Computational analysis of quantitative “omics” data

Vivien A.C. Schoonenberg

2023

Computational analysis of quantitative "omics" data

Dissertation
zur Erlangung des Grades
Doktor der Naturwissenschaften

am Fachbereich Biologie
der Johannes Gutenberg-Universität
Mainz

Vivien Antoinette Catharina Schoonenberg
geb. am 25. Oktober 1993 in Utrecht,
die Niederlande

Mainz, 2023

Dekan: Prof. Dr. Eckhard Thines

Erstgutachter:
Zweitgutachter:

Tag der mündlichen Prüfung:

"Science, like all human endeavors, is evolutionary. We build by adding to and recombining what is already there."

- Frances H. Arnold

Preface

Before you lay the collected work of my PhD, and the result of the past years. The thesis is composed of a general introduction, three chapters, and a conclusion. Each of the chapters is an original research article, printed here as it was originally published or as it is currently being prepared for peer review. For two out of the three included articles, I am the (shared) main author, and at the beginning of each chapter, there will be a statement of contribution.

Over the course of my PhD, I have been fortunate to cover a wide range of biological questions, investigating DNA damage in different model organisms. I have applied quantitative mass spectrometry as either the main method or as an added tool in an effort to answer them. My work has focused on the analysis of proteomics data and its integration with other types of (large-scale) biological data, such as transcriptomics.
In Chapter 1, General introduction, I have laid out general concepts and broader perspectives for the individual research projects or chapters within this thesis. In Chapter 2, we mapped the system's response to DNA damage over time in *Tetrahymena thermophila*, while we studied the interactome of specific DNA damage lesions across the Tree of Life in Chapter 3. Finally, in Chapter 4, we developed a computational pipeline to make positive selection analysis user-friendly.

Summary

Over the past decades, the rise of "omics" approaches has allowed for systematic, in-depth investigation of each aspect of molecular biology. It has contributed to our changed view on the on the linearity and the regulation of the informational flow of the central dogma. Different regulatory mechanisms have been identified, describing interaction and variety not only on the genetic level but also on the transcript and protein level. The development and integration of multi-omics have allowed for the uncovering of intricate molecular mechanisms underlying different phenotypic manifestations of traits at a high accuracy, in a systematic manner. With this, multi-omics is essentially the basis of network or systems biology.

In this thesis, I have utilized "omics" technologies, specifically proteomics, and the subsequent computational data analysis and integration to investigate the systematic DNA damage response in *Tetrahymena thermophila* and identify DNA damage proteins across the Tree of Life. Additionally, I co-developed a user-friendly computational pipeline for evolutionary positive selection analysis, which relies on comparative genomics and either large-scale genome sequencing or proteotranscriptomics data.

In Chapter 2, we mapped the system's response to DNA damage over time in *Tetrahymena thermophila* (Nischwitz, Schoonenberg et al., *in preparation*). To date, limited studies have combined the strength of proteomics and transcriptomics to investigate DNA damage kinetics in response to various DNA-damage treatments. Our study investigated DNA damage response (DDR) dynamics over eight hours after or during exposure to six different mutagens. We observed upregulation of previously identified DNA damage repair pathways and found novel crosstalk between DDR pathways. All treatments induced a dynamic response at both the transcript and protein levels. Using unsupervised self-organizing maps, we examined the clustering of expression profile trends to better understand the DDR. Many of the quantified proteins and transcripts exhibited damage-specific responses. We are currently employing a novel knockdown system to target a subset of PARP-related proteins to characterize their specific roles in *Tetrahymena* further.

In Chapter 3, we studied the interactome of specific DNA damage lesions across the Tree of Life, exploring the conservation of pathways responsible for repair and recognition of DNA damage lesions (Nischwitz, Schoonenberg et al., *iScience*, 2023). Due to the need for precise genome maintenance, DNA repair has been highly conserved across all domains of life. To study the shared and unique elements of the DNA damage response, we performed a phylointeractomic study to identify enriched DNA damage binders in 11 different species at the 8-oxoG and abasic lesions and at a uracil base incorporated into DNA. Our approach identified several known DNA damage factors as binders to the afore-mentioned lesions. Additionally, through orthology, network, and domain analysis, we linked 44 previously unassociated proteins to DNA repair.

Finally, in Chapter 4, we developed a computational pipeline to make positive selection analysis user-friendly (Ceron-Noriega et al., *Genome Biology and Evolution,* 2023). AlexandrusPS generates orthology relationships, sequence alignments, and phylogenetic trees with its automated process. It then performs site-specific (SSM), branch (BM), and branch-site (BSM) positive selection analyses and produces four main output files, including orthology relationships, positive selection results, and all intermediate files (sequence alignments, phylogenetic trees).

Zusammenfassung

In den letzten Jahrzehnten hat das Aufkommen der ”Omics”-Ansätze eine systematische, eingehende Untersuchung jedes Aspekts der Molekularbiologie ermöglicht. Dies hat dazu beigetragen, dass sich unsere Sichtweise auf die Linearität und die Regulierung des Informationsflusses des zentralen Dogmas geändert hat. Es wurden verschiedene Regulierungsmechanismen identifiziert, die die Interaktion und Vielfalt nicht nur auf genetischer Ebene, sondern auch auf der Ebene der Transkripte und Proteine beschreiben. Die Entwicklung und Integration von Multi-omics hat es ermöglicht, die komplexen molekularen Mechanismen, die den verschiedenen phänotypischen Ausprägungen von Merkmalen zugrunde liegen, mit hoher Genauigkeit und auf systematische Weise aufzudecken. Damit ist die Multi-omik im Wesentlichen die Grundlage der Netzwerk- oder Systembiologie.

In dieser Arbeit habe ich “omics”-Technologien, insbesondere Proteomik, und die anschließende computergestützte Datenanalyse und -integration eingesetzt, um die systematische DNA-Schadensreaktion in *Tetrahymena thermophila* zu untersuchen und DNA-Schadensproteine im phylogenetischen Stammbaum zu identifizieren. Zusätzlich habe ich eine benutzerfreundliche computergestützte Pipeline für die Analyse der evolutionären positiven Selektion mitentwickelt, die auf vergleichender Genomik und entweder groß angelegten Genomsequenzierungs- oder Proteotranskriptomikdaten beruht.

In Kapitel 2 haben wir die Reaktion von *Tetrahymena thermophila* auf DNA-Schäden im Zeitverlauf kartiert (Nischwitz, Schoonenberg et al., *in Bearbeitung*). Bisher gibt es nur wenige Studien, die die Stärken von Proteomik und Transkriptomik kombinieren, um die Kinetik von DNA-Schäden als Reaktion auf verschiedene DNA-schädigende Behandlungen zu untersuchen. In unserer Studie untersuchten wir die Dynamik der DNA-Schadensreaktion (DDR) über einen Zeitraum von acht Stunden bei der Exposition gegenüber sechs verschiedenen Mutagenen. Wir beobachteten eine Hochregulierung bereits identifizierter DNA-Schadensreparaturwege und fanden neuartige Wechselwirkungen zwischen den DDR-Wegen. Alle Behandlungen führten zu einer dynamischen Reaktion sowohl auf der Transkript- als auch auf der Proteinebene. Mithilfe von unüberwachten selbstorganisierenden Karten

untersuchten wir das Clustering von Expressionsprofiltrends, um die DDR besser zu verstehen. Viele der quantifizierten Proteine und Transkripte zeigten schädigungsspezifische Reaktionen. Wir setzen derzeit ein Knockdown-System ein, um eine Untergruppe von PARP-verwandten Proteinen herunterzuregulieren und ihre spezifische Rolle in *Tetrahymena* weiter zu charakterisieren.

In Kapitel 3 untersuchten wir das Interaktom von spezifischen DNA-Schadensläsionen im phylogenetischen Stammbaum und untersuchten die Erhaltung der Wege, die für die Reparatur und Erkennung von DNA-Schadensläsionen verantwortlich sind (Nischwitz, Schoonenberg et al., *iScience,* 2023). Weil das Genom akkurat erhalten werden muss, ist die DNA-Reparatur in allen Lebensbereichen stark konserviert. Um die gemeinsamen und einzigartigen Elemente der DNA-Schadensreaktion zu untersuchen, haben wir eine phylointeraktomische Studie durchgeführt, um in 11 verschiedenen Arten angereicherte DNA-Schadensbinder an den 8-oxoG- und abasischen Läsionen sowie an einer in die DNA eingebauten Uracil-Base zu identifizieren. Unser Ansatz identifizierte mehrere bekannte DNA-Schadensfaktoren als Binder für die oben genannten Läsionen. Darüber hinaus konnten wir durch Orthologie-, Netzwerk- und Domänenanalysen 44 bisher nicht assoziierte Proteine mit der DNA-Reparatur in Verbindung bringen.

Abschließend entwickelten wir in Kapitel 4 eine computergestützte Pipeline, um die Analyse der positiven Selektion benutzerfreundlich zu machen (Ceron-Noriega et al., *Genome Biology and Evolution*, 2023). AlexandrusPS erzeugt in einem automatisierten Prozess Orthologiebeziehungen, Sequenzalignments und phylogenetische Bäume. Anschließend führt es site-spezifische (SSM), branch-spezifische (BM) und branch-site-spezifische (BSM) Positivselektionsanalysen durch und produziert vier Hauptausgabedateien, einschließlich Orthologiebeziehungen, Positivselektionsergebnisse und alle Zwischendateien (Sequenzalignments, phylogenetische Bäume).

Table of contents

Preface vii

Summary ix

Zusammenfassung xi

1. General introduction 1
- 1.1. Central dogma of molecular biology and multi-omics . . . 2
- 1.2. Genomics and transcriptomics 3
- 1.3. Quantitative mass spectrometry-based proteomics 9
- 1.4. Proteomics and transcriptomics to study DNA damage . . . 16
- 1.5. Omics in evolutionary analysis and applications 17

2. A systems view on the DNA damage response kinetics in *Tetrahymena thermophila* 21
- 2.1. Summary 22
- 2.2. Zusammenfassung 22
- 2.3. Statement of Contribution 23
- 2.4. Abstract 24
- 2.5. Introduction 24
- 2.6. Results 25
- 2.7. Methods 34
- 2.8. Acknowledgements 37

3. DNA damage repair proteins across the Tree of Life 39
- 3.1. Summary 40
- 3.2. Zusammenfassung 40
- 3.3. Statement of Contribution 41
- 3.4. Abstract 42
- 3.5. Introduction 42
- 3.6. Results and discussion 44
- 3.7. Acknowledgements 60
- 3.8. STAR Methods 61

4. AlexandrusPS: a user-friendly pipeline for the automated detection of orthologous gene clusters and subsequent positive selection analysis 69
4.1. Summary 70
4.2. Zusammenfassung 70
4.3. Statement of Contribution 71
4.4. Abstract 72
4.5. Introduction 72
4.6. Implementation 74
4.7. Conclusion 78
4.8. Supplementary Material 79
4.9. Acknowledgements 79
4.10. Data availability 79

5. Conclusion 81

Bibliography 89

Acknowledgements 107

Appendices 109

A. Supplemental information Chapter 2 109
A.1. Supplemental tables 109
A.2. Supplemental figures 116

B. Supplemental information Chapter 3 119
B.1. Supplemental tables 119
B.2. Supplemental figures 121

C. Supplemental information Chapter 4 129

Curriculum Vitae 133

1

General introduction

1.1. Central dogma of molecular biology and multi-omics

The central dogma of molecular biology describes the unidirectional flow and transfer of information in a cell, from deoxyribonucleic acid (DNA) to ribonucleic acid (RNA) to protein. Within a cell, a gene or other segment of the DNA is transcribed into messenger RNA, which in turn is translated into protein, exercising biological function within or outside of the cell [1–3]. This simple concept between nucleic acids (DNA, RNA) and proteins was initially introduced in the early 1950s and has been the focus of various biological research in all three domains of life: eukaryotes, bacteria, and archaea. However, over the past decades, new concepts, technologies, and formulations of big data science have developed rapidly and changed our view on the linearity and the regulation of the informational flow of the central dogma [1, 4]. For example, different regulatory mechanisms have been identified, describing interaction and variety not only on the genetic level but also on the transcript and protein level (i.e., mRNA splicing, histone modifications, post-translational modifications) [3]. The rise of "omics" approaches has allowed in-depth investigation of each aspect of molecular biology instead of a more reductionist approach focusing only on one gene or a few genes and proteins [1, 5].

"Omics" can be defined as the probing and analyzing of large amounts of data representing the structure and function of an entire makeup of a given biological system (i.e., a cell or organism) at a particular level, such as gene (genomics), transcript (transcriptomics), or protein (proteomics). Following the central dogma, "omics" technologies have been used to capture static genomic alterations, temporal transcriptomic perturbations, alternative splicing, spatiotemporal proteomic dynamics, and post-translational modifications [5]. "Omics" is a rapidly developing and growing field, allowing the uncovering of the intricate molecular mechanisms underlying different phenotypic manifestations of traits in a systematic manner at a high accuracy. Thereby reinforcing that within a cell or biological system the flow of information is interconnected between levels, of which genes, transcripts, and proteins constitute the most prominent three [1, 4, 5]. Moreover, the complexity of cellular behavior and its decision-making system has driven the establishment and expansion of novel omics and associated techniques. These include epiomics to analyze modifications of the initially described three "omics" (such as epigenome, epitranscriptome, epiproteome), molecular interactomics (i.e., varied levels of interactome), and disease-associated hallmarks such as metabolome and immunome. Multi-omics integration has become a prevailing trend for constructing a comprehensive relationship between molecular signatures and phenotypic manifestations of a particular disease, tied to the aim of uncovering causality within this relationship. With that, multi-omics is essentially the basis of network or systems biology [5].

Importantly, with the large-scale and increasingly complex data generation through “omics” approaches, we now face the challenge of developing an understanding of how to analyze these different data types and how they quantitatively relate to one another and the phenotypic characteristics of the organism [6, 7].

In this thesis, I have utilized “omics” technologies, specifically proteomics, and the subsequent computational data analysis and integration to investigate the systematic DNA damage response in *Tetrahymena thermophila* and identify DNA damage proteins across the tree of life. Additionally, I co-developed a user-friendly computational pipeline for evolutionary positive selection analysis, which relies on comparative genomics and either large-scale genome sequencing or proteotranscriptomics data.

1.2. Genomics and transcriptomics

The application of omics in entire genomes, aiming to determine the (complete) genomic sequence, base order, and characterize and quantify all genes of an organism is referred to as genomics. It aids in uncovering the (inter)relationship of genomic sequences and genes and their respective influence on the organism and specific phenotypes [8].

Transcriptomics describes the study of the expression of all RNAs from a given cell population, offering a global perspective on molecular dynamic changes induced by environmental factors or pathogenic agents. The transcriptome includes many types of RNA, including protein-coding RNAs (mRNAs), long noncoding RNAs (lncRNAs), short noncoding RNAs (microRNAs, small-interfering RNAs, short noncoding RNAs, enhancer RNAs), and circular RNAs. All of these types of RNA have been indicated to affect phenotype and have been associated with different diseases (such as diabetes, cancer, and cardiovascular disease) [8].

Both the transcriptome and genome have initially been investigated through micro-array technology. This technology is based on the comparative hybridization of fluorescently labeled DNA or cDNA (in the case of transcriptome profiling) under stringent conditions to capture probes (complementary oligonucleotides). Micro-arrays allow the simultaneous analysis of tens of thousands of molecules, which revolutionized the scale and depth in which DNA and RNA could be investigated around three decades ago [9]. Transcriptome and genome analysis have continued to develop quickly, especially since next-generation sequencing (NGS) emerged. NGS allows for sequencing hundreds of millions of DNA molecules simultaneously, potentially highlighting millions of genomic or transcriptomic variants in a single experiment. Its arrival has dramatically improved the turnaround time of genome-scale experiments and their results, thereby speeding up genetic and genomic discovery and advancing the understanding of molecular mechanisms of disease and cell biology [9, 10].

The most popular next-generation sequencing technologies currently available are based on a sequencing-by-synthesis approach. Within this, we can distinguish short-read and long-read sequencing [11]. For each of these technologies, different platforms are available, with Illumina leading the current market for short-read sequencing and Oxford Nanopore Technologies (ONT) and Pacific Biosciences (PacBio) for long-read sequencing [12].

The complete workflow, sample preparation, and analysis of each available technology slightly differ, depending on the platform used.

The general steps of sample preparation for current short-read sequencing are (i) RNA or DNA fragmentation (for RNA sequencing technologies, it is common to convert RNA to cDNA before sequencing, thereby losing any RNA base modifications). Next, (ii) adapters are ligated to the template to facilitate the attachment of fragments to solid surfaces (e.g., microchips, microbeads, or nanowells) or for fragments to be circularized. Finally, (iii) templates are amplified to provide enough copies of each template to allow the sequencer to detect them. The libraries can either be sequenced only from one end (known as single-end reads) or from both ends (known as paired-end reads) [11]. In an Illumina sequencer, DNA templates with ligated adapters hybridize to a solid surface with patterned clusters of complementary adapters. Then, cluster generation begins, where thousands of copies of each fragment are generated through a process known as bridge amplification. In this process, one strand folds over, and the adapter on the end of the molecule hybridizes with another oligonucleotide in the flow cell. A polymerase incorporates nucleotides to build double-stranded DNA molecule bridges, which are denatured to leave single-stranded DNA fragments tethered to the flow cell. This process is repeated continually, producing millions of clusters of clonal template DNA fragments that can be sequenced simultaneously. Sequencing synthesis begins using reversible terminator nucleotides, which permits one nucleotide to be incorporated at a time and the representative fluorescence to be recorded as a base call by high-resolution optical imaging. Cleavage of the terminal chemical modification allows the next complementary fluorescently labeled nucleotide to be incorporated. This process is repeated for the length of the read to generate the sequence output. The read lengths of the fragments can range between 25 to 450 base pairs (bp) (Figure 1.1) [11, 13].

Illumina's short-read sequencing produces highly accurate sequencing reads, which are inexpensive and easy to generate on a massive scale. However, short reads are too short to detect more than 70% of human genome structural variation (affecting sequences longer than 50 base pairs). Additionally, more than 15% of the human genome is inaccessible via short-read sequencing due to its repeat content or GC content, which causes problems when assembling or mapping the short reads [12]. One of the solutions developed to overcome these issues is long-read sequencing. Long-read technologies can generate continuous sequences ranging from 10 kilobases (kb) to several megabases (Mb) in length directly from native DNA.

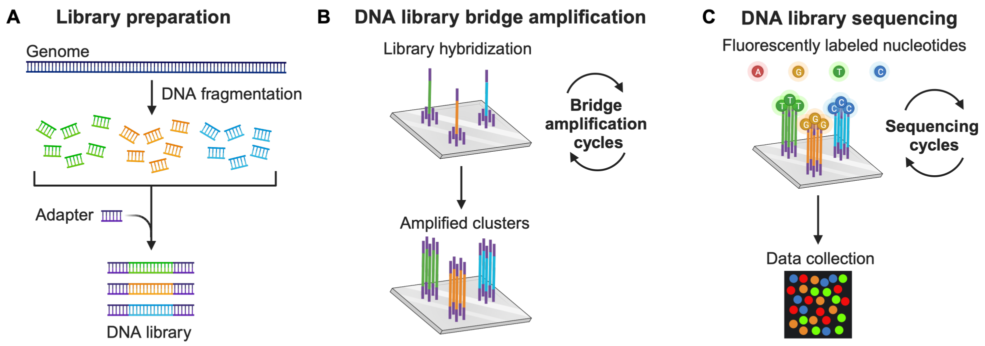


Figure 1.1.: Overview of NGS short-read sequencing on an Illumina platform. 3 steps in the workflow for sequencing: **A)** Library preparation: NGS library is prepared by fragmenting the DNA sample and ligating specialized adapters to both fragment ends. **B)** Bridge and cluster amplification: The library is loaded into a flow cell and the fragments are hybridized to the flowcell surface. Each bound fragment is amplified into a clonal cluster through bridge amplification. **C)** Sequencing: Sequencing reagents, including the fluorescently labeled nucleotides, are added and the first base is incorporated. The flow cell is imaged and the emission from each cluster is recorded. The emission wavelength and intensity are used to identify the base. This is repeated "n" times to create a read length of "n" bases. Created with BioRender.com

PacBio's single-molecule real-time (SMRT) long-read sequencing uses a circular DNA molecule template called a SMRTbell, composed of a double-stranded DNA insert with single-stranded hairpin adapters on either end. The insert can range in length from one to more than a hundred kilobases, which allows long sequencing reads to be generated. During the sequencing reaction, a DNA polymerase processes around the SMRTbell template and incorporates fluorescently labeled deoxynucleoside triphosphates into the nascent strand. After each incorporation, a laser excites the fluorophore, and a camera records the emission, as with the Illumina short-read platform (Figure 1.2A) [11, 12].

ONT long-read sequencing technology uses linear DNA molecules. They are typically one to several hundred kilobases long but can have a length of several megabases. Steps for ONT sequencing involve attaching a double-stranded DNA molecule to a sequencing adapter, which is preloaded with a motor protein. Next, the DNA mixture is loaded onto a flow cell containing nanopores embedded in a synthetic membrane. The motor protein unwinds the double-stranded DNA and, in combination with an electrical current, drives the negatively charged DNA through the nanopore. As the DNA goes through the pore, it causes characteristic disruptions to the current (based on its base composition), translated in real-time into base calls (Figure 1.2B). The throughput for nanopore sequencing is high but has much higher error rates (>15%) than short-read sequencing. SMRT long-read sequencing has a higher accuracy but is limited by high costs and lower throughput. However, these techniques will continue to be developed and improved together with others and keep expanding the boundaries of NGS, genomics, and transcriptomics research [11, 12].

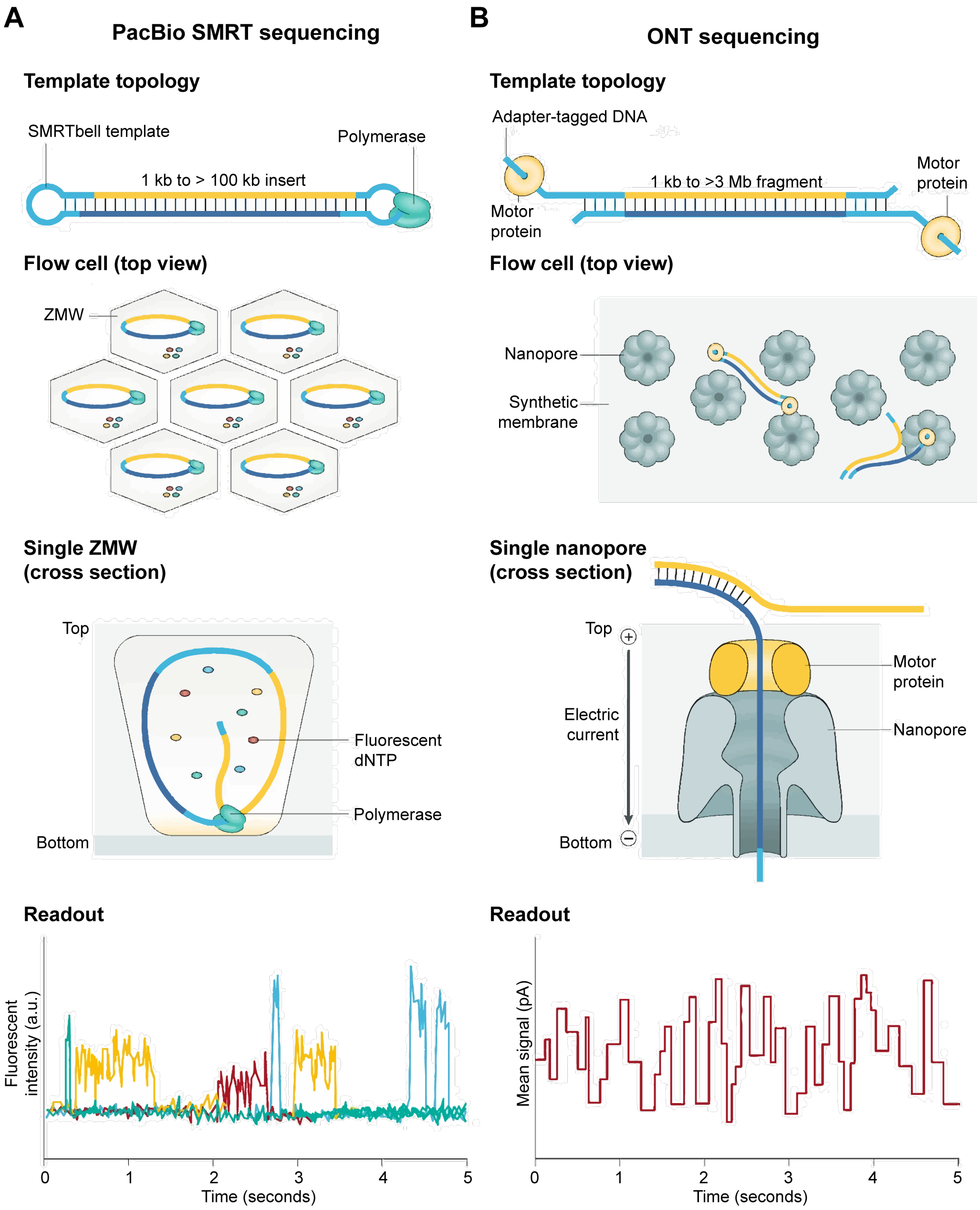


Figure 1.2.: Overview of PacBio and ONT long-read sequencing. **A)** In Pacific Biosciences (PacBio) single-molecule, real-time (SMRT) sequencing, DNA (yellow for forward strand, dark blue for reverse strand) is fragmented and ligated to hairpin adapters (light blue) to form a topologically circular molecule known as a SMRTbell. Once the SMRTbell has been generated, it is bound by a DNA polymerase and loaded onto a SMRT Cell for sequencing. Each SMRT Cell can contain up to 8 million zero-mode waveguides (ZMWs), which are chambers that hold picolitre volumes. Light penetrates the lower 20–30 nm of each well, reducing the detection volume of the well to only 20 zl (10−21 l). As the DNA mixture floods the ZMWs, the SMRTbell template and polymerase become immobilized on the bottom of the chamber. Fluorescently labelled deoxynucleoside triphosphates (dNTPs) are added to begin the sequencing reaction. As the polymerase begins to synthesize the new strand of DNA, a fluorescent dNTP is briefly held in the detection volume, and a light pulse from the bottom of the well excites the fluorophore. Unincorporated dNTPs are not typically excited by this light but, in rare cases, can become excited if they diffuse into the excitation volume, thereby contributing to noise and error in PacBio sequencing. The light emitted from the excited fluorophore is detected by a camera, which records the wavelength and relative position of the incorporated base in the nascent strand.

Figure 1.2.: Overview of PacBio and ONT long-read sequencing (continued). The phosphate-linked fluorophore is then cleaved from the nucleotide as part of the natural incorporation of the base into the new strand of DNA and released into the buffer, preventing fluorescent interference during the subsequent light pulse. The DNA sequence is determined by the changing fluorescent emission that is recorded within each ZMW, with a different colour corresponding to each DNA base (for example, green, T; yellow, C; red, G; blue, A). **B)** In Oxford Nanopore Technologies (ONT) sequencing, arbitrarily long DNA (yellow for forward strand, dark blue for reverse strand) is tagged with sequencing adapters (light blue) preloaded with a motor protein on one or both ends. The DNA is combined with tethering proteins and loaded onto the flow cell for sequencing. The flow cell contains thousands of protein nanopores embedded in a synthetic membrane, and the tethering proteins bring the DNA molecules towards these nanopores. Then, the sequencing adapter inserts into the opening of the nanopore, and the motor protein begins to unwind the double-stranded DNA. An electric current is applied, which, in concert with the motor protein, drives the negatively charged DNA through the pore at a rate of about 450 bases per second. As the DNA moves through the pore, it causes characteristic disruptions to the current, generating a readout known as a ‘squiggle’. Changes in current within the pore correspond to a particular k-mer (that is, a string of DNA bases of length k), which is used to identify the DNA sequence. Figure and caption from [@Logsdon2020].

In summary, short-read NGS has an incredibly high throughput and accuracy, making it the method of choice for standard gene expression analysis. Long-read sequencing is becoming an increasingly popular option and solution to investigate large structural variants, repeat sequences, and splice variants. Together, the rise of NGS has pushed the “omics” revolution forward by facilitating whole genome (re)sequencing projects, whole-exome sequencing, genome analyses such as large-scale detection of single nucleotide polymorphisms (SNPs) and variant calling (VC), as well as the detection of (large) DNA mutations, (i.e., insertions and deletions), and DNA methylation [9, 10]. NGS can be further used to map protein-DNA (using chromatin immunoprecipitation sequencing (ChIP-seq) or CUT&RUN) or DNA-DNA interactions (chromosome conformation capture, Hi-C) at nucleotide resolution [14, 15]. Additionally, as NGS does not rely on capture probe design, novel noncoding RNAs, splice variants, post-transcriptional modifications, and nascent RNA synthesis can be quantitatively analyzed [9].

1.2.1. Analysis of transcriptome data

During one single sequencing run, NGS analyzes millions of DNA fragments. The read lengths of the short-read fragments can range between 25 to 450 base pairs, depending on the NGS platform. For long-read sequencing, read lengths can range from 250 base pairs up to 2.3 megabases. As the throughput of NGS is incredibly high, creating data sets of up to 50 gigabases in a single run, its development raised the need for scalable and improved computational methods and algorithms to analyze this data [13]. Additionally, long-read sequencing platforms produce qualitatively different data from second-generation sequencing, thus necessitating tailored analysis tools [16].

Analysis for both long-read and short-read sequencing relies on availability of an accurate reference genome. Reference genomes can be created by layering (genome) sequencing information to combine into scaffold information. Genome assembly and its annotation can still be challenging in itself and has been proposed to be improved by the use of RNA-seq contig evidence (overlapping reads or sequence data) as well as peptide information obtained by mass-spectrometry in a proteo-transcriptomics assembly workflow [17].

A standard analysis pipeline for short-read RNA-seq consists of several steps. First, raw image data is converted into short-read sequences, known as base-calling. Generally, the short-read sequences are configured in FASTQ format, a text-based representation of every nucleotide, and assigned an associated base quality score. Next, the reads are aligned to a reference genome or transcriptome [13].
Examples of popular read-aligning tools are Bowtie, STAR, and BWA. The performance of the different aligners is usually a tradeoff between accuracy and speed. Overall, the performance is impacted by the transcriptome size, coverage, and alignment lengths. Depending on the setup and question asked from the data, it is essential to consider choosing intron-aware or splice-aware aligners [18].
The aligners produce a human-readable sequence alignment map (SAM) file and a binary version (BAM) with a smaller file size. These files enable visualization and interrogation of the sequence (read assembly and base sequence) using programs such as the Integrative Genomics Viewer (IGV) [11]. After aligning or mapping RNA-seq reads to a reference genome, the number of mapped reads is counted, and gene expression level is calculated by peak calling algorithms. Examples of such algorithms are featureCounts (from Subread), HTSeq, or Cufflinks [19].

As for short-read sequencing, the first step in any long-read read analysis is base calling. ONT base-calling is more complex than SMRT base-calling. During SMRT sequencing, successions of fluorescence flashes are recorded as a movie, similar to how this is done for short-read sequencing. Because the template is circular, the polymerase may go over both strands of the DNA fragment multiple times, resulting in a continuous long read. This read is split into subreads, where each subread corresponds to one pass over the library insert without the linker sequences. Subreads are stored as an unaligned BAM file. From aligning these subreads together, an accurate consensus circular sequence (CCS) for the insert is derived [12, 16].

Nanopore raw data are current intensity values. Base-calling of nanopore reads is an active research area where algorithms are quickly evolving (incorporating different machine learning techniques and training models). Both SMRT and ONT technologies provide lower per-read accuracy than short-read sequencing. In the case of SMRT, the circular consensus sequence quality heavily depends on the number of times the fragment is read, which results from the original fragment's length and the polymerase's longevity.

The quality of nanopore reads is independent of the length of the DNA fragment. Read quality depends on achieving optimal translocation speed of the DNA fragment through the pore, which typically decreases in the late stages of sequencing runs, negatively affecting the quality [12, 16].

The following steps for long-read analysis are the same as in short-read analysis: aligning the reads to a reference genome and read quantification (gene expression) [20]. However, a major strength of long-read sequencing is the ability to determine the full-length RNA transcripts and isoforms. It simplifies the downstream analysis by eliminating the need to reconstruct isoforms based on the error-prone assembly of short RNA-sequencing reads. There are specific long-read isoform detection tools, which work by clustering aligned and error-corrected reads into groups and collapsing these into isoforms, but the detailed implementations differ between tools [11, 16, 20].

The most common follow-up analysis for gene expression quantification is differential expression (DE) [13, 21]. A basic mean DE analysis determines whether individual genes are up or downregulated between conditions (i.e., disease states, different tissues, etc.). Typically, DE analysis is done at the gene level by collapsing all mapped read counts to single gene units [21]. Importantly, DE analysis and its interpretation should account for biases associated with the expression level or abundance of reads for a particular gene. Additionally, while it gives valuable insights, the use of DE analysis alone could lead to missing some biological complexity and context, i.e., DE genes may not be causal to a phenotype, or functional genes are not nominated as only their function changes rather than their expression level [21]. All biological components in a cell or organism work in a biological system, never truly in isolation. A change to one molecule, like a transcript or gene, might cause a perturbation in its interaction with other components in the biological system, contributing to (more prominent) phenotypic differences or effects. To capture these coordinated patterns of gene expression, we can group genes by pathway or function in pathway analyses (e.g., Kyoto Encyclopedia of Genes and Genomes (KEGG) and Gene Ontology (GO)) [21–23]. Alternatively, or as a complementary approach, gene expression profiles can be presented as a biological network, highlighting known and predicted interactions and showing that genes can be part of multiple pathways. Together, these different analyses combine multiple information types, can improve in silico-predicted interactions within a condition, and prove insights into the biological systems or conditions investigated [21].

1.3. Quantitative mass spectrometry-based proteomics

As per the central dogma, proteins represent one of the main functional entities inside and outside cells. They regulate the activity of our immune system, build cells, funnel information, and are known as the building blocks of life. As described above, we now understand that it is not solely proteins that are responsible for the phenotype, but instead that proteins are organized in functional modules and networks, carrying out cellular functions and determining phenotypes through coordinated activities in combination with other molecules (such as RNA and DNA) in the cell [24].

Proteomics investigates the functional relevance of all expressed proteins in a cell, tissue, or organism [5]. The field is a collection of various technical disciplines, including cell imaging by light and electron microscopy, array and chip experiments, and genetic readout experiments (e.g., yeast two-hybrid assay) [25].

Over the past decades, mass spectrometry (MS)-based methods have emerged for the confident and near-exhaustive identification and quantification of the proteins in a biological sample. De novo analysis of proteins or protein populations from cells or tissues can be challenging due to the high complexity of cellular proteomes and the low abundance of many proteins, necessitating highly sensitive analytical techniques [24, 25]. In short, MS relies on the measurement of charged molecules, determining their mass-to-charge ratios, and quantifying peptides by their signal intensities. It is a protein characterization technique that can determine the amino acid sequence of a protein or peptide and post-translational modification (PTM) sites and quantify them [5].

MS-based proteomics is made possible by the availability of gene and genome sequence databases. It can reveal the quantitative state of the proteome and has significantly contributed to unraveling cellular signaling networks, elucidating the dynamics of protein-protein interaction in different cellular states, and improving molecular understanding of disease mechanisms [24, 25].

1.3.1. Principles of mass spectrometry

The generic overall process of MS-based proteomics consists of digesting the protein sample into peptides (using trypsin, lysin, or another enzyme mixture), fractionating the peptides (separation with liquid chromatography (LC)), and mass spectrometry analysis. Mass spectrometry analysis includes the ionizing of peptides, measurement of the mass-to-charge ratio of these peptide ions (also precursor ions), and sequential selection of precursor ions for fragmentation through collision. Fragment ion masses are then analyzed in a second analyzer to infer the peptide sequence [25, 26]. This workflow is also known as liquid chromatography coupled tandem MS (LC-MS/MS or LC-MS2) since two generations of ions are being analyzed (precursor and fragment) [5, 27, 28].

Ionization of peptides is commonly achieved through electrospray ionization (ESI) or matrix-assisted laser desorption/ionization (MALDI). ESI can directly be combined with separation techniques like high-performance LC (HPLC), as it ionizes peptides or analytes out of a solution. MALDI ionizes the samples out of a dry, crystalline matrix via laser pulses and produces singly charged ions of peptides, thereby minimizing spectral complexity [25, 27].

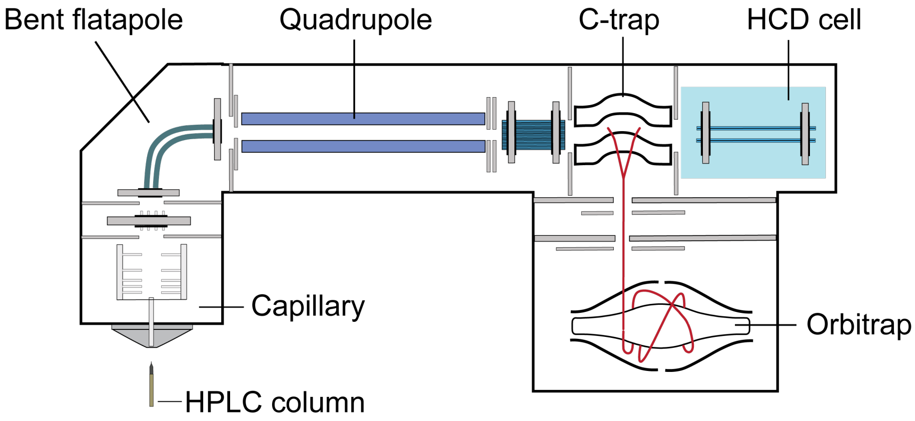


Figure 1.3.: Schematic presentation of the Q Exactive mass spectrometer (Thermo Fisher Scientific). View from above. The tip of the HPLC column is pointed at the front of the spectrometer. Liquid droplets containing peptides are sprayed from the column tip; peptides become mostly desolvated before entering the capillary, which is heated to a high temperature to help complete desolvation. Applied electromagnetic fields direct and focus the ion beam. During MS2 spectrum acquisitions, the quadrupole filters a small range of m/z values centered around the desired precursor m/z. The higher-energy collision-induced dissociation (HCD) cell is where high velocity precursors collide with gas particles, generating fragments. Peptides or fragments are collected in the C-trap for a set time ("injection time") before an applied voltage injects them into the orbitrap for mass analysis and detection (m/z and intensity measurements). Adapted from [28]

The sensitivity and resolution of a mass spectrometer depend on the mass analyzer's ability to separate ions effectively. Two of the most popular mass analyzers and detectors for proteomics are the time-of-flight (TOF) analyzer and the orbitrap [27, 28]. An example of a mass spectrometer using an orbitrap analyzer is the Q Exactive Plus (Thermo Fisher Scientific) (Figure 1.3). Here, for the first full MS scan (measurement of precursor ions m/z, MS1), ions within a wide range of m/z pass through the quadrupole (filters for specific m/z ranges), are trapped in the C-trap (stabilizing the ions using nitrogen and an electromagnetic field), before being injected into the orbitrap. The orbitrap works with a magnetic spindle, around which the ion spins from side to side. A big molecule will move slower, and a small one will move faster. Based on the image current, the exact mass of the molecule can be inferred. During a single sample measurement, as the peptides are sprayed into the mass spectrometer, MS1 spectra are acquired repeatedly. From each MS1 spectrum several MS/MS (MS2) acquisition events are triggered, which occur before the following MS1 spectrum is acquired [26–28]. Usually, in the case of discovery-based proteomics, the top N most abundant precursor ions from the MS1 scan are selected for fragmentation and MS/MS analysis. The quadrupole selects these ions for a specific m/z value, sending them through the C-trap to the higher-energy collision-induced dissociation (HCD) cell for fragmentation. Ion fragments are again stabilized in the C-trap and measured in the orbitrap.

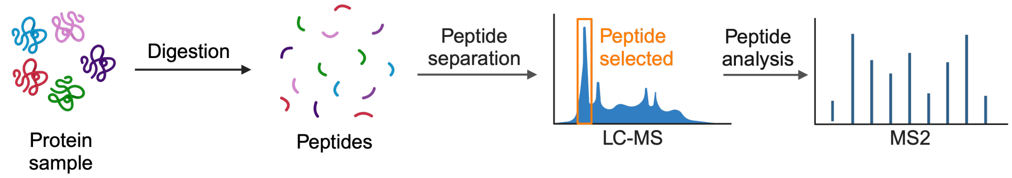


Figure 1.4.: Workflow of Data-dependent acquisition (DDA) mass spectrometry. Proteins are digested into small peptides, which are measure on the mass spectrometer. In the first MS scan (MS1), top N abundant peptides are selected for fargmentation and identification (MS2). LC-MS: liquid chromatography□mass spectroscopy. Created with BioRender.com

An example of a mass spectrometer using a TOF mass analyzer to acquire MS1 spectra rather than an orbitrap is the timsTOF (Bruker). A major difference between the timsTOF and the Q Exactive is the trapped ion mobility spectrometry (TIMS) element [28]. Briefly, the concept behind TIMS is using an electric field to hold ions stationary against a moving gas so that the drift force is compensated by the electric field and ions are separated based on their respective ion mobilities [29]. The ions are “eluted” gradually from the dual TIMS analyzer, separating different precursors. These are analyzed by a TOF mass analyzer, where ions are pulsed by an electric field and accelerated. All ions acquire the same kinetic energy and enter the flight tube, which is a field-free drift region where mass separation occurs. Ions with a lighter mass will have a shorter time of flight, whereas heavier ions will take longer to traverse the flight path toward the detector. Current time-of-flight analyzers have a reflectron device built in, which corrects for kinetic energy dispersion and spatial spread of ions that exhibit the same m/z but have varying velocities. This reflectron correction allows ions of the same m/z to arrive at the detector simultaneously. The reflectron device also increases the flight path length, improving mass resolution [30].
Each MS1 acquisition can trigger MS2 spectra before the next MS1, as with the Q Exactive. Similarly, ions will be filtered by the quadrupole and fragmented in the collision-induced dissociation (CID) cell [28]. MS2 spectra are used to identify peptides, whereas quantitation happens based on the MS1 spectrum. This setup of MS measurement, where the most abundant ions are selected for identification, is called data-dependent acquisition (DDA) (Figure 1.4) [26, 28, 31].

An alternative to DDA is data-independent acquisition (DIA). The most popular DIA methods are based on Sequential Window Acquisition of All Theoretical Mass Spectra (SWATH-MS), in which all m/z values within the MS1 range are included in fragmentation and identification. DIA allows excellent temporal resolution and can quantify proteins in complex mixtures over an extensive dynamic range, thereby overcoming the challenge of under-sampling when using DDA. For label-free quantification methods, this means greatly improved data completeness and increased proteomic depth. It offers high precision and reproducibility. However, the data generated is much larger and more complex because of the number of multiplexed MS spectra. Therefore, the database search methods developed for DDA analysis cannot be applied directly (see section "Analysis of proteome data") [28, 32].

1.3.2. Quantification methods

Mass spectrometry-based proteomics is not inherently quantitative. Proteolytic peptides have a wide range of physicochemical properties, such as size, charge, and hydrophobicity, which leads to significant differences in mass spectrometric measurement and response. Therefore, each peptide needs to be compared between experiments for accurate quantitation. When molecules only differ in their isotopic composition and have identical physical and chemical properties, they can be compared between different experiments [27, 33]. Technically, this can be achieved in one of two ways: (i) label-based quantification and (ii) label-free quantification (Figure 1.5).

For label-based quantification, peptides are labeled with groups of atoms that are identical but contain different isotopes, introducing an expected mass difference. This results in different m/z values, either of the peptides or their fragments, while conserving chemical properties such as retention time, ionizability, and fragmentation pattern [27, 28]. Different labeling methods are chemical, metabolic, or enzymatic. The most popular and widely used method for metabolic labeling is stable isotope labeling with amino acids in cell culture (SILAC). Here, the whole proteome of one condition is labeled with amino acids containing heavy isotopes (by exchanging amino acids in cell culture media or feeding organisms with labeled food, e.g., feeding *C. elegans* a heavy lysine- and arginine-labeled *E. coli* strain [34]). The natural "light" proteins are combined with the "heavy" proteins at the start of sample preparation, and the mass shift will be used in the MS1 spectra [27, 28, 33].

Examples of chemical labeling are dimethyl labeling (DML) or the introduction of tandem mass tags (TMT), an isobaric labeling method.

For DML, samples are digested as usual, with proteases such as trypsin. The derived peptides of the different samples are then labeled with isomeric dimethyl labels. A reductive amination reaction converts all primary amines (the N terminus and the side chain of lysine residues) in a peptide mixture to dimethylamines. Peptide triplets can be obtained using combinations of several isomers of formaldehyde and cyanoborohydride. They will differ by a minimum mass of 4 Da between different samples. The labeled samples are mixed and simultaneously analyzed by MS, whereby the mass difference of the dimethyl labels is found in MS1 and can be used to compare the peptide abundance in the different samples [35].

In isobaric labeling methods, such as TMT, peptides are labeled after digestion and combined before LC-MS/MS. Labeled peptides will coelute as a single peak with the same m/z value in an MS1 scan. Only fragmentation of the labeled peptides during the MS2 or MS3 in the collision-induced dissociation cell (CID or HCD) generates reporter ion peaks of differing mass, enabling quantification across samples [27, 28, 36, 37].

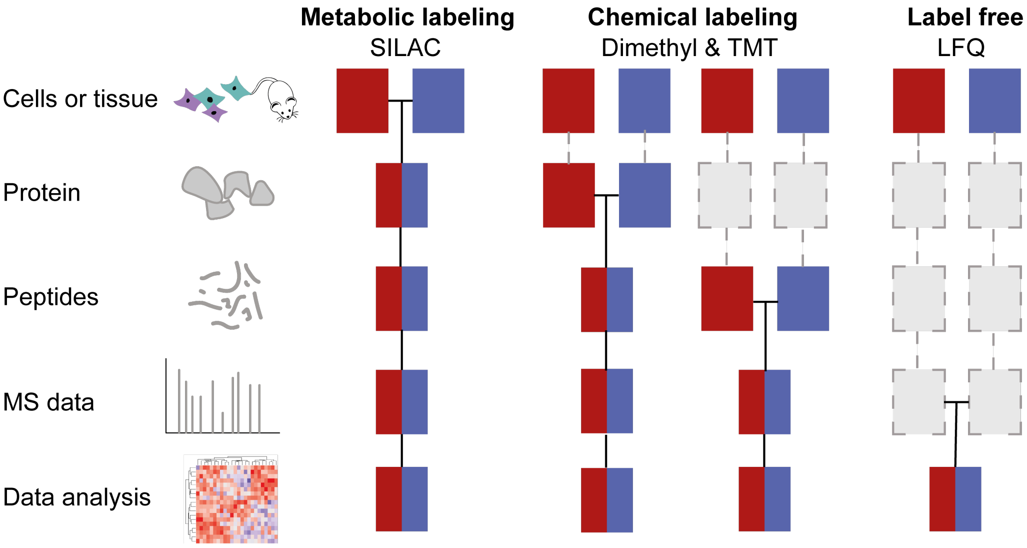


Figure 1.5.: Quantitation methods and multiplexing options for MS-based proteomics. Common quantitative mass spectrometry (MS)-based proteomics workflows and quantitation methods. Red and blue boxes boxes represent two experimental conditions. Horizontal lines indicate when samples are combined. Dashed lines indicate the points at which experimental variation and thus quantification errors can occur. Adapted from [38].

Label-free quantification (LFQ) methods aim to compare two or more samples by comparing the direct mass spectrometric signal intensity for any given peptide or using the number of acquired spectra matching a peptide (spectral counts). Intensity-based LFQ uses the MS signal response of intact peptides for quantification. Typically, this is accomplished by integrating the ion intensities of a precursor ion over its chromatographic elution profile. A precursor ion's MS1 peak is observed multiple times during its elution from the HPLC column, even if the precursor is only fragmented once during the whole run. We can use the height of this integrated chromatographic peak or the area under the curve (AUC) to measure the relative abundance of the peptide [28, 38].

In most cases, labeling-based methods for quantification will be the most accurate. Additionally, with LFQ, only one biological sample can be analyzed per MS run. For SILAC, each run can analyze two to three biological samples. With TMT, one run can analyze up to 18 biological samples [28]. However, labeling techniques typically limit direct comparison of experiments, whereas direct comparison between LFQ samples is unlimited. Additionally, there is evidence that label-free methods provide a higher dynamic range of quantification than stable isotope labeling and, therefore, may be advantageous when significant and global protein changes between experiments are observed. It is, therefore, worth considering experiment type, research question, and finances when deciding on a quantification method [27, 28, 38].

1.3.3. Analysis of proteome data

Quantitative proteomic data are complex. The raw data set produced by a single LC-MS/MS run is an extensive collection of spectra, each with a retention time, m/z values, intensities, and other metadata. Several software packages can process these data and perform peptide identification and quantification [28, 33].

One such software is MaxQuant [39]. For peptide identification, MaxQuant has implemented the Andromeda search engine, which performs the most commonly used identification approach, a database search. For this, the program needs a user-supplied reference proteome (FASTA). It will predict all peptides that could arise from the proteins in the database by enzymatic cleavage and predict the MS2 spectra of the corresponding peptide ions (precursors). These predicted peptides and their predicted spectra are compared with the experimental spectra to make peptide-spectrum matches (PSMs). After false discovery rate (FDR) control, through searching against reverse peptide sequences of the database, peptide identification is complete [28, 39].

After peptide identification, they will be matched to the proteins or genes from which they originated. This can be a somewhat ambiguous process, as protein sequences may match the same set of identified peptide sequences. If no unique identifying peptides are found, proteins will be grouped together in a protein group (PG). Then, relative PG quantities are generated using peptide intensities. Most software packages, including MaxQuant, allow quantification only with PG-unique peptides.

Notably, for LFQ, the popular MaxLFQ algorithm within MaxQuant can account for the fact that different peptides belonging to the same protein can have very different base intensities, for example, due to differing ionization efficiencies. In addition, MaxLFQ deals with missing values by taking available pairwise comparisons and using median ratios to compare and reliably estimate protein intensities. This is an important feature, as proteomic data usually are incomplete. Even the most advanced mass spectrometers can be overwhelmed by the number of peptides in a sample. As a result, only a subset of all proteins present can be identified. For protein quantification, it is mandatory to detect a protein in all experiments that will be compared, but this can be partially overcome by solutions as implemented by MaxLFQ [28, 33, 40].

The next steps in data analysis for generic (comparative) proteomic experiments consist of removing contaminants and filtering for protein groups confidently identified by two or more peptides. The data is usually log transformed to approximate a normal distribution for statistical testing and normalized to correct for inter-run technical variability. Further, as statistical methods often require complete data, missing values might be imputed or estimated, especially in the case of LFQ. Estimation can be done by averaging available values of the protein from other replicates or using related values from other proteins from the same experiment. Imputation is standard practice in proteomic data analysis, but it should be noted that estimating values will result in decreased statistical power [27, 28, 38].

1.4. Proteomics and transcriptomics to study DNA damage

Both exogenous and endogenous mutagens constantly threaten the stability of the genome. These stressors can damage the architecture and structure of the DNA, causing single-stranded breaks, double-stranded breaks, or chemical modifications to individual bases. To prevent genomic instability, a carefully orchestrated DNA damage response (DDR) functions to identify and repair damaged DNA [41]. The cellular response to DNA damage typically involves a wide range of cellular processes, such as gene expression modulation, protein and metabolic activity changes, and, in extreme cases, changes in DNA sequence or structure, all of which contribute differently to cellular phenotype [42].

The core of the cellular defense against DNA damage is formed by various DNA repair mechanisms, each with its specificity. Together, they can remove the vast majority of damage from the genome [43]. Generally, the DDR consists of a cascade of sensors, transducers, mediators, and effectors [44]. Recruitment of the appropriate sensor, transducer, mediator, and effector depends on the cell cycle, extent of damage, and type of DNA damage. They dictate which DNA repair pathway is induced [44]. Bases with minor chemical alterations that do not strongly disturb the DNA double-helix structure are substrates for Base Excision Repair (BER).

On the other hand, Nucleotide Excision Repair (NER) removes a broad spectrum of single-strand lesions that cause local helix destabilization. Two different modes of damage detection are functional in NER: transcription-coupled NER (TC-NER), which efficiently removes transcription-stalling lesions and allows fast resumption of transcription, and global genome NER (GG-NER), which localizes lesions anywhere in the genome. Lesions that are substrates for NER and BER are located only in one of the DNA strands and are removed with a “cut-and-patch”-mechanism. In these cases, the undamaged complementary strand is an accurate template for repairing the damaged strand. However, some damaging agents affect both strands, such as ionizing radiation, inducing DNA double-strand breaks (DSBs), and agents that produce inter-strand cross-links (ISCLs). These lesions are highly cytotoxic because they are more challenging to repair as the cell cannot rely on merely copying the information from the undamaged strand. Two distinct pathways, homologous recombination (HR) and non-homologous end-joining (NHEJ), repair DSBs and fall under double-strand break repair (DSBR) [43].

Until recently, all these processes and pathways were studied in isolation, neglecting the broader cellular context for challenge-response mechanism outcome [42]. As described above, the rise of omics technologies enables measuring interaction and changes at molecular resolution for genomes, proteomes, and metabolomes covering the whole cell. Quantitative mass spectrometry has aided in identifying novel factors involved in DNA damage repair and genome instability previously uncharacterized [45]. Affinity purification and proximity labeling techniques have been essential in identifying unknown factors and revealing unknown crosstalk. Global proteome measurements have allowed for a comprehensive view of the DNA damage response. It has presented an unbiased approach to studying all aspects of DNA repair rather than focusing on the previously associated candidates [46]. Omics data can help gain a systems-level understanding of dynamic cellular response mechanisms to perturbations or DNA-damaging agents [9, 42].

In Chapter 2 of this thesis, I used quantitative proteomics and transcriptomics to profile the DDR temporally. In Chapter 3, I used quantitative mass spectrometry with (DNA) affinity purification and phylointeractomic analysis to study DNA repair proteins across the Tree of Life. Together with other molecular and biochemical techniques, this has led to novel contributions to genome instability and DNA damage studies.

1.5. Omics in evolutionary analysis and applications

Not only has multi-omics integration become a trend and useful tool for constructing a comprehensive relationship between molecular signatures and phenotypic manifestations of a particular disease within a particular organism, it has also enabled more detailed study of processes ranging from subcellular to evolutionary, that drive biological organization. Subcellular and evolutionary processes, such as differential gene expression, speciation, and phenotypic plasticity, can operate over dramatically different timescales (milliseconds to billions of years) and are responsible for generating patterns of phenotypic variation [5, 47]. Whilst phenotypic variation is often studied at specific levels of biological organization to isolate processes working at a particular scale, the varying types of omics data can provide complementary inferences to link molecular and phenotypic variation to create an integrated view of evolutionary biology, ranging from molecular pathways to speciation [47]. Using evolutionary relationships between species, we can investigate both trait evolution and the impact of traits on ecological speciation rates. By applying epigenomic, transcriptomic, proteomic, and metabolomic data in a comparative framework, we can treat these molecular phenotypes as evolvable traits sorted across species [47].

Within this context, population geneticists have long sought to understand the contribution of natural selection to molecular evolution. The rates and patterns of molecular sequence evolution are estimated using comparative studies of orthologous genes. Orthologous sequences (sequences from distinct species that descended from a common ancestor) have been modified by an extensive evolutionary process with mutation rates varying by order of magnitude.
The genetic code of each organism, containing the translational key for DNA sequence into protein–amino acid sequence, is partially redundant, with multiple nucleotide triplets translating into the same amino acid. This redundancy is demonstrated by synonymous sites, for which specific changes in the coding DNA sequence do not change the amino acid sequence. Thus, the structure or function of the protein remains unchanged. Without selective forces, beneficial mutations may be selected and developed via drift effects [48]. Conversely, mutations that encode distinct amino acids (non-synonymous sites) might be selected against and disappear from the genome since they are unfavorable. Accounting for this rate variation under different levels of selective pressure can provide insight into the functional restrictions on proteins. Proteins with strict functional or structural requirements face significant purifying (negative) selective pressure, resulting in fewer amino acid modifications. Consequently, genes with a limited rate of evolution are prone to performing critical functions optimally. Unless their interaction networks are altered, the probability of improved performance is relatively low. Genes having redundant and non-central functions, as well as weaker constraints, evolve at a faster rate [49–51]. Different approaches have been proposed that use population genetics theory to quantify the rate and strength of positive selection acting in a species' genome. Methods can use patterns of between-species nucleotide divergence and within-species diversity to estimate positive selection parameters from population genomic data [52]. Ultimately, determination of the rate and strength of positive selection aids in understanding how genomes evolve, providing researchers with insights into the biological importance of genes of interest and species differences.

Chapter 4 of this thesis addresses the need for software that greatly reduces the manual input required for positive selection analyses. I co-developed a computational pipeline called AlexandrusPS, which facilitates researchers in performing correct and efficient large-scale evolutionary analysis. AlexandrusPS solves two additional challenges in the steps needed before performing the actual positive selection analysis. These are the need for accurate orthology predictions and sequence alignment. They are critical in positive selection analysis because including ancient paralogs, i.e., paralogs that have diverged during long timescales, has been shown to cause bias. The increased nonsynonymous substitution rate caused by decreased purifying selective pressure can result in two alternative fates of the gene copies. Either one of the paralogs becomes non-functional due to the lack of selective pressure and accumulation of mutations. Alternatively, in some cases, the functions and expression patterns of the gene pair may diverge substantially and give rise to novel functions or specializations in the organism called neofunctionalization [53].

In Chapter 3 of this thesis, I investigated the phylogenetic diversity in the recognition and repair of three well-established DNA lesions, primarily repaired by BER or RER. Previous literature has highlighted strong conservation of fundamental proteins in both pathways [54]. However, only by studying these pathways across the tree of life can the convergence and divergence of these different repair machinery be elucidated.

#

2

A systems view on the DNA damage response kinetics in *Tetrahymena thermophila*

Emily Nischwitz[1], Vivien A.C. Schoonenberg[1], Rachel Mullner, Susanne Zimbelmann, Douglas L. Chalker, Joshua J. Smith and Falk Butter

In preparation

---

[1] These authors contributed equally

2.1. Summary

This study combines transcriptomics and proteomics to study DNA damage kinetics across well-established treatments in the ciliate *Tetrahymena thermophila*. We treated *Tetrahymena* with six common DNA mutagens. The damaging agents used were ultraviolet light (UV, inducing nucleotide excision repair), hydrogen peroxide (HP, inducing base and nucleotide excision repair), methyl methanesulfonate (MMS, inducing base and nucleotide excision repair, hydroxyurea (HU, halting replication) ionizing radiation (IR, inducing double-stranded break repair), and cisplatin (inducing nucleotide excision repair and inter-crosslink repair). This large-scale data set of 6 treatment conditions and 7 time points (from 0 to 8 hours) integrates over 250 transcriptome and proteome measurements. We observed the upregulation of known DNA repair proteins and a global response of transcripts and proteins that have not yet been characterized. Using self-organizing maps, we classified different expression profile trends between proteins and transcripts in response to the mutagens, including PARP and PARP-related proteins. Utilizing a novel gene knockdown system in *Tetrahymena*, we are currently investigating the effect of DNA damage agents on several proteins of the PARP family, nominated by our analysis. In addition to the comprehensive analysis presented, the data can be explored via an accessible user interface at https://butterlab.imb-mainz.de/Tt_DDR/.

We are still exploring protein and transcript expression trends further through correlation analysis. Ultimately, our study identified novel candidates in the DNA damage response and provides new insights into current proteins of interest.

2.2. Zusammenfassung

In dieser Studie werden Transkriptomik und Proteomik kombiniert, um die Kinetik von DNA-Schäden bei verschiedenen etablierten Behandlungen des Ciliaten *Tetrahymena thermophila* untersucht. Wir haben *Tetrahymena* mit sechs bekannten DNA-Mutagenen behandelt. Bei den verwendeten Schadstoffen handelte es sich um ultraviolettes Licht (UV, induziert die Nukleotid-Exzisionsreparatur), Wasserstoffperoxid (HP, induziert die Basen- und Nukleotid-Exzisionsreparatur), Methylmethansulfonat (MMS, induziert die Basen- und Nukleotid-Exzisionsreparatur), Hydroxyharnstoff (HU, stoppt die Replikation), ionisierende Strahlung (IR, induziert die Reparatur von Doppelstrangbrüchen) und Cisplatin (induziert die Nukleotid-Exzisionsreparatur und die Reparatur zwischen den Querverbindungen). Dieser groß angelegte Datensatz mit 6 Behandlungsbedingungen und 7 Zeitpunkten (von 0 bis 8 Stunden) umfasst über 250 Transkriptom- und Proteom-Messungen. Wir beobachteten die Erhöhung der Expression bekannter DNA-Reparaturproteine und

eine globale Reaktion von Transkripten und Proteinen, die bisher noch nicht charakterisiert worden sind. Mithilfe von selbstorganisierenden Karten klassifizierten wir unterschiedliche Expressionsprofil-Trends zwischen Proteinen und Transkripten als Reaktion auf die Mutagene, darunter PARP und mit PARP verwandte Proteine. Unter Verwendung eines neuartigen Gen-Knockdown-Systems in *Tetrahymena* untersuchen wir derzeit die Auswirkungen von DNA-Schadstoffen auf mehrere Proteine der PARP-Familie, die durch unsere Analyse identifiziert wurden. Zusätzlich zu der hier präsentierten umfassenden Analyse können die Daten über eine zugängliche Benutzeroberfläche unter https://butterlab.imb-mainz.de/Tt_DDR/ betrachtet werden.

Wir sind noch damit befasst, die Trends bei der Expression von Proteinen und Transkripten durch Korrelationsanalysen weiter zu untersuchen. Insgesamt identifizierte unsere Studie neue Kandidaten für die DNA-Schadensreaktion und bietet neue Einsichten in aktuelle Proteine von Interesse.

2.3. Statement of Contribution

Emily Nischwitz and I led this study with the support of Falk Butter. I designed the initial large experimental setup together with Emily, which she implemented and executed with the help of Rachel Mullner and Susanne Zimbelmann. I led all aspects of the data analysis and was responsible for initial data visualization. Emily and I led in-depth data interpretation and finalization of visualization. Emily and I are writing the initial draft version of the manuscript with support from Falk Butter. We are currently completing the final analysis and biological validation experiments. We plan to submit by the end of 2023.

2.4. Abstract

A tightly regulated DNA damage response is critical to the overall integrity of the genome. Here, we combine transcriptomics and proteomics to study DNA damage kinetics across well-established treatments in the ciliate Tetrahymena thermophila. This extensive data set of 6 conditions (HU, MMS, IR, HP, cisplatin, and UV) and 7 time points (from 0 to 8 hours) integrating over 250 transcriptome and proteome measurements. We observed upregulation of known DNA repair proteins and a global dynamic of not yet characterized transcripts and proteins. Using self-organizing maps, we classify different expression profile trends in response to the treatments, including PARP and PARP-related proteins. Utilizing a novel gene knockdown system in Tetrahymena, we investigate the effect of DNA damage agents for [several] proteins of the PARP family. In addition to the comprehensive analysis presented here, the data can be explored via an accessible user interface at https://butterlab.imb-mainz.de/Tt_DDR/. Ultimately, our study identified novel candidates in the DNA damage response and provides new insights into current proteins of interest.

2.5. Introduction

Environmental genotoxic stressors create DNA damage that poses a threat to the stability and integrity of the genome. It is, therefore, critical to have a carefully regulated orchestra of DNA damage response factors and pathways [41]. DNA damage repair activity is required in all living organisms, and the dysregulation of any of these pathways has been correlated with disease [54, 55]. Primary DNA repair pathways include nucleotide excision repair (NER), base excision repair (BER), mismatch repair (MMR), homologous recombination (HR), non-homologous end joining (NHEJ), and interstrand crosslink repair (ICL) [56–60].

Exogenous mutagens can induce damage lesions that are associated with particular repair pathways. UV exposure typically results in pyrimidine (6-4) pyrimidone photoproducts ((6-4) PPs) and cis-syn cyclobutane pyrimidine dimers [61], repaired by NER. Cisplatin (CPT) causes covalent bonds between base pairs on different DNA strands, referred to as interstrand crosslinks (ICLs) [62]; NER often repairs this damage. However, there is a cell cycle-dependent compilation of various repair pathways to address this damage, including HR, NER, translesion synthesis (TLS), and, in humans, the Fanconi Anemia (FA) pathway [58, 63]. Hydrogen peroxide (HP) and methyl methanesulfonate (MMS) cause oxidative and alkylative damage, respectively [64]. Previously, BER was thought to be the primary repair pathway to resolve these lesions. However, growing evidence highlights the interdependence of both BER and NER [65, 66]. R causes direct DSBs, which are repaired by either homologous recombination (HR) or non-homologous end joining (NHEJ) [57]. In addition to these direct

DSBs, a large number of reactive oxygen species are created by IR, which BER and NER can repair. HU does not damage DNA through direct interaction with DNA or the creation of a deleterious byproduct but inhibits the enzyme ribonucleotide reductase (RR). The inhibition of RR drastically reduces the available amount of deoxynucleotide triphosphate pools, causing large degrees of replication stress [67, 68].

Each damaging agent specifically induces one of these known DNA damage repair pathways. However, a large global and temporal DNA damage response occurs in the cell. To evaluate this globally, comprehensive DDR studies utilizing omics methods are vital. Even within well-studied unicellular eukaryotic organisms, such as *S. cerevisiae*, there have been limited studies of the proteomic or transcriptomic response to DNA damage [69–75]. These studies often lack either a combined transcriptome and proteome approach or only consider a singular or limited time point in the DDR.

Here, we studied DNA damage repair kinetics from a global transcriptomic and proteomic perspective in *Tetrahymena thermophila (Tetrahymena)*, a ciliate with a unique nuclear architecture containing a macronucleus (MAC), and the germline containing micronucleus (MIC). To obtain a systematic comparative overview of the kinetics of DNA damage repair in a eukaryotic organism, we performed transcriptome and proteome measurements over 8 hours, with six well-established genotoxic treatments invoking different DNA damage repair pathways.

2.6. Results

2.6.1. Known DNA damage repair factors are differentially regulated in response to genotoxic stressors

We treated *Tetrahymena* with six well-established treatments to study DNA damage response kinetics. The damaging agents were 254 nm ultraviolet light (UV), cis-diamine platinum (II) dichloride (CPT), hydrogen peroxide (HP), methyl methanesulfonate (MMS), ionizing radiation (IR), and hydroxyurea (HU). The treatment conditions were determined either by the establishment of EC50 or from previous DNA damage studies of *Tetrahymena* [76–78] (Table A.1). To obtain transcriptome and proteome expression information, we harvested samples at 0, 1, 2, 3, 4, 6, and 8 hours (H0-H8) in quadruplicate and performed mRNA sequencing (RNA-seq) and high-resolution mass spectrometry (MS) measurements (Figure 2.1A). We measured all transcriptomes and proteomes in sets of three, two treatments paired with a non-treatment, which were collected and processed together. We could calculate changes in transcript expression and protein intensity in each drug treatment condition over a matched nontreated condition, thereby correcting for any potential batch effect ($\log_2$ fold change values, Figure A.1). After stringent filtering across treatments and time points, we quantified 20,443 transcripts and 6,551 protein groups, with 99.6% of proteins associated with transcriptome data (Figure 2.1B). We examined differential upregulation of primary DNA repair pathways to verify each treatment induced the anticipated DNA damage. We compiled gene lists for five DNA repair pathways: NER [60, 79], BER [56, 80], MMR [81–83], DSBR [84–87], ICL [88], and general DDR [41, 89]. These lists containing 130 DNA repair genes are not intended to be comprehensive but highlight previously established critical genes involved in the different DNA repair pathways (Table A.2). Across all treatments, 81.4-100% of detected pathway-associated genes were upregulated, and 82.4-96% of detected pathway-associated proteins were upregulated. Of the detected global DNA damage response transcripts and proteins, 81.4% and 88.9% were upregulated in the transcriptome and proteome datasets, respectively. We examined the amount of upregulated transcripts and proteins for each treatment with its commonly associated DNA repair pathway. Across all treatments, 68.4-96.6% of the transcripts and 33.3-92% of the proteins associated with the respective pathways were upregulated (Figure 2.1C).

The selected members of the primary DNA repair pathways were further examined for each treatment over time (Figure 2.1D). Through hierarchical clustering, we found a group of 10 transcripts (MSH6L3, RAD51, RAD4, SNML1, RAD5, RLP1, RAD5L4, RFA1, TTHERM_00316410 (Rad3 homolog), and TTHERM_00391570 (Rev3 homolog)) with a minimum 2.3 $\log_2$ fold change across all treatments, indicating an unexpected core DNA damage response that crosses MMR, DSBR, and NER pathways, as well as general responders, regardless of damage origin. The expression profiles reveal unique kinetics amongst the members and between gene and protein regulation.

2.6.2. Genotoxic stressors induce core and specific global dynamic gene expression responses

We examined the expression dynamics of both the transcriptome and proteome over time by calculating the Gini coefficient for every quantified transcript and protein (Figure 2.2). We applied a Gini coefficient filter of the 60th quantile (Gini score > 0.042) to the transcriptome to separate dynamic and stable transcripts (Figure 2.2A). In addition to our dynamicity filter, to select transcripts up- and downregulated in response to the DNA damaging agents, we applied a $\log_2$ fold filter, requiring the expression change over the time course to reach either more than 1 or less than -1 ($\log_2$ fold change) once, and a significance filter, of adjusted p-value < 0.05 (FDR). Of the 20,443 detected transcripts, we classified 8,815 as dynamic, surpassing these thresholds. Amongst them were the previously characterized DSBR gene *RAD51* and the MSH6 homolog, *MSH6L3* [76, 90, 91], whereas TTLL6B was found to be a stable transcript (Figure 2.2B).

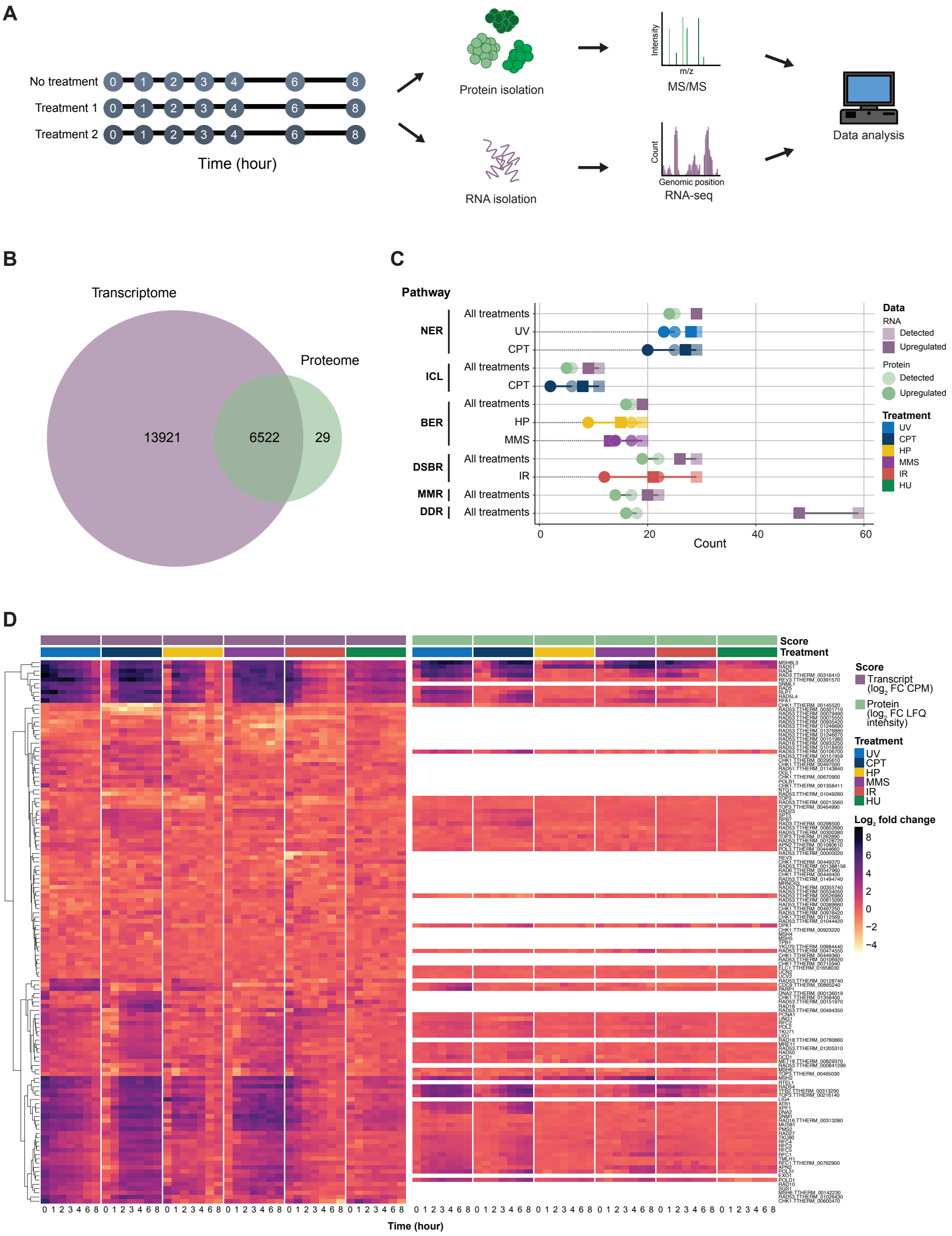


Figure 2.1.: Screen to explore the kinetics of DNA damage response in *Tetrahymena*. A) Schematic of the screen workflow. Cells were treated with a mutagenic agent, and samples were harvested incrementally over eight hours. At each time point, samples were collected for RNA sequencing and quantitative mass spectrometry processing. **B)** Venn diagram depicting the overlap between the identified transcripts and proteins. **C)** Lollipop plot of enriched DNA damage repair factors. **D)** Heat map of hierarchical clustering of DNA repair genes of interest. NER: nucleotide excision repair, ICL: inter crosslink repair, BER: base excision repair, DSBR: double-strand break repair, MMR: mismatch repair, DDR: DNA damage response, UV: ultraviolet light, CPT: cisplatin, HP: hydrogen peroxide, MMS: methyl methanesulfonate, IR: ionizing radiation, HU: hydroxyurea

There were 57 overlaps of shared dynamic transcripts among treatments, 42 of which were significantly more than expected (p-value ≤0.04, Fisher's exact test). All seven groups that contained five or more treatments were significant (p-value < 0.001) (Figure 2.2C). Of the 78 genes in these overlaps, 39 had no yeast homolog. This indicates both a strongly conserved and unique DNA damage response in *Tetrahymena*. Amongst these genes with no yeast homolog, three PARP and PARP-correlated genes were included: PCP3 (PARP12, TTHERM_00467770), PARP3 (TTHERM_00030430), and PCP5 (PZN1, TTHERM_00773650). Of the 15 PARP and PARP-correlated protein families, 12 members were dynamic in one or more treatments. While PCP3, PARP3, and PCP5 act as core responders, there are also damage-specific PARP responses. For example, PARP1, PARP7, and PCP1 only have a dynamic expression profile in response to UV. This same core and highly treatment-specific dynamic response is observed globally as well.

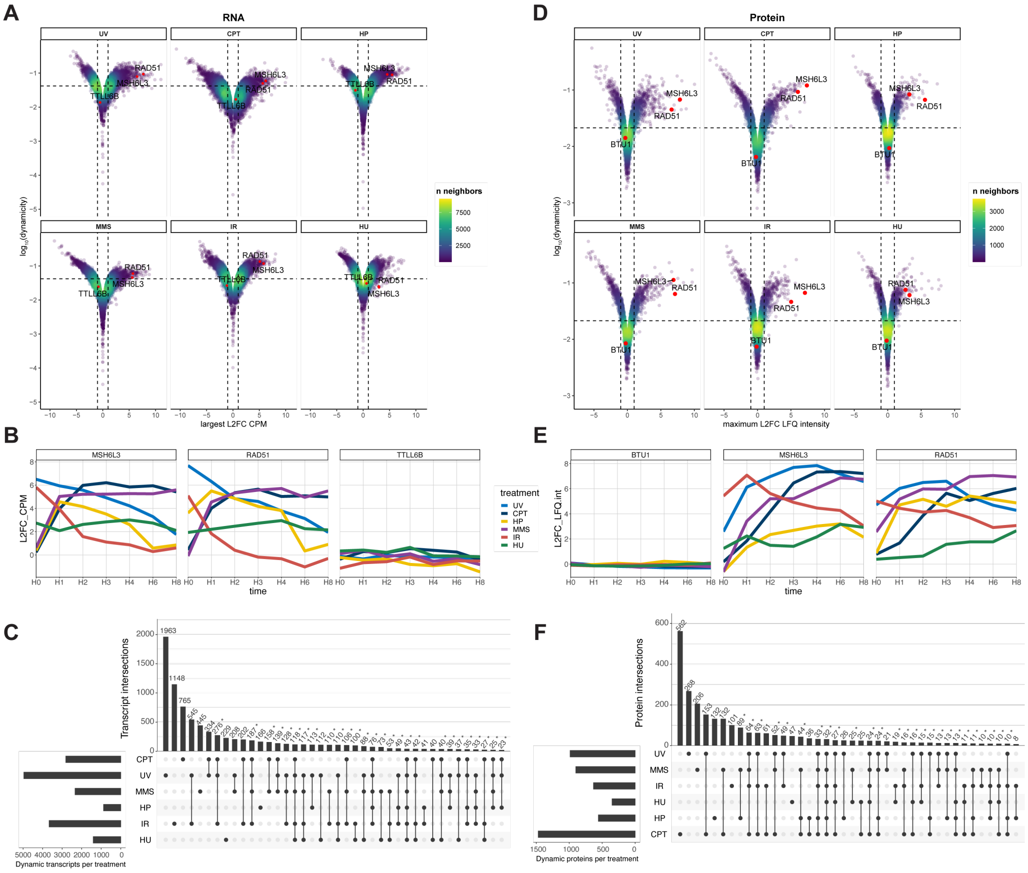


Figure 2.2.: Mutagenic treatments cause a global dynamic response. A) Volcano plots plotting dynamic transcripts for each treatment. The x-axis contains the maximal positive or negative fold change during the time course, and the y-axis contains a Gini score evaluating the dynamicity throughout the entire time course. **B)** Example line plots of the expression profiles dynamic (MSH6L3 and RAD51) and stable transcripts (TTLL6B). **C)** Upset plots of overlapping dynamic transcripts between treatments. **D)** Volcano plots plotting dynamic proteins for each treatment. The x-axis contains the maximal positive or negative fold change during the time course, and the y-axis contains a Gini score evaluating the dynamicity throughout the entire time course. **E)** Line plots of the expression profiles dynamic (MSH6L3 and RAD51) and stable proteins (BTU1). **F)** Upset plots of overlapping dynamic proteins between treatments.

We performed the same analysis with the 6,551 proteins quantified across these six treatments. Here, the Gini coefficient threshold for dynamicity was also set at the 60th quantile (Gini score > 0.021 with a minimum $\log_2$ fold change $> |1|$, and p-value < 0.05 (Welch t-test) at any point in the time course) (Figure 2.2D). We found a total of 2,582 proteins to be dynamically regulated. RAD51 and MSH6L3 were dynamic in each treatment (Figure 2.2E). Of the 57 overlapping groups of dynamically upregulated proteins, there was significantly more overlap than expected in 38 overlaps ($p<0.05$, Fisher's exact test). (Figure 2.2F). Thirty-three proteins were present in overlaps containing five or more treatments. As for the dynamic transcripts, regulated proteins across all treatments had an overrepresentation of the GO terms 'DNA repair' and 'cellular response to damage'. Of these 33 core responders, 15 proteins have no homolog in *S. cerevisiae*.

2.6.3. Early and unique transcriptional response is critical to the DNA damage response

To cluster the 8,815 dynamic and differentially regulated transcripts, we used self-organizing maps (SOMs), an unsupervised machine-learning approach. The transcripts detected across all six treatments fell into 15 distinct clusters (T1-T15); 689 transcripts could not be assigned to a cluster (Figure 2.3, Figure A.2A). All of these clusters have unique temporal and degrees of response. T1 and T2 have a variety of initial intensities but overall trends of downregulation. In contrast, T4 and T5 show variable peaks of regulation followed by downregulation. In T8, T10, T12, T13, and T14, CPT-treated cells consistently showed the most extreme differential regulation, while all other treatments showed variable degrees of differential regulation. Next, each transcript was mapped to its respective homolog(s) in *S. cerevisiae*, and functional enrichment analysis was performed using Gene Ontology (GO) for these *S. cerevisiae* homologs for each cluster (Figure 2.3B).

Cluster T10 uniquely showed an overrepresentation of genes related to 'DNA repair', 'cellular response to DNA damage stimulus', 'DNA replication', and 'DNA metabolic process'. Within this cluster's average expression profile for CPT-treated cells, there is an immediate strong and continual upregulation. This response could be due to the known long half-life of CPT [92]. At H0, the average expression profile of IR- and UV-treated cells shows strong upregulation, followed by gradual downregulation throughout the remaining time points. This also reflects these specific treatments, as UV and IR treatments had only one initial application and were not sustained in culture. This cluster also has a moderate increase in the average expression of HP-, MMS-, and HU-treated cells. Additionally, 17 DNA damage response proteins were found in this cluster (Table A.2).

Further general responses to DNA damage were found also in other clusters. All histones (T7), 20S ribosomal proteins (T2), and dense core gran-

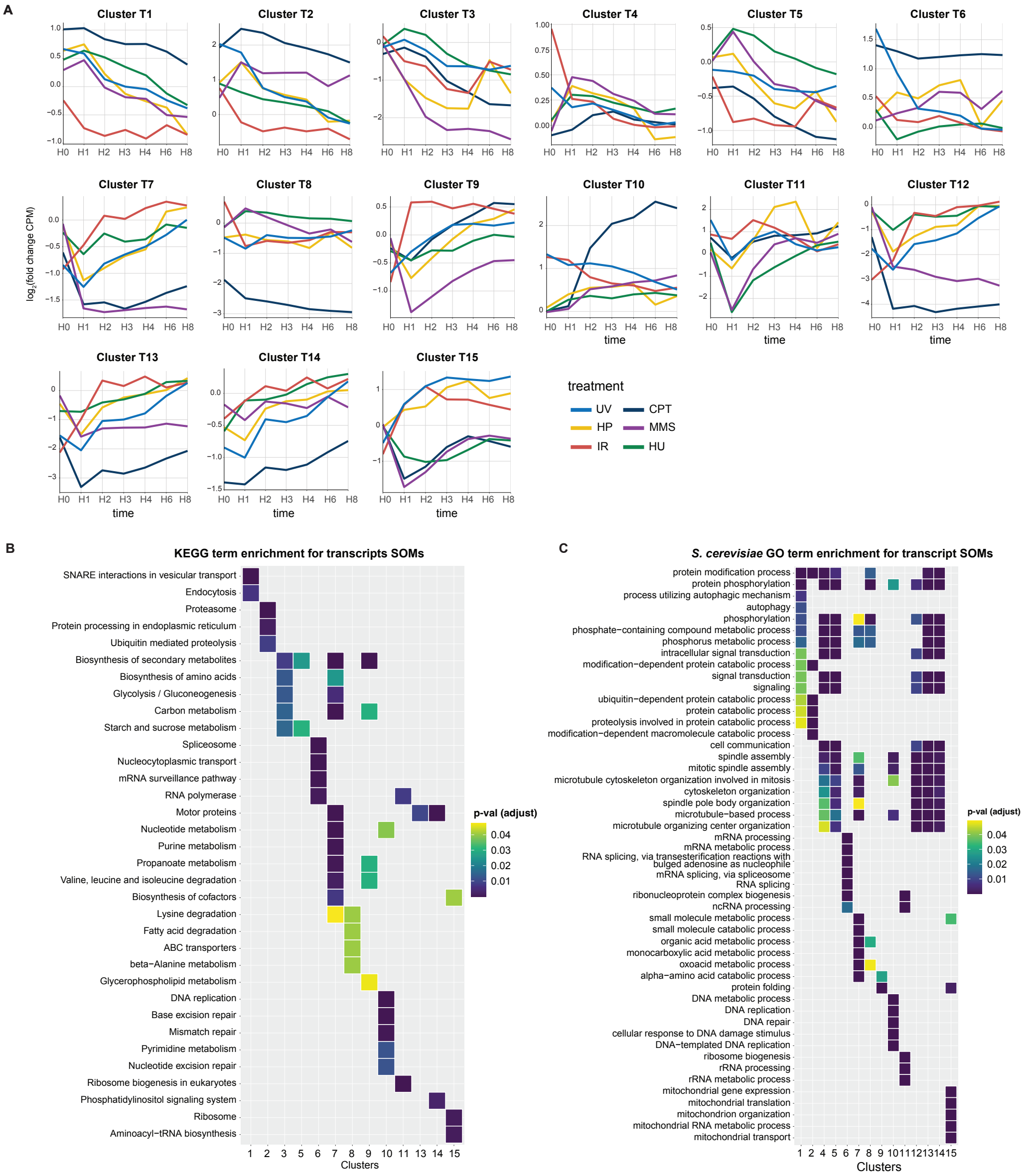


Figure 2.3.: Transcript clusters reveal dynamic DNA damage response. A) Average expression profiles for clusters of dynamic transcripts. Using an unsupervised machine learning technique, we clustered dynamic transcripts based on their shared expression profiles into 15 clusters. These line graphs are the average expression profiles for each treatment. **B)** Heat map of functional enrichment analysis using KEGG. Each row contains an over-represented KEGG term with a gradient representing the adjusted p-value. **C)** Heat map of functional enrichment analysis using GO analysis. Each gene was mapped to its respective homolog(s) in *S. cerevisiae*. Each row contains an over-represented GO term with a gradient representing the adjusted p-value.

ules (T3) clustered together showed primarily similar regulation trends for each treatment. The degree and time point of decline depended on the actual treatment, but each gene family or complex subunit behaved similarly within each treatment. In contrast, the three families of previously studied chromatin remodelers in *Tetrahymena*, the Poly-(ADP-ribose) polymerases (PARPs)/PARP-associated proteins, histone acetyltransferases (HATs), histone deacetylases (HDACs) are differentially regulated and span across multiple clusters [93–97]. The PARP and PARP-correlated proteins mediate DNA repair by chromatin modifications via ADP-ribosylation and direct binding, modification, and recruitment of DNA repair proteins [98, 99]. PARP7, PARP8, and PARP12 (T2), PCP1 (T7), PARP6 (T9), and PARP2 and PARP5 (T10) all showed unique responses to DNA damage. The histone acetylases and deacetylases are critical to changing chromatin architecture to facilitate DNA repair [97]. The histone acetylases (HATs) HAT1 (T10) and MYST2 (T2) and histone deacetylases THD4, THD17, and THD18a (T2, T3, T6, respectively) also showed greatly differential regulation. This indicates that each of the PARPs, HATs, and HDACs in *Tetrahymena* has a particular role.

In another example, within the 15 clusters, nine clusters are enriched for 'protein phosphorylation' (T1, T2, T4, T5, T8, T10, T12-14). It has been previously reported that phosphorylation plays a critical role in the processing of interstrand crosslinks, as well as preventing ICL proteins from conducting inappropriate repair [58, 100–102]. Overall, it is clear that protein phosphorylation is critical to immediate and sustained DNA damage response as a whole in *Tetrahymena*.

The transcriptome data shows specific transcriptional regulation kinetics of the DNA damage response dependent on the genotoxic stressor.

2.6.4. Protein expression over time reveals specific trends involved in DNA damage response

As for the transcriptome, we used self-organizing maps clustering 2,582 dynamically expressed proteins into seven distinct clusters (P1-P7); 202 proteins could not be assigned (Figure 2.4, Figure A.2B). Each protein included in these seven clusters was mapped to its respective homolog(s) in *S. cerevisiae*, and functional enrichment analysis was performed using GO (Figure 2.4B). Cluster P6 uniquely showed an overrepresentation of genes related to 'DNA repair', 'cellular response to DNA damage stimulus', 'DNA replication', and 'DNA metabolic process'. In P6, there were 15 known DNA damage factors, which included ATR1, RAD53/Chk1, TKU80, RAD3, DNA2, three members of the RFC complex, and TKU80. ATR1 and TKU80 are critical in DNA damage response and conjugation in *Tetrahymena* [77, 103].

In this cluster, two MMR proteins, MSH3L6 and TMLH1, were also identified. Together with MSH2, MSH3 is part of the MMR MutSβ complex, which repairs larger insertions and deletions. Alternatively, if MSH6 and MSH2 form an MMR recognition heterodimer MutSα, one to two base pair mismatches and indels are repaired. Intriguingly, MSH6 is found in P3, indicating a differential regulation of these portions of the MMR recognition complex. TMHL1 and PMS2 interact with the recognition complex to initiate cleavage events. Our clustering data suggests that TMLH1 expression profiles are more comparable to MSH3 than MSH6.

Similarly to transcriptional regulation, the chromatin remodelers within the PARP family were differentially regulated across clusters. However, this family of genes is not just being regulated at the transcriptional level. For example, PARP7, PARP8, and PARP12 all fell within T2, whereas now they are found within P6, P7, and P4, respectively.

Other gene families not directly responsible for DNA repair with a similar time-dependent transcriptional regulation fell into different clusters and showed diverging protein-level expression patterns. For example, all dynamic dense core granules clustered in T3 but were part of two different protein clusters (P2 and P3). All histones were transcriptionally regulated similarly (T33), whereas the histone proteins had different protein expression profiles (P6 and P7). Also, the DNA repair-related PARP family was differentially regulated. While the dynamic transcripts also showed a differential regulation across all clusters, some of the PARP transcripts expression profiles were clustered in the same SOM. However, there were some instances of similar protein and transcriptional regulation, such as for the 20S proteasome (T2 and P7) and transcription-related factors (T6 and P7). Generally, the differences between transcriptome and protein expression profiles indicate additional regulation at the protein level.

2.6.5. Using novel knockdown system to characterize highly dynamic DNA repair proteins and PARP-correlated proteins

Utilizing a novel knockdown system, we take advantage of the unique phenotypic characteristics of a mutant strain of Beta tubulin 1, BTU-1, in *Tetrahymena* [104]. We targeted the *btu-1* (K350M; pac^{s} loci) allele of CU522. Previously characterized due to its unique effects on macronuclear development, these mutants are sensitive to several selective pressures, including sublethal treatments of Paclitaxel. This microtubule stabilizer is used for positive selection of successful transformants [105, 106]. The novel knockdown system contains a designed hairpin flanked by two Beta tubulin arms in the base pUC118 construct. After being integrated via biolistic bombardment and undergoing positive selection, cells were screened with whole-cell PCR for successful integration. Then, to verify the successful reduction of protein levels, quantitative mass spectrometry was used to measure the reduction in protein levels.

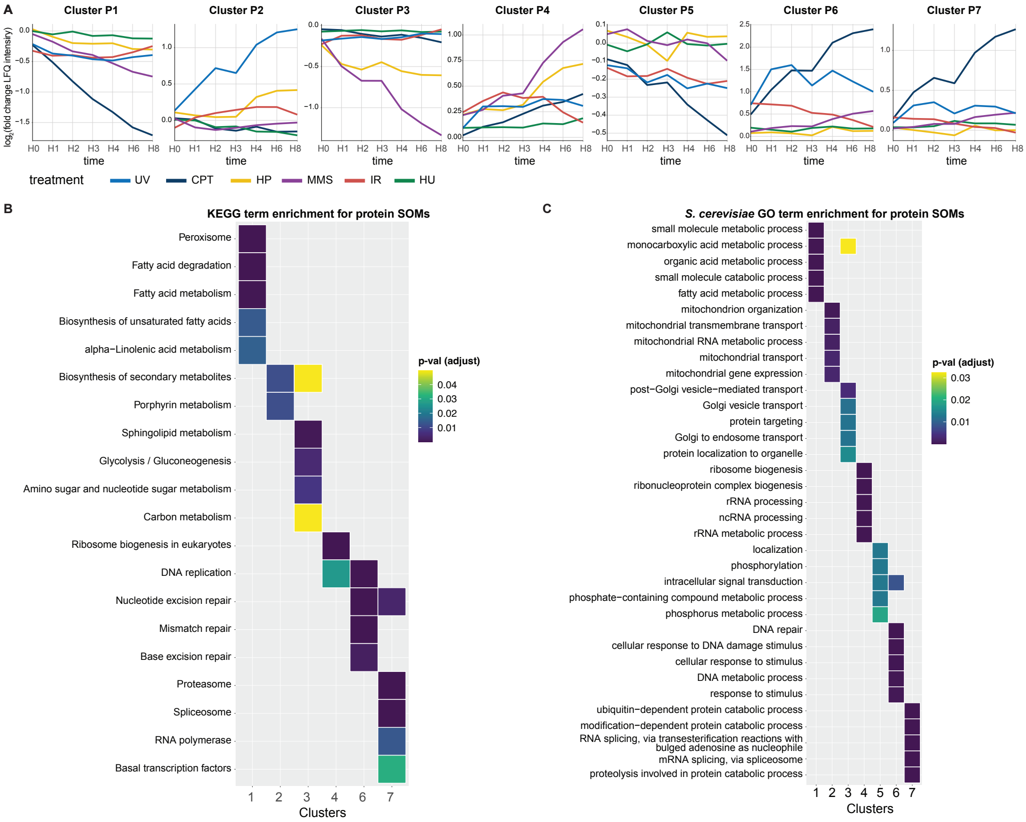


Figure 2.4.: Protein clusters reveal a specific and dynamic DNA damage response. **A)** Average expression profiles for clusters of dynamic transcripts. Using an unsupervised machine learning technique, we clustered dynamic transcripts based on their shared expression profiles into 7 different clusters. These line graphs are the average expression profiles for each treatment. **B)** Heat map of functional enrichment analysis using KEGG. Each row contains an over-represented KEGG term with a gradient representing the adjusted p-value. **C)** Heat map of functional enrichment analysis using GO analysis. Each gene was mapped to its respective homolog(s) in *S. cerevisiae*. Each row contains an overrepresented GO term with a gradient representing the adjusted p-value.

2.7. Methods

2.7.1. Cell culture

The *Tetrahymena thermophila* wildtype strain SB210 (Tetrahymena Stock Center) was used throughout the study. Cultures were grown in a medium of 2% proteose peptone (BD Biosciences), 0.2% yeast extract (BD Biosciences), 12 μM FeCl, and 1x Penicillin/Streptomycin/Funizone (Hyclone) at 30 °C at 100-150 rotations per minute.

2.7.2. Collection of *Tetrahymena* for mass spectrometry and RNA sequencing

Tetrahymena were grown to a concentration between 1.5x10^5-3x10^5 cells/ml in 500 ml cultures. Samples were treated with six different conditions, and two treatments were grouped with one nontreated group (as MMS, HP; CP, UV; and HU, IR). Details of treatments are described in Table A.1. Cells were harvested at 0, 1, 2, 3, 4, 6, and 8 hours after the initial treatment. To collect samples for later quantitative mass spectrometry, 5x10^4 cells were centrifuged at 9,400 xg for 5 minutes. The supernatant was removed, and cells were washed with 1 ml 10 mM Tris-HCl (pH=7.5) and centrifuged at 9,400xg for 5 minutes. The supernatant was discarded, leaving a total of $\sim$15 μl of cells and Tris, and 5 μl of 4x LDS (Thermo) and 2 μl of 1M DTT (Sigma) were added. Samples were heated to 90 °C for 10 minutes. Samples were stored at -20 °C until mass spectrometry sample preparation. To collect samples for later RNA sequencing (RNA-seq), 5 ml of cells were collected and centrifuged at 1,400xg for 3 minutes. The supernatant was decanted, and cells were washed with 5 ml 10 mM Tris-HCl (pH=7.5). Cells were centrifuged at 1,400xg for 3 minutes, and the supernatant was removed. The cell pellet was resuspended in 600 ul Buffer RLT (Qiagen, RNeasy mini kit), flash frozen in liquid nitrogen, and stored at -80 °C until RNA sequencing sample preparation.

2.7.3. Mass spectrometry sample preparation

LDS sample was loaded on a 4-12% NuPage NOVEX Bis-Tris gel (Thermo) and ran for 10 min at 180V in 1x MES buffer (Thermo Fisher Scientific). Samples were processed as previously described [107]. In short, the gel was stained and fixed with Coomassie Brilliant Blue G250 (Sigma Aldrich); initial destaining of the gels was done overnight with water. Gel pieces were cut, further destained with 50% EtOH / 50 mM ammonium bicarbonate (ABC) and dehydrated with acetonitrile (VWR), reduced with 10 mM DTT (Sigma) and alkylated using iodoacetamide (Sigma), and subsequently again dehydrated with acetonitrile (VWR) and digested with 1 µg of MS-grade trypsin (Sigma) at 37 °C overnight. The peptides were eluted from the gel pieces, loaded onto activated C18 material (Empore) StageTips [108], and stored at 4 °C until elution and measurement.

2.7.4. Mass spectrometry measurement

Peptides were eluted from the StageTips using 80% acetonitrile / 0.1% formic acid and concentrated prior to loading on an Easy-nLC-1200 system coupled to an Orbitrap Exploris 480 mass spectrometer (Thermo Fisher). The peptides were loaded on a 50 cm column (75 µm inner diameter, New Objective) in-house packed with ReproSil-Pur 120 C18-AQ (Dr. Maisch GmbH). We used a 103-min gradient from 3% to 40% acetonitrile with 0.1% formic acid at a flow of 250 nl/min. The mass spectrometer was operated in positive ion mode with a top 20 MS/MS data-dependent acquisition strategy of one MS full scan (scan range 300 - 1,650 m/z; 60,000 resolution; normalized AGC target 300%; max IT 28 ms) and up to twenty MS/MS scans (15,000 resolution; AGC target 100%, max IT 40 ms; isolation window 1.4 m/z) with peptide match preferred using HCD fragmentation.

2.7.5. Mass spectrometry data analysis

Raw files were analyzed using MaxQuant (version 1.6.10.43). As a search space, the *T. thermophila* protein database was used (June 2014, TGD). Oxidation and acetylation were set as variable modifications, Carbamidomethylation as fixed modification. Fast LFQ was used to calculate and normalize intensities. The minimum ratio count used was 2. Match between runs was used to match within each time point per treatment and to the time points right before and after, with a match time window of 0.7 min, match ion mobility window of 0.05, an alignment time window of 20 min, and alignment ion mobility of 1. Matching of unidentified features was deactivated. Label minimum ratio count 2 and unique + razor peptides were used for protein quantification.

2.7.6. RNA sample preparation and sequencing

Previously obtained samples were thawed on ice. RNA isolation was performed with RNeasy mini kit (Qiagen) per manufacturer instructions with the addition of the optional DNaseI on column digestion. This digestion was carried out with 3 units of DNaseI (Qiagen) per sample, and samples were digested for 15 minutes on the column at room temperature. NGS library prep was performed with Lexogen's QuantSeq 3'mRNA-Seq Library Prep Kit FWD following Lexogen's standard protocol (015UG009V0252). Libraries were prepared with a starting amount of 300 ng and amplified in 14 PCR cycles. Libraries were profiled in a High Sensitivity DNA on a

2100 Bioanalyzer (Agilent Technologies) and quantified using the Qubit dsDNA HS Assay Kit in a Qubit 2.0 Fluorometer (Life Technologies). All libraries from the two treatments and coordinating non-treatment were pooled together in equimolar ratio and sequenced on 1 NextSeq 500 high output flow cell, SR for 1x84 cycles plus 7 cycles for the index read.

2.7.7. RNA-seq analysis

All demultiplexed, raw sequencing files of each treatment set were analyzed together. Initial analysis was done through a modified version of the NGS pipeline by the bioinformatics core facility of the IMB (available at https://gitlab.rlp.net/imbforge/NGSpipe2go). For reference, "subread2rnatypes", "genebodyCov2", "rMATS", and the GO enrichment analysis were removed from the pipeline. In short, the library quality was assessed with FastQC before alignment against the *T. thermophila* genome assembly SB210 and a custom-built GTF file, which included gene annotations from *T. thermophila* (TGD, T_thermophila_June2014.gff3). Alignment was performed with STAR aligner version 2.7.3a [109]. Reads mapping to annotated features in the custom GTF file were counted with featureCounts [110]. Initial CPM counts were calculated with DESeq2 [111] in R [112].

2.7.8. Further bioinformatic analysis

All further analysis was done with scripts developed in R [112], incorporating ggplot2 for visualization [113] among other packages.

For proteome data, contaminants, reverse database hits, protein groups only identified by site, and protein groups with less than two peptides (at least one classified as unique) were removed. Additionally, only protein groups present in at least 2 out of 4 technical replicates were kept. Missing values were imputed by shifting a compressed beta distribution obtained from the LFQ intensity values to the limit of quantitation (between 0.2 and 2.5 percentile of the measured intensity distribution per sample). LFQ intensities were log2 transformed, after which fold changes for individual comparisons of time points or strains could be calculated per protein; a Welch t-test was used to calculate p-values. The general protein enrichment threshold was set to a p-value lower than 0.05 and an absolute fold change higher than 1. All calculated values can be found in the supplemental data.

For transcriptome data, transcripts that did not have any CPM value across the time points and treatments below the 25th quantile of all CPM values ($CPM < 1.673028$) were removed (Figure A.3). All CPM values were log2 transformed. Differential regulation thresholds were set at $L2FC > 1$ or < -1, and adjusted p-value $(FDR) < 0.05$.

The dynamicity of transcripts or proteins was calculated using the Gini ratio, as described before [114, 115]. Statistical testing of overlaps of dynamic genes was done with the R package SuperExactTest [116]. Functional enrichment analysis was performed using Kyoto Encyclopedia of Genes and Genomes (KEGG) [117], Gene Ontology [23], and the ClusterProfiler R package [118, 119] for statistical analysis. Terms for groups of enriched proteins were assessed for overrepresentation with a Fisher's exact test against all terms found in our complete dataset as background. The enrichment threshold was set to an adjusted (FDR) p-value < 0.05. Self-organizing map (SOM) clustering was done with the help of the Kohonen package in R [120].

All data can be explored through a user-friendly web interface at https://butterlab.imb-mainz.de/Tt_DDR. This web interface was designed and built with the use of R Shiny. All data and code for the analysis in this study was written in R and is freely available via the workflowr [121] website https://vivienschoonenberg.gitlab.io/Tetddr_wflowr/ or https://gitlab.com/vivienschoonenberg/Tetddr_wflowr.

2.7.9. Data and code availability

All raw data generated and used for this study will be deposited in GEO and ProteomeXchange. Calculated values and outcomes can be found in the supplemental data of this study.

2.8. Acknowledgements

We thank the IMB Media Lab, the Electronics Workshop, and the Proteomics Core Facility for their assistance. We thank Michal Levin and Albert Fradera-Sola for their input and advice for data analysis. Support by the IMB Genomics Core Facility and the use of its NextSeq500 (funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – INST 247/870-1 FUGG) is gratefully acknowledged.

2.8.1. Funding

This project was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) [Project-ID 393547839-SFB 1361]. VACS is supported by the Joachim Herz Stiftung through an Add-On fellowship for Interdisciplinary Life Science.

3

DNA damage repair proteins across the Tree of Life

Emily Nischwitz[1], Vivien A.C. Schoonenberg[1], Albert Fradera-Sola,
Mario Dejung, Olga Vydzhak, Michal Levin, Brian Luke, Falk Butter[2]
and Marion Scheibe[2]

iScience, April 28, 2023

---

[1] These authors contributed equally

[2] Corresponding authors

3.1. Summary

In this study, we investigate the phylogenetic diversity in recognizing and repairing three well-established DNA lesions. These lesions are 8-oxoguanine, an abasic site, and a ribonucleotide incorporated into DNA. They are primarily repaired by base excision repair (BER) and ribonucleotide excision repair (RER).

8-oxoG is formed through oxidative or alkylative damage. An abasic lesion can occur as an independent lesion or a BER intermediate. A uracil incorporated base is primarily caused by improper DNA replication and is often repaired by ribonucleotide excision repair. The three DNA damage lesions were incorporated into synthetic oligos, and affinity purifications were performed with protein extracts from 11 different species to compare against similar interactions with an undamaged synthetic oligo. The 11 species we investigated were *E. coli, B. subtilis, H. salinarum, T. brucei, T. thermophila, S. cerevisiae, S. pombe, C. elegans, H. sapiens, A. thaliana,* and *Z. mays*. Using quantitative mass spectrometry, we identified 337 binding proteins across these species. Of these proteins, 99 were previously characterized to be involved in DNA repair. We linked 44 previously unconnected proteins to DNA repair through orthology, network, and domain analysis. Together, this study presents an extensive resource for future study of the crosstalk and evolutionary conservation of DNA damage repair across all domains of life.

3.2. Zusammenfassung

In dieser Studie untersuchen wir die phylogenetische Vielfalt bei der Erkennung und Reparatur von drei bekannten DNA-Läsionen. Bei diesen Läsionen handelt es sich um 8-Oxoguanin, eine abasische Stelle und ein Ribonukleotid eingebaut in die DNA. Sie werden hauptsächlich durch Basen-Exzisionsreparatur (BER) und Ribonukleotid-Exzisionsreparatur (RER) repariert.

8-oxoG wird durch oxidative oder alkylative Schäden geformt. Eine abasische Läsion kann als unabhängige Läsion oder als BER-Zwischenprodukt auftreten. Eine Uracil inkorporierte Base wird in der Mehrheit der Fälle durch eine fehlerhafte DNA-Replikation verursacht und wird oft durch Ribonukleotid-Exzisionsreparatur repariert. Die drei DNA-Schadensläsionen wurden in synthetische Oligos eingebaut, und es wurden Aufreinigungen von Proteinextrakten aus 11 verschiedenen Spezies ausgeführt, um sie mit ähnlichen Aufreinigungen an einem unbeschädigten synthetischen Oligo zu vergleichen. Die 11 untersuchten Arten waren *E. coli, B. subtilis, H. salinarum, T. brucei, T. thermophila, S. cerevisiae, S. pombe, C. elegans, H. sapiens, A. thaliana* und *Z. mays*. Mit Hilfe der quantitativen Massenspektrometrie identifizierten wir 337 Bindungsproteine in diesen Arten. Von diesen Proteinen wurden

99 bereits zuvor als an der DNA-Reparatur beteiligt charakterisiert. Mit Hilfe von Orthologie-, Netzwerk- und Domänenanalysen konnten wir 44 Proteine mit der DNA-Reparatur in Verbindung bringen, die zuvor nicht in dieser Verbindung standen. Zusammengenommen bietet diese Studie eine umfangreiche Informationsquelle für die weitere Erforschung der Interkonnektivität und der evolutionären Konservierung der DNA-Schadensreparatur in allen Domänen des Lebens.

3.3. Statement of Contribution

Emily Nischwitz and I led this study with the support of Falk Butter and Marion Scheibe. Emily, Marion, Falk, and I contributed to the experimental design and its implementation. Emily and I led the data analysis and data visualization and provided in-depth data interpretation. Albert Fradera Sola, Mario Dejung, and Michal Levin conducted bioinformatic data analysis. Olga Vydzhak and Brian Luke provided experimental support and offered critical feedback on the manuscript. Emily and I wrote the initial draft and final version of the manuscript with support from Falk Butter and Marion Scheibe. All authors read and approved the final manuscript.

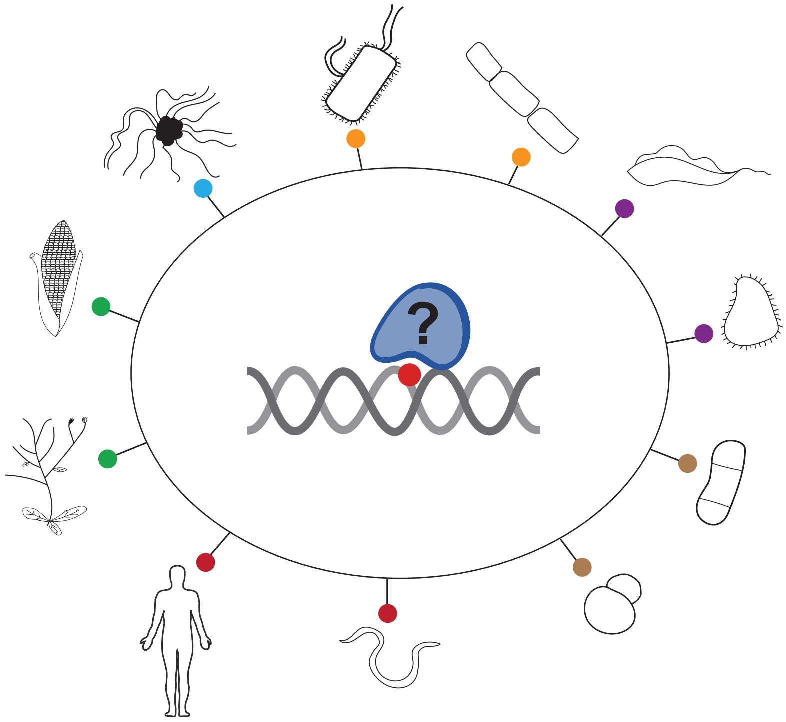


Figure 3.1.: Graphical abstract

3.4. Abstract

Genome maintenance is orchestrated by a highly regulated DNA damage response with specific DNA repair pathways. Here, we investigate the phylogenetic diversity in the recognition and repair of three well-established DNA lesions, primarily repaired by base excision repair (BER) and ribonucleotide excision repair (RER): 1) 8-oxoguanine, 2) abasic site, and 3) incorporated ribonucleotide in DNA in 11 species: *E. coli, B. subtilis, H. salinarum, T. brucei, T. thermophila, S. cerevisiae, S. pombe, C. elegans, H. sapiens, A. thaliana*, and *Z. mays*. Using quantitative mass spectrometry, we identified 337 binding proteins across these species. Of these proteins, 99 were previously characterized to be involved in DNA repair. Through orthology, network, and domain analysis, we linked 44 previously unconnected proteins to DNA repair. Our study presents a resource for future study of the crosstalk and evolutionary conservation of DNA damage repair across all domains of life.

3.5. Introduction

The stability of the genome is constantly threatened by both exogenous and endogenous mutagens. These genotoxic stressors can damage the architecture of the DNA, causing single-stranded breaks, double-stranded breaks, or chemical modifications to individual bases. These alterations may prevent the successful storage of genetic information and its transmission from one generation to the next and may potentially affect cellular fitness. To maintain genome integrity, there is a carefully orchestrated DNA damage response that functions to identify and subsequently repair damaged DNA [41]. Base excision repair (BER) and ribonucleotide excision repair (RER) represent two pathways that are responsible for resolving some of the most frequently encountered DNA lesions.

Base excision repair (BER) is primarily responsible for removing nonhelix-distorting lesions [56]. Some of the most prevalent lesions removed via the BER pathway are alkylated or oxidized bases and uracil misincorporation. The most frequent oxidative base lesion is 7,8-dihydro-8-oxoguanine (8-oxoG/8-oxoGuanine), which has been reported to occur up to 1,500 times per mammalian cell per day [122]. There is strong conservation of the BER pathway in archaea, protozoa, fungi, metazoa, and plantae [123–127]. In higher eukaryotes, the repair process generally begins with damage recognition by a DNA glycosylase, which then removes the damaged base and creates an apurinic/apyrimidinic site (AP site/abasic site). Abasic sites can be formed not only as BER intermediates but also endogenously. It has been estimated that there are up to 10,000 abasic sites arising per day in a single mammalian cell [128]. When abasic sites are generated, a 5‘-cleavage event is typically triggered by an AP endonuclease, resulting in a 3‘-hydroxyl and 5‘-deoxyribose phosphate. In single nucleotide repair, the 5‘-deoxyribose is removed primarily by DNA polymerase β and in some cases by DNA polymerase γ, and the resulting gap is then filled. If two or more nucleotides are repaired, the 3‘-hydroxyl is used for strand displacement synthesis via either DNA polymerase β or δ and ε, usually in conjunction with PCNA [129]. The previously cleaved 5‘-deoxyribose strand, often referred to as a 5‘-flap, is removed by FEN1. In both instances, the nick is sealed with ligase I or III [126]. Even more common than the generation of abasic sites is ribonucleotide misincorporation into double-stranded DNA during DNA replication. This occurs at a rate of one million sites per genome in mammalian cells, rendering it the most common endogenous DNA damage [130]. DNA polymerases have a highly conserved amino acid pocket that enforces sugar selectivity, referred to as a steric gate. While this steric gate helps polymerases prevent the entry of ribonucleotide triphosphates (rNTPs), there is still a large rate of ribonucleoside incorporation into DNA due to the imbalance of the nucleotide pools. For example, in *S. cerevisiae*, there are 30- to 200-fold more ribonucleotides than nucleotides [131]. The *S. cerevisiae* replicative polymerases α, δ and ε add approximately 1,900, 2,200, and 9,600 ribonucleotides per round of replication, respectively [132]. Across different organisms, there is a variable bias within the type of ribonucleotides incorporated into DNA. In this study, we selected rU, which in *S. cerevisiae* and *S. pombe* has comparable incorporation rates to rC and rA in nuclear genomes [133] but has thus far been studied less. When misincorporated ribonucleoside monophosphate (rNMP), also known as DNA-incorporated rNTPs, are integrated into DNA, they are most frequently repaired by RNase H2-mediated ribonucleotide excision repair (RER). RNase H2 recognizes the rNMP and incises at the 5‘-side of the ribonucleoside, leaving a 3‘-hydroxyl and 5‘-phosphate. As in BER, the 3‘-hydroxyl is used for strand displacement DNA synthesis via either DNA polymerase δ supported by PCNA or by DNA polymerase ε. The flap that is formed, beginning with the 5‘-phosphate, is removed by FEN1 or EXO1, after which the repaired strand is ligated [134, 135].

Previously, we used a phylointeractomic screen to study the evolution of proteins binding telomeres across the vertebrate lineage [136]. Here, we revisit this concept, investigating the phylogenetic diversity in the recognition and repair of three well-established DNA lesions, primarily repaired by BER or RER: 1) 8-oxoguanine, 2) an abasic site, and 3) incorporated ribonucleotide in DNA. Previous literature has highlighted strong conservation among fundamental proteins in both of these pathways [54]. However, only by studying these pathways across the tree of life can the conservation and divergence of these different repair machinery be elucidated. Including organisms across all three domains of life, this study recapitulates previous findings and reveals new candidate proteins with the potential to be involved in DNA damage repair. We provide a large resource dataset that can be used to propel new discoveries within these specific DNA repair pathways and model organisms.

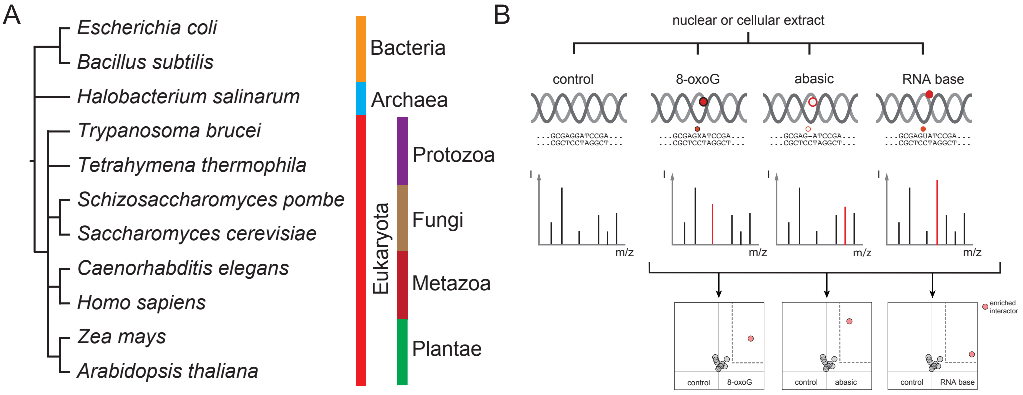


Figure 3.2.: Overview of screen for proteins interacting with DNA damage marks. A) Phylogenetic tree and overview of the eleven species included in this study. **B)** Experimental setup of the interactomics screen. Pull downs were performed for a control, and for an 8-oxoG, abasic, and RNA base lesion. Pull downs of the respective DNA damage lesion were compared to the common control to calculate enriched interaction partners passing a fold change threshold > 2 with Welch t-test p-value < 0.05 (dashed gray line).

3.6. Results and discussion

3.6.1. Wide-scale identification of proteins interacting with DNA damage marks

In this study, we selected 11 species from a broad phylogenetic range encompassing all three domains of life: *Escherichia coli* and *Bacillus sub-*

tilis (bacteria); *Halobacterium salinarum* (archaea); *Trypanosoma brucei* and *Tetrahymena thermophila* (eukaryota, protists); *Schizosaccharomyces pombe* and *Saccharomyces cerevisiae* (eukaryota, fungi); *Caenorhabditis elegans* and *Homo sapiens* (eukaryota, metazoa); *Zea mays* and *Arabidopsis thaliana* (eukaryota, plantae) (Figure 3.2A). We used oligonucleotides that were 79 bases long with three different site-specific synthesized DNA alterations, to which a biotinylated counterstrand was annealed (Table B.1). These double-stranded nucleic acid baits were immobilized on paramagnetic streptavidin beads and then incubated with protein lysates from the different species. Bound proteins were eluted from the beads and prepared for mass spectrometry measurements on a high-resolution orbitrap platform (Figure 3.2B). We quantified between 1,357 and 3,615 protein groups per species (Figure B.1A). The replicates of each single experiment showed good technical reproducibility covering similar range of LFQ protein intensities (Figure B.1B). Each of the three DNA lesions, 8-oxoG, abasic, and RNA, was compared to a common nonmodified oligonucleotide with four replicates per condition to allow the calculation of an average enrichment value (fold change) and a p-value for the reproducibility of the enrichment (Welch t-test) (Figure 3.2). Those proteins that had a $\log_2$ fold change > 2 and a p-value < 0.05 were considered enriched. Overall, we enriched 337 proteins across all lesions and species.

Table 3.1.: Overview of enriched interactors of each DNA damage lesion, per species (fold change > 2, Welch t-test p-value < 0.05). *Indicates orthology to known DNA damage repair factor, bold indicates previously known role in DNA damage repair, italics indicates no OrthoMCL orthology with the other 10 species included in the study.

Species	8-oxoG	abasic	RNA base
E. coli	**mutY, phrB**	fadJ, **nfo, phrB, polA**	**nfo, polA**
B. subtilis	**exoA, mutY, nfo,** *ydaT*, yhaZ, yisX, **yxlJ**	dinG*, *disA*, **exoA,** hupA, **mutM, nfo**, *parC*, parE, **priA,** topB*, *ydaT, ydeI, yfjM, yhaZ, yqxK,* **yxlJ**	dinG***, exoA, mutM, nfo,** topB*, *ydcG, ydeI, yfjM, yhaZ,* yisX, yusI, **yxlJ**
H. salinarum	cydB, *VNG_2525H*	**ogg**, *VNG_2498H*	**ogg**
T. brucei	GLE2, *Tb927.11.14995*, Tb927.7.1290, *Tb927.8.4240*, **Tb927.8.5510**	DRBD9, GLE2, PPL2, Tb927.10.6550, Tb927.3.5150, **Tb927.8.5510**, TOP2	DRBD9, NST4, *SET30*, *Tb927.2.6100*, **Tb927.6.1580**, **Tb927.8.5510**

Species	**8-oxoG**	**abasic**	**RNA base**
T. thermophila	PHR2*, TTHERM_000530789, *TTHERM_00145210*, TTHERM_00147470, TTHERM_00361370, TTHERM_00463150, *TTHERM_00614680*, TTHERM_00852850	APN2*, *PARP4*, PARP6, *PCP1*, *PCP2*, PHR2*	PARP6, *PCP1*, *TTHERM_00013250*
S. pombe	**myh1**	sac11, *SPAC3H8.08c*, top2	alp5, hmo1, *hpz1*, kin1, *mca1*, mlo3, *moc3*, nop12, **rfc1, rfc2, rfc3, rfc4, rfc5**, *SPAC3H8.08c*, *SPCC126.11c*
S. cerevisiae	**APN1**, ASG1, MYO4, *NUT1*, **PHR1**, POL5, *RNQ1*	**APN1**, ASG1, CMR1, **INO80**, MAK5, MYO4, *PDR1*, **PHR1**, POL5, **RFC1**, **RFC2**, **RFC3**, **RFC4**, **RFC5**, **RSC1**, *RSC58*, RSC6, **SNF2**, SWI6, TOP2	APL4, **APN1**, ASG1, CMR1, *HAP1*, **INO80**, MBP1, **MGM101**, MYO4, *OAF3*, *PDR1*, POL5, **RFC1**, **RFC2**, **RFC3**, **RFC4**, **RFC5**, **RSC1**, ***RSC30***, *RSC58*, RSC6, RSC9, SFH1, **SNF2**, **STH1**, SWI6, TOP2, *YPL245W*
C. elegans	*col-143*, **exo-3**, *hmg-5*	**apn-1**, *col-119*, *col-140*, *col-143*, *dpy-17*, **exo-3**, *F07A5.2*, *F07H5.8*, his-74, **K07C5.3**, obr-1, **parp-2**, *perm-2*, *phat-1*, *phat-2*, *T01E8.8*, *Y14H12B.2*, *Y37D8A.19*	*C27D8.2*, **exo-3**, *F07A5.2*, hmg-12, *T01E8.8*

Species	**8-oxoG**	**abasic**	**RNA base**
H. sapiens (HeLa)	**FANCI**, FERMT2, KPNA6, MYL12A, *NACC1*, PPWD1, RTRAF	**APTX**, ATP5MG, *BEND3*, **BLM**, BOP1, COQ6, DNAJC13, EXOSC3, GATAD2A, HNRNPF, HNRNPH2, **HPF1**, *ISG20L2*, **LIG3**, MRTO4, MYL12A, NAP1L1, NIP7, **NOP53**, **PARP1**, **POLB**, PPIG, *RIOX1*, RPL21, RPLP1, RPS26, *S100A8*, **UBE2N**, **XRCC1**	AHCTF1, CENPV, CHD2, *FXR1, KAT6A, MECP2*, **MPG**, *PCGF1, SAP130*, ZMYND11, *ZNF512B*
H. sapiens (HEK293)	MAX, **MUTYH**, **NTHL1**, SEPTIN11	**APTX**, CMSS1, **DDB1**, **DDB2**, DNAJC13, **LIG3**, NOC3L, **PARP2**, **PNKP**, **POLB**, **WRN**, **XPC**, **XRCC1**	AHCTF1, *APOBEC3C, BCOR, BCORL1*, BRPF1, CENPV, CHD1, CHD2, *CTCF*, GLYR1, *KAT6A*, KRI1, KRR1, **MPG**, *MSANTD7*, NIP7, NOC3L, ***NSD2***, NUP205, *PCGF1*, PITX2, RNF2, SUB1, **TRIP12**, *ZNF512B*

Species	8-oxoG	abasic	RNA base
Z. mays	B4FTT9*, *P06678*	A0A1D6F6W7*, A0A1D6JZF1*, A0A1D6K922, *A0A1D6LV91*, *A0A1D6NSE6*, A0A1D6P5Y9, *A0A804P6S3*, *B4FDA0*, B4FER3*, **B4FJC2**, **B4FQT5**, *B4FRR3*, *B4FWP8*, *B4FX14 B6SNB5*, , B6U4F1, *K7UTP1*, K7VBU4*	A0A1D6F4B6, *A0A1D6GRJ8*, A0A1D6HK01, A0A1D6HW59, *A0A1D6LV91*, A0A1D6LVY7, *A0A1D6MYU1*, *A0A1D6N2N7*, A0A1D6NSE6, A0A1D6QEP6, *A0A804MH07*, *A0A804MT25*, A0A804NRM4, *A0A804R2N8*, *B4FDA0*, B4FDW2, *B4FRR3*, *B4FX14*, B4G1M3, B4G1W8, *B6SNB5*, *B6UA70*, *C0P7N5*, C0P9C9, C4J4W6, C4J9R0, *C4JC33*, *K7UTP1*, Q6R9L4
A. thaliana	**ARP**, At1g09150, At4g32105, **At5g16990, CRYD, PHR1**, TRE1	At1g06260, At1g07080, **CRYD, MOC1, PHR1**	**ARP**, *HON5*, **MOC1**, TRE1

3.6.2. Functional enrichment and network analysis reveal novel insights into the enriched interactors

We classified the 337 enriched proteins as either ‘DNA repair’ or ‘non-DNA repair’ using the Gene Ontology term (GO:0006281) (Figure 3.3A). Of the 337 proteins, 99 were related to DNA repair, and 13 proteins were orthologs of DNA repair proteins (Figure 3.3A, Table 3.1, proteins with asterisks). Thus, our experimental conditions allowed for the identification of both known direct and indirect binders to the DNA damage lesions. Next, we used OrthoMCL to establish protein orthologies between species

[137]. The orthology group predictions are based on sequence similarity (reciprocal BLAST), normalization of interspecies differences, followed by Markov clustering. In total, the OrthoMCL database contains 70,388 ortholog groups across more than 55 species [138]. Proteins detected in our DNA damage interactome screen across eleven species belonged to 10,329 of these groups. We identified 82 proteins that possessed no OrthoMCL orthology with the other 10 species included within the study (Table 3.1, italicized protein names), four of which were repair proteins (Figure 3.3A). This suggests that in addition to finding conserved and previously established DNA repair factors, we also enriched for species specific DNA repair proteins.

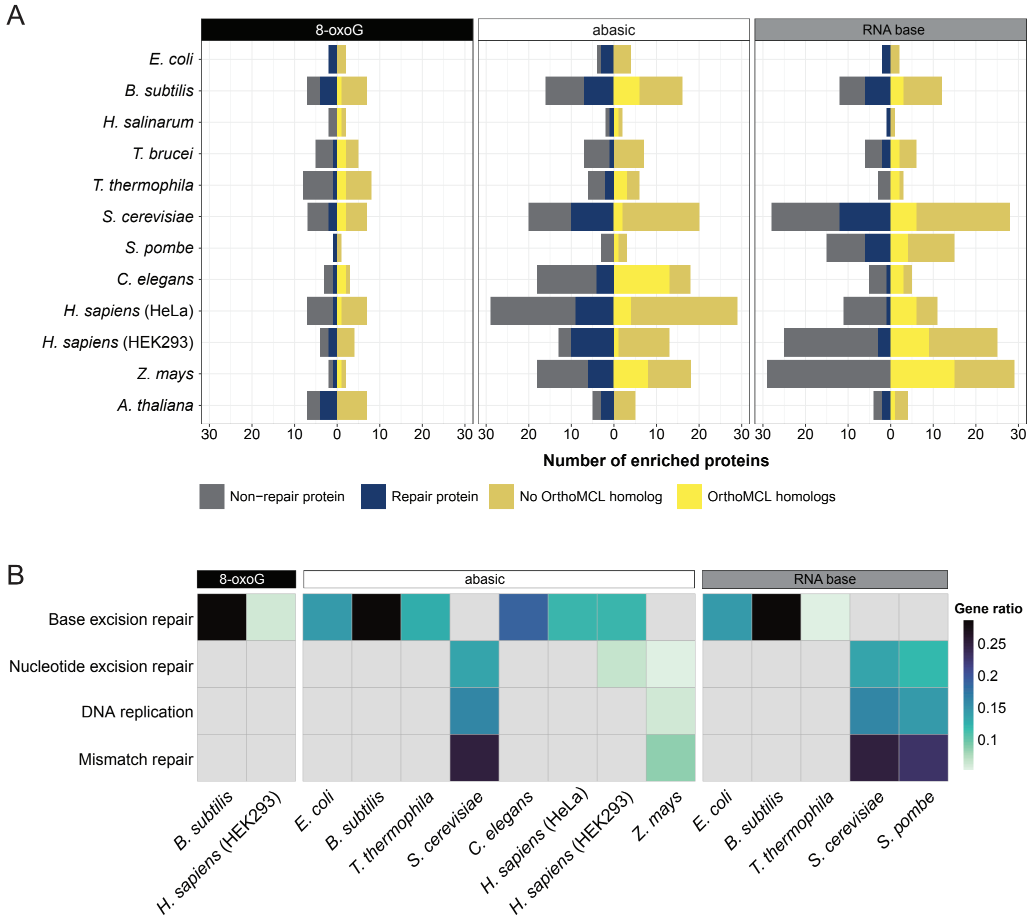


Figure 3.3.: Interactors of the DNA damage lesions per species. A) Number of proteins enriched at each lesion in each species highlighted for Gene Ontology annotation "DNA repair" (GO:0006281) (blue) and presence of orthologs in OrthoMCL (yellow). **B)** KEGG term overrepresentation of enriched proteins at each lesion across species. Conditions with no enriched KEGG terms are not shown, or presented in gray. 'Gene ratio' refers to genes in the dataset (enriched proteins at lesion) over genes in the background (whole genome).

To determine which functionalities were overrepresented, in addition to general 'DNA repair', within the interactors of 8-oxoG, abasic, and RNA lesions, we utilized both the Kyoto Encyclopedia for Genes and Genomes (KEGG) and GO [117, 139] (Table B.5). We found an overrepresentation of the KEGG term 'base excision repair' for all lesions. There was additional enrichment of ‘nucleotide excision repair’, ‘mismatch repair’, and ‘DNA replication’ (Figure 3.3B). Further interrogation of the enriched interactors of 8-oxoG showed enrichment of the GO biological processes ‘Base-excision repair’, ‘Base-excision repair ap site formation’, and ‘Photoreactive repair’ (Table B.5, Figure B.2). Within the interactors of the abasic lesion, there was enrichment of ‘DNA repair’ annotated proteins in multiple species, and there were seven more terms belonging to the parent term of ‘DNA repair’. Four DNA repair related GO terms (‘UV-damage excision repair’, ‘double-strand break repair’, ‘DNA repair’, and ‘base-excision repair’) were overrepresented among the interactors of the RNA base lesion.

To investigate the context of our enriched proteins at each of the lesions, we created lesion- and species-specific networks using previously established interactions and proteins included in the STRING database [140]. We found a total of 339 interactions across our enriched proteins and species (Figure B.3B). Of these enriched protein sets (3 lesions, 12 conditions, 36 total), ~61% had previously reported interactions among them. The largest number of known interactions (90) was found for the RNA lesion in *S. cerevisiae*. The 8-oxoG, abasic, and RNA enriched proteins exhibited 7, 187, and 151 previously established interactions, respectively. This indicates relative specificity of the 8-oxoG recognition and a more complex response resolving abasic and RNA lesions.

3.6.3. Interactors of 8-oxoG, abasic, and RNA lesions across phylogenetic branches

To establish the overlap of enriched orthologs across the included species at the 8-oxoG lesion, abasic lesion, and RNA base, we used orthology group predictions by OrthoMCL (Table B.4), only counting proteins that surpassed our enrichment threshold (Figure 3.4, Table B.10). Within the interactors of the 8-oxoG lesion, we identified protein families that were conserved in up to four species (Figure 3.4A-B, Figure B.4). The most conserved protein families were photolyases, MUTYH, and ExoIII-like and EndoIV-like AP endonucleases. Photolyases are critical repair proteins in bacteria, archaea, plantae, fungi, and animals. Despite their importance, they lost all DNA repair functionality in placental mammals [141]. The five enriched photolyases were grouped into two orthology groups (hsap_CRY1/OG6_100453 and atha_PHR1/OG6_104135). The divergence in these orthology groups indicates a specialization of the photolyases between species. It was unanticipated that photolyases would be enriched at 8-oxoG, as typically these proteins recognize and resolve pyrimidine dimers. However, with the enrichment traversing five different species, there is a strong argument to suggest that a base conversion or lesion intermediate interacts with these photolyases, and is resolved similarly across the tree of life. Other conserved interactors enriched at the 8-oxoG lesions were four members of the hsap_MUTYH group (OG6_102506). This enrichment was specific to 8-oxoG in *B.*

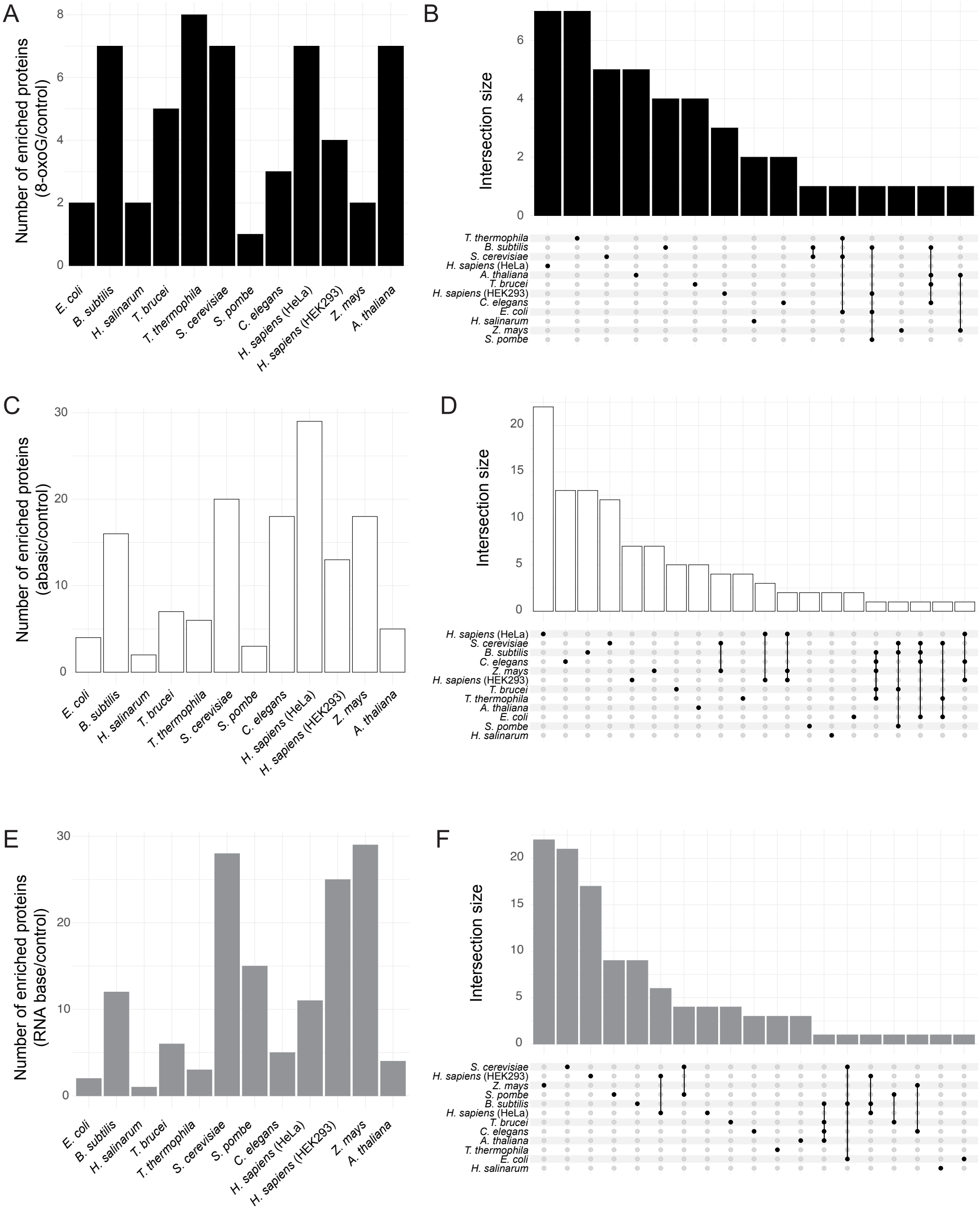


Figure 3.4.: Interactors of the different lesions across phylogenetic branches. A) Barplot of the total number of enriched proteins at 8-oxoG across species. **B)** UpSet plot showing overlap of enriched proteins at the 8-oxoG lesion for the different species based on assigned orthology groups via OrthoMCL. **C)** Barplot of the total number of enriched proteins at abasic lesions per species. **D)** UpSet plot showing overlap of enriched proteins at the abasic lesion for the different species based on assigned orthology groups via OrthoMCL. **E)** Bar plots of the total number of enriched proteins at the uracil RNA base per species. **F)** UpSet plot showing overlap of enriched proteins at the RNA base lesion for the different species based on assigned orthology groups via OrthoMCL.

subtilis, S. pombe, and *H. sapiens*, whereas mutY in *E. coli* was also bound to the abasic lesion. Although this is a well-characterized base excision repair glycosylase, it has thus far been shown primarily to bind 8-oxoG:A as opposed to the 8-oxoG:C used here. It is possible that the MUTYH orthologs generally bind to 8-oxoG due to their strong affinity, or they bind to a shared intermediate state of 8-oxoG:A and 8-oxoG:C [142].

At the abasic lesion, we found a higher degree of overlapping proteins with seven instances of three or more orthologs enriched in two or more species (hsap_DNAJC13, hsap_TOP2B, scer_PHR1, hsap_LIG3, scer_APN1, hsap_APTX1, and hsap_APEX1) (Figure 3.4C-D, Figure B.5, Table B.10). Two anticipated groups were the hsap_APEX1 (ExoIII-like) and scer_APN1 (EndoIV-like) AP endonucleases (OG6_101139 and OG6_104339, respectively), which are critical to the removal of abasic sites. Members of hsap_LIG3 and hsap_APTX1 are also critical to the BER pathway [56]. While LIG3 has been well studied in *H. sapiens*, the enriched ortholog in *C. elegans* has not been studied in the context of BER (K07C5.3, UniProt ID: Q19138). It is still unclear which ligase is involved in BER in *C. elegans* [143]. There were three homologs enriched in the hsap_APTX1 group, in HeLa and HEK cell lines and in *Z. mays*. APTX removes AMP from BER intermediates to form 3'-OH utilized by repair polymerases. A similar enrichment pattern was present in the hsap_DNAJC13 group. DNAJC13 is a heat shock protein that is critical to the heat stress response and has been associated with Parkinson's disease [144, 145]. DNAJC13 has not been studied in the context of BER.

Among the enriched proteins interacting with rU across species, members of the RFC complex were enriched in both *S. cerevisiae* and *S. pombe* (Table B.10). RFC is critical to the loading of PCNA, which is a well-established interactor of RNaseH2, an initiator of RER. Additionally, there was significant enrichment of the hsap_APEX group in *B. subtilis, T. brucei, C. elegans,* and *A. thaliana.* Additionally, proteins of the scer_APN group in *E. coli, B. subtilis,* and *S. cerevisiae* were enriched at rU. While the striking amount of enrichment of AP endonuclease was expected at the abasic and 8-oxoG lesions, this was unanticipated for the RNA lesion. There was also a noticeable enrichment of chromatin remodelers (Figure 3.4E-F, Figure B.6). In both, HeLa and HEK293 cells, PCGF1 and CHD1 were enriched. PCGF1 is part of the polycomb repressive complex 1, which is critical to epigenetic alterations repressing gene expression. Additionally, in HEK293 cells, two interactors of the polycomb repressive complex were enriched, BCOR and BCORL1 [146]. CHD1 is critical in the opening of chromatin around DNA damage lesions [147]. Within HEK293 cells, CHD2 and CTCF, which also mediate chromatin architecture in the presence of damage, were enriched [148, 149]. In *S. cerevisiae*, we observed enrichment of chromatin remodelers Ino80, Snf2, Swi6, and seven members of the Remodels the Structure of Chromatin (RSC) family (Sfh1, Sth1, Rsc1/6/9/30/58). All of the described chromatin remodelers have not yet been characterized in the misincorporated uracil from DNA but have been directly linked to the promotion of BER [150].

3.6.4. DNA damage interactors conserved across lesions

In this study, we observed potential DNA repair crosstalk through preferential binding of the same proteins at multiple lesions (Figure 4). We included two DNA damage lesions that are canonical substrates for base excision repair, 8-oxoG and abasic lesions, as well as a uracil ribonucleotide incorporated into DNA. As 8-oxoG is a common trigger for BER, and abasic lesions are a common BER intermediate, we anticipated finding joint interactors between these two lesions. Of the 55 8-oxoG interactors, 19 overlapped with the abasic interactors (Table B.11). Within this overlap, we unexpectedly found four instances of photolyases (Figure 3.5A-C, Table B.11). Additionally, in *B.* subtilis, ydaT was shared between 8-oxoG and abasic lesions (Figure 3.5D). This is an uncharacterized stress response protein that increases resistance to ethanol and low temperatures [151].

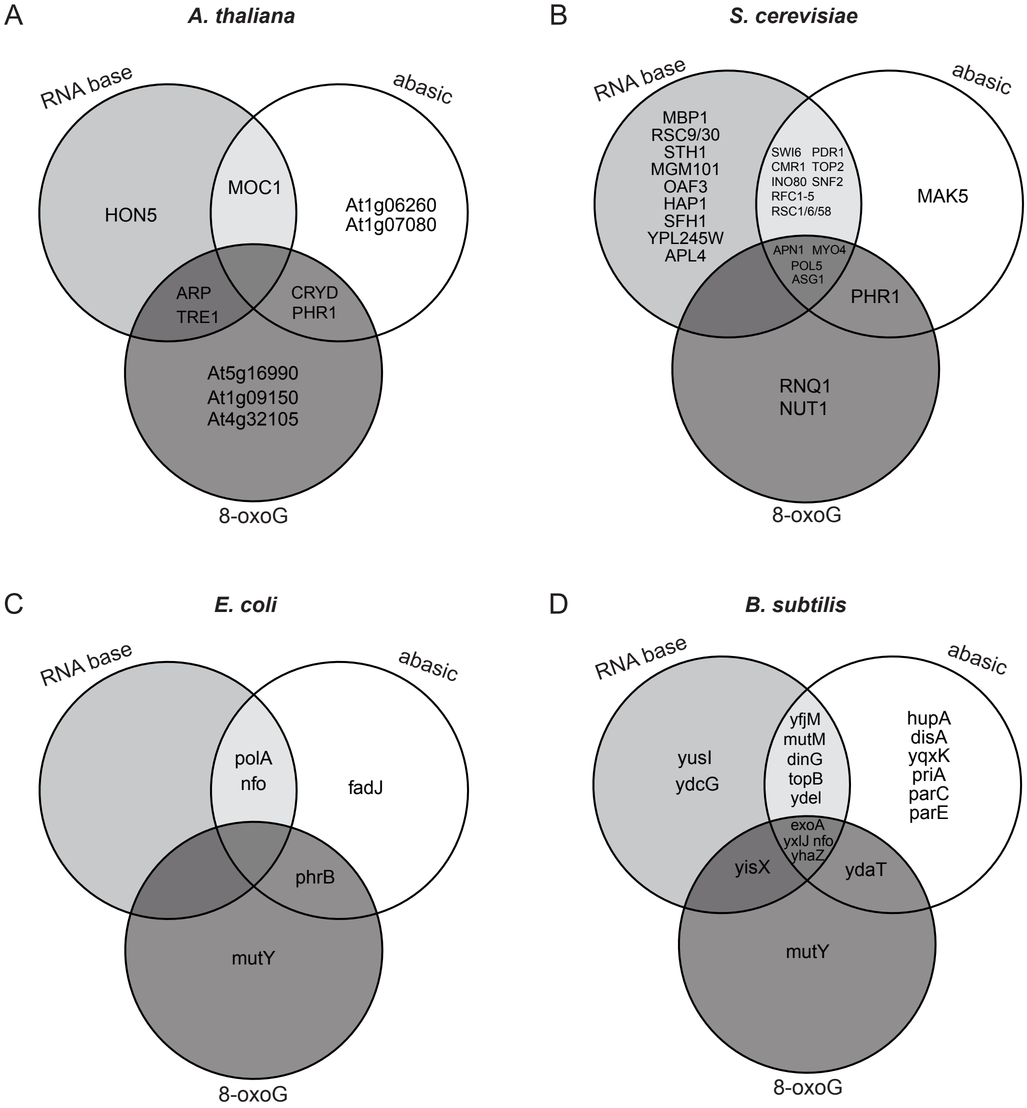


Figure 3.5.: Conserved interaction partners across the lesions. Venn diagrams showing the overlapping enriched proteins at the RNA base, abasic site, and 8-oxoG lesions for **A)** *A. thaliana,* **B)** *S. cerevisiae,* **C)** *E. coli* and **D)** *B. subtilis*. Overlap in other species is detailed in Table B.2.

There were 47 instances in which a protein was enriched both at the abasic site and rU. Such a large degree of overlap between the RNA base and abasic lesion was not initially expected. However, there has been evidence that abasic sites can occur within RNA, and are primarily resolved by APE1 and MPG [152]. In HEK and HeLa cells as well as *B. subtilis*, we enriched MPG and its ortholog yxlJ. Additionally, APE1 and APN1 orthologs were enriched in 6 of the 11 species. Thus, the removal of abasic sites from RNA may share mechanisms with uracil and abasic sites removal when incorporated into DNA. Our data also suggest that in *S. cerevisiae* the chromatin remodeling mechanisms that are needed to repair abasic sites are shared for the repair of rU (Ino80, Rsc1, Rsc6, Rsc58, Swi6 and Snf2) (Figure 3.5B). Chromatin state is a critical factor for the removal of both ribonucleotides and BER intermediates [135, 147, 150]. Beyond the overlaps between the enriched protein sets of two lesions, we also observed a notable overlap between all three lesions. In *B. subtilis, T. brucei, S. cerevisiae,* and *C. elegans*, AP endonuclease orthologs are enriched at all three lesions. In *B. subtilis*, we observed two uncharacterized glycosylases, yhaZ and yxlJ, at all three lesions (Figure 3.5D). Although ASG-1, POL5, and MYO4 are not characterized as DNA repair proteins, they were also found in all three lesions in *S. cerevisiae* (Figure 3.5B). Taken together, our screen reiterates a broader profile for DNA repair factors in the repair of 8-oxoG, abasic, and RNA lesions and a potential crosstalk between the different repair pathways (Figure 3.5, Table B.11).

3.6.5. Binding patterns by DNA repair factors are evolutionarily conserved across all domains of life

As the maintenance of genome stability is critical in each organism, many DNA damage factors are conserved in both sequence and functionality across species [54]. Across species and lesions, we enriched for classical BER-related proteins, including orthologs of the glycosylases MUTYH and MPG, deadenylase APTX, LIG3 and XRCC1, PCNA clamp loader RFC1-4, POLB, and the AP endonucleases APEX1 and Apn1 (Figure 3.6A). The APEX1/APE1 and Apn1 orthology groups represent the ExoIII-like AP exonucleases and EndoIV-like AP endonucleases, respectively. These groups of conserved AP endonucleases have been studied at length due to their evolutionary history [56, 153, 154]. Using a maximum likelihood phylogenetic tree including all AP endonucleases across the 11 species we demonstrate the potential enrichment differences between the two groups (Figure B.7). For both groups of endonucleases, we found a twofold or greater binding to 8-oxoG and abasic lesions, in eight of the eleven species. Additionally, more unexpectedly spanning both groups was the enrichment of AP endonucleases at the RNA base in six of the eleven species. While AP endonucleases have been well-characterized within BER, thus far they have been shown to play a more minor role in RER [135]. It is possible that both types of AP endonucleases play a larger role than originally anticipated.

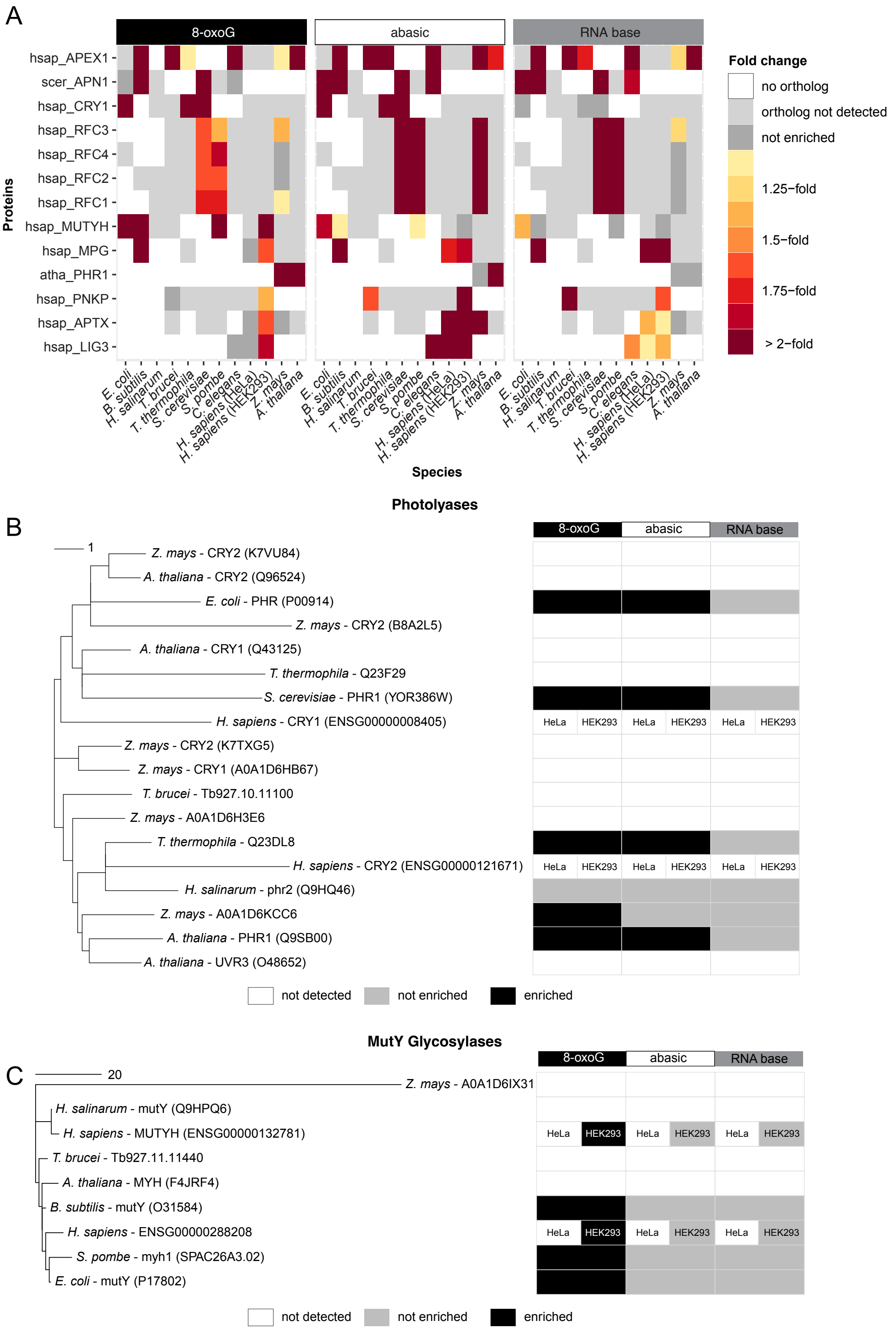


Figure 3.6.: Conservation of DNA repair orthologs across the tree of life. A) Heatmap representing enrichment levels of OrthoMCL orthology groups with GO annotation ‘DNA repair’ (GO:0006281) with two or more enriched proteins across eleven species and 8-oxoG (black), abasic (white) and RNA base (gray) lesions. The color scale represents the fold change in comparison to control samples . Abbreviations: hsap, *Homo sapiens*; scer, *S. cerevisiae*; cele, *C. elegans*; atha, *A. thaliana*; spom, *S. pombe*. **B)** Neighbor-joining phylogenetic tree of the photolyase gene family including information on detection and enrichment (fold change > 2, Welch t-test p-value < 0.05) for the different lesions. White boxes represent proteins that were not detected in the respective experiment. The scale bar in the plots indicates the number of amino acid substitutions per site. **C)** Neighbor-joining phylogenetic tree of the MUTY glycosylase gene family. Same as B.

Two additional protein families that had highly conserved enrichment patterns were the photolyases (scer_PHR1 and atha_PHR1) and MUTYH-related glycosylases (hsap_MUTYH). Despite both being DNA repair proteins, the binding of these proteins was unexpected in this particular context. Photolyases are known to have specific repair activity for cyclobutane pyrimidine dimers and 6-4 pyrimidine-pyrimidone photoproducts caused by UV light [155]. However, the *S. cerevisiae* PHR1 orthologs in *E. coli, T. thermophila,* and *S. cerevisiae* were significantly enriched at both the 8-oxoG and abasic lesions (Figure 3.6B). Both orthologs in the atha_PHR1 group were also significantly enriched at the 8-oxoG lesion. There was enrichment at the abasic lesion in *A. thaliana*, and for *Z. mays,* it was 1.9-fold, just below our threshold. As per the orthology groups, the maximum likelihood phylogenetic tree showed a clear divergence of the plant photolyases, despite their similar *in vitro* binding characteristics. We did not observe enrichment of any orthologs of PHR1 (atha_PHR1 and scer_PHR1) at the RNA lesion, which extended across all species regardless of evolutionary relation (Figure 3.6B).

MutY-related glycosylases are well characterized in the removal of 8-oxoG:A, but there are few studies showing their binding to 8-oxoG:C, which was used in this study. In an *in vitro* setting when the diffusion rate was measured, MUTYH orthologs would linger much longer at 8-oxoG:A but also have moderate stalling at 8-oxoG:C [156]. MUTYH orthologs were found to bind specifically to 8-oxoG in *E. coli, B. subtilis, S. pombe,* and *H. sapiens* (Figure 3.6C). There were no instances of detection of a MUTYH ortholog without enrichment at 8-oxoG, indicating highly specific binding which was independent of the evolutionary relation of the protein sequences. MUTYH has recently been suggested to facilitate the overall DNA damage response as a scaffolding protein [157]. While this function has been primarily explored within vertebrates, our findings indicate that its multiple functionalities might have emerged far earlier in evolution than originally estimated (Figure 3.6C).

3.6.6. Identification of uncharacterized DNA repair proteins across multiple species

In addition to the enriched known DNA repair proteins, one-third of the enriched proteins were previously not associated with the 'DNA repair' GO term (GO:0006281). We found enrichment of 35, 85 and 105 non-DNA repair classified proteins at the 8-oxoG lesion, abasic lesion, and RNA base, respectively. To investigate these proteins further, we created species-specific networks with all three DNA damage lesions using the STRING database (Figure B.8, Table B.6, Table B.7, Table B.8, Table B.9). Within these networks we marked proteins categorized as repair (triangle) and non-DNA repair proteins (circle), and indicated at which lesion they were enriched. Here, we highlighted the *S. cerevisiae*, *C. elegans*, and *T.*

thermophila networks (Figure 3.7A). Within *S. cerevisiae*, all enriched proteins contained in STRING interacted and formed one large network (Figure 3.7A). Included are five chromatin remodelers (RSC6, RSC9, RSC58, SFH1, and SWI6) that, although not characterized as DNA repair proteins, had a prominent number of interactions with both repair and non-DNA repair proteins. Within the RSC family, RSC1, RSC30, and STH1 have been classified as DNA repair proteins and are specifically linked to base excision repair [150, 158]. This suggests that the other RSC proteins likely play a role in chromatin remodeling surrounding DNA repair. Additionally, the non-DNA repair protein CMR1 had 11 interaction partners, five of which were 'DNA repair' proteins. Notably, although not included in the 'DNA repair' GO term, CMR1 has been shown to be needed to resolve genotoxic stress and has a preference binding UV lesions *in vitro* [159, 160]. For the *C. elegans* interactors, we identified three different subnetworks. Within one subnetwork, the 'DNA repair' proteins parp-2, exo-3, and apn-1 interacted with 3, 4, and 1 non-DNA repair proteins, respectively. All three proteins were mutually linked to hmg-5. Hmg-5 was studied in a *C. elegans* Parkinson's disease model, and together with nth-1, BER glycosylase, and other associated proteins reduce mitochondrial stress and oxidative damage [161]. Within the *T. thermophila* network, there were mutual interactions between APN2, identified as a DNA repair protein based on its orthology to the *S. cerevisiae* AP endonuclease, and four different PARP-related proteins, as well as TTHERM_00463150, which has not been characterized. This indicates that APN2 might orchestrate the recruitment of PARP-related proteins, or that PARP-related proteins are needed for APN2 to access DNA.

We evaluated the Pfam domains found among the enriched non-DNA repair proteins to elucidate more of their potential functionalities [162]. The two most frequently identified domains were DNA-binding domains: 1) 'protein of unknown function, DUF573' (corresponding to Interpro protein family 'GLABROUS1 enhancer-binding protein family'), which is often part of proteins associated with plant stress response, and 2) 'Fungal Zn(2)-Cys(6)' often involved in growth and metabolism [163, 164]. We assigned each Pfam domain into one of 15 categories to summarize its primary function (Table B.12). At all three lesions, the majority of domains were related to DNA repair and DNA binding (Figure 3.7B). Thus, despite the lack of categorization as DNA repair genes under the GO term 'DNA repair', there was a clear link to DNA repair functionality within these proteins. For example, we identified the 'Poly(ADP-ribose) polymerase' and 'DNA-Ligase Zn-finger region' in four different proteins. These included hpz1 in *S. pombe* and Tb927.10.6550 in *T. brucei*, which both belong to the same orthology group. The other two proteins are PARP-related proteins in *T. thermophila*, PCP1 and PCP2. We also detected the 'PARP-associated WGR domain' and a 'PARP catalytic domain' in PARP4 and PARP6 in *T. thermophila*.

Furthermore, we examined the conservation of enrichment of non-DNA repair proteins across species, to further support a role in DNA damage repair

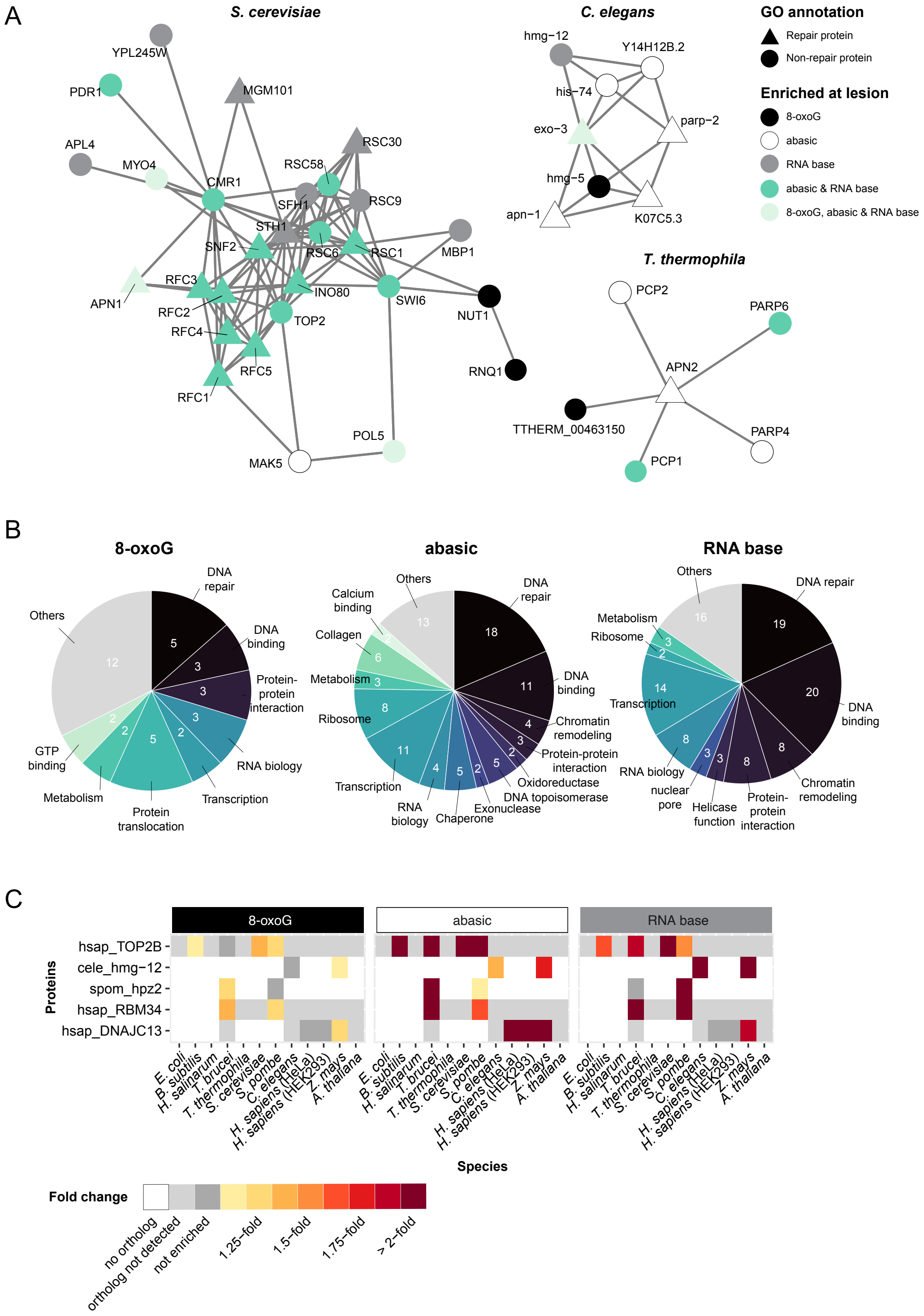


Figure 3.7.: Network, domain, and phylogenetic analyses implicate novel proteins in DNA repair. A) Networks of enriched proteins across lesions for *S. cerevisiae, C. elegans,* and *T. thermophila.* Interactions as established in the STRING database. **B)** Classification of non-DNA repair proteins based on Pfam domain annotation. The total number of proteins classified at 8-oxoG was 29, at abasic 75, and at the RNA base 74. **C)** Heatmap representing enrichment levels of OrthoMCL orthology groups without GO annotation 'DNA repair' (GO:0006281) with two or more enriched proteins across all eleven species and 8-oxoG (black), abasic (white) and RNA base (gray) lesions. The color scale represents the fold change in comparison to control samples. Abbreviations: hsap, *Homo sapiens*; cele, *C. elegans*; spom, *S. pombe.*

and recognition of lesions. We found at least five instances in which non-DNA repair genes were enriched in multiple species (Figure 3.7C). Intriguingly, some of these proteins were also identified within our domain analysis. For example, both enriched proteins in the spom_hpz2 orthology group in *T. brucei* and *S. pombe* contained a PARP-related domain. Furthermore, there was specific enrichment of the *T. brucei* ortholog (Tb927.10.6550) at the abasic lesion and the *S. pombe* ortholog (hpz1) at the RNA base. Additionally, all three proteins enriched within the hsap_DNAJC13 orthology group, have Pfam ‘DnaJ domains’. These proteins preferentially bound to the abasic lesion in both HEK293 and HeLa cell lines as well as the two paralogs in *Z. mays* (UniProt: A0A1D6K922 and A0A1D6P5Y9). The conservation of enrichment across various species in both cases suggests a very likely role in DNA repair.

Through the use of network, domain, and phylogenetics analysis, we have identified proteins that, despite not being classified as DNA repair proteins, likely have a role in the DNA damage response.

3.6.7. Conclusions and limitations of study

Performing a mass spectrometry based phylointeractomics screen across 11 species, we compared the binding capabilities of three well-established DNA damage lesions, an 8-oxoG modification, abasic site, and ribonucleotide base incorporation. We enriched 337 proteins across all lesions and selected species (Table 3.1). Of these 337 proteins, 99 were related to DNA repair, which in a proteome-wide generic screen with thousands of possible proteins strongly indicates the specificity of the experiment. Supporting the specificity even further, DNA repair-related KEGG and GO terms were overrepresented in the enriched group of proteins. Through phylogenetic analysis, we established that the enrichment of particular DNA damage proteins extends through many species. In some cases, we do not identify or enrich all expected interaction partners, which can be caused by a variety of reasons. For instance, preparation from a large range of different tissues and cellular material can lead to variation in the pool of proteins available for measurement. The lack of *in vivo* conditions, such as pH, salt concentrations, temperature, post-translational modifications and many other cellular conditions, affects DNA-protein interactions. As we did not perform cross-linking mass spectrometry, it is possible that some more transient interactions were not maintained. Furthermore, it is important to highlight the likely creation of repair intermediates in the *in vitro* pull down assays. The ability to repair 8-oxoG, abasic sites, and uracil residues *in vitro* has been previously demonstrated with human cell extract [165, 166]. However, we did find that unrepaired lesions existed in our experiment, for example, 11 out of the 24 canonical DNA repair-related proteins were uniquely enriched at the 8-oxoG lesion, suggesting that unrepaired lesions persisted.

In addition to DNA repair genes, we identified two other intriguing groups of interactors in our screen. Namely, we detected an enrichment of 82 species-specific proteins as well as proteins that have not been implicated previously in DNA repair. This group of proteins presents an avenue to study potentially unique aspects of repair or damage response in their corresponding model organism. To elucidate functionality and connection to DNA damage repair for originally non-DNA repair proteins, we utilized network, domain, and phylogenetics analysis. With this we indicated an additional 44 proteins to potentially play a role in the DNA damage response.

Our study systematically evaluates *in vitro* binding partners in both BER lesions and an RNA lesion in eleven model species across the tree of life. We recapitulate previous findings and nominate putative unknown candidates to be involved in the resolution of these lesions. Through the use of network, domain, and phylogenetics analysis, we identified a subset of non-DNA repair classified proteins to likely be involved in DNA repair. Overall, this study opens avenues for further investigation of newly identified candidates to explore key factors in the crosstalk between BER and RER DNA damage pathways.

3.7. Acknowledgements

We are indebted to Franziska Roth und Jasmin Cartano for technical support. We thank Varvara Verkhova for cultivation of *B. subtilis,* Sabrina Dietz for growing *H. salinarum* and *C. elegans,* and Markus Engstler for providing *T. brucei* cell extract. Assistance by the IMB Media Lab and Proteomics Core Facility is gratefully acknowledged.

Funding: This project was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) [Project-ID 393547839-SFB 1361]; Joachim Herz Stiftung Add-on fellowship to V.A.C.S.

3.7.1. Author contributions

Conceptualization, F.B., and M.S.; investigation, E.N., V.A.C.S., M.S.; formal analysis, E.N., V.A.C.S., A.F.-S., M.D., M.L., F.B.; visualization, E.N., V.A.C.S., A.F.-S, M.D., M.L., F.B., and M.S.; writing–original draft, E.N., V.A.C.S., F.B., and M.S.; writing–review & editing, all authors contributed; supervision: F.B. and M.S.; project administration: F.B.; funding acquisition: F.B.

3.7.2. Declaration of interests

The authors declare no competing interests.

3.8. STAR Methods

3.8.1. Resource availability

Lead contact
Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact Falk Butter (f.butter@imb.de).

Materials availability
This study did not generate new unique reagents.

Data and Code Availability

- The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD036040.
- All original code has been deposited into the GitHub repository used for the proteomics and STRING database analysis, which is available at: (https://github.com/mariodejung/DNAdamage_phylointeractome and https://github.com/AFraderaSola/DNADamage_Phylointeracome).
- Any additional information required to reanalyze the data reported in this paper is available from the lead contacts upon request.

3.8.2. Experimental model and subject details

All cultivation and growth conditions as relevant for *B. subtilis* (DSM10), E. coli (DH5α), *H. salinarum* (NRC-1), *S. cerevisiae* (BY4742α), *S. pombe* (pp265), *T. thermophila* (SB210), *C. elegans* (N2), *T. brucei* (Lister 427), *H. sapiens* cell lines (HeLa and HEK293), *A. thaliana* and *Z. mays* are included within the 'Method Details' section.

3.8.3. Method details

3.8.3.1. Cultivation and extract preparation

Bacteria: *B. subtilis (DSM10)* and *E. coli* (DH5α) were grown at 37 °C in LB medium (IMB media lab) and harvested at OD_{600}=0.7. Cell pellets were resuspended in PBB buffer (150 mM NaCl, 50 mM Tris/HCl pH 8.0, 0.5% Igepal CA-630, 10 mM $MgCl_2$, Pierce protease inhibitor EDTA free) and sonicated on with a sonifier 450 (Branson) 3 times for 45 s (cycle=70%, output level 2) with 2-minute breaks. The lysate was centrifuged at 4 °C for 15 min at 20,200 x g). The supernatant was supplemented with 10% (f.c.) glycerol (Roth), shock-frozen in liquid nitrogen and stored at -80 °C.

Archaea: *H. salinarum* strain NRC-1 was cultivated in Complex Media (4.3 M NaCl, 81 mM $MgSO_4$ x 7 H_2O, 27 mM KCl, 12 mM sodium citrate, 1% w/v oxoid peptone) at 37 °C and in light for ~52h/2.5 days and harvested at OD_{600}=0.5. The cells were pelleted at 3,500 x g for 30 min at 4 °C and washed twice in Basic Salt Solution (4.3 M NaCl, 81 mM $MgSO_4$ x 7 H_2O, 27 mM KCl, 12 mM sodium citrate) to remove the medium. After washing the cells were resuspended in 10 ml Lysis Buffer (2.1M NaCl, 50 mM Tris/HCl pH 7.5, 10 mM $MgCl_2$) and sonicated on ice using a Branson 450 sonifier 6 times for 30 s (cycle=50%, output level 2) with 1 min breaks. The sonicated lysate was cleared by centrifugation at 3,500 x g for 30 min at 4 °C and supplemented with 10% (f.c.) glycerol (Sigma) before shock-freezing in liquid nitrogen and stored at -80 °C.

Yeast: *S. cerevisiae* (BY4742α) was grown in YP medium containing 20% glucose (IMB media lab) at 37 °C until OD_{600}=0.5 and harvested by centrifugation at 20,200 x g. *S. pombe (pp265)* was cultivated in YES media at 32 °C until OD_{600}=1.0 and harvested by centrifugation. For both species, cells were lysed using 0.5 mm zirconia glass beads (Roth) in lysis buffer (100 mM NaCl, 50 mM Tris-HCl pH 7.5, 10 mM $MgCl_2$, 0.01% Igepal CA-630, 1x PMSF) at 4 °C with 3 cycles alternating between 30 s milling and 30 s cooling using a FastPrep-24 system (MP Biomedicals). The supernatant was transferred to a new tube, shock-frozen in liquid nitrogen and stored at -80 °C.

T. thermophila: A mid-log SB210 culture of $3x10^7$ cells was grown in 2% proteose peptone (BD Biosciences), 0.2% yeast extract (BD Biosciences), 12 µM ferric chloride, and 1x Penicillin/Streptomycin/Funizone (Hyclone) at 30 °C at 100-120 rotations per minute. Cells were pelleted at 1,500 x g for 3 minutes and washed in 10 mM Tris-HCl pH 7.4. Cells were transferred to a 1.5 ml centrifuge tube and centrifuged at 1,500 x g for 2 min and the supernatant was removed. Cells were resuspended in 1.2 ml lysis buffer (350 mM NaCl, 40 mM Hepes pH 7.5, 1% Triton X-100, 10% glycerol, freshly added 1 mM DTT, and 1x complete protease inhibitors [Roche]) and approximately 200 µl zirconia glass beads (Roth) were added and vortexed for 3 minutes at 4 °C. The tube was centrifuged at ≥16,000 x g at 4 °C for 5 min, and the supernatant was transferred to a new tube. The sample was centrifuged at ≥16,000 x g at 4 °C for 15 minutes. The supernatant was transferred to a new tube, shock-frozen in liquid nitrogen and stored at -80 °C.

C. elegans: Nuclear extraction was performed with N2 gravid adult worms as in (de Albuquerque et al. 2014). Worms were synchronized and grown on egg plates until they reached the gravid adult stage. Then, worms were washed with M9 buffer 3 times, pelleted, and frozen into pellets in Extraction Buffer (40 mM NaCl, 20 mM MOPS pH 7.5, 90 mM KCl, 2 mM EDTA, 0.5 mM EGTA, 10% glycerol, 2 mM DTT, and 1x complete protease inhibitors, Roche). Pellets were ground into a fine powder with a mortar and pestle. The powder was transferred to a precooled glass douncer (Kimble), and the samples were ruptured with piston B over 30 strokes. The debris was cleared twice at 200 x g for 5 minutes at 4 °C. The nuclear pellet was isolated by centrifuging at 2,000 x g for 5 minutes at 4 °C. This pellet was washed in extraction buffer twice. The nuclear pellet was resuspended in 200 µL Buffer C+ (420 mM NaCl, 20 mM Hepes/KOH pH 7.9, 2 mM $MgCl_2$, 0.2 mM EDTA, 20% glycerol, and freshly added 0.1% Igepal CA-630, 0.5 mM DTT, 1x complete protease inhibitors [Roche]). The lysate was centrifuged at 4 °C for 15 min at 20,200 x g. The supernatant was supplemented with 10% (f.c.) glycerol (Roth), shock-frozen in liquid nitrogen and stored at -80 °C.

Plants: *Z. mays and A. thaliana (Columbia)* were ground, frozen in liquid nitrogen and transferred to a liquid nitrogen precooled 50 ml steel container for cryomilling with an MM400 (Retsch) at 30 Hz for 4 min. *Z. mays* powder was resuspended in 35 ml PBB buffer (150 mM NaCl, 50 mM Tris/HCl pH 8.0, 0.5% IGEPAL-CA630, 10 mM $MgCl_2$, Pierce protease inhibitor EDTA free) and incubated on ice for 10 min. For *A. thaliana*, powder was resuspended in 30 ml Buffer A (10 mM Hepes KOH pH 7.9, 1.5 mM $MgCl_{2,}$ 10 mM KCl), incubated on ice for 10 min, and subsequently dounced with 40 strokes in a glass douncer using pestle B (Kimble). After centrifugation at 3,640 x g at 4 °C, the pellet was washed with 1x DPBS (Gibco), centrifuged again and incubated in 4-6 ml Buffer C+ (420 mM NaCl, 20 mM Hepes/KOH pH 7.9, 2 mM $MgCl_2$, 0.2 mM EDTA, 20% glycerol, and freshly added 0.1% Igepal CA-630, 0.5 mM DTT, 1x complete protease inhibitors [Roche]) for 1 hour at 4 °C on a rotation wheel. Cell fragments were removed by centrifugation at 20,200 x g and 4 °C for 60 min. The supernatant was shock-frozen in liquid nitrogen and stored at -80 °C.

Cultured cells: HeLa and HEK293 cells were grown in DMEM (Gibco) with 10% FBS (Gibco) and PennStrep (Sigma) at 37 °C with 75% relative humidity and 5% CO_2 in an incubator (Thermo). Cells were harvested, washed in 1x DPBS (Gibco), resuspended in buffer A (10 mM Hepes KOH pH 7.9, 1.5 mM $MgCl_{2,}$ 10 mM KCl) and incubated on ice for 10 min. Cells were centrifuged at 500 x g for 5 min and resuspended in Buffer A+ (10 mM Hepes KOH pH 7.9, 1.5 mM $MgCl_{2,}$ 10 mM KCl, Roche protease inhibitor EDTA free, 0.1% Igepal CA-630, 0.5 mM DTT) and then dounced with 40 strokes in a glass douncer using pestle B (Kimble). Cells were centrifuged at 2,640 x g for 15 min and the cell pellet was washed with 1x DPBS (Gibco) prior to incubation of the pellet in buffer C+ (420 mM NaCl, 20 mM Hepes/KOH pH 7.9, 2 mM $MgCl_2$, 0.2 mM EDTA, 20% glycerol, and freshly added 0.1% Igepal CA-630, 0.5 mM DTT, 1x complete protease inhibitors [Roche]) for 1 hour at 4 °C on a rotation wheel. Cell fragments were removed by centrifugation at 20,200 x g and 4 °C for 60 min. Supernatant was shock-frozen in liquid nitrogen and stored at -80 °C.

3.8.3.2. DNA pull-down experiments

Chemically synthesized oligonucleotides (Table B.1) were ordered HPLC-purified from BioSynthesis (Lewisville) and Metabion (Planegg).

For pull-down 1 nmol of single-stranded DNA lesion (or nondamaged control) oligonucleotide was annealed with 1 nmol of 5'-biotinylated counterstrand with annealing buffer (20 mM Tris-HCl pH 8.0, 10 mM $MgCl_2$, 100 mM KCl) by first heating to 85 °C for 5 min and slowly cooling to RT. The double-stranded oligonucleotides were immobilized on 250 µg streptavidin Dynabeads C1 (Thermo) and incubated with different amounts of protein extract ranging from 200-1,000 µg (200 µg: *C. elegans*, *Z. mays* and *A. thaliana*; 400 µg: HEK293 and HeLa; 500 µg: *H. salinarum*, *T. thermophila*; 800 µg: *S. cerevisiae* and 1,000 µg: *B. subtilis*, *E. coli*, *S. pombe* and *T. brucei*) in 1x PBB buffer (150 mM NaCl, 50 mM Tris-HCl pH 8.0, 0.5% Igepal CA-630, 5 mM $MgCl_2$ and 1x protease inhibitor cocktail [Roche]) rotating at 4 °C for 90 min. Protein concentrations were determined using Protein Assay Dye Reagent (Bio-Rad). All samples were prepared in quadruplicate. After incubation, unbound proteins were removed by 3 washes with PBB buffer. The Dynabeads were ultimately resuspended in 25 µl 1x LDS (Thermo) containing 100 mM DTT (Sigma) and heated to 70 °C for 10 min.

3.8.3.3. Mass spectrometry sample preparation

LDS supernatant was loaded on a 4-10% NuPage NOVEX PAGE gel (Thermo) and run for 10 min at 180 V. Samples were processed as previously described (Scherer et al. 2020). In short, gel pieces were cut, destained with 50% EtOH/50 mM ammonium bicarbonate (ABC), dehydrated with acetonitrile (VWR), reduced with 10 mM DTT (Sigma), alkylated using iodoacetamide (Sigma) and subsequently again dehydrated with acetonitrile (VWR) and digested with 1 µg of MS-grade trypsin (Sigma) at 37 °C overnight. The peptides were eluted from the gel pieces, loaded onto a StageTip [108] and stored at 4 °C until measurement.

3.8.3.4. Mass spectrometry measurement

Peptides were eluted from the StageTips using 80% acetonitrile/0.1% formic acid and concentrated prior to loading either on an uHPLC nLC-1000 system coupled to a Q Exactive Plus mass spectrometer (Thermo) or an uHPLC nLC-1200 system coupled to an Exploris 480 mass spectrometer (Thermo). The peptides were loaded on a 20 cm (Q Exactive Plus) or 50 cm (Exploris 480) column (75 µm inner diameter) in-house packed with Reprosil C18 (Dr. Maisch GmbH) and eluted with a 73- or 88-min optimized gradient increasing from 2% to 40% mixture of 80% acetonitrile/0.1% formic acid at a flow rate of 225 nl/min or 250 nl/min. The Q Exactive Plus was operated in positive ion mode with a data-dependent acquisition strategy of one MS full scan (scan range 300 - 1,650 m/z; 70,000 resolution; AGC target 3e6; max IT 20 ms) and up to ten MS/MS scans (17,500 resolution; AGC target 1e5, max IT 120 ms; isolation window 1.8 m/z) with peptide match preferred using HCD fragmentation. The Exploris 480 was operated in positive ion mode with a data-dependent acquisition strategy of one MS full scan (scan range 300 - 1,650 m/z; 60,000 resolution; normalized AGC target 300%; max IT 28 ms) and up to twenty MS/MS scans (15,000 resolution; AGC target 100%, max IT 40 ms; isolation window 1.4 m/z) with peptide match preferred using HCD fragmentation.

3.8.3.5. Mass spectrometry data analysis

MaxQuant (Version 1.6.5.0) was used to search and quantify the raw mass spectrometry files, for each species individually. Individual protein databases used as search space for MaxQuant can be found in Table B.3. Oxidation and acetylation were set as variable modifications, and carbamidomethylation was set as a fixed modification. LFQ was used to calculate and normalize intensities, without activating fast LFQ. The minimum ratio count used was 2. Match between runs was used to match within each lesion (control, abasic, 8-oxoG, RNA base), with a match time window of 0.7 min, match ion mobility window of 0.05, alignment time window of 20 min, and alignment ion mobility of 1. Matching of unidentified features was deactivated. For protein quantification we used a label minimum ratio count of 2, and unique + razor peptides for quantification.

3.8.3.6. Bioinformatics analysis and statistical analysis

MaxQuant proteinGroup results files of all species were combined into a single file, with a column “species” indicating the individual species and cell type (Table B.4). The complete dataset was filtered by removing reverse database binders, potential contaminants or proteins identified only on a modification site. Additionally, all protein groups with fewer than 2 peptides (1 unique) were filtered out. Missing LFQ values were treated as if they were below the detection limit of the mass spectrometer. Imputation was performed for each replicate of a condition individually from a beta distribution, within a range of the 0.2 and 2.5 percentile of measured intensities of the replicate. Only proteins that were present in ≥ 2 replicates of 4 per pull down condition were used to calculate enrichment values ($\log_2$ fold change, p-value by Welch t-test) (Table B.4). Gene information and annotations were downloaded [167, 168] and used to assign detected proteins to orthology groups, as per OrthoMCL [138]. Labeling of specific orthology groups for Figure 3.7 was performed based on the following hierarchy of species: hsap, scer, spom, cele, ecol, atha, bsub, halo, tbrt, tetr, and zmay. In other words, if an orthology group contained a human gene, it would be referred to as this. If not, the *S. cerevisiae* gene was taken, and so forth according to the listed hierarchy. If multiple genes of one species were present in the orthology group, the first one from the list would be selected.

Heatmap clustering was performed on a numerical matrix, where 1 was an enriched protein ($\log_2$ fold change > 2, p-value < 0.05), 0 a detected protein (i.e., not enriched but measured), and -1 a protein not detected within a species at all. To find similar clusters of proteins, we applied the complete linkage method (default setting) in hclust from the stats package in the R framework [112].

For functional enrichment analysis, terms were queried in the Gene Ontology (GO) [139] and the Kyoto Encyclopedia of Genes and Genomes (KEGG) [169] databases. Terms for a particular group of enriched proteins were tested for overrepresentation (adjusted p-value [FDR] $<$ 0.05; Fisher's exact test) against all terms found in the background (whole genome). The top three most overrepresented terms in each database were selected for graphical representation.

To determine known and predicted interactions, enriched proteins were queried in the STRING database version 11.5 [140]. Hits from text-mining and co-occurrence interaction sources were excluded. Hits with a score $>$ 150 in any of the remaining interaction sources (experiments, databases, coexpression, gene fusion and neighborhood) were included in the downstream analysis. Thus, protein-protein networks were generated with in-house scripts based on an R framework incorporating igraph [170], with the Fruchterman-Reingold force-directed layout algorithm implementation, and ggnetwork [171]. Enriched proteins were illustrated as nodes, where color indicates their associated experimental lesion and their shape indicates whether they are known repair proteins or not. STRING known and predicted interactions were visualized as edges. All networks were drawn with the spoke model.

For phylogenetic tree construction, the amino acid sequences of all orthologs from the respective OrthoMCL groups were extracted from the species specific protein sequence FASTA files (Table B.3). For AP endonucleases the OrthoMCL groups OG6_101139 and OG6_104339 were chosen to represent the group. OG6_104135 and OG6_100453 contains the Photolyase family and OG6_102506 contains the MutY Glycosylase family. Amino acid sequences of each family were aligned using Clustal Omega [172]. The phylogenetic tree was constructed from these alignments by using the neighbor-joining method of Clustal Omega with no genetic distance correction and no outgroup setting [172]. The phylogenetic tree was then exported as a Newick file for visualization in R alongside relevant mass-spectrometry binding results using the ape package [173].

Pfam analysis for proteins with no previous DNA repair associations was conducted using Pfam domain annotations downloaded from OrthoMCL [138]. To enable broader categorizations Pfam terms were classified into more general terms based on text-mining of the Pfam term description (Table B.12). These classifiers were used to detect the distribution of Pfam functions across the proteins that have not been previously annotated as DNA repair proteins.

3.8.4. Quantification and statistical analysis

All quantification and statistical analysis details and associated citations can be found in the method details in section 'Mass spectrometry data analysis' and 'Bioinformatics analysis and statistical analysis'. In short, the pulldowns performed in the analysis were performed in quadruplicate and p-value was determined by Welch t-test with an enrichment threshold of log2 fold change > 2 and p-value < 0.05. Utilizing both GO and KEGG databases, enriched proteins were tested for overrepresentation using Fisher's exact test determining an adjusted p-value (false discovery rate) < 0.05. The STRING database was used to determine previously established interactors to proteins of interest. To create phylogenetic tree the evolutionary history was inferred in MEGA X61 by using the Maximum Likelihood method and JTT matrix-based model [174]. Pfam domain annotations were downloaded from OrthoMCL [138].

#

4

AlexandrusPS: a user-friendly pipeline for the automated detection of orthologous gene clusters and subsequent positive selection analysis

Alejandro Ceron-Noriega, Vivien A.C. Schoonenberg, Falk Butter[1,2]
and Michal Levin[1,2]

Genome Biology and Evolution, October 13, 2023

---

[1] These authors contributed equally
[2] Corresponding authors

4.1. Summary

The evolution of protein sequences is influenced by the constraint of changes (purifying selection) or the fixation of alleles that confer fitness advantage (positive selection). An essential metric to detect the selection type driving such sequence evolution is the nucleic acid and amino acid substitution rate, namely, the nonsynonymous to synonymous substitution rate ratio. This measure has proven helpful in understanding different evolutionary adaptation processes by comparative genomics. Additionally, accounting for rate variation under different levels of selective pressure can provide insight into the functional restrictions of proteins.

These evolutionary analyses have benefited from the massive amounts of data from next-generation sequencing (NGS) technologies, making comparative genome-wide analyses more attainable.

In this project, we developed a user-friendly pipeline, AlexandrusPS, designed to simplify genome-wide positive selection analysis. The pipeline, implemented as a combination of Perl, R, and shell scripts running in a Linux/UNIX environment, is provided as an open-source solution and available as a Docker image to minimize the need for local installation. AlexandrusPS only requires CDS and peptide FASTA files as input. AlexandrusPS automatically generates orthology relationships, sequence alignments, and phylogenetic trees. It then performs site-specific (SSM), branch (BM), and branch-site (BSM) positive selection analyses.

4.2. Zusammenfassung

Die Evolution von Proteinsequenzen wird durch die Begrenzung von Veränderungen (reinigende Selektion) oder die Festlegung von Allelen, die einen Fitnessvorteil bieten (positive Selektion), beeinflusst. Ein wichtiges Maß zur Ermittlung der Selektionsart, die eine solche Sequenzevolution antreibt, ist die Nukleinsäure- und Aminosäure-Substitutionsrate, das heißt, das Verhältnis von nicht-synonymer zu synonymer Substitutionsrate. Dieses Maß hat sich als hilfreich für das Studium verschiedener evolutionärer Prozesse in der vergleichenden Genomik erwiesen. Zusätzlich kann die Berücksichtigung von Variationsraten bei Selektionsdruck Aufschluss über die funktionellen Einschränkungen von Proteinen geben.

Diese evolutionären Analysen haben von den riesigen Datenmengen der Next-Generation-Sequencing (NGS)-Technologien profitiert, die vergleichende Genomanalysen einfacher zugänglich machen.

In diesem Projekt haben wir eine benutzerfreundliche Pipeline, AlexandrusPS, entwickelt, die die genomweite positive Selektionsanalyse erleichtern soll. Die Pipeline, die als eine Kombination aus Perl-, R- und Shell-Skripten in einem Linux/UNIX-Umgebung implementiert ist, wird als

Open-Source-Lösung bereitgestellt und ist als Docker-Image verfügbar, um die Notwendigkeit einer lokalen Installation zu minimieren. AlexandrusPS benötigt nur CDS- und Peptid-FASTA-Dateien als Eingabe. AlexandrusPS generiert automatisch Orthologiebeziehungen, Sequenzalignments und phylogenetische Bäume. Anschließend führt es site-spezifische (SSM), branch-spezifische (BM) und branch-site-spezifische (BSM) Positivselektionsanalysen durch.

4.3. Statement of Contribution

For this project, Alejandro Ceron Noriega conceived the initial design, implemented the software, participated in debugging and software testing phases, and drafted the manuscript. I led the debugging phase, updated the software implementation and documentation, and conceived the Docker and Singularity implementation. Michal Levin and Falk Butter participated in the initial design, coordination, and manuscript drafting. All authors read and approved the final manuscript.

4.4. Abstract

The detection of adaptive selection in a systems approach considering all protein coding genes allows for the identification of mechanisms and pathways that enabled adaptation to different environments. Currently available programs for the estimation of positive selection signals can be divided into two groups. They are either easy to apply but can analyze only one gene family at a time, restricting systems analysis; or they can handle larger cohorts of gene families, but require considerable prerequisite data such as orthology associations, codon alignments, phylogenetic trees and proper configuration files. All these steps require extensive computational expertise restricting this endeavor to specialists. Here, we introduce AlexandrusPS, a high-throughput pipeline that overcomes technical challenges when conducting transcriptome-wide positive selection analyses on large sets of nucleotide and protein sequences. The pipeline streamlines (1) the execution of an accurate orthology prediction as a precondition for positive selection analysis, (2) preparing and organizing configuration files for CodeML, (3) performing positive selection analysis using CodeML and (4) generating an output that is easy to interpret, including all maximum likelihood and log likelihood test results. The only input needed from the user is the CDS and peptide FASTA files of proteins of interest. The pipeline is provided in a Docker image, requiring no program or module installation, enabling the application of the pipeline in any computing environment. AlexandrusPS and its documentation are available via GitHub (https://github.com/alejocn5/AlexandrusPS).

4.4.1. Significance

Understanding the mechanisms and pathways that enable adaptation to different environments is crucial in evolutionary biology. However, existing tools for detecting such adaptive processes in protein sequences have limitations in terms of the computational complexity and required resources. AlexandrusPS is a user-friendly containerized pipeline that streamlines positive selection analysis of protein-coding genes on a genome scale by automating key steps, providing an easily interpretable output and facilitating high-throughput analyses on a desktop computer.

4.5. Introduction

The evolution of protein sequences is influenced by the constraint of changes (purifying selection) or by the fixation of alleles that confer fitness advantage (positive selection) [175]. An essential metric to detect the selection type driving such sequence evolution is the nucleic acid and amino acid substitution rate, namely, the nonsynonymous (d_N) to synonymous (d_S) substitution rate ratio ($\omega = d_N/d_S$). This measure has proven to be useful for understanding different evolutionary processes in comparative genomics [176–190]. Such evolutionary analyses have profited from massive amounts of data derived from next-generation sequencing (NGS) technologies, making comparative genomics analyses more attainable.

The enormous quantity of such data provides a valuable resource for researchers, but as the number of genomes continues to grow, downstream analyses have become increasingly challenging in terms of the quality and amount of data that need to be processed. This problem has led to the need for the development of specialized, efficient and user-friendly bioinformatics tools that can help researchers in downstream tasks [191].

One of the most popular bioinformatics tools for applying maximum likelihood (ML) based models in evolutionary research to test the ratio between nonsynonymous and synonymous substitutions ($\omega = d_N/d_S$) for multiple orthologous protein-coding sequences is CodeML [192]. CodeML is implemented in the PAML (Phylogenetic Analysis by Maximum Likelihood) program package [175, 192]. While the program is statistically robust and highly accurate in examining selective pressure [193–196] CodeML also faces limitations: i) being executed on a single processing unit renders operations on large sets of sequences highly time-consuming, driving the need for accessibility to high-performance computers. ii) Each individual orthology group analysis needs to be separately prepared and executed by the user. iii) The execution requires a preceding accurate orthology analysis, which itself is challenging and can introduce errors to the analysis if not performed properly. iv) CodeML provides output that is difficult to interpret, especially for inexperienced users [48, 175, 197].

To support less experienced users and minimize the manual operation of CodeML, several programs have emerged: JCoDA [198], Armadillo [199], PAMLX [200], IMPACT_S [201], PSP [202], PhyleasProg [203], and Selecton [204]. These programs use graphical interfaces or web-server implementations for single-gene family analysis. However, they are not suitable for streamlined operation of CodeML for multiple analyses. Some additional software to solve these large-scale analysis challenges include VESPA [205] , IDEA [179], and POTION [206]. These programs still have certain shortcomings: i) The installation is complex. ii) They depend on large computational infrastructure such as high-performance computers (HPCs). iii) They require advanced programming skills of the user. Table C.1 provides a comprehensive comparison of the features and implementation properties of different available tools.

Here, we introduce AlexandrusPS, a high-throughput user-friendly pipeline designed to simplify the automated operation of established CodeML protocols. Containerized in a Docker image, AlexandrusPS was developed as a single command pipeline minimizing user intervention in both installation and execution. The pipeline provides a well-organized output table including all relevant results for drawing conclusions. All intermediate data, such as the results of the orthology analysis as well as multiple sequence alignments, are also retained. To enable full analysis flexibility for more experienced researchers, AlexandrusPS is an open source software and thus enables modifications of parameters in all major configuration files.

4.6. Implementation

4.6.1. AlexandrusPS: Functionality

AlexandrusPS is a pipeline consisting of Perl and R scripts called by a main bash shell script and is available as a Docker image \ (https://github.com/alejocn5/AlexandrusPS). The only input needed from the user is FASTA files of CDS and amino acid sequences of all target proteins. AlexandrusPS leverages the ProteinOrtho program [207] to discern and anticipate orthologous gene clusters (OGCs). These OGCs are selected for further investigation if they meet two criteria: First, they must encompass a minimum of three species; second, they exclusively consist of 1-to-1 orthologs, excluding any paralogs within the cluster spanning different species. The pipeline then utilizes PRANK to generate alignments and gene trees for each identified OGC. These gene trees are formatted in Nexus format initially but are subsequently converted to the dnd format, ensuring compatibility with subsequent analysis using CodeML.

To evaluate site-specific models (SSMs), the following model comparisons are performed: M0 versus M3, M1a versus M2a and M7 versus M8. For branch models (BM) ω values are estimated by evaluating M2 against a nearly neutral null model (M1a). For the branch-site model, M8a is compared with its null (M8a null) using a fixed ω assumption ($\omega = 1$). Subsequently, Bayesian empirical Bayes (BEB) analysis further identifies sites of positive selection, allowing posterior probability computation [208].

These results are then used for likelihood ratio tests (LRTs) to determine whether the models reflect diversifying selection. For this, the log-likelihood score (2ΔlnL) between any two models is calculated. Subsequently, the *P* value is determined by comparing each 2ΔlnL against the Chi-square distribution using the respective degrees-of-freedom (DoF) for each model pair. Significant LRT results (FDR < 0.05) indicate a significant difference between the two models and thus imply an evolutionary explanation for these differences.

The main workflow of AlexandrusPS Figure 4.1 is composed of four steps: i) Orthology prediction by ProteinOrtho [207]; ii) multiple amino acid sequence alignment and gene tree generation by PRANK [209] and DNA codon sequence alignment by pal2nal [210]; iii) site-specific model calculations by CodeML [192]; and iv) branch and branch-site-specific model calculations by CodeML.

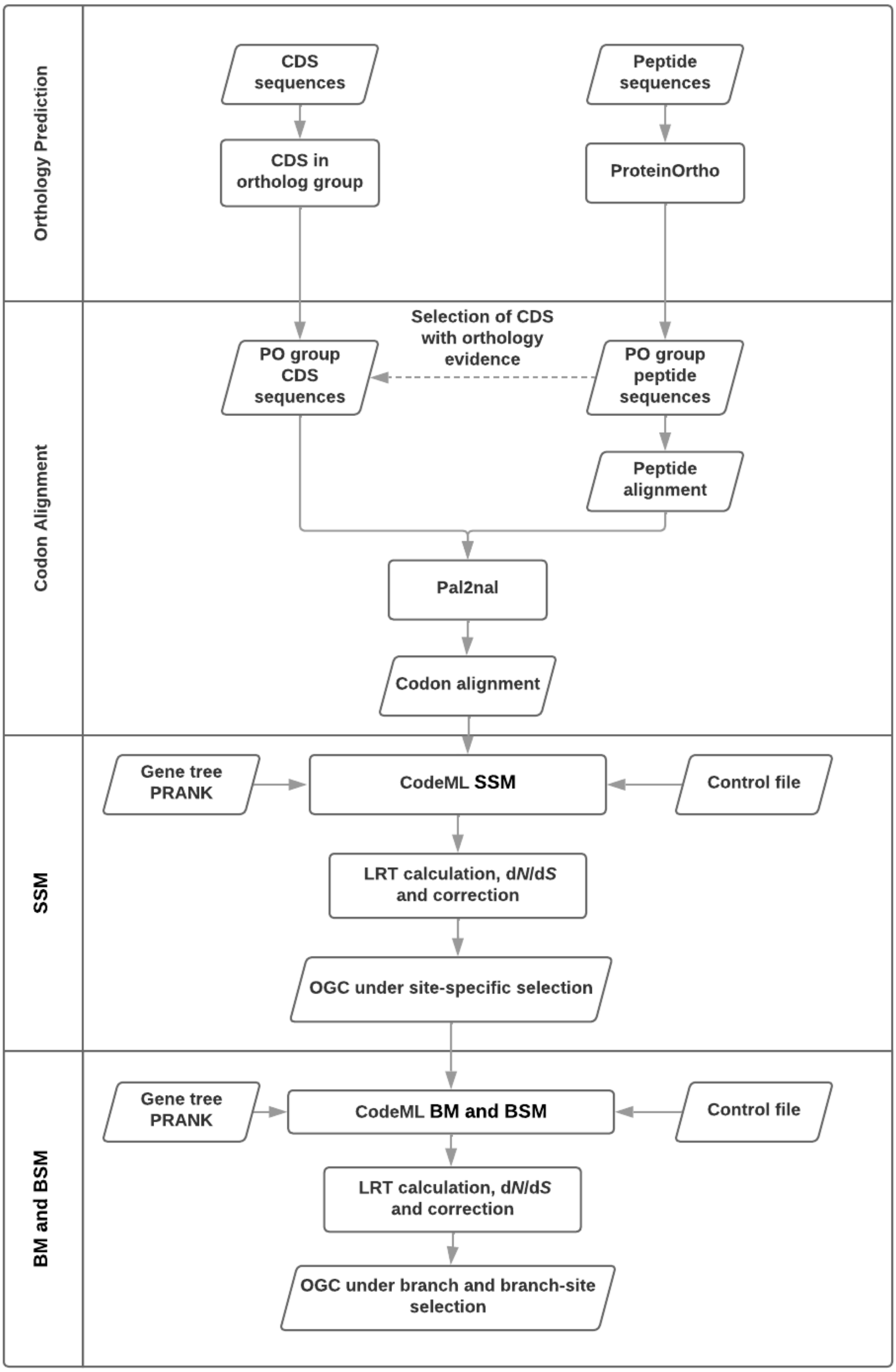


Figure 4.1.: AlexandrusPS workflow. Flowchart describing the AlexandrusPS workflow, which sequentially combines four steps to finally execute CodeML and collect results. PO = ProteinOrtho; SSM = Specific Site Model; BM = Branch Model; BSM = Branch Site Model; LRT = Likelihood Ratio Test; OGC = Orthologous Gene Cluster.

4.6.2. AlexandrusPS: Input Files

FASTA files of all proteins of interest

For each species included in the analysis two FASTA files are needed: one with the amino acid and the other with the respective CDS sequences. Both files should contain the same number of sequences and their headers must be identical. AlexandrusPS can analyze all orthologous protein groups from protein-coding genes on a genome scale across multiple species. An example dataset is provided with the pipeline to enable testing of the proper functionality of the pipeline (CDS and protein fasta files of this example dataset are also included as Supplemental Data C.1).

4.6.3. AlexandrusPS: Output Files

Site-Specific Models (SSM) The CodeML output files are parsed into a CSV file. This file contains all orthologous gene clusters (OGC) organized in rows. Columns include OGC_ID, species included in the OGC and ML results for all models with the respective metrics such as likelihood (lnL), the number of parameters (np),ω (d_N/d_S), degrees of freedom (DoF), log likelihood value (lnL), likelihood ratio tests (LRT) and positively selected sites (PSS).

Branch and branch-site models The results of the LTR-based branch and branch-site model analyses (null model (H0) and alternative model (H1) of the branch-site test) for the OGC with significant signals of site-specific diversifying selection are written into final easily interpretable results files (the final output folder containing the result files of the example dataset is included for illustration in Supplemental Data C.2). We have introduced a significant improvement in comparison to other pipelines involving CodeML. AlexandrusPS employs a more refined selection procedure testing every individual branch. Specifically, within each OGC, each individual terminal branch is sequentially designated as the foreground, with all other branches considered as background. This approach, reminiscent of the methodology employed by Anisimova and Yang in their study [211], offers numerous advantages. Concentrating on a single foreground branch alongside multiple background branches, we constrain the calculations to the count of orthologs within the examined OGC, varying from a minimum of three to a maximum of all evaluated organisms. This unbiased choice of branches for foreground and background streamlines a more unbiased analysis, ultimately enhancing the comprehensiveness of our branch and branch-site model analysis.

4.6.4. AlexandrusPS: Execution and Paralleling

Utilizing the inherent single-node architecture of CodeML, which operates on a single CPU, the parallelization process entails a series of systematic steps. Upon gathering the codon alignment, configuration files for the seven distinct CodeML models, and the phylogenetic tree specific to each Orthologous Gene Cluster (OGC) from prior stages, all relevant components are allocated to one of the accessible nodes, subsequently initializing the CodeML analysis. After extracting the values of likelihood (lnL), the number of parameters (np), ω (dN/dS), degrees of freedom (DoF), and log-likelihood value (lnL), the input and output files undergo compression. Subsequent to this, the node is freed to undertake the analysis of another OGC, continuing this iterative process until all OGCs have been analyzed. With this methodology we enable increasingly efficient processing of large volumes of OGCs with augmenting amounts of available CPUs, making the pipeline optimally adjusted to run in high-performance computing (HPC) environments. We used AlexandrusPS for a positive selection analysis including three of the nematode proteo-transcriptomes (*C. elegans, C. briggsae* and *C. inopinata*) established in [212] on a tabletop PC with 20 CPUs and on an HPC system with 128 CPUs and could reduce computation time from 12.3 hours to 2.5 hours emphasizing the added value of using the pipeline on an HPC.

4.6.5. Testing positive selection in subgroups of the phylogeny

AlexandrusPS enables positive selection analysis within OGC subgroups involving a minimum of 3 species, diverging from the conventional approach considering OGCs present in all species. This choice aims to address the potential impact of phylogenetic distances on positive selection signal dilution, which is often underestimated in large-scale analyses. Testing positive selection in subgroups relies on gene trees generated automatically by AlexandrusPS. Users should be aware that utilizing a gene tree can affect phylogenetic accuracy and positive selection detection due to distorted branch lengths, potentially leading to inaccurate substitution rate estimates and misidentification of positively selected genes. To validate positive selection signals, it is possible to confirm them with a reliable species phylogeny. After running AlexandrusPS, compressed intermediate data for each OGC can be accessed in the output folder. For validation with a species tree, replace the tree in the *.dnd.GenTree.nex file that is contained in the output/Results/<_ocgid_>.tar.gz/Orthology_Groups directory and run CodeML manually using the same config files that were already created (example config files are included in Supplemental Data C.3).

4.6.6. AlexandrusPS: Proof of principle

AlexandrusPS was successfully applied to perform a large-scale positive selection analysis using proteotranscriptomics data across 12 nematode species including 77,000 protein sequences resulting in 5,400 1-1 orthologous groups including orthologs from at least 3 species [212]. This extensive phylogenetic analysis was executed on a tabletop PC with a processor of 8 cores/16 hyperthreads (8 GB RAM each) finished within 7 days. The analysis allowed interesting new insights into the evolutionary processes of this metazoan group and uncovered evolutionary events that suggest intriguing adaptive mechanisms. Notably, *C. japonica* exhibited an exceptionally high frequency of positive selection events. Interestingly, positively selected genes in *C. japonica* are closely linked to its distinctive phoretic lifestyle, setting it apart from other Caenorhabditis species, which are predominantly free-living. In stark contrast, *C. inopinata* displayed the lowest count of positively selected protein-coding genes. This stands in sharp contrast to the findings in its sister species, *C. elegans*, where we observed an enrichment of positively selected genes associated with muscle-related functions. This discrepancy is particularly striking given the close relationship between these two species and may be attributed to the long-term cultivation of *C. elegans* in laboratory conditions. The prevalence of muscle-related functions among the positively selected genes in *C. elegans* might reflect an adaptation to distinct demands for locomotion, such as moving on two-dimensional agar plates versus navigating a three-dimensional environment in soil or on decaying fruit. Additionally, we noted widespread adaptive evolution among ribosomal proteins in seven out of the 12 species, highlighting that adaptation often occurs at fundamental gene regulatory levels rather than within highly specific functional subnetworks. Investigating these potent evolutionary changes is of significant interest and enhances our understanding of biological phenomena through in-depth phylogenetic comparisons among species that have more recently diverged.

4.7. Conclusion

AlexandrusPS is a pipeline that is available in a Docker image to avoid the need for local installation of any modules or programs. It is provided as an open-source pipeline that allows the use of various CodeML models for molecular adaptive evolution (SSM, BM, and BSM) in parallel. It can run with default parameters, as it is based on standard protocols that allow the analysis of datasets that encompass protein-coding genes on a genome scale. Users are only required to provide the CDS and peptide FASTA files of the proteins of interest. With its usage simplicity, AlexandrusPS offers distinct advantages over other programs.

AlexandrusPS automatically generates orthology relationships and identifies optimal orthology groups for positive selection analysis to avoid problems such as paralog introduction. It also generates a gene tree of each OGC and organizes, executes and extracts all pertinent information from CodeML outputs. This completely automates the analysis with no need for intervention by the user. AlexandrusPS generates four main outputs:

orthology relationships, site-specific positive selection results, branch and branch-site positive selection results, along with all intermediate files for each OGC. These intermediate files enable manual repetition of certain analyses for any individual OGC without having to repeat the entire process. AlexandrusPS allows users to run CodeML protocols on a desktop computer in an automated parallel manner, facilitating high-throughput analyses without the need for high-performance computer systems.

We successfully applied AlexandrusPS to protein-coding genes on a genome scale to investigate positive selection in a phylogeny of 12 nematode species and obtained highly interesting results [212]. We believe that this implementation will empower many more researchers to explore positive selection in any species range of interest.

4.8. Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

4.9. Acknowledgements

This project was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – GRK2526/1 – Projectnr. 407023052. VACS is supported by the Joachim Herz Stiftung through an Add-On fellowship for Interdisciplinary Life Science.

4.10. Data availability

There is no new data associated with this article. A detailed manual of the pipeline, all underlying scripts and the Docker image can be found on the AlexandrusPS GitHub page (https://github.com/alejocn5/AlexandrusPS).

5

Conclusion

The complexity of cellular behavior and its decision-making system has driven the establishment and expansion of novel “omics” and associated techniques, allowing for systematic, in-depth investigation of each aspect of molecular biology. “Omics” technologies have captured static genomic alterations, temporal transcriptomic perturbations, alternative splicing, spatiotemporal proteomic dynamics, and post-translational modifications [5]. Through this, they have contributed to our changing view on the linearity and the regulation of the informational flow of the central dogma. Regulatory mechanisms have been identified, describing interaction and variety not only on the genetic but also on the transcript and protein levels. The development and integration of multi-omics have allowed for the uncovering of intricate molecular mechanisms underlying different phenotypic manifestations of traits at high accuracy in a systematic manner, thereby forming the basis of network or systems biology [5].

In this thesis, I have utilized ”omics” technologies, specifically proteomics, and the subsequent computational data analysis and integration to investigate the systematic DNA damage response in *Tetrahymena thermophila* and identify DNA damage proteins across the Tree of Life. Additionally, I co-developed a user-friendly computational pipeline for evolutionary positive selection analysis, which relies on comparative genomics and either large-scale genome sequencing or proteotranscriptomics data.

In Chapter 2, we conducted one of the most extensive systematic studies to date, profiling the DNA damage response (DDR) in the model organism *Tetrahymena thermophila (*Nischwitz, Schoonenberg et al., *in preparation)*. Our objective was to understand the underlying DDR kinetics in *Tetrahymena*. We collected samples for transcriptome and proteome measurements over an eight-hour time course after damage induction with six different mutagens.

Here, we used Lexogen 3' mRNA QuantSeq short-read sequencing to measure the transcriptome, measured on an Illumina 500 NextSeq sequencer, even though long-read sequencing is becoming increasingly popular. In particular, long-read sequencing is a great option to resolve repetitive regions of the genome and regions with high GC content, as well as for (structural) variant discovery. In addition, nanopore sequencing is capable of label-free sequence determination of native DNA and RNA molecules without the need for amplification. It can produce long read lengths, making it optimal for de novo genome assembly. It might be interesting to perform this type of sequencing in *Tetrahymena*, as the organism has a vast genome with a high degree of repetition, making it hard to annotate and map. However, our study aimed to investigate the response to DNA-damaging agents on a large scale. As this involved a significant number of transcript measurements (9 conditions, 7 time points, 4 replicates, 252 samples), short-read sequencing was the most appropriate choice for gene expression profiling. Although short-read sequencing is considered unbiased, it is essential to note that fragmentation and library construction can introduce biases into RNA-seq results. With the sequencing of cDNA fragments, the number of reads corresponding to each transcript is proportional to the number of cDNA fragments rather than the number of transcripts. Longer transcripts will be assigned more reads as they give rise to more fragments than shorter transcripts. Thus, when carrying out differential expression analysis, the differentially expressed genes are more likely to be enriched for longer than shorter transcripts, as the statistical power is higher for longer transcripts due to the larger counts. The levels of expression (number of reads for a specific transcript) can be corrected by the transcript size. However, this information is most likely unavailable for non-model species. In this case, the correction can be performed using the contig size from the *de novo* reconstruction of the transcript (based on the reads) or the transcript sizes of a closely related model species. Nevertheless, this correction only partially solves the problem owing to the transcript size, as the sampling is always higher for longer transcripts. 3’ RNA-seq methods have been developed to minimize this bias, like the QuantSeq we used here. In the 3' RNA-seq method, mRNAs are not fragmented before reverse transcription. Instead, the cDNAs are only reverse transcribed from the 3′ end of the mRNAs, and only one copy of cDNA is generated for each transcript. Thus, when the cDNAs are sequenced, the number of reads directly reflects the number of transcripts of a specific gene, and the longer and shorter transcripts should have the same coverage of reads. The downside of 3' RNA sequencing is the loss of complete transcript information, making *de novo* transcriptome assembly impossible, and, per definition, losing information on splicing [213, 214].

We used 3' QuantSeq as it is a robust and straightforward mRNA sequencing method. As *Tetrahymena* is a well-annotated model organism, QuantSeq increases the precision in gene expression measurements as only one read per transcript is generated. Focusing on the 3′ end at lower read depths results in higher stability of differential gene expression measurements. QuantSeq is ideal for increasing the degree of multiplexing in NGS gene expression experiments and is the method of choice for accurately determining gene expression at the lowest cost [215].

For the proteome profiling in this study, we used label-free quantification, measuring 252 samples on an Orbitrap Exploris 480 mass spectrometer (Thermo Fisher Scientific) set up for LC-MS/MS data-dependent acquisition (DDA).

While chemical labeling-based quantification methods (e.g., TMT) are generally considered to possess high quantitative accuracy, they nonetheless suffer from ratio distortion and sample interference issues while being less cost-effective and offering less throughput than label-free approaches. Consequently, label-free quantification (LFQ) has been widely used in comparative quantitative experiments profiling the native and post-translationally modified proteomes [216] and is also used here. Additionally, this label-free approach allowed us to compare all the different conditions we measured (9) in a non-restricted way.

As described in Chapter 1, General Introduction, data-independent acquisition (DIA) utilizes isolation windows to co-isolate and elute fragment peptides regardless of their signal intensity, thereby providing a systematic collection of peptide fragments, as is not the case with DDA. Consequently, DIA should allow for identifying peptides with high sensitivity and improved reproducibility [217]. A major disadvantage of DIA workflows is that each MS2 scan contains multiplexed spectra from several precursor ions, making accurate identification of peptides difficult [216]. Additionally, while DIA addresses the stochasticity of precursor selection for fragmentation, it does not solve the problem of incomplete MS analysis due to the limited charge capacity of C-traps that lie upstream of Orbitraps (which have become much faster and thus have a greater analysis capacity). The limited capacity of the C-trap means that modern Orbitrap mass spectrometers only analyze <1% of available ions at the MS1 level.

For this reason, we explored a novel acquisition scheme called 'BoxCar' for this study. This method distributes the maximal charge capacity of the C-trap evenly over multiple narrow m/z segments. This limits the proportion of highly abundant species in the C-trap and greatly increases ion injection (or 'filling') times for less abundant precursor ions. A similar benefit has been observed in DIA methods, in which the instrument cycles through m/z segments to acquire fragment ion spectra of all precursors in each segment. This method has been shown to improve performance on the MS1 level drastically [218]. However, after running multiple tests with our Exploris Orbitrap setup, we could not improve sequencing depth over the established, robust DDA methods and could not achieve consistent measurements. Therefore, we decided for a label-free, DDA-based workflow, allowing us to identify 6,551 protein groups robustly.

In our data analysis, we encountered some batch effects in our proteome data. However, as we always included a matched non-treated set in the sample preparation and measurement among treatment batches, this proved to be easily corrected using the matched non-treated sample of proteins and transcripts within each treatment batch.

We performed hierarchical clustering of a curated list of DNA repair proteins to confirm the induction of DNA damage with the mutagens used. Interestingly, a distinct cluster of ten DNA damage factors, including MSH6L3 and RAD51, was upregulated at both transcript and protein levels. RAD51 and other mismatch repair proteins are critical DNA repair proteins during sexual reproduction (conjugation) in *Tetrahymena* [77, 219, 220]. We are further investigating these proteins' roles in repair, as they might mirror their role in conjugation.

To assess the kinetics of proteins and transcripts during DDR over time, we calculated their dynamicity with a Gini score. Economists initially used the Gini Index to describe inequalities in wealth distribution in populations, which varies between 0 (complete equality) and 1 (extreme inequality), and has in recent years been adopted by biologists to describe, in a simple way, the distributions of expression levels of different genes between tissues or cell lines [114, 221]. Here, using the Gini score, we can essentially collapse our temporal measurements of 8 timepoints into a single score.

While we found variability in the dynamic transcripts and proteins across treatments, we were interested in identifying treatments with overlapping dynamic responses. We found a significantly higher level of overlap than expected among three or more treatments, indicating a specific shared response alongside the overall global response. We also identified a core overlap of eight proteins between all treatments, including the aforementioned RAD51 and MSH6L3. However, there was no overlap between the 15 core dynamic transcripts and the eight core dynamic proteins, highlighting the differential regulation in transcription and protein expression.

To investigate these dynamic proteins and transcripts further, we used self-organizing maps (SOMs), an unsupervised machine learning approach, to cluster the expression profiles of all six treatments. Because of the number of treatments and time points in this data set, the SOM algorithm is an excellent way to reduce the data dimensions and get an interpretable result. The SOM algorithm is an unsupervised neural network trained to build a low-dimensional, topological map using unsupervised learning techniques. It will not cause data loss as the input data is preserved and retains the topological relations (i.e., similarity of the temporal dynamics) of the input. SOMs can handle various categorization issues while producing an insightful and practical summary of the data [222].

We determined 15 transcript and 7 protein expression profile clusters. For both, we observed complexes grouping within the same cluster, such as 20S proteasome and transcription-related factors, indicating that the clustering method successfully identifies similarly regulated complexes. In contrast, we found that PARP and PARP-correlated proteins showed specific up- or downregulation to unique treatments, and some of them displayed variable transcriptional and protein responses. However, this still needs to be confirmed through ongoing correlation analysis of the transcript and protein expression profiles. Additionally, we are currently implementing a novel knockdown system for experimental validation. We are working on examining the effects of reducing the aforementioned PARP proteins on global protein expression changes. There will likely be a compensatory DDR response, although it remains unknown whether this response will originate from other members of the PARP family or other DNA damage repair proteins. We will also assess the effects on cell survivability when these PARP proteins are reduced. Some knockdowns may exhibit sensitivity to particular DNA-damaging agents.

Ultimately, we hope this work will also serve as a resource dataset for DNA damage research. We are working on creating an accessible online database with an easy-to-use interface. We hope this propels ongoing analysis forward and opens up new research areas.

In Chapter 3 of this thesis, we studied the interactome of specific DNA damage lesions across the Tree of Life. We explored the conservation of pathways responsible for repairing and recognizing DNA damage lesions (Nischwitz, Schoonenberg, et al., *iScience*, 2023). We used a mass spectrometry-based phylointeractomics workflow, comparing the *in vitro* binding capabilities of three well-established DNA damage lesions: 8-oxoG, abasic site, and ribonucleotide incorporated into DNA. To gain a broad perspective on the repair and recognition of these lesions, we included 11 different species in our study. Previous literature has highlighted the strong conservation of fundamental proteins in the pathways that repair these lesions. Only by studying these pathways across the Tree of Life can the convergence and divergence of these different repair machinery be elucidated.

For this study, we again used a label-free quantification approach for the mass spectrometry measurements. Since we compare proteins binding to a DNA lesion over a control sequence, the study could have been a candidate for a labeling method for quantitation, such as SILAC or dimethyl labeling, allowing for accurate relative quantification; however, as we included 3 lesions, in 11 species, the label-free approach again allowed us to make unrestricted, straightforward comparisons between all conditions. SILAC labeling would not have been possible in all species tested, and dimethyl labeling would have limited the comparisons between lesions and species. With this unbiased approach, we identified several known DNA damage factors as binders to the aforementioned lesions. We enriched 337 proteins, of which 99 were related to the 'DNA repair' GO term. In addition to known DNA repair genes, we identified both species-specific and non-DNA repair proteins. These 82 species-specific proteins had no orthologs in the 10 other investigated species, offering the opportunity to study potentially unique repair or damage response aspects in their respective model organisms.

First, we focussed our study on known enriched DNA repair protein homologs, which are especially interesting for the species in which these proteins have not been characterized. Unexpectedly, we found enrichment of photolyases and MutY glycosylases, which are highly associated with DNA repair. However, the lesions included in this study are considered non-canonical targets.

Additionally, an interesting finding was the crosstalk between all three lesions. We had anticipated a high degree of overlap between the 8-oxoG and abasic lesions since they rely heavily on BER for repair. However, we also observed a high degree of overlap with proteins enriched at the uracil incorporated into the DNA. We believe this finding warrants further investigation. While a few studies have related some BER proteins to RER, more extensive research is needed [134, 135].

Next, we focused on enriched proteins not associated with the 'DNA repair' GO term. Through network, domain, and phylogenetic analysis, we identified 44 additional proteins likely to have a role in the DNA damage response.

Network analysis relied on the STRING database, connecting enriched proteins based on known interactions, revealing connection with known DNA repair proteins.

For instance, within the *S. cerevisiae* network, there was an incredibly elaborate network of chromatin remodelers. While few had DNA repair designation, many interacted with those DNA repair-associated chromatin remodelers and amongst themselves. This leads to the conclusion that a more extensive network of chromatin remodelers may be involved in DNA repair than previously thought.

Further, we used Pfam domain annotations in the non-DNA repair proteins to find potential unknown functions. We curated the major domains into categories based on their general descriptions, again inferring the previously unknown potential for DNA repair function in these proteins.

Finally, the most significant finding in our study was through phylogenetic analysis. With this, we could identify enriched proteins across species that had not been previously associated with repair. We found five instances of these orthology groups, providing strong evidence that our screen successfully discovered novel DNA repair proteins across species.

Notably, while we enriched proteins previously associated with the recognition and repair of the three studied lesions, not all previously described proteins were identified in our screen. This was not unexpected as the cells' physiological conditions, such as pH, temperature, salt concentration, etc., are highly specific and cannot consistently be replicated. Nonetheless, our ability to identify classical repair proteins reinforces the validity of our screen. Altogether, our study opens avenues for further investigation of newly identified candidates and exploration of key factors in the crosstalk between BER and RER DNA damage pathways.

Finally, in Chapter 4, we developed a computational pipeline to make positive selection analysis user-friendly (Ceron-Noriega et al., *Genome Biology and Evolution,* 2023). AlexandrusPS generates orthology relationships, sequence alignments, and phylogenetic trees with its automated process. It then performs site-specific (SSM), branch (BM), and branch-site (BSM) positive selection analyses. It produces four main output files, including orthology relationships, positive selection results, and all intermediate files (sequence alignments, phylogenetic trees).

The development of this tool came forth out of other work by our lab, in which high-throughput experimental data, such as RNA-seq and peptide evidence, was integrated to facilitate accurate protein-coding gene annotation [17, 212]. Using proteotranscriptomics leads to highly valid gene prediction even in species without a reference genome, which is crucial for conducting any evolutionary analyses such as positive selection.

AlexandrusPS implements standard CodeML protocols and aims to avoid biases in positive selection identification. Combining high-throughput omics data, creating high-quality gene annotations, and appropriate positive selection analysis (as per AlexandursPS) allows for a comprehensive evolutionary analysis. This was demonstrated in nematodes by Ceron-Noriega et al., extending the understanding gained from decades of research on *C. elegans* to a diverse range of nematode species with different life histories, modes of reproduction, and habitats. This analysis shed light on how nematode species have evolved to better adapt to their environments through changes in genes involved in stress response, detoxification, metabolism, reproduction, and development [212, 223].

Using “omics” technologies, specifically proteotranscriptomics, results in highly reliable and experimentally validated gene annotations. These annotations can advance evolutionary studies, including the analysis of positive selection and phylogeny. It underscores the importance and impact of large data sets in evolutionary analyses and is a valuable foundation for future research.

In conclusion, in this thesis, we show various applications and analyses of different ”omics” technologies, specifically proteomics. We find that proteomics and transcriptomics, and their subsequent integration, give rise to unbiased approaches to investigating large biological questions. We show that these methods create large amounts of data, requiring different approaches to multi-omics integration so that we can try to construct comprehensive relationships between molecular signatures, systems, mechanisms, and phenotypic manifestations.

Bibliography

[1] Aizat WM, Ismail I, Noor NM (2018). Recent Development in Omics Studies. In: Aizat WM, Goh H-H, Baharum SN, editors Omics Applications for Systems Biology. Cham: Springer International Publishing, 1–9.

[2] Azvolinsky A (2019). Demystifying Proteomics in Hematology.

[3] Qin H, Niu T, Zhao J (2019). Identifying Multi-Omics Causers and Causal Pathways for Complex Traits. Frontiers in Genetics 10:

[4] Franklin S, Vondriska TM (2011). Genomes, Proteomes and the Central Dogma. Circ Cardiovasc Genet, 4:576.

[5] Dai X, Shen L (2022). Advances and Trends in Omics Technology Development. Frontiers in Medicine 9:

[6] Ebrahim A, Brunk E, Tan J, O'Brien EJ, Kim D, Szubin R, et al. (2016). Multi-omic data integration enables discovery of hidden biological regularities. Nat Commun, 7:13091.

[7] Goh H-H (2018). Integrative Multi-Omics Through Bioinformatics. In: Aizat WM, Goh H-H, Baharum SN, editors Omics Applications for Systems Biology. Cham: Springer International Publishing, 69–80.

[8] Chen C, Wang J, Pan D, Wang X, Xu Y, Yan J, et al. (2023). Applications of multi-omics analysis in human diseases. MedComm, 4:e315.

[9] Derks KWJ, Hoeijmakers JHJ, Pothof J (2014). The DNA damage response: The omics era and its impact. DNA Repair, 19:214–220.

[10] Pervez MT, Hasnain MJ ul, Abbas SH, Moustafa MF, Aslam N, Shah SSM (2022). A Comprehensive Review of Performance of Next-Generation Sequencing Platforms. Biomed Res Int, 2022:3457806.

[11] Kumar KR, Cowley MJ, Davis RL (2019). Next-Generation Sequencing and Emerging Technologies. Semin Thromb Hemost, 45:661–673.

[12] Logsdon GA, Vollger MR, Eichler EE (2020). Long-read human genome sequencing and its applications. Nat Rev Genet, 21:597–614.

[13] Mutz K-O, Heilkenbrinker A, Lönne M, Walter J-G, Stahl F (2013). Transcriptome analysis using next-generation sequencing. Current Opinion in Biotechnology, 24:22–30.

[14] Belton J-M, McCord RP, Gibcus JH, Naumova N, Zhan Y, Dekker J (2012). Hi-C: A comprehensive technique to capture the conformation of genomes. Methods, 58:268–276.

[15] Skene PJ, Henikoff S (2017). An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. eLife, 6:e21856.

[16] Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q (2020). Opportunities and challenges in long-read sequencing data analysis. Genome Biology, 21:30.

[17] Levin M, Butter F (2022). Proteotranscriptomics - A facilitator in omics research. Comput Struct Biotechnol J, 20:3667–3675.

[18] Musich R, Cadle-Davidson L, Osier MV (2021). Comparison of Short-Read Sequence Aligners Indicates Strengths and Weaknesses for Biologists to Consider. Frontiers in Plant Science. doi: 10.3389/fpls.2021.657240.

[19] Sarantopoulou D, Brooks TG, Nayak S, Mrčela A, Lahens NF, Grant GR (2021). Comparative evaluation of full-length isoform quantification from RNA-Seq. BMC Bioinformatics, 22:266.

[20] Dong X, Du MRM, Gouil Q, Tian L, Jabbari JS, Bowden R, et al. (2023). Benchmarking long-read RNA-sequencing analysis tools using in silico mixtures. Nat Methods, 20:1810–1821.

[21] Jones EF, Haldar A, Oza VH, Lasseigne BN (2023). Quantifying transcriptome diversity: A review. Briefings in Functional Genomics, elad019.

[22] Kanehisa M, Goto S (2000). KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res, 28:27–30.

[23] The Gene Ontology Consortium, Aleksander SA, Balhoff J, Carbon S, Cherry JM, Drabkin HJ, et al. (2023). The Gene Ontology knowledgebase in 2023. Genetics, 224:iyad031.

[24] Schubert OT, Röst HL, Collins BC, Rosenberger G, Aebersold R (2017). Quantitative proteomics: Challenges and opportunities in basic and applied research. Nat Protoc, 12:1289–1294.

[25] Aebersold R, Mann M (2003). Mass spectrometry-based proteomics. Nature, 422:198–207.

[26] Domon B, Aebersold R (2010). Options and considerations when selecting a quantitative proteomics strategy. Nat Biotechnol, 28:710–721.

[27] Rozanova S, Barkovits K, Nikolov M, Schmidt C, Urlaub H, Marcus K (2021). Quantitative Mass Spectrometry-Based Proteomics: An Overview. In: Marcus K, Eisenacher M, Sitek B, editors Quantitative Methods in Proteomics. New York, NY: Springer US, 85–116.

[28] Shuken SR (2023). An Introduction to Mass Spectrometry-Based Proteomics. J Proteome Res, 22:2151–2171.

[29] Garabedian A, Benigni P, Ramirez CE, Baker ES, Liu T, Smith RD, et al. (2018). Towards Discovery and Targeted Peptide Biomarker Detection Using nanoESI-TIMS-TOF MS. J Am Soc Mass Spectrom, 29:817–826.

[30] Allen DR, McWhinney BC (2019). Quadrupole Time-of-Flight Mass Spectrometry: A Paradigm Shift in Toxicology Screening Applications. Clin Biochem Rev, 40:135–146.

[31] Pino LK, Rose J, O'Broin A, Shah S, Schilling B (2020). Emerging mass spectrometry-based proteomics methodologies for novel biomedical applications. Biochemical Society Transactions, 48:1953–1966.

[32] Guan S, Taylor PP, Han Z, Moran MF, Ma B (2020). Data Dependent–Independent Acquisition (DDIA) Proteomics. J Proteome Res, 19:3230–3237.

[33] Bantscheff M, Schirle M, Sweetman G, Rick J, Kuster B (2007). Quantitative mass spectrometry in proteomics: A critical review. Anal Bioanal Chem, 389:1017–1031.

[34] Larance M, Bailly AP, Pourkarimi E, Hay RT, Buchanan G, Coulthurst S, et al. (2011). Stable Isotope Labeling with Amino acids in Nematodes. Nat Methods, 8:849–851.

[35] Boersema PJ, Raijmakers R, Lemeer S, Mohammed S, Heck AJR (2009). Multiplex peptide stable isotope dimethyl labeling for quantitative proteomics. Nat Protoc, 4:484–494.

[36] Li J, Van Vranken JG, Pontano Vaites L, Schweppe DK, Huttlin EL, Etienne C, et al. (2020). TMTpro reagents: A set of isobaric labeling mass tags enables simultaneous proteome-wide measurements across 16 samples. Nat Methods, 17:399–404.

[37] Zecha J, Satpathy S, Kanashova T, Avanessian SC, Kane MH, Clauser KR, et al. (2019). TMT Labeling for the Masses: A Robust and Cost-efficient, In-solution Labeling Approach. Mol Cell Proteomics, 18:1468–1478.

[38] Bantscheff M, Lemeer S, Savitski MM, Kuster B (2012). Quantitative mass spectrometry in proteomics: Critical review update from 2007 to the present. Anal Bioanal Chem, 404:939–965.

[39] Cox J, Mann M (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol, 26:1367–1372.

[40] Cox J, Hein MY, Luber CA, Paron I, Nagaraj N, Mann M (2014). Accurate Proteome-wide Label-free Quantification by Delayed Normalization and Maximal Peptide Ratio Extraction, Termed MaxLFQ. Mol Cell Proteomics, 13:2513–2526.

[41] Ciccia A, Elledge SJ (2010). The DNA Damage Response: Making it safe to play with knives. Molecular cell, 40:179–204.

[42] Pino JC, Lubbock ALR, Harris LA, Gutierrez DB, Farrow MA, Muszynski N, et al. (2022). Processes in DNA damage response from a whole-cell multi-omics perspective. iScience, 25:105341.

[43] Giglia-Mari G, Zotter A, Vermeulen W (2011). DNA Damage Response. Cold Spring Harb Perspect Biol, 3:a000745.

[44] Molinaro C, Martoriati A, Cailliau K (2021). Proteins from the DNA Damage Response: Regulation, Dysfunction, and Anticancer Strategies. Cancers, 13:3819.

[45] Chen Z, Chen J (2021). Mass spectrometry-based protein-protein interaction techniques and their applications in studies of DNA damage repair. J Zhejiang Univ Sci B, 22:1–20.

[46] Stokes MP, Zhu Y, Farnsworth CL (2018). Mass spectrometry-based proteomic analysis of the DNA damage response. FBL, 23:597–613.

[47] Snead AA, Clark RD (2022). The Biological Hierarchy, Time, and Temporal ‘Omics in Evolutionary Biology: A Perspective. Integrative and Comparative Biology, 62:1872–1886.

[48] Steffen R, Ogoniak L, Grundmann N, Pawluchin A, Soehnlein O, Schmitz J (2022). paPAML: An improved computational tool to explore selection pressure on protein-coding sequences. Genes, 13:1090.

[49] Li WH, Wu CI, Luo CC (1985). A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol Biol Evol, 2:150–174.

[50] Liu J, Zhang Y, Lei X, Zhang Z (2008). Natural selection of protein structural and functional properties: A single nucleotide polymorphism perspective. Genome Biol, 9:R69.

[51] Montoya-Burgos JI (2011). Patterns of positive selection and neutral evolution in the protein-coding genes of Tetraodon and Takifugu. PLoS One, 6:e24800.

[52] Booker TR, Jackson BC, Keightley PD (2017). Detecting positive selection in the genome. BMC Biology, 15:98.

[53] Nembaware V, Crum K, Kelso J, Seoighe C (2002). Impact of the presence of paralogs on sequence divergence in a set of mouse-human orthologs. Genome Res, 12:1370–1376.

[54] Kovalchuk I (2016). Chapter 38 - Conserved and Divergent Features of DNA Repair: Future Perspectives in Genome Instability Research. In: Kovalchuk I, Kovalchuk O, editors Genome Stability. Boston: Academic Press, 651–666.

[55] Jackson SP, Bartek J (2009). The DNA-Damage response in human biology and disease. Nature, 461:1071–78.

[56] Beard WA, Horton JK, Prasad R, Wilson SH (2019). Eukaryotic Base Excision Repair: New Approaches Shine Light on Mechanism. Annual Review of Biochemistry, 88:137–162.

[57] Chapman JR, Taylor MRG, Boulton SJ (2012). Playing the end game: DNA double-strand break repair pathway choice. Molecular Cell, 47:497–510.

[58] Deans AJ, West SC (2011). DNA interstrand crosslink repair and cancer. Nature Reviews Cancer, 11:467–80.

[59] Li Z, Pearlman AH, Hsieh P (2016). DNA mismatch repair and the DNA damage response. DNA Repair, 38:94–101.

[60] Schärer OD (2013). Nucleotide Excision Repair in Eukaryotes. Cold Spring Harbor Perspectives in Biology, 5:a012609.

[61] Spivak G (2015). Nucleotide excision repair in humans. DNA Repair, 36:13–18.

[62] Huang H, Zhu L, Reid BR, Drobny GP, Hopkins PB (1995). Solution structure of a cisplatin-induced DNA interstrand cross-link. Science, 270:1842–45.

[63] Duan M, Ulibarri J, Liu KJ, Mao P (2020). Role of Nucleotide Excision Repair in Cisplatin Resistance. International Journal of Molecular Sciences, 21:9248.

[64] Ransy C, Vaz C, Lombès A, Bouillaud F (2020). Use of H2O2 to cause oxidative stress, the catalase issue. International Journal of Molecular Sciences, 21:9149.

[65] Fayyad N, Kobaisi F, Beal D, Mahfouf W, Ged C, Morice-Picard F, et al. (2020). Xeroderma Pigmentosum C (XPC) Mutations in Primary Fibroblasts Impair Base Excision Repair Pathway and Increase Oxidative DNA Damage. Frontiers in Genetics, 11:561687.

[66] Kumar N, Theil AF, Roginskaya V, Ali Y, Calderon M, Watkins SC, et al. (2022). Global and transcription-coupled repair of 8-oxoG is initiated by nucleotide excision repair proteins. Nature Communications, 13:974.

[67] Agrawal RK, Patel RK, shah V, Nainiwal L, Trivedi B (2014). Hydroxyurea in sickle cell disease: Drug review. Indian Journal of Hematology & Blood Transfusion, 30:91–96.

[68] Petermann E, Orta ML, Issaeva N, Schultz N, Helleday T (2010). Hydroxyurea-stalled replication forks become progressively inactivated and require two different RAD51-Mediated pathways for restart and repair. Molecular Cell, 37:492–502.

[69] Bharati AP, Kumari S, Akhtar MS (2020). Proteome analysis of saccharomyces cerevisiae after methyl methane sulfonate (MMS) treatment. Biochemistry and Biophysics Reports, 24:100820.

[70] Hauser M, Abraham PE, Barcelona L, Becker JM (2019). UV Laser-Induced, Time-Resolved Transcriptome Responses of Saccharomyces Cerevisiae. G3: Genes|Genomes|Genetics, 9:2549–60.

[71] Kim DR, Gidvani RD, Ingalls BP, Duncker BP, McConkey BJ (2011). Differential chromatin proteomics of the MMS-Induced DNA damage response in yeast. doi: 10.1186/1477-5956-9-62.

[72] Kubota T, Stead DA, Hiraga S, Have S, Donaldson AD (2012). Quantitative proteomic analysis of yeast DNA replication proteins. Methods, 57:196–202.

[73] Rodríguez-Lombardero S, Vizoso-Vázquez Á, Lombardía LJ, Manuel Becerra MIG-S, Cerdán ME (2014). Sky1 regulates the expression of sulfur metabolism genes in response to cisplatin. Microbiology, 160:1357–68.

[74] Suter B, Auerbach D, Stagljar I (2006). Yeast-based functional genomics and proteomics technologies: The first 15 years and beyond. BioTechniques, 40:625–44.

[75] Zhou C, Elia AEH, Naylor ML, Dephoure N, Ballif BA, Goel G, et al. (2016). Profiling DNA damage-induced phosphorylation in budding yeast reveals diverse signaling networks. Proceedings of the National Academy of Sciences, 113:3667–75.

[76] Campbell C, Romero DP (1998). Identification and characterization of the RAD51 gene from the ciliate tetrahymena thermophila. Nucleic Acids Research, 26:3165–72.

[77] Loidl J, Mochizuki K (2009). Tetrahymena meiotic nuclear reorganization is induced by a checkpoint Kinase–Dependent response to DNA damage. Molecular Biology of the Cell, 20:2428–37.

[78] Sandoval PY, Lee P-H, Meng X, Kapler GM (2015). Checkpoint activation of an unconventional DNA replication program in tetrahymena. PLoS Genetics, 11:1005405.

[79] Tatum D, Li S (2011). Nucleotide Excision Repair in S. cerevisiae. DNA Repair - On the Pathways to Fixing DNA Damage and Errors. doi: 10.5772/22129.

[80] Kelley MR, Kow YW, Wilson DM III (2003). Disparity between DNA base excision repair in yeast and mammals: Translational Implications1. Cancer Research, 63:549–54.

[81] Bowen N, Smith CE, Srivatsan A, Willcox S, Griffith JD, Kolodner RD (2013). Reconstitution of long and short patch mismatch repair reactions using saccharomyces cerevisiae proteins. Proceedings of the national academy of sciences of the united states of america 110. 18472–77.

[82] Chakraborty U, Alani E (2016). Understanding how mismatch repair proteins participate in the Repair/Anti-Recombination decision. FEMS Yeast Research, 16:071.

[83] Kunkel TA, Erie DA (2015). Eukaryotic mismatch repair in relation to DNA replication. Annual Review of Genetics, 49:291–313.

[84] Li X, Heyer W-D (2008). Homologous recombination in DNA repair and DNA damage tolerance. Cell Research, 18:99–113.

[85] Mathiasen DP, Lisby M (2014). Cell cycle regulation of homologous recombination in saccharomyces cerevisiae. FEMS Microbiology Reviews, 38:172–84.

[86] Pannunzio NR, Watanabe G, Lieber MR (2018). Nonhomologous DNA end-joining for repair of DNA double-strand breaks. The Journal of Biological Chemistry, 293:10512–23.

[87] Scully R, Panday A, Elango R, Willis NA (2019). DNA double-strand break repair-pathway choice in somatic mammalian cells. Nature Reviews Molecular Cell Biology, 20:698–714.

[88] Lehoczký P, McHugh PJ, Chovanec M (2007). DNA interstrand cross-link repair in saccharomyces cerevisiae. FEMS Microbiology Reviews, 31:109–33.

[89] Pizzul P, Casari E, Gnugnoli M, Rinaldi C, Corallo F, Longhese MP (2022). The DNA damage checkpoint: A tale from budding yeast. Frontiers in Genetics. doi: 10.3389/fgene.2022.995163.

[90] Marsh TC, Cole ES, Stuart KR, Campbell C, Romero DP (2000). RAD51 is required for propagation of the germinal nucleus in tetrahymena thermophila. Genetics, 154:1587–96.

[91] Marsh TC, Cole ES, Romero DP (2001). The transition from conjugal development to the first vegetative cell division is dependent on RAD51 expression in the ciliate tetrahymena thermophila. Genetics, 157:1591–98.

[92] Evans WE, Yee GC, Crom WR, Pratt CB, Green AA (1982). Clinical pharmacology of bleomycin and cisplatin. Drug Intelligence & Clinical Pharmacy, 16:448–58.

[93] Ashraf K, Nabeel-Shah S, Garg J, Saettone A, Derynck J, Gingras A-C, et al. (2019). Proteomic analysis of histones H2A/H2B and variant Hv1 in tetrahymena thermophila reveals an ancient network of chaperones. Molecular Biology and Evolution, 36:1037–55.

[94] Chalker DL, Meyer E, Mochizuki K (2013). Epigenetics of ciliates. Cold Spring Harbor Perspectives in Biology, 5:017764.

[95] Saettone A, Nabeel-Shah S, Garg J, Lambert J-P, Pearlman RE, Fillingham J (2019). Functional proteomics of nuclear proteins in tetrahymena thermophila: A review. Genes, 10:333.

[96] Slade KM, Freggiaro S, Cottrell KA, Smith JJ, Wiley EA (2011). Sirtuin-mediated nuclear differentiation and programmed degradation in tetrahymena. BMC Cell Biology, 12:40.

[97] Wahab S, Saettone A, Nabeel-Shah S, Dannah N, Fillingham J (2020). Exploring the histone acetylation cycle in the protozoan model tetrahymena thermophila. Frontiers in Cell and Developmental Biology, 8:509.

[98] Morales J, Li L, Fattah FJ, Dong Y, Bey EA, Patel M, et al. (2014). Review of poly (ADP-Ribose) polymerase (PARP) mechanisms of action and rationale for targeting in cancer and other diseases. Critical Reviews in Eukaryotic Gene Expression, 24:15–28.

[99] Sousa FG, Matuo R, Soares DG, Escargueil AE, Henriques JAP, Larsen AK, et al. (2012). PARPs and the DNA Damage Response. Carcinogenesis, 33:1433–40.

[100] Clingen PH, Wu JY-H, Miller J, Mistry N, Chin F, Wynne P, et al. (2008). Histone H2AX phosphorylation as a molecular pharmacological marker for DNA interstrand crosslink cancer chemotherapy. Biochemical Pharmacology, 76:19–27.

[101] Huang J, Zhang J, Bellani MA, Pokharel D, Gichimu J, James RC, et al. (2019). Remodeling of Interstrand Crosslink Proximal Replisomes Is Dependent on ATR, FANCM, and FANCD2. Cell Reports, 27:1794–1808 5.

[102] Lopez-Martinez D, Kupculak M, Yang, Yoshikawa Y, Liang C-C, Wu R, et al. (2019). Phosphorylation of FANCD2 inhibits the FANCD2/FANCI complex and suppresses the fanconi anemia pathway in the absence of DNA damage. Cell Reports, 27:2990–3005 5.

[103] Lin I-T, Chao J-L, Yao M-C (2012). An essential role for the DNA breakage-repair protein Ku80 in programmed DNA rearrangements in tetrahymena thermophila. Molecular Biology of the Cell, 23:2213–25.

[104] Gaertig J, Thatcher TH, Gu L, Gorovsky MA (1994). Electroporation-mediated replacement of a positively and negatively selectable beta-tubulin gene in tetrahymena thermophila. Proceedings of the National Academy of Sciences of the United States of America, 91:4549–53.

[105] Gaertig J, Gao Y, Tishgarten T, Clark TG, Dickerson HW (1999). Surface display of a parasite antigen in the ciliate tetrahymena thermophila. Nature Biotechnology, 17:462–65.

[106] Smith JJ, Yakisich JS, Kapler GM, Cole ES, Romero DP (2004). A β-Tubulin mutation selectively uncouples nuclear division and cytokinesis in tetrahymena thermophila. Eukaryotic Cell, 3:1217–26.

[107] Scherer M, Levin M, Butter F, Scheibe M (2020). Quantitative Proteomics to Identify Nuclear RNA-Binding Proteins of Malat1. International Journal of Molecular Sciences, 21:E1166.

[108] Rappsilber J, Mann M, Ishihama Y (2007). Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. Nature Protocols, 2:1896–1906.

[109] Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. (2013). STAR: Ultrafast universal RNA-Seq aligner. Bioinformatics, 29:15–21.

[110] Liao Y, Smyth GK, Shi W (2014). FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics, 30:923–30.

[111] Love MI, Huber W, Anders S (2014). Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. Genome Biology, 15:550.

[112] R Core Team R: A language and environment for statistical computing. R Foundation for Statistical Computing: Vienna, Austria; 2022.

[113] Wickham H Ggplot2: Elegant graphics for data analysis. Springer-Verlag New York; 2016.

[114] Casas-Vila N, Bluhm A, Sayols S, Dinges N, Dejung M, Altenhein T, et al. (2017). The developmental proteome of Drosophila melanogaster. Genome Res, 27:1273–1285.

[115] Damgaard C, Weiner J (2000). Describing inequality in plant size or fecundity. Ecology, 81:1139–42.

[116] Wang M, Zhao Y, Zhang B (2015). Efficient test and visualization of multi-set intersections. Scientific Reports, 5:16923.

[117] Kanehisa M, Furumichi M, Sato Y, Kawashima M, Ishiguro-Watanabe M (2023-01-06, 2023-1-6). KEGG for taxonomy-based analysis of pathways and genomes. Nucleic Acids Research, 51:D587–D592.

[118] Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, et al. (2021). clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. The Innovation, 2:100141.

[119] Yu G, Wang L-G, Han Y, He Q-Y (2012). clusterProfiler: An R package for comparing biological themes among gene clusters. OMICS: A Journal of Integrative Biology, 16:284–287.

[120] Wehrens R, Buydens LMC (2007). Self- and super-organizing maps in R: The kohonen package. Journal of Statistical Software, 21:1–19.

[121] Blischak JD, Carbonetto P, Stephens M (2019). Creating and sharing reproducible research code the workflowr way. F1000Res, 8:1749.

[122] Klungland A, Rosewell I, Hollenbach S, Larsen E, Daly G, Epe B, et al. (1999). Accumulation of premutagenic DNA lesions in mice defective in removal of oxidative base damage. Proceedings of the National Academy of Sciences, 96:13300–13305.

[123] Marshall CJ, Santangelo TJ (2020). Archaeal DNA Repair Mechanisms. Biomolecules, 10:1472.

[124] Genois M-M, Paquet ER, Laffitte M-CN, Maity R, Rodrigue A, Ouellette M, et al. (2014). DNA Repair Pathways in Trypanosomatids: From DNA Repair to Drug Resistance. Microbiology and Molecular Biology Reviews, 78:40–73.

[125] Yao S, Feng Y, Zhang Y, Feng J (2021). DNA damage checkpoint and repair: From the budding yeast Saccharomyces cerevisiae to the pathogenic fungus Candida albicans. Computational and Structural Biotechnology Journal, 19:6343–6354.

[126] Robertson AB, Klungland A, Rognes T, Leiros I (2009). DNA repair in mammalian cells: Base excision repair: The long and short of it. Cellular and molecular life sciences: CMLS, 66:981–993.

[127] Córdoba-Cañero D, Morales-Ruiz T, Roldán-Arjona T, Ariza RR (2009). Single-nucleotide and long-patch base excision repair of DNA damage in plants. The Plant Journal, 60:716–728.

[128] Lindahl T, Nyberg B (1972). Rate of depurination of native deoxyribonucleic acid. Biochemistry, 11:3610–3618.

[129] Gredilla R, Garm C, Stevnsner T (2012). Nuclear and mitochondrial DNA repair in selected eukaryotic aging model systems. Oxid Med Cell Longev, 2012:282438.

[130] Reijns MAM, Rabe B, Rigby RE, Mill P, Astell KR, Lettice LA, et al. (2012). Enzymatic removal of ribonucleotides from DNA is essential for mammalian genome integrity and development. Cell, 149:1008–1022.

[131] Nick McElhinny SA, Watts BE, Kumar D, Watt DL, Lundström E-B, Burgers PMJ, et al. (2010). Abundant ribonucleotide incorporation into DNA by yeast replicative polymerases. Proceedings of the National Academy of Sciences, 107:4949–4954.

[132] Williams JS, Lujan SA, Kunkel TA (2016). Processing ribonucleotides incorporated during eukaryotic DNA replication. Nature Reviews Molecular Cell Biology, 17:350–363.

[133] Balachander S, Gombolay AL, Yang T, Xu P, Newnam G, Keskin H, et al. (2020). Ribonucleotide incorporation in yeast genomic DNA shows preference for cytosine and guanosine preceded by deoxyadenosine. Nat Commun, 11:2447.

[134] Sassa A, Yasui M, Honma M (2019). Current perspectives on mechanisms of ribonucleotide incorporation and processing in mammalian DNA. Genes Environ, 41:3.

[135] Kellner V, Luke B (2020). Molecular and physiological consequences of faulty eukaryotic ribonucleotide excision repair. The EMBO Journal, 39:e102309.

[136] Kappei D, Scheibe M, Paszkowski-Rogacz M, Bluhm A, Gossmann TI, Dietz S, et al. (2017). Phylointeractomics reconstructs functional evolution of protein binding. Nature Communications, 8:14334.

[137] Li L, Stoeckert CJ, Roos DS (2003). OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. Genome Research, 13:2178–2189.

[138] Chen F, Mackey AJ, Stoeckert CJ, Roos DS (2006). OrthoMCL-DB: Querying a comprehensive multi-species collection of ortholog groups. Nucleic Acids Research, 34:D363–368.

[139] The Gene Ontology Consortium, Carbon S, Douglass E, Good BM, Unni DR, Harris NL, et al. (2021). The Gene Ontology resource: Enriching a GOld mine. Nucleic Acids Research, 49:D325–D334.

[140] Szklarczyk D, Gable AL, Nastou K, Lyon D, Kiirsch R, Pyysalo S, et al. (2021-01-08, 2021-1-8). The STRING database in 2021: Customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic acids research. doi: 10.1093/nar/gkaa1074.

[141] Mei Q, Dvornyk V (2015-09-09, 2015-9-9). Evolutionary history of the Photolyase/Cryptochrome superfamily in eukaryotes. PLoS ONE, 10:e0135940.

[142] Yudkina AV, Shilkin ES, Endutkin AV, Makarova AV, Zharkov DO (2019-05, 2019-5). Reading and misreading 8-oxoguanine, a paradigmatic ambiguous nucleobase. Crystals, 9:269.

[143] Elsakrmy N, Zhang-Akiyama Q-M, Ramotar D (2020). The Base Excision Repair Pathway in the Nematode Caenorhabditis elegans. Frontiers in Cell and Developmental Biology, 8:598860.

[144] Besemer AS, Maus J, Ax MDA, Stein A, Vo S, Freese C, et al. (2021). Receptor-mediated endocytosis 8 (RME-8)/DNAJC13 is a novel positive modulator of autophagy and stabilizes cellular protein homeostasis. Cellular and Molecular Life Sciences, 78:645–660.

[145] Gorenberg EL, Chandra SS (2017). The role of co-chaperones in synaptic proteostasis and neurodegenerative disease. Frontiers in Neuroscience, 11:248.

[146] Astolfi A, Fiore M, Melchionda F, Indio V, Bertuccio SN, Pession A (2019-05, 2019-5). BCOR involvement in cancer. Epigenomics, 11:835–855.

[147] Stadler J, Richly H (2017). Regulation of DNA Repair Mechanisms: How the Chromatin Environment Regulates the DNA Damage Response. International Journal of Molecular Sciences, 18:1715.

[148] Luijsterburg MS, de Krijger I, Wiegant WW, Shah RG, Smeenk G, de Groot AJL, et al. (2016). PARP1 Links CHD2-Mediated Chromatin Expansion and H3.3 Deposition to DNA Repair by Non-homologous End-Joining. Molecular Cell, 61:547–562.

[149] Tanwar VS, Jose CC, Cuddapah S (2019). Role of CTCF in DNA Damage Response. Mutation research, 780:61–68.

[150] Czaja W, Mao P, Smerdon MJ (2014). Chromatin remodelling complex RSC promotes base excision repair in chromatin of Saccharomyces cerevisiae. DNA Repair, 16:35–43.

[151] Pedreira T, Elfmann C, Stülke J (2022). The current state of *Subti* Wiki, the database for the model organism *Bacillus Subtilis*. Nucleic Acids Research, 50:D875–D882.

[152] Liu Y, Rodriguez Y, Ross RL, Zhao R, Watts JA, Grunseich C, et al. (2020-08-25, 2020-8-25). RNA abasic sites in yeast and human cells. Proceedings of the National Academy of Sciences, 117:20689–20695.

[153] Redrejo-Rodríguez M, Vigouroux A, Mursalimov A, Grin I, Alili D, Koshenov Z, et al. (2016). Structural comparison of AP endonucleases from the exonuclease III family reveals new amino acid residues in human AP endonuclease 1 that are involved in incision of damaged DNA. Biochimie, 128–129:20–33.

[154] Daley JM, Zakaria C, Ramotar D (2010). The endonuclease IV family of apurinic/apyrimidinic endonucleases. Mutation Research/Reviews in Mutation Research, 705:217–227.

[155] Kavakli IH, Baris I, Tardu M, Gül Ş, Öner H, Çal S, et al. (2017). The Photolyase/Cryptochrome family of proteins as DNA repair enzymes and transcriptional repressors. Photochemistry and Photobiology, 93:93–103.

[156] Nelson SR, Kathe SD, Hilzinger TS, Averill AM, Warshaw DM, Wallace SS, et al. (2019-04-08, 2019-4-8). Single molecule glycosylase studies with engineered 8-oxoguanine DNA damage sites show functional defects of a MUTYH polyposis variant. Nucleic Acids Research, 47:3058–3071.

[157] Raetz AG, David SS (2019-08, 2019-8). When you're strange: Unusual features of the MUTYH glycosylase and implications in cancer. DNA repair, 80:16–25.

[158] Bohm KA, Hodges AJ, Czaja W, Selvam K, Smerdon MJ, Mao P, et al. (2021-06-01, 2021-6-1). Distinct roles for RSC and SWI/SNF chromatin remodelers in genomic excision repair. Genome Research, 31:1047–1059.

[159] Gallina I, Colding C, Henriksen P, Beli P, Nakamura K, Offman J, et al. (2015-03-30, 2015-3-30). Cmr1/WDR76 defines a nuclear genotoxic stress body linking genome integrity and protein quality control. Nature Communications, 6:6533.

[160] Choi D-H, Kwon S-H, Kim J-H, Bae S-H (2012). Saccharomyces cerevisiae Cmr1 protein preferentially binds to UV-damaged DNA in vitro. Journal of Microbiology (Seoul, Korea), 50:112–118.

[161] SenGupta T, Palikaras K, Esbensen YQ, Konstantinidis G, Galindo FJN, Achanta K, et al. (2021). Base excision repair causes age-dependent accumulation of single-stranded DNA breaks that contribute to Parkinson disease pathology. Cell Reports. doi: 10.1016/j.celrep.2021.109668.

[162] Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, et al. (2021-01-08, 2021-1-8). Pfam: The protein families database in 2021. Nucleic Acids Research, 49:D412–D419.

[163] Huang J, Zhang Q, He Y, Liu W, Xu Y, Liu K, et al. (2021-08-15, 2021-8-15). Genome-wide identification, expansion mechanism and expression profiling analysis of GLABROUS1 enhancer-binding protein (GeBP) gene family in gramineae crops. International Journal of Molecular Sciences, 22:8758.

[164] Tianqiao S, Xiong Z, You Z, Dong L, Jiaoling Y, Junjie Y, et al. (2021-11-01, 2021-11-1). Genome-wide identification of Zn2Cys6 class fungal-specific transcription factors (ZnFTFs) and functional analysis of UvZnFTF1 in ustilaginoidea virens. Rice Science, 28:567–578.

[165] Parsons JL, Dianov GL (2012). In vitro base excision repair using mammalian cell extracts. Methods Mol Biol, 920:245–262.

[166] Squillaro T, Finicelli M, Alessio N, Del Gaudio S, Di Bernardo G, Melone MAB, et al. (2019). A rapid, safe, and quantitative in vitro assay for measurement of uracil-DNA glycosylase activity. J Mol Med (Berl), 97:991–1001.

[167] Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, et al. (2005). BioMart and Bioconductor: A powerful link between biological databases and microarray data analysis. Bioinformatics (Oxford, England), 21:3439–3440.

[168] Durinck S, Spellman PT, Birney E, Huber W (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nature Protocols, 4:1184–1191.

[169] Kanehisa M (2019). Toward understanding the origin and evolution of cellular organisms. Protein Science: A Publication of the Protein Society, 28:1947–1951.

[170] Csárdi G, Nepusz T (2006). The igraph software package for complex network research.

[171] Tyner S, Briatte F, Heike H (2017-06-01, 2017-6-1). Network visualization with Ggplot2. R Journal, 9:27–59.

[172] Sievers F, Higgins DG (2018-01, 2018-1). Clustal Omega for making accurate alignments of many protein sequences. Protein Science: A Publication of the Protein Society, 27:135–145.

[173] Paradis E, Schliep K (2019-02-01, 2019-2-1). Ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics, 35:526–528.

[174] Jones DT, Taylor WR, Thornton JM (1992). The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci, 8:275–282.

[175] Maldonado E, Almeida D, Escalona T, Khan I, Vasconcelos V, Antunes A (2016). LMAP: Lightweight multigene analyses in PAML. BMC bioinformatics, 17:1–11.

[176] Bast J, Parker DJ, Dumas Z, Jalvingh KM, Tran Van P, Jaron KS, et al. (2018). Consequences of asexuality in natural populations: Insights from stick insects. Molecular biology and evolution, 35:1668–1677.

[177] Chuang JH, Li H (2004). Functional bias and spatial organization of genes in mutational hot and cold regions in the human genome. PLoS biology, 2:e29.

[178] Clark AG, Glanowski S, Nielsen R, Thomas PD, Kejariwal A, Todd MA, et al. (2003). Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. Science, 302:1960–1963.

[179] Egan A, Mahurkar A, Crabtree J, Badger JH, Carlton JM, Silva JC (2008). IDEA: Interactive display for evolutionary analyses. BMC bioinformatics, 9:1–9.

[180] Fedorova ND, Khaldi N, Joardar VS, Maiti R, Amedeo P, Anderson MJ, et al. (2008). Genomic islands in the pathogenic filamentous fungus Aspergillus fumigatus. PLoS genetics, 4:e1000046.

[181] Felsenstein J, Felenstein J Inferring phylogenies. Sinauer associates Sunderland, MA; 2004.

[182] Forni G, Ruggieri AA, Piccinini G, Luchetti A (2021). BASE: A novel workflow to integrate nonubiquitous genes in comparative genomics analyses for selection. Ecology and Evolution, 11:13029–13035.

[183] Glover N, Dessimoz C, Ebersberger I, Forslund SK, Gabaldón T, Huerta-Cepas J, et al. (2019). Advances and applications in the quest for orthologs. Molecular biology and evolution, 36:2157–2164.

[184] Li C, Zhang Y, Li J, Kong L, Hu H, Pan H, et al. (2014). Two Antarctic penguin genomes reveal insights into their evolutionary history and molecular changes related to the Antarctic environment. GigaScience, 3:2047–217X.

[185] Liu A, He F, Shen L, Liu R, Wang Z, Zhou J (2019). Convergent degeneration of olfactory receptor gene repertoires in marine mammals. BMC genomics, 20:1–14.

[186] Pan D, Zhang S, Jiang J, Jiang L, Zhang Q, Liu J (2013). Genome-wide detection of selective signature in Chinese Holstein. PloS one, 8:e60440.

[187] Parker J, Tsagkogeorga G, Cotton JA, Liu Y, Provero P, Stupka E, et al. (2013). Genome-wide signatures of convergent evolution in echolocating mammals. Nature, 502:228–231.

[188] Policarpo M, Fumey J, Lafargeas P, Naquin D, Thermes C, Naville M, et al. (2021). Contrasting gene decay in subterranean vertebrates: Insights from cavefishes and fossorial mammals. Molecular biology and evolution, 38:589–605.

[189] Sánchez R, Serra F, Tárraga J, Medina I, Carbonell J, Pulido L, et al. (2011). Phylemon 2.0: A suite of web-tools for molecular evolution, phylogenetics, phylogenomics and hypotheses testing. Nucleic acids research, 39:W470–W474.

[190] Stark A, Lin MF, Kheradpour P, Pedersen JS, Parts L, Carlson JW, et al. (2007). Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. Nature, 450:219–232.

[191] Koepfli K-P, Paten B, Scientists of G10KC, O'Brien SJ (2015). The genome 10K project: A way forward. Annu Rev Anim Biosci, 3:57–111.

[192] Yang Z (2007). PAML 4: Phylogenetic analysis by maximum likelihood. Molecular biology and evolution, 24:1586–1591.

[193] Gharib WH, Robinson-Rechavi M (2013). The branch-site test of positive selection is surprisingly robust but lacks power under synonymous substitution saturation and variation in GC. Molecular biology and evolution, 30:1675–1686.

[194] Macías LG, Barrio E, Toft C (2020). GWideCodeML: A python package for testing evolutionary hypotheses at the genome-wide level. G3: Genes, Genomes, Genetics, 10:4369–4372.

[195] Yang Z, Nielsen R, Goldman N (2009). In defense of statistical methods for detecting positive selection. Proceedings of the National Academy of Sciences, 106:E95–E95.

[196] Zhai W, Nielsen R, Goldman N, Yang Z (2012). Looking for Darwin in genomic sequences—validity and success of statistical methods. Molecular biology and evolution, 29:2889–2893.

[197] Maldonado E, Khan I, Philip S, Vasconcelos V, Antunes A (2013). EASER: Ensembl easy sequence retriever. Evolutionary Bioinformatics, 9:EBO–S11335.

[198] Steinway SN, Dannenfelser R, Laucius CD, Hayes JE, Nayak S (2010). JCoDA: A tool for detecting evolutionary selection. BMC bioinformatics, 11:1–9.

[199] Lord E, Leclercq M, Boc A, Diallo AB, Makarenkov V (2012). Armadillo 1.1: An original workflow platform for designing and conducting phylogenetic analysis and simulations. PloS one, 7:e29903.

[200] Xu B, Yang Z (2013). PAMLX: A graphical user interface for PAML. Molecular biology and evolution, 30:2723–2724.

[201] Maldonado E, Sunagar K, Almeida D, Vasconcelos V, Antunes A (2014). IMPACT$_S$: Integrated multiprogram platform to analyze and combine tests of selection. PloS one, 9:e96243.

[202] Su F, Ou H-Y, Tao F, Tang H, Xu P (2013). PSP: Rapid identification of orthologous coding genes under positive selection across multiple closely related prokaryotic genomes. BMC genomics, 14:1–10.

[203] Busset J, Cabau C, Meslin C, Pascal G (2011). PhyleasProg: A user-oriented web server for wide evolutionary analyses. Nucleic acids research, 39:W479–W485.

[204] Stern A, Doron-Faigenboim A, Erez E, Martz E, Bacharach E, Pupko T (2007). Selecton 2007: Advanced models for detecting positive and purifying selection using a Bayesian inference approach. Nucleic acids research, 35:W506–W511.

[205] Webb AE, Walsh TA, O'Connell MJ (2017). VESPA: Very large-scale evolutionary and selective pressure analyses. PeerJ Computer Science, 3:e118.

[206] Hongo JA, de Castro GM, Cintra LC, Zerlotini A, Lobo FP (2015). POTION: An end-to-end pipeline for positive Darwinian selection detection in genome-scale data through phylogenetic comparison of protein-coding genes. BMC genomics, 16:1–16.

[207] Lechner M, Findeiß S, Steiner L, Marz M, Stadler PF, Prohaska SJ (2011). Proteinortho: Detection of (co-) orthologs in large-scale analysis. BMC bioinformatics, 12:1–9.

[208] Esteves P, Abrantes J, Carneiro M, Müller A, Thompson G, Van der Loo W (2008). Detection of positive selection in the major capsid protein VP60 of the rabbit haemorrhagic disease virus (RHDV). Virus Research, 137:253–256.

[209] Löytynoja A (2014). Phylogeny-aware alignment with PRANK. Multiple sequence alignment methods, 155–170.

[210] Suyama M, Torrents D, Bork P (2006). PAL2NAL: Robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic acids research, 34:W609–W612.

[211] Anisimova M, Yang Z (2007). Multiple hypothesis testing to detect lineages under positive selection that affects only a few sites. Molecular Biology and Evolution, 24:1219–1228.

[212] Ceron-Noriega A, Almeida MV, Levin M, Butter F (2023). Nematode gene annotation by machine-learning-assisted proteotranscriptomics enables proteome-wide evolutionary analysis. Genome Research. doi: 10.1101/gr.277070.122.

[213] Ma F, Fuqua BK, Hasin Y, Yukhtman C, Vulpe CD, Lusis AJ, et al. (2019). A comparison between whole transcript and 3’ RNA sequencing methods using Kapa and Lexogen library preparation methods. BMC Genomics, 20:9.

[214] Tandonnet S, Torres TT (2017). Traditional versus 3′ RNA-seq in a non-model species. Genomics Data, 11:9–16.

[215] Moll P, Ante M, Seitz A, Reda T (2014). QuantSeq 3′ mRNA sequencing for RNA quantification. Nat Methods, 11:i–iii.

[216] Mehta D, Scandola S, Uhrig RG (2022). BoxCar and Library-Free Data-Independent Acquisition Substantially Improve the Depth, Range, and Completeness of Label-Free Quantitative Proteomics. Anal Chem, 94:793–802.

[217] Ishikawa M, Konno R, Nakajima D, Gotoh M, Fukasawa K, Sato H, et al. (2022). Optimization of Ultrafast Proteomics Using an LC-Quadrupole-Orbitrap Mass Spectrometer with Data-Independent Acquisition. J Proteome Res, 21:2085–2093.

[218] Meier F, Geyer PE, Virreira Winter S, Cox J, Mann M (2018). BoxCar acquisition method enables single-shot proteomics at a depth of 10,000 proteins in 100 minutes. Nat Methods, 15:440–448.

[219] Howard-Till RA, Lukaszewicz A, Loidl J (2011). The Recombinases Rad51 and Dmc1 Play Distinct Roles in DNA Break Repair and Recombination Partner Choice in the Meiosis of Tetrahymena. PLoS Genet, 7:e1001359.

[220] Wang L, Xue Y, Yang S, Bo T, Xu J, Wang W (2023). Mismatch Repair Protein Msh2 Is Necessary for Macronuclear Stability and Micronuclear Division in Tetrahymena thermophila. IJMS, 24:10559.

[221] O'Hagan S, Wright Muelas M, Day PJ, Lundberg E, Kell DB (2018). GeneGini: Assessment via the Gini Coefficient of Reference "Housekeeping" Genes and Diverse Human Transporter Expression Profiles. Cell Syst, 6:230–244.e1.

[222] Park Y-S, Chon T-S, Bae M-J, Kim D-H, Lek S (2018). Multivariate Data Analysis by Means of Self-Organizing Maps. In: Recknagel F, Michener WK, editors Ecological Informatics: Data Management and Knowledge Discovery. Cham: Springer International Publishing, 251–272.

[223] Ceron-Noriega A, Schoonenberg VAC, Butter F, Levin M (2023). AlexandrusPS: A User-Friendly Pipeline for the Automated Detection of Orthologous Gene Clusters and Subsequent Positive Selection Analysis. Genome Biology and Evolution, 15:evad187.

Acknowledgements

A

Supplemental information Chapter 2

A.1. Supplemental tables

Table A.1.: DNA damage treatments for *Tetrahymena*. BER: base excision repair; DSBR: double strand break repair; NER: nucleotide excision repair; ICL: Interstrand crosslink repair

Treatment	Induced Repair Pathway	Treatment Concentration	Basis of concentration	Stock Preparation
Hydrogen peroxide (HP)	BER	0.66 mM	Tested EC50s	Purchased at 9.8 M (Carl-Roth)
Methyl methanesulfonate (MMS)	BER;DSBR	2.38 mM	Tested EC50s	Purchase at 11.8 M (Sigma)
Ultraviolet light (UV)	NER; ICL	100 J/m^2	Tested EC50s	UV-Crosslinker Cells treated in 10 mM Tris-HCl (pH=7.5)
Cisplatin (CP)	ICL; NER	100 ug/mL	Tested EC50s/ Loidl and Mochizuki, 2009	2 mg/ml in DMSO (Sigma)

Treatment	Induced Repair Pathway	Treatment Concentration	Basis of concentration	Stock Preparation
Hydroxyurea (HU)	ICL	20 mM	Sandoval et al., 2015	1.5 M in water (Sigma)
Ionizing radiation (IR)	DSBR	5000 rads (equivalent to 50 Grays)	Loidl and Mochizuki, 2009	Used Faxitron CellRad, Cells treated in 10 mM Tris-HCl (pH=7.5)

Table A.2.: Curated list of DNA damage genes. Inlcuding human description/gene, yeast standard ID and name, *Tetrahymena* gene and ID, and pathway classification(s). BER: base excision repair; DSBR: double strand break repair; NER: nucleotide excision repair; ICL: Interstrand crosslink repair; DDR: DNA damage response; MMR: Mismatch repair

Human Gene	Yeast Gene	Yeast Standard	*Tetrahymena* Standard	TTHERM	Repair Pathway
MSH6	MSH6	YDR097C	MSH3	00426230	MMR
XPG	EXO1	YOR033C	NA	00773520	DSBR
XPG	EXO1	YOR033C	NA	00773520	MMR
XPG	EXO1	YOR033C	NA	00773520	NER
down stream mediators	RAD53	YPL153C	NA	00000020	DDR
ERCC1	RAD10	YML095C	RAD10	00011650	NER
RAD23	RAD23	YEL037C	RAD23	00013290	NER
Mec1/ATR	DNA2	YHR164C	NA	000136019	DDR
XPF1	RAD1	YPL022W	XPF1	000160559	NER
DSIF	SPT5	YML010W	SPT5	00028580	NER
KU80	YKU80	YMR106C	TPB1	000309879	DSBR
NA	RAD5	YLR032W	RAD5	00037210	ICL
NA	SNM1/ PSO2	YMR137C	SNM1	000697499	ICL
down stream mediators	RAD53	YPL153C	NA	00075550	DDR
down stream mediators	RAD53	YPL153C	NA	00079490	DDR
down stream mediators	RAD53	YPL153C	NA	000841299	DDR
down stream mediators	RAD53	YPL153C	NA	00106700	DDR
RPA1	RFA1	YAR007C	RFA1	00106890	DSBR
RPA1	RFA1	YAR007C	RFA1	00106890	NER

Human Gene	Yeast Gene	Yeast Standard	*Tetrahymena* Standard	TTHERM	Repair Pathway
down stream mediators	RAD53	YPL153C	NA	00106920	DDR
NA	APN2	YBL019W	NA	001080610	BER
POL Epsilon	POL2	YNL262W	POL2	00112520	BER
POL Epsilon	POL2	YNL262W	POL2	00112520	NER
down stream mediators	CHK1	YBR274W	NA	00112569	DDR
MLH1	MLH1	YMR167W	TMLH1	00127000	MMR
down stream mediators	RAD53	YPL153C	NA	00128720	DDR
down stream mediators	RAD53	YPL153C	NA	00128740	DDR
down stream mediators	CHK1	YBR274W	NA	001358411	DDR
Mec1/ATR	DNA2	YHR164C	DNA2	00136030	DDR
down stream mediators	RAD53	YPL153C	NA	001388156	DDR
MSH6	MSH6	YDR097C	NA	00142230	MMR
RAD51	RAD51	YER095W	RAD51	00142330	DSBR
down stream mediators	CHK1	YBR274W	NA	00145520	DDR
MSH6	MSH6	YDR097C	MSH6L3	00150000	MMR
down stream mediators	RAD53	YPL153C	NA	00151959	DDR
down stream mediators	RAD53	YPL153C	NA	00151970	DDR
down stream mediators	RAD53	YPL153C	NA	00151980	DDR
RFC	RFC5	YBR087W	RFC5	00161180	BER
clamp loader	RFC5	YBR087W	RFC5	00161180	DDR
RFC	RFC5	YBR087W	RFC5	00161180	MMR
RFC	RFC5	YBR087W	RFC5	00161180	NER
MSH6	MSH6	YDR097C	MSH6	00194810	MMR
DNAPKCs	NA	NA	DPK1	00203010	DSBR
down stream mediators	RAD53	YPL153C	NA	00213560	DDR
RFC	RFC3	YNL290W	RFC3	00213600	BER
clamp loader	RFC3	YNL290W	RFC3	00213600	DDR
RFC	RFC3	YNL290W	RFC3	00213600	MMR
RFC	RFC3	YNL290W	RFC3	00213600	NER
TOP3	TOP3	YLR234W	NA	00216140	DSBR
TOP3	TOP3	YLR234W	NA	00216140	MMR
RAD54	RAD54	YGL163C	RAD54	00237490	DSBR
RFC	RFC2	YJR068W	RFC2	00245150	BER
clamp loader	RFC2	YJR068W	RFC2	00245150	DDR

Human Gene	**Yeast Gene**	**Yeast Standard**	***Tetrahymena* Standard**	**TTHERM**	**Repair Pathway**
RFC	RFC2	YJR068W	RFC2	00245150	MMR
RFC	RFC2	YJR068W	RFC2	00245150	NER
down stream mediators	CHK1	YBR274W	NA	00295610	DDR
MSH2	MSH2	YOL090W	MSH2	00295920	MMR
NA	RAD5	YLR032W	RAD5L4	00298220	ICL
XPD	RAD3	YER171W	NA	00298500	NER
down stream mediators	RAD53	YPL153C	NA	00300380	DDR
down stream mediators	RAD53	YPL153C	NA	00301710	DDR
CSB	RAD16	YBR114W	NA	00313280	NER
TFIIH	TFB2	YPL122C	NA	00313290	BER
TFIIH	TFB2	YPL122C	NA	00313290	NER
XPD	RAD3	YER171W	NA	00316410	NER
LIG1	CDC9	YDL164C	LIG1	00348170	BER
LIG1	CDC9	YDL164C	LIG1	00348170	DSBR
LIG1	CDC9	YDL164C	LIG1	00348170	NER
down stream mediators	RAD53	YPL153C	NA	00355740	DDR
LIGIV	DNL4	YOR005C	LIG4	00387050	DSBR
down stream mediators	RAD53	YPL153C	NA	00389660	DDR
NA	REV3	YPL167C	NA	00391570	ICL
NA	RAD5	YLR032W	RAD16	00420480	ICL
NA	SNM1/ PSO2	YMR137C	SNML1	00433640	ICL
FEN1	RAD27	YKL113C	RAD27	00437617	BER
FEN1	RAD27	YKL113C	RAD27	00437617	DSBR
NA	REV3	YPL167C	REV3	00437650	ICL
POL Delta	POL3	YDL102W	NA	00444660	BER
POL Delta	POL3	YDL102W	NA	00444660	DSBR
POL Delta	POL3	YDL102W	NA	00444660	NER
RPB7	RPB7	YDR404C	RPB7	00446180	NER
down stream mediators	CHK1	YBR274W	NA	00449360	DDR
down stream mediators	CHK1	YBR274W	NA	00449370	DDR
down stream mediators	CHK1	YBR274W	NA	00449400	DDR
TOP3	TOP3	YLR234W	NA	00464990	DSBR
TOP3	TOP3	YLR234W	NA	00464990	MMR
TOP3	TOP3	YLR234W	NA	00465030	DSBR
TOP3	TOP3	YLR234W	NA	00465030	MMR

Human Gene	**Yeast Gene**	**Yeast Standard**	***Tetrahymena* Standard**	**TTHERM**	**Repair Pathway**
down stream mediators	RAD53	YPL153C	NA	00474550	DDR
KU80	YKU80	YMR106C	TKU80	00492460	DSBR
down stream mediators	RAD53	YPL153C	NA	00494350	DDR
down stream mediators	CHK1	YBR274W	NA	00497000	DDR
down stream mediators	CHK1	YBR274W	NA	00497250	DDR
TOP3	TOP3	YLR234W	TOP3	00497920	DSBR
TOP3	TOP3	YLR234W	TOP3	00497920	MMR
9-1-1 clamp	DDC1	YHR144C	DCD1	00498180	DDR
down stream mediators	RAD53	YPL153C	NA	00526980	DDR
down stream mediators	RAD53	YPL153C	NA	00534050	DDR
NA	RAD6	YGL058W	NA	00547960	ICL
KU70	YKU70	YMR284W	TKU71	00561799	DSBR
down stream mediators	CHK1	YBR274W	NA	00600470	DDR
MUS81/ERCC4	MUS81	YDR386W	MUS81	00624870	DSBR
POL Delta	POL3	YDL102W	POLD1	00636920	BER
POL Delta	POL3	YDL102W	POLD1	00636920	DSBR
POL Delta	POL3	YDL102W	POLD1	00636920	NER
down stream mediators	RAD53	YPL153C	NA	00637100	DDR
NA	MMS2	YGL087C	UCV2	00670590	ICL
down stream mediators	CHK1	YBR274W	NA	00670900	DDR
KU70	YKU70	YMR284W	NA	00684440	DSBR
XPD	RAD3	YER171W	RTEL1	00684490	NER
POL31	POL31	YJR006W	POL31	00691170	DSBR
down stream mediators	CHK1	YBR274W	NA	00715940	DDR
apical kinases	MRE11	YMR224C	MRE11	00721450	DDR
MRE11	MRE11	YMR224C	MRE11	00721450	DSBR
RPA1	RFA1	YAR007C	RLP1	00726370	DSBR
RPA1	RFA1	YAR007C	RLP1	00726370	NER
PARP1	NA	NA	PARP1	00726460	DSBR
POL4 (pol lambda)	POL4	YCR014C	POLB1	00732550	DSBR
RFC	RFC1	YOR217W	NA	00762900	BER
RFC	RFC1	YOR217W	NA	00762900	MMR
RFC	RFC1	YOR217W	NA	00762900	NER
MSH5/MutS	MSH5	YDL154W	MSH5	00763040	MMR

Human Gene	Yeast Gene	Yeast Standard	*Tetrahymena* Standard	TTHERM	Repair Pathway
apical kinases	RAD50	YNL250W	RAD50	00773790	DDR
RAD50	RAD50	YNL250W	RAD50	00773790	DSBR
NA	RAD18	YCR066W	NA	00780660	ICL
RFC	RFC4	YOL094C	RFC4	00780750	BER
clamp loader	RFC4	YOL094C	RFC4	00780750	DDR
RFC	RFC4	YOL094C	RFC4	00780750	MMR
RFC	RFC4	YOL094C	RFC4	00780750	NER
UNG	UNG1	YML021C	UNG1	00794250	BER
NA	APN2	YBL019W	APN2	00794600	BER
down stream mediators	RAD53	YPL153C	NA	00815090	DDR
XPC	RAD4	YER162C	RAD4	00825460	NER
MET19	MET18	YIL128W	NA	00829370	NER
down stream mediators	RAD53	YPL153C	NA	00852690	DDR
MSH4/MutS	MSH4	YFL003C	MSH4	00857890	MMR
LIG1	CDC9	YDL164C	NA	00865240	BER
LIG1	CDC9	YDL164C	NA	00865240	DSBR
LIG1	CDC9	YDL164C	NA	00865240	NER
down stream mediators	CHK1	YBR274W	NA	00923220	DDR
CSB	RAD16	YBR114W	NA	00933250	NER
down stream mediators	RAD53	YPL153C	NA	00935420	DDR
RFC	RFC1	YOR217W	RFC1	00939110	BER
RFC	RFC1	YOR217W	RFC1	00939110	MMR
RFC	RFC1	YOR217W	RFC1	00939110	NER
down stream mediators	RAD53	YPL153C	NA	00976420	DDR
apical kinases	MEC1	YBR136W	ATR1	01008650	DDR
down stream mediators	RAD53	YPL153C	NA	01018400	DDR
down stream mediators	RAD53	YPL153C	NA	01026430	DDR
RecQ	SGS1	YMR190C	SGS1	01030000	DSBR
RecQL, RecQ4, RecQ5, BLM, WRN	SGS1	YMR190C	SGS1	01030000	MMR
down stream mediators	RAD53	YPL153C	NA	01044420	DDR
down stream mediators	RAD53	YPL153C	NA	01049260	DDR
NTH1	NTG1	YAL015C	NTG1	01106120	BER
PCNA	POL30	YBR088C	PCNA1	01107420	BER
PCNA	PCNA	YBR088C	PCNA1	01107420	MMR

Human Gene	**Yeast Gene**	**Yeast Standard**	***Tetrahymena* Standard**	**TTHERM**	**Repair Pathway**
PCNA	PCNA	YBR088C	PCNA1	01107420	NER
PMS2	PMS1	YNL082W	PMS2	01109940	MMR
NA	UBC13	YDR092W	UCN2	01123950	ICL
RAD51	RAD51	YER095W	NA	01143840	DSBR
XPG	EXO1	YOR033C	EXO1	01179960	DSBR
XPG	EXO1	YOR033C	EXO1	01179960	MMR
XPG	EXO1	YOR033C	EXO1	01179960	NER
down stream mediators	RAD53	YPL153C	NA	01205310	DDR
OGG1	OGG1	YML060W	OGL1	01243450	BER
down stream mediators	RAD53	YPL153C	NA	01246670	DDR
down stream mediators	RAD53	YPL153C	NA	01246690	DDR
TOP3	TOP3	YLR234W	NA	01262890	DSBR
TOP3	TOP3	YLR234W	NA	01262890	MMR
down stream mediators	CHK1	YBR274W	NA	01358400	DDR
down stream mediators	RAD53	YPL153C	NA	01376880	DDR
down stream mediators	RAD53	YPL153C	NA	01494740	DDR
ELOC	ELC1	YPL046C	NA	01658030	NER

A.2. Supplemental figures

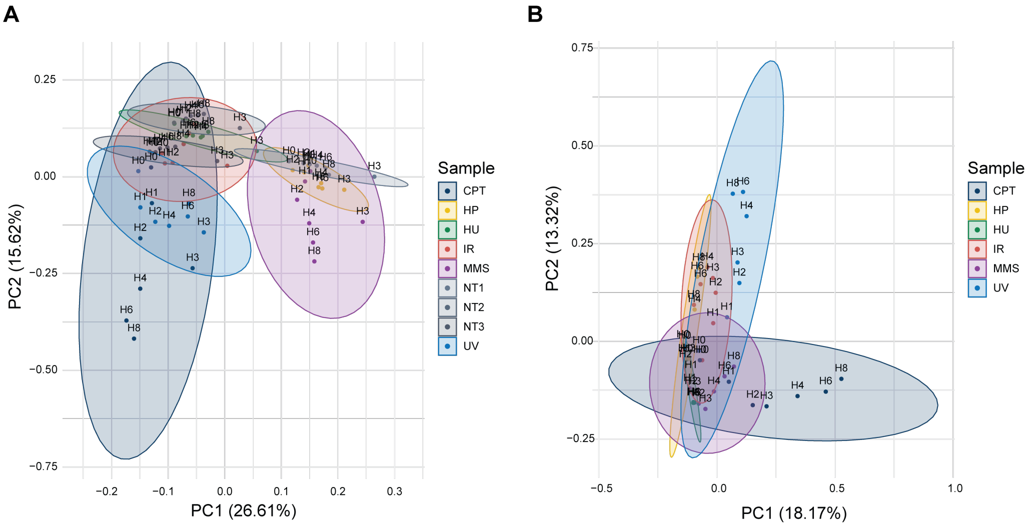


Figure A.1.: Principal component analysis (PCA) of LFQ protein data, before (A) and after (B) normalization. Sets measured together (A), were i) MMS, HP, NT1; ii) UV, CPT, NT2; iii) IR, HU, NT3.

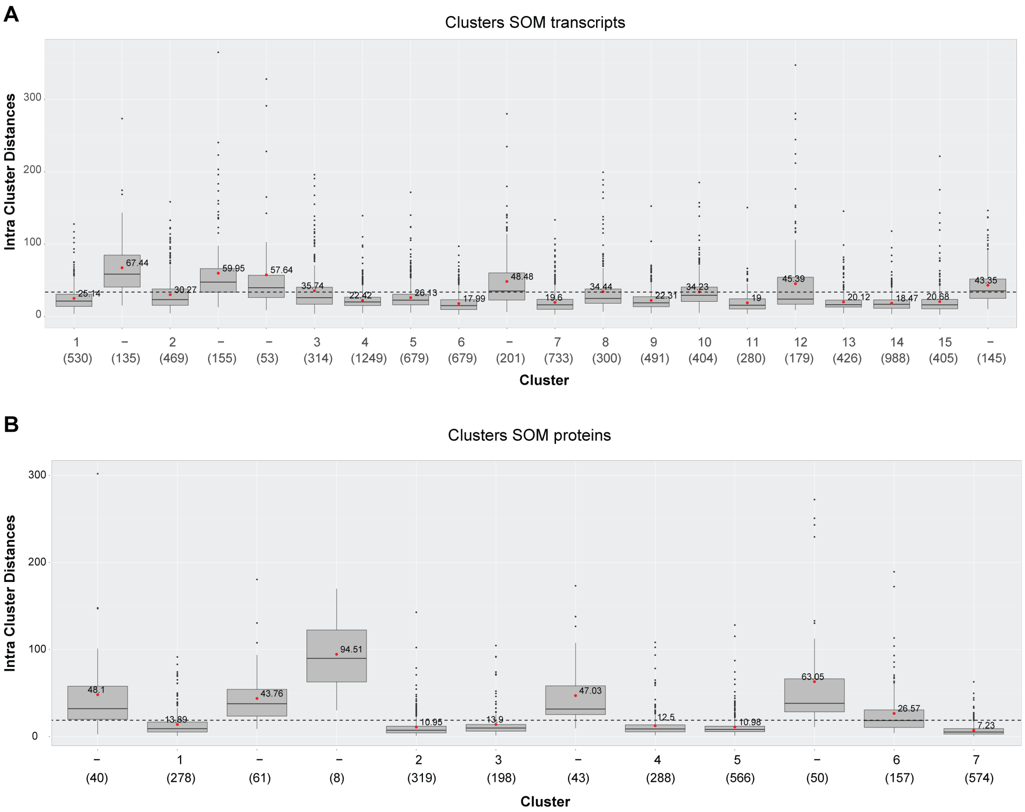


Figure A.2.: Boxplots showing the intradistance (distance between trends of proteins (A) or transcripts (B) clustered together) for each self-organzing map cluster. Clusters with a large median intradistance were excluded from further analysis (cutoff presented by dotted line). Mean values are written and presented for each cluster with a red dot

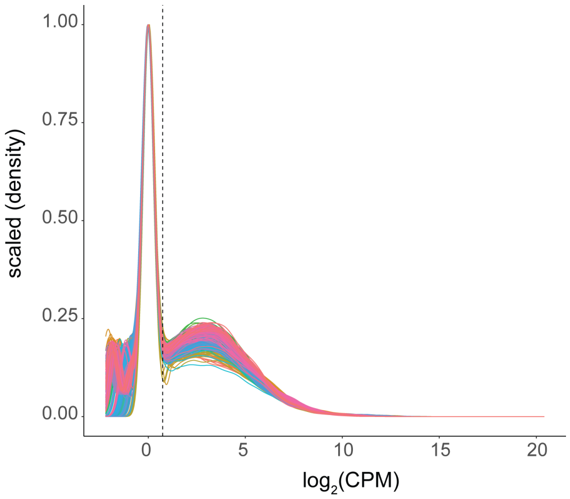


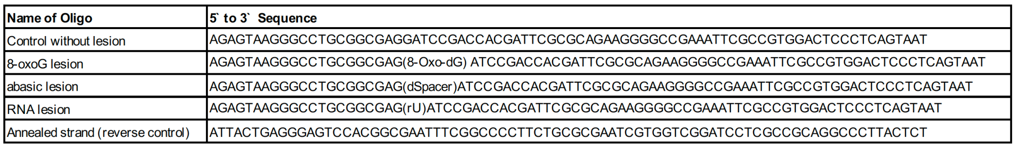
Figure A.3.: Density of the log2(CPM) values of each RNA sample measured. Transcripts that did not have any CPM value across the complete time series and in any of the treatments below the 25th quantile of all CPM values(CPM < 1.673028, dotted line) were removed

B

Supplemental information Chapter 3

B.1. Supplemental tables

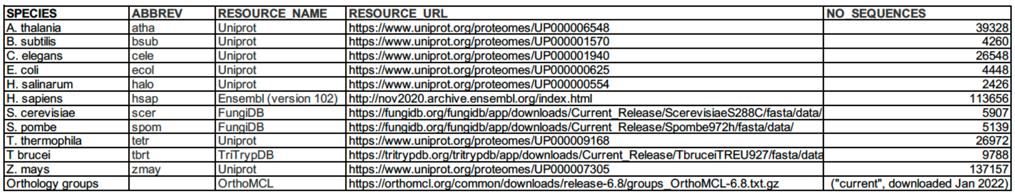
Table B.1.: Oligo sequences used in this study.



Name of Oligo	5` to 3` Sequence
Control without lesion	AGAGTAAGGGCCTGCGGCGAGGATCCGACCACGATTCGCGCAGAAGGGGCCGAAATTCGCCGTGGACTCCCTCAGTAAT
8-oxoG lesion	AGAGTAAGGGCCTGCGGCGAG(8-Oxo-dG) ATCCGACCACGATTCGCGCAGAAGGGGCCGAAATTCGCCGTGGACTCCCTCAGTAAT
abasic lesion	AGAGTAAGGGCCTGCGGCGAG(dSpacer)ATCCGACCACGATTCGCGCAGAAGGGGCCGAAATTCGCCGTGGACTCCCTCAGTAAT
RNA lesion	AGAGTAAGGGCCTGCGGCGAG(rU)ATCCGACCACGATTCGCGCAGAAGGGGCCGAAATTCGCCGTGGACTCCCTCAGTAAT
Annealed strand (reverse control)	ATTACTGAGGGAGTCCACGGCGAATTTCGGCCCCTTCTGCGCGAATCGTGGTCGGATCCTCGCCGCAGGCCCTTACTCT

Table B.2.: Identified orthology groups (as per OrthoMCL for species in this study.) \
See https://doi.org/10.1016/j.isci.2023.106778

Table B.3.: Overview of the databases used (in Supplemental Information).



SPECIES	ABBREV	RESOURCE_NAME	RESOURCE_URL	NO_SEQUENCES
A. thalania	atha	Uniprot	https://www.uniprot.org/proteomes/UP000006548	39328
B. subtilis	bsub	Uniprot	https://www.uniprot.org/proteomes/UP000001570	4260
C. elegans	cele	Uniprot	https://www.uniprot.org/proteomes/UP000001940	26548
E. coli	ecol	Uniprot	https://www.uniprot.org/proteomes/UP000000625	4448
H. salinarum	halo	Uniprot	https://www.uniprot.org/proteomes/UP000000554	2426
H. sapiens	hsap	Ensembl (version 102)	http://nov2020.archive.ensembl.org/index.html	113656
S. cerevisiae	scer	FungiDB	https://fungidb.org/fungidb/app/downloads/Current_Release/ScerevisiaeS288C/fasta/data/	5907
S. pombe	spom	FungiDB	https://fungidb.org/fungidb/app/downloads/Current_Release/Spombe972h/fasta/data/	5139
T. thermophila	tetr	Uniprot	https://www.uniprot.org/proteomes/UP000009168	26972
T brucei	tbrt	TriTrypDB	https://tritrypdb.org/tritrypdb/app/downloads/Current_Release/TbruceiTREU927/fasta/data	9788
Z. mays	zmay	Uniprot	https://www.uniprot.org/proteomes/UP000007305	137157
Orthology groups		OrthoMCL	https://orthomcl.org/common/downloads/release-6.8/groups_OrthoMCL-6.8.txt.gz	("current", downloaded Jan 2022)

Table B.4.: Complete dataset of identified and enriched proteins, including calculated values.
See https://doi.org/10.1016/j.isci.2023.106778

Table B.5.: Enrichment values for KEGG and GO biological processes.
See https://doi.org/10.1016/j.isci.2023.10677

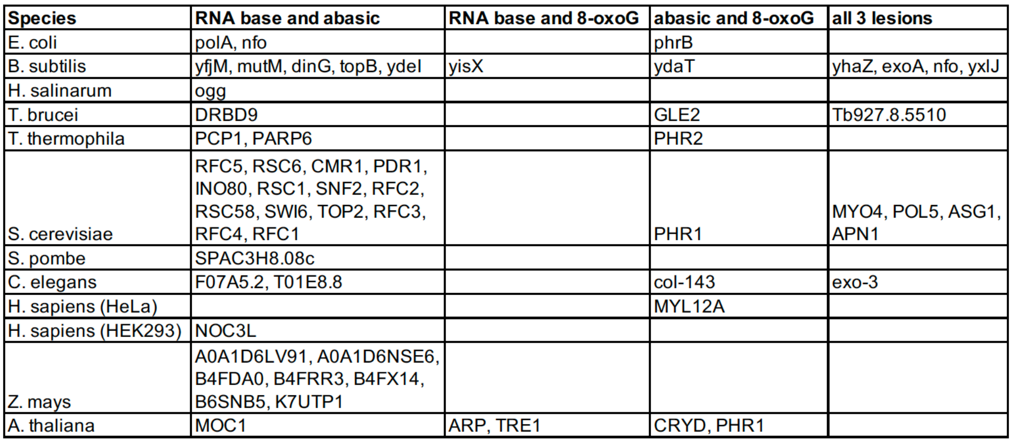
Table B.6.: Enriched proteins included in the STRING database.
See https://doi.org/10.1016/j.isci.2023.10677

Table B.7.: STRING interaction score of combined networks.
See https://doi.org/10.1016/j.isci.2023.106778

Table B.8.: Network edges in all combined networks.
See https://doi.org/10.1016/j.isci.2023.106778

Table B.9.: Network nodes in all combined networks.
See https://doi.org/10.1016/j.isci.2023.106778

Table B.10.: Overlap of interaction partners across species, per lesion.



Species	RNA base and abasic	RNA base and 8-oxoG	abasic and 8-oxoG	all 3 lesions
E. coli	polA, nfo		phrB	
B. subtilis	yfjM, mutM, dinG, topB, ydeI	yisX	ydaT	yhaZ, exoA, nfo, yxlJ
H. salinarum	ogg			
T. brucei	DRBD9		GLE2	Tb927.8.5510
T. thermophila	PCP1, PARP6		PHR2	
S. cerevisiae	RFC5, RSC6, CMR1, PDR1, INO80, RSC1, SNF2, RFC2, RSC58, SWI6, TOP2, RFC3, RFC4, RFC1		PHR1	MYO4, POL5, ASG1, APN1
S. pombe	SPAC3H8.08c			
C. elegans	F07A5.2, T01E8.8		col-143	exo-3
H. sapiens (HeLa)			MYL12A	
H. sapiens (HEK293)	NOC3L			
Z. mays	A0A1D6LV91, A0A1D6NSE6, B4FDA0, B4FRR3, B4FX14, B6SNB5, K7UTP1			
A. thaliana	MOC1	ARP, TRE1	CRYD, PHR1	

Table B.11.: Overlap of interaction partners across lesions (in Supplemental Information).
See https://doi.org/10.1016/j.isci.2023.106778

Table B.12.: Pfam domains of non-DNA repair enriched proteins and their categorization.
See https://doi.org/10.1016/j.isci.2023.106778

B.2. Supplemental figures

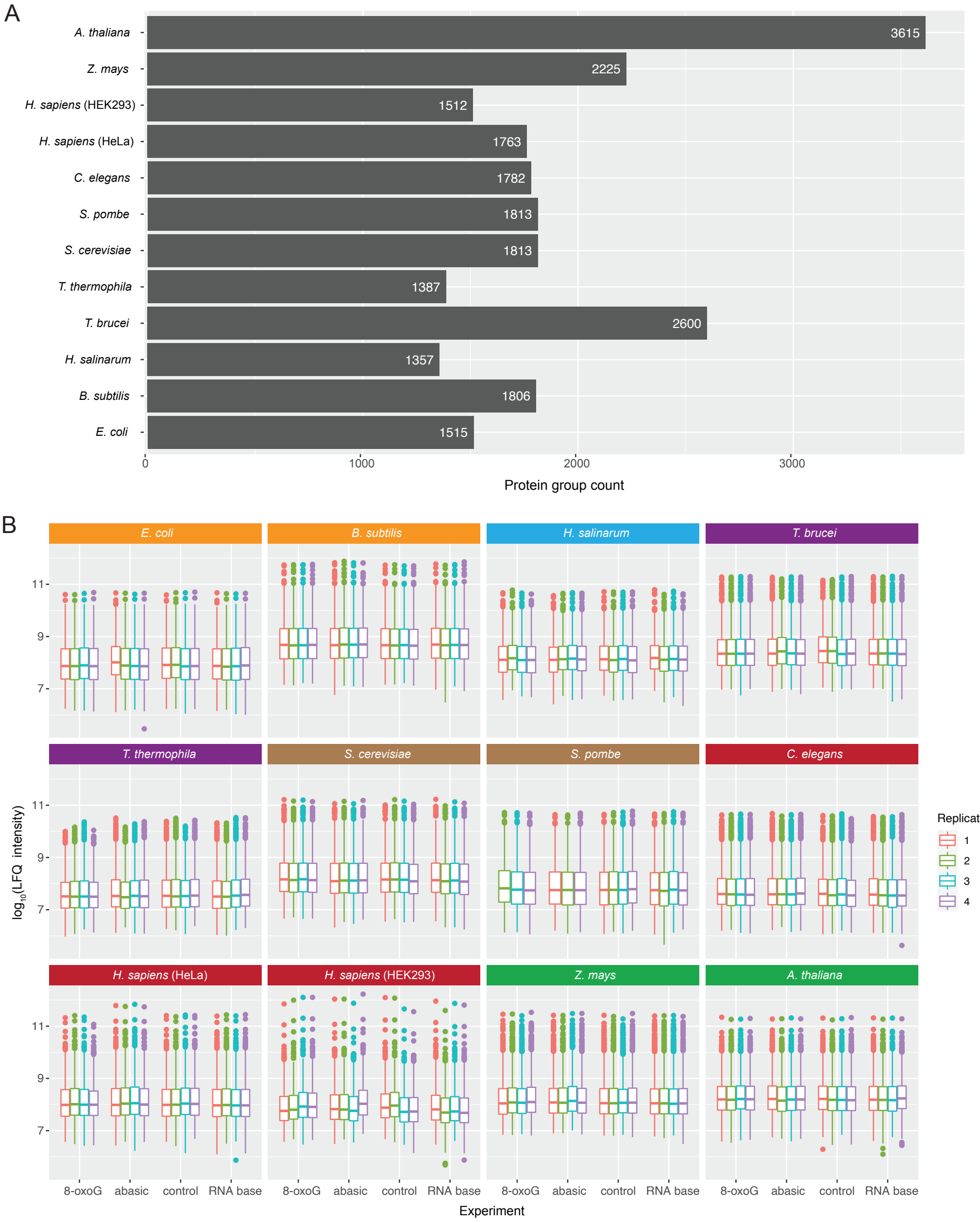


Figure B.1.: A) Barplot showing the total amount of identified, quantifiable proteins per species (see methods). **B)** Boxplot showing log10 LFQ intensity per replicate (y-axis) and experiment (x-axis) of each included species.

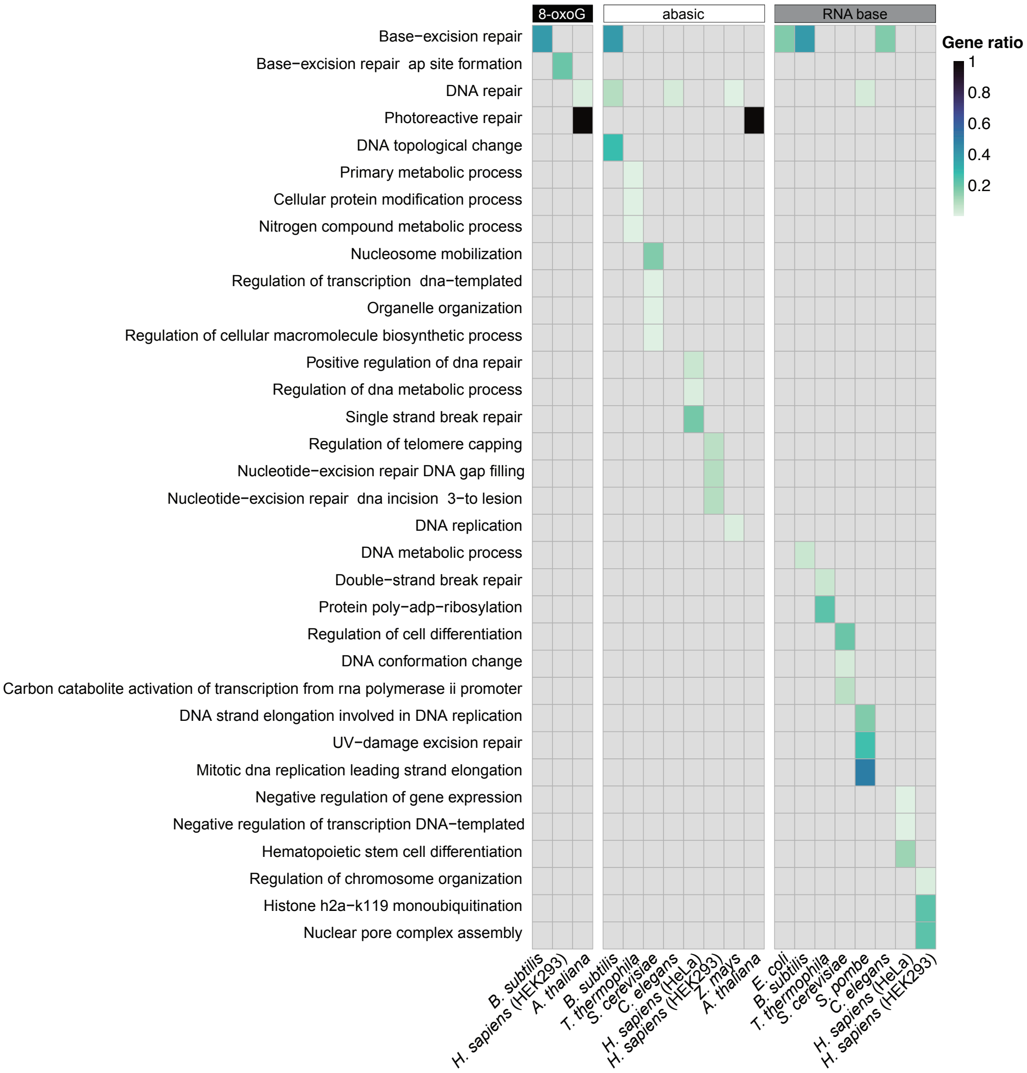


Figure B.2.: Overrepresentation of GO terms biological processes among enriched proteins at each lesion across species. Conditions with no enriched GO terms are not shown, or presented in gray. ‘Gene ratio’ refers to genes in the dataset (enriched proteins at lesion) over genes in the background (whole genome).

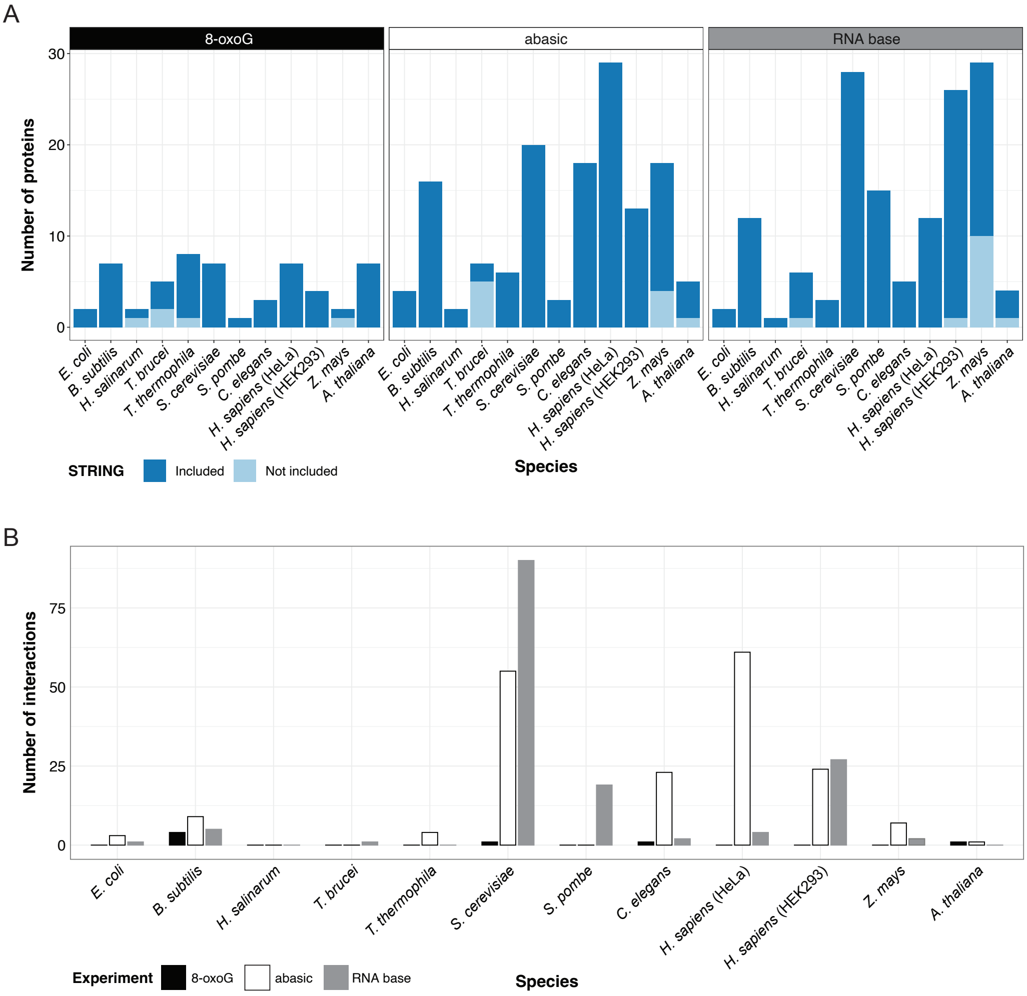


Figure B.3.: A) Barplot of enriched proteins per lesion and species that are included in the STRING database. **B)** Bar plot of number of interactions per lesion and species.

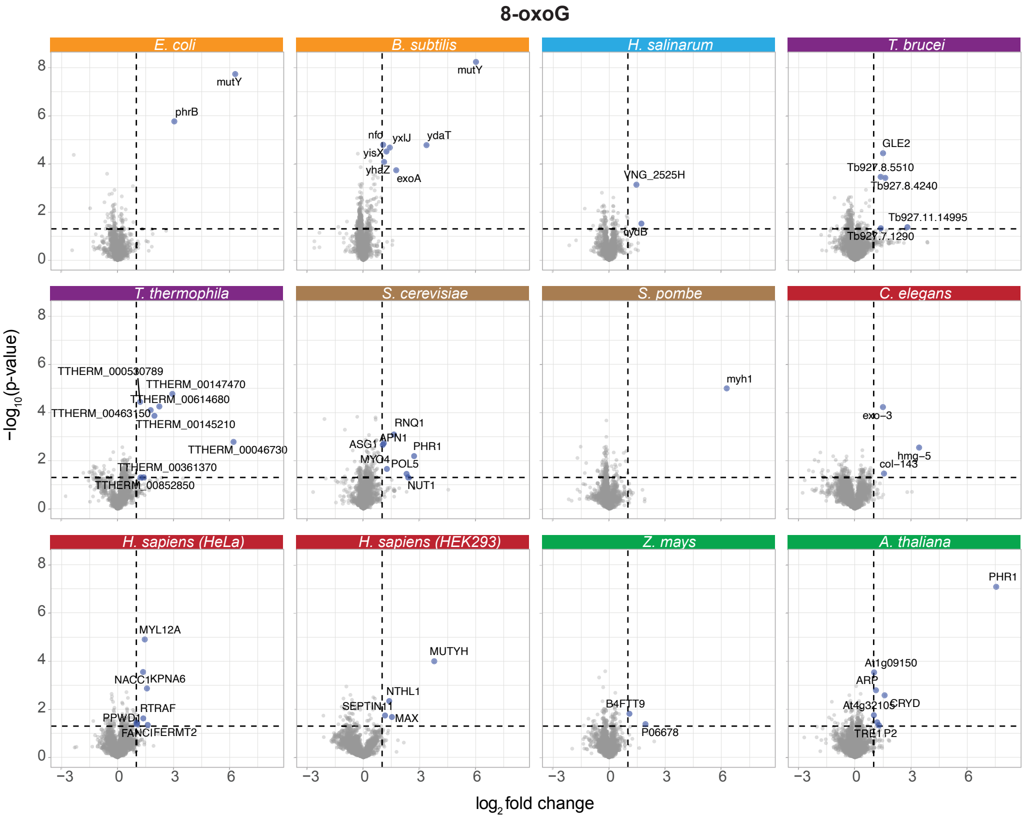


Figure B.4.: Volcano plots showing the fold change (x-axis) and p-value (y-axis) of proteins binding to the 8-oxoG lesion compared to the control for each species. Proteins with log2 fold change > 1 and p-value < 0.05 are highlighted and labeled.

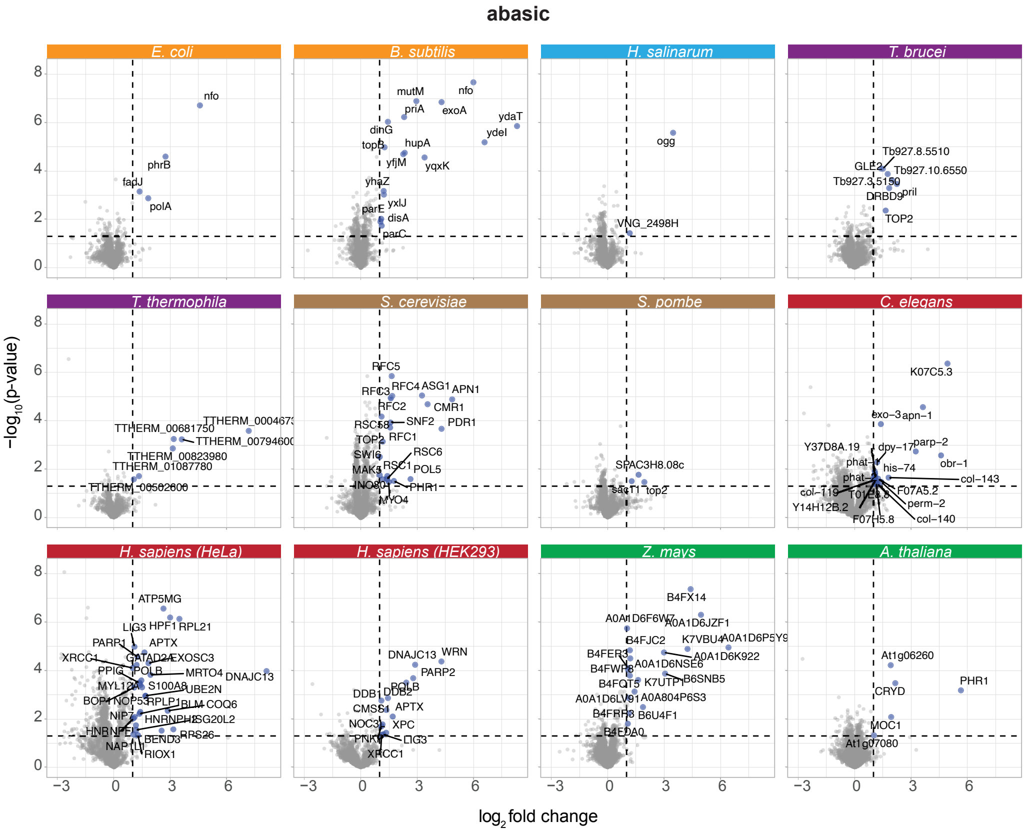


Figure B.5.: Volcano plots showing the fold change (x-axis) and p-value (y-axis) of proteins binding to the abasic lesion compared to the control for each species. Proteins with log2 fold change > 1 and p-value < 0.05 are highlighted and labeled.

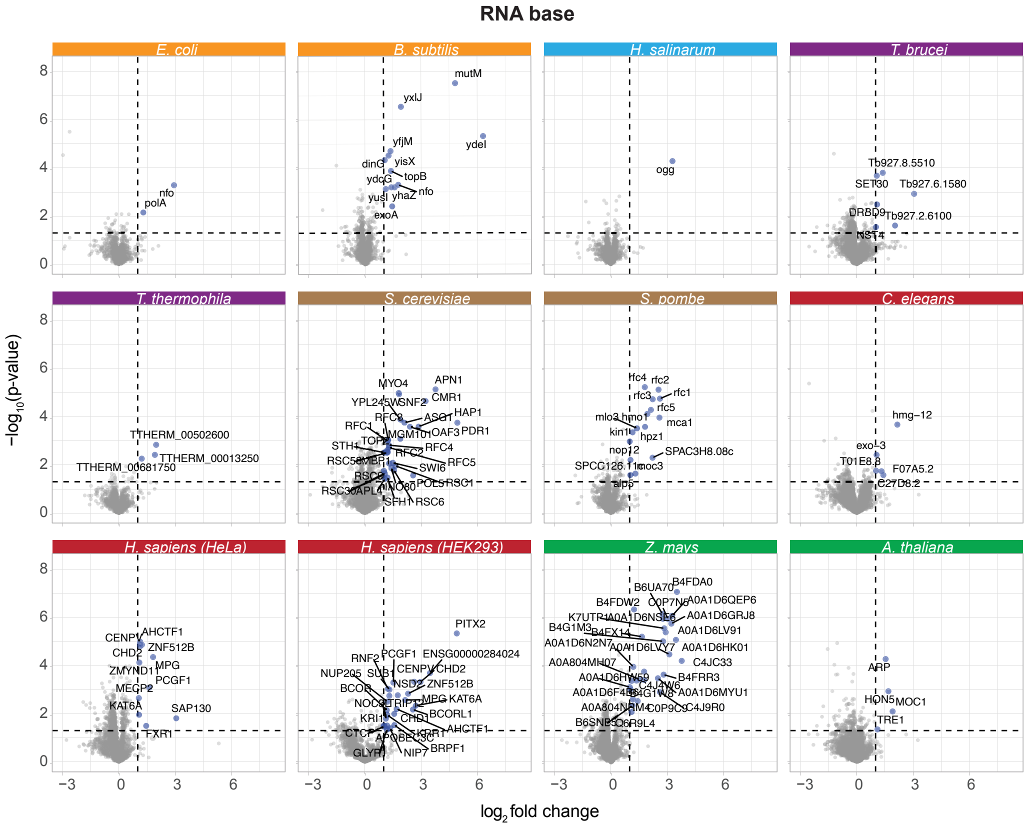
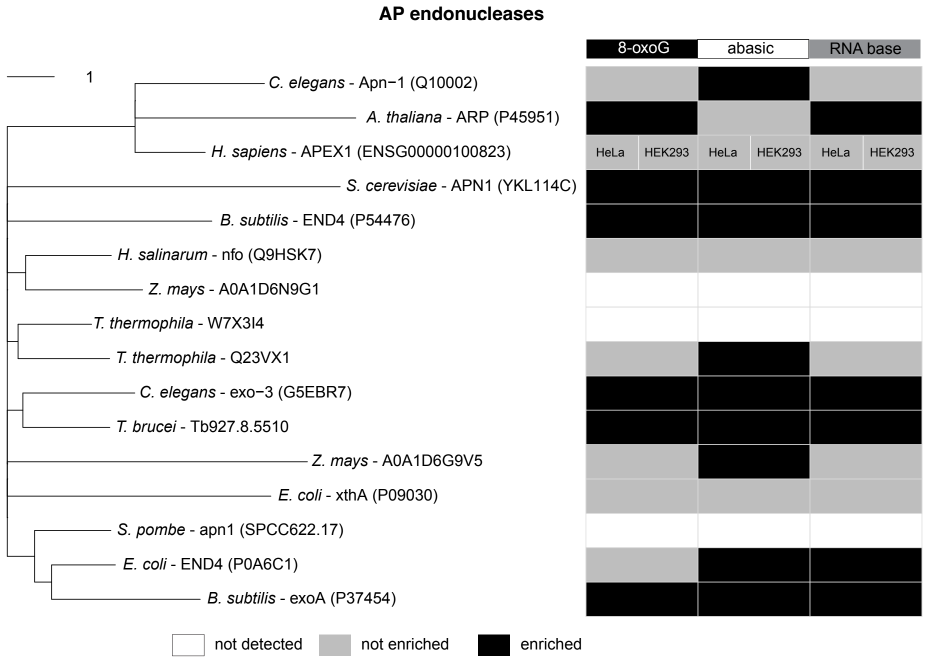


Figure B.6.: Volcano plots showing the fold change (x-axis) and p-value (y-axis) of proteins binding to the RNA base lesion compared to the control for each species. Proteins with log2 fold change > 1 and p-value < 0.05 are highlighted and labeled.


AP endonucleases
1
C. elegans - Apn−1 (Q10002)
A. thaliana - ARP (P45951)
H. sapiens - APEX1 (ENSG00000100823)
S. cerevisiae - APN1 (YKL114C)
B. subtilis - END4 (P54476)
H. salinarum - nfo (Q9HSK7)
Z. mays - A0A1D6N9G1
T. thermophila - W7X3I4
T. thermophila - Q23VX1
C. elegans - exo−3 (G5EBR7)
T. brucei - Tb927.8.5510
Z. mays - A0A1D6G9V5
E. coli - xthA (P09030)
S. pombe - apn1 (SPCC622.17)
E. coli - END4 (P0A6C1)
B. subtilis - exoA (P37454)

	8-oxoG	abasic	RNA base
C. elegans - Apn−1 (Q10002)	▨	■	▨
A. thaliana - ARP (P45951)	■	▨	■
H. sapiens - APEX1 (ENSG00000100823)	HeLa HEK293	HeLa HEK293	HeLa HEK293
S. cerevisiae - APN1 (YKL114C)	■	■	■
B. subtilis - END4 (P54476)	■	■	■
H. salinarum - nfo (Q9HSK7)	▨	▨	▨
Z. mays - A0A1D6N9G1	□	□	□
T. thermophila - W7X3I4	□	□	□
T. thermophila - Q23VX1	▨	■	▨
C. elegans - exo−3 (G5EBR7)	■	■	■
T. brucei - Tb927.8.5510	■	■	■
Z. mays - A0A1D6G9V5	▨	■	▨
E. coli - xthA (P09030)	▨	▨	▨
S. pombe - apn1 (SPCC622.17)	□	□	□
E. coli - END4 (P0A6C1)	▨	■	■
B. subtilis - exoA (P37454)	■	■	■

□ not detected ▨ not enriched ■ enriched

Figure B.7.: Neighbor-joining phylogenetic tree of photolyases; Figure 3.6B

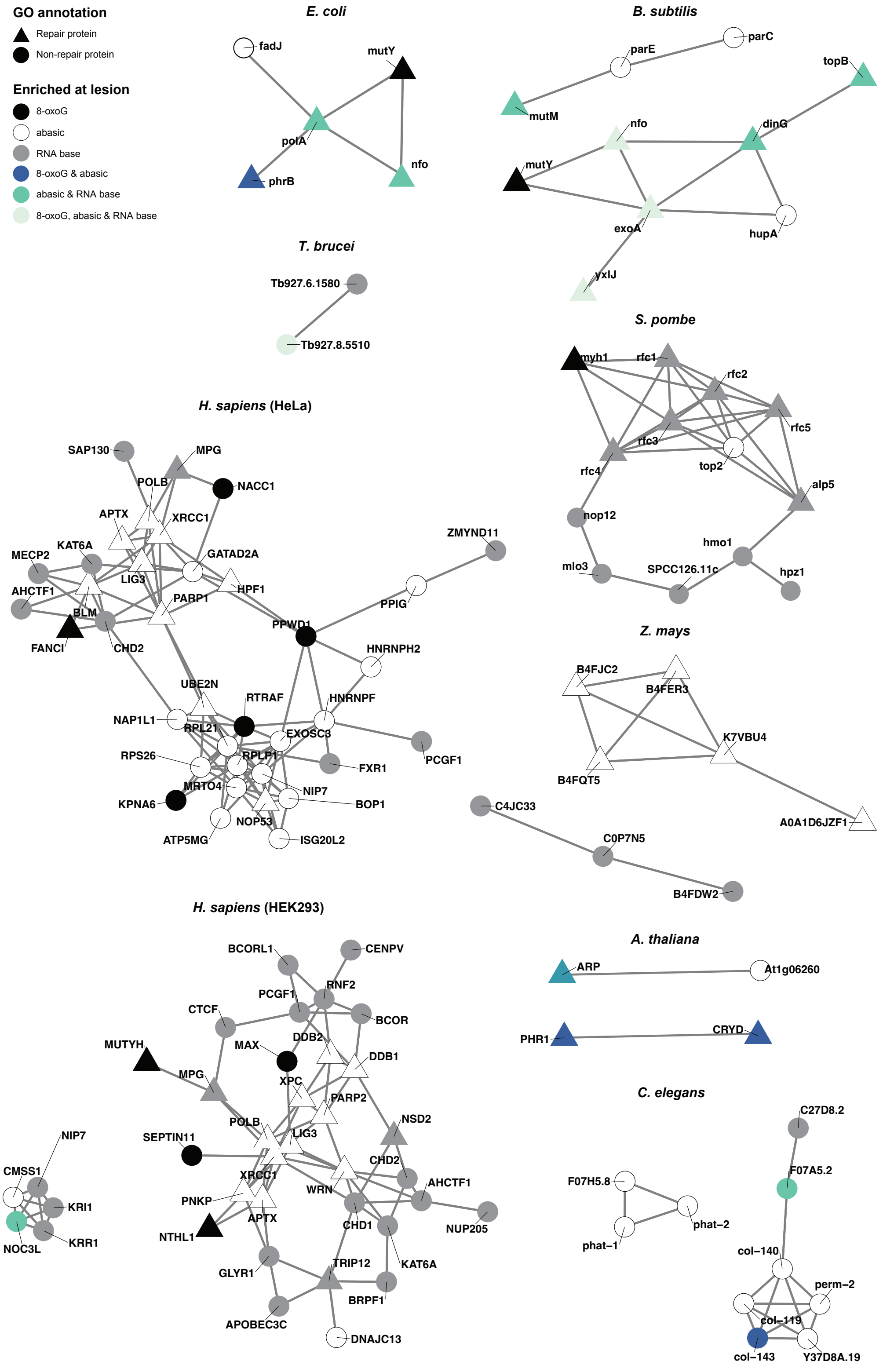


Figure B.8.: Networks of enriched proteins across lesions for *E. coli*, *B. subtilis*, *T. brucei*, *S. pombe*, *H. sapiens* (HeLa and HEK293), *Z. mays*, *A. thaliana*, and *C. elegans*. Interactions as reported in the STRING database.

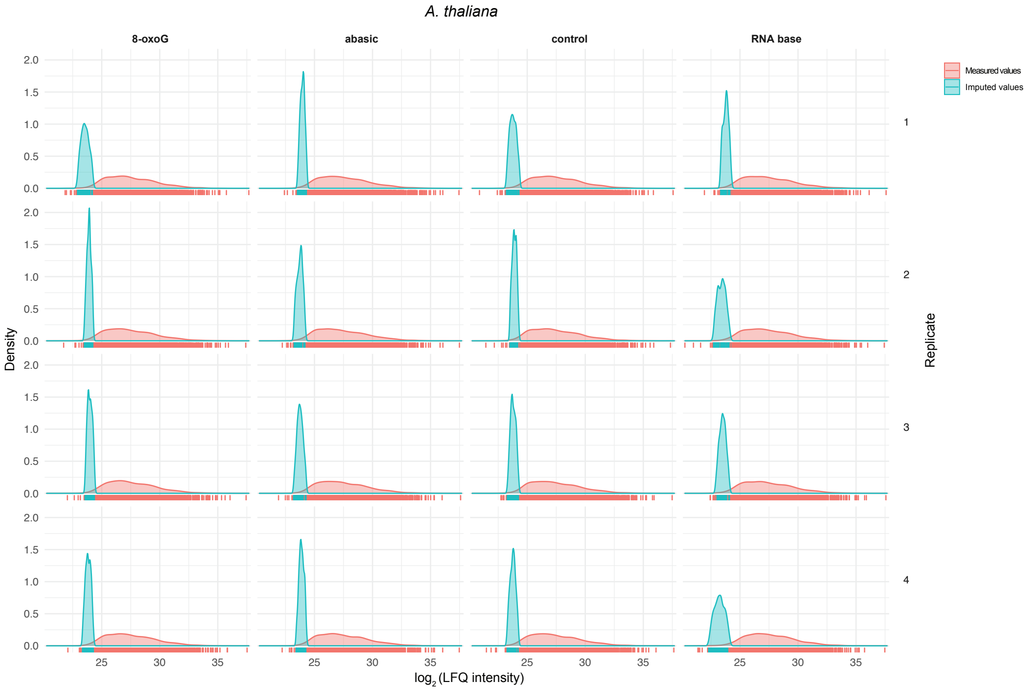


Figure B.9.: Density plot showing representative example (*A. thaliana*) of distribution of imputed log2 LFQ intensity values. Imputation was done for each replicate of the experiment (8-oxoG, abasic, control, RNA base), from a beta distribution within a range of the 0.2 and 2.5 percentile of measured intensities of the replicate (see Section 3.8). Density of imputation values shown in light blue, original measured intensities in red.

C

Supplemental information Chapter 4

Table C.1.: Comparison of different positive selection software with AlexandrusPS

Program Name	AlexandrusPS	IDEA	Armadillo	IMPACT_S	JCoDA	Armadillo	PhyleasProg	POTION	PSP	Selecton	VESPA
Multi-task CodeML	✓	×	×	×	×	×	×	✓	×	×	✓
SM Analyses	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
BM Analyses	✓	✓	✓	✓	×	✓	✓	✓	✓	✓	×
BSM Analyses	✓	×	✓	✓	×	✓	✓	✓	×	✓	✓
Manual Installation	×	✓	×	×	×	×	×	×	×	×	×
Web-Server Implementation	×	✓	✓	✓	×	✓	✓	×	✓	✓	×
Config Files for CodeML produced automatically	✓	✓	✓	✓	✓	✓	×	×	✓	✓	✓
Orthology Relations for CodeML produced automatically	✓	×	×	×	×	×	✓	✓	✓	×	×
MSA Files for CodeML produced automatically	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Phylogenetic Tree Files for CodeML produced automatically	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Automating Results Retrieval and LRTs	✓	✓	✓	✓	×	✓	✓	✓	×	×	✓
HPC Applicability	✓	✓	✓	✓	×	✓	✓	✓	✓	✓	✓
Open Source	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	×
Parallelized Processes	✓	✓	✓	✓	×	✓	✓	✓	✓	✓	✓
Docker Container	✓	×	×	×	×	×	×	×	×	×	×

Supplemental Data C.1
See https://doi.org/10.1093/gbe/evad187
Supplemental Data C.2
See https://doi.org/10.1093/gbe/evad187
Supplemental Data C.3
See https://doi.org/10.1093/gbe/evad187

Curriculum Vitae