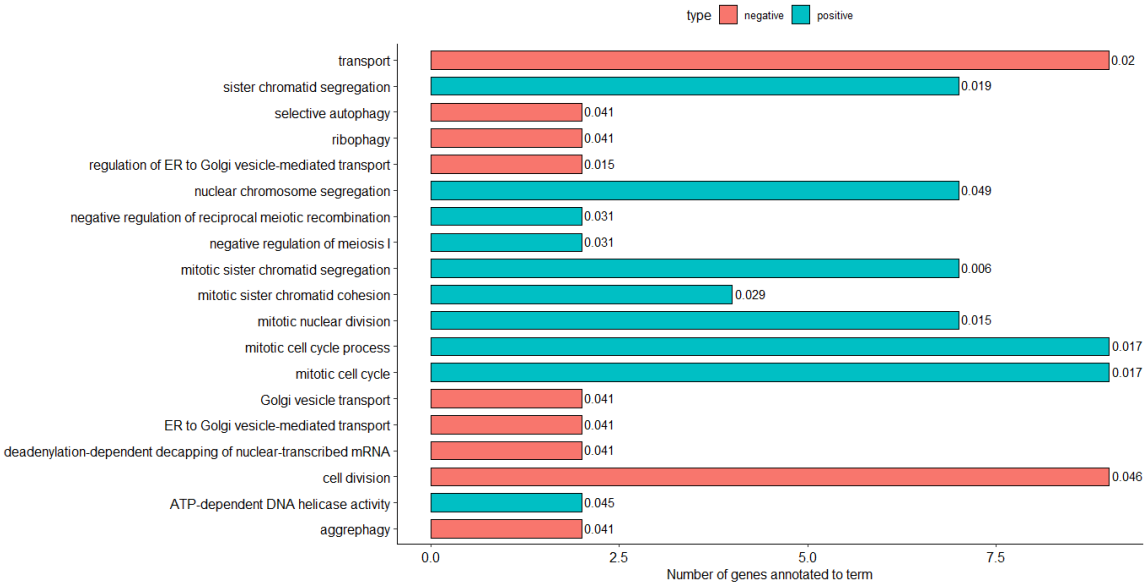

Figure 9: Results of SGA analysis of YHI9 mutant strain, presented as GO biological process and molecular function terms that were significantly enriched in the screening. Selection was performed 3 times, and interactions that displayed a growth defect or enhancement of at least 30% in 2 or more trials were used for term analysis. P-values of term enrichment were corrected against enrichment for the plate background.

Interactions with several members of the *RAD* family of genes were observed from analysis of *YHI9*- among them checkpoint protein *RAD17*, nucleotide excision repair proteins *RAD14, RAD27* and *RAD33*, and recombinational repair protein *RAD54*. Like *GAL7*, *YHI9* also interacted negatively with members of the Sirtuin family.

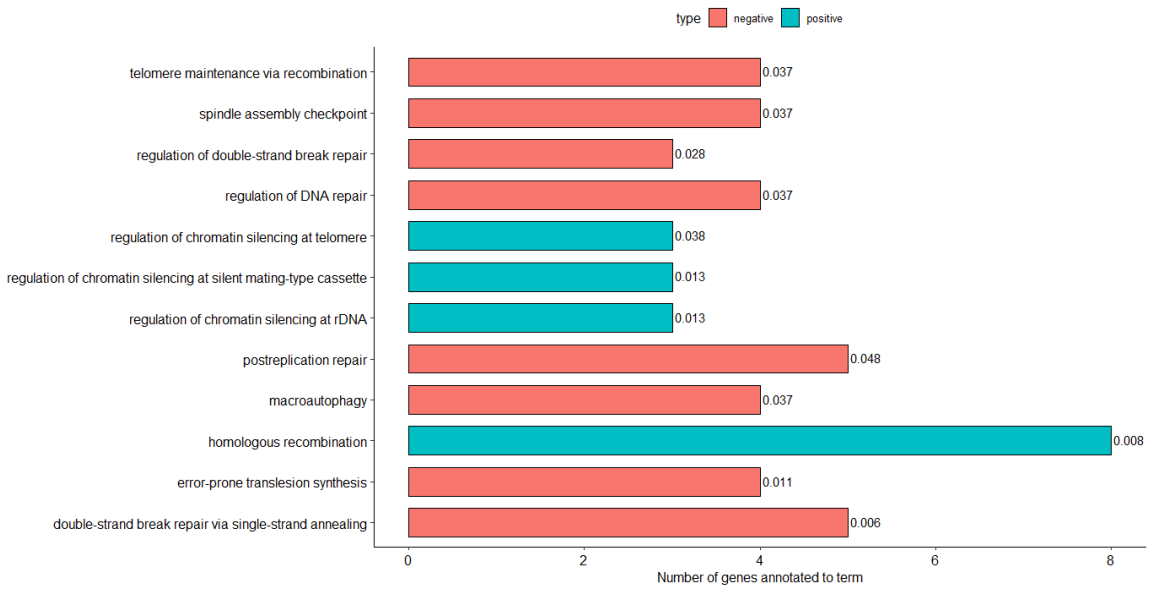

Figure 10: Results of SGA analysis of YMR130W mutant strain, presented as GO biological process and molecular function terms that were significantly enriched in the screening. Selection was performed 3 times, and interactions that displayed a growth defect or enhancement of at least 30% in 2 or more trials were used for term analysis. P-values of term enrichment were corrected against enrichment for the plate background.

YMR130W interacted with several genes involved in cell cycle damage checkpoints, including *DDC1*, *CDC73*, and cyclins associated with cell cycle progression and replication (*CLB5, CLN3)*. *MMS22*, a gene involved in replication repair as part of the Cul8- RING DSB repair complex, was identified as a negative interactor. These interactions suggest a potential role for *YMR130W* in checkpoint response.

4. Discussion

4.1 Utilizing existing and predictive data to identify novel genes in DSB repair

Though relatively young in the scope of systems biology, computational network biology is already becoming a cornerstone in biological research techniques. As discussed, the pervasive end goal of network analysis and functional biology is the completion of a 'definitive' integrated biological network, capable of data transfer across species. This requires functional characterization of all complexes and modules that underly cellular function alongside associated inter or intra- molecular interactions mediating this function. Certainly, high-throughput tools have made large scale screening and data collection easier. It is, however, difficult to imagine the concept of generating the amount of quality information required for a complete cellular network through purely experimental processes. Yes, experimental analysis remains a core tenet of systems biology, and is an essential step when validating computationally- derived data or hypotheses. The issues of researcher bias, time, cost, and inter-lab variability/ reproducibility inherent to biological research must be addressed. Computational tools do not entirely fix these problems; however, they are arguably essential if full genomes are to be understood. Annotation of the of *S. cerevisiae* functional network remains incomplete despite decades of functional research- in part, perhaps, due to lack of characterization of the full genome.

Genes or proteins with well-studied cellular roles or interactions present themselves across more databases and literature searches than their non-annotated counterparts, leading unknown genes to remain as such while literature on 'popular' genes grows.

Integration/ curation of large- scale experimental sets to create informative biological networks provides researchers with powerful investigative tools, though they too are at the mercy of data bias. GeneMANIA, for example, does an excellent job of presenting cohesive, relevant information in network form through automated weighting and data collection. The source of the data and visualization must be carefully interpreted, though, taking into consideration the nature of the program- the output is dependent almost entirely on information contained in experimental catalogues[58]. The predictions are therefore a reflection of available knowledge (it is difficult to predict functions for an uncharacterized protein for which little experimental data is available, and by extension few interactions reported) and can be highly variable depending on the weighting option used and the size and relatedness of the query data. For example, *GAL7* was initially identified through a physical interaction with *YKU80*. This interaction is reported once in the GeneMANIA dataset, from a large scale affinity- capture experiment and was assigned a low weight value[77]. Input of *GAL7* OR *YKU80* alone into the algorithm will not return this interaction in the network- much of the weight/ value attributed to edges in the network is dependent on the functional relatedness of the input genes, and in ontology-directed weighting is dependent on prevalence.

Use of a predictive interaction tool may alleviate the issues of reproducibility and lack of available experimental data. In silico prediction tools often address this by utilizing sequence data from known interactors as a base input, and experimentally confirmed data to validate the approach after predictions are generated. Results are therefore still dependent on reported data to some degree depending on the algorithm but may be less

prone to misguided interpretation than the aforementioned approaches. Algorithms ranging from identification of short, co-occurring sequence patterns (peptides/ motifs) to sophisticated machine-learning models can be applied to predict novel interactions associated with genes, complexes, or pathways of interest across a broad range of species. Recently, a predictive software was claimed to have enabled the first all-to-all run of the human protein interaction network through recognition and application of short linear motif sequences[40]. The software, PIPE, was utilized here to evaluate the potential involvement of uncharacterized genes in the NHEJ pathway. Genes *YHI9* and *YMR130W* had no functional annotation assignment, and little experimental data was available. *GAL7* had no annotations associated with NHEJ or HR, and is one of the most extensively studied genes in yeast in the context of galactose metabolism[78].
Owing to the reliability of protein-protein interactions as indicators of molecular function, and their overlap with genetic interactions, we hypothesized that the results of a PPI prediction software (PIPE) would enable confident prioritization of genes for wet lab analysis.

GAL7 belongs to a group of genes involved in galactose metabolism and by extension one of the most extensively annotated pathways in *S. cerevisiae*. A physical interaction with *YKU80*, a protein with critical involvement in NHEJ in both yeast and higher eukaryotes, was intriguing. Caution was exercised in assigning importance due to the singular nature of the observation[77]. Manual searching through the SGD catalogue revealed that several other *GAL* genes had been characterized in roles outside of their classical assignment. *GAL11* has been identified as a member of the RNA polymerase II

mediator complex, a core component of the signal transduction and recruitment machinery. It also physically interacts with proteins involved in NER and checkpoint arrest recovery. *GAL83* is under study as a kinase subunit, and physically interacts with *RAD30, RAD14,* implicating involvement in post-replication repair, NER, and repair of UV induced dimers. Additional genetic and physical interactions with key repair genes on both the known and predicted level led us to hypothesize novel function for *GAL7* in non-homologous ending joining.

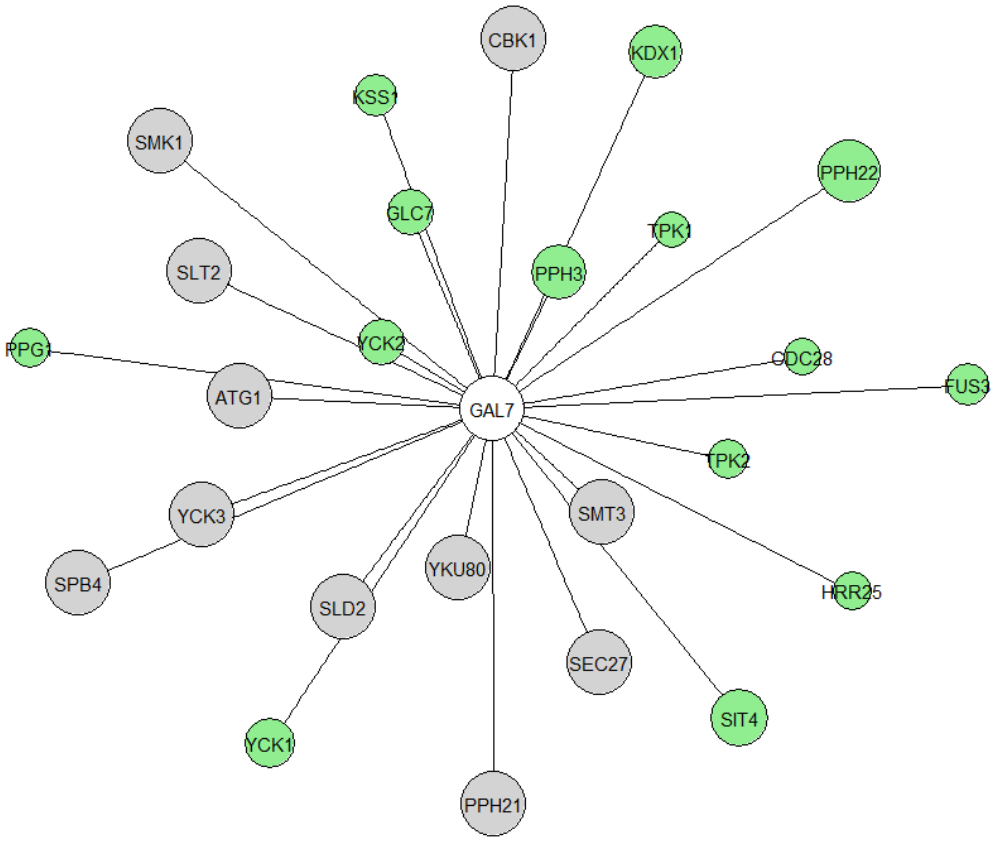

Figure 11: Protein-protein interactions for candidate gene *GAL7* identified through PIPE software[40]. Predicted (novel) interactors are indicated as green, known interactors are indicated as grey. Node size is representative of the PIPE-generated interaction score.

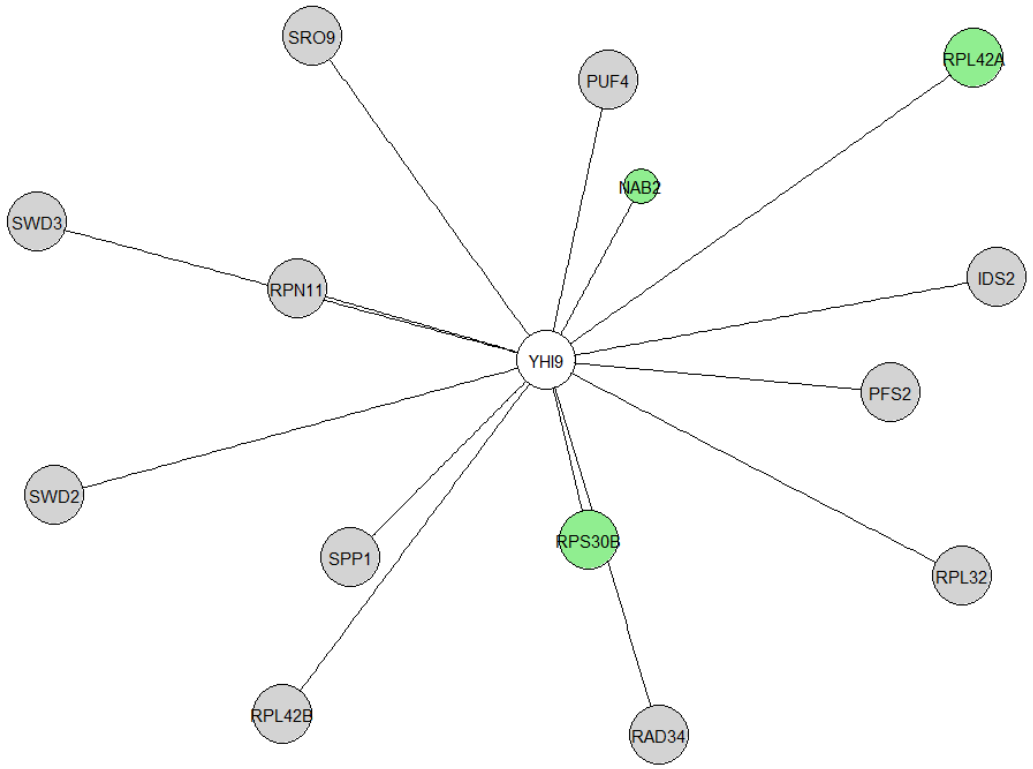

Figure 12: Protein-protein interactions for candidate gene *YHI9* identified through PIPE software[40]. Predicted (novel) interactors are indicated as green, known interactors are indicated as grey. Node size is representative of the PIPE-generated interaction score.

In contrast to the abundance of data available on *GAL7*, albeit of alternative focus, minimal data was available on *YHI9*. The initial network screen identified physical interactions with *RAD34* (nucleotide excision repair, homologous to *RAD4*), and EST1 (telomere length regulation), otherwise support for involvement in NHEJ was limited to interactions of secondary degree (e.g. no direct interaction, rather interactions with *YHI9* neighbours). Follow-up manual literature searches indicated interactions with *CDC37* (cell division) and members of the COMPASS complex (transcriptional silencing at

telomeres), indicating putative roles for *YHI9* in cellular stability and telomere maintenance.

PIPE results were not particularly supportive either. A larger list of interactors might be generated by lowering the specificity of the algorithm, but at the expense of increasing potential for false positive predictions. *YHI9* at first glance had little apparent support for NHEJ involvement and its was chosen primarily to test proof-of-concept of our approach.

YMR130W, like *YHI9*, was not well-annotated and presented itself as a truly novel candidate for involvement in DNA damage repair. A short ORF encoding a hypothetical protein, initial screening did not identify any first- degree physical interactions with well-known NHEJ or HR proteins, though co-expression was observed with NEJ1, YKU70, and members of the Sirtuin family. A study published after selection of *YMR130W* presented the gene as a member of the mitochondrial proteome, functioning as a mitochondrial import protein. PIPE screening predicted interactions with nuclear ribosomal proteins (*RUB1, RPL40A, RPL40B, NAB2*) functioning primarily in mRNA transport.

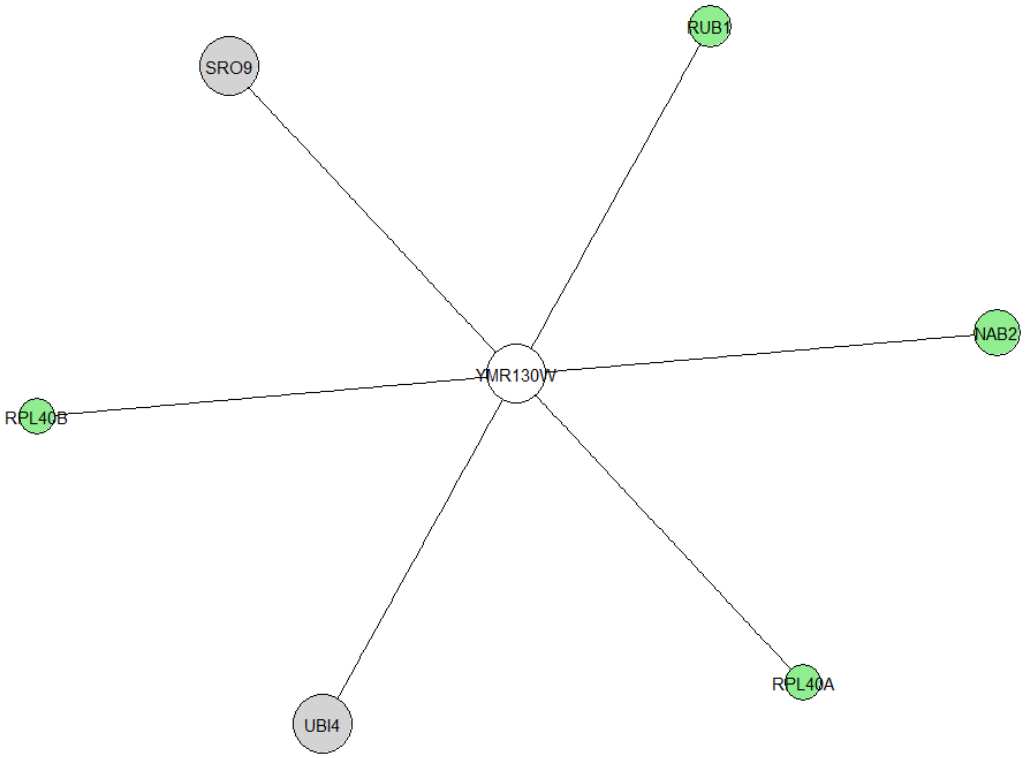

Figure 13: Protein-protein interactions for candidate gene *YMR130W* identified through PIPE software[40]. Predicted (novel) interactors are indicated as green, known interactors are indicated as grey. Node size is representative of the PIPE-generated interaction score.

4.2 Repair of 5' overhang and blunt end extrachromosomal breaks

GAL7, *YHI9*, and *YMR130W* were assayed for involvement in NHEJ with a plasmid-based repair assay. It was hypothesized that deletion of these genes would affect efficiency of NHEJ repair, following their identification through an NHEJ- enriched network. Repair of 5' overhangs in this assay is thought to be performed by simple religation due to the cohesive nature of the ends- extensive end processing is not observed and is further thought to be inhibited by the *YKU80/YKU70* heterodimer[79].

Deletion of *YHI9* led to a significant decrease in repair efficiency (68%) compared to the wild type, though the severity of impairment was not comparable to that of *yku80Δ* (approximately 6%).

To further investigate the results, the assay was also performed to assess efficiency of repair of blunt end/ non-cohesive breaks in mutant cells. Blunt ended breaks undergo extensive end processing before rejoining can occur, therefore successful repair may involve NHEJ-related factors independent of the *YKU80/YKU70* heterodimer. *ymr130w*Δ mutants displayed a potential increase in NHEJ efficiency of 2- fold compared to the wild type and *gal7*Δ cells displayed a 1.2-fold ratio of repair relative to the wild type. *yhi9*Δ mutants displayed a substantial increase in repair of blunt ends by 5.6-fold that of the wild type. It is of note that *yku80*Δ mutants display a much less severe defect in repair of non-cohesive ends than is to be expected. Past studies have suggested this is due to the protective nature of the complex- binding of the KU heterodimer is thought to prevent end resectioning and may therefore diminish repair efficiency at DSBs where HR is not possible[15,65,79]. These findings suggest a potential role for *YHI9* in the NHEJ pathway, possibly independent of c-NHEJ.

4.3 Analysis of pathway choice in repairing plasmid- based breaks

A homologous recombination assay was performed to evaluate the effect of gene deletion on repair through HR[67]. Association of members of the RAD group of genes with candidate genes in the network screening provided a broad implication in DNA repair, and we sought to assess whether *GAL7, YHI9,* and *YMR130W* were involved in repair

independent of NHEJ. All three genes showed significant impairment of HR. In wild type cells, HR events comprised 77% of total repair events, whereas HR events in *gal7Δ*, *yhi9Δ*, and *ymr130w*Δ comprised 65%, 61%, and 52% of total events. This finding indicated that our initial hypothesis of involvement in NHEJ might be expanded to consider involvement in HR or alternative repair pathways.

4.4 Sensitivity of mutants to damage- inducing treatments

Drug sensitivity assays were performed using known DNA damaging agents to assess the effect of candidate gene deletion on cellular sensitivity to various forms of chromosomal damage. Hydroxyurea acts by limiting the dNTP pool available during S phase, with cycle checkpoint mutants incurring a high incidence of collapsed replication forks. Methyl- methanesulfonate is an alkylating agent commonly used as an anti-cancer treatment, causing base mispairing events and blocks during replication[80]. MMS also has an observed increase in effect upon cells in S phase[81]. Its mechanism of action has been contested- it has been described as both a cell-cycle dependent DSB precursor and incapable of DSB induction. The latter study proposed the observation of DSB formation following MMS treatment was a combined effect resulting from sample preparation methods[80,81].

Damaging agents induce various lesions on a chromosomal level through different modes of action, therefore it is only possible to draw inferences regarding effect on overall repair mechanisms rather than specific pathways. While limited in scope and specificity of conclusions due to the nature of analysis by visual inspection, the assay resulted in several key observations. *gal7*Δ cells were highly sensitive to MMS and HU, indicating a

potential influence on NER pathway. Examination of *gal7*Δ mutants in a more specific manner would help to determine if these results are a result of direct involvement in DNA repair or a confounding effect of metabolic state on repair events.[82] While no defect was visible in *yhi9*Δ and *ymr130w*Δ mutants under exposure to HU, both appeared to decrease sensitivity to MMS in comparison to the wild type. Validation of this result has the potential to implicate both *YHI9* and *YMR130W* in cell-cycle dependent DSB repair. The results for *GAL7* indicate an influence on DNA repair in a manner divergent to that of *YHI9* or *YMR130W*.

4.5 SGA reveals genetic interactions with known genes involved in damage repair and cell cycle checkpoints

Genetic interactions are the building blocks of biological interaction networks- they can describe overarching functional relationships between genes and complexes. Negative interactions are identified through phenotypic defects conferred upon the cell when both interactors are deleted, indicating involvement in parallel pathways. The double deletion renders the cell unable to compensate for the loss of function through compensation. Positive interactions confer a less severe phenotype than expected, indicating functional redundancy and by extension involvement in the same or similar pathway. Positive interactions have additionally been shown to detect genetic relationships between genes involved in the same complex[28].

Negative interactions for *GAL7* included members of the Sirtuin family- *SIR2* performs a key role in regulating replication initiation during S phase following replication fork stall

or DSB- induced collapse. The activity of *SIR2* in DNA stability is essential in maintaining rDNA repeats and regulating use of sister chromatids as HR templates to this end. Increased sensitivity to MMS and UV light has been observed with a loss of rDNA repeats attributed to mis regulation of HR activity. Further investigation of *GAL7* in the HR pathway in this context may reveal novel relationships between cellular metabolism, senescence, and DNA repair. Positive interactors included *RAD30*, *YKU80*, *YKU70*, and *YHI9*.

YHI9 SGA analysis revealed a high number of genes enriched in binding activity- negative interactions were observed with protein binding, mismatched base binding, and double-stranded DNA binding. Interestingly, positive interactions were identified in telomere binding. Positive interaction with *YKU80* may be a source of further investigation due to function in telomeric maintenance, and the genetic interaction between the genes may be representative of involvement in the same pathway or complex. It is important to consider how these findings might be validated in the context of unassigned function- generally these enrichments would be compared against published GO terms to evaluate the reliability of the assay. *YHI9* has yet to be assigned function aside from a potential GO functional annotation for DNA topoisomerase activity, and thus it is difficult to ascertain the validity of the interactions.

Genetic interactions between *YMR130W* with *RAD30* and *MMS4* were scored as positive, consistent with the results of the HR assay. Unlike *GAL7*, negative interactions with *SIR2*

and *SIR3* were observed, and may provide broad context for the results of the drug sensitivity assay.

5. Conclusion and future directions

DNA damage is constant and unavoidable- a human cell is subject to thousands of lesions per day and must repair these lesions faithfully if it is to survive. The majority of breaks occurring in a eukaryotic cell do not threaten integrity of the cell, and are repaired efficiently and accurately[73,13]. DSBs, however, are a severe form of lesion that must be repaired to avoid gross chromosomal rearrangements, mutagenic, or oncogenic events. Loss of or compromised function in the various repair pathways employed by the cell has direct implications in cancer, disease, and cell death.

It has become clear in recent years that the eukaryotic DNA damage response is highly complex and more dynamic than previously thought- genes with novel direct functions in HR, NHEJ, alt-NHEJ or regulatory function are continually identified. The consistent growth of the DNA damage repair network indicates existence of yet- uncharacterized genes involved in DSB repair. We sought to use a systems biology approach to this phenomenon, incorporating computational tools, network analysis, and predictive software to identify novel genes involved in DNA repair in *S. cerevisiae*.

Conservation of the DSB repair pathway and its key complexes is highly conserved between yeast and human, which enables study of these novel genes across the two species and the ability to transfer resultant interaction networks. The generation of a complete human interactome relies heavily on study of model organisms- compiling a high confidence biological network in yeast is a stepping stone to performing the same feat in the highly complex human cell[83]. We utilized several computational tools to

identify 3 genes potentially involved in NHEJ repair- *GAL7, YHI9,* and *YMR130W*. *GAL7* was well characterized for its role in galactose metabolism but had no prior annotations associated with DNA repair. *YHI9* and *YMR130W*, respectively, were genes of unknown function with minimal appearances in the literature. Follow-up with wet lab analyses was successful in providing context to our computational screens and revealed potential novel functions for all three genes. Plasmid- based repair assays demonstrated varying involvement of candidates across DSB end types and pathway choice, while drug sensitivity analysis presented intriguing results regarding repair defects in the gene mutants. Genetic interactions ascertained through SGA analysis revealed relevant connections to known DNA damage repair genes- expansion of this study to increase the array of genes analyzed across damage conditions would help to further characterize *GAL7, YHI9,* and *YMR130W*.

The implications of this screening are simple- continued discovery of novel genes influencing DNA damage repair as a primary function or as a moonlighting role highlights the necessity of continued study. It also represents, at a basic level, a modern approach to solving questions in systems biology. The network screening employed in this study was simplistic and as such would be unsuitable for large scale application- the high level of precision and fine-tuning required to construct and/or integrate reliable genome- wide or cross species interaction networks should not be disregarded. The incidence of 'moonlighting', or of genes performing alternative functions outside of their canonical role, is unable to be assayed using the techniques described in this study despite the likelihood relevance. In addition, novel genes identified in DSB repair may be

involved in more recently identified (and therefore less well characterized) pathways, and assays that are able to capture more detail regarding repair function would provide strength to the findings presented in this work.

GAL7, for example, is routinely understood to be transcriptionally repressed under glucose conditions- this presents a conundrum considering the results obtained here. RT-PCR or similar assays of *GAL7* under DNA damage conditions is necessary to solve this discrepancy and elucidate the relationship between metabolic state and DNA repair. Further, none of the classical repair assays were able to identify a role for *GAL7* related to its reported physical interaction with NHEJ protein *YKU80*. Next steps might involve conditional SGA in a chemical genomics context, involvement in break repair on a chromosomal level, and generation of double mutants to narrow down the mechanisms behind the observations of this study. *YHI9* is a strong candidate for more directed analysis on a chromosomal level and in a cell cycle dependent context. *YMR130W* analyses produced results indicative of a role in HR in a context outside of traditional repair analyses- study of both *YHI9* and *YMR130W* would benefit greatly from large-scale screening such as yeast-two-hybrid or CHIP-seq to gather more information on these uncharacterized genes and place them in a genome- wide functional context. Notably, *YMR130W* is a currently listed on SGD and UniProt as a short, putative ORF (~300bp)- structural studies to confidently identify the nature of this ORF are necessary[38,66].

References

1. Reay, D. I'd whisper to my student self: you are not alone. *Nature* **557,** 160–161 (2018).
2. Ciccia, A. & Elledge, S. J. The DNA damage response: making it safe to play with knives. *Mol. Cell* **40,** 179–204 (2010).
3. Sotiriou, S. K. *et al.* Mammalian RAD52 Functions in Break-Induced Replication Repair of Collapsed DNA Replication Forks. *Mol. Cell* **64,** 1127–1134 (2016).
4. Bedford, J. S. *et al.* From DNA damage to chromosome aberrations: Joining the break. *Mutat. Res. Toxicol. Environ. Mutagen.* **756,** 5–13 (2013).
5. Tsabar, M. & Haber, J. E. Chromatin modifications and chromatin remodeling during DNA repair in budding yeast. *Curr. Opin. Genet. Dev.* **23,** 166–173 (2013).
6. Li, X. C. & Tye, B. K. Ploidy dictates repair pathway choice under DNA replication stress. *Genetics* **187,** 1031–40 (2011).
7. Kostyrko, K. & Mermod, N. Assays for DNA double-strand break repair by microhomology-based end-joining repair mechanisms. *Nucleic Acids Res.* **44,** e56–e56 (2016).
8. Chapman, J. R., Taylor, M. R. G. & Boulton, S. J. Playing the End Game: DNA Double-Strand Break Repair Pathway Choice. *Mol. Cell* **47,** 497–510 (2012).
9. Mathiasen, D. P. & Lisby, M. Cell cycle regulation of homologous recombination in *Saccharomyces cerevisiae*. *FEMS Microbiol. Rev.* **38,** 172–184 (2014).
10. Dudáš, A. & Chovanec, M. DNA double-strand break repair by homologous recombination. *Mutat. Res. Mutat. Res.* **566,** 131–167 (2004).

11. Omidi, K. Identification and characterization of novel genes involved in DNA double strand break repair process in the yeast Saccharomyces cerevisiae. (2017).

12. Jiang, G. *et al.* BRCA1-Ku80 protein interaction enhances end-joining fidelity of chromosomal double-strand breaks in the G1 phase of the cell cycle. *J. Biol. Chem.* **288,** 8966–76 (2013).

13. Lieber, M. R. The Mechanism of Double-Strand DNA Break Repair by the Nonhomologous DNA End-Joining Pathway. *Annu. Rev. Biochem.* **79,** 181–211 (2010).

14. Jessulat, M. *et al.* Spindle Checkpoint Factors Bub1 and Bub2 Promote DNA Double-Strand Break Repair by Nonhomologous End Joining. *Mol. Cell. Biol.* **35,** 2448–63 (2015).

15. Menon, V. & Povirk, L. F. End-processing nucleases and phosphodiesterases: An elite supporting cast for the non-homologous end joining pathway of DNA double-strand break repair. *DNA Repair (Amst).* **43,** 57–68 (2016).

16. Karathanasis, E. & Wilson, T. E. Enhancement of Saccharomyces cerevisiae End-Joining Efficiency by Cell Growth Stage but Not by Impairment of Recombination Elissa. *Genetics* **161,** 1015–27 (2002).

17. Hoeijmakers, J. H. J. DNA Damage, Aging, and Cancer. *N. Engl. J. Med.* **361,** 1475–1485 (2009).

18. Lieber, M. R. Pathological and Physiological Double-Strand Breaks: Roles in Cancer, Aging, and the Immune System. *Am. J. Pathol.* **153,** 1323–1332 (1998).

19. Lim, D. S. *et al.* Analysis of ku80-mutant mice and cells with deficient levels of p53. *Mol. Cell. Biol.* **20,** 3772–80 (2000).

20. Langerak, P. & Russell, P. Regulatory networks integrating cell cycle control with DNA damage checkpoints and double-strand break repair. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **366,** 3562–71 (2011).

21. Balogun, F. O., Truman, A. W. & Kron, S. J. DNA resection proteins Sgs1 and Exo1 are required for G1 checkpoint activation in budding yeast. *DNA Repair (Amst).* **12,** 751–760 (2013).

22. Ito, T. *et al.* A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. U. S. A.* **98,** 4569–74 (2001).

23. Pitre, S. *et al.* Short Co-occurring Polypeptide Regions Can Predict Global Protein Interaction Maps. *Sci. Rep.* **2,** 239 (2012).

24. Bork, P. *et al.* Protein interaction networks from yeast to human. *Curr. Opin. Struct. Biol.* **14,** 292–299 (2004).

25. Cusick, M. E., Klitgord, N., Vidal, M. & Hill, D. E. Interactome: gateway into systems biology. *Hum. Mol. Genet.* **14,** R171–R181 (2005).

26. Uetz, P. *et al.* A comprehensive analysis of protein–protein interactions inSaccharomyces cerevisiae. *Nature* **403,** 623–627 (2000).

27. Galletta, B. J. & Rusan, N. M. A yeast two-hybrid approach for probing protein-protein interactions at the centrosome. *Methods Cell Biol.* **129,** 251–77 (2015).

28. Baryshnikova, A. *et al.* Quantitative analysis of fitness and genetic interactions in yeast on a genome scale. *Nat. Methods* **7,** 1017–24 (2010).

29. Szklarczyk, D. *et al.* STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* **43,** D447–D452 (2015).

30. Copley, R. R., Doerks, T., Letunic, I. & Bork, P. Protein domain analysis in the era

of complete genomes. *FEBS Lett.* **513,** 129–134 (2002).

31. Amos-Binks, A. *et al.* Binding site prediction for protein-protein interactions and novel motif discovery using re-occurring polypeptide sequences. *BMC Bioinformatics* **12,** 225 (2011).

32. Finn, R. D. *et al.* The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* **44,** D279–D285 (2016).

33. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28,** 235–242 (2000).

34. Raghavachari, B., Tasneem, A., Przytycka, T. M. & Jothi, R. DOMINE: a database of protein domain interactions. *Nucleic Acids Res.* **36,** D656-61 (2008).

35. Ceol, A. *et al.* DOMINO: a database of domain-peptide interactions. *Nucleic Acids Res.* **35,** D557-60 (2007).

36. Mosca, R., Céol, A., Stein, A., Olivella, R. & Aloy, P. 3did: a catalog of domain-based interactions of known three-dimensional structure. *Nucleic Acids Res.* **42,** D374–D379 (2014).

37. Schultz, J., Copley, R. R., Doerks, T., Ponting, C. P. & Bork, P. SMART: a web-based tool for the study of genetically mobile domains. *Nucleic Acids Res.* **28,** 231–4 (2000).

38. Bateman, A. *et al.* UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **45,** D158–D169 (2017).

39. Chatr-aryamontri, A. *et al.* The BioGRID interaction database: 2017 update. *Nucleic Acids Res.* **45,** D369–D379 (2017).

40. Schoenrock, A. *et al.* Efficient prediction of human protein-protein interactions at

a global scale. *BMC Bioinformatics* **15,** 383 (2014).

41. Chipman, K. C. & Singh, A. K. Predicting genetic interactions with random walks on biological networks. *BMC Bioinformatics* **10,** 17 (2009).

42. Wong, S. L. *et al.* Combining biological networks to predict genetic interactions. *Proc. Natl. Acad. Sci.* **101,** 15682–15687 (2004).

43. Calzone, L., Barillot, E. & Zinovyev, A. Predicting genetic interactions from Boolean models of biological networks. *Integr. Biol.* **7,** 921–929 (2015).

44. Alberghina, L. & Cirulli, C. Proteomics and systems biology to tackle biological complexity: Yeast as a case study. *Proteomics* **10,** 4337–4341 (2010).

45. Zhao, Q. *et al.* GPS-SUMO: a tool for the prediction of sumoylation sites and SUMO-interaction motifs. *Nucleic Acids Res.* **42,** W325–W330 (2014).

46. Kumar, G. & Ranganathan, S. Network analysis of human protein location. *BMC Bioinforma. 2010 117* **11,** S9 (2010).

47. Koh, J. L. Y. *et al.* CYCLoPs: A Comprehensive Database Constructed from Automated Analysis of Protein Abundance and Subcellular Localization Patterns in Saccharomyces cerevisiae. *G3 (Bethesda).* **5,** 1223–32 (2015).

48. Goldberg, T. *et al.* LocTree3 prediction of localization. *Nucleic Acids Res.* **42,** W350-5 (2014).

49. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25,** 25–29 (2000).

50. The Gene Ontology Consortium. Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Res.* **45,** D331–D338 (2017).

51. Hafez, D. *et al.* McEnhancer: predicting gene expression via semi-supervised

assignment of enhancers to target genes. *Genome Biol.* **18,** 199 (2017).

52. Consortium, T. E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).

53. Natarajan, A., Yardimci, G. G., Sheffield, N. C., Crawford, G. E. & Ohler, U. Predicting cell-type-specific gene expression from regions of open chromatin. *Genome Res.* **22,** 1711–22 (2012).

54. Singh, R., Lanchantin, J., Robins, G. & Qi, Y. DeepChrome: deep-learning for predicting gene expression from histone modifications. *Bioinformatics* **32,** i639–i648 (2016).

55. van Dam, S. *et al.* GeneFriends: An online co-expression analysis tool to identify novel gene targets for aging and complex diseases. *BMC Genomics* **13,** 535 (2012).

56. Ge, H., Walhout, A. J. M. & Vidal, M. Integrating 'omic' information: a bridge between genomics and systems biology. *Trends Genet.* **19,** 551–60 (2003).

57. Boucher, B. & Jenna, S. Genetic interaction networks: better understand to better predict. *Front. Genet.* **4,** 290 (2013).

58. Mostafavi, S., Ray, D., Warde-Farley, D., Grouios, C. & Morris, Q. GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. *Genome Biol.* **9,** S4 (2008).

59. Moreau, Y. & Tranchevent, L.-C. Computational tools for prioritizing candidate genes: boosting disease gene discovery. *Nat. Rev. Genet.* **13,** 523–536 (2012).

60. Arnoult, N. *et al.* Regulation of DNA repair pathway choice in S and G2 phases by the NHEJ inhibitor CYREN. *Nature* **549,** 548–552 (2017).

61. Winzeler, E. A. *et al.* Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. *Science* **285,** 901–6 (1999).

62. Xiao, W. *Yeast Protocols*. **313,** (2005).

63. Chee, M. K. & Haase, S. B. New and Redesigned pRS Plasmid Shuttle Vectors for Genetic Manipulation of Saccharomycescerevisiae. *G3 (Bethesda).* **2,** 515–26 (2012).

64. Gietz, R. D. & Schiestl, R. H. High-efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method. *Nat. Protoc.* **2,** 31–34 (2007).

65. Boulton, S. & Jackson, S. P. Identification of a Saccharomyces cerevisiae Ku80 homologue: roles in DNA double strand break rejoining and in telomeric maintenance. *Nucleic Acids Res.* **24,** 4639–4648 (1996).

66. Cherry, J. M. *et al.* Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res.* **40,** D700–D705 (2012).

67. Erdemir, T., Bilican, B., Cagatay, T., Goding, C. R. & Yavuzer, U. Saccharomyces cerevisiae C1D is implicated in both non-homologous DNA end joining and homologous recombination. *Mol. Microbiol.* **46,** 947–957 (2002).

68. George, K. & George, K. How to validate an automated colony counter. *Protoc. Exch.* (2013). doi:10.1038/protex.2013.058

69. Wagih, O. *et al.* SGAtools: one-stop analysis and visualization of array-based genetic interaction screens. *Nucleic Acids Res.* **41,** W591-6 (2013).

70. Baryshnikova, A. *et al.* Synthetic Genetic Array (SGA) Analysis in Saccharomyces cerevisiae and Schizosaccharomyces pombe. *Methods Enzymol.* **470,** 145–179 (2010).

71. McKinney, J. S. *et al.* A multistep genomic screen identifies new genes required for repair of DNA double-strand breaks in Saccharomyces cerevisiae. *BMC Genomics* **14,** 251 (2013).

72. Cassani, C. *et al.* Tel1 and Rif2 Regulate MRX Functions in End-Tethering and Repair of DNA Double-Strand Breaks. *PLOS Biol.* **14,** e1002387 (2016).

73. Moore, J. K. & Haber, J. E. Cell cycle and genetic requirements of two pathways of nonhomologous end-joining repair of double-strand breaks in Saccharomyces cerevisiae. *Mol. Cell. Biol.* **16,** 2164–73 (1996).

74. Koç, A., Wheeler, L. J., Mathews, C. K. & Merrill, G. F. Hydroxyurea arrests DNA replication by a mechanism that preserves basal dNTP pools. *J. Biol. Chem.* **279,** 223–30 (2004).

75. Whalen, C. *et al.* --RNA Polymerase II Transcription Attenuation at the Yeast DNA Repair Gene, DEF1, Involves Sen1-Dependent and Polyadenylation Site-Dependent Termination. *G3 (Bethesda).* g3.200072.2018 (2018). doi:10.1534/g3.118.200072

76. Omidi, K. *et al.* Uncharacterized ORF HUR1 influences the efficiency of non-homologous end-joining repair in Saccharomyces cerevisiae. *Gene* **639,** 128–136 (2018).

77. Ho, Y. *et al.* Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature* **415,** 180–183 (2002).

78. Pannala, V. R., Bhat, P. J., Bhartiya, S. & Venkatesh, K. V. Systems biology of *GAL* regulon in *Saccharomyces cerevisiae*. *Wiley Interdiscip. Rev. Syst. Biol. Med.* **2,** 98–106 (2010).

79. Bahmed, K., Nitiss, K. C. & Nitiss, J. L. Yeast Tdp1 regulates the fidelity of nonhomologous end joining. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 4057–62 (2010).

80. Lundin, C. *et al.* Methyl methanesulfonate (MMS) produces heat-labile DNA damage but no detectable in vivo DNA double-strand breaks. *Nucleic Acids Res.* **33,** 3799–811 (2005).

81. Ma, W., Westmoreland, J. W., Gordenin, D. A. & Resnick, M. A. Alkylation Base Damage Is Converted into Repairable Double-Strand Breaks and Complex Intermediates in G2 Cells Lacking AP Endonuclease. *PLoS Genet.* **7,** e1002059 (2011).

82. Kitanovic, A. *et al.* Metabolic response to MMS-mediated DNA damage in *Saccharomyces cerevisiae* is dependent on the glucose concentration in the medium. *FEMS Yeast Res.* **9,** 535–551 (2009).

83. Greene, C. S. *et al.* Understanding multicellular function and disease with human tissue-specific networks. *Nat. Genet.* **47,** 569–576 (2015).