

# The Molecular Underpinnings of Adaptive Evolution in Diatoms

A thesis submitted for the degree of Doctor of Philosophy  
University of East Anglia, Norwich, UK

Reuben John Gilbertson

November 2023

This copy of the thesis has been supplied on condition that anyone who consults it is understood to recognise that its copyright rests with the author and that use of any information derived there-from must be in accordance with current UK Copyright Law. In addition, any quotation or extract must include full attribution.

*For all who stayed by my side throughout and didn't let me quit, I am forever grateful.*

# Abstract

Diatoms are single-celled microalgae and are responsible for approximately half of all annual marine primary production. Thus, they play major roles in biogeochemical cycles and form the base of the marine food web. They are diverse with an estimated > 100,000 species and display high levels of genetic variability within natural populations, but the driving mechanisms are understudied. Mitotic recombination has been overlooked as a driver of genetic diversity due to a lower frequency of events compared to meiotic recombination. This thesis aims to elucidate the role of mitotic recombination in maintaining genome integrity and the generation of genetic diversity in diatoms by impairing homologous recombination (HR) through CRISPR/Cas.

Mitotic recombination is carried out by a suite of DNA repair genes which are highly conserved. Through extensive phylogenetic analysis in this thesis, the predicted core genes necessary for DNA pathways to function were identified in diatom genomes. However, differences in the conservation of X-family polymerases between diatom classes highlight potential evolutionary differences in repair mechanisms. To impair HR, the core DNA repair gene – BRCA2– was knocked out in the model diatom *Thalassiosira pseudonana* for the first time. Mutant cell lines showed increased sensitivity to induced DNA damage and were unable to adapt to high-temperature stress, confirming that BRCA2 has a role in the repair of genetic mutations arising from induced or environmental stress. Comparative genomics revealed that loss of HR function in *T. pseudonana* resulted in increases in copy number variations (CNV) and copy-neutral loss of heterozygosity (LOH) events leading to fixation of mutations in mutant genomes. This increase in genomic instability allowed aneuploidy to progress from affecting small regions to whole chromosomes over time. This thesis presents novel data on the role of HR in maintaining genomic integrity and discuss its potential in the generation of genetic diversity in diatoms.

## **Access Condition and Agreement**

Each deposit in UEA Digital Repository is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of the Data Collections is not permitted, except that material may be duplicated by you for your research use or for educational purposes in electronic or print form. You must obtain permission from the copyright holder, usually the author, for any other use. Exceptions only apply where a deposit may be explicitly provided under a stated licence, such as a Creative Commons licence or Open Government licence.

Electronic or print copies may not be offered, whether for sale or otherwise to anyone, unless explicitly stated under a Creative Commons or Open Government license. Unauthorised reproduction, editing or reformatting for resale purposes is explicitly prohibited (except where approved by the copyright holder themselves) and UEA reserves the right to take immediate 'take down' action on behalf of the copyright and/or rights holder if this Access condition of the UEA Digital Repository is breached. Any material in this database has been supplied on the understanding that it is copyright material and that no quotation from the material may be published without proper acknowledgement.



## Table of Contents

Abstract.....	iii
Table of Contents.....	iv
List of Tables .....	ix
List of Figures .....	xi
List of Abbreviations .....	xv
Acknowledgements.....	xvi
1 Introduction .....	1
1.1 Introduction to Phytoplankton .....	1
1.1.1 Introduction to Diatoms.....	3
1.1.2 Introduction to the Genetic Diversity of Diatoms.....	5
1.1.2.1 Evolution of the Diatoms .....	5
1.1.2.2 Endosymbiosis.....	7
1.1.3 Introduction to <i>Thalassiosira pseudonana</i> as a Model Species.....	9
1.2 Mitotic Recombination .....	9
1.2.1 DNA Repair through Homologous Recombination .....	9
1.2.2 Generation of Genetic Diversity via Mitotic Recombination .....	12
1.3 Thesis Outline and Aims.....	15
1.3.1 Chapter 2: Materials and Methods.....	16
1.3.2 Chapter 3: DNA Repair Systems in Diatoms.....	16
1.3.3 Chapter 4: Confirmation of <i>Brca2</i> Function in Diatoms and Long-term Consequences of Knock-out.....	16
1.3.4 Chapter 5: What are the mechanisms of Adaptation in HR Deficient <i>T. pseudonana</i> ..	16
1.3.5 Chapter 6:.....	17
1.4 Diatoms and Their Microbiomes in Changing Polar Oceans.....	17
2 Materials and Methods.....	31
2.1 Cell Culture.....	31
2.1.1 Growth Conditions .....	31
2.1.1.1 Temperate diatoms.....	31
2.1.1.2 Polar diatoms .....	31
2.1.2 Media Preparation .....	31
2.1.3 Growth Curves .....	33
2.1.3.1 <i>Global Methodologies</i> .....	33
2.1.3.2 <i>Dose-response Growth Curves</i> .....	34

2.1.3.3	<i>Temperature Response Curves</i> .....	35
2.1.3.4	<i>Short experimental evolution experiment – MMS treatment</i> .....	35
2.2	Identification of DNA Damage Using the Terminal Deoxynucleotidyl Transferase dUTP Nick End Labelling (TUNEL) Assay.....	36
2.2.1	Flow Cytometry.....	37
2.3	Genome Editing in <i>T. pseudonana</i> via CRISPR/Cas.....	38
2.3.1	sgRNA Design and <i>In Vitro</i> Analysis.....	38
2.3.2	Golden Gate Cloning.....	41
2.3.3	Biolistic Transformation.....	45
2.3.4	Selection of Transformed Colonies.....	46
2.4	DNA Extraction.....	47
2.4.1	Extraction of High-Molecular-Weight (HMW) genomic DNA (gDNA).....	47
2.4.2	DNA Extraction from Small Amounts of Cells.....	48
2.5	Bioinformatics.....	48
2.5.1	DNA Repair Proteins in Diatoms.....	48
2.5.2	Phylogenetic Analysis of <i>Brca2</i> and <i>Ku70</i> .....	51
2.5.3	Variant Calling Pipeline (WGS data).....	51
2.5.3.1	FASTQC and K-mer analysis.....	51
2.5.3.2	Read Mapping to the <i>T. pseudonana</i> Reference Genome.....	51
2.5.3.3	Marking Duplicates Reads and Variant Calling.....	52
2.5.4	Calculation of Mutation Rates.....	52
2.5.5	Go Term Enrichment Analysis.....	53
2.5.6	Detection of Copy Number Variation.....	53
2.5.7	Detection of Loss of Heterozygosity (LOH).....	53
3	DNA Repair Systems in Diatoms.....	55
3.1	Introduction.....	55
3.1.1	Overview of DNA Repair Pathways and Mechanisms.....	55
3.1.1.1	Double-strand Break Repair.....	57
<i>Homologous Recombination</i> .....	57	
<i>Non-Homologous End-Joining</i> .....	58	
3.1.1.2	Single-strand Break Repair, Mismatch Repair, and Repair of DNA Adducts Crosslinks and Oxidized Bases.....	60
<i>Base Excision Repair</i> .....	60	
<i>Mismatch Repair</i> .....	61	
<i>Nucleotide Excision Repair</i> .....	61	
3.1.2	DNA Repair's Role in Adaptation and Evolution.....	62

3.1.3	Inducing DNA Damage to Study Organisms' Response .....	64
3.1.3.1	Agents Used to Induce DNA Damage.....	64
3.1.3.1.1	Zeocin .....	64
3.1.3.1.2	Methyl methanesulfonate (MMS).....	65
3.1.3.2	Dose Response Curves .....	66
3.1.3.3	Evidence of DNA Damage .....	66
3.2	Results.....	67
3.2.1	Global Review of DNA Repair Pathways in Diatoms .....	67
3.2.1.1	Genome Size Compared with Number of DNA Repair Genes.....	67
3.2.1.2	Double Strand Break Repair .....	69
3.2.1.2.1	Homologous Recombination .....	69
3.2.1.2.2	Non-Homologous End-Joining.....	71
3.2.1.3	Single-strand Break Repair, Mismatch Repair, and Repair of DNA Adducts Crosslinks and Oxidized Bases.....	75
3.2.1.3.1	Base Excision Repair .....	75
3.2.1.3.2	Mismatch Repair.....	79
3.2.1.3.3	Nucleotide Excision Repair .....	80
3.2.2	Dose Response Curves .....	83
3.2.2.1	Evidence of Induced DNA Damage .....	83
3.2.2.2	Zeocin Dose Response Curves.....	85
3.2.2.3	Methyl methanesulfonate Dose Response Curves .....	89
3.2.3	Selection of Genes to Target to Impair Double-Strand Break Repair .....	91
3.2.3.1	Brca2 – Homologous Recombination.....	91
3.2.3.2	Ku70 – Non-Homologous End-Joining.....	93
3.3	Discussion – Potential Molecular Underpinnings of Tolerance Diversity .....	96
4	CRISPR/Cas Genome Editing in <i>T. pseudonana</i> and Comparative Genomics of <i>Brca2</i> Knock Outs 98	
4.1	Introduction .....	98
4.1.1	Genome Editing: CRISPR/Cas .....	98
4.1.1.1	Golden Gate Cloning as a Method to Create CRISPR/Cas9 Constructs.....	99
4.1.2	Whole Genome Sequencing in Diatoms .....	100
4.2	Results.....	100
4.2.1	<i>In Vitro</i> CRISPR/Cas Temperature Assay .....	100
4.2.2	Golden Gate Cloning .....	103
4.2.3	Transformation and Selection of Homozygous Knock Out Cell Lines .....	105
4.2.3.1	Selection of Edited Cell Lines Through PCR.....	105

4.2.3.2	Sanger Sequencing .....	107
4.2.4	Homozygous <i>Ku70</i> Knock-Out Cultures are Unstable.....	109
4.2.5	Impaired DNA Repair in Edited Cell Lines. ....	109
4.2.5.1	Dose-Response Curves .....	109
4.2.5.2	Temperature Response Curve.....	110
4.2.6	Whole Genome Sequencing of <i>Brca2</i> <i>-/-</i> Knock-Out Cell Lines .....	114
4.2.6.1	DNA Extraction and Quality Assessment of Raw Illumina Sequencing Data .....	114
4.2.6.2	Alignment to Reference Genome .....	114
4.2.6.3	Overview of Called Variants and Initial Filtering.....	115
4.2.6.4	Copy Number Variation .....	118
4.2.6.4.1	GO Analysis of CNV Effected Genes in <i>Brca2-09</i> .....	120
4.2.6.5	Allele Balance .....	121
4.3	Discussion.....	126
5	Short-term Experimental Evolution of <i>Brca2</i> <i>-/-</i> <i>T. pseudonana</i> .....	129
5.1	Introduction .....	129
5.1.1	Experimental Evolution in Microbial Populations.....	129
5.1.2	Miotic Recombination in Diatoms and <i>Brca2</i> .....	129
5.2	Results.....	132
5.2.1	Growth Data.....	132
5.2.2	DNA Extraction, Whole-Genome Sequencing and Mapping to Reference Genome ..	136
5.2.2.1	Treated <i>Brca2</i> <i>-/-</i> Replicate A at Timepoint 1 Has a Wild-type Copy of <i>Brca2</i> ....	136
5.2.3	Filtering and Overview of Mutations .....	137
5.2.3.1	Mutation Rates.....	138
5.2.4	Loss of Heterozygosity .....	139
5.2.5	Genetic Differentiation ( $F_{ST}$ ).....	143
5.2.6	Copy Number Variation (CNV) .....	146
5.2.6.1	Genes Affected by Copy Number Variation .....	151
5.2.7	Allele Balance and Chromosome Instability .....	153
5.3	Discussion.....	156
6	General Discussion.....	159
6.1	Summary of Main Results .....	159
6.2	General Discussion.....	160
6.3	Future Work .....	162
7	References .....	165
	Appendix .....	a
	Supplementary Information for Chapter 3 .....	a

Supplementary Information for Chapter 4 ..... a  
Supplementary Information for Chapter 5 ..... i

# List of Tables

## Chapter 2: Materials and Methods

- Table 2.1: Composition of Aquil media.
- Table 2.2: Reagents and their quantity used for level 1 Golden Gate reactions.
- Table 2.3: Reagents and their quantity used for level 2 Golden Gate reactions.
- Table 2.4: Final concentrations of antibiotics added to media for removal of bacteria from diatom cultures.
- Table 2.5: Reference proteomes used for phylogenetic study of conservation of DNA repair proteins in diatoms.

## Chapter 3: DNA Repair Systems in Diatoms

- Table 3.1: EC<sub>50</sub> values for *T. pseudonana*, *F. cylindrus* and *P. tricornutum* exposed zeocin.
- Table 3.2: EC<sub>50</sub> values for *T. pseudonana*, *F. cylindrus* and *P. tricornutum* exposed methyl methanesulfonate (MMS).

## Chapter 4: CRISPR/Cas Genome Editing in *T. pseudonana* and Comparative Genomics of Brca2 Knock Outs

- Table 4.1: List of primers used to synthesise sgRNA for *in vitro* testing.
- Table 4.2: Primers used to amplify regions of *Brca2* and *Ku70* for use in band shift assay.
- Table 4.3: EC<sub>50</sub> values for wild type and *Brca2* *-/-* *T. pseudonana* strains exposed methyl methanesulfonate (MMS).
- Table 4.4: A) Equations used to model the thermal niche of *Brca2* *-/-* strains and B) results of models used.
- Table 4.5: QualiMap BamQC statistics of alignment files.
- Table 4.6: Enriched GO terms of genes with a gain in CNV for each *Brca2* *-/-* culture.
- Table 4.7: List of events of trisomy in each sample by chromosome.

## Chapter 5: Short-term evolution of *Brca2* *-/-* *T. pseudonana*

- Table 5.1: Number of SNPs and mutation rate for each sample.
- Table 5.2: Number of novel homozygous SNPs for each sample.
- Table 5.3: GO terms from regions in chromosome 18 effected by loss of heterozygosity.
- Table 5.4: Weir and Cockerham  $F_{st}$  estimates for each population of *Brca2*, and wild type culture grouped by treatment and timepoint.

- Table 5.5: Number of genes affected by CNV for each population of *Brca2* and wild type culture grouped by treatment and timepoint.
- Table 5.6: Enriched GO terms of genes affected by CNV in individual *Brca2*  $-/-$  samples.
- Table 5.7: Enriched GO terms of genes affected by CNV in individual wild type samples.

#### Appendix:

- Table A.1: Spectrophotometry results of gDNA sent to EI for sequencing.
- Table A.2: Overview of mutations called via bcftools call.
- Table A.3: Overview of unique variants called via bcftools call after filtering.
- Table A.4: Protein ID, chromosome and KOG annotation of enriched genes with an increased copy number variation in *Brca2*-09
- Table A.5: Cells/ml data of inoculated density and end density for each date cultures were subbed during short-term experimental evolution experiment.
- Table A.6: Total generations and average specific growth rate of each culture by treatment and timepoint during short-term experimental evolution experiment.
- Table A.7: Raw reads and estimated raw coverage generated by the Earlham Institute for each sample.
- Table A.8: Statistics on the alignment of Illumina reads against the reference *T. pseudonana* genome.
- Table A.9: Overview of numbers of variants and variant types for each sample before and after filtering.
- Table A.10: Populations used for  $F_{st}$  calculations.
- Table A.11: Number of genes with either a gain or loss in CNV compared with wild-type T0.

# List of Figures

## Chapter 1: General Introduction

- Figure 1.1: The biological carbon pump.
- Figure 1.2: Diagram of diatom frustule morphology.
- Figure 1.3: Diatom diversity and abundance over the past 260 million years.
- Figure 1.4: Diagram showing primary and secondary endosymbiotic events.
- Figure 1.5: Overview of the homologous recombination DNA repair pathway.
- Figure 1.6: Outcomes of mitotic recombination and potential loss of heterozygosity events.
- Figures published in 'Diatoms and Their Microbiomes in Complex and Changing Polar Oceans'

Figure 1: Diagram showing examples of adaptations of polar diatoms in response to various environmental stressors.

Figure 2: Summary of available omics data for selected polar species.

Figure 3: Conceptual diagram of processes influencing polar microbial communities.

## Chapter 2: Materials and Methods

- Figure 2.1: Diagram of TUNEL assay principle and mechanisms.
- Figure 2.2: Initial gate settings for removal of doublet cells from FACS analysis
- Figure 2.3: Location of designed cut sites (sgRNAs) for both *Brca2* and *Ku70* in relation to position of conserved domains.
- Figure 2.4: Vector map of final level 2 constructs for both A) *Brca2* and B) *Ku70*.

## Chapter 3: DNA Repair Systems in Diatoms

- Figure 3.1: Overview of major DNA repair pathways and their associated proteins.
- Figure 3.2: Pathway map of eukaryotic homologous recombination.
- Figure 3.3: Pathway map of non-homologous end-joining.
- Figure 3.4: Comparison of genome size and the number of predicted DNA repair genes.
- Figure 3.5: PCA of the relationship between genome size, gene count, protein count and number of DNA repair genes.
- Figure 3.6: Heat map showing presence of genes involved in the homologous recombination (HR) pathway in diatoms.



- Figure 3.7: Heat map showing presence of genes involved in the non-homologous end joining (NHEJ) pathway in diatoms.
- Figure 3.8: Phylogenetic tree polymerase lambda ( $\lambda$ ).
- Figure 3.9: Heat map showing presence of genes involved in the base excision repair (BER) pathway in diatoms.
- Figure 3.10: Alignment of uracil-DNA glycosylase across diatom species.
- Figure 3.11: Heat map showing presence of genes involved in the mismatch repair (MMR) pathway in diatoms.
- Figure 3.12: Heat map showing presence of genes involved in the nucleotide excision repair (NER) pathway in diatoms.
- Figure 3.13: Results of fluoresce assisted cell sorting (FACS) analysis of TUNEL stained diatoms exposed to the DNA-damaging agent, zeocin.
- Figure 3.14: Increase in cell size of *T. pseudonana* exposed to zeocin over 72 hours.
- Figure 3.15: Growth curves of *T. pseudonana*, *F. cylindrus* and *P. tricornutum* exposed to a gradient of zeocin concentrations.
- Figure 3.16: A) Growth curve of *T. pseudonana* exposed to a gradient of zeocin concentrations with B) cell size and C) Fv/Fm.
- Figure 3.17: Relative viability of *T. pseudonana*, *F. cylindrus* and *P. tricornutum* over a gradient of zeocin concentrations, shown as a dose-response curve.
- Figure 3.18 Relative viability of *T. pseudonana*, *F. cylindrus* and *P. tricornutum* over a gradient of methyl methanesulfonate (MMS) concentrations, show as a dose-response curve, and B) cell size of *T. pseudonana* exposed to different MMS concentrations.
- Figure 3.19: Heat map showing conservation of *Brca2* across all 47 reference proteomes.
- Figure 3.20: Phylogenetic tree of *Brca2*.
- Figure 3.21: Phylogenetic tree of *Ku70*.

#### Chapter 4: CRISPR/Cas Genome Editing in *T. pseudonana* and Comparative Genomics of *Brca2* Knock Outs

- Figure 4.1: Gel results of PCR amplification of programmable 20nt sequence part of sgRNAs.
- Figure 4.2: RNA products synthesised for *in vitro* analysis of sgRNA cutting efficiency run on a polyacrylamide urea gel.
- Figure 4.3: Results of *in vitro* cutting assay.
- Figure 4.4: Vector maps of final level two constructs for A) *Brca2* and B) *Ku70* and C) the results of a restriction digest to check their completeness.

- Figure 4.5: Results of for band shift assay of PCR amplified target genes (*Brca2* and *Ku70*) for primary colonies.
- Figure 4.6: Results of for band shift assay of PCR amplified target genes (*Brca2* and *Ku70*) for secondary clones.
- Figure 4.7: Alignment of sanger sequencing results of transformed secondary clones for A) *Brca2* and B) *Ku70*.
- Figure 4.8: Relative viability wild type and *Brca2* *-/-* *T. pseudonana* strains exposed methyl methanesulfonate (MMS) shown as a dose-response curve.
- Figure 4.9: Temperature response curve of *Brca2* *-/-* *T. pseudonana* strains across a gradient of temperatures.
- Figure 4.10: Venn diagram of filtered shared variants and affected genes between three *Brca2* *-/-* strains and wild type.
- Figure 4.11: Percent of total unique variants which are SNPs.
- Figure 4.12: Scatter plot of CNV in *Brca2* *-/-* strains.
- Figure 4.13: Allele balance plots for *Brca2* *-/-* and wild type strains showing changes in ploidy.

#### Chapter 5: Short-term evolution of *Brca2* *-/-* *T. pseudonana*

- Figure 5.1: Growth data of cultures throughout the short-term experimental evolution project. A) end cell density and B)  $F_v/F_m$
- Figure 5.2: Number of generations of wild type and *Brca2* *-/-* strains at timepoint 1 and timepoint 2.
- Figure 5.3: Visualisation of *Brca2* locus in wild type and selected *Brca2* strains showing presence of wild type copy of *Brca2* in one of the labelled knock out strains.
- Figure 5.4: Number of filtered variants in *Brca2* and wild type strains.
- Figure 5.5: Number of homozygous variants grouped by A) *Brca2* and wild type, B) *Brca2* treated and untreated; C) wild type treated and untreated.
- Figure 5.6: LOH events on chromosome 18.
- Figure 5.7: Circular figures showing regions of genetic differentiation ( $F_{st}$ ).
- Figure 5.8: Genes affected by copy number variations in *Brca2* and wild type strains.
- Figure 5.9: Genes affected by copy number variations in *Brca2* and wild type strains with and without treatment with MMS.
- Figure 5.10: Scatter plots showing gain of CNV events between T1 and T2.
- Figure 5.11: Allele balance plots of chromosome 23 in *Brca2* and wild type strains.

## Appendix:

- Figure A.1: Alignment of *Brca2* protein sequence showing conserved domains.
- Figure A.2: Pulse-gel electrophoresis results of gDNA sent to EI for sequencing.
- Figure A.3: Venn diagram of all called variants.
- Figure A.4: CNV plots for each sample.
- Figure A.5: Allele balance plots of T0 *Brca2*  $-/-$  and wild type.

## List of Abbreviations

BER	Base excision repair
BIR	Break-induced repair
BRCA2	Breast Cancer type 2 susceptibility protein
BWA-MEM	Burrows-Wheeler Alignment Tool
clonNAT	Nourseothricin
CNV	Copy number variation
CO	Crossover event
CRISPR	Clustered regularly interspersed short palindromic repeats
DDR	DNA damage repair
dHJ	Double Holliday junction
DMS	Dimethyl sulfide
DMSO	Dimethyl sulfoxide
DSB	Double strand break
DSBR	Double-strand break repair
EC <sub>50</sub>	Half maximal effective concentration
EGT	Endosymbiotic gene transfer
FACS	Flourescence activated cell sorting
FITC	Fluorescein isothiocyanate
$F_{ST}$	Fixation index
HJ	Holliday junction
HR	Homologous recombination
ICL	Interstranded crosslinks
INDELS	Small insertion or deletion
IR	Ionising radiation
LOH	Loss of heterozygosity
MMR	Mismatch repair
MMS	Methyl methanesulfonate
NEB	New England Biolabs
NER	Nucleotide excision repair
NHEJ	Non-homologous end joining
PCR	Polymerase chain reaction
POC	Particulate organic carbon
RNP	Ribonucleoprotein
ROS	Reactive oxygen species
rRNA	Ribosomal RNA
SDSA	Synthesis-dependent strand annealing
sgRNA	Single guide RNA
SNP	Single nucleotide polymorphism
SOW	Synthetic ocean water
SSB	Single-strand break
ssDNA	Single-stranded DNA
TUNEL	Terminal deoxynucleotidyl transferase dUTP nick end labelling

# Acknowledgements

I would like to thank my primary supervisor, Thomas Mock, for his constant support and advice throughout my PhD. Thank you for allowing me to design my experiments to explore the world of diatoms. Thank you to Amanda Hopes who taught me all I know about molecular biology in diatoms. You taught me so much and I am forever grateful for your time. And a special thank you for supplying the plasmids and necessary materials for my lab work, I couldn't have done this without you. I would also like to thank my secondary supervisor, Cock van Oosterhout for his support and advice.

Thank you to the Leverhulme Trust (RPG-2017-364) and NERC (NE\R000883\1) who supplied the funding for my work.

Thank you to the School of Environmental Sciences at UEA for their constant support. Especially from the laboratory technicians. A special thank you to Robert Utting, who was always there to offer a lending hand no matter how busy he was. He was essential in helping me complete my work and always made me smile, thank you.

I would also like to thank the members of the Mock group. In particular, thank you Krisztina Sarkozi and Nigel Belshaw for your guidance and support with my lab work throughout my PhD. Andrew Toseland, thank you so much for putting up with my constant questions about bioinformatics and showing me how to conduct the analysis that led to my results I promise I will stop bothering you (for now). Thank you to all the other members of the lab who always had interesting conversation and insights into my work.

Thank you to Darren Heavens and Richard Leggett at the Earlham Institute for advising me on DNA extraction techniques. Thank you to all at the Earlham Institute who helped with decisions on sequencing methods and ultimately sequencing my samples.

Thank you to Roy Dunford at the John Innes Centre for training me to use the FACS sorter to analyse my samples.

A special thanks to my friends and family, I would not have completed my thesis without your love and support. To both my parents, I will never be able to describe how much you helped me. I am forever grateful and will always remember your support.

Last but certainly not least, I would like to thank my wonderful wife, Victoria. Thank you for your love and care when I was at my best and staying with me through the lows. You've kept me sane and grounded throughout the entire PhD, thank you so much.

# 1 Introduction

## 1.1 Introduction to Phytoplankton

Phytoplankton are single-celled, free-floating photosynthetic organisms that are present in almost all aquatic environments from the polar oceans to inland lakes. Phytoplankton are a polyphyletic group which are classified into seven major taxonomic phyla: Bacillariophyta (diatoms), Chlorophyta (green algae), Chrsophyta (golden algae), Cryptophyta (cryptomonads), Cyanobacteria (prokaryotic blue-green algae), Dinoflagellata (dinoflagellates) and Euglenophyta (euglenids). Phytoplankton have been of critical significance to global cycles ever since the first cyanobacteria began to oxygenate the atmosphere around 2.1 – 2.4 billion years ago (Schirrmeister and Gugger, 2015). In today's ecosystems, they play major roles in both regulation of biogeochemical climate cycles and form the base of food chains.

Half of all global photosynthesis, the biological process of generating energy using sunlight and water, is conducted by marine phytoplankton (Simon et al., 2009). Phytoplankton use the energy from photosynthesis to synthesise cellular components, maintain cellular functions and to ultimately grow through asexual or sexual reproduction. Since photosynthesis relies completely on solar radiation, marine phytoplankton occupy the upper ocean layer known as the photic zone (0 - 200m deep). They account for approximately half of all global primary production despite only occupying ~1% of global biomass (Field et al., 1998). In the process, they accumulate micro and macronutrients from their surroundings and integrate them into their biomass. Notably, one of these is carbon dioxide (CO<sub>2</sub>), a greenhouse gas which is currently at the centre of global climate change and global concentrations are becoming a driver of government environment policy. Phytoplankton capture CO<sub>2</sub> and sequester it to the deep ocean as they sink as particulate organic carbon (POC), known as the biological carbon pump (Figure 1.1; Basu et al., 2018). Approximately 1% of the CO<sub>2</sub> is removed from the upper ocean and the rest is consumed by microbes and zooplankton and turned back into CO<sub>2</sub> through cellular respiration (Herndl and Reinthaler, 2013). Their impact on the natural ecosystems is significant and in turn effects

how countries are designing their environmental policies. Their impact on biogeochemical cycles is not limited to just carbon capture but also includes the cycling of dimethylsulfide (DMS) which provides cloud condensation nuclei around which clouds can form helping to increase the amount of solar radiation backscatter (Charlson et al., 1987; Levasseur et al., 1994). These are just two examples of the major role phytoplankton play in global climate cycles.

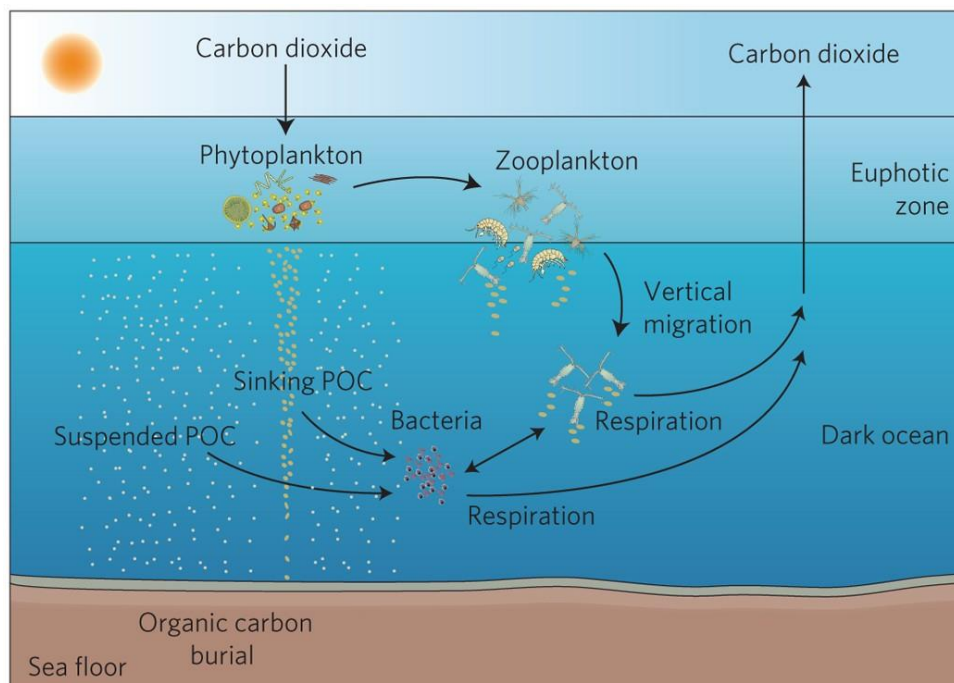


Figure 1.1. The biological carbon pump. In the euphotic zone, phytoplankton fix  $\text{CO}_2$  by using energy from the sun and splitting water. The carbon is sequestered when the particulate organic carbon (POC) sinks below the euphotic zone where it is consumed by other microbes and zooplankton (Graphic is from Herndl and Reinthaler, 2013).

Phytoplankton form the base of many food chains across the globe. For example, they support the polar ecosystem food webs which are some of the most productive on the planet (Gilbertson et al., 2022). Specifically in the Antarctic, they are the main food source for krill (*Euphausia superba*) which is the primary source of energy for larger organisms such as whales and penguins. This impact goes further than simply sustaining the food web, krill are also a key part of biogeochemical nutrient cycles due to their vertical migration and also support large-scale fisheries (Pauly and Christensen, 1995; Cavan et al., 2019). Understanding the molecular underpinnings of phytoplankton functions is critical to aid in our knowledge of the natural world.



### 1.1.1 Introduction to Diatoms

One of the major phyla of phytoplankton is the Bacillariophyta also known as diatoms. Diatoms comprise a significant proportion of the global phytoplankton community and dominate stochastic coastal environments and polar ocean microbiomes (Gilbertson et al., 2022). Diatoms are responsible for approximately 20% of global photosynthesis accounting for almost half of all phytoplankton (Nelson et al., 1995). This photosynthesis results in roughly the same amount of carbon-based molecules generated by all terrestrial rainforests put together (Field et al., 1998). The term diatom comes from the Greek 'diatomos' which means 'cut in two' and refers to their unique cell wall, or frustule. The silica deposition vesicle, a special organelle in diatoms, creates the frustule using amorphous silica (SiO<sub>2</sub>; Tesson et al., 2017). Each diatom species' frustule exhibits a unique nano-scale pattern morphology which has been used to characterise individual species (Round et al., 1990). The frustule is composed of two halves called *thecae* and the shape is often described as a petri dish. One *theca* is larger (*epitheca*) with a smaller valve inside the larger (*hypothecae*; Figure 1.2; Cox, 2014). Each theca are subdivided into a *valve* and *cingulum*. In terms of the petri dish analogy, the *valve* is the lid, and *the cingulum* is the side of the dish. The girdle band nearest the valve is termed the *valvocopula* and a common structural feature of girdle bands is a tongue-like extension called the *lingula* (Figure 1.2). The valves face each other, and the cingulum overlap to form the *girdle* (Ross et al., 1979).

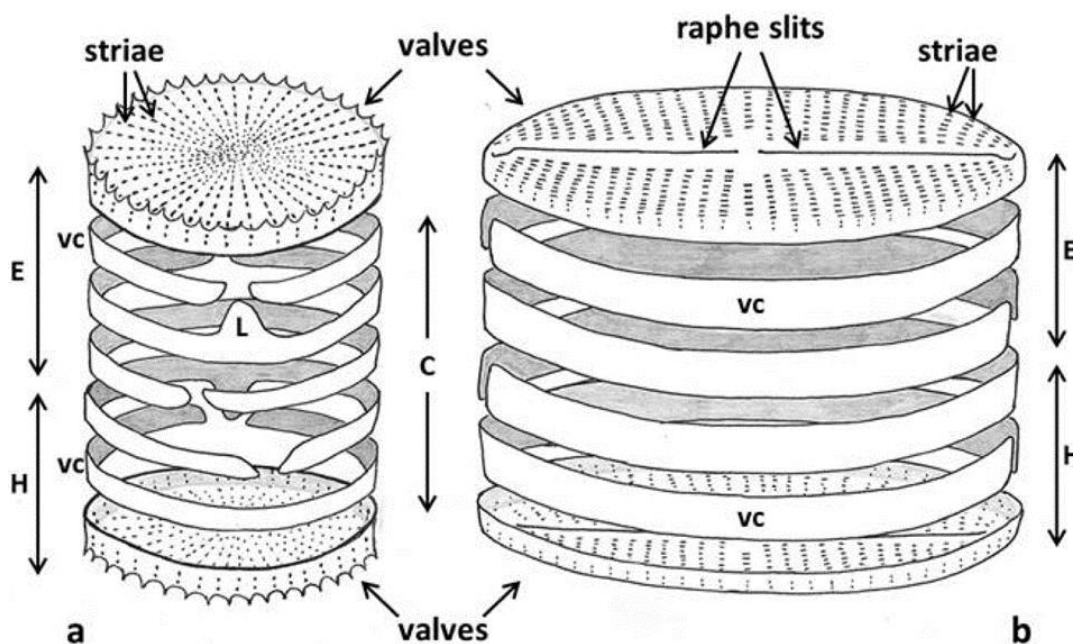


Figure 1.2. Diagram showing the structure of the diatom cell wall or frustule. a) representation of a centric diatom morphology and b) representation of a pennate diatom morphology. Abbreviations: E = epitheca; H = hypotheca; C = cingulum; vc = valvocopula and L = lingula. Figure is from Cox, 2014).

Generally, diatoms are classified into two overarching groups by the symmetry of the frustule, with centric diatoms displaying radial symmetry and non-polar symmetry in pennate diatoms. Pennate diatoms are further divided based on the presence of a raphe – a canal-like slit in the frustule that provides mobility to the cell (Ruck and Theriot, 2011). However, diatoms are such a diverse group with a vast myriad of morphologies and with more species being documented along with increasing molecular data (DNA sequencing) it is challenging to set definite boundaries of lineages. Medlin and Kaczmarska reviewed all available fossil, morphological and molecular data at the time and proposed diatoms to be classified into three classes: Coscinodiscophyceae, Mediophyceae and Bacillariophyceae (Medlin and Kaczmarska, 2004). Coscinodiscophyceae consists of primarily radially symmetrical frustules ornamented from a single central point, which are considered radial centric diatoms. Mediophyceae contains the multi and bi-polar centric diatoms. Finally, the Bacillariophyceae class is composed of bipolar diatoms usually displaying bilateral symmetry and the presence of a central sternum with or without a raphe (Medlin and Kaczmarska, 2004).

The life cycle of diatoms is split between long periods of asexual (i.e., clonal) vegetative growth lasting anywhere from months to years followed by irregular sexual reproduction. The mechanism of reproduction is linked through the size of the frustule. The frustule is rigid and is not deconstructed before asexual reproduction, and each valve becomes the new epivalve. After mitotic division the daughter cell which received the parental epivalve is the same size while the other is smaller, decreasing the cell size of the population with each division (Drebes, 1977). Once diatoms reach a critical size, they undergo sexual reproduction through the formation of an auxospore (Sánchez et al., 2019). Each diatom species has a unique critical cell size at which sexual reproduction is triggered. The mechanism of sexual reproduction is another major difference between centric and pennate diatoms. Sexual reproduction in centric diatoms is oogamous, where a mobile 'male' gamete fertilises an immobile 'female' gamete. The 'female' diatom produces one or two egg cells which are fertilised by microspores produced through meiosis in 'male' cells. In pennates, both 'male' and 'female' cells produce gametangia cells which pair and undergo meiosis. Both mechanisms produce a zygote which forms an auxospore. Division of the auxospore produces cells of the maximum cell size for the species and allows mitotic division to resume (Drebes, 1977). However, some species can restore size without auxospore production or show no decrease in size.

### 1.1.2 Introduction to the Genetic Diversity of Diatoms

#### 1.1.2.1 *Evolution of the Diatoms*

Diatom frustules which sink to the bottom of the water column have the potential to be preserved in an extensive fossil record, providing detailed insight into their evolutionary history. Frustules were once the only way to understand diatoms' evolutionary history; however, molecular techniques are now widely used in evolutionary studies. Even with molecular data and the fossil record, there are several theories on the emergence of diatoms. For example, two differing theories are from Round et al (1990), who suggested that diatoms arose from a single "naked" photosynthetic cell making them monophyletic, whereas Mann & Marchant (1990) propose that the Parmales, a polar silicified autotroph are the closest lineage to diatoms (Sims et al., 2006). Despite varying theories on their closest lineage, it is generally accepted from the fossil record in parallel with molecular data

that the radial centric lineage of diatoms first appeared as very heavily silicified autotrophs during the early Mesozoic Era (150-200 mya) (Kooistra and Medlin, 1996). The fossil record supports this as the earliest record of diatoms. In the late 1800's Rothpletz extracted diatom frustules from sediment deposits dating back to the Jurassic period, called the Liassic Boll shales. During the Jurassic period the tectonic plates were separating (Figure 1.3), creating shallow waters with high levels of nutrients from the new increased level of terrestrial runoff, providing a habitat favouring larger phytoplankton residing on the continental margins such as diatoms (Cermeno et al., 2015; Benoiston et al., 2017).

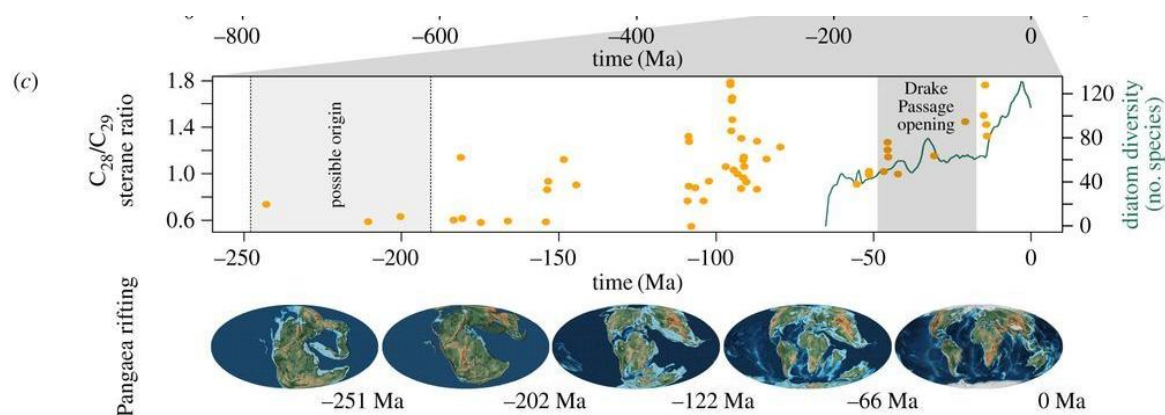


Figure 1.3. Diatom diversity and abundance over the past 260 million years. This is a portion of Figure 1 from Benoiston et al., 2017.

After diatoms emerged, they rapidly diversified and spread throughout the oceans. The taxonomy of the group has changed several times with novel discoveries and technologies. Molecular techniques have had the most influence on diatom taxonomy since C. A. Aragdh (1832) defined diatom taxonomic structure based on frustule shape and colony formation and concluded there are three families: Cymbelleae, Styllarieae, and Fragilarieae (Medlin 2016). Molecular clocks analysis of the differences in genomic deoxyribonucleic acid (DNA), ribosomal ribonucleic acid (rRNA), or amino acid sequences between the groups of interest with relative sister groups and outgroups to determine the divergence rate (Ho, 2008). Using molecular clocks, and node calibration using the fossil record, Medlin and Kaczmarska formally suggested a revision to the taxonomic structure of diatoms (Medlin and Kaczmarska, 2004). They defined three major classes of diatoms, the radial centrics,

characterised by radial symmetry and numerous discoid plastids, bipolar centrics, and pennates which exhibit bilateral symmetry (Medlin and Kaczmarska, 2004). Diatoms, a young group of organisms, have diversified rapidly and are still actively adapting to their environments. With new adaptations and eventual speciation, taxonomists are continuously looking at morphological and molecular data to correctly place diatoms in the tree of life.

#### 1.1.2.2 *Endosymbiosis*

The theory of endosymbiosis states that several eukaryotic organelles, including the mitochondria and plastids, originally evolved from free-living prokaryotes (Bodył et al., 2017). The word symbiosis is derived from Latin meaning 'living together' and is applied to instances where two distinct organisms coexist. There is a debate as to who should be credited with discovering endosymbiotic origins of eukaryotic organelles, but the Russian biologist Constantin Mereschkowsky proposed a thorough argument that some cells evolved through the fusion of two independent organisms (Mereschkowsky, 1905; Martin and Klaus, 1999). In 1967 Lynn Margulis hypothesised that the mitochondria, plastids and basal bodies of flagella were originally independent prokaryotic cells (Sagan, 1967). This paper has been credited with being the first unified theory of endosymbiotic theory and sparked global debate about the origin of these organelles (Gray., 2017). Lynn Margulis's insights helped bring about a contemporary understanding of endosymbiosis.

A primary endosymbiotic event defines a heterotrophic organism engulfing a prokaryotic cell (cyanobacterium) and retaining the cell as an organelle (Martin et al., 2015). The presence of multiple membranes surrounding mitochondria and plastids is cited as evidence of these organelles once being autonomous organisms (Falkowski, 2004). Green algae emerged from primary endosymbiosis which eventually evolved multicellularity and gave rise to terrestrial plants.

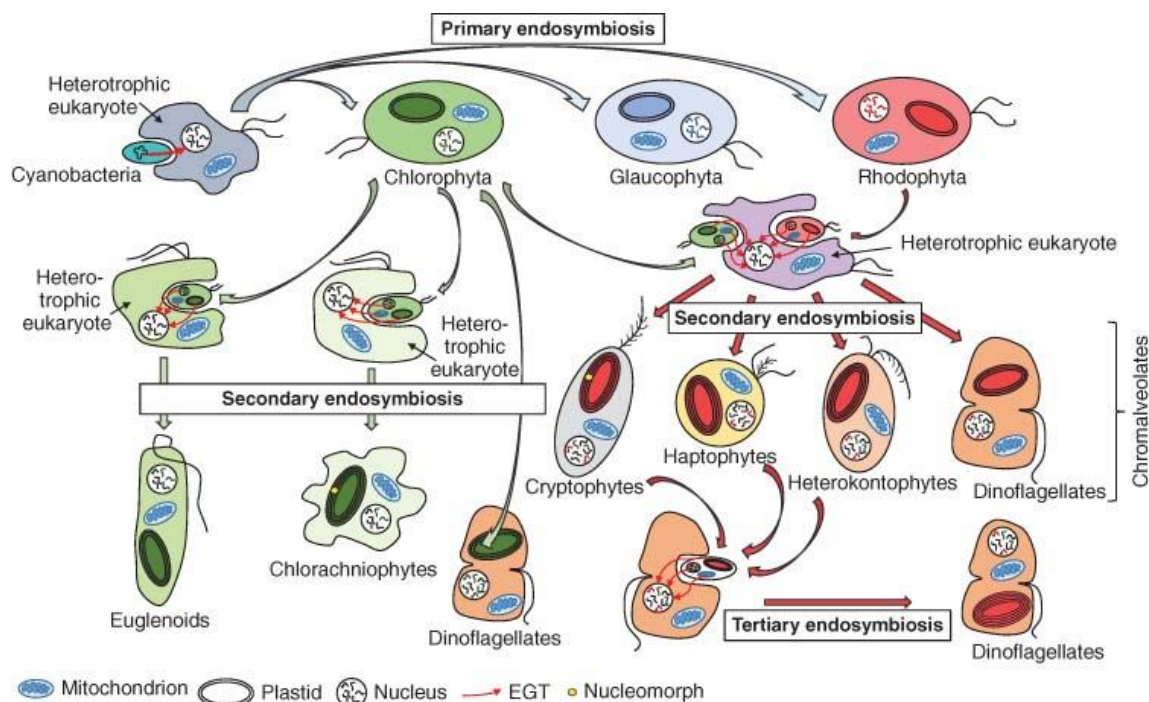


Figure 1.4. Diagram showing primary and secondary symbiotic events in phytoplankton evolution. (From Hopes and Mock, 2015)

Diatoms are part of the stramenopiles, also known as heterokonts as described in figure 1.4, a clade which encompasses an extreme diversity of organisms including brown macroalgae such as kelps and photo-mixotrophic species which both photosynthesise and acquire nutrients to phagotrophy (Dorrell and Bowler, 2017). A core event shared by all stramenopiles is a secondary endosymbiotic event between an ancestral heterotrophic stramenopile and both a green and red alga (Figure 1.4; Falkowski, 2004). As well as utilising the plastid for photosynthesis, some genes were moved to the nuclear genome through endosymbiotic gene transfer (EGT; Chan et al., 2012). This event contributed to the mix-and-match genomes found in diatoms and aided in their adaptation to diverse ecosystems allowing them to become one of the most successful phytoplankton taxa, but also the most genetically diverse (Armbrust et al., 2004; Bowler et al., 2008). There are an estimated 100,000 – 200,000 distinct species of diatoms, though only around 12,000 have been identified through molecular techniques and phenotypic traits (Mann and Vanormelingen, 2013; Malviya et al., 2016).

Diatom's chimeric genomes have accumulated genes from a variety of external sources outside of endosymbiotic events, providing them with molecular pathways and



functions not commonly found throughout phytoplankton lineages. For example, diatoms possess a complete urea cycle, previously thought to be unique to heterotrophic organisms to detoxify cellular ammonia (Armbrust et al., 2004). The urea cycle aids in the recovery of cellular functions impacted by nitrogen limitation. Polar diatoms appear to have acquired ice-binding proteins of prokaryotic origin which has aided in their adaptation to the polar oceans (Vance et al., 2019).

### 1.1.3 Introduction to *Thalassiosira pseudonana* as a Model Species

*Thalassiosira pseudonana* CCMP 1335 (Hustedt) Hasle et Heimdal is a small (3.5 ~ 10µm in length) centric diatom which has become a powerful model organism in recent research. This strain was first isolated from Moriches Bay, Forge River, Long Island, New York USA in 1958 by Guillard, R. It was the first diatom to have its whole genome sequenced and assembled into chromosomes (Armbrust et al., 2004). The availability of the genome sequence enabled genetic editing methods to be developed for *T. pseudonana* allowing the use of reverse genetics to elucidate the role of single genes and pathways (Poulsen et al., 2006; Hopes et al., 2016; Shrestha and Hildebrand, 2017; Nawaly et al., 2020; Belshaw et al., 2022). Sexual reproduction provides the potential for the generation of genetic diversity by facilitating recombination, but it has never been observed in *T. pseudonana*. Despite this this species has been found across both marine and freshwater habitats, suggesting it has adapted to a myriad of ecosystems (Alverson et al., 2011; Malviya et al., 2016).

*T. pseudonana* was used in this study given the reasons above to further the understanding of the role of mitotic recombination in the maintenance of genomic stability and the generation of genetic diversity in the context of adaptability.

## 1.2 Mitotic Recombination

### 1.2.1 DNA Repair through Homologous Recombination

DNA damage is inevitable and occurs throughout the cell cycle. A variety of external and internal factors can cause damage such as ultraviolet radiation (UV) and replication errors respectively. There are several major pathways that specialise in repairing specific types of DNA lesions. This section focuses on the most deleterious form of DNA damage,

double-strand breaks (DSBs) and their subsequent repair through the homologous recombination pathway. The other pathways are introduced in chapter 3. DSBs pose the most risk to genomes and can be created in a variety of ways resulting in breaks in both DNA strands that activates either the NHEJ or HR pathways (Ackerman and Horton, 2018). Cells respond to DNA damage through mechanisms that both scan for and repair damage bases in the genome. Damaged DNA must be dealt with quickly and efficiently to prevent them from becoming 'fixed' mutations that could be passed down to further generations. There are five major categories of DNA repair pathways: base excision repair (BER), nucleotide excision repair (NER), non-homologous end joining (NHEJ), mismatch repair (MMR) and homologous recombination (HR). Chapter three contains individual introductory sections that describe the molecular mechanisms of each pathway.

HR, which uses a homologous DNA sequence as a template to guide and repair the damaged DNA, is found throughout the domains of life (Lin et al., 2006). Recombination was first discovered in the early 1900s by the works of scientists such as William Bateson, Reginald Punnett, and Thomas Hunt Morgan, who observed links between inherited genes. Since the discovery of HR scientists have been unravelling the suite of proteins and signals that are core to DNA repair. The necessity to have a homologous sequence drove the thought that only eukaryotes could use this mechanism; however, in 1947 Tatum and Lederberg showed that *Escherichia. coli* was also capable of using HR to create genetic diversity in the absence of sexual reproduction centred around the *recA* gene during mitotic division (Reviewed Lin et al., 2006). This ability to shift sequences around and create new combinations can potentially lead to more efficient or novel proteins but can also have deleterious effects. The evolutionary advantages of homologous recombination have seen it be transferred throughout the domains of life from ancient beginnings (Lin et al., 2006).

HR is not a simple process and requires a suite of genes to successfully repair damage such as DSBs and interstranded crosslinks (ICLs). This thesis will focus on the eukaryotic HR pathway. HR can be broken down into three conceptual steps: presynapsis, synapsis, and postsynapsis (Figure 1.5; Li and Heyer, 2008). First, the MRN protein complex, composed of checkpoint proteins Rad50, NBS-1, Mre11, and ATM, scans DNA for DSBs. After finding a DSB, the *recA* eukaryotic homologue *Rad51* provides the initial and core reaction of HR by performing a homology search and invading the single-stranded DNA (ssDNA) to form a Rad51-ssDNA filament. RPA has a higher affinity for ssDNA and binds before RAD51.



Some single-strand binding proteins, such as RPA, while preparing DNA for homologous recombination by preventing secondary structures in the ssDNA, inhibit the ability of Rad51 to bind to the ssDNA. Mediator proteins, such as Rad52 and BRCA2, displace and replace RPA with RAD51 to the ssDNA (Sugiyama et al., 1997; Beernink and Morrical, 1999; Egger et al., 2002; Krogh and Symington, 2004; Symington et al., 2014) The subsequent invasion of the Rad51-ssDNA filament, and the formation of a D-loop structure with the homologous DNA strand concludes the synapse in the pathway. After the formation of the D-loop, there are several pathways resulting in different genetic outcomes (Figure 1.5). Double-strand break repair (DSBR), synthesis-dependent strand annealing (SDSA), and break-induced repair (BIR) are the three known pathways possible following the D-loop formation, there may be more possibilities that are unknown. In DSBR Rad52 catalyses the annealing of the invading ssDNA template to its complementary sequence (Sugiyama, et al., 1997), resulting in a double Holliday junction (dHJ), a complex where four strands of ss-DNA interact. Sometimes a single allele will be assigned as the template and duplicated on the other DNA strands in the HJ, known as gene conversion which contributes to copy-neutral loss of heterozygosity (LOH). Proteins, such as the ATP-dependent eukaryotic RuvB-like protein (RUVBL1) , catalyse the extension of Holliday junctions, polymerases then bind and synthesise new DNA strands, and their resulting homology depends on the active proteins and the structure of the synapse. Reshuffling of genetic material through crossover when repairing dsDNA breaks within somatic cells can be potentially deleterious by disrupting core genes and creating a LOH (Moynahan and Jasin, 1997). The SDSA pathway, in contrast to BIR, does not result in LOH as it only results in non-crossover products, conserving genetic diversity (Figure 1.5).

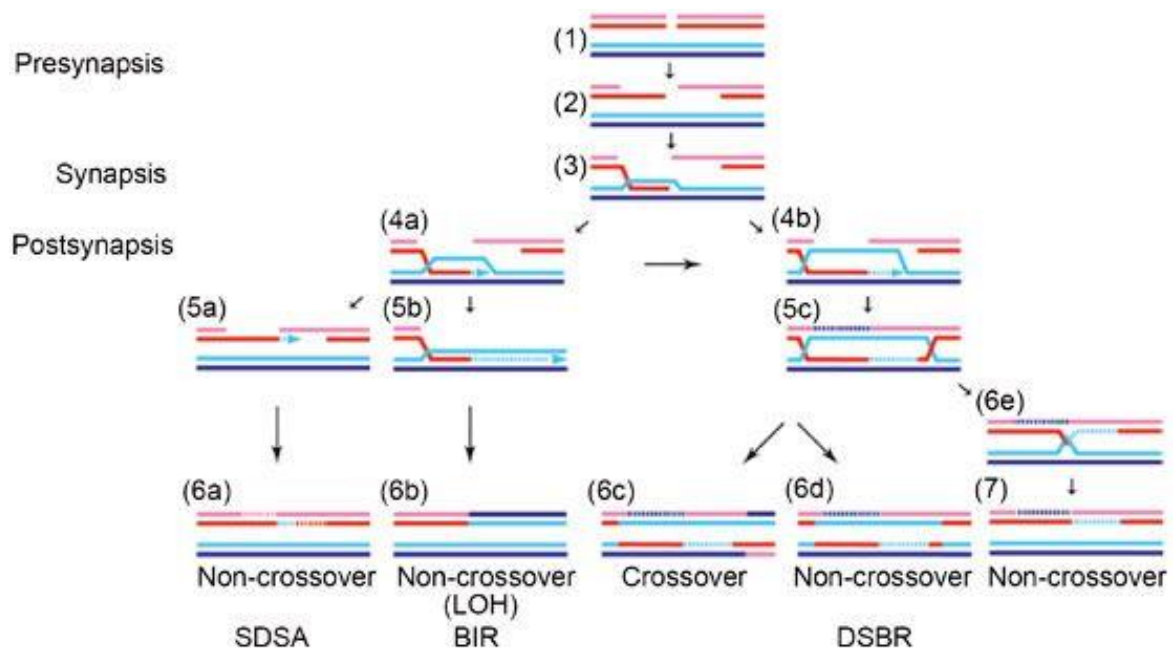


Figure 1.5 A schematic outlining the various sub-pathways of homology-directed repair of a DNA double-strand break (DSB). The three stages, Presynapsis, synapsis and postsynapsis are divided into steps. In presynapsis steps 1-2: DSB ends recognised and processed to expose ssDNA. Synapsis is step 3) the invasion of ssDNA by Rad52 and creation of D-loop. Postsynapsis steps 4a – 5a – 6a: are synthesis-dependent strand annealing (SDSA) sub pathway; steps 4a – 5b – 6b are break-induced repair (BIR) sub pathway; and steps 4b – 5c – 6c-e – 7: are double-strand break repair (DSBR). The figure is from Li and Heyer 2008.

### 1.2.2 Generation of Genetic Diversity via Mitotic Recombination

Mitotic recombination has been understudied as a driver of genetic diversity due to the low rate of events with greater research focus on base mutation rates (Yang et al., 2015), meiotic crossover events (Pazhayam and Turcotte, 2021) and gene flow (Mancera et al., 2008). The primary function of mitotic recombination is the maintenance of genomic stability through the repair of DNA DSBs (Moynahan et al., 2010). Mitotic recombination events are rare and have been estimated to have a rate of  $1 \times 10^{-5} - 10^{-6}$  per site per generation in *Arabidopsis thaliana* (Assaad and Signer 1992) with low rates also reported in yeast (Yin and Peters, 2013). With rates lower than meiosis, mitotic recombination has been overlooked, especially in organisms that do not separate germline and somatic growth such as plants, but also in diatom species which undergo irregular sexual reproduction. Mitotic recombination can also affect a more varied length of the genome in a crossover event. In yeast single events have been characterised to affect over 120kb segments while meiosis only modified 2kb (Yim et al., 2014). Given the intrinsically low event rate of mitotic recombination, it can be hard to directly study. Several experimental designs to attempt to

directly analyse mitotic recombination events include resequencing DNA repair deficient organisms (Ries et al., 2000), reporter sequences (Bulankova et al., 2021), induction of DNA damage and site-specific nucleases (Pâques and Haver et al., 1999; Mestiri et al., 2014). Genome modifications through mitotic recombination require a balance between adaptive instability and maintain chromosome integrity (Gusa and Jinks-Robertson, 2019). As reviewed in the previous section, HR repair can lead to crossover (CO) events and in diploid organisms this can lead to LOH in alleles. LOH events can vary by the length of sequence affected. In Figure 1.6 (from Gusa and Jinks-Robinson, 2019), two homologous regions are shown, one with dominant alleles (A, B, C and D) and the other holds recessive alleles (a, b, c and d). A DSB in close proximity to the B allele repaired through classic DSBR will result in a gene CO creating a region of interstitial LOH (Figure 1.6). However, this can also lead to a reciprocal CO event during replication resulting in one daughter cell with only dominant alleles beyond the DSB repair site and the other recessive. Since this affects the region from the break site to the end of the chromosome it is called terminal LOH (Figure 1.6). Lastly, BIR can result in the copying event that can extend across the repair template. The repair template replaces the damaged DNA and there is only LOH in one of the daughter cells (Figure 1.6; Gusa and Jinks-Robinson, 2019).

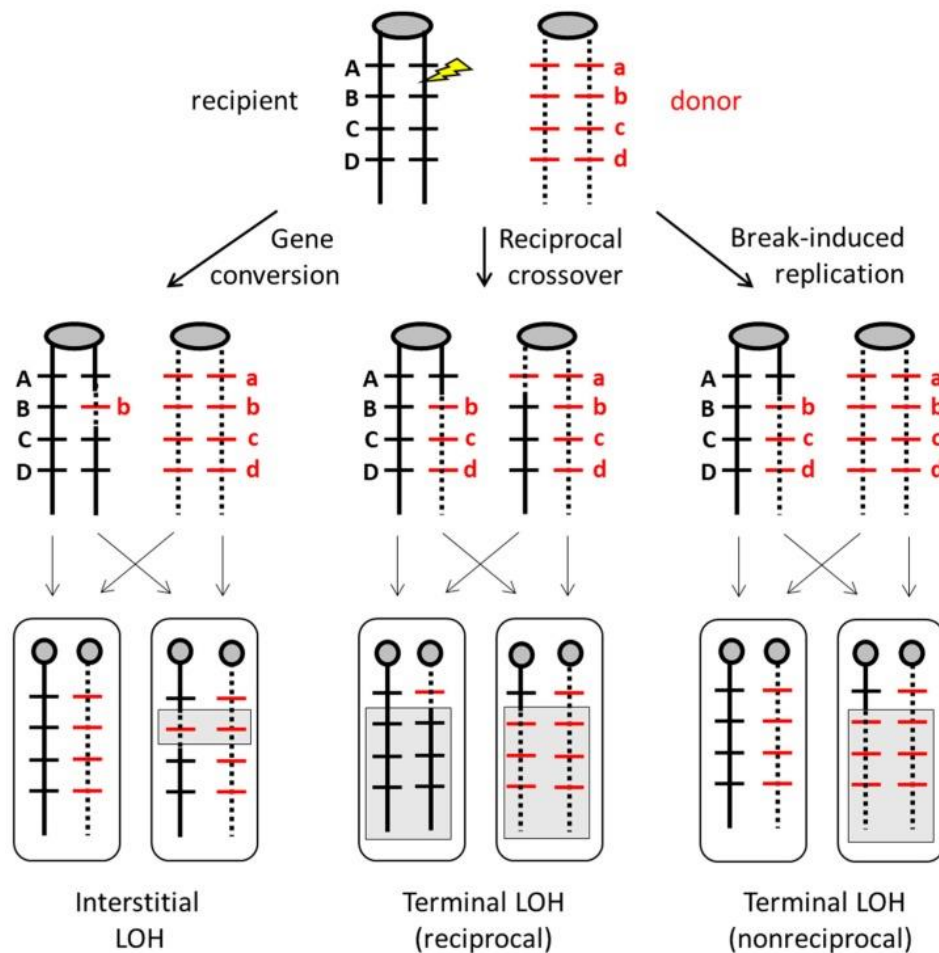


Figure 1.6. Loss of heterozygosity outcomes as a result of repair of a DSB through homologous recombination. Each oval is a centromere with the line representing sister chromatids. The DSB is represented by the yellow lightning bolt. The recipient is on the left and donor on the right. Black strands represent dominant alleles (A, B, C, and D) and red indicates recessive allele (a, b, c, and d). Thin vertical and diagonal lines show sister chromatid segregation during mitosis. LOH is represented by grey boxes in final chromosome pairings (Figure is from Gasu and Jinks-Robertson, 2019).

The requirement for a homologous sequence for HR means that the majority of HR events use a sister chromatid during replication, however, long stretches of repeating sequences are common in higher eukaryotes, especially in diatom genomes (Filloramo et al., 2021). These repeat sequences can sometimes be used by HR as a template since only a region of only 100bp has been reported to be considered homologous (Jinks-Robertson et al., 1993). Recombination between repeat sequences (ectopic recombination) usually does not alter allele frequency but it can create hybrid versions of genes potentially creating proteins with altered functions.

The understanding of the role of mitotic recombination and generation of genetic diversity in diatoms is currently an emerging research area. However as mentioned previously, diatoms display high genetic diversity even within clonal populations despite prolonged periods of asexual division (Godhe and Rynearson, 2017). A recent paper by Bulankova (et al., 2021) was the first to characterise the frequency of mitotic recombination in clonal model diatom species. They showed that the rate of mitotic recombination increased under environmentally stressful conditions, suggesting plasticity in the response of the pathway. This study put mitotic recombination in a new perspective concerning the adaptability of diatoms undergoing clonal reproduction (Bulankova et al 2021). This thesis presents novel data generated from the first HR-deficient diatoms, these results are discussed in the context of furthering the understanding of mitotic recombination in diatoms through resequencing and phenotypic analysis.

### 1.3 Thesis Outline and Aims

Diatoms are such an important and diverse group of organisms that impact all aspects of marine and freshwater ecosystems in which they are present. With the planet undergoing significant changes in climate and habitat connectivity it is crucial to further our understanding of how diatoms adapt to changing environmental conditions. In addition, diatoms can be used as a model organism to understand genome evolution in an environmentally relevant species. This thesis contributes to this effort by investigating the following overarching research question:

*What are the underpinning molecular mechanisms by which diatoms adapt under selective pressures?*

To further understand the molecular underpinnings of adaptive evolution in diatoms I, along with my supervisory team, selected a single pathway – homologous recombination – to study through whole-genome sequencing of CRISPR/Cas genetically edited knock-out cell lines. This thesis describes the methods and analysis to first identify predicted DNA repair orthologous in diatoms to select appropriate genes to study these pathways, secondly to knock out core DNA repair genes involved in repair of DNA DSBs and finally analyse whole-

genome sequencing data to reveal both long and short-term genomic changes which resulted from the loss of a core DNA repair pathway – homologous recombination.

### 1.3.1 Chapter 2: Materials and Methods

Chapter 2 outlines the materials and methods used throughout this thesis. These include methods for cell culture, phenotyping, CRISPR/Cas cloning and transformations, assay development and DNA extraction and whole-genome sequencing analysis. Any modifications to these methods (when used in more than one chapter) are detailed in the respective data chapter.

### 1.3.2 Chapter 3: DNA Repair Systems in Diatoms

This chapter provides evidence of the presence of core and accessory DNA repair orthologs in available diatom reference proteomes. This chapter aimed to understand which genes diatoms possess from each of the major DNA repair pathway to select genes for CRISPR/Cas genome editing in later chapters.

### 1.3.3 Chapter 4: Confirmation of *Brca2* Function in Diatoms and Long-term Consequences of Knock-out

Chapter 4 presents the results of genome editing via CRISPR/Cas to knock out the genes *Brca2* and *Ku70* in *T. pseudonana*. After screening and selection of transformed cell lines carrying a homozygous knock-out of each gene, the cell lines were first tested for evidence of hypersensitivity to DNA damage and the ability to grow under temperature stress. After growth under regular conditions for 10 months (approximately 200 – 250 generations) three independent *Brca2*  $-/-$  cell lines were re-sequenced and the whole-genome data was analysed to reveal genomic changes after 10 months of regular growth without *Brca2*.

### 1.3.4 Chapter 5: What are the mechanisms of Adaptation in HR Deficient *T. pseudonana*.

The experiment in this chapter was designed to reveal the initial molecular mechanisms employed by *T. pseudonana* to survive without *Brca2*. The design and analysis of this chapter were driven by the results from Chapter 4 and the evidence of global genomic instability in *T. pseudonana* grown under regular growth conditions for 10 months

(approximately 200 – 250 generations) without a functional BRCA2 protein. *Brca2*  $-/-$  and wild-type cell lines were grown for  $40 \pm 4$  generations under either regular growth conditions or induced DNA damage stress. DNA damage was created through regular exposure to the alkylating agent methyl methanesulfonate (MMS) diluted in dimethyl sulfoxide (DMSO). MMS causes replication errors which lead to DNA damage, including DSBs. Cultures under regular growth conditions were exposed to DMSO as a control. All cell lines were grown in triplicates Cell lines were resequenced before the experiment (T0), after  $20 \pm 2$  generation (T1) and finally at  $40 \pm 4$  generations. Whole-genome sequencing data was analysed with similar methods used in Chapter 4 to further the understanding of the initial mechanisms of adaptation in *T. pseudonana* with impaired HR with or without induced DNA damage stress.

### 1.3.5 Chapter 6:

Lastly, chapter 6 concludes with a summary of major findings from this research and a general discussion of the main results followed by suggestions for future research.

## 1.4 Diatoms and Their Microbiomes in Changing Polar Oceans

This final section of the review contains a peer reviewed review paper jointly written between me, Dr Emma Langan and Prof Thomas Mock. The review paper reflects on both the unique adaptations of polar diatoms and traits that have co-evolved across the polar microbiome, summarising the most current multi-disciplinary work. These data are then put into context of a region undergoing rapid disruption due to climate change and discusses how present adaptations can help researchers understand how these crucial communities will fare in the future and highlight the adaptability of polar eukaryotes. This paper has been included because it adds insights into the mechanisms behind the chimeric nature of diatom genomes and their high levels of diversity.





# Diatoms and Their Microbiomes in Complex and Changing Polar Oceans

Reuben Gilbertson<sup>1</sup>, Emma Langan<sup>1,2</sup> and Thomas Mock<sup>1\*</sup>

<sup>1</sup> School of Environmental Sciences, University of East Anglia, Norwich Research Park, Norwich, United Kingdom, <sup>2</sup> The Earlham Institute, Norwich Research Park, Norwich, United Kingdom

## OPEN ACCESS

### Edited by:

Anne D. Jungblut,  
Natural History Museum,  
United Kingdom

### Reviewed by:

Mark Moore,  
University of Southampton,  
United Kingdom  
Caroline Chénard,  
National Research Council Canada  
(NRC-CNRC), Canada

### \*Correspondence:

Thomas Mock  
t.mock@uea.ac.uk

### Specialty section:

This article was submitted to  
Extreme Microbiology,  
a section of the journal  
Frontiers in Microbiology

**Received:** 30 September 2021

**Accepted:** 23 February 2022

**Published:** 25 March 2022

### Citation:

Gilbertson R, Langan E and  
Mock T (2022) Diatoms and Their  
Microbiomes in Complex  
and Changing Polar Oceans.  
*Front. Microbiol.* 13:786764.  
doi: 10.3389/fmicb.2022.786764

Diatoms, a key group of polar marine microbes, support highly productive ocean ecosystems. Like all life on earth, diatoms do not live in isolation, and they are therefore under constant biotic and abiotic pressures which directly influence their evolution through natural selection. Despite their importance in polar ecosystems, polar diatoms are understudied compared to temperate species. The observed rapid change in the polar climate, especially warming, has created increased research interest to discover the underlying causes and potential consequences on single species to entire ecosystems. Next-Generation Sequencing (NGS) technologies have greatly expanded our knowledge by revealing the molecular underpinnings of physiological adaptations to polar environmental conditions. Their genomes, transcriptomes, and proteomes together with the first eukaryotic meta-omics data of surface ocean polar microbiomes reflect the environmental pressures through adaptive responses such as the expansion of protein families over time as a consequence of selection. Polar regions and their microbiomes are inherently connected to climate cycles and their feedback loops. An integrated understanding built on “omics” resources centered around diatoms as key primary producers will enable us to reveal unifying concepts of microbial co-evolution and adaptation in polar oceans. This knowledge, which aims to relate past environmental changes to specific adaptations, will be required to improve climate prediction models for polar ecosystems because it provides a unifying framework of how interacting and co-evolving biological communities might respond to future environmental change.

**Keywords:** diatoms, psychrophile, genomics, meta-omics, polar winter, climate change, adaptive evolution, microbiomes

## INTRODUCTION

Polar oceans are major drivers of the global carbon pump, circulation of nutrients and reflection of solar radiation (Murata and Takizawa, 2003; Sigman et al., 2010; MacGilchrist et al., 2014; Laufkötter et al., 2018; Wadham et al., 2019). Despite the global importance of both the geophysical and biological aspects of the polar oceans, they are critically understudied. In addition to



consistently low temperatures, polar oceans comprise a multitude of stressors such as extremes in solar irradiance, salinity and UV radiation (Boyd, 2002). While the Arctic and Antarctic have similar climates due to the high latitudes, each ecosystem is characterized by unique geography which shapes the local environment. The Arctic Ocean is surrounded by large shelf areas connected to land masses and annually undergoes dramatic changes in the extent of sea ice. Conversely, the Southern Ocean is a merry-go-round system characterized by strong latitudinal gradients of temperature (Oceanic fronts) isolating the Antarctic continent (Hansom and Gordon, 1998; Maksym, 2019). These fronts have recently been discovered to be responsible for ecotypic differentiation and speciation in the endemic and pelagic Southern Ocean diatom *Fragilariopsis kerguelensis* (Postel et al., 2020). Hence, they contribute to creating and maintaining diatom biodiversity in the Southern Ocean. Wherever there are fewer barriers such as in the Arctic Ocean, there is likely more exchange (e.g., gene flow) between microbial populations. Their terrestrial landscapes also differ. For example, higher vascular plants have colonized around 75% of the Arctic (Walker et al., 2005; Iversen et al., 2015), compared to a limited number of plant species across the Antarctic (Pointing et al., 2015; Singh et al., 2018). Despite these differences, most of the primary productivity and nutrient cycling for both ecosystems takes place in marine microbial communities inhabiting their associated oceans (Nelson et al., 1987; Smythe-Wright et al., 2010).

Polar microbiomes, mainly composed of diverse microalgae and their associated prokaryotes, are of particular importance in polar primary productivity and nutrient cycling and are the base of the highly productive polar food webs (Stretch et al., 1988; Harrison and Cota, 1991; Boyd, 2002; Bluhm and Gradinger, 2008; Aslam et al., 2012; Petrou et al., 2014; Hayward and Grigor, 2020). Diatoms are the most abundant and diverse group of eukaryotic polar phytoplankton and are key components of both pelagic and sea-ice habitats (Thomas and Dieckmann, 2002; Kooistra et al., 2007; Armbrust, 2009; Bracher et al., 2009; Tréguer et al., 2018; Trefault et al., 2021), as they thrive in seasonally mixed, cold and nutrient-rich water, characteristic of the polar oceans. Thus, polar oceans are their preferred environment where they outcompete many other phytoplankton groups, at least under past and current climatic conditions. Reasons for the dominance of diatoms in polar oceans include: (a) the ability to “boom and bust”, which matches the response required to thrive under extreme seasonality, (b) an increased abundance of genetic elements that enhance an adaptive response (i.e., transposable elements), (c) differential allelic expression and (d) multiple sources of genes from endosymbiotic (EGT) and horizontal gene transfer (HGT), both of which contribute to metabolic plasticity and therefore facilitate specific adaptations (Martin et al., 1998; Sjöqvist and Kremp, 2016; Mock et al., 2017). Given their prevalence in polar phytoplankton microbiomes, initial research focused on understanding their physiological mechanisms for survival under extreme conditions. The advent of genomics tools and methods has influenced later research with a focus on identifying the underpinning genes and pathways (Sackett et al., 2013; Palenik, 2015).

Molecular dating of the emergence of adaptations and radiations in polar species coincide with major geological events, such as the opening of the Drake Passage (~35 Mya), which resulted in the Antarctic Circumpolar Current, and subsequent isolation and freezing of Antarctica (Suto et al., 2012; Benoiston et al., 2017). Generally, between the Eocene and Oligocene (~34 Ma), the Earth cooled, and changes in upwelling patterns created favorable conditions for diatoms, especially in the Southern Ocean, where the evidence of the rise and sustained dominance of diatoms in the water column is found in large siliceous ooze deposits in the area (Salamy and Zachos, 1999; Dutkiewicz et al., 2015; Benoiston et al., 2017). All microbial life was affected by this global cooling, including the divergence of polar clades of *Chlamydomonas sp.* from temperate lineages (Zhang et al., 2020), the Atlantic *Chaetoceros* Explosion (ACE), and global radiation in soil diatoms (Suto et al., 2012; Pinseel et al., 2020). It has been suggested that the rise in diatom abundance and subsequently primary production, specifically *Chaetoceros*, during this time enabled organisms in higher trophic levels such as zooplankton and marine mammals to thrive and diversify (Suto et al., 2012). These data highlight that major changes in marine microbial communities are inextricably linked to climate cycles and can have far-reaching implications on the rest of the food chain (Falkowski, 1998; Suto et al., 2012). Just as diversification events such as the ACE can be observed in fossil records, the genetic code of marine organisms can show how they adapted and evolved to survive changing climates. Understanding the evolution of diverse polar diatom genomes and their associated microbes can therefore provide insight into past and current climate conditions, potentially improving models based on either traits or functional types, especially due to the rapid nature of climate change in polar ecosystems (Kwok, 2018). It appears that warming polar oceans have so far been generally beneficial for polar phytoplankton populations as consistently increasing temperatures have extended the season in which polar phytoplankton can grow but have so far been small enough to prevent significant encroachment of invasive species (Arrigo et al., 2008; Erwin, 2009). Increasing temperatures are generally resulting in prolonged stratification and growing seasons, ocean acidification and reduced upwelling creating conditions which some phytoplankton are unable to adapt to, opening niches for invasive species (Vincent, 2010; Boyd et al., 2016). However, the polar oceans are continuing to warm and the wider effect this will have on diverse polar microbiomes in the future is not well known. Unlike diatom fossil records there are no historical databases built on omics data, without a reference of the current diversity of polar microbiomes and associated functional traits, our ability to predict changes is limited. To combat this, we suggest an increase in metaomics sequencing of environmental communities from the polar regions to provide a background of current populations to relate future changes to.

However, evolutionary genomics with polar microbes is still in its infancy but will be necessary to improve predictions of key species' responses to climate change (Waldvogel et al., 2020) which could have significant effects on the food-web structure and biogeochemical cycles of elements, as seen during past major geological events (Suto et al., 2012). Even intraspecific changes



in biodiversity will have knock-on effects on the carbon cycle, as the changing polar ocean likely selects for different strains (e.g., ecotypes) with different traits impacting food web dynamics and the cycling of elements especially if keystone groups such as diatoms are affected (Field et al., 1998; Wolf et al., 2019). Thus, forecasts for how this rapid climate change will impact polar ecosystems remain incomplete unless we improve our understanding of how the polar environment has shaped the evolution and biodiversity of polar organismal communities (Bindoff et al., 2007; Steig et al., 2009; Lee, 2014; Mock et al., 2016; Murphy et al., 2016; Verde et al., 2016; Brown et al., 2019).

Consequently, to address the uncertainty of the effect future climate change will have on biodiversity change and loss in polar marine ecosystems, integrative approaches are required. We think they should be based on sequencing data because they provide comprehensive insights into microbial functional diversity and how this diversity might change due to selection driven by climate change. However, our current molecular knowledge about polar microbes is very limited because of (a) lack of diverse polar model species for cell biology and therefore fundamental insights into their biochemical adaptation and molecular evolution, and (b) a very limited number of environmental sequencing initiatives to reveal genetic and genomic biodiversity of polar microbes. The aim of this review paper, therefore, is to reflect on what we have learned so far from the omics resources currently available from limited model psychrophilic microalgae, specifically *Fragilariopsis cylindrus*, and their associated microbiomes in the context of future advances exemplified by developments in neighboring fields such as plant sciences, microbiome research, and macroecology. In addition to highlighting some of the latest advances in polar diatom molecular biology, we provide suggestions as to how to build bridges to neighboring disciplines to fill gaps in our knowledge and therefore to advance our field for tackling the challenges mentioned above.

### Using *Fragilariopsis cylindrus* as a Model to Reveal How Marine Phytoplankton Are Adapted to the Polar Climate

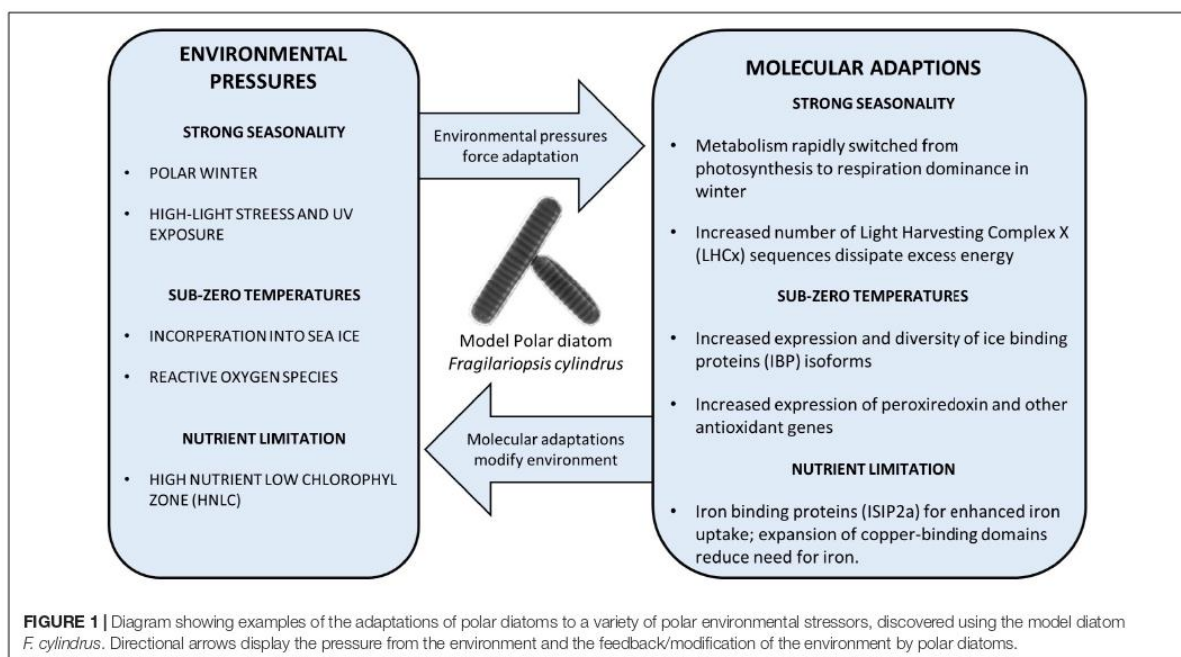
Although diatoms are the most species-rich group of algae, diatom research has immensely benefited from model species such as *Thalassiosira pseudonana* and *Phaeodactylum tricorutum* (Falcione et al., 2020). However, due to limitations given by the nature of both models (e.g., small and streamlined genomes, limited evidence for sexual reproduction, mesophiles, in culture for many decades), additional diatom species have been developed into models. One of them is *F. cylindrus*, which is the first and only eukaryotic psychrophile so far that is genetically tractable (Faktorová et al., 2020) with a sequenced genome, and grows well under laboratory conditions (Mock et al., 2017). Hence, it has been used to study physiological and molecular adaptation to polar conditions. Significant seasonality in light and the overall low temperatures are amongst the strongest selecting agents shaping the evolution and adaptation of *F. cylindrus* and likely other polar diatoms. In this section of the review paper, we will focus on molecular data gained from

the sequencing of polar diatoms, mainly *F. cylindrus*, and how this molecular data compares to that from other psychrophilic microbes, and mesophilic model diatom species.

### The Dark Polar Winter

In most temperate and polar aquatic environments, especially coastal waters, diatoms are typically one of the first taxa groups to bloom with the changing seasons by taking advantage of renewed nutrients and increased photosynthetically active radiation (PAR) (Soreide et al., 2010; Kvernvik et al., 2018). Polar diatoms are no exception, due to mechanisms allowing them to survive the polar winter with the necessary cellular components ready and waiting for the return of spring (Peters and Thomas, 1996; van de Poll et al., 2020). However, the molecular mechanisms underpinning these adaptations have been a mystery. Polar phytoplankton have been documented to employ a diverse set of strategies, such as forming resting spores, heterotrophy and respiration of stored lipids (Bunt and Lee, 1972; Reeves et al., 2011; McMinn and Martin, 2013; Schaub et al., 2017). The traditional view of a pause in biological activity has recently been challenged as more data during the winter months is collected, indicating that there is stable biological activity (Berge et al., 2015). This is suggesting a diverse polar microbial community with multiple mechanisms of survival in prolonged darkness. The addition of detailed transcriptomics and proteomics data from *F. cylindrus* cultured in complete darkness has enabled researchers to explore genetic variation, and how it could explain the physiological activity in prolonged darkness. Within 24h of the onset of darkness, there was a significant reduction in the abundance of light-harvesting protein complexes (LHCs), which transfer light energy to the photosystem reaction center (Kennedy et al., 2019). Despite the sudden widespread downregulation in the abundance of LHCs, these complexes are maintained at a low basal level in the absence of light. This coincided with the upregulation of proteins affiliated with glycolysis, the TCA cycle and the Entner–Doudoroff pathway, suggesting a coordinated shift from photosynthesis to cellular respiration to maintain essential functions for survival (Figure 1; Kennedy et al., 2019). This large response in the proteome corroborates data on differential expression in the *F. cylindrus* transcriptome in response to prolonged darkness (Mock et al., 2017). Dynamic regulation of the proteome and transcriptome to significant changes in light intensity does not appear to be exclusive to *F. cylindrus*. For example, the model psychrophile *Chlamydomonas raudensis* and the natural community around it, also maintained photosystem complexes at a basal level when cultured in dialysis bags in Lake Bonney (McMurdo Dry Valleys, Antarctica) during the transition to polar night conditions (Morgan-Kiss et al., 2016). Despite this study focusing on a different ecosystem and a distantly related species from diatoms, the same molecular response was discovered, suggesting co-adaptation to the polar climate across distinct phytoplankton communities.

After months of darkness, the sea-ice thickness begins to decline in the spring and more light can reach diatom populations in the surface ocean eventually becoming sustained high light during the summer months. The *F. cylindrus* genome encodes significantly more LHCx genes, a subset of LHC



sequences that are induced by high-light stress, than the model temperate diatoms *T. pseudonana* and *P. tricornutum*. The evolutionary expansion and upregulation of LHCx genes, known to reduce high-light stress, suggests their key role in the adaptive evolution to the polar light climate (Figure 1; Mock et al., 2017). Interestingly, the LHCx gene family is also expanded in the mesophilic alga *Aureococcus anophagefferens*, enabling this species to rapidly form blooms in high-light coastal environments (Gobler et al., 2011; Mock et al., 2017). Thus, to be prepared for high-light environments after a period of darkness, polar diatoms appear to be downregulating a subset of photosynthesis-associated genes accompanied by increasing cellular respiration. This combination of acclamatory metabolic processes enables them to survive the polar winter in an active state (Berge et al., 2015) and take full advantage of the spring and summer high-light conditions, especially in open water by upregulation of LHCx genes. These results show the usefulness of genomic data in researching how phytoplankton adapted to the physical polar environment, and the extent to which adaptations are possibly unique to specific taxa groups, giving more insight into the potential impact climate change will have on diverse polar phytoplankton communities.

## Freezing Ice-Binding Proteins

Ice-binding proteins (IBPs) are molecules that act on the interface between ice and water (Dolev et al., 2016). Their activity can reduce the freezing point and inhibit ice-crystal growth altogether. IBPs have been found in a diverse array of taxa across the world from insects to fish, with the majority found in species

living in Arctic and Antarctic ecosystems such as oceans, frozen lakes and glaciers (DeVries and Wohlschlag, 1969; Raymond, 2014; Jung et al., 2016; Vance et al., 2019). The structure of IBPs is diverse throughout the domains of life, but the general function is conserved, suggesting multiple independent origins (Dolev et al., 2016).

Many diatom species are incorporated into the sea ice each year (Eicken, 1992; Thomas and Dieckmann, 2002). When frozen into forming sea ice, diatoms are enclosed into interconnected channels and pockets filled with brine of high salinity. To maintain an aqueous habitat inside sea ice, polar diatoms are known to produce IBPs and extracellular polymeric substances (EPS) to influence the physical state of their surrounding icy environment (Hoagland et al., 1993; Bayer-Giraldi et al., 2010; Lyon and Mock, 2014; Raymond, 2014; Arrigo et al., 2017; Aslam et al., 2018; Liang et al., 2019; Vance et al., 2019). There is evidence that the IBPs they secrete help maintain the aqueous state of the brine channel system to ensure access to nutrients from the seawater underneath through diffusive and convective transport processes (Raymond et al., 2009; Dolev et al., 2016). One interesting aspect of the functional conservation of IBPs is the Domain of Unknown Function 3494 (DUF3494), which has been identified in a large range of taxa in a variety of cold habitats. Despite significant sequence divergence, all homologs share the biophysical ability to reduce the growth of ice crystals (Vance et al., 2019; Raymond et al., 2021). Phylogenetic analysis of IBP sequences does not correlate with 18S-based phylogeny, which suggests that horizontal gene transfer (HGT) may have been a vector for the DUF3494's widespread presence (Keeling and Palmer, 2008; Bayer-Giraldi et al., 2010; Raymond and Kim, 2012; Mock et al., 2017; Raymond and Morgan-Kiss, 2017; Vance et al.,







2019). HGT in sea-ice communities is considered to be facilitated by the proximity of organisms in the narrow brine-channel system in combination with a strong selection pressure imposed by the harsh environmental conditions such as high salinities and subfreezing temperatures (Raymond et al., 2009; Raymond and Kim, 2012). Their important role in the adaptive evolution to the sea-ice habitat is supported by the fact that they usually expand (e.g., gene duplications) once they have been acquired via HGT (Mock et al., 2017; Raymond and Morgan-Kiss, 2017). Hence, many polar organisms have more than a single IBP gene encoded in their genomes such as *F. cylindrus* and *Chlamydomonas sp.* ICE-L.

The genome of *F. cylindrus* encodes 11 unique IBP isoforms, several of which are significantly upregulated under freezing temperatures and elevated salinity (Figure 1; Mock et al., 2017). Isoform 11 from *F. cylindrus* (FcIBP11), despite having only moderate activity, can bind to multiple planes of ice crystals (Kondo et al., 2018). In addition to FcIBPs containing the DUF3494, predicted to inhibit ice crystallization, several proteins such as FcIBP-1, have transmembrane domains and are therefore hypothesized to protect the cell membrane from ice crystals (Mock et al., 2017). However, alternative roles are equally likely, such as sensing the formation of ice crystals and therefore initiating the appropriate physiological response to mitigate the impact of ice-crystal formation on the integrity of cellular structures.


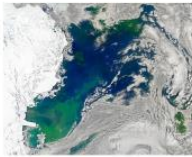
Encoding multiple isoforms of antifreeze proteins to a variety of ice planes is evident in other polar eukaryotes, such as the psychrophilic yeast *Glaciozyma antarctica* (Figure 2) and the green algae *Coccomyxa subellipsoidea* (Figure 2; Blanc et al., 2012; Firdaus-Raih et al., 2018). The genome of the psychrophilic green alga *Chlamydomonas sp.* ICE-L, the first polar green algae sequenced, also contains expanded IBP gene families. All 12 IBPs in the genome of *Chlamydomonas sp.* contain the DUF3494, and again these sequences show a closer phylogenetic relationship to that of bacteria (Zhang et al., 2020). Polar diatoms and the wider microbiome produce EPS to work in synergy with IBPs by creating disorder in ice-crystal formation leading to retention of salinity, effectively reducing the freezing point (Ewert and Deming, 2011; Krembs et al., 2011; Mykytczuk et al., 2013; Aslam et al., 2018; Raymond-Bouchard et al., 2018; Torstensson et al., 2019). By having multiple IBP isoforms with a diverse range of functions and applications to different conditions, the cell appears to be able to sense the icy environment when it forms, which likely induces diverse protection mechanisms in which IBPs play a key role.

### Radical Oxygen Species and the Role of Trace Metals

Low temperatures generally reduce enzyme kinetics. Because of slow kinetics, excess free radicals accumulate within the cell causing dangerous levels of oxidative stress which can seriously damage cellular components, including DNA and membranes (Wingsle et al., 1999). As mentioned above, *F. cylindrus* displays a divergent expression of homologous alleles under varying environmental conditions. The most divergently expressed allele set across all tested conditions, including freezing temperatures, is glutathamine s-transferase I (GST-1), a scavenger for free

A Emerging eukaryotic model species		Available data / methods
<i>Fragilariopsis cylindrus</i> (Mock et al., 2017)		<ul style="list-style-type: none"> <li>Genome [epigenome], comparative genomics</li> <li>transcriptome, proteome, genetic editing</li> </ul>
<i>Polarella glacialis</i> (Stephens et al., 2020)		<ul style="list-style-type: none"> <li>Genome, transcriptome, SL RNA Transcripts, marker genes</li> </ul>
<i>Chlamydomonas sp.</i> UW0241 (Zhang et al., 2021)		<ul style="list-style-type: none"> <li>Genome, transcriptome, comparative genomics</li> </ul>
<i>Glaciozyma antarctica</i> (Firdaus-Raih et al., 2018)		<ul style="list-style-type: none"> <li>Genome, transcriptome, comparative genomics</li> </ul>

B Natural microbial communities		Available meta omics data
Sea-ice communities		<ul style="list-style-type: none"> <li>Metatranscriptomics: Arctic and Antarctic</li> <li>Metagenomics: Arctic</li> <li>16/18S rDNA amplicons: Arctic and Antarctic</li> </ul>
Pelagic communities		<ul style="list-style-type: none"> <li>Metatranscriptomics: Arctic and Antarctic</li> <li>Metagenomics: Arctic and Antarctic</li> <li>16/18S rDNA amplicons: Arctic and Antarctic</li> <li>Metaproteomics: Antarctic</li> </ul>

**FIGURE 2 | (A)** Non-comprehensive figure showing selected model species of core polar taxa groups and a summary of available data/methods for each model system. **(B)** Polar ecosystems and available meta-omics datasets for each.

radical oxygen species (ROS). Thus, this enzyme is a good example that diverged expression of alleles might be driven by environmental pressures (Mock et al., 2017). Similarly, *F. kerguelensis* overexpressed transcripts for peroxiredoxin (PrxQ) to inhibit a diverse set of peroxides (Moreno et al., 2020). Similar responses across other psychrophiles include growth changes, upregulating expression of other antioxidant proteins such as specific catalases, overexpression of antifreeze proteins, and superoxide dismutases (Wong et al., 2019; Zhang et al., 2020). The latter converts superoxide radicals to more stable species (Miteva-Staleva et al., 2011; Scholz et al., 2014; Raymond-Bouchard et al., 2018). These data suggest that

multiple molecular strategies have evolved to cope with common environmental stressors in polar ecosystems.

The low concentrations of micronutrients, specifically iron, compared with macronutrients in the Southern Ocean results in a High Nutrient Low Chlorophyll region (HNLC) which limits phytoplankton growth (Martin et al., 1991; Pitchford and Brindley, 1999; Venables and Moore, 2010; Hassler et al., 2012). Thus, iron limitation can impose cellular stress. For example, iron limitation compounds the intracellular ROS stress due to a reduction in the capacity of the electron transport chain (Allen et al., 2008). Genomics, transcriptomics, and proteomics studies have discovered unique proteins which are responsive to micronutrient fluxes to mitigate this stress such as through FLDA2b in *F. kerguelensis* (Allen et al., 2008; Marchetti et al., 2009; Lommer et al., 2012; Bender et al., 2014; Marchetti, 2019; Moreno et al., 2020). Southern Ocean diatoms exposed to varying light intensities, and different iron concentrations, were found to have an increased number of genes associated with iron uptake compared to temperate diatoms (Mock et al., 2017; Moreno et al., 2018). Iron concentrating proteins in polar diatoms, such as iron starvation-induced protein 2a (ISIP2a) identified in *F. kerguelensis*, are significantly overexpressed in low iron conditions to increase uptake of iron from the environment (Marchetti et al., 2009; Moreno et al., 2020). In addition to increased affinity for iron, the *F. cylindrus* genome is enriched for copper-binding domains, outnumbering iron-binding domains (Mock et al., 2017). Most are of the plastocyanin/azurin-like family, possibly reducing the requirement for iron in cellular processes (Peers and Price, 2006; Mock et al., 2017). *F. kerguelensis* contains several plastocyanin isoforms which are significantly upregulated in response to low iron, specifically PCYN-2b (Moreno et al., 2020). The *F. cylindrus* genome also revealed an elevated number of zinc-binding domain-containing genes, with more than six times the number of homologous clusters compared to *P. tricornutum* and *T. pseudonana*; specifically, MYND Zinc Fingers. MYND domains in *F. cylindrus* are associated with a variety of accompanying domains with diverse functions, many of which are unknown (Laity et al., 2001; Mock et al., 2017). The evolutionary expansion of this family suggests their importance potentially in terms of regulating the activity of gene expression and/or the activity of enzymes (Mock et al., 2017). Their expansion may have been facilitated by the relatively high zinc concentrations in the Southern Ocean (Crook et al., 2011; Mock et al., 2017). Most of the expansion in this family is predicted to have originated around 30 million years ago, which therefore coincides with the opening of the Drake passage (Mock et al., 2017). Thus, the coincidence between the evolutionary expansion of zinc-binding protein genes and the cooling of the Southern Ocean begs the question about the role of zinc in the adaptation of diatoms to environmental conditions in surface polar oceans.

The studies discussed so far in this review have given insight into the viability and usefulness of polar model organisms to reveal fundamental processes underpinning adaptation and evolution, however, relying on model organisms limits our ability to understand how biodiversity changes due to environmental change. For instance, the role of microbial interactions is neither

well studied nor represented by using a monoculture of an individual model species (Wolf et al., 2019) because intra- and interspecific competition is based on complex dynamic populations and their communities shaping the evolution and adaptation of all microbes involved. For example, in the study by Wolf et al. (2019), different strains of the species *Thalassiosira hyaline* were grown in monoculture under predicted future climate conditions and then mixed with other strains to see if their physiology changed when cultivated altogether. Under elevated temperatures, the fastest-growing strain outcompeted the others when grown in mixed culture. Conversely, under ambient temperatures, its growth rate resembled that of the slowest growing strain (Wolf et al., 2019). This demonstrates that monoculture studies may not be representative of how individual species behave as part of complex communities in nature. Thus, understanding intraspecific variation between strains of the same species is important because differences in traits might shape the structure and function of food webs and therefore biogeochemical cycling of elements (Vance et al., 2019). Hence, differences in the co-occurrence of strains and species in complex microbial communities are important for considering how environmental conditions have shaped these communities (e.g., Martin et al., 2021). To address this question, meta-omics approaches have proven successful in revealing the intricacies of complex microbial communities and their activity.

## POLAR META-OMICS

### Metagenomes and Their Assembled Genomes

Metagenomics, metatranscriptomics, and metaproteomics are emerging as important tools for understanding the molecular basis of adaptation in polar microbes by allowing us to study organisms in their ecosystem and to capture their interactions in the natural environment (McCain et al., 2021). Metagenomics has been used to establish which diatom species are present in certain locations, providing a baseline for future monitoring systems. For example, the Tara Oceans project investigated the global distribution of diatoms and discovered unexpected species in polar oceans (Malviya et al., 2016; Carradec et al., 2018; Obiol et al., 2020). Similarly unexpected was that polar oceans are a particular hotspot of viral diversity with high levels of novel genes (Gregory et al., 2019). Thus, metagenomics helps to reveal the microbial biodiversity in polar oceans and how different it is when compared to metagenomes from non-polar oceans. Metagenomics has also been applied to provide insights into how polar environmental gradients (e.g., sea-ice water interface) influence microbial biodiversity. For instance, a metagenomics survey of microbial communities in the Canadian Arctic found that in sea-ice there was a higher concentration of algal genes and chlorophyll-a compared to seawater underneath, where prokaryotic genes were more dominant. Metagenomic data showed that this was largely due to diatoms. Differences in operational taxonomic units (OTUs), a group of closely related individuals, between sea-ice and sea-water samples indicated that sea-ice dwellers have



different strategies for adaptation than seawater-based microbes. Increased variability was found in the sea-ice communities, perhaps indicative of the adaptability of polar diatoms as a dominant and diverse group of microbes in the sea-ice habitat (Yergeau et al., 2017).

Sequencing of individual species such as *F. cylindrus*, *P. glacialis*, and *Chlamydomonas* sp. UW0241 (Figure 2), has advanced our fundamental understanding of polar algae (Bayer-Giraldi et al., 2010; Mock et al., 2017; Stephens et al., 2020; Zhang et al., 2021). However, their sequencing requires resources and takes time. Yet, they serve as a reference not only for fundamental research but also for analyzing phytoplankton-enriched metagenomes. The latter together with reference genomes therefore will allow for the assembly of novel algal genomes isolated from natural communities including their associated microbiomes. Insights from this work will help us to understand the similarities and differences between genomes from natural communities, rather than relying on model organisms which risks not representatively capturing the diversity of adaptive mechanisms underlying their complex intertwined co-evolution. For instance, targeted metagenomics has been used to establish the evolutionary history of picopyrnesiophyte populations in the North Atlantic, which make up approximately 25% of global pico-plankton biomass. Their dominance was found to be due to high gene density, and genes that are likely to be involved in defense and nutrient uptake (Cuvelier et al., 2010).

Metagenome-assembled genomes (MAGs) of prokaryotes have been produced using large quantities of metagenomic data from both the Arctic and Antarctic (Cao et al., 2020; Royo-Llonch et al., 2021). Increased availability of high-quality MAGs without the requirement for individual culturing and sequencing will allow researchers to study a broader, more representative range of polar microbiomes and compare between species to understand mechanisms for polar survival (Cao et al., 2020; Duncan et al., 2020). MAGs have been used to study polar bacterial and archaeal populations from the Tara Oceans dataset, providing the first compendium of Arctic prokaryotic MAGs including differences in gene enrichment between Arctic and Antarctic populations (Cao et al., 2020; Royo-Llonch et al., 2021).

Although most MAGs from polar oceans are still from prokaryotes, one of the first MAG-based datasets from natural Arctic and Atlantic microbial communities was used to reveal inter-kingdom species associations including microbial eukaryotes (Duncan et al., 2020). Besides the generation of first eukaryotic MAGs from a polar ocean (e.g., diatoms, prasinophytes), this dataset was used to identify metabolism enriched in MAG-based species associations. By identifying which protein families are enriched in selected species, comparisons of associations to a background set of phylogenetically related species known not to have associations shed first insights into shared metabolism, potentially underpinning their biotic interactions. By applying this approach, Duncan et al. (2020) found positive and negative associations between algal and bacterial MAGs including some that were only found in the Arctic (e.g., *Micromonas*

MAG associated with a Gammaproteobacteria MAG). Combined with cell isolations from the same samples, these MAG-based species associations can be empirically tested if the microbial partners can be co-cultivated under laboratory conditions.

## Metatranscriptomes and Metaproteomes

Investigation of the genes expressed across global oceans was carried out as part of the Tara Oceans project (Bork et al., 2015) and the Sea of Change project (Martin et al., 2021). The former has been used to create an ocean atlas of eukaryotic genes from temperate and tropical regions whereas the latter has been used to produce the first pole-to-pole catalog of expressed genes from microalgae and other microeukaryotes. Most of the expressed genes in the Tara Oceans Atlas were novel, indicating that eukaryotic ocean life is under-studied and therefore incompletely understood. Additionally, genes specific to the Southern Ocean were different to those found elsewhere, potentially indicating that adaptation to polar conditions results in a very different gene set to non-polar species (Carradec et al., 2018). The latter insights were corroborated and extended by the Sea of Change project, which revealed that global algal microbiomes can be largely separated into two main groups: polar and non-polar species associations. The same demarcation was found for expressed genes of the algal partners considering their biogeographic distribution from pole to pole based on a combination of sequence co-occurrence analysis and the geographical location of breakpoints in their beta diversity (Difference in species composition between neighboring assemblages) (Martin et al., 2021). Thus, it appears that there are ecosystem boundaries in sub-polar regions of the upper ocean in both hemispheres separating polar from non-polar algal microbiomes.

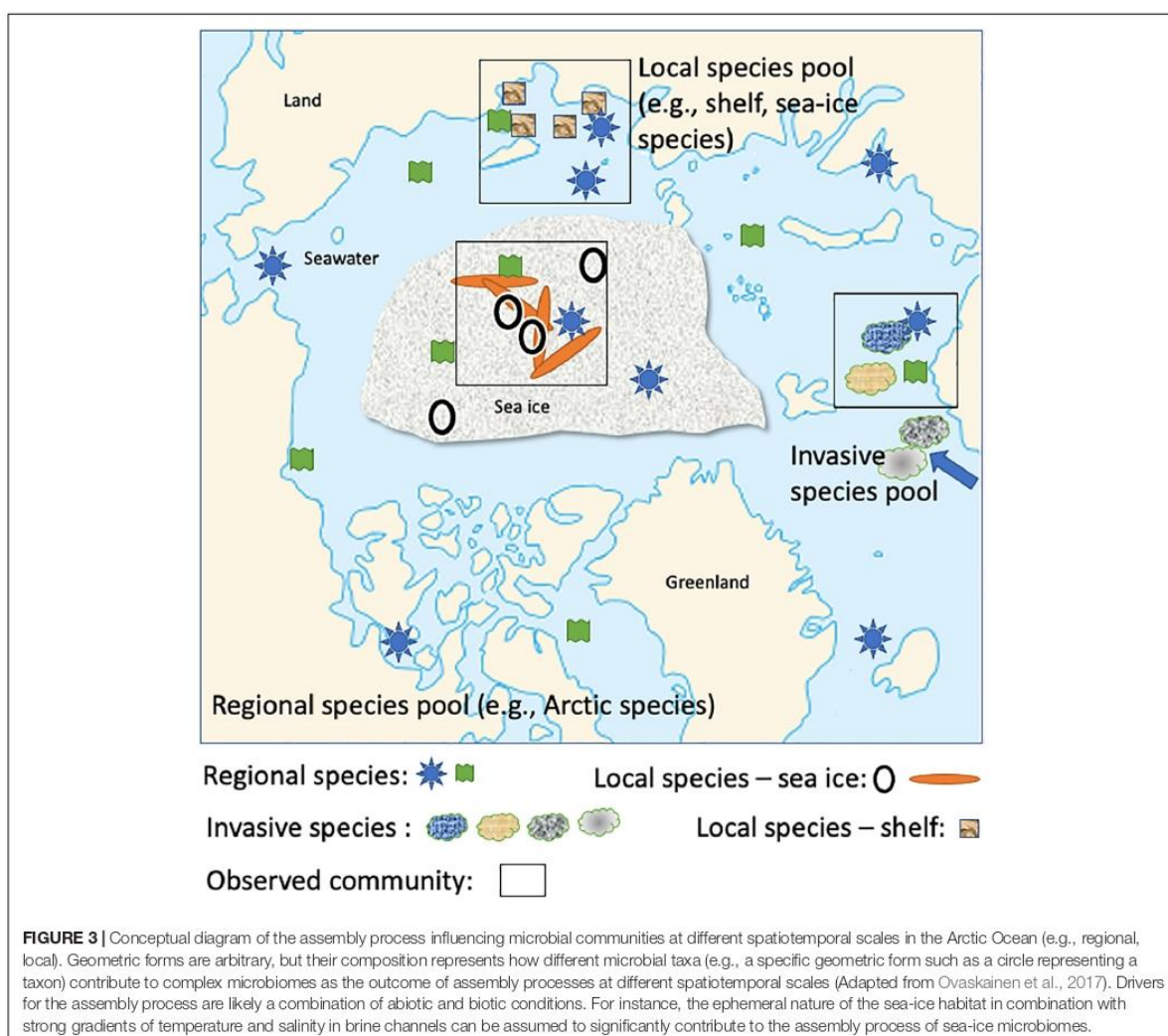
Although metaproteomics with natural polar microbial communities is still in its infancy, the first study on coastal Antarctic phytoplankton communities has been undertaken. This study by McCain et al. (2021) used proteome-level traits to identify ecological strategies for different taxa. For instance, haptophytes appear to have a lower regulatory cost than diatoms, which may explain the observed haptophyte-to-diatom bloom progression in the Ross Sea. As protein synthesis is the main energy sink in cells, the quantity of cellular proteins measured as part of complex metaproteomes relates to costs involved in their synthesis. Hence, the nature and quantity of specific proteins identified potentially provides insights into the costs of traits and trade-offs. Challenges arise, however, with incomplete taxon-specific proteomes as part of complex metaproteomes especially if reference data are missing.

## Omics-Based Monitoring and Modeling to Improve Prediction of Species' Responses to Polar Warming

Genetic monitoring using metagenomics and metatranscriptomics of mixed communities to oversee changes in populations over time can be used to study adaptations

to changing environmental conditions, showing the rate of evolution over time and how adaptations are occurring (Hansen et al., 2012; Pearson et al., 2015). However, sequence-based monitoring of microbial communities and their populations needs to be carefully designed to identify the most appropriate sites for capturing changes in the biogeography of species and their populations including alterations of gene flow and climate-driven range-shift expansions in their distribution. Reoccurring surveys with sufficient spatial sampling in a given geographic area or multiple monitoring sites with frequent sampling will be needed for a rigorous assessment of how warming in polar oceans influences microbial populations and the biogeochemical cycles they drive. Furthermore, monitoring needs to cover complete seasonal cycles to allow us to tease apart long-term ecosystem change from seasonal effects, which might be challenging as the latter are much more pronounced

at high latitudes. Although many different geographical sites have been used for sampling and monitoring polar microbial communities, two stand out: The Fram Observatory in the Arctic and long-term observation of plankton at the Western Antarctic Peninsula (Lin et al., 2021; Wietz et al., 2021). The Fram observatory is based on autonomous samplers, which can sample at discrete time intervals to cover a complete seasonal cycle. Preliminary results covering 12 months including the dark winter period have revealed that the strong seasonal dynamics including the variable sea-ice extent are the main drivers for the succession in changes of microbial biodiversity based on 16 and 18S rDNA data and subsequent analysis (e.g., alpha diversity). Furthermore, the persistent sea-ice cover appears to reduce the seasonality effect on changes in microbial biodiversity (Wietz et al., 2021). As the autonomous samplers can be redeployed, they provide a platform for long-term monitoring of microbial





communities in the Arctic Ocean. A similar monitoring system is not yet, to our knowledge, available for the Southern Ocean, although there is a ship-based multi-year monitoring programme at the Western Antarctic Peninsula (Lin et al., 2021) based on similar methods (e.g., 16/18S rDNA). Modeling has also been shown that sea ice plays a significant role in structuring microbial communities and their productivity (Lin et al., 2021). However, increasing temperatures due to warming in this region are likely to result in declining microbial diversity (Tonelli et al., 2021).

Both of these monitoring programs have provided novel and highly needed insights into how the environmental conditions of the changing polar oceans shape their microbial inventory. To further understand how biotic and abiotic forces determine the abundance and distribution of microbial taxa in a polar ocean, future studies need to provide quantitative insights into how microbial traits (e.g., freeze-thaw resistance) underpin species interactions and how they are associated with habitat characteristics (e.g., sea-ice thickness, oceanic fronts, temperature gradients, snow-cover thickness). These biotic and abiotic forces likely encompass speciation, dispersal, and biotic interactions in a complex and highly dynamic polar environment.

The responses of microbial taxa to different habitat characteristics (environmental filters) and biotic interactions (biotic filters) vary depending on taxa-specific traits including their ability to acclimate, adapt, and their level of competitiveness (Ovaskainen et al., 2017). Thus, these traits will determine which taxa colonize habitats and dominate in seasonal successions (phenology) and therefore contribute to the community assembly process. To address this fundamental question, it is instrumental to apply explanatory models which can integrate data generated for quantitative insights into how microbial communities including their networks relate to environmental conditions and larger-scale ecosystem processes (e.g., seasonality of microbial diversity linked with habitat transformations such as freezing and melting of sea ice) (Ovaskainen et al., 2017; Tikhonov et al., 2020). One approach to tackle this challenge is hierarchical joint species/taxa distribution modeling (HJSDM) (Tikhonov et al., 2020). The heart of this approach is integrating species co-occurrence in co-variation with environmental conditions, and phylogeny of species and traits. Data required for this approach can be extracted from metagenomes, metatranscriptomes, and metaproteomes. For instance, MAGs will provide a genomic catalog (e.g., Cao et al., 2020; Royo-Llonch et al., 2021) of co-occurring microbial taxa in a phylogenetic but also spatio-temporal context if based on long-term sampling. Their genes, transcripts, and proteins provide trait information for revealing co-variation with environmental conditions. If sampling covers larger geographic regions, as done for the West Antarctic Peninsula, or the central Arctic Ocean as part of MOSAiC (Multidisciplinary drifting Observatory for the Study of Arctic Climate) (Wake, 2019), we will be able to delineate processes influencing the biodiversity of microbial communities at different spatio-temporal scales (Figure 3). Generally, a spatio-temporal context is required for separating drivers of seasonal differences vs climate-driven differences in microbial community composition and therefore to make robust predictions as to how warming impacts polar ecosystems.

## DISCUSSION

The Earth's climate has been changing across the globe due to anthropogenic influences since the industrial revolution. There is no doubt of the importance of polar environments to regulate the global climate, with the Southern Ocean alone responsible for the uptake of around 40% of all anthropogenic CO<sub>2</sub> from the atmosphere (Khaliwala et al., 2009; Takahashi et al., 2009; DeVries, 2014; Long et al., 2021). As previously mentioned, polar ecosystems are experiencing more rapid changes, specifically warming, than any other biome on Earth. This alarming shift has been met with increased research over recent decades to understand the physical properties of these environments and the effect they have on food webs, with a focus on marine microbiomes, and diatoms in particular as one of the most important primary producers in polar oceans. Genome-enabled approaches have provided a step-change in our understanding of the adaptability of polar diatoms and allowed us to understand how they reacted to past climate events. However, there is still only one publicly available polar diatom genome, *F. cylindrus*, and with > 350 species of phytoplankton found in the Southern Ocean alone, using one genome to study such a diverse community is not viable. Hence, these studies need to expand as exemplified by the 100 Diatom Genomes Project<sup>1</sup> and the MOSAiC sequencing project at the United States Department of Energy Joint Genome Institute<sup>2</sup>. If monitoring programs build on these initiatives and adopt high-throughput omics approaches, we can generate the foundation for cataloging polar microbiomes through space and time. This information will help to integrate species association networks with traits to provide quantitative insights into how traits of species and their co-occurrence networks are associated with habitat characteristics. Based on this integrated approach, we potentially will be able to predict the future of microbial taxa and their diverse populations with greater confidence in warming polar oceans.

## AUTHOR CONTRIBUTIONS

RG, EL, and TM wrote the article and equally contributed to the design of the figures. All authors contributed to the article and approved the submitted version.

## FUNDING

RG and TM acknowledge funding from The Leverhulme Trust (RPG-2017-364) and the Natural Environment Research Council (NERC) (Grant Nos. NE/K004530/1 and NE/R000883/1). EL acknowledges funding from NEXUSS, a Natural Environment Research Council (NERC) funded Doctoral Training Partnership (DTP). This work also partially funded from the School of Environmental Sciences at the University of East Anglia, Norwich, United Kingdom.

<sup>1</sup> <https://jgi.doe.gov/csp-2021-100-diatom-genomes/>

<sup>2</sup> <https://jgi.doe.gov/csp-2020-arctic-ice-drift-experiment-mosaic/>



## REFERENCES

- Allen, A. E., LaRoche, J., Maheswari, U., Lommer, M., Schauer, N., Lopez, P. J., et al. (2008). Whole-cell response of the pennate diatom *Phaeodactylum tricoratum* to iron starvation. *Proc. Natl. Acad. Sci. U.S.A.* 105, 10438–10443. doi: 10.1073/pnas.0711370105
- Armburst, E. V. (2009). The life of diatoms in the world's oceans. *Nature* 459, 185–192. doi: 10.1038/nature08057
- Arrigo, K. R., van Dijken, G., and Pabi, S. (2008). Impact of a shrinking Arctic ice cover on marine primary production. *Geophys. Res. Lett.* 35:L19603. doi: 10.1029/2008GL035028
- Arrigo, K. R., van Dijken, G. L., Alderkamp, A. C., Erickson, Z. K., Lewis, K. M., Lowry, K. E., et al. (2017). Early spring phytoplankton dynamics in the Western Antarctic Peninsula. *J. Geophys. Res. Oceans* 122, 9350–9369. doi: 10.1002/2017jc013281
- Aslam, S. N., Cresswell-Maynard, T., Thomas, D. N., and Underwood, G. J. C. (2012). Production and characterization of the intra- and extracellular carbohydrates and polymeric substances (EPS) of three sea-ice diatom species, and evidence for a cryoprotective role for EPS. *J. Phycol.* 48, 1494–1509.
- Aslam, S. N., Strauss, J., Thomas, D. N., Mock, T., and Underwood, G. J. C. (2018). Identifying metabolic pathways for production of extracellular polymeric substances by the diatom *Fragilariopsis cylindrus* inhabiting sea ice. *ISME J.* 12, 1237–1251. doi: 10.1038/s41396-017-0039-z
- Bayer-Giraldi, M., Uhlig, C., John, U., Mock, T., and Valentini, K. (2010). Antifreeze proteins in polar sea ice diatoms: diversity and gene expression in the genus *Fragilariopsis*. *Environ. Microbiol.* 12, 1041–1052. doi: 10.1111/j.1462-2920.2009.02149.x
- Bender, S. J., Durkin, C. A., Berthiaume, C. T., Morales, R. L., and Armburst, E. V. (2014). Transcriptional responses of three model diatoms to nitrate limitation of growth. *Front. Mar. Sci.* 1:3. doi: 10.3389/fmars.2014.00003
- Benoiston, A. S., Ibarbalz, F. M., Bittner, L., Guidi, L., Jahn, O., Dutkiewicz, S., et al. (2017). The evolution of diatoms and their biogeochemical functions. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 372:20160397. doi: 10.1098/rstb.2016.0397
- Berge, J., Daase, M., Renaud, P. E., Ambrose, W. G., Damis, G., Last, K. S., et al. (2015). Unexpected levels of biological activity during the polar night offer new perspectives on a Warming Arctic. *Curr. Biol.* 25, 2555–2561. doi: 10.1016/j.cub.2015.08.024
- Bindoff, N. L., Willebrand, J., Artale, V., Cazenave, A., Gregory, J. M., Gulev, S., et al. (2007). "Observations: oceanic climate change and sea level," in *Climate Change 2007 The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, eds S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K. B. Averyt, et al. (Cambridge, MA: Cambridge University Press).
- Blanc, G., Agarkova, I., Grimwood, J., Kuo, A., Brueggeman, A., Dunigan, D. D., et al. (2012). The genome of the polar eukaryotic microalga *Coccomyxa subellipsoidea* reveals traits of cold adaptation. *Genome Biol.* 13:R39. doi: 10.1186/gb-2012-13-5-r39
- Bluhm, B. A., and Gradinger, R. (2008). Regional variability in food availability for arctic marine mammals. *Ecol. Appl.* 18, S77–S96. doi: 10.1890/06-0562.1
- Bork, P., Bowler, C., De Vargas, C., Gorsky, G., Karsenti, E., and Wincker, P. (2015). Tara Oceans studies plankton at Planetary scale. *Science* 348:873.
- Boyd, P. W. (2002). Review of environmental factors controlling phytoplankton processes in the Southern Ocean. *J. Phycol.* 38, 844–861. doi: 10.1046/j.1529-8817.2002.t01-1-01203.x
- Boyd, P. W., Cornwall, C. E., Davison, A., Doney, S. C., Fourquez, M., Hurd, C. L., et al. (2016). Biological responses to environmental heterogeneity under future ocean conditions. *Glob. Change Biol.* 22, 2633–2650. doi: 10.1111/gcb.13287
- Bracher, A., Vountas, M., Dinter, T., Burrows, J. P., Röttgers, R., and Peeken, I. (2009). Quantitative observation of cyanobacteria and diatoms from space using PhytoDOAS on SCIAMACHY data. *Biogeosciences* 6, 751–764. doi: 10.5194/bg-6-751-2009
- Brown, M. S., Munro, D. R., Feehan, C. J., Sweeney, C., Ducklow, H. W., and Schofield, O. M. (2019). Enhanced oceanic CO<sub>2</sub> uptake along the rapidly changing West Antarctic Peninsula. *Nat. Clim. Change* 9, 678–683. doi: 10.1038/s41558-019-0552-3
- Bunt, J. S., and Lee, C. C. (1972). Data on the composition and dark survival of four sea-ice microalgae. *Limnol. Oceanogr.* 17, 458–461.
- Cao, S., Zhang, W., Ding, W., Wang, M., Fan, S., Yang, B., et al. (2020). Structure and function of the Arctic and Antarctic marine microbiota as revealed by metagenomics. *Microbiome* 8, 1–12. doi: 10.1186/s40168-020-00826-9
- Carradec, Q., Pelletier, E., Da Silva, C., Alberti, A., Seeluthner, Y., Blanc-Mathieu, R., et al. (2018). A global ocean atlas of eukaryotic genes. *Nat. Commun.* 9, 1–13. doi: 10.1038/s41467-017-02342-1
- Croft, P. L., Baars, O., and Streu, P. (2011). The distribution of dissolved zinc in the Atlantic sector of the Southern Ocean. *Deep Sea Res. II Top. Stud. Oceanogr.* 58, 2707–2719. doi: 10.1016/j.dsr2.2010.10.041
- Cuvelier, M. L., Allen, A. E., Monier, A., McCrow, J. P., Messi'e, M., Tringe, S. G., et al. (2010). Targeted metagenomics and ecology of globally important uncultured eukaryotic phytoplankton. *Proc. Natl. Acad. Sci. U.S.A.* 107, 14679–14684. doi: 10.1073/pnas.1001665107
- DeVries, A. L., and Wohlschlag, D. E. (1969). Freezing resistance in some Antarctic fishes. *Science* 163, 1073–1075. doi: 10.1126/science.163.3871.1073
- DeVries, T. (2014). The oceanic anthropogenic CO<sub>2</sub> sink: storage, air-sea fluxes, and transports over the industrial era. *Glob. Biogeochem. Cycles* 28, 631–647. doi: 10.1002/2013GB004739
- Dolev, M. B., Braslavsky, I., and Davies, P. L. (2016). Ice-binding proteins and their function. *Annu. Rev. Biochem.* 85, 515–542. doi: 10.1146/annurev-biochem-060815-014546
- Duncan, A., Barry, K., Daum, C., Eloe-Fadrosh, E., Roux, S., Tringe, S., et al. (2020). Metagenome-assembled genomes of phytoplankton communities across the Arctic Circle. *BioRxiv* [Preprint]. doi: 10.1101/2020.06.16.154583
- Dutkiewicz, A., Müller, R. D., O'Callaghan, S., and Jónasson, H. (2015). Census of seafloor sediments in the world's ocean. *Geology* 43, 795–798. doi: 10.1130/g36883.1
- Eicken, H. (1992). The role of sea ice in structuring Antarctic ecosystems. *Polar Biol.* 12, 3–13. doi: 10.1007/BF00239960
- Erwin, D. H. (2009). Climate as a driver of evolutionary change. *Curr. Biol.* 19, R575–R583. doi: 10.1016/j.cub.2009.05.047
- Ewert, M., and Deming, J. (2011). Selective retention in saline ice of extracellular polysaccharides produced by the cold-adapted marine bacterium *Colwellia psychrerythraea* strain 34H. *Ann. Glaciol.* 52, 111–117. doi: 10.31818/172756411795931868
- Faktorová, D., Nisbet, R. E. R., Fernández Robledo, J. A., Casacuberta, E., Sudek, L., Allen, A. E., et al. (2020). Genetic tool development in marine protists: emerging model organisms for experimental cell biology. *Nat. Methods* 17, 481–494. doi: 10.1038/s41592-020-0796-x
- Falciatore, A., Jaubert, M., Bouly, J.-P., Bailleul, B., and Mock, T. (2020). Diatom molecular research comes of age: model species for studying phytoplankton biology and diversity. *Plant Cell* 32, 547–572. doi: 10.1105/tpc.19.00158
- Falkowski, P. G. (1998). Biogeochemical controls and feedbacks on ocean primary production. *Science* 281, 200–206. doi: 10.1126/science.281.5374.200
- Field, C. B., Behrenfeld, M. J., Randerson, J. T., and Falkowski, P. (1998). Primary production of the biosphere: integrating terrestrial and oceanic components. *Science* 281:237. doi: 10.1126/science.281.5374.237
- Firdaus-Raih, M., Hashim, N. H. F., Bharudin, I., Abu Bakar, M. F., Huang, K. K., Alias, H., et al. (2018). The *Glaciozyma antarctica* genome reveals an array of systems that provide sustained responses towards temperature variations in a persistently cold habitat. *PLoS One* 13:e0189947. doi: 10.1371/journal.pone.0189947
- Gobler, C. J., Berry, D. L., Dyrman, S. T., Wilhelm, S. W., Salamov, A., Lobanov, A. V., et al. (2011). Niche of harmful alga *Aureococcus anophagefferens* revealed through ecogenomics. *Proc. Natl. Acad. Sci. U.S.A.* 108, 4352–4357. doi: 10.1073/pnas.1016106108
- Gregory, A. C., Zayed, A. A., Conceição-Neto, N., Temperton, B., Bolduc, B., Alberti, A., et al. (2019). Marine DNA viral macro- and microdiversity from pole to pole. *Cell* 177, 1109–1123. doi: 10.1016/j.cell.2019.03.040
- Hansen, M. M., Olivieri, I., Waller, D. M., Nielsen, E. E., and Group, G. W. (2012). Monitoring adaptive genetic responses to environmental change. *Mol. Ecol.* 21, 1311–1329. doi: 10.1111/j.1365-294X.2011.05463.x
- Hansom, J. D., and Gordon, J. E. (1998). *Antarctic Environments and Resources: A Geographical Perspective*, 1st Edn. Abingdon: Routledge.
- Harrison, W. G., and Cota, G. F. (1991). Primary production in polar waters: relation to nutrient availability. *Polar Res.* 10, 87–104. doi: 10.3402/polar.v10i1.6730

- Hassler, C. S., Schoemann, V., Boye, M., Tagliabue, A., Rozmarynowycz, M., and McKay, R. M. L. (2012). "Iron bioavailability in the Southern Ocean," in *Oceanography and Marine Biology: An Annual Review*, eds R. N. Gibson, R. J. A. Atkinson, J. D. M. Gordon, and R. N. Hughes (Boca Raton, FL: CRC Press-Taylor & Francis Group).
- Hayward, A., and Grigor, J. (2020). The bottom of the Arctic's food web is of top importance. *Front. Young Minds* 8:122. doi: 10.3389/frym.2020.00122
- Hoagland, K. D., Rosowski, J. R., Gretz, M. R., and Roemer, S. C. (1993). Diatom Extracellular polymeric substances: function, fine structure, chemistry, and physiology. *J. Phycol.* 29, 537–566. doi: 10.1111/j.0022-3646.1993.00537.x
- Iversen, C. M., Sloan, V. L., Sullivan, P. F., Euskirchen, E. S., McGuire, A. D., Norby, R. J., et al. (2015). The unseen iceberg: plant roots in arctic tundra. *New Phytol.* 205, 34–58. doi: 10.1111/nph.13003
- Jung, W., Campbell, R. L., Gwak, Y., Kim, J. I., Davies, P. L., and Jin, E. (2016). New cysteine-rich ice-binding protein secreted from antarctic microalgae, *Chloromonas* sp. *PLoS One* 11:e0154056. doi: 10.1371/journal.pone.0154056
- Keeling, P. J., and Palmer, J. D. (2008). Horizontal gene transfer in eukaryotic evolution. *Nat. Rev. Genet.* 9, 605–618. doi: 10.1038/nrg2386
- Kennedy, F., Martin, A., Bowman, J. P., Wilson, R., and McMinn, A. (2019). Dark metabolism: a molecular insight into how the Antarctic sea-ice diatom *Fragilariopsis cylindrus* survives long-term darkness. *New Phytol.* 223, 675–691. doi: 10.1111/nph.15843
- Khatiwal, S., Primeau, F., and Hall, T. (2009). Reconstruction of the history of anthropogenic CO<sub>2</sub> concentrations in the ocean. *Nature* 462, 346–349. doi: 10.1038/nature08526
- Kondo, H., Mochizuki, K., and Bayer-Giraldi, M. (2018). Multiple binding modes of a moderate ice-binding protein from a polar microalgae. *Phys. Chem. Chem. Phys.* 20, 25295–25303. doi: 10.1039/c8cp04727h
- Koistira, W. H. C. F., Gersonde, R., Medlin, L. K., and Mann, D. G. (2007). "The origin and evolution of the diatoms: their adaptation to a planktonic existence," in *Evolution of Primary Producers in the Sea*, eds P. Falkowski and A. H. Knoll (Cambridge, MA: Academic Press), 207–249.
- Krembs, C., Eicken, H., and Deming, J. W. (2011). Exopolymer alteration of physical properties of sea ice and implications for ice habitability and biogeochemistry in a warmer Arctic. *Proc. Natl. Acad. Sci. U.S.A.* 108, 3653–3658. doi: 10.1073/pnas.1100701108
- Kvernvik, A. C., Hoppe, C. J. M., Lawrenz, E., Prasil, O., Greenacre, M., Wiktor, J. M., et al. (2018). Fast reactivation of photosynthesis in arctic phytoplankton during the polar night. *J. Phycol.* 54, 461–470. doi: 10.1111/jpy.12750
- Kwok, R. (2018). Arctic sea ice thickness, volume, and multiyear ice coverage: losses and coupled variability (1958–2018). *Environ. Res. Lett.* 13:105005. doi: 10.1088/1748-9326/aae3ec
- Laity, J. H., Lee, B. M., and Wright, P. E. (2001). Zinc finger proteins: new insights into structural and functional diversity. *Curr. Opin. Struct. Biol.* 11, 39–46. doi: 10.1016/s0959-440x(00)00167-6
- Laufkötter, C., Stern, A. A., John, J. G., Stock, C. A., and Dunne, J. P. (2018). Glacial iron sources stimulate the southern ocean carbon cycle. *Geophys. Res. Lett.* 45, 377–413. doi: 10.1029/2018GL079797
- Lee, S. (2014). A theory for polar amplification from a general circulation perspective. *Asia Pac. J. Atmos. Sci.* 50, 31–43. doi: 10.1007/s13143-014-0024-7
- Liang, Y., Koester, J. A., Liefer, J. D., Irwin, A. J., and Finkel, Z. V. (2019). Molecular mechanisms of temperature acclimation and adaptation in marine diatoms. *ISME J.* 13, 2415–2425. doi: 10.1038/s41396-019-0441-9
- Lin, Y., Moreno, C., Marchetti, A., Ducklow, H., Schofield, O., Delage, E., et al. (2021). Decline in plankton diversity and carbon flux with reduced sea ice extent along the Western Antarctic Peninsula. *Nat. Commun.* 12:4948. doi: 10.1038/s41467-021-25235-w
- Lommer, M., Specht, M., Roy, A. S., Kraemer, L., Andreson, R., Gutowska, M. A., et al. (2012). Genome and low-iron response of an oceanic diatom adapted to chronic iron limitation. *Genome Biol.* 13:20. doi: 10.1186/gb-2012-13-7-r66
- Long, M. C., Stephens, B. B., McKain, K., Sweeney, C., Keeling, R. F., Kort, E. A., et al. (2021). Strong Southern Ocean carbon uptake evident in airborne observations. *Science* 374, 1275–1280. doi: 10.1126/science.abi4355
- Lyon, B. R., and Mock, T. (2014). Polar microalgae: new approaches towards understanding adaptations to an extreme and changing environment. *Biology* 3, 56–80. doi: 10.3390/biology3010056
- MacGilchrist, G. A., Naveira Garabato, A. C., Tsubouchi, T., Bacon, S., Torres-Valdés, S., and Azetsu-Scott, K. (2014). The Arctic Ocean carbon sink. *Deep Sea Res. I Oceanogr. Res. Pap.* 86, 39–55. doi: 10.1016/j.dsr.2014.01.002
- Maksym, T. (2019). Arctic and Antarctic Sea ice change: contrasts, commonalities, and causes. *Annu. Rev. Mar. Sci.* 11, 187–213. doi: 10.1146/annurev-marine-010816-060610
- Malviya, S., Scalco, E., Audic, S., Vincent, F., Veluchamy, A., Poulain, J., et al. (2016). Insights into global diatom distribution and diversity in the world's ocean. *Proc. Natl. Acad. Sci. U.S.A.* 113, E1516–E1525. doi: 10.1073/pnas.1509523113
- Marchetti, A. (2019). A global perspective on iron and plankton through the Tara Oceans Lens. *Glob. Biogeochem. Cycles* 33, 239–242. doi: 10.1029/2019gb006181
- Marchetti, A., Parker, M. S., Moccia, L. P., Lin, E. O., Arrieta, A. L., Ribalet, F., et al. (2009). Ferritin is used for iron storage in bloom-forming marine pennate diatoms. *Nature* 457, 467–470. doi: 10.1038/nature07539
- Martin, J. H., Gordon, M., and Fitzwater, S. E. (1991). The case for iron. *Limnol. Oceanogr.* 36, 1793–1802. doi: 10.4319/lo.1991.36.8.1793
- Martin, K., Schmidt, K., Toseland, A., Boulton, C. A., Barry, K., Beszteri, B., et al. (2021). The biogeographic differentiation of algal microbiomes in the upper ocean from pole to pole. *Nat. Commun.* 12:5483. doi: 10.1038/s41467-021-25646-9
- Martin, W., Stoebe, B., Goremykin, V., Hansmann, S., Hasegawa, M., and Kowallik, K. V. (1998). Gene transfer to the nucleus and the evolution of chloroplasts. *Nature* 393:162. doi: 10.1038/30234
- McCain, J. S. P., Allen, A. E., and Bertrand, E. M. (2021). Proteomic traits vary across taxa in a coastal Antarctic phytoplankton bloom. *ISME J.* 16, 569–579. doi: 10.1038/s41396-021-01084-9
- McMinn, A., and Martin, A. (2013). Dark survival in a warming world. *Proc. Biol. Sci.* 280:20122909. doi: 10.1098/rspb.2012.2909
- Miteva-Staleva, J., Stefanova, T., Krumova, E., and Angelova, M. (2011). Growth-phase-related changes in reactive oxygen species generation as a cold stress response in Antarctic *Penicillium* strains. *Biotechnol. Biotechnol. Equ.* 25, 58–63. doi: 10.5504/bbeq.2011.0131
- Mock, T., Daines, S. J., Geider, R., Collins, S., Metodiev, M., Millar, A. J., et al. (2016). Bridging the gap between omics and earth system science to better understand how environmental change impacts marine microbes. *Glob. Change Biol.* 22, 61–75. doi: 10.1111/gcb.12983
- Mock, T., Otillar, R. P., Strauss, J., McMullan, M., Paajanen, P., Schmutz, J., et al. (2017). Evolutionary genomics of the cold-adapted diatom *Fragilariopsis cylindrus*. *Nature* 541, 536–540. doi: 10.1038/nature20803
- Moreno, C. M., Gong, W., Cohen, N. R., DeLong, K., and Marchetti, A. (2020). Interactive effects of iron and light limitation on the molecular physiology of the Southern Ocean diatom *Fragilariopsis kerguelensis*. *Limnol. Oceanogr.* 65, 1511–1531. doi: 10.1002/lno.11404
- Moreno, C. M., Lin, Y., Davies, S., Monbureau, E., Cassar, N., and Marchetti, A. (2018). Examination of gene repertoires and physiological responses to iron and light limitation in Southern Ocean diatoms. *Polar Biol.* 41, 679–696. doi: 10.1007/s00300-017-2228-7
- Morgan-Kiss, R. M., Lizotte, M. P., Kong, W., and Priscu, J. C. (2016). Photoadaptation to the polar night by phytoplankton in a permanently ice-covered Antarctic lake. *Limnol. Oceanogr.* 61, 3–13. doi: 10.1002/lno.10107
- Murata, A., and Takizawa, T. (2003). Summertime CO<sub>2</sub> sinks in shelf and slope waters of the western Arctic Ocean. *Continental Shelf Res.* 23, 753–776. doi: 10.1016/S0278-4343(03)00046-3
- Murphy, E. J., Cavanagh, R. D., Drinkwater, K. F., Grant, S. M., Heymans, J. J., Hofmann, E. E., et al. (2016). Understanding the structure and functioning of polar pelagic ecosystems to predict the impacts of change. *Proc. Biol. Sci.* 283:1646. doi: 10.1098/rspb.2016.1646
- Myktyczuk, N. C. S., Foote, S. J., Omedon, C. R., Southam, G., Greer, C. W., and Whyte, L. G. (2013). Bacterial growth at -15 degrees C: molecular insights from the permafrost bacterium *Planococcus halocryophilus* Or1. *ISME J.* 7, 1211–1226. doi: 10.1038/ismej.2013.8
- Nelson, D. M., Smith, W. O. Jr., Gordon, L. I., and Huber, B. A. (1987). Spring distributions of density, nutrients, and phytoplankton biomass in the ice edge zone of the Weddell-Scotia Sea. *J. Geophys. Res. Oceans* 92, 7181–7190.
- Obiol, A., Giner, C. R., Sánchez, P., Duarte, C. M., Acinas, S. G., and Massana, R. (2020). A metagenomic assessment of microbial eukaryotic diversity in the global ocean. *Mol. Ecol. Resour.* 20, 718–731. doi: 10.1111/1755-0998.13147
- Ovaskainen, O., Tikhonov, G., Norberg, A., Guillaume Blanchet, F., Duan, L., Dunson, D., et al. (2017). How to make more out of community data? A



- conceptual framework and its implementation as models and software. *Ecol. Lett.* 20, 561–576. doi: 10.1111/ele.12757
- Palenik, B. (2015). Molecular mechanisms by which marine phytoplankton respond to their dynamic chemical environment. *Annu. Rev. Mar. Sci.* 7, 325–340. doi: 10.1146/annurev-marine-010814-015639
- Pearson, G. A., Lago-Leston, A., Cánovas, F., Cox, C. J., Verret, F., Lasternas, S., et al. (2015). Metatranscriptomes reveal functional variation in diatom communities from the Antarctic Peninsula. *ISME J.* 9, 2275–2289. doi: 10.1038/ismej.2015.40
- Peers, G., and Price, N. M. (2006). Copper-containing plastocyanin used for electron transport by an oceanic diatom. *Nature* 441, 341–344. doi: 10.1038/nature04630
- Peters, E., and Thomas, D. N. (1996). Prolonged darkness and diatom mortality I: marine Antarctic species. *J. Exp. Mar. Biol. Ecol.* 207, 25–41. doi: 10.1016/S0022-0981(96)02520-8
- Petrou, K., Trimbom, S., Rost, B., Ralph, P. J., and Hassler, C. S. (2014). The impact of iron limitation on the physiology of the Antarctic diatom *Chaetoceros simplex*. *Mar. Biol.* 161, 925–937. doi: 10.1007/s00227-014-2392-z
- Pinsel, E., Janssens, S. B., Verleyen, E., Vanormelingen, P., Kohler, T. J., Biersma, E. M., et al. (2020). Global radiation in a rare biosphere soil diatom. *Nat. Commun.* 11:2382. doi: 10.1038/s41467-020-16181-0
- Pitchford, J. W., and Brindley, J. (1999). Iron limitation, grazing pressure and oceanic high nutrient-low chlorophyll (HNLC) regions. *J. Plankton Res.* 21, 525–547. doi: 10.1093/plankt/21.3.525
- Pointing, S., Buedel, B., Convey, P., Gillman, L., Koerner, C., Leuzinger, S., et al. (2015). Biogeography of photoautotrophs in the high polar biome. *Front. Plant Sci.* 6:692. doi: 10.3389/fpls.2015.00692
- Postel, U., Glemser, B., Salazar Alekseyeva, K., Eggert, S. L., Groth, M., Glöckner, G., et al. (2020). Adaptive divergence across Southern Ocean gradients in the pelagic diatom *Fragilariopsis kerguelensis*. *Mol. Ecol.* 29, 4913–4924. doi: 10.1111/mec.15554
- Raymond, J. A. (2014). The ice-binding proteins of a snow alga, *Chloromonas brevispina*: probable acquisition by horizontal gene transfer. *Extremophiles* 18, 987–994. doi: 10.1007/s00792-014-0668-3
- Raymond, J. A., Janech, M. G., and Fritsen, C. H. (2009). Novel ice-binding proteins from a psychrophilic Antarctic alga (Chlamydomonadaceae, Chlorophyceae). *J. Phycol.* 45, 130–136. doi: 10.1111/j.1529-8817.2008.00623.x
- Raymond, J. A., Janech, M. G., and Mangiagli, M. (2021). Ice-binding proteins associated with an Antarctic *Cyanobacterium*, *Nostoc* sp. HG1. *Appl. Environ. Microbiol.* 87:e02499-20. doi: 10.1128/AEM.02499-20
- Raymond, J. A., and Kim, H. J. (2012). Possible role of horizontal gene transfer in the colonization of Sea Ice by Algae. *PLoS One* 7:e35968. doi: 10.1371/journal.pone.0035968
- Raymond, J. A., and Morgan-Kiss, R. (2017). Multiple ice-binding proteins of probable prokaryotic origin in an antarctic Lake Alga, *Chlamydomonas* sp ICE-MDV (Chlorophyceae). *J. Phycol.* 53, 848–854. doi: 10.1111/jpy.12550
- Raymond-Bouchar, I., Tremblay, J., Altshuler, I., Greer, C. W., and Whyte, L. G. (2018). Comparative transcriptomics of cold growth and adaptive features of a eury- and steno-psychrophile. *Front. Microbiol.* 9:1565. doi: 10.3389/fmicb.2018.01565
- Reeves, S., McMinn, A., and Martin, A. (2011). The effect of prolonged darkness on the growth, recovery and survival of Antarctic sea ice diatoms. *Polar Biol.* 34, 1019–1032. doi: 10.1007/s00300-011-0961-x
- Royo-Llonch, M., Sánchez, P., Ruiz-González, C., Salazar, G., Pedrós-Alió, C., Sebastián, M., et al. (2021). Compendium of 530 metagenome-assembled bacterial and archaeal genomes from the polar Arctic Ocean. *Nat. Microbiol.* 6, 1561–1574. doi: 10.1038/s41564-021-00979-9
- Sackett, O., Petrou, K., Reedy, B., De Grazia, A., Hill, R., Doblin, M., et al. (2013). Phenotypic plasticity of Southern Ocean diatoms: key to success in the sea ice habitat? *PLoS One* 8:e81185. doi: 10.1371/journal.pone.0081185
- Salamy, K. A., and Zachos, J. C. (1999). Latest Eocene–Early Oligocene climate change and Southern Ocean fertility: inferences from sediment accumulation and stable isotope data. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* 145, 61–77. doi: 10.1016/S0031-0182(98)00093-5
- Schaub, I., Wagner, H., Graeve, M., and Karsten, U. (2017). Effects of prolonged darkness and temperature on the lipid metabolism in the benthic diatom *Navicula perminuta* from the Arctic Adventfjorden, Svalbard. *Polar Biol.* 40, 1425–1439. doi: 10.1007/s00300-016-2067-y
- Scholz, B., Rua, A., and Liebezeit, G. (2014). Effects of UV radiation on five marine microphytobenthic Wadden sea diatoms isolated from the Solthorn tidal flat (Lower Saxony, southern North Sea) – Part I: growth and antioxidative defence strategies. *Eur. J. Phycol.* 49, 68–82. doi: 10.1080/09670262.2014.889214
- Sigman, D. M., Hain, M. P., and Haug, G. H. (2010). The polar ocean and glacial cycles in atmospheric CO<sub>2</sub> concentration. *Nature* 466, 47–55. doi: 10.1038/nature09149
- Singh, J., Singh, R. P., and Khare, R. (2018). Influence of climate change on Antarctic flora. *Polar Sci.* 18, 94–101. doi: 10.1016/j.polar.2018.05.006
- Sjöqvist, C. O., and Kremp, A. (2016). Genetic diversity affects ecological performance and stress response of marine diatom populations. *ISME J.* 10, 2755–2766. doi: 10.1038/ismej.2016.44
- Smythe-Wright, D., Cunningham, S. A., Lampitt, R. A., Kent, E. C., King, B. A., Quartly, G. D., et al. (2010). “Sustained observations in the Atlantic and Southern Oceans,” in *Proceedings of OceanObs’09: Sustained Ocean Observations and Information for Society*, eds J. Hall, D. E. Harrison, and D. Stammer (Paris: European Space Agency).
- Soreide, J. E., Leu, E., Berge, J., Graeve, M., and Falk-Petersen, S. (2010). Timing of blooms, algal food quality and *Calanus glacialis* reproduction and growth in a changing Arctic. *Glob. Change Biol.* 16, 3154–3163. doi: 10.1111/j.1365-2486.2010.02175.x
- Steig, E. J., Schneider, D. P., Rutherford, S. D., Mann, M. E., Comiso, J. C., and Shindell, D. T. (2009). Warming of the Antarctic ice-sheet surface since the 1957 International Geophysical Year. *Nature* 457, 459–462. doi: 10.1038/nature07669
- Stephens, T. G., González-Pech, R. A., Cheng, Y., Mohamed, A. R., Burt, D. W., Bhattacharya, D., et al. (2020). Genomes of the dinoflagellate *Polarella glacialis* encode tandemly repeated single-exon genes with adaptive functions. *BMC Biol.* 18:56. doi: 10.1186/s12915-020-00782-8
- Stretch, J. J., Hamner, P. P., Hamner, W. M., Michel, W. C., Cook, J., and Sullivan, C. W. (1988). Foraging behavior of Antarctic krill *Euphausia superba* on sea ice microalgae. *Mar. Ecol. Prog. Ser.* 44, 131–139. doi: 10.3354/meps044131
- Suto, I., Kawamura, K., Hagimoto, S., Teraishi, A., and Tanaka, Y. (2012). Changes in upwelling mechanisms drove the evolution of marine organisms. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* 339, 39–51. doi: 10.1016/j.palaeo.2012.04.014
- Takahashi, T., Sutherland, S. C., Wanninkhof, R., Sweeney, C., Feely, R. A., Chipman, D. W., et al. (2009). Climatological mean and decadal change in surface ocean pCO<sub>2</sub>, and net sea–air CO<sub>2</sub> flux over the global oceans. *Deep Sea Res. II Top. Stud. Oceanogr.* 56, 554–577. doi: 10.1016/j.dsr2.2008.12.009
- Thomas, D. N., and Dieckmann, G. S. (2002). Antarctic sea ice — a habitat for extremophiles. *Science* 295:641. doi: 10.1126/science.1063391
- Tikhonov, G., Opedal, ØH., Abrego, N., Lehtikoinen, A., de Jonge, M. M. J., Oksanen, J., et al. (2020). Joint species distribution modelling with the r-package Hmsc. *Methods Ecol. Evol.* 11, 442–447. doi: 10.1111/2041-210X.13345
- Tonelli, M., Signori, C. N., Bendia, A., Neiva, J., Ferrero, B., Pellizari, V., et al. (2021). Climate projections for the southern ocean reveal impacts in the marine microbial communities following increases in sea surface temperature. *Front. Mar. Sci.* 8:636226. doi: 10.3389/fmars.2021.636226
- Torstenson, A., Young, J. N., Carlson, L. T., Ingalls, A. E., and Deming, J. W. (2019). Use of exogenous glycine betaine and its precursor choline as osmoprotectants in Antarctic sea-ice diatoms. *J. Phycol.* 55, 663–675. doi: 10.1111/jpy.12839
- Trefault, N., De la Iglesia, R., Moreno-Pino, M., Lopes dos Santos, A., Gériques Ribeiro, C., Parada-Pozo, G., et al. (2021). Annual phytoplankton dynamics in coastal waters from Fildes Bay, Western Antarctic Peninsula. *Sci. Rep.* 11:1368. doi: 10.1038/s41598-020-80568-8
- Tréguer, P., Bowler, C., Moriceau, B., Dutkiewicz, S., Gehlen, M., Aumont, O., et al. (2018). Influence of diatom diversity on the ocean biological carbon pump. *Nat. Geosci.* 11, 27–37. doi: 10.1038/s41561-017-0028-x
- van de Poll, W. H., Abdullah, E., Visser, R. J. W., Fischer, P., and Buma, A. G. J. (2020). Taxon-specific dark survival of diatoms and flagellates affects Arctic phytoplankton composition during the polar night and early spring. *Limnol. Oceanogr.* 65, 903–914. doi: 10.1002/lno.11355
- Vance, T. D. R., Bayer-Giraldi, M., Davies, P. L., and Mangiagli, M. (2019). Ice-binding proteins and the ‘domain of unknown function’ 3494 family. *FEBS J.* 286, 855–873. doi: 10.1111/febs.14764

- Venables, H., and Moore, M. (2010). Phytoplankton and light limitation in the Southern Ocean: learning from high-nutrient, high-chlorophyll areas. *J. Geophys. Res.* 115:C02015. doi: 10.1029/2009JC005361
- Verde, C., Giordano, D., Bellas, C. M., di Prisco, G., and Anesio, A. M. (2016). Polar marine microorganisms and climate change. *Adv. Microb. Physiol.* 69, 187–215. doi: 10.1016/bs.ampbs.2016.07.002
- Vincent, W. F. (2010). Microbial ecosystem responses to rapid climate change in the Arctic. *ISME J.* 4, 1087–1090. doi: 10.1038/ismej.2010.108
- Wadham, J. L., Hawkings, J. R., Tarasov, L., Gregoire, L. J., Spencer, R. G. M., Gutjahr, M., et al. (2019). Ice sheets matter for the global carbon cycle. *Nat. Commun.* 10:3567. doi: 10.1038/s41467-019-11394-4
- Wake, B. (2019). A drift in the Arctic. *Nat. Clim. Change* 9:733. doi: 10.1038/s41558-019-0597-3
- Waldvogel, A.-M., Feldmeyer, B., Rolshausen, G., Exposito-Alonso, M., Rellstab, C., Kofler, R., et al. (2020). Evolutionary genomics can improve prediction of species' responses to climate change. *Evol. Lett.* 4, 4–18. doi: 10.1002/evl3.154
- Walker, D. A., Raynolds, M. K., Daniëls, F. J. A., Einarsson, E., Elvebakk, A., Gould, W. A., et al. (2005). The Circumpolar Arctic vegetation map. *J. Veg. Sci.* 16, 267–282. doi: 10.1111/j.1654-1103.2005.tb02365.x
- Wietz, M., Bienhold, C., Metfies, K., Torres-Valdés, S., von Appen, W. J., Salter, I., et al. (2021). The polar night shift: annual dynamics and drivers of microbial community structure in the arctic ocean. *bioRxiv* [Preprint]. doi: 10.1101/2021.04.08.436999w
- Wingsle, G., Karpinski, S., and Hallgren, J. E. (1999). Low temperature, high light stress and antioxidant defence mechanisms in higher plants. *Phyton Ann. Rei Bot.* 39, 253–268.
- Wolf, K. K. E., Romanelli, E., Rost, B., John, U., Collins, S., Weigand, H., et al. (2019). Company matters: the presence of other genotypes alters traits and intraspecific selection in an Arctic diatom under climate change. *Glob. Change Biol.* 25, 2869–2884. doi: 10.1111/gcb.14675
- Wong, C., Boo, S. Y., Voo, C. L. Y., Zainuddin, N., and Najimudin, N. (2019). A comparative transcriptomic analysis provides insights into the cold-adaptation mechanisms of a psychrophilic yeast, *Glaciozyma antarctica* PI12. *Polar Biol.* 42, 541–553. doi: 10.1007/s00300-018-02443-7
- Yergeau, E., Michel, C., Tremblay, J., Niemi, A., King, T. L., Wyglinski, J., et al. (2017). Metagenomic survey of the taxonomic and functional microbial communities of seawater and sea ice from the Canadian Arctic. *Sci. Rep.* 7:42242.
- Zhang, X., Cvetkovska, M., Morgan-Kiss, R., Hüner, N. P. A., and Smith, D. R. (2021). Draft genome sequence of the Antarctic green alga *Chlamydomonas* sp. UWO241. *iScience* 24:102084. doi: 10.1016/j.isci.2021.102084
- Zhang, Z., Qu, C., Zhang, K., He, Y., Zhao, X., Yang, L., et al. (2020). Adaptation to extreme antarctic environments revealed by the genome of a sea ice green alga. *Curr. Biol.* 30, 3330.e7–3341.e7. doi: 10.1016/j.cub.2020.06.029

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Gilbertson, Langan and Mock. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

## 2 Materials and Methods

### 2.1 Cell Culture

#### 2.1.1 Growth Conditions

##### 2.1.1.1 *Temperate diatoms*

*Thalassiosira pseudonana* (CCMP 1335) mutant and wild type cell lines and wild type *Phaeodactylum tricornutum* (CCAP 1055) were grown in synthetic ocean water (SOW; table 2.1) based media under constant light (75 – 100  $\mu\text{mol}/\text{photons m}^{-2}\text{s}^{-1}$ ) and temperature (20°C) in a MLR-350H Versatile Environmental Test Chamber (Sanyo Electric Biomedical Co. Ltd., Japan). A constant light regime was used as the wild-type strains had been acclimated to these conditions for at least 5 years prior to the start of this research. The lab's incubator provided constant light to accommodate both standard growth and experimental growth curves. Genetically edited *T. pseudonana* cell lines were kept in SOW with half the final concentration of anhydrous and hydrous salts (Table 2.1) due to selection plates containing half-salinity SOW. Both glassware and plastic culture flasks were used. All glassware was cleaned with detergent, washed with Milli-Q purified water, left to dry and autoclaved at 121°C before use. Plastic culture flasks were purchased sterile and ready to use. All cell culture was conducted under a class II microbiological safety cabinet (Class II, Walker, Glossop, UK).

##### 2.1.1.2 *Polar diatoms*

*F. cylindrus* cultures were grown in a temperature-controlled room at 4°C. All transfers were conducted within a safety cabinet as previously described, and cultures were kept on ice for the duration of transfers to reduce heat shock. All cultures were transferred to new media at no less than  $3.5 \times 10^4$  cells/ml.

#### 2.1.2 Media Preparation

Both *T. pseudonana* and *F. cylindrus* were cultured in Aquil media (Price et al., 1989) prepared according to the Bigelow Laboratory for Ocean Sciences (NCMA) using a synthetic ocean water base enriched for macro and micro-nutrients (Table 2.1). Nutrient stocks were

filtered through a 0.22 $\mu$ m filter (Millipore) and stored in autoclaved glassware at 4°C. The SOW base was made first by diluting the appropriate mass of anhydrous and hydrous salts in Milli-Q purified water and left to stir overnight. The pH was measured and adjusted to between 7.8 and 8.0 using sodium hydroxide. Once a stable pH was achieved the SOW was passed through a filter unit comprising of a 0.45 $\mu$ m and 0.2 $\mu$ m filter (Policap TC and AS respectively; GE Healthcare UK), with an additional 0.2  $\mu$ m disposable filter (Millipore) into an autoclaved Nalgene bottle. Sterile macro and micro-nutrients, trace metal, and final vitamin stocks were then added to the media.

*Table 2.1. Composition of Aquil medium with the final concentrations of salts and nutrients. All nutrient stocks were sterilised by filtering through a 0.22 $\mu$ m filter unit (Millipore) before addition to synthetic ocean water (SOW).*

	Compound	Molar Concentration in Final Medium	Molar Concentration in Half Salinity Medium
<b>Anhydrous Salts</b>	NaCl	$4.20 \times 10^{-1}$	$2.10 \times 10^{-1}$
	Na <sub>2</sub> SO <sub>4</sub>	$2.88 \times 10^{-2}$	$1.44 \times 10^{-2}$
	KCl	$9.39 \times 10^{-3}$	$4.65 \times 10^{-3}$
	NaHCO <sub>3</sub>	$2.38 \times 10^{-3}$	$1.19 \times 10^{-3}$
	KBr	$8.40 \times 10^{-4}$	$4.20 \times 10^{-4}$
	H <sub>3</sub> BO <sub>3</sub>	$4.85 \times 10^{-5}$	$2.43 \times 10^{-5}$
	NaF	$7.15 \times 10^{-5}$	$3.58 \times 10^{-5}$
<b>Hydrous Salts</b>	MgCl <sub>2</sub> .6H <sub>2</sub> O	$5.46 \times 10^{-2}$	$2.73 \times 10^{-2}$
	CaCl <sub>2</sub> .2H <sub>2</sub> O	$1.05 \times 10^{-2}$	$5.25 \times 10^{-3}$
	SrCl <sub>2</sub> .6H <sub>2</sub> O	$6.38 \times 10^{-5}$	$3.19+ \times 10^{-5}$
<b>Major Nutrients</b>	NaH <sub>2</sub> PO <sub>4</sub> H <sub>2</sub> O	$1.00 \times 10^{-5}$	$1.00 \times 10^{-5}$
	NaNO <sub>3</sub>	$1.00 \times 10^{-4}$	$1.00 \times 10^{-4}$
	Na <sub>2</sub> SiO <sub>3</sub> 9H <sub>2</sub> O	$1.00 \times 10^{-4}$	$1.00 \times 10^{-4}$
<b>Trace Metals</b>	EDTA	$1.00 \times 10^{-4}$	$1.00 \times 10^{-4}$
	FeCl <sub>3</sub> .6H <sub>2</sub> O	$1.00 \times 10^{-6}$	$1.00 \times 10^{-6}$
	ZnSO <sub>4</sub> .7H <sub>2</sub> O	$7.97 \times 10^{-8}$	$7.97 \times 10^{-8}$
	MnCl <sub>2</sub> .4H <sub>2</sub> O	$1.21 \times 10^{-7}$	$1.21 \times 10^{-7}$
	CoCl <sub>2</sub> .6H <sub>2</sub> O	$5.03 \times 10^{-8}$	$5.03 \times 10^{-8}$
	NaMoO <sub>4</sub> .2H <sub>2</sub> O	$1.00 \times 10^{-7}$	$1.00 \times 10^{-7}$
	CuSO <sub>4</sub> .5H <sub>2</sub> O	$1.96 \times 10^{-8}$	$1.96 \times 10^{-8}$
	Na <sub>2</sub> SeO <sub>3</sub>	$1.00 \times 10^{-8}$	$1.00 \times 10^{-8}$
<b>Vitamins</b>	Thiamine (Vit. B <sub>1</sub> )	$2.97 \times 10^{-7}$	$2.97 \times 10^{-7}$
	Biotin (Vit. H)	$2.25 \times 10^{-9}$	$2.25 \times 10^{-9}$
	cyanocobalamin (Vit. B <sub>12</sub> )	$3.70 \times 10^{-10}$	$3.70 \times 10^{-10}$

### 2.1.3 Growth Curves

#### 2.1.3.1 Global Methodologies

The methodologies in this section apply to all conducted growth curves, with the following subsections reporting specific methods. All growth curves were carried out with three biological replicates under standard growth conditions unless stated otherwise. A 1 millilitre (ml) sample of each biological replicate was transferred to a 2 ml Eppendorf microcentrifuge tube. First, the maximum quantum yield of photosystem II ( $F_v/F_m$ ) was measured with a Walz Phyto-Pam Phytoplankton Analyser fluorometer (Walz GmbH, Effeltrich, Germany) using the Phyto-ED component.  $F_v/F_m$  data provides a quantification of the photosynthetic efficiency of *in vivo* algae cultures by excitation of chlorophyll at different wavelengths and using the resulting data in the following equation:

$$F_v/F_m = \frac{(F_m - F_0)}{F_m}$$

After  $F_v/F_m$  quantification, cell counts were either counted on the Beckman Coulter Counter or under a microscope, depending on the number of samples per growth curve or if mean size data was required. For the coulter counter samples were first diluted into 1% NaCl and thoroughly mixed and run through the 100-micron ( $\mu\text{m}$ ) aperture tube. Microscope counts were conducted using a haemocytometer. Technical replicates were counted for each sample and an average was taken. Raw cell counts during the exponential phase were used to calculate the specific growth rate ( $\mu$ ) using the following equation:

$$\mu(d^{-1}) = \frac{\ln(C_1) - \ln(C_0)}{t_1 - t_0}$$

Where  $C_1$  and  $C_0$  represent the cell concentration at timepoint 1 ( $t_1$ ) and timepoint 0 ( $t_0$ ) respectively. This is divided by the time between data points in days.

The number of generations between two time points the was calculated using the following equation:

$$g = \frac{\text{Log}\left(\frac{f}{i}\right)}{\text{Log}2}$$

In the above equation:  $g$  = number of generations,  $f$  = final density of cells per millilitre (cells/ml) and  $i$  = initial cell density (cells/ml).

### 2.1.3.2 Dose-response Growth Curves

Dose-response curves were conducted with three model diatom species and genetically edited *T. pseudonana* cultures to determine the half maximal effective concentration (EC<sub>50</sub>) for zeocin (Zeocin™ Selection Reagent, Invitrogen, Catalogue # R25001) and methyl methanesulfonate (MMS; Sigma-Aldrich, CAS # 66-27-3). These two reagents were chosen as they both cause DNA DSBs though through different mechanisms. Zeocin, part of the bleomycin family of antibiotics, intercalates into the DNA while MMS creates adducts and stalls replication forks. Zeocin and MMS have been used as model DNA damaging agents given their consistency in the type of mutations generated across a variety of organisms. The EC<sub>50</sub> value is the concentration of a compound that provides half of the maximal response between the control and maximum effect (Jiang and Kopp-Schneider, 2014). Given differences in each species' tolerance of compounds, the EC<sub>50</sub> specific to each species was used in all experiments to normalise the effect of each DNA-damaging agent. These growth parameters were as described in section 2.1.3.1, with the addition of zeocin or MMS in increasing concentrations in media. For negative controls the vehicle each drug was dissolved in prior to addition in the media was used; nuclease free water was for zeocin, and dimethyl sulfoxide (DMSO) for MMS. Zeocin concentration ranges differed between species due to the ability of each species to tolerate zeocin, for *T. pseudonana* zeocin concentrations ranged from 0.1 nanogram per millilitre (ng/ml) to 150 ng/ml, whereas for *F. cylindrus* and *P. tricornutum* concentrations ranged from 0.1 microgram per millilitre (µg/ml) to 150 µg/ml. However, MMS tolerance is similar across all species, so the range was 0.01 milli molar (mM) to 5mM was used for all species. Extreme concentrations of zeocin (150 ng/ml or 150 µg/ml) and MMS (5mM) which killed each species within 24 hours of exposure were used as an internal positive control to confirm the drug was active in the solution. Data was then modelled in R using the package *drda* to obtain the EC<sub>50</sub> values and



provide statistics to show differences between distinct species and genetically modified strains (Malyutina et al., 2023).

### 2.1.3.3 Temperature Response Curves

Using a gradient temperature bar developed and provided by Dr Erik Buitenhuis (UEA), temperature stress experiments were conducted using *Brca2* knock-out strains to determine if a core DNA repair enzyme has a significant role in fitness under environmentally relevant stress. Cultures were grown as stated in general growth curve methods in 55ml glass culture tubes (Pyrex Brand 9826) across a temperature gradient from 7°C to 36°C under constant light (100  $\mu$ mol/photons m<sup>-2</sup>s<sup>-1</sup>; Scalar PAR Irradiance Sensor QSL 2101, Biospherical Instruments Inc., San Diego, USA). Cultures were counted twice a day using light microscopy as well as recording Fv/Fm. The temperature optimum curve was modelled using R. Due to time constraints previously generated temperature response curves for wild-type *T. pseudonana* cultures using the same equipment were used as control data.

### 2.1.3.4 Short experimental evolution experiment – MMS treatment

In chapter 5, triplicates of *Brca2* -/- and wild type *T. pseudonana* strains were cultured under either normal growth conditions (mock treated) or under intermittent exposure to MMS (MMS treated). Light and temperature conditions were the same as previously described for *T. pseudonana* in section 2.1.1. All strains were transferred to fresh media at a concentration of 1 x 10<sup>5</sup> cells/ml every 3 days to ensure they did not reach stationary phase. Every seven days, MMS treated cultures were exposed to the EC<sub>50</sub> concentration of MMS dissolved in 30 $\mu$ l of DMSO for 1.5 hours and the mock treated cultures were exposed to 30 $\mu$ l of DMSO for 1.5 hours to control for any effects of using DMSO as a vehicle for MMS. After incubation with MMS (MMS treated) or DMSO (mock treated), all cultures were washed 3 times and transferred into fresh media at a concentration of 1 x 10<sup>5</sup> cells/ml. Four weekly treatments with either MMS or DMSO were carried out.

Subsamples of all cultures were taken after the 1<sup>st</sup>, 2<sup>nd</sup>, 3<sup>rd</sup> and 4<sup>th</sup> treatments and grown in 250ml to a density between 5.5 x 10<sup>5</sup> and 6 x 10<sup>6</sup> cells/ml and filtered under sterile conditions. Filters holding the samples were snap frozen on dry ice and stored at -80°C until

DNA extraction. DNA extractions were carried out using the methods described in section 2.4.2.

## 2.2 Identification of DNA Damage Using the Terminal

### Deoxynucleotidyl Transferase dUTP Nick End Labelling (TUNEL)

#### Assay

Terminal deoxynucleotidyl transferase dUTP nick end labelling (TUNEL) detects DNA breaks formed from DNA fragmentation by using the protein terminal-deoxy transferase (TdT) to insert fluorescently labelled nucleotides at free 3' -OH DNA ends characteristic of fragmented regions of DNA (Figure 2.1).

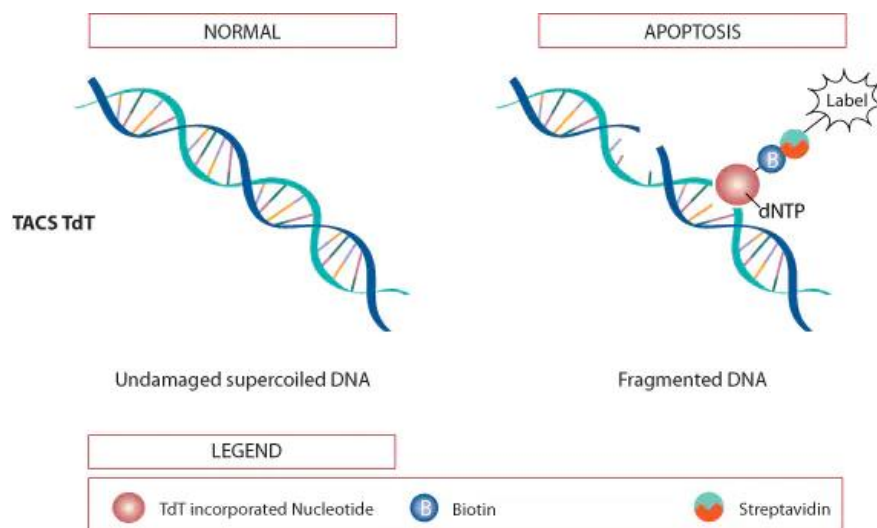


Figure 2.1. Diagram of TUNEL Assay principle. The enzyme TdT incorporates fluorescently tagged nucleotides to 3'OH ends of fragmented DNA. This signal is then able to be analysed through flow cytometry or microscopy.

The TUNEL assay was developed primarily to detect apoptotic DNA but can also be used to show excessive DNA damage in individuals and populations of cells (Gorczyca et al., 1992). The method utilises the enzyme TdT which facilitates attachment of a nucleotide modified with a fluorochrome to 3'-hydroxyl termini resulting from DNA fragmentation. Samples can be analysed by either fluorescence microscopy or flow cytometry. The

autofluorescence of diatoms proved an issue by creating significant background noise when analysing samples, so amendments were made to the manufacturer's protocol derived from Gallo et al., 2017 and Tripathi et al., 2017. The In Situ Cell Death Detection Kit, Fluorescein (ROCHE) was used in all experiments.

### 2.2.1 Flow Cytometry

*T. pseudonana* cultures were incubated with or without a DNA damaging agent and harvested via filtration and washed three times with sterile 1x phosphate-buffered saline (PBS) before fixation by incubating them in 4% formaldehyde in 1xPBS at room temperature for 20 minutes. Samples were washed with PBS and resuspended in 750µl of 70% ethanol overnight to permeabilise the membrane and dissolve pigments, after which they were washed three times with 1xPBS to remove any ethanol and resuspended in 200µl 1xPBS. Positive controls were created at this stage by incubating permeabilised samples in 10 µg/ml of DNaseI for 10 minutes. The TUNEL reaction mix was always freshly prepared immediately before addition to samples following the manufacturer's protocol. Samples were incubated in a 50µl TUNEL reaction mixture consisting of the labelling and enzyme solutions at a 10:1 ratio at 37°C for 1 hour with light shaking. Negative controls were incubated in the labelling solution only. Samples were again washed 3 times with 1xPBS and resuspended in 250µl 1xPBS for analysis via microscopy or flow cytometry.

TUNEL samples were primarily analysed via flow cytometry after confirmation of reaction via microscopy. Analysis was carried out using the Beckman CytoFLEX Flow Cytometer at the Biomedical Research Centre (BMRC) at UEA and the John Innes Centre (Norwich Research Park). Samples were first diluted in 1 ml of sterile SOW and pulse vortexed to reduce aggregation of cells. Initial gates were defined based on cell size using the forward (FSC-A::FSC-A) and side (SSC-A::SSC-A) scatter to ensure any contaminants were not analysed and to remove duplets (Figure 2.2). The forward scatter is typically considered a measure of the relative cell size, whereas side scatter is used to measure the relative complexity of the cell. The fluorescein isothiocyanate (FITC) CytoFLEX laser detected the TUNEL fluorochrome. The excitation wavelength was 488nm and detection was set to

525nm, and 100,000 events were recorded for each sample. The FlowJo™ software (BD Life Sciences) was used to analyse data and create figures.

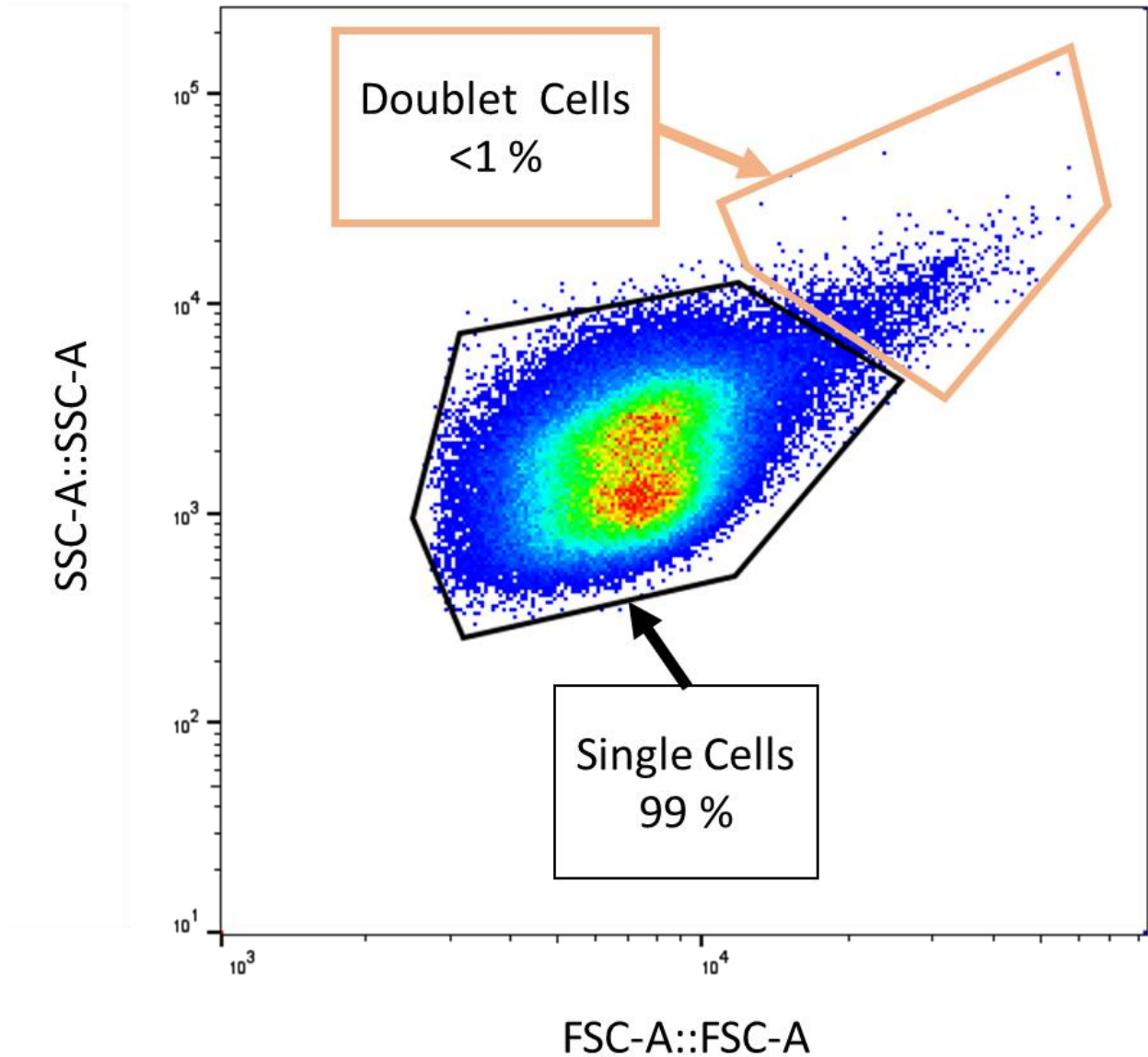


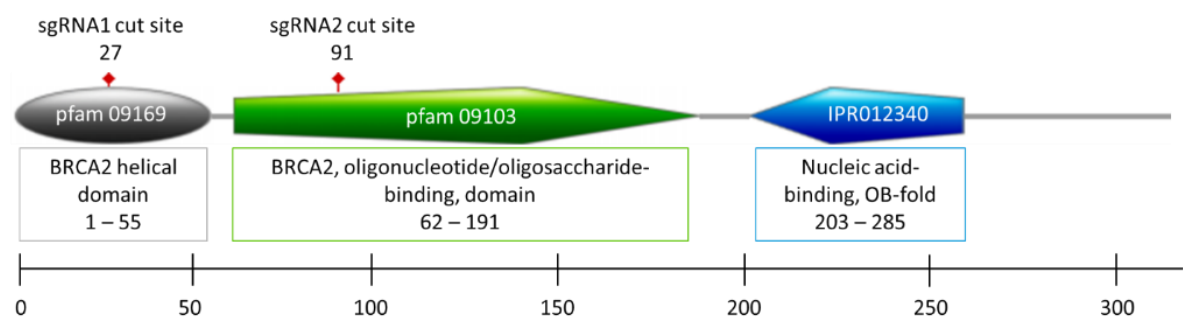
Figure 2.2. Initial gates set up using Forward scatter on the x-axis (FSC-A) and Side scatter on the y-axis, to reduce doublet cells being recorded for analysis. Both the x and y axis are log scales. Figure made with FlowJo software.

## 2.3 Genome Editing in *T. pseudonana* via CRISPR/Cas

### 2.3.1 sgRNA Design and *In Vitro* Analysis

For both *Brca2* and *Ku70*, two single guide RNAs (sgRNAs) were designed and tested that met the following criteria, A) directed to conserved domains within the coding region of the gene, B) high specificity to the target gene to reduce off-target effects, C) high cutting efficiency *in vitro* and D) reduced potential of secondary structures. Conserved domains were confirmed through protein alignments. The BRCA2 helical domain through the OB1 fold (Figure 2.3a) and Beta-Barrel domain were targeted for BRCA2 and KU70 (Figure 2.3b) respectively. Disruption of these domains and creation of frame shifts in other organisms has resulted in non-functional or significantly reduced protein function (Le et al., 2021; Ghosh et al., 2022). For each gene five possible sgRNAs were created based on the following aspects, on-target scores, melt temperature, percent of GC content, and their potential for secondary structures.

**A)**



**B)**

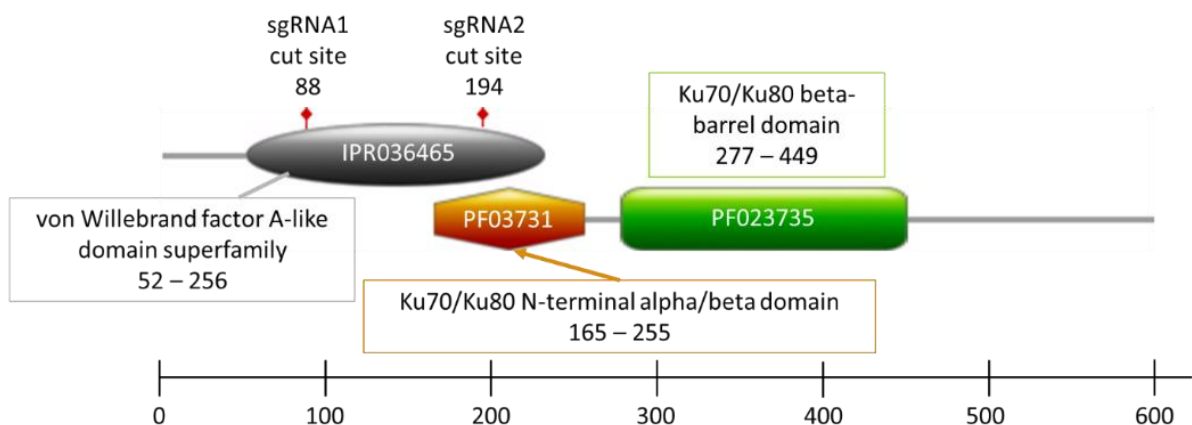


Figure 2.3. Representational figures of the location of conserved domains and Cas9 cut sites for A) *Brca2* and B) *Ku70*. Numbers for each domain show position and length of amino acids and show the site of targeted cutting for sgRNA1 and sgRNA2.

The sgRNAs were synthesized using oligos ordered from Eurofins Genomics UK (<https://www.eurofins.co.uk/genomic-services/>). The requisite components for the HiScribe T7 High Yield RNA Synthesis Kit (NEB; for short RNA transcripts) were added to the oligos via PCR. The resulting PCR templates were checked by running them on a 2% agarose gel in 1x TAE and purified using the New England Biolabs PCR Clean-Up Kit (<5 µg). The purified templates were then used to synthesize small RNA transcripts using the HiScribe T7 High Yield RNA Synthesis Kit, following the standard protocol. The reaction was incubated overnight, and the resulting products were purified using the Zymo Clean and Concentrator Kit (Zymo, Catalogue # D4033) to remove DNA contaminants. The resulting transcripts were confirmed to be the correct size and DNA free by running them on an 8% 7M urea denaturing gel (0.5x Tris/Borate/EDTA [TBE]) at 200V for 30-40 minutes. Gels were stained with SYBR Gold nucleic acid gel stain (Invitrogen CAT# S11494) and imaged on a transilluminator.

An *in vitro* temperature assay, developed by Dr Amanda Hopes, was used to test the cutting efficiency of the sgRNAs when paired with the Cas9 protein *in vitro*. This method was optimized to maximize the chances of successful gene editing *in vivo* at the temperature at which *T. pseudonana* is cultured. First, 30 nM of the RNA transcripts, 30 nM of *Streptococcus pyogenes* Cas9 protein (NEB # M0386S) and the appropriate reaction buffer were incubated at 25°C for 15 minutes to form a ribonucleoprotein (RNP) complex. Then, the linearized plasmid containing the corresponding target gene was added to the reaction and incubated at 20°C overnight (*T. pseudonana* culturing temperature). The stop buffer from Hopes et al., 2017 was added to halt the reaction at 80°C for ten minutes and the reaction was immediately loaded onto a 1x TAE agarose gel. We imaged the gel on a transilluminator to visualize the cutting efficiency of each guide RNA. Successful cutting will produce up to three bands, one for the remaining uncut plasmid, and the other two fragments are products of a double-strand break cut. ImageJ software was used to calculate the cutting efficiency of each guide RNA (Schindelin et al., 2012). The two guide RNAs with the highest cutting efficiency were selected for further use in the constructs for biolistic transformation.

### 2.3.2 Golden Gate Cloning

Golden Gate cloning is a powerful and widely used DNA assembly technique that enables the construction of complex, multi-component DNA constructs in a single reaction (Engler and Marillonnet, 2014). This method relies on the use of type IIS restriction enzymes, which cleave DNA outside their recognition sequence, generating defined overhangs that can be used for seamless, directional ligation of multiple DNA fragments. In Golden Gate cloning, the DNA fragments to be assembled are flanked by specific sets of type IIS restriction enzyme recognition sites that enable their ordered and precise assembly into a final construct. This method offers several advantages over traditional cloning methods, including high efficiency, flexibility, and scalability. By using the methods outlined in Hopes et al., 2017, final constructs were cloned over three stages.

First, the newly designed sgRNAs were directly assembled into the level 1 vector after PCR. The unique 20 nucleotide target sequence represented by the run of N's is encoded into the forward primer: aggtctcattgtNNNNNNNNNNNNNNNNNNNGTTTTAGAGCTAGAAATAGCAAG. Using the reverse primer: tggctcaagcgTAATGCCAACTTTGTACAAG, the final products were amplified from the pICH96966::AtU6p::sgRNA\_PDS scaffold (Addgene plasmid # 46966; <http://n2t.net/addgene:46966>; RRID: Addgene\_46966). PCR products were purified using the Monarch® PCR & DNA Cleanup Kit (5 µg) (New England Biosciences, Catalogue # T1030L). Two sgRNAs (sgRNA1 and sgRNA2) were made per gene to encode two cut sites to improve the chances of impairing protein functionality (Hopes et al., 2017).

Next sgRNA1 and sgRNA2 were cloned with the level 0 U6 promoter using the restriction enzyme *Bsa*I (NEB, Catalogue # R0535) and T4 DNA ligase (Promega, Catalogue # M1794) into the vectors pICH47751 (Addgene plasmid # 48002; <http://n2t.net/addgene:48002>; RRID: Addgene\_48002) and pICH47761 (Addgene plasmid # 48003; <http://n2t.net/addgene:48003>; RRID: Addgene\_48003) respectively (Table 2.2). The U6 promoter, an RNA Polymerase dependent promoter, has been used widely to control the transcription of sgRNAs due to their precise transcription start sites and consistent control of the transcript length (Long et al., 2018).

Table 2.2. Quantity of reagents used for level 1 Golden Gate assembly cloning reaction.

Reagent	Quantity
L1 Vector	40 fmol
L0 Modules/PCR Product	40 fmol/each
<i>Bsal</i>	10 U
T4 DNA Ligase	10 U
10x ligation Buffer	2 $\mu$ l
Nuclease-free water	Add up to 20 $\mu$ l

All reagents were incubated for 5 hours at 37°C for restriction digest, followed by 5 minutes at 50°C for final ligation and 10 minutes at 80°C to deactivate all enzymes, then directly transformed into NEB® 5-alpha Competent *E. coli* (High Efficiency) (Catalogue # C2987H) following the manufacturer's protocol. The transformed *E. coli* were selected on agar plates containing 50 $\mu$ g/ml carbenicillin, as the level 1 vectors contain the carbenicillin resistance cassette. Extracted plasmids were screened via restriction digest with *Xba*I.

Successfully cloned level 1 vectors (pICH47751\_TpU6:sgRNA1 and pICH47761\_TpU6:sgRNA2) were then cloned into a final level two vector, pAGM4723 (Addgene plasmid # 48015; <http://n2t.net/addgene:48015>; RRID: Addgene\_48015), together with the following components: pICH47732\_TpFCP:NAT containing the nourseothricin resistance gene for selection, pICH47742\_TpFCP:Cas9:YFP containing the Cas9 coding sequence and yellow fluorescence protein tag (YFP), and pICH41780 which links the final ligation (Table 2.3).

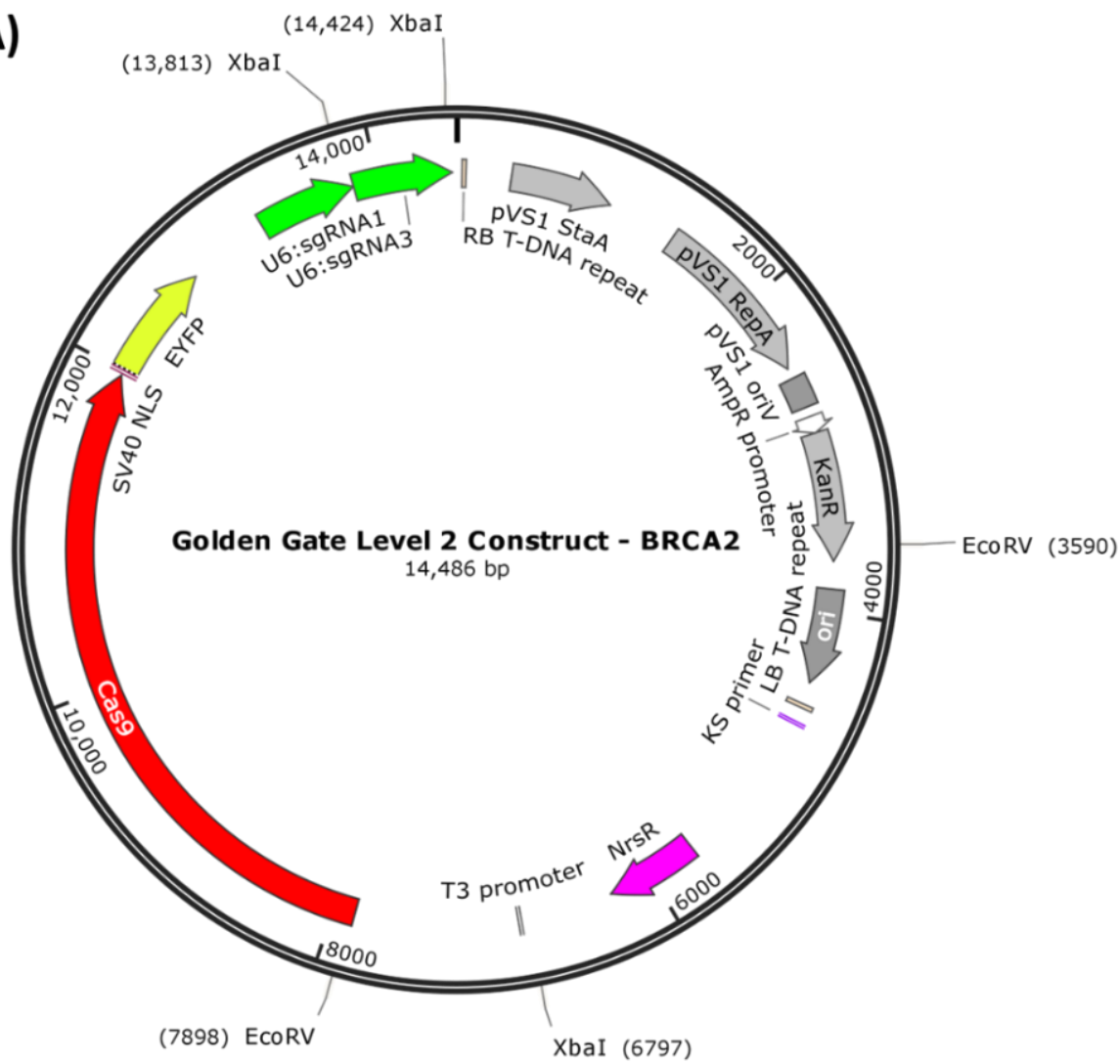


Table 2.3. Quantity of reagents used for level 2 Golden Gate assembly.

Reagent	Quantity
pICH47732_TpFCP:NAT	40 fmol
pICH47742_TpFCP:Cas9:YFP	40 fmol
pICH47751_TpU6:sgRNA1	40 fmol
pICH47761_TpU6:sgRNA2	40 fmol
pICH41780 (L4E linker)	40 fmol
pAGM4723 (L2 backbone)	40 fmol
Bpil	10 U
T4 DNA ligase	10 U
10x ligation buffer	2 $\mu$ l
Nuclease-free water	Up to 20 $\mu$ l

Same as level 1, the cloning reaction was incubated for 5 hours at 37°C, followed by 5 minutes at 50°C and 10 minutes at 80°C, then directly transformed into NEB® 5-alpha Competent *E. coli* (High Efficiency). The extracted plasmids were screened by both restriction digest with EcoRV and Sanger sequencing using primers that anneal within each inserted component. Final constructs for both *Brca2* and *Ku70* were purified and stored at -20°C (Figure 2.4).

A)



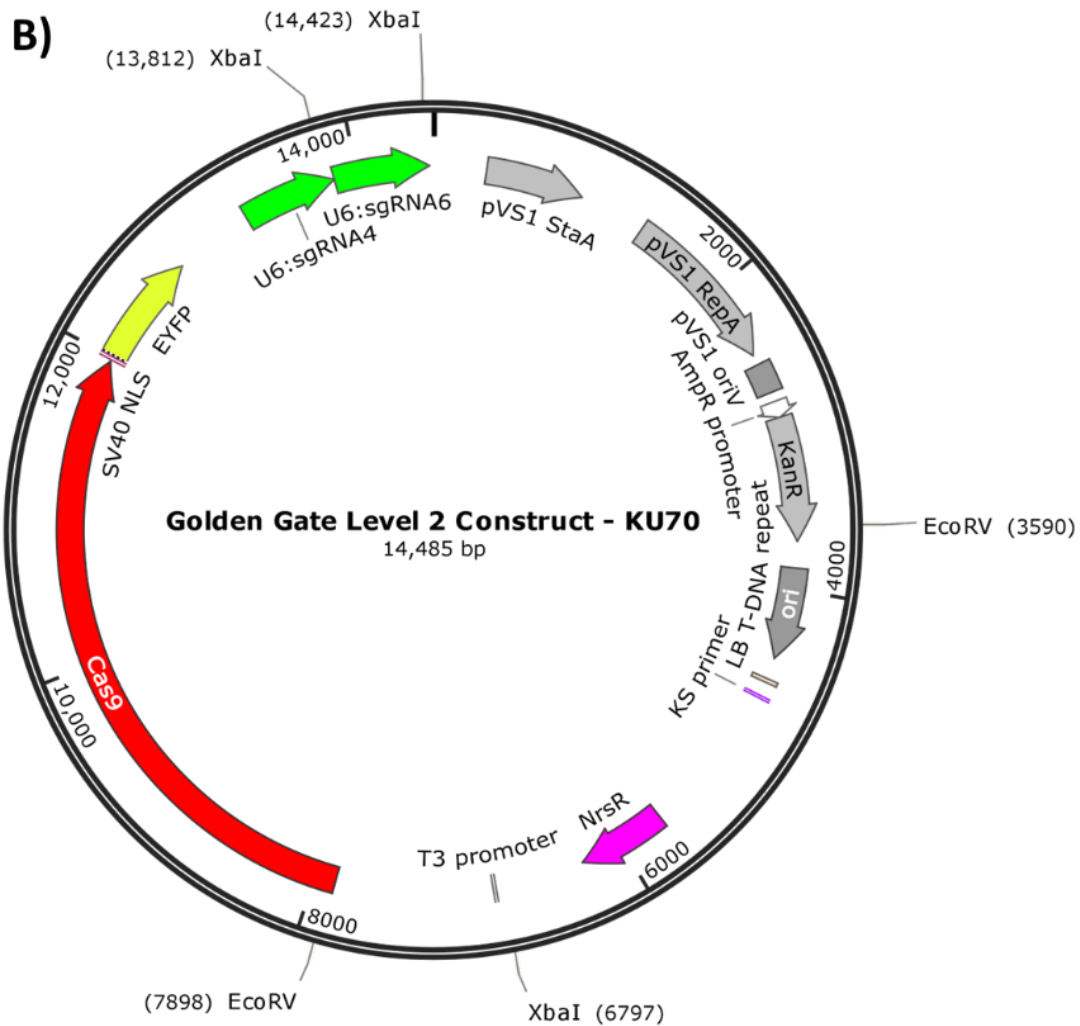


Figure 2.4. Vector map of final plasmid constructs for A) BRCA2 and B) KU70. Components cloned into vector: U6:sgRNAs (green) containing U6 promoter and sgRNA sequence, NrsR (pink) containing resistance gene to clonNAT antibiotic, and Cas9 (red) containing Cas9 protein. Eco-RV and XbaI indicate respective cut sites for restriction digest. Constructed using golden gate cloning. Maps made using SnapGene® (<https://www.snapgene.com/>).

### 2.3.3 Biolistic Transformation

Before transforming wild type *T. pseudonana* was acclimated to growth in Aquil media containing half concentrations of anhydrous and hydrous salts provided in table 2.1 (half salinity Aquil), as transformants were selected for on agar plates at half salinity. Wild-type *T. pseudonana* was grown in half salinity Aquil for at least 5 transfers in the exponential phase. After acclimation, cultures were subject to treatment with antibiotics to remove bacteria contaminants. Antibiotics were added to the media (Table 2.4) and cultures were

grown with antibiotics for two generations, after which were transferred to fresh, antibiotic-free media.

Table 2.4 Final concentration of antibiotics added into media to create axenic cultures.

Antibiotic	Concentration
Streptomycin	100 µg/ml
Penicillin G	500 µg/ml
Ampicillin	700 µg/ml
Cefotaxime	150 µg/ml

Axenic *T. pseudonana* CCMP1335 were grown to a density of  $\sim 1 \times 10^6$  cells/ml with  $5 \times 10^7$  cells filtered onto a 1.2 µm filter (Millipore) for each transformation and the filter placed on a 1.2% agar plate (shooting plates). Each transformation was done in triplicate. Tungsten (0.7mm) particles were coated with the plasmids (Figure 2.4) using 50 µl 2.5M CaCl<sub>2</sub> and 20 µl 0.1M Spermidine. The tungsten particles were washed with 100% ethanol and resuspended in a final volume of 30 µl ethanol and placed on ice until use. The components of the PDS-1000/He Particle Delivery System (Bio-Rad) were sterilised and assembled, the plasmid was loaded onto 1335 PSI rupture discs and the cells were placed 7cm below in a vacuum-sealed chamber. After transformations, the filters were aseptically transferred into 25ml of half-salinity Aquil for 24 hours under normal growth conditions. Each culture was spread onto 5 individual 0.8% agar selection plates containing 100 µg/ml of the selection antibiotic, nourseothricin (clonNAT; Jena Bioscience, Catalogue # AB-102), and grown at 20°C under constant illumination until colonies appeared (>14 days).

#### 2.3.4 Selection of Transformed Colonies

Both genes targeted by CRISPR/Cas are haplosufficient, meaning only one allele is required to carry out the function of the protein. For the protein's efficiency to be impaired, and disrupt their respective pathways, both alleles needed to be edited. The CRISPR/Cas constructs were designed to cut out sections of conserved domains within each gene (Figure 2.3), resulting in truncated gene and non-functional proteins. Single colonies were picked and resuspended in 10 µl Aquil with 100 µg/ml clonNAT; 5 µl was inoculated into 150 µl

Aquil containing 100 µg/ml clonNAT, and the remaining 5 µl lysed for colony PCR to screen for possible editing. The entire gene sequence was amplified via PCR, and products were run on an agarose gel (2% agarose for *Brca2* products; 0.8% agarose for *Ku70* products). Any potential editing resulted in two bands, with edited genes showing as smaller bands. The first colonies to appear (primary colonies) are usually mosaic colonies consisting of a mixture of populations, some edited and some with wild-type targeted genes (Hopes et al., 2017). Mosaic colonies showing evidence of editing were separately inoculated into 3 ml Aquil containing 100µg/ml clonNAT and streaked onto selective plates to obtain clonal bi-allelic (i.e., homozygous) knock-out sub-clones. PCR products from sub-clones that showed evidence of successful bi-allelic editing were sent for Sanger sequencing. The sequenced target genes were aligned using MEGA (Molecular Evolutionary Genetics Analysis) to confirm the loss of the intended portion of the gene (Tamura et al., 2021). Confirmed bi-allelic knock-out cultures were cultured continuously in half-salinity Aquil supplemented with 100µg/ml clonNAT to keep selection for mutant cell lines. The efficiency of transformations is presented as a percentage of the number of obtained transformant colonies which grew on selective media (Belshaw et al., 2022).

## 2.4 DNA Extraction

### 2.4.1 Extraction of High-Molecular-Weight (HMW) genomic DNA

#### (gDNA)

Diatom cultures were first stained with 4',6-diamidino-2-phenylindole (DAPI) to confirm they were axenic and subsequently grown up into larger volumes (>250ml). Cells in the exponential phase ( $\sim 5.0 \times 10^5$  cells/ml) were harvested through gentle filtration and separated into several pellets. The Qiagen MagAttract HMW DNA Kit (Qiagen, Catalogue # 67563) was used, but given the difficulties in obtaining HMW from diatoms, the protocols were amended. In personal communication with Dr Darren Heavens at the Earlham Institute, the protocol, "*Manual Purification of High-Molecular Weight-Genomic DNA from Gram-Negative Bacteria*" from the handbook was amended to improve DNA extraction from diatoms. The amendments were to double all reagents, increase lysis incubation time, add a third wash in PE buffer, and pulse centrifuge the final eluted DNA to pellet any

remaining salts before purification. The resulting gDNA was cleaned up via ethanol precipitation overnight and analysed through spectrophotometry (Nanodrop 2000; ThermoFisher) and pulse field gel electrophoresis (PFGE).

## 2.4.2 DNA Extraction from Small Amounts of Cells

The protocol “*Small amounts of cells, tissues, or plant leaves*” from the EasyDNA kit (Invitrogen) was used when cultures could not be grown in larger volumes due to time constraints, the number of samples or the availability of equipment. The protocol was amended to increase the quantity of gDNA extracted, by increasing lysis incubation from 10 minutes to between 5 and 8 hours. gDNA was cleaned up by ethanol precipitation overnight and quantified using spectrophotometry via the Nanodrop 2000 (ThermoFisher) and PFGE.

## 2.5 Bioinformatics

### 2.5.1 DNA Repair Proteins in Diatoms

A reference set of core and associated DNA repair genes were collated from MD Anderson ([Human DNA repair genes \(mdanderson.org\)](http://mdanderson.org)) and the Gene set enrichment analysis webpage ([GSEA \(gsea-msigdb.org\)](http://gsea-msigdb.org)). Human genes were chosen as a reference as there is a significant amount of research concerning DNA repair genes both experimentally and *in silico*, given their role in major diseases, such as cancer (Helleday et al., 2008). Well annotated proteomes (i.e., *Mus musculus*, *Arabidopsis thaliana*, and *Saccharomyces cerevisiae*) were used as internal controls to confirm the methodology was accurately predicting homologues of DNA repair genes in diatom proteomes. A total of 47 reference proteomes from across the three domains of life (Bacteria, Archaea, and Eukaryotes) were downloaded from UniProt (<https://www.uniprot.org/>) to create a robust dataset to account for phylogenetic diversity of DNA repair genes (Table 2.5). The reference proteomes selected focused on model organisms from taxonomic groups closely related to diatoms, such as green and red algae. The remaining proteomes were selected based on both their level of genome annotation to improve the prediction of DNA repair proteins and placement on the tree of life to capture the diversity in the composition of DNA repair proteins between distantly related organisms.

Table 2.5 Reference proteomes used to search for DNA repair orthologues. Species are characterised to the phylum level for taxonomy; genome size data is from the European Nucleotide Archive (<https://www.ebi.ac.uk/ena/browser/home>); and the UniProt Proteome ID is the name of the reference proteome on UniProt.

Domain	Taxonomy	Species	Genome Size Mbp	UniProt Proteome ID
Bacteria	Actinomycetota	<i>Micromonospora rosaria</i>	7.38	UP000070620
	Actinomycetota	<i>Streptomyces pluripotens</i>	7.59	UP000031501
	Alphaproteobacteria	<i>SAR11 cluster bacterium</i>	0.66	UP000050662
	Alphaproteobacteria	<i>Pelagibacteriales bacterium</i>	1.21	UP000618275
	Cyanobacteriota	<i>Prochlorococcus marinus</i>	1.69	UP000000788
	Cyanobacteriota	<i>Thermosynechococcus vestitus</i>	2.59	UP000000440
	Cyanobacteriota	<i>Gloeobacter violaceus</i>	4.66	UP000000557
	Cyanobacteriota	<i>Acaryochloris marina</i>	8.36	UP000000268
	Gammaproteobacteria	<i>Escherichia coli</i>	4.64	UP000000625
	Gammaproteobacteria	<i>Pectobacterium atrosepticum</i>	5.06	UP000007966
	Spirochaetota	<i>Treponema bryantii</i>	3.43	UP000182360
	Spirochaetota	<i>Breznakiella homolactica</i>	4.65	UP000595917
Archaea	Candidatus Thermoplasmata	<i>Picrophilus torridus</i>	1.55	UP000000438
	Euryarchaeota	<i>Pyrococcus furiosus</i>	1.91	UP000001013
	Euryarchaeota	<i>Thermococcus kodakarensis</i>	2.09	UP000000536
	Euryarchaeota	<i>Haloferax volcanii</i>	4.01	UP000008243
Eukaryote	Arthropoda	<i>Drosophila melanogaster</i>	143.73	UP000000803
	Ascomycota	<i>Saccharomyces cerevisiae</i>	12.07	UP000002311
	Bacillariophyceae	<i>Phaeodactylum tricornutum</i>	27.45	UP000000759
	Bacillariophyceae	<i>Thalassiosira pseudonana</i>	32.44	UP000001449
	Bacillariophyceae	<i>Fistulifera solaris</i>	49.74	UP000198406
	Bacillariophyceae	<i>Pseudo-nitzschia multistriata</i>	56.77	UP000291116
	Bacillariophyceae	<i>Fragilariopsis cylindrus</i>	80.54	UP000095751
	Bacillariophyceae	<i>Thalassiosira oceanica</i>	81	UP000266841
	Chlorophyta	<i>Ostreococcus tauri</i>	12.92	UP000009170
	Chlorophyta	<i>Micromonas pusilla</i>	22	UP000001876
	Chlorophyta	<i>Chlorella variabilis</i>	46.16	UP000008141
	Chlorophyta	<i>Chlamydomonas reinhardtii</i>	111.1	UP000006906
	Chlorophyta	<i>Volvox reticuliferus</i>	133.07	UP000747110
	Chordata	<i>Nematostella vectensis</i>	356.61	UP000001593
	Chordata	<i>Alligator mississippiensis</i>	2161.71	UP000050525
	Chordata	<i>Mus musculus</i>	2728.22	UP000000589
	Dinophyceae	<i>Symbiodinium microadriaticum</i>	808.23	UP000186817
	Dinophyceae	<i>Symbiodinium pilosum</i>	1089.42	UP000649617
	Dinophyceae	<i>Polarella glacialis</i>	2984.68	UP000654075
	Nematoda	<i>Caenorhabditis elegans</i>	100.27	UP000001940
	Phaeophyceae	<i>Ectocarpus sp. CCAP 1310/34</i>	242.36	UP000485383
	Placozoa	<i>Trichoplax sp. H2</i>	94.88	UP000253843

Table 2.5 continued  
from previous page

Porifera	<i>Amphimedon queenslandica</i>	166.68	UP000007879
Rhodophyta	<i>Cyanidiococcus yangmingshanensis</i>	12.07	UP000530660
Rhodophyta	<i>Galdieria sulphuraria</i>	13.71	UP000030680
Rhodophyta	<i>Porphyridium purpureum</i>	22.19	UP000324585
Rhodophyta	<i>Gracilariopsis chorda</i>	92.18	UP000247409
Stramenopiles	<i>Nannochloropsis gaditana</i>	27.59	UP000019335
Streptophyta	<i>Arabidopsis thaliana</i>	119.67	UP000006548
Streptophyta	<i>Oryza sativa subsp. japonica</i>	373.8	UP000059680
Streptophyta	<i>Zea mays</i>	2182.08	UP000007305

The reference set of genes was queried using phmmer (HMMER 3.3; Nov 2019; <http://hmmer.org/>), against diatom proteomes available through UniPot. All results were reported regardless of e-values or scores. All further analysis was conducted on filtered data, which consisted of any hits with an e-value for both the full sequence and single best domain  $< 1 \times 10^{-4}$  and where the bias was  $< 0.1$  of the full sequence score as recommended in the HMMER manual (<http://eddylab.org/software/hmmer/Userguide.pdf>). Many of these genes have either been duplicated (Henikoff and Greene, 1997), or in distantly related proteomes the conserved domains have been shuffled (Aravind et al., 1999), which meant some query sequences' best hit overlapped with that of other distinct query sequences. To resolve these issues, the reference set of DDR proteins was also queried against all proteomes using OrthoFinder (Emms and Kelly, 2019). OrthoFinder grouped hits into orthogroups, which are family of genes that are descended from a single gene at the time of the least common ancestor and includes paralogs within each species (Emms and Kelly, 2019). These orthogroups accounted for the diversity of protein copy numbers and domain shuffling within distantly related taxonomic groups.

The presence of DNA repair genes predicted through HMMER and OrthoFinder searches were used to create heat maps through the Broad Institute online tool, Morpheus (<https://software.broadinstitute.org/morpheus>). These helped to visualise the conservation of functionality across diatom species. In addition, a regression analysis was run in base R (R 4.3.1 [2021-11-01]) between the number of hits and genome size (length in bp) to determine if genome size appears to influence the number of DNA repair genes.



## 2.5.2 Phylogenetic Analysis of *Brca2* and *Ku70*

Amino acid sequences of each gene were downloaded from the genome of *T. pseudonana* through the Joint Genome Institute genome portal (<https://mycocosm.jgi.doe.gov/Thaps3/Thaps3.home.html>) and were queried against the UniProt reference proteomes (<https://UniProt.org>) using the phmmer tool (Potter et al., 2018; HMMER 3.3). Full-length fasta sequences of significant hits (e value <0.001) were downloaded from representative species from Heterokonts, Chlorophyta, Rhodophyta, Streptophytes, fungi (Oomycetes), Stramenopiles, Chordata, and bacteria. Sequences were loaded into MEGA and aligned using Multiple Sequence Comparison by Log-Expectation (MUSCLE; Edgar, 2004) with the following parameters: gap open penalty = -2.9, gap extend penalty = 0, hydrophobicity multiplier = 1.2, max iterations = 8, and using the unweighted pair group method with arithmetic mean (UPGMB; Sokal and Michener 1958; Edgar, 2004). Alignments were saved in a multiple sequence alignment (MSA) format and a neighbour-joining tree was created through MEGA, over 1000 bootstrap runs using the Poisson substitution model and a gamma distributed (G) rate of 2. Outgroups consisting of diatom sequences of *Brca2* and *Ku70* were used in the *Ku70* and *Brca2* alignments respectively.

## 2.5.3 Variant Calling Pipeline (WGS data)

### 2.5.3.1 FASTQC and K-mer analysis

Before any further analysis, the raw reads were analysed using FASTQC to assess their overall quality. Raw reads with an overall high quality with no poor reads, were passed into the gcp tool from the K-mer Analysis toolkit (KAT; Mapleson et al., 2016) for further analysis. This counts the GC nucleotides in each distinct K-mer from two FastQ files to identify any significant contaminating sequences. Contaminating sequences will appear at unexpected GC and coverage levels.

### 2.5.3.2 Read Mapping to the *T. pseudonana* Reference Genome

Raw reads were aligned to the reference *T. pseudonana* genome (Armbrust et al., 2004) using the Burrows-Wheeler Alignment Tool (BWA-MEM; Li and Durbin, 2010) which

performs local alignment using medium to long reads (70bp – 1Mbp). This command resulted in .bam files containing alignment information. The accuracy of alignments was assessed using QualiMap BamQC before variant calling (Okonechnikov et al., 2015).

### 2.5.3.3 *Marking Duplicates Reads and Variant Calling*

A variant calling script was written based on previous work done in the Mock research group with help from Dr Andrew Toseland utilising samtools (Danecek et al., 2021) to call variants. Samtools was used to filter the resulting bam files for a mapping quality of 30 and removed duplicate aligned reads to minimise false positive variants created by errors in the library preparation process. These were then piped into the samtools mpileup tool which created an overview of the mapped reads and their associated coverage at a base pair resolution and prepares the file for variant calling with bcftools call (Li, 2011). Bcftools call was used to call variants and the resulting bcf files were filtered to only report variants with a variant calling quality >30, a minimum read depth of x5, a minor allele frequency of > 0.05, a gap of at least 10bp between single nucleotide polymorphisms (SNPs) and a distance of at least 5bp from small insertions and deletions (INDELS). After variants had been called and filtered, the bcf files were converted into .vcf files for annotation using SnpEff (Cingolani et al., 2012). A SnpEff database (SnpEffdb) was created for both the *T. pseudonana* reference genome (Armbrust et al., 2004) and the Flye assembly (Filloramo et al., 2021). SnpEff was then used to annotate the called variants, notably for genes affected and predicted level of impact on genetic information.

### 2.5.4 Calculation of Mutation Rates

First, in addition to the filtering applied to the called variants above, all variants shared with a sample and the resequenced reference wild type genome were removed, leaving only novel variants. Using this number of SNPs, the mutation rate was calculated using the following formula where  $\mu$  = mutation rate.

$$\mu = \frac{\text{Number of novel SNPs}}{\text{generations} * \text{genome size}}$$

### 2.5.5 Go Term Enrichment Analysis

All GO term enrichment analysis in this thesis were conducted using gProfiler (Kolberg et al., 2023). Lists of genes for analysis were input into the gProfiler2 web-based tool – g:GO st functional profiling (<https://biit.cs.ut.ee/gprofiler/gost>). The following options were used: the organism to compare against was set to *T. pseudonana*, all GO terms and protein databases were used and the significance threshold was set to a p value of < 0.05 after Bonferroni correction.

### 2.5.6 Detection of Copy Number Variation

Copy number variation (CNV) was detected using the depth of sequencing of each sample compared to the resequenced reference wild type genome. First the location of all genes in the *T. pseudonana* genome were extracted from the genomic features file downloaded from <https://genome.jgi.doe.gov/portal/pages/dynamicOrganismDownload.jsf?organism=Thaps3>; Armbrust et al., 2004). Then the average sequencing depth for each coding region was directly compared to the average sequencing depth across the whole genome. In concordance with previous methods developed by Dr Andrew Toseland in the Mock research group (unpublished), any coding region which had either 25% more or less coverage was determined to have a gain or loss in CNV. Plots for CNV were created using ggplot2 in R.

### 2.5.7 Detection of Loss of Heterozygosity (LOH)

Regions of loss of heterozygosity were defined as regions which had a least three consecutive homozygous SNPs in resequenced genomes with the same SNPs heterozygous in the sequence wild-type reference genome. After all variants had been called using the previously mentioned methods, all INDELS and homozygous SNPs were removed from the wild-type reference genome. The locations of the remaining heterozygous SNPs was extracted as a .bed file. All variant call files were filtered to report only the locations within the .bed file. All heterozygous mutations were removed so that only SNPs which changed

from heterozygous to homozygous remained for each sample. Finally, any regions with three consecutive SNPs within close proximity to each other were defined as a LOH event.

## 3 DNA Repair Systems in Diatoms

### 3.1 Introduction

The repair of DNA lesions is a critical cell process for maintaining life (Tubbs et al., 2017). Mammalian cells are estimated to experience up to  $10^5$  DNA-damaging events per day (Lindahl and Barnes, 2000; Hoeijmakers, 2009). These can be caused by both external (i.e., ultraviolet light) and internal factors (i.e., replication errors). Most lesions are single-strand breaks (SSBs) and mismatched bases (Ciccia and Elledge, 2020), while more deleterious double-strand breaks (DSBs) occur at a much lower frequency and are generally created by exogenous factors (Mehta and Haber, 2014). DNA repair can also be a source of genetic diversity in microbial communities. Organisms have evolved finely tuned DNA repair systems which enable a degree of genomic instability at the risk of deleterious mutations (Vincent and Uphoff, 2021).

DNA damage repair (DDR) is a complex process comprising many core and accessory proteins. Some of these are transcriptionally regulated, but the vast majority are post-translationally regulated by kinases (Sirbu and Cortez, 2013). The repertoire of core DDR genes is highly conserved while the number of accessory proteins differs between distant evolutionarily taxonomic groups, but the overall function of each pathway is conserved (Nischiwitz et al., 2023). For example, multicellular eukaryotes have evolved multiple accessory proteins through gene duplications to create redundancy within DDR to increase genome stability (Nischiwitz et al., 2023). Some studies have argued the increase in the number of DNA repair genes is directly correlated to genome size (Voskarides et al., 2019), however, these studies analyse a narrow taxonomic group creating a bias and overlooking phylogenetically distant organisms (Udroiu, 2020). The genes that comprise DNA repair systems in diatoms are currently inferred through automated annotation and literature from other species with little experimental evidence of their function.

#### 3.1.1 Overview of DNA Repair Pathways and Mechanisms

DNA repair pathways are highly conserved throughout the domains of life to preserve DNA integrity (Nischiwitz et al., 2023). Depending on the type of DNA insult, the appropriate

pathway is activated (Brandsma and Gent, 2012). While certain DNA insults can always trigger specific pathways, it is important to note that other factors, such as the cell cycle, can influence the pathway used for repair. This can even result in 'competition' of pathways and sub-pathways to repair DNA damage. The major DNA repair pathways are homologous recombination (HR; Li and Heyer, 2008), non-homologous end-joining (NHEJ; Chang et al., 2017), base excision repair (BER; Krokan and Bjørås, 2013), nucleotide excision repair (NER; Schärer, 2013) and mismatch repair (MMR; Jiricny, 2006; Lindahl and Barnes, 2000; Figure 3.1). These pathways incorporate a diverse array of proteins, including ligases, helicases, recombinases, kinases, phosphatases, polymerases, topoisomerases, and demethylases (Ciccio and Elledge, 2016).

DNA repair systems are tightly regulated to ensure appropriate maintenance of DNA (Jackson and Bartek, 2009). The primary regulators of DNA repair pathways are kinases, specifically the DNA checkpoint protein complexes serine/threonine protein kinases ataxia telangiectasia (ATM) and ataxia telangiectasia Rad3-related protein (ATR), the DNA-dependent protein kinase (DNA-PK) and DNA-dependent protein kinase (PARP) family proteins (Meek et al., 2008). These proteins detect DNA damage by recognising single-stranded DNA (ssDNA) coated in replication protein A (RPA), and recruit repair proteins through phosphorylation (Karakaidos et al., 2020). Additional regulators of the DDR and cell cycle are checkpoint proteins. These can be categorised based on their function as sensors of DNA damage, transducers that relay signals or effectors which execute the checkpoint response (Bartek and Lukas, 2007). Checkpoint proteins serve to ensure the cell cycle only progresses when safe to do so by regulating the activation of effector proteins. Checkpoint proteins functions are core to multiple DNA repair pathways such as ATR while some are specific, such as the MRN complex which senses DNA DSBs. Checkpoint proteins that have roles in multiple DNA repair pathways are not exclusively described in the following subsections. Briefly, core checkpoint proteins were found in diatoms such as ATM and ATR. Not surprisingly, the number of checkpoint proteins in diatom proteomes is much smaller than the number found in multicellular organisms such as humans.

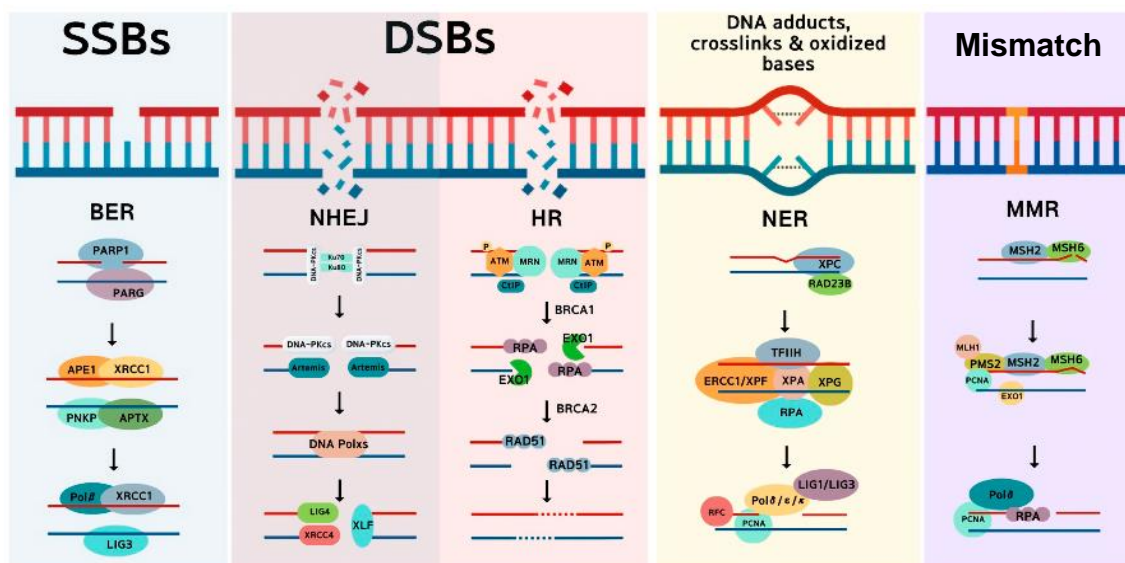


Figure 3.1. DNA repair systems and the DNA insults they primarily repair. Single-strand breaks (SSBs) and double-strand breaks (DSBs). Base Excision Repair (BER); non-homologous end-joining (NHEJ); homologous recombination (HR); nucleotide excision repair (NER); and mismatch repair (MMR). Based on figure from Moon et al., 2023.

### 3.1.1.1 Double-strand Break Repair Homologous Recombination

Homologous recombination (HR) is conducted by a suite of proteins to repair DNA double-stranded breaks (DSBs) and interstranded crosslinks (ICLs; Krejci et al., 2012). HR occurs during late S through G2 phases of the cell cycle as a homologous template is required. HR has three conceptual steps: pre-synapse, synapsis, and post-synapse (Figure 3.2; Li and Heyer, 2008). A detailed introduction and description of the HR pathway and the genes involved and their functions has been provided in Chapter 1 section 1.2. Briefly, the MRN protein complex (Qiu and Huang, 2021), composed of DNA repair protein Rad50 (RAD50), Nibrin (NBS-1), double-strand break repair protein Mre11 (MRE11), and ATM, scans DNA for double-strand breaks (DSBs). After detection the DNA is unwound by helicases and the 3' ends are resected to expose ssDNA. Rad51 is loaded onto the ssDNA which promotes the search for homologous sequences to be used as a template for repair. Once a suitable homologous sequence has been identified the two strands form a complex and polymerases synthesise a new strand using the template. The complex is then dissolved

which can result in different combinations of recombination depending on the confirmation of the complex.

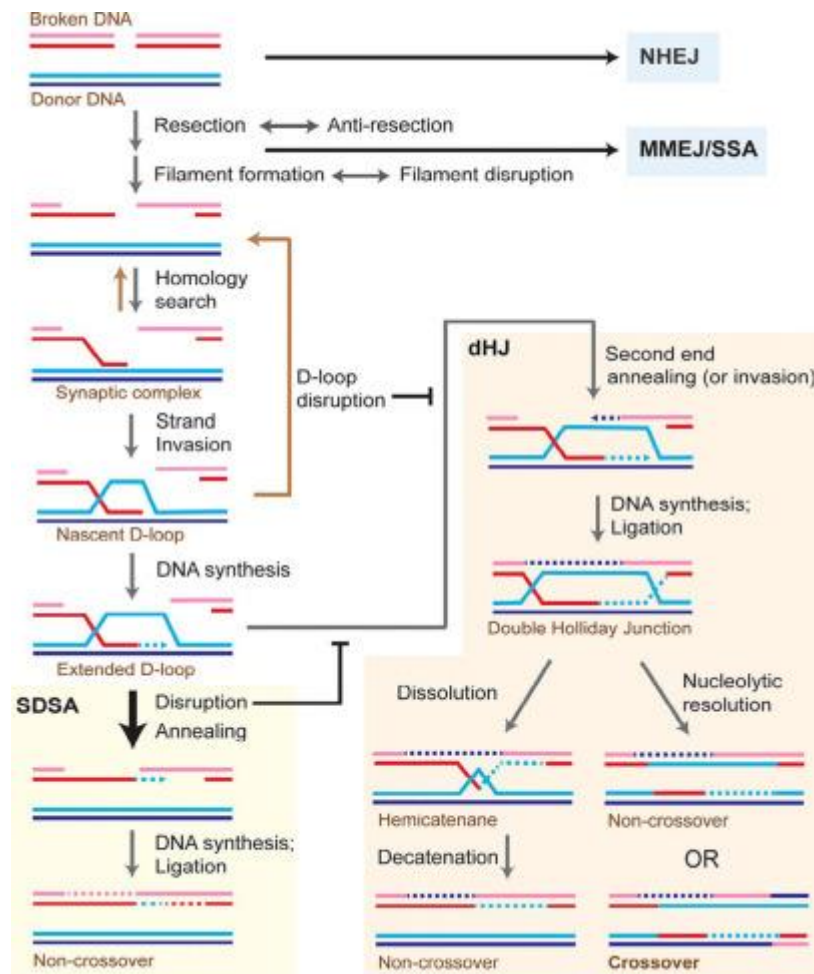


Figure 3.2. Pathway map of eukaryotic type homologous recombination from Wright et al., 2018. NHEJ = non-homologous end-joining. MMEJ/SSA = microhomology-mediated end joining/single-strand annealing. dHJ = double Holliday junction. SDSA = synthesis-dependent strand annealing.

### Non-Homologous End-Joining

Unlike HR, non-homologous end-joining (NHEJ) does not require a template for repair and occurs throughout the entire cell cycle. NHEJ requires fewer proteins, less energy to complete than HR and can be completed without a template homologous sequence. Since it can repair DNA throughout the cell cycle, it is the most frequently used DSB repair pathway in eukaryotes despite a higher probability of small insertions or deletions (Lieber, 2010).



The first stage involves the binding of proteins KU70 and KU80 (KU heterodimer) which binds to the free dsDNA ends (Ku:DNA complex) to prevent resection and loss of genetic information (Lieber, 2010; Davis and Chen., 2013). The KU heterodimer, also termed KU78, is a dimer between the proteins -x-ray repair cross-complementing 5 (XRCC5) 6 (XRCC6) in eukaryotes and is also known as KU70 and KU80 (Walker et al., 2001). Heterodimer formation is essential for function as the C-terminal region required to interact with further proteins is not present in the crystal structure of Ku70 or Ku80 alone (Walker et al., 2001). The Ku:DNA complex recruits and provides the substrate for DNA-Pkcs, X-ray cross-complementing protein 4 (XRCC4), XRCC4-like factor (XLF), Aprataxin-and-PNK-like factor (APLF), and DNA Ligase IV (DNLI4) to interact with the DNA. KU70 and KU80 directly interact with all recruited proteins. To complete repair, DNLI4, catalysed by XRCC4/XLF, ligate the ends together without polymerases filling in the gap (Figure 3.3; Davis & Chen 2013). There is no definite method known of how the NHEJ complex is disassembled, however, it is most likely through ubiquitination (Figure 3.3; Davis and Chen., 2013).

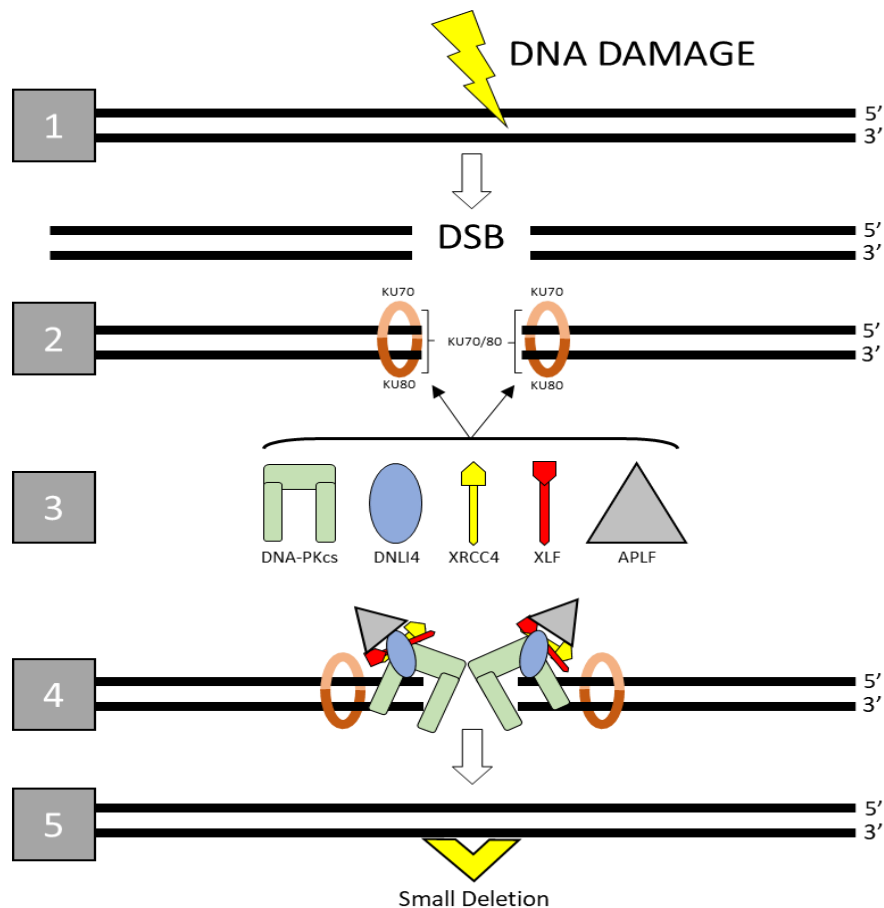


Figure 3.3: Depiction of major steps of NHEJ. 1) DNA damage and creation of DSB. 2) KU70-80 heterodimer detects and binds to free DNA ends. 3) Recruitment of further proteins. 4) Ligation and dissolution of NHEJ complex. 5) Resulting DNA ligated but with small deletion.

### 3.1.1.2 Single-strand Break Repair, Mismatch Repair, and Repair of DNA Adducts Crosslinks and Oxidized Bases

#### Base Excision Repair

The majority of DNA insults are induced by reactive oxygen species (ROS) produced by cellular metabolism which can react with DNA and create nicks and single-strand breaks (SSBs; Lindahl, 1993). The base excision repair (BER) pathway is of bacterial origin, has been characterised in all domains of life, and is the primary system of repair of SSBs. BER was first discovered by identifying uracil DNA glycosylase (UNG) in *E. coli*, a DNA glycosylase which repairs cytosines replaced by uracil after deamination (Lindahl, 1974). Briefly, the BER pathway consists of four steps after the detection of a base lesion by PARP family proteins: excision, incision, end processing, and either strand displacement leading to repair synthesis

or direct repair synthesis (Krokan and Bjørås., 2013). First, the erroneous base is removed through excision and incision by DNA glycosylases resulting in a location without a purine or pyrimidine base, an apurinic/apyrimidinic (AP) abasic site (Talpaert-Borlè, 1987; Krokan and Bjørås., 2013). An AP endonuclease then cuts out the AP site to expose 3' OH and 5' deoxyribose phosphates (Doetsch and Cunningham, 1990). Finally, the AP site is filled in by a polymerase, specifically in mammalian cells this is carried out by DNA polymerase  $\beta$  (Matsumoto et al., 1995).

### *Mismatch Repair*

Mismatch repair (MMR) is a highly conserved pathway, and its primary function is the repair mismatched bases and small insertions or deletions (INDELS) created during DNA replication (Kolodner and Marsischy, 1999; Li, 2008). MMR is a bidirectional excision and synthesis system which, in bacteria, starts with the MutS protein recognising mismatch bases within the DNA. Once MutS is bound, MutL and MutH homolog proteins are recruited. MutH searches the DNA for the nearest methylated GATC sequence on the parent strand and nicks the unmethylated daughter strand, which may be up to several kilobase pairs away. Helicases unwind the DNA at the nick site at which point endonucleases dissolve the excised DNA strand (Li, 2008). Polymerases then synthesize the daughter strand with the correct base in place. The pathway is similar in eukaryotes, with damage recognition carried out by MutS homologues, either MutS $\alpha$  or MutS $\beta$ . The individual steps following the recruitment of MutL $\alpha$  (MutL homolog) are not as well understood in eukaryotes compared to bacteria. Eukaryotes do not express a direct homolog of the bacterial MutH nuclease, and it is thought that the latent endonuclease activity of MutL $\alpha$  is activated by proliferating cell nuclear antigen (PCNA) resulting in nicks in the DNA (Liu et al., 2017). After which exonuclease 1 (EXO1) is activated to excise the newly synthesised strand. Polymerase  $\delta$  then fills in the gap with the correct nucleotide in place.

### *Nucleotide Excision Repair*

Nucleotide excision repair (NER) removes bulky adducts, such as cyclobutane dimers (CPDs) and 6-4 pyrimidine-pyrimidine dimers (6-4PP) caused by ultraviolet light (UV; Koehler et al., 1991; Sancar 2016; Vaughn and Sancar 2020). NER operates through two mechanisms: transcription-coupled repair (TCR) and global repair. TCR depends on the

recognition of a stalled RNA polymerase at the damaged site on the transcribed strand, and repairs only that strand (Kusakabe et al., 2019; Vaughn and Sancar, 2020). In global repair, proteins xeroderma pigmentosum complementing group A and C (XPA and XPC) and RPA detect the damage and recruit the general transcription factor (TFIIH) protein complex (Vaughn and Sancar, 2020). The complex encircles the damaged DNA and excision repair cross-complementing endonucleases ERCC4 and ERCC5 cut out a 24-30 bp long oligodeoxynucleotide, which is degraded by nucleases (Wood, 1997; Kemp and Sancar, 2012; Schärer, 2013). The NER pathway is completed when DNA polymerases  $\kappa$  or  $\lambda$  (Lehmann, 2011) fill the gap and the repair patch is ligated by DNA ligase I or the x-ray cross-complementing 3-ligase3 complex (XRCC3-DNLI3; Sancar, 1996; Vaughn and Sancar, 2020). In TCR, either one of the two translocases excision repair chromatin remodelling factors 5 or 6 (ERCC8 or ERCC7) bind to the stalled RNA polymerase and facilitates its bypass of smaller DNA adducts. It may also alter the protein-protein interactions that enable the recruitment of the rest of the NER proteins to finish the repair process in the same manner as global repair (Xu et al., 2017; Vaughn and Sancar, 2020).

### 3.1.2 DNA Repair's Role in Adaptation and Evolution

While DNA damage can be deleterious, genome instability—as a result of reduced DNA repair efficiency or increased mutation rate—can result in mutations that favour survival in harsh conditions (Giraud et al., 2001). This process has been referred to as the mutation for survival hypothesis (Rosenberg, 1997). The generally accepted notion is that a cost-benefit ratio exists between 'permission and repair' of DNA damage: too much DNA repair may disable the capacity of organisms to alter their proteins and adapt to changing environments, while too many mutations are incompatible with normal life (Rosenberg, 1997). DNA repair also comes at a cost to the cell, requiring energy in the form of ATP. Cells under increasingly stressful conditions not only have to deal with a higher mutation load but also maintaining critical cell processes. If the cost of DNA repair increases linearly with mutation load the cost of repair can outweigh the potential risks and benefits from mutations (Nik-Zainal and Hall, 2019). This phenomenon has been assessed *in silico* by modelling systems capable of either unlimited or restricted DNA repair, particularly in cases where mutations are predominantly non-deleterious. Models which assume an unlimited

DNA repair capacity exhibited a significantly elevated rate of apoptosis, primarily attributed to the heightened energy expenditure associated with DNA repair. Conversely, models constrained by limited rates of DNA repair managed to endure, despite the inherent risk associated with potentially deleterious mutations (Breivik and Gustav, 2004).

DNA repair, specifically DSBR, can result in genome duplication events, loss of heterozygosity (LOH) and generation of genetic diversity driving genetic differentiation (Davila., et al., 2011; Carvajal-Garcia et al., 2023). Induction of DSBs in the model plant *Arabidopsis thaliana* resulted in multiple copy number variation (CNVs) and gene duplication events (Muramoto et al., 2018). The changes within the genome translated into phenotypic changes with plants exposed to DSBs having a significantly higher dry weight of stems, leaves and seeds. Polyploid plants displayed a notable increase in the occurrence of CNVs and genomic restructuring events. This can be attributed to the redundancy in the genome, which facilitates greater genome plasticity (Muramoto et al., 2018). Importantly, these plants were edited using a heat-induced endonuclease, allowing for control over when DSBs were induced. All DSBs were induced during vegetative growth and all data was collected before meiosis, meaning all events are a consequence of DNA repair during mitotic cell cycles.

In an evolution experiment, Byrne et al., 2014 exposed four separate populations of wild-type *E. coli* to 20 rounds of ionizing radiation (IR) treatments, which damages all cellular components including proteins and DNA (Byrne et al, 2014). The intensity of IR in each treatment (3000 Gy) was enough to kill ~99% of the population, with the remaining cells left to grow after each round. The genomes of isolates from the remaining four populations after 20 rounds of exposure were sequenced (Byrne et al., 2014). Mutations in evolved strains were re-introduced in turn into the founder strain to determine which mutations increased survival. Interestingly, single mutations in core DNA repair enzymes had the greatest positive impact on survival, specifically a mutation in recombination protein A (RECA) which was present in all but one of the evolved populations (Byrne et al., 2014). The mutations in the RECA gene altered the protein in a way that it created filaments quicker but extended them slower than the wild type, allowing for slower ATP use. This would aid the cell given the amount of DNA damage and strain on its metabolism (Byrne et al., 2014). When comparing the founder strains to bacteria species known to tolerate extreme radiation, *Deinococcus radiodurans*, there were no significant differences between

the repertoire of DNA repair genes. Bacteria species adapted to high IR exposure do not simply have more DNA repair proteins but have selected for more efficient DNA repair genes under IR exposure, giving evidence that DNA repair proteins can be modified to aid in the evolution under extreme stress (Byrne et al., 2014). It should be noted that adaptation to extreme IR is a complex mechanism and is not achieved by a single molecular event (Byrne et al., 2014). There are many other molecular changes occurring during the evolution of these strains that contributed to their ability to survive extreme IR, such as amelioration of protein oxidation which allowed proteins to be more readily available to repair damaged DNA (Daly et al., 2004; Daly, 2012).

These studies have shed light on the role DNA repair systems can have in adaption over relatively short timescales. Diatoms have evolved in a vast number of niches, both aquatic and freshwater (Benoiston et al., 2017), especially raphid pennate diatoms. The expansion of raphid pennate diatoms has generally been attributed to increased mobility due to their raphe, giving them access to more habitats, which has in turn accelerated their diversification (Nakov et al., 2018). However, as was noted in the study by Byrne et al., 2014, there were a multitude of molecular adaptations paving the way for increased survival under extreme stress, including modifications to DNA repair proteins. High genetic diversity has been revealed in model diatoms that have not been observed to undergo sexual reproduction suggesting other mechanisms outside classic meiosis to generate this diversity (Bulankova et al., 2021). With its core function to manipulate DNA, DNA damage repair (DDR) enzymes may have a larger role in generating genetic diversity than previously thought. This could particularly be the case in diatoms which undergo prolonged clonal growth during seasonal blooms (Armbrust, 2009), which is followed by irregular sexual reproduction (D'Alelio, et al., 2010). However, there is currently no literature reporting on the role of DNA repair systems in diatoms' adaptation to environmental stress and their role as a driver of genetic diversity.

### 3.1.3 Inducing DNA Damage to Study Organisms' Response

#### 3.1.3.1 Agents Used to Induce DNA Damage

##### 3.1.3.1.1 Zeocin

The antibiotic phleomycin D1 (zeocin) is a member of the bleomycin family derived from *Streptomyces verticillus*. Zeocin stocks are in an inactive copper-chelated form,  $\text{Cu}^{2+}$ ,

and once it enters the cell the copper is reduced to  $\text{Cu}^+$  and removed, activating zeocin. Once activated, zeocin intercalates into the DNA inducing double-strand breaks (DSBs) primarily through reactive oxygen species (ROS; Ehrenfel et al., 1987; Chankova et al., 2007). The DSBs are repaired through the HR and NHEJ pathways. Zeocin is a widely used selection agent causing cell death through DNA damage, specifically DNA DSBs (Pfeifer et al., 1997; Benko and Zhao, 2011; Van Blokland et al., 2011). Whole genome sequencing (WGS) of *S. cerevisiae* cultures exposed to zeocin revealed an increased ratio of T-to-G or A-to-C transversions (Zheng et al., 2022). Zeocin-induced point mutations tend to alter G sites in the 5'-TGTNC-3' motif on a genome wide scale in yeast (Zheng et al., 2022). However, none of these studies explicitly show DNA-DSBs in diatoms as a result of incubation with zeocin, though it has been used as a DNA damage model in *C. reinhardtii* (Chankova et al., 2007; Čížková et al., 2019).

#### 3.1.3.1.2 Methyl methanesulfonate (MMS)

Methyl methanesulfonate (MMS) is a carcinogenic alkylating agent which methylates DNA on  $N^7$ -deoxyguanosine and  $N^3$ -deoxyadenosine, causing stalled replication forks that indirectly cause DNA double strand breaks and bulky adducts (Lundin et al., 2005). The  $N^7$ -deoxyguanosine can be non-mutagenic, but the  $N^3$ -deoxyadenosine needs to be quickly repaired or the lesion will inhibit DNA synthesis and can be lethal (Beranek, 1990). Bypass repair at the replication fork, HR, and BER pathways are required in order for cells to tolerate MMS-induced damage (Xiao et al., 1996). MMS exposure increases in the number of T > A and T > C substitutions (Volkova et al., 2020). However, the mutational signatures of genomes exposed to MMS is a result of both the initial damage and the mechanism of repair. Different DNA repair deficient backgrounds alter the signature of mutation, for example, the rate of mutations on adenines was increased by 1.5x in NER deficient *C. elegans* (Volkova et al., 2020). MMS has been used as a model DNA damage system in many studies to research the effects of DNA repair on repair mechanisms because of its predictable mechanism of damage (Lundin et al., 2005). HR deficient cells are known to have increased sensitivity to MMS exposure (Zamborszky et al., 2017). MMS was used in addition to zeocin to study the DNA repair capacity of diatoms because of its well characterised mechanism of causing DNA damage, frequently used as a model DNA damage system, and it

causes a different type of DNA damage compared to zeocin to elucidate differences between diatom species' tolerance to two separate stressors and rate of use of different DNA repair pathways.

### 3.1.3.2 Dose Response Curves

Dose-response growth curves describe the extent of an organism's response to a gradient of either concentration or timed exposure to a stimulus or stressor (Crump et al., 1976). The organism's response is monitored and modelled according to the amount of external influence applied to determine the *in vivo* effect on the fitness of the organism. The external influence can be any factor that impacts the overall fitness of an organism. Dose-response curves are the primary assay to determine toxicity of chemicals on cell lines, primarily used for pharmaceutical purposes, especially in cancer therapies (Skipper et al., 1964). They are also useful in studying the response of the photosynthetic performance of autotrophic organisms, including phytoplankton, under variable light conditions (Yang et al., 2020). Applications of these assays to understanding the relationship toxic compounds have on phytoplankton communities is increasing due to aquatic environments experiencing increasing anthropogenic influence (Yi et al., 2019, Andersson et al., 2022). In this study dose-response curves were used to determine the half-maximal effective concentration, or  $EC_{50}$  which describes the concentration at which the chemical is 50% effective (Chen et al., 2013). The  $EC_{50}$  value of each chemical was used to describe the difference in tolerance between species and to standardize experiments across species.

### 3.1.3.3 Evidence of DNA Damage

The Terminal Deoxynucleotidyl Transferase dUTP nick end labelling (TUNEL) assay was used to confirm DNA damage in *in vivo*. The TUNEL assay was developed to detect apoptosis (Gold et al., 1993). Despite its original purpose of apoptosis detection, the TUNEL assay functions to label DNA damage with fluorescent reporters that can be detected via flow cytometry and microscopy. Therefore, this function was utilised to confirm the presence of DNA damage within cell cultures resulting from exposure to DNA damaging agents *in vivo*. The enzyme terminal-deoxy transferase (TdT) catalyses the attachment of a deoxynucleic base attached to a fluorescein isothiocyanate (FITC) fluorochrome to exposed OH groups on



the 3' ends of DNA. Samples can then be analysed through flow cytometry. The amount of DNA damage can be roughly calculated using the number of cells counted and the genome size.

## 3.2 Results

### 3.2.1 Global Review of DNA Repair Pathways in Diatoms

#### 3.2.1.1 *Genome Size Compared with Number of DNA Repair Genes*

Using the methods from Chapter 2 section 2.5.1, a reference list of 221 core and accessory DNA repair genes were queried against 47 reference proteomes from bacteria, archaea and eukaryotes. The total number of returned significant hits were plotted against genome size. No linear relationship was observed, however, there is a relationship with genome size on a logarithmic scale (Figure 3.4a). Unsurprisingly there is a clear split between prokaryotic and eukaryotic composition and the number of DDR genes, with eukaryotes requiring more specificity due to the size and complexity of their genomes (Figure 3.4b). Organism complexity appears to have an impact as well on the abundance of DDR core and accessory genes (Figure 3.4c), with low complex genomes (unicellular organisms) having less DDR-associated genes. It is known that higher eukaryotes have acquired DDR genes such as DNA-PK<sub>cs</sub> and created paralogues of RAD51 (Brandsma and Gent 2012). The number of DNA repair genes correlates more with the number of genes (0.64) and proteins (0.69) compared with genome size (0.43; Pearson correlation), suggesting that the number of DNA repair genes is not linearly correlated with genome size, but rather conserved based on functionality of the DNA repair systems. When just eukaryotic data is considered, there is non-significant clustering based on cellularity (Figure 3.4b). When the data is analysed using principal component analysis (PCA) considering the genome size, gene and protein count, and number of DNA repair genes, three clusters arise which cluster based on the taxonomic domain – eukaryote, bacteria, or archaea (Figure 3.5).

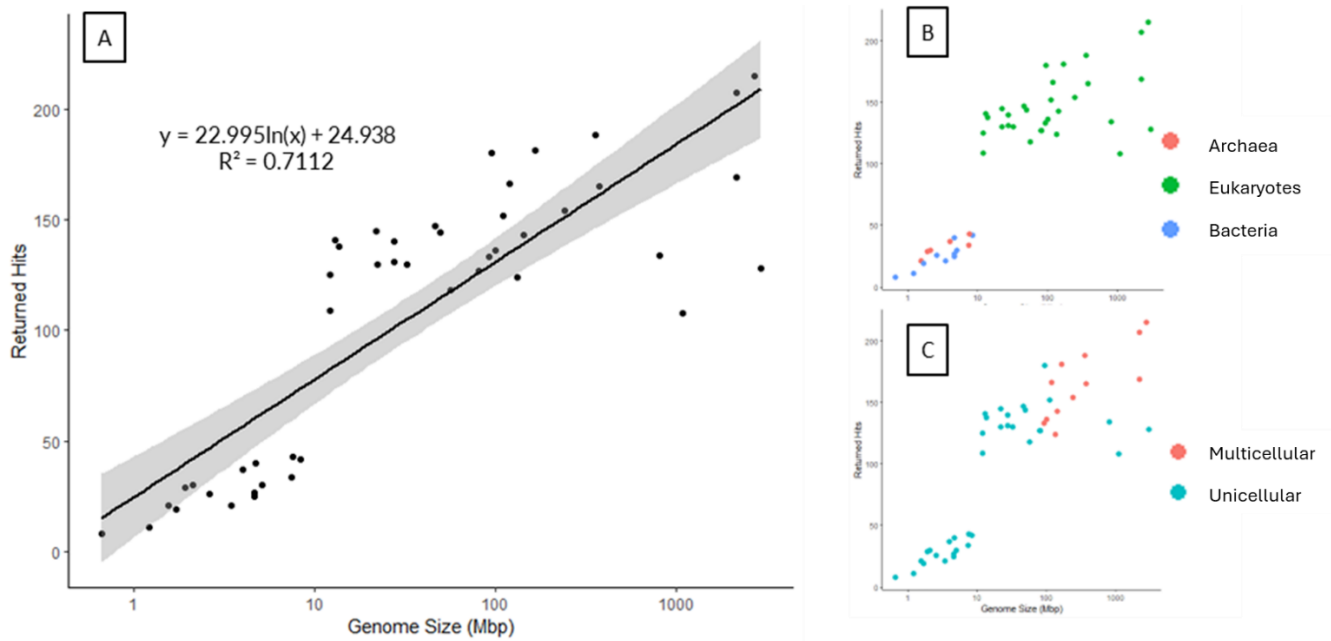
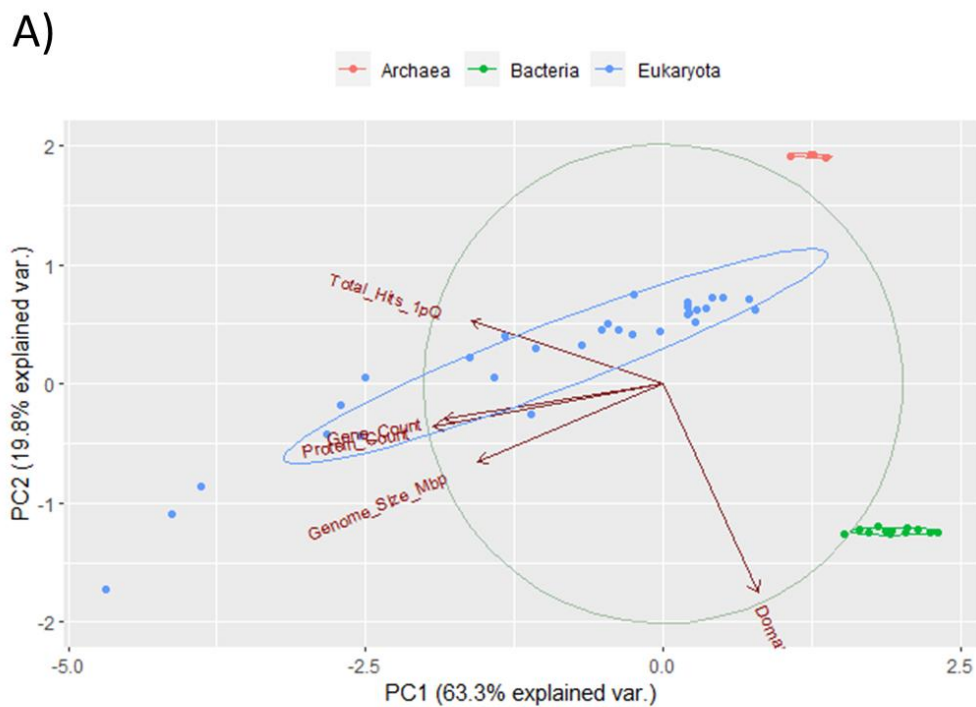


Figure 3.4. Genome size data was collected from the European Nucleotide Archive (ENA) for each of the 47 organisms used for this chapter. Returned hits represent the number of significant protein sequences showing significant similarity to the reference DDR gene set. A) Returned hits compared with genome size, logarithmic trendline was generated in R; B) Data coloured by the domain of life; C) Data coloured depending on if uni or multi-cellular. All scatter plots show the same data on the same scales.



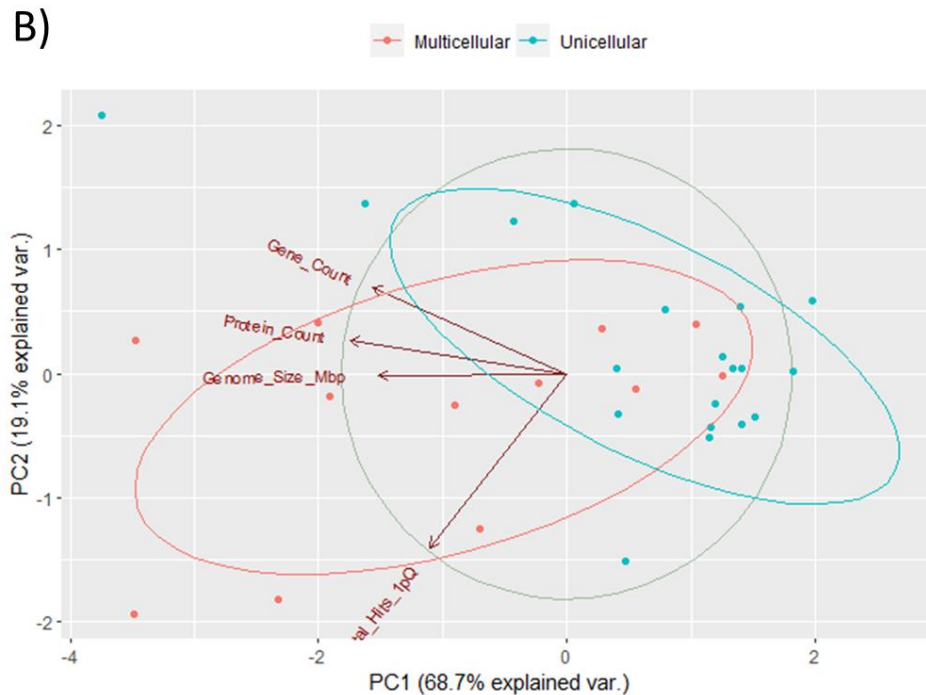


Figure 3.5. A) Principal component analysis (PCA) of genome size in relation to number of DNA repair genes returned from the reference set of 221 human DNA repair genes. Gene\_Count = number genes in genome, Protein\_Count = number of proteins in genome, Genome\_Size\_Mbp = genome size in millions of base pairs; Total\_Hits\_1pQ = number of significant ( $p < 0.001$ ) DNA repair proteins returned from HMMER 3.3 (Nov 2019); <http://hmmer.org/>, analysis. A) PCA plot of all 47 proteomes used, data clustered based on domain (Prokaryote, Eukaryote, and Archaea). B) PCA plot of only Eukaryote data, showing weak clustering based on if the organism is either multi or unicellular.

### 3.2.1.2 Double Strand Break Repair

#### 3.2.1.2.1 Homologous Recombination

A combination of data from OrthoFinder and individual searches confirmed that diatoms have the necessary genes to carry out HR (Figure 3.6). However, some proteins were not predicted to be present in select diatom proteomes. The single-strand DNA-binding protein (SSBP) and DNA topoisomerase 2-binding protein 1 (TOPB1) were not found in the reference proteomes for *F. solaris*, *T. oceanica*, or *P. multistriata*

TOPB1, in humans, consists of repeating BRCA1 C terminus (BRCT) domains which are found in a variety of proteins with diverse functions in DNA repair. BRCT domains are known to be involved in protein-protein interactions during DNA repair making them widespread in DDR genes (Glover and Williams, 2004). Given the versatility and expansion of the BRCT domain, the predicted absence of TOPB1 in *F. solaris*, *P. multistriata* or *T. oceanica* does not confirm that the function of TOPB1's has been lost. The function may be carried

out by proteins containing the required conserved domains, but not predicted due to their low sequence similarity. The same result is achieved when using a seed sequence from evolutionarily closer taxa, such as *Chlamydomonas reinhardtii* where TOPB1 has been annotated in the genome.

Similarly, all proteomes and available diatom genomes have SSPB except for *P. multistriata*. However, unlike RFA2 or RFA4, when building a HMM model this protein was not found even when using parameters that assume distant homologies. SSPB is a small protein (232 amino acids in *F. solaris* [A0A1Z5KE78\_FISSO]) containing only one domain, the single-stranded binding protein family domain (SSB; pfam00436). When using the curated HMM model of this domain to search available diatom proteomes and genomes, significant hits are returned for all diatoms besides *P. multistriata*. Given this protein is a highly conserved single copy gene core to DNA repair throughout eukaryotes including plants and green algae, it is more likely this protein exists in the *P. multistriata* genome and is not found because of availability of genetic information.

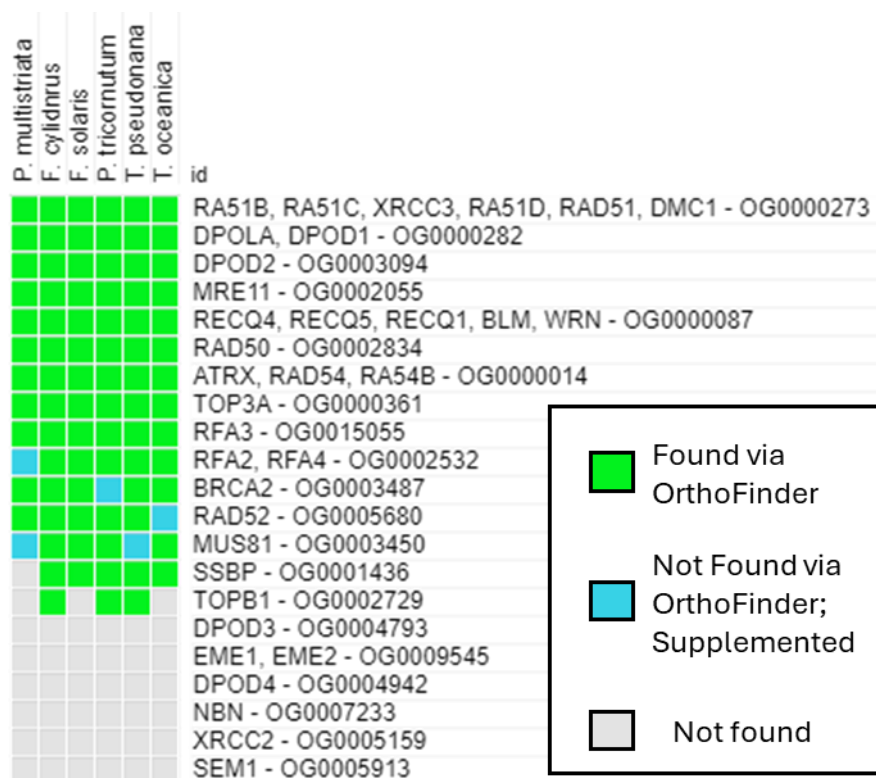


Figure 3.6. Heat map showing presence/absence of core DNA repair proteins active in the HR pathway. Each row represents an orthogroup created by OrthoFinder based on sequence similarity, and columns are by diatom

species. Genes within orthogroups are from the reference set of DNA repair proteins detailed in Chapter 2 section 2.5.1. Cells are coloured on presence or absence of significant ( $p < 0.001$ ) hits from blast searches and HMM model searches. Green indicates proteins identified by OrthoFinder, blue cells (Supplemented) indicate data that was not recognised by OrthoFinder but was detected by hmsearch through HMMER 3.3 (Nov 2019); <http://hmmer.org/>) and grey cells represent no hits found.

### 3.2.1.2.2 Non-Homologous End-Joining

A combination of data from OrthoFinder and individual searches confirmed that diatoms have the necessary genes to carry out non-homologous end-joining (NHEJ; Figure 3.7). Only one orthogroup was not present in all proteomes, OG001245, which consists of four x-family polymerase proteins: DNA nucleotidylexotransferase (TdT), polymerase mu (Pol  $\mu$ ), polymerase lambda (Pol  $\lambda$ ) and polymerase beta (Pol  $\beta$ ; figure 3.7).

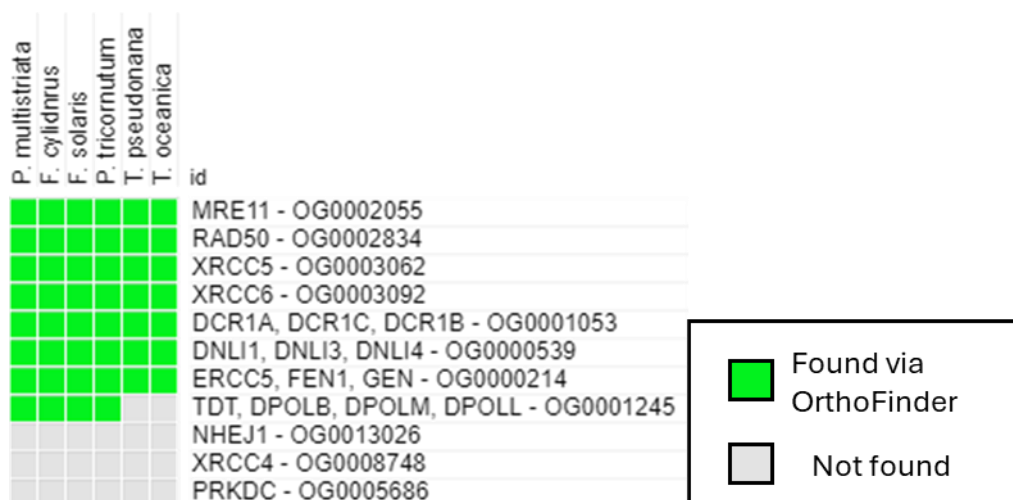


Figure 3.7; Heat map showing presence/absence of core DNA repair proteins active in the NHEJ pathway. Each row represents an orthogroup created by OrthoFinder based on sequence similarity, and columns are by diatom species. Genes within orthogroups are from the reference set of DNA repair proteins detailed in Chapter 2 section 2.5.1. Cells are coloured on presence or absence of significant ( $p < 0.001$ ) hits from blast searches and HMM model searches. Green indicates proteins identified by OrthoFinder, and grey cells represent no hits found.

X family polymerases (Pol) are a group of specialized DNA polymerases that play essential functions in DNA repair and recombination in eukaryotes (Prostova et al., 2022). They are all thought to have evolved from a common Pol  $\lambda$ -like gene involved in NHEJ, with

the three other family members arising via gene duplication (Uchiyama et al., 2009). They participate in resynthesis of missing or damaged nucleotides during the NHEJ pathway of DNA DSB repair (Yamtich and Sweasy, 2010). Vertebrates are the only eukaryotes to have all four X-family polymerases, whereas fungi, plants and eukaryotic microorganisms only have one or two (Uchiyama et al., 2009).

The orthogroup (#OG0001245) which contains the X-family polymerases: polymerase mu, polymerase lambda, polymerase beta and DNA nucleotidylexotransferase (Pol  $\mu$ , Pol  $\lambda$ , Pol  $\beta$  and TdT), was only reported to have been found in pennate diatoms (*P. tricornutum*, *P. multistriata*, *F. cylindrus* and *F. solaris*) (Figure 3.7). Despite significant ( $p < 0.001$ ) hits returned for all three polymerases, only Pol  $\lambda$  was found. The hits to Pol  $\mu$ , Pol  $\beta$ , and TdT overlap with the Pol  $\lambda$  sequence given their close similarity. With Pol  $\lambda$  having the highest similarity, and also being the only X-family polymerase found in plants and green algae (Uchiyama et al., 2004; figure 3.8). The only centric diatom found to possess a Pol  $\lambda$  homolog was *Chaetoceros tenuissimus*.

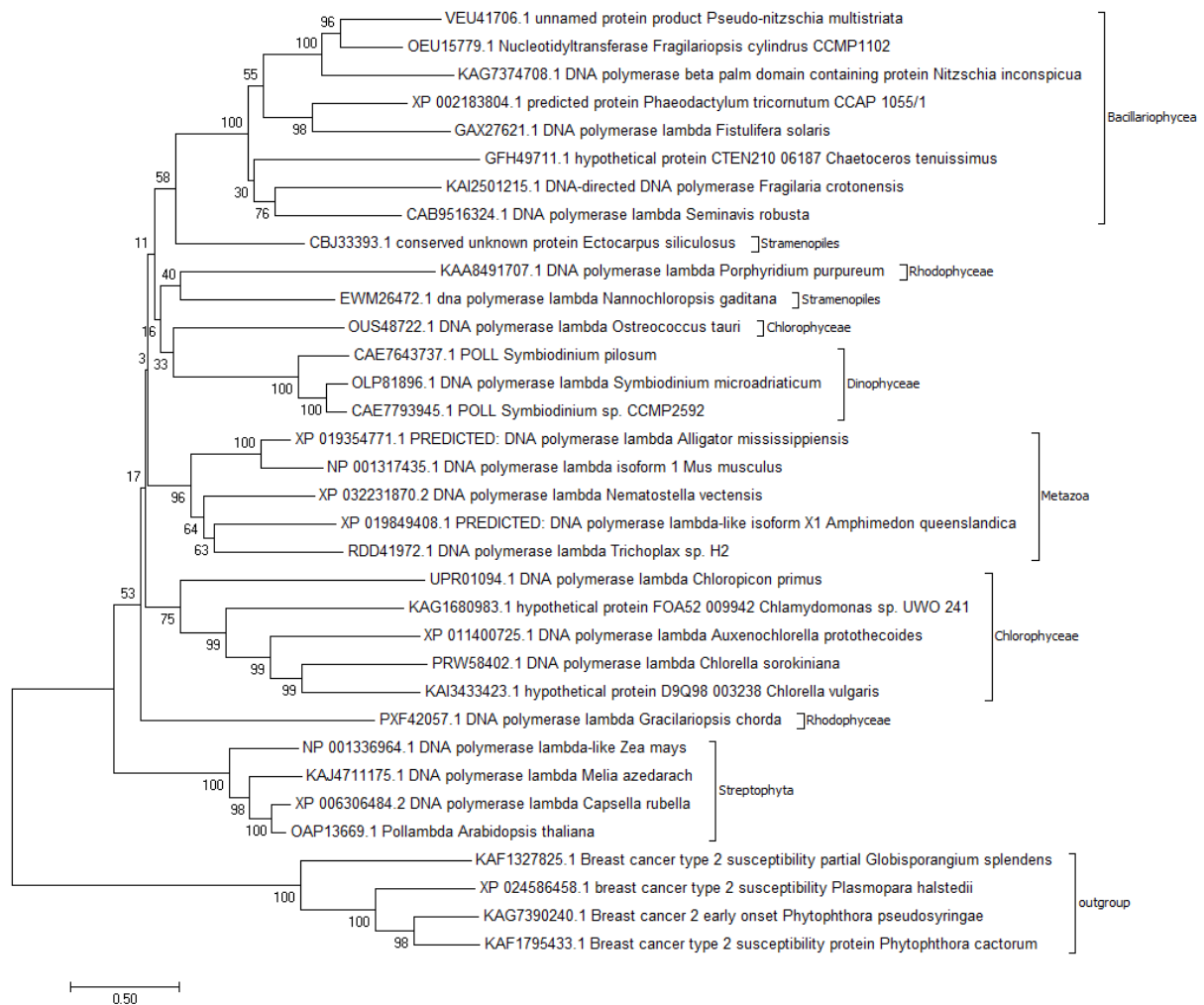


Figure 3.8. The evolutionary history of Pol  $\lambda$  was inferred using the Neighbour-Joining method (Saitou and Nei, 1987). The optimal tree with the sum of branch length = 24.56 is shown. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) are shown next to the branches (Felsenstein, 1985). The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Poisson correction method (Zuckerkand and Pauling, 1965) and are in the units of the number of amino acid substitutions per site. The rate variation among sites was modelled with a gamma distribution (shape parameter = 2). The analysis involved 34 amino acid sequences. All ambiguous positions were removed for each sequence pair. There were a total of 2368 positions in the final dataset. Evolutionary analyses were conducted in MEGA7 (Kumar et al., 2015).

The lack of X-family polymerases has also been described in *Drosophila melanogaster* (Bienstock et al., 2014) and *Caenorhabditis elegans* (Asagoshi et al., 2011; Uchiyama et al., 2009). In *C. elegans*, Asagoshi et al., 2011 discovered that polymerase theta (pol  $\theta$ ) contributed significantly to the gap-filling step in the BER pathway, providing evidence that X-family polymerases are not critical to BER function. Why centrics in the Thalassiosirales order do not have X-family polymerases compared to pennates is paradoxical, as pennate lineage diverged from centrics around 70 million years ago

(Benoiston et al., 2017). But the presence of a pol  $\lambda$  homologue in the Chaetocerotales order suggests this protein was lost in within the diatom order of Thalassiosirales. To confirm the absence of x-family polymerases from centric genomes, searches were expanded to encompass all diatom genomes available, with no genetic data derived from centrals showing evidence of x-family polymerases. Searches were conducted by downloading curated hmm profiles for the conserved domains unique to each x-family polymerase and blasted via HMMER 3.3 (Nov 2019; <http://hmmer.org/>).

One gene missing from all diatom proteomes was PRKDC (OG0005686), which encodes the DNA-dependent protein kinase catalytic subunit (DNA-PKcs). DNA-PKcs is found across all major eukaryotic phyla and is a core part of the NHEJ pathway (Figure 3.3). In addition to its role in DSB repair, DNA-PKcs is involved in other cellular functions as a checkpoint protein including metabolism regulation, apoptosis, telomere protection and maintaining genomic stability (Kumar, 2023). Despite the importance of DNA-PKcs's role in NHEJ and its wide conservation across eukaryotes, it is not found in multiple lineages. For example, DNA-PKcs has not been identified in angiosperms or in red algae (Qiu et al., 2013; Kumar, 2023). DNA-PKcs has been predicted to be present in the stramenopile *Ectocarpus siliculosus* (Lees-Miller et al., 2021). However, even when using the DNA-PKcs sequence from *E. siliculosus* as the query in this thesis, a full DNA-PKcs sequence could not be identified in diatoms. Only two of the 6 conserved domains within DNA-PKcs could be identified in diatoms using hmmsearch; the P13\_P14 kinase and FATC domains. These alone do not provide enough evidence to support the presence of a functional DNA-PKcs homologue. Lees-Miller et al., suggest that DNA-PKcs has been lost in groups of organisms that inhabit stochastic environments due to its role in DNA repair and maintenance and genomic stability. Its loss may promote decreased genomic stability and contribute to an increased ability to adapt to harsh environmental conditions (Lees-Miller et al., 2021). Whether DNA-PKcs is present or absent in diatoms will require a more in-depth analysis. If it was missing this could be evidence of the evolution of DNA repair systems as a mechanism of adaptation.



### 3.2.1.3 Single-strand Break Repair, Mismatch Repair, and Repair of DNA Adducts Crosslinks and Oxidized Bases

#### 3.2.1.3.1 Base Excision Repair

Data collected through OrthoFinder and supplemented with data from HMMER hmmsearch (<http://hmmer.org/>; Figure 3.9) show that diatoms possess the core proteins necessary for the BER pathway. The only exception is that several DNA glycosylases were not found in the initial searches and X-family polymerases were not found within centric diatom genomes.

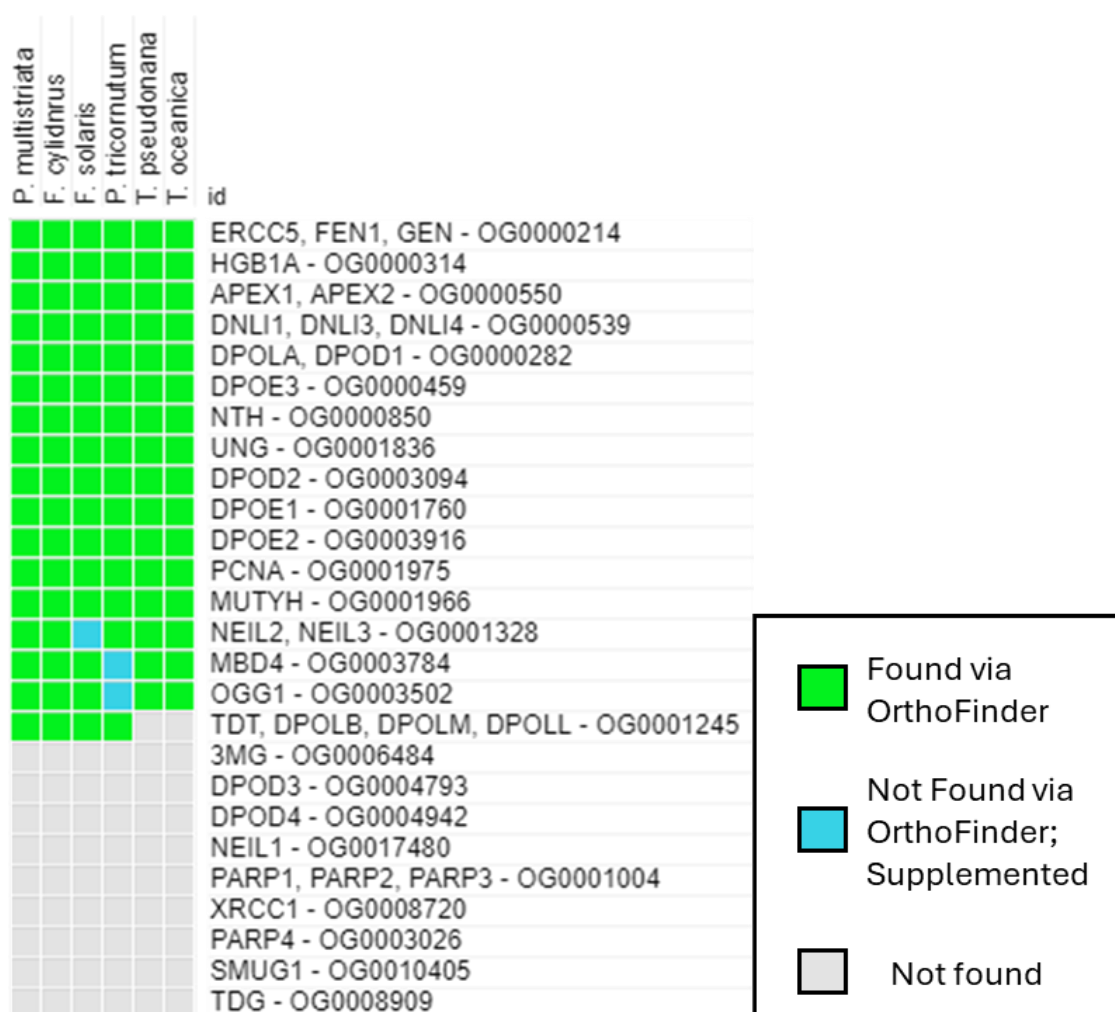


Figure 3.9. Heat map showing presence/absence of core DNA repair proteins active in the BER pathway. Each row represents an orthogroup created by OrthoFinder based on sequence similarity, and columns are by diatom species. Genes within orthogroups are from the reference set of DNA repair proteins detailed in Chapter 2 section 2.5.1. Cells are coloured on presence or absence of significant ( $p < 0.001$ ) hits from blast searches and HMM model searches. Green indicates proteins identified by OrthoFinder, blue cells (Supplemented) indicate data that was not recognised by OrthoFinder but was detected using hmmsearch, through HMMER (Nov 2019; <http://hmmer.org/>) and grey cells represent no hits found.

The orthogroups OG0003502 and OG0003784 which contain the proteins Methyl-CpG Binding Domain 4, DNA Glycosylase (MBD4) and 8-Oxoguanine DNA Glycosylase (OGG1) respectively are present in all diatom proteomes except for *P. tricornutum*. Both proteins are glycosylases responsible for excision of damaged bases (Jacobs and Schär, 2012). To confirm if these proteins do or do not exist in the *P. tricornutum* genome, a hidden Markov model profile (hmm profile) of the amino acid sequences returned for both MBD4 and OGG1 was created through HMMER3.3 (Nov 2019; <http://hmmer.org/>) using the hmmbuild tool. The hmm profiles were then queried back against the *P. tricornutum* proteome.

For MBD4, the only protein sequence with a significant e value (<0.001) was Phatr2\_43040, however it only showed 25% similarity over 25% of the hmm model and was missing the core domain containing a hallmark helix-hairpin-helix (HhH) and Gly/Pro rich loop (GPD) – HhH-GPD (pfam00730) which is part of a superfamily involved in the BER pathway (Burner et al., 2000). To find out if this domain is absent or has been moved to another protein, the domain, pfam00730, was searched against *P. tricornutum*. The domain was found in three proteins, one of which also had the MUTY glycosylase superfamily domain at the C-terminal, indicating glycosylase activity, characteristic of the function of MBD4. This suggests that *P. tricornutum* has retained the core function of MBD4, but the sequence of the protein responsible has differentiated to the point it is not considered a significant match when using blast or HMMER 3.3 (Nov 2019; <http://hmmer.org/>).

The same results were found for OGG1 when searching via blast and hmm profiles built from diatom sequences for OGG1. The primary conserved domain, 8-oxoguanine DNA glycosylase, N-terminal domain (OGG\_N; pfam07934) which is responsible for the removal of 8-oxoguanine residues on bases as a result of oxidative damage (Burner et al., 2000; Bjørås et al., 2002). Interestingly there is no evidence of the presence of this domain within the *P. tricornutum* proteome or genome, when searching through HMMER 3.3 (Nov 2019; <http://hmmer.org/>), NCBI blastp or the pfam database. However, if the protein sequence of OGG1 from any of the other 5 diatoms is blasted directly against the *P. tricornutum* genome (<https://mycocosm.jgi.doe.gov/Phatr2/Phatr2.home.html>) the protein Phatr2\_53988 is returned as a significant hit with ~65% identity over ~60% of the query sequence. But this

protein is not in the 'filtered' genes list and it appears this has been left out of the reference proteome downloaded from UniProt (<https://www.uniprot.org/proteomes/UP000000759>). This may explain why it was not found by the initial searches.

*F. solaris* was the only proteome reported to not have a protein sequence of close enough similarity to either Endonuclease VIII-like 2 or 3 (NEIL2 and NEIL3 respectively; OG0001328). These proteins, like MBD4 and OGG1, are glycosylases, responsible for removing damaged bases before repair and ligation. The core formamidopyrimidine-DNA glycosylase H2TH domain (pfam06831) has been moved to a protein with an uracil glycosylase domain (UDG; pfam03167) which facilitates removal of uracil bases from single and double-stranded DNA.

This domain shuffling appears to be conserved within subset of diatom species consisting of *F. solaris*, *Nitzchia inconspicua*, *Seminavis robusta*, *Mayamaea pseudoterrestris*, *Chaetoceros tenuissimus*, *P. multistriata* and *T. oceanica* (Figure 3.10).



Figure 3.10. Alignment of all sequences showing significant sequence similarity to *F. solaris* Uracil-DNA glycosylase (AOA12ZJPI9\_FISSO). The top 8 lines consist of the diatom species described in the text above containing both the Formamidopyrimidine-DNA glycosylase N-terminal domain (pfam 01149; show in first 2 panels) in conjunction with the Uracil DNA glycosylase superfamily domain (UNG; pfam 03167; shown in last two panels). The remaining 16 protein sequences do not have a protein sequence containing both domains, but rather they are on separate loci.

It is not surprising that all glycosylases were found in each diatom proteome despite not all having significant sequence similarity over the full sequence length. DNA glycosylases are of bacterial origin and are thought to be some of the first DNA repair proteins to have evolved (Prorok et al., 2021). The lack of significant hits over the full sequence is potentially a result of domain shuffling within and onto other protein sequences as this has been seen in divergent organisms and the reference set is from humans and plants (Aravind et al., 1999).

Like PKRDC in the HR pathway, the poly (ADP-Ribose) polymerase (PARP) was not found in reference diatom proteomes through OrtoFinder or HMMER. PARP family

proteomes is unexpected. PARP is a family of proteins which are part of DNA repair systems as well as the maintenance of genomic stability (Herceg and Wang, 2001). Its main role is to detect ssDNA breaks and initiate single-strand break repair pathways, such as BER. Previous studies have investigated the conservation of PARP proteins across eukaryotes and have not been able to identify PARP proteins in diatoms (Citarelli et al., 2010). However, another study claimed that the diatom *T. oceanica* does have two PARP family proteins, while also predicting that PARP family proteins are missing in red algae and a subset of green algae (Perina et al., 2014). Diatoms do possess the conserved poly (ADP-ribose) polymerase and DNA-ligase Zn-finger region (PF00645) that functions to detect DNA nicks, however this is not found in conjunction with the highly conserved PARP catalytic domains like model PARP genes. One of the PARP catalytic domains (PF00644) was only found in *F. cylindrus* and *T. oceanica* through HMMER (hmmsearch) using HMM models but was at a locus with no other domains. Again, the evidence here does not support the presence or absence of PARP family proteins within diatoms as it is also likely that these proteins are not appearing due to errors during genome assembly or annotation. However, like DNA-PKcs, the potential loss of PARP family proteins in diatoms provides an interesting area for debate and further research into the mechanisms of DNA repair in diatoms.

#### 3.2.1.3.2 Mismatch Repair

Searches for mismatch repair homologues via HMMER (Nov 2019; <http://hmmer.org/>) and blast searches show that diatoms share a common repertoire of genes which contains all the necessary genes for the pathway (Figure 3.11). Like other pathways, any gene missing from all diatom proteomes are only found in higher multicellular eukaryotes. Given the list of DNA repair genes was derived from humans this is not unexpected. The only proteins that could not be found via searches were single-strand binding protein (SSBP) and RFA2 or RFA4 in the *P. multistriata* genome.

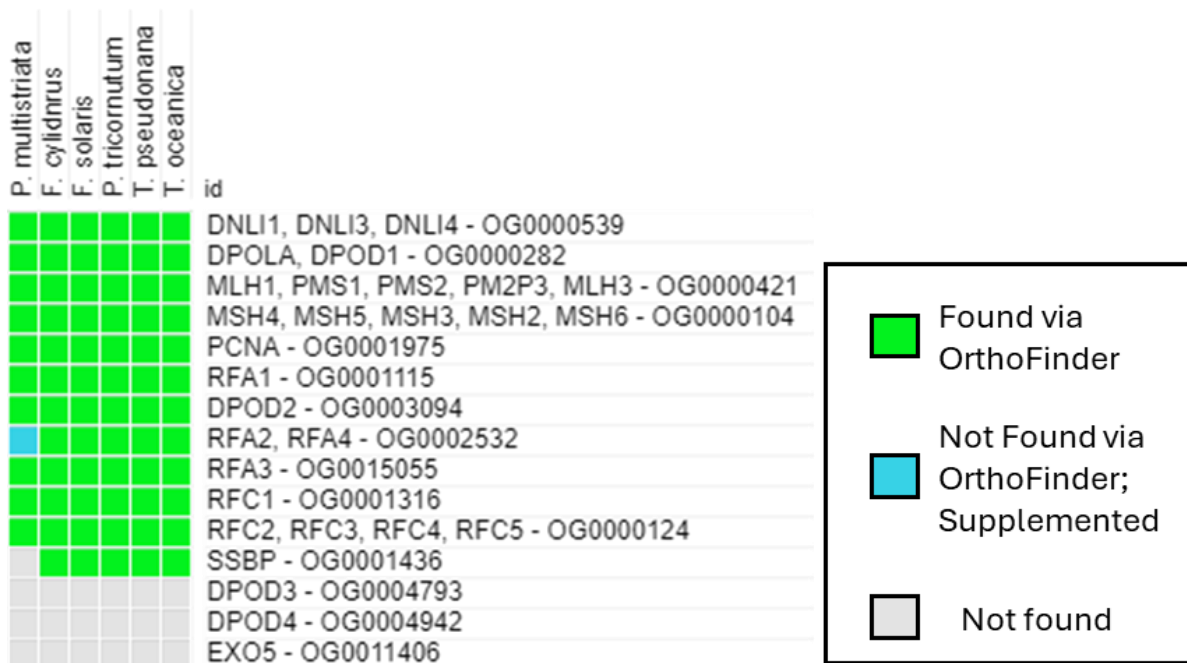


Figure 3.11. Heat map showing presence/absence of core DNA repair proteins active in the MMR pathway. Each row represents an orthogroup created by OrthoFinder based on sequence similarity, and columns are by diatom species. Genes within orthogroups are from the reference set of DNA repair proteins detailed in Chapter 2 section 2.5.1. Cells are coloured on presence or absence of significant ( $p < 0.001$ ) hits from blast searches and HMM model searches. Green indicates proteins identified by OrthoFinder, blue cells (Supplemented) indicate data that was not recognised by OrthoFinder but was detected using *hmmsearch* through HMMER (Nov 2019; <http://hmmer.org/>) and grey cells represent no hits found.

### 3.2.1.3.3 Nucleotide Excision Repair

In diatoms, all core proteins necessary to complete nucleotide excision repair (NER) are present (Figure 3.12). Genes that were not found evolved in higher eukaryotes through gene duplication events and are not found in lower eukaryotes taxa groups. The only exceptions are two subunits of the general transcription factor IIH (TFIIH), TFIIH2 and TFIIH5, of the TFIIH protein complex which creates a bubble around the damaged DNA and recruits nucleases to the site. No sequences of significant ( $e$  value  $< 0.01$ ) similarity were found in either *T. oceanica* and *P. multistriata* for both subunits, and no sequences were found for TFIIH5 in *T. pseudonana* or *F. solaris*. The complex is made up of two major sub-complexes: the core general transcription factor II (TFHII) and the CAK (Cdk activation kinase). TFHII consists of excision repair cross-complementing helicases ERCC2 and ERCC3 in combination with subunits TFIIH1, TFIIH2, TFIIH3, TFIIH4 and TFIIH5. CAK is made up of cyclin H (CCNH),



cyclin dependent kinase 7 (CDK7) and the CDK-activating kinase assembly factor (MAT1; Bedez et al., 2013). It is unlikely the missing subunits (TFIIH2 & 5) are completely missing from the genomes as all the other components to create the complex are found and are of critical importance in DNA repair and transcription. All subunits of this complex have been identified in plants, fungi, animals and protists suggesting that the subunits evolved early in the eukaryotic lineage and are in fact part of all diatom genomes. Potentially these are not found in a few of these proteomes due to poor annotation/sequencing given the amount of repeats in diatom genomes and lack of reference genomes for comparative genomics, or they have diverged significantly from known canonical isoforms. Further research would be needed to determine the reason they are not present.

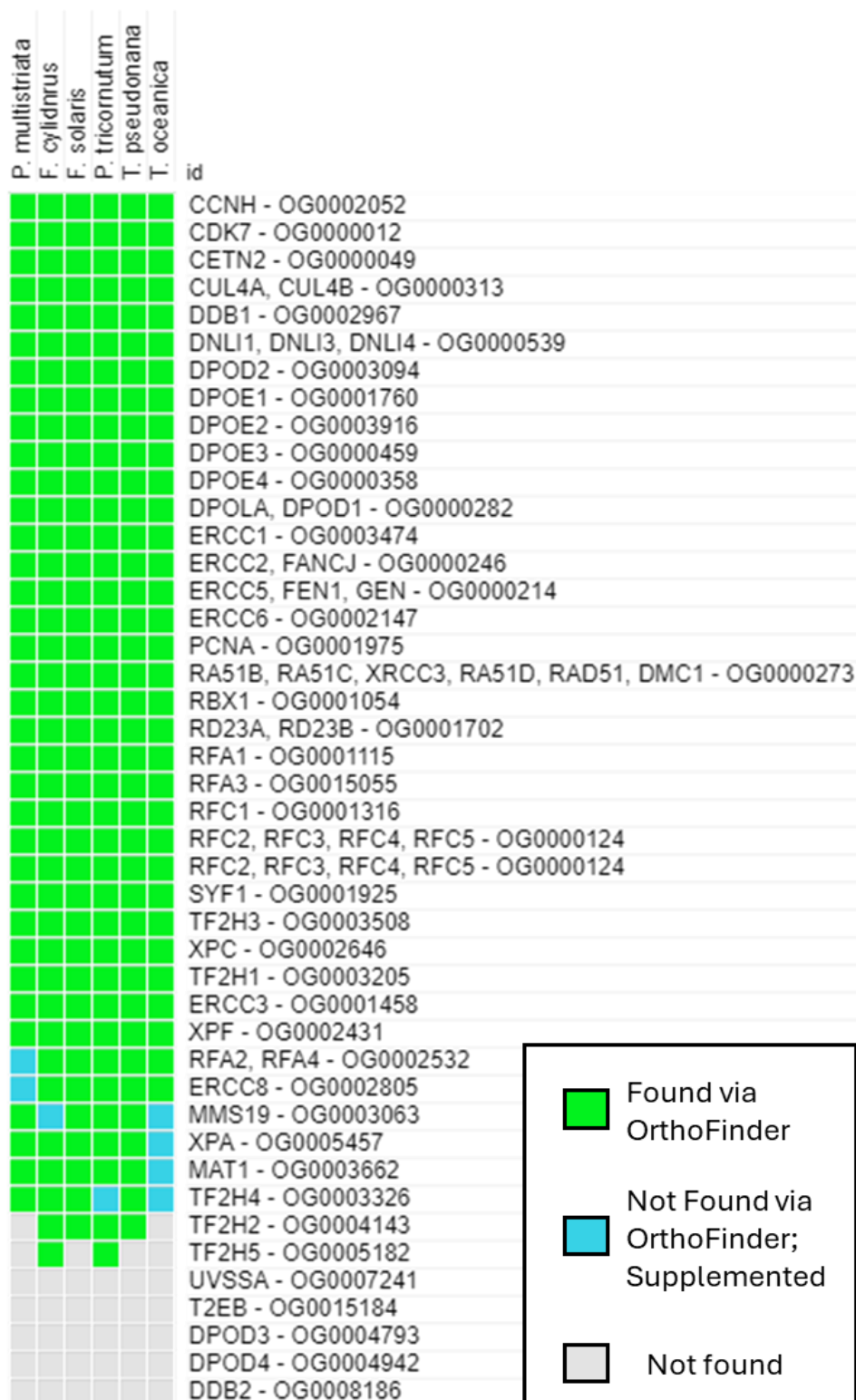


Figure 3.12; Heat map showing presence/absence of core DNA repair proteins active in the NER pathway. Each row represents an orthogroup created by OrthoFinder based on sequence similarity, and columns are by diatom species. Genes within orthogroups are from the reference set of DNA repair proteins detailed in Chapter 2 section 2.5.1. Cells are coloured on presence or absence of significant ( $p < 0.001$ ) hits from blast searches and HMM model searches. Green cells indicate proteins identified by OrthoFinder, blue cells (Supplemented) indicate data that was not recognised by OrthoFinder but was detected by HMMER (Nov 2019; <http://hmmer.org/>) hmmsearch, and grey cells represent no hits found.

### 3.2.2 Dose Response Curves

#### 3.2.2.1 Evidence of Induced DNA Damage

The Terminal deoxynucleotidyl transferase dUTP nick end labelling (TUNEL) assay was used to confirm DNA-DSBs were being created as a result of exposure to zeocin. The TUNEL assay was developed to detect apoptosis (Gold et al., 1993) by using the enzyme terminal-deoxy transferase (TdT) to catalyse the attachment of a deoxynucleic base attached to a fluorescein isothiocyanate (FITC) fluorochrome to exposed OH groups on the 3' ends of DNA. Samples can then be analysed through fluorescence-activated cell sorting (FACS) using a flow cytometer. The assay was conducted as stated in Chapter 2 (section 2.2), and samples were analysed through flow cytometry at the John Innes Centre (JIC). Samples were diluted in sterile synthetic ocean water (SOW; chapter 2 section 2.1.2) at a 1:10 ratio before being passed through the BD FACS Melody flow cytometer. First gates were set up using forward and side scatter metrics to ensure only the population of *T. pseudonana* cells were being analysed. Cultures were made axenic before FACS analysis; however, the flow cytometer was not enclosed in a safety cabinet, therefore, cultures were primarily screened for contaminating organisms. Gating was done to ensure only single events were analysed. Once the primary gates were set up, the flow cytometer FITC laser was turned on. 100,000 events were recorded resulting in two populations, TUNEL positive and negative (Figure 3.13). Negative control samples, incubated with fluorochrome reporters but not TdT, were also analysed through the same process (Figure 3.13).

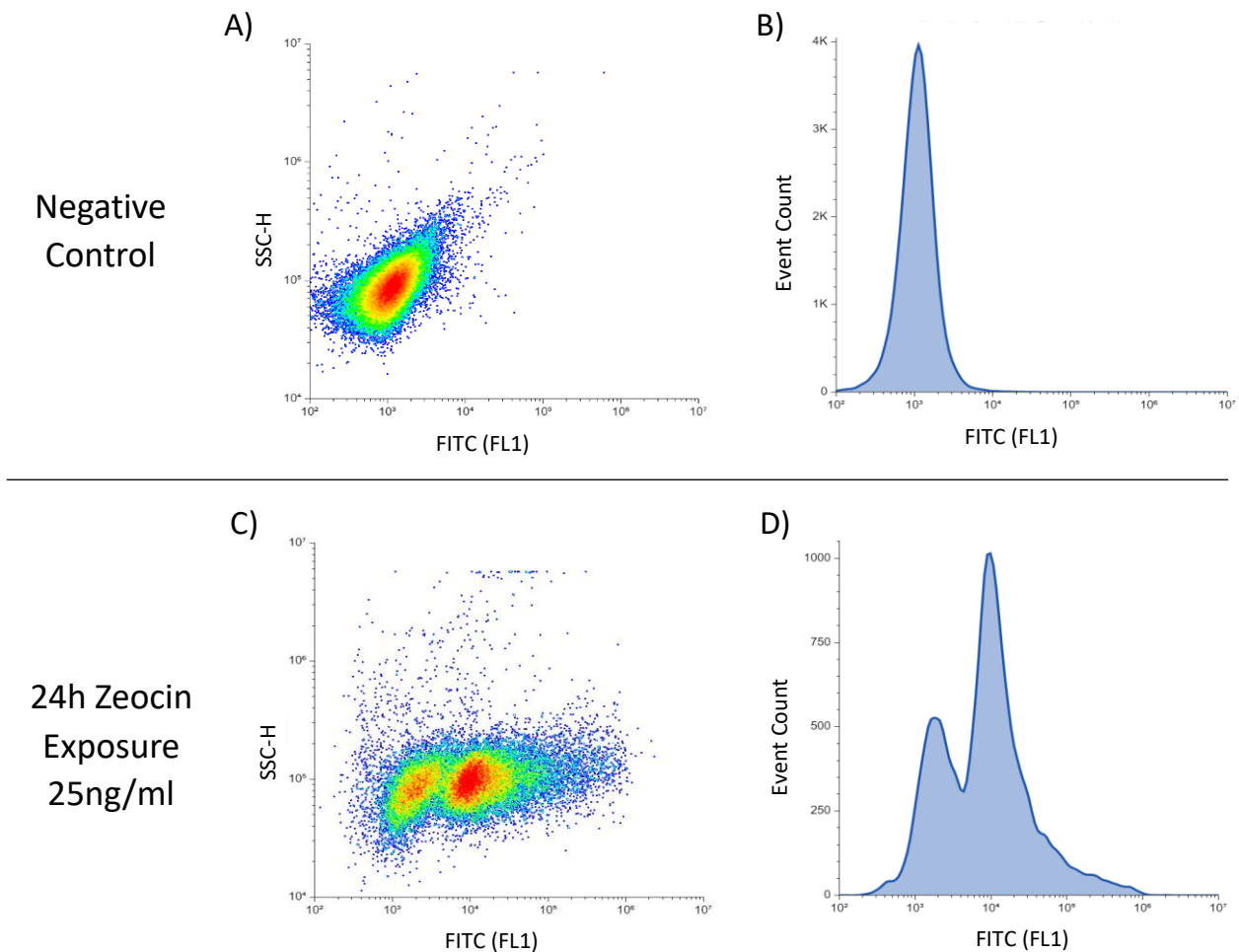


Figure 3.13. Density plots (A and C) and histograms (B and D) showing the distribution of 5 thousand events recorded from the control sample not exposed to zeocin (top panel) and samples exposed to 25ng/ml of Zeocin (bottom panel) for 24 hours. SSC-H = side scatter height; FITC(FL1) = fluorescein isothiocyanate intensity per event; Event Count = number of recorded events for a given fluorescence intensity (FITC [FL1]). Samples were analysed using a BD FACS Melody.

Interestingly, the percentage of TUNEL positive cells did not significantly increase when exposed to either 25ng/ml or 150ng/ml of zeocin for 24 hours. 24 hours of incubation was used as any time less was not enough for the EC<sub>50</sub> concentration (25ng/ml) of zeocin to be fully effective *in vivo*, but incubation over 24 hours stopped cell growth, as determined by a recovery assay (Figure 3.14). The TUNEL assay provides evidence that the loss in fitness when exposed to zeocin, was in fact due to DNA damage in the form of DSBs.

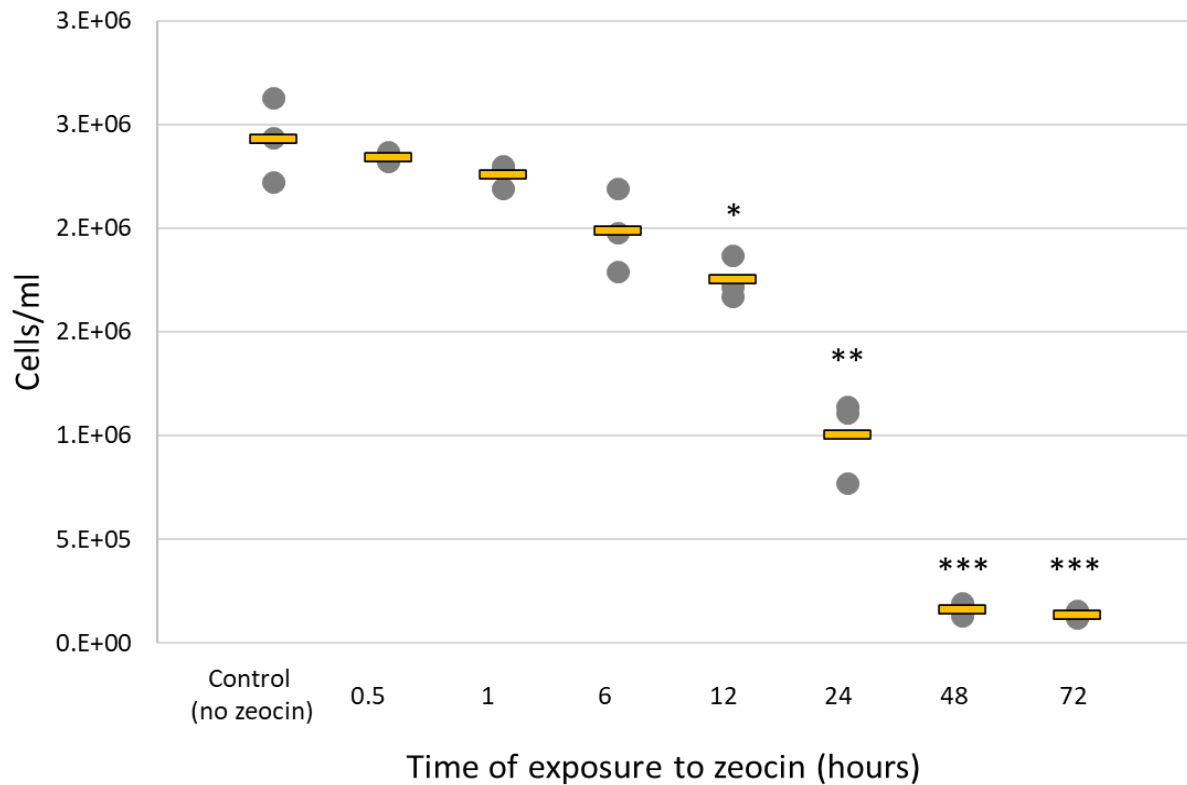


Figure 3.14: Cells/ml of cultures at 72 hours of growth after incubation with zeocin from 0.5 hours to 72 hours.  $N=3$ . The x-axis describes how long cultures were exposed to 25 ng/ml zeocin before being washed and transferred to replete media not containing zeocin. Grey dots indicate data from each biological replicate, yellow bars indicate mean, \* indicate significance compared with control (\* =  $p < 0.05$ ; \*\* =  $p < 0.005$ ; \*\*\* =  $p < 0.0005$ ).  $P$  values are results of 2-tailed t-test assuming equal variances.

### 3.2.2.2 Zeocin Dose Response Curves

Using the methods stated in Chapter 2 section 2.1.2.3, dose response curves were first conducted to find the  $EC_{50}$  concentration of zeocin for *T. pseudonana*, *P. tricornutum* and *F. cylindrus*. The first range of zeocin concentrations tested were selected based on other similar experiments with eukaryotic cultures in the literature. However, under this range of concentrations all *T. pseudonana* cultures exposed to zeocin did not grow and went extinct within three days (Figure 3.15).

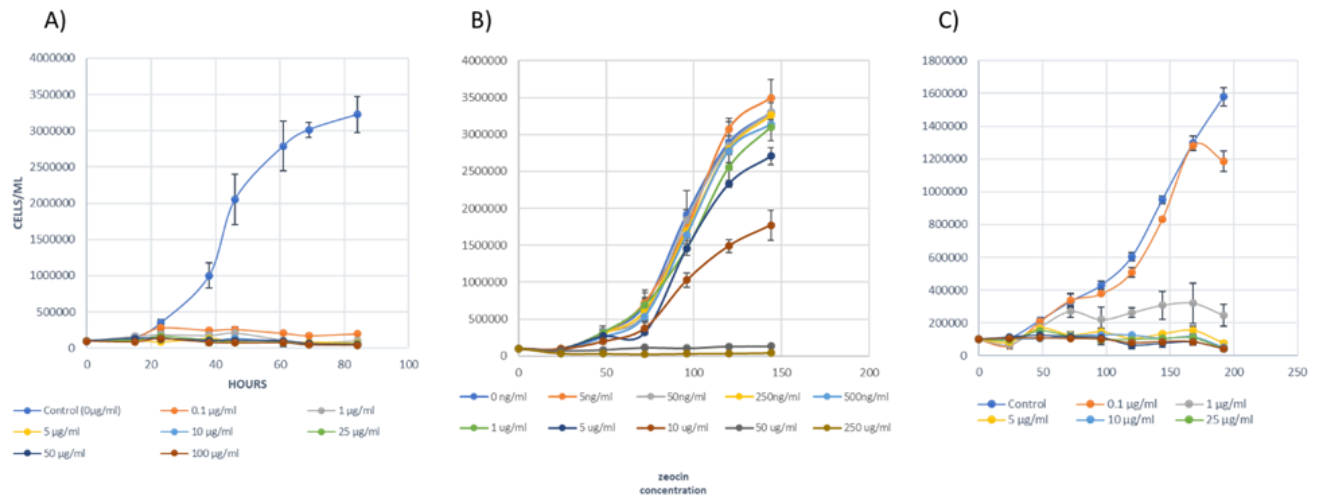


Figure 3.15. Growth curves of A) *T. pseudonana*; B) *P. tricornutum* and C) *F. cylindrus* under initial concentrations of zeocin in Aquil. Graphs display cells/ml counted using microscopy over time (hours). Zeocin concentrations are between 0 ng/ml (control) and 500 ng/ml given below each graph.

For *T. pseudonana* the range of concentrations was reduced from micrograms to nanograms per-millilitre as there was no growth observed in the first growth curves (Figure 3.15a). Cultures were grown in Aquil media containing zeocin (0 – 150ng/ml) with Fv/Fm, cells/ml counts and mean cell size measurements also taken (Figure 3.16). *T. pseudonana* cultures exposed to relatively high concentrations of zeocin showed a consistent phenotype of increased cell size (Figure 3.16b). Given that zeocin causes DNA damage, this increase in cell size was attributed to cell cycle arrest due to mutation load causing activation of cell cycle checkpoint proteins which halt replication. This was also observed in *Chlamydomonas reinhardtii* under DNA damage stress caused by zeocin. Čížková et al., 2019, reported that the DNA damage response was connected to systems blocking the cell cycle. They were able to reverse this effect with the addition of caffeine, which is known to allow cells to bypass cell cycle checkpoints, to cultures under zeocin stress. As expected, cultures were able to divide, but their viability significantly decreased due to mutation load (Čížková et al., 2019).



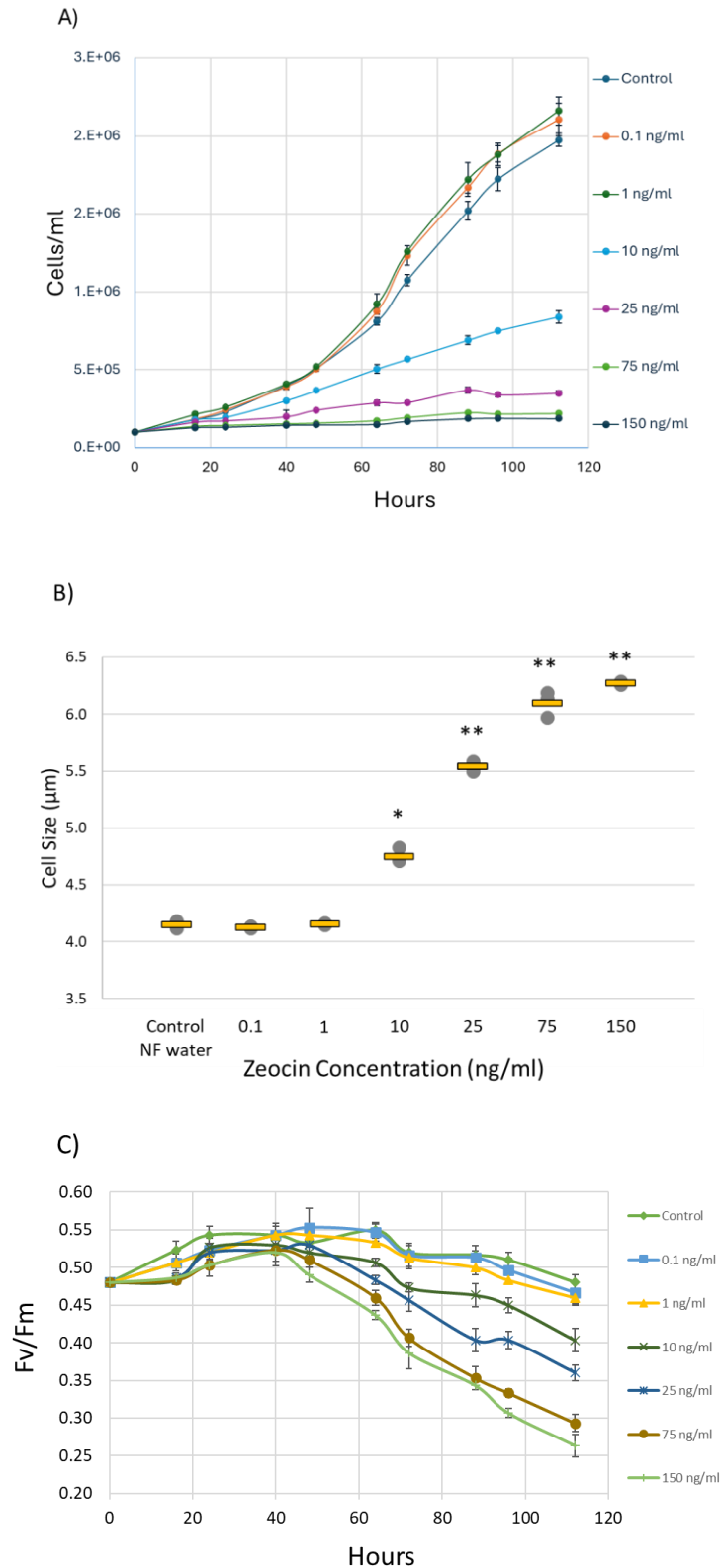


Figure 3.16. Dose response curve of *T. pseudonana* exposed to a range of zeocin concentrations. Error bars show standard deviation. For all figures  $n=3$ . A) Cells/ml; B) mean cell size, grey dots are data points for each biological replicate and yellow bars indicate the mean, \* indicate significant difference compared to control cultures (\* =  $p<0.0005$ , \*\* =  $p<0.00005$ ); and C) Fv/Fm. Zeocin was diluted in nuclease free water (NF-water) before addition to media, so NF-water was added to controls to compensate for any bias created by the drug vehicle. P values are result of a two-tailed t-test assuming equal variance.

With appropriate concentration ranges for each species, the difference between these three species' tolerance to zeocin is best shown in a dose response curve (Figure 3.17). Under zeocin stress there was a significant difference in the sensitivity of *T. pseudonana* when compared with *P. tricornutum* and *F. cylindrus* (Figure 3.17). Table 3.1 shows the EC<sub>50</sub> value of each species calculated by the R package *drda* (Malyutina et al., 2023).

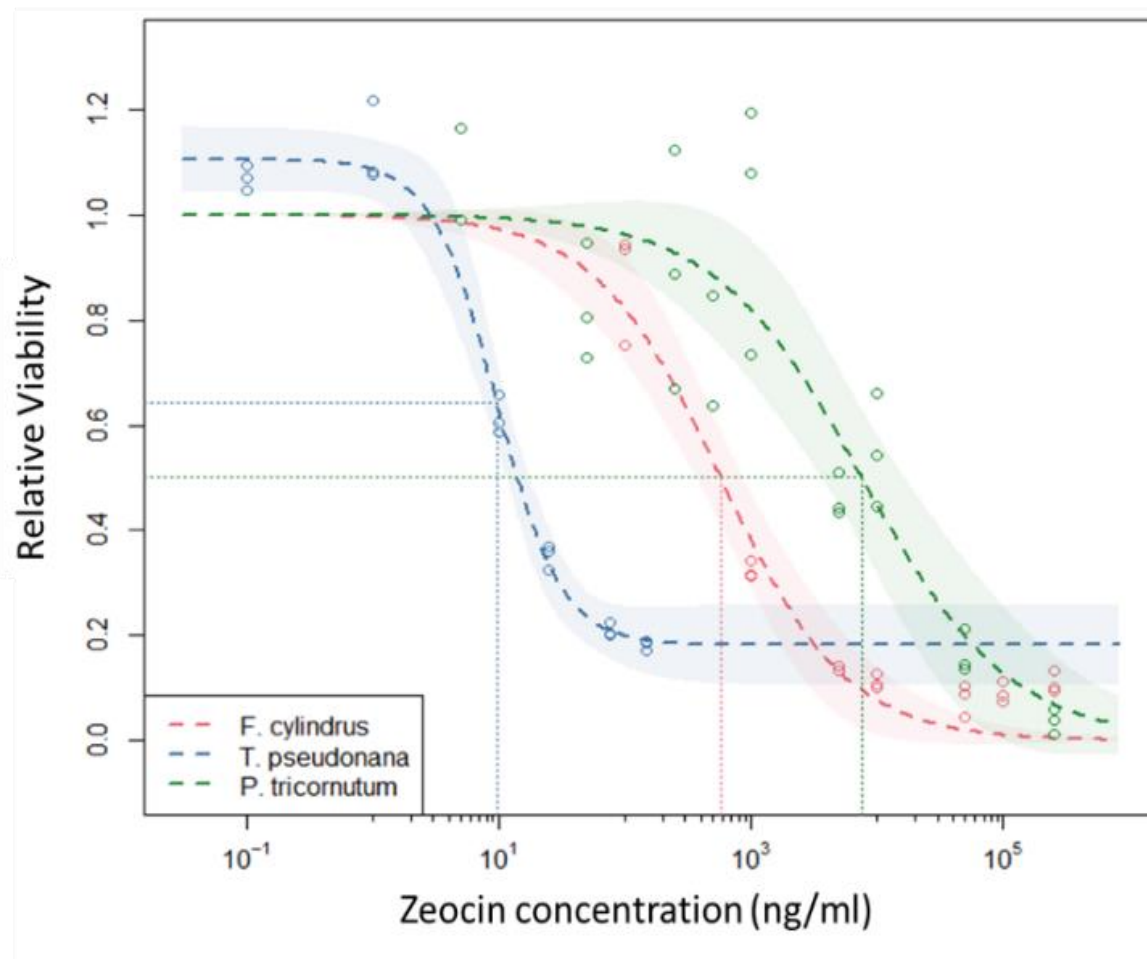


Figure 3.17. Dose response curves created through the R package *drda* for three diatom species under zeocin stress. Each curve shows the relevant viability of each species during exponential phase compared to control cultures. All samples are in triplicates. Shaded areas show 95% confidence interval. Vertical and Horizontal lines intersecting each curve show the  $EC_{50}$  value for each species. X axis is on a logarithmic scale.

Table 3.1: Half Maximal Lethal Concentration ( $EC_{50}$ ) values for *P. tricornutum*, *F. cylindrus* and *T. pseudonana* exposed zeocin.

Species	EC <sub>50</sub> (µg/ml)	Lower 95%	Upper 95%
<i>P. tricornutum</i>	7.61	4.40	13.16
<i>F. cylindrus</i>	0.58	0.41	0.79
<i>T. pseudonana</i>	0.01	0.001	0.01

### 3.2.2.3 Methyl methanesulfonate Dose Response Curves

Unlike when exposed to zeocin, all three species were able to tolerate similar concentrations of methyl methanesulfonate (MMS; Figure 3.18a), with no significant differences in their respective EC<sub>50</sub> values (Table 3.2). *T. pseudonana* again acquired an elongated cell shape with increasing concentration of MMS, suggesting cell cycle arrest due to a high DNA damage load (Figure 3.18b). However, the increase in cell size was not as significant as observed with zeocin.

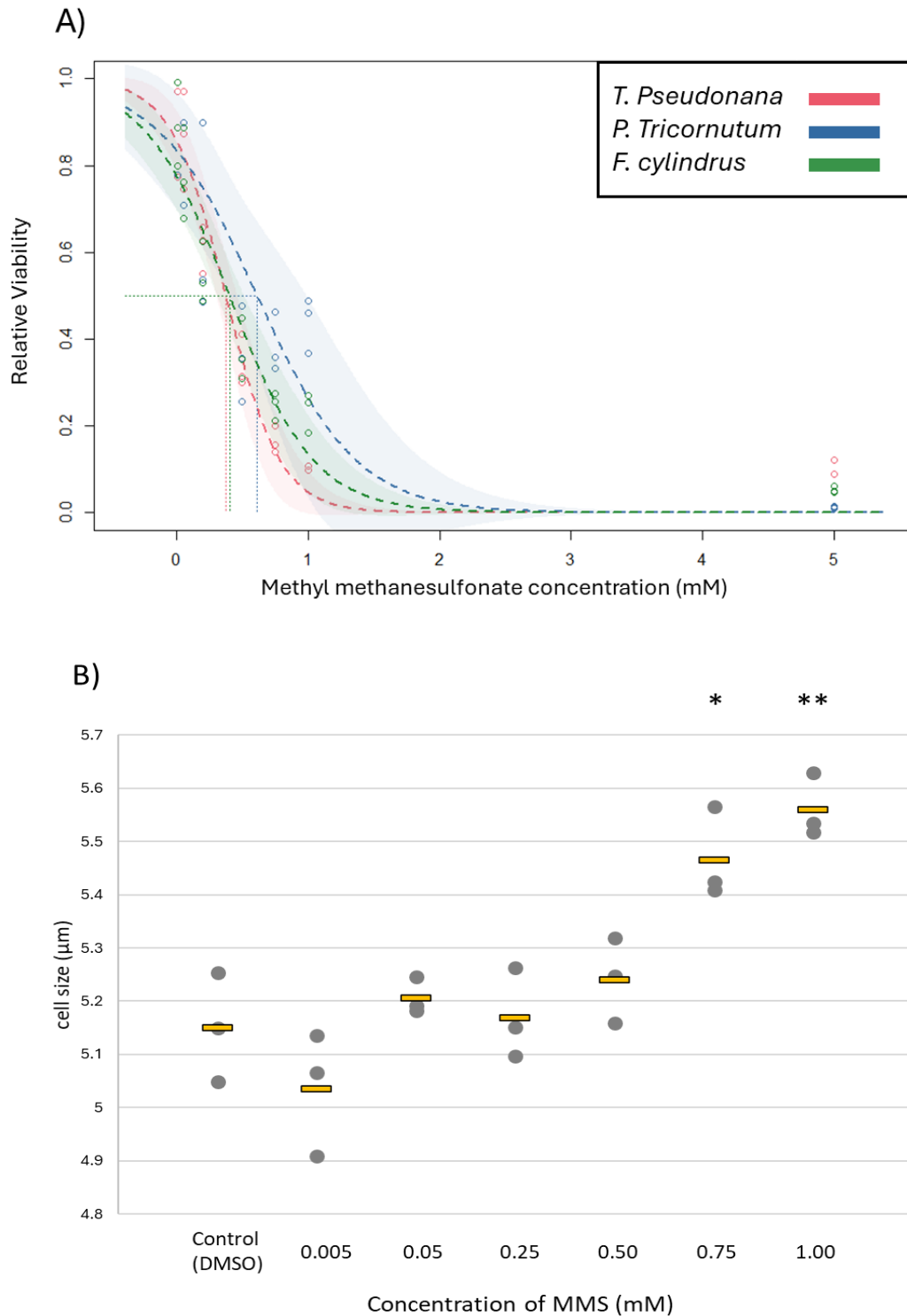


Figure 3.18. A) Dose response curves created through the R package *drda* for three diatom species exposed to methyl methanesulfonate (MMS). Each curve shows the relevant viability of each set of replicates (3) during exponential phase compared to control cultures grown under normal growth conditions. All samples are in triplicates. Shaded areas show 95% confidence interval. Vertical and horizontal lines intersecting each curve show the EC50 value for each species. B) Cell size of exponentially growing wild type *T. pseudonana* cultures exposed to increasing MMS concentrations. Control cultures were grown in the drug vehicle (DMSO) to compensate for any effect this may have on fitness. Asterisks (\*) show significant difference compared to size of control cultures ( $p < 0.05 = *$ ;  $p < 0.005 = **$ ). Grey dots are data points for each biological replicate and yellow bars indicate the mean.

Table 3.2. Half Maximal Lethal Concentration ( $EC_{50}$ ) for *P. tricornutum*, *F. cylindrus* and *T. pseudonana* exposed to MMS.

Species	$EC_{50}$ (mM)	Lower 95%	Upper 95%
<i>P. tricornutum</i>	0.61	0.40	0.82
<i>F. cylindrus</i>	0.40	0.32	0.48
<i>T. pseudonana</i>	0.37	0.31	0.44

### 3.2.3 Selection of Genes to Target to Impair Double-Strand Break Repair

#### 3.2.3.1 *Brca2* – Homologous Recombination

The BRCA2 DNA repair associated gene (*Brca2*) was selected as a target to impair the HR pathway through CRISPR/Cas in *T. pseudonana* for several reasons: A) it is present in all diatom proteomes (Figure 3.19), B) it has high conservation of core domains (Appendix Figure A.1), C) it is well conserved across eukaryotes (Figure 3.20), and D) as its role is core to the three sub pathways (Figure 3.2). *Brca2* is a DNA repair protein of eukaryotic origin and plays the major role of catalysing the removal and replacement of RPA with Rad51. The *T. pseudonana Brca2* homologue (TpBRCA2) consists of three main conserved domains, the BRCA2 helical domain which binds SEM1, the BRCA2-2\_OB1 fold which facilitates ssDNA annealing, and BRCA2-2\_OB2 fold which also binds ssDNA (Figure 2.3a in Chapter 2 section 2.3.1).

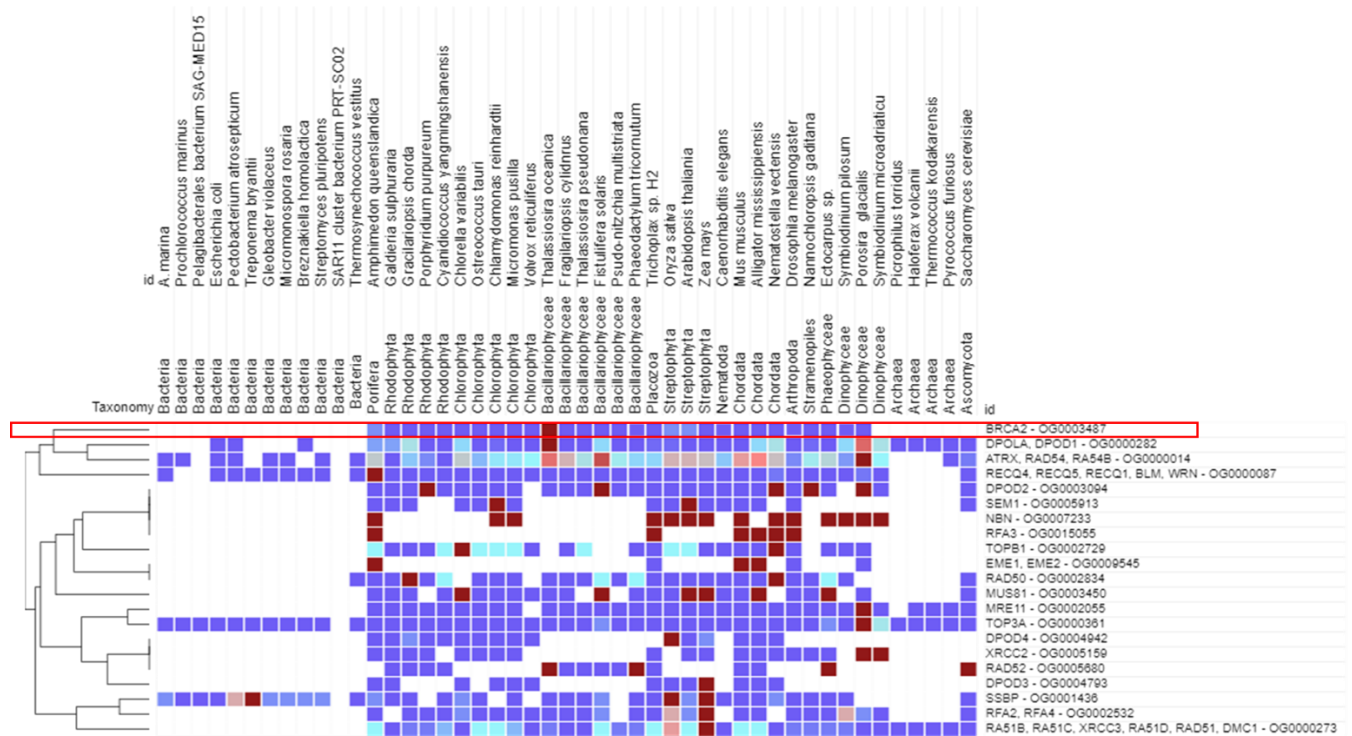


Figure 3.19. Heat map showing results of OrthoFinder search for HR related DNA repair proteins split into orthogroup families based on sequence similarity across 47 species. Cells are shaded in darker shades of red for a higher number of hits retrieved by OrthoFinder, white cells represent no hits present. The orthogroup containing *Bra2* is highlighted in red.

The sequence of TpBRCA2 was downloaded from the current genome available through the Joint Genome Institute (JGI; Armbrust et al., 2004). The protein sequence of the human BRCA2 gene (9606 NCBI) was used as a seed sequence for initial searches. Initially all gene models were searched but only the filtered gene models were considered and the gene THAPS\_263089 was determined to be a homolog of BRAC2 after reciprocal blast searches against the UniProt Reference Proteomes and subsequent alignments showed conservation of the core domains: BRCA2 helical domain (PF09169), BRCA2\_2\_OB1 fold (PF09103), BRCA2\_2\_OB2 fold (IPR048262; Appendix Figure A.1.).

There are other genes/proteins which satisfy these criteria and have been knocked out in other organisms to inhibit HR. Such as RAD51. As previously described, the function of RAD51 is to facilitate the search for a homologous locus and form a duplex initiating templated-directed repair (Krejci et al., 2012). Throughout evolution, there have been duplications of the Rad51 gene resulting in paralogues (Sullivan and Bernstein, 2018). Though their individual functions are not completely understood in every organism, they are

involved in DNA repair and cell cycle processes. For this reason, Rad51 was not chosen given the potential of paralogues and disruption of other functions within DNA repair and cell cycle regulation.

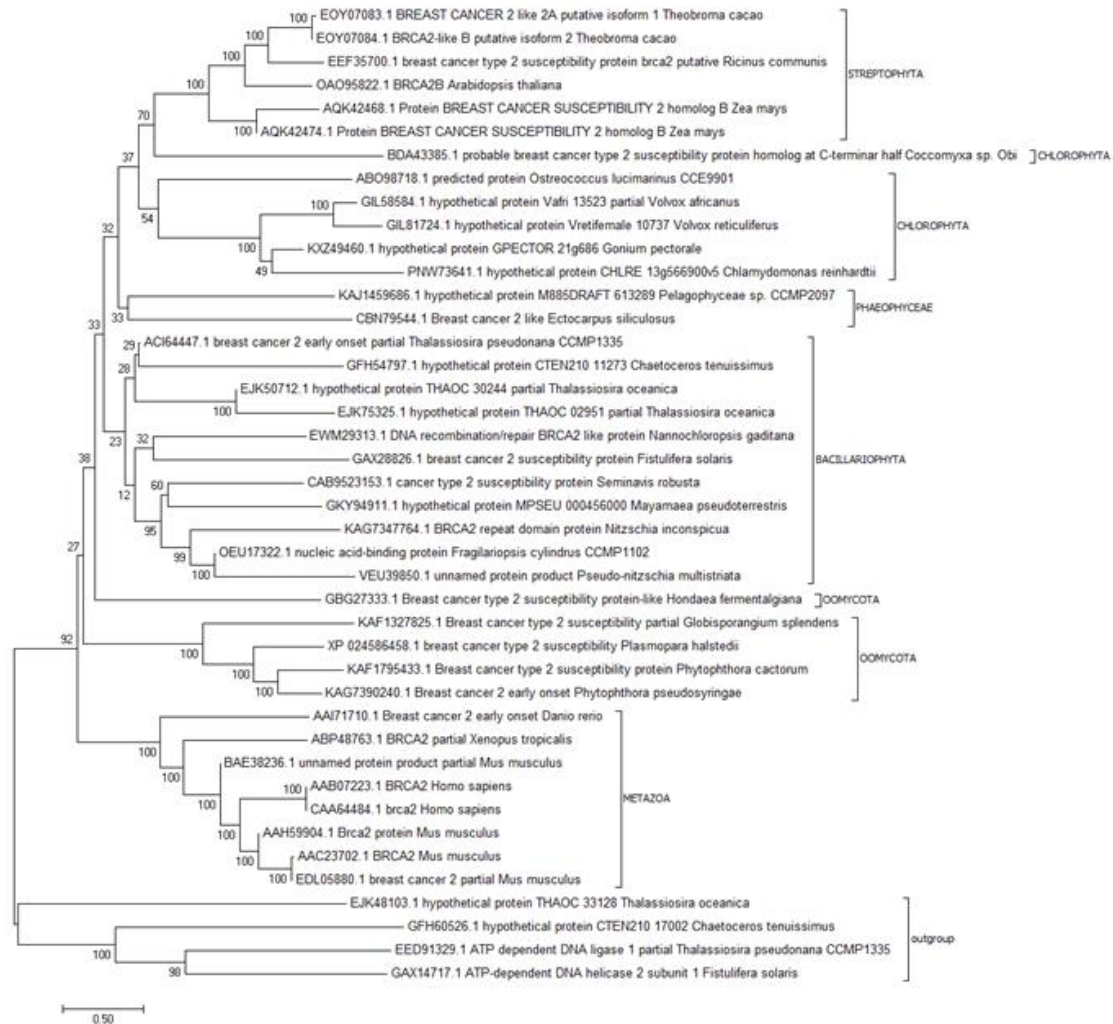


Figure 3.20. Evolutionary relationships of taxa for *Brca2*. The evolutionary history was inferred using the Neighbour-Joining method (Saitou and Nei, 1987). The optimal tree with the sum of branch length = 36.10762604 is shown. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) are shown next to the branches (Felsenstein, 1985). The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Poisson correction method (Zuckerkand and Pauling, 1965) and are in the units of the number of amino acid substitutions per site. The rate variation among sites was modelled with a gamma distribution (shape parameter = 2). The analysis involved 42 amino acid sequences. All ambiguous positions were removed for each sequence pair. There was a total of 4715 positions in the final dataset. Evolutionary analyses were conducted in MEGA7 (Kumar et al., 2016).

### 3.2.3.2 *Ku70* – Non-Homologous End-Joining

KU70 was chosen for reverse genetics via CRISPR/Cas to impair NHEJ for the same reasons as already mentioned for *Brca2* to impair HR. Unlike *Brca2*, the Ku-core domain of



*Ku70* is of prokaryotic origin and functional homologs are ubiquitous throughout all major taxa, excluding viruses (Aravind et al., 1999; Aravind and Koonin, 2001). *Ku70* was chosen as a target for several reasons; A) its conservation across all major taxa groups excluding viruses, B) it is well studied in model systems regarding DNA repair including when function is disrupted, and C) its role in the primary stage of the NHEJ pathway.

The KEGG database has automatically annotated the gene Thaps3\_263010 as the KU70 homologue in *T. pseudonana* and has been annotated as the DNA-binding subunit of a DNA-dependent protein kinase (*Ku70* autoantigen) via KOGG (KOG2327). The protein sequence of Thaps2\_263010 was downloaded and was blasted against the UniProt reference proteomes using phmmer (HMMER 3.3; Nov 2019; <http://hmmer.org/>). Significant hits from other organisms ( $p < 0.001$ ) were downloaded and aligned to confirm confirmation of known conserved domains (Figure 3.21).

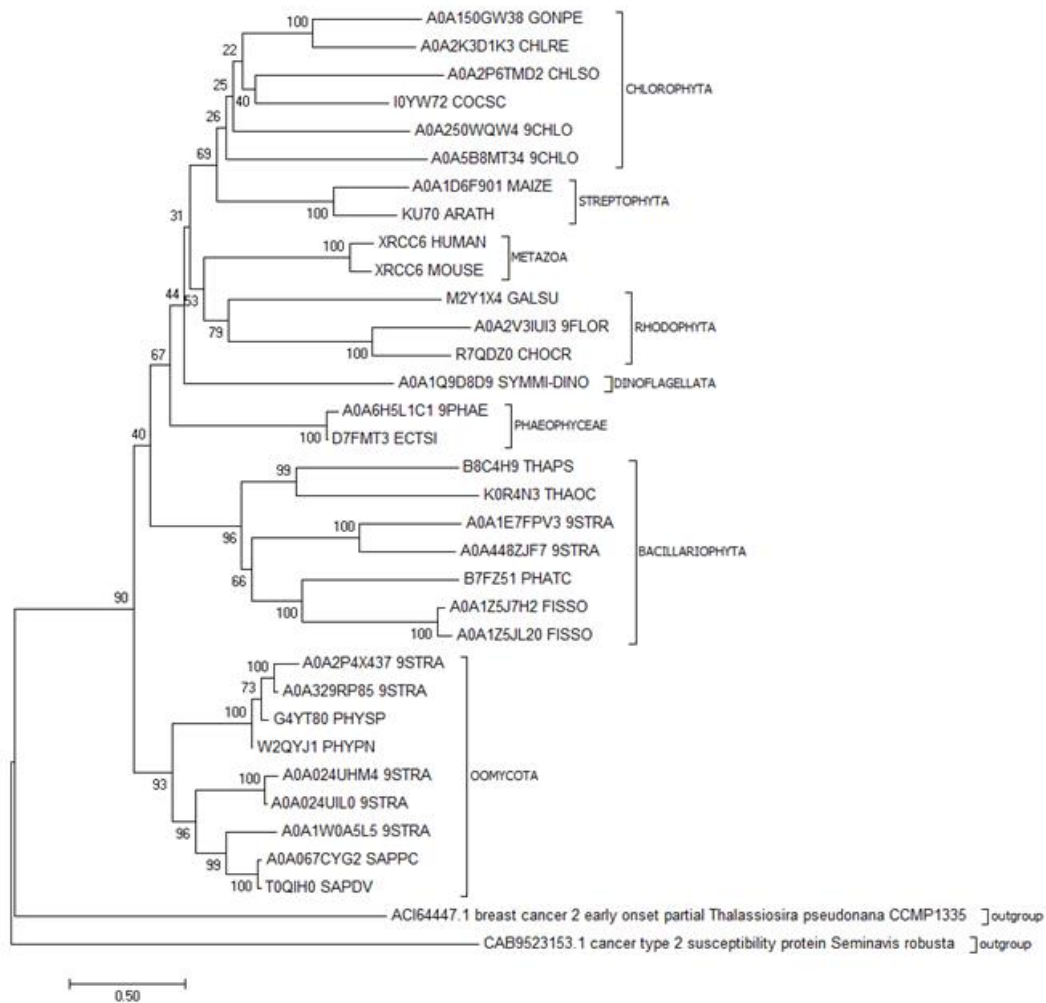


Figure 3.21. Evolutionary relationships of taxa of Ku70. The evolutionary history was inferred using the Neighbour-Joining method (Saitou and Nei, 1987). The optimal tree with the sum of branch length = 21.65789412 is shown. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) are shown next to the branches (Felsenstein, 1985). The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Poisson correction method (Zuckerkand and Pauling 1965) and are in the units of the number of amino acid substitutions per site. The rate variation among sites was modelled with a gamma distribution (shape parameter = 2). The analysis involved 34 amino acid sequences. All ambiguous positions were removed for each sequence pair. There were a total of 2811 positions in the final dataset. Evolutionary analyses were conducted in MEGA7 (Kumar et al., 2016).

Like the HR pathway, there are other genes that could have been targeted to impair the NHEJ pathway. One, is DNA ligase IV (LigIV), which ligates the nonhomologous DNA ends during the final stage of NHEJ (Wilson et al., 1997). This gene has been targeted in multiple species resulting in disruption of the NHEJ pathway and an increase in which DNA is repaired through (Schorsch et al., 2009; Angstenberger et al., 2019).

### 3.3 Discussion – Potential Molecular Underpinnings of Tolerance Diversity

There is no literature speculating on the significant difference in tolerance to zeocin between the three species, specifically *T. pseudonana*. However, it is known that zeocin is not an effective selecting agent for transforming *T. pseudonana* since the wild-type strain is vulnerable to small concentrations of zeocin. As previously mentioned zeocin damages DNA by intercalating into the DNA creating DSBs. Eukaryotic organisms primarily repair DNA DSBs via NHEJ as it requires low energy input and can be carried out at any point in the cell cycle. The only identified molecular difference in DNA repair pathways between the centric *T. pseudonana* and pennates *P. tricornutum* and *F. cylindrus* is the X-family polymerase, pol  $\lambda$ , with only pennates having copies of the protein.

All pol X polymerases have been shown to be active in the formation of DNA junctions originating at incompatible ends during NHEJ, especially in the presence of Ku:DNA complexes characteristic of NHEJ (Ma et al., 2004). Despite the apparent lack of these polymerases in Thalassiosirales order centric diatoms, they still have the core components of NHEJ (Ku, Artemis:DNA-PK<sub>cs</sub>, and XRCC4:DNA ligase IV) which are able to complete the NHEJ function (Wilson et al., 1997; Ma et al., 2004). As mentioned earlier this is not uncommon as there are other model organisms that do not possess any X-family polymerases either, such as *D. melanogaster* or *C. elegans*. However, given the evolutionary history in diatoms the situation is more complex. Molecular evidence shows the first diatoms to emerge (~240 Mya) were large radial centric species, and diatom fossils found dating as far back at 190 Mya (Kooistra and Medlin 1996; Medlin et al., 1997; Sims et al., 2006; Benoiston et al., 2017). The first pennate diatom fossils only date back to the Campanian age of the late Cretaceous period (75Mya) with molecular data also supporting the appearance of pennate around this time (Kooistra and Medlin 1996; Sims et al., 2006). Since orthologues of Pol  $\lambda$  are found in green and red algae, plants, and dinoflagellates and evidence Pol  $\lambda$  share a common origin from a *Bacillus* bacteria species would suggest Pol  $\lambda$  was lost in lineage of centric diatoms after the divergence of the pennate lineage (Uchiyama et al., 2004; Beinstock et al., 2014).

The increase in tolerance to zeocin observed in pennate diatoms may be impacted by the function of pol  $\lambda$ , as it catalyses template-dependent synthesis with deoxynucleotide triphosphate (dNTP) and ribonucleoside tri-phosphate (rNTP; McElhinny et al., 2003), functions as an error-prone polymerase (Dominguez et al., 2000), slip on the template strand (Davis et al., 2008), polymerises across a discontinuous template strand (Gu et al., 2007), and has template independent activity (Ramadan et al., 2004; Moon et al., 2007; Lieber, 2010). Incorporating this activity to the NHEJ and BER pathways increases the ability to repair lesions by addition of non-specific nucleotides to DNA overhangs increasing the chance of microhomology for more efficient repair (Gu et al., 2007; Lieber, 2010).

Despite the significant differences in tolerance to zeocin, tolerance to MMS was similar among the three species (Figure 3.18a). This may be because homologous recombination is the primary pathway to repair damage caused by MMS, and all diatom proteomes revealed conservation of the same set of core HR proteins (Figure 3.6).

Further experiments that analyse the transcriptome and proteome of diatoms under mutagenic stress are needed to confirm if Pol  $\lambda$  has a direct role in repairing damage caused by zeocin. Given the results from conducted dose response curves in this chapter, the pennate diatoms *F. cylindrus* and *P. tricornutum* display increased tolerance to zeocin induced DNA damage. Further studies using more DNA damaging agents across a more comprehensive set of species is needed to confirm this trend, but the dose-response data in conjunction with the absence of pol X family polymerases in centric diatoms suggests that modifications to DNA repair genes may have a larger role than thought in diatoms' ability to colonise such a vast array of habitats.

## 4 CRISPR/Cas Genome Editing in *T. pseudonana* and Comparative Genomics of *Brca2* Knock Outs

### 4.1 Introduction

#### 4.1.1 Genome Editing: CRISPR/Cas

Genome editing is a powerful tool which can reveal the roles of genes and proteins by creating programmed insertions, deletions or replacements of specific sequences within an organism's genome (Khalil, 2020). There are four major mechanisms by which genome editing can be achieved: meganucleases (MegNs), zinc finger nucleases (ZFNs), transcription activator-like effector nuclease (TALENs) and clustered regularly interspaced short palindromic repeats (CRISPR) with CRISPR-associated protein 9 (CRISPR/Cas9; Khalil, 2020). CRISPR is the simplest to program, most applicable and cost-effective (Li et al., 2020). It was discovered by Ishino et al in 1987 in the *E. coli* genome (Gostimskaya, 2022). The short 'spacer' sequences between palindromic repeat sequences in the *E. coli* genome are relics of past encountered viral DNA, providing a library of potentially harmful foreign DNA sequences and were found to be part of a prokaryotic 'immune' system (Marraffini and Sontheimer, 2009). In 2012, the potential of CRISPR/Cas9 as a precise genome-editing tool was first unveiled by a research collaboration led by Jennifer Doudna and Emmanuelle Charpentier, beginning a significant movement in genetic engineering (Jinek et al., 2012).

The mechanism of CRISPR/Cas9 involves the use of guide RNA (gRNA) to locate specific sequences in the genome, which is then cleaved by the nuclease Cas9. CRISPR/Cas9 is a powerful and adaptable tool because of its ability to create precise double-strand breaks through base complementarity *in vivo* (Jinek et al., 2012). This technology has been used in a wide range of research fields due to the ability to be programmed to unique DNA regions (Adli, 2018; Rasheed et al., 2022). The majority of research utilising CRISPR/Cas technology is centred around research questions regarding practical aspects such as clinical research (i.e., drug development and disease modelling), agriculture biotechnology, and treatment of genetic disorders. However, recently CRISPR/Cas genome editing has been applied to

answer questions in evolutionary biology to help understand fundamental questions in genomics.

The application of CRISPR/Cas9 genome editing in diatoms has great potential for understanding their biological functions and increasing their potential in biotechnology. A fundamental understanding of the molecular underpinnings of diatom processes aids a wide range of environmental and molecular research as they are major contributors to the global climate process (Nelson et al., 1995; Serôdio and Lavaund, 2020). As well as furthering the potential to optimize their production of valuable compounds such as recombinant proteins, pharmaceutical intermediates, and antioxidants CRISPR/Cas9 has been pivotal in increasing our understanding of fundamental cellular processes in diatoms and their wider impact on global climate and biological processes.

#### 4.1.1.1 *Golden Gate Cloning as a Method to Create CRISPR/Cas9 Constructs*

Golden Gate cloning can be used to simultaneously assemble multiple DNA fragments into a single vector or construct using Type II-S restriction enzymes (Pingoud and Jeltsch, 2001) and T4 DNA ligase (Weber et al., 2011; Engler and Marillonnet, 2014). Unlike standard Type II restriction enzymes such as *EcoRI* and *BamHI*, these enzymes cut DNA outside of their recognition sites creating non-palindromic overhangs (Pingoud and Jeltsch, 2001). With a variety of potential overhang sequences possible, multiple fragments of DNA can be assembled by using a unique combination creating scarless ligation sites. Additionally, because the final product does not have a Type IIS restriction enzyme recognition site, the correctly ligated product cannot be cut again, meaning the reaction is essentially irreversible (Engler and Marillonnet, 2014). Golden Gate cloning has been successfully used to create CRISPR/Cas constructs used to facilitate genetic editing in diatoms (Hopes, 2016; Hopes et al., 2017) and in a variety of organisms such as plants (Castel et al., 2019), yeast (Lee et al., 2015), prokaryotes (Hinz et al., 2022) and animals (Fonseca et al., 2020). Using methods described in Hopes et al., 2017 (detailed in Chapter 2) two constructs were created through golden gate cloning to create separate *T. pseudonana* homozygous knockout strains of *Brca2* and *Ku70*.

### 4.1.2 Whole Genome Sequencing in Diatoms

Historically the funding for whole genome sequencing (WGS) was exclusively for clinical research, but the significant reduction of WGS costs in the past decade has opened up sequencing technologies to every aspect of molecular biology questions (Giani et al., 2020). This has been mainly driven by short-read sequencing technology such as Illumina but recently long-read sequencing (i.e., Oxford Nanopore and PacBio HiFi) has increased the diversity in the industry and this competition has helped costs fall.

Diatoms are one of the most diverse and ecologically important phytoplankton groups in the ocean as they are responsible for approximately 20% of global carbon fixation and roughly 40% of marine primary productivity (Nelson et al., 1995). The first diatom to be sequenced was *Thalassiosira pseudonana* (Armbrust et al., 2004), which sparked the sequencing of diatom species with the following genomes sequenced at the time of writing: *Phaeodactylum tricornutum* (Bowler et al., 2008), *Thalassiosira oceanica* (Lommer et al., 2012), *Fragilaria radians* (Galachyants et al., 2015), *Fistulifera solaris* (Tanaka et al., 2015), *Cyclotella cryptica* (Traller et al., 2016), *Pseudo-nitzschia multistriata* (Basu et al., 2017), *Fragilariopsis cylindrus* (Mock et al., 2017), *Skeletonema costatum* (Ogura et al., 2018), *S. marinoi* (Johannsen et al., 2019), *Seminavis robusta* (Osuna-Cruz et al., 2020), *Nitzschia inconspicua* (Oliver et al., 2021), *Minidiscus variabilis* and *Pseudo-nitzschia multiseriata*. These studies have revealed the unexpected presence of pathways such as the urea cycle, which is uncommon for unicellular organisms, as well as providing new insights into the evolutionary biology of phytoplankton.

## 4.2 Results

### 4.2.1 *In Vitro* CRISPR/Cas Temperature Assay

Using methods published by OmicronCR® (<https://www.omicroncr.co.uk/>), sgRNAs were synthesised via PCR to combine the sequences necessary to bind to Cas9 (crRNA – tracrRNA chimera) and to direct Cas9 to cleave the genetic sequence of interest (Figure 4.1; Table 4.1; Hopes et al., 2017).



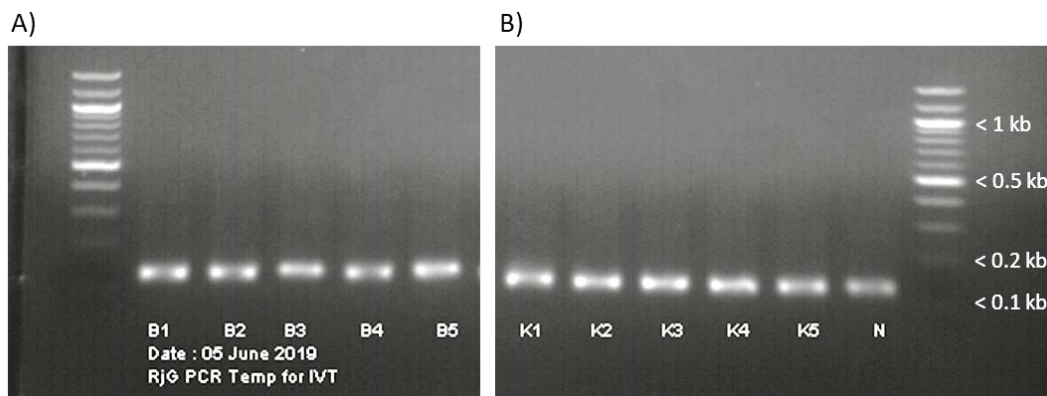


Figure 4.1. Products of PCR to synthesise programmable sgRNA 20nt sequence and CRISPR RNA sequence to be transcribed into RNA for in vitro temperature cutting assay. Products of sgRNA templates for A) BRCA2 [B1-B5], and B) KU70 [K1-K5]. The ladder is NEB 100bp (CAT #B7025), with 1 kb, 0.5 kb, 0.2 kb, and 0.1 kb bands marked.

Successful PCR products were transcribed into short RNA using the HiScribe T7 RNA Synthesis Kit (Invitrogen) and products were analysed on a 10% polyacrylamide TBE-urea gel (Figure 4.2). Successfully synthesised sgRNAs were incubated with the Cas9 protein to form an RNP complex and incubated overnight at 20°C with a linearized plasmid containing the corresponding wild-type target gene (*Brca2* or *Ku70*). After incubation, the reaction was stopped by a combination of a stop buffer and heat inactivation and run on a 0.8% agarose gel at 85 volts for 1 hour (Figure 4.3). Samples which contained a single band the same size as the control were deemed to have not combined with Cas9 to form a RNP complex necessary to cut the gene and were discarded for future use. If cutting was successful, 3 bands were present at expected sizes given the position of the target sequence within the plasmid. Each band size was measured against a 1 kb DNA ladder (NEB, CAT #N3232S) to confirm the bands were a product of digestion with sgRNA-guided CRISPR/Cas (Figure 4.3; Table 4.1). Based on these results, sgRNA's BRCA2\_sgRNA1 and BRCA2\_sgRNA3, and Ku70\_sgRNA3 and Ku70\_sgRNA6 were chosen for cloning into the final construct to be transformed into *T. pseudonana*.

Table 4.1. List of primers used to synthesise sgRNAs with CRISPR RNA to create RNP complex for the in vitro temperature assay (IVT). The T7 promoter (HiScribe) is in bold, the programmable target sequence within the genome is in red, and the region where the primers and oligo 2 (reverse primer) overlaps are underlined.

Targeted Gene	Primer ID	Primer
<i>Brca2</i>	BRCA2_sgRNA1_F_T7	<b>TAATACGACTCACTATAGGG</b> <u>AGCTAGCAGCAATGGAGAGA</u> GTTTTAGAGC TAGAAATAGCAAGTAAAAATA <b>AG</b>
	BRCA2_sgRNA2_F_T7	<b>TAATACGACTCACTATAGGG</b> <u>AATGAAGGGCCGCTACGATA</u> GTTTTAGAGC TAGAAATAGCAAGTAAAAATA <b>AG</b>
	BRCA2_sgRNA3_F_T7	<b>TAATACGACTCACTATAGGG</b> <u>TTCAAATCAAGACCGCCAAA</u> GTTTTAGAGCT AGAAATAGCAAGTAAAAATA <b>AG</b>
	BRCA2_sgRNA4_F_T7	<b>TAATACGACTCACTATAGGG</b> <u>GCAAGATGGGATGCAACTCT</u> GTTTTAGAGCTA GAAATAGCAAGTAAAAATA <b>AG</b>
	BRCA2_sgRNA5_F_T7	<b>TAATACGACTCACTATAGGG</b> <u>TTGGCGTTCATGATTGCAC</u> GTTTTAGAGCT AGAAATAGCAAGTAAAAATA <b>AG</b>
<i>Ku70</i>	Ku70_sgRNA2_F_T7	<b>TAATACGACTCACTATAGGG</b> <u>CCAAACACTCACAGCCCGAA</u> GTTTTAGAGCT AGAAATAGCAAGTAAAAATA <b>AG</b>
	Ku70_sgRNA3_F_T7	<b>TAATACGACTCACTATAGGG</b> <u>GCAACCCTAGCATGTTTGAG</u> GTTTTAGAGCTA GAAATAGCAAGTAAAAATA <b>AG</b>
	Ku70_sgRNA4_F_T7	<b>TAATACGACTCACTATAGGG</b> <u>AATGTGCAACTTGCATCCA</u> GTTTTAGAGCT AGAAATAGCAAGTAAAAATA <b>AG</b>
	Ku70_sgRNA5_F_T7	<b>TAATACGACTCACTATAGGG</b> <u>AGCATCTAAAATGATTGCAT</u> GTTTTAGAGCT AGAAATAGCAAGTAAAAATA <b>AG</b>
	Ku70_sgRNA6_F_T7	<b>TAATACGACTCACTATAGGG</b> <u>GAAGATATGGATACATTGTT</u> GTTTTAGAGCTA GAAATAGCAAGTAAAAATA <b>AG</b>
	Reverse Primer	Oligo2

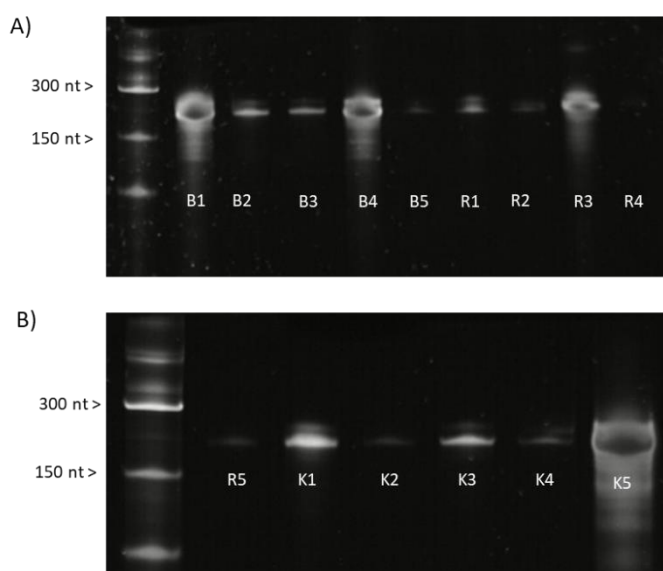


Figure 4.2 RNA products synthesised via T7 HiScribe (Invitrogen) kit for IVT assay. A) samples B1 – R4; B) Samples R5 – K5. 10% polyacrylamide TBE-urea gel. The ladder used in both gels is the Low Range single-stranded RNA (ssRNA) Ladder (NEB #N0364S). Products are expected to be 170nt in length. Samples B1-B5 are IVTs to target BRCA2, samples K1-K5 are IVTs to target Ku70. Samples R1-R5 were synthesised to target Rad52 but were not carried forward for transformation.

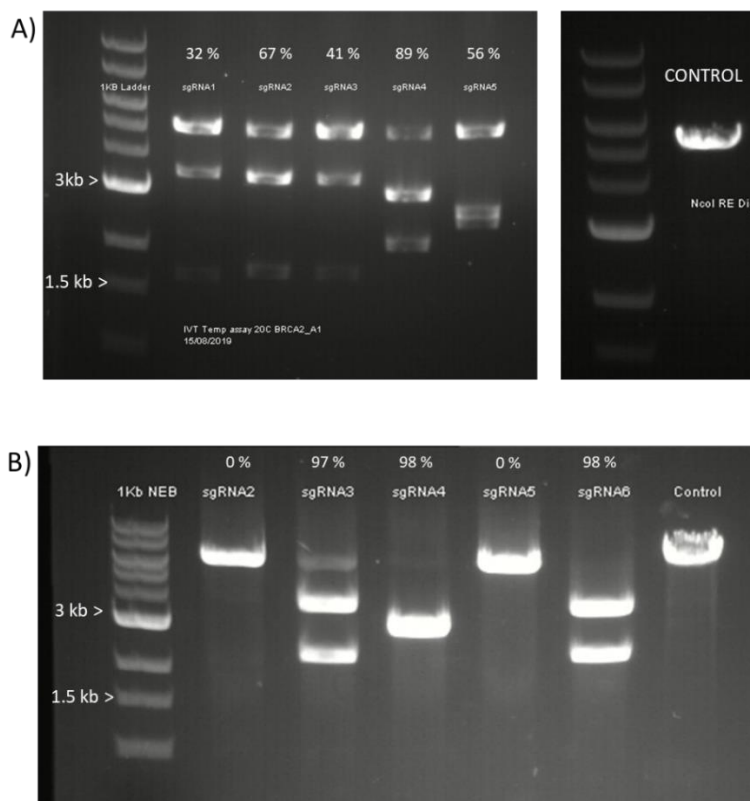


Figure 4.3. Results of the *in vitro* temperature (IVT) assay. A) Cutting assay of IVTs targeting linearised TOPO vector containing the *BRCA2* wild-type gene sequence. The second panel is a linearised plasmid incubated with CRISPR/Cas protein but without IVTs. B) IVTs cutting of linearised plasmid containing *Ku70* wild-type gene sequence. Control is incubated with CRISPR/Cas protein without IVTs. Numbers above each lane represent the percentage of linearised plasmid cut (i.e., cutting efficiency) determined by ImageJ.

#### 4.2.2 Golden Gate Cloning

Once sgRNAs had been proven successful *in vitro* they were cloned into the final level two construct for their respective target genes (*Brca2* or *Ku70*) using methods stated in Chapter 2.3. Final constructs (Figure 4.4) were confirmed to be complete via restriction digest with the endonuclease EcoRV-HF (Figure 4.4c; NEB #R3195S) and sanger sequencing of select regions using primers in Table 4.2. Successfully cloned level two vectors to target *Brca2* (B2-A1\_SC1 and B2-A1\_SC2) and *Ku70* (K2\_SC1 and K2\_SC2) were amplified through mini-prep in bacteria (Monarch Plasmid Miniprep kit: NEB # T1010S) and stored at -20°C until use in biolistic transformation.

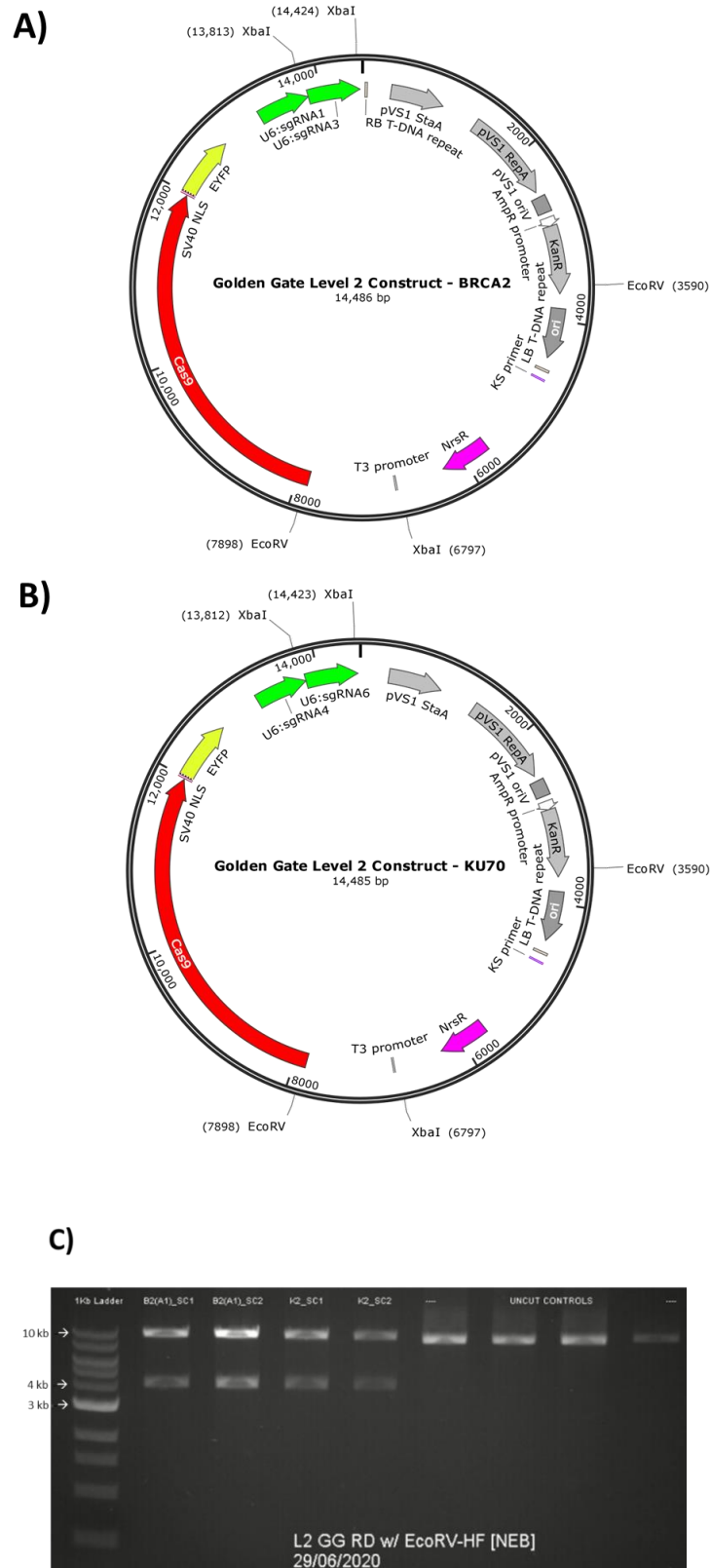


Figure 4.4. Maps of final level 2 golden gate construct for *Brca2* (A) and *Ku70* (B). In both maps, red = CRISPR/Cas9 component, green = U6-promoter:sgRNAs, and pink = Nourseothricin resistance gene (*NrsR*). C) Results of restriction digest with *EcoRV*-HF (NEB # R3195S) of final golden gate constructs. Uncut controls are constructs incubated with all components except for the restriction enzyme, *EcoRV*-HF.

## 4.2.3 Transformation and Selection of Homozygous Knock Out Cell Lines

### 4.2.3.1 Selection of Edited Cell Lines Through PCR

As stated in Chapter 2 section 2.3.4, the methods from Hopes et al., 2017 were used to select transformed cell lines. After ~14 days, 6 primary colonies appeared on plates containing cultures transformed with *Brca2* constructs with only one showing potential for editing while 24 grew after transformation with *Ku70* constructs. Lysate from subsamples was used for colony PCR with primers which amplify the targeted gene with primers shown in table 4.2. Cultures showing potential for editing resulted in two bands -wild-type and truncated size after a portion of the gene is removed from successful CRISPR/Cas editing- when run on agarose gels and were used for further selection (Figure 4.5). These are mosaic colonies, meaning they contain a mixture of populations which have differing genotypes at a single locus.

Table 4.2. Primers used to amplify section of targeted genes, *Brca2* or *Ku70*, containing region where CRISPR/Cas was targeted.

Targeted Gene	Primer ID	Primer Sequence	Wild-Type Product Size	Edited Product Size
<i>Brca2</i>	BRCA2_F1	TGGAATGTATGCACCGAATG	591 bp	394 bp
	BRCA2_R2	TCCCACTGCGATTCTCCTTT		
<i>Ku70</i>	Ku70_F1	CGCTCTCCTCCCACAATCAA	1,323 bp	626 bp
	Ku70_R2	CTACGAGGCGATACGATGGTT		

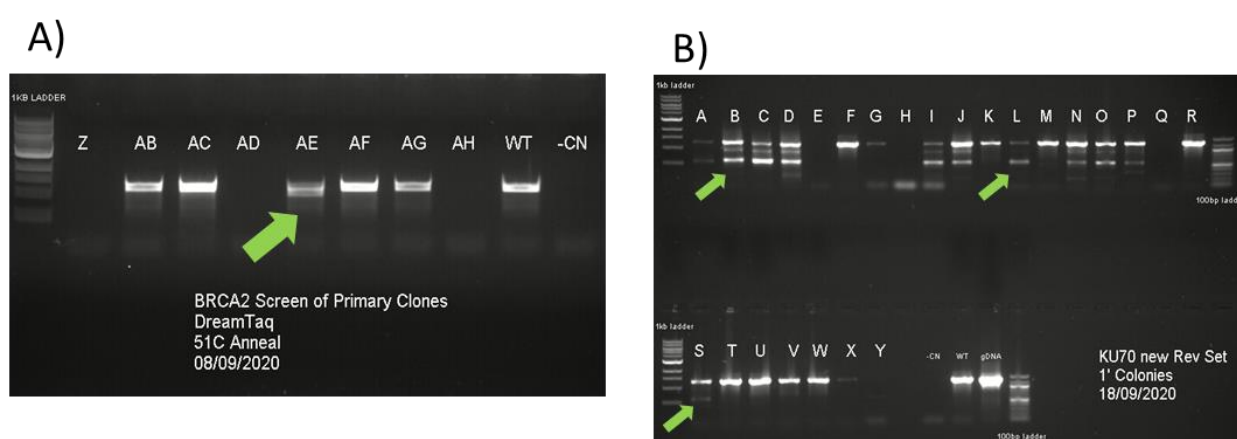
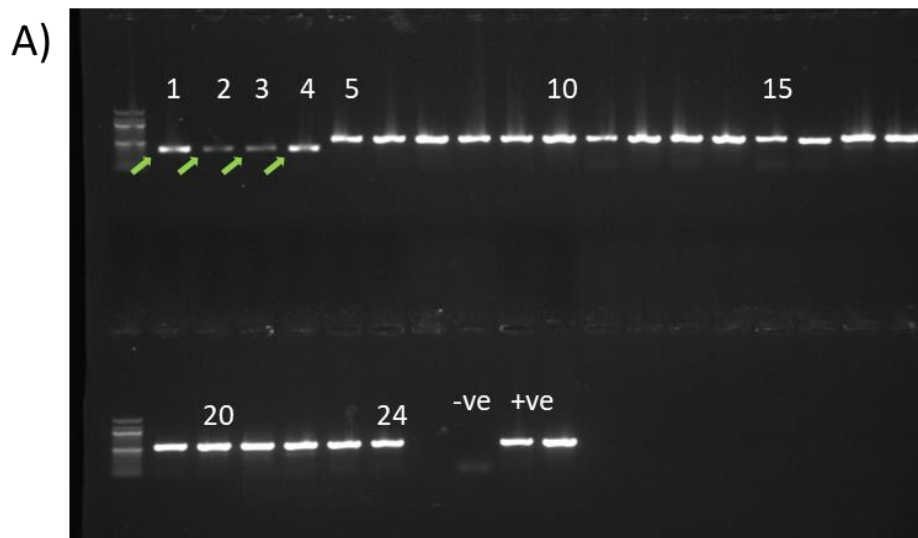


Figure 4.5. PCR of target gene; A) *Brca2*, B) *Ku70* of primary colonies selected on HS-Aquid plates containing 100  $\mu$ g/ml clonNAT. Green arrows indicate examples of mosaic colonies containing wild-type and potentially successfully edited cell populations.

Culture 'AE' was the only primary colony which showed potential editing after transformation with *Brca2* constructs which gives a transformation efficiency of 12.5%. This colony was spread onto 5 selection plates to obtain clonal cultures (secondary clones). For *Ku70*, 11 of the 25 colonies showed evidence of editing (44% of transformants). These cultures: A, B, C, D, I, J, L, N, O, P, and S were spread onto selection plates. Once secondary colonies appeared they were picked in the same manner as the primary colonies, with half the resuspended colony taken for colony PCR and the other half placed in selective media. The same primers (Table 4.2) were used to screen the secondary clones by PCR (Figure 4.6).



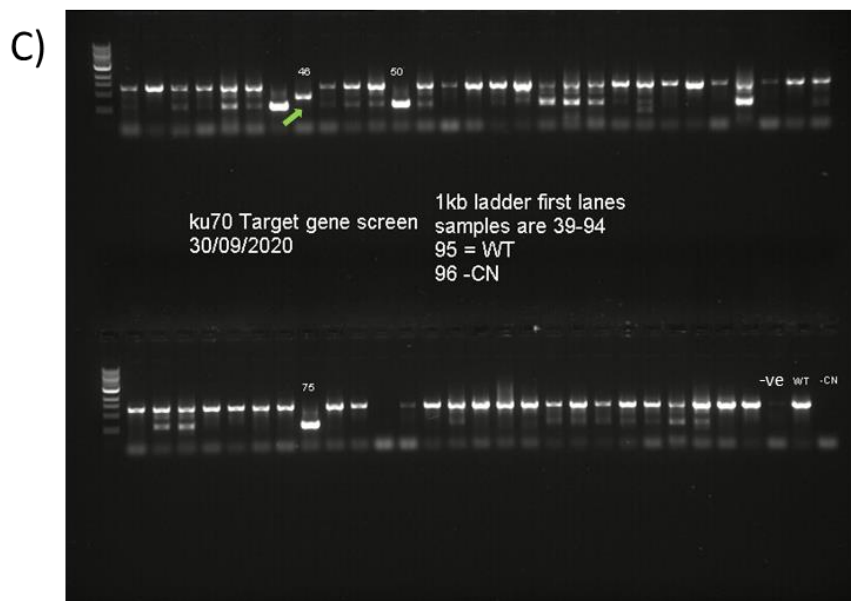


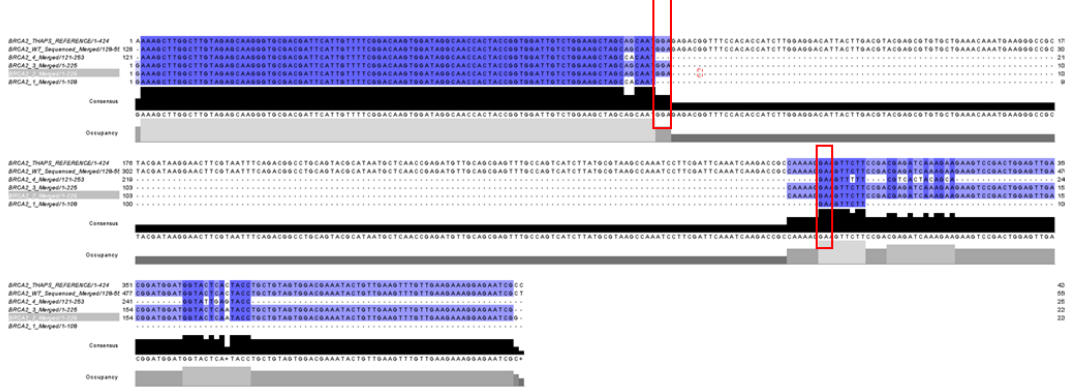
Figure 4.6. PCR of target gene in secondary clones; A) *Brca2*; B & C) *Ku70*; selected on HS-Aquil plates containing 100 µg/ml clonNAT. Green arrows indicate clonal cultures which potentially possess the edited target genes with no wild-type copies. Cultures indicated with green arrows were isolated for further screening via Sanger sequencing.

#### 4.2.3.2 Sanger Sequencing

Products of PCR from secondary clones were cleaned up using the Monarch® PCR & DNA Cleanup Kit (NEB # T1030S) and sent for Sanger sequencing through Eurofins Genomics UK (<https://www.eurofins.co.uk/genomic-services/>). Cultures were deemed to be homozygous (bi-allelic) knockouts if the sequencing results returned the gene sequence with the region between the intended Cas9 cut sites removed and with no traces of the wild-type sequence. Secondary cultures 1 – 4 were confirmed to be homozygous knockouts for the edited *Brca2* gene (*Brca2*  $-/-$ ; Figure 4.7) and cultures 15 and 46 for *Ku70* (*Ku70*  $-/-$ ; Figure 4.7).



A)



B)

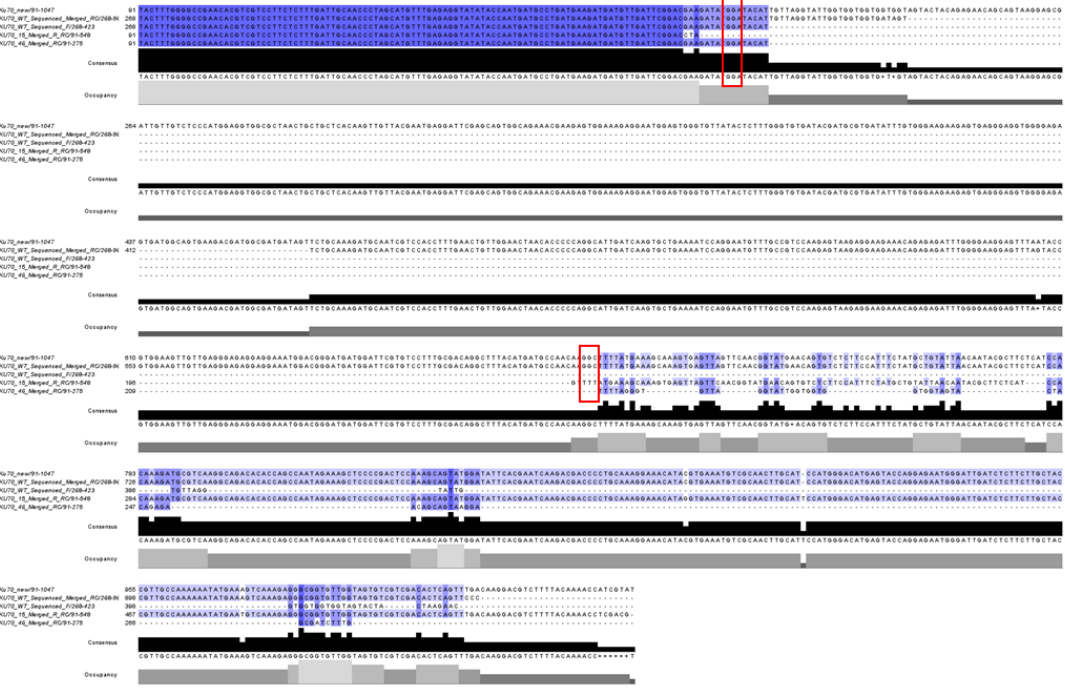


Figure 4.7. Alignment of Sanger sequencing results from PCRs amplifying the wild-type sequence of A)Brca2 and B) Ku70. A) Row 1 = Thaps3\_263089 coding sequence, row 2 = wild-type PCR product, rows 3-6 sequences from Brca2 -/- secondary clones 1-4. B) Row 1 = Thaps3\_263010 coding sequence, row 2 = wild-type PCR product, row 3-4 = sequences from Ku70 -/- secondary clones 15 and 47. Red boxes indicate PAM sequences where Cas9 was targeted to cut by sgRNAs.

#### 4.2.4 Homozygous *Ku70* Knock-Out Cultures are Unstable.

Once cultures were confirmed to contain homozygous editing at desired locations the phenotypes were then characterised. However, maintaining stable cultures of *Ku70* *-/-* cell lines proved to be challenging given the critical cellular process disturbed by editing. Previous studies have documented that disruption of core genes of the NHEJ DNA repair pathway (Gandía et al., 2016;), the proportion of DNA lesions repaired through homologous recombination is increased (Angstenberger et al., 2019) leading to an increase in gene conversion events. The *Ku70* locus in *Ku70* *-/-* cell lines became unstable because of this effect and the genetic integrity at the location of editing was compromised. Inferring changes in the genome due to the loss of *Ku70* function would be difficult to understand since the locus where *Ku70* was edited was unstable. Therefore, after discussions with my supervisory team, we decided to take forward the *Brca2* *-/-* knock-out cell lines for further experiments and resequencing.

#### 4.2.5 Impaired DNA Repair in Edited Cell Lines.

##### 4.2.5.1 Dose-Response Curves

*Brca2* *-/-* knock-out cell lines or organisms are hyper-sensitive to DNA-damaging agents such as methyl methanesulfonate (MMS), providing evidence of the essential function of *Brca2* in the HR DNA repair pathway (Sharan et al., 1997; Ding et al., 2017). The function of *Brca2*, or any DNA repair genes, in diatoms has not been experimentally proven through reverse genetics. To confirm the knockout cultures were deficient in DNA repair due to the loss of *Brca2* function, dose-response curves were conducted using the DNA-damaging agent MMS. These were run in triplicates in parallel with wild-type *T. pseudonana*, using methods and MMS concentrations stated in Chapter 2 section 2.1.3.2. As the literature predicted, *Brca2* *-/-* *T. pseudonana* cell lines are hypersensitive to MMS as shown by the significant decrease in the EC<sub>50</sub> value (Table 4.3; Figure 4.8).

Table 4.3: EC<sub>50</sub> values for wild type and *Brca2* *-/-* *T. pseudonana* cell lines exposed to the DNA damaging agent methyl methanesulfonate (MMS).

Culture	EC <sub>50</sub> (mM)	Lower 95%	Upper 95%
Wild Type	0.53	0.46	0.61
<i>Brca2</i> <i>-/-</i>	0.16	0.13	0.21

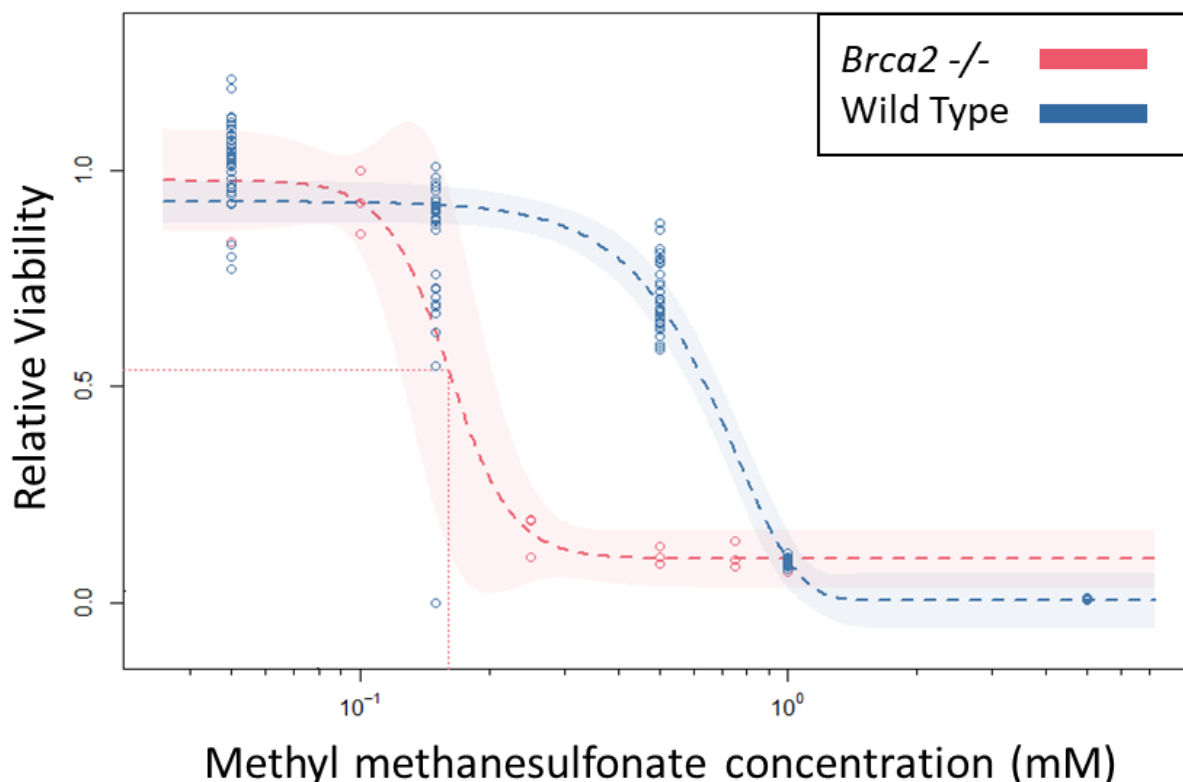


Figure 4.8. Dose-response curve of wild-type and *Brca2*  $-/-$  *T. pseudonana* cell lines exposed to increasing concentrations of MMS. The dotted lines are fitted logistic models and shaded areas represent a 95% confidence interval. Models and  $EC_{50}$  values were determined through the R package *drda* (Malyutina et al., 2023).  $n=3$ , wild-type data has been supplemented with plate reader data ( $n=12$ ).

Hyper-sensitivity to MMS confirmed BRCA2's function as a core enzyme involved in DNA repair in *T. pseudonana*, however, this does not provide insight into the role of BRCA2 in the context of environmental stressors. MMS is artificially produced and therefore not found in the natural environment and creates a higher frequency of DNA damage, specifically DSBs, than environmental factors. To understand the impact of the loss of BRCA2 function on the fitness of *T. pseudonana* in the environment, *Brca2*  $-/-$  cultures were grown under a gradient of temperatures (7.6°C – 33.9°C).

#### 4.2.5.2 Temperature Response Curve

Previous data generated in the Mock research group by Dr Katrin Schmidt (Schmidt, 2017) has described the optimum ( $T_{opt}$ ), maximum ( $T_{max}$ ) and minimum ( $T_{min}$ ) temperatures that wild-type *T. pseudonana* can survive. Further analysis by Dr Andrew Toseland has shown that the rate of mutation increases in *T. pseudonana* with increasing temperature

creating increased genomic instability (unpublished). With *Brca2* core to the HR DNA repair pathway in diatoms, and temperature stress shown to increase genomic instability, a temperature response curve was conducted to elucidate the role of *Brca2* in the adaptation of *T. pseudonana* to environmentally relevant stress. The methods used are described in Chapter 2 section 2.1.3.3 and are identical to those used by Dr Katrin Schmidt in the original experiments using wild-type cultures.

The normalised specific growth rate ( $\mu$ ) of each replicate ( $n=3$ ) for each temperature was used to model a temperature performance curve to obtain the  $T_{opt}$ ,  $T_{max}$  and  $T_{min}$  for *Brca2*  $-/-$  cultures. The specific growth rates were calculated using the R package *growthcurver* (Sprouffske and Wagner, 2016). However, for the cultures at higher temperatures which either did not grow or display typical growth, *growthcurver* was unable to calculate the specific growth rate. For these cultures, the specific growth rate was calculated using the equation in Chapter 2 section 2.1.3.1. Seven models were fit to the data using the R package *rTPC* (Table 4.4; Figure 4.9; Padfield et al., 2021). The Akaike information criterion (AIC) statistic was used to determine the model which fit the data the best (Table 4.4). AIC estimates prediction error and provides a value to the quality of a statistical model (McElreath, 2016). The model of best fit was the Beta\_2012 model from Neihaus et al., 2012 (Table 4.4).

Table 4.4. A) Equations of models fitted to data through rTPC r package; B) maximum temperature ( $T_{max}$ ), optimum temperature ( $T_{opt}$ ), minimum temperature ( $T_{min}$ ), and AIC values of each fitted model.

**A)**

**Model Name (citation) Equation**

Beta_2012 (Niehaus et al., 2012)	$rate = \frac{a \left( \frac{temp - b + \frac{c(d-1)}{d+e-2}}{c} \right)^{d-1} \cdot \left( 1 - \frac{temp - b + \frac{c(d-1)}{d+e-2}}{c} \right)^{e-1}}{\left( \frac{d-1}{d+e-2} \right)^{d-1} \cdot \left( \frac{e-1}{d+e-2} \right)^{e-1}}$
Briere_1999 (Briere et al., 1999)	$rate = a \cdot temp \cdot (temp - t_{min}) \cdot (t_{max} - temp)^{\frac{1}{b}}$
Flinn_1991 (Flinn, 1991)	$rate = \frac{1}{1 + a + b \cdot temp + c \cdot temp^2}$
Gaussain_1987 (Lynch and Gabriel, 1987)	$rate = r_{max} \cdot exp \left( -0.5 \left( \frac{ temp - t_{opt} }{a} \right)^2 \right)$
Hinshelwood_1947	$rate = a \cdot exp^{\frac{-e}{k \cdot (temp + 273.15)}} - b \cdot exp^{\frac{-e_h}{k \cdot (temp + 273.15)}}$
Joehnk_2008 (Jöhkn et al., 2008)	$rate = r_{max} \left( 1 + a \left( \left( b^{temp - t_{opt}} - 1 \right) - \frac{\ln(b)}{\ln(c)} (c^{temp - t_{opt}} - 1) \right) \right)$
Boatman_2017 (Boatman and Geider, 2017)	$rate = r_{max} \cdot \left( \sin \left( \pi \left( \frac{temp - t_{min}}{t_{max} - t_{min}} \right)^a \right) \right)^b$

**B)**

Model Name	$\mu_{max}$	$T_{max}$ (°C)	$T_{opt}$ (°C)	$T_{min}$ (°C)	AIC
Beta_2012	0.97	31.49	20.78	1.69	-31.9393
Briere_1999	0.91	33.90	19.51	5.74	-25.7286
Flinn_1991	1.00	33.90	19.70	-10.38	-22.2102
Gaussain_1987	0.97	35.22	19.70	4.18	-30.9565
Hinshelwood_1947	0.74	33.39	21.19	-31.47	19.3847
Joehnk_2008	0.85	33.26	20.06	4.89	23.1904
Boatman_2017	0.96	32.10	20.25	0	-29.0307

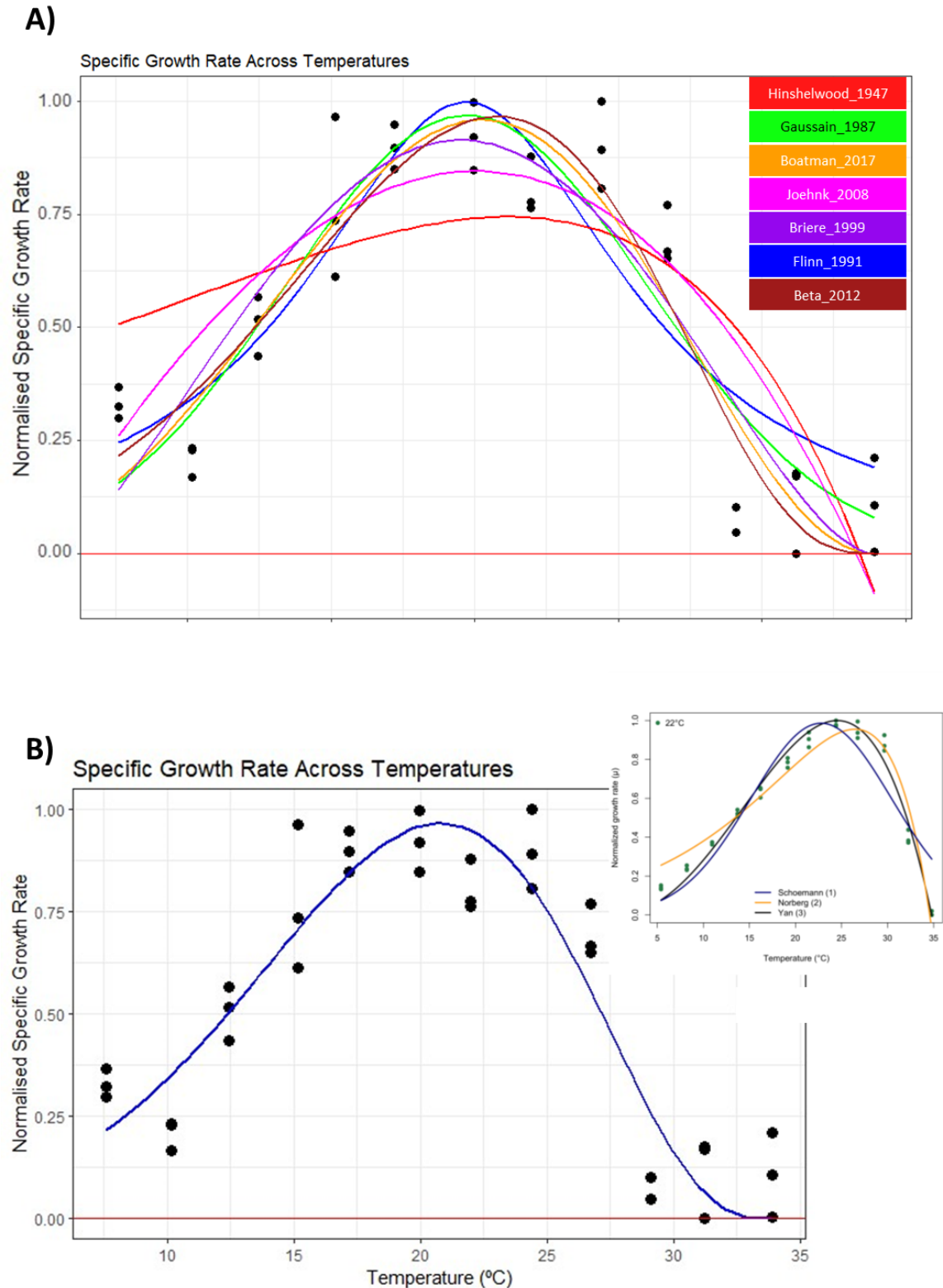


Figure 4.9. The normalised specific growth rate ( $\mu$ ) of *Brca2*  $-/-$  cultures over an increasing temperature ( $^{\circ}\text{C}$ ) gradient overlayed with models fitted through *rTPC* in R to predict:  $T_{\text{max}}$ ,  $T_{\text{opt}}$  and  $T_{\text{min}}$ . A) all seven fitted models fitted, legend describes the colour of the corresponding model; B) the Beta\_2012 (Neihaus et al., 2012) model fitted on its own with the temperature response curve for wild-type *T. pseudonana* from Schmidt 2017 overlayed in the upper right. The figure from Schmidt 2016 shows the thermal response curves of three cell lines 9C (blue line), 22 $^{\circ}\text{C}$  (black line) and 32 $^{\circ}\text{C}$  (yellow line).

The *Brca2*  $-/-$  cultures were unable to survive at temperatures above 31.58°C and their optimal growth temperature was predicted to be 20.78°C, compared to 34.57°C and 24.94°C in the wild-type respectively (Schmidt, 2017). These values show a reduction in temperature niche in DNA repair deficient *T. pseudonana* compared to the wild-type due to the loss of *Brca2* function. These data support the hypothesis that DNA repair, specifically homologous recombination during replication, is a mechanism used by *T. pseudonana* to adapt to changing environmental conditions. Further studies to support these data could use genetic engineering to systematically create cell lines deficient in the other main DNA repair pathways to elucidate the importance of each DNA repair pathway in adaptation to environmental change.

#### 4.2.6 Whole Genome Sequencing of *Brca2* $-/-$ Knock-Out Cell Lines

##### 4.2.6.1 DNA Extraction and Quality Assessment of Raw Illumina Sequencing Data

Genomic DNA (gDNA) was extracted using the Qiagen MagAttract Kit (Qiagen) as described in Chapter 2 section 2.4.1 from three independent homozygous *Brca2* knock-out cultures (*Brca2-09*, *Brca2-10* and *Brca2-11*). gDNA quality was assessed via spectrophotometry and pulse-gel electrophoresis (Appendix table A.1; Appendix Figure A.2) and sent to the Earlham Institute (EI) for library preparation and Illumina sequencing.

Results from FastQC (Andrews, 2010) on raw fastq reads showed that Illumina adapters were still present in the data. Illumina adapters were removed from raw reads using trimmomatic/0.39 (Bolger et al., 2014). Trimmed reads were piped into the mapping software bwa-mem (Li and Durbin, 2009) as stated in Chapter 2 section 2.5.3.2.

##### 4.2.6.2 Alignment to Reference Genome

All alignment files (.bam) were analysed with QuailMap BamQC (QualiMap/2.2.1; Okonechnikov et al., 2016) after alignment to the *T. pseudonana* reference genome (<https://mycocosm.jgi.doe.gov/Thaps3/Thaps3.home.html>) using bwa-mem (Li and Durbin, 2009). Table 4.5 shows alignment statistics for the three bam alignment files from *Brca2*  $-/-$  cultures and wild-type.



Table 4.5. QualiMap BamQC statistics on alignment files.

	<b>Wild type</b>	<b><i>Brca2-09</i></b>	<b><i>Brca2-10</i></b>	<b><i>Brca2-11</i></b>
Mapping quality	59.569	59.565	59.557	59.560
Number of mapped reads	31,687,174	7,663,991	6,870,092	6,864,642
Duplication rate	60.76%	14.55%	13.31%	13.51%
General error rate	0.0078	0.0056	0.0056	0.0056
Mean coverage	149.970X	36.001X	32.231X	32.187X
Standard deviation of coverage	78.557X	17.340X	15.617X	15.584X
GC percentage	46.95%	46.83%	46.83%	46.85%

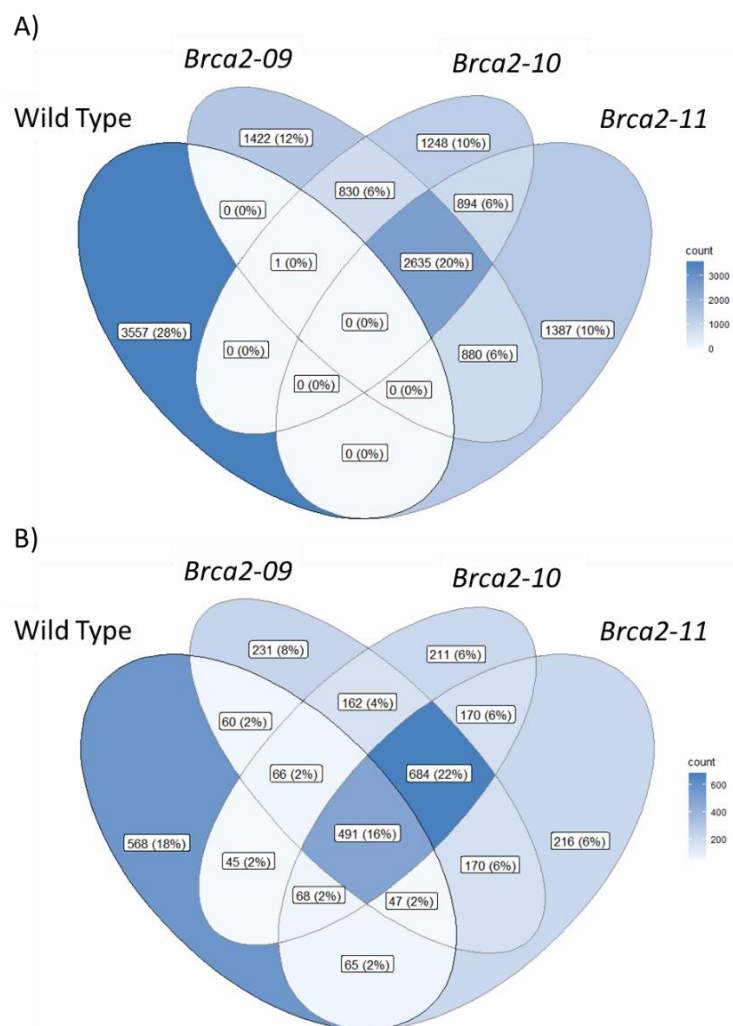
Initially, only the three *Brca2* *-/-* cultures were sent for WGS due to the uncertainty that was brought about by national lockdowns in response to the COVID-19 pandemic. However, it became apparent that simply comparing *Brca2* *-/-* sequencing data to the reference made it difficult to distinguish mutations arising due to the loss of *Brca2* and mutations unique to our lab's *T. pseudonana* CCMP 1335 strain. So, the wild type was sequenced later and through a different library prep method using PCR amplification which resulted in a higher number of reads and coverage. *Brca2* *-/-* sequencing libraries were prepared without PCR to reduce false positives arising through PCR amplification. This did result in a reduced amount of reads but maintains equivalent mean mapping quality, GC percentage, and a slightly lower general error rate than that of the wild type which was sequenced from PCR-generated libraries. Not surprisingly, the PCR-generated libraries for the wild type had significantly more duplications in the data than PCR-free libraries (*Brca2* 09-11).

#### 4.2.6.3 Overview of Called Variants and Initial Filtering

Variants were annotated using the pipeline described in Chapter 2 section 2.5.3.3. The initial variant call format (vcf) files output for each sample contained many annotated variants (Appendix Table A.2), of which most are shared between the re-sequenced wild type and *Brca2* knockouts (Appendix Figure A.3).

For further analysis, all identical variants found in both wild type and *Brca2* cultures were filtered out to create files containing 'unique variants' to each sequenced genome

(Figure 4.10). This enabled focused analysis on variants that potentially are a result of the loss of *Brca2* function. After this filtering, there are still affected genes which overlap with wild-type and mutant strains. The vcf files containing unique variants showed that overall, there was an increase in single nucleotide polymorphisms (SNPs) and small insertions and deletions (INDELS) in *Brca2*  $-/-$  compared to the wild type. SNPs also comprised a higher proportion of variants in the *Brca2* cultures compared with the wild type (Appendix Table A.3; Figure 4.11).



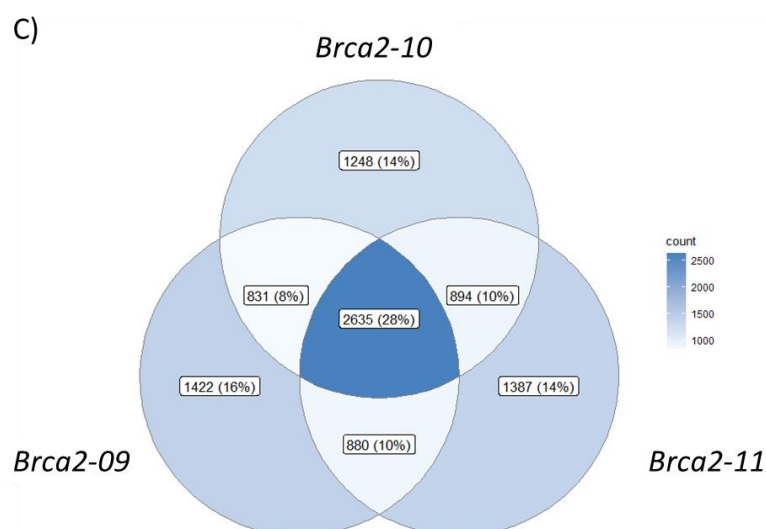


Figure 4.10; Venn diagrams generated through the R package ggplot2 of unique: A) variants called and B) genes with at least one mutation in *Brca2-09*, *Brca2-10*, *Brca2-11* and Wild Type; C) unique variants unique to *Brca2* knockouts.

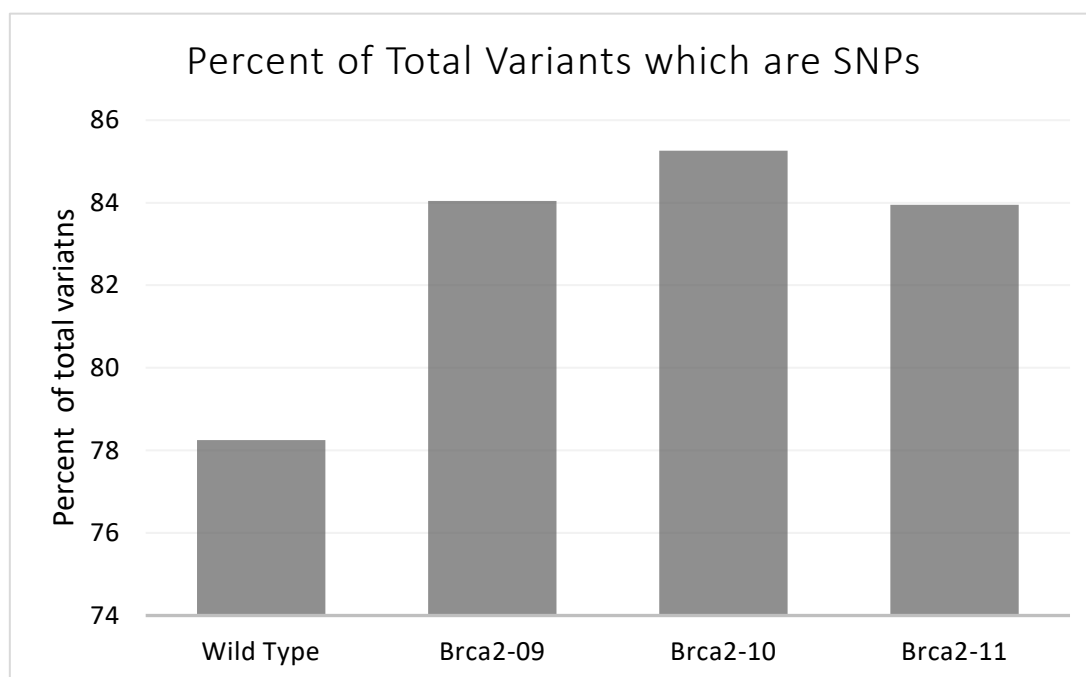


Figure 4.11. Percent of total unique variants which are SNPs.

An overall increase in mutations due to loss of *Brca2* function has been reported in a variety of model organisms, but with a primary focus on multicellular organisms due to the clinical importance of *Brca2*'s role in tumour development (Zamborszky et al., 2017; Samstein et al., 2021). Interestingly, 2,635 (28%) of variants only found in *Brca2* cultures are

shared between all three, suggesting that there are conserved mutational signatures due to the loss of *Brca2* function, but there are also strain-specific mutations. The shared variants were comprised primarily of synonymous variants or missense mutations. The most frequent 'high' impact mutation across all three cultures were frameshift mutations. This is expected when the HR pathway is disrupted. Without the HR pathway, the only other pathway to repair DNA DSBs is the NHEJ pathway. Frameshift mutations were caused by INDELS within the protein-coding sequence of genes. This has been observed when the HR pathway is disrupted leading to increased repair via the error prone NHEJ pathway which causes INDELS. From the overview of mutations, it is apparent that there was an increase in spontaneous mutations of all types across the genome without any significant increase in a specific type of mutation or location within the genome.

#### 4.2.6.4 Copy Number Variation

As discussed in the introduction section on the role of DNA repair in evolution and adaptive response, aneuploidy is a potential mechanism to create genetic diversity through redundancy. Sequencing of the first polar diatom, *F. cylindrus*, revealed a triploid genome which enables genomic and transcriptomic plasticity through genomic redundancy (Mock et al., 2017). By the time the three *Brca2*  $-/-$  cell lines were sequenced, they had returned to phenotypic fitness to that of the wild type with similar growth rates and cell size. This was achieved without repairing the *Brca2* gene. Currently, this has not been studied or observed in any organisms in which *Brca2* has been knocked out through genome editing. The question remained, how were they able to overcome the loss of a core DNA repair enzyme which had significant consequences on their fitness and morphology since no significant mutational hotspots or specific mutation types were identified?

The copy number variation (CNV) of all genes was investigated to understand if A) *T. pseudonana* had increased or decreased CNV of individual genes and if B) the affected genes potentially compensate for loss of *Brca2* to maintain genome integrity. CNV is a general term which describes repeated regions of genomic sequences and has been reported to range from seemingly no effect on physiological traits to impacting morphological variation (Wright et al., 2009; Zhang et al., 2015; Henkel et al., 2019; Pös et al., 2021). CNVs was determined by extracting the depth of sequencing for all protein coding genes and dividing this by the average sequencing depth of the chromosome it is located. The resulting ratios

were then compared to the wild type to control for any possible CNVs unique to our lab's *T. pseudonana* culture. All three *Brca2*  $-/-$  cultures showed a similar pattern of gains and losses of sequencing depth at affected genes (Figure 4.12).

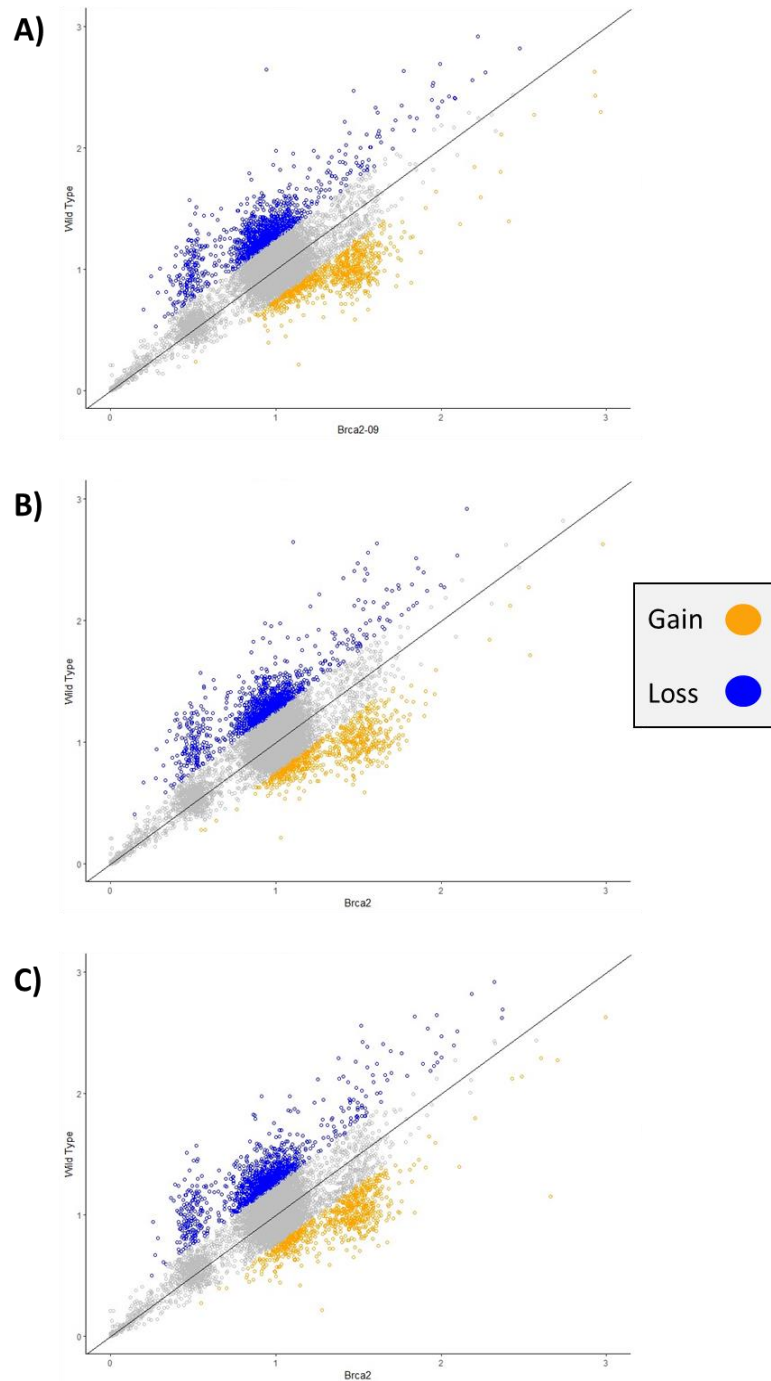


Figure 4.12. Copy number variation (CNV) within coding regions of three *Brca2*  $-/-$  cultures A) *Brca2*-09, B) *Brca2*-10 and C) *Brca2*-11. Each point represents the ratio of depth in the wild type (y-axis) to the ratio of depth of the corresponding *Brca2* culture (x-axis) of a single gene.

Interestingly, the same pattern of gains and losses in CNV was observed in wild-type *T. pseudonana* grown under high-temperature stress (Toseland, unpublished). These data complement the findings of CNV in the *Brca2* *-/-* cultures, as high temperatures are known to increase spontaneous mutation rates (Muller, 1928; Zamenhof and Greer, 1958; Waldvogel and Pfenninger, 2021). The appearance of CNV in wild-type cultures exposed to high temperatures and *Brca2* *-/-* strains suggests that altering CNV is a conserved mechanism used by *T. pseudonana* to adapt to stressful environments. And since *Brca2* *-/-* strains were unable to survive at high temperatures (Section 4.2.4.2), this suggests that DNA repair, specifically homologous recombination, is part of the mechanism by which it adapts to stressful environments.

#### 4.2.6.4.1 GO Analysis of CNV Effected Genes in *Brca2-09*

Genes which had an increase or decrease in read depth, or CNV, were submitted for GO term analysis to elucidate if certain cellular processes were significantly selected. The tool gProfiler (Kolberg et al., 2023) was used as described in Chapter 2, with a significance cut-off of an adjusted p-value of < 0.05. Most interestingly the genes which gained copy numbers in the culture *Brca2-09* are significantly enriched for genes involved in DNA repair (Table 4.6).

Table 4.6. Enriched GO terms of genes with a gain in CNV for each *Brca2* *-/-* culture.

Culture	Enriched GO Terms	GO Term ID	Adjusted p-value	Number of Genes
<i>Brca2-09</i>	DNA integrity checkpoint signalling	GO:0031570	8.41E-05	7
	mitotic DNA integrity checkpoint signalling	GO:0044774	3.51E-04	6
	cell cycle process	GO:0022402	4.23E-04	17
	mitotic cell cycle	GO:0000278	1.51E-03	13
	mitotic cell cycle process	GO:1903047	1.58E-03	12
	DNA binding	GO:0003677	4.74E-03	51
	cell cycle	GO:0007049	4.96E-03	18
	cell cycle checkpoint signalling	GO:0000075	1.46E-02	7
	negative regulation of cell cycle phase transition	GO:1901988	1.46E-02	7
	nucleic acid binding	GO:0003676	1.58E-02	89
	regulation of cell cycle phase transition	GO:1901987	2.89E-02	8
	negative regulation of cell cycle process	GO:0010948	3.21E-02	7
	negative regulation of cell cycle	GO:0045786	3.21E-02	7
	cell cycle phase transition	GO:0044770	3.90E-02	8
<i>Brca2-10</i>	cellular component organization	GO:0016043	2.31E-02	39
	organelle	GO:0043226	2.68E-02	110
	cilium	GO:0005929	4.23E-02	7
<i>Brca2-11</i>	Phosphatidylinositol signalling system	KEGG:04070	2.45E-02	6

The enrichment of genes involved in DNA repair, cell cycle processes and DNA binding in *Brca2-09* suggests that these genes were under selection to compensate for the loss of *Brca2* function. Many of the enriched proteins overlapped in their GO term annotation so duplicated protein IDs were discarded, resulting in a list of 102 individual genes. The annotations of the genes available were retrieved through both the *T. pseudonana* genome page (<https://mycocosm.jgi.doe.gov/Thaps3/Thaps3.home.html>) and searches against Ensembl databases (<https://www.ensembl.org/index.html>) through R using the package gprofiler2/0.2.2 (Kolberg et al., 2020). The vast majority were predicted and hypothetical proteins, comprising 75% of proteins. Genes with annotations are in Appendix Table A.4.

Interestingly, the vast majority are transcription factors which have been reported to accumulate mutations faster than the genes they target (Ali and Seshasayee, 2020). By increasing the copy number of transcription factors, the regulation of many genes is potentially altered to benefit the overall fitness of the cell. Core DNA repair genes from other pathways also showed an increase in copy number, including mismatch repair ATPase (MSH5) and DNA mismatch repair protein MLH3 of the mismatch DNA repair (MMR) pathway, replication factor C (RFC1) involved in base excision repair (BER) and enzymes related to apurinic/apyrimidinic endonucleases which perform the initial steps of BER. If these were the only proteins enriched it would appear that the cells are offsetting the loss of the HR repair pathway with alternative pathways, however, there are many other genes enriched, primarily transcription factors. This resembles the study discussed in Chapter 3 section 3.1.2, in which *E. coli* strains evolved under severe DNA damage stress (Byrne et al., 2014). Modification of DNA repair proteins' sequence and copy number was present alongside a multitude of other genomic changes suggesting that adaption to increased mutation rate does not depend on DNA repair alone, but it is part of a larger strategy by the cell to adapt to mutations driven by DNA damage.

#### 4.2.6.5 Allele Balance

Copy number variation results from an increase or decrease of copies of genes, which can correlate to allelic imbalance. Allele balance is a term used to describe the relative abundance of two different alleles at a particular locus in a diploid organism (Fletcher et al., 2022). Allelic imbalance is generated by aneuploidy – the presence of an abnormal number



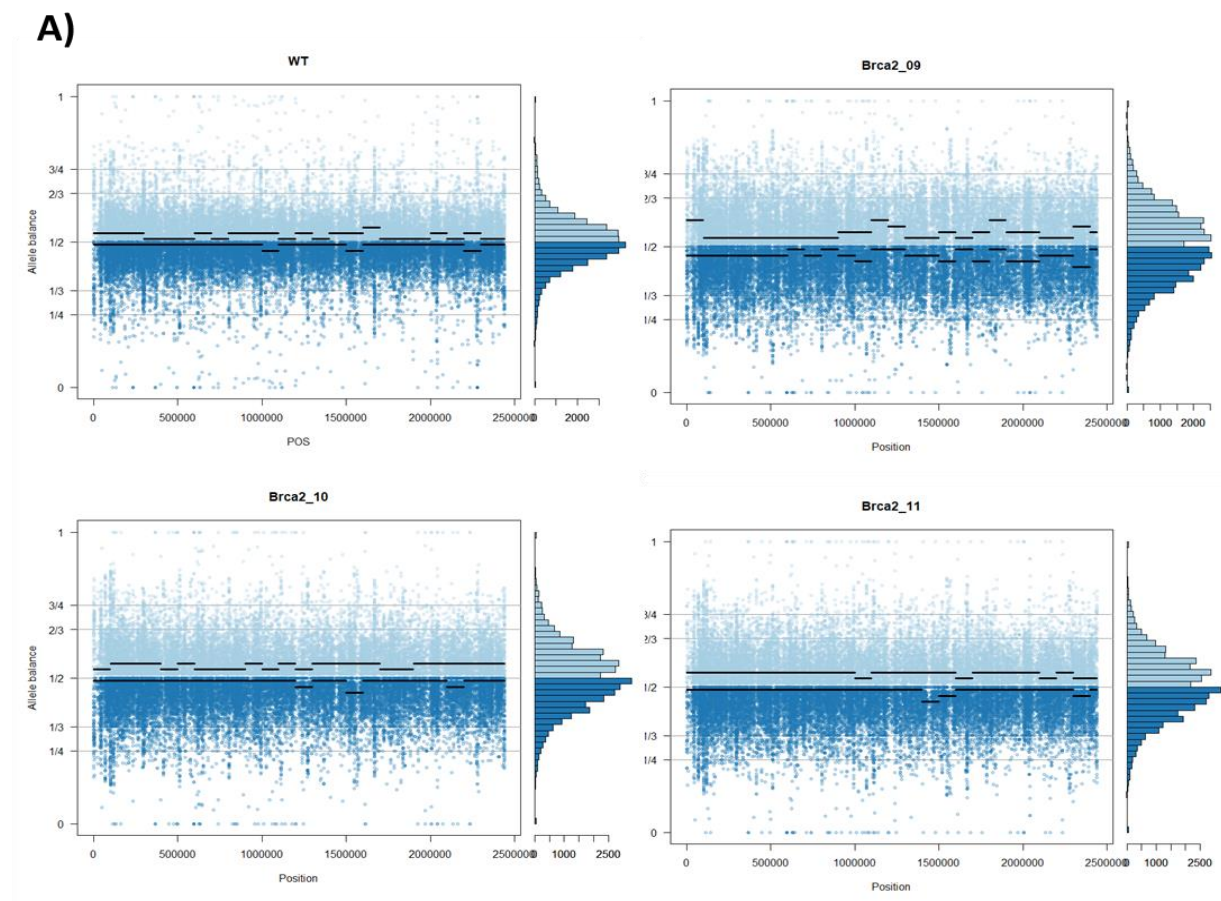
of chromosomes, which can be caused by alterations to recombination pathways (Hassold et al., 2007). As mentioned before, the HR repair pathway had been disrupted in the *Brca2*  $-/-$  mutant *T. pseudonana* cell lines. Significant disruption of recombination is known to increase the frequency of chromosome nondisjunction during meiosis in humans leading to aneuploidy (Lamb et al., 2005). As there are no studies investigating the role of recombination in genome integrity in diatoms, the methods described in Chapter 2 were used to investigate the allele balance of *Brca2*  $-/-$  cell lines.

The allele balance data was gathered from each site in the variant call files, with a minor allele frequency cut-off of 5%, meaning the allele with less depth had to at least account for 5% of all depth. At each variant site, the ploidy was calculated by the ratio of read depth of both alleles. A 1/1 ratio is expected in diploid organisms. If the major allele has a ratio of 2/3 and the minor 1/3, it is considered to be triploid, with one allele having 2 times more depth than the other. There were four classes of results, 1) all cultures had the same ploidy as the wild type, 2) ploidy change was common in all *Brca2*  $-/-$  cultures, 3) ploidy change was strain specific and 4) lack of sufficient data points to determine ploidy (Table 4.7). Trisomy refers to duplication of sequences in a diploid organism. Table 4.7 displays the extent of trisomy in affected chromosomes: 1, 10, 13, 15, 16b, 16a and 19c\_19, the remaining chromosomes had showed no evidence of trisomy.

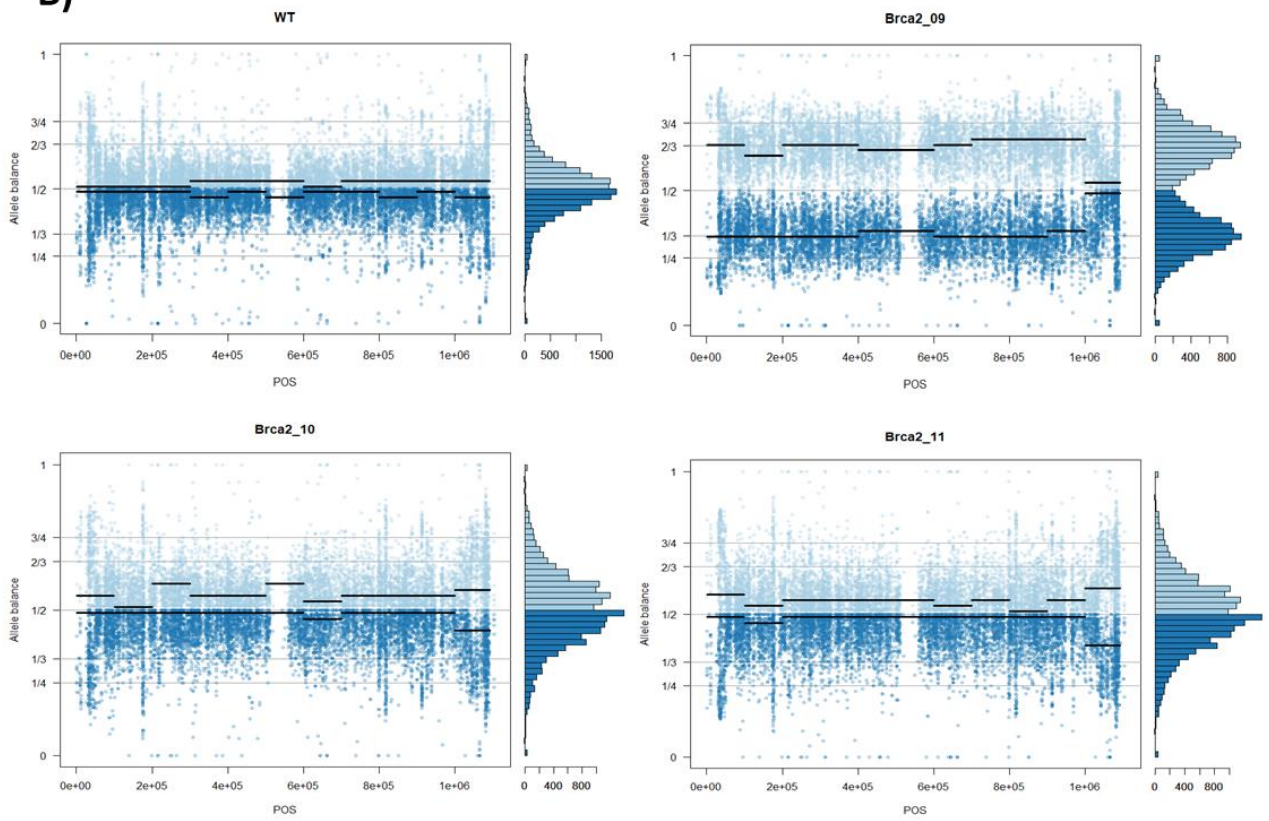
Table 4.7. List of chromosomes displaying trisomy in *Brca2*  $-/-$  strains. Whole chromosome means trisomy is across entire chromosome resulting in complete duplication. Partial means that a region of the chromosome was duplicated.

Chromosome	Wild Type	<i>Brca2-09</i>	<i>Brca2-10</i>	<i>Brca2-11</i>
16b	Diploid	Whole chromosome	Whole chromosome	Whole chromosome
16a	Diploid	Whole chromosome	Diploid	Partial
10	Diploid	Whole chromosome	Diploid	Diploid
1	Diploid	Diploid	Partial	Diploid
19c_19	Diploid	Diploid	Partial	Diploid
15	Diploid	Diploid	Diploid	Whole chromosome
13	Diploid	Diploid	Diploid	Partial
22	Diploid	Diploid	Diploid	Partial

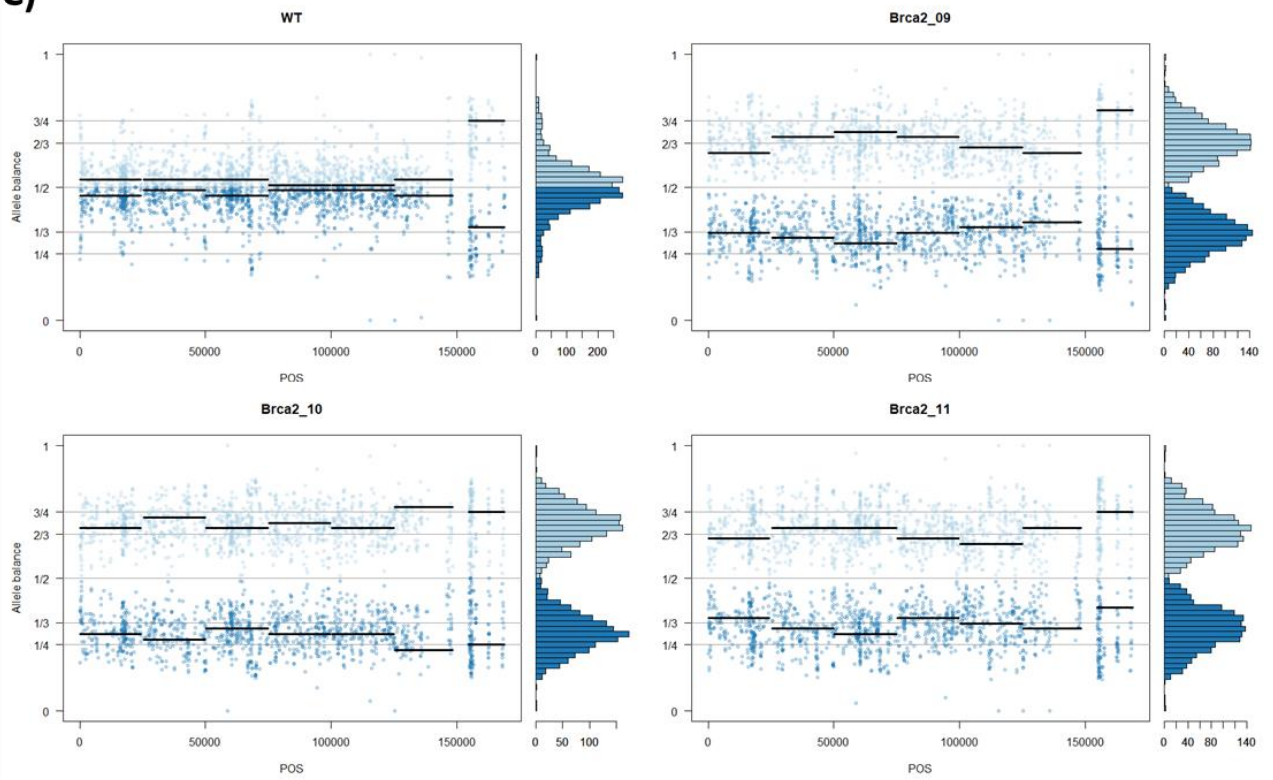
These results are visualised through allele balance plots (Figure 4.13) generated using the R package *vcfR* (Knaus and Grünwald, 2017). These figures display the depth of the major and minor alleles at each locus as points, and the histograms show the distribution of the data. For the majority of chromosomes, the *Brca2*  $-/-$  cultures were the same as the wild type, diploid (Figure 4.13a). The only chromosome that showed a ploidy change across all *Brca2*  $-/-$  cultures was chromosome 16b (Figure 4.13b). Chromosomes 16a, 10, 1, 19c\_19, 15, 13 and 22 were all affected by strain-specific changes in ploidy, either of the whole chromosome or regions (Figure 4.13c). Lastly, Figure 4.13d shows an example of chromosomes where there was not enough data to confidently measure ploidy in any of the data sets, including wild type.



B)



C)



D)

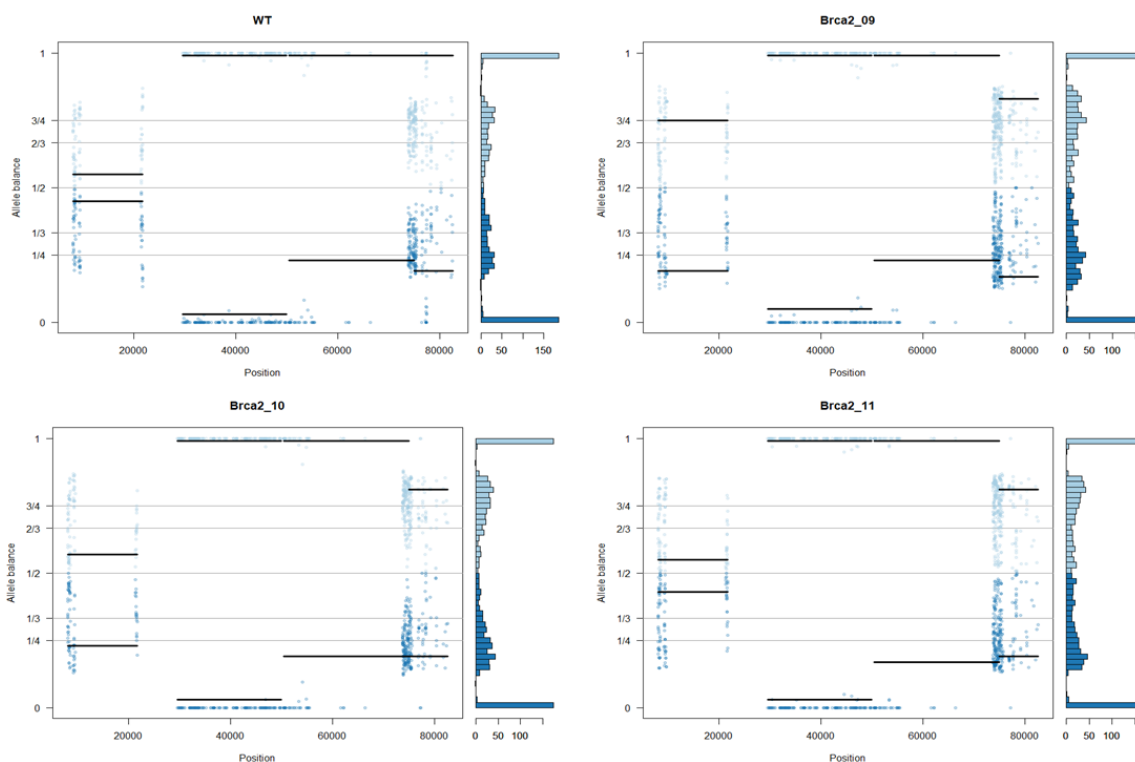


Figure 4.13. Allele balance plots generated using the R package *vcfR* (Knaus and Grünwald, 2017). All figures show the allele balance on the x-axis and the position of the chromosome on the y-axis. Each point represents the allele frequency with a window average (black line). The histograms show the distribution of allele balance; one peak = diploid and two peaks = triploid. A) All cultures show diploid signature with allele balance 1/2 for chromosome 3, B) strain-specific triploid signature in *Brca2*-09 of chromosome 10, C) triploid signature common across all *Brca2*<sup>-/-</sup> cultures in chromosome 16b and D) example of chromosome where determination of ploidy is unachievable due to lack of data points.

Similar to the CNV data, aneuploidy was also seen in wild-type *T. pseudonana* cultures grown under high-temperature stress for over 300 generations (Toseland, unpublished), however, it is more frequent in *Brca2*<sup>-/-</sup> cultures occurring in a total of 8 different chromosomes, 4 of which the whole chromosome was affected with only specific regions duplicated in the other 4.

Trisomy in chromosome 16b was the only instance common across all *Brca2*<sup>-/-</sup> strains. All 52 genes on chromosome 16b were subjected to a GO enrichment analysis. Results showed that two molecular functions and one biological process GO terms were enriched – mannosyltransferase activity (GO:0000030), alpha-1,6-mannosyltransferase activity (GO:0000009) and mannosylation (GO:0097502). Mannosylation is a type of

glycosylation: the process of attaching a carbohydrate to the hydroxyl or other functional group of a target macromolecule, usually proteins and lipids (Varki, 2017). Mannosylation, specifically, is the process of adding a mannose sugar to a tryptophan residue, resulting in a carbon-carbon bond between the second carbon of the tryptophan and the first carbon of the alpha-mannose sugar (Shcherbakova et al., 2019). These proteins act as posttranslational modifiers of proteins and lipids. They play a major role in modifying secreted and cell-surface proteins as well as influencing the folding of proteins and their subsequent ability to interact with macromolecules (Shcherbakova et al., 2019). While the number of these proteins on chromosome 16b is enriched compared to all proteins in the genome, more research is needed to confirm if aneuploidy of chromosome 16b confers a benefit in adaption under mutational stress.

### 4.3 Discussion

This is the first study to impair a DNA repair pathway through CRISPR/Cas9 genome editing in diatoms. The first research question was to confirm the function of *Brca2* in diatoms. *Brca2* is a conserved eukaryotic gene which is critical in the DNA repair pathway, homologous recombination (HR), in which it facilitates the removal and replacement of replication protein A (RPA) from the single-stranded DNA (ssDNA) with RAD51 monomers creating a RAD51-ssDNA complex. The function of *Brca2* has been the subject of intense research in the clinical field as it is a tumour suppressor (Wooster et al., 1995; Shahid et al., 2014), however, its function in phytoplankton has never been confirmed through functional genomics. The removal of a portion of the BRCA2-Helical domain (PF09169) in *T. pseudonana* via CRISPR/Cas9 generated mutant cell lines which were hypersensitive to induced DNA damage. This confirmed, as reported in other eukaryotes, that *Brca2* is critical to the homologous recombination DNA repair pathway in *T. pseudonana*. Given the level of conservation of *Brca2* in eukaryotes, this result was not unexpected, and it provides the first insight into the DNA repair mechanisms of diatoms. Further research is needed to confirm the roles of other DNA genes, for example, *Rad52*. *Rad52* is of prokaryotic origin and has the same function as *Brca2*, but the loss of *Rad52* function in eukaryotes does not result in any phenotypic changes, indicating its redundancy by *Brca2* (Martin et al., 2005; Rossi et al., 2021). Further studies are needed to elucidate the role of *Brca2* in the mitotic division of

diatoms through *in vitro* assays, proteomics and transforming labelled *Brca2* genes for analysis through microscopy.

With the function of *Brca2* confirmed, the question remained, does this function contribute to the adaptive potential of *T. pseudonana*? The adaptive ability of diatoms is evident as they are found in almost every aquatic environment, but the mechanisms behind this ability are not well understood (Malviya et al., 2016). The primary life cycle stage of diatoms is mitotic, and this research explores the hypothesis that DNA repair has a critical role to play in generating genomic diversity. The results generated in this study confirm that the loss of *Brca2* function reduces genomic stability and increases spontaneous mutation rates as seen in other organisms (Zámborszky et al., 2017). *Brca2* function was confirmed to be essential for *T. pseudonana* to adapt to increased temperature, a common environmental variable. Despite this, the changes in the genome are strikingly similar, though more frequent, to those seen in *T. pseudonana* cultures grown under high-temperature stress (Toseland, unpublished). The similarity of genomic mutational signatures suggests that the mechanism of adaptation is increasing copy numbers of genes and inducing aneuploidy when faced with increased mutation rates.

The results presented here showed an increase in aneuploidy in the *Brca2* *-/-* cultures which is also seen in wild type under high temperatures (Toseland, unpublished). This suggests that aneuploidy is a common mechanism used by *T. pseudonana* when exposed to a prolonged increase in mutation rate. Aneuploidy is widespread among unicellular eukaryotes (Yona et al., 2012; Storchova, 2018) and can create burdens on the cell at the cost of providing fitness advantages under stressful growth conditions (Berman, 2016; Ene et al., 2018). Duplicating portions of the genome creates redundancy and allows for mutations in duplicated regions to accumulate while mitigating potential deleterious effects to the organism. Aneuploidy has been reported in diatoms when assembling their genomes but not as a mechanism for adaptation (Von Dassow et al., 2008; Maeda et al., 2022). The increase in frequency of aneuploidy in *Brca2* *-/-* strains compared to wild type under temperature stress suggests that *Brca2* has a role in regulating the extent of duplications or losses of genetic information in diatoms. Loss of *Brca2* function has been attributed as the cause of aneuploidy events due to the nondisjunction and misalignment of chromosomes during mitotic division where *Brca2* function has been disrupted (Ehlén et al.,



2020; Karaayvaz-Yildirim et al., 2020). This confirms the role of *Brca2* in diatoms during mitotic recombination and division. Further research, such as RNA sequencing data to examine the expression of *Brca2* and other DNA repair enzymes in diatoms under environmental stress would increase our understanding of the role of these proteins in adaptation. These data are necessary as transcription of *Brca2* might be suppressed under stressful conditions to relax DNA repair and allow spontaneous mutations, but also increase the frequency of aneuploidy events.



## 5 Short-term Experimental Evolution of *Brca2* -/- *T. pseudonana*

### 5.1 Introduction

#### 5.1.1 Experimental Evolution in Microbial Populations

Experimental evolution explores molecular processes in organisms or populations over time in response to experimental conditions (Kawecki et al., 2012; O'Malley, 2018). By studying changes in genetic structure these projects provide mechanistic underpinnings to complement evolutionary theory. Microbes are powerful tools to answer questions related to evolutionary biology as they have relatively short generation times, can be sampled easily and cryopreserved and it is possible to strictly regulate the environment they are grown (Barrick and Lenski, 2013; Adams and Rosenzweig, 2014). Microbial experimental evolution projects typically start with a culture inoculated into media and grown under normal growth conditions as a control and under external pressures (McDonald, 2019). After the culture has grown to a high population density, it is diluted to allow continuous exponential growth. This can be repeated indefinitely and allows the population to be driven by natural selection to adapt to experimental conditions (McDonald, 2019). The synergy of whole genome sequencing and experimental evolution has provided ways to understand the adaptive significance of genetic variation (Shendure et al., 2005; Hegreness and Kishony, 2007). The expansion of sequencing technologies and reduction in costs has increased the ability to conduct 'Evolve and Resequence' projects (Turner et al., 2011).

#### 5.1.2 Mitotic Recombination in Diatoms and *Brca2*

The BRCA2 DNA repair associated gene (BRCA2) also known as the breast cancer susceptibility type 2 gene, is a tumour suppressor which is linked to breast and ovarian cancer (Wooster et al., 1995). *Brca2* is a core protein in the homology-directed repair pathway (HDR or homologous recombination – HR) which repairs DNA double-strand breaks (DSBs; Venkitaraman, 2014). The loss of *Brca2* function leads to impaired HDR and increases the frequency of DNA lesions repaired by more error-prone pathways such as non-homologous end-joining (NHEJ); increasing the frequency of mutations at sites of repair

(Gorodetska et al., 2019). The relationship between *Brca2* and genomic stability has been extensively studied in yeast and animal models to increase understanding of mutations within *Brca2* and resulting tumours (Spugnesi et al., 2013; Brown et al., 2021). However, the role of *Brca2* in diatoms has not been studied. In phytoplankton, *Brca2*'s role in DNA repair has been studied in *Chlamydomonas reinhardtii* (Ferenczi et al., 2021). This study aimed to reveal the mechanistic underpinnings of single-strand templated repair (SSTR) to enhance the efficiency of precision editing (Ferenczi et al., 2021). They found that the frequency of SSTR was significantly increased in NHEJ-deficient *C. reinhardtii ku70/80* lines suggesting that SSTR is in direct competition with NHEJ for repair. While this study confirmed the function of *Brca2* in green algae, it did not investigate genome-wide changes of knock-out strains grown over time.

As mentioned in the general introduction the number of diatom species has been estimated to be as high as 100,000 (Malviya et al., 2016). Due to the life cycle of diatoms, sexual reproduction can be an irregular occurrence and has never been documented in some species. Without regular meiosis, it is not clearly understood how diatoms can generate the genetic variation observed. One mechanism that can create novel genetic variation through the exchange of genetic sequences within the genome is mitotic recombination (Bulankova et al., 2021). The primary role of mitotic recombination (homologous recombination; HR) is to repair DNA double-strand breaks (DSBs) using homologous regions of a sister chromatid during replication (Li and Heyer, 2008; Krejci et al., 2012). Mitotic recombination can provide frequent opportunities to generate genetic diversity. The current understanding of the role of mitotic recombination and genetic diversity in diatoms was pioneered in *Phaeodactylum tricornerutum* (Bulankova et al., 2021). This study reported that *P. tricornerutum* and *Seminavis robusta* accumulated recombined haplotypes across the entire genome (Bulankova et al., 2021). This can potentially lead to novel allele combinations and generate potentially beneficial adaptations through the 'unmasking' of recessive alleles.

The studies mentioned above have significantly increased the understanding of mitotic recombination and genetic diversity. However, the use of genome editing to study the loss of a key protein in the mitotic recombination pathway and the subsequent consequences on genome stability over time has not yet been studied in diatoms. In

Chapter 4, three independent *Brca2* *-/-* *T. pseudonana* strains were sequenced which revealed the reduction of genome stability when HR was impaired. However, without multiple time points and accurate data on the number of generations, mutation rate and timescales for changes in genomic structure were unable to be answered. Therefore, a short experimental evolution project was designed using *Brca2* *-/-* and wild-type *T. pseudonana* strains. To understand the initial stages of adaptation to the loss of *Brca2*, newly transformed *Brca2* *-/-* strains were generated using the same methods and materials used in earlier transformations. The previously sequenced *Brca2* *-/-* strains from Chapter 4, had already undergone significant changes in both genome and morphology resulting in cellular fitness comparable to that of the wild-type. Using newly transformed strains meant generated data would provide insights into the initial mechanisms of adaptation to the loss of *Brca2* function. Results from a previous experimental evolution project with *T. pseudonana* reported that the rate of mutations was highest within the first 40 – 50 generations and then plateaued (Schmidt and Toseland, unpublished).

This chapter reports and discusses the results of a short-term experimental evolution experiment using of two *T. pseudonana* strains: wild-type and *Brca2* homozygous knockouts (*Brca2* *-/-*). These cultures were grown under two separate conditions: intermittent treatment with the DNA-damaging agent methyl methanesulfonate (MMS) and mock treatment with the drug vehicle, dimethyl sulfoxide (DMSO). After ~20 and ~40 generations, DNA was extracted and sent for whole-genome resequencing at the Earlham Institute. Whole-genome resequencing of exponentially growing *T. pseudonana* strains deficient in HR provided insights into the role of *Brca2* in maintaining genomic integrity. To my knowledge, this is the first study with a diploid marine diatom deficient in HR which implemented whole-genome analysis to elucidate the molecular underpinnings of the role of recombination over time with or without induced stress.

*T. pseudonana* was chosen since there are published methods of transformation, the availability of a reference genome and its short generation times (Armbrust et al., 2004; Hopes et al., 2017). Novel mutations and changes in genomic structure were compared between wild-type and *T. pseudonana* strains carrying a homozygous knock-out of *Brca2* (*Brca2* *-/-*) under MMS stress and mock treatment. Results from Chapter 4 directed the analysis of this chapter to understand the initial genomic changes over the first ~40

generations. For example, the *Brca2* *-/-* strains in Chapter 4 showed evidence of genomic instability through increased copy number variation (CNV) and events of partial and whole chromosome trisomy. By re-sequencing sooner after transformation, these data help to understand the early stages of adaption to the loss of HR in a marine diatom.

The following hypotheses were tested through the experiment described: 1) loss of *Brca2* function will result in increased spontaneous mutation rate, 2) treatment with MMS will disproportionately affect the genome of *Brca2* *-/-* strains compared with wild-type, and 3) loss of *Brca2* function leads to increased genome instability.

## 5.2 Results

### 5.2.1 Growth Data

All cultures were kept in constant exponential growth under regular growth conditions (Chapter 2 section 2.1.1.1) to mitigate potential effects on the genome under stationary phase growth. Before the project was started all cultures were treated with antibiotics to remove as much bacteria as possible and were confirmed axenic under fluorescent microscopy using the methods in Chapter 2 section 2.3.3. The growth of both the *Brca2* *-/-* and wild-type *T. pseudonana* strains was characterised before starting the experimental evolution study to ensure growth phases were known. Both wild-type and *Brca2* *-/-* strains were either treated with the DNA-damaging agent MMS or given a mock treatment of DMSO (MMS vehicle) weekly. Cultures were subbed into fresh media in between treatments and directly after treatments to ensure they maintained exponential growth. The maximum potential quantum efficiency of Photosystem II (Fv/Fm) and cell counts were conducted between and directly before treatments (Appendix Table A.5; Figure 5.1).

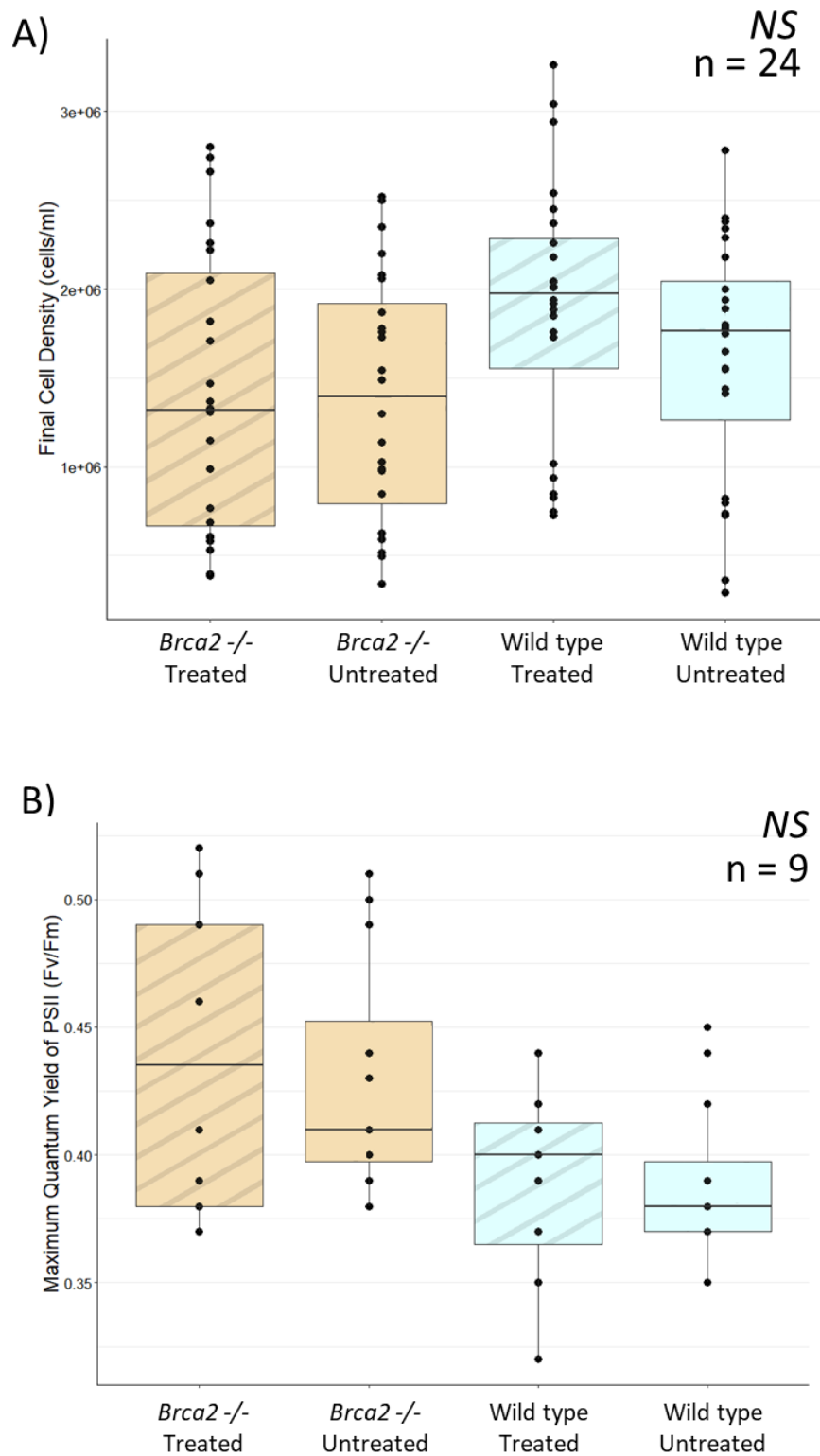


Figure 5.1. Growth data collected during diluting cultures during experimental evolution experiment. *Brca2* -/- (light yellow), wild type (blue), MMS treated (dashed boxes), DMSO mock treatment (solid boxes) A) Final cell density of cultures at each transfer; the time between inoculation and transfer 3-4 days. B) Fv/Fm of cultures at subs between treatments with MMS or DMS, no Fv/Fm was taken immediately before treatment. NS = no significance, n = number of samples.

The total number of generations was calculated using methods stated in Chapter 2. Each timepoint was roughly 20 generations of growth for all samples (Appendix Table A.6; Figure 5.2). Across both timepoints, the wild-type strains grew quicker than the *Brca2*  $-/-$  strains regardless of whether they received MMS treatment or mock treatment (DMSO).

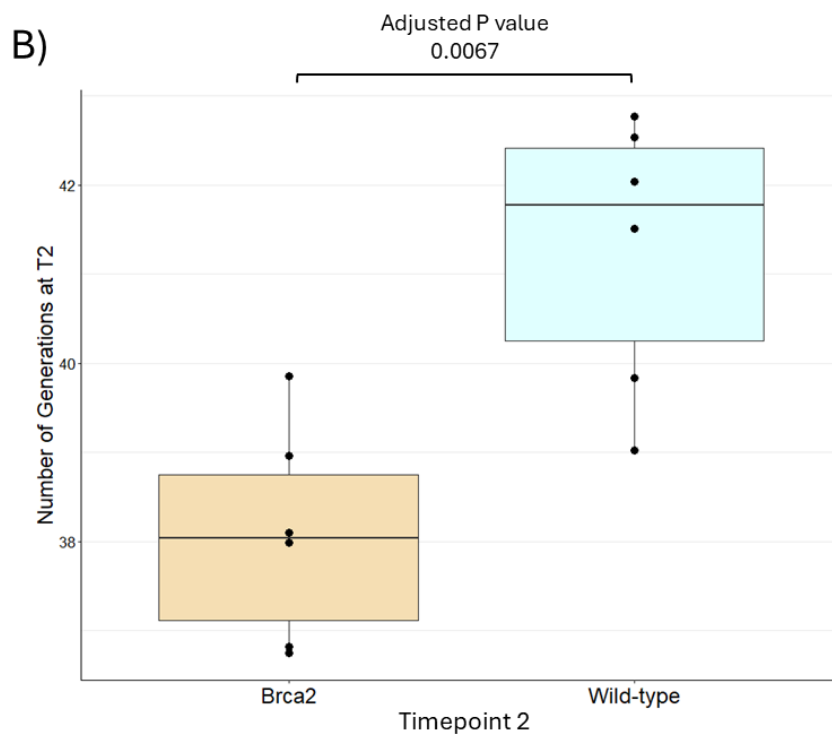
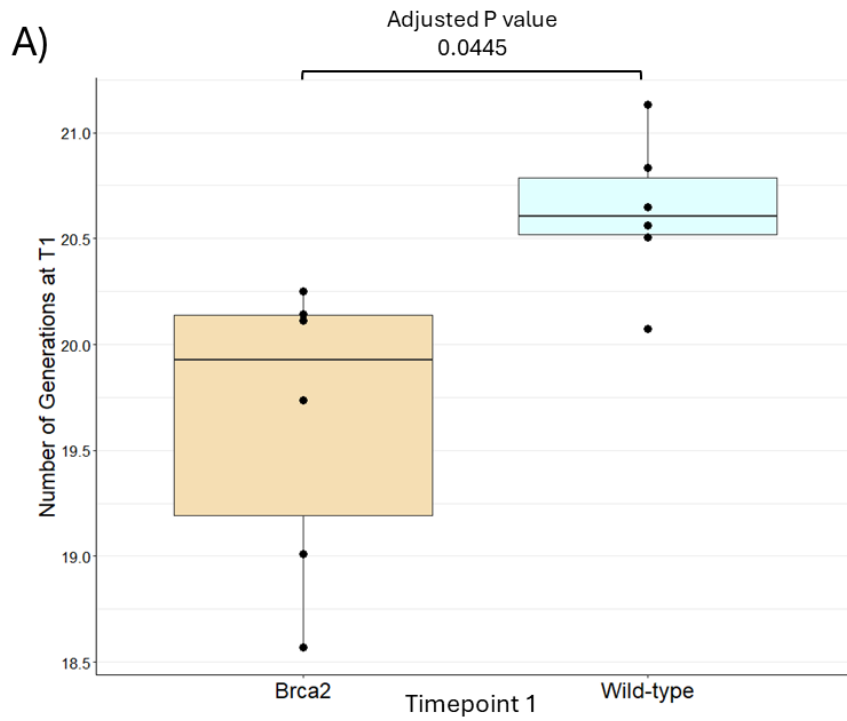




Figure 5.2. Total number of generations at each timepoint grouped by culture. A) Total generations at timepoint 1 (T1); B) total generations at timepoint 2 (T2). P values were calculated using a paired t-test assuming equal variance and adjusted using the Bonferroni method in R. Data for *Brca2* cultures is in light yellow boxes and wild type is in light blue boxes.

## 5.2.2 DNA Extraction, Whole-Genome Sequencing and Mapping to Reference Genome

Samples were taken via filtration onto sterile filters, snap-frozen on dry ice and stored at  $-80^{\circ}\text{C}$  until extraction. Samples were taken at timepoint 0 (T0) the day before the experiment was started, timepoint 1 (T1) after two weeks and timepoint 2 (T2) after four weeks. Treatment with either MMS (Treated/MMS) or DMSO (Untreated/Mock) was conducted every 7 days after the start of the experiment. DNA was extracted using the methods described in Chapter 2 section 2.4. Extracted gDNA was sent to the Earlham Institute where the libraries were prepared via their LITE Library Prep protocol and sequenced on 1 lane of the NovaSeq 6000 SP v 1.5 flow cell with 150 paired-end (PE) reads (Perez-Sepulveda et al., 2021). Appendix Table A.7 shows the results of the raw genetic data generated by the Earlham Institute. All samples were successfully sequenced except the sample for treated wild-type replicate B (Wild-type\_B\_MMS\_T1) at timepoint 1 and was discarded for all analysis. It appears this sample may have failed in either initial DNA extraction or library preparation, as the raw data generated had similar mapping statistics to other samples, but at a significantly reduced mean coverage (Appendix Table A.8).

The raw reads were quality-checked with FastQC and universal Illumina adapters were trimmed using trimmomatic before mapping to the reference *T. pseudonana* genome using the Burrows-Wheeler alignment tool (BWA-MEM; Armbrust et al., 2004; Andrews, 2015; Bolger et al., 2014). Once aligned, duplicate reads were marked and removed using Picard's tool 'markduplicates' (<http://broadinstitute.github.io/picard/>). Final alignment files were quality-checked using QualiMap BamQC (Appendix Table A.8; Okonechnikov et al., 2015).

### 5.2.2.1 Treated *Brca2* $-/-$ Replicate A at Timepoint 1 Has a Wild-type Copy of *Brca2*

Before calling mutations, each alignment file was visualised in the Integrated Genome Viewer at the *Brca2* locus on chromosome 7:427471 – 428442 to confirm that wild-

type samples had the full *Brca2* locus and *Brca2*<sup>-/-</sup> samples were edited (Robinson et al., 2011). The sample for treated *Brca2*<sup>-/-</sup> replicate A at timepoint 1 (*Brca2*<sub>A\_MMS\_T1</sub>) had the wild-type *Brca2* locus (Figure 5.3). It is possible that during sample preparation or library preparation, this sample was mislabelled, but this could not be confirmed. The sample was discarded as a *Brca2*<sup>-/-</sup> culture and was annotated separately from the other samples as it could not be determined which culture or replicate it came from.

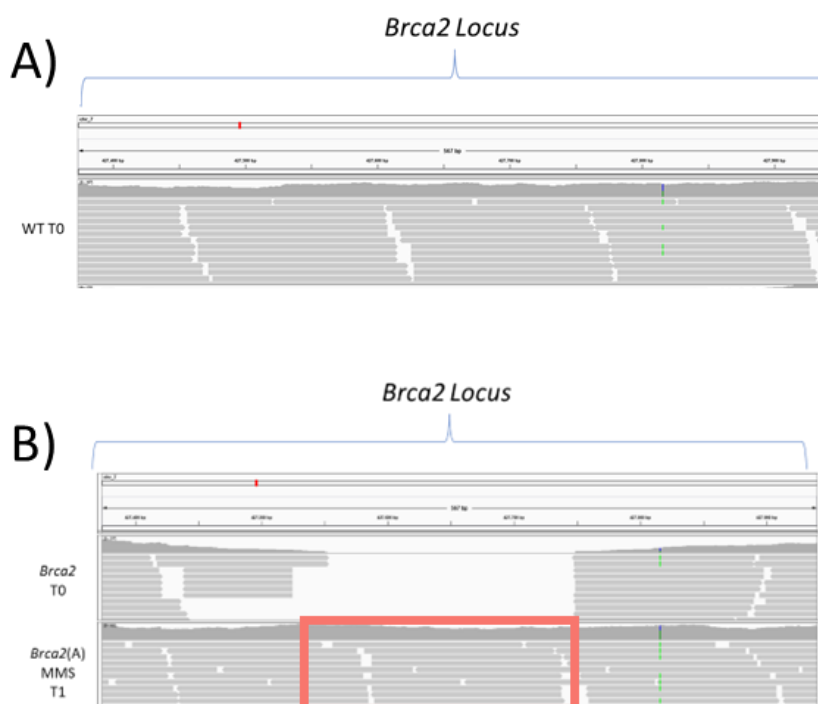


Figure 5.3. Alignments of raw reads to reference genome visualised using the Integrated Genome Viewer. Each track is at the coordinates of the CRISPR cut site in the *Brca2* locus (Chr\_7:427,375-427,943) Each track is labelled on the left with the corresponding sample ID. A) Wild-type T0; B) *Brca2* T0 and treated replicate A at T1. The red box highlights the region between the PAM recognition sites which should be removed in *Brca2*<sup>-/-</sup> cultures.

### 5.2.3 Filtering and Overview of Mutations

Using the methods described in Chapter 2 section 2.5.3.3, variants were called using bcftools call and annotated through SnpEff to describe their location with respect to coding regions and predict their impact on the genome (Cingolani et al., 2012; Danecek et al., 2021). All final variant call files (.vcfs) contained >185,000 called mutations (Appendix Table A.9). Finally, variants were filtered for a minor allele support greater than 5%. The majority of mutations were found in the re-sequenced wild-type culture from our lab and were

removed from further analysis so that only novel mutations for each sample were kept. The most common mutations were SNPs followed by INDELs. Appendix Table A.9 contains an overview of the number of variants and types for each sample before and after filtering. Generally, there was a higher number of mutations of all types found in *Brca2*  $-/-$  strains compared with the wild-type (Figure 5.4). Before the filtration of variants, there was no significant difference between the number of variants between *Brca2*  $-/-$  strains and wild-type. After filtering variants, there were significantly more ( $P=0.00142$ ) variants in *Brca2*  $-/-$  strains compared to wild-type.

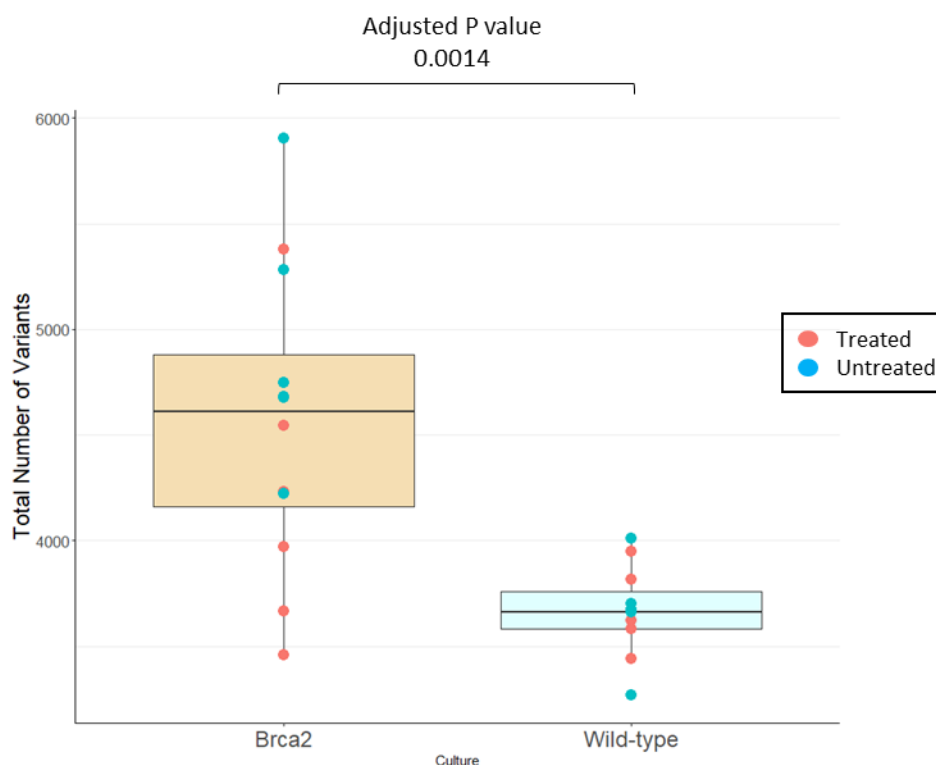


Figure 5.4; Number of all variants called after filtering grouped by *Brca2*  $-/-$  (light yellow) and wild-type (light blue) strains. Points are coloured by treatment of samples with red dots = MMS-treated samples and blue points = mock-treated (DMSO; untreated) samples. P values shown are adjusted P values and were calculated via a pairwise t-test and adjusted using the Bonferroni method in R.

### 5.2.3.1 Mutation Rates

The global mutation rate – the frequency of novel mutations across the genome – was calculated for each sample using the equation in Chapter 2. Results from Chapter 4 showed that knocking out the function of *Brca2* increased the number of spontaneous mutations, however, without data on the number of generations between knock-out and sequencing, mutation rates were unable to be calculated. In this project, each replicate was

sequenced around every 20 generations. To calculate the mutation rates, all mutations found in the T0 wild-type sequencing data were removed leaving only novel mutations. Mutation rates were generally higher in *Brca2* knock-out cultures compared with wild-type (Table 5.1). Interestingly the treatment had little effect on mutation rates.

Table 5.1. Number of SNPs, number of generations and mutation rates of each culture.

Culture	Number of SNPs	Number of Generations	Mutation Rate
<i>Brca2</i> _A_Treated_T1	2536	18.568	4.211E-06
<i>Brca2</i> _A_Treated_T2	3115	36.819	2.608E-06
<i>Brca2</i> _A_Untreated_T1	4779	20.112	7.326E-06
<i>Brca2</i> _A_Untreated_T2	3197	39.850	2.473E-06
<i>Brca2</i> _B_Treated_T1	4236	19.011	6.869E-06
<i>Brca2</i> _B_Treated_T2	2651	38.095	2.145E-06
<i>Brca2</i> _B_Untreated_T1	4154	19.738	6.488E-06
<i>Brca2</i> _B_Untreated_T2	3614	36.742	3.032E-06
<i>Brca2</i> _C_Treated_T1	3442	20.249	5.240E-06
<i>Brca2</i> _C_Treated_T2	2937	38.959	2.324E-06
<i>Brca2</i> _C_Untreated_T1	3534	20.145	5.408E-06
<i>Brca2</i> _C_Untreated_T2	3632	37.982	2.948E-06
Wild-type_A_Treated_T1	2918	20.646	4.357E-06
Wild-type_A_Treated_T2	2327	42.538	1.686E-06
Wild-type_A_Untreated_T1	2731	20.505	4.106E-06
Wild-type_A_Untreated_T2	2814	41.508	2.090E-06
Wild-type_B_Treated_T2	2541	42.765	1.832E-06
Wild-type_B_Untreated_T1	3104	20.564	4.653E-06
Wild-type_B_Untreated_T2	2664	39.836	2.062E-06
Wild-type_C_Treated_T1	2629	42.039	1.928E-06
Wild-type_C_Treated_T2	2608	42.039	1.913E-06
Wild-type_C_Untreated_T1	2726	20.835	4.033E-06
Wild-type_C_Untreated_T2	2719	39.020	2.148E-06

#### 5.2.4 Loss of Heterozygosity

Loss of heterozygosity (LOH) is a signature of genomic instability created by impaired homologous recombination (Stewart et al., 2022). Studies in breast and ovarian cancers have found patterns of LOH associated with impaired HR (Abkevich et al., 2012). Briefly, to determine if *T. pseudonana* strains carrying a homozygous knockout of *Brca2* showed signatures of LOH, all heterozygous variants in the wild-type T0 sample were isolated and

compared to the variants at the same location in all samples. LOH was confirmed if a sample returned a homozygous variant where the wild-type T0 was heterozygous for the same locus. This analysis used variants filtered for minor allele frequency (> 5%), distance from INDELs (>10bp), aligned read depth (>10x), variant quality (>35) and without INDELs to report only SNPs.

There was a total of 177,472 heterozygous wild-type T0 variants. The majority of which were common across all samples (Section 5.2.1.1.1). There were significantly more LOH events in *Brca2* *-/-* samples compared to wild-type samples (Table 5.2; Figure 5.5a). The treatment with MMS increased LOH events in both wild-type and *Brca2* *-/-* strains but only significantly impacted the number of LOH events in *Brca2* *-/-* strains (Figure 5.5b-c). *Brca2* is crucial in DNA repair processes in *T. pseudonana* when exposed to MMS (Chapter 4 section 4.2.4.1), so an increase in LOH events under DNA damage stress suggests that other error-prone DNA repair pathways were used to cope with the increased frequency of DNA insults.

Table 5.2. Number of homozygous SNPs in samples which were heterozygous in wild-type T0.

Culture	Number of Variants HET > HOM
<i>Brca2</i> <i>-/-</i> T0	918
<i>Brca2</i> <i>-/-</i> A Treated T2	912
<i>Brca2</i> <i>-/-</i> A Untreated T1	882
<i>Brca2</i> <i>-/-</i> A Untreated T2	894
<i>Brca2</i> <i>-/-</i> B Treated T1	902
<i>Brca2</i> <i>-/-</i> B Treated T2	923
<i>Brca2</i> <i>-/-</i> B Untreated T1	879
<i>Brca2</i> <i>-/-</i> B Untreated T2	888
<i>Brca2</i> <i>-/-</i> C Treated T1	897
<i>Brca2</i> <i>-/-</i> C Treated T2	899
<i>Brca2</i> <i>-/-</i> C Untreated T1	885
<i>Brca2</i> <i>-/-</i> C Untreated T2	904
Wild-type A Treated T1	83
Wild-type A Treated T2	104
Wild-type A Untreated T1	66
Wild-type A Untreated T2	91
Wild-type B Treated T2	98
Wild-type B Untreated T1	69
Wild-type B Untreated T2	131
Wild-type C Treated T1	64
Wild-type C Treated T2	131
Wild-type C Untreated T1	81
Wild-type C Untreated T2	88

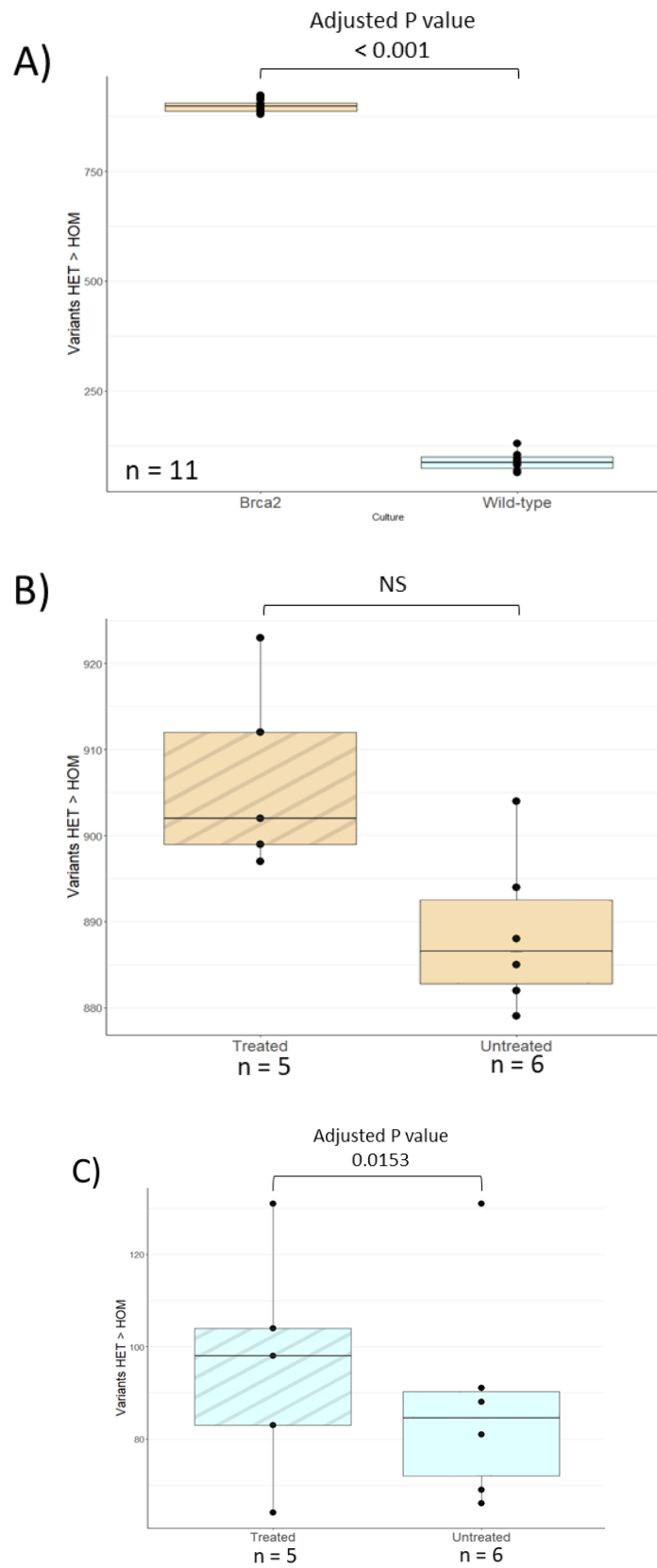
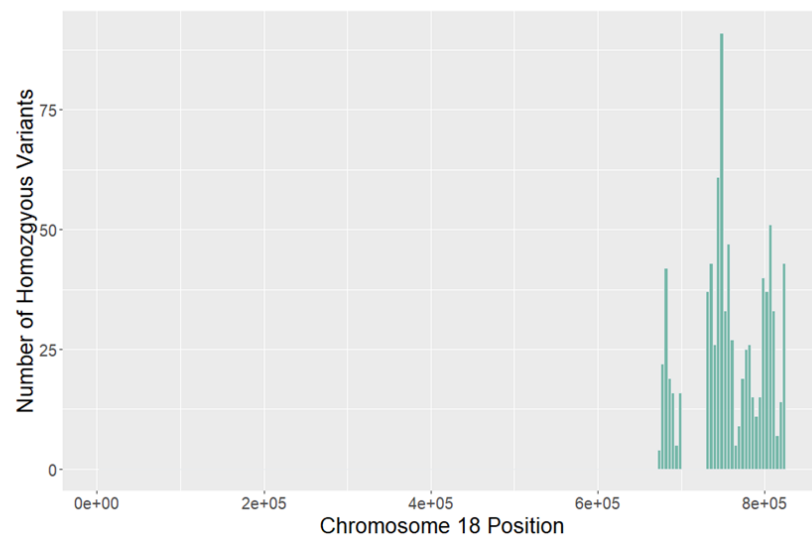


Figure 5.5. Number of homozygous variants per sample which are heterozygous in wild-type T0. N = number of samples. A) All samples grouped by either Brca2 -/- (light yellow) or wild-type (blue) regardless of treatment; B) Brca2 -/- (blue) samples grouped by treated (dashed) or untreated (solid); C) Wild type (blue) samples grouped

by treated (dashed) or untreated (solid). *P* values shown are adjusted *P* values and were calculated via a pairwise *t*-test and adjusted using the Bonferroni method in R. NS = no significance.

LOH events can be categorised into short-range (interstitial), long-range (terminal), and whole-chromosome LOH (Sui et al., 2020; Dutta et al., 2022). The majority of LOH events seen in *Brca2*  $-/-$  strains were present in *Brca2*  $-/-$  T0. Interestingly, 91.7 % of all novel homozygous mutations in *Brca2*  $-/-$  T0 occurred within chromosome 18 and clustered at the end of the long arm (Figure 5.6).

A)



B)

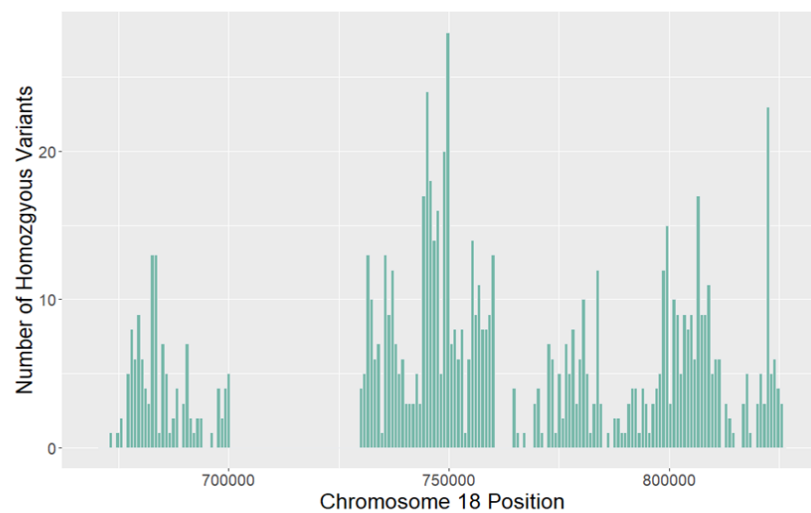


Figure 5.6. Histograms showing the location of novel homozygous SNPs in *Brca2*  $-/-$  T0 on chromosome 18. The x-axis shows the location on chromosome 18 going from the short to long arms of the chromosome.



The clustering of novel homozygous variants on chromosome 18 appears to be separated across two separate regions (673514 – 700025 and 730249 – 825349). Each of the two regions contains 8 and 33 genes respectively with 11 genes in-between that are not affected by LOH. GO term analysis of the genes in this region did not result in any enriched GO terms. All GO terms from this region are in Table 5.3. Despite there being no significantly enriched GO terms, there are more GO terms associated with either DNA (DNA methylation and DNA binding) or translation of RNA. This may have been the result of a beneficial mutation in the early stages of growth in the *Brca2* *-/-* strain used in the experiment. LOH events have been shown to promote the expression of recessive alleles potentially creating new allele combinations resulting in an increase in genetic diversity within the population (Dutta et al., 2022).

Table 5.3. GO terms from two regions at the end of chromosome 18 which are affected by LOH. The numbers below 'Chromosome 18' represent the region of the chromosome (base pairs)

Region	GO Term ID	GO Term Name
Chromosome 18: 673514 – 700025	GO:0006412	translation
	GO:0003676	nucleic acid binding
	GO:0008483	transaminase activity
	GO:0006306	DNA methylation
Chromosome 18: 730249 – 825349	GO:0051726	regulation of cell cycle
	GO:0006412	translation
	GO:0006468	protein phosphorylation
	GO:0006468	protein phosphorylation
	GO:0003700	DNA-binding transcription factor activity
	GO:0003677	DNA binding
	GO:0006886	intracellular protein transport
	GO:0009765	photosynthesis, light harvesting
	GO:0005975	carbohydrate metabolic process
GO:0007155	cell adhesion	

### 5.2.5 Genetic Differentiation ( $F_{ST}$ )

The fixation index ( $F_{ST}$ ) was calculated for all samples using the `--weir-fst-pop` tool which is part of the `vcftools` software using a sliding window of 1kb. This tool uses the Wier and Cockerham method to determine  $F_{ST}$  (Weir and Cockerham, 1984).  $F_{ST}$  is used to summarise the genetic differentiation among distinct groups or populations (Jakobsson et

al., 2013). Sewall Wright introduced the  $F$  statistic as a way of describing the genetic diversity within and between diploid populations (Wright, 1951).  $F_{ST}$  is correlated with the variance of allele frequency among populations. Small  $F_{ST}$  values mean the variance in allele frequency is similar between two populations, whereas large values mean that the allele frequencies are different (Holsinger and Weir, 2009). Each sample was grouped into populations based on their timepoint and treatment (Appendix Table A.10; Danecek et al., 2011). As mentioned earlier, both populations for *Brca2*  $-/-$  and wild-type treated at T1 are missing 1 sample each due to sequencing errors and the presence of wild-type *Brca2* allele respectively.

All populations were compared with the wild-type T0 sample similar to the LOH analysis. Table 5.4 shows the mean and weighted Weir and Cockerham  $F_{ST}$  values calculated for each population. Interestingly, all wild-type populations returned negative  $F_{ST}$  values, while all *Brca2*  $-/-$  populations had positive values. Negative  $F_{ST}$  values are usually reported as zeros and they mean there is more heterozygosity between the populations, whereas positive values mean there is a higher proportion of homozygous alleles describing genetic differentiation. There is a discussion around vcftools and the potential for negative  $F_{ST}$  values, and it is common for the  $F_{ST}$  values to be negative, but reported as zeros, for purposes here the negative values are reported.

Table 5.4. Weir and Cockerham  $F_{ST}$  estimates for each population. Mean  $F_{ST}$  is the average per site  $F_{ST}$ . Weighted  $F_{ST}$  is the sum of  $A_s$  and  $B_s$  for each window size.

Population	Number of Generations (std)	Weir and Cockerham mean $F_{ST}$ Estimate	Weir and Cockerham weighted $F_{ST}$ Estimate
<i>Brca2</i> Treated T1	19.63 (0.62)	0.0029	0.0026
<i>Brca2</i> Treated T2	37.96 (0.88)	0.0031	0.0025
<i>Brca2</i> Untreated T1	19.99 (0.18)	0.0031	0.0025
<i>Brca2</i> Untreated T2	38.19 (1.28)	0.0032	0.0026
Wild-type Treated T1	20.74 (0.09)	-0.0002	-9.66E-05
Wild-type Treated T2	42.45 (0.30)	-0.0001	-4.81E-06
Wild-type Untreated T1	20.38 (0.22)	-0.0001	4.80E-05
Wild-type Untreated T2	40.12 (1.04)	-0.0002	-3.32E-05

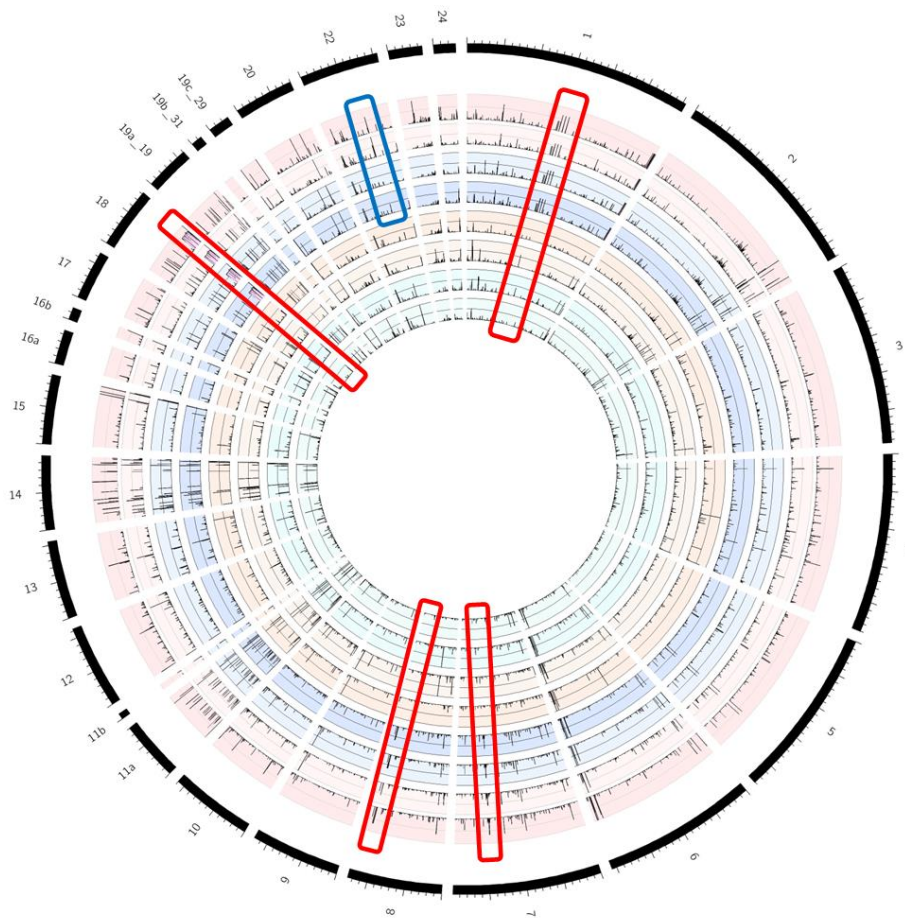


Figure 5.7. Visualisation of  $F_{ST}$  statistic in wild-type and *Brca2*  $-/-$  evolved cultures grouped into populations based on timepoint and treatment. Tracks from the inside out: 1) wild-type Untreated T1; 2) wild-type Untreated T2; 3) wild-type treated T1; 4) wild-type treated T2; 5) *Brca2*  $-/-$  untreated T1; 6) *Brca2*  $-/-$ 6; and 7) *T. pseudonana* chromosomes 1-24. Major and minor tick marks on the outside track represent 0.1 and 0.5 Mb length respectively. Red boxes highlight areas of genetic differentiation between *Brca2*  $-/-$  strains and wild type. The blue box indicates a region of genetic differentiation between treated and untreated *Brca2*  $-/-$  strains. Plot was made using Circos via the Galaxy web tool (Krzyszowski et al., 2009; Community, 2022)

Figure 5.7 shows the location and value of mean  $F_{ST}$  for each 1kb window along the *T. pseudonana* genome. Briefly, in Figure 5.7 the exterior black circle is the chromosomes of *T. pseudonana* with the major and minor ticks representing 0.1 and 0.5 Mb of length respectively and each track is a separate population. There is a higher level of genetic differentiation between *Brca2*  $-/-$  and wild-type strains than there is between treatment conditions, suggesting again that the treatment effect is minimal, whereas the loss of *Brca2* has increased the rate of genetic differentiation. The end of chromosome 18 shows the

most drastic  $F_{ST}$  values across all *Brca2*  $-/-$  strains. This is the same region in which large LOH events were detected (Section 5.2.1.1.2). The LOH event significantly changed the allele frequencies across this region and therefore resulted in large  $F_{ST}$  values.

### 5.2.6 Copy Number Variation (CNV)

Copy number variation (CNV) describes the genetic phenomenon where the number of genetic sequences at loci is either repeated or reduced resulting in an abnormal number of sequence copies (Pös et al., 2021). Exogenous environmental factors (i.e. chemical DNA mutagenesis and ultraviolet radiation) and genomic structure contribute to CNV instability (Arlt et al., 2012; Zhou et al., 2013; Chen et al., 2014). In chapter 4 *Brca2*  $-/-$  cultures which were able to grow in normal conditions for 10 months (approximately 200 – 250 generations) showed a marked increase in both gains and losses in CNV of specific genes. In this experiment both wild-type and *Brca2*  $-/-$  strains were re-sequenced after 20( $\pm$ 3) and 40( $\pm$ 3) generations, T1 and T2 respectively.

Using methods in Chapter 2, the CNV was calculated for all genes within the *T. pseudonana* reference genome in each sample. The number of genes with either an increase or decrease in CNV was generally higher in *Brca2*  $-/-$  strains than in wild-type strains (Table 5.5; Figure 5.8). However, there is only a significant difference in the number of genes with a loss in CNV when comparing all *Brca2*  $-/-$  and wild-type samples, but not genes with a gain in CNV. The overall increase in global copy number variation in *Brca2*  $-/-$  strains compared with wild-type, regardless of treatment, reflects the results of studies that show replication stress can induce CNV formation (Arlt et al., 2012). Loss of *Brca2* function is known to reduce genomic stability as gains and losses in DNA, resulting in CNVs, are commonly found in cancer tissues due to their inherent replication stress due to mitotic recombination errors (Levine and Holland, 2018; Steele et al., 2022). Data from Chapter 4 showed that the *T. pseudonana Brca2* gene is directly involved in DNA maintenance (Chapter 4 section 4.2.4.1), which provides evidence that the increase in copy number variation relative to the wild type is impacted by the loss of *Brca2* function which results in impaired mitotic recombination.

Table 5.5. Average losses and gains in CNV in samples grouped by culture and culture and treatment. SD = standard deviation of the mean. N = number of samples.

Culture	Gains in CNV		Losses in CNV		N
	Average	SD	Average	SD	
<b>By Culture</b>					
<i>Brca2</i> -/-	691.3	254.04	692.5	269.23	n=11
Wild type	530.9	337.64	432.5	281.05	n=11
<b>By Culture and Treatment</b>					
<i>Brca2</i> -/- Treated	710.0	386.51	741.4	398.04	n=5
<i>Brca2</i> UNTREATED	675.8	298.08	651.8	304.04	n=6
Wild type TREATED	441.6	354.48	380.4	336.55	n=5
Wild-type UNTREATED	605.3	359.27	475.8	261.62	n=6

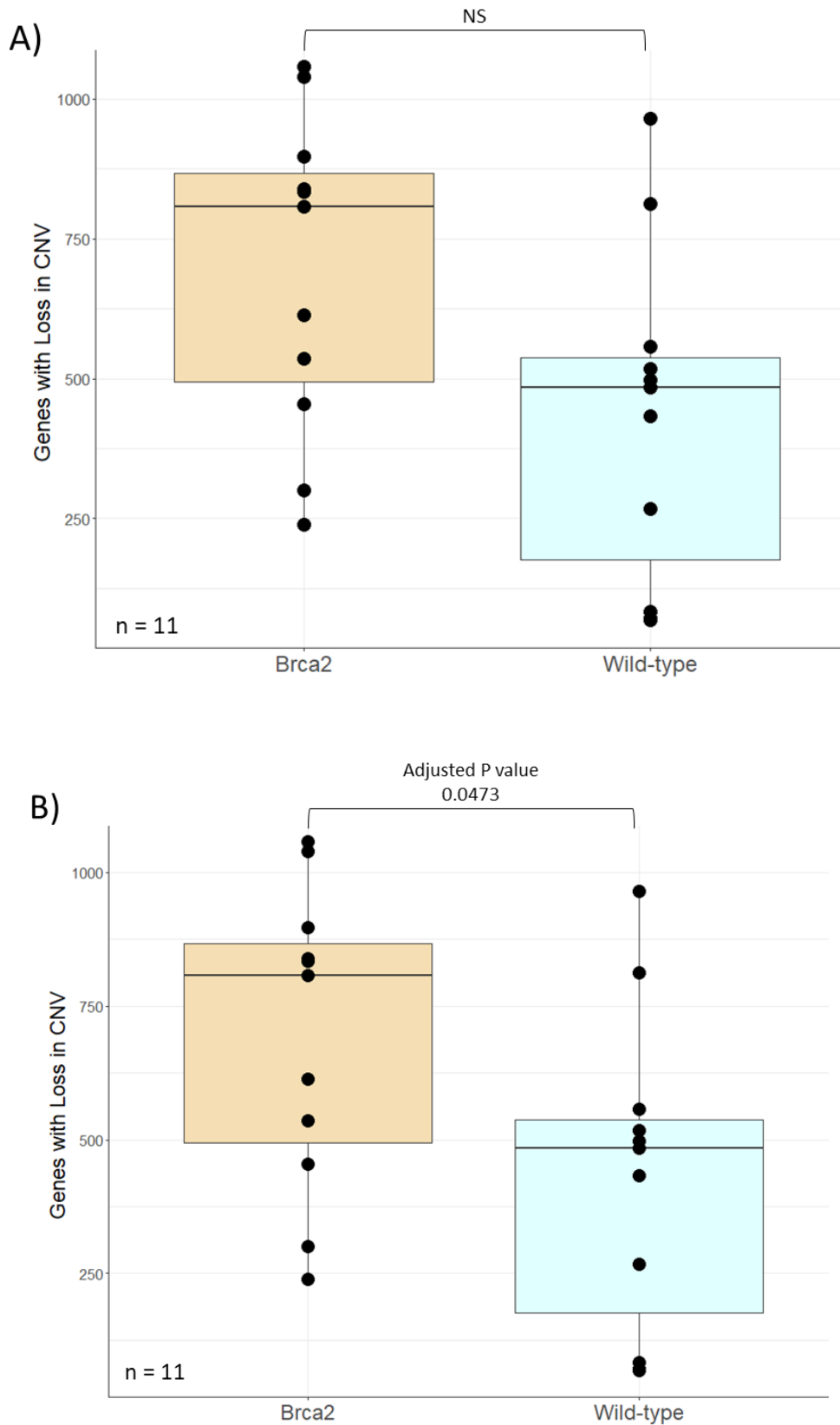
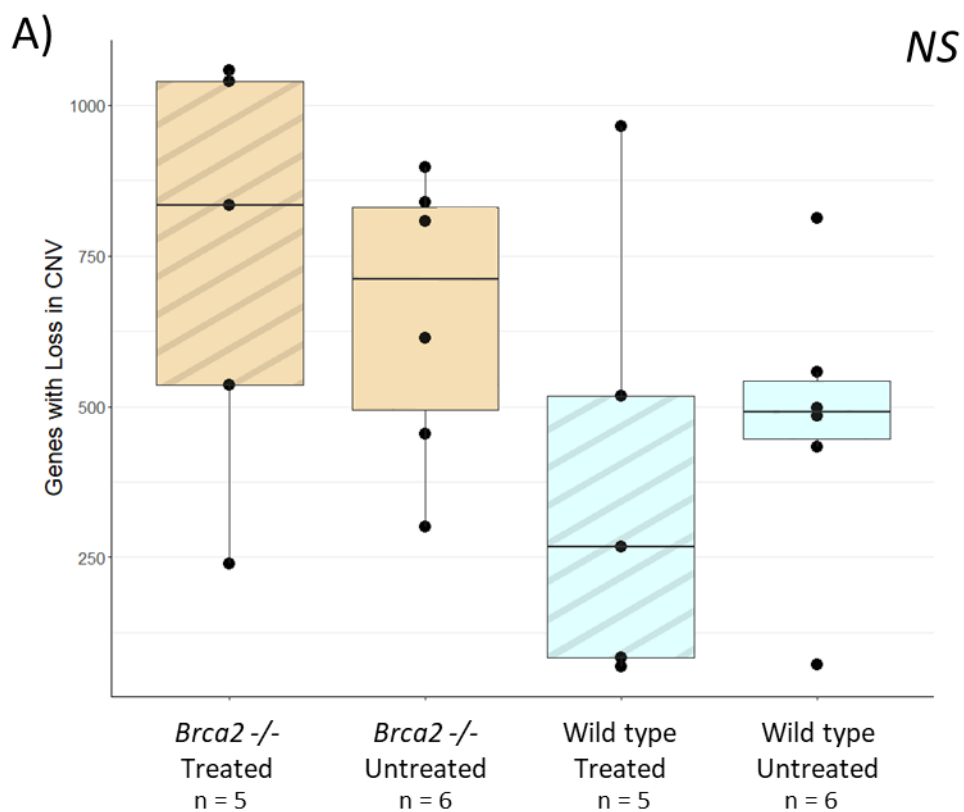


Figure 5.8. CNV in *Brca2*<sup>-/-</sup> (light yellow) strains compared to wild-type (blue). A) Number of Genes with a gain in CNV; B) Number of genes with a loss in CNV. N=11. P values shown are adjusted P values and were calculated via a pairwise t-test and adjusted using the Bonferroni method in R. NS = no significance.

The same trend is seen when viewing the results based on treatment, either with the DNA-damaging agent MMS (Treated) or with mock treatment (Untreated). There are no significant differences between any of the groups (Figure 5.9). Interestingly, the MMS treatment appears to generate a relative increase in both losses and gains in CNV in *Brca2*  $-/-$  strains, but not in the wild-type (Figure 5.9). Both wild-type and *Brca2*  $-/-$  strains received the same dose of MMS – the  $EC_{50}$  concentration for wild-type – which was proven to be lethal in *Brca2* strains under prolonged exposure. During this experiment strains were only exposed for 1.5 hours to ensure *Brca2*  $-/-$  cultures survived, and the number of CNVs found suggests that the wild-type strains could withstand the MMS treatment without the same implications in genomic instability.





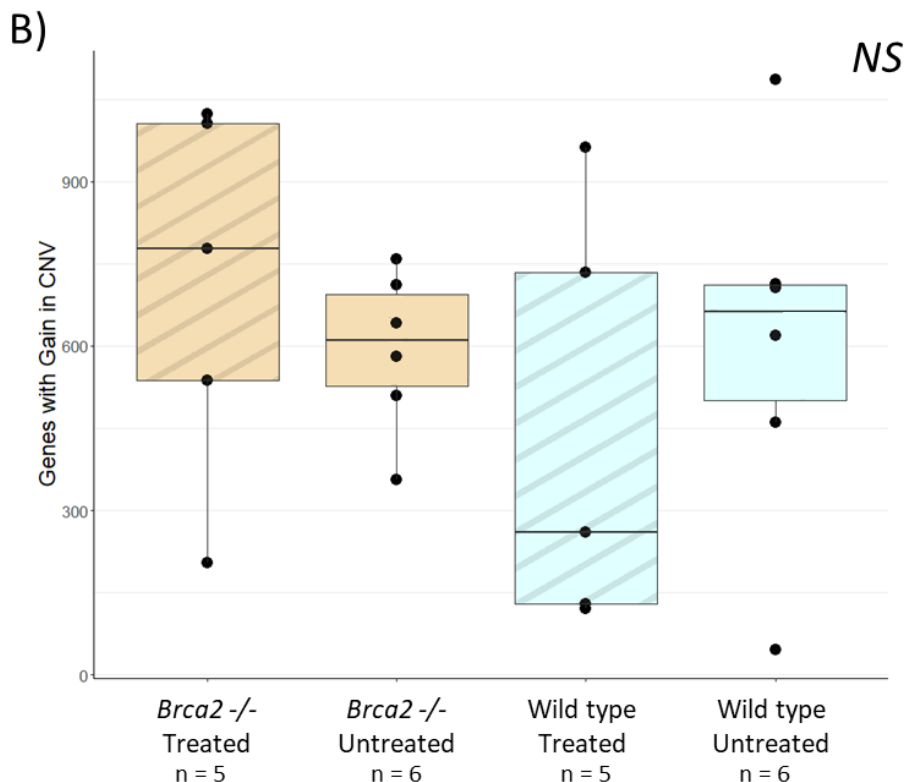


Figure 5.9. Number of CNV-affected genes in *Brca2*  $-/-$  (light yellow) and wild-type (blue) strains grouped by treatment (Treated = dashed boxes and Untreated = solid boxes). A) Number of genes which had a loss in CNV; B) Number of genes which had a gain in CNV.  $N=5$  for both treated groups and  $N=6$  for untreated groups. Adjusted  $P$  values were calculated via a pairwise  $t$ -test and adjusted using the Bonferroni method in R, however there is no significance between any groups, so NS = no significance.

Each sample was compared to the wild-type T0 sample to reduce background noise in the data due to the growth of *T. pseudonana* in general lab conditions for a prolonged period. Figure 5.10 shows the CNV profile of two samples at T1 and T2. Red points represent gains in CNV, blue points represent genes with a loss in CNV and grey represents genes with less than 25% difference in CNV compared to wild-type T0. The figures showing the CNV profile of all samples is in Appendix figure A.11. These figures suggest that *Brca2* has a direct impact on the genomic stability of *T. pseudonana*. The numbers of genes influenced by CNV are in Appendix Table A.11.

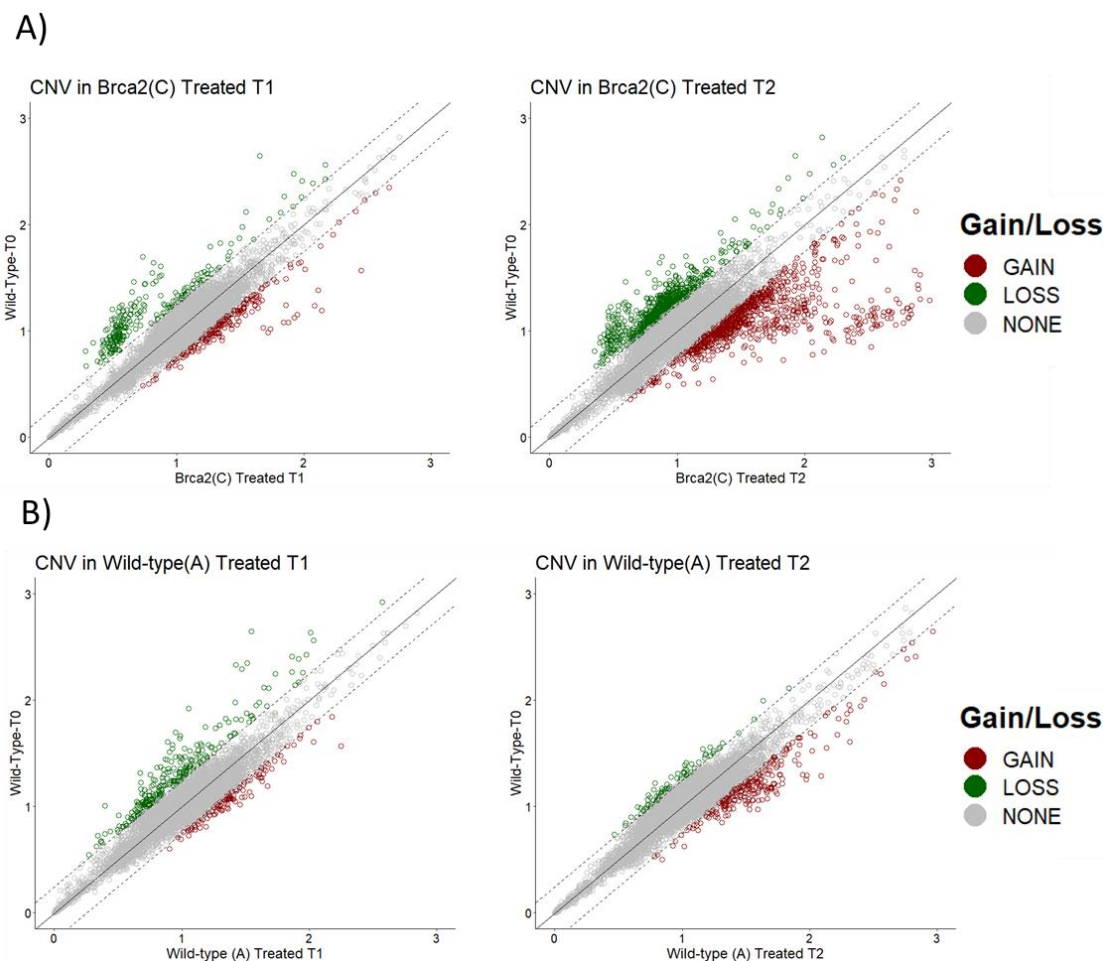


Figure 5.10. CNV variation of two samples; A) *Brca2*  $-/-$  Treated Replicate C T1 and T2 and B) Wild-type Replicate A Treated T1 and T2. Red dots indicate genes with gain in CNV while green indicates a loss in CNV. The solid line represents a 1:1 relationship with the CNV of the wild-type and the dashed lines represent the cut-off for a CNV to be considered (25% increase or decrease in CNV compared to wildtype). All figures for all samples are in Appendix figure A.4.

### 5.2.6.1 Genes Affected by Copy Number Variation

Using methods described in Chapter 2, genes affected by CNV were submitted for GO term analysis using g:Profiler online (Kolberg et al., 2023). First, any genes with a loss or gain in CNV shared across either all *Brca2*  $-/-$  strains or wild-type were identified. Resulting in 14 and 142 genes out of all 11,390 genes shared across all *Brca2*  $-/-$  strains for gains and losses respectively while no genes were shared across all wild-type samples. There were no enriched GO terms for the 14 shared genes with gains in *Brca2*  $-/-$ , but there were two biological processes enriched in the 142 shared genes with a loss in CNV shared across *Brca2*  $-/-$  samples. They were cyclic guanosine monophosphate (cGMP) biosynthesis (GO:0006182) and cGMP metabolic process (GO:004608), adjusted P values were  $2.533 \times 10^{-2}$  and  $2.533 \times$

$10^{-2}$  respectively. cGMP, a cyclic nucleotide, is derived from guanosine triphosphate (GTP and its primary function is a secondary messenger by activating protein kinases (Francis and Corbin, 1999). In diatoms, cGMPs have been reported to have a key role in the onset of sexual reproduction in pennate diatoms, regulation of the silicon cycle, acclimation to CO<sub>2</sub> and regulation of Ca<sup>+</sup> in the cell (Aline et al., 1984; Hennon et al., 2015).

Only the untreated *Brca2* *-/-* (B) sample at timepoint 2 had an enriched GO term relating to cGMP biosynthesis. There were more enriched terms among genes with a loss in CNV for both wild-type and *Brca2* *-/-* samples (Table 5.6; Table 5.7). Overall, more *Brca2* *-/-* samples contained enriched GO terms for genes affected by CNV compared with wild-type; 4 and 3 for gains in *Brca2* *-/-* and wild-type samples respectively and 7 and 5 for losses in *Brca2* *-/-* and wild-type samples respectively. In *Brca2* *-/-* samples, genes with a gain in CNV shared enriched GO terms related to gene expression. In genes with a loss in CNV, there was shared enrichment for purine metabolism and catalytic activity (Table 5.6). While in wild-type samples there were no common GO terms enriched between the individual samples (Table 5.7).

Table 5.6. Enriched GO terms of genes affected by CNV in individual *Brca2* *-/-* samples.

Culture	CNV	Source	GO term name	GO ID	adjusted P value
B2B MOCK_T1	GAIN	GO:BP	gene expression oxidoreduction-driven active transmembrane	GO:0010467	2.03E-02
B2C_MMS_T1	GAIN	GO:MF	transporter activity	GO:0015453	3.37E-02
B2C_MMS_T1	GAIN	GO:BP	tRNA metabolic process	GO:0006399	3.88E-02
B2C_MMS_T1	GAIN	GO:BP	translation	GO:0006412	4.18E-02
B2C_MMS_T1	GAIN	GO:CC	ribosomal subunit	GO:0044391	3.41E-02
B2C_MMS_T2	GAIN	GO:MF	alpha-1,6-mannosyltransferase activity	GO:0000009	1.41E-02
B2C_MMS_T2	GAIN	GO:CC	mannan polymerase complex	GO:0000136	2.96E-02
B2C_MMS_T2	GAIN	GO:CC	Golgi cis cisterna cellular nitrogen compound biosynthetic process	GO:0000137 GO:0044271	2.96E-02 1.15E-02
B2a_Mock_T1	LOSS	GO:MF	transferase activity	GO:0016740	1.90E-02
B2a_Mock_T1	LOSS	GO:MF	serine-type endopeptidase activity	GO:0004252	3.92E-02
B2a_Mock_T2	LOSS	GO:CC	guanylate cyclase complex, soluble	GO:0008074	4.99E-02
B2b_MMS_T1	LOSS	GO:CC	inner mitochondrial membrane protein complex	GO:0098800	2.29E-02
B2b_MMS_T1	LOSS	GO:CC	organelle membrane	GO:0031090	3.66E-02
B2b_MMS_T1	LOSS	GO:CC	cytoplasm	GO:0005737	4.12E-02
B2b_MMS_T1	LOSS	KEGG	Purine metabolism	KEGG:00230	1.10E-02
B2B_MMS_T2	LOSS	GO:MF	catalytic activity	GO:0003824	2.03E-02
B2B_MMS_T2	LOSS	KEGG	Purine metabolism	KEGG:00230	1.34E-02
B2B MOCK_T1	LOSS	GO:MF	catalytic activity	GO:0003824	2.57E-02
B2B MOCK_T1	LOSS	GO:CC	cytoplasm	GO:0005737	1.37E-03
B2B MOCK_T1	LOSS	GO:CC	organelle membrane	GO:0031090	4.68E-02
B2B MOCK_T1	LOSS	KEGG	Purine metabolism	KEGG:00230	8.01E-03
B2C_MMS_T1	LOSS	GO:CC	guanylate cyclase complex, soluble	GO:0008074	4.99E-02
B2C_MMS_T2	LOSS	GO:CC	cytoplasm	GO:0005737	3.09E-02

Table 5.7. Enriched GO terms of genes affected by CNV in individual wild type samples.

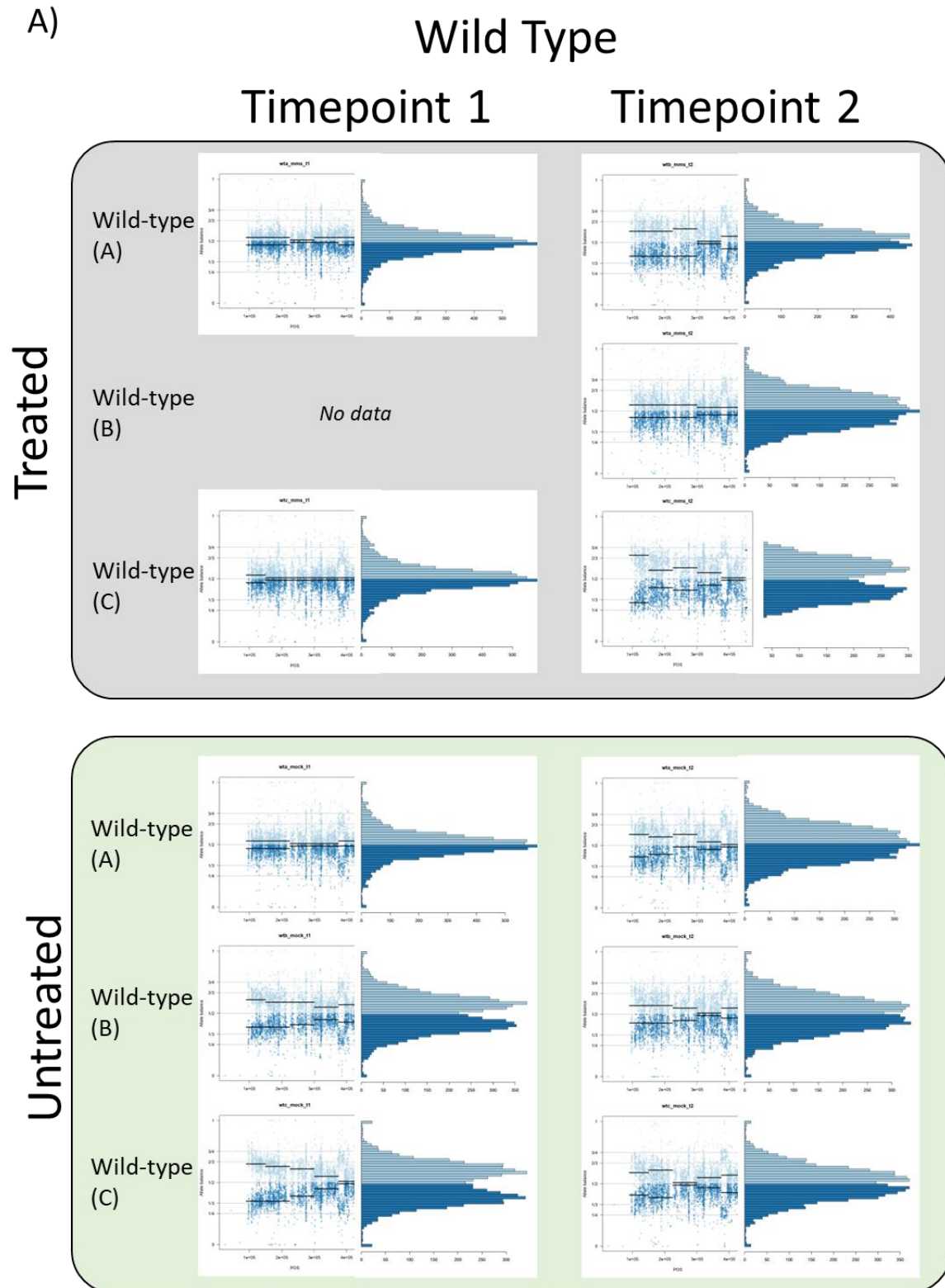
Culture	CNV	Source	GO term name	GO ID	Adjusted P value
WTA_MMS_T1	GAIN	KEGG	Riboflavin metabolism	KEGG:00740	4.99E-02
WTA_MOCK_T2	GAIN	GO:BP	positive regulation of cell cycle process	GO:0090068	8.39E-03
WTB_MOCK_T1	GAIN	KEGG	Aminoacyl-tRNA biosynthesis	KEGG:00970	4.27E-02
WTA_MMS_T1	LOSS	GO:MF	tyrosine-tRNA ligase activity	GO:0004831	4.98E-02
WTA_MMS_T2	LOSS	GO:MF	transition metal ion binding	GO:0046914	2.35E-03
WTB_MMS_T2	LOSS	GO:MF	peptidyl-prolyl cis-trans isomerase activity	GO:0003755	1.29E-02
WTB_MMS_T2	LOSS	GO:MF	cis-trans isomerase activity	GO:0016859	1.29E-02
WTB_MOCK_T2	LOSS	GO:BP	response to inorganic substance	GO:0010035	3.59E-02
WTC_MOCK_T2	LOSS	GO:MF	CDP-alcohol phosphatidyltransferase activity	GO:0017169	5.30E-03
WTC_MOCK_T2	LOSS	KEGG	Glycerophospholipid metabolism	KEGG:00564	5.00E-02

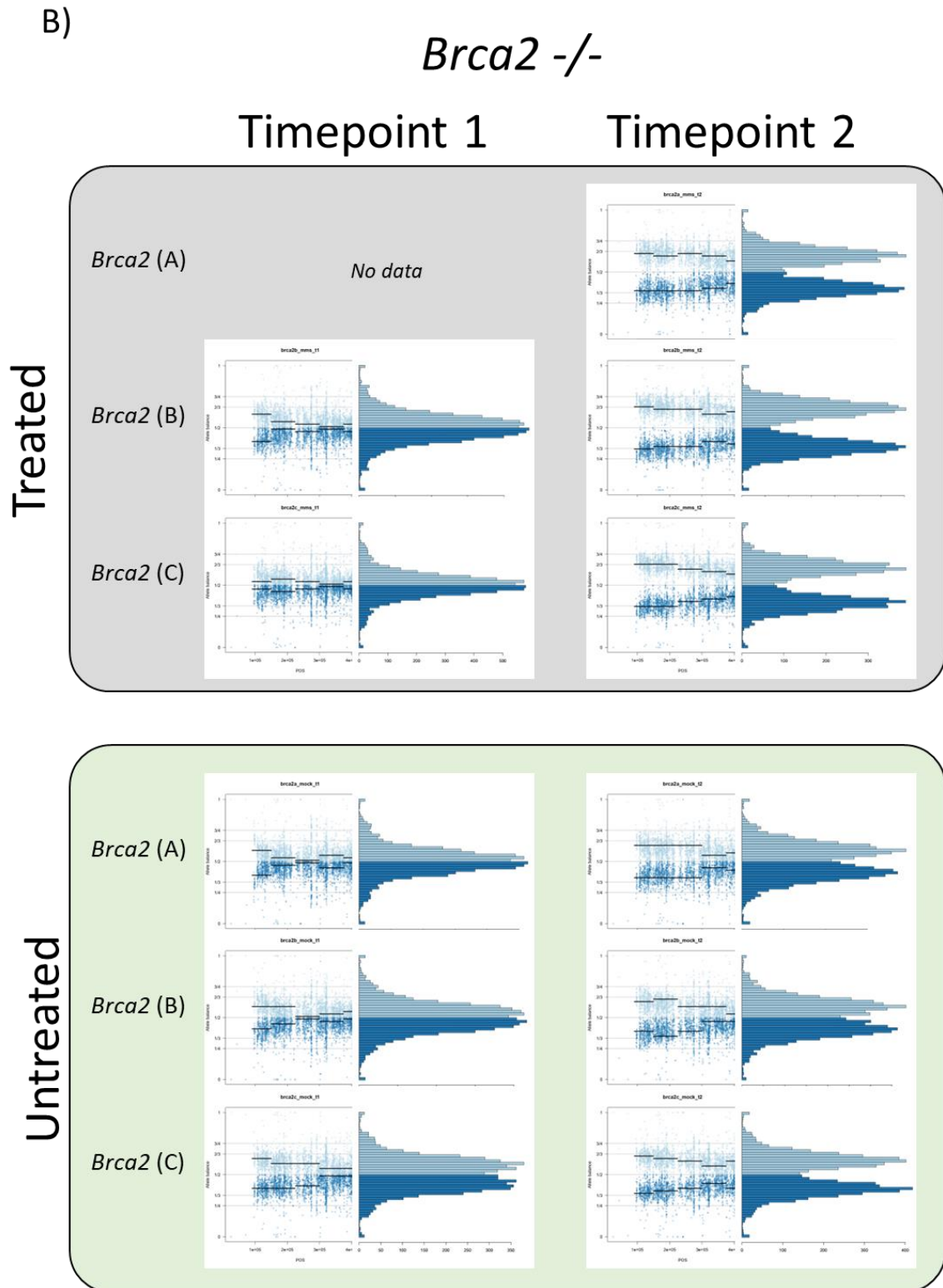
### 5.2.7 Allele Balance and Chromosome Instability

Allele balance is a term used to describe the relative abundance of alleles at a particular locus in a diploid organism (Fletcher et al., 2022). Aneuploidy defines cells with an abnormal number of chromosomes, while trisomy and partial trisomy refer to three copies of an entire chromosome or a distinct region respectively (Santaguida and Amon, 2015). Aneuploidy is a result of errors in mitotic division such as the non-disjunction of chromosomes. In Chapter 4, *Brca2*  $-/-$  strains showed evidence of trisomy in multiple chromosomes with both strain and genotype-specific trisomy evident compared to the wild-type. In this experiment, only one chromosome (chromosome 23) showed any evidence of trisomy. Interestingly, every single *Brca2*  $-/-$  replicate duplicated chromosome 23 between timepoint 1 and timepoint 2 (~19 generations  $\pm$  2 generations) regardless of treatment (Figure 5.11). Treatment with MMS slightly increased the extent of trisomy across the chromosome in *Brca2*  $-/-$  strains. Both wild type and *Brca2*  $-/-$  T0 samples showed no signs of aneuploidy.

The results from the wild-type samples suggest that chromosome 23 is not as stable as others and is prone to copy number variations and duplication. Some regions of partial trisomy reverted to a diploid state (Wild-type C Untreated), whereas others (Wild-type A and C treated) went from a diploid state to partial trisomy (Figure 5.11). Mutations in *Brca2* which disrupt the protein's function in cell-cycle regulation are known to increase

chromosome instability (Joukov et al., 2006; Thompson et al., 2010). These results, along with data from Chapter 4, suggest that *Brca2* has a role in mitotic division in diatoms and is critical to maintaining chromosome stability.







### 5.3 Discussion

This experiment used whole-genome resequencing of triplicates from both wild-type and *Brca2*  $-/-$  strains under induced DNA damage stress after ~20 and ~40 generations to identify the role of *Brca2* in the maintenance of genomic stability in *T. pseudonana*. Sequencing was successful for all but one sample, MMS treated wild-type replicate B at timepoint 1 (WT\_B\_MMS\_T1). The MMS-treated *Brca2*  $-/-$  replicate A at timepoint one also had to be discarded due to the presence of a wild-type sequence at the *Brca2* locus. This was most likely due to mislabelling of the sample between extraction and sequencing.

Unlike in Chapter 4, this experiment resequenced both *Brca2*  $-/-$  and wild type *T. pseudonana* strains and allowed for the direct comparison of both over time. As mentioned previously, loss of *Brca2* is known to increase global spontaneous mutation rate due to its essential role in DNA repair of DNA double-strand breaks (DSBs) through homologous recombination (HR; Moynahan and Jasin, 2010; Zámbořszky et al., 2017). As expected, there was a higher mutation rate across all *Brca2*  $-/-$  strains regardless of treatment compared to wild-type. However, the mutation rate estimates in this chapter are relatively high compared to those previously reported for diatoms (Krasovec et al., 2019). Though the mutation rates from Krasovec et al., 2019 were not calculated from diatoms under laboratory conditions and only undergoing exponential growth with 24 hours of light. The difference in mutation rates could be because previous studies have started from freshly isolated clonal cultures, whereas these experiments started with a population of *Brca2*  $-/-$  cells. However, this population did arise from a single clonal culture and had only been in culture for roughly 2 months before starting the short-term evolution experiment. Despite this, the result still confirms that in *Brca2*  $-/-$  strains there is an increase in mutation rate. These mutation rate estimates are similar to those reported in previous studies using *T. pseudonana* in our lab (Schmidt, 2017).

Genome instability can negatively impact the fitness and adaptive potential of microbial organisms. However, the prevailing theory is one of 'risk and reward', where a balance is formed between the generation of genetic diversity and the maintenance of genome sequences (Giraud et al., 2001). *Brca2* is well known to maintain genomic structure and in mammalian systems, it suppresses tumorigenesis which arises due to LOH events,



increase in CNVs, large chromosome rearrangements and changes in ploidy (Pawlyn et al., 2018). *Brca2* mediates genomic stability directly through homologous recombination (HR). HR can result in crossover events, LOH, gene conversion and ectopic recombination where the template is from homologous sequences at nonallelic positions (Symington et al., 2014). If the HR pathway is non-functional, error-prone DNA repair pathways are employed instead. The results from this chapter characterised increases in both CNVs and LOH events in *Brca2* strains, which potentially arose from the repair of DNA insults via the NHEJ pathway (Lieber, 2010).

Changes in ploidy across chromosome 23, termed partial and chromosome-wide trisomy, in *Brca2*  $-/-$  strains were interesting because they occurred in all 11 strains. The entire chromosome is duplicated in several replicates within 20 generations and partial trisomy was seen in the others. The MMS treatment appeared to increase the speed of trisomy. While in the wild type, there was evidence of instability within the chromosome, only partial trisomy was observed. The partial trisomy in the wild-type cultures was not influenced by the MMS treatment surprisingly and the underlying reasons why chromosome 23 was prone to trisomy is unclear. *Brca2* does not have a direct role in chromosome segregation during replication, but it does form a complex with *Brca1* which is directly involved in centrosome amplification and chromosome segregation (Xu et al., 1999). Loss of *Brca2* resulted in the formation of micronuclei in tumour cell lines (Tutt et al., 1999). Micronuclei arise due to mitotic division occurring with chromosomes which have unrepaired DNA breaks or chromosomes which failed segregation during replication (Chester et al., 1998). Again, these studies are in mammalian systems which have many more accessory DNA repair proteins which are linked to *Brca2* function, and they hypothesise that while *Brca2* may not be directly involved, the indirect consequences of its loss have knock-on effects on DNA repair and cell cycle checkpoint proteins which interact with *Brca2*. Results from Chapter 3 show that diatoms possess fewer DNA repair proteins than mammalian systems. To further understand the link between *Brca2* function and ploidy changes, it would be interesting to label BRCA2 with a fluorescent transgene to see its location during replication and use proteomics to see the changes in the abundance of all DNA repair proteins over the cell cycle.

One point to note is that throughout this chapter the treatment of MMS rarely impacted the genome in *T. pseudonana*. The treatment was mild as cultures were only exposed for a short period weekly. In future, either a more harsh or continuous treatment would increase the frequency of induced DNA damage and repair, which might increase the frequency of large-scale genomic changes and further elucidate the role of *Brca2* in diatoms.

## 6 General Discussion

### 6.1 Summary of Main Results

#### **DNA Repair Genes in Diatoms**

Through extensive phylogenetic analysis all DNA repair genes in diatoms were identified. The core genes within each DNA repair pathway were conserved across diatom species. The composition of orthologues found in diatoms were similar to those found in other related taxonomic groups, such as green and red algae with a few exceptions. Notably, the X-family polymerases, specifically polymerases lambda, were found in all proteomes from pennate diatoms, while no orthologues were identified in centric diatoms. Though not essential for DNA repair, it highlights previously unknown differences in the mechanisms of DNA repair between diatom groups.

#### **Role of *Brca2* in *T. pseudonana* and Impact on Adaptive Potential**

CRISPR/Cas mediated knockout of the highly conserved BRCA2 DNA repair associated gene (*Brca2*) in *T. pseudonana*, created mutant cell lines hypersensitive to DNA damage and unable to adapt to high temperature stress. Hypersensitivity to induced DNA damage confirmed the role of *Brca2* as an essential DNA repair gene in *T. pseudonana* and provided evidence that the homologous recombination (HR) pathway was significantly impaired. The inability to grow under high temperatures, which the wild type could withstand, showed that HR is a key mechanism in adaptation to environmental stress in diatoms.

#### **Maintenance of Genomic Stability by Homologous Recombination**

The consequences of impaired HR in *T. pseudonana* was revealed through whole-genome resequencing of mutant cell lines. Comparative genomics showed mutant cell lines; genomes to be characterised by increases in copy number variations (CNV), copy neutral LOH events and spontaneous mutations. Genes with a loss in CNV were significantly higher in *Brca2*  $-/-$  mutants as was copy number neutral LOH events where two events spanned > 100kb in chromosome 18 within 40 generations of growth. Chromosomes that were inherently unstable (Chromosome 23) were prone to trisomy in *Brca2*  $-/-$  cell lines, potentially due to the role of *Brca2* in chromosome segregation during replication. After

regular growth conditions for 10 months (roughly 250 generations) trisomy events were found throughout the genome with both strain and genotype specific events. This suggests that HR directly impacts segregation of chromosomes during division and diatoms were able to adapt to aneuploidy without HR.

## 6.2 General Discussion

Diatoms are one of the most successful phytoplankton groups and are responsible for around 50% of annual marine primary production (Nelson et al., 1995). Their unique life cycle is dominated by extended periods of clonal reproduction, especially during bloom formation in the spring and autumn months, while meiotic division is regulated by cell size. Despite this they are the most diverse phyla of phytoplankton and have adapted to almost all aquatic environments. Mitotic recombination has been overlooked as mechanisms by which they generate this diversity and there is little research to elucidate the responsible genes and molecular underpinnings of HR in diatoms. By using phylogenomics and genome editing to impair HR in *T. pseudonana* for the first time, this research will increase the understanding of the role of mitotic recombination on genome stability in diatoms.

The function of the BRCA2 has been extensively studied since mutations in *Brca2* genes were discovered to be linked with increased chances of tumour development in humans (Wooster, 1995). In eukaryotes, BRCA2 mediates homology search and strand exchange during HR DNA repair of DNA double-strand breaks (Liu et al., 2011). Mouse models without functional BRCA2 are highly sensitive to DNA damaging agents, including methyl methanesulfonate (MMS; Tutt et al., 1999). Their genomes were unstable and characterised by changes in genomic structure such as large chromosomal aberrations and aneuploidy. The same phenotypic hypersensitivity to induced DNA damage and genomic instability is also reported in *Caenorhabditis elegans* as well as multiple higher plants such as *Arabidopsis thaliana* and *Ustilago maydis* (Kojic et al., 2002; Siaud et al., 2004; Martin et al., 2005). Knocking out *Brca2* in *T. pseudonana* resulted in hypersensitivity to DNA damage through MMS and negative impacts on growth rate, confirming its role in DNA repair. However, unlike previous studies, this study also showed that *T. pseudonana* was able to adapt after loss of *Brca2* and revert to a fitness similar to that of wild type (Chapter 4), without repairing the damaged locus, however sensitivity to MMS never recovered. In

future the dose-response curves reported in this thesis could be done in parallel with cultures under a light-dark cycle. The addition of the light-dark cycle would impact the cell cycle, promoting more replication during the light period. It would be interesting to see the tolerance of DNA damage in comparison with growth under continuous light as there may be less background DNA from less UV radiation under the dark period. The dark period may also give *Brca2*  $-/-$  cell lines more time to repair DNA DSBs through NHEJ before replication.

The importance of HR in the adaptation of natural diatom populations to changing environmental conditions is not well understood. In this thesis, *Brca2*  $-/-$  *T. pseudonana* strains were unable to grow at temperatures above 30°C while wild type strains have been grown at 32°C for prolonged periods of time (Chapter 4; Schmidt, 2017). However, they could grow at lower temperatures with a similar fitness as the wild type. Temperatures above and below the normal niche for a species are known to increase spontaneous mutation rates, with higher temperatures resulting in higher mutation rates than colder temperatures (Waldvogel and Pfenninger, 2021). HR-deficient diatoms generated in this thesis had completely lost their ability to adapt and survive under high temperatures, suggesting that HR mediated DNA repair is essential to adapt to environmental stress. HR is used to repair double-strand breaks (DSBs), implying that high temperature stress increased the frequency of DSBs across the genome. But even without HR, non-homologous end joining (NHEJ) and its sub pathways would be employed to repair these DSBs. The error prone nature of NHEJ in combination with the increase in mutations due to high temperature appears halt division completely and ultimately end in cell death. Understanding the link between HR and the adaptive potential of diatoms is key to understand how they will evolve under the warming climate. Whether HR is upregulated to generate novel genetic diversity under stress, or simply has a higher rate of repair than previously thought requires further work.

The loss of HR in *T. pseudonana* resulted in an unstable genome prone to loss of heterozygosity (LOH) and duplication of entire chromosomes. Similar to the phenotype, the genomic consequences of HR deficiency are commonly seen in organisms that have lost *Brca2* and HR function (Martin et al., 2005; Abkevich et al., 2012). This further confirms *Brca2*'s in genome maintenance is highly conserved across eukaryotes. Most studies have focused on the impact of *Brca2* loss on multicellular organisms in clinical research with loss

of *Brca2* commonly resulting in embryonic lethality in animal models (Hakem et al., 1998). In unicellular organisms, such as diatoms, there is no threat of disease or tumour development. While a knock out of *Brca2* reduced fitness, it was not lethal. Given *T. pseudonana*'s quick generation time and large populations there is a greater chance of lineages adapting and surviving. Diatoms may be able to withstand increased genomic instability and adapt using both short and long-term mechanisms.

In the initial generations mutant cell lines showed a significant increase in copy number variations, but not extensive duplication events. The genes affected by CNVs were enriched for both DNA repair domains and transcription factors, suggesting a beneficial result of CNVs created as a result of genomic instability. Longer term adaptation resulted in a similar number of CNV events but saw a significant increase in duplications affecting entire chromosomes. CNV and duplication events can have potentially deleterious effects, but again the large population size and rapid growth rate in diatoms has balanced that with the potential to adapt to loss of core cellular functions. Though the mechanisms underpinning this ability are yet to be elucidated, the data presented here provide exciting prospects of the use of diatoms as model species to understand the role of mitotic recombination in the evolution of phytoplankton.

### 6.3 Future Work

The data generated in this thesis provides novel insight into the functional role of HR in diatoms and highlights its importance in adaptation to environmental stress. These results, along with the current understanding of HR in diatoms, lead to the question, how is HR regulated in diatoms? Genomes are under constant threat of DNA damage through both internal and external factors and lesions must be readily repaired. The regulation of DNA repair pathways has been extensively studied in mammalian and yeast model organisms (Branzei and Foiani, 2008). The DNA damage response (DDR) is regulated on three levels by a network of kinases. First, kinases regulate DDR enzymes through direct phosphorylation to alter their activity, second kinases modify the structure of chromatin near the damaged DNA allowing for recruitment of further proteins and lastly, they regulate the cell cycle to allow DNA repair to proceed before division (Sirbu and Cortez, 2013). DNA repair has been reported to be significantly increased in lymphocytes under stress, though there is debate

on whether this due to increased DNA damage or if the cell upregulates the DDR in response to stress. This was also found in diatoms where Bulankova (et al., 2021) reported that HR rates increased under induced stress. To confirm the regulatory mechanisms in diatoms future studies would need to examine DNA repair genes on three levels, transcription (mRNA), post-transcription (ribosome profiling) and post translation (proteomics). Elucidating the regulation of each protein would provide a map of the process by which diatoms regulate DNA repair. From this future work could then examine the response of selected proteins to different environmental stressors to reveal which DNA repair pathways are activated. By doing so this would increase the understanding of the role of DNA repair, especially mitotic recombination, in diatom adaptation.

In this thesis, BRCA2 was shown to be a core player in maintaining genomic stability, adaption to environmental stress (temperature) and in repair of DNA DSBs. However, this may not be exclusive to diatoms. One area of future work could be to mine marine metaomic databases for BRCA2 to see if there are differences in the expression between climates. It would be interesting, for example, if there was a higher presence of BRCA2 in more stressful environments, such as the polar regions or high-nutrient low-chlorophyll zones (HNLC) in response to a higher rate of DNA damage. Equally, BRCA2 and other DDR genes might be suppressed in stochastic environments allowing increased genomic instability to generate novel genetic diversity without the need for meiosis.

Additionally, the short experimental evolution experiment from chapter 5 could be expanded to either include several DNA damaging agents or increased mms concentration/exposure time. The exposure time and concentration used in this thesis was chosen to introduce DNA damage stress without significantly impacting growth, however the concentration or the exposure time could have been increased. This may have increased the signature of MMS exposure and spontaneous mutations in both the wild type and *Brca2* *-/-* cultures further elucidating the role of *Brca2* under DNA damage stress.





## 7 References

- Abkevich, V., Timms, K. M., Hennessy, B. T., Potter, J., Carey, M. S., Meyer, L. A., . . . Lanchbury, J. S. (2012). Patterns of genomic loss of heterozygosity predict homologous recombination repair defects in epithelial ovarian cancer. *British Journal of Cancer*, 107(10), 1776-1782. doi:10.1038/bjc.2012.451
- Ackerman, S., & Horton, W. (2018). Chapter 2.4 - Effects of Environmental Factors on DNA: Damage and Mutations. In B. Török & T. Dransfield (Eds.), *Green Chemistry* (pp. 109-128): Elsevier.
- Adams, J., & Rosenzweig, F. (2014). Experimental microbial evolution: history and conceptual underpinnings. *Genomics*, 104(6), 393-398. doi:10.1016/j.ygeno.2014.10.004
- Adli, M. (2018). The CRISPR tool kit for genome editing and beyond. *Nature Communications*, 9(1), 1911. doi:10.1038/s41467-018-04252-2
- Ali, F., & Seshasayee, Aswin Sai N. (2020). Dynamics of genetic variation in transcription factors and its implications for the evolution of regulatory networks in Bacteria. *Nucleic Acids Research*, 48(8), 4100-4114. doi:10.1093/nar/gkaa162
- Aline, R. F., Jr., Reeves, C. D., Russo, A. F., & Volcani, B. E. (1984). Role of Silicon in Diatom Metabolism 1: Cyclic Nucleotide Levels, Nucleotide Cyclase, and Phosphodiesterase Activities during Synchronized Growth of *Cylindrotheca fusiformis*. *Plant Physiology*, 76(3), 674-679. doi:10.1104/pp.76.3.674
- Alverson, A. J., Beszteri, B., Julius, M. L., & Theriot, E. C. (2011). The model marine diatom *Thalassiosira pseudonana* likely descended from a freshwater ancestor in the genus *Cyclotella*. *BMC Evolutionary Biology*, 11(1), 125. doi:10.1186/1471-2148-11-125
- Andersson, B., Godhe, A., Filipsson, H. L., Zetterholm, L., Edler, L., Berglund, O., & Rengefors, K. (2022). Intraspecific variation in metal tolerance modulate competition between two marine diatoms. *ISME J*, 16(2), 511-520. doi:10.1038/s41396-021-01092-9
- Andreassen, P. R., Seo, J., Wiek, C., & Hanenberg, H. (2021). Understanding BRCA2 Function as a Tumor Suppressor Based on Domain-Specific Activities in DNA Damage Responses. *Genes (Basel)*, 12(7). doi:10.3390/genes12071034
- Angstenberger, M., Krischer, J., Aktaş, O., & Büchel, C. (2019). Knock-Down of a ligIV Homologue Enables DNA Integration via Homologous Recombination in the Marine Diatom *Phaeodactylum tricornutum*. *ACS Synth Biol*, 8(1), 57-69. doi:10.1021/acssynbio.8b00234
- Aravind, L., Walker, D. R., & Koonin, E. V. (1999). Conserved domains in DNA repair proteins and evolution of repair systems. *Nucleic Acids Research*, 27(5), 1223-1242. doi:10.1093/nar/27.5.1223
- Aravind L, Koonin EV. (2001) Prokaryotic homologs of the eukaryotic DNA-end-binding protein Ku, novel domains in the Ku protein and prediction of a prokaryotic double-strand break repair system. *Genome Research*.11(8):1365-74. doi: 10.1101/gr.181001. PMID: 11483577; PMCID: PMC311082.
- Arlt, M. F., Wilson, T. E., & Glover, T. W. (2012). Replication stress and mechanisms of CNV formation. *Current Opinion in Genetics & Development*, 22(3), 204-210. doi:https://doi.org/10.1016/j.gde.2012.01.009

- Armbrust, E. V., Berges, J. A., Bowler, C., Green, B. R., Martinez, D., Putnam, N. H., . . . Rokhsar, D. S. (2004). The genome of the diatom *Thalassiosira pseudonana*: ecology, evolution, and metabolism. *Science*, 306(5693), 79-86. doi:10.1126/science.1101156
- Asagoshi, K., Lehmann, W., Braithwaite, E. K., Santana-Santos, L., Prasad, R., Freedman, J. H., . . . Wilson, S. H. (2011). Single-nucleotide base excision repair DNA polymerase activity in *C. elegans* in the absence of DNA polymerase  $\beta$ . *Nucleic Acids Research*, 40(2), 670-681. doi:10.1093/nar/gkr727
- Assaad, F. F., & Signer, E. R. (1992). Somatic and germinal recombination of a direct repeat in Arabidopsis. *Genetics*, 132(2), 553-566. doi:10.1093/genetics/132.2.553
- Barrick, J. E., & Lenski, R. E. (2013). Genome dynamics during experimental evolution. *Nature Reviews Genetics*, 14(12), 827-839. doi:10.1038/nrg3564
- Bartek, J., & Lukas, J. (2007). DNA damage checkpoints: from initiation to recovery or adaptation. *Current Opinion in Cell Biology*, 19(2), 238-245. <https://doi.org/https://doi.org/10.1016/j.ceb.2007.02.009>
- Basu, S., & Mackey, K. R. M. (2018). Phytoplankton as Key Mediators of the Biological Carbon Pump: Their Responses to a Changing Climate. *Sustainability*, 10(3), 869. Retrieved from <https://www.mdpi.com/2071-1050/10/3/869>
- Bedez, F., Linard, B., Brochet, X., Ripp, R., Thompson, J. D., Moras, D., . . . Poch, O. (2013). Functional insights into the core-TFIIH from a comparative survey. *Genomics*, 101(3), 178-186. doi:10.1016/j.ygeno.2012.11.003
- Beernink, H. T., & Morrical, S. W. (1999). RMPs: recombination/replication mediator proteins. *Trends in Biochemical Sciences*, 24(10), 385-389.
- Belshaw, N., Grouneva, I., Aram, L., Gal, A., Hopes, A., & Mock, T. (2022). Efficient gene replacement by CRISPR/Cas-mediated homologous recombination in the model diatom *Thalassiosira pseudonana*. *New Phytologist*, 238(1), 438-452. doi:https://doi.org/10.1111/nph.18587
- Benoiston, A. S., Ibarbalz, F. M., Bittner, L., Guidi, L., Jahn, O., Dutkiewicz, S., & Bowler, C. (2017). The evolution of diatoms and their biogeochemical functions. *Philos Trans R Soc Lond B Biol Sci*, 372(1728). doi:10.1098/rstb.2016.0397
- Berman, J. (2016). Ploidy plasticity: a rapid and reversible strategy for adaptation to stress. *FEMS Yeast Research*, 16(3). doi:10.1093/femsyr/fow020
- Bienstock, R. J., Beard, W. A., & Wilson, S. H. (2014). Phylogenetic analysis and evolutionary origins of DNA polymerase X-family members. *DNA Repair (Amst)*, 22, 77-88. doi:10.1016/j.dnarep.2014.07.003
- Boatman, T. G., Lawson, T., & Geider, R. J. (2017). A Key Marine Diazotroph in a Changing Ocean: The Interacting Effects of Temperature, CO<sub>2</sub> and Light on the Growth of *Trichodesmium erythraeum* IMS101. *PLoS One*, 12(1), e0168796. doi:10.1371/journal.pone.0168796
- Bodył, A., Mackiewicz, P., & Ciesála, J. (2017). Endosymbiotic Theory: Models and Challenges. In *Reference Module in Life Sciences*: Elsevier.
- Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15), 2114-2120. doi:10.1093/bioinformatics/btu170

Bowler, C., Allen, A. E., Badger, J. H., Grimwood, J., Jabbari, K., Kuo, A., . . . Grigoriev, I. V. (2008). The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes. *Nature*, 456(7219), 239-244. Retrieved from <http://www.nature.com/articles/nature07410>

Brandsma, I., & Gent, D. C. (2012). Pathway choice in DNA double strand break repair: observations of a balancing act. *Genome Integr*, 3(1), 9. doi:10.1186/2041-9414-3-9

Branzei, D., & Foiani, M. (2008). Regulation of DNA repair throughout the cell cycle. *Nature Reviews Molecular Cell Biology*, 9(4), 297-308. doi:10.1038/nrm2351

Breivik, J., & Gaudernack, G. (2004). Resolving the evolutionary paradox of genetic instability: a cost-benefit analysis of DNA repair in changing environments. *FEBS Letters*, 563(1), 7-12. doi:[https://doi.org/10.1016/S0014-5793\(04\)00282-0](https://doi.org/10.1016/S0014-5793(04)00282-0)

Briere, J.-F., Pracros, P., Le Roux, A.-Y., & Pierre, J.-S. (1999). A Novel Rate Model of Temperature-Dependent Development for Arthropods. *Environmental Entomology*, 28(1), 22-29. doi:10.1093/ee/28.1.22

Brown, A. D., Greenman, S., Claybon, A. B., & Bishop, A. J. R. (2021). BRCA2 Promotes Spontaneous Homologous Recombination *in vivo*. *Cancers (Basel)*, 13(15). doi:10.3390/cancers13153663

Bulankova, P., Sekulić, M., Jallet, D., Nef, C., van Oosterhout, C., Delmont, T. O., . . . De Veylder, L. (2021). Mitotic recombination between homologous chromosomes drives genomic diversity in diatoms. *Current Biology*, 31(15), 3221-3232.e3229. doi:<https://doi.org/10.1016/j.cub.2021.05.013>

Byrne, R. T., Klingele, A. J., Cabot, E. L., Schackwitz, W. S., Martin, J. A., Martin, J., . . . Cox, M. M. (2014). Evolution of extreme resistance to ionizing radiation via genetic adaptation of DNA repair. *eLife*, 3, e01322. doi:10.7554/eLife.01322

C.N, H. (1947). *The Chemical Kinetics of the Bacterial Cell*. Oxford University Press.

Carvajal-Garcia, J., Samadpour, A. N., Hernandez Viera, A. J., & Merrikh, H. (2023). Oxidative stress drives mutagenesis through transcription-coupled repair in bacteria. *Proc Natl Acad Sci U S A*, 120(27), e2300761120. doi:10.1073/pnas.2300761120

Castel, B., Tomlinson, L., Locci, F., Yang, Y., & Jones, J. D. G. (2019). Optimization of T-DNA architecture for Cas9-mediated mutagenesis in *Arabidopsis*. *PLoS One*, 14(1), e0204778. doi:10.1371/journal.pone.0204778

Cavan, E. L., Belcher, A., Atkinson, A., Hill, S. L., Kawaguchi, S., McCormack, S., . . . Boyd, P. W. (2019). The importance of Antarctic krill in biogeochemical cycles. *Nature Communications*, 10(1), 4742. doi:10.1038/s41467-019-12668-7

Cermeno, P., Falkowski, P. G., Romero, O. E., Schaller, M. F., & Vallina, S. M. (2015). Continental erosion and the Cenozoic rise of marine diatoms. *Proc Natl Acad Sci U S A*, 112(14), 4239-4244. doi:10.1073/pnas.1412883112

Chan, C. X., Bhattacharya, D., & Reyes-Prieto, A. (2012). Endosymbiotic and horizontal gene transfer in microbial eukaryotes: Impacts on cell evolution and the tree of life. *Mob Genet Elements*, 2(2), 101-105. doi:10.4161/mge.20110

Chang, H. H. Y., Pannunzio, N. R., Adachi, N., & Lieber, M. R. (2017). Non-homologous DNA end joining and alternative pathways to double-strand break repair. *Nature Reviews Molecular Cell Biology*, 18(8), 495-506. doi:10.1038/nrm.2017.48

- Chankova, S. G., Dimova, E., Dimitrova, M., & Bryant, P. E. (2007). Induction of DNA double-strand breaks by zeocin in *Chlamydomonas reinhardtii* and the role of increased DNA double-strand breaks rejoining in the formation of an adaptive response. *Radiation and Environmental Biophysics*, 46(4), 409-416. doi:10.1007/s00411-007-0123-2
- Charlson, R. J., Lovelock, J. E., Andreae, M. O., & Warren, S. G. (1987). Oceanic phytoplankton, atmospheric sulphur, cloud albedo and climate. *Nature*, 326(6114), 655-661. doi:10.1038/326655a0
- Chen, L., Zhou, W., Zhang, C., Lupski, J. R., Jin, L., & Zhang, F. (2014). CNV instability associated with DNA replication dynamics: evidence for replicative mechanisms in CNV mutagenesis. *Human Molecular Genetics*, 24(6), 1574-1583. doi:10.1093/hmg/ddu572
- Chen, Z., Bertin, R., & Froidi, G. (2013). EC50 estimation of antioxidant activity in DPPH assay using several statistical programs. *Food Chemistry*, 138(1), 414-420. doi:https://doi.org/10.1016/j.foodchem.2012.11.001
- Chester, N., Kuo, F., Kozak, C., O'Hara, C. D., & Leder, P. (1998). Stage-specific apoptosis, developmental delay, and embryonic lethality in mice homozygous for a targeted disruption in the murine Bloom's syndrome gene. *Genes & Development*, 12(21), 3382-3393. doi:10.1101/gad.12.21.3382
- Ciccia, A., & Elledge, S. J. (2010). The DNA Damage Response: Making It Safe to Play with Knives. *Molecular Cell*, 40(2), 179-204. doi:https://doi.org/10.1016/j.molcel.2010.09.019
- Cingolani, P., Platts, A., Wang le, L., Coon, M., Nguyen, T., Wang, L., . . . Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)*, 6(2), 80-92. doi:10.4161/fly.19695
- Citarelli, M., Teotia, S., & Lamb, R. S. (2010). Evolutionary history of the poly(ADP-ribose) polymerase gene family in eukaryotes. *BMC Evolutionary Biology*, 10(1), 308. https://doi.org/10.1186/1471-2148-10-308
- Čížková, M., Slavková, M., Vítová, M., Zachleder, V., & Bišová, K. (2019). Response of the Green Alga *Chlamydomonas reinhardtii* to the DNA Damaging Agent Zeocin. *Cells*, 8(7). doi:10.3390/cells8070735
- Cox, E. J. (2014). Diatom identification in the face of changing species concepts and evidence of phenotypic plasticity. *Journal of Micropalaeontology*, 33(2), 111-120. doi:doi:10.1144/jmpaleo2014-014
- Crump, K. S., Hoel, D. G., Langley, C. H., & Peto, R. (1976). Fundamental carcinogenic processes and their implications for low dose risk assessment. *Cancer Research*, 36(9 pt.1), 2973-2979.
- D'Alelio, D., d'Alcala, M. R., Dubroca, L., Sarno, D., Zingone, A., & Montresor, M. (2010). The time for sex: A biennial life cycle in a marine planktonic diatom. *Limnology and Oceanography*, 55(1), 106-114. doi:10.4319/lo.2010.55.1.0106
- Daly, M. J. (2012). Death by protein damage in irradiated cells. *DNA Repair*, 11(1), 12-21. doi:https://doi.org/10.1016/j.dnarep.2011.10.024
- Daly, M. J., Gaidamakova, E. K., Matrosova, V. Y., Vasilenko, A., Zhai, M., Venkateswaran, A., . . . Ghosal, D. (2004). Accumulation of Mn(II) in *Deinococcus radiodurans* facilitates gamma-radiation resistance. *Science*, 306(5698), 1025-1028. doi:10.1126/science.1103185

- Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., . . . Group, G. P. A. (2011). The variant call format and VCFtools. *Bioinformatics*, 27(15), 2156-2158. doi:10.1093/bioinformatics/btr330
- Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., . . . Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, 10(2). doi:10.1093/gigascience/giab008
- Davila, J. I., Arrieta-Montiel, M. P., Wamboldt, Y., Cao, J., Hagmann, J., Shedge, V., . . . Mackenzie, S. A. (2011). Double-strand break repair processes drive evolution of the mitochondrial genome in *Arabidopsis*. *BMC Biology*, 9(1), 64. doi:10.1186/1741-7007-9-64
- Davis, A. J., & Chen, D. J. (2013). DNA double strand break repair via non-homologous end-joining. *Transl Cancer Res*, 2(3), 130-143. doi:10.3978/j.issn.2218-676X.2013.04.02
- Davis, B. J., Havener, J. M., & Ramsden, D. A. (2008). End-bridging is required for pol  $\mu$  to efficiently promote repair of noncomplementary ends by nonhomologous end joining. *Nucleic Acids Research*, 36(9), 3085-3094.
- Ding, X., Philip, S., Martin, B. K., Pang, Y., Burkett, S., Swing, D. A., . . . Sharan, S. K. (2017). Survival of BRCA2-Deficient Cells Is Promoted by GIPC3, a Novel Genetic Interactor of BRCA2. *Genetics*, 207(4), 1335-1345. doi:10.1534/genetics.117.300357
- Doetsch, P. W., & Cunningham, R. P. (1990). The enzymology of apurinic/aprimidinic endonucleases. *Mutation Research/DNA Repair*, 236(2-3), 173-201.
- Domínguez, O., Ruiz, J. F., de Lera, T. L., García-Díaz, M., González, M. A., Kirchhoff, T., . . . Blanco, L. (2000). DNA polymerase mu (Pol  $\mu$ ), homologous to TdT, could act as a DNA mutator in eukaryotic cells. *The EMBO Journal*, 19(7), 1731-1742.
- Dorrell, R. G., & Bowler, C. (2017). Chapter Three - Secondary Plastids of Stramenopiles. In Y. Hirakawa (Ed.), *Advances in Botanical Research* (Vol. 84, pp. 57-103): Academic Press.
- Dutta, A., Dutreux, F., & Schacherer, J. (2022). Loss of Heterozygosity Spectrum Depends on Ploidy Level in Natural Yeast Populations. *Molecular Biology and Evolution*, 39(11), msac214. doi:10.1093/molbev/msac214
- Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5), 1792-1797. doi:10.1093/nar/gkh340
- Egglar, A. L., Inman, R. B., & Cox, M. M. (2002). The Rad51-dependent pairing of long DNA substrates is stabilized by replication protein A. *Journal of Biological Chemistry*, 277(42), 39280-39288. doi:10.1074/jbc.M204328200
- Ehlén, Å., Martin, C., Miron, S., Julien, M., Theillet, F.-X., Ropars, V., . . . Carreira, A. (2020). Proper chromosome alignment depends on BRCA2 phosphorylation by PLK1. *Nature Communications*, 11(1), 1819. doi:10.1038/s41467-020-15689-9
- Ehrenfeld, G. M., Shipley, J. B., Heimbrook, D. C., Sugiyama, H., Long, E. C., Van Boom, J. H., . . . Hecht, S. M. (1987). Copper-dependent cleavage of DNA by bleomycin. *Biochemistry*, 26(3), 931-942. doi:10.1021/bi00377a038
- Emms, D. M., & Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol*, 20(1), 238. doi:10.1186/s13059-019-1832-y

- Ene, I. V., Farrer, R. A., Hirakawa, M. P., Agwamba, K., Cuomo, C. A., & Bennett, R. J. (2018). Global analysis of mutations driving microevolution of a heterozygous diploid fungal pathogen. *Proc Natl Acad Sci U S A*, 115(37), E8688-e8697. doi:10.1073/pnas.1806002115
- Engler, C., & Marillonnet, S. (2014). Golden Gate cloning. *Methods Mol Biol*, 1116, 119-131. doi:10.1007/978-1-62703-764-8\_9
- Falkowski, P. G., Katz, M. E., Knoll, A. H., Quigg, A., Raven, J. a., Schofield, O., & Taylor, F. J. R. (2004). The Evolution of Modern Eukaryotic Phytoplankton. *Science*, 305(July), 354-360. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/15256663>
- Felsenstein, J. (1985). Confidence Limits on Phylogenies: An Approach Using the Bootstrap. *Evolution*, 39(4), 783-791. doi:10.2307/2408678
- Ferenczi, A., Chew, Y. P., Kroll, E., von Koppenfels, C., Hudson, A., & Molnar, A. (2021). Mechanistic and genetic basis of single-strand templated repair at Cas12a-induced DNA breaks in *Chlamydomonas reinhardtii*. *Nature Communications*, 12(1), 6751. doi:10.1038/s41467-021-27004-1
- Field, C. B., Behrenfeld, M. J., Randerson, J. T., & Falkowski, P. (1998). Primary Production of the Biosphere: Integrating Terrestrial and Oceanic Components. *Science*, 281(5374), 237. doi:10.1126/science.281.5374.237
- Filloramo, G. V., Curtis, B. A., Blanche, E., & Archibald, J. M. (2021). Re-examination of two diatom reference genomes using long-read sequencing. *BMC Genomics*, 22(1), 25. doi:10.1186/s12864-021-07666-3
- Fletcher, K., Han, R., Smilde, D., & Michelmore, R. (2022). Variance of allele balance calculated from low coverage sequencing data infers departure from a diploid state. *BMC Bioinformatics*, 23(1), 150. doi:10.1186/s12859-022-04685-z
- Flinn, P. W. (1991). Temperature-Dependent Functional Response of the Parasitoid *Cephalonomia waterstoni* (Gahan) (Hymenoptera: Bethyridae) Attacking Rusty Grain Beetle Larvae (Coleoptera: Cucujidae). *Environmental Entomology*, 20(3), 872-876. doi:10.1093/ee/20.3.872
- Fonseca, J. P., Bonny, A. R., Town, J., & El-Samad, H. (2020). Assembly of Genetic Circuits with the Mammalian ToolKit. *Bio Protoc*, 10(5), e3547. doi:10.21769/BioProtoc.3547
- Francis, S. H., & Corbin, J. D. (1999). Cyclic Nucleotide-Dependent Protein Kinases: Intracellular Receptors for cAMP and cGMP Action. *Critical Reviews in Clinical Laboratory Sciences*, 36(4), 275-328. doi:10.1080/10408369991239213
- G, D. (1977). Sexuality. *The Biology of Diatoms*. In *Botanical Monographs* (Vol. 13 pp. 250-283). London: Blackwell Scientific Publications.
- Gallo, C., d'Ippolito, G., Nuzzo, G., Sardo, A., & Fontana, A. (2017). Autoinhibitory sterol sulfates mediate programmed cell death in a bloom-forming marine diatom. *Nature Communications*, 8(1), 1292. doi:10.1038/s41467-017-01300-1
- Gandía, M., Xu, S., Font, C., & Marcos, J. F. (2016). Disruption of *ku70* involved in non-homologous end-joining facilitates homologous recombination but increases temperature sensitivity in the phytopathogenic fungus *Penicillium digitatum*. *Fungal Biol*, 120(3), 317-323. doi:10.1016/j.funbio.2015.11.001



- Ghosh, D., Nilavar, N. M., & Raghavan, S. C. (2022). A novel KU70-mutant human leukemic cell line generated using CRISPR-Cas9 shows increased sensitivity to DSB inducing agents and reduced NHEJ activity. *Biochimica et Biophysica Acta (BBA) - General Subjects*, 1866(12), 130246. <https://doi.org/https://doi.org/10.1016/j.bbagen.2022.130246>
- Giani, A. M., Gallo, G. R., Gianfranceschi, L., & Formenti, G. (2020). Long walk to genomics: History and current approaches to genome sequencing and assembly. *Computational and Structural Biotechnology Journal*, 18, 9-19. doi:<https://doi.org/10.1016/j.csbj.2019.11.002>
- Gilbertson, R., Langan, E., & Mock, T. (2022). Diatoms and Their Microbiomes in Complex and Changing Polar Oceans. *Frontiers in Microbiology*, 13. Retrieved from <https://www.frontiersin.org/articles/10.3389/fmicb.2022.786764>
- Giraud, A., Matic, I., Tenaillon, O., Clara, A., Radman, M., Fons, M., & Taddei, F. (2001). Costs and benefits of high mutation rates: adaptive evolution of bacteria in the mouse gut. *Science*, 291(5513), 2606-2608. doi:10.1126/science.1056421
- Qiu, H., Price, D. C., Weber, A. P. M., Reeb, V., Yang, E. C., Lee, J. M., Kim, S. Y., Yoon, H. S., & Bhattacharya, D. (2013). Adaptation through horizontal gene transfer in the cryptoendolithic red alga *Galdieria phlegrea*. In *Current biology: CB* (Vol. 23, Issue 19, pp. R865-6). <https://doi.org/10.1016/j.cub.2013.08.046>
- Glover, J. N. M., Williams, R. S., & Lee, M. S. (2004). Interactions between BRCT repeats and phosphoproteins: tangled up in two. *Trends in Biochemical Sciences*, 29(11), 579-585. doi:10.1016/j.tibs.2004.09.010
- Godhe, A., & Rynearson, T. (2017). The role of intraspecific variation in the ecological and evolutionary success of diatoms in changing environments. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1728), 20160399.
- Gold, R., Schmied, M., Rothe, G., Zischler, H., Breitschopf, H., Wekerle, H., & Lassmann, H. (1993). Detection of DNA Fragmentation in Apoptosis - Application of *in situ* Nick Translation to Cell-culture Systems and Tissue-Sections. *Journal of Histochemistry & Cytochemistry*, 41(7), 1023-1030. doi:10.1177/41.7.8515045
- Gorczyca, W., Bruno, S., Darzynkiewicz, R. J., Gong, J. P., & Darzynkiewicz, Z. (1992). DNA strand breaks occurring during apoptosis - their early *in situ* detection by the terminal deoxynucleotidyl transferase and nick translation assays and prevention by serine protease inhibitors. *International Journal of Oncology*, 1(6), 639-648. doi:10.3892/ijo.1.6.639
- Gorodetska, I., Kozeretska, I., & Dubrovska, A. (2019). BRCA Genes: The Role in Genome Stability, Cancer Stemness and Therapy Resistance. *J Cancer*, 10(9), 2109-2127. doi:10.7150/jca.30410
- Gostimskaya, I. (2022). CRISPR-Cas9: A History of Its Discovery and Ethical Considerations of Its Use in Genome Editing. *Biochemistry (Mosc)*, 87(8), 777-788. doi:10.1134/s0006297922080090
- Gu, J., Lu, H., Tippin, B., Shimazaki, N., Goodman, M. F., & Lieber, M. R. (2007). XRCC4:DNA ligase IV can ligate incompatible DNA ends and can ligate across gaps. *The EMBO Journal*, 26(4), 1010-1023. doi:<https://doi.org/10.1038/sj.emboj.7601559>
- Gusa, A., & Jinks-Robertson, S. (2019). Mitotic Recombination and Adaptive Genomic Changes in Human Pathogenic Fungi. *Genes (Basel)*, 10(11). doi:10.3390/genes10110901

- Hakem, R., de la Pompa, J. L., & Mak, T. W. (1998). Developmental studies of Brca1 and Brca2 knock-out mice. *J Mammary Gland Biol Neoplasia*, 3(4), 431-445. doi:10.1023/a:1018792200700
- Hassold, T., Hall, H., & Hunt, P. (2007). The origin of human aneuploidy: where we have been, where we are going. *Human Molecular Genetics*, 16(R2), R203-R208. doi:10.1093/hmg/ddm243
- Hegreness, M., & Kishony, R. (2007). Analysis of genetic systems using experimental evolution and whole-genome sequencing. *Genome Biology*, 8(1), 201. doi:10.1186/gb-2007-8-1-201
- Helleday, T., Petermann, E., Lundin, C., Hodgson, B., & Sharma, R. A. (2008). DNA repair pathways as targets for cancer therapy. *Nature Reviews Cancer*, 8(3), 193-204. doi:10.1038/nrc2342
- Henikoff, S., Greene, E. A., Pietrokovski, S., Bork, P., Attwood, T. K., & Hood, L. (1997). Gene families: the taxonomy of protein paralogs and chimeras. *Science*, 278(5338), 609-614. doi:10.1126/science.278.5338.609
- Hennon, G. M. M., Ashworth, J., Groussman, R. D., Berthiaume, C., Morales, R. L., Baliga, N. S., . . . Armbrust, E. V. (2015). Diatom acclimation to elevated CO<sub>2</sub> via cAMP signalling and coordinated gene expression. *Nature Climate Change*, 5(8), 761-765. doi:10.1038/nclimate2683
- Herceg, Z., & Wang, Z.-Q. (2001). Functions of poly(ADP-ribose) polymerase (PARP) in DNA repair, genomic integrity and cell death. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 477(1), 97-110. [https://doi.org/https://doi.org/10.1016/S0027-5107\(01\)00111-7](https://doi.org/https://doi.org/10.1016/S0027-5107(01)00111-7)
- Herndl, G. J., & Reinthaler, T. (2013). Microbial control of the dark end of the biological pump. *Nature Geoscience*, 6(9), 718-724. doi:10.1038/ngeo1921
- Hinz, A. J., Stenzler, B., & Poulain, A. J. (2022). Golden Gate Assembly of Aerobic and Anaerobic Microbial Bioreporters. *Applied and Environmental Microbiology*, 88(1), e0148521. doi:10.1128/aem.01485-21
- Hoeijmakers, J. H. J. (2009). DNA Damage, Aging, and Cancer. *New England Journal of Medicine*, 361(15), 1475-1485. doi:10.1056/NEJMra0804615
- Holsinger, K. E., & Weir, B. S. (2009). FUNDAMENTAL CONCEPTS IN GENETICS. Genetics in geographically structured populations: defining, estimating and interpreting  $F_{ST}$  *Nature Reviews Genetics*, 10(9), 639-650. doi:10.1038/nrg2611
- Hopes, A. (2017). Expanding the molecular toolbox in diatoms: developing a transformation system, CRISPR-Cas and Inverse Yeast-1-hybrid. (July). Retrieved from [https://ueaeprints.uea.ac.uk/66542/1/Final{\\\_}thesis{\\\_}corrections.pdf](https://ueaeprints.uea.ac.uk/66542/1/Final{\_}thesis{\_}corrections.pdf)
- Hopes, A., & Mock, T. (2015). Evolution of Microalgae and Their Adaptations in Different Marine Ecosystems. 1-9. doi:10.1002/9780470015902.a0023744
- Hopes, A., Nekrasov, V., Kamoun, S., & Mock, T. (2016). Editing of the urease gene by CRISPR-Cas in the diatom *Thalassiosira pseudonana*. *Plant Methods*, 12, 49. doi:10.1186/s13007-016-0148-0
- I., M. L. K. a. K. (2004). Evolution of the diatoms: V. Morphological and cytological support for the major clades and a taxonomic revision. *Phycologia*, 43(3), 247-270.
- Jackson, S. P., & Bartek, J. (2009). The DNA-damage response in human biology and disease. *Nature*, 461(7267), 1071-1078. doi:10.1038/nature08467

- Jacobs, A. L., & Schär, P. (2012). DNA glycosylases: in DNA repair and beyond. *Chromosoma*, 121(1), 1-20. doi:10.1007/s00412-011-0347-4
- Jakobsson, M., Edge, M. D., & Rosenberg, N. A. (2013). The relationship between F(ST) and the frequency of the most frequent allele. *Genetics*, 193(2), 515-528. doi:10.1534/genetics.112.144758
- Jiang, X., & Kopp-Schneider, A. (2014). Summarizing EC50 estimates from multiple dose-response experiments: A comparison of a meta-analysis strategy to a mixed-effects model approach. *Biometrical Journal*, 56(3), 493-512. doi:https://doi.org/10.1002/bimj.201300123
- Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., & Charpentier, E. (2012). A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science*, 337(6096), 816-821. doi:10.1126/science.1225829
- Jinks-Robertson, S., Michelitch, M., & Ramcharan, S. (1993). Substrate length requirements for efficient mitotic recombination in *Saccharomyces cerevisiae*. *Molecular and Cellular Biology*.
- Jiricny, J. (2006). The multifaceted mismatch-repair system. *Nature Reviews: Molecular Cell Biology*, 7(5), 335-346. doi:10.1038/nrm1907
- Jöhnk, K. D., Huisman, J., Sharples, J., Sommeijer, B., Bisser, P. M., & Stroom, J. M. (2008). Summer heatwaves promote blooms of harmful cyanobacteria. *Global Change Biology*, 14(3), 495-512. doi:https://doi.org/10.1111/j.1365-2486.2007.01510.x
- Joukov, V., Groen, A. C., Prokhorova, T., Gerson, R., White, E., Rodriguez, A., . . . Livingston, D. M. (2006). The BRCA1/BARD1 heterodimer modulates ran-dependent mitotic spindle assembly. *Cell*, 127(3), 539-552. doi:10.1016/j.cell.2006.08.053
- Karaayvaz-Yildirim, M., Silberman, R. E., Langenbucher, A., Saladi, S. V., Ross, K. N., Zarcaro, E., . . . Ellisen, L. W. (2020). Aneuploidy and a deregulated DNA damage response suggest haploinsufficiency in breast tissues of BRCA2 mutation carriers. *Sci Adv*, 6(5), eaay2611. doi:10.1126/sciadv.aay2611
- Karakaidos, P., Karagiannis, D., & Rampias, T. (2020). Resolving DNA Damage: Epigenetic Regulation of DNA Repair. *Molecules*, 25(11). doi:10.3390/molecules25112496
- Kawecki, T. J., Lenski, R. E., Ebert, D., Hollis, B., Olivieri, I., & Whitlock, M. C. (2012). Experimental evolution. *Trends in Ecology & Evolution*, 27(10), 547-560. doi:https://doi.org/10.1016/j.tree.2012.06.001
- Kemp, M. G., & Sancar, A. (2012). DNA excision repair: where do all the dimers go? *Cell Cycle*, 11(16), 2997-3002. doi:10.4161/cc.21126
- Khalil, A. M. (2020). The genome editing revolution: review. *Journal of Genetic Engineering and Biotechnology*, 18(1), 68. doi:10.1186/s43141-020-00078-y
- Knaus, B. J., & Grünwald, N. J. (2017). vcfr: a package to manipulate and visualize variant call format data in R. *Molecular Ecology Resources*, 17(1), 44-53. doi:https://doi.org/10.1111/1755-0998.12549
- Koehler, D. R., Awadallah, S. S., & Glickman, B. W. (1991). Sites of preferential induction of cyclobutane pyrimidine dimers in the nontranscribed strand of lacI correspond with sites of UV-induced mutation in *Escherichia coli*. *Journal of Biological Chemistry*, 266(18), 11766-11773.

- Kolberg, L., Raudvere, U., Kuzmin, I., Adler, P., Vilo, J., & Peterson, H. (2023). g:Profiler—interoperable web service for functional enrichment analysis and gene identifier mapping (2023 update). *Nucleic Acids Research*, 51(W1), W207-W212. doi:10.1093/nar/gkad347
- Kolodner, R. D., & Marsischky, G. T. (1999). Eukaryotic DNA mismatch repair. *Current Opinion in Genetics & Development*, 9(1), 89-96. doi:10.1016/s0959-437x(99)80013-6
- Kooistra, W. H. C. F., & Medlin, L. K. (1996). Evolution of the Diatoms (Bacillariophyta): IV. A Reconstruction of Their Age from Small Subunit rRNA Coding Regions and the Fossil Record. *Molecular Phylogenetics and Evolution*, 6(3), 391-407. doi:https://doi.org/10.1006/mpev.1996.0088
- Krasovec, M., Sanchez-Brosseau, S., & Piganeau, G. (2019). First Estimation of the Spontaneous Mutation Rate in Diatoms. *Genome Biology and Evolution*, 11(7), 1829-1837. doi:10.1093/gbe/evz130
- Krejci, L., Altmannova, V., Spirek, M., & Zhao, X. (2012). Homologous recombination and its regulation. *Nucleic Acids Research*, 40(13), 5795-5818. doi:10.1093/nar/gks270
- Krogh, B. O., & Symington, L. S. (2004). Recombination proteins in yeast. *Annual Review of Genetics*, 38, 233-271. doi:10.1146/annurev.genet.38.072902.091500
- Krokan, H. E., & Bjørås, M. (2013). Base excision repair. *Cold Spring Harb Perspect Biol*, 5(4), a012583. doi:10.1101/cshperspect.a012583
- Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., . . . Marra, M. A. (2009). Circos: an information aesthetic for comparative genomics. *Genome Research*, 19(9), 1639-1645. doi:10.1101/gr.092759.109
- Kumar, S., Stecher, G., & Tamura, K. (2016). MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Molecular Biology and Evolution*, 33(7), 1870-1874. doi:10.1093/molbev/msw054
- Kumar, K. R. R. (2023). Lost in the bloom: DNA-PKcs in green plants. *Frontiers in Plant Science*, 14. <https://www.frontiersin.org/journals/plant-science/articles/10.3389/fpls.2023.1231678>
- Kusakabe, M., Onishi, Y., Tada, H., Kurihara, F., Kusao, K., Furukawa, M., . . . Sugawara, K. (2019). Mechanism and regulation of DNA damage recognition in nucleotide excision repair. *Genes and Environment*, 41(1), 2. doi:10.1186/s41021-019-0119-6
- Lamb, N. E., Sherman, S. L., & Hassold, T. J. (2005). Effect of meiotic recombination on the production of aneuploid gametes in humans. *Cytogenetic and Genome Research*, 111(3-4), 250-255. doi:10.1159/000086896
- Le, H. P., Heyer, W.-D., & Liu, J. (2021). Guardians of the Genome: BRCA2 and Its Partners. *Genes*, 12(8). <https://doi.org/10.3390/genes12081229>
- Lee, M. E., DeLoache, W. C., Cervantes, B., & Dueber, J. E. (2015). A Highly Characterized Yeast Toolkit for Modular, Multipart Assembly. *ACS Synthetic Biology*, 4(9), 975-986. doi:10.1021/sb500366v
- Lees-Miller, J. P., Cobban, A., Katsonis, P., Bacolla, A., Tsutakawa, S. E., Hammel, et al. (2021). Uncovering DNA-PKcs ancient phylogeny, unique sequence motifs and insights for human disease. *Progress in Biophysics and Molecular Biology*, 163, 87-108. <https://doi.org/10.1016/j.pbiomolbio.2020.09.010>

- Lehmann, A. R. (2011). DNA polymerases and repair synthesis in NER in human cells. *DNA Repair (Amst)*, 10(7), 730-733. doi:10.1016/j.dnarep.2011.04.023
- Levasseur, M., Gosselin, M., & Michaud, S. (1994). A new source of dimethylsulfide (DMS) for the arctic atmosphere: ice diatoms. *Marine Biology*, 121(2), 381-387. doi:10.1007/BF00346748
- Levine, M. S., & Holland, A. J. (2018). The impact of mitotic errors on cell proliferation and tumorigenesis. *Genes & Development*, 32(9-10), 620-638. doi:10.1101/gad.314351.118
- Li, G.-M. (2008). Mechanisms and functions of DNA mismatch repair. *Cell Research*, 18(1), 85-98. doi:10.1038/cr.2007.115
- Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, 27(21), 2987-2993. doi:10.1093/bioinformatics/btr509
- Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14), 1754-1760. doi:10.1093/bioinformatics/btp324
- Li, H., Yang, Y., Hong, W., Huang, M., Wu, M., & Zhao, X. (2020). Applications of genome editing technology in the targeted therapy of human diseases: mechanisms, advances and prospects. *Signal Transduction and Targeted Therapy*, 5(1), 1. doi:10.1038/s41392-019-0089-y
- Li, X., & Heyer, W.-D. (2008). Homologous recombination in DNA repair and DNA damage tolerance. *Cell Research*, 18(1), 99-113. doi:10.1038/cr.2008.1
- Lieber, M. R. (2010). The Mechanism of Double-Strand DNA Break Repair by the Nonhomologous DNA End-Joining Pathway. In R. D. Kornberg, C. R. H. Raetz, J. E. Rothman, & J. W. Thorner (Eds.), *Annual Review of Biochemistry*, Vol 79 (Vol. 79, pp. 181-211).
- Lin, Z., Kong, H., Nei, M., & Ma, H. (2006). Origins and evolution of the recA/RAD51 gene family: evidence for ancient gene duplication and endosymbiotic gene transfer. *Proc Natl Acad Sci U S A*, 103(27), 10328-10333. doi:10.1073/pnas.0604232103
- Lindahl, T. (1974). An *N*-Glycosidase from *Escherichia coli* That Releases Free Uracil from DNA Containing Deaminated Cytosine Residues. *Proceedings of the National Academy of Sciences*, 71(9), 3649-3653. doi:doi:10.1073/pnas.71.9.3649
- Lindahl, T. (1993). Instability and decay of the primary structure of DNA. *Nature*, 362(6422), 709-715.
- Lindahl, T., & Barnes, D. E. (2000). Repair of Endogenous DNA Damage. *Cold Spring Harbor Symposia on Quantitative Biology*, 65, 127-134. doi:10.1101/sqb.2000.65.127
- Liu, D., Keijzers, G., & Rasmussen, L. J. (2017). DNA mismatch repair and its many roles in eukaryotic cells. *Mutation Research/Reviews in Mutation Research*, 773, 174-187. <https://doi.org/https://doi.org/10.1016/j.mrrev.2017.07.001>
- Long, L., Guo, D. D., Gao, W., Yang, W. W., Hou, L. P., Ma, X. N., . . . Song, C. P. (2018). Optimization of CRISPR/Cas9 genome editing in cotton by improved sgRNA expression. *Plant Methods*, 14, 85. doi:10.1186/s13007-018-0353-0
- Lundin, C., North, M., Erixon, K., Walters, K., Jenssen, D., Goldman, A. S. H., & Helleday, T. (2005). Methyl methanesulfonate (MMS) produces heat-labile DNA damage but no detectable in vivo DNA double-strand breaks. *Nucleic Acids Research*, 33(12), 3799-3811. doi:10.1093/nar/gki681

- Lynch, M., & Gabriel, W. (1987). Environmental Tolerance. *The American Naturalist*, 129(2), 283-303. Retrieved from <http://www.jstor.org/stable/2462004>
- Ma, Y., Lu, H., Tippin, B., Goodman, M. F., Shimazaki, N., Koiwai, O., . . . Lieber, M. R. (2004). A Biochemically Defined System for Mammalian Nonhomologous DNA End Joining. *Molecular Cell*, 16(5), 701-713. doi:<https://doi.org/10.1016/j.molcel.2004.11.017>
- Maeda, Y., Kobayashi, R., Watanabe, K., Yoshino, T., Bowler, C., Matsumoto, M., & Tanaka, T. (2022). Chromosome-Scale Genome Assembly of the Marine Oleaginous Diatom *Fistulifera solaris*. *Marine Biotechnology*, 24(4), 788-800. doi:10.1007/s10126-022-10147-7
- Malviya, S., Scalco, E., Audic, S., Vincent, F., Veluchamy, A., Poulain, J., . . . Bowler, C. (2016). Insights into global diatom distribution and diversity in the world's ocean. *Proceedings of the National Academy of Sciences*, 113(11), E1516-E1525. doi:10.1073/pnas.1509523113
- Malyutina, A., Tang, J., & Pessia, A. (2023). drda: An R Package for Dose-Response Data Analysis Using Logistic Functions. *Journal of Statistical Software*, 106(4), 1 - 26. doi:10.18637/jss.v106.i04
- Mancera, E., Bourgon, R., Brozzi, A., Huber, W., & Steinmetz, L. M. (2008). High-resolution mapping of meiotic crossovers and non-crossovers in yeast. *Nature*, 454(7203), 479-485. doi:10.1038/nature07135
- Mann, D. G., & Marchant, H. J. (1990). The Origins Of The Diatom And Its Life Cycle. In J. C. Green, B. S. C. Leadbeater, & W. L. Diver (Eds.), *The Chromophyte Algae: Problems and Perspectives* (pp. 0): Oxford University Press.
- Mann, D. G., & Vanormelingen, P. (2013). An inordinate fondness? The number, distributions, and origins of diatom species. *Journal of Eukaryotic Microbiology*, 60(4), 414-420.
- Mapleson, D., Garcia Accinelli, G., Kettleborough, G., Wright, J., & Clavijo, B. J. (2017). KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. *Bioinformatics*, 33(4), 574-576. doi:10.1093/bioinformatics/btw663
- Marraffini, L. A., & Sontheimer, E. J. (2008). CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. *Science*, 322(5909), 1843-1845. doi:10.1126/science.1165771
- Martin, J. S., Winkelmann, N., Petalcorin, M. I., McIlwraith, M. J., & Boulton, S. J. (2005). RAD-51-dependent and -independent roles of a *Caenorhabditis elegans* BRCA2-related protein during DNA double-strand break repair. *Molecular and Cellular Biology*, 25(8), 3127-3139. doi:10.1128/mcb.25.8.3127-3139.2005
- Martin, W., & Kowallik, K. (1999). Annotated English translation of Mereschkowsky's 1905 paper 'Über Natur und Ursprung der Chromatophoren im Pflanzenreiche'. *European Journal of Phycology*, 34(3), 287-295. doi:10.1080/09670269910001736342
- Martin, W. F., Garg, S., & Zimorski, V. (2015). Endosymbiotic theories for eukaryote origin. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1678), 20140330. doi:10.1098/rstb.2014.0330
- Matsumoto, Y., & Kim, K. (1995). Excision of deoxyribose phosphate residues by DNA polymerase  $\beta$  during DNA repair. *Science*, 269(5224), 699-702.
- McDonald, M. J. (2019). Microbial Experimental Evolution - a proving ground for evolutionary theory and a tool for discovery. *Embo Reports*, 20(8). doi:10.15252/embr.201846992

- McElhinny, S. A. N., & Ramsden, D. A. (2003). Polymerase mu is a DNA-directed DNA/RNA polymerase. *Molecular and Cellular Biology*.
- McElreath, R. (2016). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*: AIC provides a surprisingly simple estimate of the average out-of-sample deviance.
- Medlin, L. (2016). Evolution of the diatoms: Major steps in their evolution and a review of the supporting molecular and morphological evidence (Vol. 55).
- Medlin, L. K. (1997). Evolution of the diatoms-A total approach using molecules, morphology and geology Linda K. Medlin. Paper presented at the Phycologia.
- Mehta, A., & Haber, J. E. (2014). Sources of DNA double-strand breaks and models of recombinational DNA repair. *Cold Spring Harb Perspect Biol*, 6(9), a016428. doi:10.1101/cshperspect.a016428
- Mereschkowsky, C. (1905). Über natur und ursprung der chromatophoren im pflanzenreiche. *Biol Centralbl*, 25, 593.
- Mestiri, I., Norre, F., Gallego, M. E., & White, C. I. (2014). Multiple host-cell recombination pathways act in *Agrobacterium*-mediated transformation of plant cells. *The Plant Journal*, 77(4), 511-520. doi:https://doi.org/10.1111/tpj.12398
- Moon, A. F., Garcia-Diaz, M., Batra, V. K., Beard, W. A., Bebenek, K., Kunkel, T. A., . . . Pedersen, L. C. (2007). The X family portrait: Structural insights into biological functions of X family polymerases. *DNA Repair*, 6(12), 1709-1725. doi:10.1016/j.dnarep.2007.05.009
- Moon, J., Kitty, I., Renata, K., Qin, S., Zhao, F., & Kim, W. (2023). DNA Damage and Its Role in Cancer Therapeutics. *Int J Mol Sci*, 24(5). doi:10.3390/ijms24054741
- Moynahan, M. E., & Jasin, M. (1997). Loss of heterozygosity induced by a chromosomal double-strand break. *Proc Natl Acad Sci U S A*, 94(17), 8988-8993.
- Moynahan, M. E., & Jasin, M. (2010). Mitotic homologous recombination maintains genomic stability and suppresses tumorigenesis. *Nature Reviews Molecular Cell Biology*, 11(3), 196-207. doi:10.1038/nrm2851
- Muller, H. J. (1928). The Measurement of Gene Mutation Rate in *Drosophila*, Its High Variability, and Its Dependence upon Temperature. *Genetics*, 13(4), 279-357. doi:10.1093/genetics/13.4.279
- Muramoto, N., Oda, A., Tanaka, H., Nakamura, T., Kugou, K., Suda, K., . . . Ohta, K. (2018). Phenotypic diversification by enhanced genome restructuring after induction of multiple DNA double-strand breaks. *Nature Communications*, 9(1), 1995. doi:10.1038/s41467-018-04256-y
- Nakov, T., Beaulieu, J. M., & Alverson, A. J. (2018). Accelerated diversification is related to life history and locomotion in a hyperdiverse lineage of microbial eukaryotes (Diatoms, Bacillariophyta). *New Phytologist*, 219(1), 462-473. doi:10.1111/nph.15137
- Nawaly, H., Tsuji, Y., & Matsuda, Y. (2020). Rapid and precise genome editing in a marine diatom, *Thalassiosira pseudonana* by Cas9 nickase (D10A). *Algal Research-Biomass Biofuels and Bioproducts*, 47. doi:10.1016/j.algal.2020.101855
- Nelson, D. M., Tréguer, P., Brzezinski, M. A., Leynaert, A., & Quéguiner, B. (1995). Production and dissolution of biogenic silica in the ocean: Revised global estimates, comparison with regional data

- and relationship to biogenic sedimentation. *Global Biogeochemical Cycles*, 9(3), 359-372.  
doi:<https://doi.org/10.1029/95GB01070>
- Niehaus, A. C., Angilletta, M. J., Jr., Sears, M. W., Franklin, C. E., & Wilson, R. S. (2012). Predicting the physiological performance of ectotherms in fluctuating thermal environments. *Journal of Experimental Biology*, 215(Pt 4), 694-701. doi:10.1242/jeb.058032
- Nik-Zainal, S., & Hall, B. A. (2019). Cellular survival over genomic perfection. *Science*, 366(6467), 802-803. doi:10.1126/science.aax8046
- Nischwitz, E., Schoonenberg, V. A. C., Fradera-Sola, A., Dejung, M., Vydzhak, O., Levin, M., . . . Scheibe, M. (2023). DNA damage repair proteins across the Tree of Life. *iScience*, 26(6), 106778. doi:10.1016/j.isci.2023.106778
- O'Malley, M. A. (2018). The Experimental Study of Bacterial Evolution and Its Implications for the Modern Synthesis of Evolutionary Biology. *Journal of the History of Biology*, 51(2), 319-354. doi:10.1007/s10739-017-9493-8
- Okonechnikov, K., Conesa, A., & García-Alcalde, F. (2016). Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics*, 32(2), 292-294. doi:10.1093/bioinformatics/btv566
- Padfield, D., O'Sullivan, H., & Pawar, S. (2021). rTPC and nls.multstart: A new pipeline to fit thermal performance curves in r. *Methods in Ecology and Evolution*, 12(6), 1138-1143. doi:<https://doi.org/10.1111/2041-210X.13585>
- Pâques, F., & Haber, J. E. (1999). Multiple Pathways of Recombination Induced by Double-Strand Breaks in *Saccharomyces cerevisiae*. *Microbiology and Molecular Biology Reviews*, 63(2), 349-404. doi:10.1128/mubr.63.2.349-404.1999
- Pauly, D., & Christensen, V. (1995). Primary production required to sustain global fisheries. *Nature*, 374(6519), 255-257. doi:10.1038/374255a0
- Pawlyn, C., Loehr, A., Ashby, C., Tytarenko, R., Deshpande, S., Sun, J., . . . Morgan, G. J. (2018). Loss of heterozygosity as a marker of homologous repair deficiency in multiple myeloma: a role for PARP inhibition? *Leukemia*, 32(7), 1561-1566. doi:10.1038/s41375-018-0017-0
- Pazhayam, N. M., Turcotte, C. A., & Sekelsky, J. (2021). Meiotic Crossover Patterning. *Front Cell Dev Biol*, 9, 681123. doi:10.3389/fcell.2021.681123
- Perez-Sepulveda, B. M., Heavens, D., Pulford, C. V., Predeus, A. V., Low, R., Webster, H., . . . The, K. S. G. c. (2021). An accessible, efficient and global approach for the large-scale sequencing of bacterial genomes. *Genome Biology*, 22(1), 349. doi:10.1186/s13059-021-02536-3
- Perina, D., Mikoč, A., Ahel, J., Četković, H., Žaja, R., & Ahel, I. (2014). Distribution of protein poly(ADP-ribosyl)ation systems across all domains of life. *DNA Repair*, 23, 4-16. <https://doi.org/10.1016/j.dnarep.2014.05.003>
- Pfeifer, T. A., Hegedus, D. D., Grigliatti, T. A., & Theilmann, D. A. (1997). Baculovirus immediate-early promoter-mediated expression of the Zeocin resistance gene for use as a dominant selectable marker in dipteran and lepidopteran insect cell lines. *Gene*, 188(2), 183-190. doi:10.1016/s0378-1119(96)00756-1



- Pingoud, A., & Jeltsch, A. (2001). Structure and function of type II restriction endonucleases. *Nucleic Acids Research*, 29(18), 3705-3727. doi:10.1093/nar/29.18.3705
- Pös, O., Radvanszky, J., Buglyó, G., Pös, Z., Rusnakova, D., Nagy, B., & Szemes, T. (2021). DNA copy number variation: Main characteristics, evolutionary significance, and pathological aspects. *Biomed J*, 44(5), 548-559. doi:10.1016/j.bj.2021.02.003
- Potter, S. C., Luciani, A., Eddy, S. R., Park, Y., Lopez, R., & Finn, R. D. (2018). HMMER web server: 2018 update. *Nucleic Acids Research*, 46(W1), W200-w204. doi:10.1093/nar/gky448
- Poulsen, N., Chesley, P. M., & Kröger, N. (2006). Molecular genetic manipulation of the diatom *Thalassiosira pseudonana* (Bacillariophyceae). *Journal of Phycology*, 42(5), 1059-1065. doi:https://doi.org/10.1111/j.1529-8817.2006.00269.x
- Prorok, P., Grin, I. R., Matkarimov, B. T., Ishchenko, A. A., Laval, J., Zharkov, D. O., & Saparbaev, M. (2021). Evolutionary Origins of DNA Repair Pathways: Role of Oxygen Catastrophe in the Emergence of DNA Glycosylases. *Cells*, 10(7), 1591. Retrieved from https://www.mdpi.com/2073-4409/10/7/1591
- Prostova, M., Shilkin, E., Kulikova, A. A., Makarova, A., Ryazansky, S., & Kulbachinskiy, A. (2022). Noncanonical prokaryotic X family DNA polymerases lack polymerase activity and act as exonucleases. *Nucleic Acids Research*, 50(11), 6398-6413. doi:10.1093/nar/gkac461
- Qiu, H., Price, D. C., Weber, A. P. M., Reeb, V., Yang, E. C., Lee, J. M., . . . Bhattacharya, D. (2013). Adaptation through horizontal gene transfer in the cryptoendolithic red alga *Galdieria phlegrea*. In *Current biology : CB* (Vol. 23, Issue 19, pp. R865-6). https://doi.org/10.1016/j.cub.2013.08.046
- Qiu, S., & Huang, J. (2021). MRN complex is an essential effector of DNA damage repair. *J Zhejiang Univ Sci B*, 22(1), 31-37. doi:10.1631/jzus.B2000289
- Ramadan, K., Shevelev, I. V., Maga, G., & Hübscher, U. (2004). De novo DNA synthesis by human DNA polymerase  $\lambda$ , DNA polymerase  $\mu$  and terminal deoxyribonucleotidyl transferase. *Journal of Molecular Biology*, 339(2), 395-404.
- Rasheed, A., Barqawi, A. A., Mahmood, A., Nawaz, M., Shah, A. N., Bay, D. H., . . . Qari, S. H. (2022). CRISPR/Cas9 is a powerful tool for precise genome editing of legume crops: a review. *Molecular Biology Reports*, 49(6), 5595-5609. doi:10.1007/s11033-022-07529-4
- Ries, G., Heller, W., Puchta, H., Sandermann, H., Seidlitz, H. K., & Hohn, B. (2000). Elevated UV-B radiation reduces genome stability in plants. *Nature*, 406(6791), 98-101. doi:10.1038/35017595
- Rosenberg, S. M. (1997). Mutation for survival. *Current Opinion in Genetics & Development*, 7(6), 829-834. doi:https://doi.org/10.1016/S0959-437X(97)80047-0
- Ross, R., Cox, E. J., Karayeva, N., Mann, D., Paddock, T., Simonsen, R., & Sims, P. (1979). An amended terminology for the siliceous components of the diatom cell.
- Rossi, M. J., DiDomenico, S. F., Patel, M., & Mazin, A. V. (2021). RAD52: Paradigm of Synthetic Lethality and New Developments. *Frontiers in Genetics*, 12. doi:10.3389/fgene.2021.780293
- Round, F. E., Crawford, R. M., & Mann, D. G. (1990). *Diatoms: biology and morphology of the genera*: Cambridge university press.

- Ruck, E. C., & Theriot, E. C. (2011). Origin and Evolution of the Canal Raphe System in Diatoms. *Protist*, 162(5), 723-737. doi:<https://doi.org/10.1016/j.protis.2011.02.003>
- Saitou, N., & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4), 406-425. doi:[10.1093/oxfordjournals.molbev.a040454](https://doi.org/10.1093/oxfordjournals.molbev.a040454)
- Samstein, R. M., Krishna, C., Ma, X., Pei, X., Lee, K. W., Makarov, V., . . . Riaz, N. (2021). Mutations in BRCA1 and BRCA2 differentially affect the tumor microenvironment and response to checkpoint blockade immunotherapy. *Nat Cancer*, 1(12), 1188-1203. doi:[10.1038/s43018-020-00139-8](https://doi.org/10.1038/s43018-020-00139-8)
- Sancar, A. (1996). DNA excision repair. *Annual Review of Biochemistry*, 65(1), 43-81. doi:[10.1146/annurev.bi.65.070196.000355](https://doi.org/10.1146/annurev.bi.65.070196.000355)
- Sancar, A. (2016). Mechanisms of DNA Repair by Photolyase and Excision Nuclease (Nobel Lecture). *Angewandte Chemie International Edition*, 55(30), 8502-8527. doi:<https://doi.org/10.1002/anie.201601524>
- Sánchez, C., Cristóbal, G., & Bueno, G. (2019). Diatom identification including life cycle stages through morphological and texture descriptors. *PeerJ*, 7, e6770. doi:[10.7717/peerj.6770](https://doi.org/10.7717/peerj.6770)
- Santaguida, S., & Amon, A. (2015). Short- and long-term effects of chromosome mis-segregation and aneuploidy. *Nature Reviews Molecular Cell Biology*, 16(8), 473-485. doi:[10.1038/nrm4025](https://doi.org/10.1038/nrm4025)
- Schärer, O. D. (2013). Nucleotide excision repair in eukaryotes. *Cold Spring Harb Perspect Biol*, 5(10), a012609. doi:[10.1101/cshperspect.a012609](https://doi.org/10.1101/cshperspect.a012609)
- Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., . . . Cardona, A. (2012). Fiji: an open-source platform for biological-image analysis. *Nature Methods*, 9(7), 676-682. doi:[10.1038/nmeth.2019](https://doi.org/10.1038/nmeth.2019)
- Schirrmeister, B. E., Gugger, M., & Donoghue, P. C. (2015). Cyanobacteria and the Great Oxidation Event: evidence from genes and fossils. *Palaeontology*, 58(5), 769-785. doi:[10.1111/pala.12178](https://doi.org/10.1111/pala.12178)
- Schmidt, K. (2017). Thermal adaptation of *Thalassiosira pseudonana* using experimental evolution approaches. (Doctoral). University of East Anglia, (uea:63741)
- Schorsch, C., Köhler, T., & Boles, E. (2009). Knockout of the DNA ligase IV homolog gene in the sphingoid base producing yeast *Pichia ciferrii* significantly increases gene targeting efficiency. *Current Genetics*, 55(4), 381–389. <https://doi.org/10.1007/s00294-009-0252-z>
- Serôdio, J., & Lavaud, J. (2020). Diatoms and Their Ecological Importance. In W. Leal Filho, A. M. Azul, L. Brandli, A. Lange Salvia, & T. Wall (Eds.), *Life Below Water* (pp. 1-9). Cham: Springer International Publishing.
- Shahid, T., Soroka, J., Kong, E., Malivert, L., McIlwraith, M. J., Pape, T., . . . Zhang, X. (2014). Structure and mechanism of action of the BRCA2 breast cancer tumor suppressor. *Nature Structural & Molecular Biology*, 21(11), 962-968. doi:[10.1038/nsmb.2899](https://doi.org/10.1038/nsmb.2899)
- Sharan, S. K., Morimatsu, M., Albrecht, U., Lim, D. S., Regel, E., Dinh, C., . . . Bradley, A. (1997). Embryonic lethality and radiation hypersensitivity mediated by Rad51 in mice lacking *Brca2*. *Nature*, 386(6627), 804-810. doi:[10.1038/386804a0](https://doi.org/10.1038/386804a0)

- Shcherbakova, A., Preller, M., Taft, M. H., Pujols, J., Ventura, S., Tiemann, B., . . . Bakker, H. (2019). C-mannosylation supports folding and enhances stability of thrombospondin repeats. *eLife*, 8, e52978. doi:10.7554/eLife.52978
- Shendure, J., Porreca, G. J., Reppas, N. B., Lin, X., McCutcheon, J. P., Rosenbaum, A. M., . . . Church, G. M. (2005). Accurate Multiplex Polony Sequencing of an Evolved Bacterial Genome. *Science*, 309(5741), 1728-1732. doi:doi:10.1126/science.1117389
- Shrestha, R. P., & Hildebrand, M. (2017). Development of a silicon limitation inducible expression system for recombinant protein production in the centric diatoms *Thalassiosira pseudonana* and *Cyclotella cryptica*. *Microbial Cell Factories*, 16. doi:10.1186/s12934-017-0760-3
- Simon, N., Cras, A.-L., Foulon, E., & Lemée, R. (2009). Diversity and evolution of marine phytoplankton. *Comptes Rendus Biologies*, 332(2), 159-170. doi:https://doi.org/10.1016/j.crv.2008.09.009
- Sims, P., Mann, D., & Medlin, L. (2006). Evolution of the diatoms: Insights from fossil, biological and molecular data (Vol. 45).
- Sirbu, B. M., & Cortez, D. (2013). DNA damage response: three levels of DNA repair regulation. *Cold Spring Harb Perspect Biol*, 5(8), a012724. doi:10.1101/cshperspect.a012724
- Skipper, H. E., Schabel, F. M., Jr., & Wilcox, W. S. (1964). Experimental evaluation of potential anticancer agents. xiii. on the criteria and kinetics associated with "curability" of experimental leukemia. *Cancer Chemother Rep*, 35, 1-111.
- Sokal, R. R., & Michener, C. D. (1958). A statistical method for evaluating systematic relationships. *University of Kansas science bulletin*, 38, 1409-1438.
- Sprouffske, K., & Wagner, A. (2016). Growthcurver: an R package for obtaining interpretable metrics from microbial growth curves. *BMC Bioinformatics*, 17(1), 172. doi:10.1186/s12859-016-1016-7
- Spugnesi, L., Balia, C., Collavoli, A., Falaschi, E., Quercioli, V., Caligo, M. A., & Galli, A. (2013). Effect of the expression of BRCA2 on spontaneous homologous recombination and DNA damage-induced nuclear foci in *Saccharomyces cerevisiae*. *Mutagenesis*, 28(2), 187-195. doi:10.1093/mutage/ges069
- Steele, C. D., Abbasi, A., Islam, S. M. A., Bowes, A. L., Khandekar, A., Haase, K., . . . Pillay, N. (2022). Signatures of copy number alterations in human cancer. *Nature*, 606(7916), 984-991. doi:10.1038/s41586-022-04738-6
- Stewart, M. D., Merino Vega, D., Arend, R. C., Baden, J. F., Barbash, O., Beaubier, N., . . . Allen, J. (2022). Homologous Recombination Deficiency: Concepts, Definitions, and Assays. *Oncologist*, 27(3), 167-174. doi:10.1093/oncolo/oyab053
- Storchova, Z. (2018). Evolution of aneuploidy: overcoming the original CIN. *Genes & Development*, 32(23-24), 1459-1460. doi:10.1101/gad.321810.118
- Sugiyama, T., Zaitseva, E. M., & Kowalczykowski, S. C. (1997). A single-stranded DNA-binding protein is needed for efficient presynaptic complex formation by the *Saccharomyces cerevisiae* Rad51 protein. *Journal of Biological Chemistry*, 272(12), 7940-7945.
- Sui, Y., Qi, L., Wu, J. K., Wen, X. P., Tang, X. X., Ma, Z. J., . . . Petes, T. D. (2020). Genome-wide mapping of spontaneous genetic alterations in diploid yeast cells. *Proc Natl Acad Sci U S A*, 117(45), 28191-28200. doi:10.1073/pnas.2018633117

- Sullivan, M. R., & Bernstein, K. A. (2018). RAD-ical New Insights into RAD51 Regulation. In *Genes* (Vol. 9, Issue 12). <https://doi.org/10.3390/genes9120629>
- Symington, L. S., Rothstein, R., & Lisby, M. (2014). Mechanisms and Regulation of Mitotic Recombination in *Saccharomyces cerevisiae*. *Genetics*, 198(3), 795-835. doi:10.1534/genetics.114.166140
- Talpaert-Borlè, M. (1987). Formation, detection and repair of AP sites. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 181(1), 45-56. doi:[https://doi.org/10.1016/0027-5107\(87\)90286-7](https://doi.org/10.1016/0027-5107(87)90286-7)
- Tamura, K., Stecher, G., & Kumar, S. (2021). MEGA11: Molecular Evolutionary Genetics Analysis Version 11. *Molecular Biology and Evolution*, 38(7), 3022-3027. doi:10.1093/molbev/msab120
- Tesson, B., Lerch, S. J. L., & Hildebrand, M. (2017). Characterization of a New Protein Family Associated With the Silica Deposition Vesicle Membrane Enables Genetic Manipulation of Diatom Silica. *Scientific Reports*, 7(1), 13457. doi:10.1038/s41598-017-13613-8
- Thompson, S. L., Bakhoun, S. F., & Compton, D. A. (2010). Mechanisms of chromosomal instability. *Current Biology*, 20(6), R285-295. doi:10.1016/j.cub.2010.01.034
- Tripathi, A. K., Pareek, A., & Singla-Pareek, S. L. (2017). TUNEL Assay to Assess Extent of DNA Fragmentation and Programmed Cell Death in Root Cells under Various Stress Conditions. *Bio Protoc*, 7(16), e2502. doi:10.21769/BioProtoc.2502
- Tubbs, A., & Nussenzweig, A. (2017). Endogenous DNA Damage as a Source of Genomic Instability in Cancer. *Cell*, 168(4), 644-656. doi:10.1016/j.cell.2017.01.002
- Turner, T. L., Stewart, A. D., Fields, A. T., Rice, W. R., & Tarone, A. M. (2011). Population-based resequencing of experimentally evolved populations reveals the genetic basis of body size variation in *Drosophila melanogaster*. *PLoS Genetics*, 7(3), e1001336. doi:10.1371/journal.pgen.1001336
- Tutt, A., Bertwistle, D., Valentine, J., Gabriel, A., Swift, S., Ross, G., . . . Ashworth, A. (2001). Mutation in *Brca2* stimulates error-prone homology-directed repair of DNA double-strand breaks occurring between repeated sequences. *The EMBO Journal*, 20(17), 4704-4716. doi:10.1093/emboj/20.17.4704
- Uchiyama, Y., Kimura, S., Yamamoto, T., Ishibashi, T., & Sakaguchi, K. (2004). Plant DNA polymerase  $\lambda$ , a DNA repair enzyme that functions in plant meristematic and meiotic tissues. *European Journal of Biochemistry*, 271(13), 2799-2807. doi:<https://doi.org/10.1111/j.1432-1033.2004.04214.x>
- Uchiyama, Y., Takeuchi, R., Kodera, H., & Sakaguchi, K. (2009). Distribution and roles of X-family DNA polymerases in eukaryotes. *Biochimie*, 91(2), 165-170. doi:<https://doi.org/10.1016/j.biochi.2008.07.005>
- Udroiu, I. (2020). Is the number of DNA repair genes associated with evolution rate and size of genomes? *Human Genomics*, 14(1), 12. doi:10.1186/s40246-020-00259-3
- Vance, T. D. R., Bayer-Giraldi, M., Davies, P. L., & Mangiagalli, M. (2019). Ice-binding proteins and the 'domain of unknown function' 3494 family. *FEBS Journal*, 286(5), 855-873. doi:10.1111/febs.14764
- Varki, A. (2016). Biological roles of glycans. *Glycobiology*, 27(1), 3-49. doi:10.1093/glycob/cww086

- Vaughn, C. M., & Sancar, A. (2020). Mechanisms and Maps of Nucleotide Excision Repair. In M. Dizdaroglu, R. S. Lloyd, M. Dizdaroglu, & R. S. Lloyd (Eds.), *DNA Damage, DNA Repair and Disease: Volume 2* (pp. 0): The Royal Society of Chemistry.
- Venkitaraman, A. R. (2014). Cancer suppression by the chromosome custodians, BRCA1 and BRCA2. *Science*, 343(6178), 1470-1475. doi:10.1126/science.1252230
- Vincent, M. S., & Uphoff, S. (2021). Cellular heterogeneity in DNA alkylation repair increases population genetic plasticity. *Nucleic Acids Research*, 49(21), 12320-12331. doi:10.1093/nar/gkab1143
- Volkova, N. V, Meier, B., González-Huici, V., Bertolini, S., Gonzalez, S., Vöhringer, H., Abascal, F., Martincorena, I., Campbell, P. J., Gartner, A., & Gerstung, M. (2020). Mutational signatures are jointly shaped by DNA damage and repair. *Nature Communications*, 11(1), 2169. <https://doi.org/10.1038/s41467-020-15912-7>
- Von Dassow, P., Petersen, T. W., Chepurinov, V. A., & Virginia Armbrust, E. (2008). Inter- and intraspecific relationships between nuclear DNA content and cell size in selected members of the centric diatom genus *Thalassiosira* (Bacillariophyceae). *Journal of Phycology*, 44(2), 335-349. doi:https://doi.org/10.1111/j.1529-8817.2008.00476.x
- Voskarides, K., Dweep, H., & Chrysostomou, C. (2019). Evidence that DNA repair genes, a family of tumor suppressor genes, are associated with evolution rate and size of genomes. *Human Genomics*, 13(1), 26. doi:10.1186/s40246-019-0210-x
- Waldvogel, A. M., & Pfenninger, M. (2021). Temperature dependence of spontaneous mutation rates. *Genome Research*, 31(9), 1582-1589. doi:10.1101/gr.275168.120
- Walker, J. R., Corpina, R. A., & Goldberg, J. (2001). Structure of the Ku heterodimer bound to DNA and its implications for double-strand break repair. *Nature*, 412(6847), 607-614. doi:10.1038/35088000
- Weber, E., Engler, C., Gruetzner, R., Werner, S., & Marillonnet, S. (2011). A Modular Cloning System for Standardized Assembly of Multigene Constructs. *PLoS One*, 6(2), e16765. doi:10.1371/journal.pone.0016765
- Weir, B. S., & Cockerham, C. C. (1984). Estimating f-statistics for the analysis of population structure. *Evolution*, 38(6), 1358-1370. doi:10.1111/j.1558-5646.1984.tb05657.x
- Wilson, T. E., Grawunder, U., & Lieber, M. R. (1997). Yeast DNA ligase IV mediates non-homologous DNA end joining. *Nature*, 388(6641), 495-498.
- Wood, R. D. (1997). Nucleotide excision repair in mammalian cells. *Journal of Biological Chemistry*, 272(38), 23465-23468. doi:10.1074/jbc.272.38.23465
- Wooster, R., Bignell, G., Lancaster, J., Swift, S., Seal, S., Mangion, J., . . . Micklem, G. (1995). Identification of the breast cancer susceptibility gene BRCA2. *Nature*, 378(6559), 789-792. doi:10.1038/378789a0
- Wright, S. (1951). The genetical structure of populations. *Ann Eugen*, 15(4), 323-354. doi:10.1111/j.1469-1809.1949.tb02451.x

- Wright, W. D., Shah, S. S., & Heyer, W.-D. (2018). Homologous recombination and the repair of DNA double-strand breaks. *Journal of Biological Chemistry*, 293(27), 10524-10535. doi:<https://doi.org/10.1074/jbc.TM118.000372>
- Xiao, W., Chow, B. L., & Rathgeber, L. (1996). The repair of DNA methylation damage in *Saccharomyces cerevisiae*. *Current Genetics*, 30(6), 461-468. doi:10.1007/s002940050157
- Xu, J., Lahiri, I., Wang, W., Wier, A., Cianfrocco, M. A., Chong, J., . . . Wang, D. (2017). Structural basis for the initiation of eukaryotic transcription-coupled DNA repair. *Nature*, 551(7682), 653-657. doi:10.1038/nature24658
- Xu, X., Weaver, Z., Linke, S. P., Li, C., Gotay, J., Wang, X.-W., . . . Deng, C.-X. (1999). Centrosome Amplification and a Defective G2–M Cell Cycle Checkpoint Induce Genetic Instability in BRCA1 Exon 11 Isoform–Deficient Cells. *Molecular Cell*, 3(3), 389-395. doi:[https://doi.org/10.1016/S1097-2765\(00\)80466-9](https://doi.org/10.1016/S1097-2765(00)80466-9)
- Yamtich, J., & Sweasy, J. B. (2010). DNA polymerase Family X: Function, structure, and cellular roles. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, 1804(5), 1136-1150. doi:<https://doi.org/10.1016/j.bbapap.2009.07.008>
- Yang, S., Wang, L., Huang, J., Zhang, X., Yuan, Y., Chen, J.-Q., . . . Tian, D. (2015). Parent–progeny sequencing indicates higher mutation rates in heterozygotes. *Nature*, 523(7561), 463-467. doi:10.1038/nature14649
- Yang, W. (2000). Structure and function of mismatch repair proteins. *Mutation Research*, 460(3-4), 245-256. doi:10.1016/s0921-8777(00)00030-6
- Yi, X., Chi, T., Li, Z., Wang, J., Yu, M., Wu, M., & Zhou, H. (2019). Combined effect of polystyrene plastics and triphenyltin chloride on the green algae *Chlorella pyrenoidosa*. *Environmental Science and Pollution Research*, 26(15), 15011-15018. doi:10.1007/s11356-019-04865-0
- Yim, E., O'Connell, K. E., St. Charles, J., & Petes, T. D. (2014). High-Resolution Mapping of Two Types of Spontaneous Mitotic Gene Conversion Events in *Saccharomyces cerevisiae*. *Genetics*, 198(1), 181-192. doi:10.1534/genetics.114.167395
- Yin, Y., & Petes, T. D. (2013). Genome-Wide High-Resolution Mapping of UV-Induced Mitotic Recombination Events in *Saccharomyces cerevisiae*. *PLoS Genetics*, 9(10), e1003894. doi:10.1371/journal.pgen.1003894
- Yona, A. H., Manor, Y. S., Herbst, R. H., Romano, G. H., Mitchell, A., Kupiec, M., . . . Dahan, O. (2012). Chromosomal duplication is a transient evolutionary solution to stress. *Proc Natl Acad Sci U S A*, 109(51), 21010-21015. doi:10.1073/pnas.1211150109
- Zámborszky, J., Szikriszt, B., Gervai, J. Z., Pipek, O., Póti, Á., Krzystanek, M., . . . Szűts, D. (2017). Loss of BRCA1 or BRCA2 markedly increases the rate of base substitution mutagenesis and has distinct effects on genomic deletions. *Oncogene*, 36(6), 746-755. doi:10.1038/onc.2016.243
- Zamenhof, S., & Greer, S. (1958). Heat as an Agent producing High Frequency of Mutations and Unstable Genes in *Escherichia coli*. *Nature*, 182(4635), 611-613. doi:10.1038/182611a0
- Zhou, W., Zhang, F., Chen, X., Shen, Y., Lupski, J. R., & Jin, L. (2013). Increased genome instability in human DNA segments with self-chains: homology-induced structural variations via replicative mechanisms. *Human Molecular Genetics*, 22(13), 2642-2651. doi:10.1093/hmg/ddt113

Zuckerkandl, E., & Pauling, L. (1965). Molecules as documents of evolutionary history. *Journal of Theoretical Biology*, 8(2), 357-366. doi:10.1016/0022-5193(65)90083-4

# Appendix

## Supplementary Information for Chapter 3

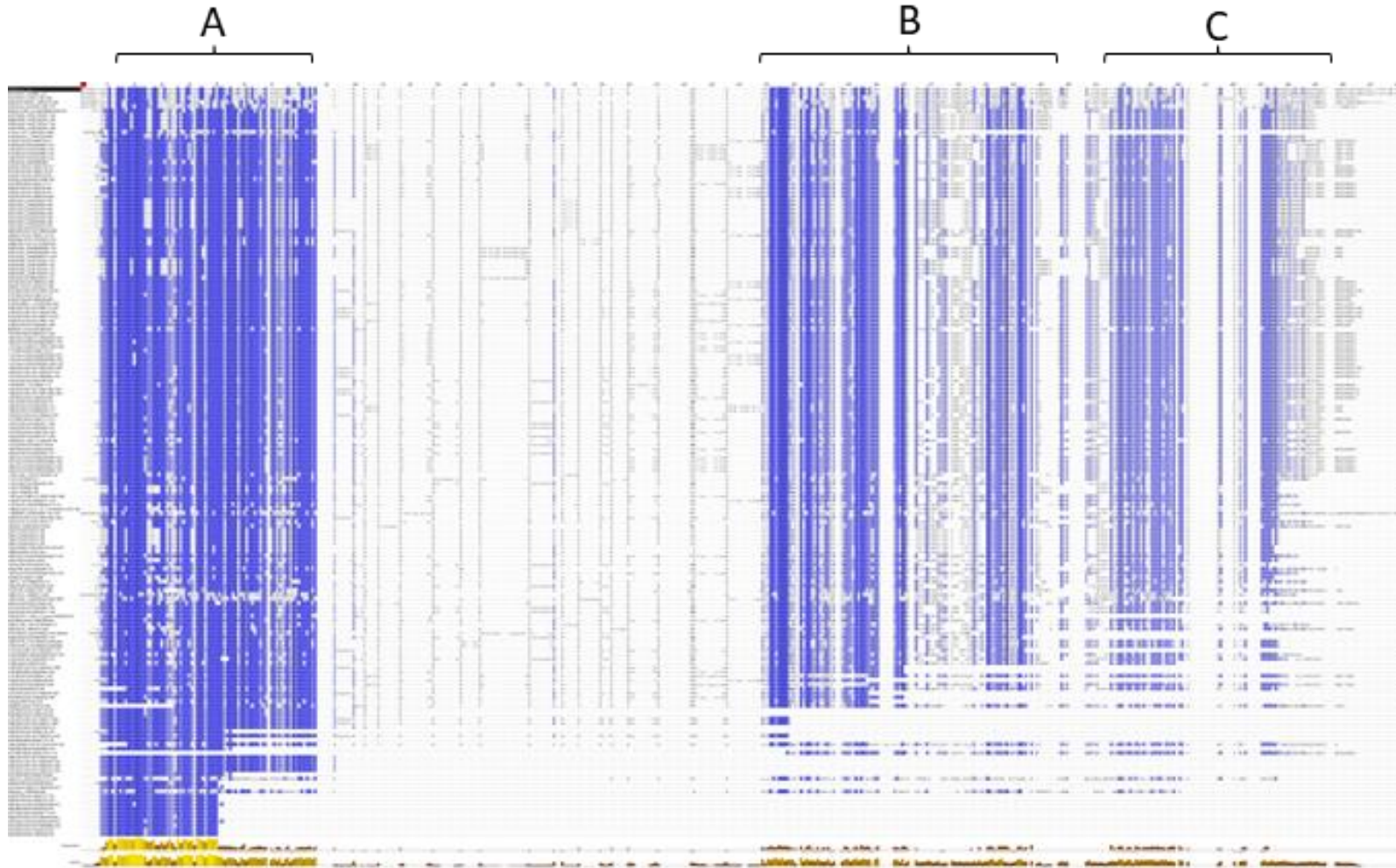




Figure A.1. Alignment of 1301 BRCA2 sequences resulting from a phmmer search against the UniProt reference proteomes using the protein sequence of THAPS\_263089 as the query sequence. Conservation of residues is shown by colour, with darker shades of blue having a higher conservation. A) BRCA2\_Helical Domain, B) BRCA2\_2\_OB1 fold, C) BRCA2\_2\_OB2 fold. TpBRCA2 sequence is on the first line of the alignment. Aligned sequences were retrieved from a blasp search on NCBI. Each row is a unique BRCA2 sequence. The final row (gold) shows the degree of conservation across each site.

## Supplementary Information for Chapter 4

Table A.1. Spectrophotometry results of gDNA sent to EI for sequencing.

<b>Culture ID</b>	<b>260/280</b>	<b>260/230</b>	<b>ng/<math>\mu</math>l</b>
<i>Brca2-09</i>	1.90	2.03	64.06
<i>Brca2-10</i>	1.86	2.07	111.61
<i>Brca2-11</i>	1.83	19.93	82.91

Table A.2. Overview of mutations called via *bcftools call* (Li, 2011) and annotation with *SnPEff* (Cingolani et al., 2012). Number of records = total number of variants called; SNPs = single nucleotide polymorphisms, indels = small insertions and deletions. Data generated via *bcftools stats*.

	<b><i>Brca2-09</i></b>	<b><i>Brca2-10</i></b>	<b><i>Brca2-11</i></b>	<b>Wild Type</b>
Number of records	243401	242072	242667	244712
Number of SNPs	220374	219576	220103	220374
Number of indels	23027	22496	22564	24338
Number of multiallelic sites	954	937	960	1096
Number of multiallelic SNP sites	44	49	50	46

Table A.3. Overview of unique variants called via *bcftools call* (Li, 2011) and annotation with *SnpEff* (Cingolani et al., 2012) by type reported for wild type and *Brca2* knock-out cultures (*Brca2-09*, *Brca2-10*, and *Brca2-11*). Number of records = total number of variants called; SNPs = single nucleotide polymorphisms, indels = small insertions and deletions. Data generated via *bcftools stats*.

	<b>Wild Type</b>	<b><i>Brca2-09</i></b>	<b><i>Brca2-10</i></b>	<b><i>Brca2-11</i></b>
number of SNPs	2810	7849	5607	4900
number of indels	747	1447	935	896
number of multiallelic sites	26	24	18	22
number of multiallelic SNP sites	8	19	16	19

Table A.4. Protein ID, chromosome and KOG annotation of enriched genes with an increased copy number variation in *Brca2-09*

Protein Id	Chromosome	KOG Description
269534	chr_10	Transcription elongation factor SPT6
261829	chr_3	ATP-dependent DNA helicase
263669	chr_10	3'-5' DNA helicase
264865	chr_22	Predicted RNA-binding protein containing PIN domain and involved in translation or RNA processing
261894	chr_3	Transcription factor, Myb superfamily
268331	chr_2	WD40-repeat-containing subunit of the 18S rRNA processing complex
262266	chr_4	Myb-like DNA-binding domain
263002	chr_6	Heat shock transcription factor
263630	chr_10	Putative tRNA binding domain
269871	chr_16a	5'-3' exonuclease
269503	chr_10	DNA-binding transcription factor activity
263720	chr_10	Sister chromatid cohesion complex Cohesin, subunit RAD21/SCC1
269474	chr_10	Heat shock transcription factor
263644	chr_10	DNA topoisomerase type II
268689	chr_4	DNA topoisomerase type II
264853	chr_22	Transcription factor, Myb superfamily
268221	chr_2	Myb-like DNA-binding domain
41829	chr_10	Translation elongation factor EF-1 alpha/Tu
36454	chr_10	50S ribosomal protein L1
36811	chr_12	Histones H3 and H4
42599	chr_16a	RNA Helicase
36339	chr_10	Isoleucyl-tRNA synthetase
36502	chr_10	Predicted mRNA cap-binding protein related to eIF-4E
39953	chr_3	ATP-dependent RNA helicase pitchoune
36434	chr_10	Uncharacterized conserved protein
36420	chr_10	DNA topoisomerase type II (double strand cut, ATP-hydrolyzing) activity
36506	chr_10	tRNA pseudouridine synthase
36511	chr_10	tRNA pseudouridylate synthases
36345	chr_10	Regulation of DNA-templated transcription Box H/ACA snoRNP component, involved in ribosomal RNA pseudouridylation
37442	chr_14	
37795	chr_16a	Predicted alpha-helical protein, potentially involved in replication/repair
35041	chr_6	Predicted alpha-helical protein, potentially involved in replication/repair
36522	chr_10	DNA topoisomerase type II
36366	chr_10	Apurinic/apyrimidinic endonuclease and related enzymes
37786	chr_16a	DEAD/DEAH box helicase
39924	chr_3	DNA topoisomerase type II
25168	chr_16a	Uncharacterized
33161	chr_3	Mitochondrial polypeptide chain release factor
27352	chr_3	Core histone H2A/H2B/H3/H4
24189	chr_10	Transcriptional activator FOSB/c-Fos and related bZIP transcription factors

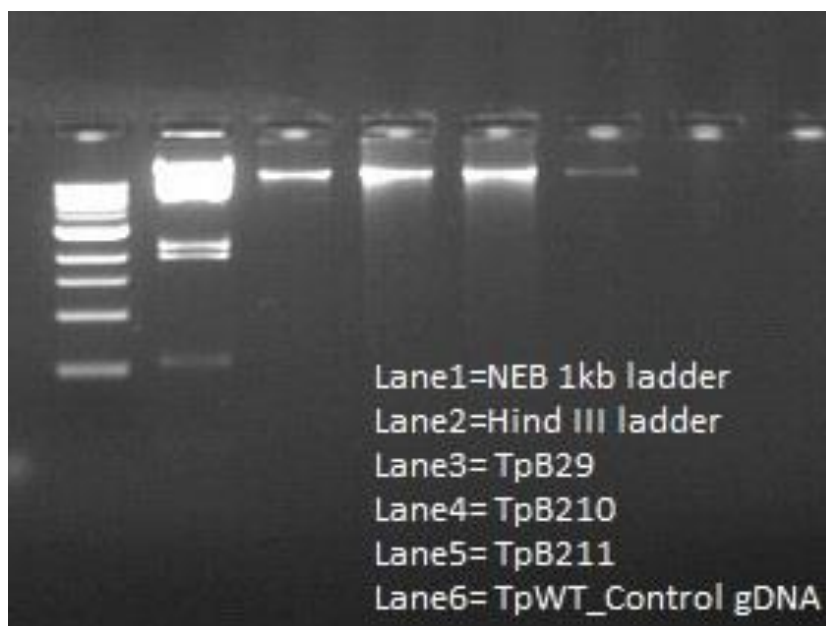
*Table A.4*  
*continued from*  
*previous page*

31404	chr_1	5'-3' exonuclease XRN1/KEM1/SEP1 involved in DNA strand exchange and mRNA turnover
24112	chr_10	Heat shock transcription factor
21432	chr_2	FOG: RRM domain
24120	chr_10	Nucleolar GTPase/ATPase p130
24514	chr_12	Uncharacterized conserved protein
25187	chr_16b	Protein kinase ATM/Tel1, involved in telomere length regulation and DNA repair
21989	chr_3	Transcriptional activator FOSB/c-Fos and related bZIP transcription factors
24163	chr_10	phenylalanine---tRNA ligase
24118	chr_10	Zuotin and related molecular chaperones (DnaJ superfamily), contains DNA-binding domains
25120	chr_16a	HAT (Half-A-TPR) repeat-containing protein
21026	chr_1	Splicing factor 3a, subunit 1
24106	chr_10	regulation of DNA-templated transcription
21706	chr_3	Exosomal 3'-5' exoribonuclease complex, subunit Rrp4
24178	chr_10	Transcription factor MEIS1 and related HOX domain proteins
25124	chr_16a	Uncharacterized
25140	chr_16a	regulation of DNA-templated transcription
21959	chr_3	FOG: RRM domain
25144	chr_16a	Transcription initiation factor TFIID, subunit BDF1 and related bromodomain proteins
21973	chr_3	Telomerase elongation inhibitor/RNA maturation protein PINX1
23025	chr_6	60S ribosomal protein L10A
20814	chr_1	60S ribosomal protein L10A
21929	chr_3	DNA methylation
25142	chr_16a	Chromatin remodelling factor subunit and related transcription factors
24126	chr_10	Uncharacterized conserved protein
22120	chr_4	DNA-binding transcription factor activity
21882	chr_3	DNA mismatch repair protein - MLH3 family
21941	chr_3	Chromatin-associated protein Dek and related proteins, contains SAP DNA binding domain
19897	chr_22	Transcription factor, Myb superfamily
17845	chr_3	MOT2 transcription factor
16039	chr_10	Mismatch repair ATPase MSH5 (MutS family)
11943	chr_22	Uncharacterized
11305	chr_19b_31	Uncharacterized
10402	chr_16a	mRNA splicing factor ATP-dependent RNA helicase
10352	chr_16a	Nucleolar GTPase/ATPase p130
10348	chr_16a	Uncharacterized
10349	chr_16a	Transcription initiation factor IIE, alpha subunit
10360	chr_16a	Transcriptional activator FOSB/c-Fos and related bZIP transcription factors
10336	chr_16a	Uncharacterized
8535	chr_10	Splicing coactivator SRm160/300, subunit SRm160 (contains PWI domain)

*Table A.4*  
*continued from*  
*previous page*

8507	chr_10	Splicing coactivator SRm160/300, subunit SRm160 (contains PWI domain) Transcription elongation factor TFIIIS/Cofactor of enhancer-binding protein
8433	chr_10	Sp1
8521	chr_10	Pseudouridylate synthase
8541	chr_10	Transcriptional coactivator CAPER (RRM superfamily)
8542	chr_10	Replication factor C, subunit RFC1 (large subunit)
8405	chr_10	von Willebrand factor and related coagulation proteins
8552	chr_10	1-phosphatidylinositol-3-phosphate 5-kinase
8259	chr_10	Nuclear receptor coregulator SMRT/SMRTER, contains Myb-like domains
8410	chr_10	DNA binding
8497	chr_10	Transcriptional activator FOSB/c-Fos and related bZIP transcription factors Heterochromatin-associated protein HP1 and related CHROMO domain proteins
6012	chr_6	
4041	chr_3	Centromere-associated protein HEC1
3973	chr_3	Transcription factor E2F/dimerization partner (TDP)-like proteins
3197	chr_3	Heat shock transcription factor
3184	chr_3	Histone H4
2728	chr_2	Nucleosome-binding factor SPN, POB3 subunit
3698	chr_3	Heat shock transcription factor
3719	chr_3	Predicted ATP-dependent RNA helicase
2157	chr_2	Uncharacterized conserved protein
3030	chr_2	Integrase core domain; DNA Recombination
3702	chr_3	Uncharacterized
3349	chr_3	Checkpoint 9-1-1 complex, HUS1 component
3124	chr_3	Chromosome condensation complex Condensin, subunit H

---



*Figure A.2. Pulse-gel electrophoresis results of gDNA sent to EI for sequencing. Samples run on 0.8% agarose gel for 1 hour.*

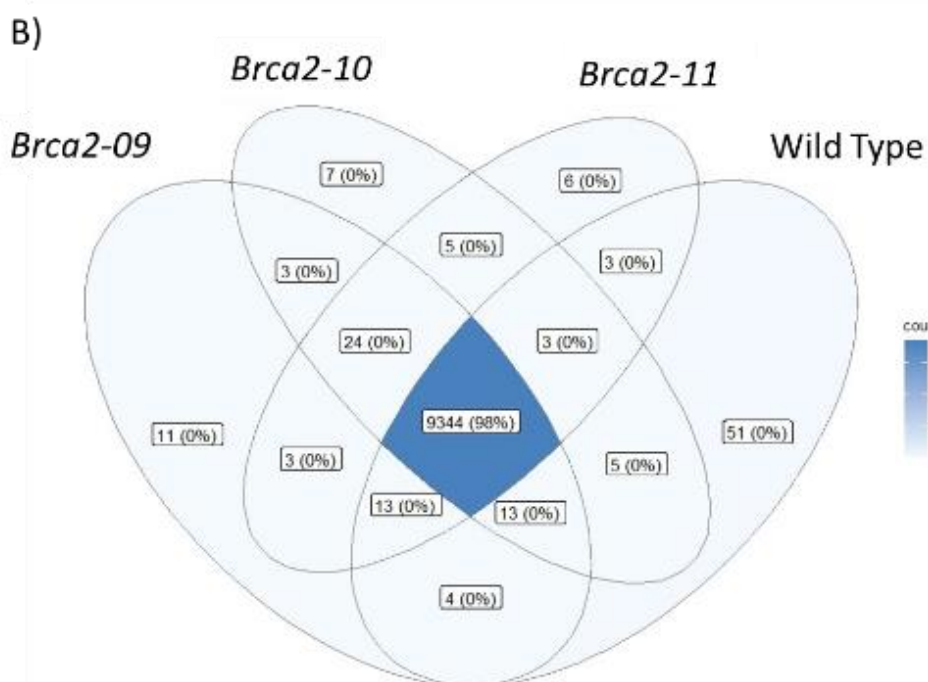
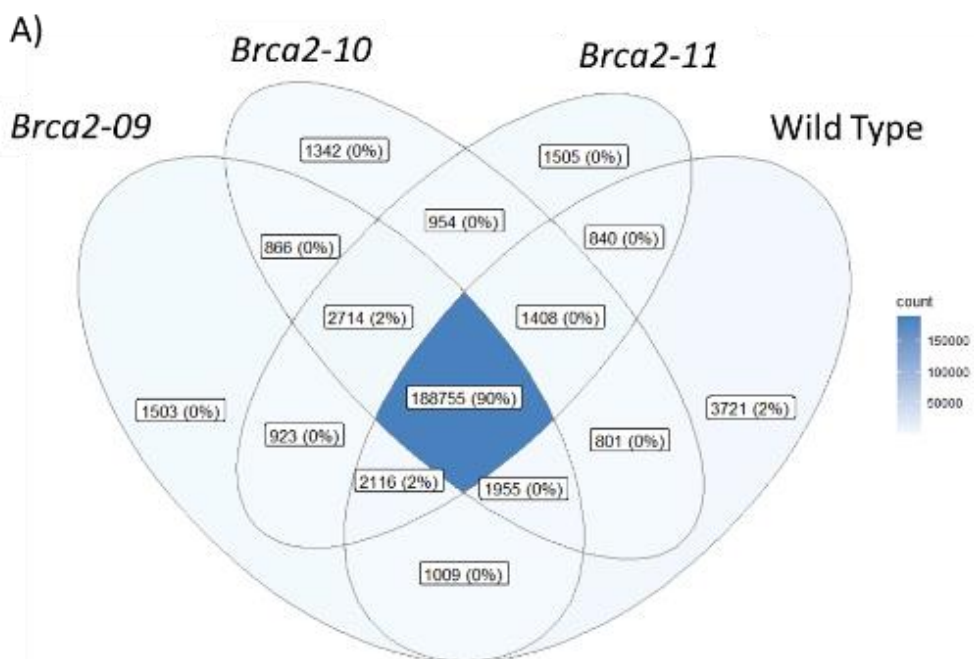


Figure A.3. Venn diagrams generated via R package ggplot2 of all: A) variants called and B) genes with at least one mutation in Brca2-09, Brca2-10, Brca2-11 and Wild Type



## Supplementary Information for Chapter 5

Table A.5. Cells/ml data of inoculated density and end density for each date cultures were subbed. A) Wild-type Untreated replicates; B) wild-type treated replicates; C) Brca2 -/- untreated replicates; D) Brca2 treated replicates.

**A)**

Date	WT(A) Untreated		WT(B) Untreated		WT(C) Untreated	
	Inoculum	End density	Inoculum	End density	Inoculum	End density
21/05/2022	50000	2340000	50000	2180000	50000	1780000
25/05/2022	50000	1940000	50000	1650000	50000	1550000
27/05/2022	50000	1800000	50000	2290000	50000	2400000
31/05/2022	35000	796667	35000	823333	35000	730000
03/06/2022	50000	740000	50000	363333	50000	290000
07/06/2022	35000	2780000	35000	2000000	35000	2380000
10/06/2023	50000	1750000	50000	1890000	50000	1440000
13/06/2023	35000	1788334	35000	1411667	35000	1555000

**B)**

Date	WT(A) Treated		WT(B) Treated		WT(C) Treated	
	Inoculum	End density	Inoculum	End density	Inoculum	End density
21/05/2022	50000	2180000	50000	2540000	50000	2014000
25/05/2022	50000	1850000	50000	2010000	50000	1920000
27/05/2022	50000	2450000	50000	2370000	50000	2260000
31/05/2022	35000	726667	35000	830000	35000	936667
03/06/2022	50000	1020000	50000	850000	50000	750000
07/06/2022	35000	3040000	35000	3260000	35000	2940000
10/06/2023	50000	2040000	50000	1760000	50000	1730000
13/06/2023	35000	1883334	35000	2045000	35000	1938334

**C)**

Date	BRCA2(A) Untreated		BRCA2(B) Untreated		BRCA2(C) Untreated	
	Inoculum	End density	Inoculum	End density	Inoculum	End density
21/05/2022	50000	2220000	50000	1820000	50000	2800000
25/05/2022	50000	2050000	50000	1310000	50000	1310000
27/05/2022	50000	2740000	50000	2660000	50000	2370000
31/05/2022	35000	397500	35000	603333	35000	583333
03/06/2022	50000	606667	50000	386667	50000	690000
07/06/2022	35000	2260000	35000	1370000	35000	1710000
10/06/2023	50000	1470000	50000	770000	50000	530000
13/06/2023	35000	1328750	35000	986667	35000	1146667

**D)**

Date	BRCA2(A) Treated		BRCA2(B) Treated		BRCA2(C) Treated	
	Inoculum	End density	Inoculum	End density	Inoculum	End density
21/05/2022	50000	850000	50000	1140000	50000	2060000
25/05/2022	50000	1760000	50000	1870000	50000	1780000
27/05/2022	50000	2200000	50000	1730000	50000	2520000
31/05/2022	35000	516667	35000	626667	35000	590000
03/06/2022	50000	343333	50000	496667	50000	343333
07/06/2022	35000	2080000	35000	2350000	35000	2500000
10/06/2023	50000	1030000	50000	980000	50000	990000
13/06/2023	35000	1298334	35000	1488334	35000	1545000

Table A.6. Total generations and average specific growth rate of each culture by treatment and timepoint. The average specific growth rate is the average of each specific growth rate calculation for each culture (n=4) generated between each timepoint.

<b>Culture</b>	<b>Total Generations</b>	<b>Average Specific Growth Rate</b>
<i>Brca2</i> -/- Treated T1	18.57	1.04
<i>Brca2</i> -/- Treated T2	36.82	0.97
<i>Brca2</i> -/- Untreated T1	20.11	1.12
<i>Brca2</i> -/- Untreated T2	39.85	1.05
<i>Brca2</i> -/- Treated T1	19.01	1.05
<i>Brca2</i> -/- Treated T2	38.10	1.01
<i>Brca2</i> -/- Untreated T1	19.74	1.10
<i>Brca2</i> -/- Untreated T2	36.74	0.91
<i>Brca2</i> -/- Treated T1	20.25	1.12
<i>Brca2</i> -/- Treated T2	38.96	0.99
<i>Brca2</i> -/- Untreated T1	20.14	1.11
<i>Brca2</i> -/- Untreated T2	37.98	0.95
Wild-type Treated T1	20.65	1.14
Wild-type Treated T2	42.54	1.17
Wild-type Untreated T1	20.50	1.11
Wild-type Untreated T2	41.51	1.12
Wild-type Treated T1	21.13	1.16
Wild-type Treated T2	42.77	1.16
Wild-type Untreated T1	20.56	1.13
Wild-type Untreated T2	39.84	1.03
Wild-type Treated T1	20.84	1.14
Wild-type Treated T2	42.04	1.13
Wild-type Untreated T1	20.08	1.11
Wild-type Untreated T2	39.02	1.01

Table A.7. Raw reads and estimated raw coverage generated by the Earlham Institute for each sample.

<b>Culture</b>	<b>Replicate</b>	<b>Treatment</b>	<b>Timepoint</b>	<b>Number of Paired Reads</b>	<b>Estimated Raw Coverage</b>
Wild-Type	N/A	N/A	T0	20,079,071	178 x
Wild-Type	A	MMS	T1	19,485,958	173 x
Wild-Type	B	MMS	T1	742,324	7 x
Wild-Type	C	MMS	T1	17,380,753	154 x
Wild-Type	A	Mock	T1	18,006,596	160 x
Wild-Type	B	Mock	T1	14,998,380	133 x
Wild-Type	C	Mock	T1	18,146,370	161 x
Wild-Type	A	MMS	T2	17,744,227	157 x
Wild-Type	B	MMS	T2	14,846,672	132 x
Wild-Type	C	MMS	T2	14,916,971	132 x
Wild-Type	A	Mock	T2	15,070,543	134 x
Wild-Type	B	Mock	T2	13,694,880	122 x
Wild-Type	C	Mock	T2	16,661,578	148 x
<i>Brca2</i>	N/A	N/A	T0	19,457,012	173 x
<i>Brca2</i>	A	MMS	T1	19,591,522	174 x
<i>Brca2</i>	B	MMS	T1	17,988,964	160 x
<i>Brca2</i>	C	MMS	T1	22,762,775	202 x
<i>Brca2</i>	A	Mock	T1	17,460,181	155 x
<i>Brca2</i>	B	Mock	T1	16,255,880	144 x
<i>Brca2</i>	C	Mock	T1	14,955,732	133 x
<i>Brca2</i>	A	MMS	T2	17,296,850	154 x
<i>Brca2</i>	B	MMS	T2	16,059,342	143 x
<i>Brca2</i>	C	MMS	T2	16,769,435	149 x
<i>Brca2</i>	A	Mock	T2	16,074,876	143 x
<i>Brca2</i>	B	Mock	T2	16,124,163	143 x
<i>Brca2</i>	C	Mock	T2	21,691,819	193 x

Table A.8. Statistics on the alignment of Illumina reads against the reference *T. pseudonana* genome (Armburst et al., 2004). Statistics were generated through the tool bamQC from QualiMap (Okonechnikov et al., 2015).

Sample	Mean Mapping Quality	Number of Mapped Reads	GC %	Duplication Rate	General Error Rate	Mean Coverage	Standard Deviation of Coverage
#	59.5672	31,383,819	46.91%	64.33%	0.0076	149.7302X	71.5853X
<i>Brca2</i> -/- A Treated T1	59.5687	31,464,119	47.03%	62.52%	0.0083	149.7362X	75.5552X
<i>Brca2</i> -/- A Treated T2	59.5706	26,928,833	47.07%	61.30%	0.0085	128.1054X	64.0191X
<i>Brca2</i> -/- A Untreated T1	59.5829	28,232,472	47.21%	61.93%	0.0087	130.4958X	57.5316X
<i>Brca2</i> -/- A Untreated T2	59.5646	25,388,684	47.02%	58.09%	0.0094	120.5096X	59.0767X
<i>Brca2</i> -/- B Treated T1	59.5761	29,076,371	47.08%	61.33%	0.0077	134.9744X	62.0323X
<i>Brca2</i> -/- B Treated T2	59.5654	25,636,329	46.96%	61.46%	0.0086	122.2395X	69.3971X
<i>Brca2</i> -/- B Untreated T1	59.5806	25,725,394	47.17%	61.52%	0.0079	120.7617X	53.6619X
<i>Brca2</i> -/- B Untreated T2	59.5676	25,039,521	47.07%	58.36%	0.0094	118.4658X	56.9277X
<i>Brca2</i> -/- C Treated T1	59.5717	36,764,223	46.98%	64.70%	0.0071	173.1987X	78.8230X
<i>Brca2</i> -/- C Treated T2	59.5696	26,136,392	47.13%	58.98%	0.0091	124.4782X	64.8289X
<i>Brca2</i> -/- C Untreated T1	59.5749	23,989,000	47.21%	58.71%	0.0094	113.8031X	53.2426X
<i>Brca2</i> -/- C Untreated T2	59.5731	30,940,229	46.98%	62.12%	0.0073	145.0512X	71.1527X
Wild-type T0	59.5660	31,950,786	46.96%	63.87%	0.0085	152.3538X	79.5627X
Wild-type A Treated T1	59.5793	33,131,437	46.98%	62.78%	0.0081	157.9718X	72.1989X
Wild-type A Treated T2	59.5631	28,437,432	46.91%	63.90%	0.0082	135.6206X	80.5106X
Wild-type A Untreated T1	59.5700	28,806,285	47.01%	60.80%	0.0077	137.1966X	70.6504X
Wild-type A Untreated T2	59.5716	23,296,238	47.08%	65.18%	0.0088	111.0806X	63.7213X
Wild-type B Treated T1	59.5901	1,177,182	47.12%	26.81%	0.0082	5.5387X	4.2183X
Wild-type B Treated T2	59.5629	23,513,667	46.92%	64.11%	0.0091	112.0934X	73.8581X
Wild-type B Untreated T1	59.5716	23,811,063	47.10%	56.70%	0.0102	113.0066X	66.1292X
Wild-type B Untreated T2	59.5608	21,165,622	46.86%	66.05%	0.0085	100.9291X	68.9661X
Wild-type C Treated T1	59.5627	27,720,560	46.95%	60.47%	0.0076	132.0506X	71.8130X
Wild-type C Treated T2	59.5656	22,995,619	46.97%	66.33%	0.0090	109.5736X	73.6085X
Wild-type C Untreated T1	59.5696	28,269,084	46.98%	60.22%	0.0078	134.4851X	85.7242X
Wild-type C Untreated T2	59.5656	25,208,632	47.04%	62.95%	0.0079	120.1611X	71.4526X

Table A.9. Overview of numbers of variants and variant types for each sample before and after filtering. Wild-type T0 and wild-type B Treated T1 are not included after filtration since all variants from wild-type T0 were filtered out and due to the failed sequencing of wild-type B treated T1. Statistics were generated using bcftools stats (Danecek et al., 2021).

Sample	BEFORE FILTERING					AFTER FILTERING				
	number of variants:	number of SNPs:	number of indels:	number of multiallelic sites:	number of multiallelic SNP sites:	number of variants:	number of SNPs:	number of indels:	number of multiallelic sites:	number of multiallelic SNP sites:
Brca2 T0	200193	182869	17324	424	19	3679	2582	1097	84	18
Brca2_A Treated T1	201150	183784	17366	408	24	3460	2536	924	70	16
Brca2_A Treated T2	200643	183406	17237	418	28	4232	3115	1117	86	26
Brca2_A Untreated T1	202973	185562	17411	416	30	5906	4779	1127	89	34
Brca2_A Untreated T2	200986	183749	17237	398	24	4221	3197	1024	72	22
Brca2_B Treated T1	202431	184994	17437	424	32	5378	4236	1142	91	31
Brca2_B Treated T2	199666	182592	17074	411	27	3665	2651	1014	61	22
Brca2_B Untreated T1	202086	184723	17363	408	26	5282	4154	1128	81	25
Brca2_B Untreated T2	201498	184209	17289	421	33	4682	3614	1068	76	23
Brca2_CTreated T1	201442	184048	17394	429	28	4546	3442	1104	85	27
Brca2_C Treated T2	200473	183267	17206	425	27	3971	2937	1034	80	26
Brca2_C Untreated T1	201347	184055	17292	420	28	4679	3534	1145	82	24
Brca2_C Untreated T2	201667	184299	17368	418	24	4746	3632	1114	89	31
Wild type_A Treated T1	201165	183827	17338	435	28	3949	2918	1031	89	22
Wild type_A Treated T2	200354	183048	17306	411	24	3267	2327	940	71	15
Wild type_A Untreated T1	201363	183903	17460	409	22	3676	2731	945	76	19
Wild type_A Untreated T2	200650	183357	17293	408	26	3816	2814	1002	67	14
Wild type_B Treated T1	22181	21193	988	37	13	N/A	N/A	N/A	N/A	N/A
Wild type_B Treated T2	200069	182968	17101	404	20	3441	2541	900	68	18
Wild type_B Untreated T1	201503	184262	17241	399	25	4012	3104	908	67	17
Wild type_B Untreated T2	199772	182780	16992	396	18	3580	2664	916	75	22
Wild type_C Untreated T1	201303	183855	17448	421	21	3622	2629	993	74	19
Wild type_C Treated T1	199977	182914	17063	415	28	3580	2608	972	72	20
Wild type_C Treated T2	201041	183711	17330	417	22	3662	2726	936	74	18
Wild type_C Untreated T2	200831	183484	17347	405	17	3702	2719	983	84	21
Wild type T0	200605	183213	17392	421	23	N/A	N/A	N/A	N/A	N/A

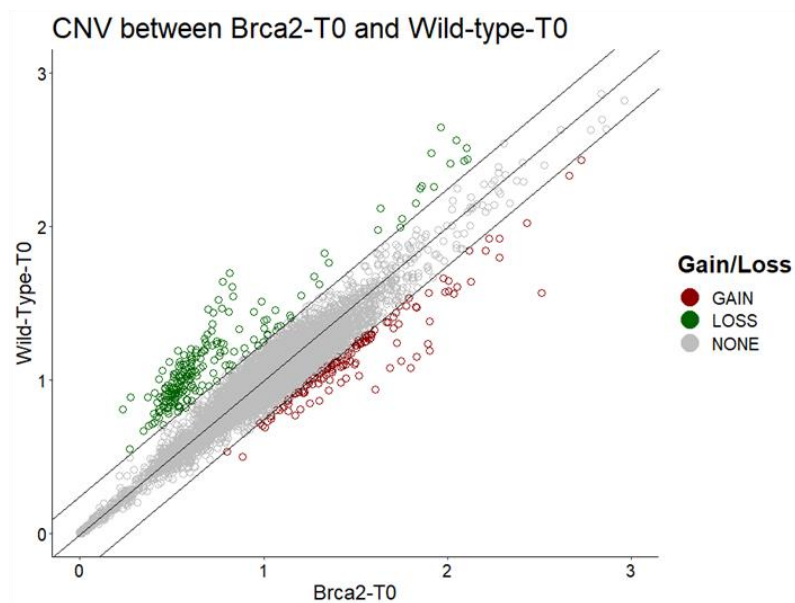
Table A.10. Populations used for *Fst* calculations, and the samples included in each population.

<b>Population</b>	<b>Samples</b>
<i>Brca2</i> -/- Treated T1	<i>Brca2</i> _B_MMS_T1; <i>Brca2</i> _C_MMS_T1
<i>Brca2</i> -/- Treated T2	<i>Brca2</i> _A_MMS_T2; <i>Brca2</i> _B_MMS_T2; <i>Brca2</i> _C_MMS_T2
<i>Brca2</i> -/- Untreated T1	<i>Brca2</i> _A MOCK_T1; <i>Brca2</i> _B MOCK_T1; <i>Brca2</i> _C MOCK_T1
<i>Brca2</i> -/- Untreated T2	<i>Brca2</i> _A MOCK_T2; <i>Brca2</i> _B MOCK_T2; <i>Brca2</i> _C MOCK_T2
Wild type Treated T1	WT_A_MMS_T1; WT_C_MMS_T1
Wild type Treated T2	WT_A_MMS_T2; WT_B_MMS_T2; WT_C_MMS_T2
Wild type Untreated T1	WT_A MOCK_T1; WT_B MOCK_T1; WT_C MOCK_T1
Wild type Untreated T2	WT_A MOCK_T2; WT_B MOCK_T2; WT_C MOCK_T2

Table A.11. Number of genes with either a gain or loss in CNV compared with wild-type T0.

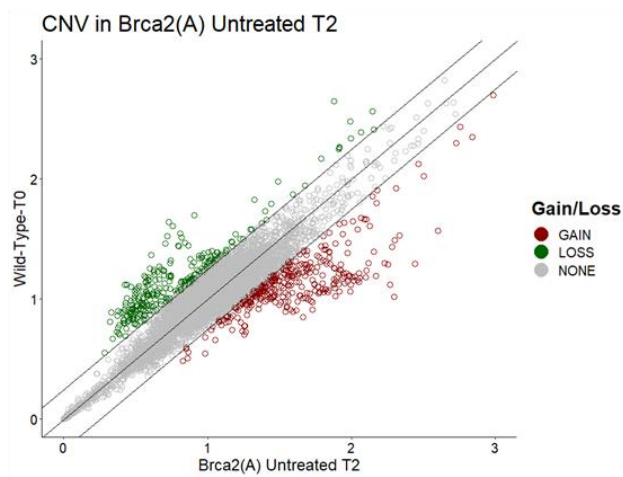
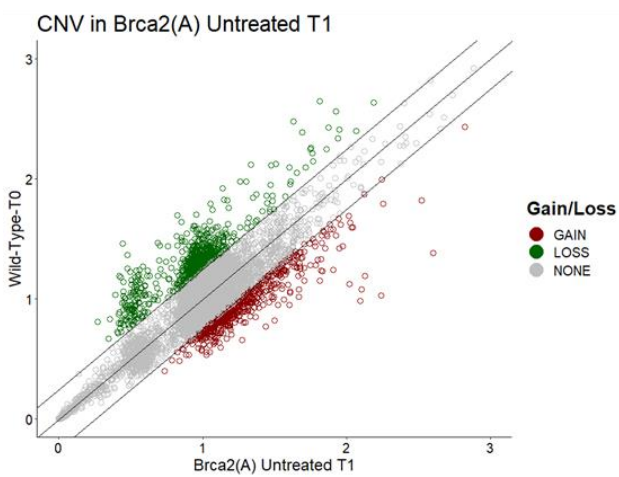
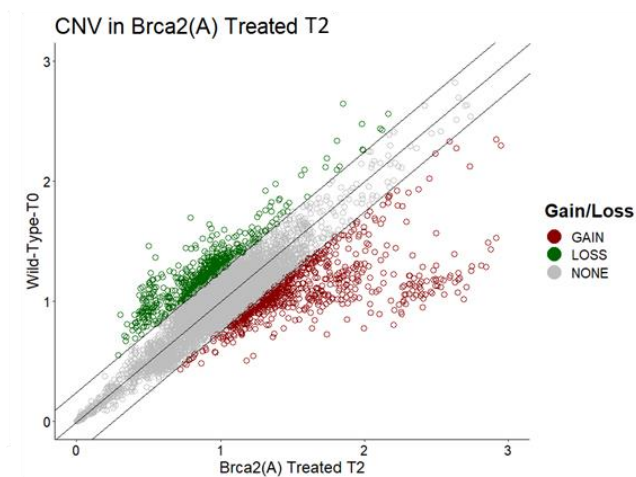
Culture	Gain in CNV	Loss in CNV
<i>Brca2</i> -/- T0	134	208
<i>Brca2</i> -/- A Treated T1	41	99
<i>Brca2</i> -/- A Treated T2	777	834
<i>Brca2</i> -/- A Untreated T1	580	838
<i>Brca2</i> -/- A Untreated T2	356	301
<i>Brca2</i> -/- B Treated T1	537	536
<i>Brca2</i> -/- B Treated T2	1007	1040
<i>Brca2</i> -/- B Untreated T1	759	897
<i>Brca2</i> -/- B Untreated T2	510	454
<i>Brca2</i> -/- C Treated T1	205	239
<i>Brca2</i> -/- C Treated T2	1024	1058
<i>Brca2</i> -/- C Untreated T1	642	808
<i>Brca2</i> -/- C Untreated T2	711	613
Wild-type A Treated T1	130	267
Wild-type A Treated T2	260	84
Wild-type A Untreated T1	45	72
Wild-type A Untreated T2	460	484
Wild-type B Treated T2	735	518
Wild-type B Untreated T1	619	433
Wild-type B Untreated T2	707	497
Wild-type C Treated T1	120	68
Wild-type C Treated T2	963	965
Wild-type C Untreated T1	1087	812
Wild-type C Untreated T2	714	557

A)



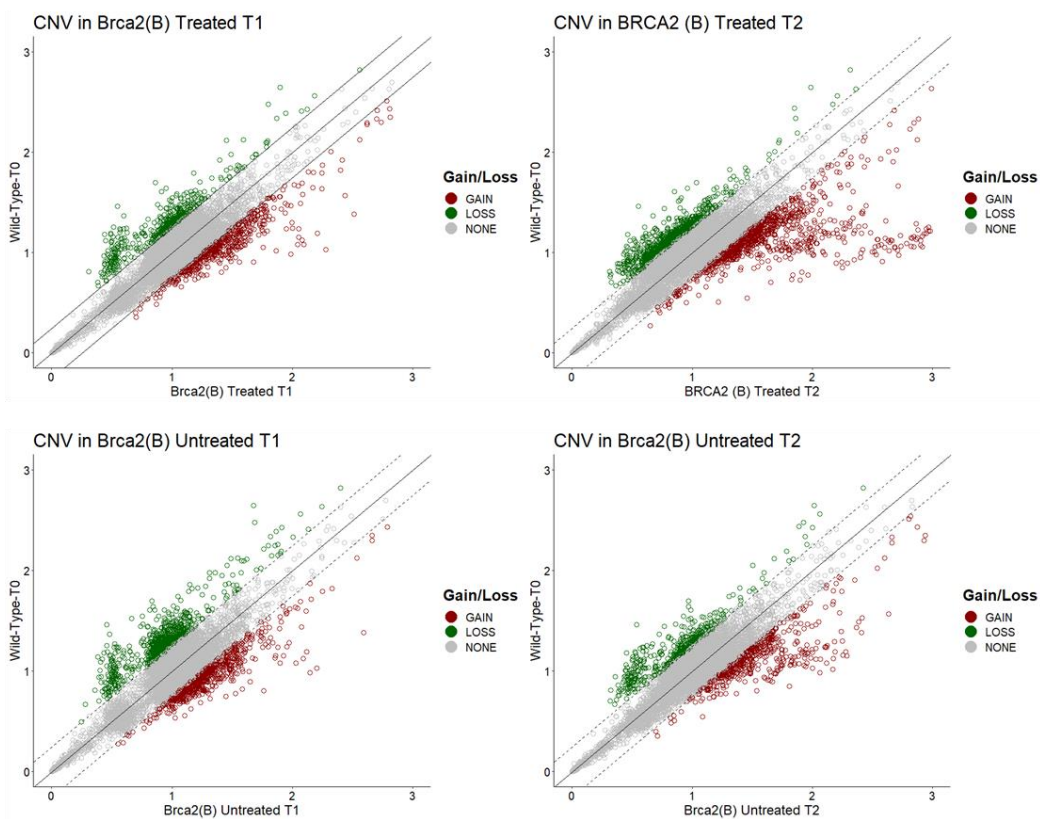
B)

*No data*

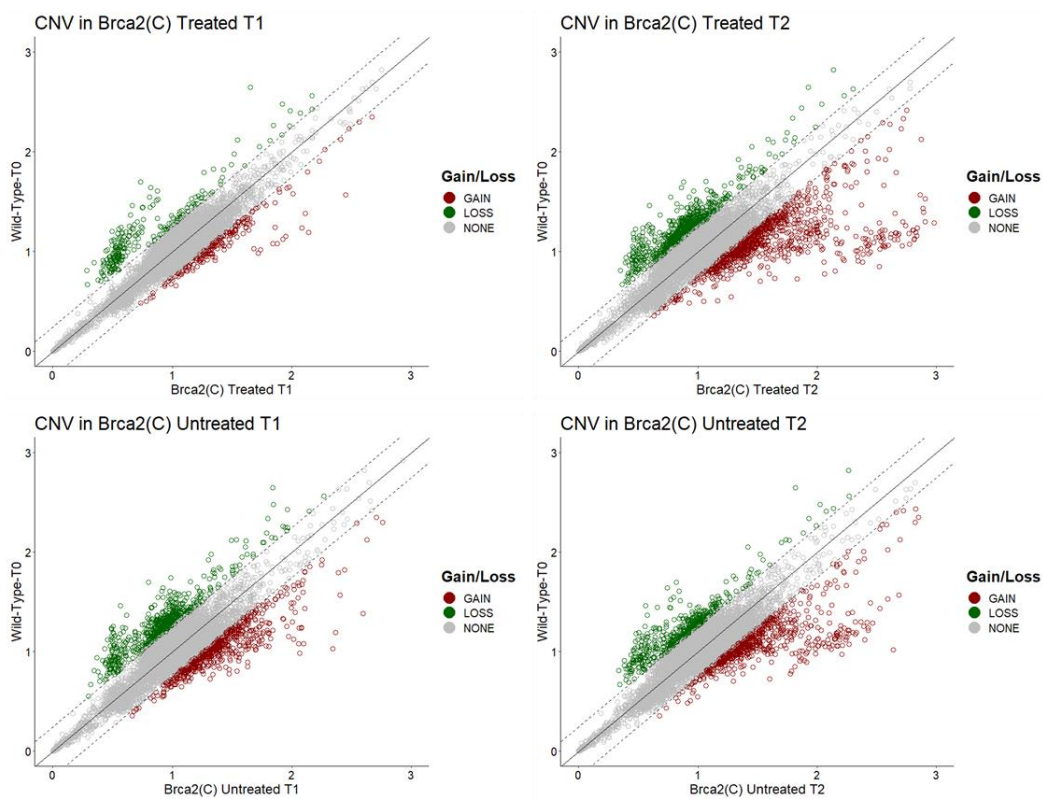




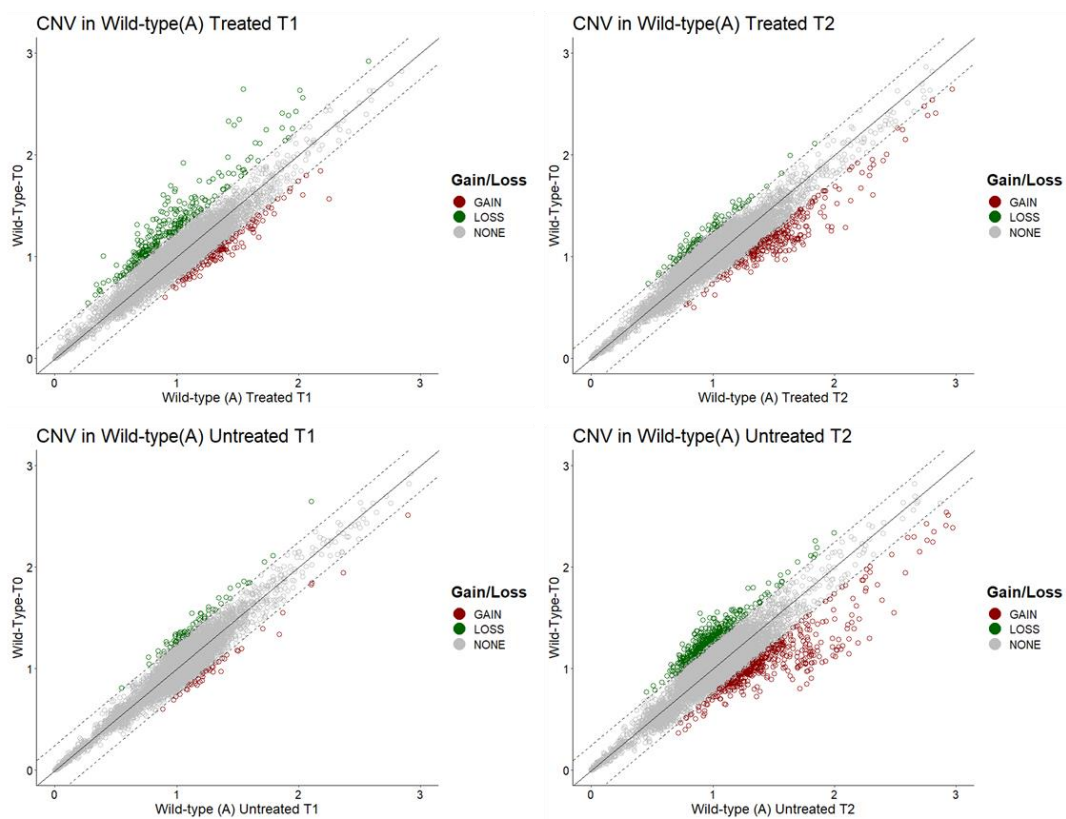
C)



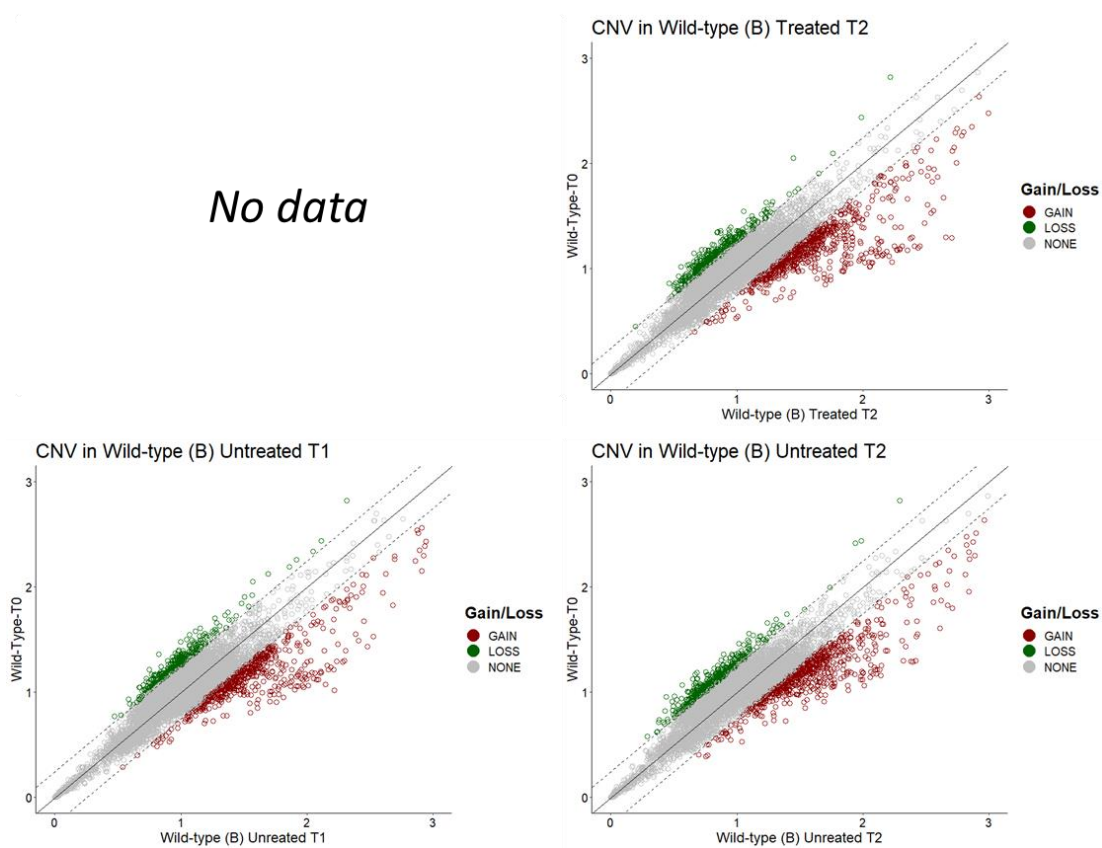
D)



E)



F)



G)

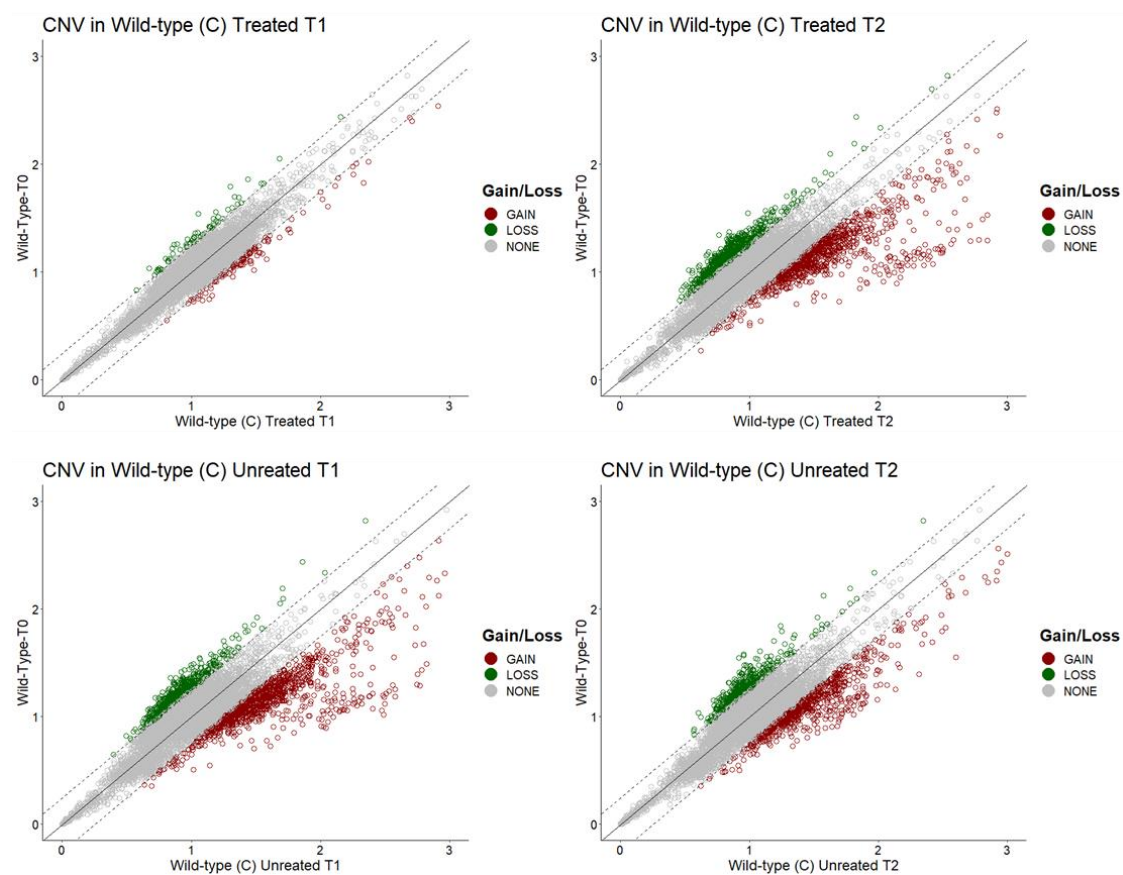


Figure A.4. CNV variation of each sample compared with wild-type Red dots indicate genes with gain in CNV while green indicates a loss in CNV. The solid line represents a 1:1 relationship with the CNV of the wild-type and the dashed lines represent the cut-off for a CNV to be considered (25% increase or decrease in CNV compared to wildtype). A) T0. A) Brca2<sup>-/-</sup> T0. Each of the following sets of four figures is ordered in the following way: top row right to left = treated T1 and T2, bottom row right to left = untreated T1 and T2. B) Brca2<sup>-/-</sup> replicate A; C) Brca2<sup>-/-</sup> replicate B; D) Brca2<sup>-/-</sup> replicate C; E) wild-type replicate A; F) wild-type replicate B; G) wild-type replicate C.

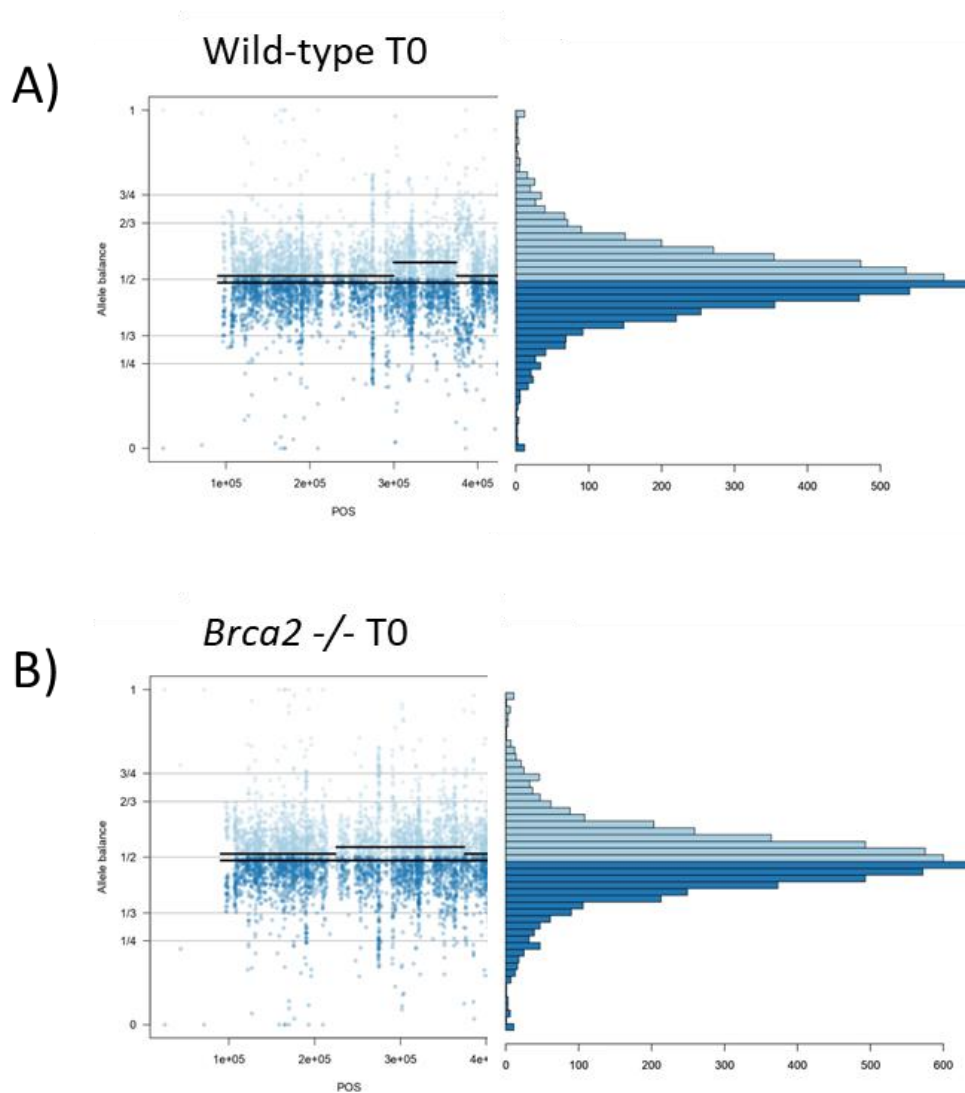


Figure A.5. Allele balance plots of A) wild type T0 and B) *Brca2* -/- T0. Plots show the allele frequency for all variants called before filtering. The y-axis shows the proportion of read depth for both alleles of each variant.  $\frac{1}{2}$  = Both alleles cover 50% of reads for the specific variant.  $\frac{2}{3}$  = the major allele has two times the coverage of the minor allele. The x-axis is the position across chromosome 23, from 1 – 454954bp.