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

Authors:

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

Affiliations:

[1] Department of Biochemistry, Microbiology and Immunology, University of Ottawa

[2] Ottawa Institute of Systems Biology, University of Ottawa

[3] Biopharmaceutical Sciences Undergraduate Program, University of Ottawa

* Corresponding Author: kin.chan@uottawa.ca

† Equal Contribution

Current affiliation:

§ Department of Chemistry and Biomolecular Sciences, University of Ottawa

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, as they are known to damage DNA and cause mutations. 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 mostly C/G → A/T, C/G → T/A, and T/A → C/G substitutions, each in similar proportions. We also 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 cancer mutational signature of unknown etiology. We also noted that acetaldehyde treatment caused an excess number of deletion events longer than four bases while formaldehyde did not. This latter result could be a distinguishing feature between the mutational patterns of these two simple aldehydes. These findings shed new light on the mutagenic characteristics of two important carcinogens.

Introduction

DNA is constantly damaged by intracellular processes (1) and exposure to exogenous damaging agents (2–4). Intracellular DNA damaging processes include: oxidation (5,6); glycosidic bond breakage (7–11); single- and double-stranded breaks (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); and tobacco (20). Mutations are also thought to result from ionization or isomerization (i.e., tautomerization) of DNA bases, which can cause changes in base pairing properties (21–24).

These processes do not affect all bases equally. Each of the four bases has its own unique set of chemically reactive moieties (amines, carbonyls, labile ring atoms) (25). For any DNA damaging process, base(s) with moieties that readily react will be damaged more frequently than bases without such moieties. Local sequence context can also be a determinant of vulnerability to damage. Mutational signatures reflect this specificity: these signatures arise naturally because each DNA damaging process specifically affects certain bases in specific contexts more frequently than others (26).

Mutational signatures are inferred using a non-negative matrix factorization (NMF) algorithm (27). NMF takes a mutational dataset as input. It initiates by essentially guessing a solution set of constituent signatures, and then computes the error when reconstructing the original dataset. NMF then tries a slightly different solution set and recomputes the error. This process loops until finding an optimal solution set that minimizes reconstruction error. A globally stable solution set is found when many

different initial conditions (i.e., first guesses) all converge to yield that solution set. NMF analysis can extract reproducible, recurrent patterns of mutations, which reflect distinct DNA damaging processes.

Previously, the International Agency for Research on Cancer (IARC) named a number of high-priority carcinogens that required further research (28). 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 (29,30). Formaldehyde is also produced endogenously in cells, as a major metabolic by-product from amino acid catabolism, resulting in high concentrations of ~100 μM in blood (30). Acetaldehyde is a reactive compound that humans are commonly exposed as a result of ethanol consumption. Like formaldehyde, acetaldehyde is also classified as a known human carcinogen (31). Alcohol consumption (and the ensuing generation of acetaldehyde) is associated with higher risk of multiple types of cancer, including: head and neck; esophageal; liver; breast; and colorectal (32).

Understanding the mutagenic characteristics of formaldehyde and acetaldehyde remain important research questions, which can provide valuable insights into the role of these common small aldehydes in carcinogenesis and cancer mutagenesis. Previous attempts to define the mutational patterns induced by formaldehyde and acetaldehyde (e.g., (33,34)) 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 of unknown etiology, namely SBS signature 40.

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 yeast strain (ySR127) that forms long regions of sub-telomeric single-stranded DNA (ssDNA) when shifted to 37°C. 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. This mutagenesis system is well suited to studying weak mutagens, as ssDNA is more prone to mutation than double-stranded DNA. The system was used previously to study the mutagenic properties of bisulfite and human APOBEC3G cytidine deaminase (35); abasic sites (36); as well as human APOBEC3A and APOBEC3B cytidine deaminases (37).

We treated cells with increasing concentrations of formaldehyde or acetaldehyde. 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%. 10 mM formaldehyde resulted in less than 20% viability. Formaldehyde-induced inactivation of *CAN1* was detected for as little as 2 mM treatment (see Figure 1b). Mutagenesis plateaued from 4 to 10 mM formaldehyde, with mutation frequencies of at least 1.5×10^{-3}.

Cells were 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. There 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 cells mutagenized by 100 mM acetaldehyde do not survive.

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

We collected mutagenized isolates for Illumina whole genome sequencing to determine what kinds of mutations are induced by either formaldehyde or acetaldehyde treatment. The genomes mutagenized by either small aldehyde were compared to control genomes that were untreated. Analysis of the untreated control genomes revealed a mutational profile 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). By comparison, formaldehyde and acetaldehyde treatment both caused a relative excess of C/G > A/T transversions (see Figure 2b and 2c). This may be a common characteristic of mutagenesis caused by small aldehydes.

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 numbers 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 additional 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 (see Figure 3c). This is a distinguishing property of acetaldehyde-induced DNA damage in our system.

Formaldehyde and acetaldehyde produce distinct mutational profiles

To investigate the mutational properties of the small aldehydes further, we plotted their mutational profiles in the 96-channel format of the COSMIC mutational signatures. The mutational profile of formaldehyde is similar to untreated controls (cosine similarity = 0.924), but the excess of C/G > A/T transversions is evident (see Figure 4a). The mutational profile of acetaldehyde is more dissimilar to the profile of untreated controls (cosine similarity = 0.855), again with a noticeable excess of C/G > A/T substitutions (see Figure 4b). When comparing the formaldehyde and acetaldehyde profiles directly, the cosine similarity value is 0.864, showing some similarities but also clear differences in the C/G > T/A channels especially (see Figure 4c).

A recent paper described mutational profiles obtained in mice that were genetically deleted for genes important in aldehyde detoxification, ADH5 and ALDH2 (34). To compare our mutational profiles derived from mutagenized yeast genomes to these profiles from mice, we first normalized for differences in trinucleotide abundances to obtained corrected mutational profiles (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 profiles retained high similarity to the original profiles in yeast (cosine similarity values > 0.95).

When we compared the various mutational profiles, we noticed that cosine similarity values between the yeast and mouse profiles are relatively low (see Table 1).

These values are higher when comparing the formaldehyde profiles in yeast to the mouse profiles. A closer examination of these profiles suggests that there are likely to be mutations from other sources mixed in the mouse profiles, e.g. from SBS1 (deamination of 5-methylcytosine at CpG motifs, see Figure 4f). We also noted differences among the various mouse profiles: while the ones from Adh5$^{-/-}$ and Aldh2$^{-/-}$ Adh5$^{-/-}$ had cosine similarity = 0.887, the Aldh2$^{-/-}$ profile was dissimilar (cosine similarity < 0.8 vs. the other two profiles). This is consistent with the hypothesis that there are other mutation sources mixed in with the mouse profiles.

Formaldehyde mutational profile resembles COSMIC SBS signature 40

We then investigated whether these mutational profiles might shed light on the etiology of COSMIC mutational signatures. We started with the mouse profiles published by Dingler et al. (34) and found 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 profiles, they are possibly obscured by being in a mixture of multiple signatures.

Comparison of the acetaldehyde profile from yeast vs. known COSMIC signatures also yielded, at best, cosine similarity of 0.777 to SBS40, a signature of unknown etiology (see Figure 5b). Since we had full control of the mutagenesis experiments using the yeast system, it does not seem likely that other mutagenic processes are obscuring the acetaldehyde-induced pattern. We conclude that the acetaldehyde profile is not a plausible match for any known COSMIC signature at this point.

Finally, we compared the formaldehyde profile to the COSMIC signatures, finding that the closest match is to SBS40, with cosine similarity = 0.890 (see Figures 5c and 5d). The second closest match was to SBS5 (cosine similarity = 0.870, 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 (38). An SBS40-like mutational pattern would require a separate explanation, which would be the addition of exogenous formaldehyde to our experimental system. As such, a plausible etiology for SBS40 in cancers is the mutagenicity of endogenously generated 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 are rare and it is difficult to discern a reliable mutational profile using these few mutations (33). 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 much larger (~3 gigabases). Leveraging these advantages of the ssDNA mutagenesis reporter, we were able to define the mutational profiles of both formaldehyde and acetaldehyde.

The two small aldehydes share some similar characteristics, but also have their differences. Both induce dose-dependent increases in mutagenesis as lower concentrations. But whereas formaldehyde-induced mutagenesis essentially plateaus from 4 mM up to 10 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 higher concentrations. They both cause an excess of C/G > A/T transversions. Interestingly, acetaldehyde induces an excess of insertion variants of five or more bases, but formaldehyde does not. These various mutagenic characteristics of formaldehyde and acetaldehyde reflect their chemical similarities and differences.

The 96-channel mutational profiles of formaldehyde and acetaldehyde revealed further differences between the two aldehydes. Whereas the acetaldehyde profile did not particularly resemble any known COSMIC signature, new mutational signatures might be

revealed as more cancers are sequenced. As alcohol consumption is associated with multiple cancer types and it is thought that acetaldehyde from alcohol detoxification would damage DNA, associated mutational signature(s) may yet be discovered in the future. The formaldehyde profile was similar to SBS signature 40. SBS40 is currently of unknown etiology, but is known to be present in at least 28 cancer types (26), making it the third most common mutational signature in cancers. The high prevalence of SBS40 suggests an endogenous origin for the underlying DNA damage that is present in different cell types throughout the body. When all of the available information is taken into consideration, mutagenesis from endogenously generated formaldehyde is a plausible candidate for the etiology of SBS40.

Comparison with mutational profiles from mice deleted for aldehyde detoxification genes revealed that those profiles are likely mixtures of mutations from different mutagenic processes, and not just from DNA damage due to excess endogenous aldehydes. The contribution from SBS1 (C/G > T/A at CpG motifs) was obvious to see. Mutagenesis from other sources interferes with making an accurate inference of the aldehyde-specific mutagenesis. This is another challenge when using systems for mutational detection that are not (and perhaps can not) be properly controlled to factor out mutagenesis from other sources. We have shown that deployment of a sensitive mutagenesis detection system where the experimenters have direct and complete control over the mutation induction can shed important new light on mutagenesis in cancers.

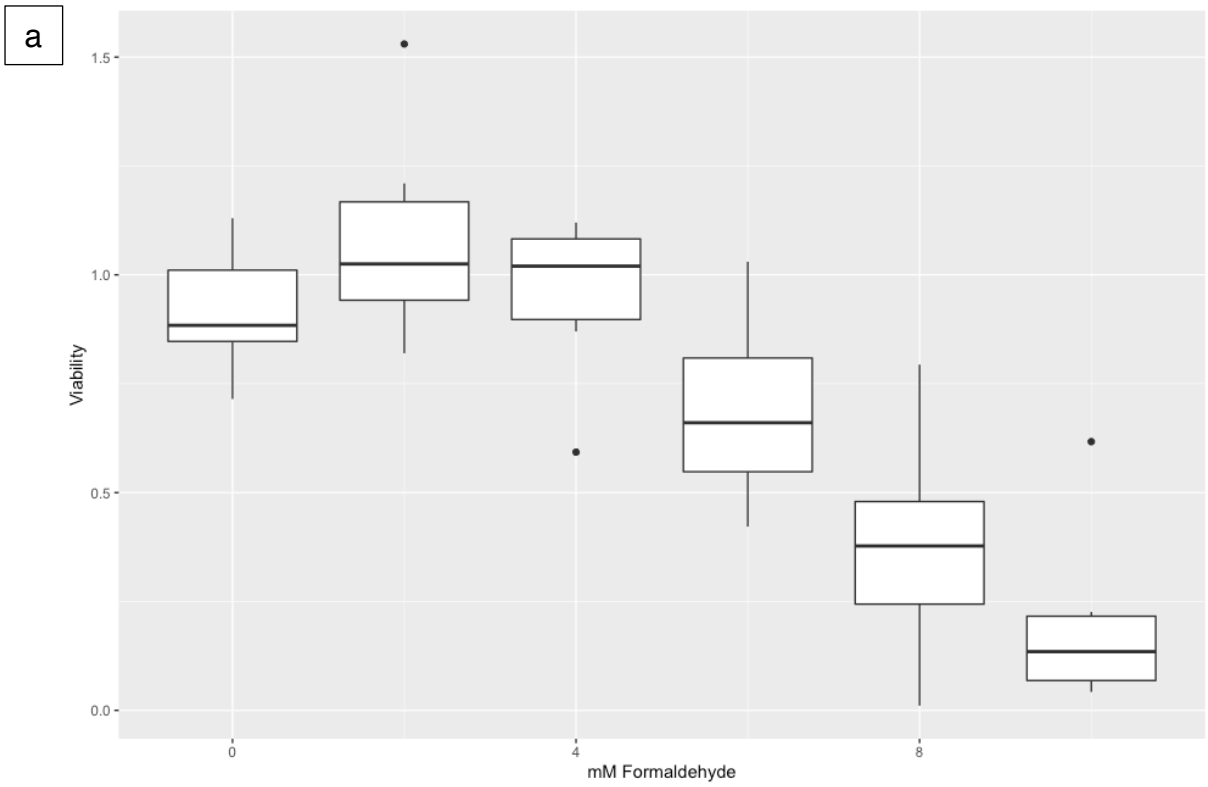
Table 1: Cosine similarity values between mutational profiles of mutagenized yeast (with correction for trinucleotide frequencies in mouse) and of mice deficient for aldehyde detoxification genes from (34).

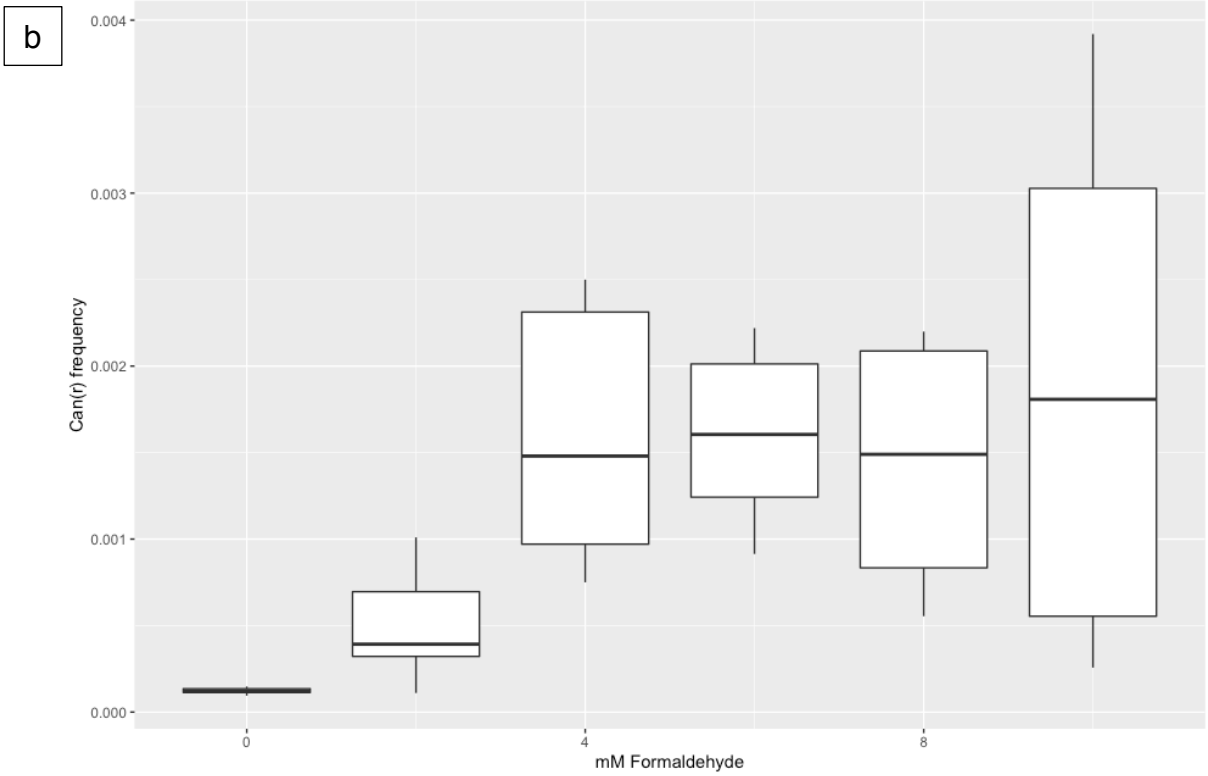
	Mouse Aldh2$^{-/-}$	Mouse Adh5$^{-/-}$	Mouse Aldh2$^{-/-}$ Adh5$^{-/-}$
Yeast FA, corrected	0.654	0.722	0.748
Yeast AA, corrected	0.609	0.618	0.656

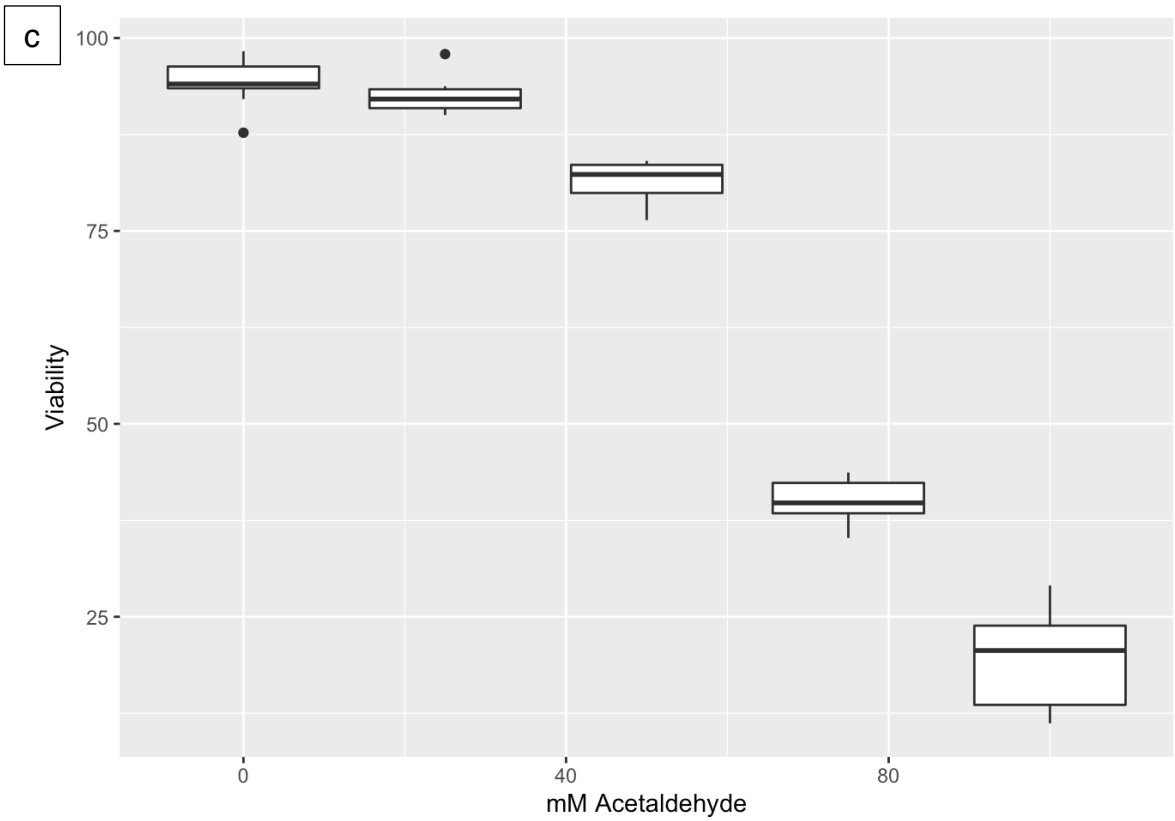
Table 2: Cosine similarity values among mice deficient for aldehyde detoxification genes from (34).

	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

Figure 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.




b
0.004
0.003
0.002
0.001
0.000
Can(r) frequency
0
4
8
mM Formaldehyde


c
100
75
50
25
Viability
0
40
80
mM Acetaldehyde

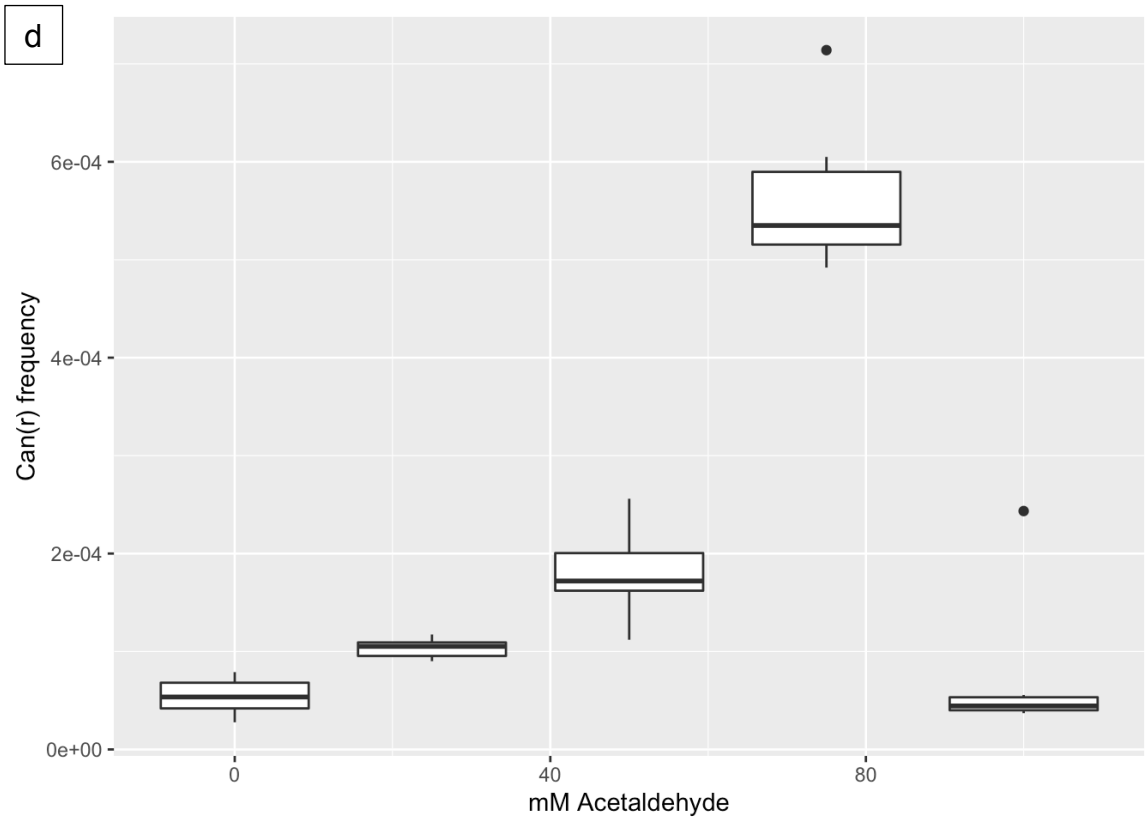
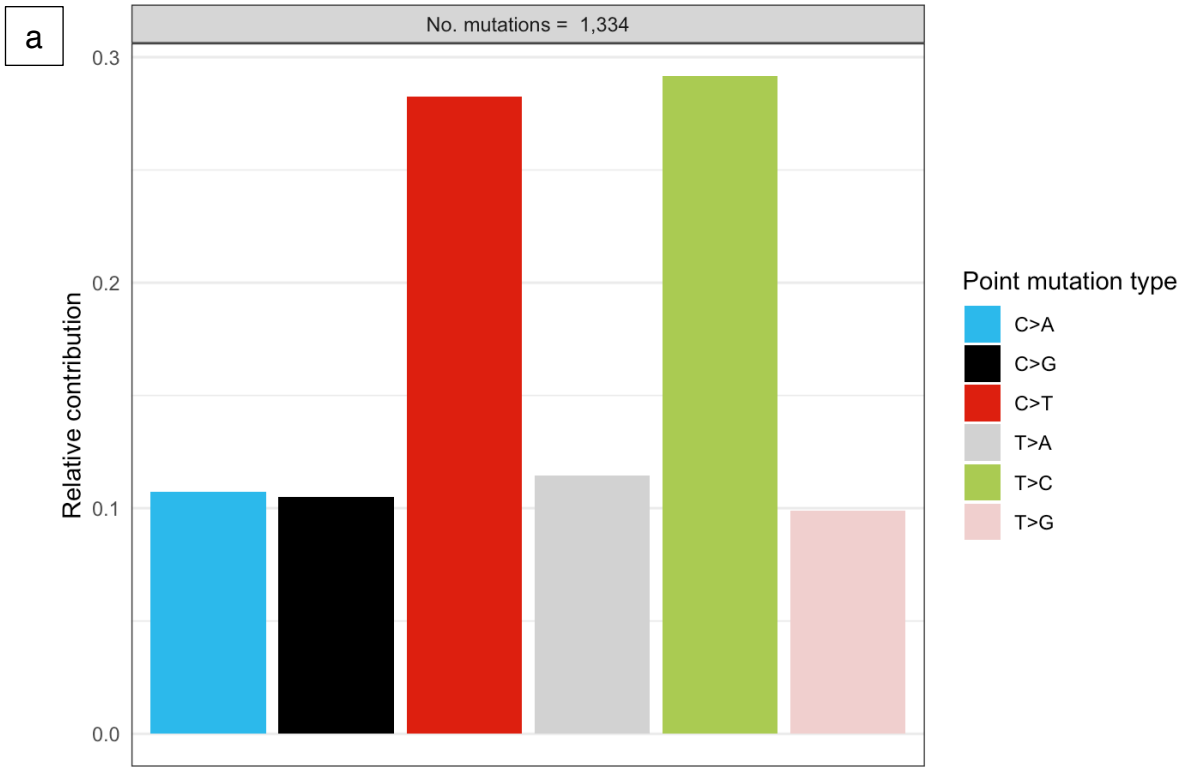
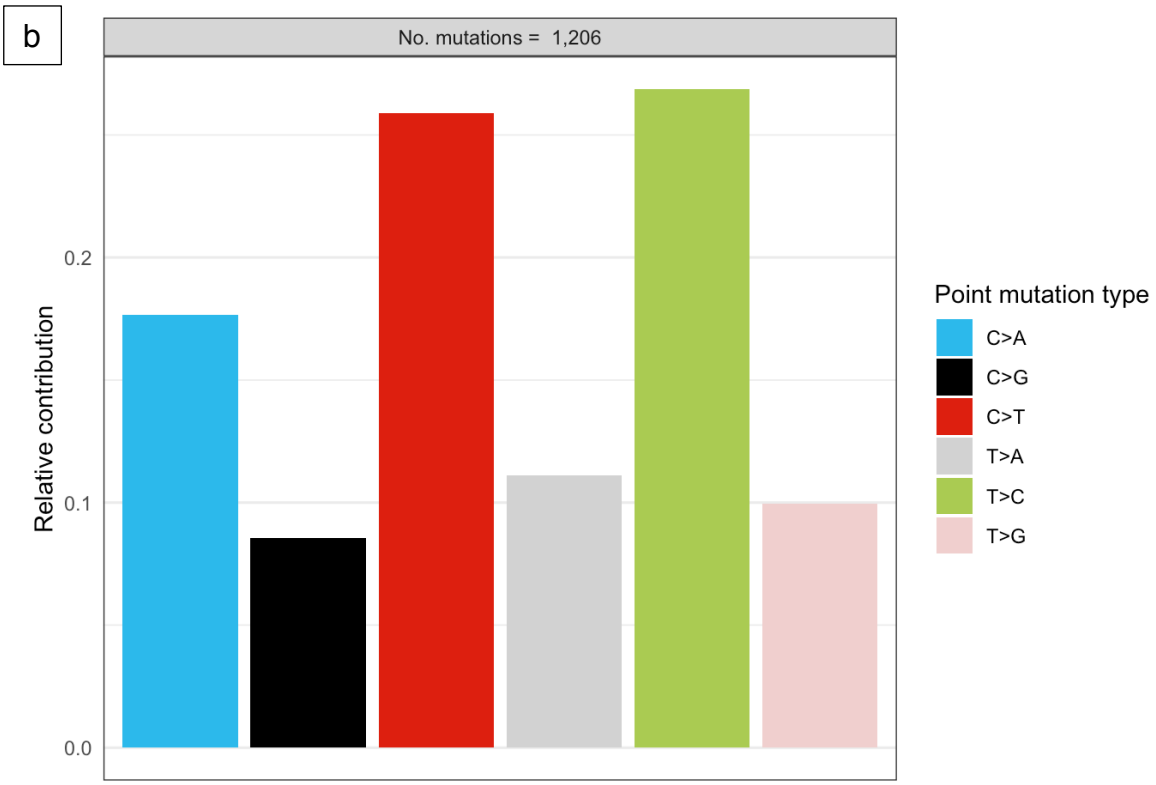

d
6e-04
4e-04
2e-04
0e+00
Can(r) frequency
0
40
80
mM Acetaldehyde

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




b
No. mutations = 1,206
Relative contribution
0.0
0.1
0.2
Point mutation type
C>A
C>G
C>T
T>A
T>C
T>G

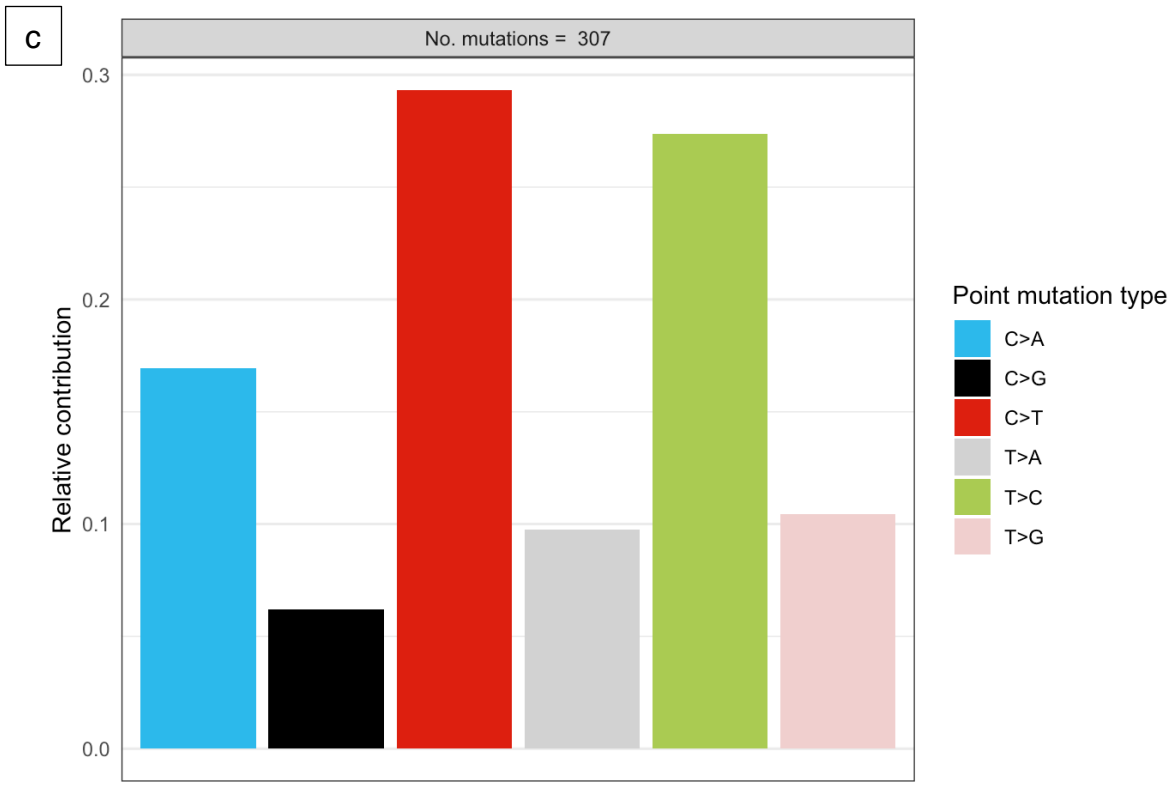
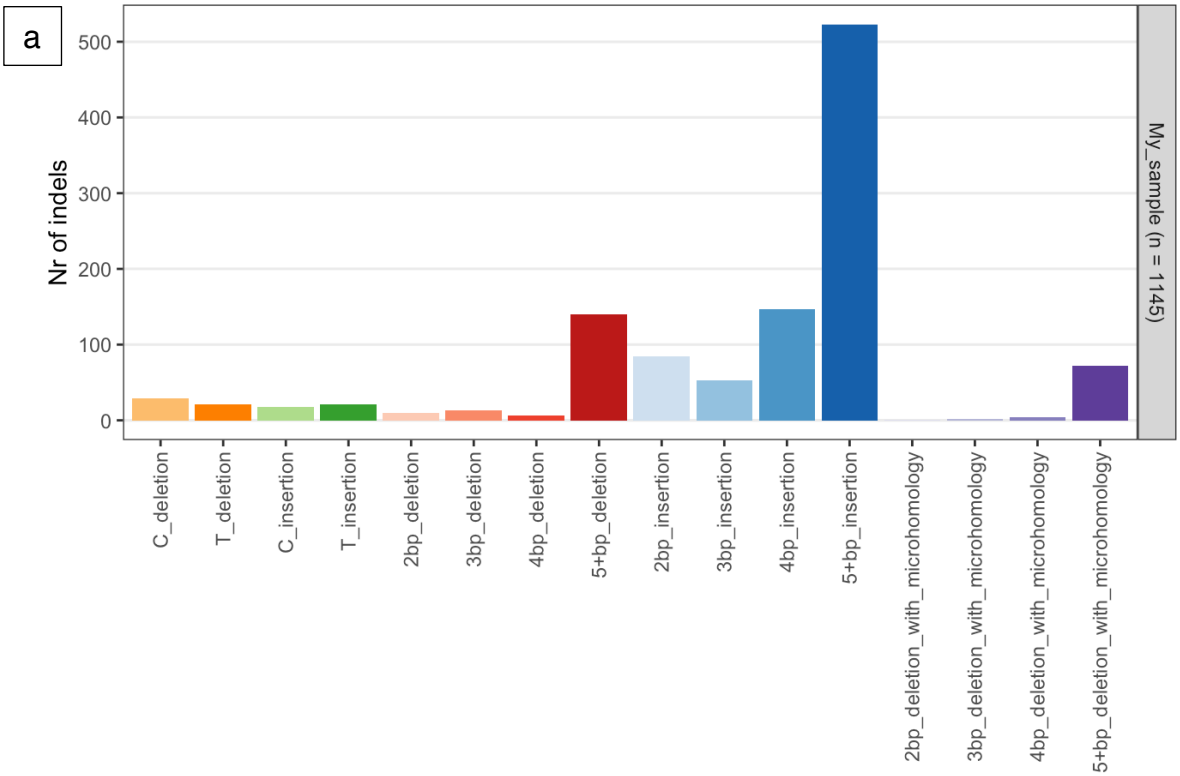

c
No. mutations = 307
Relative contribution
0.0
0.1
0.2
0.3
Point mutation type
C>A
C>G
C>T
T>A
T>C
T>G

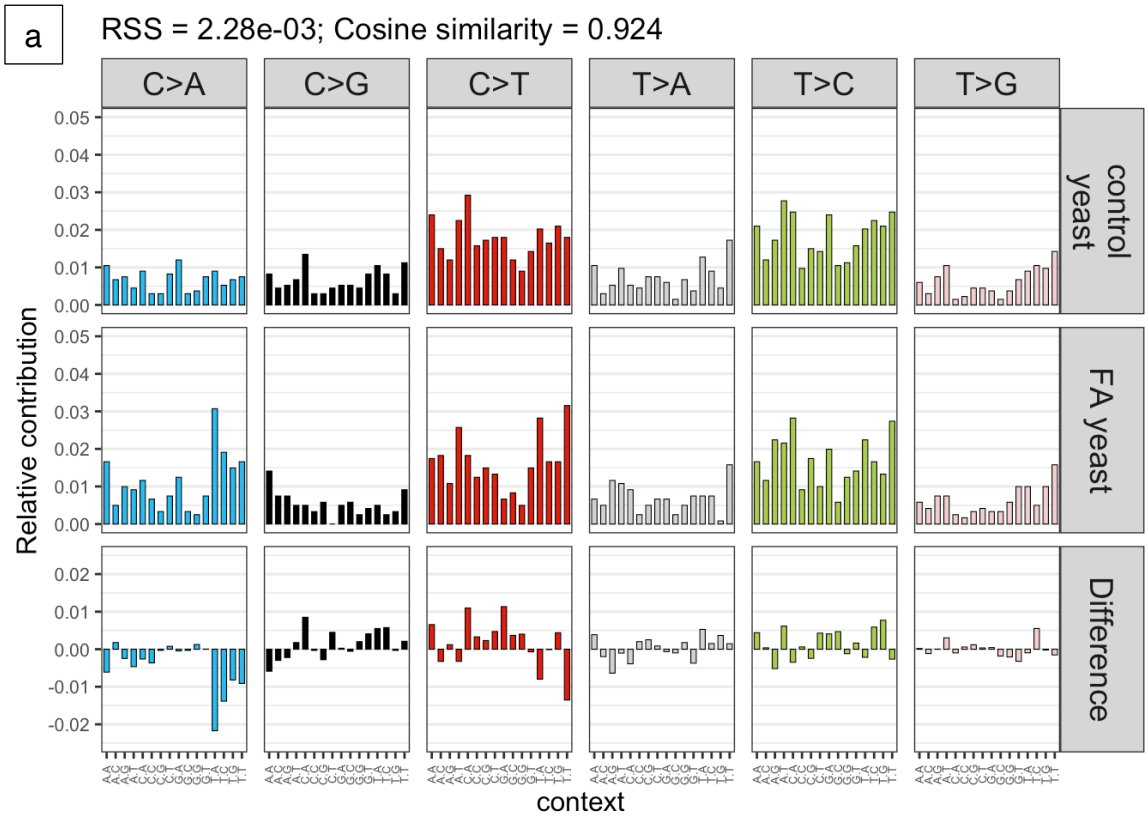
Figure 3: Small indels from (a) untreated controls, (b) formaldehyde, and (c) acetaldehyde.

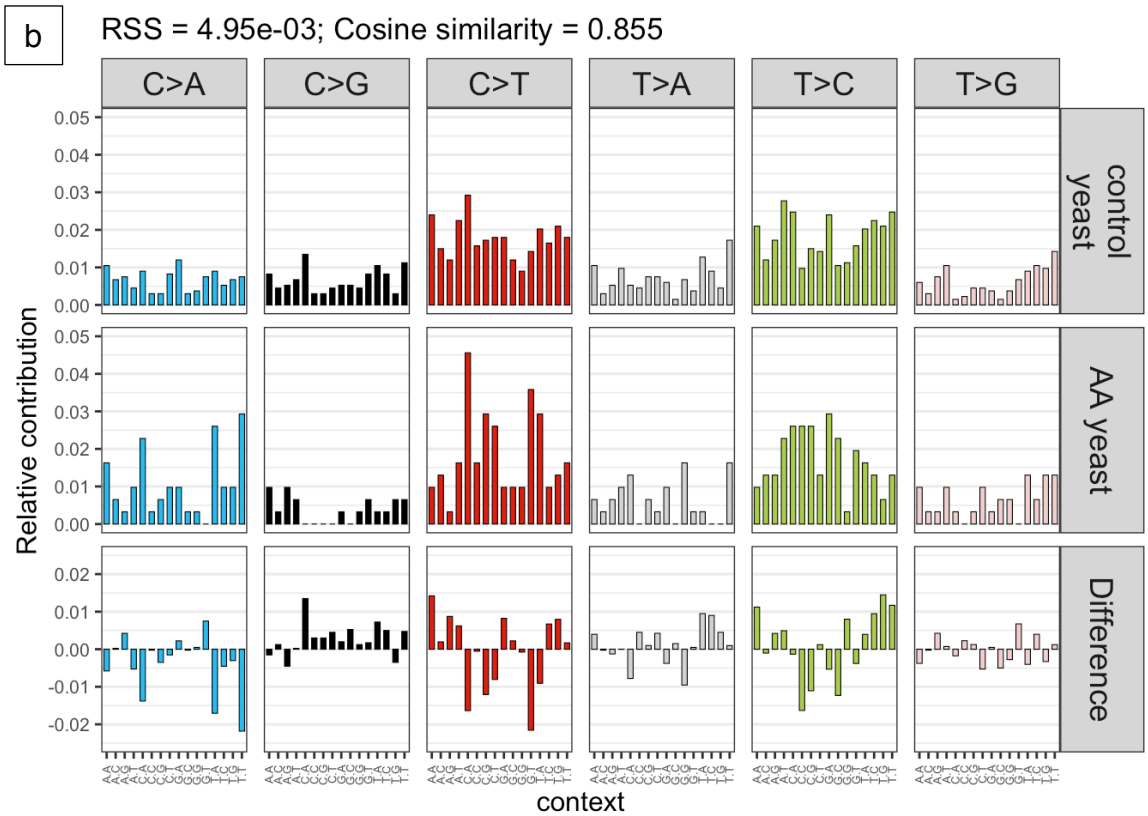


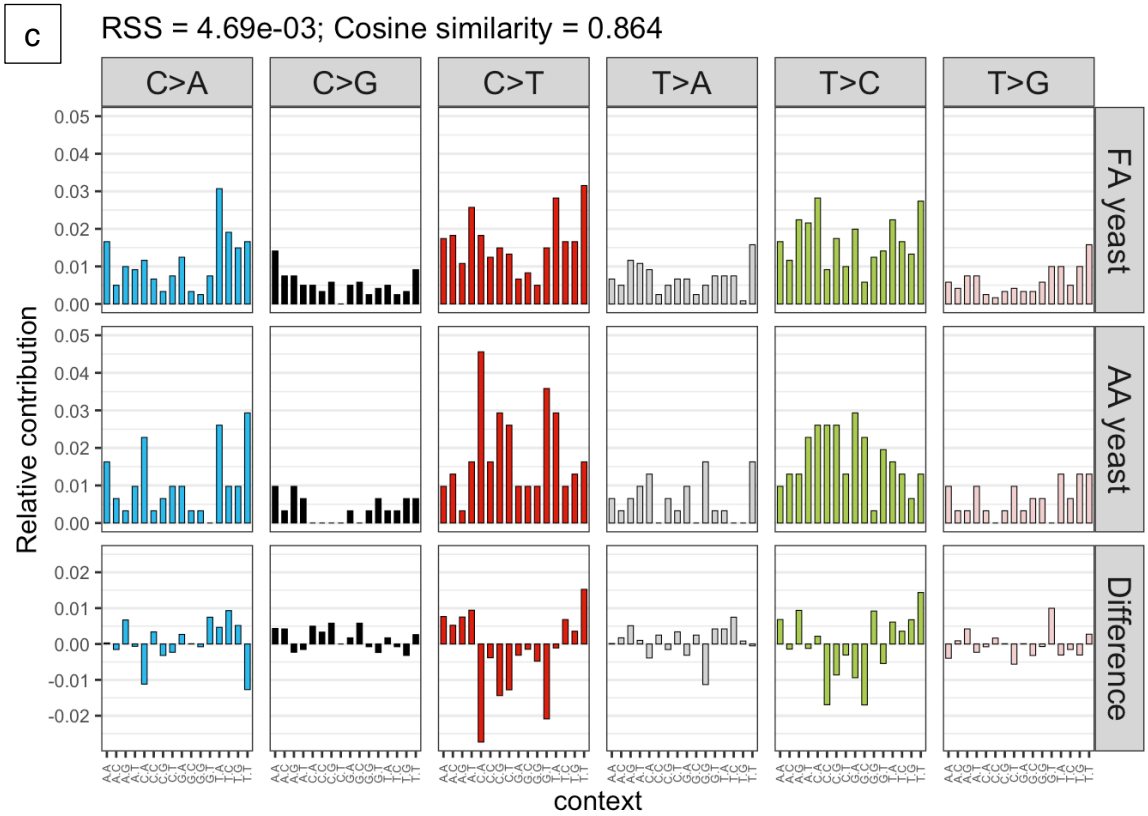
b
Nr of indels
600
400
200
0
C_deletion
T_deletion
C_insertion
T_insertion
2bp_deletion
3bp_deletion
4bp_deletion
5+bp_deletion
2bp_insertion
3bp_insertion
4bp_insertion
5+bp_insertion
2bp_deletion_with_microhomology
3bp_deletion_with_microhomology
4bp_deletion_with_microhomology
5+bp_deletion_with_microhomology
My_sample (n = 1302)

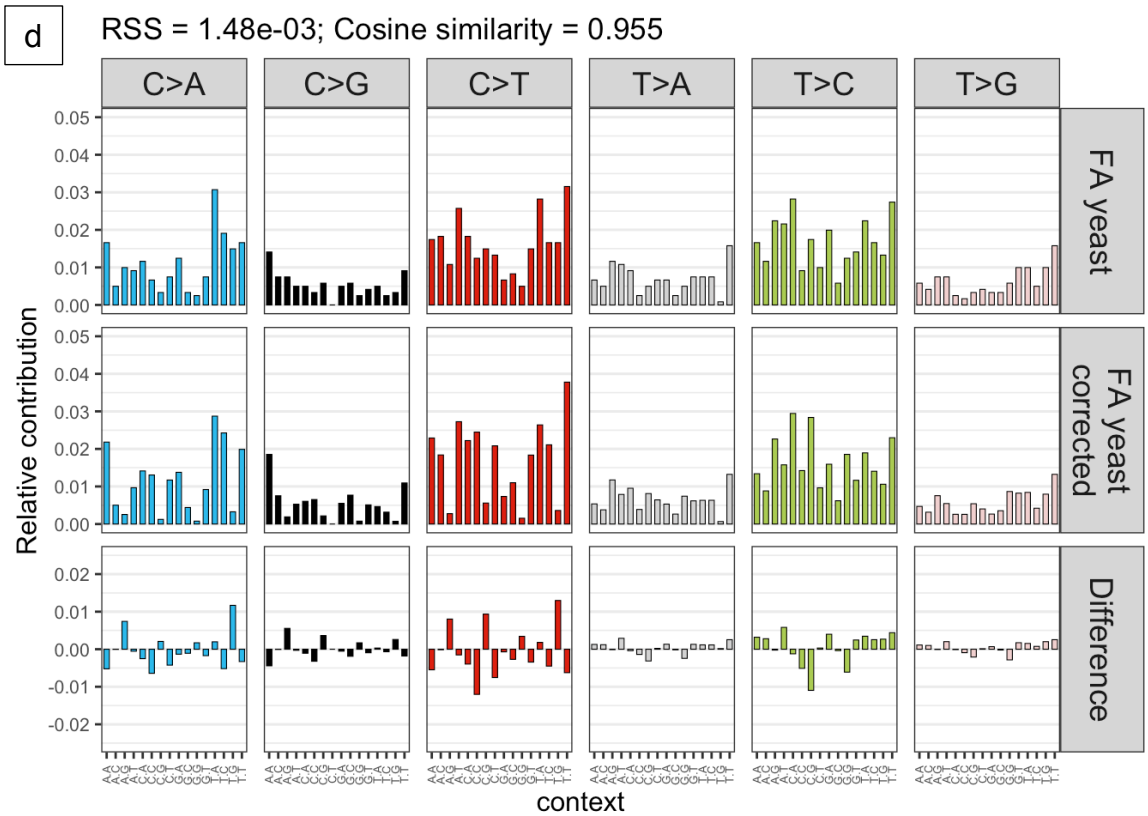
c
Nr of indels
100
75
50
25
0
C_deletion
T_deletion
C_insertion
T_insertion
2bp_deletion
3bp_deletion
4bp_deletion
5+bp_deletion
2bp_insertion
3bp_insertion
4bp_insertion
5+bp_insertion
2bp_deletion_with_microhomology
3bp_deletion_with_microhomology
4bp_deletion_with_microhomology
5+bp_deletion_with_microhomology
My_sample (n = 253)

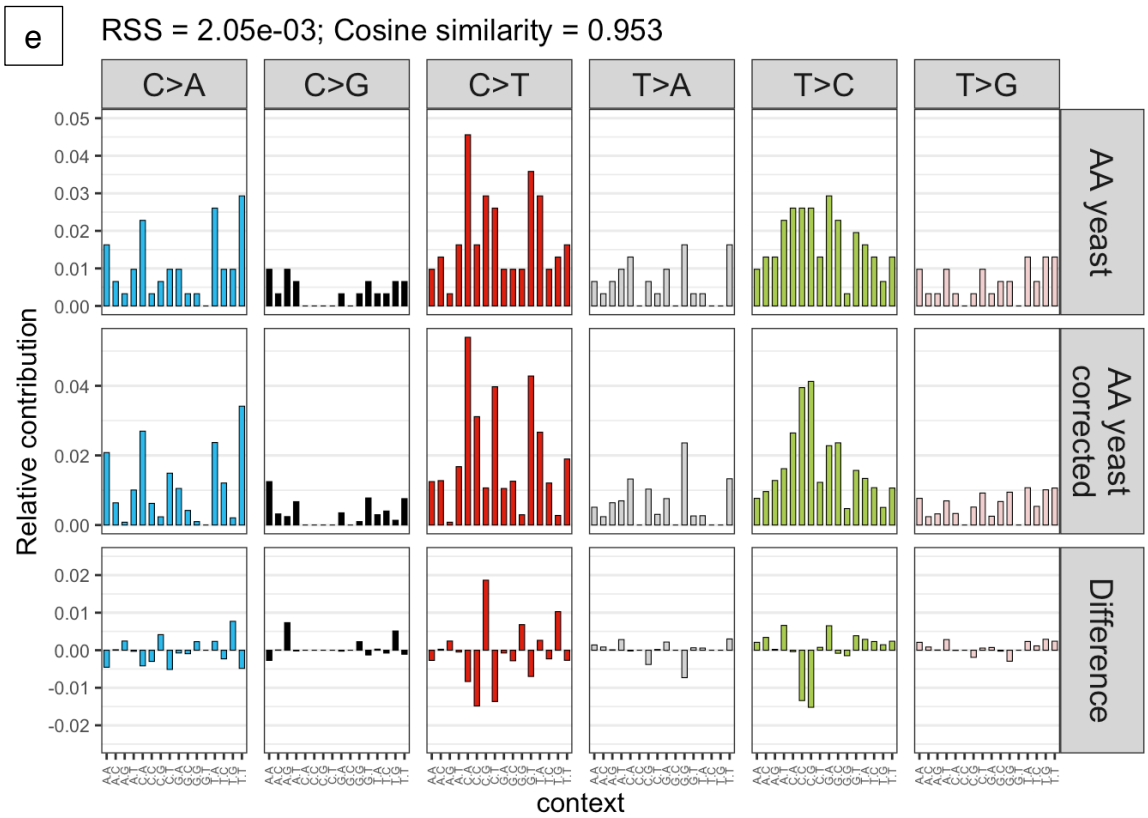
Figure 4: Comparisons of mutational profiles between (a) untreated controls and formaldehyde; (b) untreated controls and acetaldehyde; (c) formaldehyde and acetaldehyde; (d) formaldehyde with and without correction for trinucleotide frequencies in mouse; and (e) acetaldehyde with and without correction for trinucleotide frequencies in mouse.




b
RSS = 4.95e-03; Cosine similarity = 0.855
C>A
C>G
C>T
T>A
T>C
T>G
control yeast
AA yeast
Difference
Relative contribution
context


c
RSS = 4.69e-03; Cosine similarity = 0.864
C>A
C>G
C>T
T>A
T>C
T>G
FA yeast
AA yeast
Difference
Relative contribution
context


d
RSS = 1.48e-03; Cosine similarity = 0.955
C>A
C>G
C>T
T>A
T>C
T>G
FA yeast
FA yeast corrected
Difference
Relative contribution
context


e
RSS = 2.05e-03; Cosine similarity = 0.953
C>A
C>G
C>T
T>A
T>C
T>G
AA yeast
AA yeast corrected
Difference
Relative contribution
context

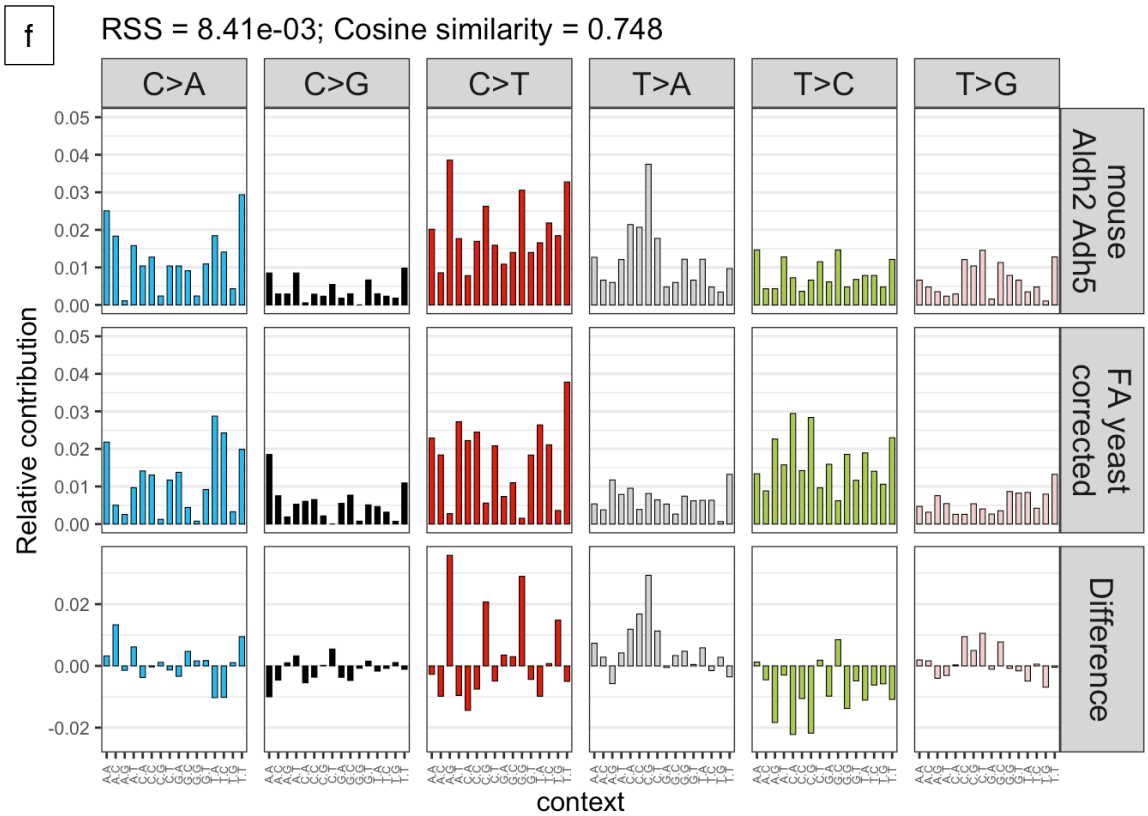
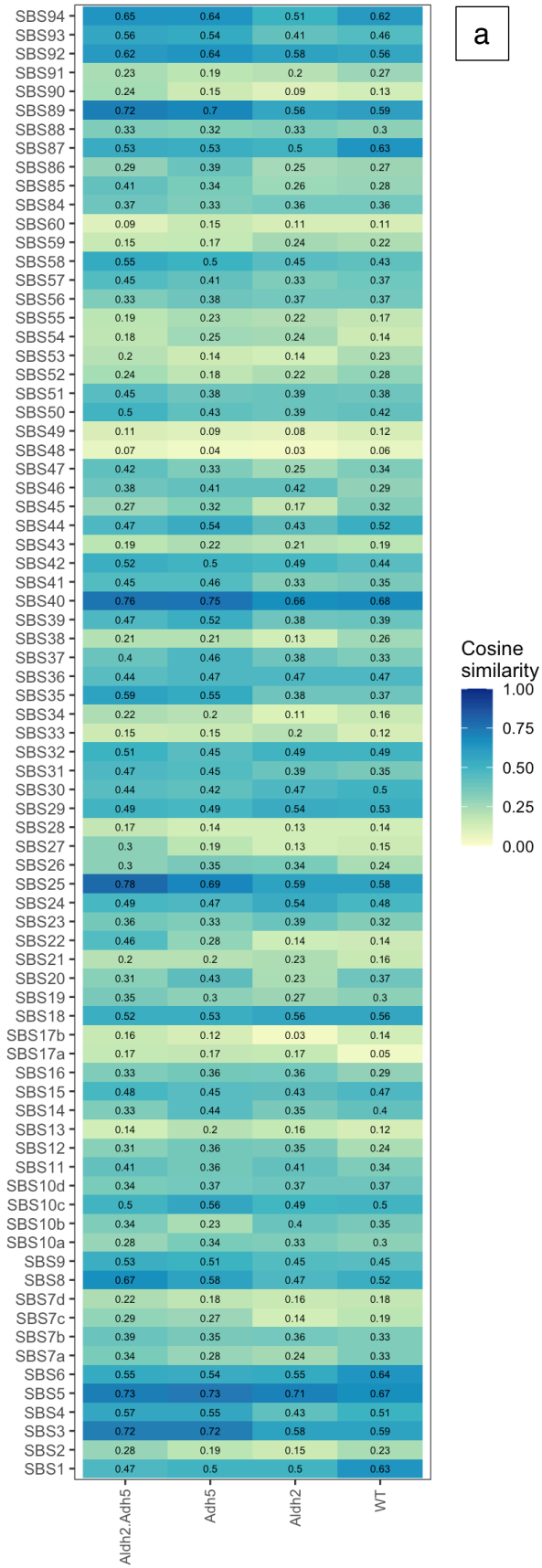
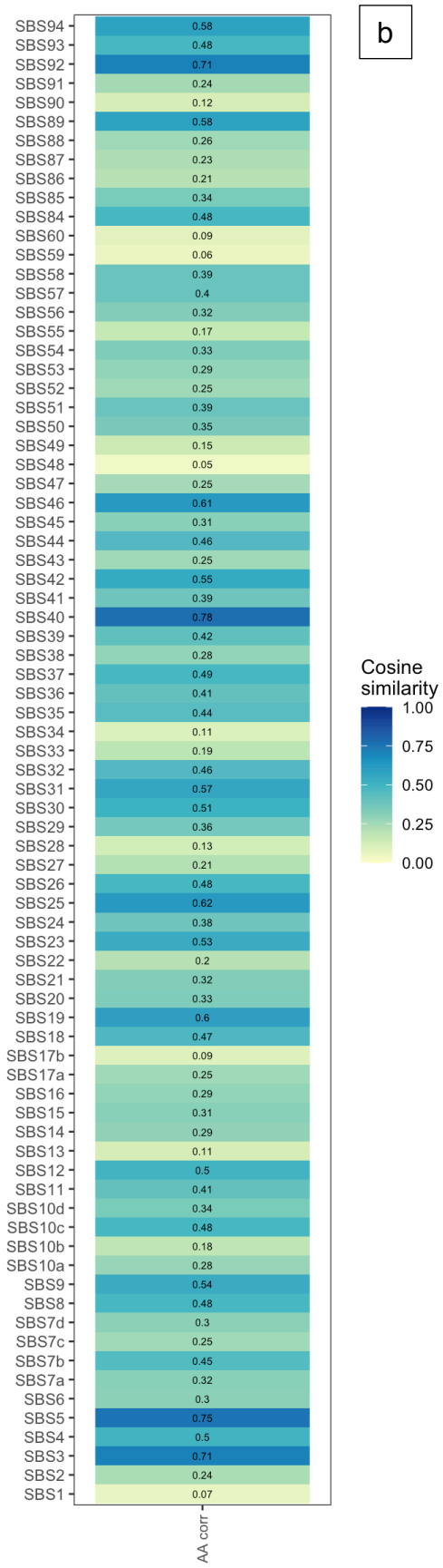
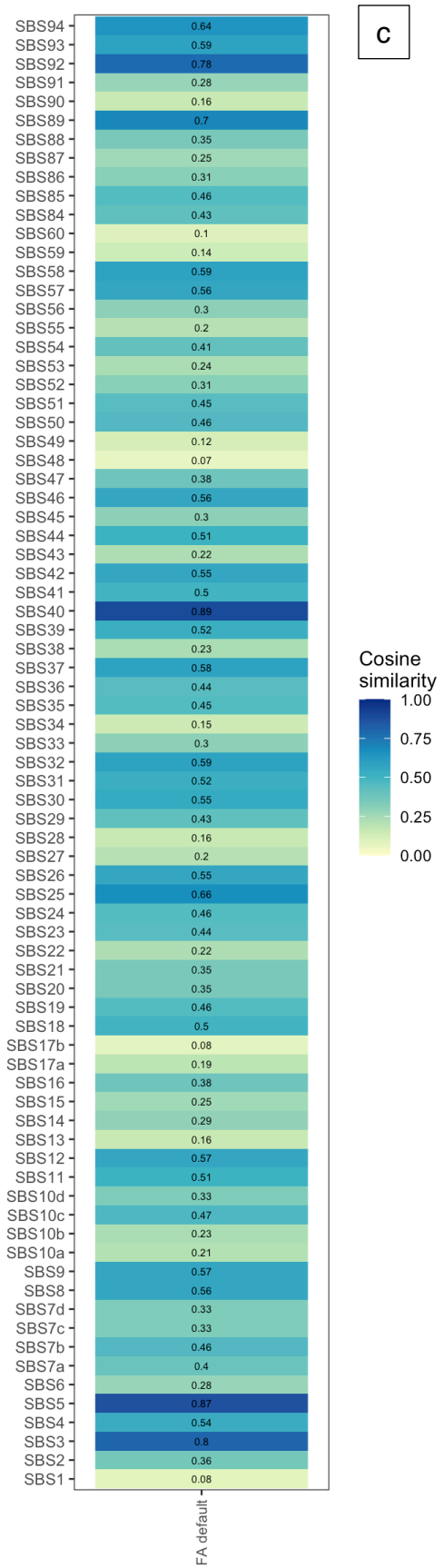
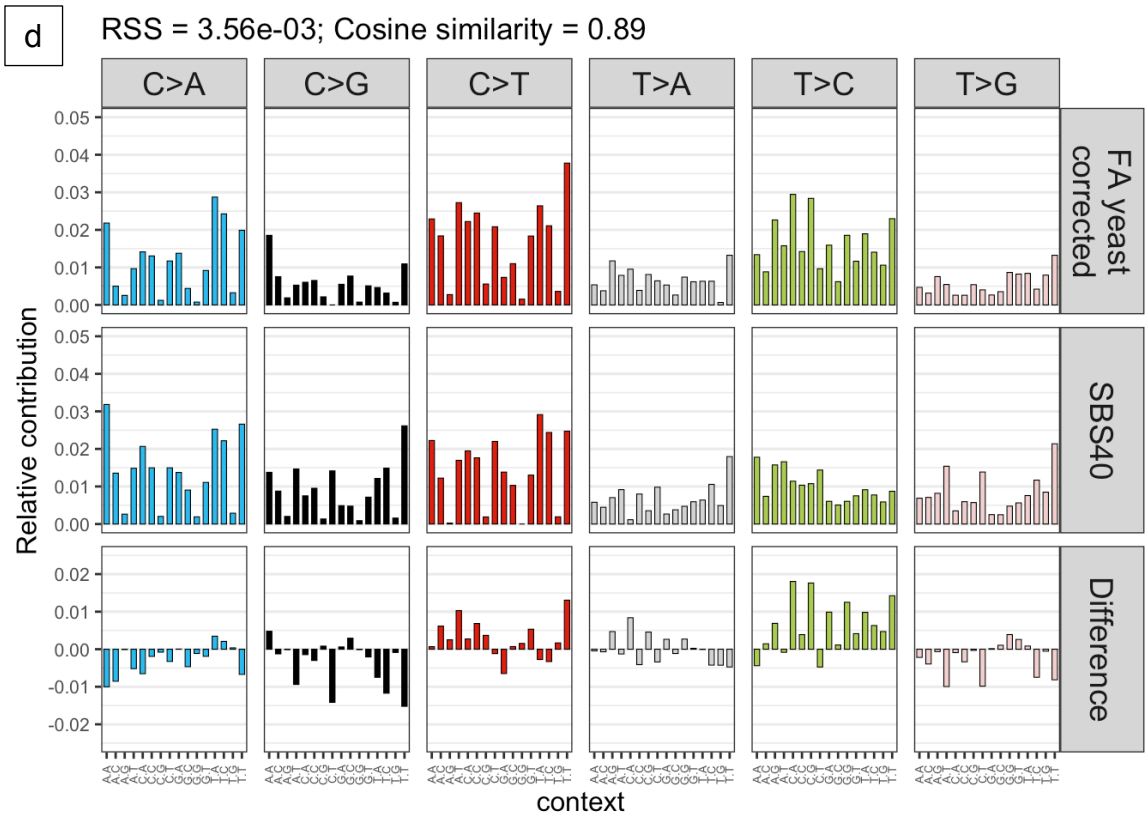

f
RSS = 8.41e-03; Cosine similarity = 0.748
C>A
C>G
C>T
T>A
T>C
T>G
mouse Aldh2 Adh5
FA yeast corrected
Difference
Relative contribution
context

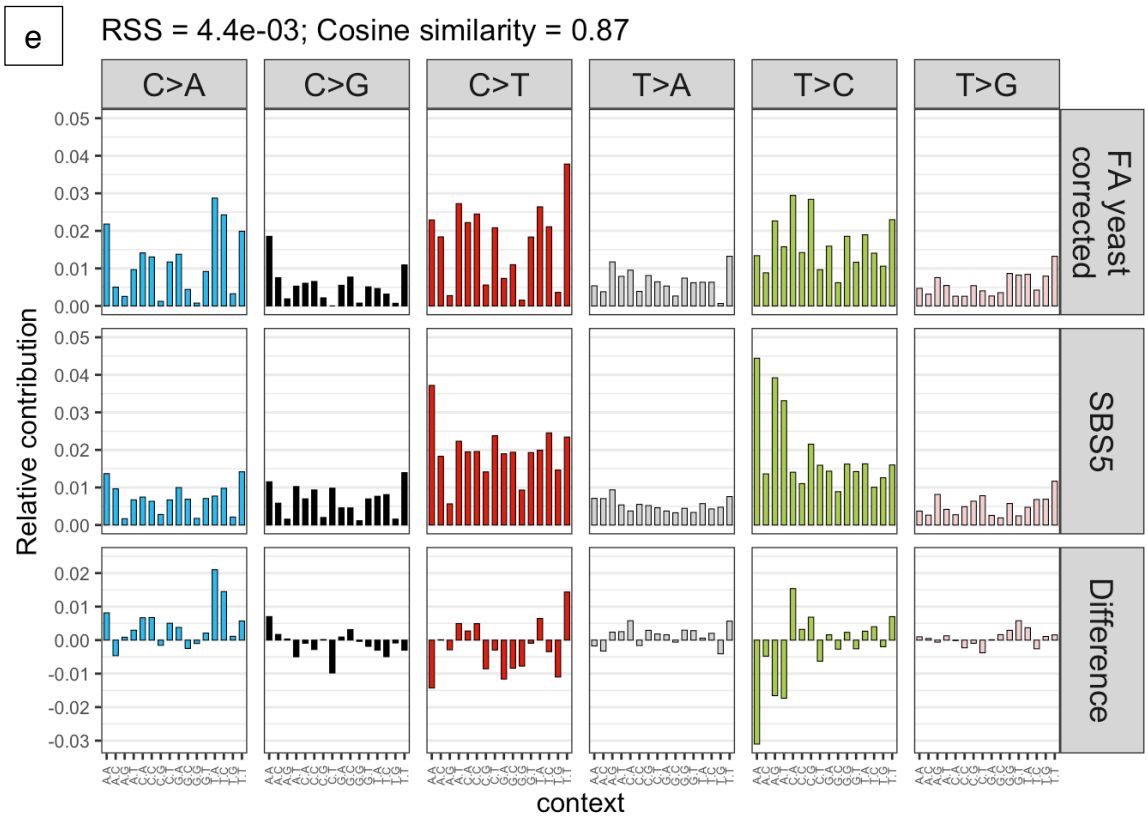
Figure 5: (a) Cosine similarity values comparing mouse mutational profiles from Dingler et al. vs. COSMIC SBS signatures. (b) Cosine similarity values comparing acetaldehyde mutational profile in yeast (with trinucleotide abundance correction) vs. COSMIC SBS signatures. (c) Cosine similarity values comparing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


b
SBS94 0.58
SBS93 0.48
SBS92 0.71
SBS91 0.24
SBS90 0.12
SBS89 0.58
SBS88 0.26
SBS87 0.23
SBS86 0.21
SBS85 0.34
SBS84 0.48
SBS60 0.09
SBS59 0.06
SBS58 0.39
SBS57 0.4
SBS56 0.32
SBS55 0.17
SBS54 0.33
SBS53 0.29
SBS52 0.25
SBS51 0.39
SBS50 0.35
SBS49 0.15
SBS48 0.05
SBS47 0.25
SBS46 0.61
SBS45 0.31
SBS44 0.46
SBS43 0.25
SBS42 0.55
SBS41 0.39
SBS40 0.78
SBS39 0.42
SBS38 0.28
SBS37 0.49
SBS36 0.41
SBS35 0.44
SBS34 0.11
SBS33 0.19
SBS32 0.46
SBS31 0.57
SBS30 0.51
SBS29 0.36
SBS28 0.13
SBS27 0.21
SBS26 0.48
SBS25 0.62
SBS24 0.38
SBS23 0.53
SBS22 0.2
SBS21 0.32
SBS20 0.33
SBS19 0.6
SBS18 0.47
SBS17b 0.09
SBS17a 0.25
SBS16 0.29
SBS15 0.31
SBS14 0.29
SBS13 0.11
SBS12 0.5
SBS11 0.41
SBS10d 0.34
SBS10c 0.48
SBS10b 0.18
SBS10a 0.28
SBS9 0.54
SBS8 0.48
SBS7d 0.3
SBS7c 0.25
SBS7b 0.45
SBS7a 0.32
SBS6 0.3
SBS5 0.75
SBS4 0.5
SBS3 0.71
SBS2 0.24
SBS1 0.07
AA cor
Cosine similarity
1.00
0.75
0.50
0.25
0.00


c
SBS94 0.64
SBS93 0.59
SBS92 0.78
SBS91 0.28
SBS90 0.16
SBS89 0.7
SBS88 0.35
SBS87 0.25
SBS86 0.31
SBS85 0.46
SBS84 0.43
SBS60 0.1
SBS59 0.14
SBS58 0.59
SBS57 0.56
SBS56 0.3
SBS55 0.2
SBS54 0.41
SBS53 0.24
SBS52 0.31
SBS51 0.45
SBS50 0.46
SBS49 0.12
SBS48 0.07
SBS47 0.38
SBS46 0.56
SBS45 0.3
SBS44 0.51
SBS43 0.22
SBS42 0.55
SBS41 0.5
SBS40 0.89
SBS39 0.52
SBS38 0.23
SBS37 0.58
SBS36 0.44
SBS35 0.45
SBS34 0.15
SBS33 0.3
SBS32 0.59
SBS31 0.52
SBS30 0.55
SBS29 0.43
SBS28 0.16
SBS27 0.2
SBS26 0.55
SBS25 0.66
SBS24 0.46
SBS23 0.44
SBS22 0.22
SBS21 0.35
SBS20 0.35
SBS19 0.46
SBS18 0.5
SBS17b 0.08
SBS17a 0.19
SBS16 0.38
SBS15 0.25
SBS14 0.29
SBS13 0.16
SBS12 0.57
SBS11 0.51
SBS10d 0.33
SBS10c 0.47
SBS10b 0.23
SBS10a 0.21
SBS9 0.57
SBS8 0.56
SBS7d 0.33
SBS7c 0.33
SBS7b 0.46
SBS7a 0.4
SBS6 0.28
SBS5 0.87
SBS4 0.54
SBS3 0.8
SBS2 0.36
SBS1 0.08
FA default
Cosine similarity
1.00
0.75
0.50
0.25
0.00


d
RSS = 3.56e-03; Cosine similarity = 0.89
C>A
C>G
C>T
T>A
T>C
T>G
FA yeast corrected
SBS40
Difference
Relative contribution
context


e
RSS = 4.4e-03; Cosine similarity = 0.87
C>A
C>G
C>T
T>A
T>C
T>G
FA yeast corrected
SBS5
Difference
Relative contribution
context

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 kit 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 (35).

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 are off-white while Can^r Ade^- colonies turn red or pink). Care was taken to handle cells gently, as they were 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

Can^r Ade^- mutants from formaldehyde and acetaldehyde treatment experiments were collected and reporter gene loss of function phenotypes were verified as described previously (35). Briefly, Can^r red 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 selected for sequencing. Can^r Ade^- Ura^- mutants were avoided because those isolates sometimes turn out to be telomere truncations.

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 (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).

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. 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).
22. 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).
23. 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).
24. Kimsey, I. J. *et al.* Dynamic basis for dG•dT misincorporation via tautomerization and ionization. *Nature* **554,** 195 (2018).
25. Alberts, B. *et al.* in *Mol. Biol. Cell* (W.W. Norton & Co., 2014).
26. Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578,** 94–101 (2020).
27. 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).
28. International Agency for Research on Cancer. *Identification of research needs to resolve the carcinogenicity of high-priority IARC carcinogens*. **42,** (2010).
29. International Agency for Research on Cancer. in *Chem. Agents Relat. Occup.* **100F,** 401–436 (2012).
30. National Toxicology Program. in *Rep. Carcinog. Thirteen. Ed.* (U.S. Department of Health and Human Services, Public Health Service, 2014).
31. 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).
32. Alcohol and Cancer Risk Fact Sheet - National Cancer Institute. (2021). at <https://www.cancer.gov/about-cancer/causes-prevention/risk/alcohol/alcohol-fact-sheet>
33. Kucab, J. E. *et al.* A Compendium of Mutational Signatures of Environmental Agents. *Cell* (2019). doi:10.1016/j.cell.2019.03.001
34. 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).
35. 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).
36. 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).
37. 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).
38. 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

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).