

Characterization of the Mutational Signature of Formaldehyde
to Elucidate Its Role in Carcinogenesis

Reena Michelle Fabros

Supervisor: Dr. Kin Chan

Thesis submitted to the University of Ottawa
in partial fulfillment of the requirements for the
Master's Degree in Biochemistry
with Specialization in Bioinformatics

Department of Biochemistry, Microbiology & Immunology
Faculty of Medicine
University of Ottawa

Abstract

Cancer is a genetic disease caused by the accumulation of multiple mutations in essential genes controlling normal proliferation. Genomic analysis of cancer samples revealed that somatic mutations form distinct patterns on the DNA called mutational signatures, which were produced by different mutational processes that have been operative throughout the development of a cancer cell. Out of 77 deciphered mutational signatures from cancer genomes, 29 of them have unknown aetiologies. The current strategy to elucidate these unknown signatures is through mutational signature characterization of various mutagens from model systems. One such mutagen with an ambiguous mutational signature is formaldehyde. This work describes the first specific characterization of the mutational patterns induced by formaldehyde. Using yeast genomics data, formaldehyde mutagenesis was shown to form C→A transversions over TCn trinucleotide motifs. Moreover, it induces an excess of single nucleotide indels with a preference for adenine insertions and cytosine deletions. Due to a limited sample size, the contribution of formaldehyde mutagenesis in relevant cancers cannot be accurately depicted, however, the formaldehyde-specific pattern of C→A transversions can be observed from these cancer genomes. These results ultimately present a likely mechanism for formaldehyde mutagenesis on the nucleotide bases.

Acknowledgements

I would like to give special thanks to Dr. Kin Chan, whose leadership and supervision throughout my graduate school experience has been incredibly valuable. His wisdom and passion for the field is an inspiration to a young researcher like me, who oftentimes have had numerous concerns about our rapidly changing environment. It was indeed an honour to have worked with a very encouraging and driven mentor.

To my Thesis Advisory Committee members, Dr. Kristin Baetz and Dr. Jonathan Lee: truly, their support and counsel was helpful in improving and advancing this project successfully.

To my colleague and friend Salma Alasmar: without her contribution, this research project would not have been made to fruition. Her enthusiasm and compassion were truly appreciated especially during trying times. And to other Chan lab members, Nathan Liang, Bledar Xhialli and Suzana Gelova: I am very thankful for their incredible insight and constructive feedback. To my fellow trainees, Alexander Pelletier and Justin Chitpin: their assistance and ceaseless camaraderie throughout my programming woes have kept me from indelibly suffering through sleepless nights. I wish them all the best in their endeavours!

And finally, to my family: their endless support and loving kindness have been profoundly meaningful and incredibly encouraging through challenging times. I could not have been more blessed to be part of a household who values the significance of hard work and higher education.

Table of Contents

Abstract	ii
Acknowledgements	iii
Table of Contents.....	iv
List of Abbreviations.....	vi
List of Figures.....	viii
List of Tables.....	x
1. Introduction	1
1.1 Mutagenesis and Mutational Signatures.....	1
1.2 Mutational Signatures in Cancer Genomes	2
1.3 A Known Crosslinking Agent called Formaldehyde.....	3
1.4 Formaldehyde's Impact on Public Health	5
1.5 Budding Yeast as a Model System to Study Mutations	7
1.6 Rationale and Research Objectives.....	9
2. Methodology.....	11
2.1 Highly Sensitive Yeast Model System, ySR127	11
2.2 Exogenous Formaldehyde Treatment (High, Acute Exposure).....	13
2.3 Mutant Screening and Selection.....	17
2.4 The Auxin-Inducible Degron (AID or Degron) System on ySR127.....	19
2.5 Construction of Genetically Modified Yeasts	20
2.6 Endogenous Formaldehyde Treatment (Low, Chronic Exposure)	23
2.7 Statistical Analyses for the Exogenous and Endogenous Formaldehyde Experiments	25
2.8 Single Base Substitution Signature Analysis of Formaldehyde-treated Yeast Genomes using the MutationalPatterns Software.....	26
2.9 Indel Signature Analysis of Formaldehyde-treated Yeast Genomes	29
2.10 Single Base Substitution Signature Analysis of Head and Neck Cancer Genomes using the MutationalPatterns software	30

3.	Results.....	31
3.1	Generating Mutants for Signature Analysis while Examining the Effects of Formaldehyde in the Yeast Model System.....	31
3.1.1	Exogenous Formaldehyde Treatment (High, Acute Exposure)	31
3.1.2	Endogenous Formaldehyde Treatment (Low, Chronic Exposure)	40
3.2	Computational Characterization of the Mutational Signature of Formaldehyde Using Yeast Genomics Data.....	50
3.2.1	Single Base Substitution (SBS) Signature	50
3.2.2	Insertion/Deletion (ID) Signature.....	57
3.3	Significance of the Formaldehyde Single Base Substitution Signature by Examining Cancer Data	64
3.3.1	Formaldehyde SBS Signature vs. COSMIC SBS Signatures.....	64
3.3.2	Deciphering Signatures from Head and Neck Cancer Genomes.....	74
4.	Discussion.....	80
4.1	Mutagenic Effects of Exogenous Formaldehyde (High, Acute Exposure).....	80
4.2	Mutagenic Effects of Endogenous Formaldehyde (Low, Chronic Exposure)	82
4.3	The Putative Single Base Substitution Signature of Formaldehyde.....	85
4.4	Indel Mutational Patterns of Formaldehyde	90
4.5	Relevance of the Formaldehyde SBS Signature in Cancer Genomics Data.....	93
4.6	Conclusions.....	96
	References	98

List of Abbreviations

ADE	Adenine-dropout media
<i>ADE2</i>	ADEnine requiring gene
AID	Auxin-Inducible Degron
CAN	Canavanine-supplemented media, with low adenine
<i>CAN1</i>	CANavanine resistance gene
<i>cdc13-1</i>	Temperature-sensitive mutation at <i>CDC13</i> gene
COSMIC	Catalogue of Somatic Mutations in Cancer
CosSim	Cosine Similarity
DMSO	Dimethyl sulfoxide
DNA	Deoxyribonucleic Acid
EDTA	Ethylenediaminetetraacetic Acid
FA	Formaldehyde
FGSH	S-formylglutathione
Head-SCC	Head and Neck Squamous Cell Carcinoma
HMGSH	S-hydroxymethyl-glutathione
IAA	Indole-3-acetic acid
IARC	International Agency for Research on Cancer
ID	Indel (or Insertions and Deletions)
LiAc-TE	Lithium Acetate, Tris, EDTA
NMF	Non-negative matrix factorization
PCAWG	Pan-Cancer Analysis of Whole Genomes

PCR	Polymerase Chain Reaction
ROS	Reactive Oxygen Species
SBS	Single base substitution
SC	Synthetic complete media
SCF-TIR1	Skp1 – Cull1 – Fbox Transport Inhibitor Response I
<i>SFA1</i>	Sensitive to Formaldehyde I gene
sfa1-degron	Degron-tagged SFA1 gene
sfa1-degron/yjl068c Δ	Degron-tagged SFA1 gene with yjl068c Δ knockout
SNP	Single Nucleotide Polymorphism
ssDNA	Single stranded DNA
STRAH	Short Tandem Repeats Analysis of Hotspot Zones
TCGA	The Cancer Genome Atlas
TLS	Translesion synthesis
URA	Uracil-dropout media
<i>URA3</i>	URAcil requiring gene
<i>YJL068C</i>	S-formylglutathione hydrolase gene
YPDA	Rich-media i.e. yeast extract, peptone, D-glucose, and 0.55% adenine sulfate
YPG	Glycerol media i.e. yeast, peptone and glycerol

List of Figures

Introduction

Figure 1.1. The Glutathione-dependent Formaldehyde Oxidation Pathway in <i>Saccharomyces cerevisiae</i>	9
---	---

Methodology

Figure 2.1. Schematic Diagram of our Highly Sensitive Yeast Model, ySR127	12
Figure 2.2. Experimental Procedure for the High, Acute Treatment of Exogenous Formaldehyde	16
Figure 2.3. Standard Procedure for Mutant Screening and Selection	18
Figure 2.4. Genetic Modification of ySR127 using the Delitto Perfetto System to Create Single-Gene Knockouts of SFA1 and YJL068C	21
Figure 2.5. Construction of Auxin-Inducible Degron (AID) strains	22
Figure 2.6. Experimental Procedure for the Low, Chronic Treatment of Endogenous Formaldehyde using Genetically Modified Yeasts	24
Figure 2.7. Vector Diagram of the Cosine Similarity Algorithm with respect to Mutational Signature Analysis	28

Results

Figure 3.1.1. Effects of Exogenous Formaldehyde Treatment to Viability under nitrogen (anaerobic) and room air (aerobic)	34
Figure 3.1.2. Effects of Exogenous Formaldehyde Treatment to Can ^R Frequency or Single-Gene Inactivation Frequency.	38
Figure 3.1.3. Effects of Exogenous Formaldehyde Treatment to Can ^R Ade ⁻ Frequency or Double-Gene Inactivation Frequency	39

Figure 3.1.4. Effects of Endogenous Formaldehyde Treatment to Can ^R Frequency in Single Knockout Strains.....	43
Figure 3.1.5. Results of the Double Knockout Phenotype Leading to the Use of an Alternative Approach.....	44
Figure 3.1.6. Mutation Frequency of Pre-growth Cultures (day 0 samples) of <i>sfa1-deg/yjl068cΔ</i> Strains.....	47
Figure 3.1.7. Effects of Endogenous Formaldehyde on <i>sfa1-deg/yjl068cΔ</i> Strains after Auxin Treatment and Temperature Shift.....	48
Figure 3.1.8. Formaldehyde Sensitivity of All Genetically Modified Strains.	49
Figure 3.2.1. Signature Extraction by Non-Negative Matrix Factorization (NMF) Algorithm.....	53
Figure 3.2.2. The Putative Single Base Substitution (SBS) Formaldehyde and No Formaldehyde Signatures Generated from 10 Different Solutions of the Non-Negative Matrix Factorization (NMF) Algorithm.	55
Figure 3.2.3. Indel Types By Length for No FA and FA Treatments.....	60
Figure 3.2.4. Comparison of 1-Base Insertions Between No Formaldehyde and Formaldehyde Treatments	61
Figure 3.2.5. Putative ID Signatures of FA and No FA Exposures.....	62
Figure 3.3.1. Cosine Similarity Values of the Formaldehyde and No Formaldehyde SBS Signatures versus 67 COSMIC SBS Signatures.....	67
Figure 3.3.2A. The Combination of COSMIC Signatures SBS5, SBS40 and SBS52 to Reconstruct the Putative Formaldehyde SBS Signature.....	68
Figure 3.3.2B. Mutational Profiles of COSMIC Signatures SBS5, SBS40 and SBS52	69
Figure 3.3.3. Indel Analysis of Samples containing COSMIC Signature SBS52 and Samples lacking SBS52.	72

Figure 3.3.4. Distribution of 1-base and 2-base Indel Mutation Types among SBS52+ and SBS52- samples	73
Figure 3.3.5. Extraction of Signatures from Head and Neck Squamous Cell Carcinoma (Head-SCC) Genomes	75
Figure 3.3.6. Cosine Similarities (cossim) of Head and Neck Cancer (Head-SCC) Signatures vs. 67 COSMIC Single-Base Substitution (SBS) Signatures and the Putative Formaldehyde Signature	78
Figure 3.3.7. Reconstruction of Head and Neck Cancer (Head-SCC) Signature A	79

Discussion

Figure 4.1. COSMIC Signatures SBS5 and SBS40	89
Figure 4.2. Mutational Profiles of Each Untreated and FA-treated Data Sets	90
Figure 4.3. Heatmap Depicting Cosine Similarities of COSMIC SBS Signatures Against Each Other	97

List of Tables

Table 2.1. Optimized number of cells plated for each treatment.....	15
Table 2.2. Selection criteria for DNA sequencing candidates.....	17

1. Introduction

1.1 Mutagenesis and Mutational Signatures

Genomic DNA is the source for perpetuation of all known living organisms, including some viruses. Although preserving genetic information is integral in the maintenance of life, the DNA is exposed to constant assaults by various exogenous and endogenous agents[1, 2]. DNA damaging agents can be physical, chemical, or enzymatic, such as ultraviolet radiation[3–5], ionizing radiation[1, 6], tobacco mutagens[2, 5, 7], the APOBEC family of proteins[5, 8–10], and DNA replication errors[2, 5]. Cells can activate various repair pathways[2, 11] as a response to DNA lesions, such as base substitutions, bulky adducts, insertions and deletions, and chromosomal rearrangements[12, 13].

Although cells have robust DNA damage response mechanisms[5], oftentimes, erratic exposures and changing microenvironments can alter cellular repair functions[14–17] leading to increased base lesions. When multiple DNA damage persists, programmed cell death or apoptosis is activated to remove cells with extensive genomic instability[2, 16]. In some cases, however, genes that are essential for maintaining normal biological processes retain damages that become permanent DNA modifications, which could then result in the uncontrolled proliferation of cells[17, 18].

Although unrepaired lesions are a major risk for genomic integrity, those that lead to permanent modifications are also important factors for genetic variation and evolution of species[19]. However, in humans, these permanent chemical modifications called

mutations contribute to aging[20, 21], development of certain diseases[22, 23] and carcinogenesis[21, 24]. Interestingly, these mutations form unique patterns on the DNA, which can be distinctly attributed to a particular mutational process[5, 25–27]. As nucleotide bases are highly reactive entities, each base is uniquely susceptible to chemical reactions with other molecules[2, 28]. This is the basic principle of the formation of a mutational signature, in which DNA mutations observed in a genome conform to a pattern that is characteristic of the mutagenic agent or process responsible for inducing those mutations[5, 29, 30].

1.2 Mutational Signatures in Cancer Genomes

While driver mutations are considered important markers for tumorigenesis and carcinogenesis, recent studies have shown that passenger mutations are also relevant in assessing cancer initiation[5, 24]. These ‘bystander’ mutations are the product of various biological processes that have occurred throughout the development of a cancer cell. Therefore, passenger mutations bear significant historical information of a cancer genome[5].

Mutational signatures, or patterns of mutations, were first discovered in cancer genomes by analyzing their somatic mutations using an algorithm called the non-negative matrix factorization (abbreviated as NMF)[25, 29]. Because many mutational processes had been operative in a single cancer cell, its genome has become a complex landscape of overlapping mutations[5, 30]. Consequently, a sequenced cancer genome will display a composite of multiple mutational signatures, which renders observation of individual

patterns particularly unfeasible[5, 25, 29]. The NMF algorithm was used to extract unique patterns of mutations from all types of cancers, which led to the remarkable discovery of various mutagenic processes and agents that contributed to carcinogenesis[25, 26, 29]. Although several processes have been known to produce some of these mutational signatures, there remains a large portion of signatures with unidentified causes[26, 31].

The Catalogue of Somatic Mutations In Cancer or COSMIC is a collection of mutational signatures that were deciphered from somatic mutations in cancer samples using data from the Pan-Cancer Analysis of Whole Genomes (PCAWG) group, The Cancer Genome Atlas (TCGA) group, and two other whole genome and whole exome sequencing databases[26, 31]. There are over 70 signatures reported in COSMIC, subdivided into three types – 49 Single Base Substitution (SBS) with 18 possible sequencing artifacts, 11 Double Base Substitution (DBS), and 17 Small Insertion and Deletion (ID) signatures[26, 31]. Of these 77 mutational signatures, 29 of them have unknown aetiologies, which implores for testing on model systems to characterize the signatures of diverse environmental and biological mutagenic agents[25].

1.3 A Known Crosslinking Agent called Formaldehyde

Although known for its genotoxic properties[32–36], formaldehyde’s mutational signature remains ambiguous[27, 36]. Formaldehyde (FA) is a colourless, flammable chemical with a pungent odour[35]. It is a highly reactive aldehyde (chemical formula = CH_2O) widely used in industry, mainly for the production of resins with urea, phenol, melamine and polyacetal[35, 36]. These can then be used as adhesives for several wood

products, such as plywood, particleboard, and fiberboard; pulp-and-paper; plastics and coatings; and textile finishing. It is also used as an intermediate in the manufacture of industrial chemicals; and in aqueous form known as formalin, it is used as a disinfectant and preservative in mortuaries and medical laboratories[35]. It is estimated that over 20 million tons of formaldehyde is produced globally each year[36].

Apart from commercial production, formaldehyde also exists as a naturally occurring substance in most living systems and the environment[37, 38]. Formaldehyde has been established as a normal by-product of various metabolic activities, such as methanol oxidation[39], folate-derivative breakdown[37, 40], methylamine deamination[41–43], histone demethylation[37, 44], and amino acid metabolism[45–47]. Endogenous formaldehyde was estimated to be 0.05 – 0.10 mM in the blood of healthy individuals, but intracellular levels can reach up to 0.20 – 0.50 mM[48]. Accidental or intentional intake of environmental sources can further increase physiological formaldehyde levels.

Humans are further exposed to formaldehyde through consumption of fruits and some foods like coffee, codfish, meat, poultry, and maple syrup[35, 38]. Commonly used commercial products also contain or release formaldehyde such as fertilizers, fumigants, paints, cosmetics, cleaning agents and toiletries. It is also a by-product of combustion reactions and found in smokes and fumes of vehicles, burning wood, coal, tobacco, kerosene and natural gas. While “formaldehyde is a natural component of ambient air”[35], anthropogenic sources and activities contribute to increased levels in both indoor and outdoor air. Although humans are generally exposed to less formaldehyde outdoors (about

less than 0.001 mg/m^3 in rural regions and 0.02 mg/m^3 in urban areas), they are typically exposed to $0.02 - 0.06 \text{ mg/m}^3$ indoors, which can vary depending on the dwelling's source materials, the age of these materials, human activities (such as cooking or composting), ventilation, temperature and humidity[35, 49]. In general, formaldehyde exposure is higher in residential homes or workplaces[35, 50], especially if they were newly built or recently renovated[35, 49, 51].

1.4 Formaldehyde's Impact on Public Health

In Canada, the average person's residential exposure to formaldehyde is "generally below the levels recommended by Health Canada" (100 ppb or $123 \text{ }\mu\text{g/m}^3$ for short-term exposures; and 40 ppb or $50 \text{ }\mu\text{g/m}^3$ for long-term exposures)[49]. On the other hand, occupational exposures are at least 3 ppm or higher (or $\geq 3.7 \text{ mg/m}^3$)[36, 51], which are reported short-term exposures for embalmers, pathologists and paper workers. Based on epidemiological studies, high, acute formaldehyde exposures (such as occupational exposures) could result in the development of certain cancers[36], while reports of carcinogenesis from low, chronic exposures (such as residential exposures) are essentially zero[49]. Both human and animal studies have shown several evidences regarding formaldehyde's carcinogenicity, and this led the International Agency of Research on Cancer (IARC) to classify formaldehyde as a high priority carcinogen (or group I carcinogen)[35, 36].

IARC concluded that formaldehyde exposure can cause the development of sinonasal cancer, nasopharyngeal cancer, and leukemia[36]. The first evidence of

carcinogenicity for formaldehyde was observed in rats when high levels of formaldehyde inhalation induced the formation of squamous cell carcinomas[52]. Genotoxic effects were also observed in multiple *in vitro* systems, in which formaldehyde was shown to form DNA strand breaks, chromosomal aberrations, micronuclei, and crosslinks between DNA and protein (reviewed extensively by IARC[35, 36]). Since formaldehyde exists at a relatively high concentration in human blood (approximately 0.10 mM), this alludes to the role of endogenous formaldehyde as a possible carcinogen. In fact, studies on mice[53], human cells[54] and chicken cells[55] demonstrated that physiological concentrations of formaldehyde can cause significant damages to their blood cells when protective enzymes responsible for formaldehyde detoxification and damage repair were removed. In addition, the known formaldehyde genotoxic effects, such as those described here, were observed in mice blood cells when, again, these protective enzymes were absent[56].

Although the consequences of formaldehyde exposure are well-recognized, there is little understanding regarding its mutagenesis at the nucleotide level. There have been many inferences on formaldehyde's mechanism of action towards DNA mutagenesis[14, 32, 57, 58]. For instance, it has been hypothesized that formaldehyde most likely reacts with amino groups of biological components, such as amino acids or amino purines. Studies on different model systems[32, 35, 59, 60] demonstrated that a chemical analog of formaldehyde may be responsible for its mutagenic effect. However, this compendium of results suggests that formaldehyde acts as a weak mutagen and its mutagenesis vary from model to model. To address this issue, our strategy is to study formaldehyde mutagenesis

on a highly sensitive model system, and analyze these mutations through mutational signature analysis.

1.5 Budding Yeast as a Model System to Study Mutations

Saccharomyces cerevisiae is a well-established model eukaryote since numerous of its key cellular, metabolic and signaling pathways have been identified and extensively characterized[61, 62]. Furthermore, these biological processes are conserved within several eukaryotic species including humans[63, 64], which renders this model organism very attractive for toxicological studies to understand susceptibility in humans[65, 66].

To study the mutations of a weak mutagen like formaldehyde, our yeast model has been designed to efficiently detect mutations. Because some genomic processes require the DNA to be in single stranded form transiently (e.g. replication, transcription and recombination), the nucleotide bases are more readily reactive to various molecules during this period as their “protection” from their hydrogen bond with bases from the second strand is temporarily dismantled[2, 28]. Since single strand DNA (ssDNA) is much more susceptible to lesions[67], our yeast strain was genetically engineered to readily form long stretches of ssDNA, which increases mutation induction by two or three orders of magnitude[68, 69]. This highly sensitive model system is ideal for mutational signature analysis because accurate signature characterization demands for a significant amount of mutation data[25, 29].

To optimize for the highest number of mutations on our yeast strain, ySR127, it was exposed to various concentrations of formaldehyde for three hours: 2 – 10 mM. This

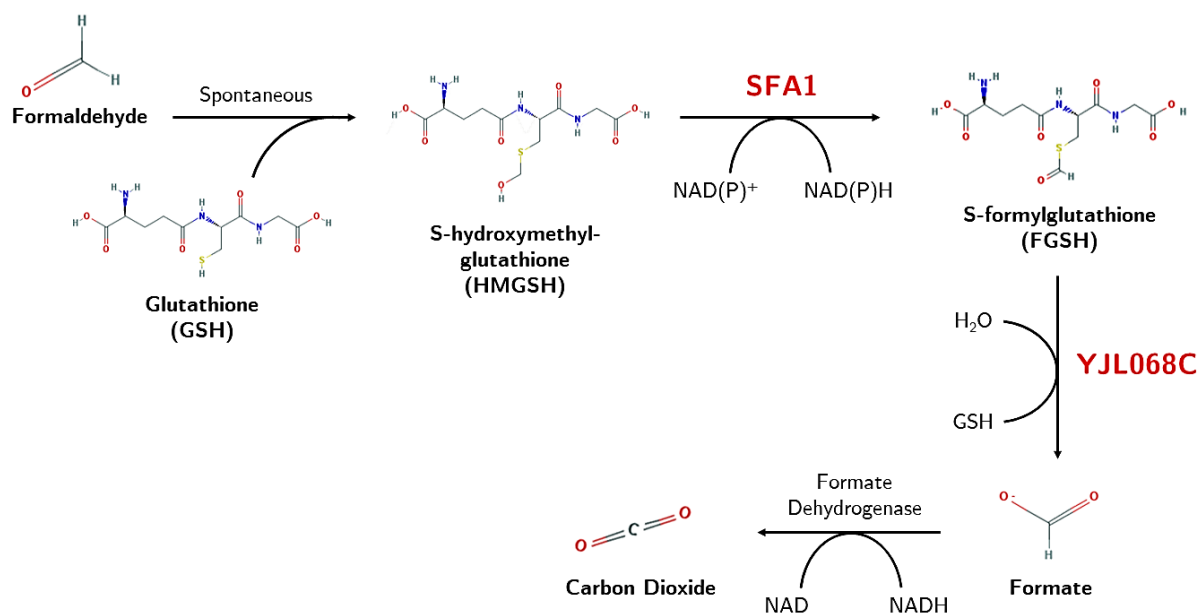
high, acute treatment should allow for multiple mutation induction on the yeast genomes, which will then be subjected to whole genome sequencing and analyzed for mutational patterns. Since these concentrations are not physiological levels, I also tested for the effects of endogenous formaldehyde by knocking out essential dehydrogenases through homologous recombination and measuring the cells' mutation frequencies thereafter.

In *Saccharomyces cerevisiae*, formaldehyde was thought to be detoxified through a glutathione-dependent system[70] via two key formaldehyde dehydrogenases – *SFA1* and *YJL068C*. Null mutants of these dehydrogenases were viable but are highly sensitive to formaldehyde[71–73]. *SFA1* codes for a protein that belongs to class III alcohol dehydrogenases, which are bifunctional enzymes containing an alcohol dehydrogenase and a glutathione-dependent formaldehyde dehydrogenase. Chemicals such as formaldehyde, ethanol and methyl methanesulfonate induces the expression of Sfa1p[74]. It was shown that overexpression of this protein causes an increased resistance to formaldehyde[71, 74]. On the other hand, *YJL068C* codes for an esterase that contains a S-formylglutathione (FGSH) hydrolase activity[73].

In the glutathione-dependent pathway of formaldehyde oxidation in yeast (see Figure 1.1), it is believed to begin with the spontaneous reaction of formaldehyde and glutathione to form S-hydroxymethyl-glutathione (HMGS). HMGS becomes a substrate for Sfa1p that is then oxidized to FGSH, which is hydrolyzed by Yjl068cp to form glutathione and formate. Formate can be further oxidized by a formate dehydrogenase to form carbon dioxide, which is released by the cell. While there might be other key players involved in formaldehyde detoxification in budding yeast[65, 75], Sfa1p

and Yjl068cp were the only enzymes directly implicated in formaldehyde metabolism thus far, and therefore, were of main interest in this study.

Figure 1.1. The Glutathione-dependent Formaldehyde Oxidation Pathway in Saccharomyces cerevisiae[70]



1.6 Rationale and Research Objectives

Because mutational signatures can be used “as prognostic indicators, as predictors of therapeutic sensitivity or as targets of disease control”[5], they are valuable benchmarks for improving cancer detection and prevention, as well as overall public health. Given the ubiquitous nature of formaldehyde, providing a stronger evidence for its mutational capacity is of great importance to further curb environmental exposures.

Although a recent study was able to characterize the mutational signatures of various environmental mutagens, some of their results were inconclusive, including the

formaldehyde mutational signature[27]. This was possibly due to the weak mutagenic nature of formaldehyde, which their model system was unable to precisely detect. Our highly sensitive yeast strain, on the other hand, allows us to detect and induce multiple mutations at a frequency of at least 1×10^{-3} mutations per cell[68].

My hypothesis is that the formaldehyde mutational signature can be effectively characterized using our yeast model, and this signature can be used to determine formaldehyde's contribution in cancer development. To address this, the specific research objectives are as follows:

1. Generate yeast mutants through exogenous exposure of formaldehyde while examining the chemical's mutagenic effects ([Section 3.1](#));
2. Characterize the mutational signature of formaldehyde using yeast genomics data ([Section 3.2](#)); and
3. Determine the prevalence of formaldehyde in cancers using its putative mutational signature ([Section 3.3](#)).

In detail, Objective 1 explores the effects of high-acute and low-chronic exposures of formaldehyde. Objective 2 examines the mutational patterns of single nucleotide variants and insertion/deletion variants from yeast genomes. And finally, Objective 3 will assess the relevance of the yeast-derived mutational signature by comparing its similarity to known cancer signatures, as well as analyze Head and Neck cancer data as specific formaldehyde-related cancers are encapsulated within this group, namely nose and throat cancers.

2. Methodology

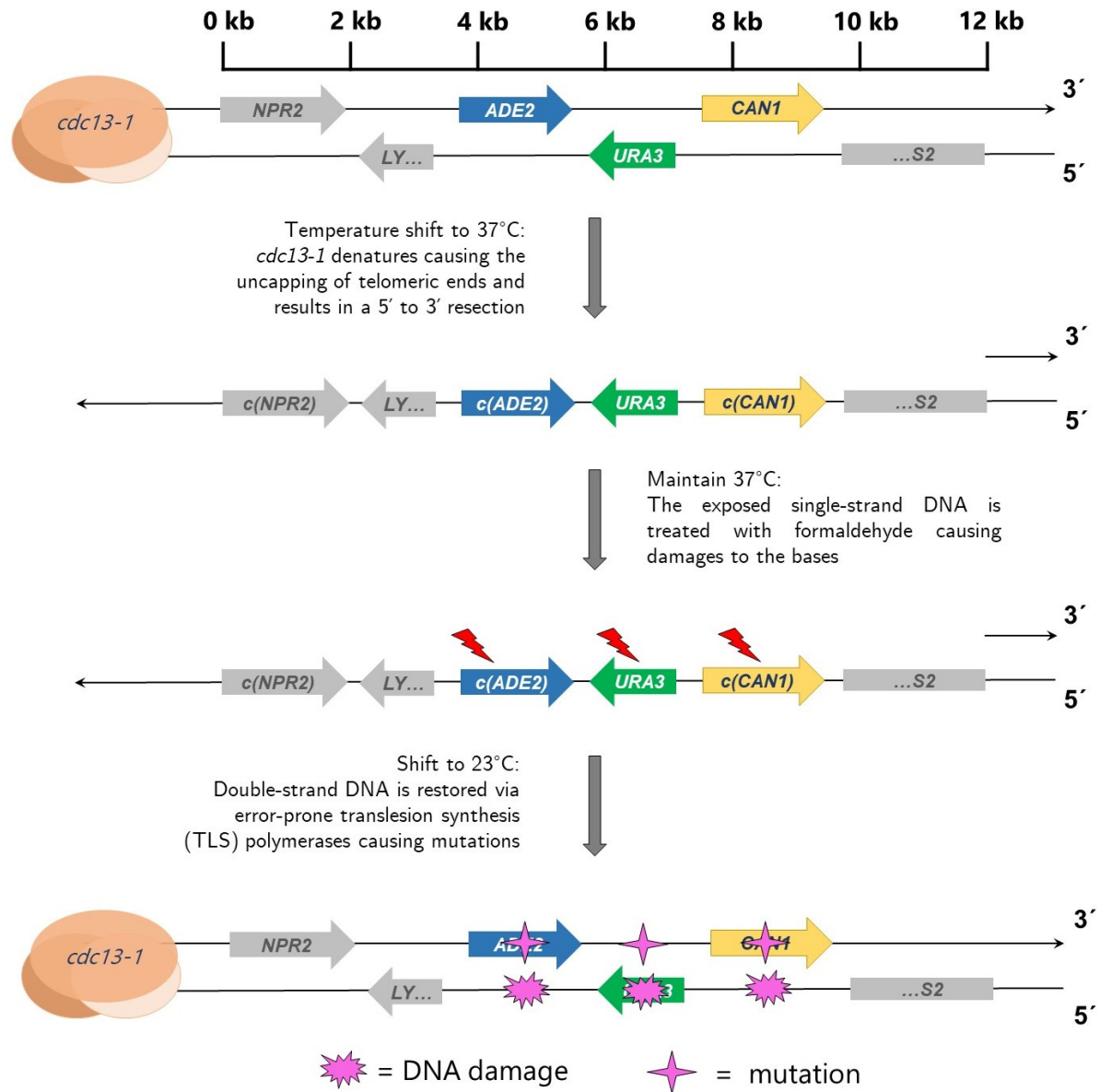
2.1 Highly Sensitive Yeast Model System, ySR127[68]

Our *Saccharomyces cerevisiae* was genetically engineered so that three genes – *CAN1*, *ADE2*, *URA3* – were translocated from their native loci to the left arm of the subtelomeric region of Chromosome V (see Figure 2.1). These genes serve as the reporter system for detecting multiple mutation induction in our haploid yeast cell. Furthermore, the temperature-sensitive mutation at *CDC13* (*cdc13-1*), which codes for the protective telomeric ends of chromosomes, plays an integral role in mutation accumulation. Upon shifting to a restrictive temperature (i.e. 37°C), the *cdc13-1* mutation will cause the protective complex to dissociate and uncap the telomeres, which then results in an 5' to 3' strand resection producing a long single strand DNA (ssDNA) overhang. This extensive strand resection (of at least 30 kb) encompasses the entire reporter system.

While maintaining the temperature at 37°C, cells are exposed to formaldehyde for three hours. The temperature is then shifted back to a more permissive temperature, 23°C, to allow the restoration of the double strand DNA configuration. Since damages were experienced by the ssDNA, the error-prone TLS polymerases will bypass the formaldehyde-induced lesions, oftentimes by introducing an incorrect base. If these erroneous bases are left unrepaired or repaired incorrectly, they become mutations. The induction of multiple mutations through the loss of function of reporter genes can be detected by plating on selective media.

Figure 2.1. Schematic Diagram of our Highly Sensitive Yeast Model, *ySR127*

(Adapted from Chan et al., 2012)



2.2 Exogenous Formaldehyde Treatment (High, Acute Exposure)

A single colony isolate of *Saccharomyces cerevisiae* ySR127 was cultured in 5 ml of YPDA media for three days at 23°C (see Figure 2.2, steps 1a – 1b). On the first day of the experiment, the cell concentration of the three-day culture (or pre-growth) was measured using the hemocytometer. Approximately 500 cells were plated on synthetic complete (SC) media to assess viability and 1×10^7 cells were plated on canavanine-supplemented (CAN) media to examine mutation frequency (see Figure 2.2, steps 2a – 2b). This provides a baseline measurement of viability and mutation frequency of non-treated cells. Next, 3.50 ml of YPDA media was aliquoted into six fresh tubes with 0.50 ml of the pre-growth culture. These tubes were labelled “0”, “2”, “4”, “6”, “8” and “10” to denote the final concentrations of formaldehyde to be added later in millimolar (mM). They are then sealed tight with a septum and equilibrated at atmospheric condition with either room air (79% nitrogen, 21% oxygen) or nitrogen for one hour at 23°C. At the end of the hour, the temperature was shifted to 37°C and incubation continued for three hours with continuous gas flow (see Figure 2.2, steps 3 – 4). After three hours, 1 ml of formaldehyde was added into the gas-tight tubes and the atmospheric condition was equilibrated for 10 more minutes under 37°C. Finally, the gas feed was disconnected, and formaldehyde exposure continued for three more hours at 37°C (see Figure 2.2, step 5).

After three hours of exogenous exposure of formaldehyde, the cell concentration of one millilitre samples were measured using the hemocytometer to plate 500 cells on SC and 5×10^5 to 1.5×10^6 cells on CAN media (see Figure 2.2, step 6). Plating volumes and

cell numbers were optimized for each treatment (see Table 2.1) in order to minimize clumps when counting colony forming units using the ImageJ plugin Colony Counter[76].

After plating the appropriate number of cells, they were incubated at 23°C for about five to seven days, and are then transferred to 4°C for about five to seven more days to allow for a stronger resolution of the red phenotype of mutants growing on CAN media while minimizing the growth of the colonies any further (explained on Section 2.3 and Figure 2.3). Colony growth on the CAN media signifies a mutation on the *CAN1* reporter gene, and the red phenotype denotes a second mutation found on the *ADE2* reporter gene. These red colonies are then randomly selected and screened for further genomic analysis.

Two parameters are calculated in order to examine the effect of formaldehyde mutagenesis at each concentration: viability and mutation frequency. Viability (growth on SC) was analyzed by dividing the plating efficiency of treated cells by the plating efficiency of pre-growth (or non-treated cells). Plating efficiency was calculated by dividing the actual number of colonies on the SC media by the expected number of colonies (i.e. the number of cells plated). Mutation frequency for a treatment was calculated by dividing the plating efficiency on CAN to the plating efficiency on SC.

$$PLATING\ EFFICIENCY = \frac{Total\ number\ of\ colonies}{Number\ of\ cells\ plated}$$

$$VIABILITY = \frac{Plating\ efficiency\ of\ treated\ cells\ on\ SC\ media}{Plating\ efficiency\ of\ pregrowth\ on\ SC\ media}$$

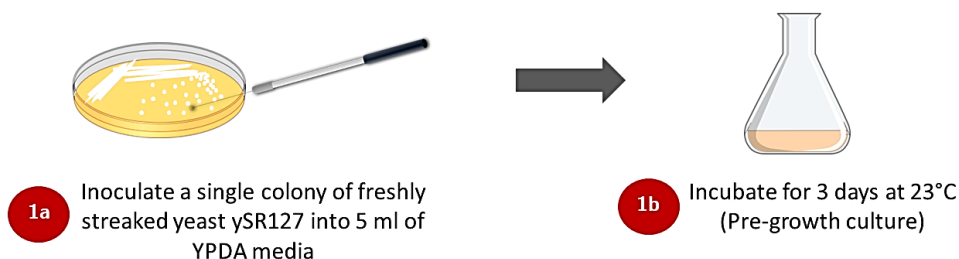
$$MUTATION\ FREQUENCY = \frac{Plating\ efficiency\ on\ CAN\ media}{Plating\ efficiency\ on\ SC\ media}$$

Table 2.1. Optimized number of cells plated for each treatment

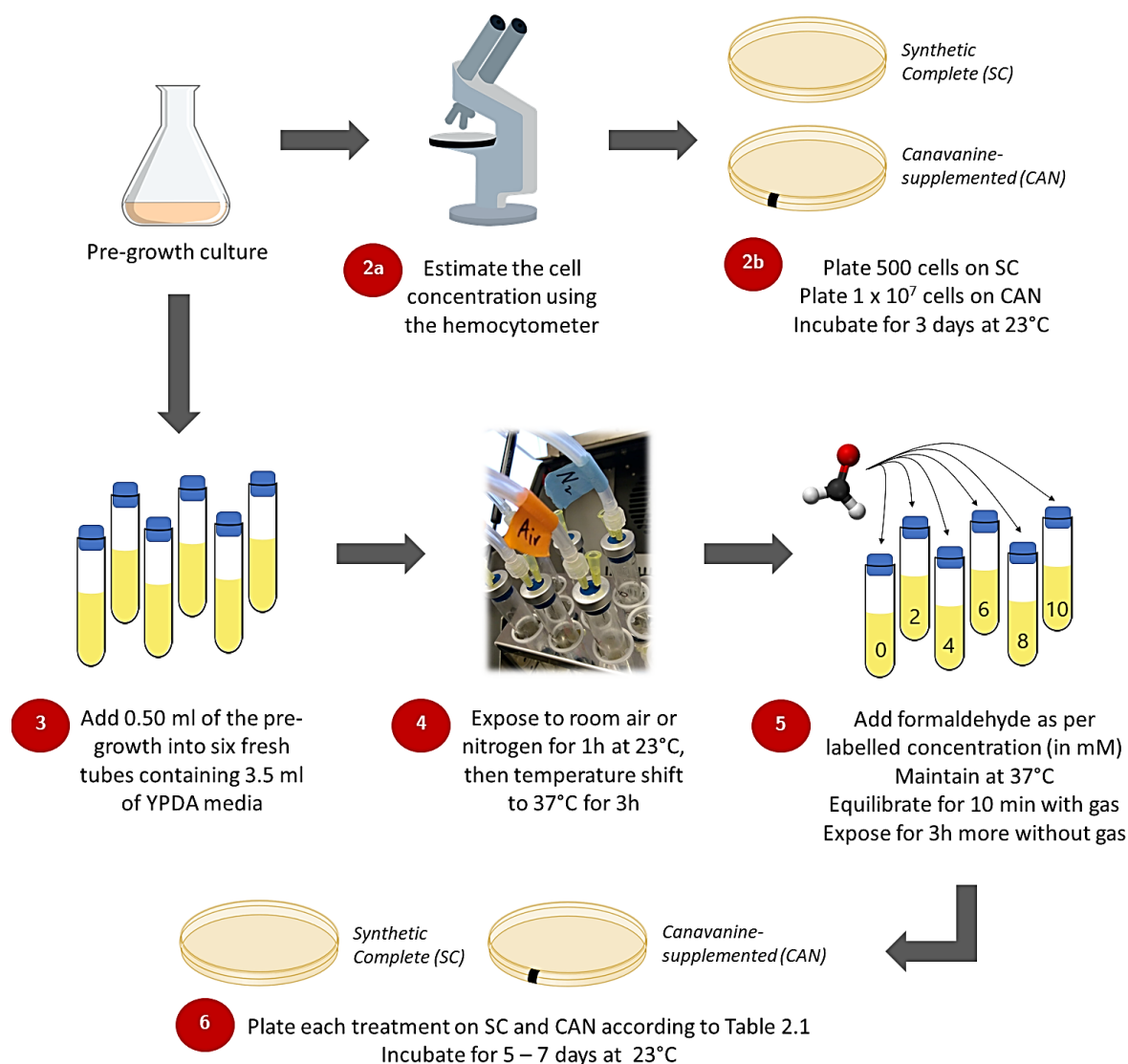
Formaldehyde concentration	Number of cells plated on SC media	Number of cells plated on CAN media
0 mM	500 cells	1.0 x 10 ⁶ cells
2 mM	500 cells	5.0 x 10 ⁵ cells
4 mM	500 cells	5.0 x 10 ⁵ cells
6 mM	500 cells	5.0 x 10 ⁵ cells
8 mM	500 cells	5.0 x 10 ⁵ cells
10 mM	500 cells	1.5 x 10 ⁶ cells

Figure 2.2. Experimental Procedure for the High, Acute Treatment of Exogenous Formaldehyde

Experiment preparation:



On the day of the experiment (formaldehyde exposure):



2.3 Mutant Screening and Selection

The randomly selected red colonies from CAN plates are streaked onto YPDA media (yeast extract, peptone, D-glucose, and 0.55% adenine sulfate) to isolate single colonies. These single colonies are then patched out on a fresh YPDA plate and replica plated onto various selection media – synthetic complete (SC) media to test for viability; glycerol (YPG) media to assess mitochondrial activity; canavanine-supplemented (CAN) media to detect *CAN1* gene inactivation; adenine-dropout (ADE) media to detect *ADE2* gene inactivation; and uracil-dropout (URA) media to detect *URA3* gene inactivation. Cells displaying the phenotype described in Table 2.2 are selected as ideal candidates for DNA sequencing and further genomic analyses.

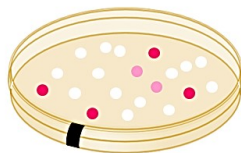
Table 2.2. Selection criteria for DNA sequencing candidates

Genotype	Description	Indication on selection media
Can ^r	Canavanine-resistant or <i>CAN1</i> inactivation	Growth on CAN media
Ade ⁻	<i>ADE2</i> inactivation	No growth on ADE media
Ura ⁺	Functional <i>URA3</i>	Growth on URA media
Rho ⁺	Normal metabolism	Growth on YPG media

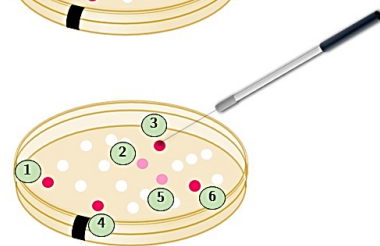
Figure 2.3. Standard Procedure for Mutant Screening and Selection

After five to seven days of incubation at 23°C...

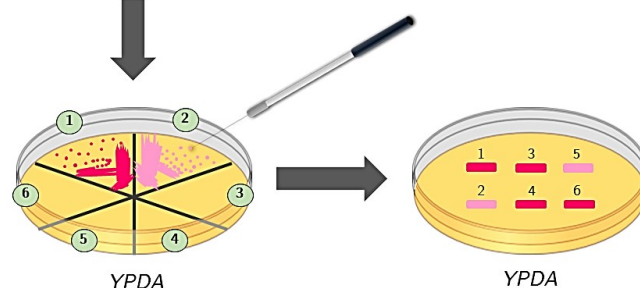
- 1 Transfer CAN plates to 4°C for 5 – 7 days to allow stronger resolution of the red colonies



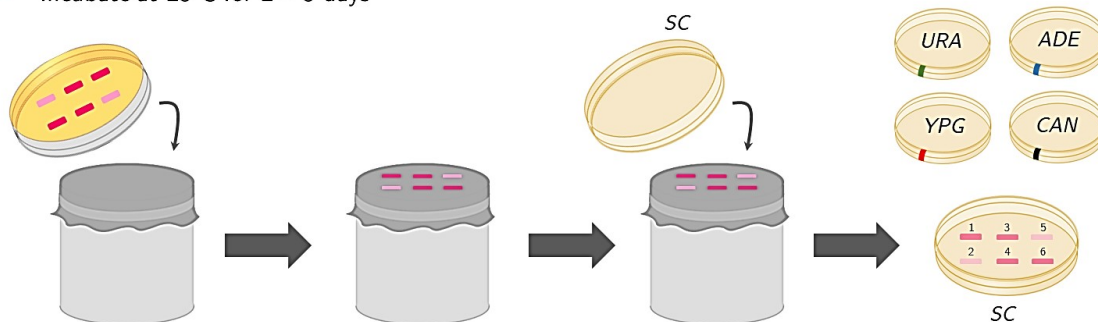
- 2 Pick the red colonies and streak them onto fresh YPDA media to isolate for single colonies
Incubate at 23°C for 3 – 5 days



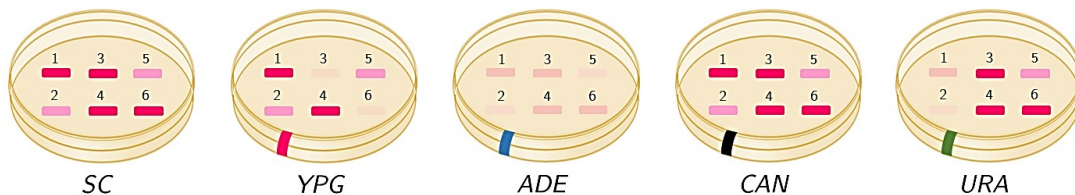
- 3 Patch out the single colonies onto another fresh YPDA media
Incubate at 23°C for 2 – 3 days



- 4 Replica plate onto various selection media
Incubate at 23°C for 2 – 3 days



- 5 Select mutants with the correct phenotype: **Can⁺ Ade⁻ Ura⁺ Rho⁺**
(In this example, mutants 4 and 5 are good candidates for sequencing)



2.4 The Auxin-Inducible Degron (AID or Degron) System on ySR127

To study the effects of endogenous formaldehyde in our yeast, our strategy was to construct single and double knockouts of essential formaldehyde dehydrogenases. While the construction of single knockouts was successful, I was unable to produce viable double knockout strains and therefore, the AID system was utilized as an alternative to the “double knockout” strain. The Auxin-Inducible Degron (AID), or simply Degron, System is a biological mechanism used to control gene expression[77]. The basic principle is the utilization of the SCF-TIR1 (Skp1 – Cul1 – Fbox Transport Inhibitor Response I) complex to promote a ubiquitylation response under the presence of an auxin inducer that leads to protein degradation. Since yeast has its own form of SCF, a *TIR1* gene needed to be introduced as well as a degron tag on the protein of interest[78]. Because it was found to be the most efficient in protein depletion, the *TIR1* gene of *Oryza sativa* (rice)[77] was introduced into ySR127 using the methods described by Papagiannakis et al. (2017)[79]. The degron tag was then introduced after successful strains with the *OsTIR1* gene have been generated. Three different auxin-inducible degrons have been developed – AID (plasmid pSM409); mini-AID or mAID (plasmid pSM412); and 3x mini-AID or 3mAID (plasmid pMK154)[78]. All three degron tags were tested in the endogenous formaldehyde experiments (see Section 2.6). The introduction of degron tags in OsTIR1 strains (i.e. yeast strain yKC522) are described in the next section, Section 2.5.

2.5 Construction of Genetically Modified Yeasts

The construction of knockout and degon-tagged yeasts were adapted from the *Delitto Perfetto* system by Storici and Resnick (2006)[80] and Auxin-Inducible Degron (AID) protocol by Nishimura and Kanemaki (2014)[78] respectively, modified for the ySR127 strain. Primers for the single-knockout and degon constructs were designed using the SnapGene software[81] and oligos were synthesized by ThermoFisher Scientific[82]. Primers contain at least 60 base-pairs upstream (5' primer) or downstream (3' primer) of the gene of interest, as well as the primer-tag of the antibiotic marker (see Primer Design on Figures 2.4 and 2.5).

Genetic construction begins with the amplification of the antibiotic or selection marker plasmids using the synthesized primers via PCR (as shown in Figures 2.4 and 2.5, step 1). The PCR products, which is now referred to as a “gene replacement cassette”, were purified and concentrated using the BioBasic EZ-10 Spin Column PCR Products Purification Kit (Catalogue number: BS363/BS36/BS664). The cassette was used to delete the gene-of-interest or introduce the degon tag adjacent to the gene-of-interest (see Figures 2.4/2.5, step 2). Both transformations of yeast strains ySR127 and yKC522 – to produce knockout and degon strains respectively – involved using lithium acetate and Tris-EDTA buffer solutions.

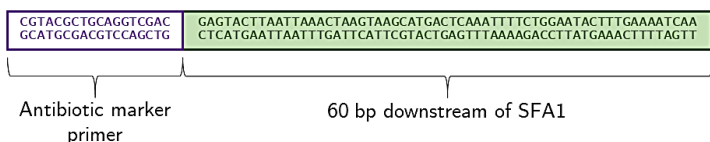
Figure 2.4. Genetic Modification of *ySR127* using the *Delitto Perfetto* System[80] to Create Single-Gene Knockouts of *SFA1* and *YJL068C*

Primer Design:

SFA1 5' primer



SFA1 3' primer



YJL068C 5' primer

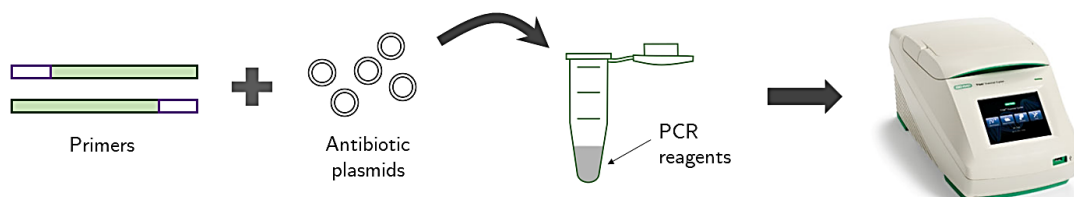


YJL068C 3' primer

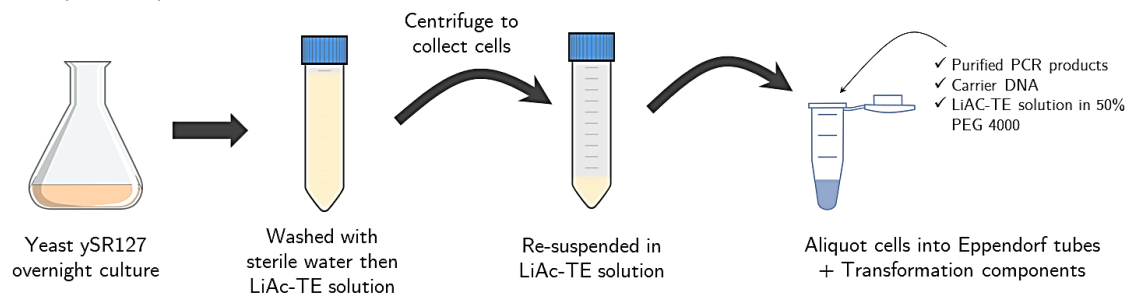


Transformation of *ySR127*:

- 1 PCR amplification of selection marker (antibiotic) using constructed primers to create a gene replacement cassette (PCR product).



- 2 Prepare the yeast *ySR127* culture for transformation using a Lithium Acetate – Tris – EDTA (LiAc-TE) solution.



- 3 Introduce cassettes into cells via heat shock treatment then plate on YPDA media before replica plating onto selection media. Verify constructs via PCR.

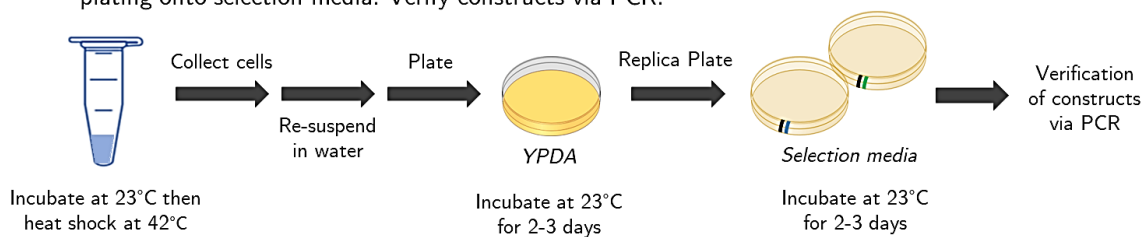
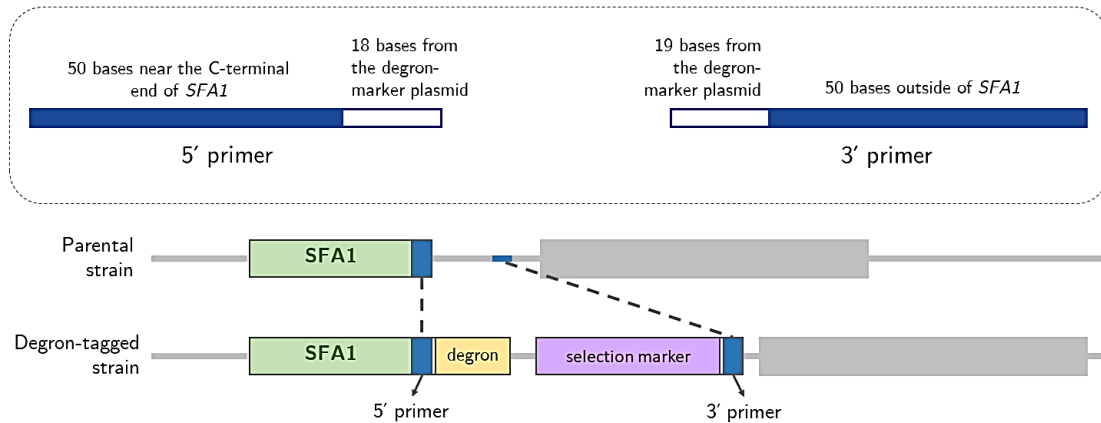


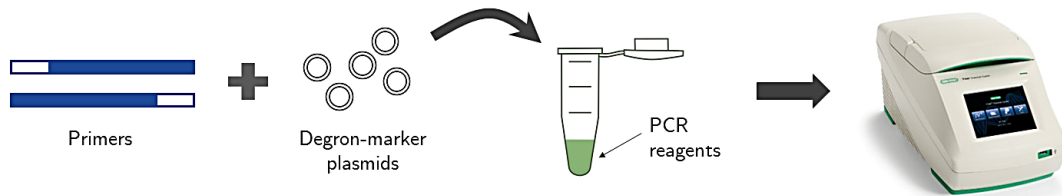
Figure 2.5. Construction of Auxin-Inducible Degron (AID) strains (Adapted from Nishimura and Kanemaki, 2014)

Primer Design:

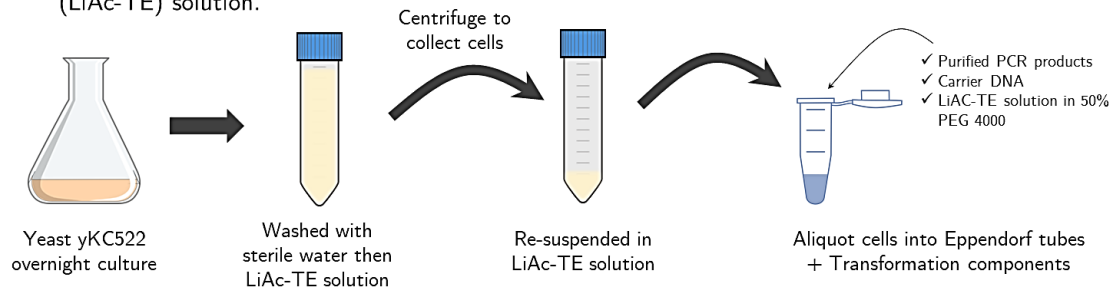


Transformation of yKC522:

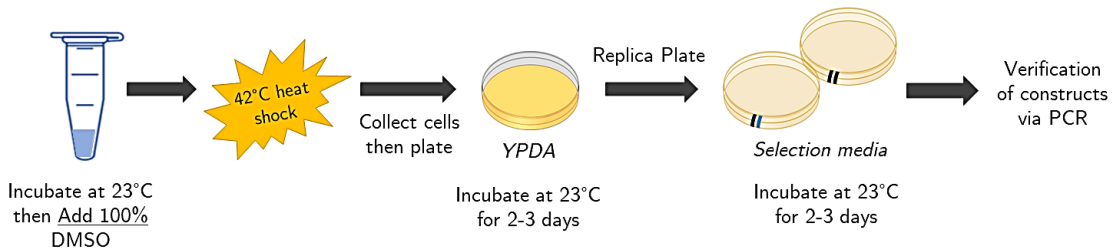
- 1 PCR amplification of degron-marker tags using the constructed primers (PCR products).



- 2 Prepare the yeast yKC522 culture for transformation using a Lithium Acetate – Tris – EDTA (LiAc-TE) solution.



- 3 Introduce degron tags into cells via heat shock treatment then plate on YPDA media before replica plating onto selection media. Verify constructs via PCR.



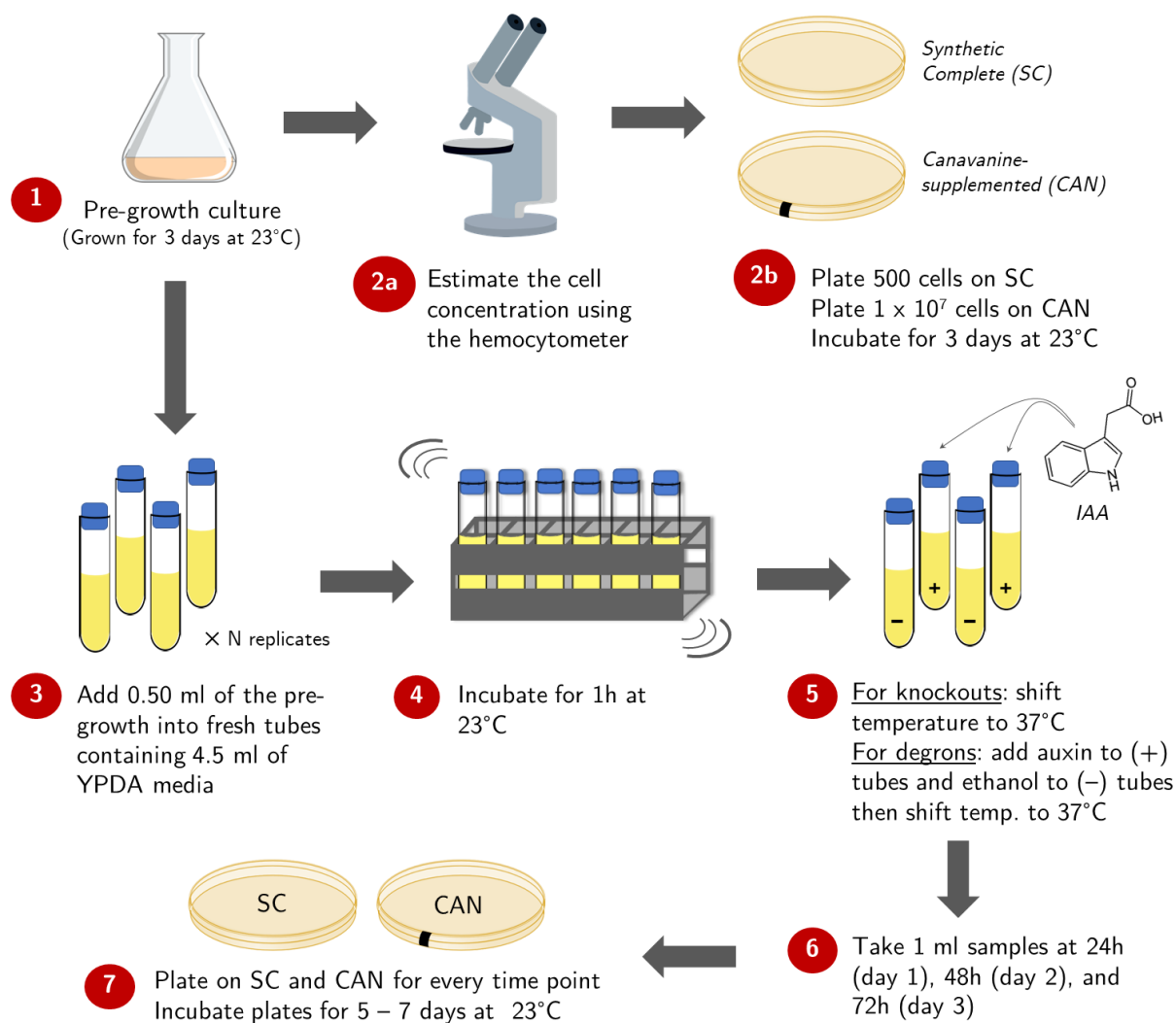
2.6 Endogenous Formaldehyde Treatment (Low, Chronic Exposure)

For the single-knockout strains: A single colony isolate of *sfa1Δ* single knockout or *yjl068cΔ* single knockout was cultured in 5 ml of YPDA media for three days at 23°C (see Figure 2.6, step 1). On the third day, 500 cells were plated on synthetic complete (SC) media and 1×10^7 cells on canavanine-supplemented media (CAN) (see Figure 2.6, steps 2a – 2b). These served as day 0 or “no treatment” measurements. Next, 0.50 ml of the three-day culture was aliquoted into fresh tubes containing 4.5 ml of YPDA media and incubated for one hour at 23°C (see Figure 2.6, steps 3 – 4), before the temperature was shifted to 37°C (see Figure 2.6, step 5 for knockouts). Incubation then resumed for three more days (approximately 72 hours in total) with continuous gas flow. One-millilitre samples were taken every 24 hours – days 1, 2 and 3 – and plated onto SC and CAN media to assess viability and mutation frequency, respectively (see Figure 2.6, steps 6 – 7). A native ySR127 strain served as a control for this experiment.

For the degron strains: A single colony isolate of a *sfa1*-degron or *sfa1*-degron/*yjl068cΔ* strain (i.e. *SFA1* was tagged with a degron and *YJL068C* was deleted) was cultured in 5 ml of YPDA media for three days at 23°C (see Figure 2.6, step 1). On the third day, samples were plated as above for day 0 or “no treatment” measurements (Figure 2.6, steps 2a – 2b). Next, 0.50 ml of the three-day culture was aliquoted into two fresh tubes containing 4.5 ml of YPDA media and incubated for one hour at 23°C (Figure 2.6, steps 3 – 4). At the end of the hour, one tube received a dose of auxin, indole-3-acetic acid or IAA, to a final concentration of 500 μ M. Since the auxin was dissolved in ethanol, the other tube (negative control) received a dose of ethanol (final concentration of 500

μM) in anticipation of the alcohol's confounding effects. The temperature was shifted to 37°C and incubation resumed for three more days (approximately 72 hours in total) (see Figure 2.6, step 5 for degrens). One-millilitre samples were taken every 24 hours and plated onto SC and CAN. A native ySR127 strain also served as a control for this experiment (Figure 2.6, steps 6 – 7).

Figure 2.6. Experimental Procedure for the Low, Chronic Treatment of Endogenous Formaldehyde using Genetically Modified Yeasts



2.7 Statistical Analyses for the Exogenous and Endogenous Formaldehyde Experiments

Viability and mutation frequency data (i.e. Can^r and Can^r Ade⁻ frequencies) for exogenous formaldehyde exposures were plotted using interquartile range. Assuming the data follow a parametric distribution with unequal variances, Welch's t-test was used to assess differences in mutagenicity under aerobic (room air) and anaerobic (nitrogen) metabolism for each concentration (0 to 10 mM FA). In addition, Tukey's multiple comparisons test was used to examine significant differences between concentrations for each oxygenation condition (see Figures 3.1.1, 3.1.2, and 3.1.3). The alpha value was set to 0.05 for all tests.

In the endogenous formaldehyde exposures, mutation frequencies between wild-type, *yjl068c* Δ and *sfa1* Δ were also compared using Welch's t-test (see Figure 3.1.4). Mutation frequencies of pre-growth cultures and day-to-day auxin treatments of degtron-knockout strains were assessed using the Tukey's multiple comparisons test (see Figures 3.1.6 and 3.1.7). The alpha value was set to 0.05 for all tests.

2.8 Single Base Substitution Signature Analysis of Formaldehyde-treated Yeast Genomes using the MutationalPatterns Software

Successful candidates screened from the exogenous formaldehyde experiments (see Section 2.2) were used for the mutational signature analysis of formaldehyde. First, DNA was extracted using zymolyase, potassium acetate and 2-propanol. Sequencing was performed by Genome Quebec using the Illumina HiSeqX technology generating paired-end reads of 150 bases. Genome data was mapped using ySR127 as a reference genome (GenBank Assembly Accession: GCA_001051215.1) through Bowtie2 (version 2.3.4.1)[83] and SAMtools (version 1.9)[84]. Variants were called and filtered using BCFtools (version 1.9)[85] generating a final mutation file (called a VCF or variant call format file). Filtering for unique variants ensures that all mutations analyzed can be uniquely attributed to formaldehyde mutagenesis. Unique variant calls were selected using the following criteria:

- a. Quality Score – variant calls with Phred quality scores of 30 and above, which signifies a base call accuracy of at least 99.9%
- b. Depth Score – calls reported from 10 or more reads, i.e. coverage with ≥ 10 reads
- c. Redundancy – calls that were reported more than once across all samples were discarded

The collection of VCF files was merged into one and then analyzed using an R package called MutationalPatterns[86] to extract mutational signatures. Untreated yeast genomes were also included in the analysis for comparison.

MutationalPatterns utilizes an algorithm called non-negative matrix factorization or NMF to deconvolute patterns of mutations within a set of data. The NMF algorithm

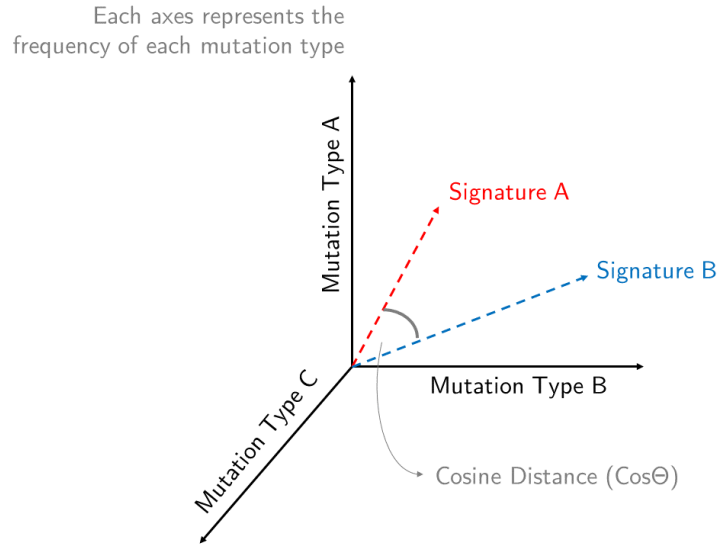
requires a ‘seed’ number and a ‘factorization rank’ number, r , to begin its calculation. A ‘seed’ number is any random number that serves as a starting point for the algorithm to initialize its calculation. The ‘factorization rank’, r , number defines the number of signatures required to approximate the profiles of the final mutational signatures. In other words, it is the expected number of signatures within a given data set. Since this number is unknown, “a common way of deciding on r is to try different values, compute some quality measure of the results, and choose the best value according to this quality criteria” [87]. This is called an NMF rank survey. There are several different quality measurements that can be used to choose a r value, but the most common approach is the cophenetic correlation coefficient[88]. The ‘best’ value for r is the smallest number for which the coefficient starts decreasing. Performing 30 – 50 runs of the NMF algorithm given a range of r values is considered sufficient to obtain a robust estimate of the factorization rank[87].

MutationalPatterns can also assess the similarity of two mutational signatures by calculating their cosine similarity and residual sums of squares. Much of the analysis reported here uses the cosine similarity value to interpret the significance of results. Cosine similarity is a metric that measures the similarity of two vectors through their cosine distance[89]. In mutational signature analysis, each signature is treated as a vector projected over axes of 96 mutation types, as per the COSMIC classification of single base substitutions (see Figure 2.7). The smaller the cosine angle between the two signatures, the higher their cosine similarity with 0 being the lowest and 1 being the highest value. Thus, a cosine similarity value of one ($\text{cossim} = 1$) signifies identical signatures.

Altogether, the NMF algorithm was used to decipher the putative formaldehyde single base substitution (SBS) signature, and this signature was compared to those found in COSMIC (Catalogue of Somatic Mutations in Cancer), which is the collection of known mutational signatures from cancer genomes[31], using cosine similarity.

Figure 2.7. Vector Diagram of the Cosine Similarity Algorithm with respect to Mutational Signature Analysis

Each 96-mutation type is represented as their own axis or dimension, and the vector mutational signatures are plotted over this multi-dimensional space. The cosine distance of their angle will determine their degree of similarity – the smaller the cosine angle, the more similar the two signatures are.



2.9 Insertion/Deletion Signature Analysis of Formaldehyde-treated Yeast Genomes

Indel data from formaldehyde-treated and untreated yeast genomes were also filtered using the criteria enumerated from the previous section (Section 2.8). In general, indel (ID) signature analysis was performed using base R functions as well as the STRAH package (Short Tandem Repeats Analysis of Hotspot Zones)[90].

Indel classification was based on indel lengths and sequence motif repetitions. First, insertion and deletion lengths of formaldehyde-treated and untreated (or no formaldehyde) yeast genomic data were plotted as bar graphs to visualize their overall distribution. Next, the sequence context of each of these indel lengths were classified based on their number of occurrences within the sequence in which they were located. Since an insertion or deletion is usually a repeat unit of its target sequence, the STRAH package was used for searching and tallying matches of tandem repeats within a sequence context[91]. Overall, the indel classification described here yields 50 possible mutation types – classified by the indel length, the number of repeat units in the sequence context, and whether it is a deletion or insertion (see Figure 3.2.5).

Cosine similarity was also calculated between the putative ID signatures of formaldehyde and no formaldehyde treatments to assess their uniqueness. Similar to the calculations performed in the SBS analysis, each putative ID signature were treated as a vector projected over axes of the 50 possible mutation types. The closer the cossim value is to 1, the more similar the signatures are.

2.10 Single Base Substitution Signature Analysis of Head and Neck Cancer Genomes using the MutationalPatterns software

The MutationalPatterns package was used to perform a *de novo* extraction of signatures from head and neck squamous cell carcinoma (Head-SCC) cancers. Nasopharyngeal and sinonasal cancer samples, which were strongly associated with formaldehyde exposure, are grouped within the head and neck cancer data[92]. The entire Head-SCC dataset was analyzed because there was no unique identifier for each of the cancer types encompassed within this group to allow selection for specific cancer types of interest.

De novo extraction began with conducting an NMF rank survey using Brunet's method of NMF and cophenetic coefficient as the quality metric. After 50 iterations, a cophenetic coefficient value of 0.990 can be achieved with factorization rank number of three, $r = 3$. This is the highest cophenetic coefficient value that could be attained given all the possible r solutions, and thus, a three-signature NMF extraction was performed for the Head-SCC dataset. After NMF characterization, the mutational signatures were compared to the COSMIC signatures as well as the putative formaldehyde signature deciphered from yeast genomic data using cosine similarity.

3. Results

3.1 Generating Mutants for Signature Analysis while Examining the Effects of Formaldehyde in the Yeast Model System

3.1.1 Exogenous Formaldehyde Treatment (High, Acute Exposure)

A large amount of mutation data is required in order to confidently decipher a mutational signature[29]. Our model system can produce hundreds of mutations within a single genome and has been demonstrated to successfully characterize mutational signatures for known mutagens[8, 68, 93]. The first objective of this research seeks to address not only the extent of formaldehyde’s mutagenicity, but also the concentration at which the most amount of mutations can be generated sufficient for mutational signature analysis.

Two parameters were used to examine the mutagenic effect of formaldehyde: viability and mutation frequency. Viability can be measured by colony growth on synthetic complete (SC) media and mutation frequency on canavanine-supplemented (CAN) media. Within the context of this research, ‘mutation frequency’ is interpreted as the number of gene inactivating mutations. Two kinds of mutation frequencies were analyzed in this experiment to understand the rate of multiple mutation induction of formaldehyde within cells – the single-gene inactivation frequency (i.e. *CAN1* gene

inactivation, referred to as Can^{R} frequency) and the double-gene inactivation frequency (i.e. *CAN1* and *ADE2* gene inactivation, referred to as $\text{Can}^{\text{R}}\text{Ade}^-$ frequency).

To begin, I exposed our yeast to a range of formaldehyde (FA) concentrations – 2 mM, 4 mM, 6 mM, 8 mM and 10 mM – including a ‘no formaldehyde’ treatment for control (0 mM), to gauge the extent of its mutagenic effects. Moreover, exposures were conducted under two atmospheric conditions – oxygen-rich (room air; 79% nitrogen, 21% oxygen) and oxygen-deprived (nitrogen) conditions. This is because preliminary data from our group revealed that aerobic respiration vs. anaerobic respiration generated different effects in mutagenesis, where the number of mutations under aerobic respiration are higher than those under anaerobic respiration. Therefore, the effects of formaldehyde mutagenesis could also vary between the two oxygenation conditions.

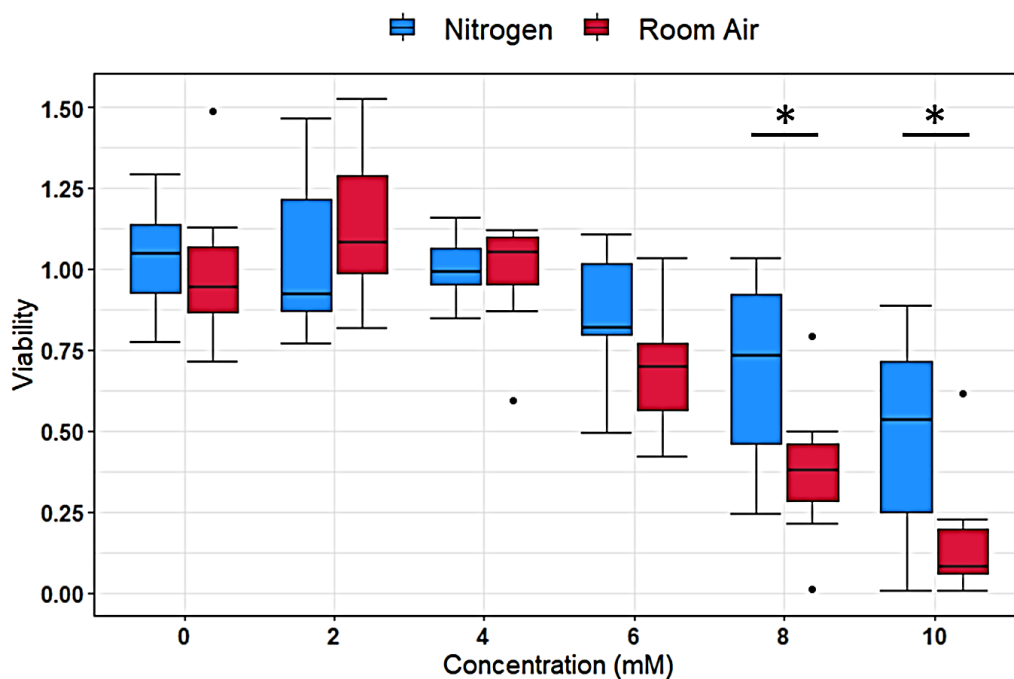
As expected, when no formaldehyde (0 mM FA) is present, cell viability remains relatively the same under both oxygenation conditions. On the other hand, viability was significantly affected after formaldehyde treatment, especially at 8 mM and 10 mM FA in both room air and nitrogen (see Figure 3.1.1). The difference between 0 mM and 8 mM FA was significant in both room air ($p < 0.001$) and nitrogen ($p < 0.05$) by Tukey’s test. Likewise, the difference between 0 mM and 10 mM was significant under both conditions ($p < 0.001$ by Tukey’s test). Viability decreased as much as 47% under 10 mM FA in nitrogen and 92% in room air. Moreover, the decrease in viability in room air is about twice as low as the decrease in nitrogen under these incredibly high formaldehyde exposures.

Although there was no difference in viability between room air and nitrogen under 0 mM FA, there was an observable difference in mutation frequencies between the two oxygenation conditions (see Figure 3.1.2). The mutation frequency under aerobic metabolism is five times higher than anaerobic metabolism. This implies that there exist chemical agents within aerobic metabolism that are sufficiently mutagenic, and added pressure from exogenous mutagens, such as formaldehyde, may further increase the amount of mutations in the genome.

Figure 3.1.1. Effects of Exogenous Formaldehyde Treatment to Viability under nitrogen (anaerobic) and room air (aerobic)

In nitrogen, viability was reduced to ~73% under 8 mM FA and ~53% under 10 mM FA exposure. In room air, it decreased to ~38% and ~8% under 8 mM and 10 mM FA respectively, which is approximately two times less than FA exposure in nitrogen.

* denotes significance at $p < 0.05$ by Welch's t-test for $n = 8$ biological replicates for each treatment.



The Can^R frequency or single-gene inactivation frequency is the rate of inactivating mutation induction on the *CAN1* reporter gene. This was measured by dividing the cell plating efficiency on canavanine-supplemented (CAN) media by the cell plating efficiency on synthetic complete (SC) media (see Section 2.2). *CAN1* codes for an arginine permease, which transports arginine into the cell. Canavanine is a toxic analog of arginine, in which Can1p is also capable of transporting into the cell. If normal cells are exposed to canavanine, it can be incorporated in proteins in lieu of arginine, rendering those proteins non-functional. When a mutation occurs in *CAN1*, canavanine can no longer be introduced into the cell, making these cells resistant to canavanine (Can^R). Thus, colonies growing on CAN media effectively has a mutation on their *CAN1* gene making them canavanine-resistant.

Under formaldehyde exposure, Can^R frequency occurs about an order of magnitude compared to no formaldehyde exposure on average (see Figure 3.1.2). Other than the 10 mM FA exposure, the differences in Can^R frequencies between room air and nitrogen under 0 – 8 mM FA exposure is very significant. Moreover, despite having no differences in viability under 0 – 6 mM FA room air and nitrogen exposures, single-gene inactivation events varied under each of these concentrations with mutation frequencies occurring five times higher in room air than nitrogen on average. It is interesting that even though a significant difference can be observed between room air and nitrogen viabilities at 10 mM FA, the difference in their mutation frequencies is not statistically significant.

The $\text{Can}^{\text{R}}\text{Ade}^-$ frequency or double-gene inactivation frequency examines mutation induction in the *CAN1* and *ADE2* genes. *ADE2* codes for the protein responsible for synthesizing adenine in the yeast cell. If a mutation inactivates the *ADE2* gene (Ade^-), cells display a red pigment indicating adenine deprivation. Therefore, colonies that survived on CAN media that were red or pink in color are indicative of the $\text{Can}^{\text{R}}\text{Ade}^-$ phenotype, i.e. canavanine-resistant cells with an inactivated *ADE2* gene. This implies that these cells have at least two mutations in their genome.

Overall, double-gene inactivation events occur at a frequency that is two orders of magnitude lower than single-gene inactivation events (see Figure 3.1.3). Under no formaldehyde exposure (0 mM FA), the difference in $\text{Can}^{\text{R}}\text{Ade}^-$ frequency between room air and nitrogen is significant. And while double-gene inactivation frequencies were generally higher when exposed to formaldehyde, only under 4 mM and 6 mM FA was there a statistically significant difference between room air and nitrogen.

Looking at the overall trend of both mutation frequencies (Can^{R} and $\text{Can}^{\text{R}}\text{Ade}^-$ frequencies), it seems that there is a steady increase from 0 mM to 4 mM, then plateaus at 6 mM, and eventually decreases slightly at 10 mM. This may be related to cell viability where cells become less in number when formaldehyde concentration increases. Since mutation frequency calculations were dependent on the number of viable cells, we deduced that the decreasing number of cells when exposed to 6 – 10 mM FA limited our ability to accurately measure mutation frequencies at increasing FA concentrations. While we expected mutation frequency to steadily increase as concentrations increase, we were restrained by the number of cells that have survived these higher concentrations. Thus,

the mutation frequencies reflected at 6 – 10 mM FA could quite possibly be the maximum mutational burden our yeast can tolerate without compromising their viability. Nonetheless, these results indicate that formaldehyde can induce multiple mutations at a rate that is approximately one or two orders of magnitude higher than no formaldehyde exposure.

As a whole, my findings indicate that our yeast model is capable of tolerating formaldehyde toxicity and mutagenicity of up to 6 mM, after which its viability starts to decrease. Viability remains relatively steady from 0 – 4 mM and only starts to decrease after 6 mM. Mutation frequency only seems to increase from 0 – 4 mM and starts to plateau after 6 mM. Moreover, room air exposure generates a higher mutation frequency than nitrogen exposure. Thus, the maximum mutational load our yeast can bear is when it is exposed to about 6 mM FA under aerobic conditions. Using this ideal condition, more than 2500 mutants were generated and screened from this experiment in which 208 samples were sent for whole-genome sequencing and computationally analyzed for the putative formaldehyde mutational signature (see Section 3.2.1).

While the goal of exogenous formaldehyde exposure is to optimize for mutation induction in our yeast model as well as gauge for immediate mutagenic effects, the concentrations used in these experiments are far from physiological levels. Thus, we also examined formaldehyde mutagenicity under a low, chronic exposure as it is biologically. To do this, endogenous formaldehyde treatments were performed through the deletion of essential formaldehyde dehydrogenases within our yeast model, which is discussed in the next section (Section 3.1.2).

Figure 3.1.2. Effects of Exogenous Formaldehyde Treatment to Can^R Frequency or Single-Gene Inactivation Frequency

Can^R frequency refers to the occurrence of *CAN1* gene inactivating events, implying that the cells have at least one mutation. The single-gene inactivation frequency at both oxygenation condition increases from 0 – 4 mM then begins to plateau at higher concentrations (6 – 10 mM). Other than 10 mM FA, the difference between room air and nitrogen exposure at every concentration is significantly different. * denotes significance at $p < 0.05$, ** denotes significance at $p < 0.01$, and *** denotes significance at $p < 0.001$ by Welch's t-test for $n = 8$ biological replicates for each treatment.

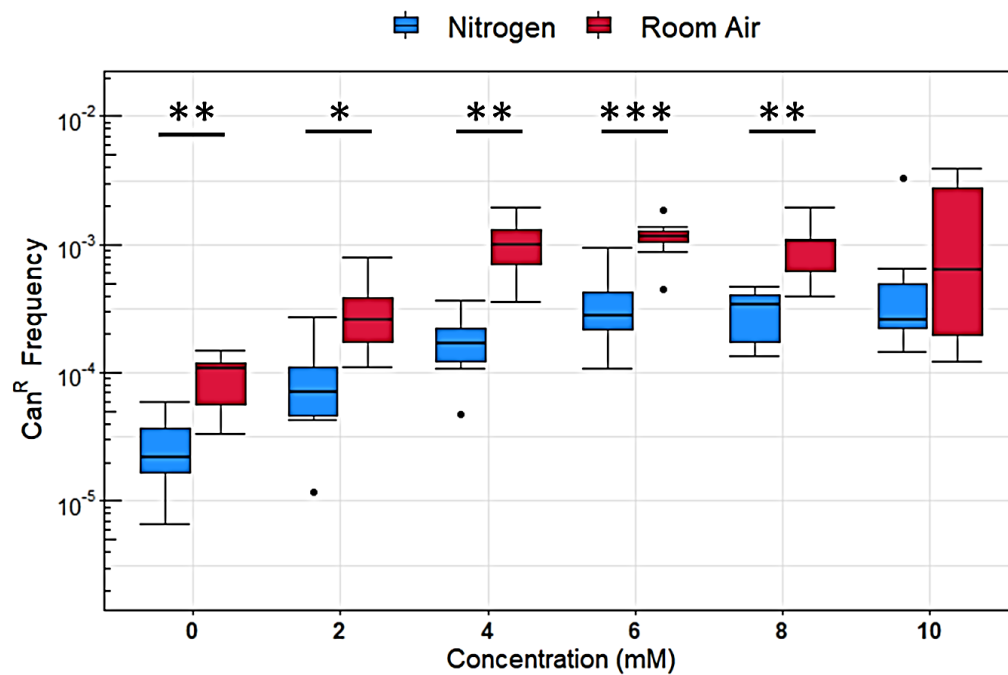
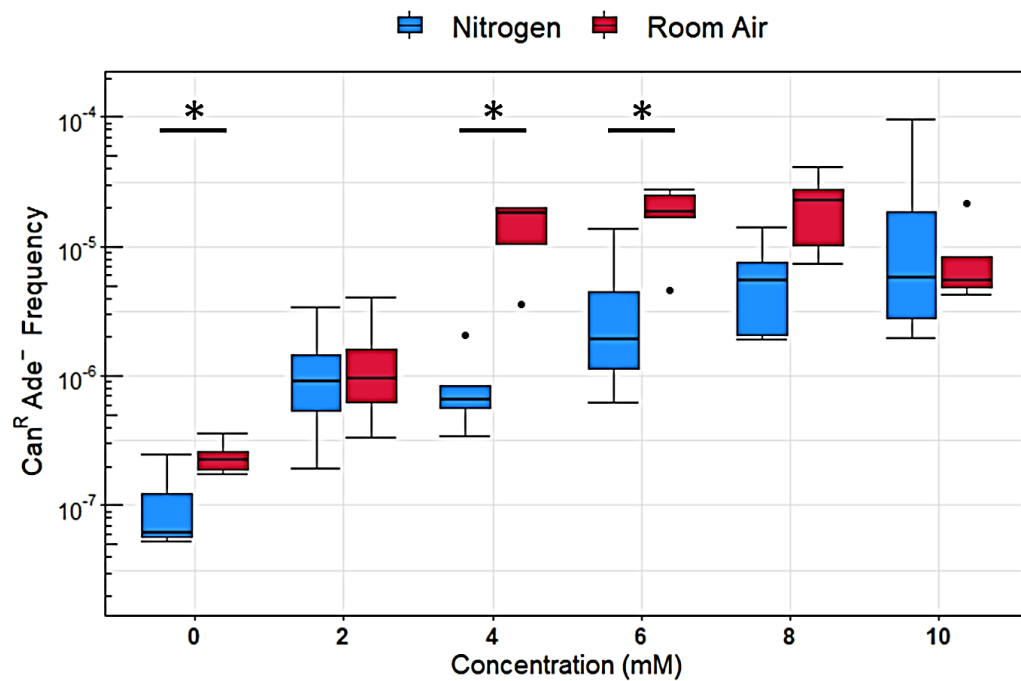


Figure 3.1.3. Effects of Exogenous Formaldehyde Treatment to $Can^R Ade^-$ Frequency or Double-Gene Inactivation Frequency

$Can^R Ade^-$ frequency refers to the occurrence of *CAN1* and *ADE2* gene inactivating events, implying that the cells have at least two mutations. Significant differences can only be observed at 0 mM, 4 mM and 6 mM FA. * denotes significance at $p < 0.05$ by Welch's t-test for $n = 5$ biological replicates for each treatment.



3.1.2 Endogenous Formaldehyde Treatment (Low, Chronic Exposure)

Since mice knockout studies[53–55] have demonstrated that formaldehyde within the blood is a potential carcinogenic risk, I also performed a knockout study using our yeast model to validate the mutagenicity of endogenous formaldehyde. The essential formaldehyde dehydrogenases in yeast are *SFA1* and *YJL068C*. Using the same measurement parameters from the exogenous formaldehyde treatment, I assessed mutation induction through cell viability and Can^R frequency (hereafter referred to as ‘mutation frequency’). The Can^RAde⁻ frequency could not be accurately measured in this assay due to limited numbers of red colonies, indicating that double gene inactivating events in this type of exposure is quite infrequent.

Single knockout strains *sfa1Δ* and *yjl068cΔ* were constructed using a standard yeast genetic modification technique (i.e. homologous recombination). Pre-growth cultures were plated on CAN media to assess basal mutation frequencies. Initial replicates (n = 3) of pre-growth cultures showed that *sfa1Δ* null mutants have unusually high background mutations compared to *yjl068cΔ* mutants. However, subsequent biological replicates (n = 10) of both null mutants tested in conjunction with wild-type (ySR127) revealed contrary results. It appears that the absence of Yjl068cp results in a mutation frequency that is significantly different from wild-type (p=0.0069 by Welch’s t-test), while the absence of Sfa1p does not affect mutation frequency (see Figure 3.1.4). Moreover, the deletion of *YJL068C* seems to result in a lower mutation frequency than that of wild-type, which is suggestive of Yjl068cp activity in contributing to baseline mutations when it is present within the cell. Although these results are statistically significant, the fold-change

difference between wild-type and *yjl068cΔ* is quite non-substantial. Furthermore, comparison of day-to-day mutation frequencies between null mutants and wild-type hardly varied from each other. Days one to three mutation frequencies were around 10^{-4} and 10^{-5} for all strains ($n = 2$ biological replicates; data not shown).

Consequently, I attempted to construct a *sfa1Δ yjl068cΔ* double knockout expecting a much drastic effect can be observed. However, this genotype produced very tiny colonies even after almost a week of incubation at 23°C (see Figure 3.1.5A), which suggested that deletion of both these formaldehyde dehydrogenases can result in a growth defect in cells. Therefore, instead of completely knocking out both genes, we decided to utilize the auxin-inducible degron (AID) system as an alternative approach to simulating a ‘double-knockout’ phenotype in our yeast. The AID system, or degron, is used to conditionally deplete a protein of interest under the presence of an auxin inducer[77]. Auxins are plant hormones that control gene expression during growth and development. This system requires a ubiquitin ligase complex called SCF-TIR1 (Skp1 – Cul1 – Fbox Transport Inhibitor Response I) complex in order to recognize the degron-fused protein and signal the rapid degradation of the protein (see Figure 3.1.5B). This system was applied in yeast and other animal cells by Nishimura et al. (2009) to deplete endogenous proteins of interest. They developed this system so that proteins are depleted only in the presence of an auxin over a wide range of temperatures, such as 37°C. Proteins can be recovered after transferring cells to a suitable medium without auxin, making this technique a reversible and optimizable system.

Using their protocol for generating my own degron-fused strains[78], *SFA1* was tagged with a degron (*sfa1-deg*) and *YJL068C* was deleted (*yjl068cΔ*), generating *sfa1-deg/yjl068cΔ* strains. The reason for exclusively tagging *SFA1* and not both proteins with the degron was mainly due to the results of the initial $n = 3$ biological replicates that suggested *sfa1Δ* generated higher background mutations, while *yjl068cΔ* background mutations were inconsequential. Quantifying mutation frequencies from a strain that generates high baseline mutations could confound our results and therefore, controlling the expression of *SFA1* was the best option; *SFA1* became the ideal candidate for the degron system given the data available. As mentioned previously, the *sfa1-deg/yjl068cΔ* strains also required the SCF-TIR1 complex, and thus, the *TIR1* gene of *Oryza sativa* (rice), *OsTIR1*, was also introduced into the cells using the protocol outlined in Papagiannakis et al. (2017).

Figure 3.1.4. Effects of Endogenous Formaldehyde Treatment to Can^R Frequency in Single Knockout Strains

These are mutation frequencies after three days of growth at 23°C (i.e. pre-growth). Initially, *sfa1Δ* exhibited a significantly higher mutation frequency than wild-type (n = 3 biological replicates). Instead, subsequent replicates revealed that mutation frequencies between *yjl068cΔ* and wild-type were significantly different, while *sfa1Δ* and wild-type frequencies were not. This result, however, is quite inconsequential because the fold-change difference is very little between *yjl068cΔ* and wild-type mutation frequencies.

** denotes significance at p<0.01 by Welch's t-test for n = 10 biological replicates.

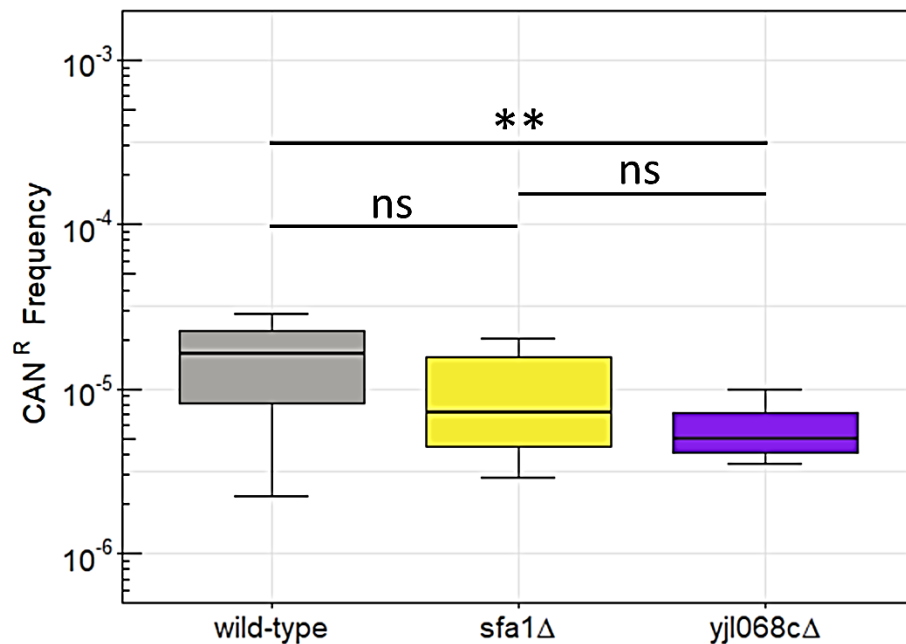
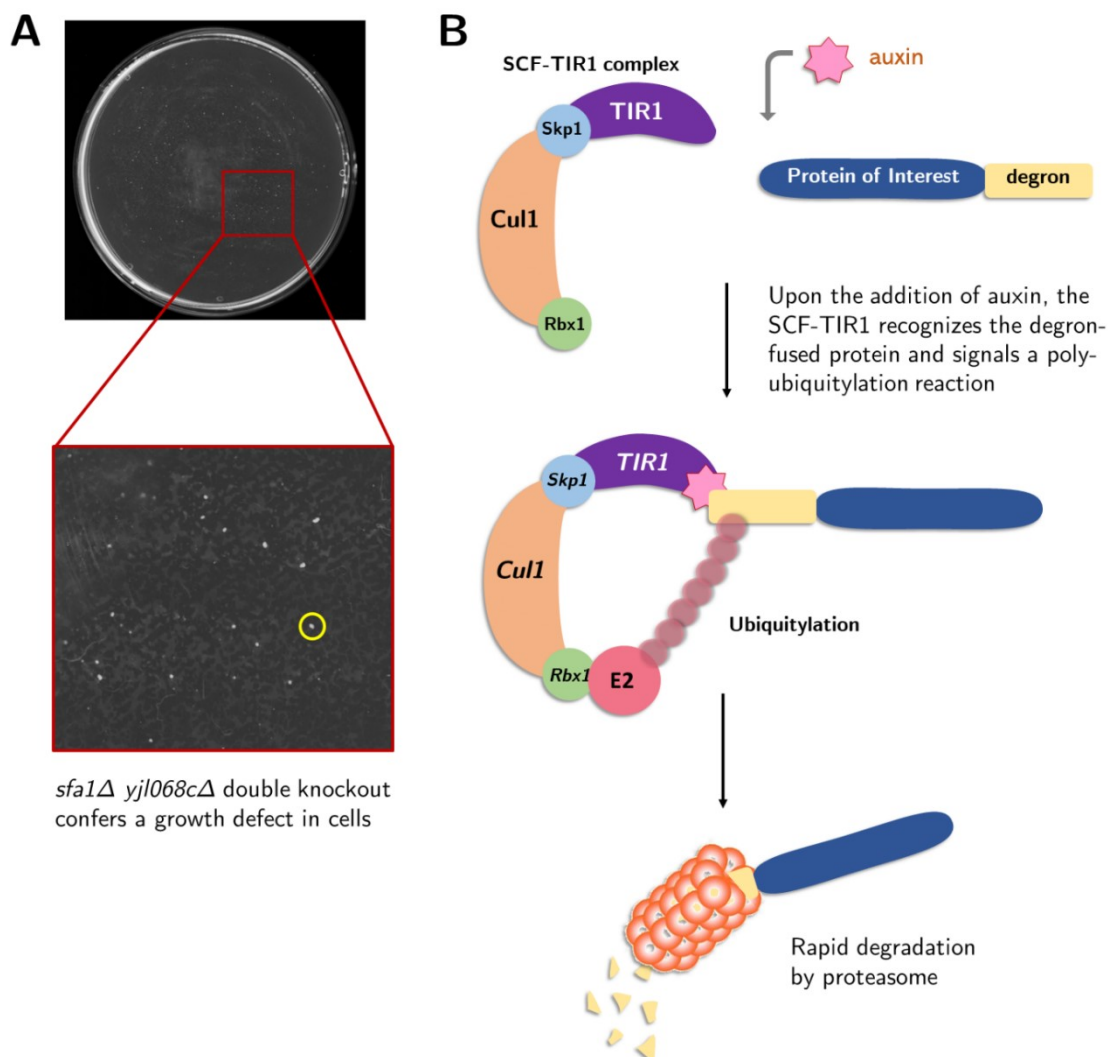


Figure 3.1.5. Results of the Double Knockout Phenotype Leading to the Use of an Alternative Approach

(A) Genetic modification of wild-type ySR127 strain to *sfa1Δ yjl068cΔ* double knockout through homologous recombination. The red box shows an enlargement of a section of the antibiotic selection plate. The yellow circle highlights one tiny colony of the double knockout strain. (B) Schematic diagram of the mechanism of the auxin-inducible degron (AID) system. The degron system was used as an alternative approach to produce ‘double-knockouts’ of *SFA1* and *YJL068C*.



The mutagenic effect of endogenous formaldehyde was examined in the constructed *sfa1-deg/yjl068cΔ* strains simply by growing them in rich-media for three days and measuring their mutation frequency day by day. The wild-type (ySR127) and OsTIR1 strains (i.e. yKC522, which is wild-type containing the SCF-TIR1 complex without the degron system) were also tested alongside the *sfa1-deg/yjl068cΔ* cells as controls. Three different versions of degron tags were tested in this assay – AID, mini-AID (mAID) and 3x mini-AID (3mAID) – and all three tags displayed similar mutation frequencies. Therefore, the result summarized in Figures 3.1.6 and 3.1.7 for *sfa1-deg/yjl068cΔ* strains is the mean data from all three degron tags.

The background mutations (i.e. mutation frequency at day 0) of *sfa1-deg/yjl068cΔ* were three times higher than wild-type (p=0.028 by Tukey’s test), while the background mutations between the controls are not significantly different (see Figure 3.1.6). However, upon temperature shift to 37°C for three days, there is no clear difference in mutation frequencies observed between each strain with or without the presence of auxin (see Figure 3.1.7). It is important to note that day 0 samples have not been treated with auxin as these samples were grown for three days under 23°C, which means that Sfa1p remained present within the cells and was not depleted prior to measuring mutation frequency on day 0. Sfa1p should theoretically have been depleted after 30 minutes following the addition of auxin before the temperature shift to 37°C on day 1[77]. Overall, these results do validate the fact that yeast is quite tolerant to endogenous formaldehyde since the threshold for formaldehyde toxicity in *S. cerevisiae* seems to lie between 1 mM and 5 mM[94]. Although there seems to be no remarkable effect in mutation frequency after the

endogenous treatment of formaldehyde, all of the different genetically modified backgrounds (i.e. single knockout and degon-knockout strains) do express sensitivity to formaldehyde after a 24-hour exogenous exposure to 2 mM FA at 23°C (see Figure 3.1.8). Samples treated with exogenous formaldehyde were much slower in growth as their doubling time seem to be delayed for about six hours. The controls (wild-type and OsTIR1) also showed sensitivity to formaldehyde, but the difference in sensitivity between the controls and yeast constructs remains to be investigated. Furthermore, more biological replicates of this assay are required to accurately quantify the difference in cell densities of FA-treated and untreated samples between controls and genetically modified strains since only one replicate has been examined so far.

Overall, the examination of endogenous formaldehyde treatment through gene manipulation studies needs further investigation in order to solidify the chemical's mutagenic effects within the cell. And while these genomes are also of great interest in signature analysis, unfortunately, endogenous exposure could not produce as many mutations as we would expect under exogenous or high, acute exposure. This was evidenced by the limited number of red colonies produced in this low, chronic treatment, which is our indication of multiple mutation induction. Hence, as prefaced in the previous section, only the exogenously treated yeast genomes were further processed and sequenced for mutational signature analysis of formaldehyde.

Figure 3.1.6. Mutation Frequency of Pre-growth Cultures (day 0 samples) of *sfa1-deg/yjl068cΔ* Strains

Pre-growth cultures were grown for three days at 23°C. ‘Wild-type’ refers to the ySR127 strain; ‘OsTIR1’ refers to the strain containing only the *TIR1* complex; and ‘*sfa1-deg/yjl068cΔ*’ refers to the degron-tagged *SFA1* and *yjl068cΔ* knockout strain. The background mutations of *sfa1-deg/yjl068cΔ* is significantly higher than wild-type.

* denotes significance at $p < 0.05$ by Tukey’s test for $n = 5$ biological replicates.

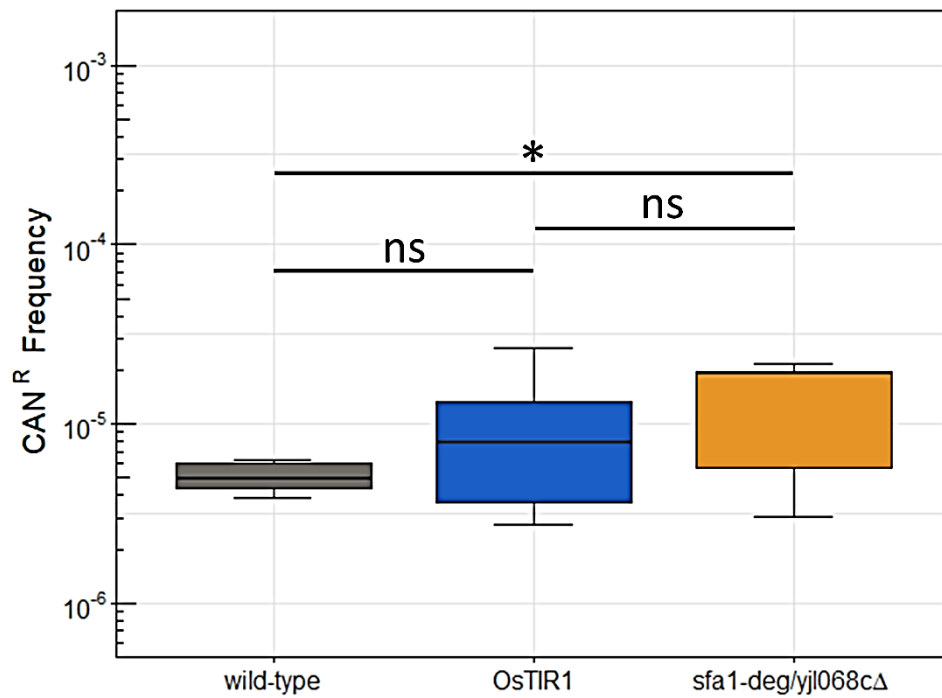


Figure 3.1.7. Effects of Endogenous Formaldehyde on $sfa1\text{-deg}/yjl068c\Delta$ Strains after Auxin Treatment and Temperature Shift

‘With auxin’ samples were treated with auxin (indole-3-acetic acid, IAA), while ‘no auxin’ samples were treated with ethanol (because the auxin was diluted in ethanol). Three different degron tags were tested – AID, mAID and 3mAID – which all showed similar Can^R frequencies. Therefore, the $sfa1\text{-deg}/yjl068c\Delta$ data refers to the combined data of all three different degron-knockout backgrounds. No significant differences were observed between ‘with auxin’ and ‘no auxin’ treated samples and between strains by Tukey’s test for $n = 5$ biological replicates.

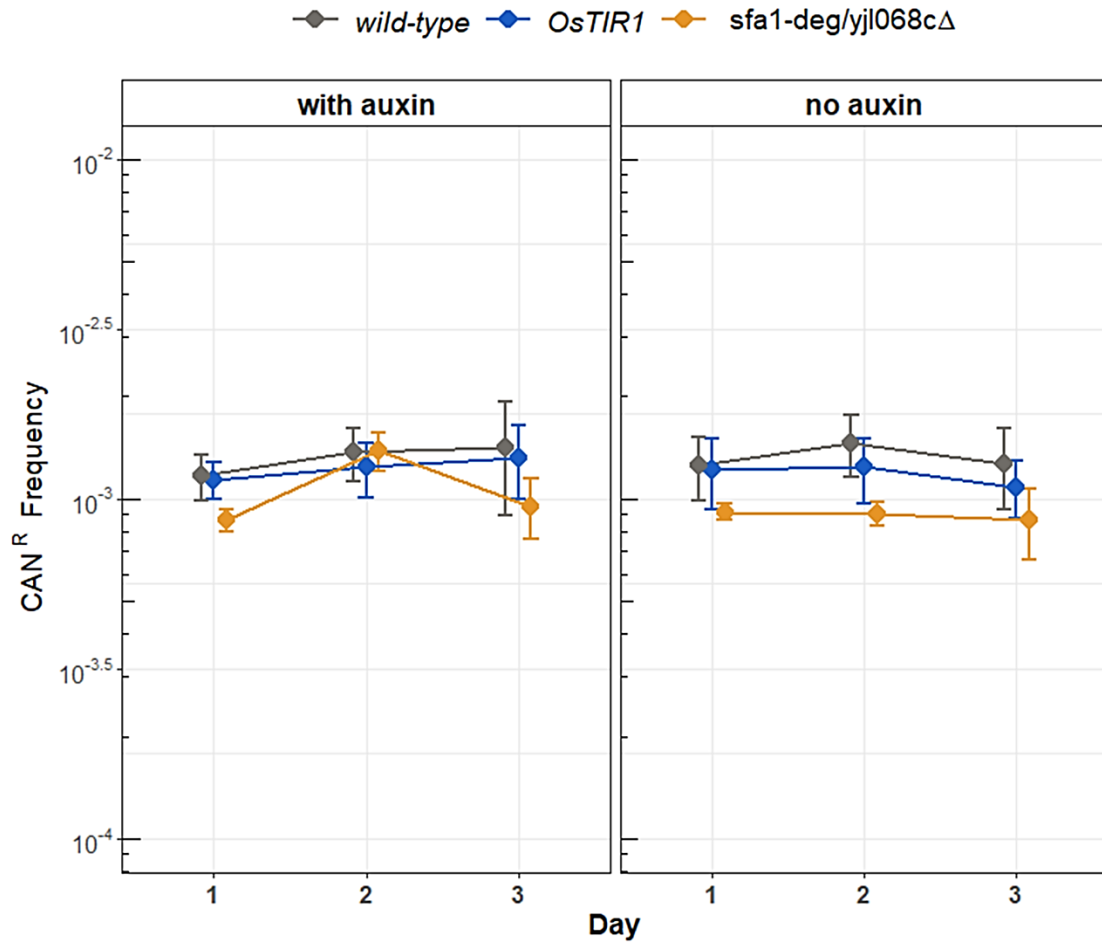
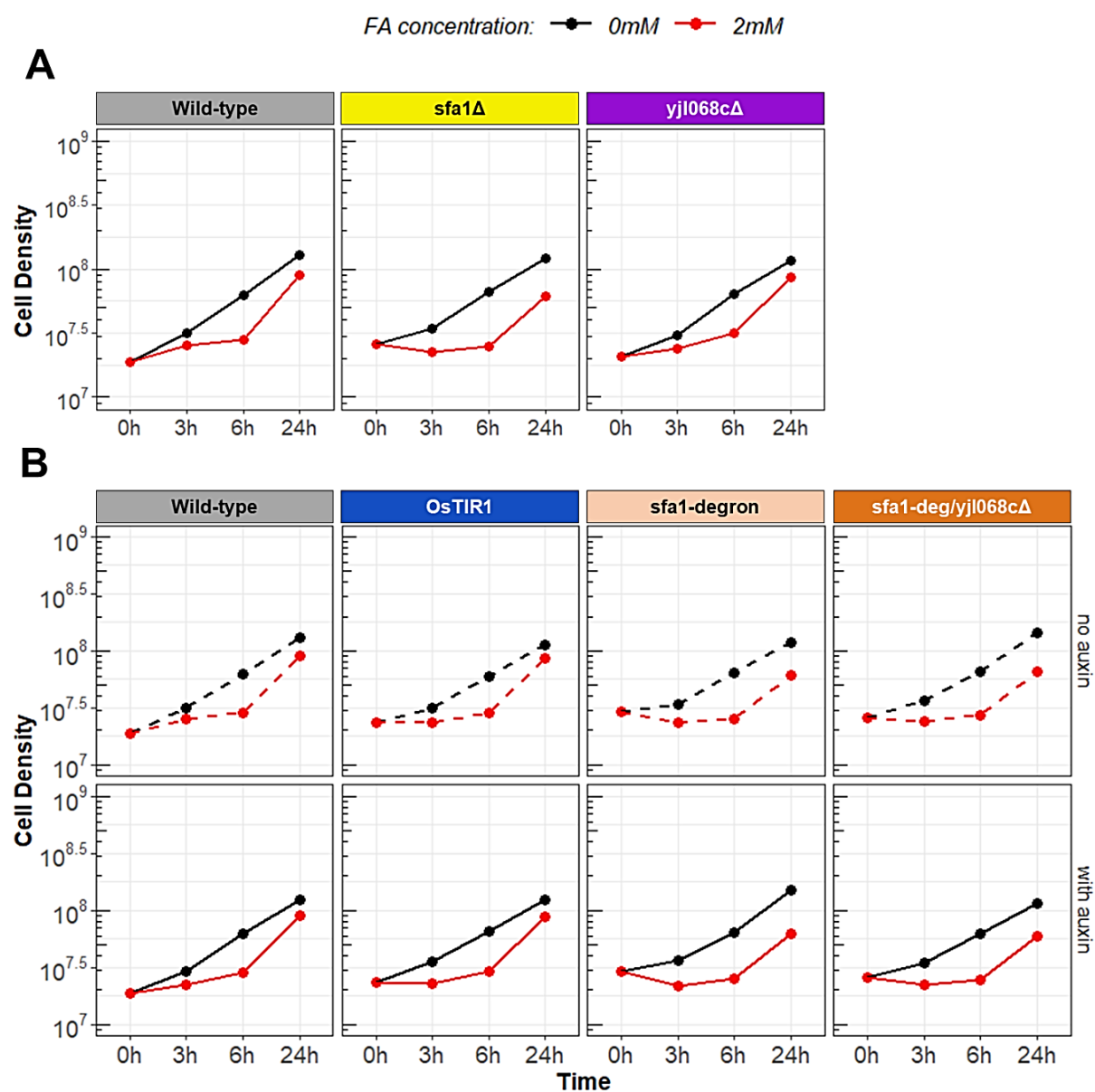


Figure 3.1.8. Formaldehyde Sensitivity of All Genetically Modified Strains

(A) Exogenous formaldehyde exposure for single knockout strains. (B) Exogenous formaldehyde exposure for degron strains treated with and without auxin. Treatment was performed under 23°C. All genetically modified strains including wild-type showed sensitivity to formaldehyde. This assay requires more replicates to effectively compare the difference in sensitivity between wild-type and constructs ($n = 1$ biological replicate).



3.2 Computational Characterization of the Mutational Signature of Formaldehyde Using Yeast Genomics Data

3.2.1 Single Base Substitution (SBS) Signature

Yeast genomes from the high, acute formaldehyde exposure experiments were extracted and sequenced to begin this analysis. Sequenced genomes were then mapped to a reference genome in order to identify specific mutations or variants found within the treated yeast genomes (see Section 2.8 for details). The variants were then filtered for quality score, depth score, indels and redundancy (also elaborated in Section 2.8). This stringent filtering measure ensures that variant calls processed by the non-negative matrix factorization (NMF) algorithm can be uniquely attributed to formaldehyde mutagenesis only. In total, I have analyzed three different data sets of formaldehyde-treated genomes with 208 genomes containing 1681 unique mutations, along with two sets of non-formaldehyde-treated genomes[93] with 250 genomes containing 1620 unique mutation calls.

Using the NMF function from the MutationalPatterns package[86], the single base substitution signature analysis began with an NMF rank survey (explained in Section 2.8), in which I ran 50 iterations of the algorithm (i.e. Brunet’s method of NMF[88]) for factorization rank numbers 2 to 5, $r = 2$ to $r = 5$, simulating two- to five-signature solutions within the formaldehyde and non-formaldehyde data sets (see Figure 3.2.1A). The cophenetic coefficient started decreasing at $r = 2$ (coefficient = 0.94) and starts to

increase at $r = 3$, which means that extracting two signatures from my data set will give the most meaningful results.

A single base substitution signature is characterized over six main types of base substitutions – $C \rightarrow A$, $C \rightarrow G$, $C \rightarrow T$, $T \rightarrow A$, $T \rightarrow C$, and $T \rightarrow G$ – with each subdivided into 16 unique trinucleotide motifs thereby, classifying 96 possible mutation types. After performing 100 iterations of the NMF algorithm using $r = 2$, I have deciphered signatures for formaldehyde and no formaldehyde treatment, which were remarkably different from each other based on their cosine similarity value (see Figure 3.2.1B). Since identical signatures produce a cosine similarity (cossim) value of 1, Signature A and Signature B produced a cossim value of 0.731, which signifies two very different patterns of mutations.

The defining difference between the two signatures could be owed to the high frequency of $C \rightarrow A$ transversions in TCn motifs observed in Signature B (where the ‘middle C’ is the mutated base, and ‘n’ denotes any nucleotide base).

In addition, relative attributions of each signature were analyzed within the different datasets – EM18 and EM19 for untreated samples (i.e. no formaldehyde treatment), and FA-J19, FA-M18 and FA-N18 for treated samples (i.e. formaldehyde treatment). Signature A was mostly found in untreated samples while Signature B contributed the most in mutations found in formaldehyde-treated samples (see Figure 3.2.1C), and thus, Signature B may possibly be the unique formaldehyde signature. As validation, I plotted the mutational profiles of the ‘untreated’ data set and ‘FA-treated’ data set and calculated their cosine similarities against Signature A and Signature B, respectively. I have chosen a cossim value of 0.90 as the minimum requirement to denote

close similarity between two signatures. Signature A and the ‘untreated’ mutational profile gave a cossim value of 0.996, while Signature B and the ‘FA-treated’ mutational profile gave a cossim value of 0.99. This confirms that Signature B is the putative formaldehyde single base substitution (SBS) signature, and Signature A can be aptly named the ‘no formaldehyde’ SBS signature.

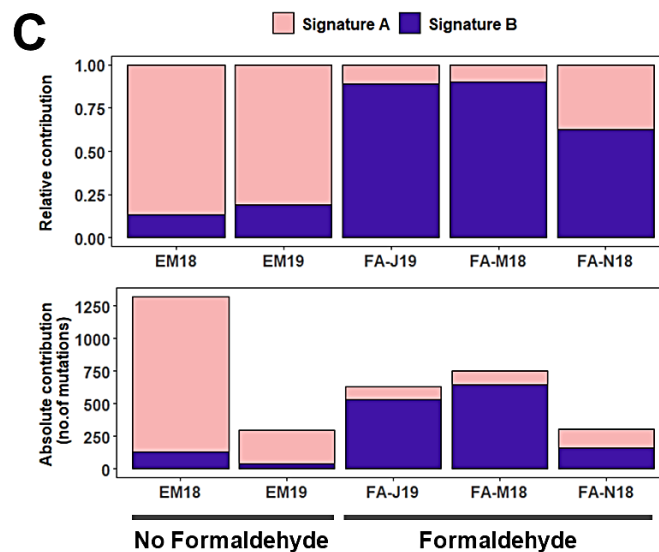
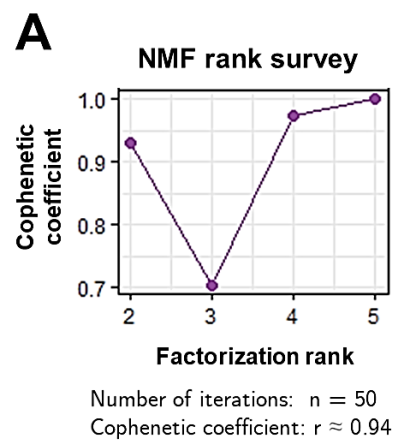
Because the seed number was manually initiated by the user, this may have affected the way NMF performed. Thus, to verify the robustness of the signature solutions deciphered by the NMF algorithm, I decided to generate 10 different solutions, each of which started with a different seed number, randomized by a random number generator, and was calculated for 10 iterations. The mean and standard deviation of the 10 different solutions were then used to generate Signature A and Signature B. This strategy resulted in solutions that have virtually little to no variation between solutions (see Figure 3.2.2), demonstrating that regardless of the number of iterations or seed number, the putative formaldehyde mutational signature and no formaldehyde signature are reproducible solutions by NMF. Furthermore, the cosine similarity values of these multiple-iterated signature solutions to the mutational profiles of FA-treated and untreated datasets were 0.9897 for Signature B vs. FA-treated and 0.9957 for Signature A vs. untreated, which further validates the previous observation for the one-solution signature profiles of Signature A and Signature B.

Figure 3.2.1. Signature Extraction by Non-Negative Matrix Factorization (NMF) Algorithm

(A) The NMF rank survey was run for 50 iterations to solve for the optimal number of mutational signatures that exist within the data set. Based on the cophenetic coefficient measure of quality, factorization rank $r = 2$ will give the most robust solution for NMF.

(B) Using $r = 2$, two signatures were extracted from the given data set (two no formaldehyde and three formaldehyde data sets). The main difference between the two signatures lie on the frequency of TCn→TAn transversions found exclusively in the formaldehyde data sets. The cosine similarity value of these two signatures is 0.731 denoting remarkably different patterns of mutations.

(C) The relative contribution (top) and absolute contribution (bottom) of each signature is plotted for each data set. Signature A is mostly found in untreated samples (data sets EM18 and EM19), while Signature B is mostly found in formaldehyde-treated samples (data sets FA-J19, FA-M18, and FA-N18).



B Signatures extracted by non-negative matrix factorization (NMF) algorithm using **rank = 2**

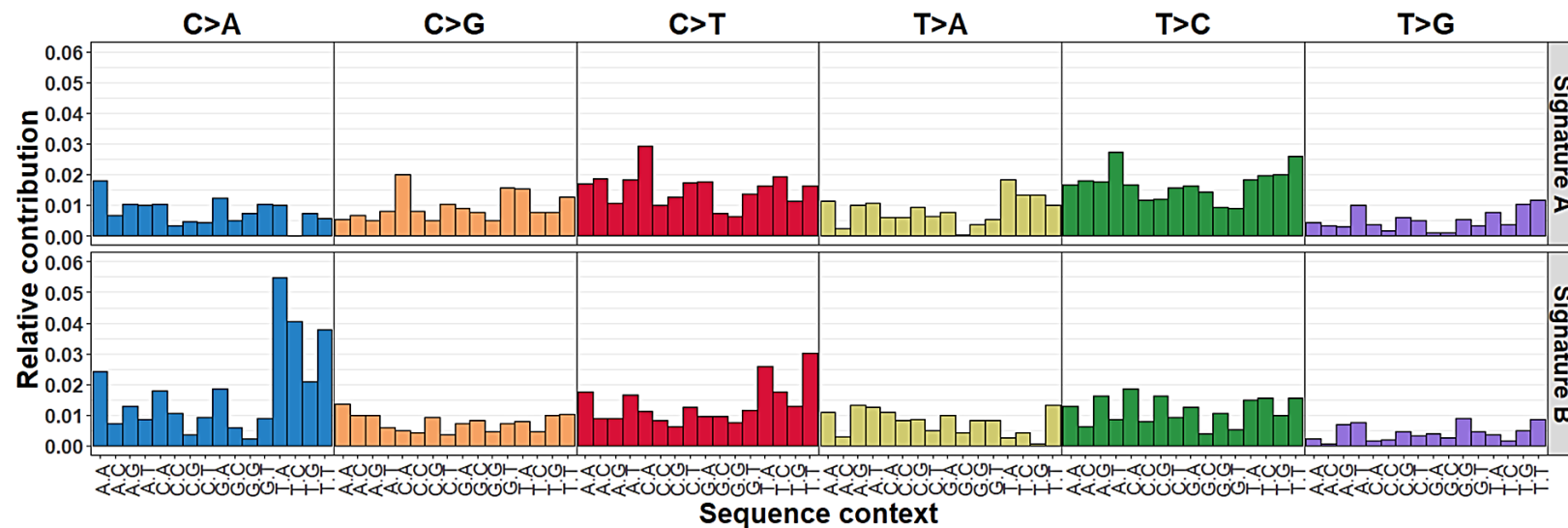
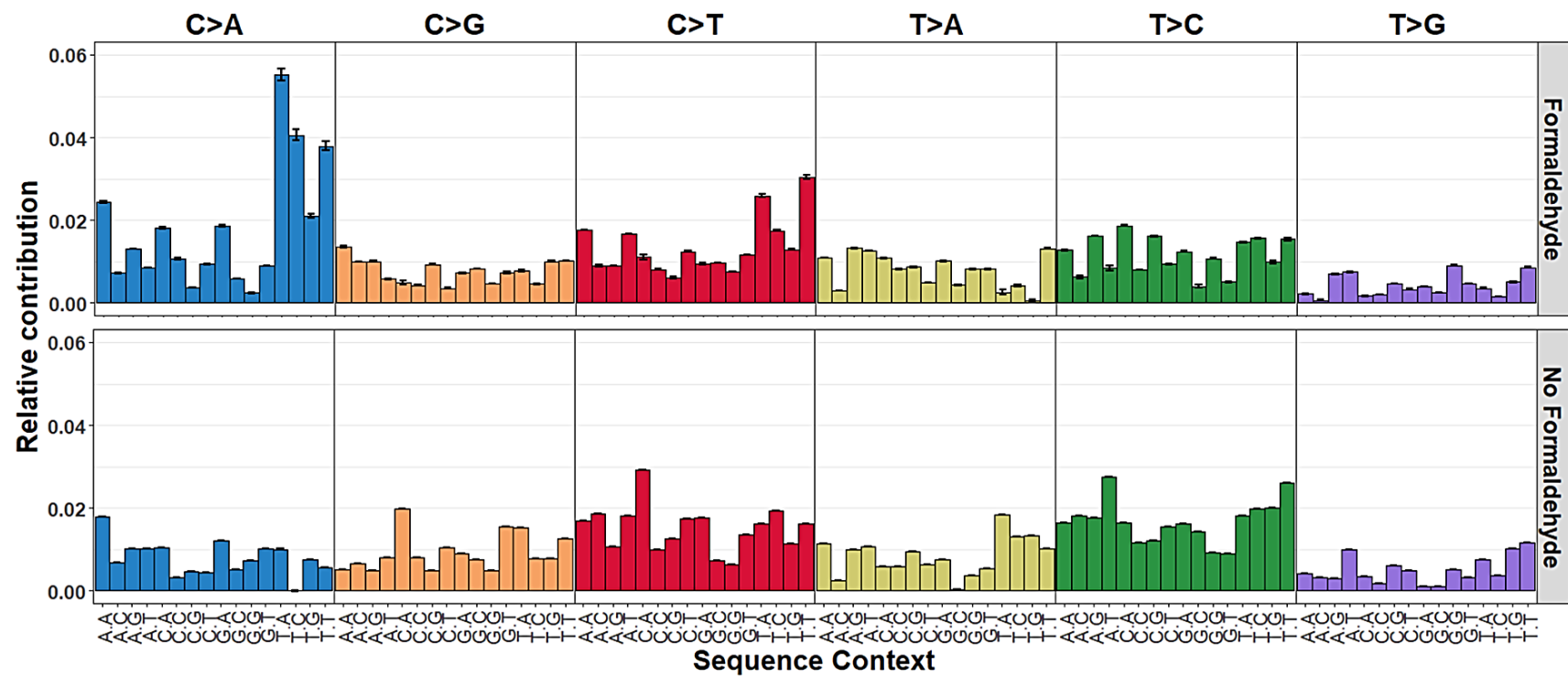


Figure 3.2.2. The Putative Single Base Substitution (SBS) Formaldehyde and No Formaldehyde Signatures Generated from 10 Different Solutions of the Non-Negative Matrix Factorization (NMF) Algorithm

The FA SBS signature was generated using 208 genomes containing 1681 unique mutations, while the No FA SBS signature was derived from 250 yeast genomes with 1620 unique mutation calls. The height of each bar in each signature represents the ‘mean’ value of the 10 different NMF solutions for that mutation type, and the error bars represent the ‘standard deviation’. The algorithm initially produced Signature A (bottom) and Signature B (top) as solutions. Cosine similarity calculations against the formaldehyde and no formaldehyde dataset confirms that Signature B is the putative formaldehyde signature (cossim = 0.9897), while Signature A is the putative no formaldehyde signature (cossim = 0.9957). The cosine similarity of these two signatures to each other is 0.737, which denotes vastly different signatures signifying that the two treatments generate their own unique pattern of mutations.



3.2.2 Insertion/Deletion (ID) Signature

Although they exist in relatively small numbers, indels were also important indicators for formaldehyde’s mutagenesis. In total, there are 473 unique indel variants within 208 formaldehyde-treated genomes and 714 unique indels within 250 untreated genomes. These were also filtered by quality score, depth score, and redundancy to ensure uniqueness of each call (see Section 2.8). The major challenge in this analysis, however, is mainly due to the varying characteristics of these indels. Indels can be short (of 1 to 5 bases) or long (of 6 or more bases) in length; they can occur in repetitive elements (such as homopolymers or short tandem repeats), or non-repetitive elements (such as microhomologies); and the length of repeat elements may also influence their occurrence. While Alexandrov and colleagues (2020) have identified and classified 17 unique small insertion and deletion (ID) signatures from cancer genomes using an indel classification of 83 mutation types, “there is no single intuitive and naturally constrained set of ID mutation types as there arguably are for single base substitutions”[31]. Therefore, I have devised my own indel classification for simplicity as it reflects the available data set on hand.

To begin this analysis, I compared the indels from each treatment based on their sizes (see Figure 3.2.3). The frequency of each deletion length seems to be the same between treatments, unlike insertions where there is an observable excess of 1-base insertions from the formaldehyde-treated samples compared to the untreated samples. Indeed, the cosine similarity of the deletions is 0.965 denoting quite similar distributions

between treatments, while insertions has a cossim value of 0.834. One area of dissimilarity could be the difference in 1-base insertions.

Further analysis of the 1-base insertions revealed that formaldehyde treatment caused an increase in A insertions by about 10% (see Figure 3.2.4). The frequency of other 1-base insertions (C, G and T), on the other hand, seems to have decreased by at least 5% in the presence of the mutagen. C deletions also seem to be favoured by formaldehyde treatment over all types of 1-base deletions. Generally, when comparing the overall distribution of indels between the two sets of data, the untreated samples have relatively equal frequencies of single nucleotide indels, while formaldehyde exposure causes fluctuating frequencies of 1-base insertions and deletions with clear favor over A-base insertions and C-base deletions.

While the goal is to create an ID signature that resembles the COSMIC classification, I have decided to take a similar but simpler approach in classifying these formaldehyde indels due to their complex nature. First, indels were grouped based on their length – one base (1bp), two bases (2bp), three to four bases (3-4bp), five to nine bases (5-9bp), and 10 or more bases (>10bp). Close examination of the immediate 5' and 3' sequences (i.e. 100 bases upstream and downstream of the indel) revealed that most variants commonly occurred at repetitive elements, and typically, the indel is a repeat unit within these homopolymer or short tandem repeat elements. Thus, the indel classification reported here defines an indel mutation type by its size and the number of repeat units within its location. A repeat unit size of '1' or '0' designates a non-repeating unit for deletions and insertions, respectively. Figure 3.2.5 shows the two putative indel

signatures for formaldehyde and no formaldehyde treatments generated using this indel classification of 50 possible mutation types.

The FA ID signature shows that deletions of >5 bases in length occur frequently even when they are not repeated within the sequence. Although 5-9bp deletions at non-repeating units also occur without the mutagen (i.e. no formaldehyde), this indel type occurs at a higher frequency with the mutagen. Interestingly, 3-4bp insertions in 5 or more repeat units and other long base insertions (>5bp at any number of repeat units) occur frequently under no formaldehyde exposure, while this occurs less with formaldehyde treatment. Overall, the putative formaldehyde and no formaldehyde indel signatures produced a cossim value of 0.792 denoting remarkably different patterns of mutations from each other. Formaldehyde exposure, therefore, generates a unique pattern of mutation of single-base substitutions and indels quite different from no formaldehyde exposure.

Figure 3.2.3. Indel Types By Length for No Formaldehyde and Formaldehyde Treatments

Deletion types between No Formaldehyde (246 deletions) and Formaldehyde (265 deletions) do not differ much, however, insertion types between the two treatment (203 FA insertions and 469 No FA insertions) have a clear difference at 1-base insertions. Formaldehyde-treated samples (upper left panel) bears a clear excess of 1-base insertions compared to untreated samples (lower left panel), which implies a preference for 1-base insertions under formaldehyde mutagenesis. ‘Relative contribution’ refers to the fraction of a particular insertion or deletion type found within all the insertions or deletions in the dataset.

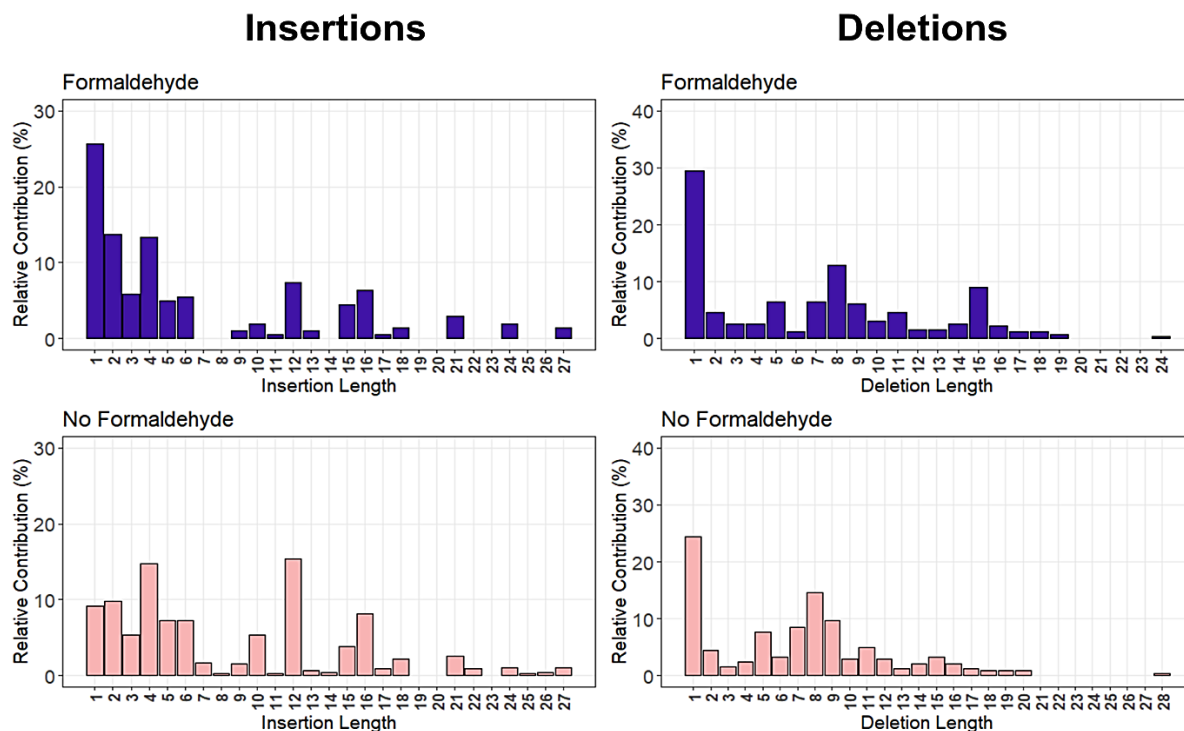


Figure 3.2.4. Comparison of 1-Base Insertions Between No Formaldehyde and Formaldehyde Treatments

Further analysis of the 1-base insertions showed a clear difference in pattern between the two treatments, where formaldehyde exposure favors A base insertions more frequently than other base insertions and favors C deletions over other base deletions. The preference for 1-base insertion and deletion events under no formaldehyde treatment remains relatively constant across all types of bases.

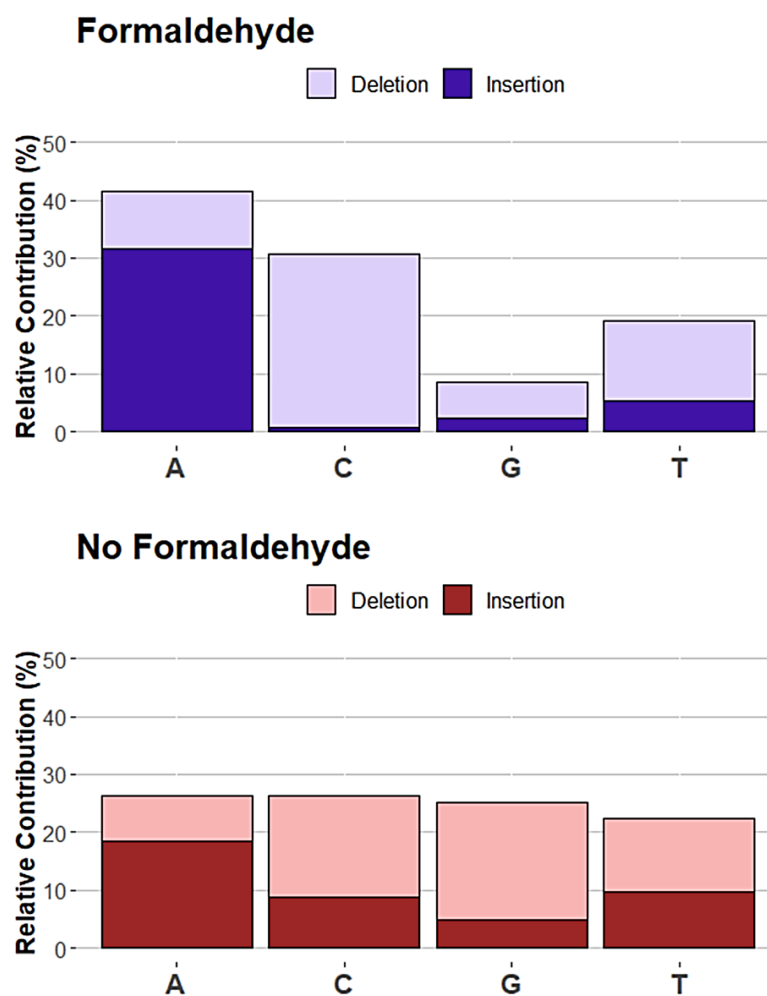
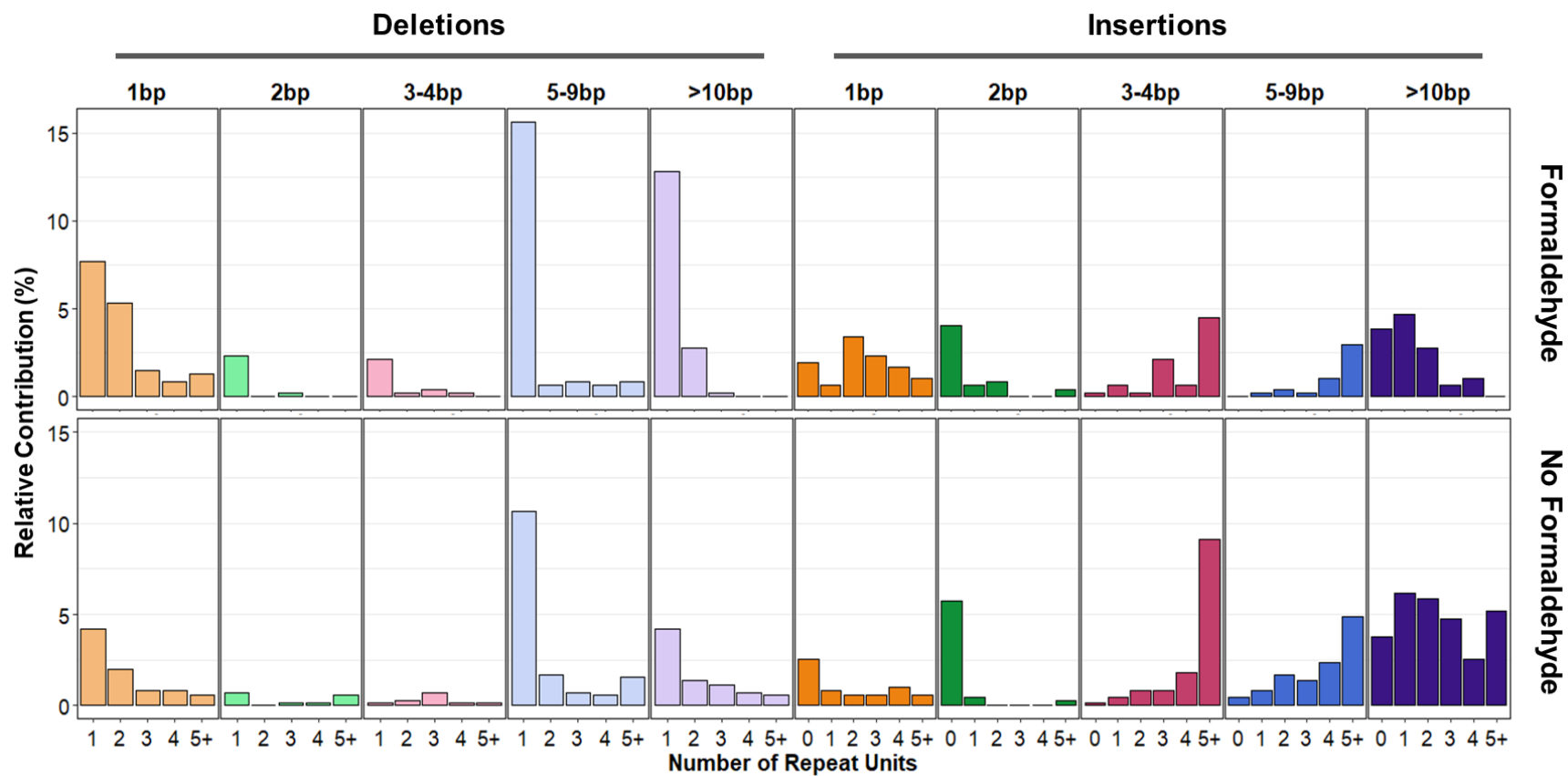


Figure 3.2.5. Putative Indel Signatures of Formaldehyde and No Formaldehyde Exposures

Each mutational profile was derived from 468 indel calls for Formaldehyde and 714 indel calls for No Formaldehyde. These signatures were generated simply by calculating the relative contribution of each mutation type. Deletions and insertions are classified into five main indel types, which is defined by their length, and further subdivided based on the nature of the sequence where they are located – repeating or non-repeating units. A repeat unit of 1 and 0 are considered “non-repeating units” for deletions and insertions respectively.



3.3 Significance of the Formaldehyde Single Base Substitution Signature by Examining Cancer Data

3.3.1 Formaldehyde SBS Signature vs. COSMIC SBS Signatures

To begin assessing the relevance of formaldehyde in cancers, cosine similarity values were calculated between the deciphered formaldehyde SBS signature against all 67 COSMIC SBS signatures to assess similarity (i.e. the 49 Single Base Substitution (SBS) signatures and 18 suspected sequencing artefacts). This strategy would not only provide an initial corroboration of an unknown cancer signature (or signatures) to a well-known mutagen but will also verify the significance of formaldehyde mutagenesis in cancers. Figure 3.3.1 summarizes the cosine similarities of all COSMIC SBS signatures versus the putative formaldehyde and no formaldehyde SBS signatures.

To reiterate, I determined closely similar signatures by using a minimum cossim value of 0.90. The formaldehyde signature showed the closest resemblance to SBS40 with a cossim value of 0.843, however, this cosine similarity is relatively low to confidently make this association. This result was not surprising as a quick visual examination of the formaldehyde signature compared to all COSMIC SBS signatures showed no obvious matching general patterns. As noted earlier, the most defining feature of the formaldehyde SBS signature was the abundance of C→A transversions in TCn motifs, which is a pattern observed in some COSMIC SBS signatures but not SBS40 (see Figure 3.3.2B for SBS40). This led us to suspect the possibility that the formaldehyde SBS signature may be a composite of two or more signatures.

To confirm this, I performed a reconstruction analysis where combinations of COSMIC SBS signatures were assessed to determine which composition best explains the patterns observed in the formaldehyde SBS signature. Reconstruction was conducted by combining two or three COSMIC signatures together then calculating the cosine similarity value of this new composite signature versus the formaldehyde signature. The highest cosine similarity value of a pairwise reconstruction was 0.899, which was achieved through the combination of COSMIC SBS40 and SBS52. This cossim value is just below my set threshold for similarity hence, I decided to examine a triple-wise combination as well. The triple-wise combination of COSMIC SBS25, SBS40 and SBS52 generated the highest cosine similarity value of 0.916. Although the additional signature made the cossim value increase quite incremental, we considered the triple-wise combination to be a sufficient explanation for the complex profile of the putative formaldehyde SBS signature as we were trying to determine the lowest possible number of combinations that would generate the most reasonable cossim value. This implies that formaldehyde exposure possibly generates mutations that conform to patterns observed in COSMIC SBS25, SBS40 and SBS52 altogether.

For context on these COSMIC SBS signatures: SBS25 has only been identified in Hodgkin's cell lines that were treated with chemotherapy, and thus, chemotherapy treatment is suspected as its possible aetiology[26]; SBS40 is considered a 'clock-like' signature where its mutational burden increases with age[31]; and SBS52 is labelled as a possible sequencing artefact. Although the combination of SBS25, SBS40 and SBS52 is the most optimal mathematical solution to explain the superimposed signatures found

within the putative formaldehyde SBS signature, we believe that the triple-wise combination of SBS5, SBS40 and SBS52 provides the most reasonable explanation biologically (cossim value = 0.907; see Figure 3.3.2A). COSMIC SBS5 currently remains as an unknown signature, however, a recent study from our group has found that SBS5 may be the mutational pattern of intrinsic biochemical processes within humans[93]. COSMIC SBS5-like patterns were also observed in other organisms such as the seven model species (including budding yeast) and a subset of mammals, plants and insect species, which further led us to believe that these processes are well-conserved among and between species. Initially, the combination of SBS5, SBS40 and SBS52 generated the highest cosine similarity value (of 0.918) amongst all possible triple-wise combination of COSMIC signatures, in which filtering for depth score was not included in the variant filtering criteria. The slightly lower cossim value in this re-analysis is quite possibly the consequence of the additional filtering layer. Nevertheless, the triple-wise combination of SBS5, SBS40 and SBS52 still generated a high cossim value (of 0.907) despite the most conservative filtering approach used here.

While examining the attributions of SBS5, SBS40 and SBS52 among all mutations observed in the formaldehyde and no formaldehyde data sets, SBS52 mutations were most exclusively found within the formaldehyde data set (see Figure 3.3.2B, right panel). Considering that the pattern of SBS52 (mostly C→A transversions in TCn motifs) is the defining characteristic that differentiates the formaldehyde signature from the no formaldehyde signature, we presume that SBS52 could be the formaldehyde-specific signature (see Figure 3.3.2B, left panel).

Figure 3.3.1. Cosine Similarity

Values of the Formaldehyde and No

Formaldehyde SBS Signatures

versus 67 COSMIC SBS Signatures

The formaldehyde SBS Signature is most

similar to SBS40 with a cossim value of

0.843, and the no formaldehyde SBS

signature is most similar to SBS5 with a

cossim value of 0.876.

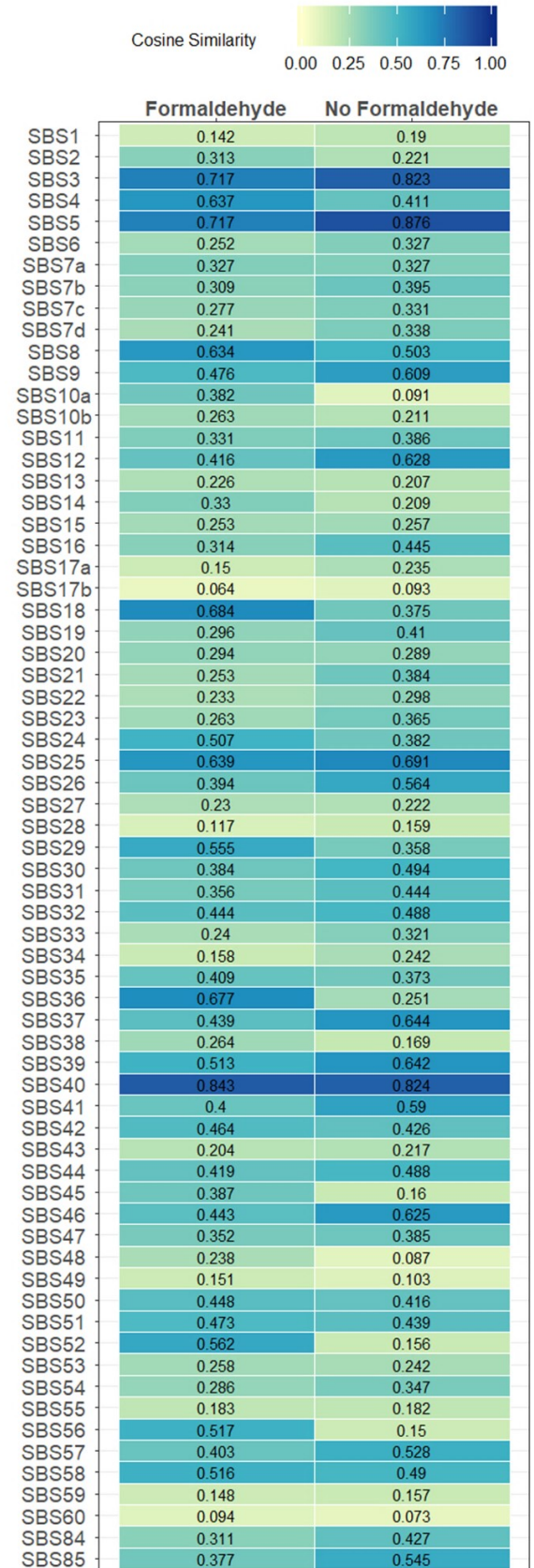


Figure 3.3.2A. The Combination of COSMIC Signatures SBS5, SBS40 and SBS52 to Reconstruct the Putative Formaldehyde SBS Signature

This triple-wise reconstruction generated a cosine similarity value of 0.907 (residual sums of squares, $RSS = 3.26e-03$, between the formaldehyde and reconstructed signatures), which best explains the composite nature of the putative formaldehyde SBS signature biologically. This also implies that SBS5, SBS40 and SBS52 could be the mutational patterns that form under formaldehyde exposure in aerobic metabolism.

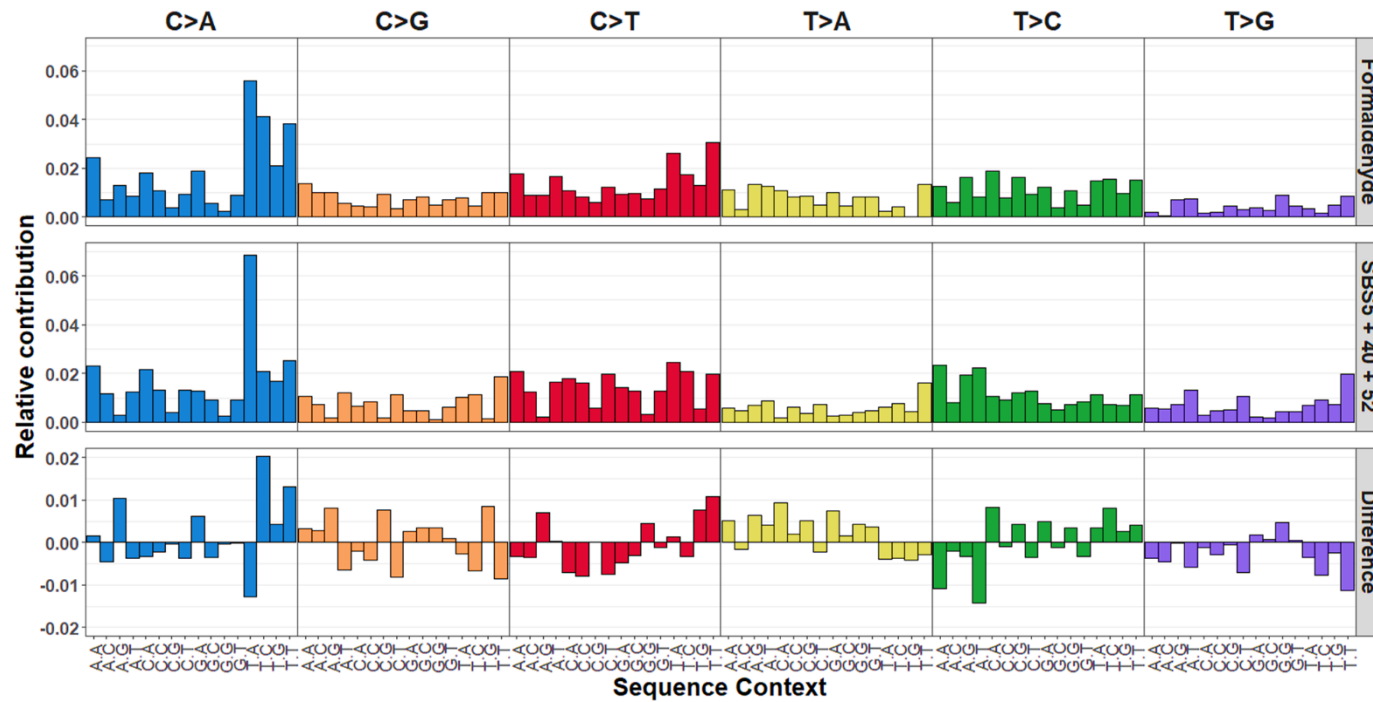
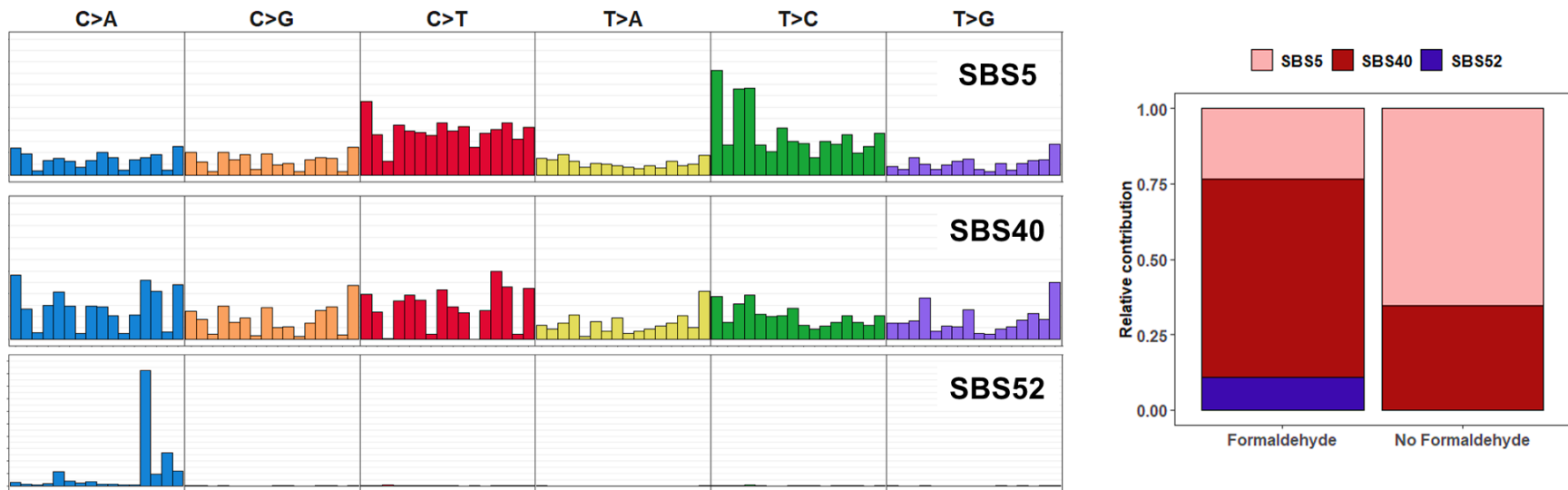


Figure 3.3.2B. Mutational Profiles of COSMIC Signatures SBS5, SBS40 and SBS52

COSMIC SBS5, SBS40 and SBS52 (left panel) were signatures deciphered from somatic mutations in cancer genomes. Both SBS5 and SBS40 have unknown aetiologies. Although SBS52 is labelled as a possible sequencing artefact, we suspect that this could be the formaldehyde-specific signature because of its characteristic pattern of C→A transversions in TC_n motifs, which is a critical feature that differentiates formaldehyde from no formaldehyde exposure. The relative contribution of each COSMIC signature (right panel) also shows that SBS52 can only be found in formaldehyde-treated yeast genomes.



For this reason, I further analyzed samples with SBS52 attributions for indel patterns to verify if they would resemble the indel pattern observed in my formaldehyde-treated genomes. There are 271 samples containing SBS52 across all available cancer genome databases: Pan-cancer Analysis of Whole Genomes (PCAWG), The Cancer Genome Atlas (TCGA), and other non-PCAWG whole genome and whole exome sequencing data sets. There were a couple limiting factors for this indel analysis, however. First, raw data was not available for the PCAWG and TCGA data sets, while raw data for non-PCAWG whole genome and whole exome sequencing data were available[26]. Second, not all reported SBS52-containing samples (referred to as “SBS52+” hereafter) have indel variant calls. In total, only 81 out of 271 available samples were analyzed for indels. These remaining 81 SBS52+ samples collectively contain 1482 insertions and 8054 deletions. Included in this analysis was a subset of randomly selected samples lacking SBS52 (referred to as “SBS52–”), which I had curated to collectively contain the same amount of insertions and deletions as those of the SBS52+ samples (i.e. 1482 insertions and 8054 deletions) for even comparison. I had expected SBS52+ samples to also display an excess of 1-base insertions as that observed in the formaldehyde-treated yeast genomes. Instead, SBS52+ samples showed an excess of 2-base insertions and deletions, while SBS52– samples contained a high frequency of 1-base insertions and deletions (see Figure 3.3.3).

I extended this analysis by plotting the frequencies of every possible 2-base and 1-base indel mutation type for both groups to assess if there is a particular mutation type that is favored in each group. Out of the 16 possible combinations of 2-base indels, SBS52+

seem to show slightly higher frequencies of CA/GT deletions and overall, a large proportion of its indels were characterized by deletions over insertions (see Figure 3.3.4, right panels). The calculated cosine similarity between the 2-base indel distribution of SBS52+ and SBS52- however, is 0.922, which signifies that 2-base indel mutation frequencies between the two groups are not different from each other. The 1-base indel mutation patterns also appeared dissimilar between the two groups, where SBS52- samples displayed an excess of A/T indels, while SBS52+ showed almost equal frequencies across all 1-base indel types. Contrarily, their cosine similarity value of 0.938 implies a high similarity in distribution.

Despite these results, it remains true that 2-base indels dominate among SBS52+ samples over other indel lengths with over 6000 2-base indel calls in the data set compared to about 1 – 900 calls of other indel variants. This high frequency of 2-base indels seem reminiscent of the microsatellite instability phenotype found in several cancer tumours[95]. This hypermutator phenotype is predominantly observed in tumours with defective DNA mismatch repair, which leads to frameshift mutations due to DNA slippage on these short repetitive sequences. This observation is also reflected on the formaldehyde-treated yeast genomes as several of the indel calls (≥ 2 -base indels) seem to occur over varying lengths of repetitive sequences of at least two or more bases.

Figure 3.3.3. Indel Analysis of Samples containing COSMIC Signature SBS52 and those lacking SBS52

SBS52-containing (SBS52+) samples have a large proportion of 2-base insertions and deletions among all its mutations, while SBS52-lacking (SBS52-) samples have an excess of 1-base indels. The SBS52- data were a collection of 10 random subsets of samples, and thus, plotted by their mean (as the height of each bar) and standard deviation (as error bars).

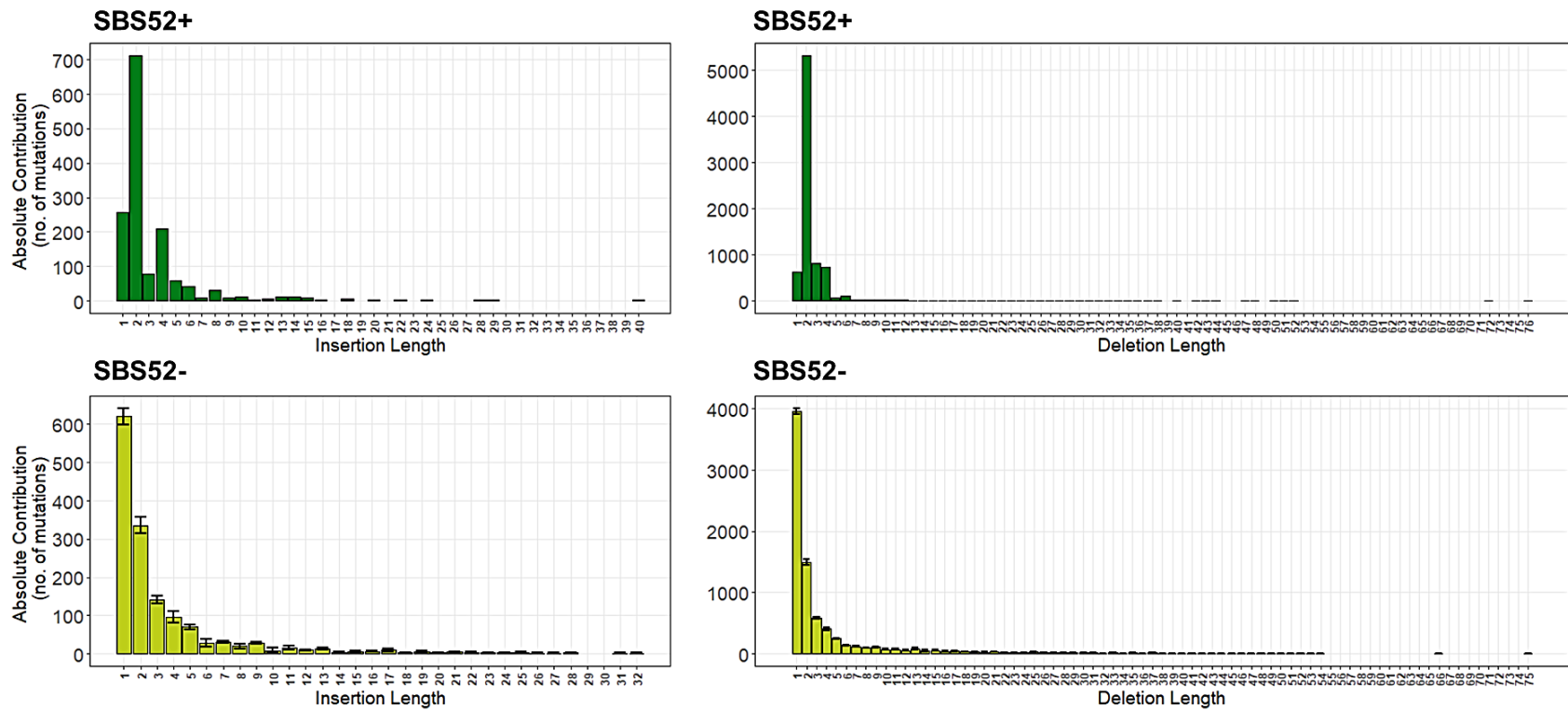
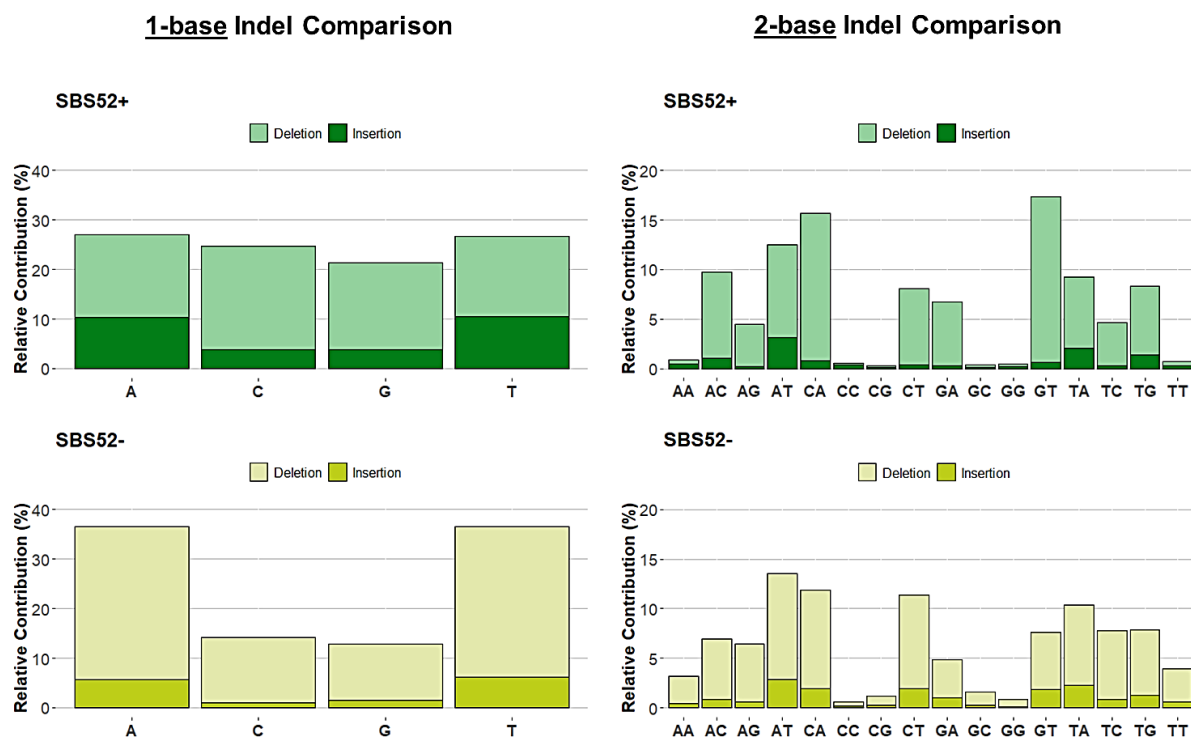


Figure 3.3.4. Distribution of 1-base and 2-base Indel Mutation Types among SBS52+ and SBS52- samples

While SBS52+ samples display almost equal frequencies across all types of 1-base indel mutations (top, left panel), SBS52- samples display an excess of A/T indels (bottom, left panel). However, the cosine similarity of their 1-base indel distribution is 0.938, signifying similar indel distributions. With 2-base indel mutations, although SBS52+ samples seem to display a high frequency of CA/GT indels than SBS52-, the overall distribution between the two groups are quite similar to each other as well (cossim value = 0.922).



3.3.2 Deciphering Signatures from Head and Neck Cancer Genomes

Another approach in determining the relevance of formaldehyde mutagenesis in cancers is to analyze formaldehyde-associated cancer genomes, such as nasopharyngeal and sinonasal cancers, which has been grouped within the head and neck cancer genomic data set[92]. Since COSMIC summarizes all mutational signatures found in all cancer types, a *de novo* extraction specifically of head and neck cancer samples will highlight signatures that are found exclusively under this cancer type. Consequently, this strategy will provide a better understanding of mutational processes operative in these cancers, as well as elucidate formaldehyde attribution within these samples.

Collectively, there are 798 head and neck squamous cell carcinoma (abbreviated as “Head-SCC”) samples from all four databases – PCAWG, TCGA, non-PCAWG whole genome and whole exome sequencing data. Similar to the process of deciphering signatures from the formaldehyde-treated yeast genomes, the first step is to conduct an NMF rank survey to assess the optimal number of signatures that can be extracted from the data set. After 50 simulated iterations of the NMF algorithm, a cophenetic coefficient of 0.990 can be achieved with a factorization rank number, $r = 3$. This means that the most robust solution for NMF to uniquely identify different signatures from the Head-SCC data set is if three signatures were extracted (see Figure 3.3.5A). Thus, using $r = 3$, I have characterized Signatures A, B and C from the Head-SCC data set (see Figure 3.3.5B) and calculated their cosine similarities to the 67 COSMIC SBS signatures for classification (see Figure 3.3.6).

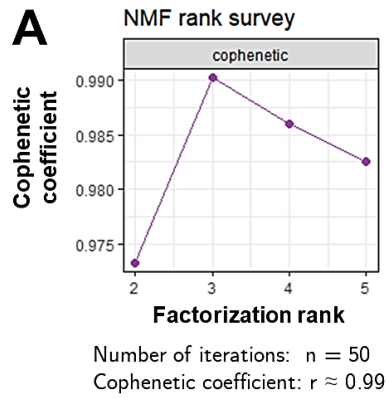
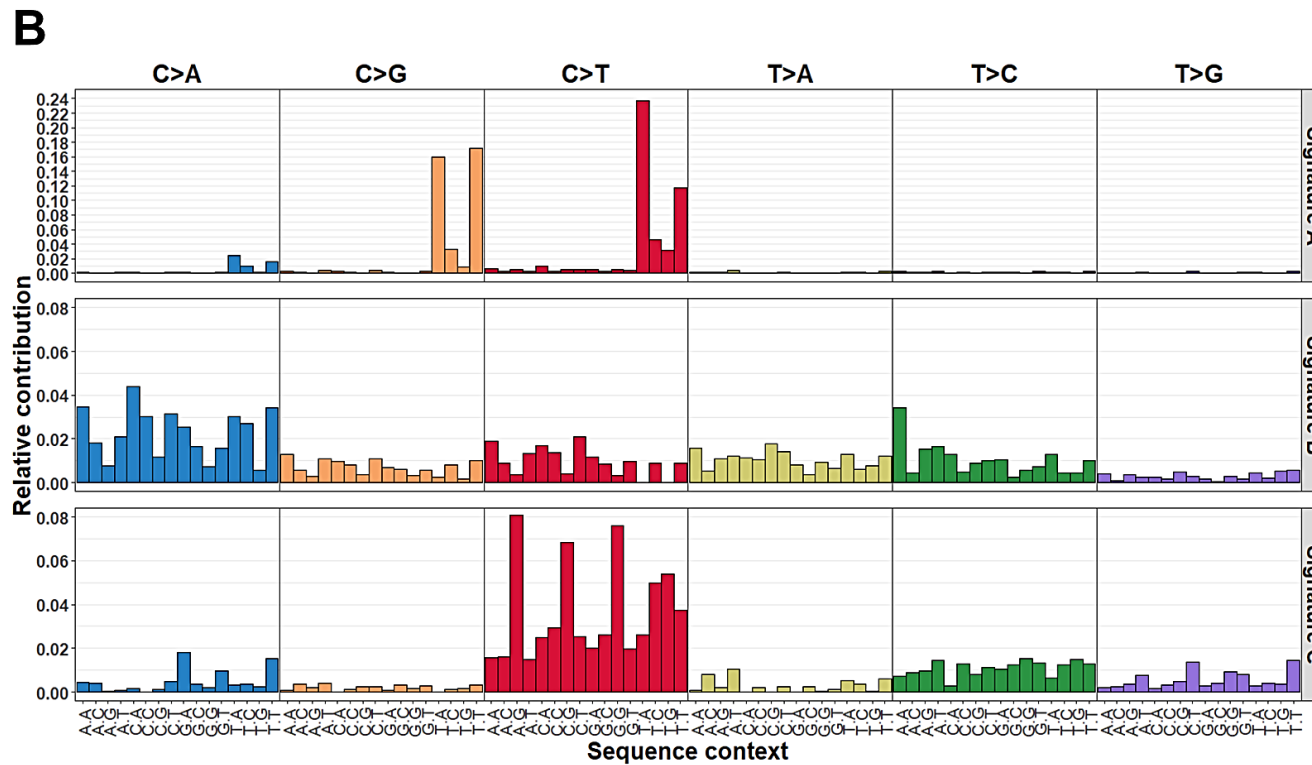


Figure 3.3.5. Extraction of Signatures from Head and Neck Squamous Cell Carcinoma (Head-SCC) Genomes

A) The NMF rank survey result suggests factorization rank number, $r = 3$ will provide the most robust solution for signature extraction. B) Based on the rank survey result, these are the three signatures that were uniquely extracted from the Head-SCC data set.



None of the Head-SCC extracted signatures have similarities higher than 0.90 to any of the COSMIC SBS signatures. Based on their highest cossim values, Head-SCC Signature A bears the most similarity to SBS2 with $\text{cossim} = 0.748$; Signature B is most similar to SBS4 with $\text{cossim} = 0.864$; and Signature C is most similar to SBS6 with $\text{cossim} = 0.802$. Interestingly, the pattern of Signature A seems to be a composite of SBS2 and SBS13, and reconstruction analysis confirms this to be the case (see Figure 3.3.7, $\text{cossim value} = 0.997$). Both SBS2 and SBS13 are attributed to the activity of the APOBEC family of cytidine deaminases (APOBEC is the abbreviation for “Apolipoprotein B mRNA editing Enzyme, Catalytic polypeptide-like”). SBS2 is characterized by C→T transversions in TCn motifs, while SBS13 is characterized by C→G transitions and some C→A transversions in TCn motifs. SBS6 is characterized by mostly C→T transversions found in all types of trinucleotide motifs and is associated with a defective DNA mismatch repair system. SBS4 is mostly C→A transversions in all types of trinucleotide motifs with low frequencies of other base substitution types. It is formed likely due to DNA damage caused by tobacco smoke mutagens.

When the putative formaldehyde signature was compared to all Head-SCC signatures, it generated the highest cossim value of 0.811 with Head-SCC Signature B. However, the suspected formaldehyde-specific pattern of C→A transversions in TCn motifs is most conspicuous in the Head-SCC Signature A though at a very low frequency. If indeed Head-SCC Signature B is COSMIC SBS4, then it is possible that formaldehyde-specific mutations may also be superimposed in Head-SCC Signature B since tobacco

smoke mutagens also include formaldehyde[96]. However, these are difficult assumptions to establish given the poor cossim values, which may be conducive of the limited number of samples and mutations found in all Head-SCC data available. Nevertheless, some reassurance remains as formaldehyde mutational patterns can be seemingly observed in head and neck cancer genomes even with such poor resolution.

Figure 3.3.6. Cosine Similarities (cossim) of Head and Neck Cancer (Head-SCC) Signatures vs. 67 COSMIC Single-Base Substitution (SBS) Signatures

and the Putative

Formaldehyde Signature

Head-SCC Signature A is most similar to SBS2 with cossim = 0.748; Head-SCC Signature B is most similar to SBS4 with cossim = 0.864; and Head-SCC Signature C is most similar to SBS6 with cossim = 0.802. The putative formaldehyde signature is “most similar” to Signature B with cossim = 0.811.

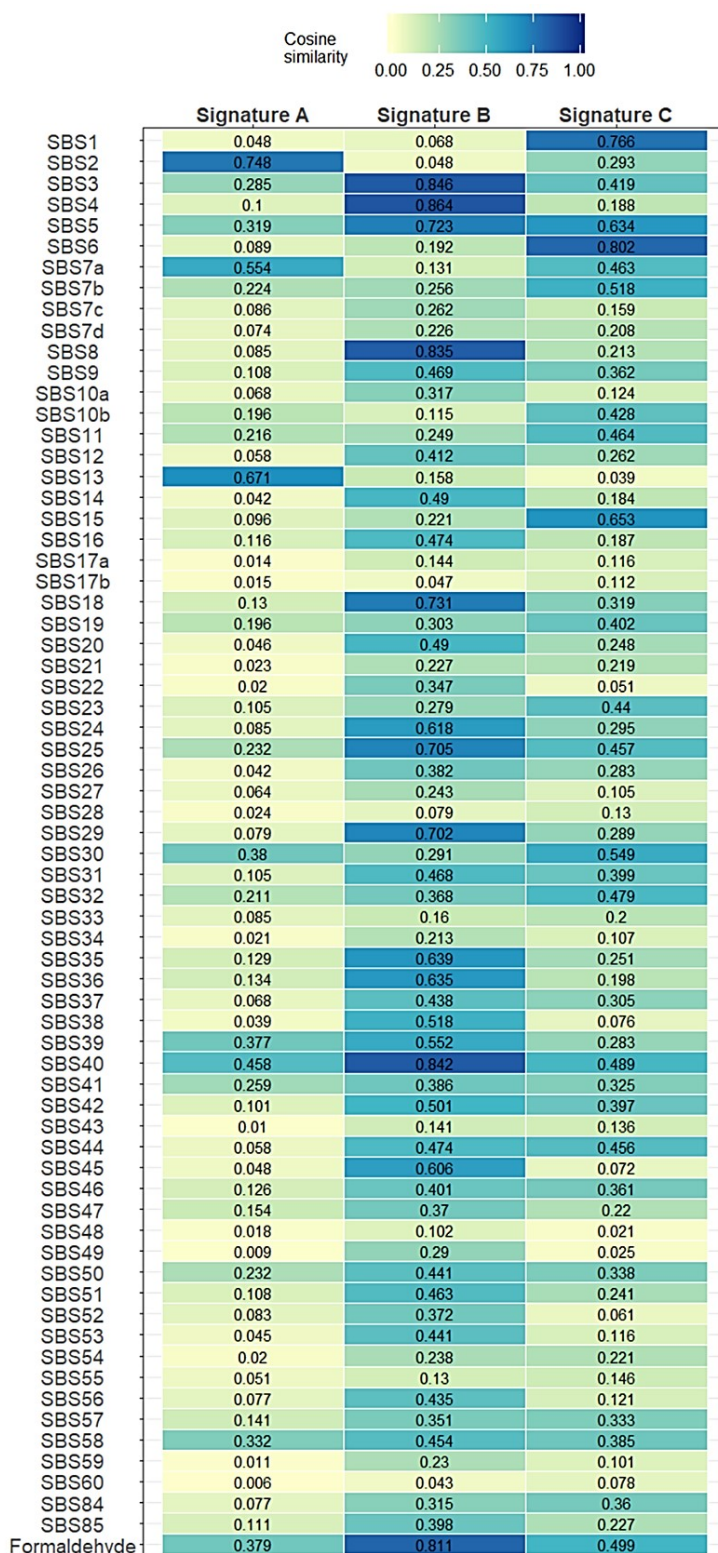
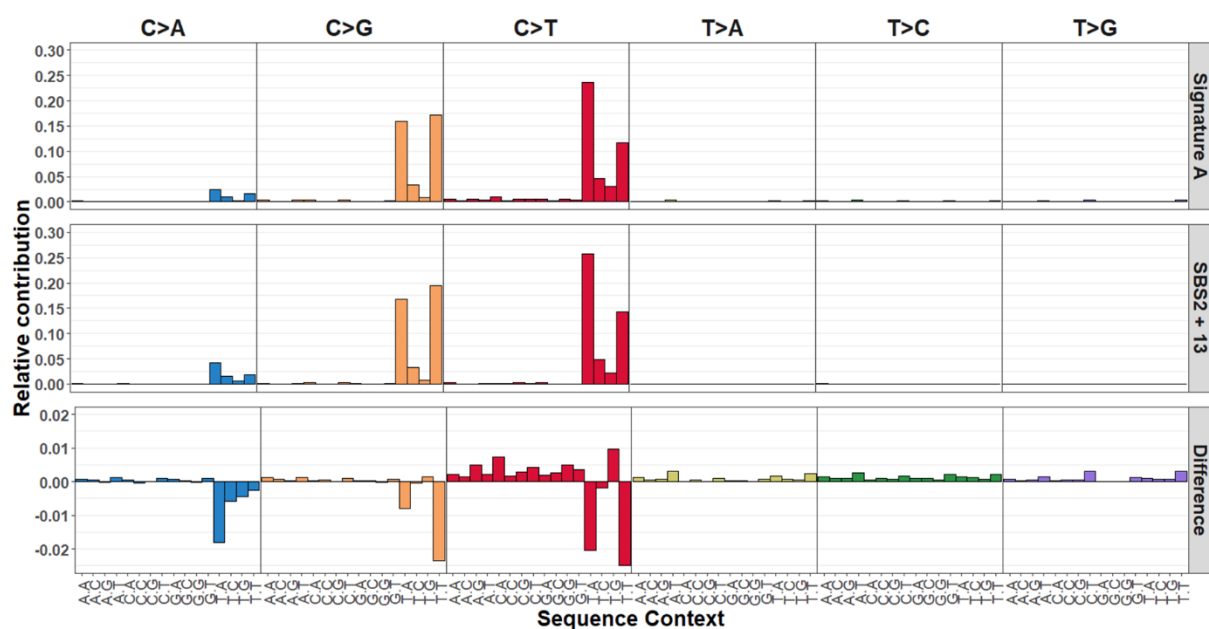


Figure 3.3.7. Reconstruction of Head and Neck Cancer (Head-SCC)

Signature A

Head-SCC Signature A seems to be a composite of COSMIC SBS2 and SBS13, which are both signatures of the APOBEC family of cytidine deaminases. The cosine similarity of this reconstruction is 0.997, which confirms this hypothesis.



4. Discussion

4.1 Mutagenic Effects of Exogenous Formaldehyde (High, Acute Exposure)

Significant differences in mutagenesis can be observed between aerobic and anaerobic respiration even without formaldehyde exposure (0 mM FA). Naturally, when cells were exposed to formaldehyde in either oxygenation state, the treatment drastically decreased cell viability and further increased mutation induction, especially in higher concentrations. It is quite interesting that under 0 – 4 mM FA, viability remained stable regardless of the oxygenation state, but mutation frequencies steadily increased from one concentration to the next. This could be an indication that our highly sensitive yeast cells are capable of tolerating a mutation rate of at least 9.83×10^{-4} mutations per nucleotide base without compromising their viability. However, mutations well-beyond this frequency resulted in cells either unable to bear the additional mutational burden or possibly develop positive epistasis contributing to increased fitness thereby increasing cell viability[97]. Another possibility could be that exposure was variable between cells, and the number of mutations accrued depended on their proliferative capacity after mutagenesis[17, 97]. This may explain the observed mutagenic effects under 6 – 10 mM FA where a few cells were still able to survive despite the large amount of formaldehyde-induced mutations in their genomes. Altogether, it is clear that our yeast cells reach their maximum mutational capacity without significantly affecting viability at 6 mM FA, where a mutation frequency of about 1.155×10^{-3} mutations per base is generated.

The results in this assay also support the toxic and mutagenic nature of formaldehyde as previous studies have shown[60, 75, 98, 99]. Moreover, the severity of its exposure depended on the oxygenation state. Under aerobic respiration, by-products of endogenous metabolism generate reactive oxygen species (ROS) and oxygen-derived free radicals that are known to cause DNA damage[100, 101]. Their highly reactive properties allow them to form different kinds of oxidative damage in DNA, such as base modifications, bulky adducts and strand breaks[102]. In addition, other key players may be involved in producing basal mutations within cells, such as glucose metabolism[93], in which a 6% increase in glucose concentration can lead to a five-fold increase in mutation frequency.

With such highly toxic and mutagenic agents already present within cells, it is highly plausible for mutations to further increase upon exposure to environmental mutagens. Based on my observations, the coupled DNA damaging effect of formaldehyde and/or its derivatives with oxidative chemical species can result in increased toxicity and mutagenicity under aerobic respiration. These damages can be bulky adducts, insertions and deletions, and inter- or intra-strand crosslinks, which can be efficiently repaired via nucleotide excision repair, mismatch repair and inter-strand crosslink repair, respectively. However, extensive mutation accumulation induced by multiple damaging agents could have inevitably jeopardized some of these DNA repair pathways resulting in persistent base lesions and decreased repair capacity[103]. In addition, while DNA normally remains in its double-stranded conformation, it may transiently exist in its single-stranded form (ssDNA) during routine genomic processes, such as replication, which is much more susceptible to damages. In cases where insufficient protection of these ssDNA in replication

forks or frequent replication stress can trigger strand breakage or persist long stretches of ssDNA[104], the genome may further accumulate mutations given that it continuously experiences insults from various sources. Therefore, my findings underscore the importance of curbing environmental exposures so as not to compromise the diverse DNA repair proteins as well as genome stability.

4.2 Mutagenic Effects of Endogenous Formaldehyde (Low, Chronic Exposure)

Knowing that exogenous formaldehyde is quite mutagenic, it then becomes important to consider if physiological levels are also sufficient in inducing mutations. The effects of endogenous formaldehyde were examined through passaging of cells in aerobic respiration with essential dehydrogenases either deleted or conditionally depleted. Although initial results suggested that the absence of *SFA1* would cause an increase in mutation frequency even under favourable conditions, subsequent experiments showed conflicting results. Not only was the mutation frequency of *yjl068cΔ* single knockout significantly different from wild-type, the genotypic effect also seem to suggest that the absence of *YJL068C* will result in a lower mutation frequency. Statistical significance can be observed between the two strains; however, their fold-change difference weakens the evidence for biological significance. Furthermore, contrary to a previous assertion, it is unclear whether *YJL068C* presents a possible role in contributing to basal mutations in yeast since one study observed “no special behaviour of yeast cells” lacking the *YJL068C* ORF[105].

One possible reason for the lowered mutagenic effect observed in *yjl068c* Δ strains is that formate could be the mutagenic form of formaldehyde. Without *YJL068C*, cells could not carry out the final oxidation step producing formate, which may be one of the contributors to basal mutations observed in wild-type. Since other forms of formaldehyde also exist within cells – such as methanediol, the hydrated form of formaldehyde – it is important to consider the pathways involved in the formation of these derivatives as they may also be key contributors of intrinsic mutations within cells.

Interestingly, simultaneous deletion of *SFA1* and *YJL068C* produced detrimental effects on cell growth. This led us to use the auxin-inducible degron approach to control the ‘double-knockout’ phenotype. Given that our initial results revealed *SFA1* deletion caused higher mutation frequencies and *YJL068C* deletion had no significant effects, we decided to tag *SFA1* and delete *YJL068C* instead of tagging both dehydrogenases with the degron. Pre-growth cultures of the degron/knockout strains exhibited a three-fold increase in mutation frequency than wild-type. Upon the addition of auxin and shift to a restrictive temperature, however, mutation frequencies between controls and degron/knockout strains remained relatively the same. Mutation frequencies between auxin treatments and day-to-day comparisons were also similar across all strains.

As the results show steady mutation frequencies at around 10^{-3} per day (see Figure 3.1.7), this may indicate that the number of mutations within cells have already reached saturation after one day of exposure. It is possible to observe incremental increases in mutation frequency if measurements were perhaps taken every 2 or 3 hours as opposed to every 24 hours. This hypothesis of mutation saturation at around 10^{-3} is also supported

by observations from the high, acute exposures, in which mutation frequencies can only be measured up to this point with respect to the high concentrations (6 mM to 10 mM).

Another explanation for this observation may be due to the fact that yeast possesses numerous stress response enzymes. It is possible that other enzymes were also responsible for detoxifying endogenous formaldehyde even without the presence of Sfa1p or Yjl068cp. Toxicogenomic studies[65, 106] discovered several other genes in yeast that were induced upon formaldehyde exposure. Moreover, any resulting damages caused by formaldehyde, such as DNA-protein crosslinks, can be effectively corrected by corresponding repair enzymes[14, 107]. For these reasons, it is highly likely that yeast is quite tolerant to physiological concentrations of formaldehyde since natural selection has enabled them to develop highly adaptive biological systems[108]. Given the seemingly obscure dynamics of formaldehyde tolerance and mutagenicity within cells, it may be worthwhile to also examine the effects of endogenous formaldehyde with the absence or inactivation of these other relevant genes in yeast as well as their homologues in human or mammalian cells for comparison.

The disadvantage in this experimental design was the fact that a direct comparison between pre-growth (day 0) and treated cultures (days 1 to 3) would be misleading because of the change in temperature. Since the temperature shift was necessary to increase mutation induction in our yeast model, the auxin treatment needed to be under 37°C to effectively observe the effects of the combined absence of Sfa1p and Yjl068cp. However, our yeast cannot be cultured under 37°C because cells experience a cell cycle arrest in the G₂ phase at this temperature due to strand resections on all telomeric ends that are

perceived as DNA damage[68]. Alternatively, given that the degtron system was optimally designed for a range of temperatures, maintaining pre-growth temperature (23°C) during auxin treatment would be a reasonable approach to provide a much fair comparison on day-to-day mutagenesis. Conversely, this will be at the expense of inducing multiple mutations in the DNA especially since formaldehyde is a weak mutagen.

4.3 The Putative Single Base Substitution Signature of Formaldehyde

Since the exogenous formaldehyde-treated yeast genomes accrued the greatest number of mutations within a single genome, they were the ideal candidates for signature analysis. Lots and lots of mutations are required in order to effectively characterize the formaldehyde mutational signature. I have also included genomic data sets of untreated yeast genomes for comparison. In total, five different data sets were used for characterizing unique mutational signatures: three formaldehyde-treated yeast genomics data, and two untreated yeast genomics data. To examine potential differences between each set, the variant call format (VCF) files, which summarizes the variants or mutations in a sample genome, from a single data set were merged into one file, curating five VCF files altogether. By and large, the non-negative matrix factorization (NMF) algorithm was able to decipher two single base substitution (SBS) signatures from this collection, which can be uniquely identified to the ‘no formaldehyde’ treatment and ‘formaldehyde’ treatment.

Previously, the untreated genomic data sets (or ‘no formaldehyde’ treated cells) were analyzed by our group as having close similarity to COSMIC SBS5 (cossim = 0.906)[93]. Whereas the yeast-derived ‘no formaldehyde’ signature has a direct matching

pattern to a known signature from human cancers, the putative yeast-derived formaldehyde SBS signature appeared to be a composite signature of several signatures. This resulted in poor cosine similarities when the formaldehyde SBS signature was compared against all known mutational signatures. A quick visual examination of its profile revealed COSMIC SBS5-like base substitutions (see Figure 4.1 for SBS5), however, the high frequency of C→A/G→T transversions at TCn/AGn trinucleotide motifs on the FA SBS signature was most likely the differentiating pattern from the no FA SBS signature. Notably, the cosine similarity of the ‘no formaldehyde’ signature reported here versus SBS5 (cossim = 0.876) does not correspond with the previously reported cosine similarity value (cossim = 0.906). One possible explanation could be due to the merging of VCF files per data set, which was interpreted by the NMF algorithm as a collection of five samples only as opposed to a collection of over 400 genomes. This is inherently the nature of the NMF algorithm, where more sample data will produce lower residual error and generate high resolution signatures[109]. The major drawback is overfitting whereby resulting signatures become highly similar and could be less indicative of distinct mutagenic processes, as was observed in the recent update of COSMIC signatures[26]. Interestingly, when the mutational profiles of each data set were plotted, they each resemble the base substitution patterns exhibited by their associated mutational signature (see Figure 4.2). The formaldehyde-like pattern was observed exclusively in all formaldehyde-treated genomic data sets. This implies that despite the limitations of the NMF algorithm and sparsity of samples, the formaldehyde base substitution patterns are uniquely identifiable.

The composite nature of the FA SBS signature was not surprising as the genomes from which it was extracted from were exposed to formaldehyde in either conditions, i.e. samples from exogenous treatment under room air and nitrogen were pooled together. Since COSMIC SBS5 seems to be the underlying ‘endogenous metabolism’ signature, we reasoned that we may decipher at least two signatures from the formaldehyde-treated yeast genomes, which are probably SBS5 and the formaldehyde-specific signature. As results have shown (see Section 3.3.1), I was able to extract two unique signatures pertaining to formaldehyde and no formaldehyde treatments, however, the FA SBS signature was still a composite of three signatures.

On a previous reconstruction analysis, the combination of SBS5, SBS40 and SBS52 gave the most similar reconstruction profile to the putative FA signature ($\text{cossim} = 0.918$). SBS40 is quite similar to SBS5 – it is another clock-like signature that is found in multiple types of cancers and exhibits a ‘flat’ pattern. Because of this close similarity, whether SBS40 is a genuine signature or not remains quite ambiguous[26]. In the reconstruction analysis reported here, in which another filtering layer (depth score) was included, SBS40 and SBS52 were still found to be involved in the mutational profile of the FA SBS signature, however, SBS5 had been replaced by SBS25. SBS25 was derived from Hodgkin lymphoma cell lines treated with chemotherapy, but no primary cancers were available for this type. It was therefore suggested that SBS25 was likely formed due to chemotherapy treatment, which induces transcriptional strand bias mutations of T→A transitions. The discovery of this signature within the FA SBS signature is most likely the consequence of NMF overfitting[109] as SBS25-like mutations occur at very low

frequencies on the FA SBS signature. Therefore, we believe that the most reasonable biological explanation for the signatures observed in the FA SBS signature is still the combination of SBS5, SBS40, and SBS52. While the cosine similarity value of this triple-wise combination is lower than previous (cossim = 0.907), it is still a reasonable cosine similarity value to argue that SBS5 is more involved in the formaldehyde signature than SBS25 given that the mutation patterns of SBS5 are reproducible within the untreated (no formaldehyde) genomes[93].

In addition, SBS52-like base substitutions seem to reflect the formaldehyde pattern in our yeast genomics data. More specifically, the high frequency of TCA→TAA substitution appears to be a consensus pattern between SBS52 and the formaldehyde signature. However, SBS52 has been deemed a possible sequencing artefact because it was derived from very few cancer samples. The large-scale study by Kucab and colleagues (2019)[27] that involved the characterization of mutational signatures of various environmental mutagens also identified a formaldehyde signature from human-induced pluripotent stem cells. Although their results were inconclusive, their reported formaldehyde signature also reflects this high frequency of TCA→TAA substitutions observed in both SBS52 and our formaldehyde-treated yeast genomes.

By and large, although I was successful in identifying the specific formaldehyde mutational pattern, technical constraints suggested that more mutation data is required to generate a much highly resolved signature. Nevertheless, this result does provide a promising opportunity for further biochemical investigations to verify formaldehyde’s preference for C→A/G→T transversions, which has not been discovered previously.

Figure 4.1. COSMIC Signatures SBS5 and SBS40

SBS5 was first discovered[29] before additional mutation data became available allowing the extraction of SBS40 as well[26]. Although their cosine similarity value is 0.83, a relatively high cossim value, Alexandrov et al. (2020) still suspects whether SBS40 is a genuine signature or not.

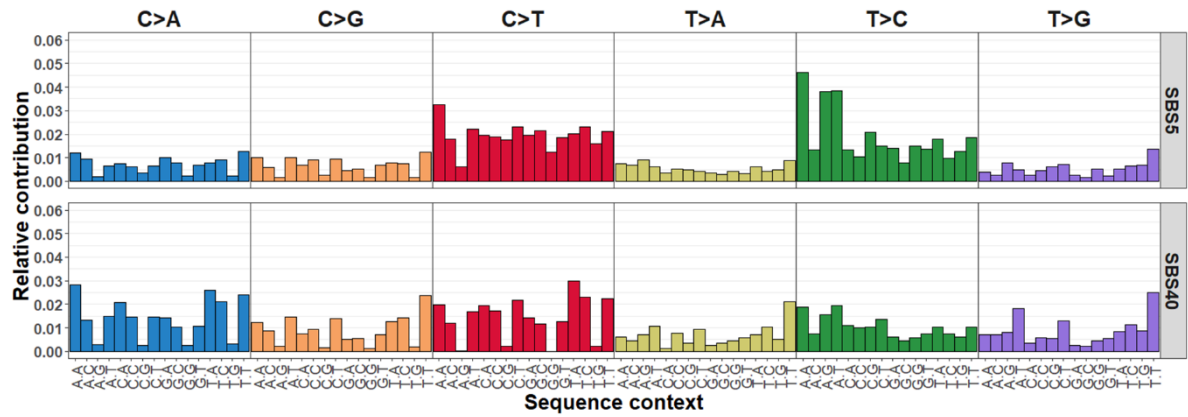
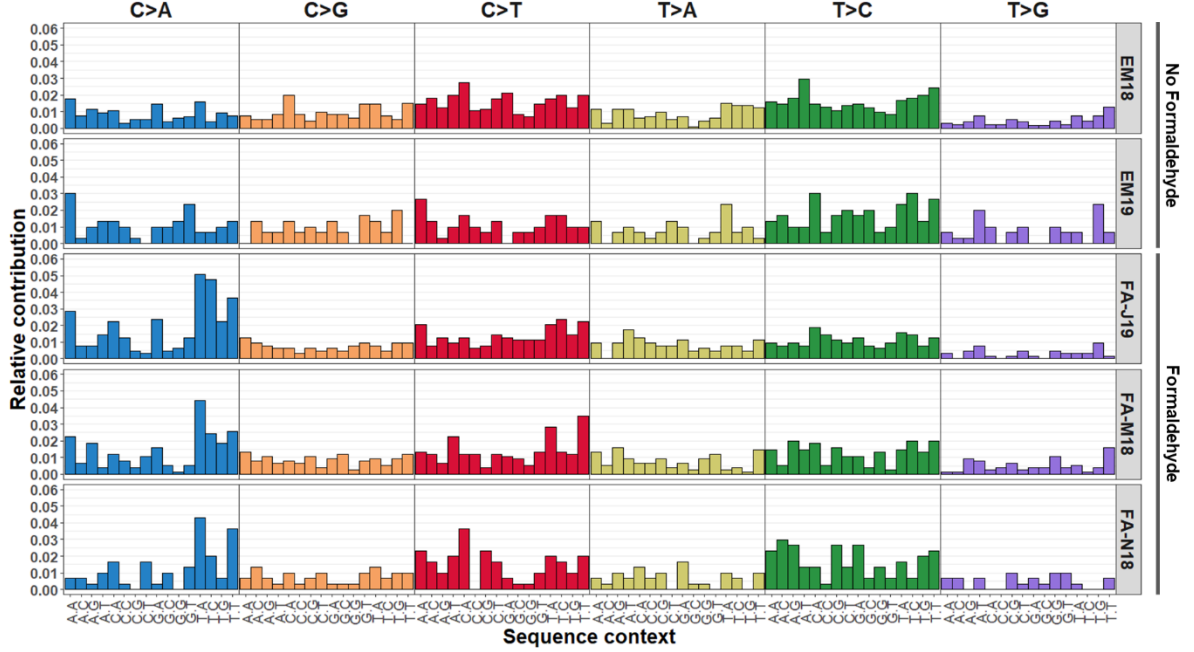


Figure 4.2. *Mutational Profiles of Each Untreated and Formaldehyde-treated Data Sets*

The formaldehyde-specific pattern of mostly C→A transversions in TCn motifs can only be observed in the formaldehyde data sets in variable proportions.



4.4 Indel Mutational Patterns of Formaldehyde

Comparisons of indel sizes showed high frequencies of 1-base insertions under formaldehyde exposure compared to no formaldehyde exposure, while the distribution of base deletions were relatively the same between treatments. Generally, indel size distributions follow a common pattern that has been observed among genomes of several species, such as humans and other mammals[110, 111], plants[112], bacteria[113] and yeasts[114, 115]. This was described as a power law distribution because the frequency of an indel length decreases as the length increases. A close examination of these species'

genomes revealed that almost 50% of all indels are single nucleotide indels (or 1bp indels) with high preference for deletions over insertions[116]. Contrary to this pattern, the untreated yeast genomes (i.e. ySR127 budding yeast strain) has a deletion to insertion ratio of 0.52 indicating that insertions are more frequent than deletions. Moreover, their insertions distribution does not follow the power law distribution, unlike the formaldehyde-treated yeast genomic data, which somewhat follows the widespread indel distribution with a deletion to insertion ratio of 1.31. The deletion pattern of each treatment also slightly follows power law, however, it is clear that longer deletion sizes are a common occurrence on both conditions. These are most likely telomere truncation events resulting from extensive oxidative stress[2]. Moreover, the high frequency of long insertions may be erroneous attempts of the homologous recombination pathway, which is the most common double strand repair mechanism in yeast, to repair the exposed telomeric ends[103, 117].

Although data is limited, the formaldehyde single nucleotide indel pattern presented here could be a genuine formaldehyde-specific pattern. Additionally, 1-base deletion events are more prone under non-repeating units, while 1-base insertion events occur frequently on homology sequences. Previous studies have implicated formaldehyde mutagenesis towards preferential targeting of adenine and guanine bases to form chemical derivatives or adducts[32, 58, 99]. However, it is unclear from these studies whether the mutagenic effect result in base substitutions or insertions/deletions. The indel pattern described here can compliment these former observations to infer a possible mechanism of formaldehyde mutagenesis.

To extend the indel length analysis, I also analyzed these indels based on surrounding repeat units as a pretext for examining their sequence context. One reason for using this approach instead of characterizing these indels based on the immediate 3' and 5' flanking regions is because it is unclear from the variant call format or VCF file as to where the indel has occurred. For example, the VCF file may report 'CCC' as an insertion at a particular position in a chromosome, however, it is uncertain if this insertion occurred at the 3' or 5' side of its reference point. Thus, following previously established strategies[26, 27], I have decided to also classify the sequence context of the formaldehyde indels through repeat units.

Given the complexity of the reported formaldehyde indels, which made computational characterization quite inefficient, I have taken a slightly conservative approach in analyzing these sequence contexts. Two general observations were critical in designing my classification approach: first, the sequence landscape where these indels occur are usually repetitive; and second, the reported indel is generally a repeat unit of the repetitive sequence. Taking into consideration the variability of short tandem repeat moieties and microhomology regions, some mismatches were allowed depending on the indel length. Obviously, some of the indels were not repeat units or perfect repeat units of their incidental location, which is especially true for longer insertions and deletions. Some of the long deletions for formaldehyde that were classified as frequent events in non-repetitive sequences could be one of several possibilities:

- they are indeed novel deletions from non-repeating units, which may be a result of severe formaldehyde bulky adduct formation;

- they are deletions that has occurred near a repetitive sequence, which further implicates the preference of formaldehyde mutagenesis over repetitive sequences[14];
or
- they are found over microhomology regions, which could be evidence for microhomology-mediated deletions characteristic of DNA double-strand breaks.

Verification of these different scenarios require a more appropriate and stringent indel classification criteria as well as additional indel data. For this reason, it is difficult to make relevant inferences on the characterized indel pattern of formaldehyde exposure given its ambiguity over longer indels. Moreover, the comparison metric used here – cosine similarity – is susceptible to significant shifts in data resulting in similarity values that may be favourable or less favourable. For this reason, comparison of signatures through residual sum of squares (RSS) or Pearson’s correlation could be better alternatives as these methods analyze the variation between the two signatures providing a more stable assessment of their ‘relatedness’, and a significance value is also given for this comparison.

4.5 Relevance of the Formaldehyde SBS Signature in Cancer Genomics Data

The COSMIC mutational signatures have been considered statistically reliable because they were analyzed and extracted over millions and millions of cancer mutation data. However, the authors preserve caution over their generality and authenticity as these signatures may have been potentially influenced by confounding factors as well as the mathematical approach they used[26], hence, the mutational landscape of a signature

can be subjected to computational adjustments upon the addition of more mutation data. As mentioned earlier, additional data provides the NMF algorithm increased ability to confidently recognize unique signatures and effectively decrease stochastic noise and signature overlap. The caveat would be over-extraction resulting in the characterization of some signatures that resemble previously identified signatures. This was demonstrated through the cosine similarity analysis of all COSMIC signatures where it shows that several of them have highly similar mutational profiles from each other (see Figure 4.3).

One approach to improve NMF-based strategies is to use a Bayesian approach in which a set of *a priori* data is used for the initial analysis of signatures, and a second set to correct and stabilize the initial set[118]. Unlike the conventional NMF strategy, Bayesian NMF infers for the optimal number of signatures (i.e. factorization rank number, r) from the data rather than requiring the user to manually input this value. The Bayesian variant also automatically generates signatures that is balanced between “data fidelity” and “model complexity”[119–121]. Another approach is to use a mixed-membership model where it uses an Expectation Maximization (EM) algorithm to first, estimate hidden activities (or latent variables) within the data set, and then, optimize the parameters of the model to best explain the mutational spectra within the given data set[118, 122, 123]. This approach also inherently selects the optimal number of signatures within the data set using probabilistic model selection. It is quite reassuring that despite the computational constraints of conventional NMF used by Alexandrov et al.[26, 29], its characterization of mutational signatures operative in cancer samples have been helpful starting points for identifying causes for cancer initiation.

While the deciphered formaldehyde SBS signature from yeast genomics data was not directly similar to any known mutational signature, the suspected formaldehyde-specific pattern can be observed from some cancer genomics data. Although considered a sequencing artefact, SBS52 may effectively be the formaldehyde-specific signature despite its low abundance on various cancer types. Interestingly, the SBS52+ samples are a slight exception to the power law distribution of indels since the highest frequency of indels observed are two base indels, and it also deviates from the formaldehyde indel pattern observed in yeast. These SBS52+ samples, however, display a dinucleotide indel pattern similar to SBS52- samples pattern (see Figure 3.3.4, right panels), and it seems that the only indication of difference between SBS52+ and SBS52- samples is related to the number of dinucleotide mutations found from each data set. Again, because the comparison metric used is cosine similarity, using a residual sum of squares or Pearson's correlation approach may provide a more reasonable assessment to the association of these data sets' patterns.

In addition to SBS52+ samples, a low frequency of the formaldehyde-specific pattern can also be observed in one of the *de novo* extracted signature from head and neck squamous cell carcinoma samples – Signature A. However, because of the small sample size causing signature overlap, it is quite difficult to make this association. Cosine similarity comparisons also implicated Signature B as being the ‘most similar’ signature to the formaldehyde SBS signature out of the three Head-SCC *de novo* extracted signatures (cossim = 0.811). Whether the putative formaldehyde SBS signature is indeed Signature B, or the suspected FA pattern can be observed from Signature A needs further

investigation. Therefore, as previously asserted, more mutation data can improve the accuracy of the putative formaldehyde mutational signature as well as the *de novo* extracted Head-SCC signatures. Furthermore, analysis of specific formaldehyde-related cancers would provide a much more reasonable assessment regarding formaldehyde’s prevalence in cancers, instead of analyzing the entire head and neck cancer data. Unlike nose and throat cancers (nasopharyngeal and sinonasal cancers), which are relevant formaldehyde cancers, other cancer types encompassed in this group may effectively have lower frequencies of formaldehyde mutational patterns, and therefore, would be considered background or noise as opposed to other mutational patterns upon signature extraction.

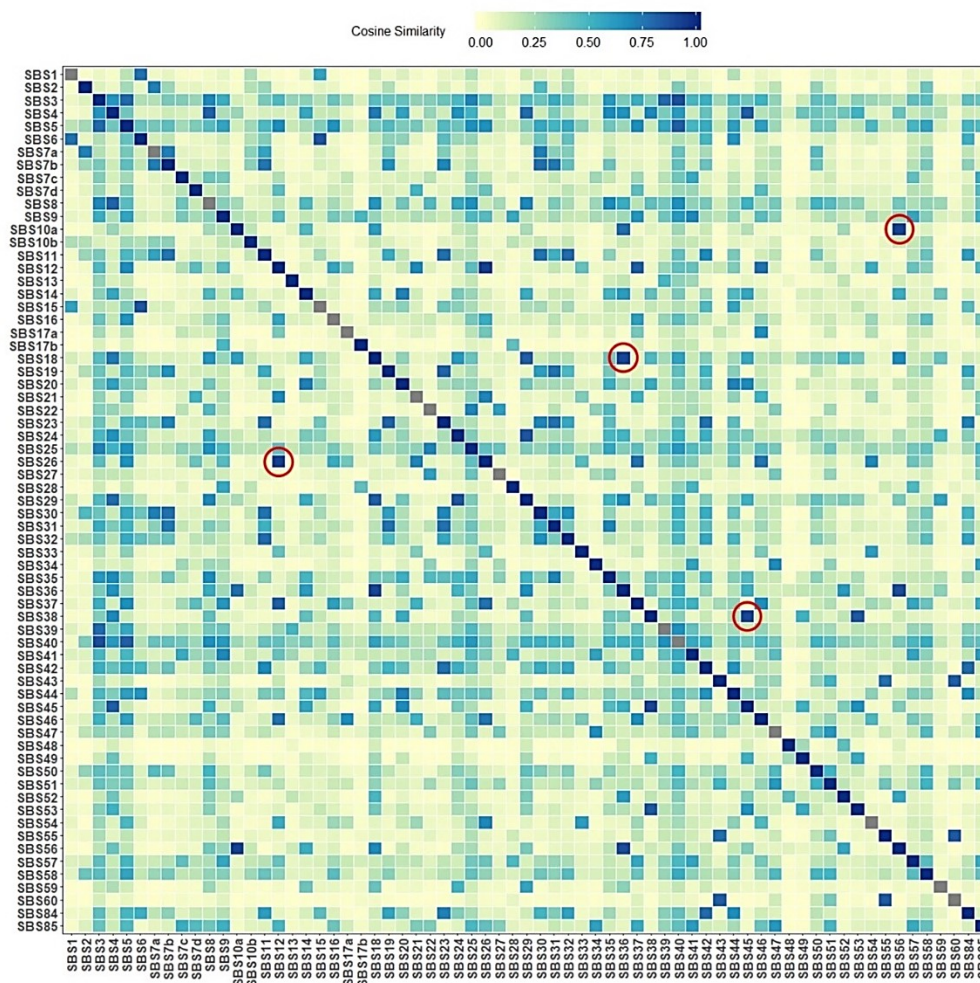
4.6 Conclusions

Formaldehyde is a mutagenic entity capable of inducing various DNA modifications. To characterize these mutations, the NMF algorithm was used to decipher unique signatures within formaldehyde-treated yeast genomes. However, NMF can only provide a best estimate of the mutational landscape within the context of the given data. The results suggest that additional mutation data may increase the accuracy of signature extraction and reduce residual noise within characterized signatures. While the approaches described here had several challenges, there are some successes in characterizing formaldehyde-specific base substitution and indel patterns, which has never been described before. Moreover, the formaldehyde-specific pattern of TCA→TAA transversions discovered in yeast has been observed in other biological systems despite the ambiguity of results from these models. Indeed, more elaborate investigations are still needed to further

verify these formaldehyde mutational patterns. Nonetheless, this work has further elucidated a probable mechanism for formaldehyde mutagenesis on the nucleotide level.

Figure 4.3. Heatmap Depicting Cosine Similarities of COSMIC SBS Signatures Against Each Other

The red circles highlight four comparisons that produced a cosine similarity value greater than 0.90: SBS12 vs. SBS26 = 0.93; SBS10a vs. SBS56* = 0.92; SBS36 vs. SBS18 = 0.91; SBS45* vs. SBS38 = 0.91. Signatures marked with * are reported as possible sequencing artefacts.



References

1. Friedberg EC, Walker GC, Siede W, Wood RD (2005) DNA Repair and Mutagenesis. American Society for Microbiology Press
2. Chatterjee N, Walker GC (2017) Mechanisms of DNA damage, repair and mutagenesis. *Environ Mol Mutagen* 58:235–263. <https://doi.org/10.1002/em.22087>
3. Dumaz N, Drougard C, Sarasin A, Daya-Grosjean L (1993) Specific UV-induced mutation spectrum in the p53 gene of skin tumors from DNA-repair-deficient xeroderma pigmentosum patients. *Proc Natl Acad Sci U S A* 90:10529–10533
4. Gentil A, Le Page F, Margot A, et al (1996) Mutagenicity of a unique thymine-thymine dimer or thymine-thymine pyrimidine pyrimidone (6-4) photoproduct in mammalian cells. *Nucleic Acids Res* 24:1837–1840. <https://doi.org/10.1093/nar/24.10.1837>
5. Helleday T, Eshtad S, Nik-Zainal S (2014) Mechanisms underlying mutational signatures in human cancers. *Nat Rev Genet* 15:585. <https://doi.org/10.1038/nrg3729>
6. Desouky O, Ding N, Zhou G (2015) Targeted and non-targeted effects of ionizing radiation. *J Radiat Res Appl Sci* 8:247–254. <https://doi.org/10.1016/j.jrras.2015.03.003>
7. Gichner T (2003) DNA damage induced by indirect and direct acting mutagens in catalase-deficient transgenic tobacco: Cellular and acellular Comet assays. *Mutat Res Toxicol Environ Mutagen* 535:187–193. [https://doi.org/10.1016/S1383-5718\(02\)00320-0](https://doi.org/10.1016/S1383-5718(02)00320-0)
8. Chan K, Roberts SA, Klimczak LJ, et al (2015) An APOBEC3A hypermutation signature is distinguishable from the signature of background mutagenesis by APOBEC3B in human cancers. *Nat Genet* 47:1067–1072. <https://doi.org/10.1038/ng.3378>
9. Burns MB, Leonard B, Harris RS (2015) APOBEC3B: pathological consequences of an innate immune DNA mutator. *Biomed J* 38:102–110
10. Petljak M, Alexandrov LB, Brummel JS, et al (2019) Characterizing Mutational Signatures in Human Cancer Cell Lines Reveals Episodic APOBEC Mutagenesis. *Cell* 176:1282–1294.e20. <https://doi.org/10.1016/j.cell.2019.02.012>
11. Boiteux S, Jinks-Robertson S (2013) DNA Repair Mechanisms and the Bypass of DNA Damage in *Saccharomyces cerevisiae*. *Genetics* 193:1025–1064. <https://doi.org/10.1534/genetics.112.145219>
12. Types of DNA Lesions. http://utminers.utep.edu/rwebb/html/types_of_dna_lesions.html. Accessed 31 Oct 2017
13. Mao P, Wyrick JJ (2019) Organization of DNA damage, excision repair, and mutagenesis in chromatin: A genomic perspective. *DNA Repair* 81:102645. <https://doi.org/10.1016/j.dnarep.2019.102645>
14. Grogan D, Jinks-Robertson S (2012) Formaldehyde-induced mutagenesis in *Saccharomyces cerevisiae*: molecular properties and the roles of repair and bypass systems. *Mutat Res* 731:92–98. <https://doi.org/10.1016/j.mrfmmm.2011.12.004>

15. Yuan J, Narayanan L, Rockwell S, Glazer PM (2000) Diminished DNA Repair and Elevated Mutagenesis in Mammalian Cells Exposed to Hypoxia and Low pH. *Cancer Res* 60:4372–4376
16. Deans AJ, West SC (2011) DNA interstrand crosslink repair and cancer. *Nat Rev Cancer* 11:467–480. <https://doi.org/10.1038/nrc3088>
17. Ames B N, Shigenaga M K, Gold L S (1993) DNA lesions, inducible DNA repair, and cell division: three key factors in mutagenesis and carcinogenesis. *Environ Health Perspect* 101:35–44. <https://doi.org/10.1289/ehp.93101s535>
18. National Cancer Institute (2007) What Is Cancer? <https://www.cancer.gov/about-cancer/understanding/what-is-cancer>. Accessed 3 May 2020
19. Ferris E, Abegglen LM, Schiffman JD, Gregg C (2018) Accelerated Evolution in Distinctive Species Reveals Candidate Elements for Clinically Relevant Traits, Including Mutation and Cancer Resistance. *Cell Rep* 22:2742–2755. <https://doi.org/10.1016/j.celrep.2018.02.008>
20. Kujoth GC, Hiona A, Pugh TD, et al (2005) Mitochondrial DNA mutations, oxidative stress, and apoptosis in mammalian aging. *Science* 309:481–484. <https://doi.org/10.1126/science.1112125>
21. Risques RA, Kennedy SR (2018) Aging and the rise of somatic cancer-associated mutations in normal tissues. *PLoS Genet* 14:. <https://doi.org/10.1371/journal.pgen.1007108>
22. Lestienne P (1992) Mitochondrial DNA mutations in human diseases: a review. *Biochimie* 74:123–130. [https://doi.org/10.1016/0300-9084\(92\)90035-D](https://doi.org/10.1016/0300-9084(92)90035-D)
23. Richard P, Manley JL (2017) R Loops and Links to Human Disease. *J Mol Biol* 429:3168–3180. <https://doi.org/10.1016/j.jmb.2016.08.031>
24. Alexandrov LB, Jones PH, Wedge DC, et al (2015) Clock-like mutational processes in human somatic cells. *Nat Genet* 47:1402–1407. <https://doi.org/10.1038/ng.3441>
25. Alexandrov LB, Nik-Zainal S, Wedge DC, et al (2013) Signatures of mutational processes in human cancer. *Nature* 500:415–421. <https://doi.org/10.1038/nature12477>
26. Alexandrov LB, Kim J, Haradhvala NJ, et al (2020) The repertoire of mutational signatures in human cancer. *Nature* 578:94–101. <https://doi.org/10.1038/s41586-020-1943-3>
27. Kucab JE, Zou X, Morganella S, et al (2019) A Compendium of Mutational Signatures of Environmental Agents. *Cell* 177:821–836.e16. <https://doi.org/10.1016/j.cell.2019.03.001>
28. Lindahl T (1993) Instability and decay of the primary structure of DNA. *nature* 362:709–715
29. Alexandrov LB, Nik-Zainal S, Wedge DC, et al (2013) Deciphering Signatures of Mutational Processes Operative in Human Cancer. *Cell Rep* 3:246–259. <https://doi.org/10.1016/j.celrep.2012.12.008>
30. Alexandrov LB, Stratton MR (2014) Mutational signatures: the patterns of somatic mutations hidden in cancer genomes. *Curr Opin Genet Dev* 24:52–60. <https://doi.org/10.1016/j.gde.2013.11.014>

31. COSMIC: Signatures of Mutational Processes in Human Cancer.
<http://cancer.sanger.ac.uk/cosmic/signatures>. Accessed 14 Oct 2017
32. Alderson T (1985) Formaldehyde-induced mutagenesis: A novel mechanism for its action. *Mutat Res Genet Toxicol* 154:101–110. [https://doi.org/10.1016/0165-1110\(85\)90022-3](https://doi.org/10.1016/0165-1110(85)90022-3)
33. Moerman DG, Baillie DL (1981) Formaldehyde mutagenesis in the nematode *Caenorhabditis elegans*. *Mutat Res* 80:273–279. [https://doi.org/10.1016/0027-5107\(81\)90100-7](https://doi.org/10.1016/0027-5107(81)90100-7)
34. Merk O, Speit G (1998) Significance of formaldehyde-induced DNA–protein crosslinks for mutagenesis. *Environ Mol Mutagen* 32:260–268. [https://doi.org/10.1002/\(SICI\)1098-2280\(1998\)32:3<260::AID-EM9>3.0.CO;2-M](https://doi.org/10.1002/(SICI)1098-2280(1998)32:3<260::AID-EM9>3.0.CO;2-M)
35. International Agency for Research on Cancer (2006) Formaldehyde, 2-butoxyethanol and 1-tert-butoxypropan-2-ol. In: *IARC Monographs on the Evaluation of Carcinogenic Risks to Humans*. pp 1–478
36. International Agency for Research on Cancer (2012) Formaldehyde. In: *IARC Monographs on the Evaluation of Carcinogenic Risks to Human*. pp 401–435
37. Reingruber H, Pontel LB (2018) Formaldehyde metabolism and its impact on human health. *Curr Opin Toxicol* 9:28–34. <https://doi.org/10.1016/j.cotox.2018.07.001>
38. Tulpule K, Dringen R (2013) Formaldehyde in brain: an overlooked player in neurodegeneration? *J Neurochem* 127:7–21. <https://doi.org/10.1111/jnc.12356>
39. Lee E-S, Chen H, Hardman C, et al (2008) Excessive S-adenosyl-l-methionine-dependent methylation increases levels of methanol, formaldehyde and formic acid in rat brain striatal homogenates: Possible role in S-adenosyl-l-methionine-induced Parkinson's disease-like disorders. *Life Sci* 83:821–827. <https://doi.org/10.1016/j.lfs.2008.09.020>
40. Burgos-Barragan G, Wit N, Meiser J, et al (2017) Mammals divert endogenous genotoxic formaldehyde into one-carbon metabolism. *Nature* 548:549. <https://doi.org/10.1038/nature23481>
41. Boor PJ, Trent MB, Lyles GA, et al (1992) Methylamine metabolism to formaldehyde by vascular semicarbazide-sensitive amine oxidase. *Toxicology* 73:251–258. [https://doi.org/10.1016/0300-483X\(92\)90067-O](https://doi.org/10.1016/0300-483X(92)90067-O)
42. O'Sullivan J, Unzeta M, Healy J, et al (2004) Semicarbazide-Sensitive Amine Oxidases: Enzymes with Quite a Lot to Do. *NeuroToxicology* 25:303–315. [https://doi.org/10.1016/S0161-813X\(03\)00117-7](https://doi.org/10.1016/S0161-813X(03)00117-7)
43. Yu PH, Wright S, Fan EH, et al (2003) Physiological and pathological implications of semicarbazide-sensitive amine oxidase. *Biochim Biophys Acta BBA - Proteins Proteomics* 1647:193–199. [https://doi.org/10.1016/S1570-9639\(03\)00101-8](https://doi.org/10.1016/S1570-9639(03)00101-8)
44. Hou H, Yu H (2010) Structural insights into histone lysine demethylation. *Curr Opin Struct Biol* 20:739–748. <https://doi.org/10.1016/j.sbi.2010.09.006>
45. Oro' J, Kimball A, Fritz R, Master F (1959) Amino acid synthesis from formaldehyde and hydroxylamine. *Arch Biochem Biophys* 85:115–130. [https://doi.org/10.1016/0003-9861\(59\)90455-2](https://doi.org/10.1016/0003-9861(59)90455-2)

46. Blakley RL (1954) The interconversion of serine and glycine: role of pteroylglutamic acid and other cofactors. *Biochem J* 58:448–462
47. Wang R, Zeng Z, Liu T, et al (2016) A novel formaldehyde metabolic pathway plays an important role during formaldehyde metabolism and detoxification in tobacco leaves under liquid formaldehyde stress. *Plant Physiol Biochem* 105:233–241. <https://doi.org/10.1016/j.plaphy.2016.04.028>
48. Dorokhov YL, Sheshukova EV, Bialik TE, Komarova TV (2018) Human Endogenous Formaldehyde as an Anticancer Metabolite: Its Oxidation Downregulation May Be a Means of Improving Therapy. *BioEssays* 40:1800136. <https://doi.org/10.1002/bies.201800136>
49. Health Canada (2012) Formaldehyde in Indoor Air. In: aem. <https://www.canada.ca/en/health-canada/services/environmental-workplace-health/reports-publications/air-quality/formaldehyde-indoor-air-environment-workplace-health.html>. Accessed 2 May 2020
50. National Toxicology Program (2016) Formaldehyde. In: Report on Carcinogens, 14th ed. Department of Health and Human Services
51. National Toxicology Program: 14th Report on Carcinogens. In: Natl. Toxicol. Program NTP. <https://ntp.niehs.nih.gov/go/roc14>. Accessed 3 May 2020
52. Meng F, Bermudez E, McKinzie PB, et al (2010) Measurement of tumor-associated mutations in the nasal mucosa of rats exposed to varying doses of formaldehyde. *Regul Toxicol Pharmacol* 57:274–283. <https://doi.org/10.1016/j.yrtph.2010.03.007>
53. Pontel LB, Rosado IV, Burgos-Barragan G, et al (2015) Endogenous Formaldehyde Is a Hematopoietic Stem Cell Genotoxin and Metabolic Carcinogen. *Mol Cell* 60:177–188. <https://doi.org/10.1016/j.molcel.2015.08.020>
54. Langevin F, Crossan GP, Rosado IV, et al (2011) Formaldehyde catabolism is essential in cells deficient for the Fanconi anemia DNA-repair pathway. *Nat Struct Mol Biol* 18:1432. <https://doi.org/10.1038/nsmb.2173>
55. Ridpath JR, Nakamura A, Tano K, et al (2007) Cells Deficient in the FANC/BRCA Pathway Are Hypersensitive to Plasma Levels of Formaldehyde. *Cancer Res* 67:11117–11122. <https://doi.org/10.1158/0008-5472.CAN-07-3028>
56. Garaycochea JI, Crossan GP, Langevin F, et al (2018) Alcohol and endogenous aldehydes damage chromosomes and mutate stem cells. *Nature* 553:171–177. <https://doi.org/10.1038/nature25154>
57. Auerbach C, Moutschen-Dahmen M, Moutschen J (1977) Genetic and cytogenetical effects of formaldehyde and related compounds. *Mutat Res Genet Toxicol* 39:317–361. [https://doi.org/10.1016/0165-1110\(77\)90011-2](https://doi.org/10.1016/0165-1110(77)90011-2)
58. Swenberg JA, Moeller BC, Lu K, et al (2013) Formaldehyde Carcinogenicity Research: 30 Years and Counting for Mode of Action, Epidemiology, and Cancer Risk Assessment. *Toxicol Pathol* 41:181–189. <https://doi.org/10.1177/0192623312466459>
59. de Serres FJ, Brockman HE, Hung CY (1988) Effect of the homokaryotic state of the uvs-2 allele in *Neurospora crassa* on formaldehyde-induced killing and ad-3 mutation. *Mutat Res Mol Mech Mutagen* 199:235–242. [https://doi.org/10.1016/0027-5107\(88\)90251-5](https://doi.org/10.1016/0027-5107(88)90251-5)

60. de Serres FJ, Brockman HE (1999) Comparison of the spectra of genetic damage in formaldehyde-induced ad-3 mutations between DNA repair-proficient and -deficient heterokaryons of *Neurospora crassa*. *Mutat Res Mutat Res* 437:151–163. [https://doi.org/10.1016/S1383-5742\(99\)00081-2](https://doi.org/10.1016/S1383-5742(99)00081-2)
61. Duina AA, Miller ME, Keeney JB (2014) Budding Yeast for Budding Geneticists: A Primer on the *Saccharomyces cerevisiae* Model System. *Genetics* 197:33–48. <https://doi.org/10.1534/genetics.114.163188>
62. Saccharomyces Genome Database | SGD. <https://www.yeastgenome.org/>. Accessed 4 May 2020
63. Kachroo AH, Laurent JM, Yellman CM, et al (2015) Systematic humanization of yeast genes reveals conserved functions and genetic modularity. *Science* 348:921–925. <https://doi.org/10.1126/science.aaa0769>
64. Liu W, Li L, Ye H, et al (2017) From *Saccharomyces cerevisiae* to human: The important gene co-expression modules. *Biomed Rep* 7:153–158. <https://doi.org/10.3892/br.2017.941>
65. North M, Gaytán BD, Romero C, et al (2016) Functional Toxicogenomic Profiling Expands Insight into Modulators of Formaldehyde Toxicity in Yeast. *Front Genet* 7:. <https://doi.org/10.3389/fgene.2016.00200>
66. Zeyl C (2000) Budding yeast as a model organism for population genetics. *Yeast* 16:773–784. [https://doi.org/10.1002/1097-0061\(20000615\)16:8<773::AID-YEA599>3.0.CO;2-1](https://doi.org/10.1002/1097-0061(20000615)16:8<773::AID-YEA599>3.0.CO;2-1)
67. Roberts SA, Sterling J, Thompson C, et al (2012) Clustered Mutations in Yeast and in Human Cancers Can Arise from Damaged Long Single-Strand DNA Regions. *Mol Cell* 46:424–435. <https://doi.org/10.1016/j.molcel.2012.03.030>
68. Chan K, Sterling JF, Roberts SA, et al (2012) Base damage within single-strand DNA underlies in vivo hypermutability induced by a ubiquitous environmental agent. *PLoS Genet* 8:e1003149. <https://doi.org/10.1371/journal.pgen.1003149>
69. Yang Y, Sterling J, Storici F, et al (2008) Hypermutability of Damaged Single-Strand DNA Formed at Double-Strand Breaks and Uncapped Telomeres in Yeast *Saccharomyces cerevisiae*. *PLoS Genet* 4:. <https://doi.org/10.1371/journal.pgen.1000264>
70. Saccharomyces Genome Database *Saccharomyces cerevisiae* formaldehyde oxidation II (glutathione-dependent). In: SGD Yeast Pathw. <https://pathway.yeastgenome.org/YEAST/NEW-IMAGE?type=PATHWAY&object=PWY-1801-1>. Accessed 4 May 2020
71. Gömpel-Klein P, Mack M, Brendel M (1989) Molecular characterization of the two genes SNQ and SFA that confer Hyperresistance to 4-nitroquinoline-N-oxide and formaldehyde in *Saccharomyces cerevisiae*. *Curr Genet* 16:65–74. <https://doi.org/10.1007/BF00393397>
72. Wehner E, Brendel M (1993) Formaldehyde lacks genotoxicity in formaldehyde-hyperresistant strains of the yeast *Saccharomyces cerevisiae*. *Mutat Res Mol Mech Mutagen* 289:91–96. [https://doi.org/10.1016/0027-5107\(93\)90134-2](https://doi.org/10.1016/0027-5107(93)90134-2)
73. Degraassi G, Uotila L, Klima R, Venturi V (1999) Purification and Properties of an Esterase from the Yeast *Saccharomyces cerevisiae* and Identification of the Encoding Gene. *Appl Environ Microbiol* 65:3470–3472

74. Wehner EP, Rao E, Brendel M Molecular structure and genetic regulation of SFA, a gene responsible for resistance to formaldehyde in *Saccharomyces cerevisiae*, and characterization of its protein product. 8
75. Yasokawa D, Murata S, Iwahashi Y, et al (2010) Toxicity of Methanol and Formaldehyde Towards *Saccharomyces cerevisiae* Assessed by DNA Microarray Analysis. *Appl Biochem Biotechnol* 160:1685–1698. <https://doi.org/10.1007/s12010-009-8684-y>
76. Viera B (2009) Colony Counter. In: ImageJ. <https://imagej.nih.gov/ij/plugins/colony-counter.html>. Accessed 4 May 2020
77. Nishimura K, Fukagawa T, Takisawa H, et al (2009) An auxin-based degron system for the rapid depletion of proteins in nonplant cells. *Nat Methods* 6:917–922. <https://doi.org/10.1038/nmeth.1401>
78. Nishimura K, Kanemaki MT (2014) Rapid Depletion of Budding Yeast Proteins via the Fusion of an Auxin-Inducible Degron (AID). *Curr Protoc Cell Biol* 64:20.9.1–20.9.16. <https://doi.org/10.1002/0471143030.cb2009s64>
79. Papagiannakis A, Niebel B, Wit EC, Heinemann M (2017) Autonomous Metabolic Oscillations Robustly Gate the Early and Late Cell Cycle. *Mol Cell* 65:285–295. <https://doi.org/10.1016/j.molcel.2016.11.018>
80. Storici F, Resnick MA (2006) The Delitto Perfetto Approach to In Vivo Site-Directed Mutagenesis and Chromosome Rearrangements with Synthetic Oligonucleotides in Yeast. In: *Methods in Enzymology*. Academic Press, pp 329–345
81. SnapGene software (from Insightful Science; available at snapgene.com)
82. ThermoFisher Scientific Custom DNA Oligos - CA. <https://www.thermofisher.com/ca/en/home/life-science/oligonucleotides-primers-probes-genes/custom-dna-oligos.html>. Accessed 4 May 2020
83. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9:357–359. <https://doi.org/10.1038/nmeth.1923>
84. Li H, Handsaker B, Wysoker A, et al (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25:2078–2079. <https://doi.org/10.1093/bioinformatics/btp352>
85. Li H (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinforma Oxf Engl* 27:2987–2993. <https://doi.org/10.1093/bioinformatics/btr509>
86. Blokzijl F, Janssen R, Bostel R van, Cuppen E (2019) MutationalPatterns: Comprehensive genome-wide analysis of mutational processes. Bioconductor version: Release (3.10)
87. Gaujoux R, Seoighe C (2010) A flexible R package for nonnegative matrix factorization. *BMC Bioinformatics* 11:367. <https://doi.org/10.1186/1471-2105-11-367>
88. Brunet J-P, Tamayo P, Golub TR, Mesirov JP (2004) Metagenes and molecular pattern discovery using matrix factorization. *Proc Natl Acad Sci U S A* 101:4164–4169. <https://doi.org/10.1073/pnas.0308531101>

89. Han J, Kamber M, Pei J (2012) 2 - Getting to Know Your Data. In: Han J, Kamber M, Pei J (eds) *Data Mining (Third Edition)*. Morgan Kaufmann, Boston, pp 39–82
90. Hermann P (2020) STRAH (Short Tandem Repeats Analysis of Hotspot Zones)
91. Heissl A, Betancourt AJ, Hermann P, et al (2018) Length asymmetry and heterozygosity strongly influences the evolution of poly-A microsatellites at meiotic recombination hotspots. *bioRxiv* 431841. <https://doi.org/10.1101/431841>
92. National Cancer Institute (2013) Head and Neck Cancers. <https://www.cancer.gov/types/head-and-neck/head-neck-fact-sheet>. Accessed 4 May 2020
93. Gelova SP, Doherty KN, Alasmar S, Chan K (2020) Intrinsic base substitution patterns in diverse species reveal links to cancer and metabolism. *bioRxiv* 758540. <https://doi.org/10.1101/758540>
94. Mack M, Gompel-Klein P, Haase E, et al (1988) Genetic characterization of hyperresistance to formaldehyde and 4-nitroquinoline-N-oxide in the yeast *Saccharomyces cerevisiae*. *MGG Mol Gen Genet* 211:260–265. <https://doi.org/10.1007/BF00330602>
95. Cortes-Ciriano I, Lee S, Park W-Y, et al (2017) A molecular portrait of microsatellite instability across multiple cancers. *Nat Commun* 8:. <https://doi.org/10.1038/ncomms15180>
96. Pfeifer GP, Denissenko MF, Olivier M, et al (2002) Tobacco smoke carcinogens, DNA damage and p53 mutations in smoking-associated cancers. *Oncogene* 21:7435–7451. <https://doi.org/10.1038/sj.onc.1205803>
97. Lee MB, Dowsett IT, Carr DT, et al (2019) Defining the impact of mutation accumulation on replicative lifespan in yeast using cancer-associated mutator phenotypes. *Proc Natl Acad Sci* 116:3062–3071. <https://doi.org/10.1073/pnas.1815966116>
98. Johnsen RC, Baillie DL (1988) Formaldehyde mutagenesis of the eT1 balanced region in *Caenorhabditis elegans*: Dose—response curve and the analysis of mutational events. *Mutat Res Mol Mech Mutagen* 201:137–147. [https://doi.org/10.1016/0027-5107\(88\)90120-0](https://doi.org/10.1016/0027-5107(88)90120-0)
99. Yu R, Lai Y, Hartwell HJ, et al (2015) Formation, Accumulation, and Hydrolysis of Endogenous and Exogenous Formaldehyde-Induced DNA Damage. *Toxicol Sci* 146:170–182. <https://doi.org/10.1093/toxsci/kfv079>
100. Pluskota-Karwatka D (2008) Modifications of nucleosides by endogenous mutagens–DNA adducts arising from cellular processes. *Bioorganic Chem* 36:198–213. <https://doi.org/10.1016/j.bioorg.2008.04.002>
101. Turrens JF (2003) Mitochondrial formation of reactive oxygen species. *J Physiol* 552:335–344. <https://doi.org/10.1111/j.1469-7793.2003.00335.x>
102. Møller P, Wallin H (1998) Adduct formation, mutagenesis and nucleotide excision repair of DNA damage produced by reactive oxygen species and lipid peroxidation product. *Mutat Res Mutat Res* 410:271–290. [https://doi.org/10.1016/S1383-5742\(97\)00041-0](https://doi.org/10.1016/S1383-5742(97)00041-0)
103. Hakem R (2008) DNA-damage repair; the good, the bad, and the ugly. *EMBO J* 27:589–605. <https://doi.org/10.1038/emboj.2008.15>

104. Tubbs A, Sridharan S, van Wietmarschen N, et al (2018) Dual Roles of Poly(dA:dT) Tracts in Replication Initiation and Fork Collapse. *Cell* 174:1127-1142.e19. <https://doi.org/10.1016/j.cell.2018.07.011>
105. Vandenbol M, Portetelle D (1999) Disruption of six ORFs on *Saccharomyces cerevisiae* chromosome X: the YJL069c gene of unknown function is essential to cell viability. *Yeast* 15:1411-1417. [https://doi.org/10.1002/\(SICI\)1097-0061\(19990930\)15:13<1411::AID-YEA463>3.0.CO;2-M](https://doi.org/10.1002/(SICI)1097-0061(19990930)15:13<1411::AID-YEA463>3.0.CO;2-M)
106. Yasokawa D, Iwahashi H (2010) Toxicogenomics using yeast DNA microarrays. *J Biosci Bioeng* 110:511-522. <https://doi.org/10.1016/j.jbiosc.2010.06.003>
107. Klages-Mundt NL, Li L (2017) Formation and Repair of DNA-Protein Crosslink Damage. *Sci China Life Sci* 60:1065-1076. <https://doi.org/10.1007/s11427-017-9183-4>
108. Carreto L, Eiriz MF, Gomes AC, et al (2008) Comparative genomics of wild type yeast strains unveils important genome diversity. *BMC Genomics* 9:524. <https://doi.org/10.1186/1471-2164-9-524>
109. Ramazzotti D, Lal A, Liu K, et al (2018) De Novo Mutational Signature Discovery in Tumor Genomes using SparseSignatures. *bioRxiv* 384834. <https://doi.org/10.1101/384834>
110. Gu X, Li W-H (1995) The size distribution of insertions and deletions in human and rodent pseudogenes suggests the logarithmic gap penalty for sequence alignment. *J Mol Evol* 40:464-473. <https://doi.org/10.1007/BF00164032>
111. Fan Y, Wang W, Ma G, et al (2007) Patterns of Insertion and Deletion in Mammalian Genomes. *Curr Genomics* 8:370-378. <https://doi.org/10.2174/138920207783406479>
112. Yamane K, Yano K, Kawahara T (2006) Pattern and Rate of Indel Evolution Inferred from Whole Chloroplast Intergenic Regions in Sugarcane, Maize and Rice. *DNA Res* 13:197-204. <https://doi.org/10.1093/dnares/dsl012>
113. Danneels B, Pinto-Carbó M, Carlier A (2018) Patterns of Nucleotide Deletion and Insertion Inferred from Bacterial Pseudogenes. *Genome Biol Evol* 10:1792-1802. <https://doi.org/10.1093/gbe/evy140>
114. Zhu YO, Siegal ML, Hall DW, Petrov DA (2014) Precise estimates of mutation rate and spectrum in yeast. *Proc Natl Acad Sci* 111:E2310. <https://doi.org/10.1073/pnas.1323011111>
115. Farlow A, Long H, Arnoux S, et al (2015) The Spontaneous Mutation Rate in the Fission Yeast *Schizosaccharomyces pombe*. *Genetics* 201:737-744. <https://doi.org/10.1534/genetics.115.177329>
116. Cartwright RA (2009) Problems and Solutions for Estimating Indel Rates and Length Distributions. *Mol Biol Evol* 26:473-480. <https://doi.org/10.1093/molbev/msn275>
117. Jasin M, Rothstein R (2013) Repair of Strand Breaks by Homologous Recombination. *Cold Spring Harb Perspect Biol* 5:. <https://doi.org/10.1101/cshperspect.a012740>
118. Omichessan H, Severi G, Perduca V (2019) Computational tools to detect signatures of mutational processes in DNA from tumours: A review and empirical comparison of performance. *PLoS ONE* 14:. <https://doi.org/10.1371/journal.pone.0221235>

119. Tan VYF, Févotte C (2013) Automatic relevance determination in nonnegative matrix factorization with the β -divergence. *IEEE Trans Pattern Anal Mach Intell* 35:1592–1605. <https://doi.org/10.1109/TPAMI.2012.240>
120. Kasar S, Kim J, Improgo R, et al (2015) Whole-genome sequencing reveals activation-induced cytidine deaminase signatures during indolent chronic lymphocytic leukaemia evolution. *Nat Commun* 6:8866. <https://doi.org/10.1038/ncomms9866>
121. Kim J, Mouw KW, Polak P, et al (2016) Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. *Nat Genet* 48:600–606. <https://doi.org/10.1038/ng.3557>
122. Shiraishi Y, Tremmel G, Miyano S, Stephens M (2015) A Simple Model-Based Approach to Inferring and Visualizing Cancer Mutation Signatures. *PLOS Genet* 11:e1005657. <https://doi.org/10.1371/journal.pgen.1005657>
123. Fischer A, Illingworth CJ, Campbell PJ, Mustonen V (2013) EMu: probabilistic inference of mutational processes and their localization in the cancer genome. *Genome Biol* 14:R39. <https://doi.org/10.1186/gb-2013-14-4-r39>