Analyses of Mutational Patterns Induced by Formaldehyde and Acetaldehyde Reveal Similarity to a Common Mutational Signature

Running Title: Mutational patterns of small aldehydes

Authors:

Mahanish J. Thapa[1,2†], Reena M. Fabros[1,2†], Salma Alasmar[3§], and Kin Chan[1,2*]

Affiliations & addresses:

[1] Department of Biochemistry, Microbiology and Immunology, University of Ottawa, 451 Smyth Rd, Ottawa, Ontario, Canada K1H 8M5

[2] Ottawa Institute of Systems Biology, University of Ottawa, 451 Smyth Rd, Ottawa, Ontario, Canada K1H 8M5

[3] Biopharmaceutical Sciences Undergraduate Program, University of Ottawa, 150 Louis-Pasteur Pvt, Ottawa, Ontario, Canada K1N 6N5

* Corresponding Author: kin.chan@uottawa.ca

† Equal Contribution

Current affiliation & address:

§ Department of Chemistry and Biomolecular Sciences, University of Ottawa, 150 Louis-Pasteur Pvt, Ottawa, Ontario, Canada K1N 6N5

Author email addresses:

M.J.T.: mthap068@uottawa.ca

R.M.F.: fabros.reenamichelle@gmail.com

S.A.: salas051@uottawa.ca

Keywords: mutagenesis, formaldehyde, acetaldehyde, genome instability, mutational pattern, mutational signature

Abstract

Formaldehyde (CH_2O) and acetaldehyde (C_2H_4O) are reactive small molecules produced endogenously in cells as well as being environmental contaminants. Both of these small aldehydes are classified as human carcinogens, since they are known to damage DNA and exposure is linked to cancer incidence. However, the mutagenic properties of formaldehyde and acetaldehyde remain incompletely understood, at least in part because they are relatively weak mutagens. Here, we use a highly sensitive yeast genetic reporter system featuring controlled generation of long single-stranded DNA regions to show that both small aldehydes induced mutational patterns characterized by predominantly C/G → A/T, C/G → T/A, and T/A → C/G substitutions, each in similar proportions. We observed an excess of C/G → A/T transversions when compared to mock-treated controls. Many of these C/G → A/T transversions occurred at T<u>C</u>/<u>G</u>A motifs. Interestingly, the formaldehyde mutational pattern resembles single base substitution (SBS) signature 40 from the Catalog of Somatic Mutations in Cancer (COSMIC). SBS40 is a mutational signature of unknown etiology. We also noted that acetaldehyde treatment caused an excess of deletion events longer than four bases while formaldehyde did not. This latter result could be another distinguishing feature between the mutational patterns of these simple aldehydes. These findings shed new light on the characteristics of two important, commonly occurring mutagens.

Introduction

Genomic DNA is constantly damaged by intracellular processes (1) and exposure to exogenous damaging agents (2–4). There are many different types of DNA damage. Intracellular DNA damaging processes include, for example: oxidation of nitrogenous bases (5,6); glycosidic bond breakage, which releases a nitrogenous base from its deoxyribose sugar (7–11); single- and double-stranded breaks of the sugar-phosphate backbone (11–13); base alkylation (11,14,15); cytosine deamination to uracil (11,16); and deamination of 5-methylcytosine to thymine (17–19). Examples of exogenous DNA damage include: ultraviolet (UV) light (3); ionizing radiation (4); tobacco (20); aristolochic acid (21); and aflatoxin (22). Mutations are also thought to result from spontaneous ionization or isomerization (i.e., tautomerization) of DNA bases, which can alter base pairing characteristics (23–26).

It is important to note that these processes do not affect all bases equally. Each of the four nitrogenous bases has its own distinct set of chemically reactive moieties (e.g, amines, carbonyls, or labile ring atoms) (27). For any given DNA damaging process or agent, the base(s) with moieties that readily react will be damaged more frequently than bases without such reactive moieties. Local sequence context can also be a key determinant of vulnerability to damage. Mutational signatures are recurrent patterns of base changes that reflect these forms of specificity: the signatures arise naturally because each particular mutagenic process or DNA damaging agent is more likely to affect certain bases in specific contexts more frequently than others (28).

Mutational signatures typically are inferred using a non-negative matrix factorization (NMF) algorithm (29). NMF takes a mutational dataset as input. It initiates

by essentially guessing a solution set of constituent signatures with estimated contributions from each putative signature, and then computes the error when attempting to reconstruct the original dataset using that solution set. NMF then tries a slightly different solution set and recomputes the error. This process loops until finding an optimal solution set that stably minimizes reconstruction error. A globally stable solution set is found when different initial conditions all converge to yield that solution set.

NMF analysis can extract reproducible, recurrent patterns of mutations, which often reflect distinct mutagenic processes or DNA damaging agents. There are many mutational signatures with well-established etiologies, including: single base substitution signature 1 (SBS1) from deamination of 5-methylcytosine at CpG motifs; SBS2 and SBS13 from enzymatic deamination of cytosine at TC motifs by APOBEC deaminases; SBS3 from deficiencies in homologous recombination DNA repair; SBS4 and SBS29 from tobacco smoking and chewing habits, respectively; SBS6, SBS15, SBS21, SBS26, and SBS44 from various deficiencies in DNA mismatch repair; SBS7 from ultraviolet light exposure; SBS10 from mutation of DNA polymerase epsilon; SBS18 from reactive oxygen species; SBS30 and SBS36 from DNA base excision repair deficiencies; and so forth (28). About one-third of currently defined mutational signatures remain of unknown etiology (30).

Previously, the International Agency for Research on Cancer (IARC) named a number of high-priority carcinogens that required further research to fill significant gaps in knowledge (31). Among these high-priority carcinogens are two small aldehyde compounds, formaldehyde (CH_2O) and acetaldehyde (C_2H_4O). Formaldehyde is classified as a known human carcinogen by IARC, based in part on evidence of

occupational exposure being associated with nasal and nasopharyngeal cancers (32,33). Formaldehyde is also produced endogenously in cells, as a major metabolic by-product from amino acid metabolism, resulting in high concentrations of up to ~100 μM in human blood (33). Acetaldehyde is a reactive compound that humans are commonly exposed to as a result of ethanol consumption. Like formaldehyde, acetaldehyde is also classified as a known human carcinogen (34). Alcohol consumption (and the ensuing generation of acetaldehyde by oxidation of ethanol) is associated with higher risk of multiple types of cancer, including: head and neck; esophageal; liver; breast; and colorectal (35).

Understanding the mutagenic characteristics of formaldehyde and acetaldehyde remain important research questions, which can provide valuable insights into the possible roles of these common small aldehydes in cancer mutagenesis and carcinogenesis. Previous attempts to define the mutational patterns induced by formaldehyde and acetaldehyde (e.g., (36,37)) have been rather inconclusive, with no demonstrated link to defined mutational signatures in cancers. Here, we report a more detailed understanding of the mutational characteristics of both formaldehyde and acetaldehyde, and show that the mutational pattern induced by formaldehyde is similar to a common cancer mutational signature that is currently of unknown etiology, namely single base substitution signature 40.

Materials and Methods

Reagents and Consumables

Bacto peptone (product code 211677) and yeast extract (212750) were purchased from Becton, Dickinson and Co. (Franklin Lakes, New Jersey). Canavanine (C9758), adenine sulfate dihydrate (AD0028), formaldehyde (F8775), and acetaldehyde (W200344) were purchased from MilliporeSigma (St. Louis, Missouri). Formaldehyde and acetaldehyde solutions were stored in gas-tight tubes in the dark under nitrogen atmosphere. Agar (FB0010), glucose (GB0219), hygromycin (BS725), PCR purification spin column kit (BS654), agarose (D0012), and Tris-Borate-EDTA (TBE) buffer (A0026) were purchased from BioBasic (Markham, Ontario). G418 sulfate (450-130) was purchased from Wisent (St-Bruno, Québec). Q5 PCR kits were purchased from New England Biolabs Canada (Whitby, Ontario). Gas-tight glass tubes with septa (2048-18150) and accessories (2048-11020 and 2048-10020) were purchased from Bellco Glass Inc. (Vineland, New Jersey).

Yeast Genetics and Mutagenesis

Mutagenesis experiments used the ySR127 yeast strain, a *MAT*α haploid bearing the *cdc13-1* temperature sensitive allele. In addition, ySR127 has a cassette of three reporter genes (*CAN1*, *URA3*, and *ADE2*) near the de novo left telomere of chromosome V. These three genes had been deleted from their native loci. Details about ySR127 were described previously (38) and the strain is available upon request.

Formaldehyde mutagenesis experiments were initiated by inoculating single colonies separately into 5 mL of YPDA rich media (2% Bacto peptone, 1% Bacto yeast extract, 2% glucose, supplemented with 0.001% adenine sulfate) in round bottom glass tubes. Cells were grown at permissive temperature (23°C) for three days. Then, cultures were diluted ten-fold into fresh media in gas-tight glass tubes, shifted to restrictive temperature (37°C), and shaken gently at 150 RPM for three hours, with syringe needles inserted through the septa to enable gas exchange. After a three-hour temperature shift, aliquots of formaldehyde stock solution diluted in media were injected into each tube to obtain the reported final concentrations. Samples were then shaken gently at 150 RPM at 37°C for three more hours, in completely sealed gas-tight tubes, to prevent escape of formaldehyde. When formaldehyde treatment was complete, cells were collected by syringe, lightly centrifuged, washed in water, and plated (using a turntable and cell spreader) onto synthetic complete media to assess survival and onto canavanine-containing media with 0.33x adenine to select for mutants (Can^r colonies were off-white while Can^r Ade^- colonies turn red or pink). Care was taken to handle cells gently throughout, as they were quite fragile. Further details of this plating procedure were described in detail previously (39).

Acetaldehyde mutagenesis experiments were carried out similarly. We found that we could simplify the acetaldehyde experiments by using tightly sealed 50 mL polypropylene tubes for the temperature shift and mutagen treatment, presumably because acetaldehyde is less volatile than formaldehyde and does not require as fastidious gas-tight containment. Similar results were obtained for acetaldehyde treatment when

using either type of tubes. Statistical analyses and data visualizations were done using base R version 4.1 (40) and tidyverse package version 1.3 (41).

Illumina Whole Genome Sequencing and Data Analyses

Can^r Ade^- mutants from formaldehyde and acetaldehyde treatment experiments were collected and reporter gene loss of function phenotypes were verified as described previously (38). Briefly, Can^r red/pink mutants were streaked on YPDA plates. A single colony from each streak was patched onto YPDA. Patches were then replica plated onto glycerol, adenine dropout, canavanine, and uracil dropout media. Mutants that grew on glycerol (i.e., were respiration competent), and were Can^r Ade^- Ura^+ were considered suitable for sequencing. Can^r Ade^- Ura^- mutants were avoided because those isolates sometimes turn out to be telomere truncations. Mutants from 4, 6, and 8 mM formaldehyde exposure were chosen for sequencing as these had similar induced mutation frequencies. For acetaldehyde, mutants from 75 mM treatment were selected for sequencing, as this concentration was most mutagenic.

Illumina library preparation and WGS were outsourced to Genome Québec (McGill University, Montréal) or performed on an Illumina MiSeq in our lab. Sequencing reads were uploaded to the NCBI SRA (National Center for Biotechnology Information Sequence Read Archive), accessions PRJNA839792 and PRJNA574140. Bowtie2 version 2.3.5.1 (42), SAMtools 1.9 (43), and bcftools 1.9 (44) were used to map the Illumina reads and call variants. Variants with quality score < 30 and/or with sequencing coverage < 10 were filtered out. The resulting VCF files were passed to MutationalPatterns version 3.6.3 (45) for further analysis and visualization. Other

numerical and statistical analyses, and data visualizations were done using base R version 4.1 (40) and tidyverse package version 1.3 (41).

For trinucleotide frequency correction, the Biostrings package version 2.38.0 (46) was used to extract trinucleotide counts for the ySR127 yeast and mm10 mouse reference genomes. Following the convention for reporting mutational signatures, counts for each trinucleotide motif centered on C or T were summed with the counts of their respective reverse complements. The proportion of each trinucleotide was then calculated. To infer the expected pattern in mouse, the frequency of each of the 96 channels of a yeast mutational pattern was multiplied by the ratio of corresponding trinucleotide proportions in mouse vs. in yeast. For example, if a given trinucleotide motif is half as abundant in mouse as in yeast, the corresponding expected frequency of mutations in mouse would be scaled by a factor of 0.5 relative to the observed frequency in yeast data.

Results

Formaldehyde- and acetaldehyde-induced mutagenesis

We began by assessing mutagenesis and toxicity induced by addition of formaldehyde or acetaldehyde. These experiments were done using a haploid yeast strain (ySR127) that forms long regions of sub-telomeric single-stranded DNA (ssDNA) when shifted to 37°C due to the *cdc13-1* temperature sensitive point mutation (47). At 37°C the cdc13-1 protein dissociates from telomeres, triggering enzymatic resection of unprotected chromosome ends, which in turn activates the DNA damage checkpoint to arrest cells in G_2 (47). The reporter genes *CAN1*, *ADE2*, and *URA3* had been deleted from their native loci and reintroduced to the left sub-telomeric region of chromosome V (38). This mutagenesis system is very well suited to studying weak mutagens, as ssDNA is more prone to mutation than double-stranded DNA and repair using the complementary strand is not possible. This latter point is an important consideration, since DNA lesions induced by formaldehyde in duplex DNA are potential substrates for nucleotide excision repair (48). The ssDNA system was used previously to study the mutagenic properties of bisulfite and human APOBEC3G cytidine deaminase (38); abasic sites (49); reactive oxygen species (50); human APOBEC3A and APOBEC3B cytidine deaminases (51); and alkylating agents (52).

We treated temperature-shifted cells with increasing concentrations of formaldehyde or acetaldehyde. Care was taken to seal the formaldehyde-treated samples in gas-tight tubes; otherwise, the formaldehyde would simply volatilize into the gaseous phase and escape into the atmosphere. Increasing concentrations of formaldehyde

resulted in lower viability (see Figure 1a). While lower concentrations are relatively well tolerated, 8 mM formaldehyde reduced viability below 50%. Formaldehyde-induced inactivation of *CAN1* was detected from as little as 2 mM treatment (median gene inactivation frequency of 3.3×10^{-4}, see Figure 1b). Mutagenesis plateaued from 4 to 8 mM formaldehyde, with median mutation frequencies of ~1.5×10^{-3}. Mock treated cells (i.e., 0 mM formaldehyde) had median mutation frequency of only 1.2×10^{-4}. These results show that when the experiments are set up properly to contain the mutagen, formaldehyde is clearly mutagenic to our ssDNA model system.

Cells were considerably more tolerant of higher concentrations of acetaldehyde. We tested concentrations from 25 to 100 mM. Cells treated with lower concentrations (25 and 50 mM) retained high viability, but higher concentrations induced significant lethality (see Figure 1c). Unlike formaldehyde, the mutagenesis induced by acetaldehyde did not show a plateau. Instead, here was a gradual increase in *CAN1* inactivation frequency when treated with 25 and 50 mM acetaldehyde (see Figure 1d). Mutation frequency peaked at over 5×10^{-4} when cells were treated with 75 mM acetaldehyde. Interestingly, treatment with 100 mM acetaldehyde did not result in detectable mutagenesis while viability was reduced to below 25%. This suggests that the cells which sustained high levels of DNA damage by 100 mM acetaldehyde likely suffered considerable cytotoxic damage as well and did not survive.

<u>Formaldehyde and acetaldehyde both induce an excess of C/G > A/T transversions</u>

We collected mutagenized isolates for Illumina whole genome sequencing to determine what kinds of genetic variants were induced by either formaldehyde or

acetaldehyde treatment. The genomes mutagenized by either small aldehyde were compared to control genomes that were not treated by either. Analysis of the untreated control genomes revealed a mutational pattern where C/G > T/A and T/A > C/G transitions outnumbered the four types of transversions (namely C/G > A/T, C/G > G/C, T/A > A/T, and T/A > G/C, see Figure 2a), similar to what we had observed previously (53). By comparison, formaldehyde and acetaldehyde treatment both caused a relative increase of C/G > A/T transversions (see Figure 2b and 2c). While these substitutions accounted for 11% of the mutational spectrum in untreated controls, this fraction rose to about 17% in the aldehyde-mutagenized genomes. This increase is a common characteristic of mutagenesis caused by small aldehydes in regions of single-stranded DNA.

Acetaldehyde induces deletions of five or more bases, but formaldehyde does not

We also analyzed short insertions and deletions (indels) to determine if treatment with either small aldehyde can induce these genetic changes. The profile of short indels in untreated controls consists mainly of insertions of five or more bases, with smaller proportions of shorter insertions as well as deletions of five or more bases (see Figure 3a). The profile of short indels in formaldehyde-mutated genomes is essentially the same as in untreated control genomes, i.e., we did not find evidence that formaldehyde induces a higher proportion of any type of indels (see Figure 3b). In contrast, there was a notable difference in the acetaldehyde-induced profile of indels: an excess of deletions of five or more bases was observed (24% in acetaldehyde vs. 12% in controls, see Figure

3c). This is a distinguishing property of acetaldehyde-induced DNA damage in the ssDNA system.

Formaldehyde and acetaldehyde produce distinct mutational patterns

To investigate the mutational properties of the small aldehydes in more detail, we plotted their mutational profiles in the 96-channel format of the COSMIC mutational signatures. By this convention, all substitutions are reported as originating from a pyrimidine base, i.e., same as the mutation spectra reported above. In addition, the 96-channel profiling features trinucleotide motifs consisting of the mutated base, flanked by an adjacent base 5′ and 3′. Cosine similarity is a metric for comparing mutational patterns, yielding a maximum value of exactly 1 for two identical patterns (29). The mutational pattern of formaldehyde is similar to untreated controls (cosine similarity = 0.93), but the excess of C/G > A/T transversions is nonetheless evident (see Figure 4a). The mutational pattern of acetaldehyde is more dissimilar vs. the profile of untreated controls (cosine similarity = 0.868), but again with a noticeable excess of C/G > A/T substitutions (see Figure 4b). When comparing the formaldehyde and acetaldehyde profiles directly to one another, the cosine similarity value is 0.882, showing some similarities but also clear differences in the C/G > T/A channels especially (see Figure 4c).

A recent high profile paper described mutational patterns obtained in mice that were genetically deleted for genes important in aldehyde detoxification, ADH5 and ALDH2, thus leading to buildup of endogenous aldehydes (37). To compare our mutational patterns derived from mutagenized yeast genomes to these profiles from mice,

we first adjusted for differences in trinucleotide abundances between the two species to obtained corrected mutational patterns (see Figure 4d and 4e). A main difference between the yeast and mouse genomes is the lower abundance of CpG motifs in the latter. Nonetheless, the corrected mutational patterns retained high similarity to the original (uncorrected) patterns in yeast (cosine similarity values > 0.95).

When we compared the various mutational patterns, we noticed that cosine similarity values are relatively low when comparing between the corrected yeast patterns we derived and the mouse patterns from Dingler et al. (37) (see Table 1). These values are somewhat higher when comparing the formaldehyde pattern in yeast to the mouse patterns. A closer examination of these profiles suggests that there are likely to be mutations from other sources mixed in the mouse patterns, e.g. from SBS1 (deamination of 5-methylcytosine at CpG motifs, see Figure 4f). We also noted some differences among the various mouse patterns themselves: while the ones from Adh5$^{-/-}$ and Aldh2$^{-/-}$ Adh5$^{-/-}$ had cosine similarity = 0.887, the Aldh2$^{-/-}$ profile was noticeably more dissimilar (cosine similarity < 0.8 vs. the other two profiles). This is consistent with the likelihood that there are other mutation sources mixed in with the mouse mutational patterns, which may be confounding interpretation of a hypothesized pattern induced by excess endogenous aldehydes.

Formaldehyde mutational pattern resembles COSMIC SBS signature 40

We then investigated whether these mutational patterns might shed light on the etiology of any known COSMIC mutational signatures. We started with the mouse profiles published by Dingler et al. (37) and confirmed that none of the mouse profiles

showed a particularly close resemblance to any known COSMIC signature (see Figure 5a). All of those cosine similarity values were < 0.8, suggesting that if there are bona fide COSMIC signatures within the mouse mutational patterns, they are possibly obscured by being in a mixture of multiple signatures.

Comparison of the corrected acetaldehyde pattern from yeast vs. known COSMIC signatures also yielded, at best, cosine similarity of 0.79 to SBS40, a signature of unknown etiology (see Figure 5b). Since we had better direct control of the induced mutagenesis experiments using the yeast system with exogenously applied mutagen, it does not seem as likely that other mutagenic processes are obscuring the acetaldehyde-induced pattern. We conclude that the acetaldehyde pattern we obtained is not a plausible match for any known COSMIC signature at this point.

Finally, we compared the formaldehyde pattern to the COSMIC signatures, finding that the closest match is to SBS40, with cosine similarity = 0.9 (see Figures 5c and 5d). The second closest match was to SBS5 (cosine similarity = 0.864, see Figures 5c and 5e). We previously studied an SBS5-like mutational pattern in yeast and showed that similar patterns are widely conserved in many species. Moreover, the SBS5-like pattern is due to error-prone translesion DNA synthesis in the absence of added mutagens and increases with increasing sugar metabolism (53). An SBS40-like mutational pattern would require a separate explanation, which would be the addition of exogenous formaldehyde to our experimental system. As such, we propose that a plausible etiology for SBS40 in cancers is the mutagenicity of formaldehyde.

Discussion

In this paper, we report the use of a sensitive ssDNA-based mutagenesis reporter system to characterize the mutagenic properties of two small aldehydes, formaldehyde and acetaldehyde. This system is especially well suited for investigating chemical agents with relatively weak mutagenicity. A challenge of using conventional mutagenesis systems to study weak mutagens is that induced mutations can be rare and it can be difficult to discern a reliable mutational pattern using relatively few mutations (36,37). In addition to being a more sensitive reporter system, it is also considerably more cost-efficient to sequence compact yeast genomes (each ~12 megabases) than mammalian genomes which are much larger (~3 gigabases). By applying a correction to account for different abundances of trinucleotide motifs, we can use data from the sequencing of mutagenized yeast to infer the expected mutational pattern in another species. Another key advantage is the single-stranded configuration of the DNA precludes repair that requires a complementary strand. By sidestepping intervention from DNA repair processes, the ssDNA system can provide, in effect, a purer readout of the effects of mutagenesis per se. Leveraging these advantages of the ssDNA mutagenesis reporter system, we were able to infer the mutational patterns of both formaldehyde and acetaldehyde.

The two small aldehydes share some similar mutagenicity characteristics, but also have their differences. Both induce dose-dependent increases in mutagenesis at lower concentrations. But whereas formaldehyde-induced mutagenesis essentially plateaus from 4 mM up to 8 mM, acetaldehyde-induced mutagenesis peaks at 75 mM and then drops sharply at the even higher concentration of 100 mM. Both aldehydes induce

significant cytotoxicity at the higher end of their respective ranges of tested concentrations, but yeast are able to tolerate considerably higher doses of acetaldehyde overall. Yeast are presumably evolved to cope with significantly higher concentrations of acetaldehyde, since it is an abundant intermediate in ethanol production from fermentation (54). Both aldehydes cause an excess of C/G > A/T transversions, which is consistent with previous reports showing preferential adduct formation and mutagenesis at guanines (55–61). Interestingly, acetaldehyde induces an excess of deletion variants of five or more bases in our system, but formaldehyde does not, consistent with previous reports (55,62). These various mutagenic characteristics of formaldehyde and acetaldehyde reflect their chemical similarities and differences.

The 96-channel mutational patterns of formaldehyde and acetaldehyde revealed further differences between the two compounds. Whereas the acetaldehyde pattern did not particularly resemble any known COSMIC signature, new mutational signatures will be revealed as more cancer samples are sequenced and analyzed. Since alcohol consumption is associated with multiple cancer types and it is thought that the acetaldehyde from alcohol detoxification would surely damage DNA (63), associated mutational signature(s) may yet be discovered in the future. The formaldehyde pattern we obtained was similar to SBS signature 40. SBS40 is currently of unknown etiology, but it is known to be present in at least 28 cancer types (28), making SBS40 the third most common mutational signature in cancers. The high prevalence of SBS40 hints at an endogenous origin for the underlying DNA damage that is present in different cell types throughout the body. Since formaldehyde is produced endogenously and exists at steady state concentrations in humans in the range of tens of micromolar (33), it would fit this

profile. When all of the available information is taken into consideration, mutagenesis from endogenously generated formaldehyde emerges as a plausible candidate for the etiology of SBS40.

Comparison with mutational patterns from mice deleted for aldehyde detoxification genes suggest that those profiles are likely mixtures of mutations from different mutagenic processes, and not just from DNA damage due to accumulation of excess endogenous aldehydes. For example, the contribution from SBS1 (C/G > T/A at CpG motifs) was quite noticeable. Mutagenesis from other sources likely interferes with making an accurate inference of the aldehyde-associated mutagenesis. This is perhaps another significant challenge when using systems for mutational detection that are not (and maybe can not) be properly controlled to factor out mutagenesis from other sources. Deployment of more specialized and sensitive mutagenesis detection systems where the experimenters have more direct control over the mutation induction can continue to play an important role in shining new light on mutagenesis.

Acknowledgments

We thank S. Gelova, B. Xhialli, and N. Liang for their critical feedback on this work. The authors gratefully acknowledge support from a Tier 2 Canada Research Chair, an NSERC Discovery Grant, an Ontario Early Researcher Award, and uOttawa startup funding to K.C. Author contributions are as follows: All authors performed experiments, and analyzed/discussed the data; K.C. conceived and supervised the study, and wrote the draft manuscript. The authors declare there are no competing interests.

References

1. De Bont, R. & van Larebeke, N. Endogenous DNA damage in humans: a review of quantitative data. *Mutagenesis* **19,** 169–185 (2004).
2. Irigaray, P. & Belpomme, D. Basic properties and molecular mechanisms of exogenous chemical carcinogens. *Carcinogenesis* **31,** 135–148 (2010).
3. Ikehata, H. & Ono, T. The Mechanisms of UV Mutagenesis. *J. Radiat. Res. (Tokyo)* **52,** 115–125 (2011).
4. Keszenman, D. J., Kolodiuk, L. & Baulch, J. E. DNA damage in cells exhibiting radiation-induced genomic instability. *Mutagenesis* **30,** 451–458 (2015).
5. Ames, B. N., Shigenaga, M. K. & Hagen, T. M. Oxidants, antioxidants, and the degenerative diseases of aging. *Proc. Natl. Acad. Sci.* **90,** 7915 (1993).
6. Helbock, H. J. *et al.* DNA oxidation matters: the HPLC-electrochemical detection assay of 8-oxo-deoxyguanosine and 8-oxo-guanine. *Proc. Natl. Acad. Sci. U. S. A.* **95,** 288–293 (1998).
7. Lindahl, T. & Nyberg, B. Rate of depurination of native deoxyribonucleic acid. *Biochemistry* **11,** 3610–3618 (1972).
8. Lindahl, T. Instability and decay of the primary structure of DNA. *Nature* **362,** 709–715 (1993).
9. Nakamura, J. *et al.* Highly Sensitive Apurinic/Apyrimidinic Site Assay Can Detect Spontaneous and Chemically Induced Depurination under Physiological Conditions. *Cancer Res.* **58,** 222 (1998).
10. Lindahl, T. in *DNA Repair Process.* (Symposia Specialists, 1977).
11. Tice, R. R. & Setlow, R. B. in *Handb. Biol. Aging* (Van Nostrand Reinhold, 1985).
12. Haber, J. E. DNA recombination: the replication connection. *Trends Biochem. Sci.* **24,** 271–275 (1999).
13. Vilenchik, M. M. & Knudson, A. G. Endogenous DNA double-strand breaks: Production, fidelity of repair, and induction of cancer. *Proc. Natl. Acad. Sci.* **100,** 12871 (2003).
14. Kadlubar, F. F. *et al.* Comparison of DNA adduct levels associated with oxidative stress in human pancreas. *Mutat. Res. Mol. Mech. Mutagen.* **405,** 125–133 (1998).
15. VanderVeen, L. A., Hashim, M. F., Shyr, Y. & Marnett, L. J. Induction of frameshift and base pair substitution mutations by the major DNA adduct of the endogenous carcinogen malondialdehyde. *Proc. Natl. Acad. Sci.* **100,** 14247 (2003).
16. Saparbaev, M. K. & Zharkov, D. O. in *Ref. Module Life Sci.* (Elsevier, 2017). doi:10.1016/B978-0-12-809633-8.06481-5
17. Greenblatt, M. S., Bennett, W. P., Hollstein, M. & Harris, C. C. Mutations in the p53 tumor suppressor gene: clues to cancer etiology and molecular pathogenesis. *Cancer Res.* **54,** 4855–4878 (1994).
18. Neddermann, P. *et al.* Cloning and Expression of Human G/T Mismatch-specific Thymine-DNA Glycosylase. *J. Biol. Chem.* **271,** 12767–12774 (1996).
19. Sassa, A., Kanemaru, Y., Kamoshita, N., Honma, M. & Yasui, M. Mutagenic consequences of cytosine alterations site-specifically embedded in the human genome. *Genes Environ.* **38,** 17 (2016).
20. Alexandrov, L. B. *et al.* Mutational signatures associated with tobacco smoking in

human cancer. *Science* **354,** 618 (2016).
21. Moriya, M. *et al.* TP53 Mutational signature for aristolochic acid: an environmental carcinogen. *Int. J. Cancer* **129,** 1532–1536 (2011).
22. Letouzé, E. *et al.* Mutational signatures reveal the dynamic interplay of risk factors and cellular processes during liver tumorigenesis. *Nat. Commun.* **8,** 1315 (2017).
23. Russo, N., Toscano, M., Grand, A. & Jolibois, F. Protonation of thymine, cytosine, adenine, and guanine DNA nucleic acid bases: Theoretical investigation into the framework of density functional theory. *J. Comput. Chem.* **19,** 989–1000 (1998).
24. Podolyan, Y., Gorb, L. & Leszczynski, J. Protonation of Nucleic Acid Bases. A Comprehensive Post-Hartree−Fock Study of the Energetics and Proton Affinities. *J. Phys. Chem. A* **104,** 7346–7352 (2000).
25. Masoodi, H. R., Bagheri, S. & Abareghi, M. The effects of tautomerization and protonation on the adenine–cytosine mismatches: a density functional theory study. *J. Biomol. Struct. Dyn.* **34,** 1143–1155 (2016).
26. Kimsey, I. J. *et al.* Dynamic basis for dG•dT misincorporation via tautomerization and ionization. *Nature* **554,** 195 (2018).
27. Alberts, B. *et al.* in *Mol. Biol. Cell* (W.W. Norton & Co., 2014).
28. Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578,** 94–101 (2020).
29. Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J. & Stratton, M. R. Deciphering Signatures of Mutational Processes Operative in Human Cancer. *Cell Rep.* **3,** 246–259 (2013).
30. COSMIC | SBS - Mutational Signatures. at <https://cancer.sanger.ac.uk/signatures/sbs/>
31. International Agency for Research on Cancer. *Identification of research needs to resolve the carcinogenicity of high-priority IARC carcinogens*. **42,** (2010).
32. International Agency for Research on Cancer. in *Chem. Agents Relat. Occup.* **100F,** 401–436 (2012).
33. National Toxicology Program. in *Rep. Carcinog. Thirteen. Ed.* (U.S. Department of Health and Human Services, Public Health Service, 2014).
34. Secretan, B. *et al.* A review of human carcinogens--Part E: tobacco, areca nut, alcohol, coal smoke, and salted fish. *Lancet Oncol.* **10,** 1033–1034 (2009).
35. Alcohol and Cancer Risk Fact Sheet - National Cancer Institute. (2021). at <https://www.cancer.gov/about-cancer/causes-prevention/risk/alcohol/alcohol-fact-sheet>
36. Kucab, J. E. *et al.* A Compendium of Mutational Signatures of Environmental Agents. *Cell* (2019). doi:10.1016/j.cell.2019.03.001
37. Dingler, F. A. *et al.* Two Aldehyde Clearance Systems Are Essential to Prevent Lethal Formaldehyde Accumulation in Mice and Humans. *Mol. Cell* **80,** 996-1012.e9 (2020).
38. Chan, K. *et al.* Base Damage within Single-Strand DNA Underlies In Vivo Hypermutability Induced by a Ubiquitous Environmental Agent. *PLoS Genet* **8,** e1003149 (2012).
39. Chan, K. in *Genome Instab. Methods Protoc.* (eds. Muzi-Falconi, M. & Brown, G. W.) 33–42 (Springer New York, 2018). doi:10.1007/978-1-4939-7306-4_4
40. R Core Team. R: The R Project for Statistical Computing. at <https://www.r-

project.org/>
41. Wickham, H. *et al.* Welcome to the tidyverse. *J Open Source Softw* **4,** 1686 (2019).
42. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Meth* **9,** 357–359 (2012).
43. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
44. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27,** 2987–2993 (2011).
45. Blokzijl, F., Janssen, R., van Boxtel, R. & Cuppen, E. MutationalPatterns: comprehensive genome-wide analysis of mutational processes. *Genome Med.* **10,** 33 (2018).
46. Pagès, H., Aboyoun, P., Gentleman, R. & DebRoy, S. *Biostrings: Efficient manipulation of biological strings*. at <https://bioconductor.org/packages/Biostrings>
47. Garvik, B., Carson, M. & Hartwell, L. Single-stranded DNA arising at telomeres in cdc13 mutants may constitute a specific signal for the RAD9 checkpoint. *Mol. Cell. Biol.* **15,** 6128–6138 (1995).
48. Grogan, D. & Jinks-Robertson, S. Formaldehyde-induced mutagenesis in Saccharomyces cerevisiae: Molecular properties and the roles of repair and bypass systems. *Mutat. Res. Mol. Mech. Mutagen.* **731,** 92–98 (2012).
49. Chan, K., Resnick, M. A. & Gordenin, D. A. The choice of nucleotide inserted opposite abasic sites formed within chromosomal DNA reveals the polymerase activities participating in translesion DNA synthesis. *DNA Repair* **12,** 878–889 (2013).
50. Degtyareva, N. P. *et al.* Oxidative stress-induced mutagenesis in single-strand DNA occurs primarily at cytosines and is DNA polymerase zeta-dependent only for adenines and guanines. *Nucleic Acids Res.* **41,** 8995–9005 (2013).
51. Chan, K. *et al.* An APOBEC3A hypermutation signature is distinguishable from the signature of background mutagenesis by APOBEC3B in human cancers. *Nat Genet* **47,** 1067–1072 (2015).
52. Saini, N. *et al.* Mutation signatures specific to DNA alkylating agents in yeast and cancers. *Nucleic Acids Res.* (2020). doi:10.1093/nar/gkaa150
53. Gelova, S. P., Doherty, K. N., Alasmar, S. & Chan, K. Intrinsic base substitution patterns in diverse species reveal links to cancer and metabolism. *bioRxiv* 758540 (2020). doi:10.1101/758540
54. Matsufuji, Y. *et al.* Acetaldehyde tolerance in Saccharomyces cerevisiae involves the pentose phosphate pathway and oleic acid biosynthesis. *Yeast* **25,** 825–833 (2008).
55. Yasui, M. *et al.* Translesional synthesis on a DNA template containing N(2)-methyl-2′-deoxyguanosine catalyzed by the Klenow fragment of Escherichia coli DNA polymerase I. *Nucleic Acids Res.* **29,** 1994–2001 (2001).
56. Crosby, R. M. *et al.* Molecular analysis of formaldehyde-induced mutations in human lymphoblasts and e. coli. *Environ. Mutagen.* **12,** 155–166 (1988).
57. Ohta, T., Watanabe-Akanuma, M., Tokishita, S. & Yamagata, H. Mutation spectra of chemical mutagens determined by Lac+ reversion assay with Escherichia coli WP3101P–WP3106P tester strains. *Mutat. Res. Toxicol. Environ. Mutagen.* **440,** 59–74 (1999).

58. Liu, X., Lao, Y., Yang, I.-Y., Hecht, S. S. & Moriya, M. Replication-Coupled Repair of Crotonaldehyde/Acetaldehyde-Induced Guanine−Guanine Interstrand Cross-Links and Their Mutagenicity. *Biochemistry* **45,** 12898–12905 (2006).
59. Upton, D. C. *et al.* Replication of N2-Ethyldeoxyguanosine DNA Adducts in the Human Embryonic Kidney Cell Line 293. *Chem. Res. Toxicol.* **19,** 960–967 (2006).
60. Stein, S., Lao, Y., Yang, I.-Y., Hecht, S. S. & Moriya, M. Genotoxicity of acetaldehyde- and crotonaldehyde-induced 1,N2-propanodeoxyguanosine DNA adducts in human cells. *Mutat. Res. Toxicol. Environ. Mutagen.* **608,** 1–7 (2006).
61. Upton, D. C. *et al.* Mutagenesis by exocyclic alkylamino purine adducts in Escherichia coli. *Mutat. Res. Mol. Mech. Mutagen.* **599,** 1–10 (2006).
62. Garaycoechea, J. I. *et al.* Alcohol and endogenous aldehydes damage chromosomes and mutate stem cells. *Nature* (2018). at <http://dx.doi.org/10.1038/nature25154>
63. International Agency for Research on Cancer. *Personal Habits and Indoor Combustions*. **100E,** (2012).

1 **Figures**

2

3 **Figure 1:** (a) Viability and (b) *CAN1* inactivation frequency of yeast treated with 0, 2, 4,

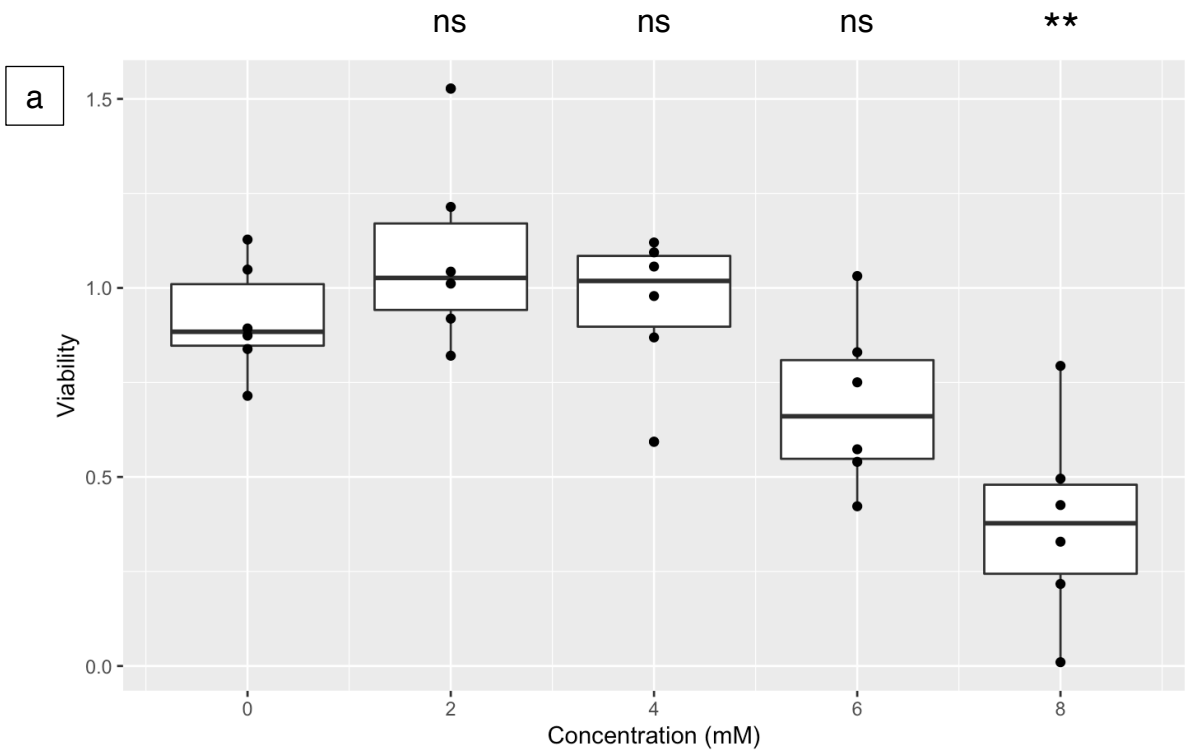
4 6, or 8 mM formaldehyde. (c) Viability and (d) *CAN1* inactivation frequency of yeast

5 treated with 0, 25, 50, 75, or 100 mM acetaldehyde. Data are from six biological

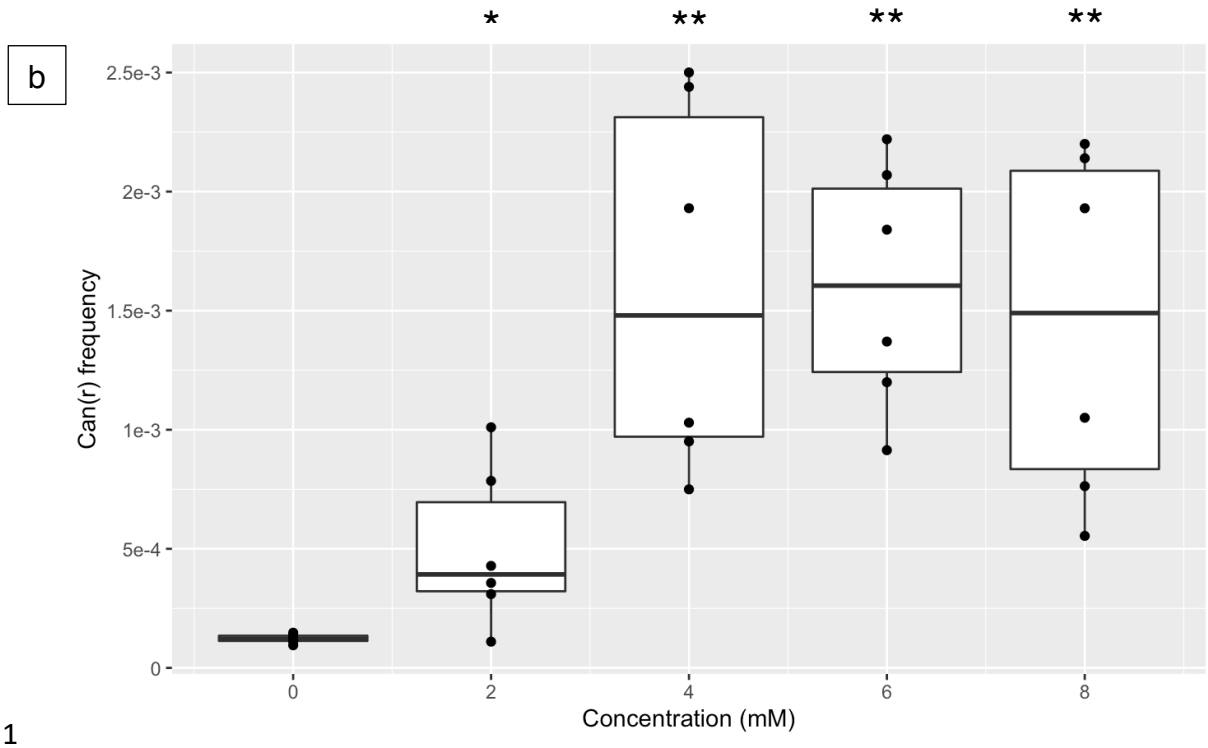
6 replicates for each aldehyde. * denotes P < 0.05, ** denotes P < 0.01, *** denotes P <

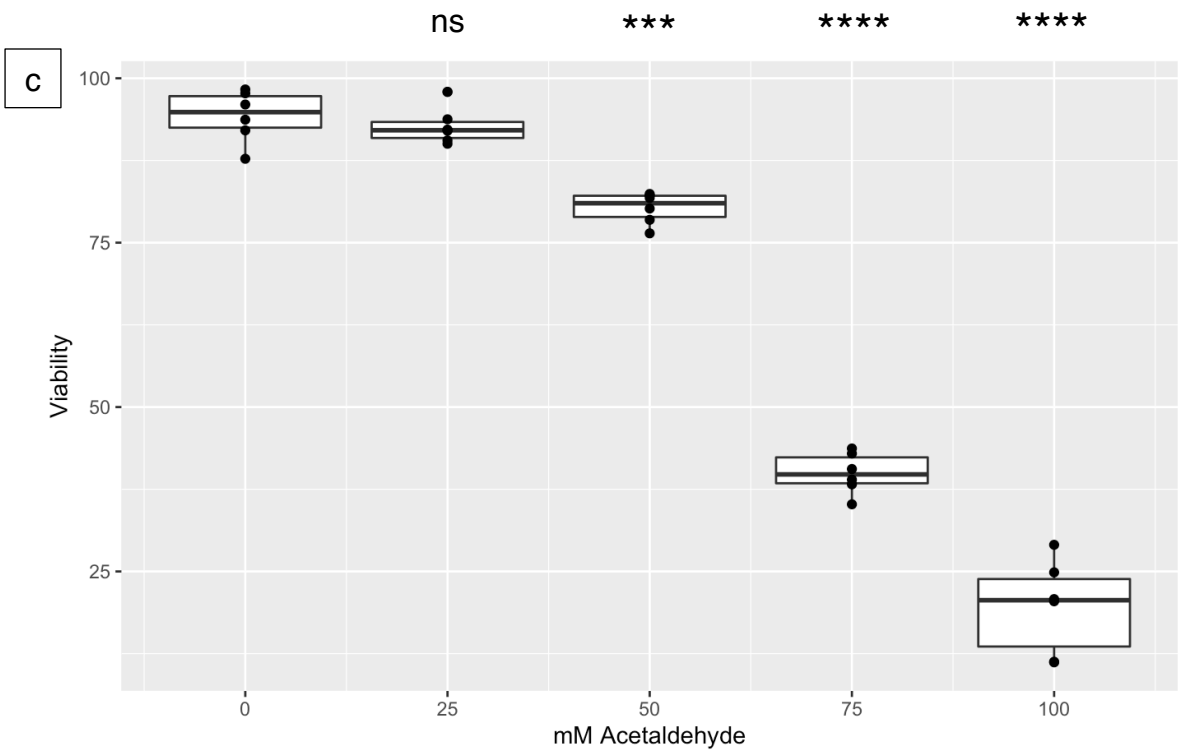
7 0.001, **** denotes P < 0.0001, and ns denotes no significant difference by paired t-test.

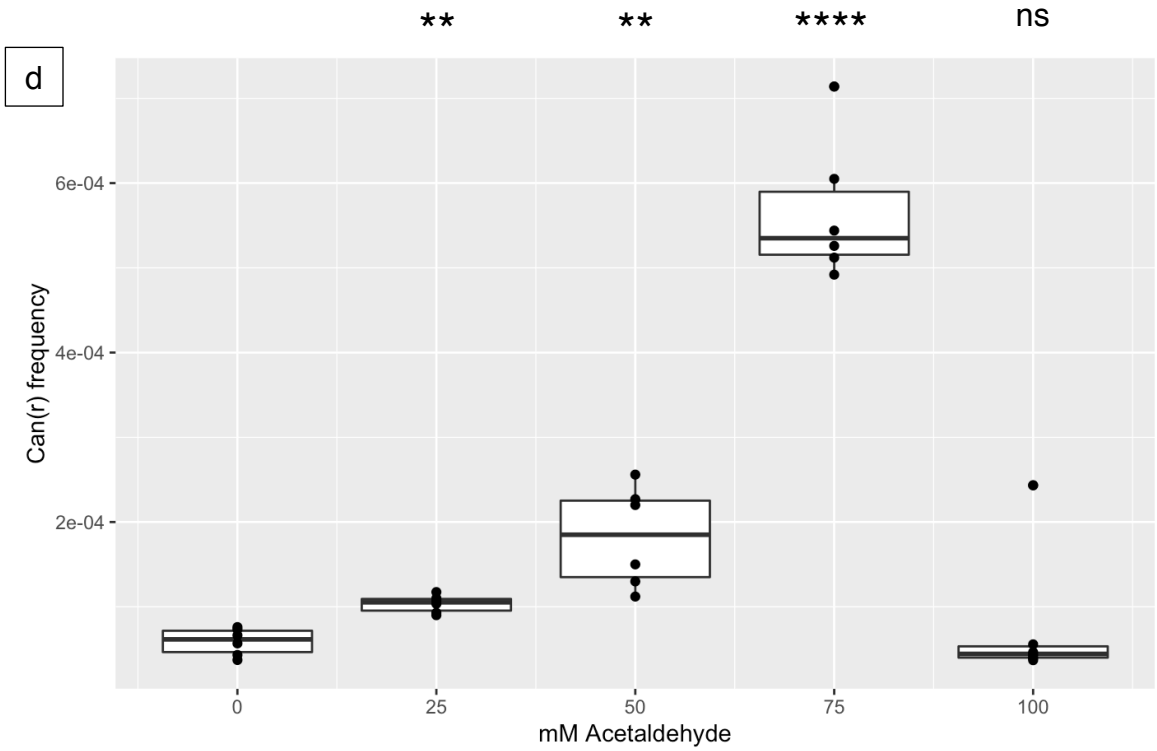
8



9


b
*
**
**
**
2.5e-3
2e-3
1.5e-3
1e-3
5e-4
0
Can(r) frequency
0
2
4
6
8
Concentration (mM)

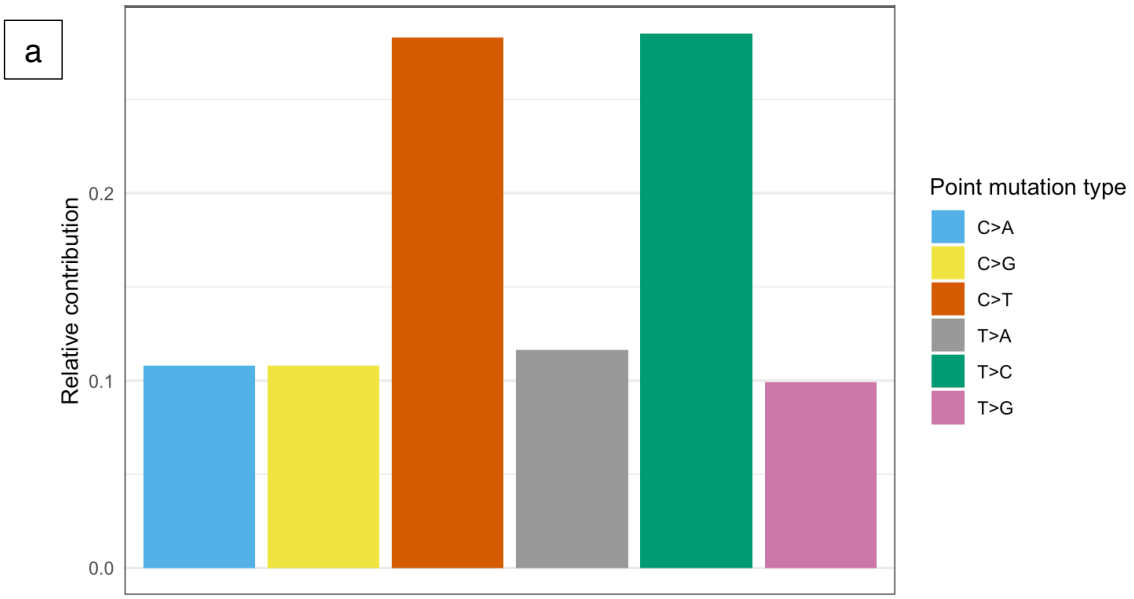

ns
***
****
****
c
100
75
50
25
Viability
0
25
50
75
100
mM Acetaldehyde


**
**
****
ns
d
6e-04
4e-04
2e-04
Can(r) frequency
0
25
50
75
100
mM Acetaldehyde

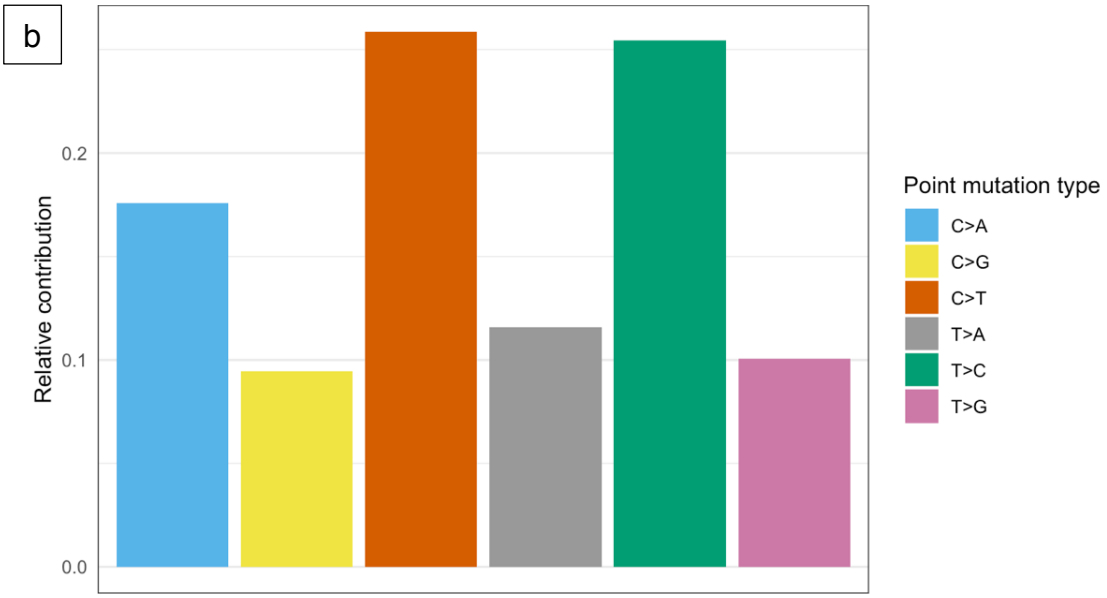
1 **Figure 2:** Base substitution types for (a) untreated controls, (b) formaldehyde, and (c)

2 acetaldehyde. Treatment with either aldehyde caused a higher proportion of C/G > A/T

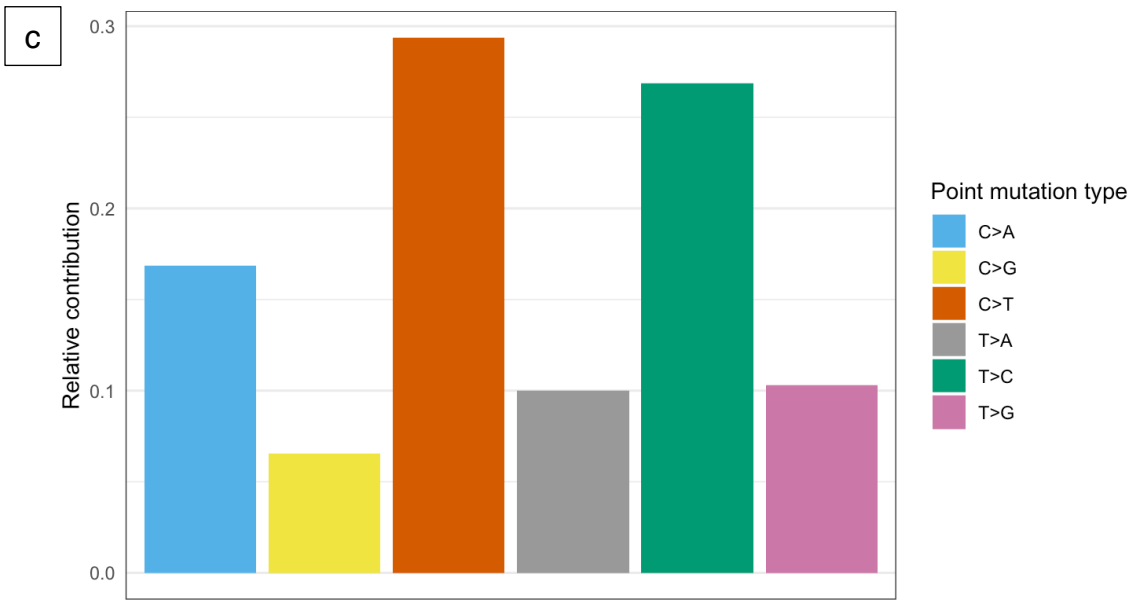
3 transversions.



4



5


c
0.3
0.2
0.1
0.0
Relative contribution
Point mutation type
C>A
C>G
C>T
T>A
T>C
T>G

1

2

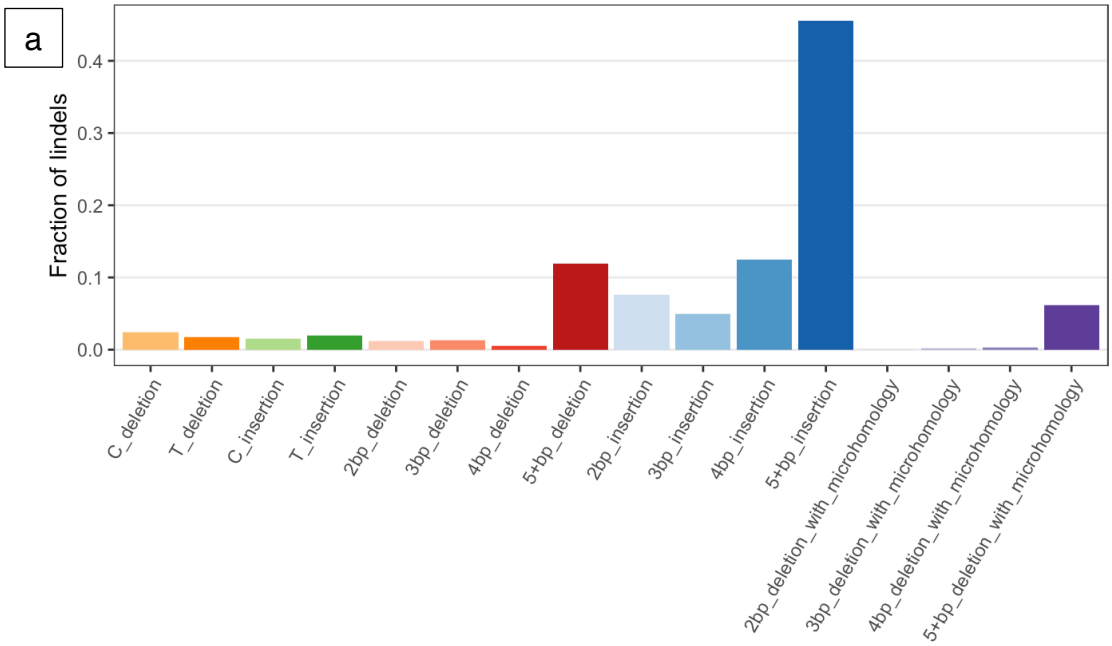
1 **Figure 3:** Small indels from (a) untreated controls, (b) formaldehyde, and (c)

2 acetaldehyde. The different categories comprise: single base deletions or insertions at

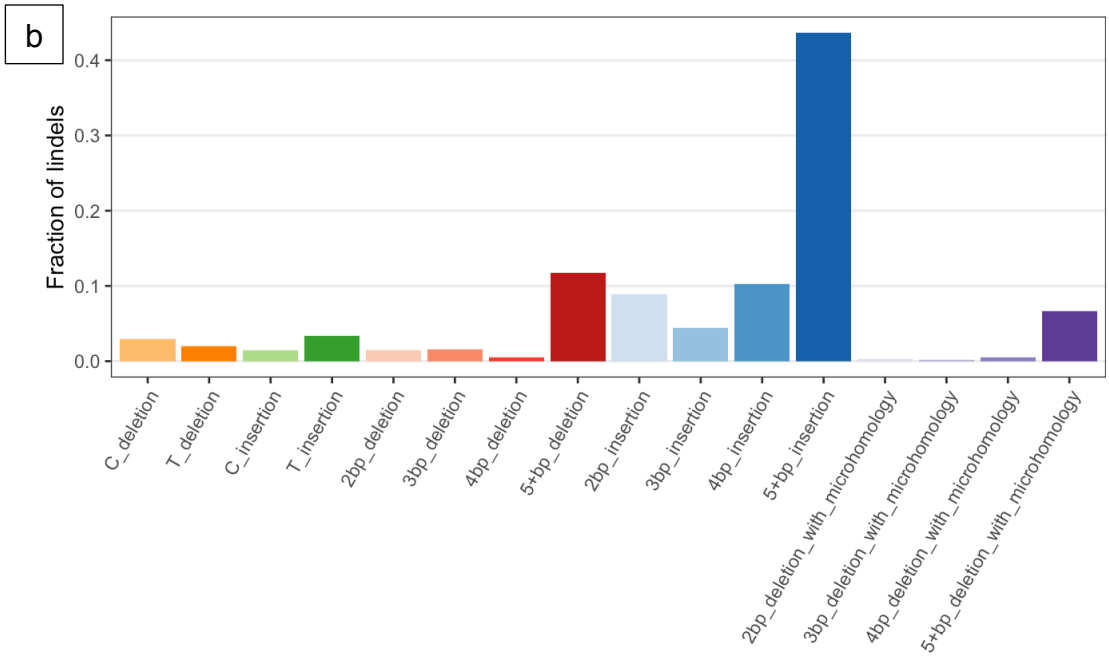
3 C/G or T/A base pairs; 2, 3, 4, or 5+ base pair deletions or insertions; and 2, 3, 4, or 5+

4 base pair deletions with microhomology at break points. Acetaldehyde treatment induces

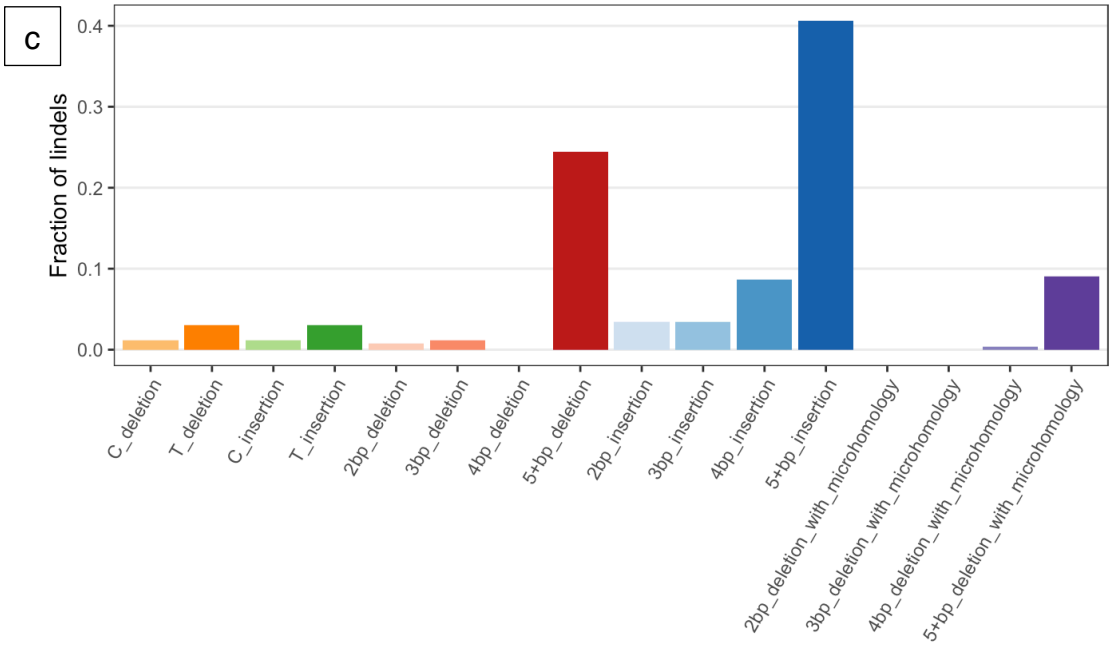
5 an increased proportion of 5+ base pair deletions (without microhomology).



6

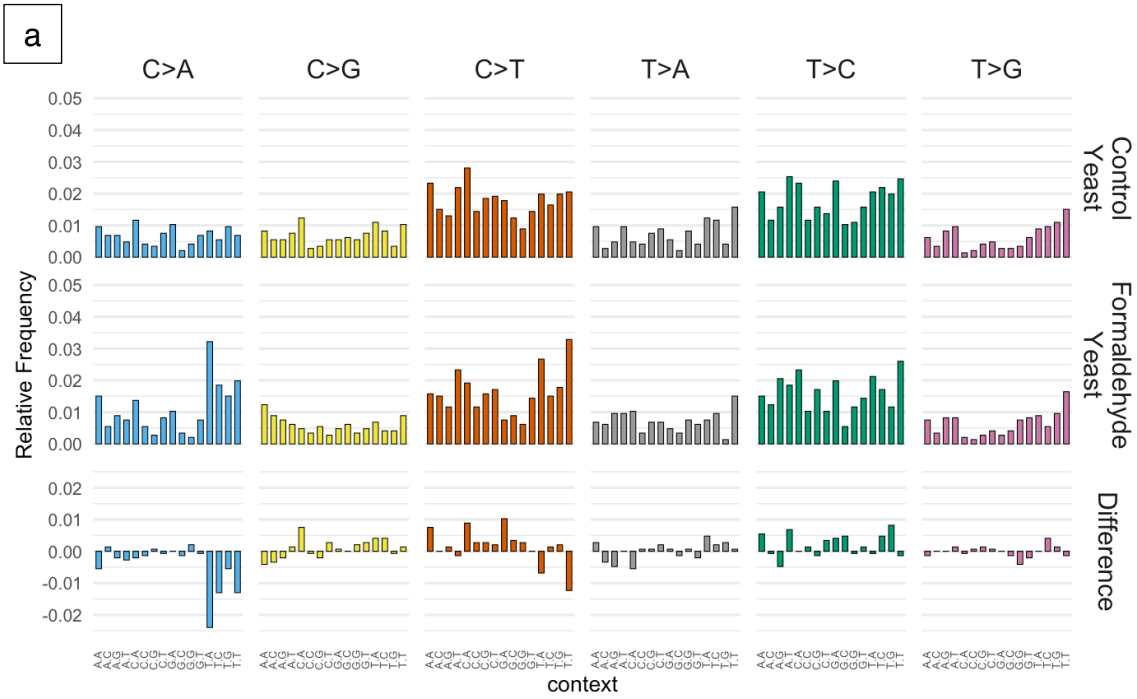

b
Fraction of lindels
0.0
0.1
0.2
0.3
0.4
C_deletion
T_deletion
C_insertion
T_insertion
2bp_deletion
3bp_deletion
4bp_deletion
5+bp_deletion
2bp_insertion
3bp_insertion
4bp_insertion
5+bp_insertion
2bp_deletion_with_microhomology
3bp_deletion_with_microhomology
4bp_deletion_with_microhomology
5+bp_deletion_with_microhomology

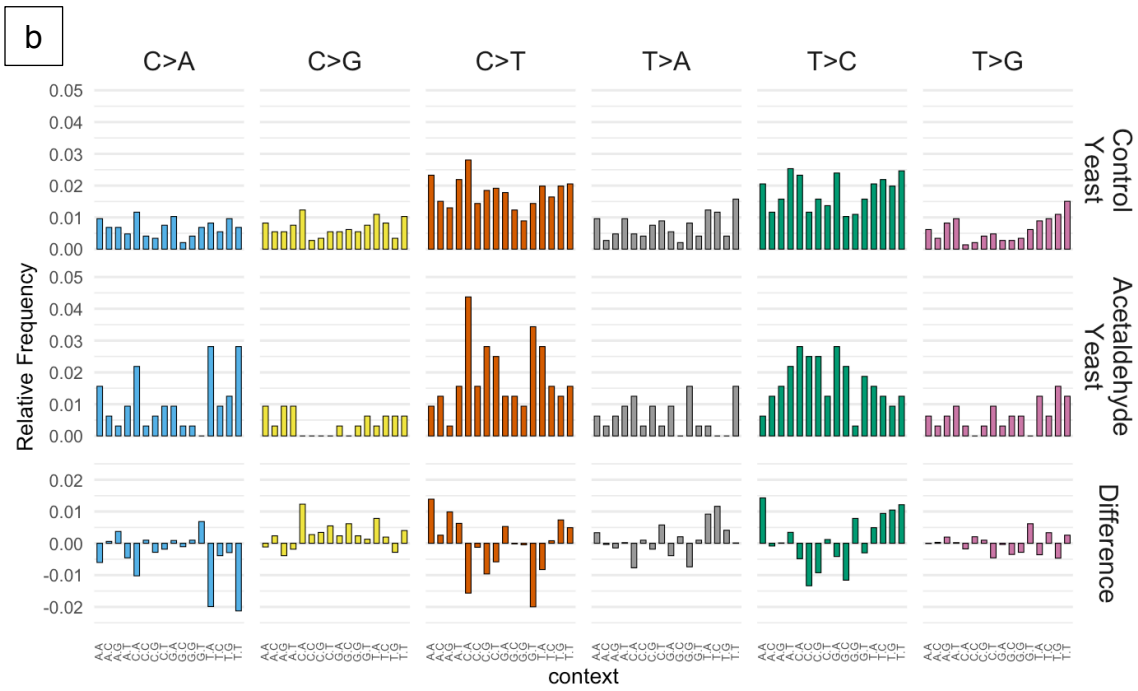
1

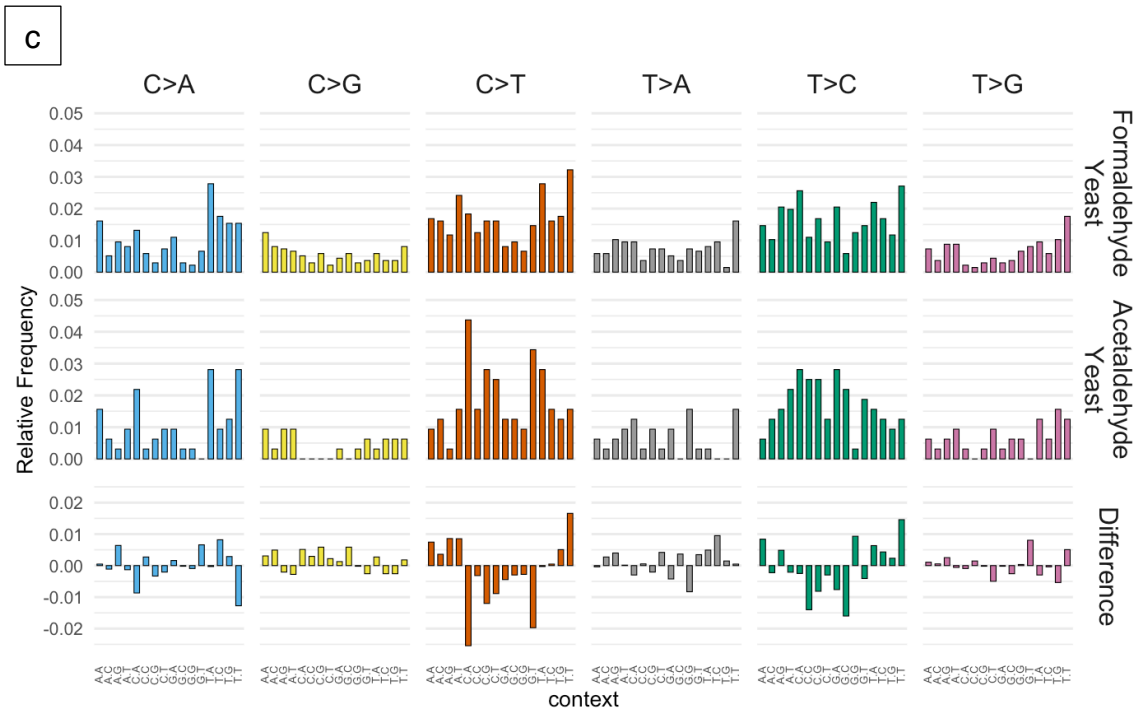

c
Fraction of lindels
0.0
0.1
0.2
0.3
0.4
C_deletion
T_deletion
C_insertion
T_insertion
2bp_deletion
3bp_deletion
4bp_deletion
5+bp_deletion
2bp_insertion
3bp_insertion
4bp_insertion
5+bp_insertion
2bp_deletion_with_microhomology
3bp_deletion_with_microhomology
4bp_deletion_with_microhomology
5+bp_deletion_with_microhomology

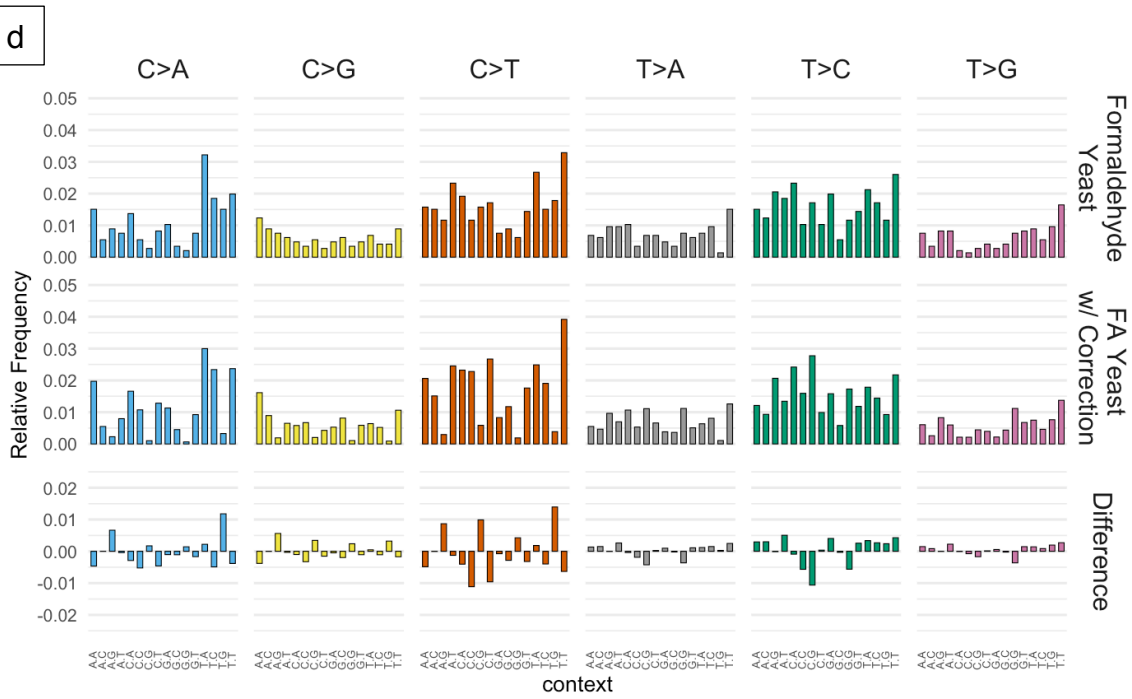
2

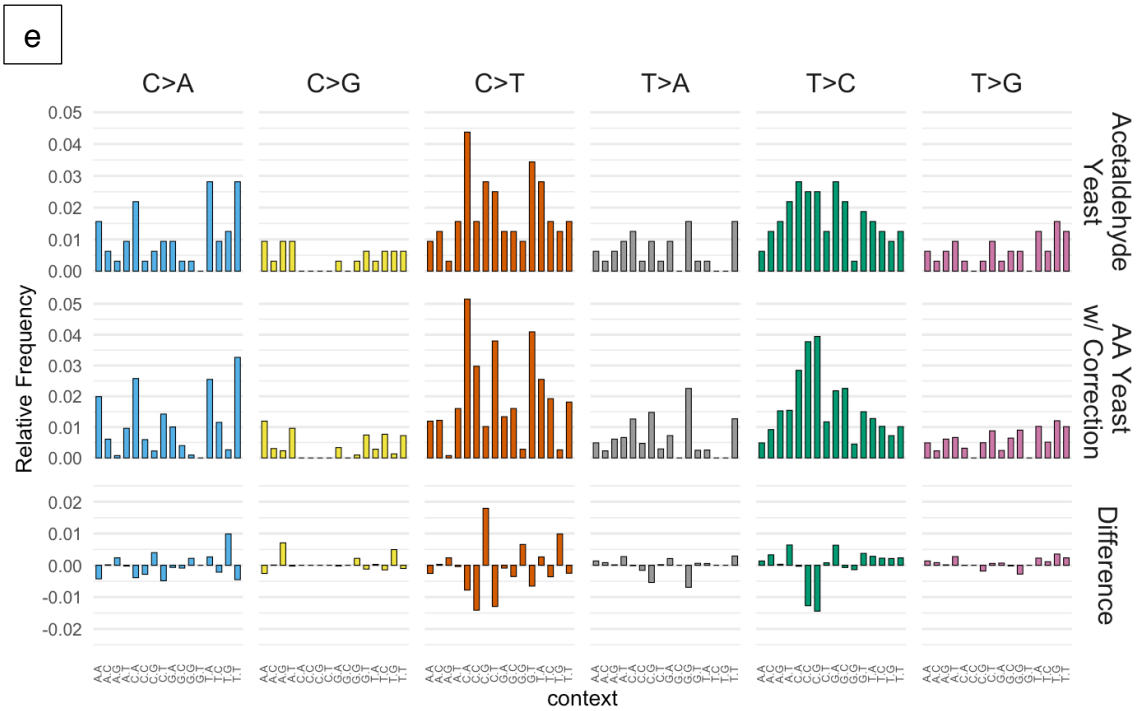
Figure 4: Comparisons of mutational patterns between (a) untreated controls and formaldehyde; (b) untreated controls and acetaldehyde; (c) formaldehyde and acetaldehyde; (d) formaldehyde with and without correction for trinucleotide frequencies in mouse; (e) acetaldehyde with and without correction for trinucleotide frequencies in mouse; and (f) Aldh2 Adh5 deficient mouse cells and trinucleotide frequency corrected yeast treated with formaldehyde.

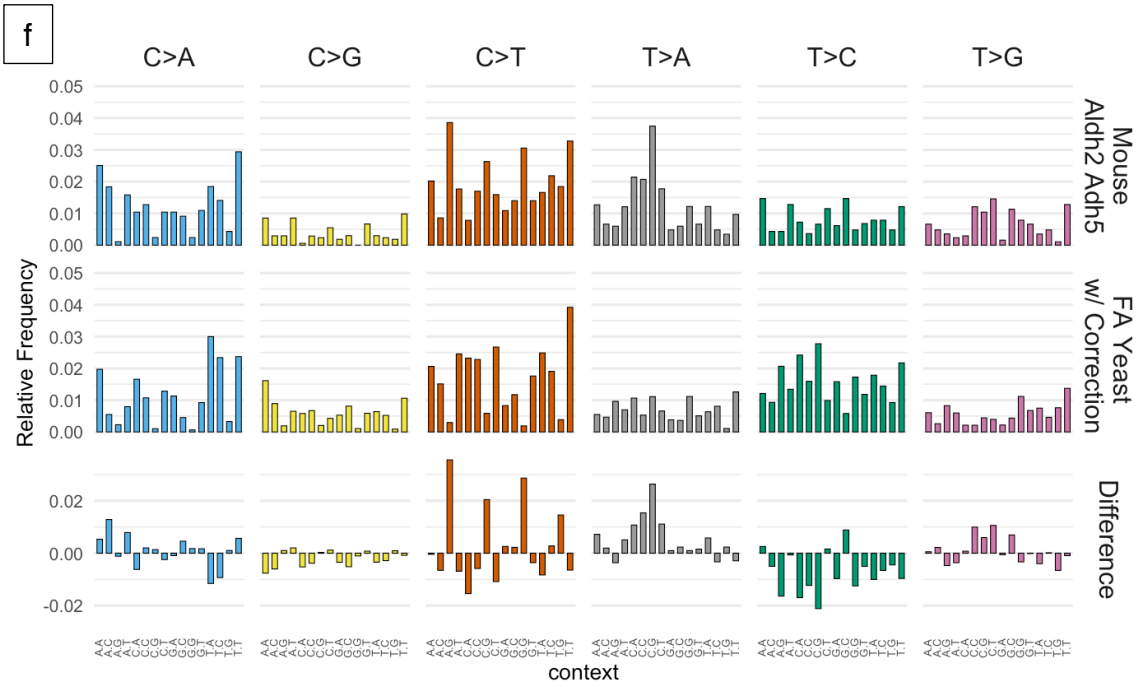



b
C>A
C>G
C>T
T>A
T>C
T>G
Control Yeast
Acetaldehyde Yeast
Difference
Relative Frequency
context


c
C>A
C>G
C>T
T>A
T>C
T>G
Formaldehyde Yeast
Acetaldehyde Yeast
Difference
Relative Frequency
context

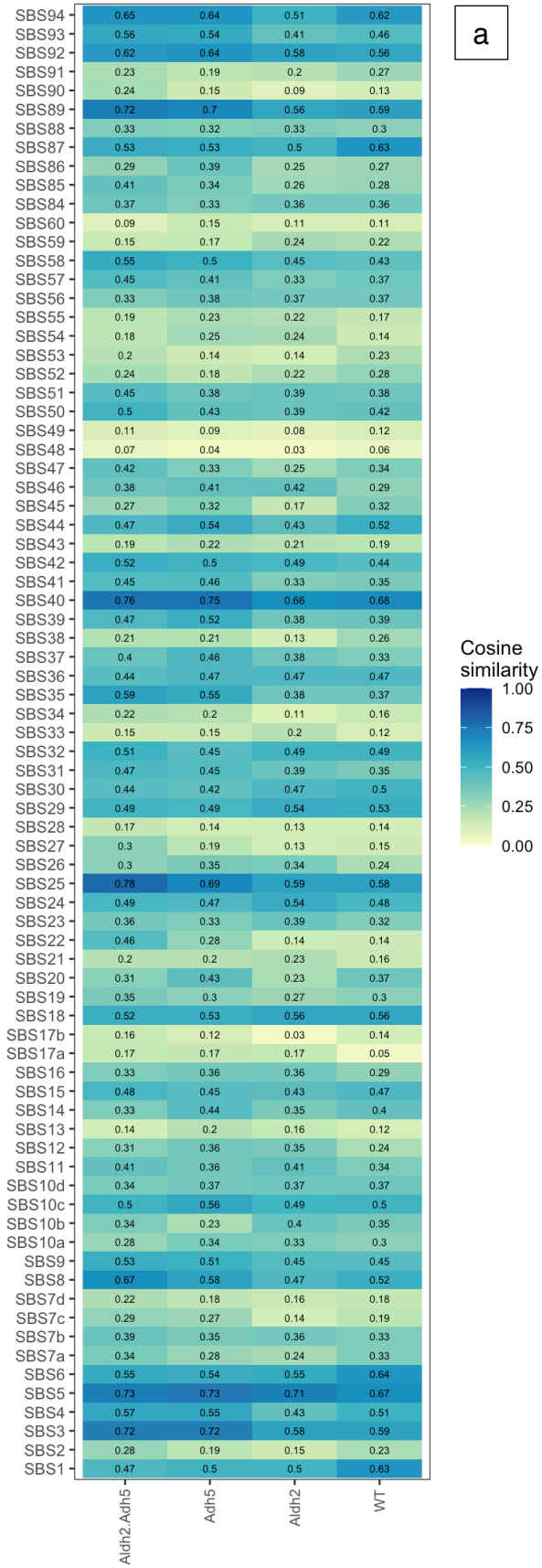

d
C>A
C>G
C>T
T>A
T>C
T>G
Formaldehyde Yeast
FA Yeast w/ Correction
Difference
Relative Frequency
context


e
C>A
C>G
C>T
T>A
T>C
T>G
Acetaldehyde Yeast
AA Yeast w/ Correction
Difference
Relative Frequency
context

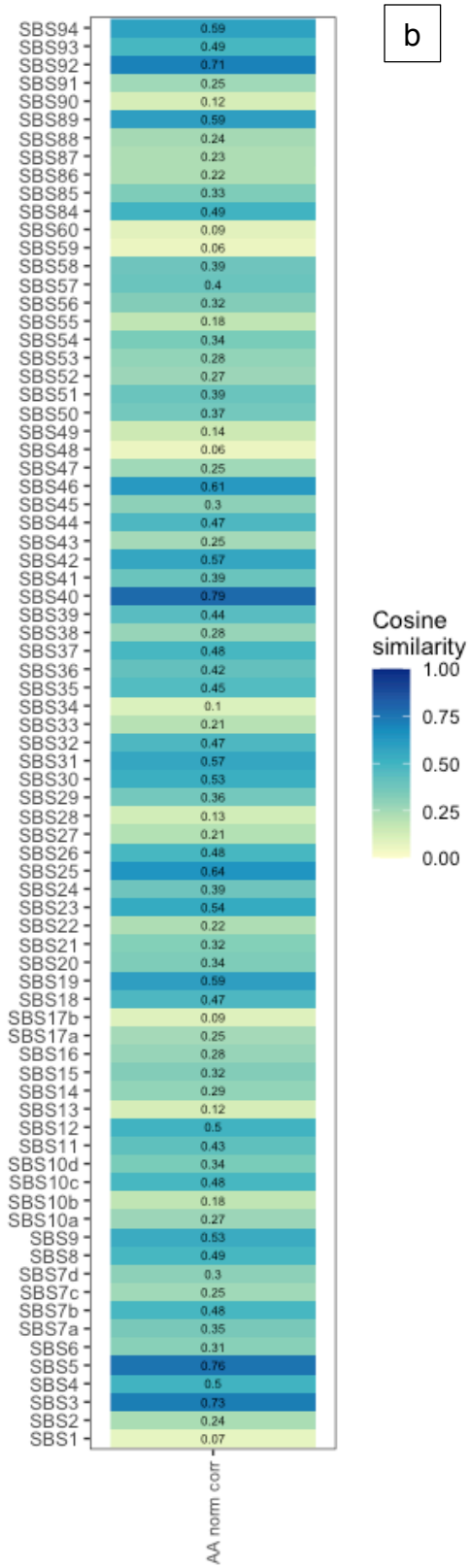

f
C>A
C>G
C>T
T>A
T>C
T>G
Mouse Aldh2 Adh5
FA Yeast w/ Correction
Difference
Relative Frequency
0.05
0.04
0.03
0.02
0.01
0.00
-0.02
context

1

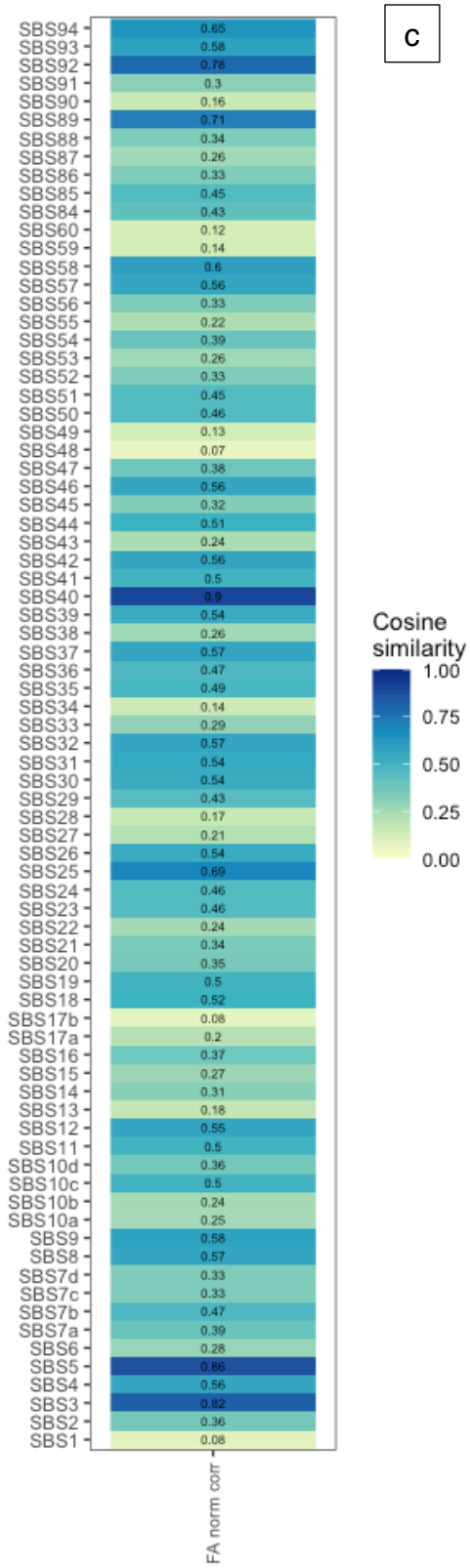
Figure 5: (a) Cosine similarity values comparing mouse mutational patterns from Dingler et al. vs. COSMIC SBS signatures. (b) Cosine similarity values comparing acetaldehyde mutational pattern in yeast (with trinucleotide abundance correction) vs. COSMIC SBS signatures. (c) Cosine similarity values comparing corrected formaldehyde mutational pattern in yeast vs. COSMIC SBS signatures. (d) Comparison of corrected formaldehyde mutational pattern in yeast vs. SBS signature 40. (e) Comparison of corrected formaldehyde mutational pattern in yeast vs. SBS signature 5.


a
SBS94 0.65 0.64 0.51 0.62
SBS93 0.56 0.54 0.41 0.46
SBS92 0.62 0.64 0.58 0.56
SBS91 0.23 0.19 0.2 0.27
SBS90 0.24 0.15 0.09 0.13
SBS89 0.72 0.7 0.56 0.59
SBS88 0.33 0.32 0.33 0.3
SBS87 0.53 0.53 0.5 0.63
SBS86 0.29 0.39 0.25 0.27
SBS85 0.41 0.34 0.26 0.28
SBS84 0.37 0.33 0.36 0.36
SBS60 0.09 0.15 0.11 0.11
SBS59 0.15 0.17 0.24 0.22
SBS58 0.55 0.5 0.45 0.43
SBS57 0.45 0.41 0.33 0.37
SBS56 0.33 0.38 0.37 0.37
SBS55 0.19 0.23 0.22 0.17
SBS54 0.18 0.25 0.24 0.14
SBS53 0.2 0.14 0.14 0.23
SBS52 0.24 0.18 0.22 0.28
SBS51 0.45 0.38 0.39 0.38
SBS50 0.5 0.43 0.39 0.42
SBS49 0.11 0.09 0.08 0.12
SBS48 0.07 0.04 0.03 0.06
SBS47 0.42 0.33 0.25 0.34
SBS46 0.38 0.41 0.42 0.29
SBS45 0.27 0.32 0.17 0.32
SBS44 0.47 0.54 0.43 0.52
SBS43 0.19 0.22 0.21 0.19
SBS42 0.52 0.5 0.49 0.44
SBS41 0.45 0.46 0.33 0.35
SBS40 0.76 0.75 0.66 0.68
SBS39 0.47 0.52 0.38 0.39
SBS38 0.21 0.21 0.13 0.26
SBS37 0.4 0.46 0.38 0.33
SBS36 0.44 0.47 0.47 0.47
SBS35 0.59 0.55 0.38 0.37
SBS34 0.22 0.2 0.11 0.16
SBS33 0.15 0.15 0.2 0.12
SBS32 0.51 0.45 0.49 0.49
SBS31 0.47 0.45 0.39 0.35
SBS30 0.44 0.42 0.47 0.5
SBS29 0.49 0.49 0.54 0.53
SBS28 0.17 0.14 0.13 0.14
SBS27 0.3 0.19 0.13 0.15
SBS26 0.3 0.35 0.34 0.24
SBS25 0.78 0.69 0.59 0.58
SBS24 0.49 0.47 0.54 0.48
SBS23 0.36 0.33 0.39 0.32
SBS22 0.46 0.28 0.14 0.14
SBS21 0.2 0.2 0.23 0.16
SBS20 0.31 0.43 0.23 0.37
SBS19 0.35 0.3 0.27 0.3
SBS18 0.52 0.53 0.56 0.56
SBS17b 0.16 0.12 0.03 0.14
SBS17a 0.17 0.17 0.17 0.05
SBS16 0.33 0.36 0.36 0.29
SBS15 0.48 0.45 0.43 0.47
SBS14 0.33 0.44 0.35 0.4
SBS13 0.14 0.2 0.16 0.12
SBS12 0.31 0.36 0.35 0.24
SBS11 0.41 0.36 0.41 0.34
SBS10d 0.34 0.37 0.37 0.37
SBS10c 0.5 0.56 0.49 0.5
SBS10b 0.34 0.23 0.4 0.35
SBS10a 0.28 0.34 0.33 0.3
SBS9 0.53 0.51 0.45 0.45
SBS8 0.67 0.58 0.47 0.52
SBS7d 0.22 0.18 0.16 0.18
SBS7c 0.29 0.27 0.14 0.19
SBS7b 0.39 0.35 0.36 0.33
SBS7a 0.34 0.28 0.24 0.33
SBS6 0.55 0.54 0.55 0.64
SBS5 0.73 0.73 0.71 0.67
SBS4 0.57 0.55 0.43 0.51
SBS3 0.72 0.72 0.58 0.59
SBS2 0.28 0.19 0.15 0.23
SBS1 0.47 0.5 0.5 0.63
Aldh2.Adh5
Adh5
Aldh2
WT
Cosine similarity
1.00
0.75
0.50
0.25
0.00

1

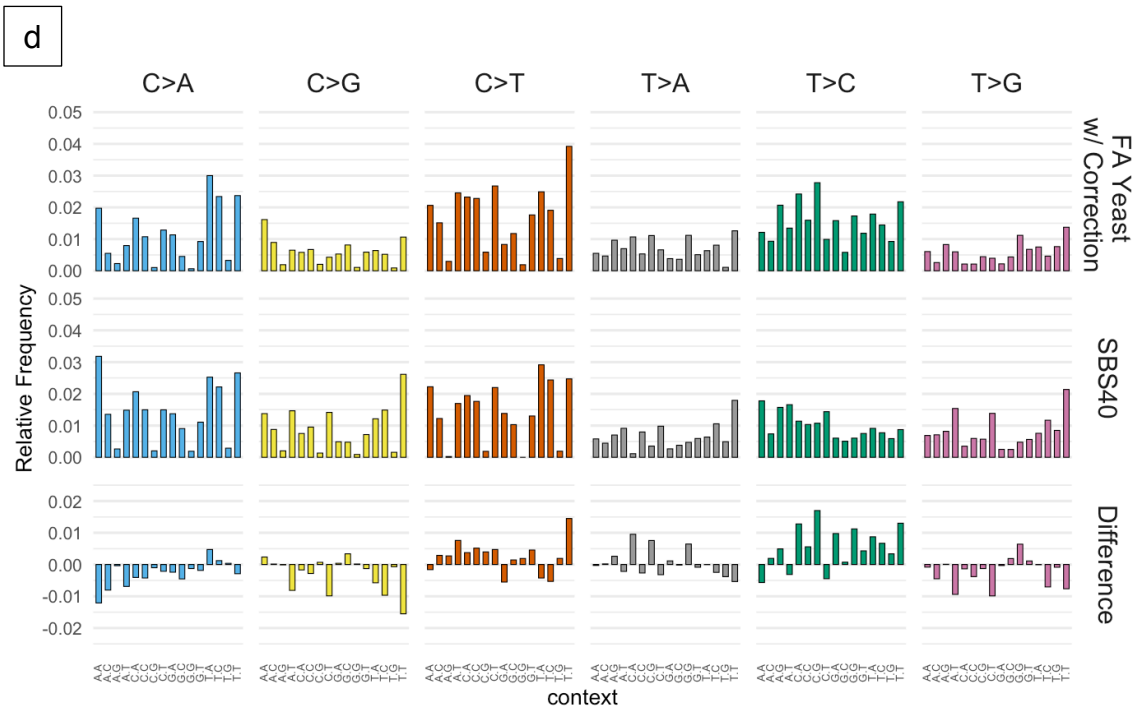

b
SBS94 0.59
SBS93 0.49
SBS92 0.71
SBS91 0.25
SBS90 0.12
SBS89 0.59
SBS88 0.24
SBS87 0.23
SBS86 0.22
SBS85 0.33
SBS84 0.49
SBS60 0.09
SBS59 0.06
SBS58 0.39
SBS57 0.4
SBS56 0.32
SBS55 0.18
SBS54 0.34
SBS53 0.28
SBS52 0.27
SBS51 0.39
SBS50 0.37
SBS49 0.14
SBS48 0.06
SBS47 0.25
SBS46 0.61
SBS45 0.3
SBS44 0.47
SBS43 0.25
SBS42 0.57
SBS41 0.39
SBS40 0.79
SBS39 0.44
SBS38 0.28
SBS37 0.48
SBS36 0.42
SBS35 0.45
SBS34 0.1
SBS33 0.21
SBS32 0.47
SBS31 0.57
SBS30 0.53
SBS29 0.36
SBS28 0.13
SBS27 0.21
SBS26 0.48
SBS25 0.64
SBS24 0.39
SBS23 0.54
SBS22 0.22
SBS21 0.32
SBS20 0.34
SBS19 0.59
SBS18 0.47
SBS17b 0.09
SBS17a 0.25
SBS16 0.28
SBS15 0.32
SBS14 0.29
SBS13 0.12
SBS12 0.5
SBS11 0.43
SBS10d 0.34
SBS10c 0.48
SBS10b 0.18
SBS10a 0.27
SBS9 0.53
SBS8 0.49
SBS7d 0.3
SBS7c 0.25
SBS7b 0.48
SBS7a 0.35
SBS6 0.31
SBS5 0.76
SBS4 0.5
SBS3 0.73
SBS2 0.24
SBS1 0.07
AA norm corr
Cosine similarity
1.00
0.75
0.50
0.25
0.00

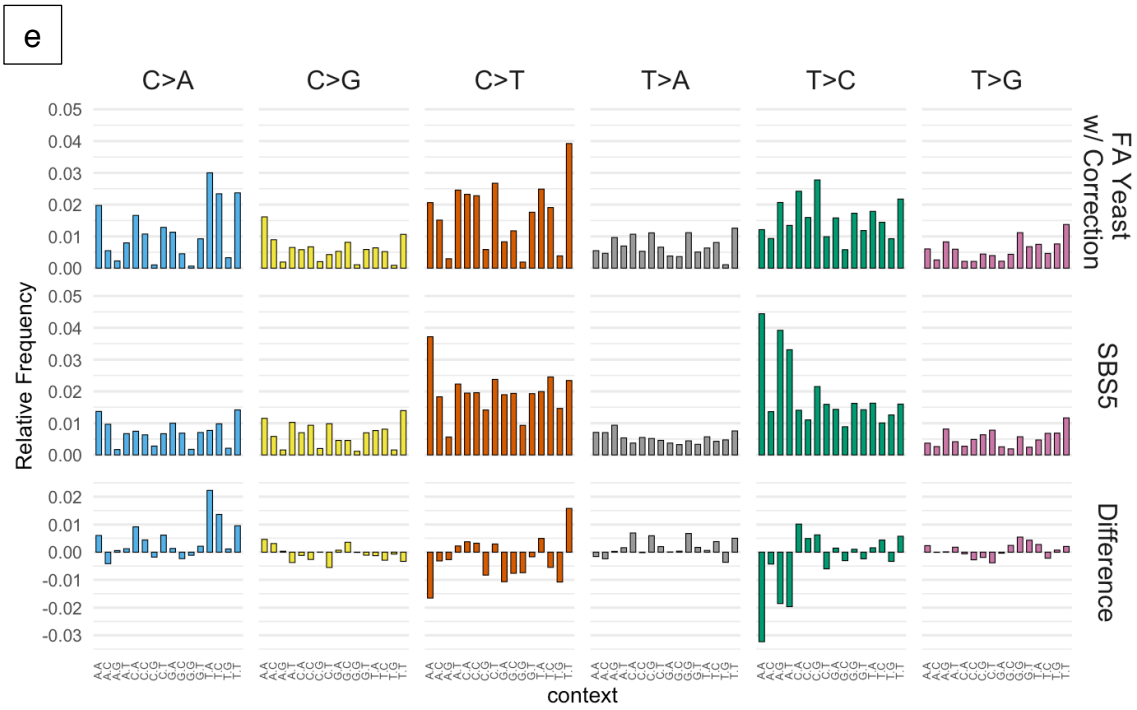
1


c
SBS94 0.65
SBS93 0.58
SBS92 0.78
SBS91 0.3
SBS90 0.16
SBS89 0.71
SBS88 0.34
SBS87 0.26
SBS86 0.33
SBS85 0.45
SBS84 0.43
SBS60 0.12
SBS59 0.14
SBS58 0.6
SBS57 0.56
SBS56 0.33
SBS55 0.22
SBS54 0.39
SBS53 0.26
SBS52 0.33
SBS51 0.45
SBS50 0.46
SBS49 0.13
SBS48 0.07
SBS47 0.38
SBS46 0.56
SBS45 0.32
SBS44 0.51
SBS43 0.24
SBS42 0.56
SBS41 0.5
SBS40 0.9
SBS39 0.54
SBS38 0.26
SBS37 0.57
SBS36 0.47
SBS35 0.49
SBS34 0.14
SBS33 0.29
SBS32 0.57
SBS31 0.54
SBS30 0.54
SBS29 0.43
SBS28 0.17
SBS27 0.21
SBS26 0.54
SBS25 0.69
SBS24 0.46
SBS23 0.46
SBS22 0.24
SBS21 0.34
SBS20 0.35
SBS19 0.5
SBS18 0.52
SBS17b 0.08
SBS17a 0.2
SBS16 0.37
SBS15 0.27
SBS14 0.31
SBS13 0.18
SBS12 0.55
SBS11 0.5
SBS10d 0.36
SBS10c 0.5
SBS10b 0.24
SBS10a 0.25
SBS9 0.58
SBS8 0.57
SBS7d 0.33
SBS7c 0.33
SBS7b 0.47
SBS7a 0.39
SBS6 0.28
SBS5 0.86
SBS4 0.56
SBS3 0.82
SBS2 0.36
SBS1 0.08
FA norm corr
Cosine similarity
1.00
0.75
0.50
0.25
0.00

1

2


d
C>A
C>G
C>T
T>A
T>C
T>G
FA Yeast w/ Correction
SBS40
Difference
Relative Frequency
context


e
C>A
C>G
C>T
T>A
T>C
T>G
FA Yeast w/ Correction
SBS5
Difference
Relative Frequency
context

1 **Tables**

2

3 **Table 1:** Cosine similarity values between mutational profiles of (FA) formaldehyde- and

4 (AA) acetaldehyde-mutagenized yeast (with correction for trinucleotide frequencies in

5 mouse) and of mice deficient for aldehyde detoxification genes from (37).

	Mouse Aldh2$^{-/-}$	Mouse Adh5$^{-/-}$	Mouse Aldh2$^{-/-}$ Adh5$^{-/-}$
Yeast FA, corrected	0.658	0.735	0.767
Yeast AA, corrected	0.617	0.633	0.673

6

7

8 **Table 2:** Cosine similarity values among mice deficient for aldehyde detoxification genes

9 from (37).

	Mouse WT	Mouse Aldh2$^{-/-}$	Mouse Adh5$^{-/-}$	Mouse Aldh2$^{-/-}$ Adh5$^{-/-}$
Mouse WT	1	0.832	0.845	0.838
Mouse Aldh2$^{-/-}$		1	0.780	0.774
Mouse Adh5$^{-/-}$			1	0.887
Mouse Aldh2$^{-/-}$ Adh5$^{-/-}$				1

10