


Analyses of mutational patterns induced by formaldehyde and acetaldehyde reveal similarity to a common mutational signature

Mahanish J. Thapa,^{1,2,†} Reena M. Fabros,^{1,2,†} Salma Alasmar,^{3,4} Kin Chan ^{1,2,*}

¹Department of Biochemistry, Microbiology and Immunology, University of Ottawa, Ottawa, ON K1H 8M5, Canada,

²Ottawa Institute of Systems Biology, University of Ottawa, Ottawa, ON K1H 8M5, Canada,

³Biopharmaceutical Sciences Undergraduate Program, University of Ottawa, Ottawa, ON K1N 6N5, Canada

⁴Present address: Department of Chemistry and Biomolecular Sciences, University of Ottawa, 150 Louis-Pasteur Pvt, Ottawa, ON K1N 6N5, Canada.

*Corresponding author: Department of Biochemistry, Microbiology and Immunology, University of Ottawa, 451 Smyth Rd, Ottawa, ON K1H 8M5, Canada. Email: kin.chan@uottawa.ca

†These authors contributed equally to this work.

Abstract

Formaldehyde and acetaldehyde 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_C/G_A motifs. Interestingly, the formaldehyde mutational pattern resembles single base substitution signature 40 from the Catalog of Somatic Mutations in Cancer. Single base substitution signature 40 is a mutational signature of unknown etiology. We also noted that acetaldehyde treatment caused an excess of deletion events longer than 4 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 2 important, commonly occurring mutagens.

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

Introduction

Genomic DNA is constantly damaged by intracellular processes (De Bont and van Larebeke 2004) and exposure to exogenous damaging agents (Irigaray and Belpomme 2010; Ikehata and Ono 2011; Keszenman et al. 2015). There are many different types of DNA damage. Intracellular DNA damaging processes include, for example oxidation of nitrogenous bases (Ames et al. 1993; Helbock et al. 1998); glycosidic bond breakage, which releases a nitrogenous base from its deoxyribose sugar (Lindahl and Nyberg 1972; Lindahl 1977; Tice and Setlow 1985; Lindahl 1993; Nakamura et al. 1998); single- and double-stranded breaks of the sugar-phosphate backbone (Tice and Setlow 1985; Haber 1999; Vilenchik and Knudson 2003); base alkylation (Tice and Setlow 1985; Kadlubar et al. 1998; VanderVeen et al. 2003); cytosine deamination to uracil (Tice and Setlow 1985; Sapparbaev and Zharkov 2017); and deamination of 5-methylcytosine to thymine (Greenblatt et al. 1994; Neddermann et al. 1996; Sassa et al. 2016). Examples of exogenous DNA damage include: ultraviolet (UV) light (Ikehata and Ono 2011); ionizing

radiation (Keszenman et al. 2015); tobacco (Alexandrov et al. 2016); aristolochic acid (Moriya et al. 2011); and aflatoxin (Letouzé et al. 2017). Mutations are also thought to result from spontaneous ionization or isomerization (i.e. tautomerization) of DNA bases, which can alter base pairing characteristics (Russo et al. 1998; Podolyan et al. 2000; Masoodi et al. 2016; Kimsey et al. 2018).

It is important to note that these processes do not affect all bases equally. Each of the 4 nitrogenous bases has its own distinct set of chemically reactive moieties (e.g. amines, carbonyls, or labile ring atoms) (Alberts et al. 2014). 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 (Alexandrov et al. 2020).

Received: June 03, 2022. Accepted: August 22, 2022

© The Author(s) 2022. Published by Oxford University Press on behalf of Genetics Society of America.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Mutational signatures typically are inferred using a nonnegative matrix factorization (NMF) algorithm (Alexandrov et al. 2013). 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 UV 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 (Alexandrov et al. 2020). About one-third of currently defined mutational signatures remain of unknown etiology (Alexandrov et al. 2020).

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 (International Agency for Research on Cancer 2010). Among these high-priority carcinogens are 2 small aldehyde compounds, formaldehyde (CH₂O) and acetaldehyde (C₂H₄O). Formaldehyde is classified as a known human carcinogen by IARC, based in part on the evidence of occupational exposure being associated with nasal and nasopharyngeal cancers (International Agency for Research on Cancer 2012a; National Toxicology Program 2014). 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 (National Toxicology Program 2014). Acetaldehyde is a reactive compound that humans are commonly exposed to as a result of ethanol consumption, as the initial step of ethanol detoxification is oxidation to acetaldehyde. Like formaldehyde, acetaldehyde is also classified as a known human carcinogen (Secretan et al. 2009). Alcohol consumption is associated with higher risk of multiple types of cancer, including: head and neck; esophageal; liver; breast; and colorectal (International Agency for Research on Cancer 2012b). Acetaldehyde associated with alcohol consumption is thought to be causative for cancers of the esophagus and the upper aerodigestive tract (including head and neck), i.e. at sites of highest direct exposure (International Agency for Research on Cancer 2012b).

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. Kucab et al. 2019; Dingler et al. 2020) 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, NJ). Canavanine (C9758), adenine sulfate dihydrate (AD0028), formaldehyde (F8775), and acetaldehyde (W200344) were purchased from MilliporeSigma (St. Louis, MO). 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 buffer (A0026) were purchased from BioBasic (Markham, ON). G418 sulfate (450-130) was purchased from Wisent (St-Bruno, QC). Q5 PCR kits were purchased from New England Biolabs Canada (Whitby, ON). Gas-tight glass tubes with septa (2048-18150) and accessories (2048-11020 and 2048-10020) were purchased from Bellco Glass Inc. (Vineland, NJ).

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 3 reporter genes (CAN1, URA3, and ADE2) near the de novo left telomere of chromosome V. These 3 genes had been deleted from their native loci. Details about ySR127 were described previously (Chan et al. 2012) 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 3 days. Then, cultures were diluted 1:10 into fresh media in gas-tight glass tubes, shifted to restrictive temperature (37°C), and shaken gently at 150 RPM for 3 h, with syringe needles inserted through the septa to enable gas exchange. After a 3-h 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 3 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 0.33 \times 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 (Chan 2018).

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 (R Core Team 2020) and tidyverse package version 1.3 (Wickham et al. 2019).

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 was verified as described previously (Chan et al. 2012). 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 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, 8, and 10 mM formaldehyde exposure were chosen for sequencing as these had high induced mutation frequencies. For acetaldehyde, mutants from 75 mM treatment were selected for sequencing, as this concentration was most mutagenic. No-aldehyde controls were isolated similarly, except that they were Can^r Ade⁻ mutants from 24-h temperature shifts without added mutagen. This longer shift was necessary for controls to acquire more mutations for analysis. Shorter temperature shift without added mutagen would have yielded fewer variants, and sequencing many more control genomes to compensate was not practicable due to budgetary constraints. Twenty-four-hour shifts in the presence of mutagen also were not possible, resulting in very high lethality.

Illumina library preparation and WGS were outsourced to Genome Québec (McGill University, Montréal) or performed on an Illumina MiSeq in our lab. Bowtie2 version 2.3.5.1 (Langmead and Salzberg 2012), SAMtools 1.9 (Li et al. 2009), and bcftools 1.9 (Li 2011) were used to map the Illumina reads and call variants. The ySR127 reference sequence was obtained soon after strain construction and represents that genome in an unmutated state, so the variants acquired from each treatment condition can be easily identified. This reference sequence was previously released publicly on NCBI (Chan et al. 2015). To map reads to the ySR127 reference and create a sorted BAM file, we ran the following command on each sample: “bowtie2 -local -x ySR127 -1 sample_R1.fastq.gz -2 sample_R2.fastq.gz | samtools view -bS | samtools sort -o sample.bam.” To call variants and output to a BCF file: “bcftools mpileup -Ou -f ySR127.fa sample.bam | bcftools call -p ploidy 1 -v -c -Ou -o sample.bcf.” Variants with quality score <30 and/or with sequencing coverage <10 were filtered out: “bcftools view sample.bcf -e ‘INFO/DP < 10’ | bcftools view -e ‘QUAL < 30’ | bcftools view -Ov -o sample.vcf.” VCF file for each sample was then compressed and indexed: “bgzip -c sample.vcf > sample.vcf.gz” and “tabix -p vcf sample.vcf.gz.” Sample VCF files were merged to create a unified VCF: “bcftools merge -m none -Ov -o merge.vcf *.vcf.gz,” where * is a wild card variable for the sample names. In this way, if the same variant is found in multiple samples, they were combined into 1 unique variant. The resulting unified VCF files were passed to MutationalPatterns version 3.6.3 (Blokzijl et al. 2018) for further analysis and visualization. Other numerical and statistical analyses, and data visualizations were done using base R version 4.1 (R Core Team 2020) and tidyverse package version 1.3 (Wickham et al. 2019).

For trinucleotide frequency correction, the Biostrings package version 2.38.0 (Pagès et al. 2022) 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 the addition of formaldehyde or acetaldehyde. These experiments were done using a haploid yeast strain (ySR127) that forms long regions of subtelomeric single-stranded DNA (ssDNA) when shifted to 37°C due to the *cdc13-1* temperature sensitive point mutation (Garvik et al. 1995). 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 G₂ (Garvik et al. 1995). The reporter genes *CAN1*, *ADE2*, and *URA3* had been deleted from their native loci and reintroduced to the left subtelomeric region of chromosome V (Chan et al. 2012). 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 (Grogan and Jinks-Robertson 2012). The ssDNA system was used previously to study the mutagenic properties of bisulfite and human APOBEC3G cytidine deaminase (Chan et al. 2012); abasic sites (Chan et al. 2013); reactive oxygen species (Degtyareva et al. 2013); human APOBEC3A and APOBEC3B cytidine deaminases (Chan et al. 2015); and alkylating agents (Saini et al. 2020).

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 Fig. 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 Fig. 1b). Mutagenesis plateaued from 4 to 8 mM formaldehyde, with median mutation frequencies of $\sim 1.5 \times 10^{-3}$. Mutagenesis peaked at 10 mM formaldehyde exposure (median mutation frequency = 2.7×10^{-3}), but with a steep decrease in viability. 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 Fig. 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 Fig. 1d). Mutation frequency peaked at over 5×10^{-4} when cells were treated with

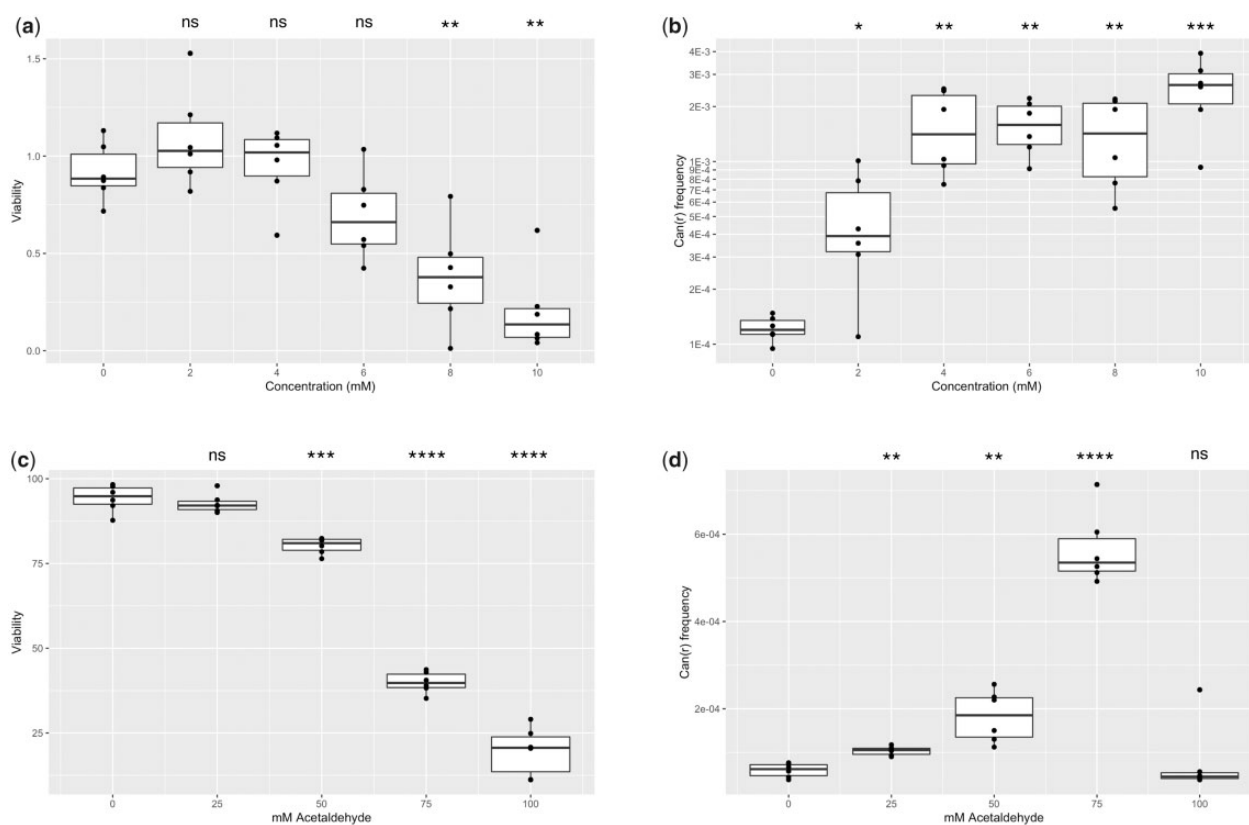


Fig. 1. a) Viability and b) CAN1 inactivation frequency of yeast treated with 0, 2, 4, 6, 8, or 10 mM formaldehyde. c) Viability and d) CAN1 inactivation frequency of yeast treated with 0, 25, 50, 75, or 100 mM acetaldehyde. Data are from 6 biological replicates for each aldehyde. * denotes $P < 0.05$, ** denotes $P < 0.01$, *** denotes $P < 0.001$, **** denotes $P < 0.0001$, and ns denotes no significant difference by paired t-test.

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.

Formaldehyde and acetaldehyde induce an excess of $C/G \geq A/T$ transversions

We collected mutagenized isolates for Illumina whole genome sequencing to determine what kinds of genetic variants were induced by either formaldehyde (119 genomes) or acetaldehyde (17 genomes) treatment. Total numbers of variants for each sequenced genome and variant calls are reported in [Supplementary Tables 2 and 3](#), respectively. As one would expect, there were mutational hotspots that were mutated recurrently in different samples. Constructing a mutational profile by tallying the number of occurrences (and recurrences) at each site would likely not be a good representation of intrinsic mutational preference, per se. Recurrence could be due to a trinucleotide being susceptible to mutation, but it might also be due to selection effects. Instead, we aggregated data across all samples in each data set and counted mutated motifs: If a treatment does preferentially mutate a trinucleotide motif, multiple instances of that motif at different genomic loci would be mutated. On the other hand, if mutation at a particular instance of a trinucleotide is observed recurrently but there are few other instances of that trinucleotide being mutated at other loci, then selection is quite possible. We adopted our analytical approach to minimize possible distorting effects of selection.

The genomes mutagenized by either small aldehyde were compared to control genomes that were not treated by either. Analysis of the 69 control genomes revealed a mutational pattern where $C/G > T/A$ and $T/A > C/G$ transitions outnumbered the 4 types of transversions (namely $C/G > A/T$, $C/G > G/C$, $T/A > A/T$, and $T/A > G/C$, see [Fig. 2a](#)), similar to what we had observed previously ([Gelova et al. 2020](#)). By comparison, formaldehyde and acetaldehyde treatment both caused a relative increase of $C/G > A/T$ transversions (see [Fig. 2, b and c](#)). 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 ssDNA.

Since ssDNA should be enriched near the chromosome ends, most variants should map in such regions. To check this, we constructed genome-wide rainfall plots for controls, formaldehyde-, and acetaldehyde-treated isolates (see [Fig. 2, d-f](#), respectively). These graphs show the number of base pairs between adjacent mutations. Consistent with expectation, variants tended to cluster near chromosome ends. Higher-resolution views showing individual chromosomes are available in [Supplementary Figs 1, 2, and 3](#).

Acetaldehyde induces deletions of 5 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 5 or more bases, with smaller

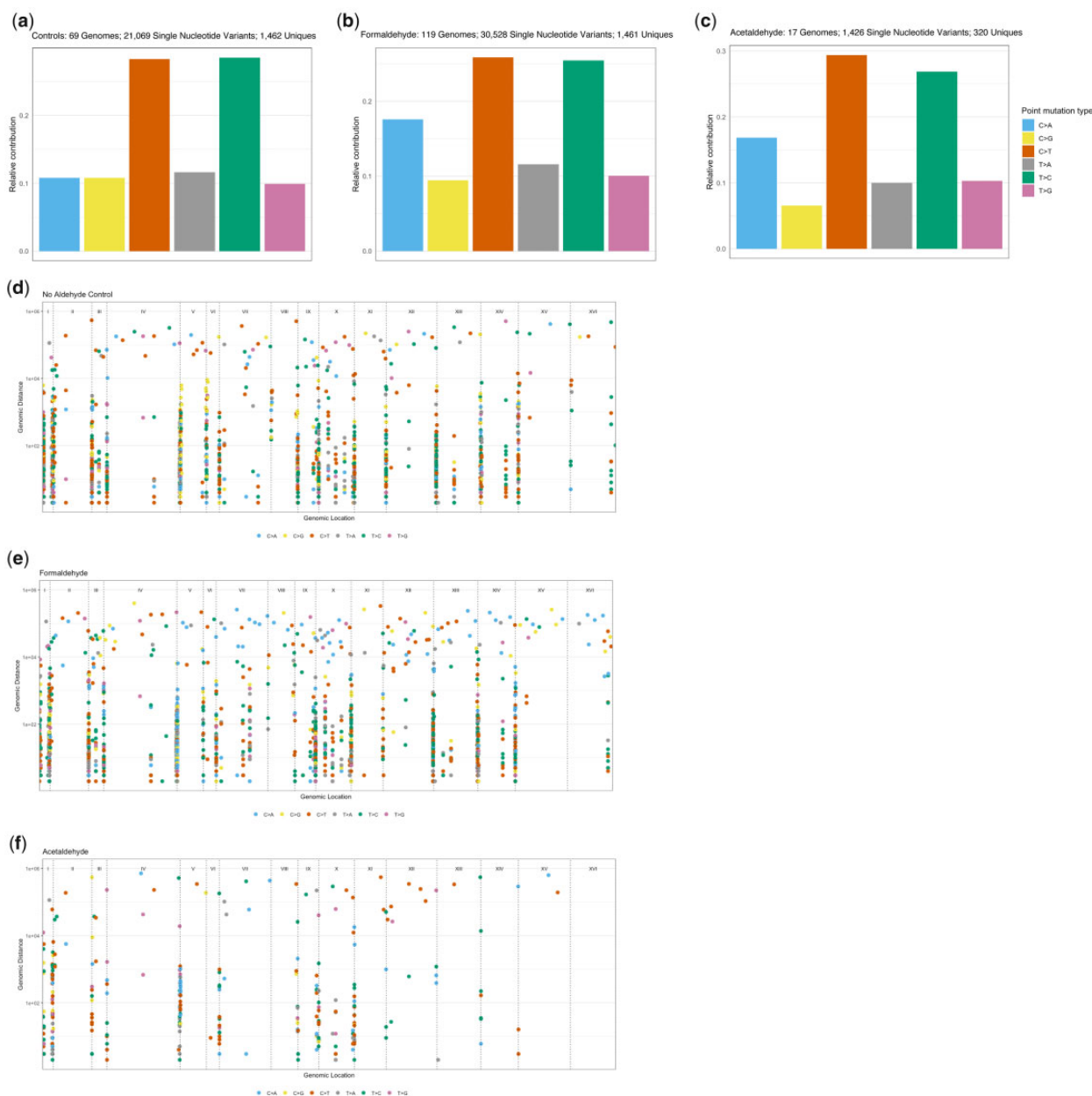


Fig. 2. Base substitution types for (a) controls, (b) formaldehyde, and (c) acetaldehyde. Treatment with either aldehyde caused a higher proportion of C/G > A/T transversions. Rainfall plots, showing distance between adjacent mutations, show that most cluster near chromosome ends where ssDNA is enriched, for (d) controls, (e) formaldehyde, and (f) acetaldehyde. Total numbers of sequenced genomes, total numbers of variant calls, and number of unique variants are reported (if the same variant occurs in multiple samples, it is counted as 1 unique).

proportions of shorter insertions as well as deletions of 5 or more bases (see Fig. 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 Fig. 3b). In contrast, there was a notable difference in the acetaldehyde-induced profile of indels: an excess of deletions of 5 or more bases was observed (24% in acetaldehyde vs. 12% in controls, see Fig. 3c). This is a distinguishing property of acetaldehyde-induced DNA damage in the ssDNA system. Plotting these data while grouping by number of repeat units adjacent to each indel confirmed the excess of these deletions from acetaldehyde treatment (compare Fig. 3, d-f). The most frequent events were deletion of a single unit. Deletions were less frequent as the number of repeat units

increased, likely because longer tandem sets of repeats were simply more rare.

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

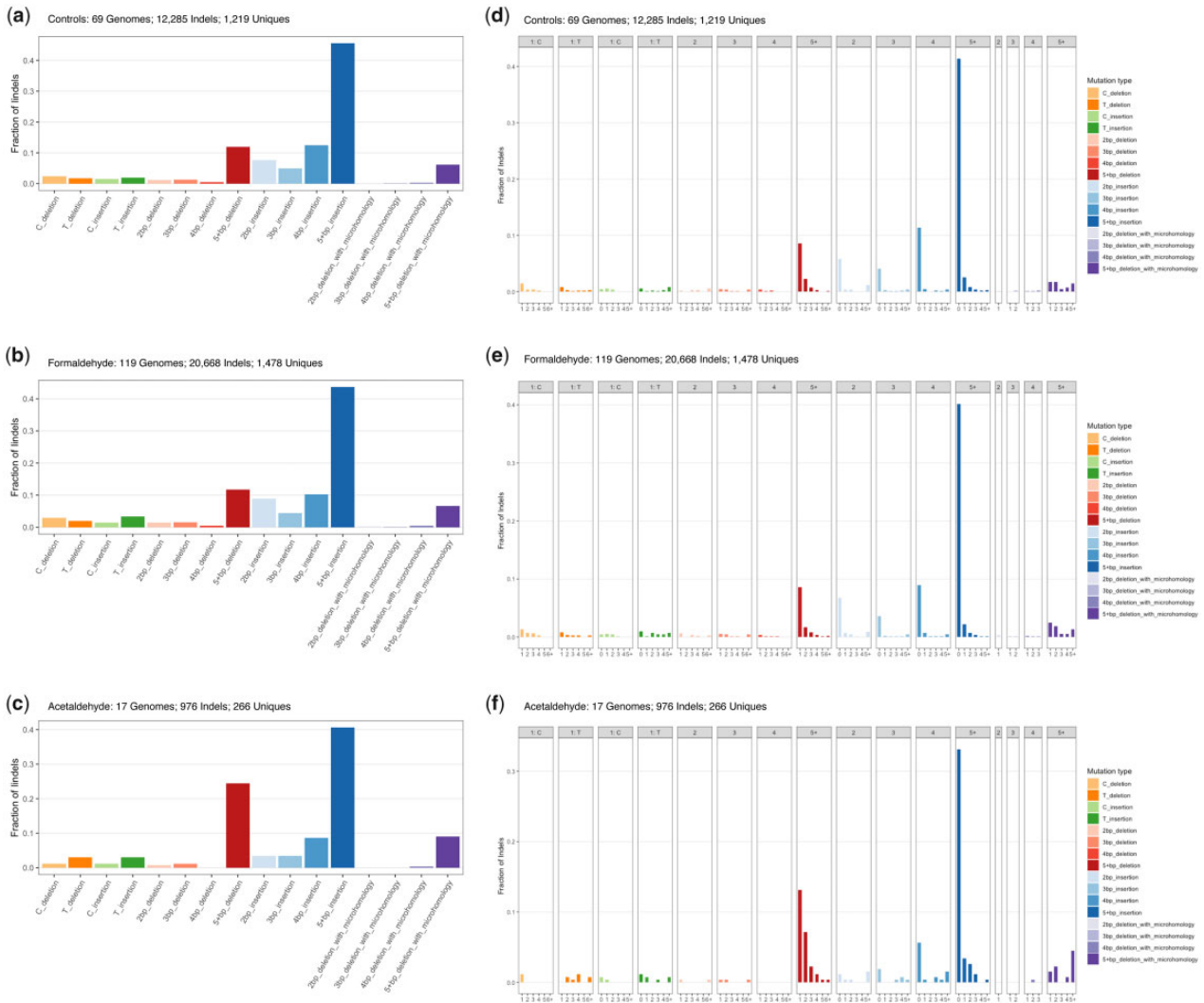


Fig. 3. Small indels from (a) no-aldehyde controls, (b) formaldehyde, and (c) acetaldehyde. The different categories comprise: single base deletions or insertions at C/G or T/A base pairs; 2, 3, 4, or 5+ base pair deletions or insertions; and 2, 3, 4, or 5+ base pair deletions with microhomology at break points. Acetaldehyde treatment induces an increased proportion of 5+ base pair deletions (without microhomology). The same small indel data, plotted showing number of repeat units from (d) no-aldehyde controls, (e) formaldehyde, and (f) acetaldehyde. For the single-nucleotide indels, the number of repeat units is the length of a homopolymer run. For indels of dinucleotide, trinucleotide, or greater length, the number of repeat units indicates how many copies of the inserted or deleted unit are immediately adjacent to the site of the indel. Total numbers of sequenced genomes, total numbers of indel calls, and number of unique indels are reported (if the same indel occurs in multiple samples, it is counted as 1 unique).

2 identical patterns (Alexandrov et al. 2013). 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 Fig. 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 Fig. 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 Fig. 4c).

A recent study 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 (Dingler et al. 2020). 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 2 species to obtain

corrected mutational patterns (see Fig. 5, a and b). Applying this adjustment is necessary to obtain the corrected mutational pattern for a more accurate comparison between species. 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. (2020) (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 from mouse 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 Fig. 5c). We also noted some differences among the various mouse patterns themselves: while the ones from Adh5^{-/-} and

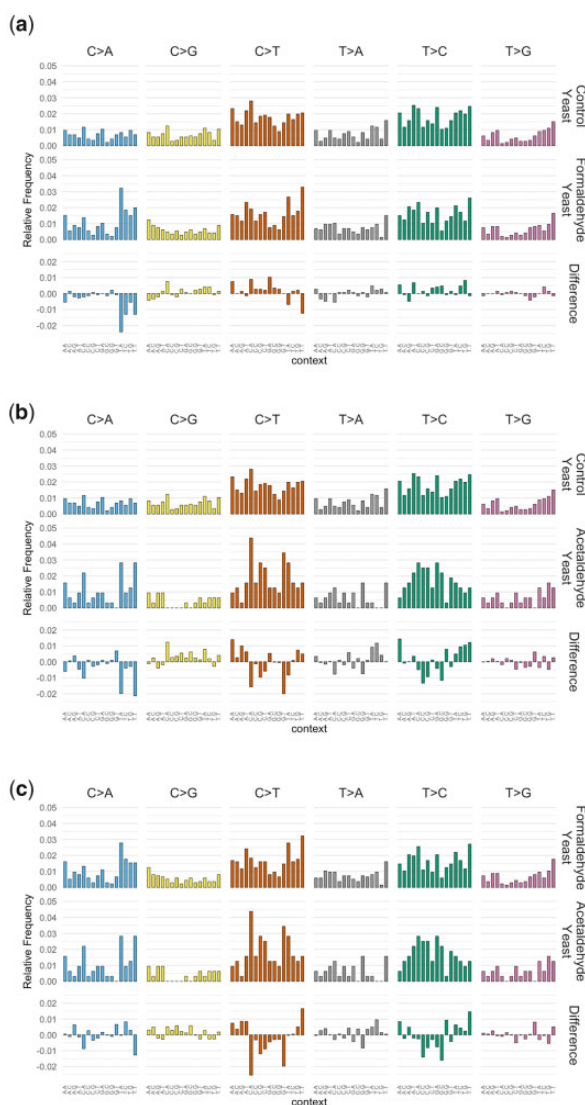


Fig. 4. Comparisons of mutational patterns between (a) controls and formaldehyde; (b) controls and acetaldehyde; and (c) formaldehyde and acetaldehyde.

$Aldh2^{-/-} Adh5^{-/-}$ had cosine similarity = 0.887, the $Aldh2^{-/-}$ profile was noticeably more dissimilar (cosine similarity < 0.8 vs. the other 2 profiles, see Table 2). 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. (2020) and confirmed that none of the mouse profiles showed a particularly close resemblance to any known COSMIC signature (see Fig. 6a and Supplementary Table 4). 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.

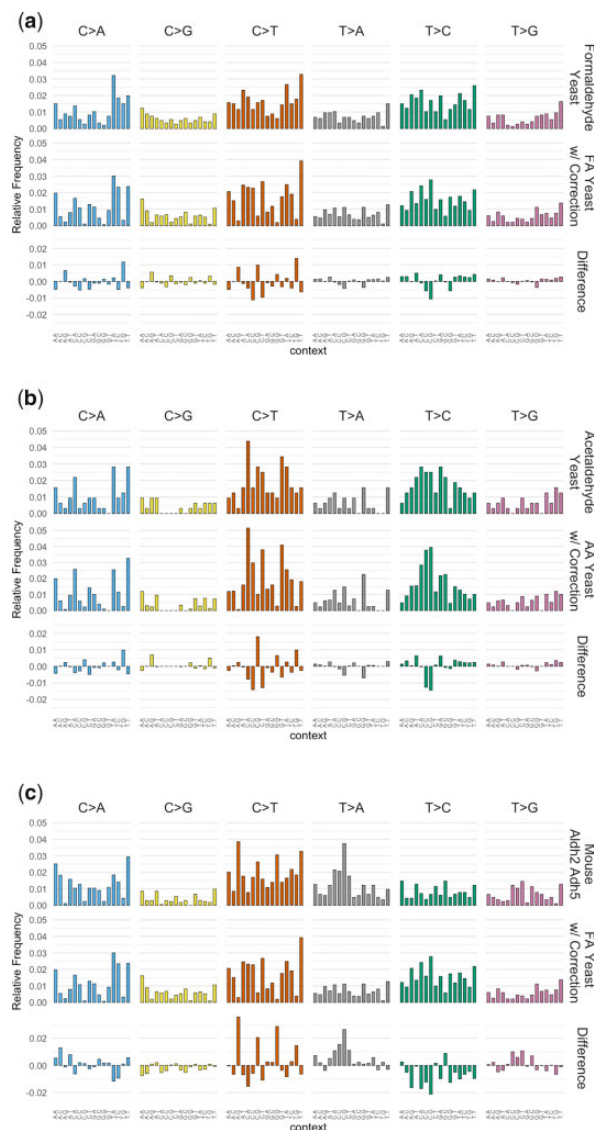


Fig. 5. Comparisons of mutational patterns between (a) formaldehyde with and without correction for trinucleotide frequencies in mouse; (b) acetaldehyde with and without correction for trinucleotide frequencies in mouse; and (c) $Aldh2 Adh5$ -deficient mouse cells and trinucleotide frequency corrected yeast treated with formaldehyde.

Table 1. Cosine similarity values between mutational profiles of (FA) formaldehyde- and (AA) acetaldehyde-mutagenized yeast (with correction for trinucleotide frequencies in mouse) and of mice deficient for aldehyde detoxification genes from (Dingler et al. 2020).

| | 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 |

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 Fig. 6a and Supplementary Table 4). 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 Fig. 6, a and b). The second closest match was to SBS5 (cosine similarity=0.864, see Fig. 6, a

and c). We previously studied an SBS5-like mutational pattern in yeast and showed that similar patterns are widely conserved in many species. The no-aldehyde control mutational pattern was indeed SBS5-like (cosine similarity=0.907, see Fig. 6a and Supplementary Table 4). 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 (Gelova et al. 2020). 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.

Table 2. Cosine similarity values among mice deficient for aldehyde detoxification genes from (Dingler et al. 2020).

| | 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 |

Discussion

In this article, we report the use of a sensitive ssDNA-based mutagenesis reporter system to characterize the mutagenic properties of 2 small aldehydes, formaldehyde and acetaldehyde.

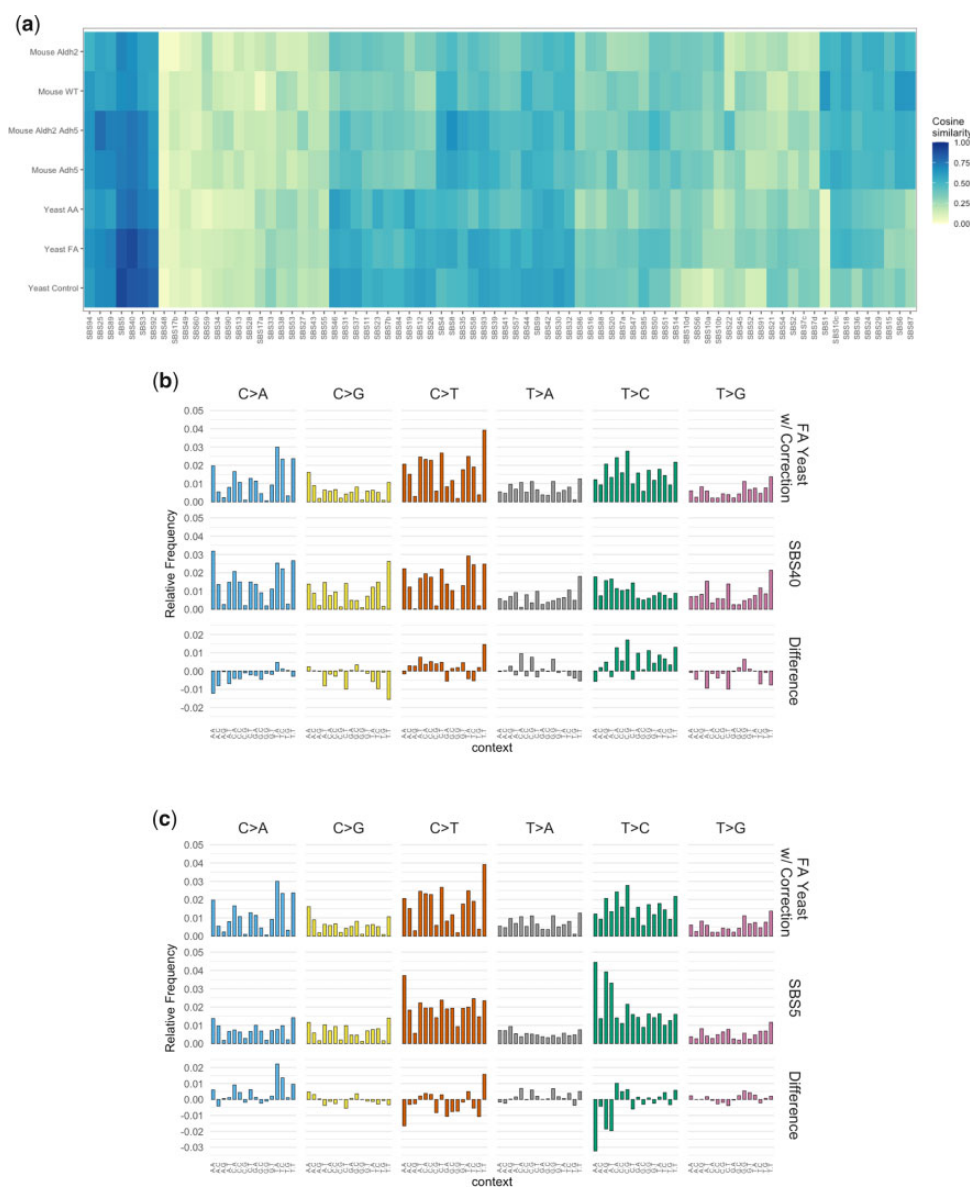


Fig. 6. a) Cosine similarity heatmap with hierarchical clustering comparing COSMIC SBS signatures vs. mouse mutational patterns from Dingler et al.; and vs. trinucleotide abundance-corrected mutational patterns in yeast from acetaldehyde, formaldehyde, and no-aldehyde control. b) Comparison of corrected formaldehyde mutational pattern in yeast vs. SBS signature 40. c) Comparison of corrected formaldehyde mutational pattern in yeast vs. SBS signature 5.

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 (Kucab et al. 2019; Dingler et al. 2020). In addition to being a more sensitive reporter system, it is considerably more cost-efficient to sequence compact yeast genomes (each ~12 Mb) than mammalian genomes, which are much larger (~3 Gb). 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 that the single-stranded configuration of the DNA precludes repair processes requiring 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. When conventional systems for studying mutagenesis do not yield clear-cut results, an ssDNA-enriched assay system can be a useful complementary approach.

It is also important to acknowledge the limitations of this system. First, the initial identification of isolates of interest requires selection for reporter gene inactivation. This selection will necessarily reveal recurrent mutational hotspot mutations when isolates are sequenced (Rogozin and Pavlov 2003). To avoid bias to a mutational pattern due to selection, it is possible to filter out variant calls that map to the reporter genes, although this could mean discarding a significant fraction of variants. Alternatively, it is possible to essentially count mutated motifs: if a mutagen does preferentially mutate a given trinucleotide, then multiple instances of that trinucleotide would be mutated at different genomic loci, as opposed to a recurrent hotspot due to selection. Another limitation is the haploidy of the system. While this facilitates identification of isolates enriched for ssDNA exposure, there is a tradeoff that haploids are not as buffered against potentially deleterious variants as diploids.

The 2 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 (Matsufuji et al. 2008). 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 (Crosby et al. 1988; Ohta et al. 1999; Yasui et al. 2001; Liu et al. 2006; Stein et al. 2006; Upton et al. 2006a,b). Interestingly, acetaldehyde induces an excess of deletion variants of 5 or more bases in our system, but formaldehyde does not, consistent with previous reports (Yasui et al. 2001; Garaycochea et al. 2018). These various mutagenic characteristics of formaldehyde and acetaldehyde reflect their chemical similarities and differences. A limitation of this study is that relatively few acetaldehyde-mutagenized genomes were sequenced, due to budgetary

constraints. Despite this, the considerations just discussed lend credence to overall validity of the findings.

The 96-channel mutational patterns of formaldehyde and acetaldehyde revealed further differences between the 2 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 (International Agency for Research on Cancer 2012b), 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 (Alexandrov et al. 2020), 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 (National Toxicology Program 2014), 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 cannot) 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.

Data availability

Sequencing reads were uploaded to the NCBI SRA (National Center for Biotechnology Information Sequence Read Archive), accessions PRJNA839792 and PRJNA574140. Details on each sequencing sample are listed in [Supplementary Table 1](#). The ySR127 reference genome is available on NCBI Assembly (accession GCA_001051215.1).

[Supplemental material](#) is available at G3 online.

Acknowledgments

The authors thank S. Gelova, B. Xhialli, and N. Liang for their critical feedback on this work.

Author contributions

All authors performed experiments and analyzed/discussed the data; KC conceived and supervised the study and wrote the draft manuscript.

Funding

The authors gratefully acknowledge funding support from a Tier 2 Canada Research Chair (950-231842), a Natural Sciences and Engineering Research Council of Canada Discovery Grant (05973/RGPIN/2017), an Ontario Early Researcher Award (ER17-13-013), and uOttawa startup funding to KC.

Conflicts of interest

None declared.

Literature cited

- Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P. DNA, chromosomes and genomes. In: *Molecular Biology of the Cell*. 6th ed. New York: W.W. Norton & Co; 2014. pp. 173–236.
- Alexandrov LB, Ju YS, Haase K, Van Loo P, Martincorena I, Nik-Zainal S, Totoki Y, Fujimoto A, Nakagawa H, Shibata T, et al. Mutational signatures associated with tobacco smoking in human cancer. *Science*. 2016;354(6312):618–622. doi:10.1126/science.aag0299.
- Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, Wu Y, Boot A, Covington KR, Gordenin DA, Bergstrom EN, et al.; PCAWG Consortium. The repertoire of mutational signatures in human cancer. *Nature*. 2020;578(7793):94–101. doi:10.1038/s41586-020-1943-3.
- Alexandrov LB, Nik-Zainal S, Wedge DC, Campbell PJ, Stratton MR. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep*. 2013;3(1):246–259. doi:10.1016/j.celrep.2012.12.008.
- Ames BN, Shigenaga MK, Hagen TM. Oxidants, antioxidants, and the degenerative diseases of aging. *Proc Natl Acad Sci U S A*. 1993;90(17):7915–7922. doi:10.1073/pnas.90.17.7915.
- Blokzijl F, Janssen R, van Boxtel R, Cuppen E. Mutational Patterns: comprehensive genome-wide analysis of mutational processes. *Genome Med*. 2018;10(1):33. doi:10.1186/s13073-018-0539-0.
- Chan K. Molecular genetic characterization of mutagenesis using a highly sensitive single-stranded DNA reporter system in budding yeast. In: M Muzi-Falconi, GW Brown, editors. *Genome Instability: Methods and Protocols*. New York, NY: Springer; 2018. p. 33–42. https://doi.org/10.1007/978-1-4939-7306-4_4.
- Chan K, Resnick MA, Gordenin DA. The choice of nucleotide inserted opposite abasic sites formed within chromosomal DNA reveals the polymerase activities participating in translesion DNA synthesis. *DNA Repair (Amst)*. 2013;12(11):878–889. doi:10.1016/j.dnarep.2013.07.008.
- Chan K, Roberts SA, Klimczak LJ, Sterling JF, Saini N, Malc EP, Kim J, Kwiatkowski DJ, Fargo DC, Mieczkowski PA, et al. An APOBEC3A hypermutation signature is distinguishable from the signature of background mutagenesis by APOBEC3B in human cancers. *Nat Genet*. 2015;47(9):1067–1072.
- Chan K, Sterling JF, Roberts SA, Bhagwat AS, Resnick MA, Gordenin DA. Base damage within single-strand DNA underlies in vivo hypermutability induced by a ubiquitous environmental agent. *PLoS Genet*. 2012;8(12):e1003149. doi:10.1371/journal.pgen.1003149.
- Crosby RM, Richardson KK, Craft TR, Benforado KB, Liber HL, Skopek TR. Molecular analysis of formaldehyde-induced mutations in human lymphoblasts and *E. coli*. *Environ Mol Mutagen*. 1988;12(2):155–166. doi:10.1002/em.2860120202.
- De Bont R, van Larebeke N. Endogenous DNA damage in humans: a review of quantitative data. *Mutagenesis*. 2004;19(3):169–185. doi:10.1093/mutage/geh025.
- Degtyareva NP, Heyburn L, Sterling J, Resnick MA, Gordenin DA, Doetsch PW. 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*. 2013;41(19):8995–9005. doi:10.1093/nar/gkt671.
- Dingler FA, Wang M, Mu A, Millington CL, Oberbeck N, Watcham S, Pontel LB, Kamimae-Lanning AN, Langevin F, Nadler C, et al. Two aldehyde clearance systems are essential to prevent lethal formaldehyde accumulation in mice and humans. *Mol Cell*. 2020;80(6):996–1012.e9. doi:10.1016/j.molcel.2020.10.012.
- Garaycochea JI, Crossan GP, Langevin F, Mulderrig L, Louzada S, Yang F, Guilbaud G, Park N, Roerink S, Nik-Zainal S, et al. Alcohol and endogenous aldehydes damage chromosomes and mutate stem cells. *Nature*. 2018;553(7687):171–177. <http://dx.doi.org/10.1038/nature25154>.
- 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*. 1995;15(11):6128–6138. doi:10.1128/MCB.15.11.6128.
- Gelova SP, Doherty KN, Alasmar S, Chan K. Intrinsic base substitution patterns in diverse species reveal links to cancer and metabolism. *bioRxiv* 758540. doi:10.1101/758540, 2020.
- Greenblatt MS, Bennett WP, Hollstein M, Harris CC. Mutations in the p53 tumor suppressor gene: clues to cancer etiology and molecular pathogenesis. *Cancer Res*. 1994;54(18):4855–4878.
- Grogan D, Jinks-Robertson S. Formaldehyde-induced mutagenesis in *Saccharomyces cerevisiae*: molecular properties and the roles of repair and bypass systems. *Mutat Res*. 2012;731(1–2):92–98. doi:10.1016/j.mrfmmm.2011.12.004.
- Haber JE. DNA recombination: the replication connection. *Trends Biochem Sci*. 1999;24(7):271–275. doi:10.1016/S0968-0004(99)01413-9.
- Helbock HJ, Beckman KB, Shigenaga MK, Walter PB, Woodall AA, Yeo HC, Ames BN. DNA oxidation matters: the HPLC-electrochemical detection assay of 8-oxo-deoxyguanosine and 8-oxo-guanine. *Proc Natl Acad Sci U S A*. 1998;95(1):288–293.
- Ikehata H, Ono T. The mechanisms of UV mutagenesis. *J Radiat Res*. 2011;52(2):115–125. doi:10.1269/jrr.10175.
- International Agency for Research on Cancer. Identification of Research Needs to Resolve the Carcinogenicity of High-Priority IARC Carcinogens. Lyons, France: IARC Technical Publications; 2010.
- International Agency for Research on Cancer. Formaldehyde. In: *Chemical Agents and Related Occupations*. Vol. 100F. Lyons, France: IARC Monographs on the Evaluation of Carcinogenic Risks to Humans; 2012a. p. 401–436.
- International Agency for Research on Cancer. Personal Habits and Indoor Combustions. Lyons, France: IARC Monographs on the Evaluation of Carcinogenic Risks to Humans; 2012b.
- Irigaray P, Belpomme D. Basic properties and molecular mechanisms of exogenous chemical carcinogens. *Carcinogenesis*. 2010;31(2):135–148. doi:10.1093/carcin/bgp252.
- Kadlubar FF, Anderson KE, Häussermann S, Lang NP, Barone GW, Thompson PA, MacLeod SL, Chou MW, Mikhailova M, Plastaras J, et al. Comparison of DNA adduct levels associated with oxidative stress in human pancreas. *Mutat Res Mol Mech Mutagen*. 1998;405(2):125–133. doi:10.1016/S0027-5107(98)00129-8.
- Keszenman DJ, Kolodiuk L, Baulch JE. DNA damage in cells exhibiting radiation-induced genomic instability. *Mutagenesis*. 2015;30(3):451–458. doi:10.1093/mutage/gev006.
- Kimsey IJ, Szymanski ES, Zahurancik WJ, Shakya A, Xue Y, Chu C-C, Sathyamoorthy B, Suo Z, Al-Hashimi HM. Dynamic basis for dG•dT misincorporation via tautomerization and ionization. *Nature*. 2018;554(7691):195–201.
- Kucab JE, Zou X, Morganella S, Joel M, Nanda AS, Nagy E, Gomez C, Degasperi A, Harris R, Jackson SP, et al. A compendium of

- mutational signatures of environmental agents. *Cell*. 2019;177(4):821–836.e16. doi:10.1016/j.cell.2019.03.001.
- Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods*. 2012;9(4):357–359.
- Letouzé E, Shinde J, Renault V, Couchy G, Blanc J-F, Tubacher E, Bayard Q, Bacq D, Meyer V, Semhoun J, et al. Mutational signatures reveal the dynamic interplay of risk factors and cellular processes during liver tumorigenesis. *Nat Commun*. 2017;8(1):1315. doi:10.1038/s41467-017-01358-x.
- Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*. 2011;27(21):2987–2993. doi:10.1093/bioinformatics/btr509.
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. The sequence alignment/map format and SAMtools. *Bioinformatics*. 2009;25(16):2078–2079. doi:10.1093/bioinformatics/btp352.
- Lindahl T. DNA repair enzymes acting on spontaneous lesions in DNA. In: *DNA Repair Processes*. Miami: Symposia Specialists; 1977. pp. 225–240.
- Lindahl T. Instability and decay of the primary structure of DNA. *Nature*. 1993;362(6422):709–715. doi:10.1038/362709a0.
- Lindahl T, Nyberg B. Rate of depurination of native deoxyribonucleic acid. *Biochemistry*. 1972;11(19):3610–3618. doi:10.1021/bi00769a018.
- Liu X, Lao Y, Yang I-Y, Hecht SS, Moriya M. Replication-coupled repair of crotonaldehyde/acetaldehyde-induced guanine–guanine interstrand cross-links and their mutagenicity. *Biochemistry*. 2006;45(42):12898–12905. doi:10.1021/bi060792v.
- Masoodi HR, Bagheri S, Abareghi M. The effects of tautomerization and protonation on the adenine–cytosine mismatches: a density functional theory study. *J Biomol Struct Dyn*. 2016;34(6):1143–1155. doi:10.1080/07391102.2015.1072734.
- Matsufuji Y, Fujimura S, Ito T, Nishizawa M, Miyaji T, Nakagawa J, Ohyama T, Tomizuka N, Nakagawa T. Acetaldehyde tolerance in *Saccharomyces cerevisiae* involves the pentose phosphate pathway and oleic acid biosynthesis. *Yeast*. 2008;25(11):825–833. doi:10.1002/yea.1637.
- Moriya M, Slade N, Brdar B, Medverec Z, Tomic K, Jelaković B, Wu L, Truong S, Fernandes A, Grollman AP. TP53 mutational signature for aristolochic acid: an environmental carcinogen. *Int J Cancer*. 2011;129(6):1532–1536. doi:10.1002/ijc.26077.
- Nakamura J, Walker VE, Upton PB, Chiang S-Y, Kow YW, Swenberg JA. Highly sensitive apurinic/aprimidinic site assay can detect spontaneous and chemically induced depurination under physiological conditions. *Cancer Res*. 1998;58(2):222–225.
- National Toxicology Program. Formaldehyde. In: *Report on Carcinogens*. 13th ed. Research Triangle Park, NC: U.S. Department of Health and Human Services, Public Health Service; 2014.
- Neddermann P, Gallinari P, Lettieri T, Schmid D, Truong O, Hsuan JJ, Wiebauer K, Jiricny J. Cloning and expression of human G/T mismatch-specific thymine-DNA glycosylase. *J Biol Chem*. 1996;271(22):12767–12774. doi:10.1074/jbc.271.22.12767.
- 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*. 1999;440(1):59–74. doi:10.1016/S1383-5718(99)00005-4.
- Page's H, Aboyoun P, Gentleman R, DebRoy S. Biostrings: Efficient Manipulation of Biological Strings; 2022. [accessed 2022 September 13]. <https://bioconductor.org/packages/Biostrings>.
- 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*. 2000;104(31):7346–7352. doi:10.1021/jp000740u.
- R Core Team. R: The R Project for Statistical Computing; 2020 [accessed 2020 Mar 11]. <https://www.r-project.org/>.
- Rogozin IB, Pavlov YI. Theoretical analysis of mutation hotspots and their DNA sequence context specificity. *Mutat Res*. 2003;544(1):65–85. doi:10.1016/S1383-5742(03)00032-2.
- 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*. 1998;19(9):989–1000. doi:10.1002/(SICI)1096-987X(19980715)19:9<989::AID-JCC1>3.0.CO;2-F.
- Saini N, Sterling JF, Sakofsky CJ, Giacobone CK, Klimczak LJ, Burkholder AB, Malc EP, Mieczkowski PA, Gordenin DA. Mutation signatures specific to DNA alkylating agents in yeast and cancers. *Nucleic Acids Res*. 2020;48:3692–3707. <https://doi.org/10.1093/nar/gkaa150>.
- Saparbav MK, Zharkov DO. Glycosylase repair. In: *Reference Module in Life Sciences*. Elsevier; 2017. [accessed 2022 September 13]. <http://www.sciencedirect.com/science/article/pii/B9780128096338064815>.
- Sassa A, Kanemaru Y, Kamoshita N, Honma M, Yasui M. Mutagenic consequences of cytosine alterations site-specifically embedded in the human genome. *Genes Environ*. 2016;38(1):17. doi:10.1186/s41021-016-0045-9.
- Secretan B, Straif K, Baan R, Grosse Y, El Ghissassi F, Bouvard V, Benbrahim-Tallaa L, Guha N, Freeman C, Galichet L, et al.; WHO International Agency for Research on Cancer Monograph Working Group. A review of human carcinogens—Part E: tobacco, areca nut, alcohol, coal smoke, and salted fish. *Lancet Oncol*. 2009;10(11):1033–1034. doi:10.1016/s1470-2045(09)70326-2.
- Stein S, Lao Y, Yang I-Y, Hecht SS, Moriya M. Genotoxicity of acetaldehyde- and crotonaldehyde-induced 1,N2-propanodeoxyguanosine DNA adducts in human cells. *Mutat Res*. 2006;608(1):1–7. doi:10.1016/j.mrgentox.2006.01.009.
- Tice RR, Setlow RB. DNA repair and replication in aging organisms and cells. In: Finch CE, Schneider EL, editors. *Handbook of the Biology of Aging*. New York: Van Nostrand Reinhold; 1985. pp. 173–224.
- Upton DC, Wang X, Blans P, Perrino FW, Fishbein JC, Akman SA. Replication of N2-ethyldeoxyguanosine DNA adducts in the human embryonic kidney cell line 293. *Chem Res Toxicol*. 2006a;19(7):960–967. doi:10.1021/tx060084a.
- Upton DC, Wang X, Blans P, Perrino FW, Fishbein JC, Akman SA. Mutagenesis by exocyclic alkylamino purine adducts in *Escherichia coli*. *Mutat Res*. 2006b;599(1–2):1–10. doi:10.1016/j.mrfmmm.2005.12.014.
- VanderVeen LA, Hashim MF, Shyr Y, Marnett LJ. Induction of frameshift and base pair substitution mutations by the major DNA adduct of the endogenous carcinogen malondialdehyde. *Proc Natl Acad Sci U S A*. 2003;100(24):14247–14252. doi:10.1073/pnas.2332176100.
- Vilenchik MM, Knudson AG. Endogenous DNA double-strand breaks: production, fidelity of repair, and induction of cancer. *Proc Natl Acad Sci U S A*. 2003;100(22):12871–12876. doi:10.1073/pnas.2135498100.
- Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, Grolmund G, Hayes A, Henry L, Hester J, et al. Welcome to the tidyverse. *JOSS*. 2019;4(43):1686. doi:10.21105/joss.01686.
- Yasui M, Matsui S, Ihara M, Laxmi YRS, Shibutani S, Matsuda T. 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*. 2001;29(9):1994–2001.