From Aldehyde-induced DNA damage to Pan-cancer biology: Linking Mutational signatures to their biological processes

Mahanish Jung Thapa

Supervisor: Dr. Kin Chan

Thesis submitted to the University of Ottawa

in partial fulfillment of the requirements

for the Doctorate in Philosophy degree in Biochemistry Specialization in Bioinformatics

Ottawa Institute of Systems Biology

Department of Biochemistry, Microbiology and Immunology

Faculty of Medicine

University of Ottawa

Abstract

Cancer is mainly driven by the long-term accumulation of somatic mutations in normal cells. The cause of these mutations in different cancers can range from exogenous and endogenous chemicals to defective biological processes. The genomic analysis of mutational landscapes in different cancer types has revealed many distinct patterns of mutations collectively known as mutational signatures. The characterization of these mutational signatures is in progress; however, many signatures of unknown etiology remain. Although the International Agency for Research in Cancer (IARC) has classified both formaldehyde and acetaldehyde as human carcinogens, there is still minimal comprehension of their mutagenesis at the nucleotide level. We used a highly sensitive yeast genetic reporter system, which features the generation of single-stranded DNA (ssDNA), to demonstrate formaldehyde- and acetaldehyde-induced mutations. My former lab colleague, Reena Fabros, researched formaldehyde mutagenesis, while I focused on acetaldehyde, and computationally analyzed the mutational signature induced by formaldehyde and acetaldehyde. I assessed the cell viability and mutation frequency of acetaldehyde-induced yeast mutants, and isolated and sequenced their genomes. My findings revealed that the relative contribution of C->A, C->T, and T->C substitutions is higher in formaldehyde and acetaldehyde yeast compared to the untreated control yeast. Most importantly, the formaldehyde-induced mutational signature resembles a single base substitution SBS40 of unknown etiology from Catalogue of Somatic Mutations in Cancer (COSMIC) database. Moreover, acetaldehyde induced an excess of deletion events longer than five bases, while formaldehyde did not. Despite deciphering many of these mutational signatures from different model systems, a significant research gap remains for

many mutational signatures with unknown or partially understood etiology. I used 52 curated cancer datasets of 19 different cancer types from the cancer genomics cBioPortal database. After reconstructing all the mutational signatures from mutation data, I used Gene Set Enrichment Analysis (GSEA) and Gene Ontology (GO) enrichment analysis to analyze the reconstructed mutational signatures and gene expression data and identify correlated biological processes. The common GO terms across male-only, female-only, and mixed sample datasets were found in most signatures. The results of many signatures are consistent with the known proposed etiology of the COSMIC signatures. For those signatures of unknown etiology, SBS8, SBS16, SBS28, and SBS41 are linked to DNA repair; SBS12, SBS19, SBS33, and SBS37 are associated with immune function; SBS17a/17b are linked to reactive oxygen species damage response; SBS5 with cell division; and SBS40 is linked to xenobiotic metabolism. The GO semantic similarity showed a range of values for different signatures. The correlation between age and mutation count varied for different mutational signatures. These findings could contribute to a better understanding of formaldehyde- and acetaldehyde-induced mutational signature and show how different mutational signatures vary not only by different related biological processes but also across different age groups.

Acknowledgement

I would like to give special thanks to my supervisor, Dr. Kin Chan, for his continuous guidance and support throughout my PhD studies. I am lucky to be mentored by such a person in the field of cancer research, whose expertise in both wet-lab techniques and computational dry lab was crucial in shaping my project. Dr. Chan's encouragement and academically insightful feedback have always helped me to push my boundaries and achieve things that I have never dreamed of. I feel like a more competent and independent researcher, and now I have a lot to offer to my country, Nepal, in the field of cancer research, and all credit goes to Dr. Chan. His non-academic suggestions related to my health and communication skills have developed my overall personality. Also, thank you for supporting me financially throughout my PhD studies.

I would like to thank my Thesis Advisory Committee members, Dr. Adam Rudner, Dr. Barbara Vanderhyden, Dr. Mathieu Lavallée-Adam, and Dr. Arvind Mer, for their continuous feedback and support in advancing my PhD project in a constructive way. I would also like to thank Dr. Alexandre Blais for providing me with the R script to repurpose it for the GSEA analysis of pan-cancer datasets.

I would like to thank all my lab colleagues, notably Dr. Suzana Gelova, Ghadir Makki, and Reena Fabros, who helped me during my hard times struggling with wet and dry lab work. I am also thankful to other lab members, Josie Jabbour, Aws Almir Ahmad, and Vénus Béka Donia Badié, for their insightful feedback and encouragement. It was a blessing working with you all. I would also like to thank my friend Hossein Davarinejad for his academic suggestions. Best wishes for all your future endeavors.

Most importantly, I would like to thank my family, friends, and all my seniors from Nepal, notably Hemanta Mainali, Salyan Bhattarai, and my friend Sabin Bhandari, who has been supporting me continuously from the start of my PhD studies until now. I hope to get such support continuously even after completing my PhD studies.

Table of Content

Chapter 3: Characterization of formaldehyde- and acetaldehyde-induced mutational signatures

Chapter 6: Supplementary results

List of Abbreviations

AA	Acetaldehyde
AHD5	Alcohol dehydrogenase 5
ALDH2	Aldehyde dehydrogenase 2
APOBEC	Apolipoprotein B mRNA Editing Catalytic Polypeptide-like
BER	Base-excision repair
BP	Biological Process
BH	Benjamaini-Hochber
COSMIC	Catalogue of Somatic Mutations in Cancer
CN	Copy Number
DAPK1	Death-Associated protein Kinase 1
DBS	Double base substitution
DPC	DNA-protein crosslink
DSB	Double Strand Break
DSBR	Double Strand Break Repair
EGFR	Epidermal Growth Factor Receptor
EMS	Ethyl Methanesulfonate
EMT	Epithelial-mesenchymal transition
ES	Enrichment score

FA	Formaldehyde
FDR	False discovery rate
FGF	Fibroblast Growth Factor
FGH	S-formylglutathione
fgsea	Fast GSEA
FPKM	Fragments per Kilobase of transcript per million mapped
GRCh37	Genome Reference Consortium Human Build 37
GO	Gene Ontology
GSEA	Gene Set Enrichment Analysis
hg19	Human Genome version 19
HMGSH	S-hydroxymethylglutathione
HR	Homologous Recombination
HSCs	Hematopoietic stem cells
IARC	International Agency for Research in Cancer
ID/INDELS	Insertions and Deletions
iPSC	induced Pluripotent Stem cell
KRAS	Kirsten Rat Sarcoma Viral Oncogene Homolog
MAF	Mutation Annotation Format
MMR	Mismatch repair

MSI	Microsatellite Instability
MSigDB	Molecular Signatures Database
MYC	Myelocytomatosis
NER	Nucleotide excision repair
NES	Normalized Enrichment Score
NHEJ	Non-homologous end joining
N^2-HOMe-dG	N^2-hydroxymethyldeoxyguanosine
NMF	Non-negative Matrix Factorization
PAH	Polycyclic Aromatic Hydrocarbon
PARP	poly (ADP-ribose) polymerase
PCAWG	Pan-Cancer Analysis of Whole Genome
POLD1	Proofreading-deficient 1
POLE	Polymerase epsilon
pol ζ	Polymerase zeta
PTEN	Phosphatase and Tensin Homolog
REV1	Reversionless 1
RNA-seq	RNA Sequencing
RNS	Reactive Nitrogen Species
ROS	Reactive Oxygen Species

RPKM	Reads per Kilobase of transcript per million mapped
RSEM	RNA-Seq by Expectation Maximization
SAM	S-adenosylmethionine
SBS	Single Base Substitution
ssDNA	Single-Stranded DNA
SSBR	Single Strand Break Repair
SV	Structural variants
TCGA	The Cancer Genome Atlas
THF	Tetrahydrofolate
TLS	Translesion Synthesis
TPM	Transcripts per million
UADT	Upper Aerodigestive Tract
UV	Ultra Violet
VEGF	Vascular Endothelial Growth Factor

List of Figures

Chapter 1: General Introduction

Chapter 2: Mutagenic Properties of Acetaldehyde and Formaldehyde

Chapter 3: Characterization of formaldehyde- and acetaldehyde-induced mutational signatures

List of Tables

Chapter 2: Mutagenic Properties of Acetaldehyde and Formaldehyde

Chapter 3: Characterization of formaldehyde- and acetaldehyde-induced mutational signatures

Chapter 4: Pan-Cancer Gene Set Enrichment and Gene Ontology Analyses

Chapter 5: Conclusions

Chapter 1: General Introduction

1.1 Copyright and License Policies

1. For figure 1.1 NMF algorithm

Alexandrov et al., "Deciphering signatures of mutational processes operative in human cancer," *Cell Rep*, vol. 3, no. 1, pp. 246–259, 2013, doi: 10.1016/j.celrep.2012.12.008.

2. For figures 1.2 SBS4 and 1.3 SBS29

so no separate written permission is required.

Here is the link of COSMIC terms and conditions.

https://www.cosmickb.org/terms/

Alexandrov et al., "Deciphering signatures of mutational processes operative in human cancer," *Cell Rep*, vol. 3, no. 1, pp. 246–259, 2013, doi: 10.1016/j.celrep.2012.12.008.

3. For figure 1.4 GSEA

This figure includes two publicly available resources.

a. **For the right panel**

Subramanian *et al.*, "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 102, no. 43, pp. 15545–15550, Oct. 2005, doi: 10.1073/pnas.0506580102.

b. For the left panel

The image is reproduced from the GenePattern GSEA module documentation, version 14 (https://www.genepattern.org/modules/docs/GSEA/14/). The images and the content on the GenePattern site are released under the Broad Institute's open-source license, which

allows its figures and content to be free for all uses, academic or commercial.

The original paper for GenePattern is

Riech et al., GenePattern 2.0. *Nat Genet*. 2006;38(5):500-501. doi:10.1038/ng0506-500.

1.2 Cancer and carcinogenesis

Cancer is a complex disease characterized by the uncontrolled division and proliferation of abnormal cells. These cells invade the surrounding tissues and metastasize to distant parts of the body via the blood and lymphatic systems, significantly impairing different types of tissues leading to multisystem failure and death. Cancer cells have several distinct features that differentiate them from normal cells, such as uncontrolled growth and division, resistance to apoptosis (programmed cell death), angiogenesis, genome instability, and immune evasion (1). There are different types of cancers categorized by their tissue of origin. Carcinomas include cancers that arise from epithelial cells or the lining of organs, such as lung, breast, and colorectal cancers. Sarcomas arise from connective tissues such as bone, cartilage, muscles, and fats, and include cancers such as osteosarcoma and liposarcoma (2). Another type of cancer, leukemias, arises from blood-forming tissues (bone marrow) where white blood cells proliferate abnormally. The lymphatic system such as, lymph nodes and lymphoid tissues, is affected by certain lymphoma cancers, such as Hodgkin and non-Hodgkin lymphoma. Brain and spinal cord cancers include gliomas and meningiomas (3).

Carcinogenesis is a complex multi-step process where normal cells change into malignant, abnormal cells through genetic and epigenetic alterations, inflammation, and immune dysfunction over a long period of time, affecting the key cellular processes such as cell division, DNA repair mechanisms, and apoptosis, ultimately leading to cancer development (4). The multiple steps involved in carcinogenesis are initiation, promotion, and progression at the molecular and cellular level (5). In the initiation step, the DNA reacts with physical and chemical human carcinogens such as ultraviolet light, formaldehyde in tobacco smoke, and

acetaldehyde in alcohol, leading to the formation of DNA adducts and errors during replication (6). The endogenous intermediate products, such as reactive oxygen species (ROS) and reactive nitrogen species (RNS), produced during metabolic processes cause oxidative stress and DNA damage, such as 8-oxo-guanine formation (7). Point mutations, gene amplifications, and chromosomal translocations in proto-oncogenes such as *KRAS (Kirsten Rat Sarcoma Viral Oncogene Homolog)*, *MYC* (MYC proto-oncogene derived from Myelocytomatosis), and *EGFR (Epidermal Growth Factor Receptor)* transform into oncogenes by constitutive activation of growth-promoting signals (8). The mutations that would cause loss of function of tumor suppressor genes such as *TP53, RB1,* and *APC* disable the ability to stop cell proliferation and induce cell death (9). Also, mutations in DNA repair genes such as *MLH1, MSH2,* important for mismatch repair (MMR), or *BRCA1/BRCA2*, important for homologous recombination repair mechanisms, compromise the overall DNA repair mechanisms, increase the mutation burden, leading to genomic instability (10).

In the promotion step, the clonal expansion of all mutated cells is stimulated by promoters such as inflammation, growth factors, and hormones such as estrogen which promote breast malignant cell proliferation (11). Epigenetic alterations such as DNA methylation and histone modifications alter gene expression, promoting cancer cell survival and proliferation (12). For example, hypermethylation of tumor suppressor gene promoters silences the expression of genes such as *DAPK1 (Death-Associated protein Kinase 1)* and *PTEN (Phosphatase and Tensin Homolog),* promoting uncontrolled cell growth (13). Inflammatory mediators such as cytokines and chemokines activate different types of proliferation pathways, such as NF-κB

(Nuclear Factor kappa-light-chain-enhancer of activated B cells) and STAT3 (Signal Transducer and Activator of Transcription 3) (14).

Finally, in the progression step, the accumulation of genetic and epigenetic changes, high mutation burden, unstable chromosomes, and impaired DNA repair systems with cell cycle checkpoint dysregulation drives the malignancy into tumor phenotypes (15). In tumor cells, the genomic instability can lead to upregulation of hTERT (human telomerase reverse transcriptase) which is a subunit of telomerase, that in turn circumvents senescence and contributes to the cell immortality (16). To fulfill all tumor cells nutrient and oxygen level requirements, blood vessel growth is promoted by angiogenic factors such as VEGF (Vascular Endothelial Growth Factor) and FGF (Fibroblast Growth Factor) (17). The invasive properties of all tumor cells are enhanced by epithelial-mesenchymal transition (EMT), which gives them the freedom to break away, move, and invade other normal tissues. They even escape the immune cells' attacks by reducing antigen presentation and exploiting the immune checkpoints (18).

1.3 Human Carcinogens

Human carcinogens are different types of physical, chemical and biological agents that can induce DNA damage in the form of lesions by chemically reacting with DNA strands, mutations, or disrupting biological pathways leading to cancer. Different types of human carcinogens damage human DNA. Many of these human carcinogens are categorized based on their origin, such as exogenous carcinogens, which come from external sources, whereas endogenous carcinogens are generated internally within the human body (19).

Exogenous carcinogens can damage the DNA by promoting mutations or interfering with key cellular processes such as DNA repair. Exogenous chemical DNA-damaging carcinogens include benzo[a]pyrene and nitrosamines found in tobacco smoke associated with DNA adducts formation, which has been linked to lung and oral cancer (19,20). Benzene, found in industrial solvents and gasoline fumes, is associated with DNA strand breaks and is linked with leukemia (21). Aflatoxins from moldy grains and nuts induce guanine adducts which mispairs with adenine leading to G->T transversion in the *TP53* gene are linked to liver cancer (22). Mineral fibers such as asbestos used in construction are associated with physical irritation and chronic inflammation, which in turn can lead to mesothelioma and lung cancer (23). Arsenic found in groundwater is associated with oxidative stress, and DNA methylation is linked to skin and lung cancer (24). Physical agent, ultraviolet (UV) radiation, has been associated with the formation of pyrimidine dimers and DNA damage and can lead to skin cancer (25). Ionizing radiation, such as X-rays and gamma rays, associated with DNA double breaks and mutations can be linked to leukemia, thyroid and breast cancers (26).

Endogenous carcinogens arise from normal physiological processes such as cellular metabolism or inflammatory responses and cause DNA damage and mutations, contributing to cancer development (27). Examples include ROS such as superoxide and hydrogen peroxide, which are the byproducts of cellular metabolism and cause oxidative DNA damage by forming 8-oxoguanine lesions, contributing to multiple cancer types (28). RNS such as nitric oxide and peroxynitrite, which are produced during inflammation can damage DNA by deamination, mutations, and strand breaks, potentially resulting in stomach, liver, and colon cancers (29). Excess hormone levels of estrogen can form DNA adducts, which damage the

DNA, contributing to breast and ovarian cancer (30). Excess levels of androgens can stimulate the increased growth of prostate epithelial cells, leading to prostate cancer (31). Bile acids such as deoxycholic acid induce ROS and inflammation in the gut lining, leading to colon and liver cancers (32). Lipid peroxidation releases products, such as malondialdehyde and 4-HNE (4-Hydroxy-2-nonenal), forming DNA adducts and mutagenesis, leading to liver cancer (33).

1.4 Aldehydes as human carcinogens

Among many human carcinogens, aldehydes represent a significant group and are found in both exogenous and endogenous forms. They are widespread and are highly reactive. DNA chemically reacts with different types of aldehydes, which modify its structure and function (34). These reactive compounds have a (-CHO) group, which forms covalent bonds with DNA and proteins, leading to different forms of DNA damage such as point mutations, DNA-protein crosslinks, interstrand and intrastrand crosslinks, DNA adducts, single and double-strand breaks. Mutagenesis induced by aldehydes is thought to lead to cancer (35).

The International Agency for Research on Cancer (IARC) has declared different types of carcinogen groups. The categorization of different groups is based on scientific judgment that reflects the strength of the evidence derived from studies in humans and experimental animals (36,37). Out of many aldehydes, we picked formaldehyde (FA) and acetaldehyde (AA) because each has widespread human exposure (38). AA is found in tobacco smoke and can lead to stomach and oral cavity cancers (38,39). AA is produced endogenously during ethanol metabolism that forms DNA adducts, leading to liver and esophageal cancers (40). Exposure to FA found in industrial smoke, vehicle exhaust, and disinfectants via inhalation or skin contact can lead to

mouth and nasopharyngeal cancers (41). FA produced during DNA and histone demethylation can crosslink with protein leading to lung and nasal epithelial cancers (42). FA has strong epidemiological and experimental evidence as a human carcinogen. The endogenous FA is in the range of tens of micromolar in human blood and is considered as group 1 carcinogenic agent by IARC, because there is sufficient evidence of carcinogenicity in humans (43,44). AA is in group 2B as it is a possible carcinogen to humans because there is limited evidence of carcinogenicity in humans, while ethanol-derived AA is in group 1 because there is sufficient evidence of human carcinogenicity (36). Despite numerous studies elaborating the genotoxic effects of FA and AA, the mutagenicity is still less understood, and results are inconclusive.

Details about AA and FA mutagenic properties can be found in Chapter 2.

1.5 Mutagenesis and Mutational signatures

Genomic DNA have the genetic information within the chromosomes of an organism. It has all the inherited genetic information important for growth, survival, and reproduction (45). Mutagenesis is a process of creating mutations, which can be permanent modifications or heritable changes in DNA, leading to altered protein functions and phenotypes. There are two types of mutagenesis (46). Spontaneous mutagenesis can occur naturally and is driven by the intrinsic chemical instability of DNA, metabolism, and unavoidable replication errors leading to depurination, which is the loss of purine bases adenine (A) and guanine (G), creating an abasic site and the insertion of random base opposite to the abasic site, and deamination, which changes cytosine to thymine, which is a ubiquitous process linked to aging (47). The generation of ROS during respiration, oxidizing guanine to 8-oxo-deoxyguanosine, miscoding

guanine to thymine transversion, and replicative polymerases inserting wrong nucleotides or slips on repeats when proofreading or if MMR is defective, can result in single base substitutions and indels (insertions and deletions) (48). In contrast, induced mutagenesis can occur due to external mutagens, which can increase the mutation burden in an organism. Mutagens damage the DNA by chemically reacting with it and forming DNA adducts, single- and double-strand breaks, DNA-protein crosslinks, or disrupting the replication or repair mechanisms (49). For example, alkylating agents Ethyl Methanesulfonate (EMS) and anti-cancer drug temozolomide can add a methyl group to O6-guanine, yielding G->A transitions (50); Polycyclic Aromatic Hydrocarbon (PAH) in tobacco smoke can form bulky adducts with DNA, leading to G->T transversion (51); ionizing radiation, such as X-rays and gamma rays can cause double-strand breaks or large deletions (52).

Cells activate different repair pathways in response to these DNA damages, and when the damage in the DNA is high, cell death or apoptosis is activated to remove the damaged cells (53). On the other hand, if genes that are important for maintaining normal biological processes are mutated, this could result in uncontrolled division of cells (54,55). Unrepaired lesions cause the genomic stability and cause genetic variation, which can contribute to carcinogenesis and to other diseases (56). Mutational signatures are unique patterns of genetic changes in the DNA that are attributed to distinct mutational processes. Each nucleotide base present in the DNA is susceptible to chemical reactions or mutations caused in reaction with other molecules, which sets a basic principle of the generation of mutational signatures, where the set of DNA mutations observed in a genome results from

superimposing the characteristic patterns of the mutagens or the mutational processes responsible for inducing mutations (57–59).

Although the concept of DNA-damaging agents causing footprints of somatic mutations was around for decades, mutational signatures as a genome-wide concept came into light only when whole-genome sequencing became practical. In 2010, the first cancer genome project (Sanger Cancer Genome Project) was completed using melanoma cells and tobacco smoking-linked lung cancer cells. The DNA in melanoma cells has UV-driven C->T transitions, whereas the malignant lung cells have C->A transversions (60). These distinct mutational patterns set a foundation that a single genome can have multiple mutation processes. In 2012, a group of scientists generated catalogs of somatic mutations from 21 whole-genome sequenced breast cancers (59). The somatic mutations were analyzed using a non-negative matrix factorization (NMF) algorithm to identify the mutational signatures (57). NMF decomposes the patterns of mutations found in the mutational catalog of cancer into individual signatures. This algorithm also estimates the number of mutations contributing to each signature (**Figure 1.1**). Reconstruction error, or non-systematic analysis error, is the difference between the original cancer mutational dataset and the decomposed mutational signatures and contributions (57). This work was the first demonstration of separating overlapping mutational processes found in the same tumor. *Alexandrov et al. (2013)* extended the NMF framework to use on 30 different cancer types across 7,042 tumors, extracting 21 different mutational signatures. Out of these 21 signatures, many of them were linked to certain etiologies, such as SBS1 was associated with age and SBS4 with tobacco smoke (61).

The COSMIC (Catalogue Of Somatic Mutations In Cancer) database in 2015 curated all different mutational signatures with associated etiology, and the number of mutational signatures has been expanded throughout the following years (62,63). The data sources used are all pan-cancer datasets, such as Pan-Cancer Analysis of Whole Genome (PCAWG) and The Cancer Genome Atlas (TCGA), and many exposure-specific studies, which may capture signatures that are not present in pan-cancer datasets. All these latest reference sets of mutational signatures were extracted using SigProfiler using the NMF algorithm as found in the research studies done by *Alexandrov et al. (2020)* using 2,780 whole-genome variants produced by the PCAWG network. Moreover, the reproducibility of all the results was assessed on 1,865 whole genomes and 19,184 exomes (64–66).

According to the latest release, v3.4, in October 2023, there are six different sets of mutational signatures. Altogether, there are 69 single base substitution (SBS) signatures, 19 double base substitution (DBS) signatures, and 23 small insertions and deletions (ID) signatures, usually between 1 and 50 base pairs, several copy number (CN), structural variants (SV) signatures, and RNA-SBS signatures (67). The profile of each SBS signature has been displayed where each signature has six substitution types: C->A, C->G, C->T, T->A, T->C, and T->G, and all these substitutions are referred to by the pyrimidine of the mutated Watson-Crick base pair. 96 possible mutation types were produced by examining the immediate bases at 5′ and 3′ positions for each type of substitution (6 types of substitution × 4 types of 5′ base × 4 types of 3′ base). All SBS mutational signatures were reported based on the 96 possible trinucleotide contexts (67,68). Out of all these SBS mutational signatures, some signatures, such as SBS1 and SBS5, are clock-like signatures. SBS1 has C->T transitions at CpG dinucleotides associated with spontaneous

deamination of 5-methylcytosine during normal DNA replication (69), and SBS5 has a flat signature with a mixture of all six single base substitutions, although its precise etiology is still unknown (70). Both signatures are ubiquitously present in all cancer types, reflecting the endogenous, low-level, continuous biological processes, and the mutation burden accumulates with time (71). Both SBS2 and SBS13 are associated with the activity of the APOBEC family of cytidine deaminases when APOBEC3A/3B converts cytosine to uracil in the same trinucleotide sequence context 5'-TpCN-3', where "T" stands for thymine, "p" for phosphodiester bond, "C" stands for cytosine, which is deaminated to uracil and ultimately mutated, and "N" stands for any base (A, C, G, or T) (72). But the dominant mutation types are C->T (transition) for SBS2 and C->G (transversion) for SBS13, respectively. SBS2 mutational patterns may be generated either directly by DNA replication across uracil before base-excision repair (BER) acts or after the use of BER to excise uracil, creating abasic sites, that are bypassed by error-prone translesion synthesis (TLS) polymerases (73). SBS13 mutational patterns are likely generated by the bypass of error-prone TLS polymerases such as REV1 (Reversionless 1), on the abasic sites created by BER (59,74). Previous studies with limited evidence suggested that the activation of APOBEC enzymes in cancer may be due to previous viral infection, tissue inflammation, or the movement of retrotransposon (75). Both SBS2 and SBS13 are found together in the same samples (76).

Although both signatures, SBS4 (**Figure 1.2**) and SBS29 (**Figure 1.3**) are tobacco-related and are rich in C->A transversions, they have different sources of tobacco carcinogens, exposure routes, and DNA lesions. SBS4 is induced by tobacco carcinogens such as PAHs found in tobacco smoke and is found in smoking-related tumor sites (70). In contrast, SBS29 is induced by tobacco nitrosamines and is predominantly present in tobacco-chewing oral sites (71,77). These differences are

important for inferring the lifestyle exposures and for separating different etiologies in mutational signature studies (Figures 1.2 and 1.3). Seven different signatures, such as SBS6, SBS14, SBS15, SBS20, SBS21, SBS26, and SBS44, are all linked to defective DNA mismatch repair (MMR) with or without microsatellite instability (MSI), because all have the same root lesion and replication errors that a functional MMR repairs. When any key protein of MMR, such as MLH1, MSH2, MSH3, MSH6, and PMS2, is either missing or silenced, the repair cycle fails, and different categories of mutations accumulate to give distinct mutational spectra. SBS6 is dominated by C->T mutations in the CpG context (70,78); SBS14 is dominated by C->A mutations in CCN contexts (70,79); SBS15 is dominated by C->T mutations in the GCN context (70,78); SBS20 is dominated by C->A mutations in CCN context and a mix of little C->T mutations found in the GCN trinucleotide context (70,78); SBS21 is dominated by T->C mutations in GTN and TTN trinucleotide contexts (70,78); SBS26 is dominated by T->C mutations in ATN, GTN, CTN, and TTN trinucleotide contexts (71,77,78); SBS44 is dominated by C->T in GCN & ACN contexts, C->A in in CCT, CCA & CCG contexts, and a wide plateau of T->C mutations in A/T-rich triplets (80). In addition to MMR and MSI, signatures SBS14 is associated with POLE (polymerase epsilon) mutation (81), and SBS20 is associated with POLD1 (proofreading-deficient) mutations (82). Altogether, all these mutational signatures provide information on exposures to diverse carcinogenic and mutagenic agents, or the results of damaged cellular pathways to understand the foundation of mutational processes of cancer (62,67–71).

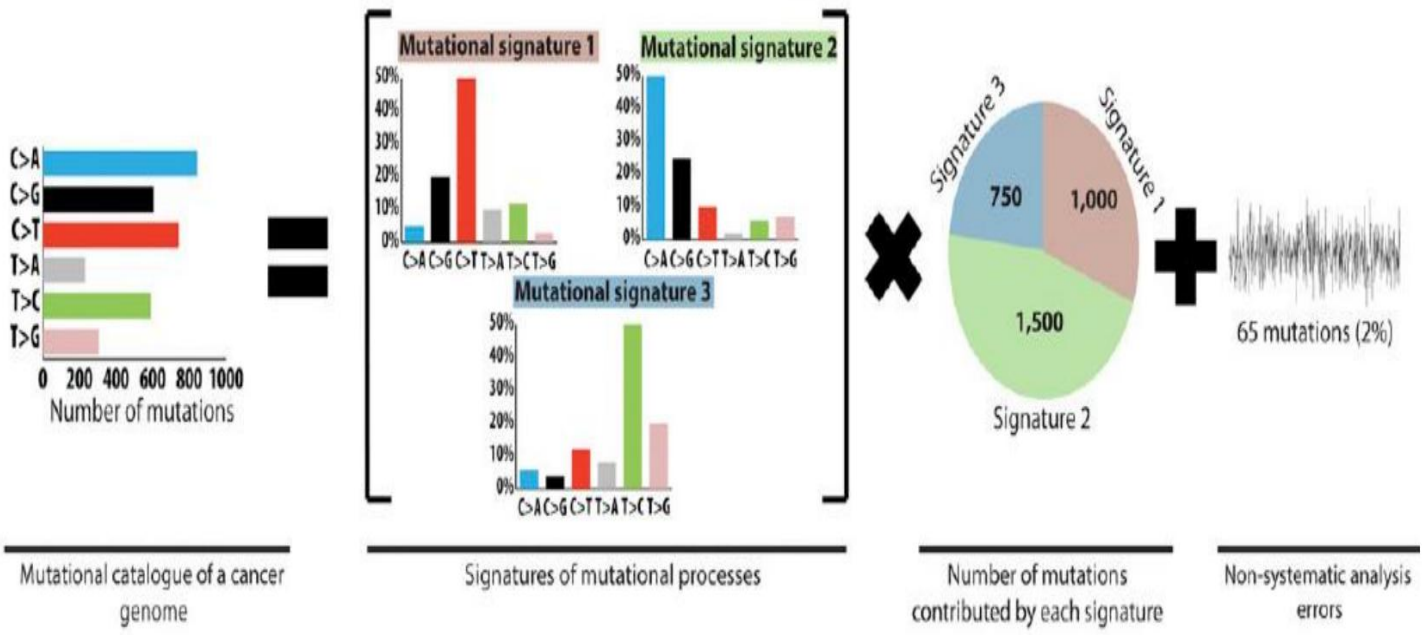


Figure 1.1: Non-negative Matrix Factorization (NMF) algorithm

The NMF is used in modeling mutational signatures of mutational processes operative in the cancer genome. It is done such that the cancer genome is a linear superimposition of the three mutational signatures with the contribution of each signature (57).

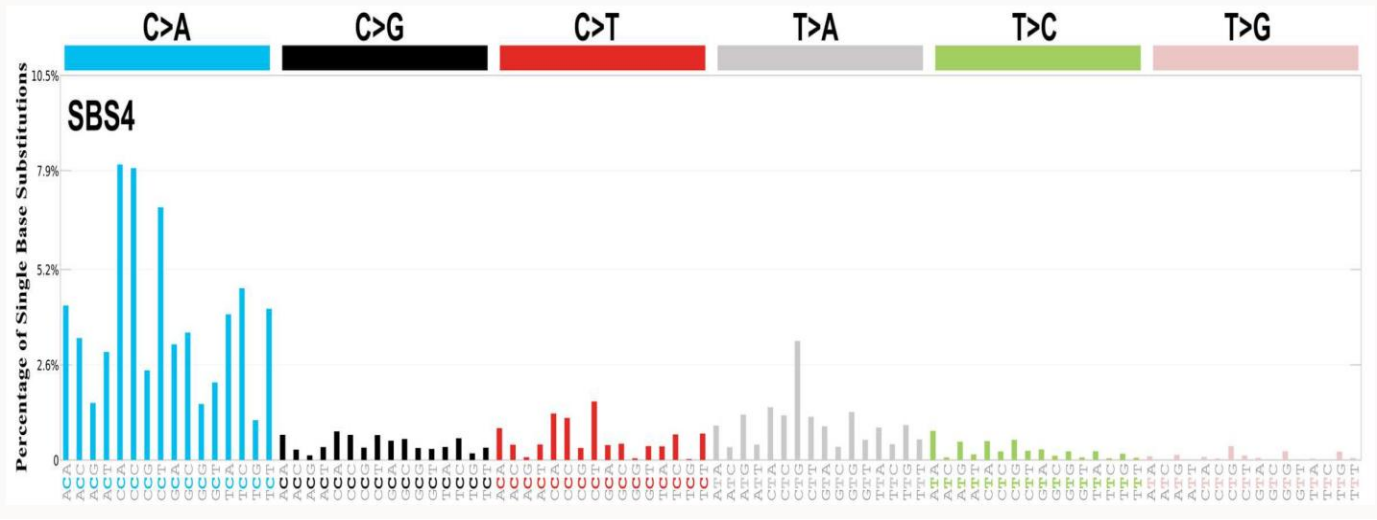


Figure 1.2: Single Base Substitution 4 (SBS4) (70)

On the top, there are six single base substitution. In the x-axis, there are 96 trinucleotide context and in y-axis, the percentage of each single base substitution in each trinucleotide context. It is associated with tobacco smoking. Its profile is similar to the mutational spectrum observed in experimental systems exposed to tobacco carcinogens such as benzo[a]pyrene

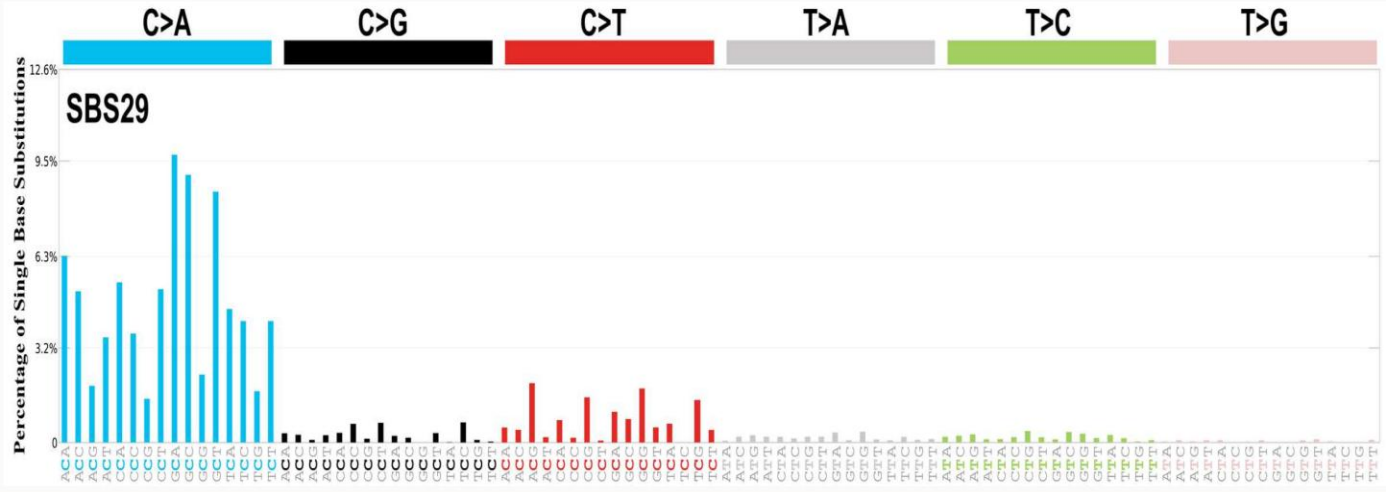


Figure 1.3: Single Base Substitution 29 (SBS29) (71,77)
On the top, there are six single base substitution. In the x-axis, there are 96 trinucleotide context and in y-axis, the percentage of each single base substitution in each trinucleotide context. It is associated with tobacco chewing.

In a recent study with human iPSC (induced pluripotent stem cell lines), the authors investigated many chemical agents, including AA and FA, but could not induce a reliable mutational signature (58). Similarly, in another study, double knockout mice where ALDH2 (Aldehyde dehydrogenase 2) and AHD5 (Alcohol dehydrogenase 5) enzymes for aldehyde detoxification were removed, the mice developed leukemia but could not produce robust mutational signatures (83). It may be possible due to the weak mutagenic property of both aldehydes for which their model system was not sensitive enough to detect all the mutations.

Details about AA- and FA-induced mutational signatures can be found in Chapter 3.

1.6 Budding yeast as a model organism

Saccharomyces cerevisiae has often been a useful choice of eukaryotic model organism to study human disease. The availability of full 12 Mb genome information, and the identification of cellular and molecular pathways sharing essential DNA replication and repair pathways found in humans, make it a suitable model organism (84). Moreover, its facile

genetics makes it suitable for mutational, carcinogenic, and toxicological studies (85). Genomic DNA is required to be in single-stranded form transiently to serve templating functions, but during this process, the nucleotides are open to react with various mutagenic molecules, increasing the magnitude of mutagenesis (86). Our temperature-sensitive yeast strains were genetically engineered to readily form long stretches of single-stranded DNA (ssDNA) to generate a high number of mutations even from weak mutagens such as AA and FA, and this significant number of mutations is required for mutational signature analysis (87). The whole genome sequencing of many cancer samples shows the presence of many mutation clusters upon the formation of ssDNA, and mutagenic agents like FA or AA can attack long stretches of ssDNA. Clusters of multiple mutations are generated when DNA damages are not repaired before replication, allowing error-prone DNA TLS polymerases to incorporate the wrong bases across these lesions (74,86).

Yeast strains have been used in many scientific studies to assess the toxicity of both AA and FA and in a few studies on their mutagenicity (87–92). In a recent study, a sensitive yeast reporter assay was used to show the mutagenic properties of AA on single-stranded DNA (ssDNA), where AA-induced mutations depend on TLS by DNA polymerase zeta (pol ζ), showing predominantly G->T mutations (93). In another study, yeast has all repair systems, such as NER, HR, and TLS, orthologous to human repair systems, so deletion of these repair genes leads to AA-induced interstrand crosslinks and DPCs (DNA-protein crosslinks), showing the role of these repair systems in acetaldehyde tolerance and contributing to the prevention of genomic instability (92). In another experiment, cis-/trans-DDP (Dichlorodiammineplatinum(II)) and FA mutagen-sensitivity test were carried out in haploid

yeast, knocking out NER genes *(rad1Δ, rad2Δ)* and interstrand cross-link repair nuclease (*smn1/pso2Δ*), including DNA-protein crosslink repair, to show how these pathways contribute to survival and lesion processing (94). Yeast offered a precise test for DPC repair because the deletion of the *WSS1* gene (a metalloprotease important for protecting against DPCs) resulted in immediate survival defects in a FA-induced mutagenesis experiment (95). Altogether, yeast has been proven to be an effective tool for carrying out different types of mutagenesis experiments (See Chapter 2 and 3) (44,49,87,96–98).

1.7 Decoding pan-cancer biology for biological pathway insights

Many mutational signatures in the PCAWG project have been associated with endogenous factors, spontaneous DNA chemical changes, and defects in DNA repair pathways (68); for example, SBS1 is associated with spontaneous deamination of methylated cytosines (59); SBS2/13 is associated with the activity of the APOBEC (Apolipoprotein B mRNA Editing Catalytic Polypeptide-like) family of cytidine deaminases (59); and SBS6/14/15 has been associated with defective DNA mismatch repair (79,99). Exogenous chemical, biological, or environmental exposure has also been associated with mutational signatures; for example, SBS4 is associated with tobacco smoking (70), all subsets of SBS7 are associated with UV light exposure (100), and SBS24 is associated with aflatoxin exposure (77,101).

However, in this PCAWG project, there were a significant number of signatures that do not have known etiology because the underlying mutational processes have not been clearly identified or experimentally validated. For example, SBS5, SBS8, SBS12, SBS16, and many other signatures are of unknown etiology (57,68). The plausible reasons behind the unknown etiology of these mutational signatures may be the complex mutational processes arising

from the combined effect of endogenous and exogenous mutations (102), low availability of these signatures in tumor cohorts (103), or rare mutational processes found in low populations. These unknown etiology signatures may appear across many cancer types but without a clear link to any established mutational signatures (104).

The integrative approach of using Gene Set Enrichment Analysis (GSEA) (105) and Gene Ontology (GO) enrichment analysis (106) to mutational signatures may be a powerful tool to understand complex mutational processes reshaping different biological pathways associated with tumor growth and vulnerabilities that mutational count alone cannot uncover (107). The NMF algorithm is used in the mutational signature deconvolution step that splits each tumor sample's mutational catalog into mutational signatures showing the highest similarity with COSMIC signatures (57). This step quantifies the DNA damage by physical or chemical agents to endogenous APOBEC activity, MMR deficiency to oxidative stress, and many other sources of DNA damage that affect the tumor's genome (108). The mutational burden is correlated with gene expression across all the samples present in the cohort, which gives a ranked list of genes that provide information about the overall gene expression landscape and identify potential highly expressed expressed genes for each mutational signature (109).

There are thousands of correlated genes, so instead of inspecting each corresponding correlation, we run GSEA analysis on the list of ranked genes of each mutational signature. GSEA can test predefined gene sets (e.g., ontology gene sets, hallmark gene sets, Reactome) and list them non-randomly at the top or bottom of the list (110,111). The upregulated gene sets are at the top, and downregulated gene sets are at the bottom of the list, with statistical significance (112). The GO enrichment analysis then converts these upregulated and

downregulated gene sets into GO terms that have biological narratives. The final biological narrative may reveal a clear and coherent cause-and-effect relationship where mutational signatures explain the type of DNA damage, while GSEA/GO analysis summarizes the cell response activity (112,113). This combination of mutational signatures with GSEA/GO will push the boundaries beyond mutational landscapes to highlight signature-specific biological associations and vulnerabilities, and may generate meaningful hypotheses that can be further tested with clinical data for disease diagnosis and treatment (107).

1.7.1 Gene Set Enrichment Analysis (GSEA)

GSEA is a computational technique that assesses whether a previously defined set of genes exhibits statistically significant and concordant differences between two biological phenotypes (like normal versus disease state) (112). The GSEA strategy was first introduced by the Broad Institute, and the fundamental concept behind GSEA is to focus on the collective action of a pre-defined group of genes rather than individual genes (114). Each group of genes (gene sets) is defined based on prior biological knowledge from multiple experiments and validation. Individual gene sets are associated with specific biological pathways, molecular functions, cellular components, chromosome location, or regulation (110). In cancer studies, the mutational processes are large and complex. The single-gene analysis cannot capture the coordinated complex changes because all the small changes by a single gene don't pass the statistical cut-off, so they seem unimportant (115). In contrast, the large coordinated changes by the group of genes can be significant and are important for the comprehensive study of mutational processes in cancer (116). It is applied to genomics and RNA-seq expression data

to identify the set of genes that has statistically significant variations in the expression matrix reflecting the genomic phenotype from upregulated to downregulated correlation (117).

We derived the mutation and RNA-Seq expression data for our computational analysis from the same samples within each cancer dataset. The MutationalPatterns package was used to reconstruct the mutational signatures. Pearson's correlation was calculated between the reconstructed mutational signature and expression data. After ranking all the genes based on Pearson's correlation coefficient, for GSEA, we repurposed an R script written by Dr. Alexandre Blais, and associated libraries were used. When GSEA is applied, it evaluates predefined sets of genes from a database, for example, the Molecular Signatures Database (MSigDB) (118), linked to the mutational signature, and ranks those gene sets to determine if those gene sets have been upregulated or downregulated based on positive or negative values of enrichment score (ES) (119). The ES is calculated by performing a random walk down the ranked list of genes. The running sum statistics increase when a gene is encountered in the ranked list of genes that is a part of the gene set, and the running sum statistics decrease when a gene is encountered in the ranked list of genes that is not a part of the gene set. The ES of each gene set is the maximum deviation from zero encountered during the walk. Each ES value of multiple gene sets is normalized to make ES scores comparable across many gene sets of different sizes. If the gene set is significantly enriched at the top of the list, the running sum statistics will show a high peak representing the high ES. The multiple gene sets requires conducting multiple hypothesis testing to calculate the false discovery rate (FDR) for each individual gene set (114,120).

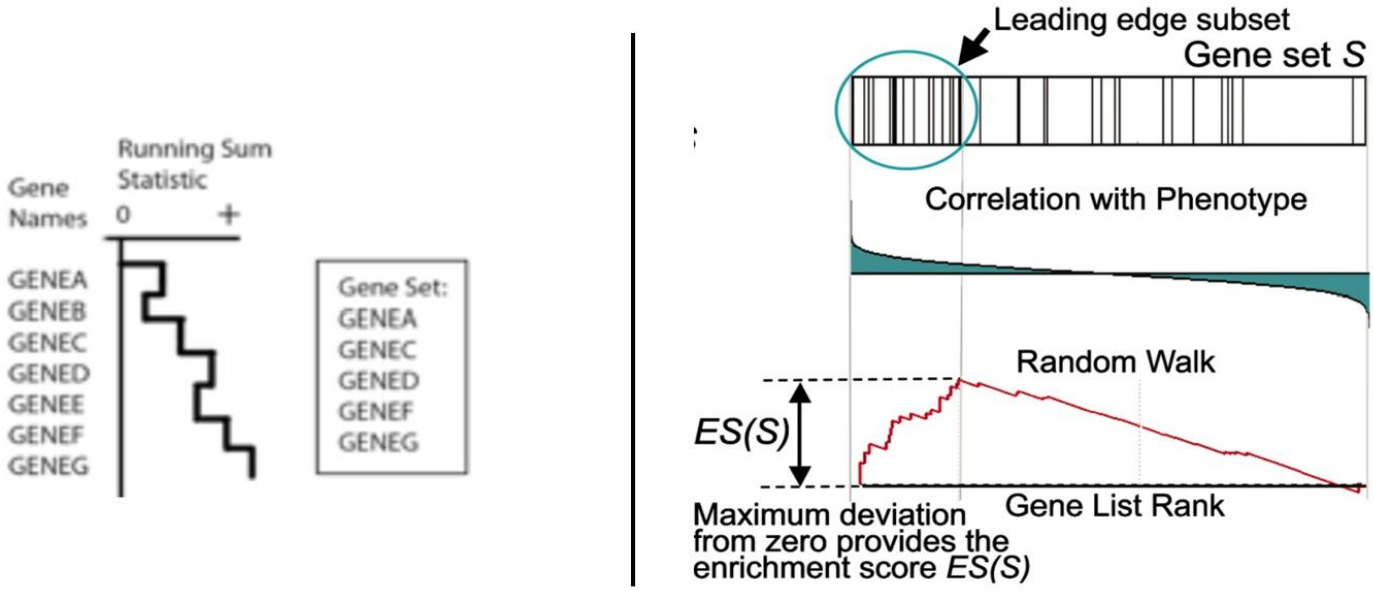


Figure 1.4: Gene Set Enrichment Analysis (GSEA) (114,120) (121)
In the right side, the random walk is over the gene list rank to find the genes in the ranked genes and in Gene set S. If there is same gene in ranked gene list and in Gene set S, the running sum statistics increase and if there is different gene, the running sum statistics decrease. The leading edge subset has core enrichment genes. In the left side, the running sum statistics increases for only those genes present in the gene set and in the gene list such as Gene A, C, D, E and F.

Here in (**Figure 1.4**), on the left side, we have a ranked list of genes, that is based on the correlation coefficient with respect to the presence of a mutational signature. The gene set has genes, such as gene A, gene C, gene D, gene F, and gene G. The random walk is performed in the ranked list of genes, and when “gene A” is encountered, the running sum statistic is increased because “gene A” is present in the gene set. In contrast, when “gene B” is encountered, the running sum statistics is decreased because “gene B” is not present in the gene set. The same principle applies down towards the “gene G” (119,122,123).

On the right side, GSEA has a three-panel enrichment plot. On the top panel, gene set S is the predefined gene set where all genes are arranged by a ranking statistic from the magnitude of correlation with the phenotype. Each vertical line present in the gene set S represents each gene position. The leading edge subset is the subset of core genes present within the gene set

that contributes most to the enrichment score (122). These are the first genes that are encountered as the random walk is performed before the running sum statistic reaches its maximum value, and biologically the most relevant genes crucial for the overrepresentation of the gene set S. The ES reflects the degree of the overrepresentation of each gene present in the gene set S at the top or at the bottom of the ranked gene list (112).

Then, multiple hypothesis testing is done by permutation test to measure the statistical significance of all enrichment scores of multiple gene sets (124). The generated nominal p-values estimate the statistical significance of each gene set's ES; adjusted p-values represent the p-values after correction for multiple hypotheses, such as FDR (125); and q-values estimate the proportion of false positives among significant results.

1.7.2 Gene Ontology (GO) Enrichment Analysis

GO enrichment analysis is performed to interpret the gene sets into biologically relevant biological processes, molecular functions, or cellular components (126). The focus of our research is on biological events, so my work is only towards the biological processes category. After GSEA, the significantly upregulated gene sets go through GO analysis. Each GO term defines particular biological processes. Multiple child GO terms can be related to single-parent GO terms. Moreover, each of these GO terms has a unique identification number (127). I used the "enrichGO" function to find enriched GO terms (113,128) and used an R package "rrvgo" to identify parent biological GO terms (129,130) to understand the broader biological context of mutational signatures.

The “enrichGO” function walks through the Gene Ontology hierarchy (131) and correlates the leading-edge subset genes of each gene set with biological process terms, applying a hypergeometric/Fisher test that returns p-, adjusted p-, and q-values for the statistical significance. The raw output is full of redundant child GO terms, so to trim all these child GO terms, the rrvgo package was used to simplify the redundancy of GO term sets based on semantic similarity by grouping similar GO terms (132,133). At first, it calculates the semantic similarity matrix to quantify the closeness between two GO terms in the GO hierarchy, and next it hierarchically clusters the child GO terms to a single parent term with statistical significance (130). For example, child GO terms such as “positive regulation of cell cycle G2/M phase transition” and “mitotic G2 DNA damage checkpoint signaling”, with GO term identification numbers GO:1902751 and GO:0007095, respectively, can have the same parent GO term “chromosome segregation”, with GO term identification number GO:0007059. Both these child terms and parent parents are associated with cell division, which is correlated with clock-like mutational signatures signatures SBS1 and SBS5 (71). In another example, child GO terms such as “positive regulation of defense response to virus by host” with GO term identification number GO:0002230 and “antiviral innate immune response” with GO term identification number GO:0140374 can have the same parent term “defense response to virus”, with GO term identification number GO:0051607 linked to mutational signatures SBS2/13 associated with known etiology of APOBEC enzyme activation halting viral DNA (59).

The importance of understanding lesser-known or signatures of unknown etiology may reveal an association with novel human carcinogens or endogenous pathways. It can shed light on DNA repair deficiencies and mutagenic properties of certain biological or chemical mutagens

(119). The discovery of the etiology behind mutational signatures may help to understand the tumor subtype classification and improve early diagnosis (134). The identification of the underlying biological mechanisms behind the unknown etiology of mutational signatures can enrich our knowledge of mutational signatures, which may help to understand the exposure level, prevention strategies, and targeted therapeutic approaches (135).

Details about pan-cancer GSEA and GO analysis to uncover biological processes of mutational signatures can be found in Chapter 4.

1.8 References

1. Hornberg JJ, Bruggeman FJ, Westerhoff HV, Lankelma J. Cancer: A Systems Biology disease. Biosystems [Internet]. 2006 Feb 1;83(2):81–90. Available from: https://www.sciencedirect.com/science/article/pii/S0303264705001176

2. Saggioro M, D'Angelo E, Bisogno G, Agostini M, Pozzobon M. Carcinoma and Sarcoma Microenvironment at a Glance: Where We Are. Front Oncol. 2020;10:76.

3. Foon K, Todd R 3d. Immunologic classification of leukemia and lymphoma. Blood [Internet]. 1986 Jul 1 [cited 2025 Jun 23];68(1):1–31. Available from: https://doi.org/10.1182/blood.V68.1.1.1

4. Pitot HC. The molecular biology of carcinogenesis. Cancer [Internet]. 1993 [cited 2025 Jun 23];72(S3):962–70. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/1097-0142%2819930801%2972%3A3%20%3C962%3A%3AAID-CNCR2820721303%3E3.0.CO%3B2-H

5. Barrett JC. Mechanisms of multistep carcinogenesis and carcinogen risk assessment. Environ Health Perspect [Internet]. 1993 Apr [cited 2025 Jun 23];100:9–20. Available from: https://ehp.niehs.nih.gov/doi/10.1289/ehp.931009

6. Ignatowicz E, Woźniak A, Kulza M, Seńczuk-Przybyłowska M, Cimino F, Piekoszewski W, et al. Exposure to alcohol and tobacco smoke causes oxidative stress in rats. Pharmacol Rep [Internet]. 2013 Jul 1 [cited 2025 Jun 23];65(4):906–13. Available from: https://doi.org/10.1016/S1734-1140(13)71072-7

7. Ding W, Hudson LG, Liu KJ. Inorganic arsenic compounds cause oxidative damage to DNA and protein by inducing ROS and RNS generation in human keratinocytes. Mol Cell Biochem [Internet]. 2005 Nov 1 [cited 2025 Jun 23];279(1):105–12. Available from: https://doi.org/10.1007/s11010-005-8227-y

8. Swierczynski J, Hebanowska A, Sledzinski T. Role of abnormal lipid metabolism in development, progression, diagnosis and therapy of pancreatic cancer. World J Gastroenterol [Internet]. 2014 Mar 7 [cited 2025 Jun 23];20(9):2279–303. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3942833/

9. Joyce C, Rayi A, Kasi A. Tumor-Suppressor Genes. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2025 [cited 2025 Jun 23]. Available from: http://www.ncbi.nlm.nih.gov/books/NBK532243/

10. Kim J, Jeong K, Jun H, Kim K, Bae JM, Song MG, et al. Mutations of TP53 and genes related to homologous recombination repair in breast cancer with germline BRCA1/2 mutations. Hum Genomics [Internet]. 2023 Jan 6 [cited 2025 Jun 23];17(1):2. Available from: https://doi.org/10.1186/s40246-022-00447-3

11. Russo IH, Russo J. Role of Hormones in Mammary Cancer Initiation and Progression. J Mammary Gland Biol Neoplasia [Internet]. 1998 Jan 1 [cited 2025 Jun 23];3(1):49–61. Available from: https://doi.org/10.1023/A:1018770218022

12. Vaissière T, Sawan C, Herceg Z. Epigenetic interplay between histone modifications and DNA methylation in gene silencing. Mutation Research/Reviews in Mutation Research [Internet]. 2008 Jul 1 [cited 2025 Jun 23];659(1):40–8. Available from: https://www.sciencedirect.com/science/article/pii/S138357420800032X

13. Wei Z, Li P, He R, Liu H, Liu N, Xia Y, et al. DAPK1 (death associated protein kinase 1) mediates mTORC1 activation and antiviral activities in CD8+ T cells. Cell Mol Immunol [Internet]. 2021 Jan [cited 2025 Jun 23];18(1):138–49. Available from: https://www.nature.com/articles/s41423-019-0293-2

14. Dąbek J, Kułach A, Gąsior Z. Nuclear factor kappa-light-chain-enhancer of activated B cells (NF-κB): a new potential therapeutic target in atherosclerosis? Pharmacological Reports [Internet]. 2010 Sep 1 [cited 2025 Jun 24];62(5):778–83. Available from: https://www.sciencedirect.com/science/article/pii/S1734114010703388

15. Zhou J, Zhou XA, Zhang N, Wang J. Evolving insights: how DNA repair pathways impact cancer evolution. Cancer Biology & Medicine [Internet]. 2020 Nov 15 [cited 2025 Jun 24];17(4):805–27. Available from: https://www.cancerbiomed.org/content/17/4/805

16. Urquidi V, Tarin D, Goodison S. Role of Telomerase in Cell Senescence and Oncogenesis. Annual Review of Medicine [Internet]. 2000 Feb 1 [cited 2025 Jun 24];51(Volume 51, 2000):65–79. Available from: https://www.annualreviews.org/content/journals/10.1146/annurev.med.51.1.65

17. An Overview of Angiogenesis and Chemical and Physiological Angiogenic Factors: Short Review - Lorestan University of Medical Sciences [Internet]. [cited 2025 Jun 24]. Available from: http://eprints.lums.ac.ir/4476/

18. Guarino M. Epithelial–mesenchymal transition and tumour invasion. The International Journal of Biochemistry & Cell Biology [Internet]. 2007 Jan 1 [cited 2025 Jun 24];39(12):2153–60. Available from: https://www.sciencedirect.com/science/article/pii/S1357272507002488

19. Hussain SP, Harris CC. Molecular epidemiology and carcinogenesis: endogenous and exogenous carcinogens. Mutation Research/Reviews in Mutation Research [Internet]. 2000 Apr 1 [cited 2025 Jun 24];462(2):311–22. Available from: https://www.sciencedirect.com/science/article/pii/S1383574200000156

20. Jiang X, Wu J, Wang J, Huang R. Tobacco and oral squamous cell carcinoma: A review of carcinogenic pathways. Tobacco induced diseases. 2019;17:29.

21. Verma N, Pandit S, Gupta PK, Kumar S, Kumar A, Giri SK, et al. Occupational health hazards and wide spectrum of genetic damage by the organic solvent fumes at the workplace: A critical appraisal. Environ Sci Pollut Res [Internet]. 2022 May 1 [cited 2025 Jun 24];29(21):30954–66. Available from: https://doi.org/10.1007/s11356-022-18889-6

22. Liu Y, Chang CCH, Marsh GM, Wu F. Population attributable risk of aflatoxin-related liver cancer: Systematic review and meta-analysis. European Journal of Cancer. 2012 Sep 1;48(14):2125–36.

23. Heintz NH, Janssen-Heininger YMW, Mossman BT. Asbestos, Lung Cancers, and Mesotheliomas. American journal of respiratory cell and molecular biology. 2010 Feb 1;42(2):133–9.

24. Oberoi S, Barchowsky A, Wu F. The Global Burden of Disease for Skin, Lung, and Bladder Cancer Caused By Arsenic in Food. Cancer epidemiology, biomarkers & prevention. 2014;23(7):1187–94.

25. Narayanan DL, Saladi RN, Fox JL. Review: Ultraviolet radiation and skin cancer. International Journal of Dermatology. 2010 Sep 1;49(9):978–86.

26. Dincer Y, Sezgin Z. Medical Radiation Exposure and Human Carcinogenesis-Genetic and Epigenetic Mechanisms. Biomedical and Environmental Sciences [Internet]. 2014 Sep 1 [cited 2025 Jun 24];27(9):718–28. Available from: https://www.sciencedirect.com/science/article/pii/S0895398814600945

27. Hussain SP, Harris CC. Molecular epidemiology and carcinogenesis: endogenous and exogenous carcinogens. Mutation Research/Reviews in Mutation Research [Internet]. 2000 Apr 1 [cited 2025 Jun 24];462(2):311–22. Available from: https://www.sciencedirect.com/science/article/pii/S1383574200000156

28. Liou GY, and Storz P. Reactive oxygen species in cancer. Free Radical Research. 2010 Jan 1;44(5):479–96.

29. Kruk J, Y. Aboul-Enein H. Reactive Oxygen and Nitrogen Species in Carcinogenesis: Implications of Oxidative Stress on the Progression and Development of Several Cancer Types. Mini-Reviews in Medicinal Chemistry. 2017;17(11):904–19.

30. Cavalieri EL, and Rogan EG. Depurinating Estrogen–DNA Adducts in the Etiology and Prevention of Breast and Other Human Cancers. Future Oncology. 2010 Jan 18;6(1):75–91.

31. Chatterjee B. The role of the androgen receptor in the development of prostatic hyperplasia and prostate cancer. Mol Cell Biochem. 2003 Nov;253(1–2):89–101.

32. Jia W, Xie G, Jia W. Bile acid–microbiota crosstalk in gastrointestinal inflammation and carcinogenesis. Nat Rev Gastroenterol Hepatol [Internet]. 2018 Feb [cited 2025 Jun 24];15(2):111–28. Available from: https://www.nature.com/articles/nrgastro.2017.119

33. Ayala A, Muñoz MF, Argüelles S. Lipid Peroxidation: Production, Metabolism, and Signaling Mechanisms of Malondialdehyde and 4-Hydroxy-2-Nonenal. Oxidative Medicine and Cellular Longevity [Internet]. 2014 [cited 2025 Jun 24];2014(1):360438. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1155/2014/360438

34. Brooks BR, Klamerth OL. Interaction of DNA with Bifunctional Aldehydes. European journal of biochemistry. 1968 Jul 1;5(2):178–82.

35. Vijayraghavan S, Saini N. Aldehyde-Associated Mutagenesis–Current State of Knowledge. Chemical research in toxicology. 2023 Jul 17;36(7):983–1001.

36. IARC. IARC monographs on the identification of carcinogenic hazards to humans. 2022;1–132.

37. World Health Organization. International agency for research in cancer. IARC Monographs on the Evaluation of the Carcinogenic Risk of Chemicals to Man. 1994;

38. Miligi L, Piro S, Airoldi C, Di Rico R, Ricci R, Paredes Alpaca RI, et al. Formaldehyde and Acetaldehyde Exposure in "Non-Traditional" Occupational Sectors: Bakeries and Pastry Producers. IJERPH [Internet]. 2023 Jan 21 [cited 2025 Sep 5];20(3):1983. Available from: https://www.mdpi.com/1660-4601/20/3/1983

39. Stornetta A, Guidolin V, Balbo S. Alcohol-derived acetaldehyde exposure in the oral cavity. Cancers (Basel). 2018;10(1):20.

40. Salaspuro M. Interrelationship between Alcohol, Smoking, Acetaldehyde and Cancer. In: Acetaldehyde-Related Pathology: Bridging the Trans-Disciplinary Divide. 2006. p. 80–96. (Novartis Foundation Symposia).

41. Roush GC, Walrath J, Stayner LT, Kaplan SA, Flannery JT, Blair A. Nasopharyngeal cancer, sinonasal cancer, and occupations related to formaldehyde: a case-control study. Journal of the National Cancer Institute. 1987 Dec;79(6):1221–4.

42. Zhong W, Que Hee SS. Formaldehyde-induced DNA adducts as biomarkers of in vitro human nasal epithelial cell exposure to formaldehyde. Mutation Research/Genetic Toxicology and Environmental Mutagenesis. 2004 Sep 12;563(1):13–24.

43. Kawanishi et al. Genotoxicity of formaldehyde: molecular basis of DNA damage and mutation. Frontiers in Environmental Science. 2014 Sep;2.

44. Thapa MJ, Chan K. The mutagenic properties of formaldehyde and acetaldehyde: Reflections on half a century of progress. Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis. 2025 Jan 1;830:111886.

45. Burge CB, Karlin S. Finding the genes in genomic DNA. Current Opinion in Structural Biology [Internet]. 1998 Jun 1 [cited 2025 Jun 24];8(3):346–54. Available from: https://www.sciencedirect.com/science/article/pii/S0959440X98800699

46. Anderson P. Chapter 2 Mutagenesis. In: Epstein HF, Shakes DC, editors. Methods in Cell Biology [Internet]. Academic Press; 1995 [cited 2025 Jun 24]. p. 31–58. (Cuenorhubditis elegans: Modern Biologcal Analysis of an Organism; vol. 48). Available from: https://www.sciencedirect.com/science/article/pii/S0091679X08613825

47. Smith KC. Spontaneous mutagenesis: Experimental, genetic and other factors. Mutation Research/Reviews in Genetic Toxicology [Internet]. 1992 Aug 1 [cited 2025 Jun 24];277(2):139–62. Available from: https://www.sciencedirect.com/science/article/pii/016511109290002Q

48. Yudkina AV, Shilkin ES, Endutkin AV, Makarova AV, Zharkov DO. Reading and Misreading 8-oxoguanine, a Paradigmatic Ambiguous Nucleobase. Crystals [Internet]. 2019 May [cited 2025 Jun 24];9(5):269. Available from: https://www.mdpi.com/2073-4352/9/5/269

49. Kelberg EP, Kovaltsova SV, Alekseev SYu, Fedorova IV, Gracheva LM, Evstukhina TA, et al. *HIM1*, a new yeast *Saccharomyces cerevisiae* gene playing a role in control of spontaneous and induced mutagenesis. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis [Internet]. 2005 Oct 15 [cited 2025 Jun 24];578(1):64–78. Available from: https://www.sciencedirect.com/science/article/pii/S0027510705001430

50. Rao V, Kumar G, Vibhavari RJA, Nandakumar K, Thorat ND, Chamallamudi MR, et al. Temozolomide Resistance: A Multifarious Review on Mechanisms Beyond O-6-Methylguanine-DNA Methyltransferase. CNS & Neurological Disorders - Drug Targets (Formerly Current Drug Targets - CNS & Neurological Disorders) [Internet]. 2023 Jul 1 [cited 2025 Jun 24];22(6):817–31. Available from: https://www.benthamdirect.com/content/journals/cnsnddt/10.2174/1871527321666220404180944

51. Pfeifer GP, Denissenko MF, Olivier M, Tretyakova N, Hecht SS, Hainaut P. Tobacco smoke carcinogens, DNA damage and p53 mutations in smoking-associated cancers. Oncogene [Internet]. 2002 Oct [cited 2025 Jun 24];21(48):7435–51. Available from: https://www.nature.com/articles/1205803

52. Vignard J, Mirey G, Salles B. Ionizing-radiation induced DNA double-strand breaks: A direct and indirect lighting up. Radiotherapy and Oncology [Internet]. 2013 Sep 1 [cited 2025 Jun 24];108(3):362–9. Available from: https://www.sciencedirect.com/science/article/pii/S0167814013002910

53. Ames et al. DNA lesions, inducible DNA repair, and cell division: three key factors in mutagenesis and carcinogenesis. 1993;

54. Ferris et al. Accelerated Evolution in Distinctive Species Reveals Candidate Elements for Clinically Relevant Traits, Including Mutation and Cancer Resistance. 2018;

55. Richard P MJ. R Loops and Links to Human Disease. 2017;

56. Helleday et al. Mechanisms underlying mutational signatures in human cancers. Nat Rev Genet. 2014;15(9):585–98.

57. Alexandrov et al. Deciphering signatures of mutational processes operative in human cancer. Cell Rep. 2013;3(1):246–59.

58. Kucab JE, Zou X, Morganella S, Joel M, Nanda AS, Nagy E, et al. A Compendium of Mutational Signatures of Environmental Agents. Cell. 2019;177(4):821-836.e16.

59. Nik-Zainal S, Alexandrov LB, Wedge DC, Van Loo P, Greenman CD, Raine K, et al. Mutational processes molding the genomes of 21 breast cancers. Cell. 2012;149(5):979–93.

60. Wheeler DA, Wang L. From human genome to cancer genome: The first decade. Genome Res [Internet]. 2013 Jul [cited 2025 Jun 24];23(7):1054–62. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3698498/

61. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin AV, et al. Signatures of mutational processes in human cancer. Nature [Internet]. 2013 Aug 22 [cited 2025 Jun 24];500(7463):415–21. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3776390/

62. Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N, Boutselakis H, et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res [Internet]. 2015 Jan 28 [cited 2025 Jun 24];43(Database issue):D805–11. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4383913/

63. COSMIC. Catalogue Of Somatic Mutations In cancer. 2019.

64. COSMIC. Catalogue Of Somatic Mutations In cancer. 2020.

65. COSMIC. Catalogue Of Somatic Mutations In cancer. 2021.

66. COSMIC. Catalogue Of Somatic Mutations In cancer. 2022.

67. COSMIC. Catalogue Of Somatic Mutations In cancer. 2023.

68. Alexandrov et al. The repertoire of mutational signatures in human cancer. Nature. 2020 Feb;578(7793):94–101.

69. Nik-Zainal S, Alexandrov LB, Wedge DC, Van Loo P, Greenman CD, Raine K, et al. Mutational Processes Molding the Genomes of 21 Breast Cancers. Cell [Internet]. 2012 May 25 [cited 2025 Jun 24];149(5):979–93. Available from: https://www.cell.com/cell/abstract/S0092-8674(12)00528-4

70. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin AV, et al. Signatures of mutational processes in human cancer. Nature [Internet]. 2013 Aug [cited 2025 Jun 24];500(7463):415–21. Available from: https://www.nature.com/articles/nature12477

71. Alexandrov LB, Jones PH, Wedge DC, Sale JE, Campbell PJ, Nik-Zainal S, et al. Clock-like mutational processes in human somatic cells. Nat Genet [Internet]. 2015 Dec [cited 2025 Jun 24];47(12):1402–7. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4783858/

72. Salter JD, Bennett RP, Smith HC. The APOBEC Protein Family: United by Structure, Divergent in Function. Trends Biochem Sci [Internet]. 2016 Jul [cited 2025 Jun 26];41(7):578–94. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4930407/

73. Vaziri C, Rogozin IB, Gu Q, Wu D, Day TA. Unravelling roles of error-prone DNA polymerases in shaping cancer genomes. Oncogene [Internet]. 2021 Dec [cited 2025 Jun 26];40(48):6549–65. Available from: https://www.nature.com/articles/s41388-021-02032-9

74. Chan K, Roberts SA, Klimczak LJ, Sterling JF, Saini N, Malc EP, et al. An APOBEC3A hypermutation signature is distinguishable from the signature of background mutagenesis by APOBEC3B in human cancers. Nature genetics. 2015;47(9):1067–72.

75. Olson ME, Harris RS, Harki DA. APOBEC Enzymes as Targets for Virus and Cancer Therapy. Cell Chemical Biology [Internet]. 2018 Jan 18 [cited 2025 Jun 26];25(1):36–49. Available from: https://www.sciencedirect.com/science/article/pii/S2451945617303884

76. Wang Y, Robinson PS, Coorens THH, Moore L, Lee-Six H, Noorani A, et al. APOBEC mutagenesis is a common process in normal human small intestine. Nat Genet [Internet]. 2023 Feb [cited 2025 Jun 26];55(2):246–54. Available from: https://www.nature.com/articles/s41588-022-01296-5

77. Nik-Zainal S, Davies H, Staaf J, Ramakrishna M, Glodzik D, Zou X, et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature [Internet]. 2016 Jun [cited 2025 Jun 24];534(7605):47–54. Available from: https://www.nature.com/articles/nature17676

78. Meier B, Volkova NV, Hong Y, Schofield P, Campbell PJ, Gerstung M, et al. Mutational signatures of DNA mismatch repair deficiency in C. elegans and human cancers. Genome research. 2018 May;28(5):666–75.

79. Hodel KP, Sun MJS, Ungerleider N, Park VS, Williams LG, Bauer DL, et al. POLE Mutation Spectra Are Shaped by the Mutant Allele Identity, Its Abundance, and Mismatch Repair Status. Molecular cell. 2020 Jun 18;78(6):1166-1177.e6.

80. Drost J, van Boxtel R, Blokzijl F, Mizutani T, Sasaki N, Sasselli V, et al. Use of CRISPR-modified human stem cell organoids to study the origin of mutational signatures in cancer. Science. 2017 Oct 13;358(6360):234–8.

81. Hodel KP, Sun MJS, Ungerleider N, Park VS, Williams LG, Bauer DL, et al. POLE mutation spectra are shaped by the mutant allele identity, its abundance and mismatch repair status. Mol Cell [Internet]. 2020 Jun 18 [cited 2025 Jun 28];78(6):1166-1177.e6. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8177757/

82. Gola M, Stefaniak P, Godlewski J, Jereczek-Fossa BA, Starzyńska A. Prospects of POLD1 in Human Cancers: A Review. Cancers (Basel) [Internet]. 2023 Mar 22 [cited 2025 Jun 28];15(6):1905. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10047664/

83. Dingler FA, Wang M, Mu A, Millington CL, Oberbeck N, Watcham S, et al. Two Aldehyde Clearance Systems Are Essential to Prevent Lethal Formaldehyde Accumulation in Mice and Humans. Molecular cell. 2020 Dec 17;80(6):996-1012.e9.

84. Dultz E, Tjong H, Weider E, Herzog M, Young B, Brune C, et al. Global reorganization of budding yeast chromosome conformation in different physiological conditions. J Cell Biol [Internet]. 2016 Feb 1 [cited 2025 Jun 24];212(3):321–34. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4748577/

85. Mohammadi et al. Scope and limitations of yeast as a model organism for studying human tissue-specific pathways. BMC Systems Biology. 2015 Dec 29;9(1):96.

86. 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–e1003149.

87. Roberts SA SJ. Clustered Mutations in Yeast and in Human Cancers Can Arise from Damaged Long Single-Strand DNA Regions. 2012;

88. Aranda A, del Olmo M lí. Response to acetaldehyde stress in the yeast *Saccharomyces cerevisiae* involves a strain-dependent regulation of several ALD genes and is mediatedby the general stress response pathway. Yeast. 2003;20(8):747–59.

89. de Smidt O, du Preez JC, Albertyn J. Molecular and physiological aspects of alcohol dehydrogenases in the ethanol metabolism of Saccharomyces cerevisiae. FEMS yeast research. 2012;12(1):33–47.

90. Tillonen et al. Role of yeasts in the salivary acetaldehyde production from ethanol among risk groups for ethanol-associated oral cavity cancer. Alcoholism, clinical and experimental research. 1999 Aug;23(8):1409–15.

91. Matsufuji Y, Fujimura S, Ito T, Nishizawa M, Miyaji T, Nakagawa J, et al. Acetaldehyde tolerance in Saccharomyces cerevisiae involves the pentose phosphate pathway and oleic acid biosynthesis. Yeast. 2008;25(11):825–33.

92. Noguchi C, Grothusen G, Anandarajan V, Martínez-Lage García M, Terlecky D, Corzo K, et al. Genetic controls of DNA damage avoidance in response to acetaldehyde in fission yeast. Cell Cycle. 2017 Jan 2;16(1):45–58.

93. Vijayraghavan et al. Acetaldehyde makes a distinct mutation signature in single-stranded DNA. Nucleic Acids Res. 2022 Jul 22;50(13):7451–64.

94. Wilborn F, Brendel M. Formation and stability of interstrand cross-links induced by cis- and trans-diamminedichloroplatinum (II) in the DNA of Saccharomyces cerevisiae strains differing in repair capacity. Current Genetics. 1989 Dec 1;16(5):331–8.

95. Stingele et al. A DNA-dependent protease involved in DNA-protein crosslink repair. Cell. 2014 Jul 17;158(2):327–38.

96. Chen D, Gervai JZ, Póti Á, Németh E, Szeltner Z, Szikriszt B, et al. BRCA1 deficiency specific base substitution mutagenesis is dependent on translesion synthesis and regulated by 53BP1. Nature Communications. 2022 Jan 11;13(1):226.

97. Magaña-Schwencke et al. Biochemical analysis of damage induced in yeast by formaldehyde I. Induction of single-strand breaks in DNA and their repair. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis. 1978 May 1;50(2):181–93.

98. Saini N, Sterling JF, Sakofsky CJ, Giacobone CK, Klimczak LJ, Burkholder AB, et al. Mutation signatures specific to DNA alkylating agents in yeast and cancers. Nucleic Acids Res. 2020;48(7):3692–707.

99. Lemmens BBLG, Tijsterman M. DNA double-strand break repair in Caenorhabditis elegans. Chromosoma. 2011 Feb 1;120(1):1–21.

100. Hayward NK, Wilmott JS, Waddell N, Johansson PA, Field MA, Nones K, et al. Whole-genome landscapes of major melanoma subtypes. Nature. 2017 May 1;545(7653):175–80.

101. Alexandrov, Jones PH, Wedge DC, Sale JE, Campbell PJ, Nik-Zainal S, et al. Clock-like mutational processes in human somatic cells. Nature genetics. 2015 Dec;47(12):1402–7.

102. Wong JKL, Aichmüller C, Schulze M, Hlevnjak M, Elgaafary S, Lichter P, et al. Association of mutation signature effectuating processes with mutation hotspots in driver genes and non-coding regions. Nature Communications. 2022 Jan 10;13(1):178.

103. Koh G, Degasperi A, Zou X, Momen S, Nik-Zainal S. Mutational signatures: emerging concepts, caveats and clinical applications. Nature Reviews Cancer. 2021 Oct 1;21(10):619–37.

104. Hu X, Xu Z, De S. Characteristics of mutational signatures of unknown etiology. NAR Cancer. 2020;2(3):zcaa026.

105. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences. 2005 Oct 25;102(43):15545–50.

106. Pesquita. Semantic Similarity in the Gene Ontology. Methods in molecular biology (Clifton, NJ). 2017;1446:161–73.

107. Chen H, Chong W, Teng C, Yao Y, Wang X, Li X. The immune response-related mutational signatures and driver genes in non-small-cell lung cancer. Cancer Sci [Internet]. 2019 Aug [cited 2025 Jun 27];110(8):2348–56. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6676111/

108. Jin SG, Padron F, Pfeifer GP. UVA Radiation, DNA Damage, and Melanoma. ACS omega. 2022 Sep 20;7(37):32936–48.

109. Reimand J, Isserlin R, Voisin V, Kucera M, Tannus-Lopes C, Rostamianfar A, et al. Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. Nature Protocols. 2019 Feb;14(2):482–517.

110. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell systems. 2015 Dec 23;1(6):417–25.

111. Yu G. Gene ontology semantic similarity analysis using GOSemSim. Stem Cell Transcriptional Networks: Methods and Protocols. 2020;207–15.

112. Shi J, Walker MG. Gene Set Enrichment Analysis (GSEA) for Interpreting Gene Expression Profiles. http://www.eurekaselect.com [Internet]. [cited 2025 Jun 24]; Available from: https://www.eurekaselect.com/article/23167

113. Yu G. Chapter 6 GO enrichment analysis | Biomedical Knowledge Mining using GOSemSim and clusterProfiler [Internet]. [cited 2025 Jun 24]. Available from: https://yulab-smu.top/biomedical-knowledge-mining-book/clusterprofiler-go.html

114. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA [Internet]. 2005 Oct 25 [cited 2025 Jul 4];102(43):15545–50. Available from: https://pnas.org/doi/full/10.1073/pnas.0506580102

115. Nesline MK, Subbiah V, Previs RA, Strickland KC, Ko H, DePietro P, et al. The Impact of Prior Single-Gene Testing on Comprehensive Genomic Profiling Results for Patients with Non-Small Cell Lung Cancer. Oncol Ther [Internet]. 2024 Jun 1 [cited 2025 Jun 27];12(2):329–43. Available from: https://doi.org/10.1007/s40487-024-00270-x

116. Lu W, Wang Q, Liu L, Luo W. Exploring the mystery of colon cancer from the perspective of molecular subtypes and treatment. Sci Rep [Internet]. 2024 May 13 [cited 2025 Jun 27];14(1):10883. Available from: https://www.nature.com/articles/s41598-024-60495-8

117. Ma Y, Sun S, Shang X, Keller ET, Chen M, Zhou X. Integrative differential expression and gene set enrichment analysis using summary statistics for scRNA-seq studies. Nat Commun [Internet]. 2020 Mar 27 [cited 2025 Jun 27];11(1):1585. Available from: https://www.nature.com/articles/s41467-020-15298-6

118. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. Bioinformatics [Internet]. 2011 Jun 15 [cited 2025 Jun 24];27(12):1739–40. Available from: https://doi.org/10.1093/bioinformatics/btr260

119. Fang Z, Liu X, Peltz G. GSEApy: a comprehensive package for performing gene set enrichment analysis in Python. Bioinformatics (Oxford, England). 2023;39(1):btac757.

120. GenePattern - GSEALeadingEdgeViewer (v5) [Internet]. [cited 2025 Jun 24]. Available from: https://www.genepattern.org/modules/docs/GSEALeadingEdgeViewer/5/

121. Reich M, Liefeld T, Gould J, Lerner J, Tamayo P, Mesirov JP. GenePattern 2.0. Nat Genet [Internet]. 2006 May [cited 2025 Jul 4];38(5):500–1. Available from: https://www.nature.com/articles/ng0506-500

122. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles | PNAS [Internet]. [cited 2025 Jun 24]. Available from: https://www.pnas.org/doi/abs/10.1073/pnas.0506580102

123. Full article: Age-related mutational signature negatively associated with immune activity and survival outcome in triple-negative breast cancer [Internet]. [cited 2025 Jun 24]. Available from: https://www.tandfonline.com/doi/full/10.1080/2162402X.2020.1788252

124. Permutation – based statistical tests for multiple hypotheses | Source Code for Biology and Medicine [Internet]. [cited 2025 Jun 24]. Available from: https://link.springer.com/article/10.1186/1751-0473-3-15

125. Gene Set Enrichment Analysis in Zebrafish Embryos Is Susceptible to False-Positive Results in the Absence of Differentially Expressed Genes - John DH Stead, Hyojin Lee, Andrew Williams, Sergio A Cortés Ramírez, Ella Atlas, Jan A Mennigen, Jason M O'Brien, Carole Yauk, 2025 [Internet]. [cited 2025 Jun 24]. Available from: https://journals.sagepub.com/doi/full/10.1177/11779322251321071

126. Zhou T, Yao J, Liu Z. Gene Ontology, Enrichment Analysis, and Pathway Analysis. In: Liu Z (John), editor. Bioinformatics in Aquaculture [Internet]. 1st ed. Wiley; 2017 [cited 2025 Jun 24]. p. 150–68. Available from: https://onlinelibrary.wiley.com/doi/10.1002/9781118782392.ch10

127. Grossmann S, Bauer S, Robinson PN, Vingron M. Improved detection of overrepresentation of Gene-Ontology annotations with parent–child analysis. Bioinformatics [Internet]. 2007 Nov 15 [cited 2025 Jun 24];23(22):3024–31. Available from: https://doi.org/10.1093/bioinformatics/btm440

128. Da Ros LU, Bastiani MAD, Lobato LW, Zimmer ER. Transcriptomics similarities between Alzheimer's Disease and Cardiovascular Disease. Alzheimer's & Dementia [Internet]. 2023 Dec [cited 2025 Jun 24];19(S15):e079678. Available from: https://alz-journals.onlinelibrary.wiley.com/doi/10.1002/alz.079678

129. Sayols S. rrvgo: a Bioconductor package for interpreting lists of Gene Ontology terms. microPublication biology. 2023;2023.

130. REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms | PLOS One [Internet]. [cited 2025 Jun 24]. Available from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0021800

131. Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Research [Internet]. 2004 Jan 1 [cited 2025 Jun 24];32(suppl_1):D258–61. Available from: https://doi.org/10.1093/nar/gkh036

132. Sevilla JL, Segura V, Podhorski A, Guruceaga E, Mato JM, Martinez-Cruz LA, et al. Correlation between Gene Expression and GO Semantic Similarity. IEEE/ACM Trans Comput Biol and Bioinf [Internet]. 2005 Oct [cited 2025 Jun 24];2(4):330–8. Available from: http://ieeexplore.ieee.org/document/1541985/

133. Yu G, Li F, Qin Y, Bo X, Wu Y, Wang S. GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. Bioinformatics (Oxford, England). 2010;26(7):976–8.

134. Hoeck et al. Portrait of a cancer: mutational signature analyses for cancer diagnostics. 2019;

135. Brady SW, Gout AM, Zhang J. Therapeutic and prognostic insights from the analysis of cancer mutational signatures. Trends in Genetics. 2022;38(2):194–208.

Chapter 2: Mutagenic Properties of Acetaldehyde and Formaldehyde

2.1 Copyright and License Policies

M. J. Thapa and K. Chan, “The mutagenic properties of formaldehyde and acetaldehyde: Reflections on half a century of progress,” *Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis*, vol. 830, p. 111886, Jan. 2025, doi: https://doi.org/10.1016/j.mrfmmm.2024.111886.

2.2 Rationale for the literature survey of AA and FA

The trajectory of my thesis began with the investigation of the mutagenicity of small and weak yet reactive aldehydes, AA and FA, because the genotoxicity of these two aldehydes in the form of DNA adducts, DPCs, inter- and intra-strand crosslinks has been strongly documented (1–4), but their mutagenic properties and its consequences have been less understood. In this published review paper entitled “The mutagenic properties of formaldehyde and acetaldehyde: Reflections on half a century of progress”, a critical literature survey was done on the mutagenicity, clastogenicity, carcinogenicity, and point mutations of AA and FA by various research groups on multiple experimental systems, demonstrating consensus results. The blind spot of mutagenic properties was interesting because these weak aldehyde mutagens demonstrated potential to shape the somatic mutational landscape by increasing the mutational burden and increase the risk of cancer (1,5,6). Thus, this review paper has provided a foundation of knowledge that one may come across while designing experiments to generate AA- and FA- induced mutational spectra.

The mutagenic properties of formaldehyde and acetaldehyde: reflections on half a century of progress

Mahanish Jung Thapa and Kin Chan*

*Corresponding Author: kin.chan@uottawa.ca

Department of Biochemistry, Microbiology and Immunology

Ottawa Institute of Systems Biology

University of Ottawa Faculty of Medicine

451 Smyth Road

Ottawa, Ontario K1H 8M5

Canada

2.4 Abstract

Formaldehyde and acetaldehyde are reactive, small compounds that humans are exposed to routinely, variously from endogenous and exogenous sources. Both small aldehydes are classified as human carcinogens. Investigation of the DNA damaging properties of these two compounds began some 50 years ago. In this review, we summarize progress in this field since its inception over half a century ago, distilling insights gained by the collective efforts of many research groups while highlighting areas for future directions. Over the decades, general consensus about aspects of the mutagenicity of formaldehyde and acetaldehyde has been reached. But other characteristics of formaldehyde and acetaldehyde remain incompletely understood and require additional investigation. These include crucial details about the mutational signature(s) induced and possible mechanistic role(s) during carcinogenesis.

2.5 Introduction

Aldehydes are a group of organic compounds that have a carbonyl group (C=O) linked to a hydrogen atom (R-CHO). These molecules are found in both natural and synthetic substances which are important for biological and commercial processes (7). The two simplest aldehydes, formaldehyde and acetaldehyde, have been of particular interest because of their prevalence and possible health effects (8).

Formaldehyde (CH_2O, abbreviated as FA) is the smallest aldehyde. It is used in many industrial applications. FA is an important ingredient for the production of disinfectants and preservatives (9). FA is also a major by-product from mainly the catabolism of serine, which feeds into one-carbon metabolism involving tetrahydrofolate (THF), which forms 5,10-methylene-tetrahydrofolate (5,10-CH_2-THF) and serves as an intermediate carrier for one-carbon units. However, 5,10-CH2-THF can undergo spontaneous decomposition to release FA (10). FA also results from oxidative demethylation reactions catalyzed by members of three families of demethylases (11,12). Additional details of formaldehyde metabolism are discussed in Section 2.10. The endogenous biochemistry of FA results in concentrations typically determined to be in the range of tens of micromolar in blood (see references in **Table 2.1**). While there are some measurements in the low micromolar range, the large majority of studies point to blood concentrations of FA >10 μM. The International Agency for Research on Cancer (IARC) classifies formaldehyde as a bona fide Group 1 human carcinogen, with sufficient evidence of carcinogenicity in both humans and model systems (13).

Species	[FA] in reported units	[FA] in µM	References
Human	2.61 µg/g	92.1 µM	(14)
Human	~0.4 - 0.6 µg/mL	~13 - 20 µM	(15)
Human	1.34 µg/mL	44.6 µM	(16)
Human	~0.073 - 0.095 mM	~73 - 95 µM	(17)
Human	~0.076 mM	~76 µM	(18)
Human	71.9 µM	71.9 µM	(19)
Human	1.52 µg/mL	50.6 µM	(20)
Mouse	~20 - 25 µM	~20 - 25 µM	(21)
Mouse	~0.078 mM	~78 µM	(18)
Mouse	4 µM	4 µM	(22)
Rat	2.24 µg/g	79.1 µM	(14)
Rat	< 0.3 µg/mL	< 10 µM	(23)
Rat	≤ 0.20 µg/mL	≤ 6.7 µM	(24)
Rat	2.71 mg/L	90.3 µM	(25)
Rat	~0.06 - 0.07 mM	~60 - 70 µM	(26)
Rhesus monkey	2.42 µg/g	85.4 µM	(27)
Yucatan minipig	1.98 ng/µL	65.9 µM	(28)

Table 2.1: Concentrations of formaldehyde measured in the blood of healthy humans and wild-type mammalian species.

The next simplest aldehyde, acetaldehyde (CH_3CHO, abbreviated as AA), is less reactive than FA since the presence of a methyl group renders the carbonyl carbon atom less electrophilic. AA is found in plants and fruits, cosmetics, and tobacco smoke. AA is also generated endogenously via the metabolism of ethanol (29). IARC classifies AA in Group 2B as a possible human carcinogen, because of limited evidence in humans but sufficient evidence in experimental animals (8). But crucially, AA associated with the consumption of alcoholic beverages is categorized into Group 1 (i.e., as a known human carcinogen) because there is sufficient human epidemiological evidence in oesophageal, as well as head and neck cancers (30).

Both exogenous and endogenous forms of AA and FA can damage DNA by chemically reacting to yield pro-mutagenic adducts, accounting for observed mutagenic and potentially carcinogenic

effects (31). Both aldehydes can produce variously: point mutations; single- and double-strand DNA breaks; bulky adducts; and inter- and intrastrand crosslinks (32) (see **Figure 2.1**). AA and FA have both been linked to DNA damage and mutations that may have an impact on the development of diseases, including most notably cancers (33,34). In this review, we survey and compile the historical evidence in the literature up to the recent investigations into the mutagenicity of AA and FA.

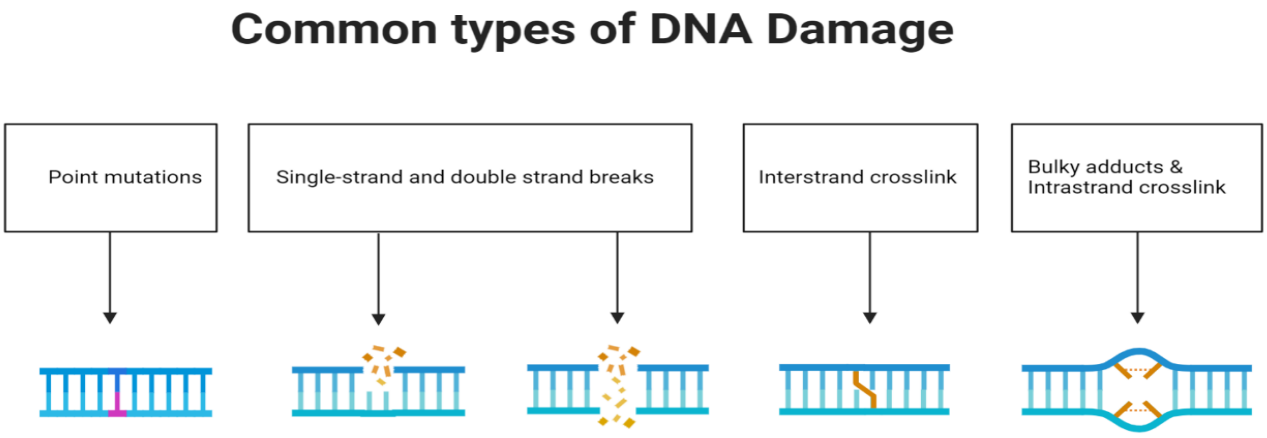


Figure 2.1: Exogenous and endogenous aldehyde-induced DNA damage.

2.6 Ethanol and acetaldehyde metabolism in humans

Humans are exposed to low levels of exogenous AA in ambient air and the diet, but the main route of exposure is via alcoholic beverages (35). In humans, almost all consumed ethanol is metabolized in the liver. In the oxidative pathway, ethanol is metabolized to toxic AA by the action of NAD^+-dependent alcohol dehydrogenase (ADH). AA is then further metabolized to non-toxic acetate by the action of aldehyde dehydrogenase (ALDH) enzymes (see **Figure 2.2**). ADH and ALDH enzymes play an important role in the accumulation and metabolism of AA. There are five ADH enzymes that convert ethanol to AA (36). Polymorphisms in the genes

coding for these enzymes affect the levels of AA produced (29,37). There are two major forms of ALDH, cytoplasmic and mitochondrial, encoded by the ALDH1 and ALDH2 genes, respectively. Of the two, mitochondrial ALDH2 plays the major role in AA metabolism to acetic acid (38).

AA is also produced from ethanol by bacterial-mediated oxidation in the upper aerodigestive tract (UADT) and large intestine, resulting in further exposure to AA. The microbes in the oral microflora show ADH activity under aerobic conditions. Out of many bacterial genera, *Neisseria* species produced the highest amount of AA while other species like *Rothia mucilaginosa*, *Streptococcus mitis* and *Prevotella histicola* also showed the ability to produce AA (39).

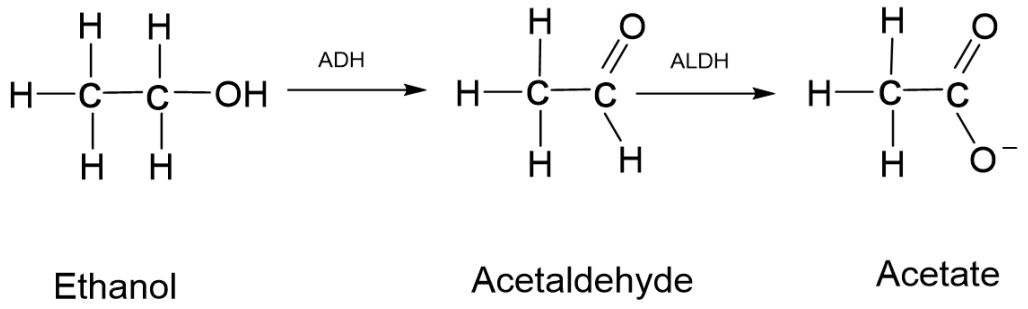


Figure 2.2: Ethanol metabolism in humans.
Alcohol dehydrogenases (ADHs) metabolize ethanol to form toxic AA. ADH is encoded by a family of genes, and mutations in these genes lead to different ADH enzyme isoforms. AA is further metabolized to form non-toxic acetate by the action of aldehyde dehydrogenases (ALDHs). Mutations in the ALDH genes impact an individual's ability to metabolize alcohol.

2.7 Cancer epidemiology associated with ethanol-driven acetaldehyde

In 2020, some 731,300 or 4.1% of new cancer diagnoses around the world were attributed to ethanol consumption, with many such cases being ethanol-driven AA-associated UADT cancers (40,41). About 76.7% of alcohol-related cancer cases were in males, and likely involved AA from ethanol. These include larynx, pharynx, lip, oral cavity and oesophagus cancers. The remaining

23.3% of alcohol-related cancers were in females, where ethanol-driven AA-associated cancers include lip, oral cavity, and breast cancer, with the latter being predominant (see **Figure 2.3**) (29,40,42).

AA exposure in the UADT occurs when the oral flora contributes to salivary AA production when consuming ethanol. This carcinogenicity of local AA is observed in East Asian alcohol drinkers who carry a single point mutation in the ALDH2 gene, resulting in two- to three-fold higher AA concentrations in saliva (43) and some 30-fold higher accumulation of AA in blood (43,44). About 28% to 45% of East Asian populations have the ALDH2*2 (rs671G>A) dominant negative allele, while the proportion is much less in Caucasians. When the activity of ALDH2 is reduced, the accumulation of AA is increased, so alcohol drinkers with the ALDH2*2 variant allele have a higher risk of the oesophageal and UADT cancers (45). Moreover, oropharyngeal and larynx mucosa have relatively low intrinsic ALDH activity, limiting the ability to detoxify AA and leading to increased local AA exposure with cancer risk (46–50). Indeed, a number of studies have reported that alcohol-associated AA exposure has detrimental effects in the oral cavity of humans, which results in UADT cancers associated with ALDH2 polymorphism (51–54). Moreover, recent meta-analyses showed that the ALDH2*2 variant may be related to cancer incidence overall in Asians, but especially for oesophageal, and head and neck, cancers (55,56). Similarly, there is consensus that ALDH2*2 correlates with increased risk of stomach cancer (57), although there were two studies among Chinese populations that found no statistical correlation between ALDH2*2 and stomach cancer (58,59). While the effects of ALDH2*2 on UADT cancer risk are well established, correlations between ALDH activity/deficiency and other cancer types associated with alcohol consumption are either negative or inconclusive. Most

studies investigating liver cancers found no correlation with ALDH2*2. Studies on breast cancer, including a very recent one (60), also found no correlation. And analyses on colorectal cancers have been rather mixed, with no clear-cut consensus. These results in the literature have been summarized previously (45). A more recent study among the Chinese population largely agreed with the consensus results across these cancer types (59).

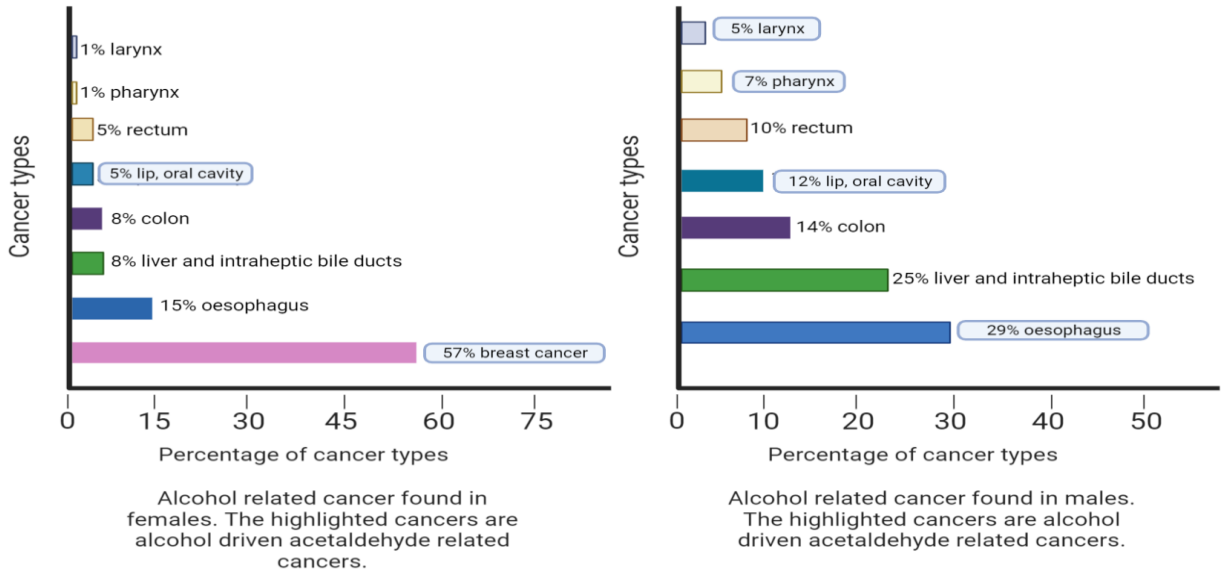


Figure 2.3: Alcohol-associated cancer types found in males and females in 2020.
The oesophagus and breast cancer cases were most frequent in males and females, respectively. Other alcohol-associated cancer types include liver, colon, pharynx, larynx, lip and rectum in both sexes (61).

2.8 Acetaldehyde-induced DNA damage

2.8.1 Acetaldehyde-induced DNA adducts

Mixtures of various adducts are readily generated when AA reacts with DNA. The major DNA adducts, N^2-ethylidene-2'-deoxyguanosine (N^2-ethylidene-dG) and N^2-ethyl-2'-deoxyguanosine (N^2-ethyl-dG) are formed when a molecule of AA reacts with the exocyclic amino group of guanine. An important minor DNA adduct is formed when two molecules of AA react with guanine to form 1,N^2-propano-2'-deoxyguanosine (1,N^2-propano-dG) (62). Other minor adducts

including N^2-(2,6-Dimethyl-1,3-dioxan-4-yl)-deoxyguanosine (N^2-Dio-dG), as well as α-*S*- and α-*R*-methyl-γ-hydroxy-1,N^2-propano-2'-deoxyguanosine (α-*S*-Me-γ-OH-PdG and α-*R*-Me-γ-OH-PdG) have been reported (53). These and other AA-induced DNA adducts are detectable in many experimental systems (summarized in **Table 2.2**).

Experimental systems	**DNA adducts**	**References**
DNA and dG	N^2-ethylidene-dG, 1,N^2-propano-dG, N^2-dimethyldioxane-dG	(62)
DNA and dG	N^2-ethylidene-dG, N^2-ethyl-dG	(63,64)
DNA	N^2-ethyl-dG, 1,N^2-propano-dG	(65–67)
DNA packaged with histones	1,N^2-propano-dG	(68,69)
HL-60 human leukemia cell line	1,N^2-propano-dG, N^2-ethyl-dG	(65)
Human leukocytes	N^2-ethylidene-dG	(70)
Human oral mucosa	N^2-ethylidene-dG	(71)
Human oral mucosa	22 adducts of various abundances	(72)
Human lung and liver	1,N^2-propano-dG	(73)
Human peripheral blood cells	N^2-ethylidene-dG	(74–76)
Human peripheral blood cells from alcoholic donors	N^2-ethylidene-dG, N^2-ethyl-dG, α-*S*-Me-γ-OH-PdG, α-*R*-Me-γ-OH-PdG	(53,77)
IMR-90 human normal fibroblast cell line	1,N^2-propano-dG, 1,N^2-etheno-dG	(78)
Mouse esophagus (with human ALDH2*2 knock-in)	N^2-ethylidene-dG	(79)
Mouse: multiple organs of ALDH2 knockout	N^2-ethylidene-dG	(80–84)
Rat brain and lung	1,N^2-propano-dG	(85)
Rat brain, liver, and lung	1,N^2-propano-dG, 1,N^2-etheno-2'-dG	(86)
Rat respiratory/olfactory epithelia	N^2-ethyl-dG	(87)
Rat, various tissues	N^2-ethylidene-dG	(88)
Rhesus monkey oral mucosa	N^2-ethylidene-dG, 1,N^2-propano-2'-dG	(89)
TK6 human lymphoblast cell line	N^2-ethylidene-dG	(90)

Table 2.2: Acetaldehyde-induced DNA adducts found in different experimental systems. Major adducts are listed first. Some studies were targeted specifically to detect certain adducts, while other studies sought to identify as many adducts as possible.

Studies using systems ranging from purified DNA, cell lines, animal models, and human donors generally agree in terms of the major adducts produced by reacting AA with DNA, namely N^2-ethylidene-dG, N^2-ethyl-dG, and 1,N^2-propano-dG. There is considerably more variance with respect to detection of minor adducts, e.g., one study using a sensitive mass spectrometry approach reported as many as 22 adducts in total (72). Additional investigations are needed to resolve the identity of minor adducts more conclusively. It has been suggested that base excision repair (BER) could play a role in tolerance to AA, perhaps by repairing these types of base damage, which are relatively compact (91). BER is a repair pathway that relies on specific glycosylases to recognize particular kinds of base damage, followed by excision of the damaged base, and repair synthesis (recently reviewed in (92)). If the adducted base causes a distortion of the DNA double helical geometry, then nucleotide excision repair (NER) can recognize the damage. The damage would be excised by cleaving the sugar-phosphate backbone 5' and 3' of the damage site. Repair synthesis then fills in the single-strand gap (recently reviewed in (93)). There is evidence in yeasts and human cells that both BER and NER contribute to repair of AA-induced base damage (91,94,95).

2.8.2 Acetaldehyde-induced crosslinks

Another common type of damage induced by AA is crosslinking. DNA-protein and interstrand crosslinks have been observed in many experimental systems, while intrastrand crosslinks have been found in a few in vitro studies (summarized in **Table 2.3**). Interstrand crosslinks are mainly repaired by the Fanconi anemia pathway, named after the Swiss physician who first described the syndrome (Fanconi anemia) when this pathway is mutated (96). Fanconi anemia is a

congenital genome instability syndrome resulting in morphological and bone marrow abnormalities, as well as cancer predisposition. The Fanconi anemia pathway consists of some 22 proteins that enable recognition and "unhooking" of the crosslink, as well as facilitating repair utilizing strand invasion (reviewed in (97)). An independent repair pathway requiring replication fork convergence that breaks the AA-induced crosslink has also been described (98). Repair of DNA-protein crosslinks can involve multiple mechanisms, including cleaving the crosslink, or pathways coupled to replication, proteolysis, or nucleolysis. These have been reviewed recently (99). Intrastrand crosslinks of nearby bases are subject to NER (100).

Experimental systems	**Types of crosslinks**	**References**
Cancer cells deficient for Fanconi anemia pathway	Interstrand crosslinks	(101)
Chinese hamster ovary cells	DNA-protein crosslinks	(102,103)
DNA	Unspecified DNA crosslinks	(104)
DNA	Intrastrand crosslinks	(105,106)
DNA and dG	Interstrand crosslinks	(62)
DNA mixed with histones	DNA-protein crosslinks	(107,108)
Fission yeast	Interstrand and DNA-protein crosslinks	(91)
Human bronchial epithelial cells and fibroblasts	DNA-protein crosslinks	(109)
Human bronchial epithelial cells and fibroblasts	Interstrand crosslinks, DNA-protein crosslinks	(110)
Human fibroblasts	DNA-protein crosslinks	(102)
Human fibroblast cell lines	Intrastrand and interstrand crosslinks	(111)
Human leucocytes	Unspecified DNA crosslinks	(112)
Human lymphocytes, gastric and colonic mucosa cells	Unspecified DNA crosslinks	(113)
Human lymphoma cells	DNA-protein crosslinks	(114)
Rat nasal mucosa	DNA-protein crosslinks	(115,116)

Table 2.3: AA-induced crosslinks in different experimental systems.

2.8.3 Acetaldehyde-induced strand breaks

AA can also cause strand breaks in DNA. The vast majority of studies in a broad range of systems have found induction of double-strand breaks (DSBs), while some authors also reported evidence for single-strand breaks (SSBs) (summarized in **Table 2.4**). Molecular mechanisms for DNA strand break repair have been reviewed extensively by many authors, e.g., (32). DSBs are repaired by either homologous recombination (HR) or non-homologous end joining (NHEJ) (117), while SSBs can be repaired via a distinct repair mechanism (118).

Experimental systems	DNA strand breaks	References
A549 human alveolar basal adenocarcinoma cell line	Double-strand breaks	(119)
Cancer cells deficient for Fanconi anemia pathway	Double-strand breaks	(101)
Chinese hamster embryonic diploid cells	Double-strand breaks	(120)
DLD1 human colorectal adenocarcinoma cell line	Double-strand breaks	(121)
DNA	Double-strand breaks	(122)
Human bronchial epithelial cells and fibroblasts	Single-strand breaks	(109)
Human lung fibroblasts	Double strand breaks	(123)
Human lymphocytes	Single-strand and double-strand breaks	(124)
Human neurons in culture	Double-strand breaks	(125)
Mouse bone marrow	Double-strand breaks	(126)
Rat astrocytes	Single-strand breaks	(127)
Rat brain cells	Double-strand breaks	(125)
Rat skin fibroblasts	Double-strand breaks	(128)
U2OS human osteosarcoma cell line	Double-strand breaks	(129)
V79 Chinese hamster lung fibroblast cell line	Double-strand breaks	(130)

Table 2.4: AA-induced DNA strand breaks in different experimental systems.

2.8.4 Acetaldehyde-induced mutations

While DNA repair mechanisms can be quite efficient, when they are unable to restore the original base pairing(s), AA-induced mutations can result. Further, when DNA damage escapes repair (e.g., in single-stranded regions), then error-prone TLS polymerases can bypass the damage, often creating mutations in the process (131). Results from the literature are summarized in **Table 2.5**. Multiple systems using site-specific AA-induced lesions generally agree that G > T transversions are the most common substitutions (132–135). Consistent with this, two recent studies in budding yeast using genome sequencing confirmed that G > T transversions are predominant (5,136). Additionally, forward mutagenesis reporters using the *HPRT* and *TP53* genes reported mostly G > A transitions (137–139) and/or large deletions (140–142). A yeast study also found evidence for AA-induced deletions of 5 or more bases (5). Taken altogether, the consensus view from the literature is that AA tends to induce point mutations at G's (especially G > T transversions) and deletions. This is consistent with adducted G's being mutagenic, as confirmed by multiple studies using site-specific adducts on plasmids (98,133–135,143).

Experimental systems	**Mutation types**	**References**
Budding yeast	Predominantly G > T	(136)
Budding yeast	Predominantly G > T, deletions of ≧ 5 bases	(5)
E. coli DNA polymerase I Klenow fragment copying site-specific N^2-ethyl-dG	G > C	(143)
E. coli replicating plasmid with site-specific G-G interstrand crosslink	G > T	(132)
E. coli replicating plasmid with site-specific N^2-ethyl-dG	Single base deletions, G > T substitutions	(133)
Human DNA polymerase eta replicating plasmid treated with AA	GG > TT tandem substitutions	(144)

Human fibroblast cell lines, nucleotide excision repair proficient and deficient	GG > TT tandem substitutions	(111)
Human embryonic kidney cell line 293 replicating plasmid with site-specific N^2-ethyl-dG	Single base deletions, point mutations (mostly G > T), large deletions	(134)
Human fibroblasts, budding yeast	Mostly G > A substitutions at TP53 locus	(137,138)
Human immortalized XPA cells replicating plasmid with site-specific 1,N^2-propano-dG	Mostly G > T substitutions	(135)
Human lymphocytes	Large deletions at HPRT locus	(140,141)
Human lymphocytes	Large deletions and unspecified point mutations at HPRT locus	(142)
Human lymphocytes	Mostly G > A substitutions at HPRT locus	(139)
Xenopus egg extract repairing site-specific interstrand crosslink	Predominantly G > T	(98)

Table 2.5: AA-induced mutations in different experimental systems.

2.9 Acetaldehyde-related carcinogenesis in model species

Many studies have investigated the carcinogenicity of ethanol and/or AA in rodents (summarized in **Table 2.6**). Most protocols used long-term exposure to ethanol via drinking water, revealing increased incidence of multiple cancer types. While chronic ingestion of ethanol has multiple effects, a contribution of AA toward mutagenesis is entirely plausible. Indeed, a study which administered AA itself via drinking water induced a variety of cancer types reminiscent of those observed in the long-term ethanol studies (145). Although not related to alcohol consumption, AA by inhalation did induce upper respiratory cancers in hamsters and rats as well (146–149). Taken together, there is ample evidence implicating AA as being carcinogenic in model rodents.

Species (sex)	Exposures	Cancer sites	References
Hamsters (both)	AA by inhalation	Nasal mucosa and larynx	(146,147)
Mice (both)	Ethanol in drinking water	Intestine	(150)
Mice (female)	Ethanol in drinking water	Mammary	(151)
Mice (male)	Ethanol in drinking water	Liver	(152)
Mice deficient for mismatch repair (both)	Ethanol in drinking water	Colon	(153)
Mice deficient for tumor suppressor APC (male)	Ethanol in drinking water	Intestine	(154)
Rats (female)	AA in drinking water	Hematopoietic, mammary, uterus	(145)
Rats (male)	AA in drinking water	Head, hematopoietic, pancreas	(145)
Rats (both)	AA by inhalation	Nasal mucosa	(148,149)
Rats (male)	Ethanol in drinking water	Adrenal, liver, pancreas, pituitary	(155)
Rats (female)	Ethanol in drinking water	Hematopoietic, oral	(156)
Rats (male)	Ethanol in drinking water	Head and neck, oral, stomach, testes	(156)
Rats bred to prefer alcohol (male)	Ethanol in drinking water	Liver	(157)

Table 2.6: AA-related carcinogenesis in model species

2.10 Formaldehyde metabolism

Humans are ubiquitously exposed to FA from exogenous and (especially) endogenous sources. Typically, humans are exposed to very low levels of exogenous FA (in the parts per billion range) from ambient air and from food consumption. Indoor FA concentrations can be higher than the outdoors environment due to off-gassing from household items and/or construction materials. Moreover, occupational exposures during work activities that use FA or FA-based agents can be

much higher, in the parts per million range. These include multiple manufacturing industries; histology, pathology, or biology labs; and funeral home and related services (158).

FA is also generated endogenously on a continuing basis, resulting in steady state intracellular concentrations thought to be in the range of tens to hundreds of micromolar (see **Table 2.7**). But ratiometric fluorescence experiments with cultured cells do yield considerably lower estimates of intracellular FA concentration (25 ± 19 μM) (159). A major source of endogenous FA is the one-carbon (1-C) cycle after the enzymatic cleavage of serine by its cognate hydroxymethyltransferases (160) to generate glycine and FA. FA can also be introduced into 1-C metabolism via cleavage of glycine (161) and choline metabolism (162).The FA then condenses with tetrahydrofolate (THF), the active form of the co-factor folate (vitamin B9) to yield 5,10-CH_2-THF. This compound is essentially a transient carrier for the 1-C unit from FA, which is converted to a reactive methyl group on S-adenosylmethionine (SAM). SAM is a universal methyl donor for many biochemical reactions. 5,10- CH_2-THF can also undergo spontaneous oxidative decomposition, releasing FA as a product (163). Free FA can be scavenged by glutathione to form S-hydroxymethylglutathione (HMGSH), which is oxidized into S-formylglutathione by the action of ADH5 (alcohol dehydrogenase 5, also formerly called ADH3) using NAD(P)+ as co-factor (164). S-formylglutathione undergoes hydrolysis by S-formylglutathione hydrolase (FGH) yielding formate and regenerating reduced glutathione (165) (see **Figure 2.4**).

Experimental system	[FA] in reported units	[FA] in µM	References
Human brain	0.047 mM/g [sic]	47 µM	(166)
Human brain	0.286 mM	286 µM	(167)
Human cortex	~0.39 mM	~390 µM	(17)
Human hippocampus	~0.22 mM	~220 µM	(17)
Mouse brain	0.19 - 0.22 mM/g [sic]	190 - 220 µM	(166)
Mouse brain	~0.33 - 0.43 mM	~330 - 430 µM	(17)
Mouse brain	~78 - 85 µmol/kg	~83 - 90 µM	(168)
Mouse brain	~12 - 14 µM	~12 - 14 µM	(169)
Mouse brain	~0.02 mM	~20 µM	(170)
Mouse brain	0.2705 mM	270.5 µM	(167)
Mouse brain	~0.3 mM	~300 µM	(171)
Mouse cerebellum	~0.45 mM	~450 µM	(172)
Mouse hippocampus	~0.35 mM	~350 µM	(18)
Mouse kidney	~1.2 mM	~1200 µM	(172)
Mouse liver	~0.45 mM	~450 µM	(172)
Mouse muscle	~0.25 mM	~250 µM	(172)
Mouse spleen	~0.5 mM	~500 µM	(172)
Rat brain	0.097 µmol/g	103 µM	(173)
Rat hippocampus	~0.27 - 0.52 mM	~270 - 520 µM	(174)
Rat hippocampus	~0.23 - 0.45 mM	~230 - 450 µM	(17)
Rat hippocampus	~0.22 - 0.48 mM	~220 - 480 µM	(175)
Rat hippocampus	~0.3 mM	~300 µM	(176)
Rat liver	0.201 µmol/g	213 µM	(173)
Rat liver	~0.23 - 0.48 mM	~230 - 480 µM	(26)
Rat nose	0.42 µmol/g	445 µM	(173)
Rat testes	0.28 µmol/g	297 µM	(173)

Table 2.7: Formaldehyde concentrations measured in tissues of humans and wild-type mammalian species.

ADH5 mitigates against FA toxicity and supports nucleic acid metabolism by providing 1-C units for nucleotide biosynthesis. As a result, human cells deficient in serine catabolism have a growth defect in the absence of the nucleotide precursors hypoxanthine, thymidine, and formate. Growth is further impaired in ADH5-deficient cells, suggesting that the conversion of

endogenous formaldehyde into formate provides 1-C units for nucleotide synthesis (10). Deletion of ADH5 in mice leads to increased levels of formaldehyde-induced DNA adducts in the kidney, liver, and bone marrow, confirming the importance of ADH5 in formaldehyde detoxification (177–180).

Another potential source of FA is via the action of demethylases. In the context of epigenetic regulation, demethylase enzymes remove the methyl groups from substrates like histones and DNA. These demethylases belong to three protein families. The Jumonji (JmjC) enzymes catalyze most histone demethylations (181); AlkB removes alkyl (including methyl) groups from DNA (182); and lysine-specific demethylases, which can also demethylate histones (183). While the molecular mechanism of each demethylase family is distinct from the others, the common net result is release of FA. Hence, oxidative demethylations are another source of FA in mammalian cells.

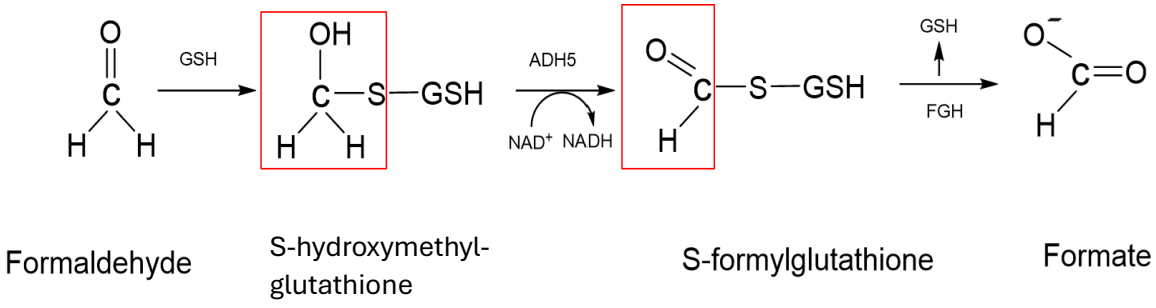


Figure 2.4: Formaldehyde detoxification
Formaldehyde reacts with glutathione (GSH) to form S-hydroxymethylglutathione, which is converted in turn to S-formylglutathione by Alcohol dehydrogenase 5 (ADH5). Finally, non-toxic formate is generated via S-formylglutathione hydrolase (FGH) activity.

2.11 Cancer epidemiology associated with formaldehyde

The epidemiology of FA-induced cancers has been studied extensively. Large-scale studies have been particularly useful to elucidate the association between occupational exposure to FA and cancer mortality. For example, analyses of a large cohort of 25,619 workers in FA-utilizing industries reported increased mortality from nasopharyngeal cancers (184) and myeloid leukemias (185). Similarly, an analysis of 6,808 funeral home workers found that embalmers had higher FA-associated mortality from myeloid leukemias (186). Another study of 11,039 garment workers confirmed this correlation (187). An analysis of 14,014 male workers in FA-utilizing industries also showed increased mortality from lung cancers (188).

Studies on increased cancer risk were also carried out, as mortality data would miss cases that did not result in death. A meta-analysis which pooled data from 12 studies reported that FA exposure correlated with increased risk of sinonasal cancers (189). Occupational exposure to FA was also correlated with increased risk of nasopharyngeal cancers (190). A meta-analysis of 26 studies investigating occupational FA exposures reported 15 that were consistent with increased risk of leukemias, in particular myeloid leukemias (191).

More recent analyses have also confirmed links between occupational FA exposures and cancers (192–195). A mainstream consensus has solidified in support of linking FA exposure to sinonasal cancers, nasopharyngeal, and myeloid leukemias, although this majority view has been challenged for the latter cancer types by some authors who have been funded by chemical industry trade groups. It is possible that the hematopoietic system is especially vulnerable, even to endogenous FA, because hematopoietic stem cells (HSCs) must replicate and divide

throughout an individual's lifetime. This would make HSCs more likely to acquire mutations than postmitotic cells. Another possibility is that HSCs have intrinsically lower DNA repair capacity because they exist in a low oxygen niche, so there is less need to defend against oxidative damage. This hypothesis could also help explain why HSCs are more sensitive to endogenous FA than other tissues.

2.12 Formaldehyde-induced DNA damage

2.12.1 Formaldehyde-induced DNA adducts

When FA reacts with DNA, the most common products are hydroxymethyl monoadducts. Such adducts can be detected by workflows using chromatography followed by mass spectrometry. The most abundant adduct is N^2-hydroxymethyldeoxyguanosine or N^2-HOMe-dG, which is observed in the majority of studies using *in vitro* and cellular systems (see **Table 2.8** and references therein). Next in abundance is N^6-HOMe-dA, i.e., N^6-hydroxymethyldeoxyadenosine, which is found in somewhat fewer studies. Finally, N^4-hydroxymethyldeoxycytidine (N^4-HOMe-dC) is a relatively minor adduct that was reported in a few papers. Whereas AA adducts are predominantly at N2 of guanine, FA adducts are more diverse. Since FA is produced endogenously on a continuing basis, steady-state concentrations of >10 μM are found in blood (180). It is therefore not surprising that adducts from endogenous FA can exceed those from exposure to exogenous FA (196). BER is thought to be a major pathway for repair of FA-induced monoadducts, which are relatively small, but NER may also be involved (197–200). As FA and AA are very similar, there is likely considerable similarity in utilization of repair pathways to process lesions induced by either small aldehyde in DNA.

Experimental systems	DNA adducts	References
Chinese hamster ovary cells	N^6-HOMe-dA	(201)
DNA	N^6-HOMe-dA, N^4-HOMe-dC, N^2-HOMe-dG	(202)
DNA	18 adducts	(203)
DNA, deoxyribonucleosides	N^6-HOMe-dA, N^4-HOMe-dC, N^2-HOMe-dG	(201)
HeLa human cervical cancer cell line	N^2-HOMe-dG, N^6-HOMe-dA	(204)
Human nasal epithelial cells	N^6-HOMe-dA, N^2-HOMe-dG, N^4-HOMe-dC	(205)
Human leukocytes	N^6-HOMe-dA	(206)
Macaque bone marrow and nasal mucosa cells	N^2-HOMe-dG	(207)
Rat bone marrow, brain, kidney, liver, lung, spleen, thymus, white blood cells	N^2-HOMe-dG, N^6-HOMe-dA	(208)
Rat bone marrow, brain, kidney, liver, lung, lymph nodes, spleen, thymus, trachea, white blood cells	N^2-HOMe-dG	(209)
Rat blood cells, bone marrow, cerebellum, hippocampus, liver, lung, nasal mucosa, olfactory bulb, trachea	N^2-HOMe-dG	(210)
Rat nasal mucosa	N^2-HOMe-dG, N^6-HOMe-dA	(196,211)
Rat nasal mucosa	N^2-HOMe-dG	(212)

Table 2.8: FA-induced DNA adducts found in different experimental systems.

Another possible indirect means for FA to cause DNA damage is via the generation of oxidative damage (213–216). Indeed, 8-hydroxyguanine or 8-oxoguanine is commonly formed under conditions of oxidative damage and known to cause G > T transversions (217), which are observed after FA exposure (see Section 2.12.4 below).

2.12.2 Formaldehyde-induced crosslinks

FA is well known as a crosslinking agent, leading to DNA damage and subsequent mutagenesis (see **Table 2.9** and references therein). Similar to detection of monoadducts, crosslinks can be identified using chromatography followed by mass spectrometry (33,218). DNA-protein crosslinks (DPCs) have received a great deal of attention, being investigated in a very large

number of studies across a broad range of experimental systems. DPC formation is typically initiated by the reaction of FA with a primary amine group (e.g., in lysine), leading to a Schiff base (-N=CH_2), which then can react with a nucleophile in DNA (219). Repair of DPCs can be carried out by multiple means, including: direct hydrolysis of the crosslink; replication-coupled repair; and nucleolytic or proteolytic mechanisms (99).

DNA-DNA crosslinks are also formed commonly by reaction with FA. Interstrand crosslinks form readily upon exposure to FA. A number of crosslinked dinucleotides have been identified by mass spectrometry, including dG-CH_2-dG, dG-CH_2-dA, dA-CH_2-hydroxymethyl-2'-dC, dA-CH_2-dA, dG-CH_2-dC, and dA-CH_2-dC (218). The Fanconi anemia pathway is the major mechanism for repair of interstrand crosslinks, including those induced by FA (97,197), while NER is thought to be a major repair pathway for intrastrand crosslinks (220).

Experimental systems	**Crosslinks**	**References**
A549 human lung cell line	DNA-protein crosslinks	(221–225)
Budding yeast	DNA-protein crosslinks	(226–228)
Budding yeast mitochondria	DNA-protein crosslinks	(229)
Chinese hamster ovary cells	DNA-protein crosslinks	(230,231,103,232,233)
Chinese hamster ovary cells	Interstrand crosslinks	(103)
DNA	Interstrand crosslinks	(234–236,218)
DNA with histones	DNA-protein crosslinks	(237–240)
DT40 chicken lymphoma cell line	DNA-protein crosslinks	(241)
E. coli	DNA-protein crosslinks	(242,243)
E. coli	Interstrand crosslinks	(242)
Fission yeast	DNA-protein and interstrand crosslinks	(199)
HEK293 human embryonic kidney cell line	DNA-protein crosslinks	(244)
HeLa human cervical cancer cell line	DNA-protein crosslinks	(236,244)
Human bronchial epithelial cells	DNA-protein crosslinks	(245–248)
Human fibroblasts	DNA-protein crosslinks	(249,247,240)
Human kidney cell line	DNA-protein crosslinks	(240)

Human leukocytes	DNA-protein crosslinks	(250)
Human lung cell line	DNA-protein crosslinks	(240)
Human lymphoblasts	DNA-protein crosslinks	(251,252)
Human lymphocytes	DNA-protein crosslinks	(236,253–258)
Human lymphosarcoma cells	DNA-protein crosslinks	(259,260)
Human nasal epithelial cells	DNA-protein crosslinks	(221)
Human normal and DNA repair deficient cell lines	DNA-protein crosslinks	(261,232)
Human peripheral blood lymphocytes	Interstrand crosslinks	(236)
Human white blood cells	DNA-protein crosslinks	(262)
Jurkat human T lymphocyte cell line	DNA-protein crosslinks	(263)
Monkey blood cells, bone marrow, liver, nasal mucosa	DNA-protein crosslinks	(33)
Monkey cells infected with simian virus 40	DNA-protein crosslinks	(239,264)
Mouse bone marrow, kidney, liver, testes	DNA-protein crosslinks	(213)
Mouse bone marrow mesenchymal stem cells	DNA-protein crosslinks	(265)
Mouse embryonic fibroblasts	DNA-protein crosslinks	(266)
Mouse hepatocytes	Interstrand crosslinks	(236)
Mouse leukemia cells	DNA-protein crosslinks	(267,268)
Rat aorta endothelial cells	DNA-protein crosslinks	(269)
Rat bladder transitional cells	DNA-protein crosslinks	(270)
Rat blood cells, bone marrow, cerebellum, hippocampus, liver, lung, nasal mucosa, olfactory bulb, trachea	DNA-protein crosslinks	(210)
Rat blood cells, bone marrow, nasal mucosa	DNA-protein crosslinks	(33)
Rat nasal mucosa cells	DNA-protein crosslinks	(271,272,115,273–276)
Rat nasal mucosa cells	Interstrand crosslinks	(196)
Rat sarcoma cells	DNA-protein crosslinks	(277)
Rat tracheal epithelial cells	DNA-protein crosslinks	(278–280)
Rhesus monkey respiratory tract epithelial cells	DNA-protein crosslinks	(281)
V79 Chinese hamster lung fibroblast cell line	DNA-protein crosslinks	(282–285)

Table 2.9: FA-induced crosslinks in different experimental systems.

Both DNA-protein and interstrand crosslinks are thought to be especially deleterious lesions, as unrepaired crosslinks would block important DNA processes, i.e., replication and transcription. As such, the pathways dedicated to repairing DNA-protein and interstrand crosslinks are vital for maintaining genomic stability and are studied extensively by many research groups around the world (recently reviewed in (99,286–288)).

2.12.3 Formaldehyde-induced strand breaks

FA readily induces DNA strand breaks in a range of experimental cell systems. SSBs are the most commonly investigated (see **Table 2.10** and references therein). DSBs are also reported, but in much fewer papers than for SSBs. DSBs are repaired by two main pathways, HR and NHEJ. Repair by HR can be error-free while NHEJ is generally error-prone (117). SSB repair is mediated by a separate pathway, where poly (ADP-ribose) polymerase (PARP) proteins play a key role in damage recognition and recruitment of repair factors (118).

Experimental systems	DNA strand breaks	References
A549 human lung cell line	Double-strand breaks	(289,290,119)
A549 human lung cell line	Single-strand breaks	(291)
Budding yeast	Single-strand breaks	(292)
C18 rat tracheal epithelial cell line	Single-strand breaks	(280)
Chinese hamster lung fibroblast cell lines	Double-strand breaks	(293,294)
Chinese hamster ovary cells	Double-strand breaks	(233)
Chinese hamster ovary cells	Single-strand breaks	(295,296)
HeLa human cervical cancer cell line	Single-strand breaks	(236)
Human bronchial epithelial cells	Single-strand breaks	(245–248,109,297)
Human fibroblasts	Single-strand breaks	(249,246,247,109,298)
Human fibroblast mitochondria	Double-strand breaks	(299)
Human hematopoietic stem cells	Single-strand breaks	(300)
Human lymphocytes	Single-strand breaks	(236,258)
Mouse hepatocytes	Single-strand breaks	(301)
Mouse bone marrow mesenchymal stem cells	Single-strand breaks	(265)

Mouse leukemia cells	Single-strand breaks	(267)
Rat hepatocytes	Single-strand breaks	(301)
Rat sarcoma cells	Single-strand breaks	(277,259)
Rat tracheal epithelial cells	Single-strand breaks	(279)
RPE-1 human retinal pigment epithelial cell lines	Double-strand breaks	(302)

Table 2.10: FA-induced DNA strand breaks in different experimental systems.

2.12.4 Formaldehyde-induced mutations

Incorrect repair or error-prone bypass of the above-mentioned forms of damage can result in a myriad of DNA sequence changes (see **Table 2.11** and references therein). Mutagenesis by FA has been studied in many model organisms and cell systems. The most frequently reported sequence changes involve relatively large-scale deletions, insertions, and other rearrangements. This is consistent with FA's activity as a crosslinking agent. Point mutations have also been studied in multiple systems, yielding a consensus of mostly G > T transversions being induced by FA, indicating that adducted G's are mutagenic. The one exception to this consensus view was a study in Chinese hamster ovary cells, which reported mostly substitutions at A:T base pairs (303). Nonetheless, the overall mutagenic properties of FA and AA are quite similar, which is not surprising, given the chemical similarity of these simple aldehydes.

Experimental systems	**Mutation types**	**References**
Budding yeast	Predominantly G > T	(5)
Budding yeast frameshift reporter	Mostly insertions, some deletions, duplications	(304)
C. elegans	Unspecified point mutations, large deletions	(305–308)
Chinese hamster ovary cells	Point mutations, predominantly at A:T base pairs	(303)
E. coli	Large deletions, large insertions, point mutations	(309)
E. coli	Predominantly G > T	(310,311)
E. coli	Dinucleotide repeat instability	(312)

Fruit flies	Chromosomal rearrangements	(313)
Fruit flies	Deletions	(314–318)
Grasshopper neuroblasts	Chromosomal rearrangements	(319)
Hamster embryo cells	Chromosomal rearrangements	(320)
HeLa human cervical cancer cell line	Chromosomal rearrangements	(321)
Human fibroblasts	Chromosomal rearrangements	(298,322)
Human lymphoblastoid cell lines	Chromosomal rearrangements	(252)
Human lymphocytes	Chromosomal rearrangements	(323–327)
Human myeloid progenitor cells	Chromosomal rearrangements	(328)
L5178Y mouse lymphoma cell line	Chromosomal rearrangements	(329)
Mouse sperm cells	Simple repeat instability	(330)
Neurospora crassa	Unspecified point mutations, deletions	(331)
Rat nasal squamous cell carcinomas	Small number of point mutations at TP53, mostly G > T or C > A	(332)
Salmonella typhimurium	G > T	(311)
TK6 human lymphoblastoid cell line	Large deletions	(309)

Table 2.11: FA-induced mutations in different experimental systems.

2.13 Formaldehyde-related carcinogenesis in model species

The carcinogenicity of FA has been tested on rodents (mostly rats). FA was administered by inhalation in the majority of studies, with a smaller number using administration via drinking water (see **Table 2.12** and references therein). FA by inhalation induced cancers of the nasal mucosa, including in both hamsters and rats, entirely consistent with epidemiological results in humans. FA in drinking water induced cancers of the gastrointestinal tract, as well as at distal sites, i.e., the lymphohematopoietic system and testes (145). But it is interesting to note that none of the FA by inhalation studies reported induced cancers at distal sites, so the correlation in humans between occupational FA exposure (via inhalation) and myeloid leukemias has not been recapitulated in animal models, at least thus far.

Species (sex)	Exposures	Cancer sites	References
Hamsters (male)	FA by inhalation	Nasal mucosa	(333)
Rats (both)	FA by inhalation	Nasal mucosa	(334–336)
Rats (both)	FA in drinking water	Small intestine	(337,145)
Rats (both)	FA in drinking water	Stomach	(145)
Rats (both)	FA in drinking water	Lymphohematopoietic system	(145)
Rats (male)	FA by inhalation	Nasal mucosa	(338–340,332,341,342)
Rats (male)	FA in drinking water	Stomach	(343)
Rats (male)	FA in drinking water	Testes	(145)

Table 2.12: FA-related carcinogenesis in different model species

2.14 Mutational signatures of acetaldehyde and formaldehyde

As we have mentioned, studies in multiple experimental systems have revealed certain common features of AA- and FA-induced mutagenesis (see Tables 2.5 and 2.11, and references therein). Indeed, it has been noted that the tandem GG > TT dinucleotide substitution pattern suggests AA as a plausible mutagen responsible for double base substitution (DBS) signature 2 (344,345) from the Catalog of Somatic Mutations in Cancer (COSMIC) (346).

Multiple studies have reported alcohol-linked signatures similar to SBS (single base substitution) signature 16 from COSMIC version 2 (347–351). It has been hypothesized that these signatures might be attributable to AA, but only correlations were demonstrated, without experimental follow-up to investigate molecular mechanisms. Additionally, SBS16 in COSMIC version 3 is quite different from version 2 of this signature, so it is not clear to what extent the conclusions from these papers still hold.

Nonetheless, with the proliferation of high throughput sequencing technologies, large mutational data sets from experimental studies are increasingly prevalent. It should therefore be possible to infer the patterns of induced mutagenesis more fully by carefully controlled

experiments. For example, one study investigating the patterns of mutations induced by many chemical agents in induced pluripotent stem cells included data on AA and FA (345). But because these small aldehydes are relatively weak mutagens, their induced mutational patterns could not be resolved reliably. Similarly, another study in mice knocked out for two key enzymes (ADH5 and ALDH2) for aldehyde detoxification (leading to accumulation of endogenous aldehydes). These double knockout mice did exhibit bone marrow failure, susceptibility to developing leukemia, and shortened lifespan. But this study did not to produce a robust mutational signature significantly different from background mutagenesis in this system (22).

As a workaround for the weak mutagenicity of small aldehydes, a high-sensitivity mutagenesis detection system featuring ssDNA enrichment in budding yeast has been utilized. A recent study using this system showed that AA induced an excess of gCn > gAn (or nGc > nTc) substitutions, consistent with previous literature (136). gCn > gAn mutations were also enriched in cancers associated with alcohol or tobacco use, consistent with possible exposure to AA (136).

We also recently deployed the same system to investigate both AA and FA (5). Both small aldehydes induced an excess of C/G > A/T transversions, and overall similar mutational patterns (5). Interestingly, the pattern of mutations from FA acting on ssDNA was the closest match to date to SBS signature 40. SBS40 is currently of unknown etiology, but it is the third most common mutational signature, present in at least 28 cancer types (346). Since FA is produced endogenously and present in relatively high intracellular concentrations, we propose that endogenous FA is a plausible candidate etiology for SBS40.

These results in an ssDNA-enriched system are also consistent with observations obtained in studies cited throughout this article using experimental systems with cycling cells that replicate their DNA. Replication is a vulnerable window for accruing DNA damage, as long stretches of ssDNA can be exposed, making mutagenesis more likely. More complex scenarios are also plausible. For example, aldehydes can potentially damage dNTPs, which can be misincorporated during replication. This can lead to mutation fixation, or if handled incorrectly during attempted repair, can lead to strand breaks during replication. Such DNA configurations would need to be rescued by replication fork restart mechanisms that can lead to large scale genomic changes, i.e., chromosomal rearrangements, if repair is not carried out correctly.

2.15 Conclusions

Investigations into the genotoxicity and mutagenicity of FA and AA have a long history, by now spanning over five decades. Work by many labs on a wide range of experimental systems has yielded broad consensus results, e.g., on carcinogenicity, mutagenicity, clastogenicity, the predominance of G > T substitutions, and so forth. But the mutational signatures of FA and AA remain to be further ascertained in more experimental systems. Moreover, the true extent of FA and AA's mutagenicity in cancers (and possibly in normal cells) is not well understood. Further investigations will be necessary to answer these important questions.

2.16 Acknowledgments

This work was supported by Natural Sciences and Engineering Research Council of Canada Grant 05973/RGPIN/2017 and Ontario Early Researcher Award ER17-13-013 to KC.

2.17 References

1. Frontiers | Genotoxicity of formaldehyde: molecular basis of DNA damage and mutation [Internet]. [cited 2025 Jun 24]. Available from: https://www.frontiersin.org/journals/environmental-science/articles/10.3389/fenvs.2014.00036/full

2. Kayani MA, Parry JM. The *in vitro* genotoxicity of ethanol and acetaldehyde. Toxicology in Vitro [Internet]. 2010 Feb 1 [cited 2025 Jun 24];24(1):56–60. Available from: https://www.sciencedirect.com/science/article/pii/S0887233309002598

3. Blasiak J, Trzeciak A, Malecka-Panas E, Drzewoski J, Wojewódzka M. *In vitro* genotoxicity of ethanol and acetaldehyde in human lymphocytes and the gastrointestinal tract mucosa cells. Toxicology in Vitro [Internet]. 2000 Aug 1 [cited 2025 Jun 24];14(4):287–95. Available from: https://www.sciencedirect.com/science/article/pii/S0887233300000229

4. Combined effects of co-exposure to formaldehyde and acrolein mixtures on cytotoxicity and genotoxicity in vitro | Environmental Science and Pollution Research [Internet]. [cited 2025 Jun 24]. Available from: https://link.springer.com/article/10.1007/s11356-018-2584-z

5. Thapa MJ, Fabros RM, Alasmar S, Chan K. Analyses of mutational patterns induced by formaldehyde and acetaldehyde reveal similarity to a common mutational signature. G3 Genes|Genomes|Genetics. 2022 Sep 8;12:jkac238.

6. Vijayraghavan S, Saini N. Aldehyde-Associated Mutagenesis–Current State of Knowledge. Chem Res Toxicol [Internet]. 2023 Jul 17 [cited 2025 Jun 24];36(7):983–1001. Available from: https://doi.org/10.1021/acs.chemrestox.3c00045

7. Sinharoy P, McAllister SL, Vasu M, Gross ER. Environmental Aldehyde Sources and the Health Implications of Exposure. Advances in experimental medicine and biology. 2019/08/02 ed. 2019;1193:35–52.

8. International Agency for Research on Cancer. Re-evaluation of some organic chemicals, hydrazine and hydrogen peroxide. IARC monographs on the evaluation of carcinogenic risks to humans. 1999/10/03 ed. 1999;71 Pt 1, Pt 2, Pt 3(Pt 1):1–1554.

9. Salthammer T, Mentese S, Marutzky R. Formaldehyde in the Indoor Environment. Chemical Reviews. 2010 Apr 14;110(4):2536–72.

10. Burgos-Barragan G, Wit N, Meiser J, Dingler FA, Pietzke M, Mulderrig L, et al. Mammals divert endogenous genotoxic formaldehyde into one-carbon metabolism. Nature. 2017 Aug 16;548:549.

11. Kooistra SM, Helin K. Molecular mechanisms and potential functions of histone demethylases. Nature Reviews Molecular Cell Biology. 2012 May 1;13(5):297–311.

12. Fedeles BI, Singh V, Delaney JC, Li D, Essigmann JM. The AlkB Family of Fe(II)/α-Ketoglutarate-dependent Dioxygenases: Repairing Nucleic Acid Alkylation Damage and Beyond. J Biol Chem. 2015 Aug 21;290(34):20734–42.

13. International Agency for Research on Cancer. Chemical agents and related occupations. IARC monographs on the evaluation of carcinogenic risks to humans. 2012/11/30 ed. 2012;100(Pt F):9–562.

14. Heck HD, Casanova-Schmitz M, Dodd PB, Schachter EN, Witek TJ, Tosun T. Formaldehyde (CH2O) concentrations in the blood of humans and Fischer-344 rats exposed to CH2O under controlled conditions. American Industrial Hygiene Association journal. 1985/01/01 ed. 1985 Jan;46(1):1–3.

15. Szarvas T, Szatlóczky E, Volford J, Trézl L, Tyihák E, Rusznák I. Determination of endogenous formaldehyde level in human blood and urine by dimedone-14C radiometric method. Journal of Radioanalytical and Nuclear Chemistry. 1986 Oct 1;106(6):357–67.

16. Luo W, Li H, Zhang Y, Ang CYW. Determination of formaldehyde in blood plasma by high-performance liquid chromatography with fluorescence detection. Journal of Chromatography B: Biomedical Sciences and Applications. 2001 Apr 5;753(2):253–7.

17. Tong Z, Han C, Luo W, Wang X, Li H, Luo H, et al. Accumulated hippocampal formaldehyde induces age-dependent memory decline. Age (Dordr). 2013 Jun;35(3):583–96.

18. Tan T, Zhang Y, Luo W, Lv J, Han C, Hamlin JNR, et al. Formaldehyde induces diabetes-associated cognitive impairments. FASEB J. 2018 Jul;32(7):3669–79.

19. Kumaravel S, Wu SH, Chen GZ, Huang ST, Lin CM, Lee YC, et al. Development of ratiometric electrochemical molecular switches to assay endogenous formaldehyde in live cells, whole blood and creatinine in saliva. Biosens Bioelectron. 2021 Jan 1;171:112720.

20. Wei Y, Wang M, Liu H, Niu Y, Wang S, Zhang F, et al. Simultaneous determination of seven endogenous aldehydes in human blood by headspace gas chromatography–mass spectrometry. Journal of Chromatography B. 2019 Jun 15;1118–1119:85–92.

21. Shindyapina AV, Komarova TV, Sheshukova EV, Ershova NM, Tashlitsky VN, Kurkin AV, et al. The Antioxidant Cofactor Alpha-Lipoic Acid May Control Endogenous Formaldehyde Metabolism in Mammals. Front Neurosci. 2017;11:651.

22. Dingler FA, Wang M, Mu A, Millington CL, Oberbeck N, Watcham S, et al. Two Aldehyde Clearance Systems Are Essential to Prevent Lethal Formaldehyde Accumulation in Mice and Humans. Molecular Cell. 2020 décembre;80(6):996-1012.e9.

23. Maejima K, Suzuki T, Numata H, Maekawa A, Nagase S, Ishinishi N. Recovery from changes in the blood and nasal cavity and/or lungs of rats caused by exposure to methanol-fueled engine exhaust. J Toxicol Environ Health. 1993 Jul;39(3):323–40.

24. Maejima K, Suzuki T, Numata H, Maekawa A, Nagase S, Ishinishi N. Subchronic (12-week) inhalation toxicity study of methanol-fueled engine exhaust in rats. J Toxicol Environ Health. 1994 Mar;41(3):315–27.

25. Kleinnijenhuis AJ, Staal YCM, Duistermaat E, Engel R, Woutersen RA. The determination of exogenous formaldehyde in blood of rats during and after inhalation exposure. Food and Chemical Toxicology. 2013 Feb 1;52:105–12.

26. Tong Z, Han C, Qiang M, Wang W, Lv J, Zhang S, et al. Age-related formaldehyde interferes with DNA methyltransferase function, causing memory loss in Alzheimer's disease. Neurobiol Aging. 2015 Jan;36(1):100–10.

27. Casanova M, d'A. Heck H, Everitt JI, Harrington WW, Popp JA. Formaldehyde concentrations in the blood of rhesus monkeys after inhalation exposure. Food and Chemical Toxicology. 1988 Jan 1;26(8):715–6.

28. Kim YH, Park J. Development of a Simple and Powerful Analytical Method for Formaldehyde Detection and Quantitation in Blood Samples. Journal of Analytical Methods in Chemistry. 2020 Jan 1;2020(1):8810726.

29. Seitz HK, Stickel F. Molecular mechanisms of alcohol-mediated carcinogenesis. Nature Reviews Cancer. 2007 Aug 1;7(8):599–612.

30. International Agency for Research on Cancer. Personal habits and indoor combustions. IARC monographs on the evaluation of carcinogenic risks to humans. 2012/12/01 ed. 2012;100(Pt E):1–538.

31. Medeiros MHG. DNA damage by endogenous and exogenous aldehydes. Journal of the Brazilian Chemical Society. 2019;30:2000–9.

32. Chatterjee N, Walker GC. Mechanisms of DNA damage, repair, and mutagenesis. Environmental and Molecular Mutagenesis. 2017 May 9;58(5):235–63.

33. Lai Y, Yu R, Hartwell HJ, Moeller BC, Bodnar WM, Swenberg JA. Measurement of Endogenous versus Exogenous Formaldehyde–Induced DNA–Protein Crosslinks in Animal Tissues by Stable Isotope Labeling and Ultrasensitive Mass Spectrometry. Cancer Res. 2016 May 1;76(9):2652.

34. Weng M wen, Lee HW, Park SH, Hu Y, Wang HT, Chen LC, et al. Aldehydes are the predominant forces inducing DNA damage and inhibiting DNA repair in tobacco smoke carcinogenesis. Proc Natl Acad Sci USA. 2018 Jul 3;115(27):E6152.

35. National Toxicology Program. Acetaldehyde. In: Report on Carcinogens, Fifteenth Edition. Research Triangle Park, NC: U.S. Department of Health and Human Services, Public Health Service; 2021.

36. Crabb DW, Liangpunsakul S. Acetaldehyde generating enzyme systems: roles of alcohol dehydrogenase, CYP2E1 and catalase, and speculations on the role of other enzymes and

processes. Novartis Foundation symposium. 2007/06/27 ed. 2007;285:4–16; discussion 16-22, 198–9.

37. Kenechukwu CO. Ethanol. In: Simon George T, editor. Psychology of Health [Internet]. Rijeka: IntechOpen; 2019 [cited 2022 Oct 17]. p. Ch. 4. Available from: https://doi.org/10.5772/intechopen.79861

38. Crabb DW, Matsumoto M, Chang D, You M. Overview of the role of alcohol dehydrogenase and aldehyde dehydrogenase and their variants in the genesis of alcohol-related pathology. Proceedings of the Nutrition Society. 2007/03/07 ed. 2004;63(1):49–63.

39. Moritani K, Takeshita T, Shibata Y, Ninomiya T, Kiyohara Y, Yamashita Y. Acetaldehyde production by major oral microbes. Oral Diseases. 2015 Sep 1;21(6):748–54.

40. Rumgay H, Shield K, Charvat H, Ferrari P, Sornpaisarn B, Obot I, et al. Global burden of cancer in 2020 attributable to alcohol consumption: a population-based study. The Lancet Oncology. 2021;22(8):1071–80.

41. Stornetta A, Guidolin V, Balbo S. Alcohol-Derived Acetaldehyde Exposure in the Oral Cavity. Cancers. 2018;10(1).

42. Anantharaman D, Marron M, Lagiou P, Samoli E, Ahrens W, Pohlabeln H, et al. Population attributable risk of tobacco and alcohol for upper aerodigestive tract cancer. Oral Oncology. 2011 Aug 1;47(8):725–31.

43. Väkeväinen S, Tillonen J, Agarwal DP, Srivastava N, Salaspuro M. High Salivary Acetaldehyde After a Moderate Dose of Alcohol in ALDH2-Deficient Subjects: Strong Evidence for the Local Carcinogenic Action of Acetaldehyde. Alcoholism: Clinical and Experimental Research. 2000 Jun 1;24(6):873–7.

44. Chen YC, Peng GS, Tsao TP, Wang MF, Lu RB, Yin SJ. Pharmacokinetic and pharmacodynamic basis for overcoming acetaldehyde-induced adverse reaction in Asian alcoholics, heterozygous for the variant ALDH2*2 gene allele. Pharmacogenet Genomics. 2009 Aug;19(8):588–99.

45. Chang JS, Hsiao JR, Chen CH. ALDH2 polymorphism and alcohol-related cancers in Asians: a public health perspective. Journal of Biomedical Science. 2017 Mar 3;24(1):19.

46. Bagnardi V, Rota M, Botteri E, Tramacere I, Islami F, Fedirko V, et al. Light alcohol drinking and cancer: a meta-analysis. Annals of oncology : official journal of the European Society for Medical Oncology. 2012/08/23 ed. 2013 Feb;24(2):301–8.

47. Nieminen MT, Salaspuro M. Local Acetaldehyde—An Essential Role in Alcohol-Related Upper Gastrointestinal Tract Carcinogenesis. Cancers. 2018;10(1).

48. Salaspuro M. Key role of local acetaldehyde in upper GI tract carcinogenesis. Best Practice & Research Clinical Gastroenterology. 2017 Oct 1;31(5):491–9.

49. Salaspuro M. Local Acetaldehyde: Its Key Role in Alcohol-Related Oropharyngeal Cancer. Visceral Medicine. 2020;36(3):167–74.

50. Tramacere I, Negri E, Bagnardi V, Garavello W, Rota M, Scotti L, et al. A meta-analysis of alcohol drinking and oral and pharyngeal cancers. Part 1: overall results and dose-risk relation. Oral oncology. 2010/05/07 ed. 2010 Jul;46(7):497–503.

51. Koyanagi YN, Ito H, Oze I, Hosono S, Tanaka H, Abe T, et al. Development of a prediction model and estimation of cumulative risk for upper aerodigestive tract cancer on the basis of the aldehyde dehydrogenase 2 genotype and alcohol consumption in a Japanese population. European journal of cancer prevention : the official journal of the European Cancer Prevention Organisation (ECP). 2016/02/11 ed. 2017 Jan;26(1):38–47.

52. Lachenmeier DW, Salaspuro M. ALDH2-deficiency as genetic epidemiologic and biochemical model for the carcinogenicity of acetaldehyde. Regulatory toxicology and pharmacology : RTP. 2017/03/05 ed. 2017 Jun;86:128–36.

53. Matsuda T, Yabushita H, Kanaly RA, Shibutani S, Yokoyama A. Increased DNA Damage in ALDH2-Deficient Alcoholics. Chem Res Toxicol. 2006 Oct 1;19(10):1374–8.

54. Yokoyama A, Kamada Y, Imazeki H, Hayashi E, Murata S, Kinoshita K, et al. Effects of ADH1B and ALDH2 Genetic Polymorphisms on Alcohol Elimination Rates and Salivary Acetaldehyde Levels in Intoxicated Japanese Alcoholic Men. Alcoholism: Clinical and Experimental Research. 2016 Jun 1;40(6):1241–50.

55. Cai Q, Wu J, Cai Q, Chen EZ, Jiang ZY. Association between Glu504Lys Polymorphism of ALDH2 Gene and Cancer Risk: A Meta-Analysis. PLOS ONE. 2015 Feb 13;10(2):e0117173.

56. Zuo W, Zhan Z, Ma L, Bai W, Zeng S. Effect of ALDH2 polymorphism on cancer risk in Asians: A meta-analysis. Medicine (Baltimore). 2019 Mar;98(13):e14855.

57. Kang SJ, Shin CM, Sung J, Kim N. Association Between ALDH2 Polymorphism and Gastric Cancer Risk in Terms of Alcohol Consumption: A Meta-Analysis. Alcohol: Clinical and Experimental Research. 2021 Jan 1;45(1):6–14.

58. Cao HX, Li SP, Wu JZ, Gao CM, Su P, Liu YT, et al. Alcohol dehydrogenase-2 and aldehyde dehydrogenase-2 genotypes, alcohol drinking and the risk for stomach cancer in Chinese males. Asian Pac J Cancer Prev. 2010;11(4):1073–7.

59. Im PK, Yang L, Kartsonaki C, Chen Y, Guo Y, Du H, et al. Alcohol metabolism genes and risks of site-specific cancers in Chinese adults: An 11-year prospective study. International Journal of Cancer. 2022 May 15;150(10):1627–39.

60. Mori T, Okamoto Y, Mu A, Ide Y, Yoshimura A, Senda N, et al. Lack of impact of the ALDH2 rs671 variant on breast cancer development in Japanese BRCA1/2-mutation carriers. Cancer Med. 2023 Mar;12(6):6594–602.

61. Rumgay H, Murphy N, Ferrari P, Soerjomataram I. Alcohol and Cancer: Epidemiology and Biological Mechanisms. Nutrients. 2021;13(9).

62. Wang M, McIntee EJ, Cheng G, Shi Y, Villalta PW, Hecht SS. Identification of DNA Adducts of Acetaldehyde. Chem Res Toxicol. 2000 Nov 1;13(11):1149–57.

63. Wang M, Yu N, Chen L, Villalta PW, Hochalter JB, Hecht SS. Identification of an Acetaldehyde Adduct in Human Liver DNA and Quantitation as N2-Ethyldeoxyguanosine. Chem Res Toxicol. 2006 Feb 1;19(2):319–24.

64. Singh R, Sandhu J, Kaur B, Juren T, Steward WP, Segerbäck D, et al. Evaluation of the DNA damaging potential of cannabis cigarette smoke by the determination of acetaldehyde derived N2-ethyl-2'-deoxyguanosine adducts. Chem Res Toxicol. 2009 Jun;22(6):1181–8.

65. Inagaki S, Esaka Y, Deyashiki Y, Sako M, Goto M. Analysis of DNA adducts of acetaldehyde by liquid chromatography–mass spectrometry. Journal of Chromatography A. 2003 Feb 14;987(1):341–7.

66. Murakami H, Horiba R, Iwata T, Miki Y, Uno B, Sakai T, et al. Progress in a selective method for the determination of the acetaldehyde-derived DNA adducts by using HILIC-ESI-MS/MS. Talanta. 2018 Jan 15;177:12–7.

67. Leung EMK, Deng K, Wong TY, Chan W. Determination of DNA adducts by combining acid-catalyzed hydrolysis and chromatographic analysis of the carcinogen-modified nucleobases. Analytical and Bioanalytical Chemistry. 2016 Jan 1;408(3):953–61.

68. Sako M, Inagaki S, Esaka Y, Deyashiki Y. Histones accelerate the cyclic 1,N2-propanoguanine adduct-formation of DNA by the primary metabolite of alcohol and carcinogenic crotonaldehyde. Bioorg Med Chem Lett. 2003 Oct 20;13(20):3497–8.

69. Inagaki S, Esaka Y, Goto M, Deyashiki Y, Sako M. LC-MS study on the formation of cyclic 1,N2-propano guanine adduct in the reactions of DNA with acetaldehyde in the presence of histone. Biol Pharm Bull. 2004 Mar;27(3):273–6.

70. Singh R, Gromadzinska J, Mistry Y, Cordell R, Juren T, Segerbäck D, et al. Detection of acetaldehyde derived N2-ethyl-2'-deoxyguanosine in human leukocyte DNA following alcohol consumption. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis. 2012 Sep 1;737(1):8–11.

71. Balbo S, Meng L, Bliss RL, Jensen JA, Hatsukami DK, Hecht SS. Kinetics of DNA Adduct Formation in the Oral Cavity after Drinking Alcohol. Cancer Epidemiology, Biomarkers & Prevention. 2012 Mar 28;21(4):601–8.

72. Guidolin V, Carlson ES, Carrà A, Villalta PW, Maertens LA, Hecht SS, et al. Identification of New Markers of Alcohol-Derived DNA Damage in Humans. Biomolecules. 2021;11(3).

73. Zhang S, Villalta PW, Wang M, Hecht SS. Analysis of Crotonaldehyde- and Acetaldehyde-Derived 1,N2-Propanodeoxyguanosine Adducts in DNA from Human Tissues Using Liquid Chromatography Electrospray Ionization Tandem Mass Spectrometry. Chem Res Toxicol. 2006 Oct 1;19(10):1386–92.

74. Chen L, Wang M, Villalta PW, Luo X, Feuer R, Jensen J, et al. Quantitation of an Acetaldehyde Adduct in Human Leukocyte DNA and the Effect of Smoking Cessation. Chem Res Toxicol. 2007 Jan 1;20(1):108–13.

75. Balbo S, Hashibe M, Gundy S, Brennan P, Canova C, Simonato L, et al. N2-Ethyldeoxyguanosine as a Potential Biomarker for Assessing Effects of Alcohol Consumption on DNA. Cancer Epidemiology, Biomarkers & Prevention. 2008 Nov 6;17(11):3026–32.

76. Balbo S, Meng L, Bliss RL, Jensen JA, Hatsukami DK, Hecht SS. Time course of DNA adduct formation in peripheral blood granulocytes and lymphocytes after drinking alcohol. Mutagenesis. 2012 Jul 1;27(4):485–90.

77. Yukawa Y, Muto M, Hori K, Nagayoshi H, Yokoyama A, Chiba T, et al. Combination of ADH1B*2/ALDH2*2 polymorphisms alters acetaldehyde-derived DNA damage in the blood of Japanese alcoholics. Cancer Science. 2012 Sep 1;103(9):1651–5.

78. Garcia CCM, Angeli JPF, Freitas FP, Gomes OF, de Oliveira TF, Loureiro APM, et al. [13C2]-Acetaldehyde Promotes Unequivocal Formation of 1,N2-Propano-2'-deoxyguanosine in Human Cells. J Am Chem Soc. 2011 Jun 22;133(24):9140–3.

79. Hirohashi K, Ohashi S, Amanuma Y, Nakai Y, Ida T, Baba K, et al. Protective effects of Alda-1, an ALDH2 activator, on alcohol-derived DNA damage in the esophagus of human ALDH2*2 (Glu504Lys) knock-in mice. Carcinogenesis. 2020 Apr 22;41(2):194–202.

80. Matsuda T, Matsumoto A, Uchida M, Kanaly RA, Misaki K, Shibutani S, et al. Increased formation of hepatic N2-ethylidene-2'-deoxyguanosine DNA adducts in aldehyde dehydrogenase 2 -knockout mice treated with ethanol. Carcinogenesis. 2007;28(11):2363–6.

81. Nagayoshi H, Matsumoto A, Nishi R, Kawamoto T, Ichiba M, Matsuda T. Increased formation of gastric N(2)-ethylidene-2’-deoxyguanosine DNA adducts in aldehyde dehydrogenase-2 knockout mice treated with ethanol. Mutation Research/Genetic Toxicology and Environmental Mutagenesis. 2009 Feb 19;673(1):74–7.

82. Oyama T, Nagayoshi H, Matsuda T, Oka M, Isse T, Yu HS, et al. Effects of acetaldehyde inhalation in mitochondrial aldehyde dehydrogenase deficient mice (Aldh2-/-). Front Biosci (Elite Ed). 2010 Jun 1;2(4):1344–54.

83. Yu HS, Oyama T, Matsuda T, Isse T, Yamaguchi T, Tanaka M, et al. The effect of ethanol on the formation of N2-ethylidene-dG adducts in mice: implications for alcohol-related carcinogenicity of the oral cavity and esophagus. Biomarkers. 2012 May 1;17(3):269–74.

84. Yukawa Y, Ohashi S, Amanuma Y, Nakai Y, Tsurumaki M, Kikuchi O, et al. Impairment of aldehyde dehydrogenase 2 increases accumulation of acetaldehyde-derived DNA damage in the esophagus after ethanol ingestion. Am J Cancer Res. 2014;4(3):279–84.

85. Sanchez AB, Garcia CCM, Freitas FP, Batista GL, Lopes FS, Carvalho VH, et al. DNA Adduct Formation in the Lungs and Brain of Rats Exposed to Low Concentrations of [13C2]-Acetaldehyde. Chem Res Toxicol. 2018 May 21;31(5):332–9.

86. Garcia CCM, Batista GL, Freitas FP, Lopes FS, Sanchez AB, Gutz IGR, et al. P59 - Quantification of DNA adducts in lungs, liver and brain of rats exposed to acetaldehyde. Free Radical Biology and Medicine. 2014 Oct 1;75:S41.

87. Liu CW, Hsiao YC, Hoffman G, Lu K. LC–MS/MS Analysis of the Formation and Loss of DNA Adducts in Rats Exposed to Vinyl Acetate Monomer through Inhalation. Chem Res Toxicol. 2021 Mar 15;34(3):793–803.

88. Hsiao YC, Liu CW, Hoffman G, Fang C, Lu K. Molecular Dosimetry of DNA Adducts in Rats Exposed to Vinyl Acetate Monomer. Toxicological Sciences. 2022 Feb 1;185(2):197–207.

89. Balbo S, Juanes RC, Khariwala S, Baker EJ, Daunais JB, Grant KA. Increased levels of the acetaldehyde-derived DNA adduct N2-ethyldeoxyguanosine in oral mucosa DNA from Rhesus monkeys exposed to alcohol. Mutagenesis. 2016 Sep 1;31(5):553–8.

90. Moeller BC, Recio L, Green A, Sun W, Wright FA, Bodnar WM, et al. Biomarkers of exposure and effect in human lymphoblastoid TK6 cells following [13C2]-acetaldehyde exposure. Toxicol Sci. 2013 May;133(1):1–12.

91. Noguchi C, Grothusen G, Anandarajan V, Martínez-Lage García M, Terlecky D, Corzo K, et al. Genetic controls of DNA damage avoidance in response to acetaldehyde in fission yeast. Cell Cycle. 2017 Jan 2;16(1):45–58.

92. Gohil D, Sarker AH, Roy R. Base Excision Repair: Mechanisms and Impact in Biology, Disease, and Medicine. Int J Mol Sci. 2023 Sep 16;24(18).

93. Zhang X, Yin M, Hu J. Nucleotide excision repair: a versatile and smart toolkit. Acta Biochim Biophys Sin (Shanghai). 2022 May 25;54(6):807–19.

94. Porcher L, Vijayraghavan S, McCollum J, Mieczkowski PA, Saini N. Multiple DNA repair pathways prevent acetaldehyde-induced mutagenesis in yeast. United States; 2024.

95. Yamazaki K, Iguchi T, Kanoh Y, Takayasu K, Ngo TTT, Onuki A, et al. Homologous recombination contributes to the repair of acetaldehyde-induced DNA damage. Cell Cycle. 2024 Feb;23(4):369–84.

96. Lichtman MA, Spivak JL, Boxer LA, Shattil SJ, Henderson ES, editors. - Commentary on and reprint of Fanconi G, Familiäre infantile perniziosaartige Anämie (perniziöses Blutbild und Konstitution) [Familial infantile pernicious-like anemia (pernicious blood picture and constitution)], in Jahrbuch für Kinderheilkunde (1927) 117:257–280. In: Hematology [Internet]. San Diego: Academic Press; 2000. p. 127–66. Available from: https://www.sciencedirect.com/science/article/pii/B9780124485105501060

97. Semlow DR, Walter JC. Mechanisms of Vertebrate DNA Interstrand Cross-Link Repair. Annu Rev Biochem. 2021 Jun 20;90:107–35.

98. Hodskinson MR, Bolner A, Sato K, Kamimae-Lanning AN, Rooijers K, Witte M, et al. Alcohol-derived DNA crosslinks are repaired by two distinct mechanisms. Nature. 2020 Mar 1;579(7800):603–8.

99. Weickert P, Stingele J. DNA–Protein Crosslinks and Their Resolution. Annu Rev Biochem. 2022 Jun 21;91(1):157–81.

100. O'Donovan A, Davies AA, Moggs JG, West SC, Wood RD. XPG endonuclease makes the 3' incision in human DNA nucleotide excision repair. Nature. 1994 Sep 1;371(6496):432–5.

101. Ghosh S, Sur S, Yerram SR, Rago C, Bhunia AK, Hossain MZ, et al. Hypersensitivities for Acetaldehyde and Other Agents among Cancer Cells Null for Clinically Relevant Fanconi Anemia Genes. The American Journal of Pathology. 2014 Jan 1;184(1):260–70.

102. Olin KL, Cherr GN, Rifkin E, Keen CL. The effects of some redox-active metals and reactive aldehydes on DNA-protein cross-links in vitro. Toxicology. 1996 Jun 17;110(1–3):1–8.

103. Lorenti Garcia C, Mechilli M, Proietti De Santis L, Schinoppi A, Kobos K, Palitti F. Relationship between DNA lesions, DNA repair and chromosomal damage induced by acetaldehyde. Mutat Res. 2009 Mar 9;662(1–2):3–9.

104. Ristow H, Obe G. Acetaldehyde induces cross-links in DNA and causes sister-chromatid exchanges in human cells. Mutation Research/Genetic Toxicology. 1978 Sep 1;58(1):115–9.

105. Sonohara Y, Yamamoto J, Tohashi K, Takatsuka R, Matsuda T, Iwai S, et al. Acetaldehyde forms covalent GG intrastrand crosslinks in DNA. Scientific Reports. 2019 Jan 24;9(1):660.

106. Tsuruta H, Sonohara Y, Tohashi K, Aoki Shioi N, Iwai S, Kuraoka I. Effects of acetaldehyde-induced DNA lesions on DNA metabolism. Genes and Environment. 2020 Jan 6;42(1):2.

107. Kuykendall JR, Bogdanffy MS. Efficiency of DNA-histone crosslinking induced by saturated and unsaturated aldehydes in vitro. Mutat Res. 1992 Oct;283(2):131–6.

108. Kuykendall JR, Bogdanffy MS. Reaction kinetics of DNA-histone crosslinking by vinyl acetate and acetaldehyde. Carcinogenesis. 1992 Nov;13(11):2095–100.

109. Grafström RC, Sundqvist K, Dypbukt JM, Harris CC. Pathobiological effects of aldehydes in cultured human bronchial cells. IARC Sci Publ. 1987;(84):443–5.

110. Grafström RC, Dypbukt JM, Sundqvist K, Atzori L, Nielsen I, Curren RD, et al. Pathobiological effects of acetaldehyde in cultured human epithelial cells and fibroblasts. Carcinogenesis. 1994;15(5):985–90.

111. Matsuda T, Kawanishi M, Matsui S, Yagi T, Takebe H. Specific tandem GG to TT base substitutions induced by acetaldehyde are due to intra-strand crosslinks between adjacent guanine bases. Nucleic Acids Research. 1998 Apr 1;26(7):1769–74.

112. Lambert B, Chen Y, He SM, Sten M. DNA cross-links in human leucocytes treated with vinyl acetate and acetaldehyde in vitro. Mutation Research/DNA Repair Reports. 1985 Nov 1;146(3):301–3.

113. Blasiak J, Trzeciak A, Malecka-Panas E, Drzewoski J, Wojewódzka M. In vitro genotoxicity of ethanol and acetaldehyde in human lymphocytes and the gastrointestinal tract mucosa cells. Toxicology in Vitro. 2000 Aug 1;14(4):287–95.

114. Costa M, Zhitkovich A, Harris M, Paustenbach D, Gargas M. DNA-protein cross-links produced by various chemicals in cultured human lymphoma cells. J Toxicol Environ Health. 1997 Apr 11;50(5):433–49.

115. Lam CW, Casanova M, Heck HD. Decreased Extractability of DNA from proteins in the rat nasal mucosa after acetaldehyde exposure. Toxicological Sciences. 1986;6(3):541–50.

116. Kuykendall JR, Taylor ML, Bogdanffy MS. Cytotoxicity and DNA-protein crosslink formation in rat nasal tissues exposed to vinyl acetate are carboxylesterase-mediated. Toxicol Appl Pharmacol. 1993 Dec;123(2):283–92.

117. Ali A, Xiao W, Babar ME, Bi Y. Double-Stranded Break Repair in Mammalian Cells and Precise Genome Editing. Genes. 2022;13(5).

118. Caldecott KW. DNA single-strand break repair and human genetic disease. Trends in Cell Biology. 2022 Sep 1;32(9):733–45.

119. Zhang S, Chen H, Wang A, Liu Y, Hou H, Hu Q. Assessment of genotoxicity of four volatile pollutants from cigarette smoke based on the in vitro γH2AX assay using high content screening. Environ Toxicol Pharmacol. 2017 Oct;55:30–6.

120. Furnus CC, Ulrich MA, Terreros MC, Dulout FN. The induction of aneuploidy in cultured Chinese hamster cells by propionaldehyde and chloral hydrate. Mutagenesis. 1990 Jul 1;5(4):323–6.

121. Tacconi EM, Lai X, Folio C, Porru M, Zonderland G, Badie S, et al. BRCA1 and BRCA2 tumor suppressors protect against endogenous acetaldehyde toxicity. EMBO Molecular Medicine. 2017 Oct 1;9(10):1398–414.

122. Rajasinghe H, Jayatilleke E, Shaw S. DNA cleavage during ethanol metabolism: role of superoxide radicals and catalytic iron. Life Sci. 1990;47(9):807–14.

123. Hande V, Teo K, Srikanth P, Wong JSM, Sethu S, Martinez- Lopez W, et al. Investigations on the new mechanism of action for acetaldehyde-induced clastogenic effects in human lung fibroblasts. Mutation Research/Genetic Toxicology and Environmental Mutagenesis. 2021 Jan 1;861–862:503303.

124. Singh NP, Khan A. Acetaldehyde: genotoxicity and cytotoxicity in human lymphocytes. Mutation research/DNA repair. 1995 Jul 1;337(1):9–17.

125. Rulten SL, Hodder E, Ripley TL, Stephens DN, Mayne LV. Alcohol induces DNA damage and the Fanconi anemia D2 protein implicating FANCD2 in the DNA damage response pathways in brain. Alcohol Clin Exp Res. 2008 Jul;32(7):1186–96.

126. Garaycoechea JI, Crossan GP, Langevin F, Mulderrig L, Louzada S, Yang F, et al. Alcohol and endogenous aldehydes damage chromosomes and mutate stem cells. Nature [Internet]. 2018; Available from: http://dx.doi.org/10.1038/nature25154

127. Signorini-Allibe N, Gonthier B, Lamarche F, Eysseric H, Barret L. Chronic consumption of ethanol leads to substantial cell damage in cultured rat astrocytes in conditions promoting acetaldehyde accumulation. Alcohol Alcohol. 2005 Jun;40(3):163–71.

128. Bird RP, Draper HH, Basrur PK. Effect of malonaldehyde and acetaldehyde on cultured mammalian cells: Production of micronuclei and chromosomal aberrations. Mutation Research/Genetic Toxicology. 1982 May 1;101(3):237–46.

129. Matsuzaki K, Kumatoriya K, Tando M, Kometani T, Shinohara M. Polyphenols from persimmon fruit attenuate acetaldehyde-induced DNA double-strand breaks by scavenging acetaldehyde. Scientific Reports. 2022 Jun 18;12(1):10300.

130. Kotova N, Vare D, Schultz N, Gradecka Meesters D, Stępnik M, Grawé J, et al. Genotoxicity of alcohol is linked to DNA replication-associated damage and homologous recombination repair. Carcinogenesis. 2013 Feb 1;34(2):325–30.

131. Prakash S, Johnson RE, Prakash L. Eukaryotic translesion synthesis DNA polymerases: specificity of structure and function. Annu Rev Biochem. 2005 Jun 1;74(1):317–53.

132. Liu X, Lao Y, Yang IY, Hecht SS, Moriya M. Replication-Coupled Repair of Crotonaldehyde/Acetaldehyde-Induced Guanine−Guanine Interstrand Cross-Links and Their Mutagenicity. Biochemistry. 2006 Oct 1;45(42):12898–905.

133. Upton DC, Wang X, Blans P, Perrino FW, Fishbein JC, Akman SA. Mutagenesis by exocyclic alkylamino purine adducts in Escherichia coli. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis. 2006 Jul 25;599(1):1–10.

134. 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. 2006 Jul 1;19(7):960–7.

135. Stein S, Lao Y, Yang IY, Hecht SS, Moriya M. Genotoxicity of acetaldehyde- and crotonaldehyde-induced 1,N2-propanodeoxyguanosine DNA adducts in human cells. Mutation Research/Genetic Toxicology and Environmental Mutagenesis. 2006 Sep 19;608(1):1–7.

136. Vijayraghavan S, Porcher L, Mieczkowski PA, Saini N. Acetaldehyde makes a distinct mutation signature in single-stranded DNA. Nucleic Acids Res. 2022 Jul 22;50(13):7451–64.

137. Paget V, Lechevrel M, Sichel F. Acetaldehyde-induced mutational pattern in the tumour suppressor gene TP53 analysed by use of a functional assay, the FASAY (functional analysis of separated alleles in yeast). Mutat Res. 2008 Mar 29;652(1):12–9.

138. Paget V, Lechevrel M, André V, Le Goff J, Pottier D, Billet S, et al. Benzo[a]pyrene, Aflatoxine B1 and Acetaldehyde Mutational Patterns in TP53 Gene Using a Functional Assay: Relevance to Human Cancer Aetiology. PLOS ONE. 2012 Feb 3;7(2):e30921.

139. Noori P, Hou SM. Mutational spectrum induced by acetaldehyde in the HPRT gene of human T lymphocytes resembles that in the p53 gene of esophageal cancers. Carcinogenesis. 2001 Nov 1;22(11):1825–30.

140. He SM, Lambert B. Acetaldehyde-induced mutation at the hprt Locus in Human Lymphocytes In Vitro. Environmental and Molecular Mutagenesis. 1990;16(2):57–63.

141. Hou SM. Novel types of mutation identified at the hprt locus of human T-lymphocytes. Mutat Res. 1994 Jul 1;308(1):23–31.

142. Lambert B, Andersson B, Bastlova T, Hou SM, Hellgren D, Kolman A. Mutations induced in the hypoxanthine phosphoribosyl transferase gene by three urban air pollutants: acetaldehyde, benzo[a]pyrene diolepoxide, and ethylene oxide. Environ Health Perspect. 1994 Oct;102 Suppl 4(Suppl 4):135–8.

143. Terashima I, Matsuda T, Fang TW, Suzuki N, Kobayashi J, Kohda K, et al. Miscoding Potential of the N2-Ethyl-2‘-deoxyguanosine DNA Adduct by the Exonuclease-Free Klenow Fragment of Escherichia coli DNA Polymerase I. Biochemistry. 2001 Apr 1;40(13):4106–14.

144. Sonohara Y, Takatsuka R, Masutani C, Iwai S, Kuraoka I. Acetaldehyde induces NER repairable mutagenic DNA lesions. Carcinogenesis. 2022 Jan 1;43(1):52–9.

145. Soffritti M, Belpoggi F, Lambertin L, Lauriola M, Padovani M, Maltoni C. Results of long-term experimental studies on the carcinogenicity of formaldehyde and acetaldehyde in rats. Ann N Y Acad Sci. 2002 Dec;982:87–105.

146. Feron VJ, Kruysse A, Woutersen RA. Respiratory tract tumours in hamsters exposed to acetaldehyde vapour alone or simultaneously to benzo(a)pyrene or diethylnitrosamine. European Journal of Cancer and Clinical Oncology. 1982 Jan 1;18(1):13–31.

147. Feron VJ, Kuper CF, Spit BJ, Reuzel PG, Woutersen RA. Glass fibers and vapor phase components of cigarette smoke as cofactors in experimental respiratory tract carcinogenesis. Carcinog Compr Surv. 1985;8:93–118.

148. Woutersen RA, Appelman LM, Van Der Heijden CA. Inhalation toxicity of acetaldehyde in rats II. Carcinogenicity study: Interim results after 15 months. Toxicology. 1984 May 14;31(2):123–33.

149. Woutersen RA, Appelman LM, Van Garderen-Hoetmer A, Feron VJ. Inhalation toxicity of acetaldehyde in rats. III. Carcinogenicity study. Toxicology. 1986 Oct 31;41(2):213–31.

150. Müller MF, Zhou Y, Adams DJ, Arends MJ. Effects of long-term ethanol consumption and Aldh1b1 depletion on intestinal tumourigenesis in mice. The Journal of Pathology. 2017 Apr 1;241(5):649–60.

151. Watabiki T, Okii Y, Tokiyasu T, Yoshimura S, Yoshida M, Akane A, et al. Long-term ethanol consumption in ICR mice causes mammary tumor in females and liver fibrosis in males. Alcohol Clin Exp Res. 2000 Apr;24(4 Suppl):117S-122S.

152. Beland FA, Benson RW, Mellick PW, Kovatch RM, Roberts DW, Fang JL, et al. Effect of ethanol on the tumorigenicity of urethane (ethyl carbamate) in B6C3F1 mice. Food and Chemical Toxicology. 2005 Jan 1;43(1):1–19.

153. Cerretelli G, Zhou Y, Müller MF, Adams DJ, Arends MJ. Ethanol-induced formation of colorectal tumours and precursors in a mouse model of Lynch syndrome. J Pathol. 2021 Dec;255(4):464–74.

154. Roy HK, Gulizia JM, Karolski WJ, Ratashak A, Sorrell MF, Tuma D. Ethanol promotes intestinal tumorigenesis in the MIN mouse. Cancer Epidemiol Biomarkers Prev. 2002 Nov;11(11):1499–502.

155. Radike MJ, Stemmer KL, Bingham E. Effect of ethanol on vinyl chloride carcinogenesis. Environ Health Perspect. 1981 Oct;41:59–62.

156. Soffritti M, Belpoggi F, Cevolani D, Guarino M, Padovani M, Maltoni C. Results of long-term experimental studies on the carcinogenicity of methyl alcohol and ethyl alcohol in rats. Ann N Y Acad Sci. 2002 Dec;982:46–69.

157. Yip-Schneider MT, Doyle CJ, McKillop IH, Wentz SC, Brandon-Warner E, Matos JM, et al. Alcohol induces liver neoplasia in a novel alcohol-preferring rat model. Alcohol Clin Exp Res. 2011 Dec;35(12):2216–25.

158. National Toxicology Program. Formaldehyde. In: Report on Carcinogens, Fifteenth Edition. Research Triangle Park, NC: U.S. Department of Health and Human Services, Public Health Service; 2021.

159. Brewer TF, Burgos-Barragan G, Wit N, Patel KJ, Chang CJ. A 2-aza-Cope reactivity-based platform for ratiometric fluorescence imaging of formaldehyde in living cells. Chem Sci. 2017 May 1;8(5):4073–81.

160. Garrow TA, Brenner AA, Whitehead VM, Chen XN, Duncan RG, Korenberg JR, et al. Cloning of human cDNAs encoding mitochondrial and cytosolic serine hydroxymethyltransferases and chromosomal localization. Journal of Biological Chemistry. 1993 Jun 5;268(16):11910–6.

161. Kikuchi G, Motokawa Y, Yoshida T, Hiraga K. Glycine cleavage system: reaction mechanism, physiological significance, and hyperglycinemia. Proc Jpn Acad Ser B Phys Biol Sci. 2008;84(7):246–63.

162. Porter DH, Cook RJ, Wagner C. Enzymatic properties of dimethylglycine dehydrogenase and sarcosine dehydrogenase from rat liver. Archives of Biochemistry and Biophysics. 1985 Dec 1;243(2):396–407.

163. Morellato AE, Umansky C, Pontel LB. The toxic side of one-carbon metabolism and epigenetics. Redox biology. 2021/01/09 ed. 2021 Apr;40:101850.

164. Staab CA, Ålander J, Morgenstern R, Grafström RC, Höög JO. The Janus face of alcohol dehydrogenase 3. Chemico-Biological Interactions. 2009 Mar 16;178(1):29–35.

165. Uotila L, Koivusalo M. Purification and Properties of S-Formylglutathione Hydrolase from Human Liver. Journal of Biological Chemistry. 1974 Dec 10;249(23):7664–72.

166. Tong Z, Zhang J, Luo W, Wang W, Li F, Li H, et al. Urine formaldehyde level is inversely correlated to mini mental state examination scores in senile dementia. Neurobiol Aging. 2011 Jan;32(1):31–41.

167. Yue X, Zhang Y, Xing W, Chen Y, Mu C, Miao Z, et al. A Sensitive and Rapid Method for Detecting Formaldehyde in Brain Tissues. Anal Cell Pathol (Amst). 2017;2017:9043134.

168. Qiang M, Xiao R, Su T, Wu BB, Tong ZQ, Liu Y, et al. A novel mechanism for endogenous formaldehyde elevation in SAMP8 mouse. J Alzheimers Dis. 2014;40(4):1039–53.

169. Li T, Su T, He Y, Lu J, Mo W, Wei Y, et al. Brain Formaldehyde is Related to Water Intake behavior. Aging Dis. 2016 Oct;7(5):561–84.

170. Mei Y, Duan C, Li X, Zhao Y, Cao F, Shang S, et al. Reduction of Endogenous Melatonin Accelerates Cognitive Decline in Mice in a Simulated Occupational Formaldehyde Exposure Environment. International Journal of Environmental Research and Public Health. 2016;13(3).

171. Zhang J, Yue X, Luo H, Jiang W, Mei Y, Ai L, et al. Illumination with 630 nm Red Light Reduces Oxidative Stress and Restores Memory by Photo-Activating Catalase and Formaldehyde Dehydrogenase in SAMP8 Mice. Antioxid Redox Signal. 2019 Apr 10;30(11):1432–49.

172. Yao D, He Q, Bai S, Zhao H, Yang J, Cui D, et al. Accumulation of formaldehyde causes motor deficits in an in vivo model of hindlimb unloading. Commun Biol. 2021 Aug 19;4(1):933.

173. d'A. Heck H, White EL, Casanova-Schmitz M. Determination of formaldehyde in biological tissues by gas chromatography/mass spectrometry. Biomedical Mass Spectrometry. 1982 Aug 1;9(8):347–53.

174. Tong Z, Han C, Luo W, Li H, Luo H, Qiang M, et al. Aging-associated excess formaldehyde leads to spatial memory deficits. Scientific Reports. 2013 May 9;3(1):1807.

175. Mei Y, Jiang C, Wan Y, Lv J, Jia J, Wang X, et al. Aging-associated formaldehyde-induced norepinephrine deficiency contributes to age-related memory decline. Aging Cell. 2015 Aug;14(4):659–68.

176. Ai L, Tan T, Tang Y, Yang J, Cui D, Wang R, et al. Endogenous formaldehyde is a memory-related molecule in mice and humans. Commun Biol. 2019;2:446.

177. Deltour L, Foglio MH, Duester G. Metabolic Deficiencies in Alcohol Dehydrogenase Adh1, Adh3, and Adh4 Null Mutant Mice: Overlapping Roles of Adh1 and Adh4 in Ethanol Clearance and Metabolism of Retinol to Retinoic Acid. Journal of Biological Chemistry. 1999 Jun 11;274(24):16796–801.

178. Rosado IV, Langevin F, Crossan GP, Takata M, Patel KJ. Formaldehyde catabolism is essential in cells deficient for the Fanconi anemia DNA-repair pathway. Nature Structural & Molecular Biology. 2011 Dec 1;18(12):1432–4.

179. Pontel LB, Rosado IV, Burgos-Barragan G, Garaycoechea JI, Yu R, Arends MJ, et al. Endogenous Formaldehyde Is a Hematopoietic Stem Cell Genotoxin and Metabolic Carcinogen. Molecular Cell. 2015 Oct 1;60(1):177–88.

180. Reingruber H, Pontel LB. Formaldehyde metabolism and its impact on human health. Current Opinion in Toxicology. 2018 Jun 1;9:28–34.

181. Tsukada Y, Fang J, Erdjument-Bromage H, Warren ME, Borchers CH, Tempst P, et al. Histone demethylation by a family of JmjC domain-containing proteins. Nature. 2005/12/20 ed. 2006 Feb 16;439(7078):811–6.

182. Aravind L, Koonin EV. The DNA-repair protein AlkB, EGL-9, and leprecan define new families of 2-oxoglutarate- and iron-dependent dioxygenases. Genome Biology. 2001 Feb 19;2(3):research0007.1.

183. Shi Y, Lan F, Matson C, Mulligan P, Whetstine JR, Cole PA, et al. Histone Demethylation Mediated by the Nuclear Amine Oxidase Homolog LSD1. Cell. 2004 Dec 29;119(7):941–53.

184. Hauptmann M, Lubin JH, Stewart PA, Hayes RB, Blair A. Mortality from Solid Cancers among Workers in Formaldehyde Industries. American Journal of Epidemiology. 2004 Jun 15;159(12):1117–30.

185. Beane Freeman LE, Blair A, Lubin JH, Stewart PA, Hayes RB, Hoover RN, et al. Mortality From Lymphohematopoietic Malignancies Among Workers in Formaldehyde Industries: The National Cancer Institute Cohort. JNCI: Journal of the National Cancer Institute. 2009 May 20;101(10):751–61.

186. Hauptmann M, Stewart PA, Lubin JH, Beane Freeman LE, Hornung RW, Herrick RF, et al. Mortality From Lymphohematopoietic Malignancies and Brain Cancer Among Embalmers Exposed to Formaldehyde. JNCI: Journal of the National Cancer Institute. 2009 Dec 16;101(24):1696–708.

187. Pinkerton LE, Hein MJ, Stayner LT. Mortality among a cohort of garment workers exposed to formaldehyde: an update. Occup Environ Med. 2004 Mar 1;61(3):193.

188. Coggon D, Harris EC, Poole J, Palmer KT. Extended Follow-Up of a Cohort of British Chemical Workers Exposed to Formaldehyde. JNCI: Journal of the National Cancer Institute. 2003 Nov 5;95(21):1608–15.

189. Luce D, Leclerc A, Bégin D, Demers PA, Gérin M, Orlowski E, et al. Sinonasal cancer and occupational exposures: a pooled analysis of 12 case–control studies. Cancer Causes & Control. 2002 Mar 1;13(2):147–57.

190. Vaughan TL, Stewart PA, Teschke K, Lynch CF, Swanson GM, Lyon JL, et al. Occupational exposure to formaldehyde and wood dust and nasopharyngeal carcinoma. Occupational and environmental medicine. 2000/05/16 ed. 2000 Jun;57(6):376–84.

191. Zhang L, Steinmaus C, Eastmond DA, Xin XK, Smith MT. Formaldehyde exposure and leukemia: a new meta-analysis and potential mechanisms. Mutat Res. 2009 Jun;681(2–3):150–68.

192. Kwon SC, Kim I, Song J, Park J. Does formaldehyde have a causal association with nasopharyngeal cancer and leukaemia? Annals of Occupational and Environmental Medicine. 2018 Jan 31;30(1):5.

193. Song Y, Cheng W, Li H, Liu X. The global, regional, national burden of nasopharyngeal cancer and its attributable risk factors (1990-2019) and predictions to 2035. Cancer Med. 2022 Nov;11(22):4310–20.

194. Zhang R, He Y, Wei B, Lu Y, Zhang J, Zhang N, et al. Nasopharyngeal Carcinoma Burden and Its Attributable Risk Factors in China: Estimates and Forecasts from 1990 to 2050. Int J Environ Res Public Health. 2023 Feb 8;20(4).

195. Heidari-Foroozan M, Saeedi Moghaddam S, Keykhaei M, Shobeiri P, Azadnajafabad S, Esfahani Z, et al. Regional and national burden of leukemia and its attributable burden to risk factors in 21 countries and territories of North Africa and Middle East, 1990-2019: results from the GBD study 2019. J Cancer Res Clin Oncol. 2023 Jul;149(8):4149–61.

196. Lu K, Collins LB, Ru H, Bermudez E, Swenberg JA. Distribution of DNA Adducts Caused by Inhaled Formaldehyde Is Consistent with Induction of Nasal Carcinoma but Not Leukemia. Toxicological Sciences. 2010 Aug 1;116(2):441–51.

197. Ridpath JR, Nakamura A, Tano K, Luke AM, Sonoda E, Arakawa H, et al. Cells Deficient in the FANC/BRCA Pathway Are Hypersensitive to Plasma Levels of Formaldehyde. Cancer Research. 2007 Dec 3;67(23):11117–22.

198. Kawanishi M, Matsuda T, Yagi T. Genotoxicity of formaldehyde: molecular basis of DNA damage and mutation. Frontiers in Environmental Science [Internet]. 2014;2. Available from: https://www.frontiersin.org/articles/10.3389/fenvs.2014.00036

199. Anandarajan V, Noguchi C, Oleksak J, Grothusen G, Terlecky D, Noguchi E. Genetic investigation of formaldehyde-induced DNA damage response in Schizosaccharomyces pombe. Curr Genet. 2020 Jun;66(3):593–605.

200. Rieckher M, Gallrein C, Alquezar-Artieda N, Bourached-Silva N, Vaddavalli PL, Mares D, et al. Distinct DNA repair mechanisms prevent formaldehyde toxicity during development, reproduction and aging. Nucleic Acids Res. 2024 Jun 19;gkae519.

201. Beland FA, Fullerton NF, Heflich RH. Rapid isolation, hydrolysis and chromatography of formaldehyde-modified DNA. Journal of Chromatography B: Biomedical Sciences and Applications. 1984 Jun 8;308:121–31.

202. Zhong W, Que Hee SS. Comparison of UV, Fluorescence, and Electrochemical Detectors for the Analysis of Formaldehyde-Induced DNA Adducts. Journal of Analytical Toxicology. 2005 Apr 1;29(3):182–7.

203. Chang YJ, Cooke MS, Chen YR, Yang SF, Li PS, Hu CW, et al. Is high resolution a strict requirement for mass spectrometry-based cellular DNA adductomics? Chemosphere. 2021 Jul;274:129991.

204. Lu K, Craft S, Nakamura J, Moeller BC, Swenberg JA. Use of LC-MS/MS and Stable Isotopes to Differentiate Hydroxymethyl and Methyl DNA Adducts from Formaldehyde and Nitrosodimethylamine. Chem Res Toxicol. 2012 Mar 19;25(3):664–75.

205. Zhong W, Que Hee SS. Formaldehyde-induced DNA adducts as biomarkers of in vitro human nasal epithelial cell exposure to formaldehyde. Mutation Research/Genetic Toxicology and Environmental Mutagenesis. 2004 Sep 12;563(1):13–24.

206. Wang M, Cheng G, Balbo S, Carmella SG, Villalta PW, Hecht SS. Clear Differences in Levels of a Formaldehyde-DNA Adduct in Leukocytes of Smokers and Nonsmokers. Cancer Research. 2009 Sep 14;69(18):7170–4.

207. Moeller BC, Lu K, Doyle-Eisele M, McDonald J, Gigliotti A, Swenberg JA. Determination of N2-Hydroxymethyl-dG Adducts in the Nasal Epithelium and Bone Marrow of Nonhuman Primates Following 13CD2-Formaldehyde Inhalation Exposure. Chemical Research in Toxicology. 2011 Feb 18;24(2):162–4.

208. Lu K, Gul H, Upton PB, Moeller BC, Swenberg JA. Formation of Hydroxymethyl DNA Adducts in Rats Orally Exposed to Stable Isotope Labeled Methanol. Toxicological Sciences. 2012 Mar 1;126(1):28–38.

209. Yu R, Lai Y, Hartwell HJ, Moeller BC, Doyle-Eisele M, Kracko D, et al. Formation, Accumulation, and Hydrolysis of Endogenous and Exogenous Formaldehyde-Induced DNA Damage. Toxicological Sciences. 2015 Jul 1;146(1):170–82.

210. Leng J, Liu CW, Hartwell HJ, Yu R, Lai Y, Bodnar W, et al. Evaluation of inhaled low-dose formaldehyde-induced DNA adducts and DNA–protein cross-links by liquid chromatography–tandem mass spectrometry. Archives of Toxicology. 2019;93:763–73.

211. Cheng G, Wang M, Upadhyaya P, Villalta PW, Hecht SS. Formation of Formaldehyde Adducts in the Reactions of DNA and Deoxyribonucleosides with α-Acetates of 4-(Methylnitrosamino)-1-(3-pyridyl)-1-butanone (NNK), 4-(Methylnitrosamino)-1-(3-pyridyl)-1-butanol (NNAL), and N-Nitrosodimethylamine (NDMA). Chem Res Toxicol. 2008 Mar 17;21(3):746–51.

212. Lu K, Moeller B, Doyle-Eisele M, McDonald J, Swenberg JA. Molecular dosimetry of N2-hydroxymethyl-dG DNA adducts in rats exposed to formaldehyde. Chem Res Toxicol. 2011 Feb 18;24(2):159–61.

213. Ye X, Ji Z, Wei C, McHale CM, Ding S, Thomas R, et al. Inhaled formaldehyde induces DNA-protein crosslinks and oxidative stress in bone marrow and other distant organs of exposed mice. Environ Mol Mutagen. 2013 Dec;54(9):705–18.

214. Yu GY, Song XF, Liu Y, Sun ZW. Inhaled formaldehyde induces bone marrow toxicity via oxidative stress in exposed mice. Asian Pac J Cancer Prev. 2014;15(13):5253–7.

215. Ghelli F, Bellisario V, Squillacioti G, Panizzolo M, Santovito A, Bono R. Formaldehyde in Hospitals Induces Oxidative Stress: The Role of GSTT1 and GSTM1 Polymorphisms. Toxics. 2021 Jul 30;9(8).

216. Umansky C, Morellato AE, Rieckher M, Scheidegger MA, Martinefski MR, Fernández GA, et al. Endogenous formaldehyde scavenges cellular glutathione resulting in redox disruption and cytotoxicity. Nature Communications. 2022 Feb 8;13(1):745.

217. Cheng KC, Cahill DS, Kasai H, Nishimura S, Loeb LA. 8-Hydroxyguanine, an abundant form of oxidative DNA damage, causes G-T and A-C substitutions. Journal of Biological Chemistry. 1992 Jan 5;267(1):166–72.

218. Hu CW, Chang YJ, Cooke MS, Chao MR. DNA Crosslinkomics: A Tool for the Comprehensive Assessment of Interstrand Crosslinks Using High Resolution Mass Spectrometry. Anal Chem. 2019 Dec 3;91(23):15193–203.

219. Hoffman EA, Frey BL, Smith LM, Auble DT. Formaldehyde Crosslinking: A Tool for the Study of Chromatin Complexes*. Journal of Biological Chemistry. 2015 Oct 30;290(44):26404–11.

220. Enderle J, Dorn A, Puchta H. DNA- and DNA-Protein-Crosslink Repair in Plants. International Journal of Molecular Sciences. 2019;20(17).

221. Speit G, Schmid O, Neuss S, Schütz P. Genotoxic effects of formaldehyde in the human lung cell line A549 and in primary human nasal epithelial cells. Environ Mol Mutagen. 2008 May;49(4):300–7.

222. Speit G, Neuss S, Schmid O. The human lung cell line A549 does not develop adaptive protection against the DNA-damaging action of formaldehyde. Environ Mol Mutagen. 2010 Mar;51(2):130–7.

223. Zhang BY, Shi YQ, Chen X, Dai J, Jiang ZF, Li N, et al. Protective effect of curcumin against formaldehyde-induced genotoxicity in A549 Cell Lines. J Appl Toxicol. 2013 Dec;33(12):1468–73.

224. Shi YQ, Chen X, Dai J, Jiang ZF, Li N, Zhang BY, et al. Selenium pretreatment attenuates formaldehyde-induced genotoxicity in A549 cell lines. Toxicol Ind Health. 2014 Nov;30(10):901–9.

225. Mu H, Liu Q, Niu H, Wang D, Tang J, Duan J. Autophagy promotes DNA-protein crosslink clearance. Mutat Res Genet Toxicol Environ Mutagen. 2016 Feb;797:21–5.

226. Wilborn F, Brendel M. Formation and stability of interstrand cross-links induced by cis- and trans-diamminedichloroplatinum (II) in the DNA of Saccharomyces cerevisiae strains differing in repair capacity. Curr Genet. 1989 Dec;16(5–6):331–8.

227. de Graaf B, Clore A, McCullough AK. Cellular pathways for DNA repair and damage tolerance of formaldehyde-induced DNA-protein crosslinks. DNA Repair. 2009 Oct 2;8(10):1207–14.

228. Stingele J, Schwarz MS, Bloemeke N, Wolf PG, Jentsch S. A DNA-dependent protease involved in DNA-protein crosslink repair. Cell. 2014 Jul 17;158(2):327–38.

229. Kaufman BA, Newman SM, Hallberg RL, Slaughter CA, Perlman PS, Butow RA. In organello formaldehyde crosslinking of proteins to mtDNA: identification of bifunctional proteins. Proc Natl Acad Sci U S A. 2000 Jul 5;97(14):7772–7.

230. Miller CA 3rd, Costa M. Analysis of proteins cross-linked to DNA after treatment of cells with formaldehyde, chromate, and cis-diamminedichloroplatinum(II). Mol Toxicol. 1989 Winter;2(1):11–26.

231. Miller CA 3rd, Costa M. Immunodetection of DNA-protein crosslinks by slot blotting. Mutat Res. 1990 Apr;234(2):97–106.

232. Nakano T, Katafuchi A, Matsubara M, Terato H, Tsuboi T, Masuda T, et al. Homologous Recombination but Not Nucleotide Excision Repair Plays a Pivotal Role in Tolerance of DNA-Protein Cross-links in Mammalian Cells. Journal of Biological Chemistry. 2009 Oct 2;284(40):27065–76.

233. Kumari A, Lim YX, Newell AH, Olson SB, McCullough AK. Formaldehyde-induced genome instability is suppressed by an XPF-dependent pathway. DNA Repair. 2012 Mar 1;11(3):236–46.

234. Chaw YFM, Crane LE, Lange P, Shapiro R. Isolation and identification of cross-links from formaldehyde-treated nucleic acids. Biochemistry. 1980 Nov 1;19(24):5525–31.

235. Huang H, Hopkins PB. DNA interstrand cross-linking by formaldehyde: nucleotide sequence preference and covalent structure of the predominant cross-link formed in synthetic oligonucleotides. Journal of the American Chemical Society. 1993;115(21):9402–8.

236. Liu Y, Li CM, Lu Z, Ding S, Yang X, Mo J. Studies on formation and repair of formaldehyde-damaged DNA by detection of DNA-protein crosslinks and DNA breaks. Frontiers in bioscience : a journal and virtual library. 2005/09/09 ed. 2006 Jan 1;11:991–7.

237. Ohba Y, Morimitsu Y, Watarai A. Reaction of formaldehyde with calf-thymus nucleohistone. Eur J Biochem. 1979 Oct;100(1):285–93.

238. Sewell BT, Bouloukos C, von Holt C. Formaldehyde and glutaraldehyde in the fixation of chromatin for electron microscopy. J Microsc. 1984 Oct;136(Pt 1):103–12.

239. Solomon MJ, Varshavsky A. Formaldehyde-mediated DNA-protein crosslinking: a probe for in vivo chromatin structures. Proc Natl Acad Sci U S A. 1985 Oct;82(19):6470–4.

240. Quievryn G, Zhitkovich A. Loss of DNA-protein crosslinks from formaldehyde-exposed cells occurs through spontaneous hydrolysis and an active repair process linked to proteosome function. Carcinogenesis. 2000/07/27 ed. 2000 Aug;21(8):1573–80.

241. Nakano T, Shoulkamy MI, Tsuda M, Sasanuma H, Hirota K, Takata M, et al. Participation of TDP1 in the repair of formaldehyde-induced DNA-protein cross-links in chicken DT40 cells. PLoS One. 2020;15(6):e0234859.

242. Wilkins RJ, Macleod HD. Formaldehyde induced DNA—Protein crosslinks in Escherichia coli. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis. 1976 Jul 1;36(1):11–6.

243. Schouten JP. Hybridization selection of covalent nucleic acid-protein complexes. 2. Cross-linking of proteins to specific Escherichia coli mRNAs and DNA sequences by formaldehyde treatment of intact cells. J Biol Chem. 1985 Aug 15;260(17):9929–35.

244. Ruggiano A, Vaz B, Kilgas S, Popović M, Rodriguez-Berriguete G, Singh AN, et al. The protease SPRTN and SUMOylation coordinate DNA-protein crosslink repair to prevent genome instability. Cell Rep. 2021 Dec 7;37(10):110080.

245. Fornace AJJ, Lechner JF, Grafstrom RC, Harris CC. DNA repair in human bronchial epithelial cells. Carcinogenesis. 1982;3(12):1373–7.

246. Grafstrom RC, Fornace AJJ, Autrup H, Lechner JF, Harris CC. Formaldehyde damage to DNA and inhibition of DNA repair in human bronchial cells. Science. 1983 Apr 8;220(4593):216–8.

247. Grafstrom RC, Fornace AJ, Harris CC. Repair of DNA damage caused by formaldehyde in human cells. Cancer Res. 1984 Oct;44(10):4323–7.

248. Saladino AJ, Willey JC, Lechner JF, Grafstrom RC, LaVeck M, Harris CC. Effects of formaldehyde, acetaldehyde, benzoyl peroxide, and hydrogen peroxide on cultured normal human bronchial epithelial cells. Cancer Res. 1985 Jun;45(6):2522–6.

249. Fornace AJJ. Detection of DNA single-strand breaks produced during the repair of damage by DNA-protein cross-linking agents. Cancer Res. 1982 Jan;42(1):145–9.

250. Zeller J, Högel J, Linsenmeyer R, Teller C, Speit G. Investigations of potential susceptibility toward formaldehyde-induced genotoxicity. Arch Toxicol. 2012 Sep;86(9):1465–73.

251. Craft TR, Bermudez E, Skopek TR. Formaldehyde mutagenesis and formation of DNA-protein crosslinks in human lymphoblasts in vitro. Mutat Res. 1987 Jan;176(1):147–55.

252. Ren X, Ji Z, McHale CM, Yuh J, Bersonda J, Tang M, et al. The impact of FANCD2 deficiency on formaldehyde-induced toxicity in human lymphoblastoid cell lines. Archives of Toxicology. 2013 Jan 1;87(1):189–96.

253. Shaham J, Bomstein Y, Melzer A, Ribak J. DNA-Protein Crosslinks and Sister Chromatid Exchanges as Biomarkers of Exposure to Formaldehyde. Int J Occup Environ Health. 1997 Apr;3(2):95–104.

254. Shaham J, Bomstein Y, Gurvich R, Rashkovsky M, Kaufman Z. DNA-protein crosslinks and p53 protein expression in relation to occupational exposure to formaldehyde. Occup Environ Med. 2003 Jun;60(6):403–9.

255. Andersson M, Agurell E, Vaghef H, Bolcsfoldi G, Hellman B. Extended-term cultures of human T-lymphocytes and the comet assay: a useful combination when testing for genotoxicity in vitro? Mutat Res. 2003 Sep 9;540(1):43–55.

256. Schmid O, Speit G. Genotoxic effects induced by formaldehyde in human blood and implications for the interpretation of biomonitoring studies. Mutagenesis. 2007 Jan 1;22(1):69–74.

257. Zeller J, Ulrich A, Mueller JU, Riegert C, Neuss S, Bruckner T, et al. Is individual nasal sensitivity related to cellular metabolism of formaldehyde and susceptibility towards formaldehyde-induced genotoxicity? Mutat Res. 2011 Jul 14;723(1):11–7.

258. Lin D, Guo Y, Yi J, Kuang D, Li X, Deng H, et al. Occupational exposure to formaldehyde and genetic damage in the peripheral blood lymphocytes of plywood workers. Journal of occupational health. 2013/05/08 ed. 2013;55(4):284–91.

259. O'Connor PM, Fox BW. Comparative studies of DNA cross-linking reactions following methylene dimethanesulphonate and its hydrolytic product, formaldehyde. Cancer Chemother Pharmacol. 1987;19(1):11–5.

260. O'Connor PM, Fox BW. Isolation and characterization of proteins cross-linked to DNA by the antitumor agent methylene dimethanesulfonate and its hydrolytic product formaldehyde. J Biol Chem. 1989 Apr 15;264(11):6391–7.

261. Speit G, Schütz P, Merk O. Induction and repair of formaldehyde-induced DNA–protein crosslinks in repair-deficient human cell lines. Mutagenesis. 2000 Jan 1;15(1):85–90.

262. Shaham J, Bomstein Y, Meltzer A, Kaufman Z, Palma E, Ribak J. DNA--protein crosslinks, a biomarker of exposure to formaldehyde--in vitro and in vivo studies. Carcinogenesis. 1996 Jan;17(1):121–5.

263. Saito Y, Nishio K, Yoshida Y, Niki E. Cytotoxic effect of formaldehyde with free radicals via increment of cellular reactive oxygen species. Toxicology. 2005 Jun 1;210(2):235–45.

264. Permana PA, Snapka RM. Aldehyde-induced protein-DNA crosslinks disrupt specific stages of SV40 DNA replication. Carcinogenesis. 1994 May;15(5):1031–6.

265. She Y, Li Y, Liu Y, Asai G, Sun S, He J, et al. Formaldehyde induces toxic effects and regulates the expression of damage response genes in BM-MSCs. Acta Biochim Biophys Sin (Shanghai). 2013 Dec;45(12):1011–20.

266. Stingele J, Bellelli R, Alte F, Hewitt G, Sarek G, Maslen SL, et al. Mechanism and Regulation of DNA-Protein Crosslink Repair by the DNA-Dependent Metalloprotease SPRTN. Mol Cell. 2016 Nov 17;64(4):688–703.

267. Ross WE, Shipley N. Relationship between DNA damage and survival in formaldehyde-treated mouse cells. Mutation research. 1980/11/01 ed. 1980 Nov;79(3):277–83.

268. Ross WE, McMillan DR, Ross CF. Comparison of DNA damage by methylmelamines and formaldehyde. J Natl Cancer Inst. 1981 Jul;67(1):217–21.

269. Lin Z, Luo W, Li H, Zhang Y. The effect of endogenous formaldehyde on the rat aorta endothelial cells. Toxicol Lett. 2005 Nov 15;159(2):134–43.

270. Wang A, Robertson JL, Holladay SD, Tennant AH, Lengi AJ, Ahmed SA, et al. Measurement of DNA damage in rat urinary bladder transitional cells: Improved selective harvest of transitional cells and detailed Comet assay protocols. Mutation Research/Genetic Toxicology and Environmental Mutagenesis. 2007 Dec 1;634(1):51–9.

271. Casanova-Schmitz M, Heck HD. Effects of formaldehyde exposure on the extractability of DNA from proteins in the rat nasal mucosa. Toxicol Appl Pharmacol. 1983 Aug;70(1):121–32.

272. Lam CW, Casanova M, Heck HD. Depletion of nasal mucosal glutathione by acrolein and enhancement of formaldehyde-induced DNA-protein cross-linking by simultaneous exposure to acrolein. Arch Toxicol. 1985 Dec;58(2):67–71.

273. Casanova M, Heck H d'A. Further studies of the metabolic incorporation and covalent binding of inhaled [3H]- and [14C]formaldehyde in Fischer-344 rats: Effects of glutathione depletion. Toxicology and Applied Pharmacology. 1987 Jun 15;89(1):105–21.

274. Heck H d'A., Casanova M. Isotope effects and their implications for the covalent binding of inhaled [3H]- and [14C]formaldehyde in the rat nasal mucosa. Toxicology and Applied Pharmacology. 1987 Jun 15;89(1):122–34.

275. Casanova M, Deyo DF, Heck HD. Covalent binding of inhaled formaldehyde to DNA in the nasal mucosa of Fischer 344 rats: analysis of formaldehyde and DNA by high-performance liquid chromatography and provisional pharmacokinetic interpretation. Fundam Appl Toxicol. 1989 Apr;12(3):397–417.

276. Casanova M, Morgan KT, Gross EA, Moss OR, Heck HA. DNA-protein cross-links and cell replication at specific sites in the nose of F344 rats exposed subchronically to formaldehyde. Fundam Appl Toxicol. 1994 Nov;23(4):525–36.

277. Bedford P, Fox BW. The role of formaldehyde in methylene dimethanesulphonate-induced DNA cross-links and its relevance to cytotoxicity. Chem Biol Interact. 1981 Dec;38(1):119–28.

278. Cosma GN, Wilhite AS, Marchok AC. The detection of DNA-protein cross-links in rat tracheal implants exposed in vivo to benzo[a]pyrene and formaldehyde. Cancer Lett. 1988 Oct;42(1–2):13–21.

279. Cosma GN, Jamasbi R, Marchok AC. Growth inhibition and DNA damage induced by benzo[a]pyrene and formaldehyde in primary cultures of rat tracheal epithelial cells. Mutat Res. 1988 Sep;201(1):161–8.

280. Cosma GN, Marchok AC. Benzo[a]pyrene- and formaldehyde-induced DNA damage and repair in rat tracheal epithelial cells. Toxicology. 1988 Oct;51(2–3):309–20.

281. Casanova M, Morgan KT, Steinhagen WH, Everitt JI, Popp JA, Heck H d'A. Covalent binding of inhaled formaldehyde to DNA in the respiratory tract of rhesus monkeys: Pharmacokinetics, rat-to-monkey interspecies scaling, and extrapolation to man. Fundamental and Applied Toxicology. 1991 Aug 1;17(2):409–28.

282. Merk O, Speit G. Significance of formaldehyde-induced DNA-protein crosslinks for mutagenesis. Environ Mol Mutagen. 1998;32(3):260–8.

283. Merk O, Speit G. Detection of crosslinks with the comet assay in relationship to genotoxicity and cytotoxicity. Environ Mol Mutagen. 1999;33(2):167–72.

284. Hu Y, Kabler SL, Tennant AH, Townsend AJ, Kligerman AD. Induction of DNA–protein crosslinks by dichloromethane in a V79 cell line transfected with the murine glutathione-S-transferase theta 1 gene. Mutation Research/Genetic Toxicology and Environmental Mutagenesis. 2006 Sep 5;607(2):231–9.

285. Speit G, Schütz P, Högel J, Schmid O. Characterization of the genotoxic potential of formaldehyde in V79 cells. Mutagenesis. 2007 Nov 1;22(6):387–94.

286. Wei X, Peng Y, Bryan C, Yang K. Mechanisms of DNA−protein cross-link formation and repair. Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics. 2021 Aug 1;1869(8):140669.

287. Peake JD, Noguchi E. Fanconi anemia: current insights regarding epidemiology, cancer, and DNA repair. Hum Genet. 2022 Dec;141(12):1811–36.

288. Li N, Chen H, Wang J. DNA damage and repair in the hematopoietic system. Acta Biochim Biophys Sin (Shanghai). 2022 Jan 25;54(6):847–57.

289. Vock EH, Lutz WK, Ilinskaya O, Vamvakas S. Discrimination between genotoxicity and cytotoxicity for the induction of DNA double-strand breaks in cells treated with aldehydes and diepoxides. Mutation Research/Genetic Toxicology and Environmental Mutagenesis. 1999 Apr 26;441(1):85–93.

290. Yoshida I, Ibuki Y. Formaldehyde-induced histone H3 phosphorylation via JNK and the expression of proto-oncogenes. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis. 2014 Dec 1;770:9–18.

291. Zhang S, Chen H, Wang A, Liu Y, Hou H, Hu Q. Combined effects of co-exposure to formaldehyde and acrolein mixtures on cytotoxicity and genotoxicity in vitro. Environ Sci Pollut Res Int. 2018 Sep;25(25):25306–14.

292. Magaña-Schwencke N, Ekert B, Moustacchi E. Biochemical analysis of damage induced in yeast by formaldehyde. I. Induction of single-strand breaks in DNA and their repair. Mutat Res. 1978 May;50(2):181–93.

293. Noda T, Takahashi A, Kondo N, Mori E, Okamoto N, Nakagawa Y, et al. Repair pathways independent of the Fanconi anemia nuclear core complex play a predominant role in mitigating formaldehyde-induced DNA damage. Biochemical and Biophysical Research Communications. 2011 Jan 7;404(1):206–10.

294. Zhang R, Kang KA, Piao MJ, Kim KC, Lee NH, You HJ, et al. Triphlorethol-A Improves the Non-Homologous End Joining and Base-Excision Repair Capacity Impaired by Formaldehyde. Journal of Toxicology and Environmental Health, Part A. 2011 Apr 29;74(12):811–21.

295. Cantoni O, Costa M. Analysis of the induction of alkali sensitive sites in the DNA by chromate and other agents that induce single strand breaks. Carcinogenesis. 1984 Sep;5(9):1207–9.

296. Graves RJ, Green T. Mouse liver glutathione S-transferase mediated metabolism of methylene chloride to a mutagen in the CHO/HPRT assay. Mutation Research/Genetic Toxicology. 1996 Mar 1;367(3):143–50.

297. Grafström RC. In vitro studies of aldehyde effects related to human respiratory carcinogenesis. Mutation Research/Reviews in Genetic Toxicology. 1990 May 1;238(3):175–84.

298. Kumari A, Owen N, Juarez E, McCullough AK. BLM protein mitigates formaldehyde-induced genomic instability. DNA Repair. 2015 Apr 1;28:73–82.

299. Nadalutti CA, Stefanick DF, Zhao ML, Horton JK, Prasad R, Brooks AM, et al. Mitochondrial dysfunction and DNA damage accompany enhanced levels of formaldehyde in cultured primary human fibroblasts. Scientific Reports. 2020 Mar 27;10(1):5575.

300. She Y, Zhao X, Wu P, Xue L, Liu Z, Zhu M, et al. Astragalus polysaccharide protects formaldehyde-induced toxicity by promoting NER pathway in bone marrow mesenchymal stem cells. Folia Histochem Cytobiol. 2021;59(2):124–33.

301. Graves RJ, Coutts C, Eyton-Jones H, Green T. Relationship between hepatic DNA damage and methylene chloride-induced hepatocarcinogenicity in B6C3F1 mice. Carcinogenesis. 1994 May;15(5):991–6.

302. Gao Y, Guitton-Sert L, Dessapt J, Coulombe Y, Rodrigue A, Milano L, et al. A CRISPR-Cas9 screen identifies EXO1 as a formaldehyde resistance gene. Nat Commun. 2023 Jan 24;14(1):381.

303. Graves RJ, Trueman P, Jones S, Green T. DNA sequence analysis of methylene chloride-induced HPRT mutations in Chinese hamster ovary cells: comparison with the mutation spectrum obtained for 1, 2-dibromoethane and formaldehyde. Mutagenesis. 1996 May 1;11(3):229–33.

304. Grogan D, Jinks-Robertson S. Formaldehyde-induced mutagenesis in Saccharomyces cerevisiae: Molecular properties and the roles of repair and bypass systems. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis. 2012 Mar 1;731(1):92–8.

305. Moerman DG, Baillie DL. Formaldehyde mutagenesis in the nematode Caenorhabditis elegans. Mutation research. 1981/02/01 ed. 1981 Feb;80(2):273–9.

306. Johnsen RC, Baillie DL. Formaldehyde mutagenesis of the eT1 balanced region in Caenorhabditis elegans: dose-response curve and the analysis of mutational events. Mutat Res. 1988 Sep;201(1):137–47.

307. Clark DV, Baillie DL. Genetic analysis and complementation by germ-line transformation of lethal mutations in the unc-22 IV region of Caenorhabditis elegans. Mol Gen Genet. 1992 Mar;232(1):97–105.

308. Schein JE, Marra MA, Benian GM, Fields C, Baillie DL. The use of deficiencies to determine essential gene content in the let-56-unc-22 region of Caenorhabditis elegans. Genome. 1993 Dec;36(6):1148–56.

309. Crosby RM, Richardson KK, Craft TR, Benforado KB, Liber HL, Skopek TR. Molecular analysis of formaldehyde-induced mutations in human lymphoblasts and e. coli. Environmental Mutagenesis. 1988 Jan 1;12(2):155–66.

310. Ohta T, Watanabe-Akanuma M, Tokishita S ichi, Yamagata H. Mutation spectra of chemical mutagens determined by Lac+ reversion assay with Escherichia coli WP3101P–WP3106P tester strains. Mutation Research/Genetic Toxicology and Environmental Mutagenesis. 1999 Mar 15;440(1):59–74.

311. Ohta T, Watanabe-Akanuma M, Yamagata H. A comparison of mutation spectra detected by the Escherichia coli Lac+ reversion assay and the Salmonella typhimurium His+ reversion assay. Mutagenesis. 2000 Jul 1;15(4):317–23.

312. Wang W, Xu J, Xu L, Yue B, Zou F. The instability of (GpT)n and (ApC)n microsatellites induced by formaldehyde in Escherichia coli. Mutagenesis. 2007 Sep 1;22(5):353–7.

313. Slizynska H. Cytological analysis of storage effects on various types of complete and mosaic change induces in Drosophila chromosomes by some chemical mutagens. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis. 1973 Aug 1;19(2):199–213.

314. O'Donnell J, Gerace L, Leister F, Sofer W. Chemical selection of mutants that affect alcohol dehydrogenase in Drosophila. II. Use of 1-pentyne-3-ol. Genetics. 1975 Jan;79(1):73–83.

315. O'Donnell J, Mandel HC, Krauss M, Sofer W. Genetic and cytogenetic analysis of the Adh region in Drosophila melanogaster. Genetics. 1977 Jul;86(3):553–66.

316. Benyajati C, Place AR, Sofer W. Formaldehyde mutagenesis in Drosophila Molecular analysis of ADH-negative mutants. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis. 1983 Sep 1;111(1):1–7.

317. Place AR, Benyajati C, Sofer W. Molecular consequences of two formaldehyde-induced mutations in the alcohol dehydrogenase gene of Drosophila melanogaster. Biochem Genet. 1987 Oct;25(9–10):621–38.

318. Le L, Ayer S, Place AR, Benyajati C. Analysis of formaldehyde-induced Adh mutations in Drosophila by RNA structure mapping and direct sequencing of PCR-amplified genomic DNA. Biochem Genet. 1990 Aug;28(7–8):367–87.

319. Dowd MA, Gaulden ME, Proctor BL, Seibert GB. Formaldehyde-induced acentric chromosome fragments and chromosome stickiness in Chortophaga neuroblasts. Environ Mutagen. 1986;8(3):401–11.

320. Hikiba H, Watanabe E, Barrett JC, Tsutsui T. Ability of fourteen chemical agents used in dental practice to induce chromosome aberrations in Syrian hamster embryo cells. J Pharmacol Sci. 2005 Jan;97(1):146–52.

321. Tan SLW, Chadha S, Liu Y, Gabasova E, Perera D, Ahmed K, et al. A Class of Environmental and Endogenous Toxins Induces BRCA2 Haploinsufficiency and Genome Instability. Cell. 2018 Apr 11;169(6):1105-1118.e15.

322. Levy S, Nocentini S, Billardon C. Induction of cytogenetic effects in human fibroblast cultures after exposure to formaldehyde or X-rays. Mutation Research Letters. 1983 Mar 1;119(3):309–17.

323. Schmid E, Göggelmann W, Bauchinger M. Formaldehyde-induced cytotoxic, genotoxic and mutagenic response in human lymphocytes and Salmonella typhimurium. Mutagenesis. 1986 Nov;1(6):427–31.

324. Jakab MG, Klupp T, Besenyei K, Biró A, Major J, Tompa A. Formaldehyde-induced chromosomal aberrations and apoptosis in peripheral blood lymphocytes of personnel working in pathology departments. Mutation Research/Genetic Toxicology and Environmental Mutagenesis. 2010 Apr 30;698(1):11–7.

325. Santovito A, Schilirò T, Castellano S, Cervella P, Bigatti MP, Gilli G, et al. Combined analysis of chromosomal aberrations and glutathione S-transferase M1 and T1 polymorphisms in pathologists occupationally exposed to formaldehyde. Arch Toxicol. 2011 Oct;85(10):1295–302.

326. Costa S, Carvalho S, Costa C, Coelho P, Silva S, Santos LS, et al. Increased levels of chromosomal aberrations and DNA damage in a group of workers exposed to formaldehyde. Mutagenesis. 2015 Jul 1;30(4):463–73.

327. Ghelli F, Cocchi E, Buglisi M, Squillacioti G, Bellisario V, Bono R, et al. The role of phase I, phase II, and DNA-repair gene polymorphisms in the damage induced by formaldehyde in pathologists. Scientific Reports. 2021 May 18;11(1):10507.

328. Lan Q, Smith MT, Tang X, Guo W, Vermeulen R, Ji Z, et al. Chromosome-wide aneuploidy study of cultured circulating myeloid progenitor cells from workers occupationally exposed to formaldehyde. Carcinogenesis. 2015 Jan 1;36(1):160–7.

329. Speit G, Merk O. Evaluation of mutagenic effects of formaldehyde in vitro: detection of crosslinks and mutations in mouse lymphoma cells. Mutagenesis. 2002 May 1;17(3):183–7.

330. Liu YR, Zhou Y, Qiu W, Zeng JY, Shen LL, Li AP, et al. Exposure to Formaldehyde Induces Heritable DNA Mutations in Mice. Journal of Toxicology and Environmental Health, Part A. 2009 Jun 3;72(11–12):767–73.

331. de Serres FJ, Brockman HE. Comparison of the spectra of genetic damage in formaldehyde-induced ad-3 mutations between DNA repair-proficient and -deficient heterokaryons of Neurospora crassa. Mutation Research/Reviews in Mutation Research. 1999 Sep 1;437(2):151–63.

332. Recio L, Sisk S, Pluta L, Bermudez E, Gross EA, Chen Z, et al. p53 mutations in formaldehyde-induced nasal squamous cell carcinomas in rats. Cancer Res. 1992 Nov 1;52(21):6113–6.

333. Albert RE, Sellakumar AR, Laskin S, Kuschner M, Nelson N, Snyder CA. Gaseous formaldehyde and hydrogen chloride induction of nasal cancer in the rat. J Natl Cancer Inst. 1982 Apr;68(4):597–603.

334. Swenberg JA, Kerns WD, Mitchell RI, Gralla EJ, Pavkov KL. Induction of squamous cell carcinomas of the rat nasal cavity by inhalation exposure to formaldehyde vapor. Cancer Res. 1980 Sep;40(9):3398–402.

335. Kerns WD, Pavkov KL, Donofrio DJ, Gralla EJ, Swenberg JA. Carcinogenicity of formaldehyde in rats and mice after long-term inhalation exposure. Cancer Res. 1983 Sep;43(9):4382–92.

336. Morgan KT, Jiang XZ, Starr TB, Kerns WD. More precise localization of nasal tumors associated with chronic exposure of F-344 rats to formaldehyde gas. Toxicol Appl Pharmacol. 1986 Feb;82(2):264–71.

337. Soffritti M, Maltoni C, Maffei F, Biagi R. Formaldehyde: an experimental multipotential carcinogen. Toxicol Ind Health. 1989 Oct;5(5):699–730.

338. Chang JC, Gross EA, Swenberg JA, Barrow CS. Nasal cavity deposition, histopathology, and cell proliferation after single or repeated formaldehyde exposures in B6C3F1 mice and F-344 rats. Toxicol Appl Pharmacol. 1983 Apr;68(2):161–76.

339. Sellakumar AR, Snyder CA, Solomon JJ, Albert RE. Carcinogenicity of formaldehyde and hydrogen chloride in rats. Toxicol Appl Pharmacol. 1985 Dec;81(3 Pt 1):401–6.

340. Feron VJ, Bruyntjes JP, Woutersen RA, Immel HR, Appelman LM. Nasal tumours in rats after short-term exposure to a cytotoxic concentration of formaldehyde. Cancer Letters. 1988 Feb 1;39(1):101–11.

341. Monticello TM, Swenberg JA, Gross EA, Leininger JR, Kimbell JS, Seilkop S, et al. Correlation of regional and nonlinear formaldehyde-induced nasal cancer with proliferating populations of cells. Cancer Res. 1996 Mar 1;56(5):1012–22.

342. Kamata E, Nakadate M, Uchida O, Ogawa Y, Suzuki S, Kaneko T, et al. Results of a 28-month chronic inhalation toxicity study of formaldehyde in male Fisher-344 rats. J Toxicol Sci. 1997 Aug;22(3):239–54.

343. Takahashi M, Hasegawa R, Furukawa F, Toyoda K, Sato H, Hayashi Y. Effects of ethanol, potassium metabisulfite, formaldehyde and hydrogen peroxide on gastric carcinogenesis in rats after initiation with N-methyl-N'-nitro-N-nitrosoguanidine. Jpn J Cancer Res. 1986 Feb;77(2):118–24.

344. Chen JM, Férec C, Cooper DN. Patterns and Mutational Signatures of Tandem Base Substitutions Causing Human Inherited Disease. Human Mutation. 2013 Aug 1;34(8):1119–30.

345. Kucab JE, Zou X, Morganella S, Joel M, Nanda AS, Nagy E, et al. A Compendium of Mutational Signatures of Environmental Agents. Cell [Internet]. 2019; Available from: http://www.sciencedirect.com/science/article/pii/S0092867419302636

346. Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, Wu Y, et al. The repertoire of mutational signatures in human cancer. Nature. 2020 février;578(7793):94–101.

347. Chang J, Tan W, Ling Z, Xi R, Shao M, Chen M, et al. Genomic analysis of oesophageal squamous-cell carcinoma identifies alcohol drinking-related mutation signature and genomic alterations. Nature communications. 2017/05/27 ed. 2017 May 26;8:15290.

348. Letouzé E, Shinde J, Renault V, Couchy G, Blanc JF, Tubacher E, et al. Mutational signatures reveal the dynamic interplay of risk factors and cellular processes during liver tumorigenesis. Nature Communications. 2017 Nov 3;8(1):1315.

349. Li XC, Wang MY, Yang M, Dai HJ, Zhang BF, Wang W, et al. A mutational signature associated with alcohol consumption and prognostically significantly mutated driver genes in esophageal squamous cell carcinoma. Ann Oncol. 2018;29(4):938–44.

350. Hoes L, Dok R, Verstrepen KJ, Nuyts S. Ethanol-Induced Cell Damage Can Result in the Development of Oral Tumors. Cancers. 2021;13(15).

351. Wei R, Li P, He F, Wei G, Zhou Z, Su Z, et al. Comprehensive analysis reveals distinct mutational signature and its mechanistic insights of alcohol consumption in human cancers. Briefings in Bioinformatics. 2021 May 1;22(3):bbaa066.

Chapter 3: Characterization of formaldehyde- and acetaldehyde-induced mutational signatures

3. 1 Copyright and License Policies

M. J. Thapa, R. M. Fabros, S. Alasmar, and K. Chan, "Analyses of mutational patterns induced by formaldehyde and acetaldehyde reveal similarity to a common mutational signature," *G3 Genes|Genomes|Genetics*, vol. 12, p. jkac238, Sep. 2022, doi: 10.1093/g3journal/jkac238

3.2 Rationale for research

After learning about all the experimental designs on different organisms and human tissues with their possible limitations and considering the consistency of results in the last 50 years (1), I worked on my first project to characterize the FA- and AA-induced mutational signatures. In this published research paper, “Analyses of mutational patterns induced by formaldehyde and reveal similarity to a common mutational signature” (2), a well-characterized, highly sensitive budding yeast genetic reporter system sharing DNA replication and repair mechanisms as in human cells was used to capture thousands of the independent mutations after controlled AA exposure (3). Also, experiments done by my previous labmate Reena Fabros, using FA exposure on the same yeast model system added additional confidence to carry out my experiments. The temperature sensitivity and haploid genetics of the yeast provide flexibility to change the double-stranded DNA into single-stranded DNA and produces a high volume of mutations even from weak mutagens like AA and FA. Whole genome sequencing identified all the mutations present in the yeast genome, and computational analysis was done using multiple bioinformatics tools to generate mutational signatures. These FA- and AA-induced mutational signatures were compared to COSMIC signatures (4). The differences in the FA- and AA-induced mutational patterns in our findings may fill the void highlighted in the review paper, and laid a foundation for experimental design and computational framework to derive mutational signatures from other weak mutagens.

Analyses of Mutational Patterns induced by Formaldehyde and Acetaldehyde reveal similarity to a common Mutational Signature

Running Title: Mutational patterns of small aldehydes

Authors:

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

Affiliations & addresses:

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

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

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

* Corresponding Author: kin.chan@uottawa.ca

† Equal Contribution

Current affiliation & address:

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

Author email addresses:

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

R.M.F.: rfabr010@uottawa.ca

S.A.: salas051@uottawa.ca

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

3.4 Abstract

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

3.5 Introduction

Genomic DNA is constantly damaged by intracellular processes (5) and exposure to exogenous damaging agents (6–8). There are many different types of DNA damage. Intracellular DNA damaging processes include, for example: oxidation of nitrogenous bases (9,10); glycosidic bond breakage, which releases a nitrogenous base from its deoxyribose sugar (11–15); single- and double-stranded breaks of the sugar-phosphate backbone (15–17); base alkylation (15,18,19); cytosine deamination to uracil (15,20); and deamination of 5-methylcytosine to thymine (21–23). Examples of exogenous DNA damage include: ultraviolet (UV) light (7); ionizing radiation (8); tobacco (24); aristolochic acid (25); and aflatoxin (26). Mutations are also thought to result from spontaneous ionization or isomerization (i.e., tautomerization) of DNA bases, which can alter base pairing characteristics (27–30).

It is important to note that these processes do not affect all bases equally. Each of the four nitrogenous bases has its own distinct set of chemically reactive moieties (e.g, amines, carbonyls, or labile ring atoms) (31). 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 (32).

Mutational signatures typically are inferred using a non-negative matrix factorization (NMF) algorithm (33). NMF takes a mutational dataset as input. It initiates by essentially guessing a solution set of constituent signatures with estimated contributions from each putative signature, and then computes the error when attempting to reconstruct the original dataset using that solution set. NMF then tries a slightly different solution set and recomputes the error. This process loops until finding an optimal solution set that stably minimizes reconstruction error. A globally stable solution set is found when different initial conditions all converge to yield that solution set.

NMF analysis can extract reproducible, recurrent patterns of mutations, which often reflect distinct mutagenic processes or DNA damaging agents. There are many mutational signatures with well-established etiologies, including: single base substitution signature 1 (SBS1) from deamination of 5-methylcytosine at CpG motifs; SBS2 and SBS13 from enzymatic deamination of cytosine at TC motifs by APOBEC deaminases; SBS3 from deficiencies in homologous recombination DNA repair; SBS4 and SBS29 from tobacco smoking and chewing habits, respectively; SBS6, SBS15, SBS21, SBS26, and SBS44 from various deficiencies in DNA mismatch repair; SBS7 from ultraviolet light exposure; SBS10 from mutation of DNA polymerase epsilon; SBS18 from reactive oxygen species; SBS30 and SBS36 from DNA base excision repair deficiencies; and so forth (32). About one-third of currently defined mutational signatures remain of unknown etiology (32).

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 (34). 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 (35,36). 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 (36). 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 (37). Alcohol consumption is associated with higher risk of multiple types of cancer, including: head and neck; oesophageal; liver; breast; and colorectal (38). Acetaldehyde associated with alcohol consumption is thought to be causative for cancers of the oesophagus and the upper aerodigestive tract (including head and neck), i.e., at sites of highest direct exposure (38).

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., (39,40)) 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.

3.6 Materials and Methods

3.6.1 Reagents and Consumables

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

3.6.2 Yeast Genetics and Mutagenesis

Mutagenesis experiments used the ySR127 yeast strain, a *MATα* 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 (3) and the strain is available upon request.

Formaldehyde mutagenesis experiments were initiated by inoculating single colonies separately into 5 mL of YPDA rich media (2% Bacto peptone, 1% Bacto yeast extract, 2% glucose, supplemented with 0.001% adenine sulfate) in round bottom glass tubes. Cells were grown at

permissive temperature (23°C) for three days. Then, cultures were diluted ten-fold into fresh media in gas-tight glass tubes, shifted to restrictive temperature (37°C), and shaken gently at 150 RPM for three hours, with syringe needles inserted through the septa to enable gas exchange. After a three-hour temperature shift, aliquots of formaldehyde stock solution diluted in media were injected into each tube to obtain the reported final concentrations. Samples were then shaken gently at 150 RPM at 37°C for three more hours, in completely sealed gas-tight tubes, to prevent escape of formaldehyde. When formaldehyde treatment was complete, cells were collected by syringe, lightly centrifuged, washed in water, and plated (using a turntable and cell spreader) onto synthetic complete media to assess survival and onto canavanine-containing media with 0.33x adenine to select for mutants (Can^r colonies were off-white while Can^r Ade^- colonies turn red or pink). Care was taken to handle cells gently throughout, as they were quite fragile. Further details of this plating procedure were described in detail previously (41).

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 (42) and tidyverse package version 1.3 (43).

3.6.3 Illumina Whole Genome Sequencing and Data Analyses

Can^r Ade^- mutants from formaldehyde and acetaldehyde treatment experiments were collected and reporter gene loss of function phenotypes were verified as described previously (3). Briefly, Can^r red/pink mutants were streaked on YPDA plates. A single colony from each streak was patched onto YPDA. Patches were then replica plated onto glycerol, adenine dropout, canavanine, and uracil dropout media. Mutants that grew on glycerol (i.e., were respiration competent), and were Can^r Ade^- Ura^+ were considered suitable for sequencing. Can^r Ade^- Ura^- mutants were avoided because those isolates sometimes turn out to be telomere truncations. Mutants from 4, 6, 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-hour 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. 24-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 (44), SAMtools 1.9 (45), and bcftools 1.9 (46) 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 (47). 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 --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 were 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 one unique variant. The resulting unified VCF files were passed to MutationalPatterns version 3.6.3 (48) for further analysis and visualization. Other numerical and statistical analyses, and data visualizations were done using base R version 4.1 (42) and tidyverse package version 1.3 (43).

For trinucleotide frequency correction, the Biostrings package version 2.38.0 (49) 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.

3.7 Results

3.7.1 Formaldehyde- and acetaldehyde-induced mutagenesis

We began by assessing mutagenesis and toxicity induced by addition of formaldehyde or acetaldehyde. These experiments were done using a haploid yeast strain (ySR127) that forms long regions of sub-telomeric single-stranded DNA (ssDNA) when shifted to 37°C due to the *cdc13-1* temperature sensitive point mutation (50). At 37°C the cdc13-1 protein dissociates from telomeres, triggering enzymatic resection of unprotected chromosome ends, which in turn activates the DNA damage checkpoint to arrest cells in G_2 (50). 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 (3). 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 (51). The ssDNA system was used previously to study the mutagenic properties of bisulfite and human APOBEC3G cytidine deaminase (3); abasic sites (52); reactive oxygen species (53); human APOBEC3A and APOBEC3B cytidine deaminases (47); and alkylating agents (54).

We treated temperature-shifted cells with increasing concentrations of formaldehyde or acetaldehyde. Care was taken to seal the formaldehyde-treated samples in gas-tight tubes; otherwise, the formaldehyde would simply volatilize into the gaseous phase and escape into the atmosphere. Increasing concentrations of formaldehyde resulted in lower viability (see **Figure 3.1A**). While lower concentrations are relatively well tolerated, 8 mM formaldehyde reduced viability below 50%. Formaldehyde-induced inactivation of *CAN1* was detected from as little as 2 mM treatment (median gene inactivation frequency of 3.3×10^{-4}, see **Figure 3.1B**). Mutagenesis plateaued from 4 to 8 mM formaldehyde, with median mutation frequencies of ~1.5×10^{-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 **Figure 3.1C**). Unlike formaldehyde, the mutagenesis induced by acetaldehyde did not show a plateau. Instead, here was a gradual increase in *CAN1* inactivation frequency when treated with 25 and 50 mM acetaldehyde (see **Figure 3.1D**). Mutation frequency peaked at over 5×10^{-4} when cells were treated with 75 mM acetaldehyde. Interestingly, treatment with 100 mM acetaldehyde did not result in detectable mutagenesis while viability was reduced to below 25%. This

suggests that the cells which sustained high levels of DNA damage by 100 mM acetaldehyde likely suffered considerable cytotoxic damage as well and did not survive.

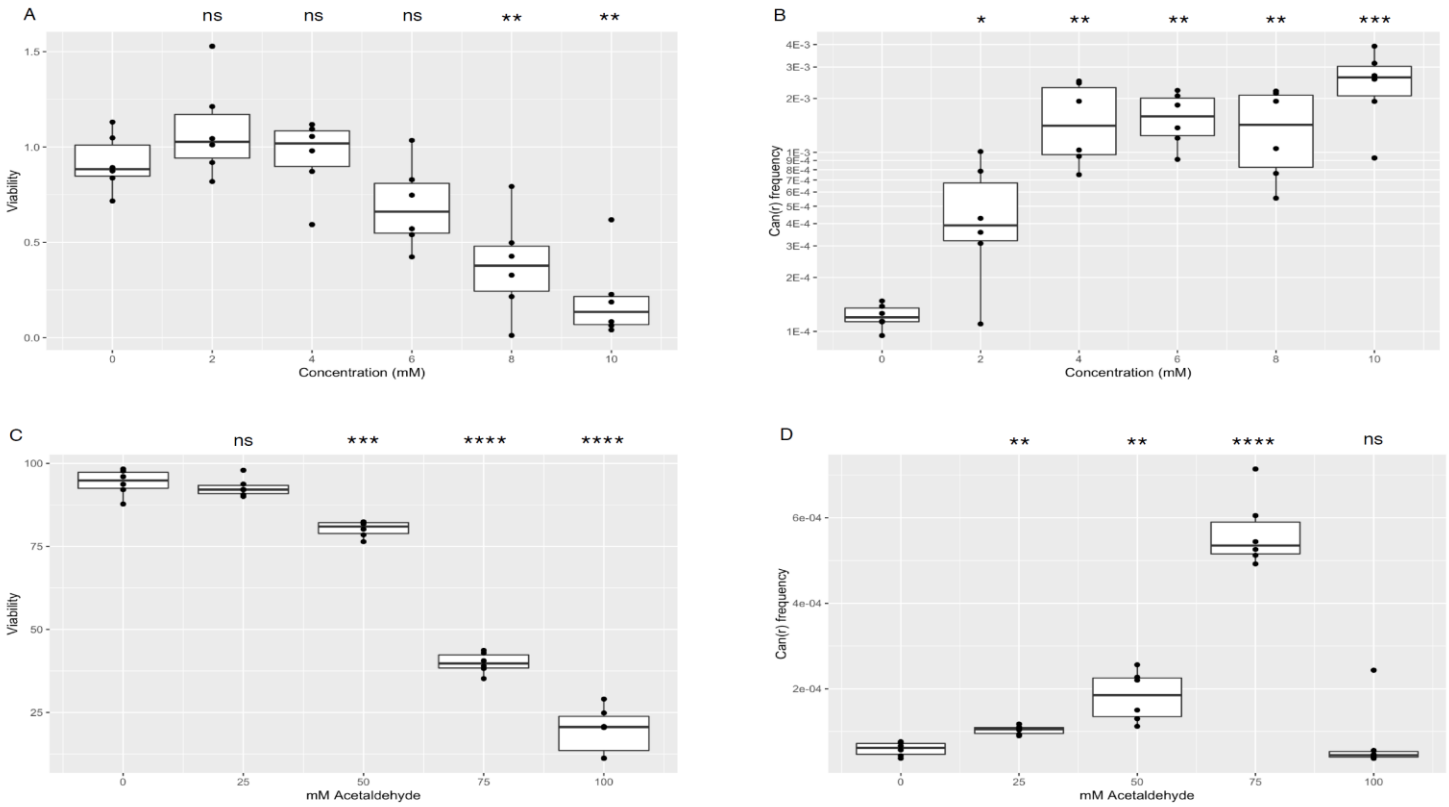


Figure 3.1 Viability and *CAN1* inactivation of yeast treated with AA and FA at different concentration
a) Viability and b) *CAN1* inactivation frequency of yeast treated with 0, 2, 4-, 6-, 8-, or 10-mM formaldehyde. c) Viability and d) *CAN1* inactivation frequency of yeast treated with 0, 25-, 50-, 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. The paired t-test was used because same biological replicate was used for untreated and treated samples.

3.7.2 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 genetic variants were induced by either formaldehyde (119 genomes) or acetaldehyde (17 genomes) treatment. 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 four types of transversions (namely C/G > A/T, C/G > G/C, T/A > A/T, and T/A > G/C, see **Figure 3.2A**), similar to what we had observed previously (55). By comparison, formaldehyde and acetaldehyde treatment both caused a relative increase of C/G > A/T transversions (see **Figure 3.2B and 3.2C**). While these substitutions accounted for 11% of the mutational spectrum in untreated controls, this fraction rose to about 17% in the aldehyde-mutagenized genomes. This increase is a common characteristic of mutagenesis caused by small aldehydes in regions of single-stranded DNA. 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 **Figure 3.3D, 3.3E, and 3.3F**, respectively. These graphs

show the number of base pairs between adjacent mutations. Consistent with expectation, variants tended to cluster near chromosome ends.

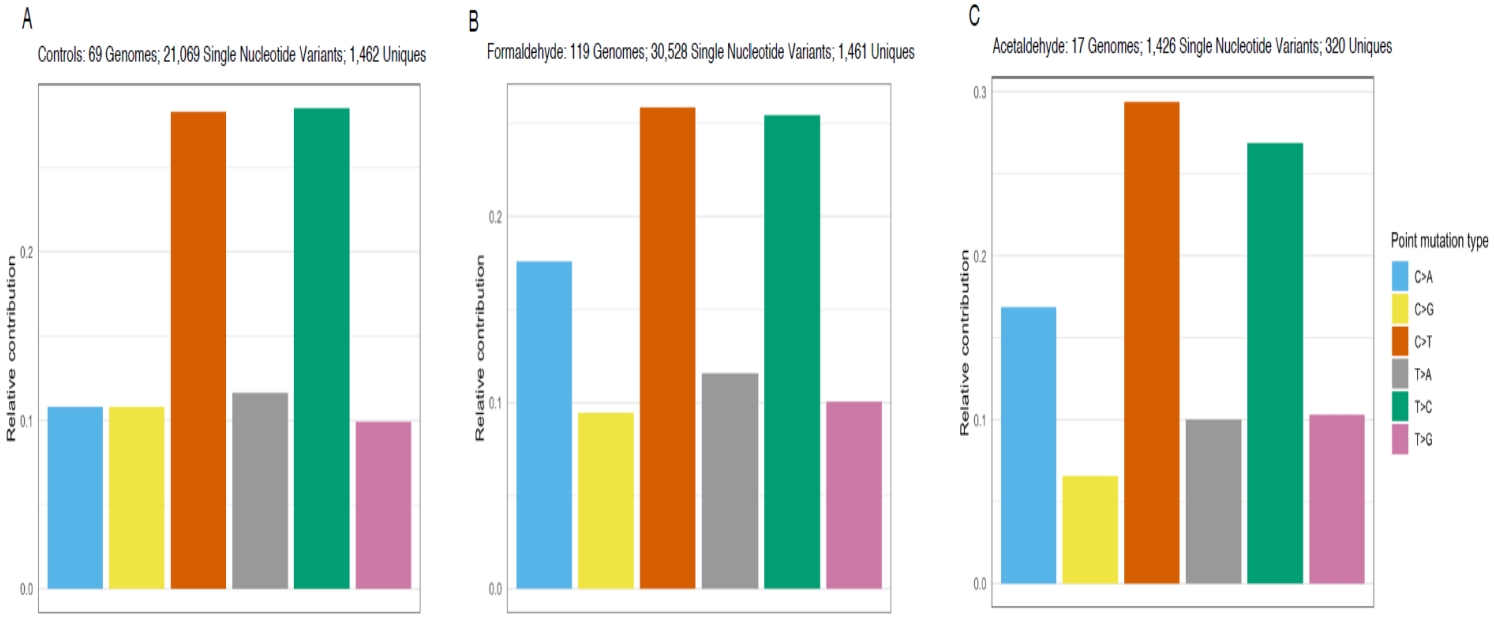


Figure 3.2: Base substitution types for (a) controls, (b) formaldehyde, and (c) acetaldehyde. Treatment with either aldehyde caused a higher proportion of C/G > A/T transversions. Six biological replicates was used. Chi-Square test showed that the mutation spectrum differs significantly between groups ($\chi^2 = 351.8$, p-value < 2.2e-16). Both aldehydes AA and FA alter the C>A substitution spectrum higher compared to the control.

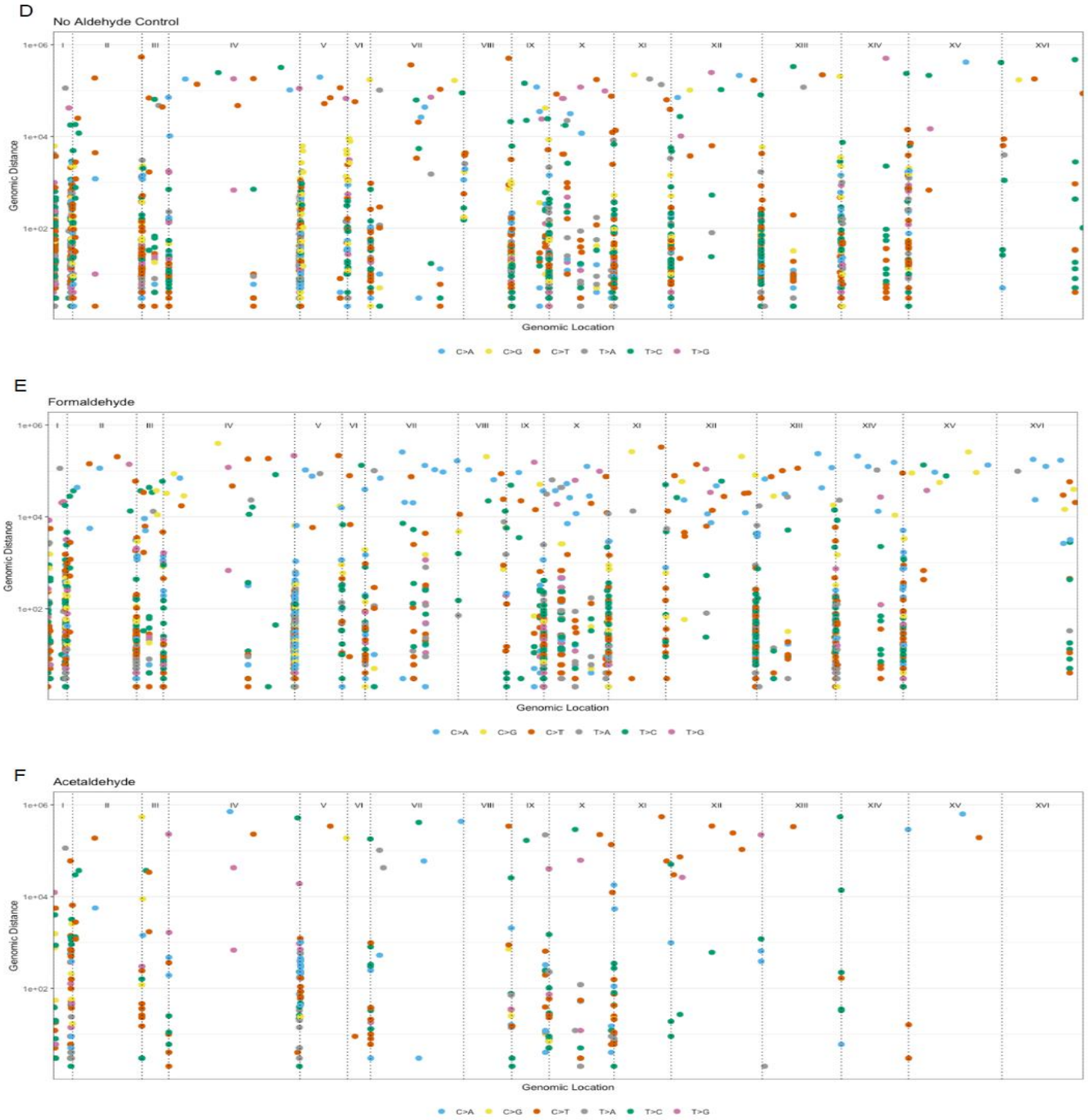


Figure 3.3: Rainfall plots
Rainfall plots showing distance between adjacent mutations, show that most cluster near chromosome ends where ssDNA is enriched, for (d) controls, (e) formaldehyde, and (f) acetaldehyde. Total numbers of sequenced genomes, total numbers of variant calls, and number of unique variants are reported (if the same variant occurs in multiple samples, it is counted as 1 unique).

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

We also analyzed short insertions and deletions (indels) to determine if treatment with either small aldehyde can induce these genetic changes. The profile of short indels in untreated controls consists mainly of insertions of five or more bases, with smaller proportions of shorter insertions as well as deletions of five or more bases (see **Figure 3.4A**). The profile of short indels in formaldehyde-mutated genomes is essentially the same as in untreated control genomes, i.e., we did not find evidence that formaldehyde induces a higher proportion of any type of indels (see **Figure 3.4B**). In contrast, there was a notable difference in the acetaldehyde-induced profile of indels: an excess of deletions of five or more bases was observed (24% in acetaldehyde vs. 12% in controls, see **Figure 3.4C**). 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 **Figures 3.5D, 3.5E, and 3.5F**). 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.

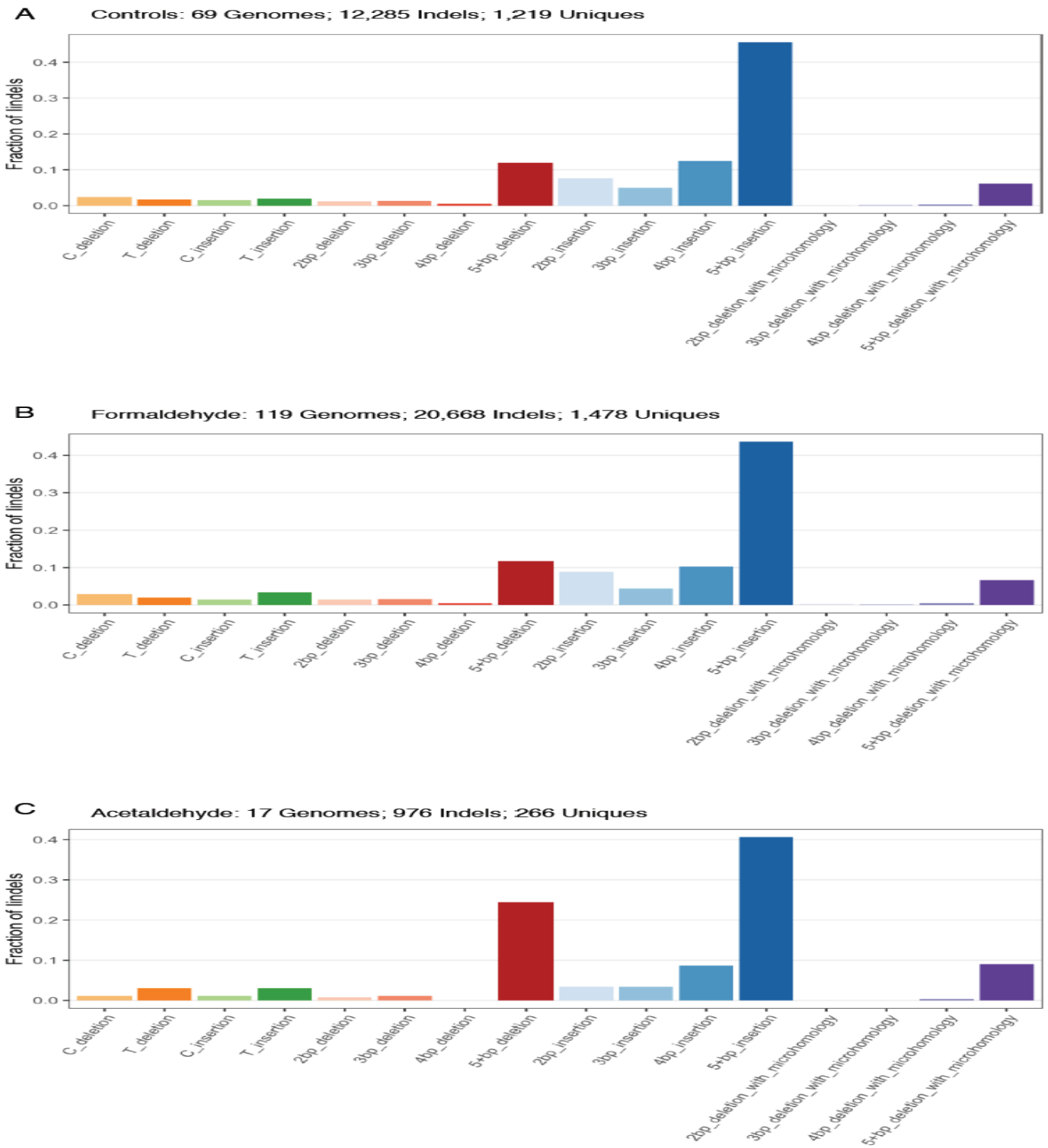


Figure 3.4: Small indels from (a) no-aldehyde controls, (b) formaldehyde, and (c) acetaldehyde.
The different categories comprise: single base deletions or insertions at C/G or T/A base pairs; 2, 3, 4, or 5+ base pair deletions or insertions; and 2, 3, 4, or 5+ base pair deletions with microhomology at break points. Acetaldehyde treatment induces an increased proportion of 5+ base pair deletions (without microhomology).

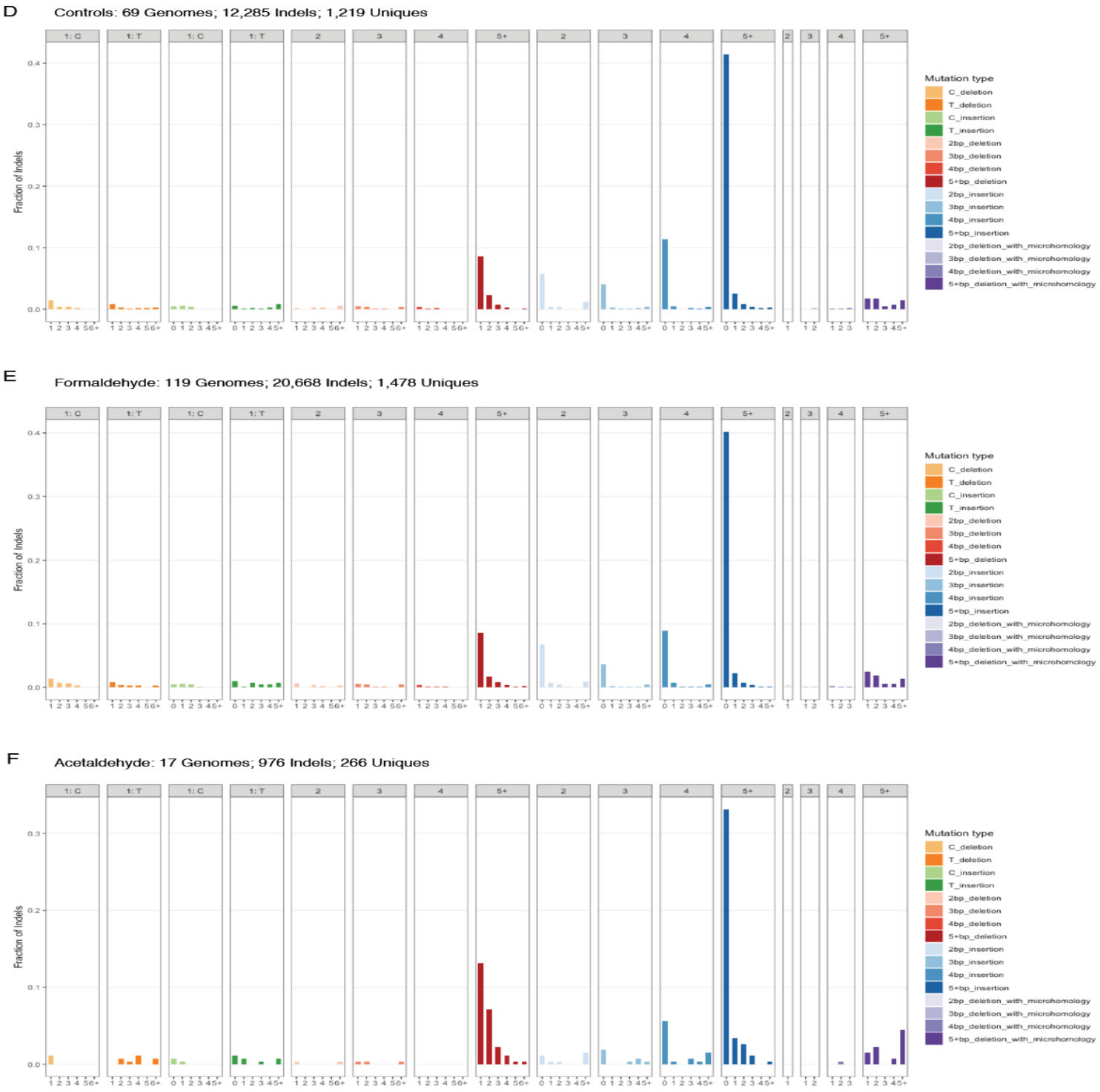


Figure 3.5: The same small indel data, plotted showing number of repeat units from (d) no-aldehyde controls, (e) formaldehyde, and (f) acetaldehyde.
For the single-nucleotide indels, the number of repeat units is the length of a homopolymer run. For indels of dinucleotide, trinucleotide, or greater length, 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).

3.7.4 Formaldehyde and acetaldehyde produce distinct mutational patterns

To investigate the mutational properties of the small aldehydes in more detail, we plotted their mutational profiles in the 96-channel format of the COSMIC mutational signatures. By this convention, all substitutions are reported as originating from a pyrimidine base, i.e., same as the mutation spectra reported above. In addition, the 96-channel profiling features trinucleotide motifs consisting of the mutated base, flanked by an adjacent base 5' and 3'. Cosine similarity is a metric for comparing mutational patterns, yielding a maximum value of exactly 1 for two identical patterns (33). The mutational pattern of formaldehyde is similar to untreated controls (cosine similarity = 0.93), but the excess of C/G > A/T transversions is nonetheless evident (see **Figure 3.6A**). The mutational pattern of acetaldehyde is more dissimilar vs. the profile of untreated controls (cosine similarity = 0.868), but again with a noticeable excess of C/G > A/T substitutions (see **Figure 3.6B**). When comparing the formaldehyde and acetaldehyde profiles directly to one another, the cosine similarity value is 0.882, showing some similarities but also clear differences in the C/G > T/A channels especially (see **Figure 3.6C**).

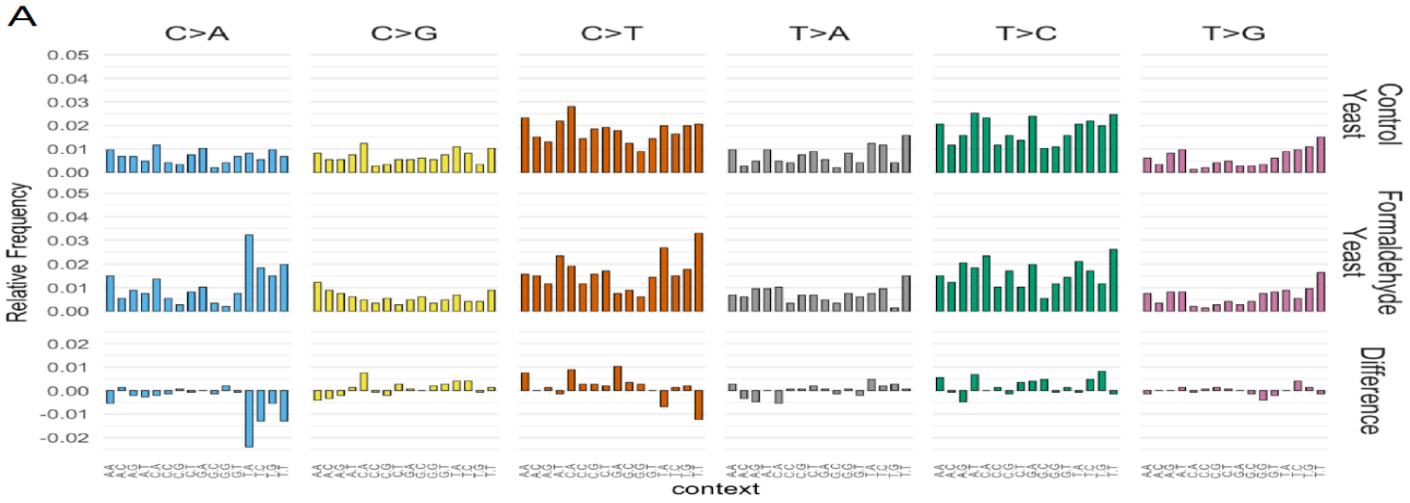


Figure 3.6: Comparisons of mutational patterns between (a) controls and formaldehyde; (b) controls and acetaldehyde; and (c) formaldehyde and acetaldehyde.

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 (40). To compare our mutational patterns derived from mutagenized yeast genomes to these profiles from mice, we first adjusted for differences in trinucleotide abundances between the two species to obtained corrected mutational patterns (see **Figure 3.7A and 3.7B**). 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).

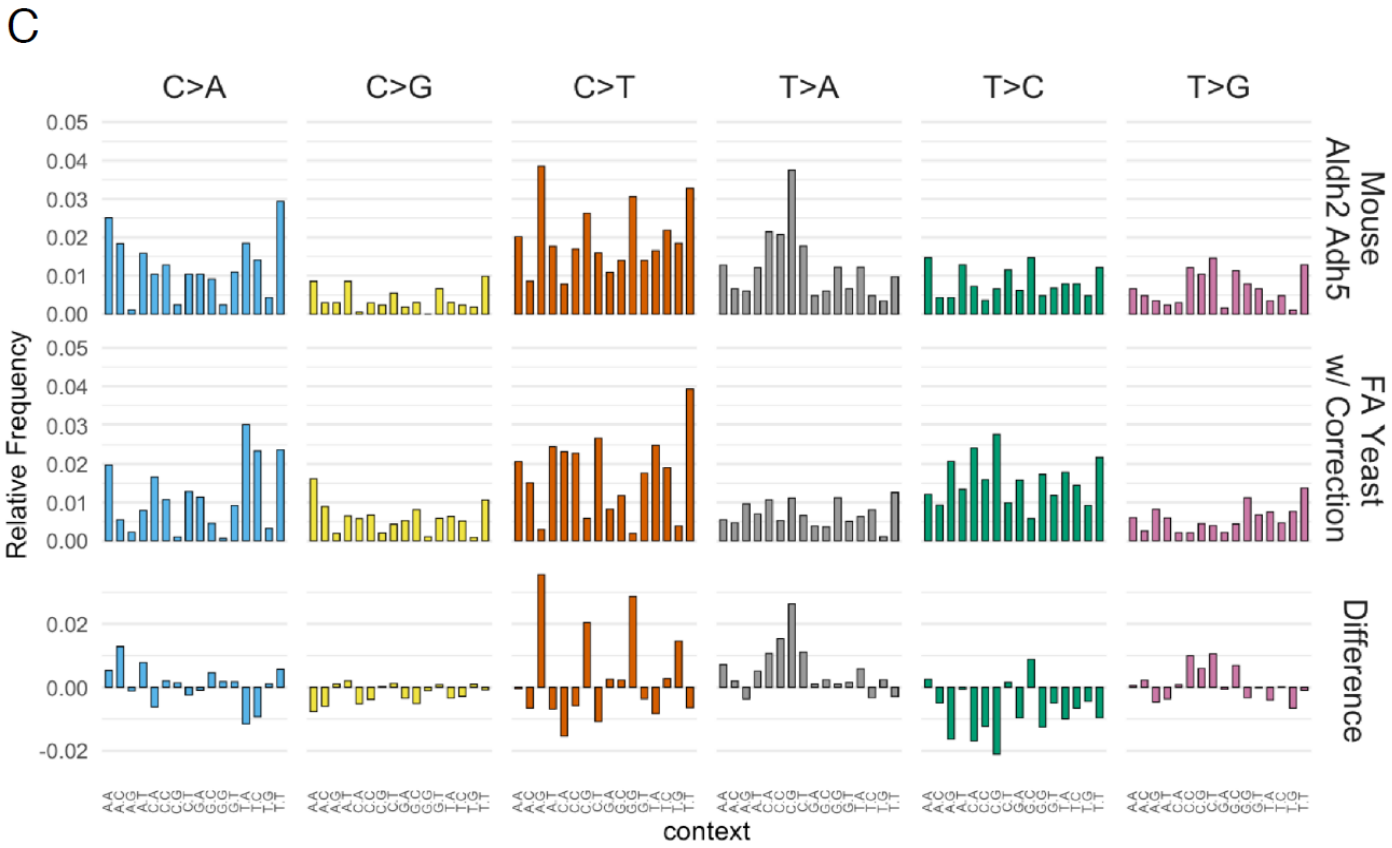


Figure 3.7: Comparisons of mutational patterns between formaldehyde and acetaldehyde treated yeast with and without correction of trinucleotides with Aldh2 Adh5-deficient mouse cells
Comparisons of mutational patterns between (a) formaldehyde with and without correction for trinucleotide frequencies in yeast; (b) acetaldehyde with and without correction for trinucleotide frequencies in yeast; and (c) Aldh2 Adh5-deficient mouse cells and trinucleotide frequency corrected yeast treated with formaldehyde

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. (40) (see **Table 3.1**). 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 **Figure 3.7C**). We also noted some differences among the various mouse patterns themselves: while the ones from Adh5$^{-/-}$ and Aldh2$^{-/-}$ Adh5$^{-/-}$ had cosine similarity =

0.887, the Aldh2$^{-/-}$ profile was noticeably more dissimilar (cosine similarity < 0.8 vs. the other two profiles, see **Table 3.2**). 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.

Table 3.1: Cosine similarity values

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 (40).

	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

Table 3.2: Cosine similarity values among mice deficient for aldehyde detoxification genes from (40).

	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

3.7.5 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. (40) and confirmed that none of the mouse profiles showed a particularly close resemblance to any known COSMIC signature (see **Figure 3.8A**). All of those cosine similarity values were < 0.8, suggesting that if there are bona fide COSMIC signatures within the mouse mutational patterns, they are possibly obscured by being in a mixture of multiple signatures. Comparison of the corrected acetaldehyde pattern from yeast vs. known COSMIC signatures also yielded, at best, cosine similarity of 0.79 to SBS40, a signature of unknown etiology (see **Figure 3.8A**). Since we had better direct control of the induced mutagenesis experiments using the yeast system with exogenously applied mutagen, it does not seem as likely that other mutagenic processes are obscuring the acetaldehyde-induced pattern. We conclude that the acetaldehyde pattern we obtained is not a plausible match for any known COSMIC signature at this point.

Finally, we compared the formaldehyde pattern to the COSMIC signatures, finding that the closest match is to SBS40, with cosine similarity = 0.9 (see **Figures 3.8A and 3.9A**). The second closest match was to SBS5 (cosine similarity = 0.864, see **Figures 3.8A and 3.9B**). 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 **Figure 3.8A**). 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 (55). 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.

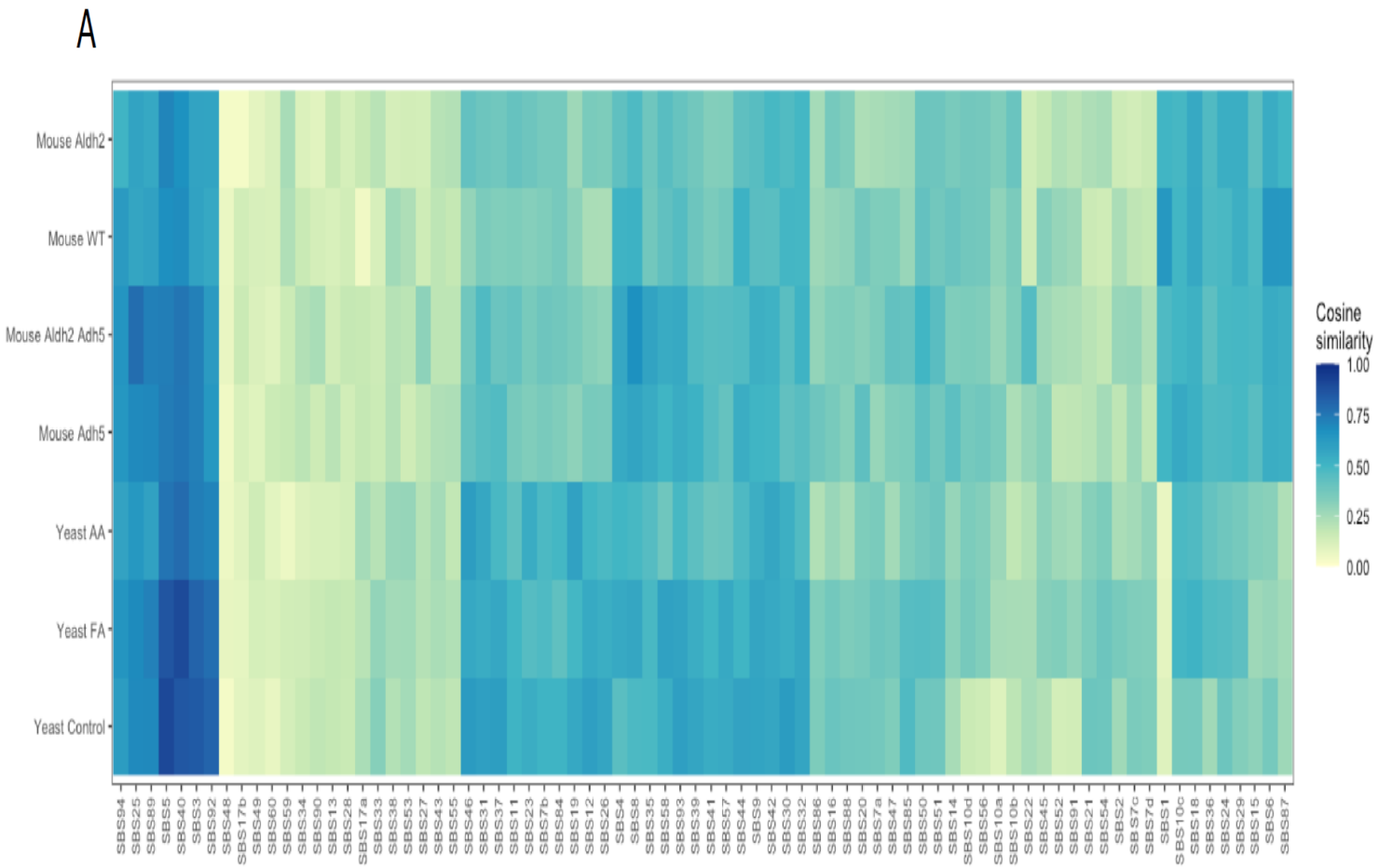


Figure 3.8: Cosine similarity heatmap
a) Cosine similarity heatmap with hierarchical clustering comparing COSMIC SBS signatures vs. mouse mutational patterns from Dingler et al.; and vs. trinucleotide abundance-corrected mutational patterns in yeast from acetaldehyde, formaldehyde, and no-aldehyde control.

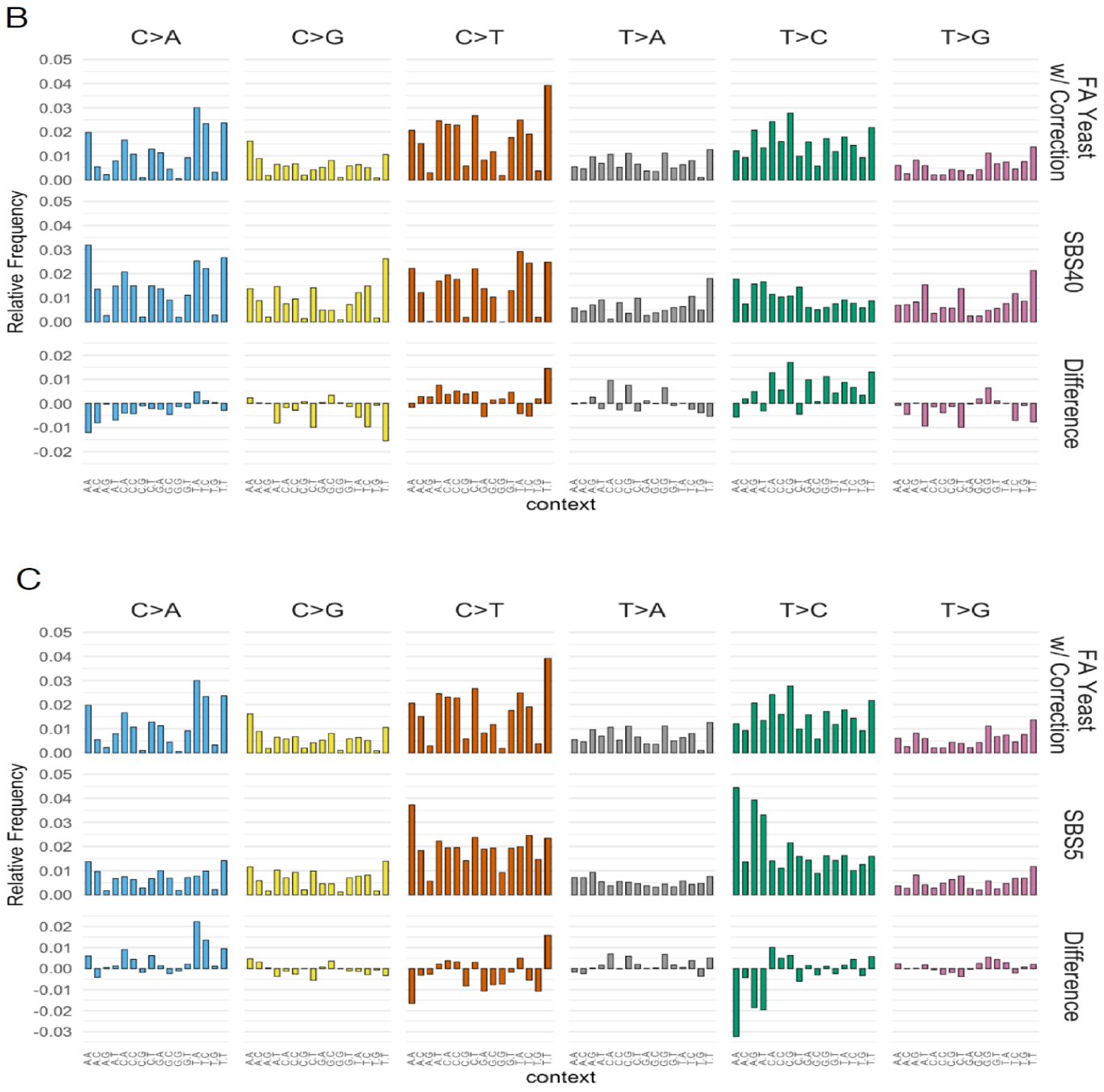


Figure 3.9: Comparison of corrected formaldehyde mutational pattern in yeast vs SBS5 and SBS40
a) Comparison of corrected formaldehyde mutational pattern in yeast vs. SBS signature 40. b) Comparison of corrected formaldehyde mutational pattern in yeast vs. SBS signature 5.

3.8 Discussion

In this paper, we report the use of a sensitive ssDNA-based mutagenesis reporter system to characterize the mutagenic properties of two small aldehydes, formaldehyde and acetaldehyde. This system is especially well suited for investigating chemical agents with relatively weak mutagenicity. A challenge of using conventional mutagenesis systems to study weak mutagens is that induced mutations can be rare and it can be difficult to discern a reliable mutational pattern using relatively few mutations (39,40). In addition to being a more sensitive reporter system, it is also considerably more cost-efficient to sequence compact yeast genomes (each ~12 megabases) than mammalian genomes which are much larger (~3 gigabases). By applying a correction to account for different abundances of trinucleotide motifs, we can use data from the sequencing of mutagenized yeast to infer the expected mutational pattern in another species. Another key advantage is the single-stranded configuration of the DNA precludes repair that requires a complementary strand. By sidestepping intervention from DNA repair processes, the ssDNA system can provide, in effect, a purer readout of the effects of mutagenesis per se. Leveraging these advantages of the ssDNA mutagenesis reporter system, we were able to infer the mutational patterns of both formaldehyde and acetaldehyde. 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 (56). 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 trade-off that haploids are not as buffered against potentially deleterious variants as diploids.

The two small aldehydes share some similar mutagenicity characteristics, but also have their differences. Both induce dose-dependent increases in mutagenesis at lower concentrations. But whereas formaldehyde-induced mutagenesis essentially plateaus from 4 mM up to 8 mM, acetaldehyde-induced mutagenesis peaks at 75 mM and then drops sharply at the even higher concentration of 100 mM. Both aldehydes induce significant cytotoxicity at the higher end of their respective ranges of tested concentrations, but yeast are able to tolerate considerably higher doses of acetaldehyde overall. Yeast are presumably evolved to cope with significantly higher concentrations of acetaldehyde, since it is an abundant intermediate in ethanol production from fermentation (57). 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 (58–64). Interestingly, acetaldehyde induces an excess of deletion variants of five or more bases in our system, but formaldehyde does not, consistent with previous reports (58,65). 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 two compounds. Whereas the acetaldehyde pattern did not particularly resemble any known COSMIC signature, new mutational signatures will be revealed as more cancer samples are sequenced and analyzed. Since alcohol consumption is associated with multiple cancer types and it is thought that the acetaldehyde from alcohol detoxification would surely damage DNA (38), 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 (32), 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 (36), it would fit this profile. When all of the available information is taken into consideration, mutagenesis from endogenously generated formaldehyde emerges as a plausible candidate for the etiology of SBS40.

Comparison with mutational patterns from mice deleted for aldehyde detoxification genes suggest that those profiles are likely mixtures of mutations from different mutagenic processes, and not just from DNA damage due to accumulation of excess endogenous aldehydes. For example, the contribution from SBS1 (C/G > T/A at CpG motifs) was quite noticeable. Mutagenesis from other sources likely interferes with making an accurate inference of the

aldehyde-associated mutagenesis. This is perhaps another significant challenge when using systems for mutational detection that are not (and maybe can not) be properly controlled to factor out mutagenesis from other sources. Deployment of more specialized and sensitive mutagenesis detection systems where the experimenters have more direct control over the mutation induction can continue to play an important role in shining new light on mutagenesis.

3.9 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 Supplemental Table 1. The ySR127 reference genome is available on NCBI Assembly (accession GCA_001051215.1).

3.10 Acknowledgments

We thank S. Gelova, B. Xhialli, and N. Liang for their critical feedback on this work. Author contributions are as follows: All authors performed experiments, and analyzed/discussed the data; K.C. conceived and supervised the study, and wrote the draft manuscript.

3.11 Funding

The authors gratefully acknowledge funding support from a Tier 2 Canada Research Chair, an NSERC Discovery Grant, an Ontario Early Researcher Award, and uOttawa startup funding to K.C.

3.12 Competing Interests

The authors declare there are no competing interests.

3.13 References

1. Thapa MJ, Chan K. The mutagenic properties of formaldehyde and acetaldehyde: Reflections on half a century of progress. Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis. 2025 Jan 1;830:111886.

2. Thapa et al. Analyses of mutational patterns induced by formaldehyde and acetaldehyde reveal similarity to a common mutational signature. G3 Genes|Genomes|Genetics. 2022;12(11):jkac238.

3. 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 Dec 13;8(12):e1003149.

4. COSMIC. Catalogue Of Somatic Mutations In cancer. 2021.

5. De Bont R, van Larebeke N. Endogenous DNA damage in humans: a review of quantitative data. Mutagenesis. 2004 May 1;19(3):169–85.

6. Irigaray P, Belpomme D. Basic properties and molecular mechanisms of exogenous chemical carcinogens. Carcinogenesis. 2010 Feb 1;31(2):135–48.

7. Ikehata H, Ono T. The Mechanisms of UV Mutagenesis. Journal of Radiation Research. 2011 Mar 1;52(2):115–25.

8. Keszenman DJ, Kolodiuk L, Baulch JE. DNA damage in cells exhibiting radiation-induced genomic instability. Mutagenesis. 2015 May 1;30(3):451–8.

9. Ames BN, Shigenaga MK, Hagen TM. Oxidants, antioxidants, and the degenerative diseases of aging. Proc Natl Acad Sci USA. 1993 Sep 1;90(17):7915.

10. Helbock HJ, Beckman KB, Shigenaga MK, Walter PB, Woodall AA, Yeo HC, et al. DNA oxidation matters: the HPLC-electrochemical detection assay of 8-oxo-deoxyguanosine and 8-oxo-guanine. Proceedings of the National Academy of Sciences of the United States of America. 1998 Jan 6;95(1):288–93.

11. Lindahl T, Nyberg B. Rate of depurination of native deoxyribonucleic acid. Biochemistry. 1972 Sep 12;11(19):3610–8.

12. Lindahl T. Instability and decay of the primary structure of DNA. Nature. 1993 Apr 22;362(6422):709–15.

13. Nakamura J, Walker VE, Upton PB, Chiang SY, Kow YW, Swenberg JA. Highly Sensitive Apurinic/Apyrimidinic Site Assay Can Detect Spontaneous and Chemically Induced Depurination under Physiological Conditions. Cancer Res. 1998 Jan 15;58(2):222.

14. Lindahl T. DNA repair enzymes acting on spontaneous lesions in DNA. In: DNA Repair Processes. Miami: Symposia Specialists; 1977.

15. Tice RR, Setlow RB. DNA repair and replication in aging organisms and cells. In: Handbook of the Biology of Aging. New York: Van Nostrand Reinhold; 1985.

16. Haber JE. DNA recombination: the replication connection. Trends in Biochemical Sciences. 1999 Jul 1;24(7):271–5.

17. Vilenchik MM, Knudson AG. Endogenous DNA double-strand breaks: Production, fidelity of repair, and induction of cancer. Proc Natl Acad Sci USA. 2003 Oct 28;100(22):12871.

18. Kadlubar FF, Anderson KE, Häussermann S, Lang NP, Barone GW, Thompson PA, et al. Comparison of DNA adduct levels associated with oxidative stress in human pancreas. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis. 1998 Sep 20;405(2):125–33.

19. 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 USA. 2003 Nov 25;100(24):14247.

20. Saparbaev MK, Zharkov DO. Glycosylase Repair. In: Reference Module in Life Sciences [Internet]. Elsevier; 2017. Available from: http://www.sciencedirect.com/science/article/pii/B9780128096338064815

21. Greenblatt MS, Bennett WP, Hollstein M, Harris CC. Mutations in the p53 tumor suppressor gene: clues to cancer etiology and molecular pathogenesis. Cancer research. 1994;54(18):4855–78.

22. Neddermann P, Gallinari P, Lettieri T, Schmid D, Truong O, Hsuan JJ, et al. Cloning and Expression of Human G/T Mismatch-specific Thymine-DNA Glycosylase. Journal of Biological Chemistry. 1996 May 31;271(22):12767–74.

23. Sassa A, Kanemaru Y, Kamoshita N, Honma M, Yasui M. Mutagenic consequences of cytosine alterations site-specifically embedded in the human genome. Genes and Environment. 2016 Sep 1;38(1):17.

24. Alexandrov LB, Ju YS, Haase K, Van Loo P, Martincorena I, Nik-Zainal S, et al. Mutational signatures associated with tobacco smoking in human cancer. Science. 2016 Nov 4;354(6312):618.

25. Moriya M, Slade N, Brdar B, Medverec Z, Tomic K, Jelaković B, et al. TP53 Mutational signature for aristolochic acid: an environmental carcinogen. International Journal of Cancer. 2011 Mar 16;129(6):1532–6.

26. Letouzé E, Shinde J, Renault V, Couchy G, Blanc JF, Tubacher E, et al. Mutational signatures reveal the dynamic interplay of risk factors and cellular processes during liver tumorigenesis. Nature Communications. 2017 Nov 3;8(1):1315.

27. 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. Journal of Computational Chemistry. 1998 Jul 15;19(9):989–1000.

28. 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 Aug 1;104(31):7346–52.

29. Masoodi HR, Bagheri S, Abareghi M. The effects of tautomerization and protonation on the adenine–cytosine mismatches: a density functional theory study. Journal of Biomolecular Structure and Dynamics. 2016 Jun 2;34(6):1143–55.

30. Kimsey IJ, Szymanski ES, Zahurancik WJ, Shakya A, Xue Y, Chu CC, et al. Dynamic basis for dG•dT misincorporation via tautomerization and ionization. Nature. 2018 Jan 31;554:195.

31. 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.

32. Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, Wu Y, et al. The repertoire of mutational signatures in human cancer. Nature. 2020 Feb 1;578(7793):94–101.

33. Alexandrov LB, Nik-Zainal S, Wedge DC, Campbell PJ, Stratton MR. Deciphering Signatures of Mutational Processes Operative in Human Cancer. Cell Reports. 2013 Jan 31;3(1):246–59.

34. International Agency for Research on Cancer. Identification of research needs to resolve the carcinogenicity of high-priority IARC carcinogens. 2010. (IARC Technical Publications; vol. 42).

35. International Agency for Research on Cancer. Formaldehyde. In: Chemical Agents and Related Occupations. Lyons, France; 2012. p. 401–36. (IARC Monographs on the Evaluation of Carcinogenic Risks to Humans; vol. 100F).

36. National Toxicology Program. Formaldehyde. In: Report on Carcinogens, Thirteenth Edition. Research Triangle Park, NC: U.S. Department of Health and Human Services, Public Health Service; 2014.

37. Secretan B, Straif K, Baan R, Grosse Y, El Ghissassi F, Bouvard V, et al. A review of human carcinogens--Part E: tobacco, areca nut, alcohol, coal smoke, and salted fish. Lancet Oncol. 2009 Nov;10(11):1033–4.

38. International Agency for Research on Cancer. Personal Habits and Indoor Combustions. 2012. (IARC Monographs on the Evaluation of Carcinogenic Risks to Humans; vol. 100E).

39. Kucab JE, Zou X, Morganella S, Joel M, Nanda AS, Nagy E, et al. A Compendium of Mutational Signatures of Environmental Agents. Cell [Internet]. 2019; Available from: http://www.sciencedirect.com/science/article/pii/S0092867419302636

40. Dingler FA, Wang M, Mu A, Millington CL, Oberbeck N, Watcham S, et al. Two Aldehyde Clearance Systems Are Essential to Prevent Lethal Formaldehyde Accumulation in Mice and Humans. Molecular Cell. 2020 Dec 17;80(6):996-1012.e9.

41. Chan K. Molecular Genetic Characterization of Mutagenesis Using a Highly Sensitive Single-Stranded DNA Reporter System in Budding Yeast. In: Muzi-Falconi M, Brown GW, editors.

Genome Instability: Methods and Protocols [Internet]. New York, NY: Springer New York; 2018. p. 33–42. Available from: https://doi.org/10.1007/978-1-4939-7306-4_4

42. R Core Team. R: The R Project for Statistical Computing [Internet]. 2020 [cited 2020 Mar 11]. Available from: https://www.r-project.org/

43. Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, et al. Welcome to the tidyverse. J Open Source Softw. 2019 Nov 21;4(43):1686.

44. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Meth. 2012 Apr;9(4):357–9.

45. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078–9.

46. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011 Nov 1;27(21):2987–93.

47. Chan K, Roberts SA, Klimczak LJ, Sterling JF, Saini N, Malc EP, et al. An APOBEC3A hypermutation signature is distinguishable from the signature of background mutagenesis by APOBEC3B in human cancers. Nat Genet. 2015 Sep;47(9):1067–72.

48. Blokzijl F, Janssen R, van Boxtel R, Cuppen E. MutationalPatterns: comprehensive genome-wide analysis of mutational processes. Genome Medicine. 2018 Apr 25;10(1):33.

49. Pagès H, Aboyoun P, Gentleman R, DebRoy S. Biostrings: Efficient manipulation of biological strings [Internet]. 2022. Available from: https://bioconductor.org/packages/Biostrings

50. Garvik B, Carson M, Hartwell L. Single-stranded DNA arising at telomeres in cdc13 mutants may constitute a specific signal for the RAD9 checkpoint. Molecular and Cellular Biology. 1995 Nov 1;15(11):6128–38.

51. Grogan D, Jinks-Robertson S. Formaldehyde-induced mutagenesis in Saccharomyces cerevisiae: Molecular properties and the roles of repair and bypass systems. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis. 2012 Mar 1;731(1):92–8.

52. 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. 2013 Nov;12(11):878–89.

53. 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 Research. 2013 Oct 1;41(19):8995–9005.

54. Saini N, Sterling JF, Sakofsky CJ, Giacobone CK, Klimczak LJ, Burkholder AB, et al. Mutation signatures specific to DNA alkylating agents in yeast and cancers. Nucleic Acids Research

[Internet]. 2020 Mar 5 [cited 2020 Jun 3];(gkaa150). Available from: https://doi.org/10.1093/nar/gkaa150

55. Gelova SP, Doherty KN, Alasmar S, Chan K. Intrinsic base substitution patterns in diverse species reveal links to cancer and metabolism. bioRxiv. 2020;758540. (in revision).

56. Rogozin IB, Pavlov YI. Theoretical analysis of mutation hotspots and their DNA sequence context specificity. Mutation Research/Reviews in Mutation Research. 2003 Sep;544(1):65–85.

57. Matsufuji Y, Fujimura S, Ito T, Nishizawa M, Miyaji T, Nakagawa J, et al. Acetaldehyde tolerance in Saccharomyces cerevisiae involves the pentose phosphate pathway and oleic acid biosynthesis. Yeast. 2008 Nov 1;25(11):825–33.

58. 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 Research. 2001 May 1;29(9):1994–2001.

59. Crosby RM, Richardson KK, Craft TR, Benforado KB, Liber HL, Skopek TR. Molecular analysis of formaldehyde-induced mutations in human lymphoblasts and e. coli. Environmental Mutagenesis. 1988 Jan 1;12(2):155–66.

60. Ohta T, Watanabe-Akanuma M, Tokishita S ichi, Yamagata H. Mutation spectra of chemical mutagens determined by Lac+ reversion assay with Escherichia coli WP3101P–WP3106P tester strains. Mutation Research/Genetic Toxicology and Environmental Mutagenesis. 1999 Mar 15;440(1):59–74.

61. Liu X, Lao Y, Yang IY, Hecht SS, Moriya M. Replication-Coupled Repair of Crotonaldehyde/Acetaldehyde-Induced Guanine−Guanine Interstrand Cross-Links and Their Mutagenicity. Biochemistry. 2006 Oct 1;45(42):12898–905.

62. 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. 2006 Jul 1;19(7):960–7.

63. Stein S, Lao Y, Yang IY, Hecht SS, Moriya M. Genotoxicity of acetaldehyde- and crotonaldehyde-induced 1,N2-propanodeoxyguanosine DNA adducts in human cells. Mutation Research/Genetic Toxicology and Environmental Mutagenesis. 2006 Sep 19;608(1):1–7.

64. Upton DC, Wang X, Blans P, Perrino FW, Fishbein JC, Akman SA. Mutagenesis by exocyclic alkylamino purine adducts in Escherichia coli. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis. 2006 Jul 25;599(1):1–10.

65. Garaycoechea JI, Crossan GP, Langevin F, Mulderrig L, Louzada S, Yang F, et al. Alcohol and endogenous aldehydes damage chromosomes and mutate stem cells. Nature [Internet]. 2018; Available from: http://dx.doi.org/10.1038/nature25154

Chapter 4: Pan-cancer GSEA and GO enrichment analyses

4.1 Copyright and License Policies

1. **For table 1**

   L. B. Alexandrov *et al.*, "The repertoire of mutational signatures in human cancer," *Nature*, vol. 578, no. 7793, pp. 94–101, Feb. 2020, doi: 10.1038/s41586-020-1943-3.

   The content of this paper (Figure 3: The number of mutations contributed by each of the mutational signature to the PCAWG tumors) is being reproduced in entirety or in part in this chapter. This paper was published under the terms of the open access Creative Commons Attribution 4.0 International (CC BY 4.0) license (https://creativecommons.org/licenses/by/4.0/deed.en), which allows unrestricted use, distribution and reproduction of its content in any medium, provided the authors and source are credited. No separate permission letter is required.

   Copyright © 2020, L.B. Alexandrov et al.

2. **For Figure 1**

   The content of the COSMIC database is being reproduced in entirety or in part based on COSMIC, a part of the Wellcome Sanger Institute, release v3.4 (accessed July 3, 2025), under the COSMIC license terms and conditions. The clause 2.1.1 of the COSMIC Academic License, "Publication Enablement", would allow a non-exclusive, royalty-free right to disclose data from COSMIC for non-commercial academic use, provided with an acknowledgement and specific release version number. There are no changes in the figures of SBS4 and SBS29 except minor resizing to accommodate them in the page layout so no separate written permission is required.

Here is the link of COSMIC's terms and conditions

https://www.cosmickb.org/terms/

Alexandrov et al., "Deciphering signatures of mutational processes operative in human cancer," *Cell Rep*, vol. 3, no. 1, pp. 246–259, 2013, doi: 10.1016/j.celrep.2012.12.008.

Subramanian *et al.*, "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 102, no. 43, pp. 15545–15550, Oct. 2005, doi: 10.1073/pnas.0506580102.

4.2 From Yeast to Human Tumor Genomes: Mutational signatures link

In my first doctoral project, a well-characterized haploid yeast genetic reporter system was used to understand the mutagenic properties of weak aldehydes AA and FA and characterize their associated mutational signatures. The FA-induced mutational signature resembles SBS40. In contrast, AA induced deletion of more than five base pairs, which FA did not (1).

Those yeast findings raised many questions for the remainder of my doctoral studies. The yeast mutagenesis experiment yielded two distinct mutational signatures: FA reproduces a mutational signature like SBS40, and AA produces long deletions, emphasizing the subtle chemical differences in these two aldehydes, translating into qualitatively different mutational patterns. Moreover, yeast shares a similar DNA replication and repair system as in human cells (2,3), so my results were established as a foundation to continue my research in mutational signatures. If the yeast genome can be used to generate a FA-induced mutational signature that resembles SBS40 (4), perhaps analyzing human tumor genomes rich in these mutational signatures may reveal more of the possible biological events associated with endogenous or exogenous mutagens.

The catalogue of mutational signatures keeps on expanding, but SBS40, the third most common COSMIC mutational signature that is found in at least one-third of cancers, still does not have a known etiology (4). Moreover, when I checked the latest release of mutational signatures in the COSMIC database, there were many other mutational signatures of unknown etiology, such as SBS5, SBS8, SBS12, SBS16, SBS17a, SBS17b, SBS19, and more (5). Thus, stepping beyond the eukaryotic haploid yeast genome to human cancer genomes makes it possible to investigate further the possible etiologies of unknown signatures and add more information to the signatures of known etiology.

All our questions cannot be answered by using only the human mutational landscape, so RNA-seq transcriptome data was also used together to investigate the relationship between them and their associated biological processes. Moreover, a large volume of mutational catalogue and transcriptome data was required to reconstruct the mutational signatures and correlate them with the transcriptomics expression data; therefore, 52 curated cancer datasets from the cBioPortal database (6) were used, where each cancer dataset has at least 100 samples. The GSEA and GO enrichment analysis methodology (7,8) was applied to find the biological themes associated with each of the mutational signature. GSEA identifies core genes in a gene set that link them with presence of mutational signatures. GO enrichment analysis identifies biological processes from those enriched gene sets associated with mutational signatures.

Pan-Cancer Gene Set Enrichment and Gene Ontology Analyses Uncover Biological Processes Linked to Mutational Signatures

Authors:

Mahanish Jung Thapa[1,2], Alexandre Blais[1,2], and Kin Chan[1,2*]

Affiliations:

[1] Department of Biochemistry, Microbiology and Immunology, University of Ottawa
[2] Ottawa Institute of System Biology, University of Ottawa

* Corresponding Author: kin.chan@uottawa.ca

4.4 Abstract

The somatic mutations that are found in each cancer genome are caused by multiple mutational processes, each of which forms a distinct pattern on the DNA as a "mutational signature". Despite deciphering many of these mutational signatures from different model systems, there is still a big research gap for many mutational signatures with unknown or partially understood etiology. An R package MutationalPatterns, was used to apply non-negative matrix factorization (NMF) of the mutation counts to reconstruct different mutational signatures for different cancer types to quantify the contribution of each signature. After reconstructing all the mutational signatures, we used Gene Set Enrichment Analysis (GSEA) and Gene Ontology (GO) enrichment analysis to analyze the mutation and gene expression data of 52 curated cancer datasets from the cancer genomics cBioPortal database to identify biological processes correlated with mutational signatures. The results of many signatures are consistent with the known proposed etiology of the COSMIC signatures. For those signatures of unknown etiology, SBS17a/17b is associated with reactive oxygen species, SBS8 is linked to nucleotide excision repair, SBS37 is linked to immune function, and SBS40 is linked to xenobiotic metabolism. Furthermore, common GO terms across male-only, female-only, and mixed sample datasets were found in most signatures. The age and mutation count showed different correlations for different mutational signatures. The GO semantic similarity showed a range of values for different signatures. These findings show how mutational signatures vary not only by different related biological processes but also across different sexes and age groups.

Keywords: Mutational signatures, GSEA, sex differences in cancer, age differences in cancer, GO semantic similarity

4.5 Introduction

4.5.1 DNA damage

DNA damage is the alteration in the chemical structure of DNA by different endogenous and exogenous factors (9). Endogenous sources include reactive oxygen species (ROS) produced during cellular processes, which cause oxidative DNA damage, replication errors that occur during DNA replication when incorrect bases are included in the opposite DNA strand, and alkylating agents can add either an ethyl or methyl group, leading to DNA base chemical modifications (10). Transposable elements can result in insertional mutagenesis (11), replication stress can cause fork collapse (12), and spontaneous hydrolytic processes like deamination and depurination continuously degrade DNA (13). Exogenous sources include: ultraviolet (UV) radiation which forms pyrimidine dimers, changing the structure of DNA and blocking transcription and translation; aldehydes found in smoke and tobacco; mycotoxins like aflatoxin B1 (14); drugs like cisplatin and temozolomide (15); pollutants in the air, heavy metals like arsenic and cadmium (16); and many other chemical agents that can create lesions in the DNA, leading to DNA base mutations (17). Different types of DNA damage occur from these endogenous and exogenous factors, such as DNA strand breaks that occur either in single or double DNA strands. Single-strand breaks (SSBs) are breaks where one strand is cut, whereas double-strand breaks (DSBs) have a cut in both strands (18). Chemicals such as formaldehyde, acetaldehyde and malondialdehyde react with DNA bases to form DNA adducts, N^2-hydroxymethyldeoxyguanosine or N^2-HOMe-dG, which is observed in the majority of studies using in vitro and cellular systems (19). DNA crosslinking occurs within the same DNA strand as intrastrand crosslinking and between opposite strands as interstrand crosslinks that prevent the separation of DNA strands during biological processes (20).

4.5.2 DNA repair

Cells have evolved different DNA repair mechanisms to repair the DNA damage and ensure the integrity of the genome. Common DNA repair mechanisms include base excision repair (BER) (21), nucleotide excision repair (NER) (22), mismatch repair (MMR) (23), single-strand break repair (SSBR) (24), and double-strand break repair (DSBR) (25). BER involves the action of DNA glycosylases that recognize the damaged DNA and initiates repair using the other strand as template (26). NER repairs the lesions created by bulky DNA adducts and cross-links by removing the long fragments containing damaged DNA and synthesizing a new DNA strand using the undamaged DNA as a template (27). MMR repairs the DNA base mismatches and the mutations that escaped from the proofreading activity of DNA polymerases during DNA replication. MMR recognizes the mismatches, degrades the error-containing strands, and synthesizes the correct strands (28). SSBs occur from endogenous ROS and errors in the activity of the DNA topoisomerase enzyme, halting transcription and disrupting DNA replication, and SSBR is accomplished by its own distinct pathway (29). DSBR can be repaired by two possible pathways: Homologous Recombination (HR) and Non-homologous End Joining (NHEJ). HR needs a homologous DNA template to perform repair, but NHEJ does not (30). Crosslink repair is an important cellular process that removes DNA crosslinks. Interstrand crosslink repair is complex and requires the coordination of multiple pathways such as HR, Fanconi anemia (FA), and translesion synthesis (TLS) (31) unlike Intrastrand crosslink repair that uses NER (20).

4.5.3 When Repair Fails: The Origin of Mutations and Mutational Signatures

DNA repair mechanisms could be quite (but not 100%) efficient against endogenous and exogenous DNA damage. When repair mechanisms are either saturated from a large amount of DNA damage or compromised by inherent deficiencies in repair pathways, lesions can persist. Occurrence of such unrepaired or incorrectly repaired DNA leads to mutations such as base substitutions (9), insertions and deletions (indels) (32), genome rearrangements (33), and copy number changes (34) potentially affecting oncogenes and tumor suppressor genes.

The footprints of all these different types of mutational processes form a specific pattern known as mutational signatures. Mutational Signatures were discovered in cancer genomes by analyzing somatic mutations using a non-negative matrix factorization algorithm (NMF) (35). COSMIC (Catalogue of Somatic Mutations in Cancer) is a database of different mutational signatures with associated etiology (36). Single base substitution (SBS) is the most numerous types of mutational signatures found in the COSMIC database. The profile of each signature has six substitution types: C>A, C>G, C>T, T>A, T>C, and T>G, and all these substitutions are referred to by the pyrimidine of the mutated Watson-Crick base pair. 96 possible mutation types were produced by examining the immediate bases at 5′ and 3′ positions for each type of substitution (6 types of substitution × 4 types of 5′ base × 4 types of 3′ base). All the mutational signatures were reported based on the 96 possible trinucleotide contexts (4,37). Together, these mutational signatures give information on exposures to diverse carcinogenic/mutagenic agents and the results of damaged cellular pathways to understand the foundation of the mutational process for cancer (38). From **Table 4.1**, each of the mutational signature has been associated with different forms of DNA damage and cellular context where repair mechanisms have failed. Despite the etiology of these COSMIC signatures ranges from chemical mutagens of tobacco

found in SBS4 and SBS29 to specific mutators such as APOBEC activity in SBS2 and SBS13 to POLE (Polymerase Epsilon) mutations in SBS10a and SBS10b to mismatch repair deficiencies in SBS6, SBS14, SBS15, SBS20, SBS21, and SBS26, nonetheless many of these signatures are less well studied or have unknown etiology (5,36,39–41). In order to investigate and understand systematically all the biological pathways associated with mutational signatures, we used gene set enrichment analysis (GSEA) and gene ontology (GO) enrichment analysis to identify the correlation between mutational signatures and expression data, providing a powerful tool to reveal the biological events associated with individual mutational signatures characterizing the complex mutational landscapes.

COSMIC Signatures	Proposed Etiology
SBS1	Deamination of 5-methylcytosine (42)
SBS2, SBS13	APOBEC activity (43)
SBS3	Defective HR DNA repair; BRCA1/2 mutation (44)
SBS4	Tobacco smoking (45)
SBS5	Unknown (35)
SBS6, SBS14, SBS15, SBS20, SBS21, SBS26, SBS44	Defective DNA mismatch repair (46–48)
SBS7a, SBS7b, SBS7c, SBS7d	Ultraviolet light exposure (45,49,50)
SBS8	Unknown (35)
SBS9	In part, polymerase η activity (37)
SBS10a, SBS10b	POLE mutation (4,51)
SBS11	Temozolomide treatment (37,52)
SBS12	Unknown (37)
SBS16	Unknown (37,53)
SBS17a, SBS17b	Unknown (54,55)
SBS18	ROS (37,52)
SBS19	Unknown (37)
SBS22	Aristolochic acid exposure (56,57)
SBS23	Unknown (58)
SBS24	Aflatoxin exposure (58,59)
SBS25	Chemotherapy (58)
SBS28	Unknown (46,58)
SBS29	Tobacco Chewing (58,60)
SBS30, SBS36	Defective base excision repair (47,58,61,62)
SBS31, SBS35	Platinum treatment (63)
SBS32	Azathioprine treatment (64)
SBS33	Unknown (4)
SBS34	Unknown (4)
SBS37	Unknown (4)
SBS38	Indirect effect of ultraviolet light (4)
SBS40	Unknown (4)
SBS41	Unknown (4)
SBS42	Haloalkane exposure (65)

Table 4.1: COSMIC signatures with their proposed etiology (4,5)

4.5.4 Gene Set Enrichment Analysis (GSEA)

Gene Set Enrichment Analysis (GSEA) is a computational approach to determine if a predefined set of genes shows statistically significant and concordant differences between two biological states, such as different phenotypes, disease conditions, or experimental conditions. GSEA is distinct from more conventional gene expression analysis techniques, which concentrate on specific differentially expressed genes (7). Rather, GSEA considers gene sets found, e.g., in the

molecular signature database (MSigDB) (66,67) or other classification systems, as a whole, which can reveal subtle variations in gene expression patterns and provides a more comprehensive understanding of the underlying biological mechanisms (7,68). Researchers can find gene sets that are either significantly enriched or depleted in samples with a given mutational signature by applying GSEA to gene expression data in the context of single base substitution (SBS) mutational signatures found in COSMIC (5,36). This may highlight the associated molecular mechanisms that underlie the mutational processes for that particular mutational signature. In contrast to conventional chemotherapy, the development of better targeted therapies can be guided by the identification of mutational patterns and altered gene expression patterns in different cancer types (69).

4.5.5 Gene Ontology (GO) Enrichment Analysis

Following GSEA, Gene Ontology (GO) enrichment analysis is used for further understanding how the broader biological significance of enriched gene sets. This method identifies GO terms that are statistically overrepresented in a given gene list relative to a reference list and overall provide a structured hierarchy of standardized GO terms for categories such as biological processes, cellular components, and molecular functions including curated and predicted gene annotations. Biological processes GO annotations are the most commonly used resource for GO enrichment analysis (68).

Researchers can associate COSMIC mutational signatures with gene sets. The next step involves performing a GO enrichment analysis to understand the biological themes among the genes correlated with specific mutational processes (70). This entails finding the mutated genes that match specific mutational signatures, annotating these genes with GO terms, and performing enrichment analysis to find GO terms that are significantly over-represented in these genes

compared to a reference gene list. This facilitates identifying putative biological processes, cellular components, or molecular functions that are specifically connected to the mutation mechanisms (68,71).

Overall, through the pan-cancer application of GSEA and GO enrichment analysis, our study intends to find a framework for understanding the biological process that correlates with these mutational signatures. This provides insights into how different mutational processes shape cancer phenotype across multiple cancer genomes.

4.5.6 Methods

The curated data for this study was sourced from the cBioPortal database, which is an open-access resource for cancer genomics datasets (72,73). We used 52 cancer datasets of 19 cancer types **(Table 4.2)** with mutation and RNA Sequencing (RNA-Seq) data. Mutation data provide genetic changes found in all different tumor samples (74). The RNA-Seq holds the gene expression profile of tumor samples quantifying the amount of mRNA present in thousands of genes (75). The human reference genome GRCh37/hg19 (Genome Reference Consortium Human Build 37/Human Genome version 19) or GRCh38/hg38 was used to align sequencing reads of both mutation and RNA-Seq data (76). The mutation and RNA-seq data were derived from the same samples within the 52 individual cancer datasets; however, the samples differed between datasets. The integration of mutation and RNA-Seq data correlate the gene expression changes with mutational signatures and uncover biological mechanisms for individual mutational signatures involved in cancer progression (77). The methodology is illustrated graphically in the **Figure 4.1**.

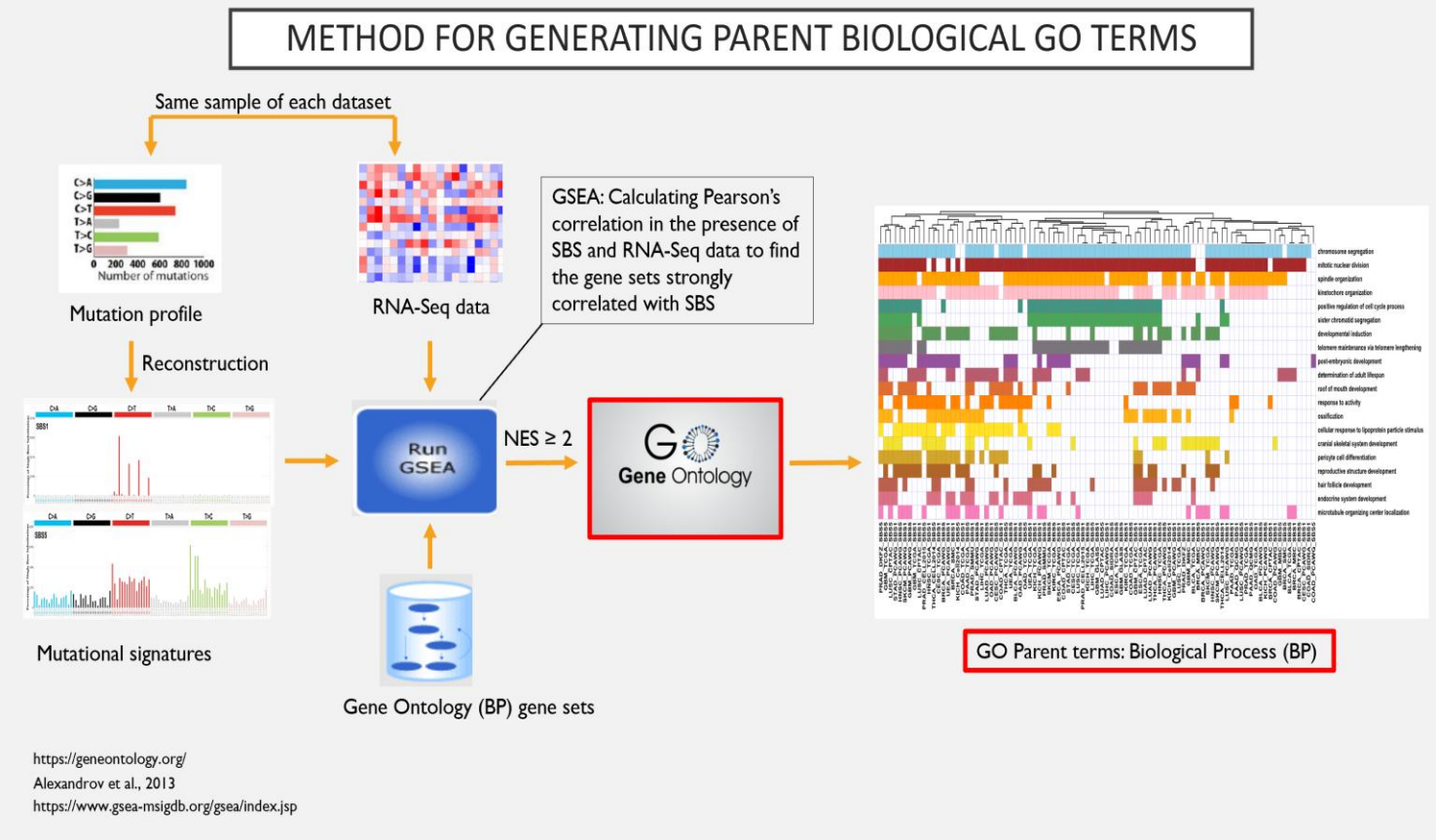


Figure 4.1: Approach for generating Parent GO terms using GSEA and GO enrichment analysis
The mutation and expression (RNA-Seq) data were derived from the same samples of cancer datasets. The reconstruction of mutational signatures was done from mutation data. The Pearson correlation was calculated between the reconstructed signature and RNA-Seq data to find genes strongly correlated with the signature, and the GSEA was used to find the enriched gene sets and GO analysis to find the parent GO terms.

4.6.1 Data Source

Abbreviation	Cancer Types
BLCA	Bladder Urothelial Carcinoma
BRCA	Breast Cancer
CESC	Cervical Squamous Cell Carcinoma
COAD	Colorectal Adenocarcinoma
ESCA	Esophageal Adenocarcinoma
GBM	Glioblastoma Multiform
HNSC	Head and Neck Squamous cell Carcinoma
KICH	Kidney Chromophobe
KIRC	Kidney Renal Clear Cell Carcinoma
LICH	Liver Hepatocellular Carcinoma
LUAD	Lung Adenocarcinoma
LUSC	Lung Squamous Cell Carcinoma
OAD	Ovarian Adenocarcinoma
PAAD	Pancreatic Adenocarcinoma
PRAD	Prostate Adenocarcinoma
SKCM	Skin Cutaneous Melanoma
STAD	Stomach Adenocarcinoma
THCA	Thyroid Carcinoma
UEA	Uterine Endometrioid Carcinoma

Table 4.2: List of cancer types

4.6.2 Computational Tools

The maftools (78) and MutationalPatterns (79) are R/Bioconductor packages which are widely used in the cancer genomics field to examine the mutational genomic sequencing data and interpret the mutational signatures found in the data for different types of cancer. All the mutation data are in text-delimited MAF (Mutation Annotation Format). MAF is a structured, tab-separated mainly used to store and display somatic mutations found in tumor samples. Each row of the MAF file corresponds to a distinct mutational event and each column provides detailed annotation (80). It is easily readable by maftools and MutationalPatterns and can directly parse and analyze these MAF fliles for mutational signature analysis. Each cancer type has a certain set of mutational signatures, so each COSMIC mutational signature was

reconstructed, along with the relative contribution of each reconstructed signature for each sample of the 52 cancer datasets.

The RNA-Seq data found in the cBioportal are in one or more of the following normalized formats such as FPKM (Fragments per Kilobase of transcript per million mapped reads), RPKM (Reads per Kilobase of transcript per million mapped reads), TPM (Transcripts per million), and RSEM (RNA-Seq by Expectation Maximization), and were used for quantifying gene expression (81). FPKM is used for paired-end RNA-Seq data and normalizes read counts by both gene length in kilobases (kb) and total number of mapped fragments (in millions) (82). RPKM is used for single-end RNA-Seq data and is like FPKM but uses reads instead of fragments (83). TPM normalizes the read counts across gene length and total transcript abundance (84). RSEM estimates genes and isoform expression levels and generate normalized expression values such as TPM or FPKM (81,85). All these RNA-Seq data are in text-delimited format where each column of RNA-Seq data has headers such as Hugo Symbol (gene symbol), Entrez ID (gene numerical identifier) for each gene and samples name and each row includes the expression values for each gene present in that sample.

For Gene Set Enrichment Analysis (GSEA), we repurposed an R script written by Dr. Alexandre Blais, and associated libraries were used (7). The GO analysis and visualization were done using an R script and loading associated libraries and the Age and Mutation correlation was evaluated using general linear regression (58,86–88).

4.6.3 Reconstruction of SBS signatures

Different cancer types have different set of mutational signatures (36). In cancer genomics, the NMF (Non-Negative Matrix Factorization) was used to identify distinct mutational signatures associated with many biological events and quantify their contribution in all tumor samples as

part of its factorization of the input matrix of mutation spectra (89). The results are biologically interpretable because NMF based approaches are not just an algorithmic process but also consider the previous evidence of biological plausibility for each set of signatures present in each cancer types, literature on DNA damage and repair, biological evidence from experimentally determined mutational signatures (4).

We started by using the maftools package to load the mutation data in MAF format from the individual cancer dataset. With the help of NMF present in MutationalPatterns package, trinucleotide mutation count matrix was created, where rows represent mutations in specific trinucleotide contexts, and columns represent samples of each cancer dataset, subsequently decomposing the mutation count matrix into mutational signatures and their relative contribution with reference to human genome (BSgenome.Hsapiens.UCSC.hg19) or (BSgenome.Hsapiens.UCSC.hg38) based on their NCBI Build (35). The trinucleotide matrix was then exported and cleaned to guarantee the proper labeling and formatting of mutation types and tumor samples.

After preparing the mutation matrix, we selected the known set of COSMIC signatures for every cancer type (4), and general or default signature refitting was done to refit the mutation data against these COSMIC signatures in each cancer type (90). The refitting step quantifies the contribution of each signature per sample which is normalized to calculate the relative contribution. This normalized contribution will guide to find the dominant mutational processes associated with each signature in tumor samples (79). Subsequently the quality of the signature refitting was done through cosine similarity which measures how closely the reconstructed mutation profiles match the original observed mutation profiles. The filtration of these similarity values for each sample can be done at thresholds ($\geq$0.9 or $\geq$0.8) to identify highly

reliable signature attributions for each cancer type dataset. These refined datasets was used for identifying plausible biological mechanisms associated with each signature (4,79,91). Altogether, mutational data from different cancer datasets were used to reconstruct COSMIC signatures with their relative contribution for each cancer sample. The previous report also showed the reconstruction of mutational signatures in these cancer samples from different cancer types in the Pan-Cancer Analysis of Whole Genomes (PCAWG) (4).

4.6.4 GSEA Analysis

The biological pathways linked to each of the mutational signature were investigated using GSEA by corelating each signature to the gene expression RNA-Seq data present in different cancer cohorts (7). The RNA-Seq data matrix was cleaned by retaining “Hugo Symbol” and removing “Entrez Gene ID” column and was transposed so that the samples were grouped in rows and genes in each column aligning the structure with the mutational signature matrix. The cleaned, transposed RNA-Seq matrix was then filtered to retain only those samples with matched signature matrix.

A Pearson’s correlation (92) was computed in the presence of a given reconstructed mutational signature contribution and RNA-Seq gene expression values for the filtered same sample set. Genes were then ranked based on their correlation values to the individual signature, with positively correlated genes at the top of the list and negatively correlated genes at the bottom. Now this ranked list serves as a input for the GSEA (93). The biological process (BP) gene ontology gene sets were used to determine the gene sets strongly correlated in the presence of a given signature (7,94,95). To avoid small-sample biasness and for robust statistical robustness, gene sets were filtered to retain only those gene sets that have 10-500 genes present in our dataset.

GSEA analysis was done implementing the GSEA function from the clusterprofiler package (96) with the input gene list, gene set database, benjamaini-hochber method for the p-value adjustment to control the false discovery rate “pAdjustMethod = BH”, which is the estimated proportion of false positives identified as significantly enriched gene sets among all gene sets and BH adjusts for multiple testing by helping to control enriched pathways likely due to random chance (68), and minimum and maximum gene set size “minGSSize = 10”, and “maxGSSize = 500” to prevent testing very small gene sets which may produce unstable enrichment score due to random variation and excludes big gene sets that may dominate the results by masking specific biological signals improving the biological interpretability of the enrichment analysis (97) and 10,000 simple pemutations “nPermSimple = 10,000” to ensure better statistical analysis (98).

The enrichment analysis was done using the Fast GSEA (fgsea) algorithm for efficient computation (99). The resulting output has gene set ID and size, normalized enrichment scores (NES), p-values, and adjusted p-values, q -values, leading edge subset and core enrichment. Each of these correlated gene sets has its specific enrichment score with the normalized enrichment score (NES). The NES is the enrichment score normalized across gene sets, which accounts for the size of the set to compare different sizes of gene sets. I ordered the correlated gene sets in descending order based on NES values (100). It assesses with an NES cut-off value of 2 whether a gene set is overrepresented at the top or bottom of a ranked list of genes. A positive value of NES at the top of the list indicates that the genes in the set tend to be upregulated, whereas a negative value at the bottom of the list indicates downregulation in correlation with the signature (7,101). Leading edge is the percentage of genes in a gene set

that contributes most to the enrichment score and core enrichment is the list of gene names that are the part of leading edge (102).

4.6.5 GO Enrichment Analysis

After GSEA, to understand the deeper insights into the biological associations of genes driving mutational signatures-associated expression changes in cancer datasets, we performed GO enrichment analysis mainly focusing on the core enrichment genes of the upregulated gene sets identified through GSEA (103). These genes were extracted from the core enrichment column of the GSEA results, where each gene symbols were delimited by forward slashes (/). The string of gene symbols was transformed into a gene list using “strsplit” function for further analysis (104).

We used “enrichGO” function from the clusterprofiler package, which helps in the overrepresentation of GO terms. The biological process (BP) ontology category was used to investigate the biological pathways associated with individual mutational signature (87). A separate GO enrichment analysis was done on the core enrichment genes of each upregulated GSEA gene set output. The background universe against which the GO enrichment was tested (105) has all the genes present in the expression RNA-Seq data of that dataset, extracted by assigning the column name Symbol for gene symbols of the expression RNA-Seq matrix to a “gene_info” data frame. The human gene annotation database “org.Hs.eg.db” with gene identifiers provided with HGNC gene symbol format “keyType = SYMBOL” (106) was used for the GO enrichment analysis. Those GO terms that have Benjamini-Hochberg adjusted p-value (FDR) < 0.05 and q-value of <0.01 were considered statistically significant ensuring better GO enrichment output (87,107). All the upregulated gene sets from the GSEA output were subjected in a loop. The GO enrichment output for each upregulated gene set was stored as an

"S4 object" (108) which is an structured way of storing complex data that has multiple layers such as genes, stats, parameters and annotations and subsequently converted into data frames. All the GO enrichment results were concatenated into a single result table using "rbind" function (109), providing a long list of GO biological Process (BP) terms enriched across all core enrichment genes present in each upregulated gene sets.

Furthermore, the biological process GO terms were refined using semantic similarity-based reduction of the GO term list using the "GOSemSim" and "rrvgo" packages (110,111). This step helps in removing the redundancy by clustering similar GO terms and identifying the parent GO term for wider biological interpretation. The GO enrichment output was loaded and then extracted the list of all GO term IDs from the enrichment table and the function "calculateSimMatrix" from the "rrvgo" package was used to compute the semantic similarity scores between all pairs of GO terms (111). This function uses GO ontology structure using "GO.db" package (112,113), human gene ontology database "org.Hs.eg.db" (114) and Resnik-based relative information method "method = Rel" (115) to quantify the similar GO terms within the biological process ontology "ont = BP" (111). A scoring vector was prepared by assigning each GO term based on negative log10 of the p-value which will provide more weight to more statistically significant GO terms (116). Both semantic similarity matrix and scoring vector were passed through the reduceSimMatrix( ) function to group the similar GO terms and provide the same parent GO term for similar child GO terms highlighting the main biological themes associated with the individual signature (86,87,111). A binary matrix was created where "1" means that the GO term is present in the dataset and "0" means that the GO term is absent. A heatmap was created to find the top 20 common biological GO terms for each of the mutational signature across all types of cancer datasets (117).

4.6.6 Sex-based mutational process

The sex differences in mutation processes could be the cause of differences in mutation density and tumor evolution features (118). The GSEA and GO analysis were done in all cancer datasets having both male and female samples, and separating male-only samples from the female-only samples. This analysis will help to investigate the gender-based mutational process differences for signatures of different cancer types. The common parent GO terms across all these sets of cancer data, and the unique parent GO terms for male-only and female-only, were evaluated.

4.6.7 GO semantic similarity

It is a computational approach to quantify the functional similarity between genes, gene products, or biological processes based on their associated GO terms (119). This method takes advantage of GO's hierarchical structure, which links terms through parent-child relationships that represent biological distinctiveness (120,121). GO semantic similarity can be used in cancer genomics to compare the biological processes associated with mutational signatures in different datasets (e.g., PCAWG vs. TCGA) for the same cancer type (122). This aids in determining whether the biological processes associated with a signature are consistent across studies or are influenced by biases unique to a certain dataset.

We used the semantic similarity analysis of parent GO terms using the “GOSemSim” package. This analysis focused on parent-level biological process (BP) GO terms which were previously derived from redundancy reduction “rrvgo” package (110,111). The parent GO terms show high level biological categories summarizing enriched biological pathways associated with mutational signatures. The annotation resources such as “GO.db”, “org.Hs.eg.db”, and initialized semantic data with “godata”, specifying the ontology “ont = BP” for biological process

terms were loaded to precompute the ontology structure and the information required for pairwise similarity calculations (123). For each individual signature present in two cancer datasets, we listed the parent GO terms associated with each signature.

The similarity between enriched parent GO terms of two cancer datasets were evaluated using function such as “mgoSim” from GOSemSim (110). This function uses the Wang’s method considering the structure of GO graph and measures the semantic similarity between two sets of enriched parents GO terms based on shared ancestry and hierarchy (8). We computed pairwise similarity scores between each signature present in two datasets of same cancer types. At the end, we consider all the pairwise semantic similarity scores for each signature and build a bar plot to see the range of median values of semantic scores for each signature.

4.6.8 Correlation of Age and Mutational Signatures

The relationship between mutations attributed to mutational signatures found in multiple cancer types and the age of diagnosis was explored using correlation analysis based on the reconstructed signatures. Mutational signatures were reconstructed applying NMF to trinucleotide mutation matrix derived from mutation data (79). The proportion and log10 transformation of the total mutational count per gigabase was calculated respectively for each sample to determine the presence of mutational signatures in different cancer types (124). Each reconstructed mutational signature was tested to evaluate the relationship between signature and age at diagnosis of the patient. The general linear regression model was used to find the linear dependencies between the mutations found in each of the mutational signature across all the samples for each cancer type and the age of diagnosis of these samples (58). The independent variable was patient’s age at diagnosis and the dependent variable was either proportional contribution or the log10-transformed mutation count per gigabase of a given

signature. This technique will help to find the association of specific mutational processes of signatures correlating with age. The p-values and 95% confidence interval bands around the regression line illustrate the trends and guide to identify the age-related mutational signatures, further confirming and extending the known associations (clock-like signatures SBS1 and SBS5) (58), and guide to investigate the age-dependent mutational processes for other signatures important for tumor development.

4.7 Results

We used GSEA and GO enrichment analysis to reveal the biological associations of many signatures which were consistent with previous studies and can validate our computational approach of identifying biological pathways for different mutational signatures present across multiple cancer datasets. We started with SBS2 and SBS13, which are of known etiology and well understood.

4.7.1 SBS2/SBS13

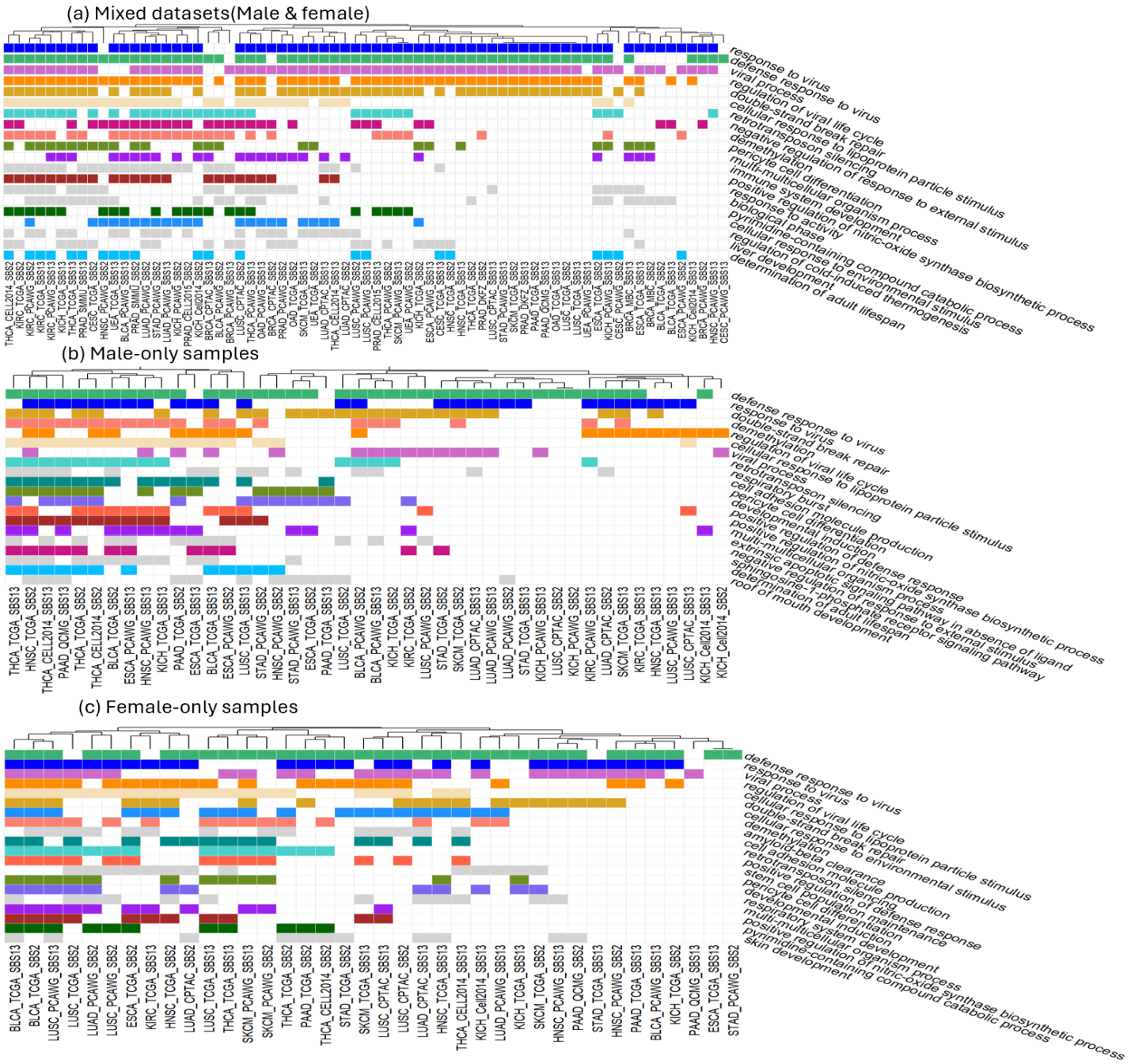


Figure 4.2: GSEA and GO analysis identified top 20 parent GO terms associated with SBS2/13
In the x-axis, these are all different cancer datasets and in y-axis, the top 20 parent GO terms for SBS2 and SBS13. The colored GO terms are commonly found across (a), (b), and (c) or between any two. The light gray GO terms are unique for (a), (b) and (c) respectively.

Both SBS2 and SBS13 are attributed to the APOBEC (apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like) family of cytidine deaminases based on similarities in the sequence context of cytosine mutations caused by APOBEC enzymes in experimental systems. SBS2 is characterized by a

high number of C>T mutations, whereas SBS13 is associated with C>G mutations. The activation of APOBEC cytidine deaminases in cancer may be due to viral infection, retrotransposon jumping, or tissue inflammation (5,36,125).

In **Figure 4.2**, all three datasets (a, mixed), (b, male samples only), and (c, female samples only) have common GO terms “response to virus”, “regulation of viral life cycle”, “defense response to virus”, “viral process”, and “double-strand break repair” as important characteristics of SBS2 and SBS13. Interestingly, APOBEC3A and APOBEC3B can hypermutate nuclear DNA, which will lead to double-strand breaks (43,126). To repair them, double strand break repair has presumably been upregulated. These results are all consistent with the proposed etiology. The enrichment of other GO terms less common across mixed datasets, male-only samples, and female-only samples such as “retrotransposon silencing”, “cellular response to lipoprotein stimulus”, “demethylation”, “pericyte cell differentiation”, “multi multicellular organism process”, and “positive regulation of nitric oxide synthase biosynthetic process” suggest a combination of inflammatory, metabolic, and epigenetic factors are more loosely correlated with APOBEC mutagenesis (127–131).

Besides the common GO terms associated with viral infection, defense response and double-strand break repair, other GO terms such as “cell adhesion molecule production” and developmental induction” were also common between male-only and female-only samples suggesting conserved biological roles of APOBEC-associated mutational signatures across sexes. Cell adhesion molecules (CAM) production helps in maintaining immune surveillance and tissue architecture, but when dysregulated, it often leads to cancer progression (132). Developmental induction reflects the signaling processes that guide cell fate and tissue patterning, which may be reactivated during tumorigenesis (133). The core biological processes affected by APOBEC mutagenesis are fundamentally conserved, although the presence of some GO terms in either male-only or female-

only samples may point to subtle sex-specific differences at the gene expression level.

4.7.2 Other signatures of known etiology/characteristics

Application of our analytical approach provides further validation. We found SBS1 and SBS5 (134) to be associated with cell division where top GO terms “chromosome segregation”, “mitotic nuclear division”, “spindle organization”, kinetochore organization”, “sister chromatid segregation”, and “positive regulation of cell cycle process”, common across mixed cancer datasets, male-only samples, and female-only samples are consistent with the previous findings of other researchers for SBS1 and SBS5, both as clock-like signatures where mutations accumulate with aging, although the etiology of SBS5 is still unknown (58) (see Supp. results 1, page no. 214, 241, 261). Our analysis also showed the association of SBS3 with top GO terms “double-strand break repair”, “double-strand break repair via homologous recombination” (HR) (135) found across mixed cancer datasets, male-only and female-only matching with previous findings of identifying SBS3 as a predictor of defective homologous recombination repair associated with BRCA1 and BRCA2 mutations (see Supp. results 1, page no. 216, 243, 263). We found that SBS4 and SBS29 (136) were associated with top GO term “nucleotide excision repair (NER)” common across mixed cancer datasets, male and female-only samples, which is consistent with the upregulation of NER to repair the DNA damage
caused by bulky DNA adducts released from tobacco smoking in SBS4 (137) and tobacco chewing in SBS29 (138) (see Supp. results 1, page no. 217, 244, 264).
Furthermore, our analysis found that signatures such as SBS6, SBS14, SBS15, SBS20, SBS21, SBS26, and SBS44 (139) were associated with top GO terms “mismatch repair” (MMR) common across mixed cancer datasets, male-only and female-only samples consistent with previous findings (see Supp. results 1, page no. 218, 245, and 265). Signature 9 have been associated

with immune function and our analysis showing correlation with top GO term “somatic diversification of immunoglobins” matches with previous findings (37) (see Supp. Results 1, page no. 212). Another signature SBS10a and 10b have been also identified with top GO term “mismatch repair” common across mixed cancer datasets, male-only and female-only samples (see Supp. Results 1, page no. 222, 248, 268). The proposed etiology for both SBS10a and SBS10b signatures is associated with polymerase epsilon (POLE) exonuclease domain mutations (4). Both signatures SBS10a and SBS10b are indicative of defective proofreading by POLE, which leads to replication errors that accumulate if not corrected by MMR. Tumors with POLE mutation often show a high mutation rate due to the combined effect of defective POLE and possibly compromised MMR (46). SBS11 is associated with temozolomide treatment and mismatch repair which is consistent with our findings (140) (see Supp. Results 1, page no. 223, 249, 269).

Signatures SBS7a, SBS7b, SBS7c and SBS7d (141) commonly found in skin cancer showed top GO terms such as “cellular response to UV”, “UV-damage excision repair”, and “response to UV-A” common across mixed cancer datasets, male-only, and female-only samples matches with previous findings of its etiology associated with UV-related DNA damage (see Supp. Results 1, page no. 219, 246, 266). SBS38, found only in ultraviolet exposed skin cancer (4) have top GO terms such as “response to UV”, and “skin development” from our analysis further validates our methodology (see Supp results 1, page no. 238).

For SBS18, top GO terms such as “response to reactive oxygen species, “response to oxidative stress”, “super-oxide metabolic process”, and “nitric-oxide metabolic process”, and for SBS30 resulting from *NTHL1* (Nth Like DNA Glycosylase 1) mutations (142), the top GO term “base-

excision repair" common across mixed datasets, male-only samples and female-only samples (see Supp. Results 1, page no. 227, 232, 253, 257, 273, 277) were consistent with previous results. Other signatures SBS22 (Aristolochic acid exposure) (45), SBS24 (Aflatoxin exposure) (59), SBS31/35 (Platinum drugs treatment) (63), demonstrate links to NER (see Supp. Results 1, page no. 229, 230, 234, 254, 255, 258, 274, 275, 278) matching previous findings which further validate our systematic computational approach of generating biological pathways for different mutational signatures present across multiple cancer datasets.

4.7.3 SBS17a/17b

The proposed etiology of SBS17a is still unknown, and few previous studies related to SBS17b have been linked to damage caused by ROS (reactive oxygen species) and fluorouracil (5FU) chemotherapy. Although both SBS17a and SBS17b are similar, SBS17a is dominated by T>G mutations, and SBS17b is dominated by T>C mutations at the NTT trinucleotide context (5,55). According to our analysis, we found that out of the top 20 GO terms in **Figure 4.3**, the top common GO terms for SBS17a/17b across (a, mixed datasets), (b, male-only samples) and (c, female-only samples) —such as "response to reactive oxygen species", "response to oxidative stress", "superoxide metabolic process", and "nitric oxide metabolic process"—suggest that these mutational signatures are likely driven by ROS, oxidative DNA damage and inflammation (54,55). These processes are fundamental to all cells and not sex-specific, explaining their consistent enrichment across different sample groups. Together, this observation may broaden the biological relevance of SBS17a and SBS17b associated with ROS and oxidative DNA damage which has implications not only for cancer biology but also the heterogenic nature of tumors which may co-opt or dysregulate systematic homeostatic processes.

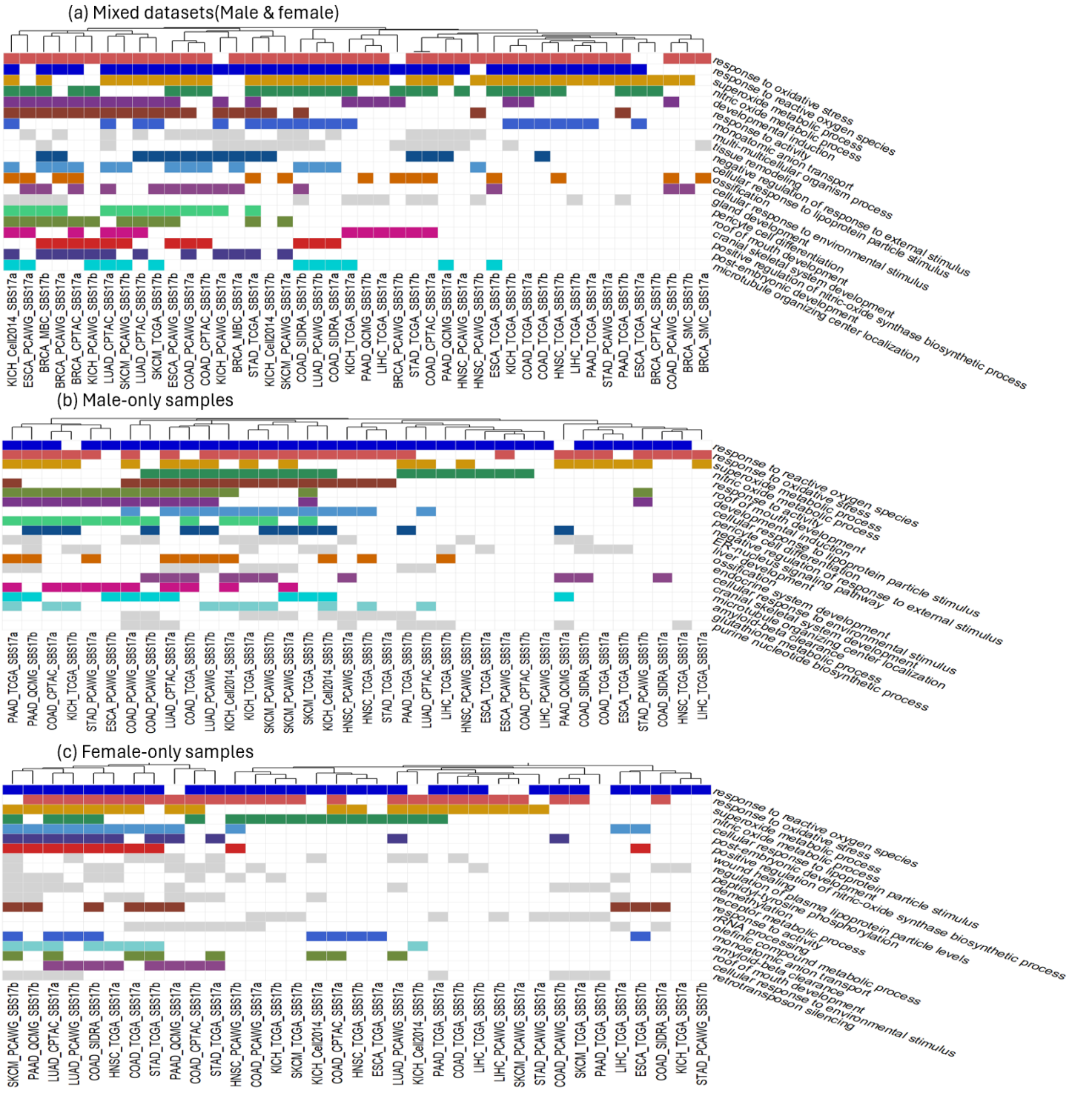


Figure 4.3: GSEA and GO analysis identified top 20 parent GO terms associated with SBS17a/17b
In the x-axis, these are all different cancer datasets and in y-axis, the top 20 parent GO terms for SBS17a and SBS17b. The colored GO terms are commonly found across (a), (b), and (c) or between any two. The light gray GO terms are unique for (a), (b) and (c) respectively.

4.7.4 SBS40

The proposed etiology for SBS40 is still unknown (36). In **Figure 4.4**, out of the top 20 parent GO terms, “xenobiotic metabolic process”, “xenobiotic catabolic process”, and “cellular response to xenobiotic stimulus” are common across (a, mixed datasets), (b, male-only samples) and (c, female-only samples). Xenobiotics are chemical substances that are foreign to the biological system, including drugs, pollutants, and other environmental toxins. Persistent exposure to compounds like formaldehyde and the resulting DNA damage could potentially lead to mutations characterized by SBS40 (1). This might include both direct DNA adduct formation and indirect effects such as oxidative stress (143). The other enriched GO terms common across many male, female, and mixed cancer datasets include “chromosome segregation”, and nuclear division”, similar to SBS1 and SBS5. The presence of these GO terms highlight conserved, general biological processes such as detoxification pathways and cell division, consistent with the common occurrence of SBS40 across many cancer types.

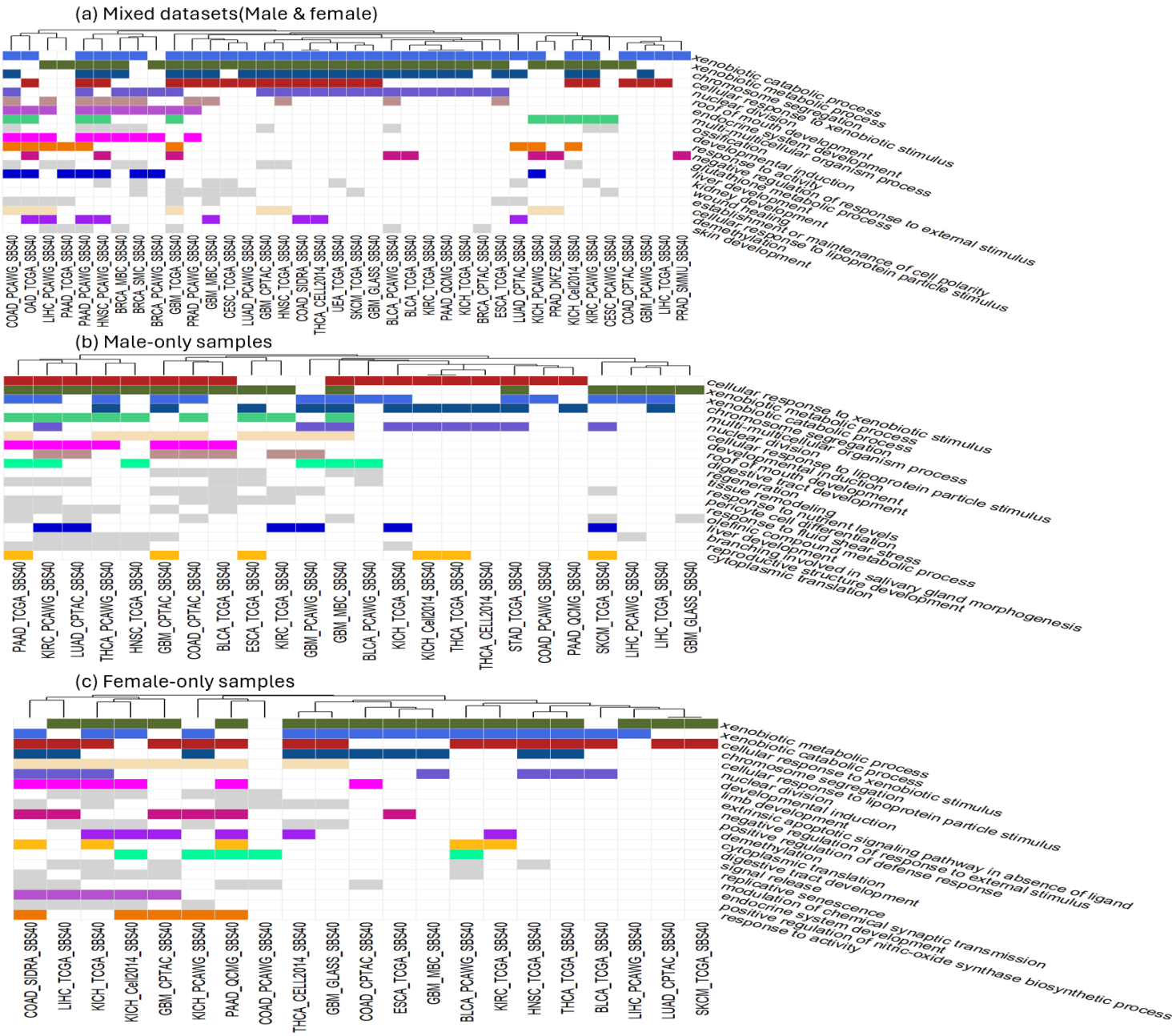


Figure 4.4: GSEA and GO analysis identified top 20 parent GO terms associated with SBS40
In the x-axis, these are all different cancer datasets and in y-axis, the top 20 parent GO terms for SBS40. The colored GO terms are commonly found across (a), (b), and (c) or between any two. The light gray GO terms are unique for (a), (b) and (c) respectively.

4.7.5 SBS37

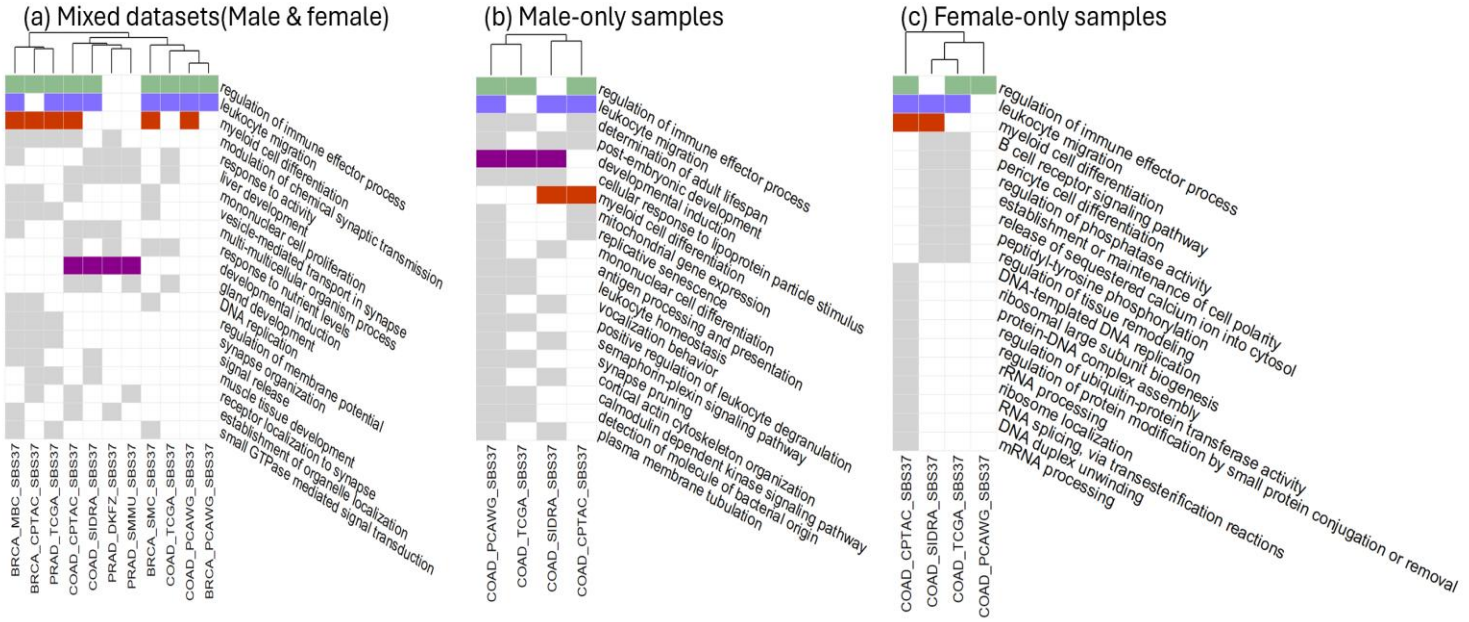


Figure 4.5: GSEA and GO analysis identified top 20 parent GO terms associated with SBS37
In the x-axis, these are all different cancer datasets and in y-axis, the top 20 parent GO terms for SBS37. The colored GO terms are commonly found across (a), (b), and (c) or between any two. The light gray GO terms are unique for (a), (b) and (c) respectively.

SBS37 is a COSMIC mutational signature with unknown etiology, but it has been observed in various cancer types at moderate levels (4,5). While its precise biological cause remains unclear, SBS37 is linked to immune-related or inflammatory processes, as per our analysis (**see Figure 4.5**). The GO terms regulation of “immune effector process”, “leukocyte migration”, and “myeloid cell differentiation” (144,145) are consistently enriched across (a, mixed datasets, (b, male-only samples) and (c, female-only samples), suggesting that SBS37 may arise in immune-infiltrated tumor microenvironments (146) or tissues undergoing immune modulation, making it broadly relevant regardless of sex. The immune cell dynamics particularly involving myeloid lineage cells such as neutrophils and macrophages (147) play an important role in inflammation, immune surveillance and creating tumor microenvironment (148). Leukocyte migration (149) and immune effector function are important for adaptive and innate immune responses (150) and may indicate the accumulation of SBS37 mutation during chronic

inflammation or immune infiltration. The fact that these associations with SBS37 persist independent of sex may represent the fundamental immunological link. Unlike most signatures, there are few common GO terms between males and females for SBS37.

4.7.6 SBS8

The etiology of SBS8 is still unknown (5,37). Based on our analysis, we found a plausible common top GO term “nucleotide-excision repair” (**see Figure 4.6**) found across (a, mixed datasets, (b, male-only samples) and (c, female-only samples). It may be linked to bulky DNA adducts or oxidative damage that require NER mechanisms for repair (27,37). Another possibility if NER function (but not expression) is impaired, then lesions may persist and become substrates for error-prone replication accumulating mutational patterns like SBS8. The consistent enrichment of NER expression in both sexes highlights a core DNA repair correlation with mutation accumulation in SBS8-positive samples. Similarly to SBS37, there are few common GO terms between males and females for SBS8. The low frequency of SBS8 in both male and female samples may be linked to specific cellular conditions or DNA damage associations which may partially impair the NER pathway reflecting localized or transient repair failures together with other mutational processes. The fact that the SBS8 is seen in few samples independent of sex, may indicate that the mutagenic driver exposure or process is not widespread, rather pointing to tissue-specific or sporadic events (151).

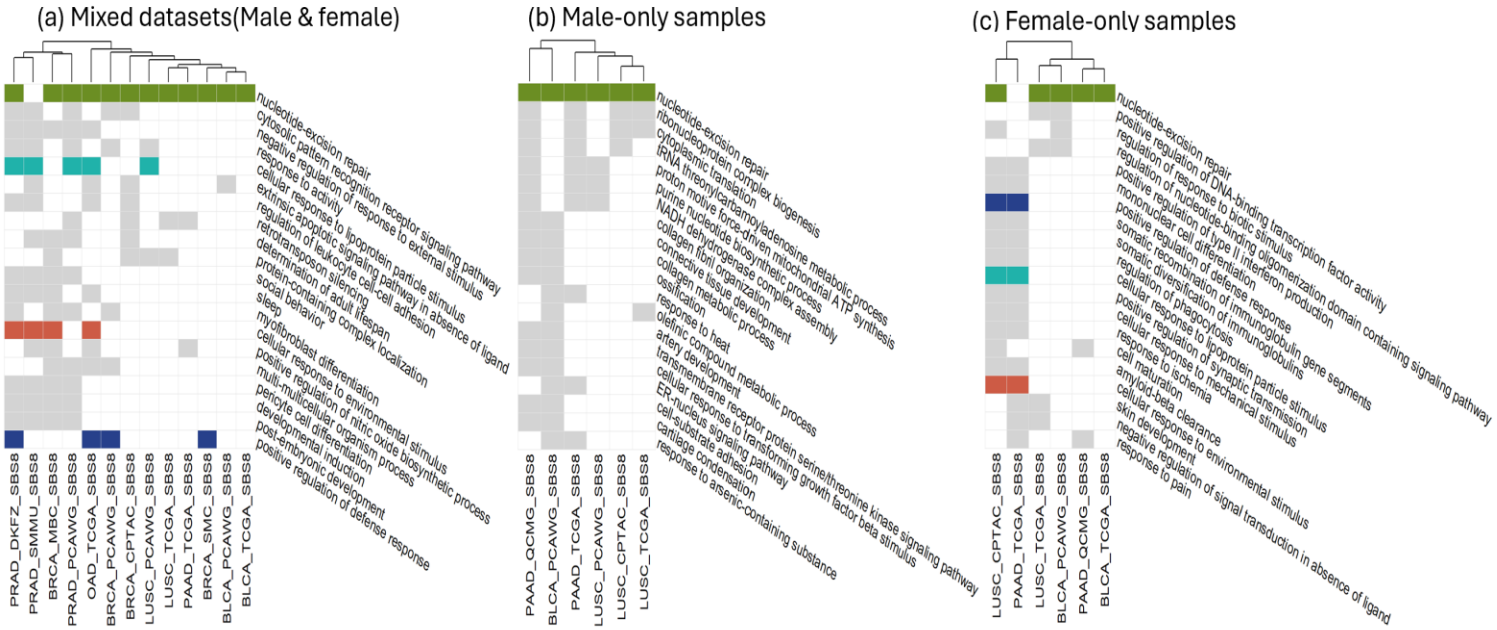


Figure 4.6: GSEA and GO analysis identified top 20 parent GO terms associated with SBS8
In the x-axis, these are all different cancer datasets and in y-axis, the top 20 parent GO terms for SBS8. The colored GO terms are commonly found across (a), (b), and (c) or between any two. The light gray GO terms are unique for (a), (b) and (c) respectively.

4.7.7 Other signatures of unknown etiology

Other signature such as SBS12, SBS19, SBS33 (37) of unknown etiology have top GO terms such as “B-cell proliferation”, “immune system development”, “somatic diversification of immune receptors” from our analysis, which suggest plausible association with immune cells in various biological processes important for understanding the potential role of these signatures (see Supp. results 1, page no. 224, 228, 233, 250, 270). Signatures such as SBS16 (unknown etiology) but with few evidence of NER (37), and SBS41 (4) of unknown etiology demonstrating top GO terms association with NER from our analysis (see Supp. Results 1, page no. 225, 240, 251, 271) may suggest the attack of endogenous or exogenous mutagenic agents damaging the DNA at a nucleotide level. SBS28 also of unknown etiology showing top GO term mismatch repair (see Supp. results 1, page no. 231, 256, 276) matches with previous few studies where SBS28 were found in POLE-deficient samples showing strong replication strand bias (152).

4.7.8 GO semantic similarity

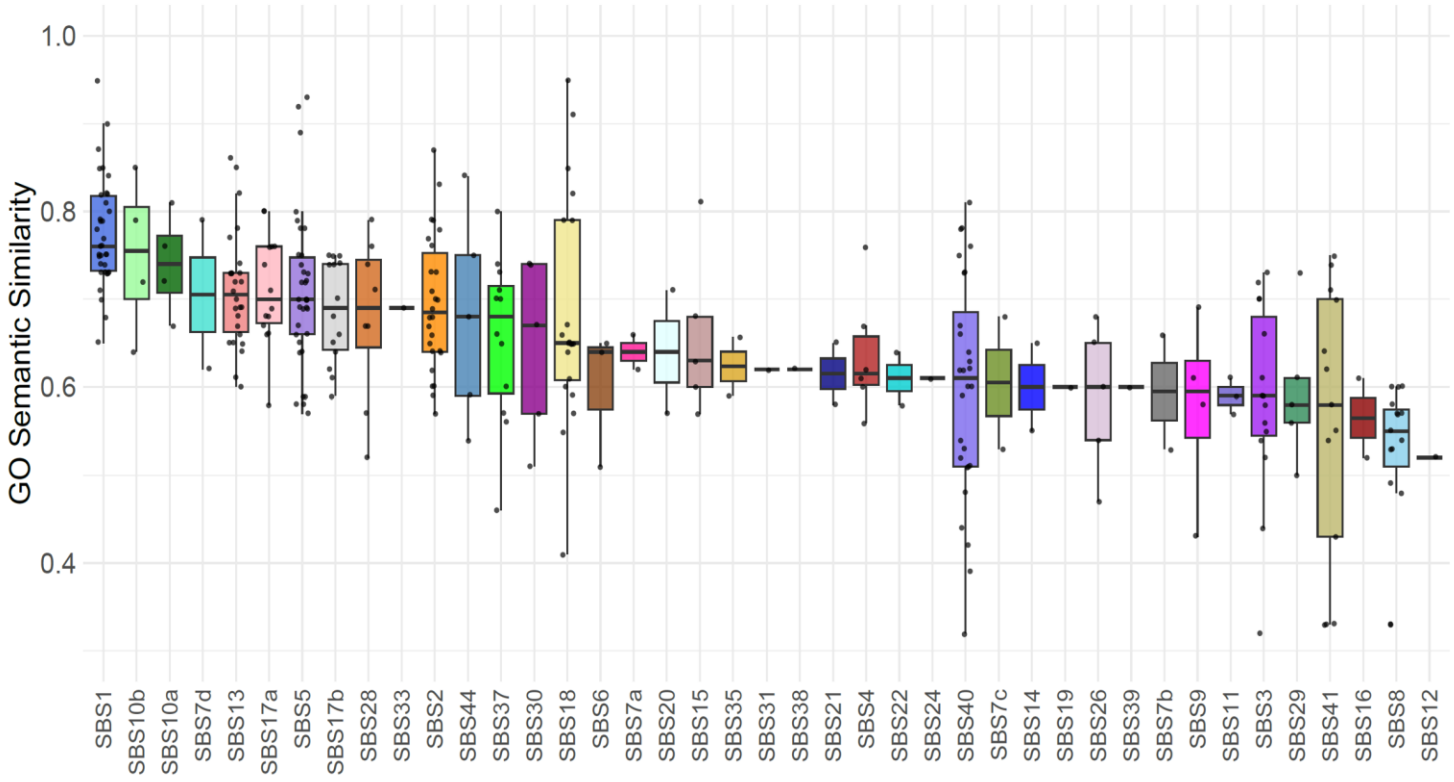


Figure 4.7: GO semantic similarity of different signatures across different cancer datasets
In the x-axis, there are different signatures, and in the y-axis, the GO semantic similarity for each signature and is ordered based on median values.

The GO semantic similarity (153) is a quantitative measure where higher value indicates higher functional similarity among the biological processes associated with each signature. In this **figure 4.7**, we have calculated the GO semantic similarity of each signature respectively present across multiple cancer datasets. The identified enriched biological process GO terms for each signature were used as input and for each set of GO terms per specific signature, pairwise semantic similarity was calculated using GOSemSim package across cancer datasets. The pairwise semantic similarity uses the best match average method between two sets of GO terms correlated to the same signature (i.e., analyzing two cancer datasets), and reduces the similarity matrix to a single value (110). The plot here shows correlation between signature and gene expression across different cancer datasets.

Pairwise GO semantic similarity median values range from 0.55 to 0.75. As we noted, the top

few GO terms are enriched in common among most datasets for each SBS, but the less frequently enriched GO terms are more variable. The semantic similarity quantifies this variability. Signatures like SBS1, SBS2, SBS5, SBS10a/b, SBS13, and SBS17a/b are among the most consistent by this analysis. Other signatures show somewhat higher variability among datasets. Possible sources of this variability are addressed below in the Discussion section.

4.7.9 Correlation between age of diagnosis and mutations attributed to signatures

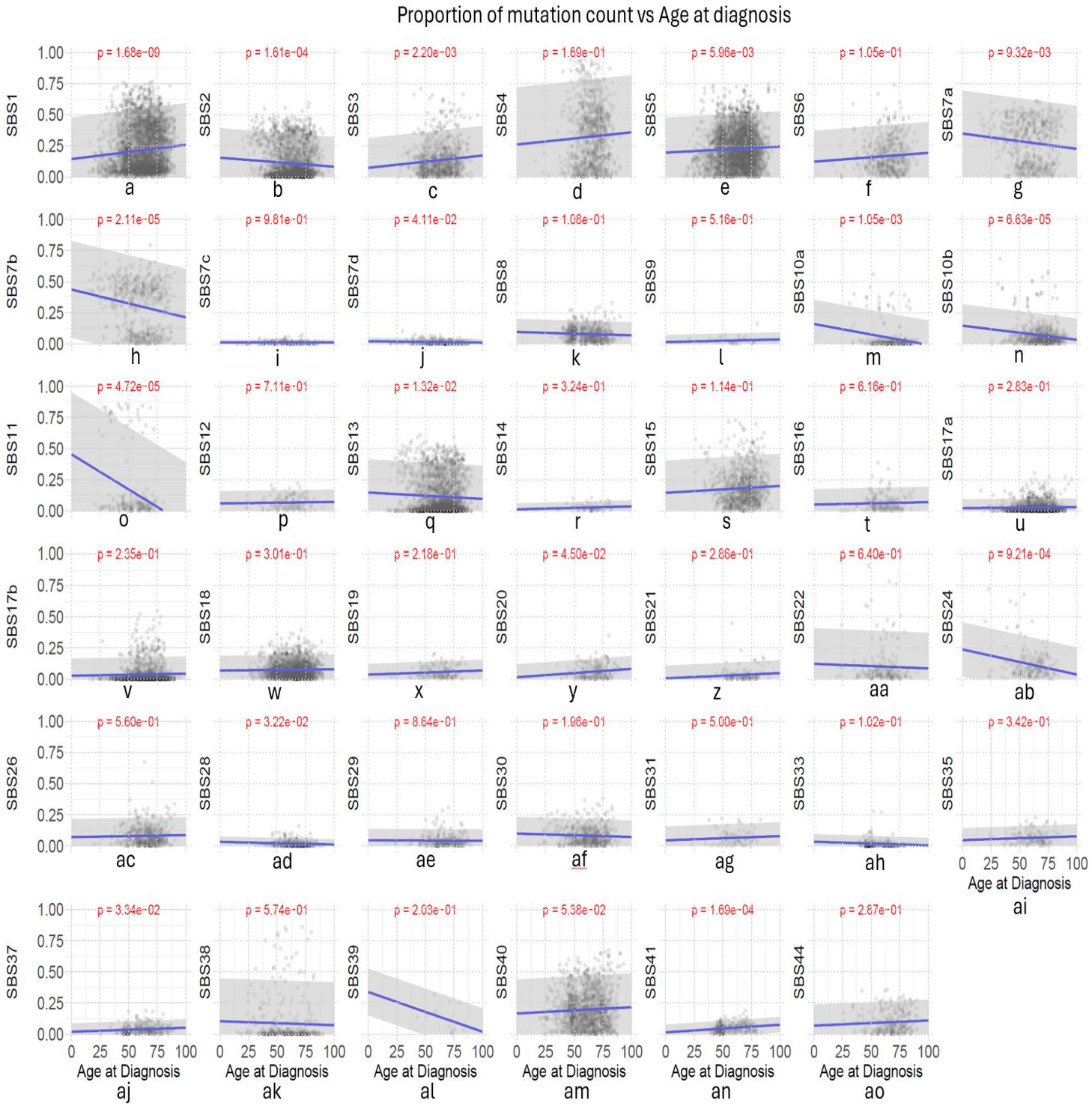


Figure 4.8: Signature-specific correlation between proportion of mutation count and age at diagnosis
The x-axis shows the age at cancer diagnosis (years), and the y-axis shows the proportion of mutation count attributed to mutational signatures respectively. Each point represents the proportion of somatic mutations for each cancer sample at a given age of diagnosis. The blue line is the fitted regression line. 95% confidence intervals are shown in light gray shading. The p-value indicates the statistical significance of the correlation. The letter at the bottom of each graph refer to each individual SBS signature.

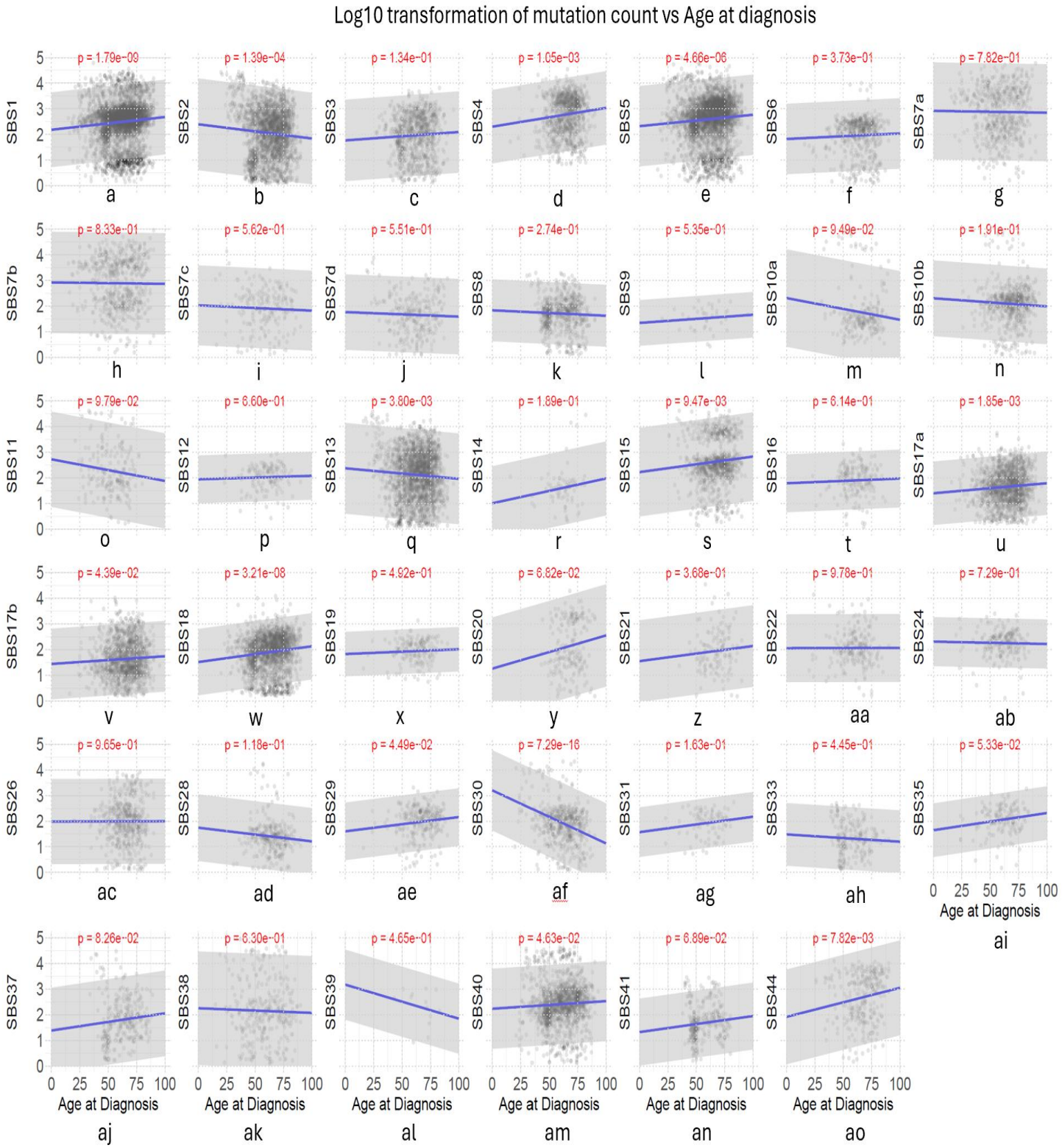


Figure 4.9: Signature-specific correlation between somatic mutations with log10 transformation and age at diagnosis

The x-axis shows the age at cancer diagnosis (years), and the y-axis shows the somatic mutations per gigabase (with log10 transformation) attributed to mutational signatures respectively. Each point represents the proportion of somatic mutations for each cancer sample at a given age of diagnosis. The blue line is the fitted regression line. 95% confidence intervals are shown in light gray shading. The p-value indicates the statistical significance of the correlation. The letter at the bottom of the each graph refer to the individual signature.

It has been shown previously that SBS1 and SBS5 mutation counts correlate with patient age at diagnosis (58). Taking advantage of larger sample numbers for more statistical power, we carried out similar analyses on the pan-cancer cBioPortal datasets.

For SBS2, the regression trend shows a negative slope, and the p-value of 1.61e-04 (**Figure 4.8, panel b**), and 1.39e-04 (**Figure 4.9, panel b**) is highly significant for proportion of mutation count and log transformation of mutation count respectively. The proportion of SBS2 mutations tends to decrease with increasing age at diagnosis. The negative correlation suggests that the APOBEC activity (the biological source of SBS2) may be at a high level in younger individuals, or younger patients may have greater APOBEC-driven mutagenesis in their tumors. For SBS13, the regression trend also shows a negative slope. The p-value of 1.32e-02 (**Figure 4.8, panel q**) and 3.80e-03 (**Figure 4.9, panel q**) for proportion of mutation count and log transformation of mutation count is still significant but somewhat less than SBS2. Like SBS2, the SBS13 mutational burden also decreases with age of diagnosis. SBS13, despite being an APOBEC-associated signature, may be at a later stage of APOBEC-driven mutagenesis, when APOBEC-induced lesions are handled by error-prone repair (154,155).

For SBS5, the regression trend shows a positive correlation between SBS5 and age of diagnosis, with a statistically significant p-value of 5.96e-03 for proportion of mutation count (**Figure 4.8, panel e**) and 4.66e-06 for log transformation of mutation count (**Figure 4.9, panel e**). This confirms the well-established classification of SBS5 as a “clock-like” mutational signature, meaning that as people age, their SBS5 mutational burden rises (58). SBS5 has a flat mutational pattern and is frequently observed alongside SBS1, another clock-like mutational signature, and is ubiquitous across cell types (134). For SBS1, the linear regression shows a

strong positive correlation for both proportion and log transformation having p-values of 1.68e-09 (**Figure 4.8, panel a**) and 1.79e-09 (**Figure 4.9, panel a**) respectively of mutation count reflecting SBS1 as a stable clock-like signature (134). For SBS40, the regression trend shows a slight positive slope, and the p-value of 5.38e-02 for proportion of mutation count (**Figure 4.8, panel am**) and 4.63e-02 for log transformation of mutation count (**Figure 4.9, panel am**) is borderline statistically significant. There is a weak positive correlation between SBS40 mutational burden and age of diagnosis, which rule out SBS40 as another bona fide clock-like mutational signature. It is often present at low or moderate levels in multiple cancer types. SBS40, despite showing some superficial similarities in shape and distribution with SBS5, did not exhibit a strong association with age, setting it apart from SBS5 (4).SBS3, associated with double-strand break via homologous recombination (HR) defect (135) shows positive regression trend of p-value 2.20e-03 for proportion of mutation count (**Figure 4.8, panel c**) which may reflect the increased HR-deficient cases (BRCA1/BRCA2 mutations) in older patients (156) but no significant positive regression trend for log transformation of mutation count with p-value of 0.134 demonstrate no time-dependent accumulation of mutation (**Figure 4.9, panel c**) suggesting no correlation of age and SBS3. Signatures such as SBS6, SBS14, SBS15, SBS20, SBS21, SBS26 and SBS44 linked with mismatch repair deficiency (MMR) (139) is a heterogenous mix of slight positive regression trend or no correlation with age either from proportion or log transformation of mutation count method (**Figure 4.8 and 4.9, panel f, r, s, y, z, ac, ao**) because MMR deficiency favors the mutations in microsatellite regions resulting from the specific loss of MMR function due to genetic and epigenetic activation in MMR genes such as MLH1, MSH2 and mutations do not steadily accumulate over time (157). SBS17a, SBS17b, SBS37 signatures of unknown etiology and ROS associated SBS18 (158), including tobacco carcinogen associated

signatures such as SBS4 and SBS29 may accumulate DNA damage but not steadily over age, so linear regression has either slight positive trend or no significant correlation with age. The slight positive correlation for SBS17a/17b, SBS18, and SBS37, and may be reflecting changes to oxidative damage to tissue environment and immune activity (159,160) and for SBS4 and SBS29 of environmental origin may be from individual variability in tobacco smoking history and tumor heterogeneity which are all context dependent (161) (**Figure 4.8 and 4.9, panel d, u, v, w, ae, aj**). SBS9 (associated with immunoglobulin gene hypermutation), SBS12, SBS16, SBS19 (35), and SBS41 of unknown etiology (4), and SBS31 and SBS35 associated with platinum drug (63) have either flat or slightly positive trend which may reflect the tumor-type specific processes or immune function and repair systems for SBS9/12/16/19/41 (162) and specific therapeutic exposures for SBS31/35 rather than time dependent accumulation of mutation (4) (**Figure 8 and 9, panel l, p, t, x, ag, ai, an**).

SBS11 is associated with temozolomide chemotherapy (140), SBS22 with aristolochic acid exposure (56), SBS24 (aflatoxin exposure) (59), SBS33 and SBS39 of unknown etiology (4) have either flat or negative correlation with age which may reflect the exposure of chemotherapy agents such as temozolomide in younger glioma patients, and the use of aristolochic and aflatoxin may be exposed early in life especially in endemic regions (163,164) and for SBS33/39 may be possibly tumor-type specific processes (**Figure 4.8 and 4.9, panel o, aa, ab, ah, al**). Signatures such as SBS10a/b (associated with POLE mutations (165)), SBS28 of unknown etiology but found in most samples with SBS10a/b (166), SBS30 associated with *NTHL1* mutations (142), and all subsets of SBS7 (SBS7a, SBS7b, SBS7c and SBS7d), and SBS38 associated with UV-damage (141) show little correlation with age (**Figure 4.8 and 4.9, panel g, h, I, j, m, n, ad, af, a**j).

4.8 Discussion

In this article, the approach that we apply identified key findings that narrow down plausible etiologies of different mutational signatures. We report the top 20 parent Gene Ontology (GO) terms associated with different signatures where SBS2, SBS8, SBS13, SBS17a, SBS17b, SBS37 and SBS40 are in the main text of paper and for other signatures such as SBS1, SBS3, SBS4, SBS5, SBS6, SBS7a, SBS7b, SBS7c, SBS7d, SBS9, SBS10a, SBS10b, SBS12, SBS14, SBS15, SBS16, SBS18, SBS19, SBS20, SBS21, SBS22, SBS24, SBS26, SBS28, SBS29, SBS30, SBS31, SBS33, SBS35, SBS37, SBS38, SBS39, SBS41, SBS44 are in the supplementary section identified in mixed cancer datasets, male samples only, and female samples only. We also identify overall common associated GO terms and the GO terms unique to males and females. In addition, the GO semantic similarity of all signatures was calculated. Furthermore, we also reported the correlation of signatures with the age of diagnosis.

The mutation profiles with transcriptome information from the same cancer samples were used to infer biological processes related to mutational signatures. To find strongly enriched GO biological process terms, we reconstructed signatures (79) and used GSEA and GO analysis by correlating each signature with gene expression (167). We additionally mapped enriched terms to their parent GO categories to eliminate repetition and improve interpretability. By using this approach, we were able to identify biological processes that are closely linked to specific mutational signatures. It is crucial to remember that mutational signature attribution uncertainty when signatures have similar shapes (168) may affect the results, even if this method offers insightful information about the possible functional impact of mutational events.

4.8.1 Comparison of signature-associated GO terms with previous studies

We identified that both SBS2 and SBS13 in 15 cancer types such as thyroid, kidney renal cell carcinoma, kidney chromophobe, pancreas, cervical, head and neck, uterine, bladder, lung adenocarcinoma, lung squamous cell carcinoma, stomach, prostate, breast, ovary, skin, and esophageal. Our analysis forwarded that multiple enriched GO terms associated with viral response pathways found across all types of cancers that has SBS2 and SBS13 reflecting the well-established role as APOBEC-induced mutational signatures (169). Our results for SBS2/13 match with previous findings and confirm that our computational approach using GSEA and GO analysis works for identifying biological process GO terms for mutational signatures.

Our analysis showed that clock-like signatures such as SBS1 and SBS5 has common GO terms associated with cell division in all 19 cancer types used for our analysis (58), SBS3 found in cancer types such as breast, lung adenocarcinoma, uterine, stomach, esophageal, head and neck, ovarian, and prostate has common GO terms associated with defective double-strand break repair via homologous recombination (HR) mainly due to mutation on BRCA1/2 genes but are not common in all cancer types (135), SBS4 and SBS29 are found across cancer types such as lung, kidney, bladder, and liver in our analysis associated with NER particularly in the context of tobacco-induced exogenous DNA damage (136), SBS9 found in breast and liver cancers was associated with somatic diversification of immunoglobulins. Signatures SBS6, SBS14, SBS15, SBS20, SBS21, SBS26, SBS44 found in colorectal, liver, pancreas, ovary, stomach, and uterine cancer from our analysis has defective MMR (170) and SBS10a, SBS10b found in colorectal and uterine arising from POLE exonuclease activity (171) has common GO term also linked to MMR. The subsets of SBS7 found in skin and head and neck cancer, and SBS38 only in skin cancer from our analysis have been found to be associated with UV damage (172), SBS18 found in pancreas, breast, prostate, cervical, esophageal, stomach, liver, ovary, head and neck, colorectal, and lung

adenocarcinoma in our analysis has enriched GO terms associated with ROS damage, SBS30 found in glioblastoma, pancreas, breast, and liver is associated with base excision repair (142). Other signatures such as SBS16 (37) found in liver and head and neck cancers, SBS22 (152) found in kidney renal cell carcinoma and liver cancer, SBS24 (173) in liver cancer, SBS31 found in liver cancer, SBS35 in ovary and liver cancer (174) found to be associated with different chemicals which may release DNA adducts so these signatures may be linked to NER. All these results are consistent with previous findings and confirm the validity of our approach.

Our analysis showed that SBS17a/17b of unknown etiology found in kidney, esophageal, breast, lung adenocarcinoma, skin, esophageal, colorectal, stomach, pancreas, head and neck, and liver has been linked to GO terms associated with ROS damage similar to previous few studies (55). Similarly, the etiology of SBS8 (found in prostate, breast, ovary, lung squamous, pancreas and bladder), and SBS41 (found in breast, stomach, prostate, ovary and stomach), and SBS37 (found in breast, prostate, and colorectal) is also unknown, and our analysis found that the SBS8 and SBS41 may be from bulky DNA adduct damage (175) at a nucleotide level, so the common enriched GO term is NER (176) and for SBS37, the enriched GO terms reflects the plausible etiology associated with immune cells (177). The etiology of SBS40 remains unknown, but our analysis in colorectal, ovary, liver, pancreas, head and neck, breast, glioblastoma, lung, prostate, cervical, liver, colorectal, thyroid, skin, bladder, kidney and breast cancer suggests a plausible link to common enriched GO terms associated with xenobiotic exposure supporting prior findings that exposure to formaldehyde may induce SBS40-like mutation patterns (1).

4.8.2 GO semantic similarity (and variability)

The GO semantic similarity analysis supports the validity of our computational framework. Certain signatures such as SBS1, SBS5, SBS2, SBS13, SBS10a, SBS10b have higher consistency in their biological associations across different cancer datasets suggesting their robust link to defined biological processes such as aging, DNA repair and immune functions (92). The alignment of our results with established etiology suggests that the workflow we used from the GSEA to GO analysis and semantic similarity can capture meaningful biological associations. Signatures such as SBS3, SBS7a, SBS7b, SBS7c, SBS7d, SBS11, SBS16, SBS17a, SBS17b, SBS18, SBS19, SBS20, SBS21, SBS22, SBS24, SBS31, SBS33, SBS35, SBS37, SBS38, SBS39, SBS40, and SBS44 showed moderate consistency suggesting their degree of tissue-specific biological processes whilst maintaining core biological processes such as DNA damage response and repair (4). In contrast, signatures SBS4, SBS6, SBS8, SBS9, SBS12, SBS14, SBS15, SBS26, SBS28, SBS29, SBS30, and SBS41, with lower semantic similarity may be reflected by biological context-dependent influenced by environmental exposure, tissue-specific differences and repair deficiencies (178). Despite this variability, our approach does identify the top GO term(s) for each SBS, which highlights the key biological processes consistently correlated with specific mutational signatures, and presumably their etiologies.

4.8.3 Age-associated trends in mutational signatures

In agreement with the previous findings by Alexandrov et al. (2015) about increased somatic mutations with age across most cancer types for clock-like signature SBS5 was supported by statistically significant p-values and 95% confidence interval of our findings for the SBS5. Although the etiology is still unknown, SBS5 may result from constant, endogenous mutational processes. In contrast, SBS40 displayed a weak positive trend with age, so it is unlikely to be a

true clock-like signature consistent with previous results (179). The etiology of SBS40 is still unknown (4), although the mutagenesis from endogenously generated formaldehyde may be a plausible source for SBS40 (1).

We also observed a significant negative correlation between mutational burden and age of diagnosis for both SBS2 and SBS13. Our results were consistent with previous results where the observed somatic mutations would decrease with age for non-clock-like signatures SBS2 and SBS13 across most cancer types and showing negative correlation (58). According to studies, tumors with viral infections, such as HPV in cervical and head-and-neck cancers, have increased APOBEC activity and are frequently identified at an earlier age (180,181). The statistical significance of SBS13 is still lower than SBS2, suggesting that SBS13 may be more stochastic or context-dependent, such as occurring in bursts or specific cancer types (125). SBS2 and SBS13 are frequently seen together, which suggests that the APOBEC activity is present or has occurred in the past.

Consistent with the findings from Alexandrov et al. (2015), our results also demonstrated that the SBS1 as a stable clock-like signature has a strong positive correlation with age. SBS3 (42) displayed a slight positive regression trend or a flat trend indicating no stable age correlation. Mismatch repair signatures such as SBS6, SBS14, SBS15, SBS20, SBS21, SBS26, SBS44 (170) and other signatures SBS17a, SBS17b of unknown etiology and SBS18 associated with ROS damage (182), including tobacco carcinogen DNA damage signatures SBS4 and SBS29, SBS9, SBS12, SBS16, SBS19, SBS31, SBS35, and SBS41 also displayed minimal or no correlation with age excluding its potential to be a clock-like signature (183). Meanwhile, SBS10a, SBS10b, SBS28, SBS8, SBS30 and the subsets of SBS7, SBS38, SBS11, SBS22, SBS24, SBS33, SBS39 showed either flat or negative correlation may be likely due to early mutagenic events or environmental

exposure or chemotherapy treatment rather than steady accumulation of mutation over age (184,185).

4.9 Future directions

All our findings can be further validated through wet-lab experiments and dry-lab analysis. In wet-lab experiments, different model organisms or human cell lines can be used for future research stimulating different mutagenic agents and conditions together with sequencing to track the mutations patterns. The understanding of mutational signatures and their associated mutational process can be further aided using CRISPR-modified cell lines (186). In the dry-lab analysis, integrating multi-omics data such as molecular and immune profiling (187) and metabolomics (188) could highlight the broader biological context of these signatures. The clinical features, tumor stage, patient exposures and treatment exposures data may further refine the etiological models of population-based assessment (189). These approaches can improve the biological interpretation of mutational signatures and strengthen their potential utility as the indicator of both endogenous and exogenous mutational processes.

Supplementary section

In the supplementary section, the results for all the signatures and their top 20 GO terms is available in supplementary results 1.

4.10 References

1. Thapa et al. Analyses of mutational patterns induced by formaldehyde and acetaldehyde reveal similarity to a common mutational signature. G3 Genes|Genomes|Genetics. 2022;12(11):jkac238.

2. Kachroo AH, Vandeloo M, Greco BM, Abdullah M. Humanized yeast to model human biology, disease and evolution. Disease models & mechanisms. 2022/06/07 ed. 2022 Jun 1;15(6).

3. Aggarwal M, Brosh RM Jr. Functional analyses of human DNA repair proteins important for aging and genomic stability using yeast genetics. DNA repair. 2012/02/22 ed. 2012 Apr 1;11(4):335–48.

4. Alexandrov, et al. The repertoire of mutational signatures in human cancer. Nature. 2020 Feb;578(7793):94–101.

5. COSMIC. Catalogue Of Somatic Mutations In cancer. 2023;

6. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Science signaling. 2013 Apr 2;6(269):pl1.

7. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences. 2005 Oct 25;102(43):15545–50.

8. Yu G. Gene ontology semantic similarity analysis using GOSemSim. Stem Cell Transcriptional Networks: Methods and Protocols. 2020;207–15.

9. Chatterjee N, Walker GC. Mechanisms of DNA damage, repair, and mutagenesis. Environ Mol Mutagen. 2017;58(5):235–63.

10. Friedberg EC, McDaniel LD, Schultz RA. The role of endogenous and exogenous DNA damage and mutagenesis. Current Opinion in Genetics & Development. 2004 Feb 1;14(1):5–10.

11. Chen P, Michel AH, Zhang J. Transposon insertional mutagenesis of diverse yeast strains suggests coordinated gene essentiality polymorphisms. Nature Communications. 2022 Mar 21;13(1):1490.

12. Bétous R, Goullet de Rugy T, Pelegrini AL, Queille S, de Villartay JP, Hoffmann JS. DNA replication stress triggers rapid DNA replication fork breakage by Artemis and XPF. PLoS Genet. 2018;14(7):e1007541.

13. Briggs AW, Stenzel U, Johnson PLF, Green RE, Kelso J, Prüfer K, et al. Patterns of damage in genomic DNA sequences from a Neandertal. Proceedings of the National Academy of Sciences. 2007;104(37):14616–21.

14. Szabó RT, Kovács-Weber M, Balogh KM, Mézes M, Kovács B. Effect of aflatoxin B1 and sterigmatocystin on DNA repair genes in common carp. Aquatic Toxicology. 2024 Nov 1;276:107076.

15. Silvani A, Eoli M, Salmaggi A, Lamperti E, Maccagnano E, Broggi G, et al. Phase II Trial of Cisplatin Plus Temozolomide, in Recurrent and Progressive Malignant Glioma Patients. Journal of Neuro-Oncology. 2004 Jan 1;66(1):203–8.

16. Mourón SA, Golijow CD, Dulout FN. DNA damage by cadmium and arsenic salts assessed by the single cell gel electrophoresis assay. Mutation Research/Genetic Toxicology and Environmental Mutagenesis. 2001 Nov 15;498(1):47–55.

17. Hakem R. DNA-damage repair; the good, the bad, and the ugly. The EMBO journal. 2008 Feb 20;27(4):589–605.

18. Ma A, Dai X. The relationship between DNA single-stranded damage response and double-stranded damage response. Cell Cycle. 2018;17(1):73–9.

19. Moeller BC, Lu K, Doyle-Eisele M, McDonald J, Gigliotti A, Swenberg JA. Determination of N2-Hydroxymethyl-dG Adducts in the Nasal Epithelium and Bone Marrow of Nonhuman Primates Following 13CD2-Formaldehyde Inhalation Exposure. Chemical research in toxicology. 2011 Feb 18;24(2):162–4.

20. Hashimoto S, Anai H, Hanada K. Mechanisms of interstrand DNA crosslink repair and human disorders. Genes and Environment. 2016 May 1;38(1):9.

21. Golato T, Brenerman B, McNeill DR, Li J, Sobol RW, Wilson DM. Development of a Cell-Based Assay for Measuring Base Excision Repair Responses. Scientific Reports [Internet]. 2017 Oct 11;7(1):13007. Available from: https://doi.org/10.1038/s41598-017-12963-7

22. Marteijn JA, Lans H, Vermeulen W, Hoeijmakers JHJ. Understanding nucleotide excision repair and its roles in cancer and ageing. Nature Reviews Molecular Cell Biology [Internet]. 2014 Jul 1;15(7):465–81. Available from: https://doi.org/10.1038/nrm3822

23. Li GM. Mechanisms and functions of DNA mismatch repair. Cell Research [Internet]. 2008 Jan 1;18(1):85–98. Available from: https://doi.org/10.1038/cr.2007.115

24. Caldecott KW. Single-strand break repair and genetic disease. Nature Reviews Genetics [Internet]. 2008 Aug 1;9(8):619–31. Available from: https://doi.org/10.1038/nrg2380

25. Huang R, Zhou PK. DNA damage repair: historical perspectives, mechanistic pathways and clinical translation for targeted cancer therapy. Signal Transduction and Targeted Therapy. 2021 Jul 9;6(1):254.

26. Krokan HE, Bjørås M. Base excision repair. Cold Spring Harbor perspectives in biology. 2013 Apr 1;5(4):a012583.

27. Marteijn et al. Understanding nucleotide excision repair and its roles in cancer and ageing. Nature reviews Molecular cell biology. 2014 Jul;15(7):465–81.

28. Li GM. Mechanisms and functions of DNA mismatch repair. Cell Research. 2008 Jan 1;18(1):85–98.

29. Abbotts R, Wilson DM. Coordination of DNA single strand break repair. Free radical biology & medicine. 2017 Jun;107:228–44.

30. Scully et al. DNA double-strand break repair-pathway choice in somatic mammalian cells. Nature Reviews Molecular Cell Biology. 2019 Nov 1;20(11):698–714.

31. Schubert L, Hendriks IA, Hertz EPT, Wu W, Sellés-Baiget S, Hoffmann S, et al. SCAI promotes error-free repair of DNA interstrand crosslinks via the Fanconi anemia pathway. EMBO reports. 2022;23(4):e53639.

32. Hao Q, Zhan C, Lian C, Luo S, Cao W, Wang B, et al. DNA repair mechanisms that promote insertion-deletion events during immunoglobulin gene diversification. Science immunology. 2023 Mar 31;8(81):eade1167.

33. Yoshioka et al. Genomic Instability and Cancer Risk Associated with Erroneous DNA Repair. Int J Mol Sci. 2021 Nov 12;22(22).

34. Izumi T. Analysis of Copy Number Variation of DNA Repair/Damage Response Genes in Tumor Tissues. Methods in molecular biology (Clifton, NJ). 2023;2701:231–42.

35. Alexandrov, et al. Deciphering signatures of mutational processes operative in human cancer. Cell Rep. 2013 Jan 31;3(1):246–59.

36. COSMIC. Catalogue Of Somatic Mutations In cancer. 2021.

37. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin AV, et al. Signatures of mutational processes in human cancer. Nature. 2013 Aug 1;500(7463):415–21.

38. Helleday et al. Mechanisms underlying mutational signatures in human cancers. Nat Rev Genet. 2014;15(9):585–98.

39. COSMIC. Catalogue Of Somatic Mutations In cancer. 2022;

40. COSMIC. Catalogue Of Somatic Mutations In cancer. 2019;

41. COSMIC. Catalogue Of Somatic Mutations In cancer. 2020;

42. Nik-Zainal S, Alexandrov LB, Wedge DC, Van Loo P, Greenman CD, Raine K, et al. Mutational processes molding the genomes of 21 breast cancers. Cell. 2012;149(5):979–93.

43. Chan K, Roberts SA, Klimczak LJ, Sterling JF, Saini N, Malc EP, et al. An APOBEC3A hypermutation signature is distinguishable from the signature of background mutagenesis by APOBEC3B in human cancers. Nature genetics. 2015;47(9):1067–72.

44. Zámborszky J, Szikriszt B, Gervai JZ, Pipek O, Póti Á, Krzystanek M, et al. Loss of BRCA1 or BRCA2 markedly increases the rate of base substitution mutagenesis and has distinct effects on genomic deletions. Oncogene. 2017 Feb 1;36(6):746–55.

45. Nik-Zainal et al. The genome as a record of environmental exposure. 2015;

46. Hodel KP, Sun MJS, Ungerleider N, Park VS, Williams LG, Bauer DL, et al. POLE Mutation Spectra Are Shaped by the Mutant Allele Identity, Its Abundance, and Mismatch Repair Status. Molecular cell. 2020 Jun 18;78(6):1166-1177.e6.

47. Drost J, van Boxtel R, Blokzijl F, Mizutani T, Sasaki N, Sasselli V, et al. Use of CRISPR-modified human stem cell organoids to study the origin of mutational signatures in cancer. Science. 2017 Oct 13;358(6360):234–8.

48. Meier B, Volkova NV, Hong Y, Schofield P, Campbell PJ, Gerstung M, et al. Mutational signatures of DNA mismatch repair deficiency in C. elegans and human cancers. Genome research. 2018 May;28(5):666–75.

49. Hayward NK, Wilmott JS, Waddell N, Johansson PA, Field MA, Nones K, et al. Whole-genome landscapes of major melanoma subtypes. Nature. 2017 May 1;545(7653):175–80.

50. Saini N, Roberts SA, Klimczak LJ, Chan K, Grimm SA, Dai S, et al. The Impact of Environmental and Endogenous Damage on Somatic Mutation Load in Human Skin Fibroblasts. PLoS Genet. 2016;12(10):e1006385.

51. Li HD, Cuevas I, Zhang M, Lu C, Alam MM, Fu YX, et al. Polymerase-mediated ultramutagenesis in mice produces diverse cancers with high mutational load. J Clin Invest. 2018 31;128(9):4179–91.

52. Kucab JE, Zou X, Morganella S, Joel M, Nanda AS, Nagy E, et al. A Compendium of Mutational Signatures of Environmental Agents. Cell. 2019;177(4):821-836.e16.

53. Otlu B, Alexandrov LB. Evaluating topography of mutational signatures with SigProfilerTopography. Genome Biology [Internet]. 2025 May 20;26(1):134. Available from: https://doi.org/10.1186/s13059-025-03612-8

54. Secrier M, Li X, de Silva N, Eldridge MD, Contino G, Bornschein J, et al. Mutational signatures in esophageal adenocarcinoma define etiologically distinct subgroups with therapeutic relevance. Nature genetics. 2016 Oct 1;48(10):1131–41.

55. Christensen S, Van der Roest B, Besselink N, Janssen R, Boymans S, Martens JWM, et al. 5-Fluorouracil treatment induces characteristic T>G mutations in human cancer. Nature Communications. 2019 Oct 8;10(1):4571.

56. Hoang M, Chen CH, Sidorenko V, He J, Dickman K, Yun BH, et al. Mutational Signature of Aristolochic Acid Exposure as Revealed by Whole-Exome Sequencing. Science translational medicine. 2013 07;5:197ra102.

57. Poon SL, Pang ST, McPherson JR, Yu W, Huang KK, Guan P, et al. Genome-wide mutational signatures of aristolochic acid and its application as a screening tool. Science translational medicine. 2013 Aug 7;5(197):197ra101.

58. Alexandrov, Jones PH, Wedge DC, Sale JE, Campbell PJ, Nik-Zainal S, et al. Clock-like mutational processes in human somatic cells. Nature genetics. 2015 Dec;47(12):1402–7.

59. Chawanthayatham S, Valentine CC, Fedeles BI, Fox EJ, Loeb LA, Levine SS, et al. Mutational spectra of aflatoxin B1 in vivo establish biomarkers of exposure for human hepatocellular carcinoma. Proceedings of the National Academy of Sciences. 2017 Apr 11;114(15):E3101–9.

60. Li L, Jiang D, Liu H, Guo C, Zhao R, Zhang Q, et al. Comprehensive proteogenomic characterization of early duodenal cancer reveals the carcinogenesis tracks of different subtypes. Nature Communications [Internet]. 2023 Mar 29;14(1):1751. Available from: https://doi.org/10.1038/s41467-023-37221-5

61. Viel A, Bruselles A, Meccia E, Fornasarig M, Quaia M, Canzonieri V, et al. A Specific Mutational Signature Associated with DNA 8-Oxoguanine Persistence in MUTYH-defective Colorectal Cancer. eBioMedicine. 2017;20:39–49.

62. Pilati C, Shinde J, Alexandrov LB, Assié G, André T, Hélias-Rodzewicz Z, et al. Mutational signature analysis identifies deficiency in colorectal cancers and adrenocortical carcinomas. The Journal of Pathology. 2017 May 1;242(1):10–5.

63. Boot A, Huang MN, Ng AWT, Ho SC, Lim JQ, Kawakami Y, et al. In-depth characterization of the cisplatin mutational signature in human cell lines and in esophageal and liver tumors. Genome research. 2018 May;28(5):654–65.

64. Inman GJ, Wang J, Nagano A, Alexandrov LB, Purdie KJ, Taylor RG, et al. The genomic landscape of cutaneous SCC reveals drivers and a novel azathioprine associated mutational signature. Nature Communications. 2018 Sep 10;9(1):3667.

65. Mimaki S, Totsuka Y, Suzuki Y, Nakai C, Goto M, Kojima M, et al. Hypermutation and unique mutational signatures of occupational cholangiocarcinoma in printing workers exposed to haloalkanes. Carcinogenesis. 2016;37(8):817–26.

66. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. Bioinformatics (Oxford, England). 2011;27(12):1739–40.

67. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell systems. 2015 Dec 23;1(6):417–25.

68. Reimand J, Isserlin R, Voisin V, Kucera M, Tannus-Lopes C, Rostamianfar A, et al. Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. Nature Protocols. 2019 Feb;14(2):482–517.

69. Marwaha et al. A guide for the diagnosis of rare and undiagnosed disease: beyond the exome. Genome Medicine. 2022 Feb 28;14(1):23.

70. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature Protocols. 2009 Jan 1;4(1):44–57.

71. Reimand J, Arak T, Adler P, Kolberg L, Reisberg S, Peterson H, et al. g:Profiler-a web server for functional interpretation of gene lists (2016 update). Nucleic Acids Res. 2016 Jul 8;44(W1):W83-9.

72. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer discovery. 2012 May;2(5):401–4.

73. Samur MK. RTCGAToolbox: A New Tool for Exporting TCGA Firehose Data. PLOS ONE. 2014;9(9):e106397.

74. Mendiratta G, Ke E, Aziz M, Liarakos D, Tong M, Stites EC. Cancer gene mutation frequencies for the U.S. population. Nature Communications. 2021 Oct 13;12(1):5961.

75. Koch CM, Chiu SF, Akbarpour M, Bharat A, Ridge KM, Bartom ET, et al. A Beginner's Guide to Analysis of RNA Sequencing Data. American journal of respiratory cell and molecular biology. 2018 Aug;59(2):145–57.

76. Haeussler M, Zweig AS, Tyner C, Speir ML, Rosenbloom KR, Raney BJ, et al. The UCSC Genome Browser database: 2019 update. Nucleic Acids Res. 2019 Jan 8;47(D1):D853-d858.

77. Jiang L, Yu H, Guo Y. Modeling the relationship between gene expression and mutational signature. Quantitative biology (Beijing, China). 2023 Mar;11(1):31–43.

78. Mayakonda A, Lin DC, Assenov Y, Plass C, Koeffler HP. Maftools: efficient and comprehensive analysis of somatic variants in cancer. Genome research. 2018 Nov;28(11):1747–56.

79. Blokzijl, et al. MutationalPatterns: comprehensive genome-wide analysis of mutational processes. Genome Medicine [Internet]. 2018 Apr 25;10(1):33. Available from: https://doi.org/10.1186/s13073-018-0539-0

80. Kanagasabai R, Choo KH, Ranganathan S, Baker CJO. A WORKFLOW FOR MUTATION EXTRACTION AND STRUCTURE ANNOTATION. Journal of Bioinformatics and Computational Biology. 2007 Dec 1;05(06):1319–37.

81. Zhao Y, Li MC, Konaté MM, Chen L, Das B, Karlovich C, et al. TPM, FPKM, or Normalized Counts? A Comparative Study of Quantification Measures for the Analysis of RNA-seq Data from the NCI Patient-Derived Models Repository. Journal of Translational Medicine. 2021 Jun 22;19(1):269.

82. Tong C, Wang X, Yu J, Wu J, Li W, Huang J, et al. Comprehensive analysis of RNA-seq data reveals the complexity of the transcriptome in Brassica rapa. BMC Genomics. 2013 Oct 7;14(1):689.

83. Li P, Piao Y, Shon HS, Ryu KH. Comparing the normalization methods for the differential analysis of Illumina high-throughput RNA-Seq data. BMC Bioinformatics. 2015 Oct 28;16(1):347.

84. Jin H, Wan YW, Liu Z. Comprehensive evaluation of RNA-seq quantification methods for linearity. BMC Bioinformatics. 2017 Mar 22;18(4):117.

85. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011 Aug 4;12(1):323.

86. Tomczak A, Mortensen JM, Winnenburg R, Liu C, Alessi DT, Swamy V, et al. Interpretation of biological experiments changes with evolution of the Gene Ontology and its annotations. Scientific reports. 2018 Mar 23;8(1):5115.

87. Yu et al. clusterProfiler: an R package for comparing biological themes among gene clusters. Omics : a journal of integrative biology. 2012 May;16(5):284–7.

88. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. The Innovation. 2021;2(3).

89. Wu Y, Chua EHZ, Ng AWT, Boot A, Rozen SG. Accuracy of mutational signature software on correlated signatures. Scientific reports. 2022 Jan 10;12(1):390.

90. Manders F, Brandsma AM, de Kanter J, Verheul M, Oka R, van Roosmalen MJ, et al. MutationalPatterns: the one stop shop for the analysis of mutational processes. BMC Genomics. 2022 Feb 15;23(1):134.

91. Pei G, Hu R, Dai Y, Zhao Z, Jia P. Decoding whole-genome mutational signatures in 37 human pan-cancers by denoising sparse autoencoder neural network. Oncogene. 2020 Jul 1;39(27):5031–41.

92. Thatikonda, Islam SMA, Autry RJ, Jones BC, Gröbner SN, Warsow G, et al. Comprehensive analysis of mutational signatures reveals distinct patterns and molecular processes across 27 pediatric cancers. Nature Cancer. 2023 Feb;4(2):276–89.

93. Zhang R, Li Q, Fu J, Jin Z, Su J, Zhang J, et al. Comprehensive analysis of genomic mutation signature and tumor mutation burden for prognosis of intrahepatic cholangiocarcinoma. BMC Cancer. 2021 Feb 3;21(1):112.

94. Chen, et al. The immune response-related mutational signatures and driver genes in non-small-cell lung cancer. Cancer science. 2019 Aug;110(8):2348–56.

95. Korotkevich G, Sukhov V, Budin N, Shpak B, Artyomov MN, Sergushichev A. Fast gene set enrichment analysis. bioRxiv : the preprint server for biology. 2021;060012.

96. Xu S, Hu E, Cai Y, Xie Z, Luo X, Zhan L, et al. Using clusterProfiler to characterize multiomics data. Nature Protocols. 2024 Nov 1;19(11):3292–320.

97. Mooney MA, Wilmot B. Gene set analysis: A step-by-step guide. American journal of medical genetics Part B, Neuropsychiatric genetics : the official publication of the International Society of Psychiatric Genetics. 2015 Oct;168(7):517–27.

98. Tiong KL, Yeang CH. MGSEA – a multivariate Gene set enrichment analysis. BMC Bioinformatics. 2019 Mar 18;20(1):145.

99. Bull C, Byrne RM, Fisher NC, Corry SM, Amirkhah R, Edwards J, et al. Dual gene set enrichment analysis (dualGSEA); an R function that enables more robust biological discovery and pre-clinical model alignment from transcriptomics data. Scientific reports. 2024 Dec 4;14(1):30202.

100. Fang Z, Liu X, Peltz G. GSEApy: a comprehensive package for performing gene set enrichment analysis in Python. Bioinformatics (Oxford, England). 2023;39(1):btac757.

101. Hong G, Zhang W, Li H, Shen X, Guo Z. Separate enrichment analysis of pathways for up- and downregulated genes. Journal of the Royal Society, Interface. 2014 Mar 6;11(92):20130950.

102. Fleming DS, Miller LC. Leading edge analysis of transcriptomic changes during pseudorabies virus infection. Genomics Data. 2016 Dec 1;10:104–6.

103. Zhou T, Yao J, Liu Z. Gene Ontology, Enrichment Analysis, and Pathway Analysis. In: Bioinformatics in Aquaculture. 2017. p. 150–68.

104. Boehmke BC. Dealing with Character Strings. In: Boehmke PDBC, editor. Data Wrangling with R. Cham: Springer International Publishing; 2016. p. 41–54.

105. Paczkowska M, Barenboim J, Sintupisut N, Fox NS, Zhu H, Abd-Rabbo D, et al. Integrative pathway enrichment analysis of multivariate omics data. Nature Communications. 2020;11(1):735.

106. Borgmästars E. Functional analysis of circulating microRNAs in pancreatic cancer. 2018.

107. Wu H, Jiang W, Ji G, Xu R, Zhou G, Yu H. Exploring microRNA target genes and identifying hub genes in bladder cancer based on bioinformatic analysis. BMC Urology. 2021 Jun 10;21(1):90.

108. Bates DM, DebRoy S. Converting a large R package to S4 classes and methods. In 2003. p. 2.

109. Huang R, Xu W, Liverani S, Hiltbrand D, Stapleton AE. A case study of r performance analysis and optimization. In: Proceedings of the Practice and Experience on Advanced Research Computing: Seamless Creativity. 2018. p. 1–6.

110. Yu G, Li F, Qin Y, Bo X, Wu Y, Wang S. GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. Bioinformatics (Oxford, England). 2010;26(7):976–8.

111. Sayols S. rrvgo: a Bioconductor package for interpreting lists of Gene Ontology terms. microPublication biology. 2023;2023.

112. Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S, et al. AmiGO: online access to ontology and annotation data. Bioinformatics (Oxford, England). 2009;25(2):288–9.

113. Li Y. GO/KEGG enrichment analysis on gene Lists from Rice (Oryza Sativa). Bio-101. 2022;12:e4446.

114. Bruixola G, Martín-Arana J, Gimeno-Valiente F, Seguí V, Catalá-Senent JF, Carbonell-Asins JA, et al. 207 ctDNA-based WES enhances actionable alterations detection in locally advanced head and neck cancer. Radiotherapy and Oncology. 2024;192:S55–8.

115. Sánchez D, Batet M. Semantic similarity estimation in the biomedical domain: An ontology-based information-theoretic perspective. Journal of Biomedical Informatics. 2011 Oct 1;44(5):749–59.

116. Heazlewood J, Durek P, Hummel J, Selbig J, Weckwerth W, Walther D, et al. PhosPhAt : A database of phosphorylation sites in Arabidopsis thaliana and a plant-specific phosphorylation site predictor. Nucleic Acids Res. 2008 01;36:D1015-21.

117. Harbig T, Paz M, Nieselt K. GO-Compass: Visual Navigation of Multiple Lists of GO terms. Computer Graphics Forum. 2023 27;42:271–81.

118. Li CH, Prokopec SD, Sun RX, Yousif F, Schmitz N, Al-Shahrour F, et al. Sex differences in oncogenic mutational processes. Nature Communications. 2020 Aug 28;11(1):4330.

119. Zhao C, Wang Z. GOGO: An improved algorithm to measure the semantic similarity between gene ontology terms. Scientific reports. 2018 Oct 10;8(1):15107.

120. Yon Rhee S, Wood V, Dolinski K, Draghici S. Use and misuse of the gene ontology annotations. Nature Reviews Genetics. 2008 Jul 1;9(7):509–15.

121. Hassan H, Shanak S. GOTrapper: a tool to navigate through branches of gene ontology hierarchy. BMC Bioinformatics. 2019 Jan 11;20(1):20.

122. Chen, et al. Establishing a consensus for the hallmarks of cancer based on gene ontology and pathway annotations. BMC Bioinformatics [Internet]. 2021 Apr 6;22(1):178. Available from: https://doi.org/10.1186/s12859-021-04105-8

123. Gene Ontology C. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res. 2004;32(suppl_1):D258–61.

124. Chen H, Chong W, Yang X, Zhang Y, Sang S, Li X, et al. Age-related mutational signature negatively associated with immune activity and survival outcome in triple-negative breast cancer. Oncoimmunology. 2020 Jun 30;9(1):1788252.

125. Petljak et al. Characterizing Mutational Signatures in Human Cancer Cell Lines Reveals Episodic APOBEC Mutagenesis. Cell. 2019 Mar 7;176(6):1282-1294.e20.

126. Mussil B, Suspène R, Aynaud MM, Gauvrit A, Vartanian JP, Wain-Hobson S. Human APOBEC3A Isoforms Translocate to the Nucleus and Induce DNA Double Strand Breaks Leading to Cell Stress and Death. PLOS ONE. 2013;8(8):e73641.

127. Feng et al. Deamination-independent restriction of LINE-1 retrotransposition by APOBEC3H. Scientific reports. 2017 Sep 7;7(1):10881.

128. Schutsky et al. APOBEC3A efficiently deaminates methylated, but not TET-oxidized, cytosine bases in DNA. Nucleic Acids Res. 2017 Jul 27;45(13):7655–65.

129. von Wronski MA, Hirano KI, Cagen LM, Wilcox HG, Raghow R, Thorngate FE, et al. Insulin increases expression of apobec-1, the catalytic subunit of the apolipoprotein B mRNA editing complex in rat hepatocytes. Metabolism. 1998 Jul 1;47(7):869–73.

130. Smith NJ, Reddin I, Policelli P, Oh S, Zainal N, Howes E, et al. Differentiation signals induce APOBEC3A expression via GRHL3 in squamous epithelia and squamous cell carcinoma. The EMBO journal. 2025 Jan;44(1):1–29.

131. Mehta et al. IFN-α and lipopolysaccharide upregulate APOBEC3 mRNA through different signaling pathways. J Immunol. 2012;189(8):4088–103.

132. Okegawa et al. The role of cell adhesion molecule in cancer progression and its application in cancer therapy. Acta biochimica Polonica. 2004;51(2):445–57.

133. Jablonka et al. Commentary: Induction and selection of variations during cancer development. International Journal of Epidemiology. 2006;35(5):1163–5.

134. Spisak et al. Disentangling sources of clock-like mutations in germline and soma. bioRxiv : the preprint server for biology. 2023 Sep 12;

135. Wu et al. Mutational signature assignment heterogeneity is widespread and can be addressed by ensemble approaches. Briefings in Bioinformatics. 2023 Sep 22;24(6).

136. Ernst et al. Tobacco Smoking-Related Mutational Signatures in Classifying Smoking-Associated and Nonsmoking-Associated NSCLC. Journal of Thoracic Oncology. 2023;18(4):487–98.

137. Torrens L, Moody S, de Carvalho AC, Kazachkova M, Abedi-Ardekani B, Cheema S, et al. The complexity of tobacco smoke-induced mutagenesis in head and neck cancer. Nature Genetics [Internet]. 2025 Apr 1;57(4):884–96. Available from: https://doi.org/10.1038/s41588-025-02134-0

138. van den Heuvel GRM, Kroeze LI, Ligtenberg MJL, Grünberg K, Jansen EAM, von Rhein D, et al. Mutational signature analysis in non-small cell lung cancer patients with a high tumor mutational burden. Respiratory Research [Internet]. 2021 Nov 24;22(1):302. Available from: https://doi.org/10.1186/s12931-021-01871-0

139. Lózsa, Németh E, Gervai JZ, Márkus BG, Kollarics S, Gyüre Z, et al. DNA mismatch repair protects the genome from oxygen-induced replicative mutagenesis. Nucleic Acids Res. 2023 Nov 10;51(20):11040–55.

140. Crisafulli G, Sartore-Bianchi A, Lazzari L, Pietrantonio F, Amatu A, Macagno M, et al. Temozolomide Treatment Alters Mismatch Repair and Boosts Mutational Burden in Tumor and Blood of Colorectal Cancer Patients. Cancer discovery. 2022 Jul 6;12(7):1656–75.

141. Mundra, Dhomen N, Rodrigues M, Mikkelsen LH, Cassoux N, Brooks K, et al. Ultraviolet radiation drives mutations in a subset of mucosal melanomas. Nature Communications. 2021 Jan 11;12(1):259.

142. Sangiorgi et al. Base-Excision Repair Mutational Signature in Two Sebaceous Carcinomas of the Eyelid. Genes. 2023 Nov 8;14(11).

143. Vijayraghavan S, Saini N. Aldehyde-Associated Mutagenesis–Current State of Knowledge. Chemical research in toxicology. 2023 Jul 17;36(7):983–1001.

144. Lv J, Zhang C, Liu X, Gu C, Liu Y, Gao Y, et al. An aging-related immune landscape in the hematopoietic immune system. Immunity & Ageing. 2024 Jan 2;21(1):3.

145. Ma F, Wang S, Xu L, Huang W, Shi G, Sun Z, et al. Single-cell profiling of the microenvironment in human bone metastatic renal cell carcinoma. Communications Biology. 2024 Jan 12;7(1):91.

146. Marchais A, Marques da Costa ME, Job B, Abbas R, Drubay D, Piperno-Neumann S, et al. Immune Infiltrate and Tumor Microenvironment Transcriptional Programs Stratify Pediatric Osteosarcoma into Prognostic Groups at Diagnosis. Cancer Research [Internet]. 2022 Mar 15 [cited 2025 Aug 9];82(6):974–85. Available from: https://doi.org/10.1158/0008-5472.CAN-20-4189

147. Silva MT, Correia-Neves M. Neutrophils and Macrophages: the Main Partners of Phagocyte Cell Systems. Frontiers in Immunology [Internet]. 2012;Volume 3-2012. Available from: https://www.frontiersin.org/journals/immunology/articles/10.3389/fimmu.2012.00174

148. Van et al. Functional states of myeloid cells in cancer. Cancer cell. 2023 Mar 13;41(3):490–504.

149. Lam P ying, Huttenlocher A. Interstitial leukocyte migration in vivo. Current Opinion in Cell Biology [Internet]. 2013 Oct 1;25(5):650–8. Available from: https://www.sciencedirect.com/science/article/pii/S0955067413000811

150. Zhao Y, Ting KK, Coleman P, Qi Y, Chen J, Vadas M, et al. The Tumour Vasculature as a Target to Modulate Leucocyte Trafficking. Cancers (Basel). 2021 Apr 6;13(7).

151. Amgalan, et al. Influence network model uncovers relations between biological processes and mutational signatures. Genome Medicine. 2023 Mar 6;15(1):15.

152. Otlu B, Díaz-Gay M, Vermes I, Bergstrom EN, Zhivagui M, Barnes M, et al. Topography of mutational signatures in human cancer. Cell Rep. 2023 Aug 29;42(8):112930.

153. Pesquita. Semantic Similarity in the Gene Ontology. Methods in molecular biology (Clifton, NJ). 2017;1446:161–73.

154. McGranahan N, Favero F, de Bruin EC, Birkbak NJ, Szallasi Z, Swanton C. Clonal status of actionable driver events and the timing of mutational processes in cancer evolution. Science translational medicine. 2015 Apr 15;7(283):283ra54-283ra54.

155. Nik-Zainal S, Van Loo P, Wedge DC, Alexandrov LB, Greenman CD, Lau KW, et al. The Life History of 21 Breast Cancers. Cell. 2012 May 25;149(5):994–1007.

156. Supek F, Ferrer-Torres P. Joint inference of mutational signatures from indels and single-nucleotide substitutions reveals prognostic impact of homologous recombination deficiency in tumors. 2024;

157. Moufarrij S, Gazzo A, Rana S, Selenica P, Abu-Rustum NR, Ellenson LH, et al. Concurrent POLE hotspot mutations and mismatch repair deficiency/microsatellite instability in endometrial cancer: A challenge in molecular classification. Gynecologic Oncology. 2024 Dec 1;191:1–9.

158. Abbas, et al. Mutational signature dynamics shaping the evolution of oesophageal adenocarcinoma. Nature Communications. 2023 Jul 15;14(1):4239.

159. Manaka Y, Kusumoto-Matsuo R, Matsuno Y, Asai H, Yoshioka K ichi. Single base substitution signatures 17a, 17b, and 40 are induced by γ-ray irradiation in association with increased reactive oxidative species. Heliyon. 2024;10(6).

160. Shen, Kamath-Loeb AS, Kohrn BF, Loeb KR, Preston BD, Loeb LA. A high-resolution landscape of mutations in the BCL6 super-enhancer in normal human B cells. Proceedings of the National Academy of Sciences of the United States of America. 2019 Dec 3;116(49):24779–85.

161. Kim YA, Hodzic E, Amgalan B, Saslafsky A, Wojtowicz D, Przytycka TM. Mutational signatures as sensors of environmental exposures: analysis of smoking-induced lung tissue remodeling. Biomolecules. 2022;12(10):1384.

162. Gavarró LM, Couturier DL, Markowetz F. A Dirichlet-multinomial mixed model for determining differential abundance of mutational signatures. BMC Bioinformatics. 2025;26:59.

163. Lukinich-Gruia AT, Nortier J, Pavlović NM, Milovanović D, Popović M, Drăghia LP, et al. Aristolochic acid I as an emerging biogenic contaminant involved in chronic kidney diseases: A comprehensive review on exposure pathways, environmental health issues and future challenges. Chemosphere. 2022;297:134111.

164. Pickova D, Ostry V, Toman J, Malir F. Aflatoxins: History, significant milestones, recent data on their toxicity and ways to mitigation. Toxins. 2021;13(6):399.

165. Huang et. Gene Mutational Clusters in the Tumors of Colorectal Cancer Patients With a Family History of Cancer. Frontiers in oncology. 2022;12:814397.

166. Singh et al. Mutational signature SBS8 predominantly arises due to late replication errors in cancer. Communications Biology. 2020 Aug 3;3(1):421.

167. Chang LY, Lee MZ, Wu Y, Lee WK, Ma CL, Chang JM, et al. Gene set correlation enrichment analysis for interpreting and annotating gene expression profiles. Nucleic Acids Res. 2024;52(3):e17–e17.

168. Senkin S. MSA: reproducible mutational signature attribution with confidence based on simulations. BMC Bioinformatics. 2021 Nov 4;22(1):540.

169. Roberts SA, Lawrence MS, Klimczak LJ, Grimm SA, Fargo D, Stojanov P, et al. An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. Nature genetics. 2013 Sep 1;45(9):970–6.

170. Fang et al. Deficiency of replication-independent DNA mismatch repair drives a 5-methylcytosine deamination mutational signature in cancer. Science Advances. 2021 Nov 5;7(45):eabg4398.

171. Ma, Riaz N, Samstein RM, Lee M, Makarov V, Valero C, et al. Functional landscapes of POLE and POLD1 mutations in checkpoint blockade-dependent antitumor immunity. Nature genetics. 2022 Jul;54(7):996–1012.

172. Zhivagui, Hoda A, Valenzuela N, Yeh YY, Dai J, He Y, et al. DNA damage and somatic mutations in mammalian cells after irradiation with a nail polish dryer. Nature Communications. 2023 Jan 17;14(1):276.

173. Bai J, Ma K, Xia S, Geng R, Shen C, Jiang L, et al. Pan-cancer mutational signature surveys correlated mutational signature with geospatial environmental exposures and viral infections. Computational and Structural Biotechnology Journal. 2023 Jan 1;21:5413–22.

174. Pich O, Muiños F, Lolkema MP, Steeghs N, Gonzalez-Perez A, Lopez-Bigas N. The mutational footprints of cancer therapies. Nature genetics. 2019 Dec;51(12):1732–40.

175. Yimit et al. Differential damage and repair of DNA-adducts induced by anti-cancer drug cisplatin across mouse organs. Nature Communications. 2019 Jan 18;10(1):309.

176. Cai, et al. Nucleotide excision repair efficiencies of bulky carcinogen-DNA adducts are governed by a balance between stabilizing and destabilizing interactions. Biochemistry. 2012 Feb 21;51(7):1486–99.

177. Luster et al. Immune cell migration in inflammation: present and future therapeutic targets. Nature Immunology. 2005 Dec 1;6(12):1182–90.

178. Peng, Zhang X, Hui W, Lu J, Li Q, Liu S, et al. Improving the measurement of semantic similarity by combining gene ontology and co-functional network: a random walk based approach. BMC Systems Biology. 2018 Mar 19;12(Suppl 2):18.

179. Creaney J, Patch AM, Addala V, Sneddon SA, Nones K, Dick IM, et al. Comprehensive genomic and tumour immune profiling reveals potential therapeutic targets in malignant pleural mesothelioma. Genome Medicine. 2022 May 30;14(1):58.

180. Zapatka M, Borozan I, Brewer DS, Iskar M, Grundhoff A, Alawi M, et al. The landscape of viral associations in human cancers. Nature genetics. 2020 Mar 1;52(3):320–30.

181. Wang, Robinson PS, Coorens THH, Moore L, Lee-Six H, Noorani A, et al. APOBEC mutagenesis is a common process in normal human small intestine. Nature genetics. 2023 Feb;55(2):246–54.

182. Zhivagui M, Hoda A, Valenzuela N, Yeh YY, Dai J, He Y, et al. DNA damage and somatic mutations in mammalian cells after irradiation with a nail polish dryer. Nature Communications [Internet]. 2023 Jan 17;14(1):276. Available from: https://doi.org/10.1038/s41467-023-35876-8

183. Ren. Somatic mutations in aging and disease. GeroScience. 2024 Oct;46(5):5171–89.

184. Cédri, Suurenbroek LC, Kleisman MM, Antić Ž, Lelieveld SH, Yeong M, et al. Mutational mechanisms in multiply relapsed pediatric acute lymphoblastic leukemia. Leukemia. 2024 Nov 1;38(11):2366–75.

185. Choi, Holowatyj AN, Du M, Chen Z, Wen W, Schultz N, et al. Distinct Genomic Landscapes in Early-Onset and Late-Onset Endometrial Cancer. JCO precision oncology. 2022 Feb;6:e2100401.

186. Chehelgerdi M, Chehelgerdi M, Khorramian-Ghahfarokhi M, Shafieizadeh M, Mahmoudi E, Eskandari F, et al. Comprehensive review of CRISPR-based gene editing: mechanisms, challenges, and applications in cancer therapy. Molecular Cancer [Internet]. 2024 Jan 9;23(1):9. Available from: https://doi.org/10.1186/s12943-023-01925-5

187. Lyons YA, Wu SY, Overwijk WW, Baggerly KA, Sood AK. Immune cell profiling in cancer: molecular approaches to cell-specific identification. npj Precision Oncology [Internet]. 2017 Aug 15;1(1):26. Available from: https://doi.org/10.1038/s41698-017-0031-0

188. Taunk K, Jajula S, Bhavsar PP, Choudhari M, Bhanuse S, Tamhankar A, et al. The prowess of metabolomics in cancer research: current trends, challenges and future perspectives. Molecular and Cellular Biochemistry [Internet]. 2025 Feb 1;480(2):693–720. Available from: https://doi.org/10.1007/s11010-024-05041-w

189. Shieh Y, Eklund M, Sawaya GF, Black WC, Kramer BS, Esserman LJ. Population-based screening for cancer: hope and hype. Nature Reviews Clinical Oncology [Internet]. 2016 Sep 1;13(9):550–65. Available from: https://doi.org/10.1038/nrclinonc.2016.50

Chapter 5: Conclusion and Future directions

5.1 Conclusion

My doctoral thesis begins with a literature survey on two small yet reactive aldehydes, AA and FA, which are generated endogenously and exogenously via various sources, including tobacco smoking, alcohol consumption, and industrial emissions (1–4). According to IARC, both of these aldehydes are classified as human carcinogens and are recognized as genotoxic compounds, but the mutagenic properties were not well documented (5,6). In our first published review paper, "The mutagenic properties of formaldehyde and acetaldehyde: Reflections on half a century of progress", I summarized all the studies in the last 50 years related to a broad spectrum of AA- and FA-induced DNA damage, such as DNA adducts, interstrand and intrastrand crosslinks, DPCs, single- and double-strand breaks, and predominance of G->T mutations. The collective efforts by many research groups on diverse model organisms and analytical methods yielded a broad consensus on the genotoxicity and mutagenicity of both AA and FA (7). But the genomic sequencing of AA- and FA-treated cells to find mutational signatures in normal and malignant cells was still inconclusive (8,9). Further investigations are required, integrating additional experimental platforms and multi-omics analysis to understand DNA damage and repair mechanisms in the mutational landscape, including carcinogenic potential, to uncover the role of these weak AA and FA aldehydes. Thus, this review paper (7) sets out a clear experimental challenging question to characterize AA- and FA-induced mutational signatures.

To answer this question, I used *Saccharomyces cerevisiae* to produce a large number of mutations in a single-stranded DNA (10). Using temperature-sensitive yeast reporter strain

ySR127, mutational patterns were identified from both AA and FA (11). These mutational patterns have predominantly C/G -> A/T, C/G -> T/A, and T/A -> C/G single base substitutions, where mostly C/G -> A/T mutations occurred at TC/GA motifs. Furthermore, we also observed that C/G -> A/T mutations were high in AA- and FA-induced mutational patterns compared to mock-treated controls. We found two distinct mutational patterns from AA- and FA-induced yeast genomics. The FA-induced mutational signature resembles SBS40, a COSMIC signature of unknown etiology and the third most common signature found in one-third of cancer types (12,13) whereas the AA-induced mutational signature did not resemble closely any COSMIC signature. But the AA treatment caused deletion events longer than 5 bases, while FA treatment did not. These striking features of both AA- and FA-induced mutational signatures have been published in the paper "Analyses of mutational patterns induced by formaldehyde and reveal similarity to a common mutational signature" (14). The use of yeast genetics to derive AA- and FA-induced mutational signatures with high confidence raised important questions. If the characterization of AA- and FA-induced mutational signatures is possible in the eukaryotic yeast genome, then why not in the human cancer genome enriched with all these mutational signatures? Also, many of these mutational signatures found in the COSMIC database do not have known etiology (12,13,15,16), which highlights the gaps that cannot be answered by mutational catalogs only.

Using RNA-seq transcriptomics data along with mutational genomics data to characterize the less-understood mutational signatures and signatures of unknown etiology with their associated biological themes was the next logical step to be followed. The computational framework that I used for reconstruction of mutational signatures to GSEA and GO enrichment analysis (17,18),

(19,20) was validated from the results that I produced for signatures of known etiology. All my results for mutational signatures of known etiology having common GO terms in mixed datasets (male and female), male-only, and female-only datasets showed consensus with previous results (13,16,21,22).

In total, out of 41 mutational signatures, 29 signatures of known etiology were consistent with the results from my analysis, and the remaining 12 signatures of unknown etiology were also analyzed to discover correlated biological processes, which narrows down possible etiologies. This computational framework establishes an integrative pipeline to bridge the gaps between mutational signatures and biological pathways. In **Table 5.1**, most of the mutational signatures that demonstrated a link to DNA repair pathways because the mutational patterns are shaped by the initial DNA damage and then by the different DNA repair pathways, making these DNA repair GO terms rise on top (23). When repair pathways such as homologous recombination, mismatch, and base-excision repair are defective, and fail to repair DNA damage, they generate distinctive mutational patterns found in SBS3 (22), SBS6 (24,25), and SBS30 (26–28), respectively. Other mutational signatures have specific etiologies. It is widely understood that the mutational signatures indicate the root cause of mutational processes by exogenous and endogenous mutagenic agents, responsible for the DNA damage and original DNA lesion, so all subsets of SBS7 (SBS7a/7b/7c/7d) mutational patterns reflect the DNA damage caused by UV rays (27,29,30). The similarity in the mutational patterns for different cancer types may lead to use of similar treatments. Poly(ADP-ribose) polymerase inhibitors, drugs against *BRCA1/BRCA2* mutants responsible for SBS3 mutational patterns, may be used in all cancer types having the SBS3 mutational signature (31). Mutational signatures such as SBS1/5, SBS2/13, SBS10a/10b,

and SBS17a/17b from different cancer datasets have high semantic similarity (32), and clustering together with their respective subtypes may further validate our computational pipeline to measure the mutational signatures associations with biological pathways. In addition, the correlation between age and mutational signatures demonstrated the potential of my computational framework to compare between previously reported time-dependent mutational processes in SBS1 and SBS5 versus other mutational patterns induced endogenously or exogenously by mutagenic agents (27).

The 12 signatures of unknown etiology that I worked on (SBS5, SBS8, SBS16, SBS41, SBS28, SBS12, SBS19, SBS33, SBS37, SBS17a/17b, SBS40) yielded results related to existing literature on these signatures. Previous studies showed early replication time bias for SBS5 across multiple tissues due to continuous low-level DNA damage and DNA repair errors during S-phase (33), suggesting a possible link to the upregulation of GO terms related to cell division from my analysis. Another study done in soft tissue sarcomas demonstrated that SBS5, found in homologous recombination repair deficiency sarcoma cells lacking RAD1 formation (a key player for homologous recombination repair) after DNA damage were linked to biological events involved in G2/M checkpoints and cell-cycle linked targets of MYC and E2F transcription factors (34). The cell cycle biological pathways for SBS5 are just a part of a broader picture. Previous studies done in breast, urothelial, and lymphoma cell lines showed that the low-fidelity REV1-dependent TLS has been associated with SBS5 (35), and in budding yeast, sugar metabolism and error-prone TLS reflect the SBS5-linked mutagenesis (36). TLS DNA polymerase kappa is operative in non-dividing neurons (37), and TLS DNA polymerase zeta is mutagenic in quiescent yeast cells (38). The basal low-level DNA damage is found in all cell types, either in normal or

tumor cells, and in dividing or non-dividing cells, including post-mitotic cells like neurons and oocytes, reflecting the ubiquitous and clock-like nature of SBS5, like SBS1 (27,39). The subsets of SBS17 (17a/17b) are associated with ROS based on my analysis. Previous findings supporting my analysis demonstrated that the ϒ-ray irradiation caused ROS formation leading to SBS17a/17b-like mutational patterns in the mouse embryonic fibroblast cells when the mutational patterns were compared between unirradiated and irradiated cells (40). In a few other studies, 5-fluorouracil was linked to ROS production (41,42), and when treating intestinal organoids followed by whole genome sequencing, the induced mutational patterns were SBS17-like (43). The linking of SBS40 with xenobiotics exposure from my analysis has been backed by previous findings. One of previous the findings is in Chapter 2, where FA-treated yeast cells generated mutational patterns close to SBS40 (14). The tris(chloropropyl) phosphate-treated B6C3F1/N mice develop liver cancer, and whole-exome sequencing of those tumors has SBS40-like mutational signatures (44). In another studies done in proteogenic cohorts of multiple cancer types, having a high SBS40 mutational burden showed a link to detoxifying enzymes creating chances of correlation between SBS40 and xenobiotics exposure (45–48).

SBS8, SBS16, and SBS41 were linked to NER, while SBS28 was linked to MMR from my analysis. Previous studies showed that SBS8 mutational pattern was enriched in late heterochromatin regions where late replication errors persist in NER-deficient tumor genomes (49). In another study, NER-deficient mice generated a mutational signature resembling SBS8, and in breast tumor genomes, those predicted to be NER-deficient tumors have SBS8 mutational pattern, supporting my current analysis of the SBS8 biological link (50). SBS16 has been linked to transcription-coupled DNA damage (51), where RNA polymerase II recruits NER on the active

transcribed strand to repair bulky DNA adducts (52,53). Moreover, previous studies suggested a link of SBS16 with inefficient NER involved in African American esophageal cancers, providing support to my results on SBS16 (54). Recent studies suggested SBS41-like mutational patterns derived from the endogenous colibactin genotoxic compound found in *E. coli* after the inactivation of NER (55,56). This result has been backed by the similar DNA damage at the nucleotide level in colorectal cancers (57), supporting my interpretation of SBS41. SBS28-linked mutational pattern has replication strand bias in POLE-deficient samples of colorectal, uterine, lung, and stomach cancers, impairing the proofreading function of polymerase linked with the mismatch repair mechanism (56). In another study in *S. cerevisiae*, POLE mutations resulting in SBS28-like mutational pattern has been associated with mismatch repair, lending support to my findings of MMR for SBS28 (58).

Signatures SBS12, SBS19, SBS33, and SBS37 have been linked with immune function from my analysis. Both SBS12 and SBS19 demonstrated high transcriptional-strand bias with mutations on the transcribed strand (56). SBS12 appears to contribute less than 20% of mutations to liver cancer and displays higher T-cell infiltration in response to programmed death 1 (PD-1) blockade in a few studies (59), whereas SBS19, linked with carcinogenic cobalt, modestly increases the mutational load, changing the tumor immune and inflammation microenvironment in mouse models, further strengthening my findings related to SBS12 and SBS19 (60). SBS33 is one of the sporadic signatures found in few samples (61) and has the thymine-specific strand bias (21), where AID off-targeting may deaminate on the non-template strand of transcribed DNA, where C->U events are fixed as C->T mutations on the untranscribed strand in the germinal center B cells (62), providing preliminary support to my analysis. SBS37 has been associated with AID

activity in long-read and circulating tumor DNA in multiple myeloma and diffuse large B-cell lymphoma cases where a high percentage of mutation was found in immunoglobulin loci (63). In another study, the mutation in the *BCL6* super enhancer locus in human B cells, an off-target of somatic hypermutation, has patterns close to SBS37, which may reinforce its division in the immune-linked signature consistent with my results (64).

Thus, the trajectory of my thesis from the literature review on the mutagenic properties of AA and FA (7) to yeast mutagenesis for generating AA- and FA-induced mutational signatures (14) to GSEA and GO analysis of cancer datasets using mutation and RNA-seq data demonstrates that all those experimental designs for yeast mutagenesis to the computational framework used for analyzing yeast genomics and human genomics, including human transcriptomics of cancer datasets, may have been able to interpret the biological processes associated with each mutational signature (17,32). Together, all my findings on mutational signatures and their associated etiology may be crucial for understanding the connection between mutational patterns and specific environmental exposures, inherited defects, or biological processes enabling better cancer prevention, diagnosis, and targeted treatments (65).

Signatures	Plausible Etiology
SBS3, SBS4, SBS6, SBS8, SBS10a/10b, SBS11, SBS14, SBS15, SBS16, SBS20, SBS21, SBS22, SBS24, SBS26, SBS28, SBS29, SBS30, SBS31, SBS35, SBS39, SBS41, SBS44	DNA repair
SBS1/5	Cell division
SBS2/13, SBS9, SBS12, SBS19, SBS33, SBS37	Immune function
SBS7a/7b/7c/7d, SBS38	UV light response
SBS17a/17b, SBS18	ROS response
SBS40	Xenobiotics exposure

Table 5.1: Mutational Signatures with plausible etiology
The blue-colored signatures are of unknown etiology and the remaining signatures are of known etiology. Maximum signatures have DNA repair as their plausible etiology and other have specific etiology.

5.2 Overall Summary

The course of my research started with understanding the carcinogenicity and mutagenicity of two aldehydes, FA and AA over half decades (7). But the results of FA and AA induced mutational signatures were still inconclusive (8,9), so we introduced a haploid yeast genetics system based on single stranded mutagenesis to generate the FA and AA-induced mutational signatures. Interestingly, FA-induced mutational signature resembles COSMIC signature SBS40 of unknown etiology, whereas AA treatment shows an excess of deletion events longer than 5 bp characterizing both aldehydes (14). Building on these findings, I extended my study from yeast to humans because mutations in tumor cells are much more complex and diverse. The integrated approach of two computational GSEA and GO analysis was used to analyze human

mutation, and RNA-seq expression data to unfold the biological processes associated with each of the signature. Altogether, this workflow provided a bridge between mutational patterns associated with environmental exposure of aldehydes such as FA and AA in yeast to interpretable biological pathways found in humans.

5.3 Future directions

Going forward, to limit the biological gap between humans and yeast, we may use liquid-biopsy samples at three different stages, before and after treatment, and at the relapse stage to see how the sets of mutational signatures in different tumor cells emerge before the treatment, fade, or switch into different mutational signatures under the influence of chemotherapy (66). In addition, coupling it with single-cell and spatial genomics sequencing may track the distinct mutational patterns resulting from DNA damage and repair mechanisms in the tumor microenvironment (67,68). The haploid yeast may be replaced with human near-haploid HAP1 cells (69) and patient-derived organoids, and murine xenografts to generate thousands of mutations to strengthen the intensity of various mutagenic agents derived mutational spectra close to COSMIC signatures and test the drug vulnerabilities. The integration of epigenetics landscape with mutational landscape further expands the knowledge of mutation patterns, because epigenetic modifications such as DNA methylation, histone modifications, and chromosome accessibility influences DNA repair pathways shaping different mutation type and their frequencies (70). This may refine the etiological inferences; for example, clock-like signatures of SBS1/5 mutations rise quicker in the late-replicating heterochromatin regions (71), while SBS2/13 mutations accumulate in early-replicating euchromatin regions (72), further confirming the differences in the genomic distribution and DNA repair pathways. The spatial

footprints of each signature with known etiology may be used to trace the biological causes of other signatures with unknown etiologies. The addition of clinical details can add more confidence to these mutational signatures and their inferred etiologies. Overall, an atlas of DNA damage and repair, including genomic distribution and clinical details, may be helpful to understand the cancer genome diversity (73).

5.4 References

1. Swenberg JA, Lu K, Moeller BC, Gao L, Upton PB, Nakamura J, et al. Endogenous versus Exogenous DNA Adducts: Their Role in Carcinogenesis, Epidemiology, and Risk Assessment. Toxicological Sciences. 2011;120(suppl_1):S130–45.

2. Bachand et al. Epidemiological studies of formaldehyde exposure and risk of leukemia and nasopharyngeal cancer: a meta-analysis. Critical Reviews in Toxicology. 2010;40(2):85–100.

3. Balbo S, Brooks PJ. Implications of Acetaldehyde-Derived DNA Adducts for Understanding Alcohol-Related Carcinogenesis. In: Vasiliou V, Zakhari S, Seitz HK, Hoek JB, editors. Springer International Publishing; 2015. p. 71–88.

4. Chen et al. Quantitation of an acetaldehyde adduct in human leukocyte DNA and the effect of smoking cessation. Chemical research in toxicology. 2007 Jan;20(1):108–13.

5. IARC. Chemical agents and related occupations. IARC monographs on the evaluation of carcinogenic risks to humans. 2012b;100(Pt F):9–562.

6. IARC. IARC monographs on the identification of carcinogenic hazards to humans. 2022;1–132.

7. Thapa MJ, Chan K. The mutagenic properties of formaldehyde and acetaldehyde: Reflections on half a century of progress. Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis. 2025 Jan 1;830:111886.

8. Dingler FA, Wang M, Mu A, Millington CL, Oberbeck N, Watcham S, et al. Two Aldehyde Clearance Systems Are Essential to Prevent Lethal Formaldehyde Accumulation in Mice and Humans. Molecular cell. 2020 Dec 17;80(6):996-1012.e9.

9. Kucab JE, Zou X, Morganella S, Joel M, Nanda AS, Nagy E, et al. A Compendium of Mutational Signatures of Environmental Agents. Cell. 2019;177(4):821-836.e16.

10. Boiteux S JRS. DNA Repair Mechanisms and the Bypass of DNA Damage in Saccharomyces cerevisiae. 2013;

11. 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–e1003149.

12. Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N, Boutselakis H, et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res [Internet]. 2015 Jan 28 [cited 2025 Jun 24];43(Database issue):D805–11. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4383913/

13. COSMIC. Catalogue Of Somatic Mutations In cancer. 2023;

14. Thapa MJ, Fabros RM, Alasmar S, Chan K. Analyses of mutational patterns induced by formaldehyde and acetaldehyde reveal similarity to a common mutational signature. G3 Genes|Genomes|Genetics. 2022;jkac238.

15. COSMIC. Catalogue Of Somatic Mutations In cancer. 2022.

16. COSMIC. Catalogue Of Somatic Mutations In cancer. 2021.

17. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences. 2005 Oct 25;102(43):15545–50.

18. Chang LY, Lee MZ, Wu Y, Lee WK, Ma CL, Chang JM, et al. Gene set correlation enrichment analysis for interpreting and annotating gene expression profiles. Nucleic Acids Res. 2024;52(3):e17–e17.

19. Zhou T, Yao J, Liu Z. Gene Ontology, Enrichment Analysis, and Pathway Analysis. In: Liu Z (John), editor. Bioinformatics in Aquaculture [Internet]. 1st ed. Wiley; 2017 [cited 2025 Jun 24]. p. 150–68. Available from: https://onlinelibrary.wiley.com/doi/10.1002/9781118782392.ch10

20. Sayols S. rrvgo: a Bioconductor package for interpreting lists of Gene Ontology terms. microPublication biology. 2023;2023.

21. Alexandrov et al. The repertoire of mutational signatures in human cancer. Nature. 2020 Feb;578(7793):94–101.

22. Nik-Zainal S, Alexandrov LB, Wedge DC, Van Loo P, Greenman CD, Raine K, et al. Mutational processes molding the genomes of 21 breast cancers. Cell. 2012;149(5):979–93.

23. Volkova NV, Meier B, González-Huici V, Bertolini S, Gonzalez S, Vöhringer H, et al. Mutational signatures are jointly shaped by DNA damage and repair. Nat Commun. 2020 May 1;11(1):2169.

24. Alexandrov et al. Deciphering signatures of mutational processes operative in human cancer. Cell Rep. 2013;3(1):246–59.

25. Meier B, Volkova NV, Hong Y, Schofield P, Campbell PJ, Gerstung M, et al. Mutational signatures of DNA mismatch repair deficiency in C. elegans and human cancers. Genome research. 2018 May;28(5):666–75.

26. Drost J, van Boxtel R, Blokzijl F, Mizutani T, Sasaki N, Sasselli V, et al. Use of CRISPR-modified human stem cell organoids to study the origin of mutational signatures in cancer. Science. 2017 Oct 13;358(6360):234–8.

27. Alexandrov, Jones PH, Wedge DC, Sale JE, Campbell PJ, Nik-Zainal S, et al. Clock-like mutational processes in human somatic cells. Nature genetics. 2015 Dec;47(12):1402–7.

28. Nik-Zainal S, Davies H, Staaf J, Ramakrishna M, Glodzik D, Zou X, et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature [Internet]. 2016 Jun [cited 2025 Jun 24];534(7605):47–54. Available from: https://www.nature.com/articles/nature17676

29. Hayward NK, Wilmott JS, Waddell N, Johansson PA, Field MA, Nones K, et al. Whole-genome landscapes of major melanoma subtypes. Nature. 2017 May 1;545(7653):175–80.

30. Saini N, Roberts SA, Klimczak LJ, Chan K, Grimm SA, Dai S, et al. The Impact of Environmental and Endogenous Damage on Somatic Mutation Load in Human Skin Fibroblasts. PLoS Genet. 2016;12(10):e1006385.

31. Lee J m., Ledermann JA, Kohn EC. PARP Inhibitors for BRCA1/2 mutation-associated and BRCA-like malignancies. Ann Oncol [Internet]. 2014 Jan [cited 2025 Jun 29];25(1):32–40. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3868320/

32. Yu G. Gene ontology semantic similarity analysis using GOSemSim. Stem Cell Transcriptional Networks: Methods and Protocols. 2020;207–15.

33. Yaacov A, Rosenberg S, Simon I. Mutational signatures association with replication timing in normal cells reveals similarities and differences with matched cancer tissues. Sci Rep [Internet]. 2023 May 15 [cited 2025 Jul 5];13(1):7833. Available from: https://www.nature.com/articles/s41598-023-34631-9

34. Planas-Paz L, Pliego-Mendieta A, Hagedorn C, Aguilera-Garcia D, Haberecker M, Arnold F, et al. Unravelling homologous recombination repair deficiency and therapeutic opportunities in soft tissue and bone sarcoma. EMBO Mol Med [Internet]. 2023 Feb 13 [cited 2025 Jul 10];15(4):e16863. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10086583/

35. Petljak M, Dananberg A, Chu K, Bergstrom EN, Striepen J, von Morgen P, et al. Mechanisms of APOBEC3 mutagenesis in human cancer cells. Nature [Internet]. 2022 [cited 2025 Jul 11];607(7920):799–807. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9329121/

36. Gelova SP, Doherty KN, Alasmar S, Chan K. Intrinsic base substitution patterns in diverse species reveal links to cancer and metabolism. Genetics [Internet]. 2022 Sep 23 [cited 2025 Jul 11];222(3):iyac144. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9630983/

37. Zhuo M, Gorgun MF, Englander EW. Translesion Synthesis DNA Polymerase Kappa Is Indispensable for DNA Repair Synthesis in Cisplatin Exposed Dorsal Root Ganglion Neurons. Mol Neurobiol [Internet]. 2018 Mar 1 [cited 2025 Jul 11];55(3):2506–15. Available from: https://doi.org/10.1007/s12035-017-0507-5

38. Long LJ, Lee PH, Small EM, Hillyer C, Guo Y, Osley MA. Regulation of UV damage repair in quiescent yeast cells. DNA Repair (Amst) [Internet]. 2020 Jun [cited 2025 Jul 11];90:102861. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7891302/

39. Spisak N, Manuel M de, Milligan W, Sella G, Przeworski M. The clock-like accumulation of germline and somatic mutations can arise from the interplay of DNA damage and repair. PLOS Biology [Internet]. 2024 Jun 17 [cited 2025 Jul 5];22(6):e3002678. Available from: https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.3002678

40. Manaka Y, Kusumoto-Matsuo R, Matsuno Y, Asai H, Yoshioka K ichi. Single base substitution signatures 17a, 17b, and 40 are induced by γ-ray irradiation in association with increased

reactive oxidative species. Heliyon [Internet]. 2024 Mar 15 [cited 2025 Jul 7];10(6):e28044. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10965518/

41. Focaccetti C, Bruno A, Magnani E, Bartolini D, Principi E, Dallaglio K, et al. Effects of 5-Fluorouracil on Morphology, Cell Cycle, Proliferation, Apoptosis, Autophagy and ROS Production in Endothelial Cells and Cardiomyocytes. PLOS ONE [Internet]. 2015 Feb 11 [cited 2025 Jul 7];10(2):e0115686. Available from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0115686

42. Negrei C, Hudita A, Ginghina O, Galateanu B, Voicu SN, Stan M, et al. Colon Cancer Cells Gene Expression Signature As Response to 5- Fluorouracil, Oxaliplatin, and Folinic Acid Treatment. Front Pharmacol [Internet]. 2016 Jun 23 [cited 2025 Jul 7];7. Available from: https://www.frontiersin.org/journals/pharmacology/articles/10.3389/fphar.2016.00172/full

43. Christensen S, Van der Roest B, Besselink N, Janssen R, Boymans S, Martens JWM, et al. 5-Fluorouracil treatment induces characteristic T>G mutations in human cancer. Nat Commun [Internet]. 2019 Oct 8 [cited 2025 Jul 7];10(1):4571. Available from: https://www.nature.com/articles/s41467-019-12594-8

44. Program NT. Multiomics Evaluation of B6C3F1/N Mouse Hepatocellular Carcinomas Arising Spontaneously or Following Chronic Exposure to Tris(chloropropyl) Phosphate. In: NTP Technical Report on the Toxicology and Carcinogenesis Studies of an Isomeric Mixture of Tris(chloropropyl) Phosphate Administered in Feed to Sprague Dawley (Hsd:Sprague Dawley® SD®) Rats and B6C3F1/N Mice: Technical Report 602 [Internet] [Internet]. National Toxicology Program; 2023 [cited 2025 Jul 7]. Available from: https://www.ncbi.nlm.nih.gov/books/NBK592942/

45. Clark DJ, Dhanasekaran SM, Petralia F, Pan J, Song X, Hu Y, et al. Integrated Proteogenomic Characterization of Clear Cell Renal Cell Carcinoma. Cell [Internet]. 2019 Oct 31 [cited 2025 Jul 7];179(4):964-983.e31. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7331093/

46. Zhang Y, Chen F, Chandrashekar DS, Varambally S, Creighton CJ. Proteogenomic characterization of 2002 human cancers reveals pan-cancer molecular subtypes and associated pathways. Nat Commun [Internet]. 2022 May 13 [cited 2025 Jul 7];13(1):2669. Available from: https://www.nature.com/articles/s41467-022-30342-3

47. Wamsley NT, Wilkerson EM, Guan L, LaPak KM, Schrank TP, Holmes BJ, et al. Targeted Proteomic Quantitation of NRF2 Signaling and Predictive Biomarkers in HNSCC. Mol Cell Proteomics [Internet]. 2023 Sep 15 [cited 2025 Jul 7];22(11):100647. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10587640/

48. Cornejo P, Vargas R, Videla LA. Nrf2-regulated phase-II detoxification enzymes and phase-III transporters are induced by thyroid hormone in rat liver. Biofactors. 2013;39(5):514–21.

49. Singh VK, Rastogi A, Hu X, Wang Y, De S. Mutational signature SBS8 predominantly arises due to late replication errors in cancer. Commun Biol [Internet]. 2020 Aug 3 [cited 2025 Jul 5];3(1):421. Available from: https://www.nature.com/articles/s42003-020-01119-5

50. Jager M, Blokzijl F, Kuijk E, Bertl J, Vougioukalaki M, Janssen R, et al. Deficiency of nucleotide excision repair is associated with mutational signature observed in cancer. Genome Res [Internet]. 2019 Jul [cited 2025 Jul 6];29(7):1067–77. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6633256/

51. Haradhvala NJ, Polak P, Stojanov P, Covington KR, Shinbrot E, Hess JM, et al. Mutational Strand Asymmetries in Cancer Genomes Reveal Mechanisms of DNA Damage and Repair. Cell [Internet]. 2016 Jan 28 [cited 2025 Jul 6];164(3):538–49. Available from: https://www.cell.com/cell/abstract/S0092-8674(15)01714-6

52. Hanawalt PC, Spivak G. Transcription-coupled DNA repair: two decades of progress and surprises. Nat Rev Mol Cell Biol [Internet]. 2008 Dec [cited 2025 Jul 6];9(12):958–70. Available from: https://www.nature.com/articles/nrm2549

53. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin AV, et al. Signatures of mutational processes in human cancer. Nature [Internet]. 2013 Aug [cited 2025 Jul 6];500(7463):415–21. Available from: https://www.nature.com/articles/nature12477

54. Erkizan HV, Sukhadia S, Natarajan TG, Marino G, Notario V, Lichy JH, et al. Exome sequencing identifies novel somatic variants in African American esophageal squamous cell carcinoma. Sci Rep [Internet]. 2021 Jul 20 [cited 2025 Jul 6];11:14814. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8292420/

55. Thakur BK, Malaisé Y, Martin A. Unveiling the Mutational Mechanism of the Bacterial Genotoxin Colibactin in Colorectal Cancer. Molecular Cell [Internet]. 2019 Apr 18 [cited 2025 Jul 6];74(2):227–9. Available from: https://www.sciencedirect.com/science/article/pii/S1097276519302771

56. Otlu B, Díaz-Gay M, Vermes I, Bergstrom EN, Zhivagui M, Barnes M, et al. Topography of mutational signatures in human cancer. Cell Rep [Internet]. 2023 Aug 29 [cited 2025 Jul 6];42(8):112930. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10507738/

57. Dziubańska-Kusibab PJ, Berger H, Battistini F, Bouwman BAM, Iftekhar A, Katainen R, et al. Colibactin DNA-damage signature indicates mutational impact in colorectal cancer. Nat Med [Internet]. 2020 Jul [cited 2025 Jul 6];26(7):1063–9. Available from: https://www.nature.com/articles/s41591-020-0908-2

58. Strauss JD, Pursell ZF. Replication DNA polymerases, genome instability and cancer therapies. NAR Cancer [Internet]. 2023 Sep 1 [cited 2025 Jul 6];5(3):zcad033. Available from: https://doi.org/10.1093/narcan/zcad033

59. Wu Q, Wang L, Tsui SKW. Mutational signatures representative transcriptomic perturbations in hepatocellular carcinoma. Front Genet [Internet]. 2022 Aug 23 [cited 2025 Jul 6];13:970907. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9445436/

60. Riva L, Pandiri AR, Li YR, Droop A, Hewinson J, Quail MA, et al. The mutational signature profile of known and suspected human carcinogens in mice. Nat Genet [Internet]. 2020 Nov 1 [cited 2025 Jul 6];52(11):1189–97. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7610456/

61. Li Z, Liang H, Zhang S, Luo W. A practical framework RNMF for exploring the association between mutational signatures and genes using gene cumulative contribution abundance. Cancer Med [Internet]. 2022 May 16 [cited 2025 Jul 7];11(21):4053–69. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9636515/

62. Reynaud CA, Weill JC. Predicting AID off-targets: A step forward. J Exp Med [Internet]. 2018 Mar 5 [cited 2025 Jul 7];215(3):721–2. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5839771/

63. Hosoya H, Carleton M, Tanaka K, Sworder B, Syal S, Sahaf B, et al. Deciphering response dynamics and treatment resistance from circulating tumor DNA after CAR T-cells in multiple myeloma. Nat Commun [Internet]. 2025 Feb 20 [cited 2025 Jul 7];16(1):1824. Available from: https://www.nature.com/articles/s41467-025-56486-6

64. Shen JC, Kamath-Loeb AS, Kohrn BF, Loeb KR, Preston BD, Loeb LA. A high-resolution landscape of mutations in the BCL6 super-enhancer in normal human B cells. Proc Natl Acad Sci U S A [Internet]. 2019 Dec 3 [cited 2025 Jul 7];116(49):24779–85. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6900602/

65. Van Hoeck A, Tjoonk NH, van Boxtel R, Cuppen E. Portrait of a cancer: mutational signature analyses for cancer diagnostics. BMC Cancer [Internet]. 2019 May 15 [cited 2025 Jun 29];19(1):457. Available from: https://doi.org/10.1186/s12885-019-5677-2

66. Hollizeck S, Wang N, Wong SQ, Litchfield C, Guinto J, Ftouni S, et al. Unravelling mutational signatures with plasma circulating tumour DNA. Nat Commun [Internet]. 2024 Nov 14 [cited 2025 Jun 29];15(1):9876. Available from: https://www.nature.com/articles/s41467-024-54193-2

67. Funnell T, O'Flanagan CH, Williams MJ, McPherson A, McKinney S, Kabeer F, et al. Single-cell genomic variation induced by mutational processes in cancer. Nature [Internet]. 2022 Dec [cited 2025 Jun 29];612(7938):106–15. Available from: https://www.nature.com/articles/s41586-022-05249-0

68. Pan Y, Fei L, Wang S, Chen H, Jiang C, Li H, et al. Integrated analysis of single-cell, spatial and bulk RNA-sequencing identifies a cell-death signature for predicting the outcomes of head and neck cancer. Front Immunol [Internet]. 2024 Nov 7 [cited 2025 Jun 29];15:1487966. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11578999/

69. Llargués-Sistac G, Bonjoch L, Castellvi-Bel S. HAP1, a new revolutionary cell model for gene editing using CRISPR-Cas9. Front Cell Dev Biol [Internet]. 2023 Mar 3 [cited 2025 Jun 29];11:1111488. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10020200/

70. Miller JL, Grant PA. The Role of DNA Methylation and Histone Modifications in Transcriptional Regulation in Humans. Subcell Biochem [Internet]. 2013 [cited 2025 Jun 29];61:289–317. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6611551/

71. Caballero M, Boos D, Koren A. Cell-type specificity of the human mutation landscape with respect to DNA replication dynamics. Cell Genom [Internet]. 2023 May 2 [cited 2025 Jun 29];3(6):100315. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10300547/

72. Kazanov MD, Roberts SA, Polak P, Stamatoyannopoulos J, Klimczak LJ, Gordenin DA, et al. APOBEC-Induced Cancer Mutations Are Uniquely Enriched in Early-Replicating, Gene-Dense, and Active Chromatin Regions. Cell Rep. 2015 Nov 10;13(6):1103–9.

73. Liao J, Bai J, Pan T, Zou H, Gao Y, Guo J, et al. Clinical and genomic characterization of mutational signatures across human cancers. International Journal of Cancer [Internet]. 2023 [cited 2025 Jun 29];152(8):1613–29. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/ijc.34402

Chapter 6: Supplementary results

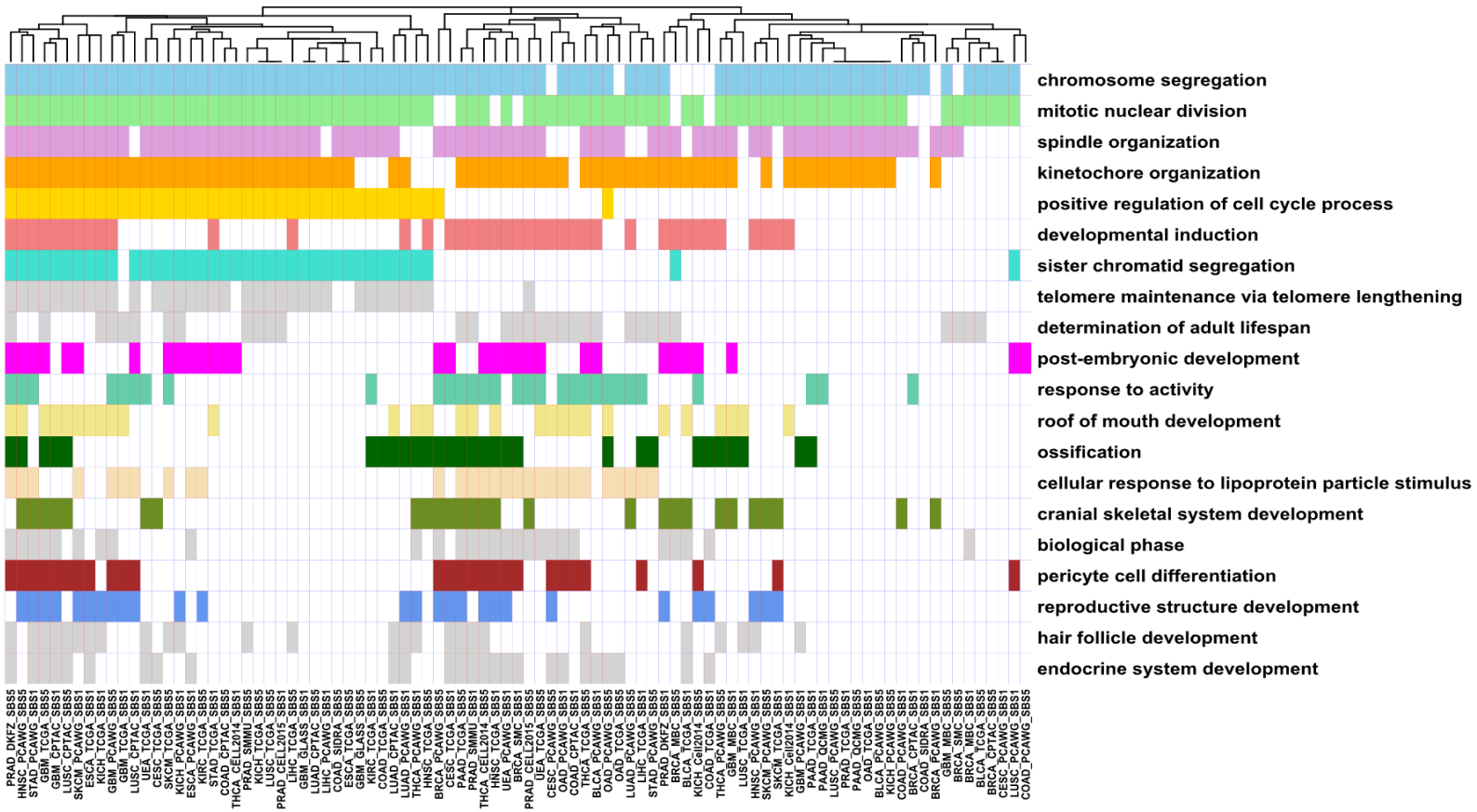
Supplementary section

In the supplementary section, the results for Chapter 4 of all the signatures and their top 20 GO terms are available in supplementary results 1.

Supplementary Results 1

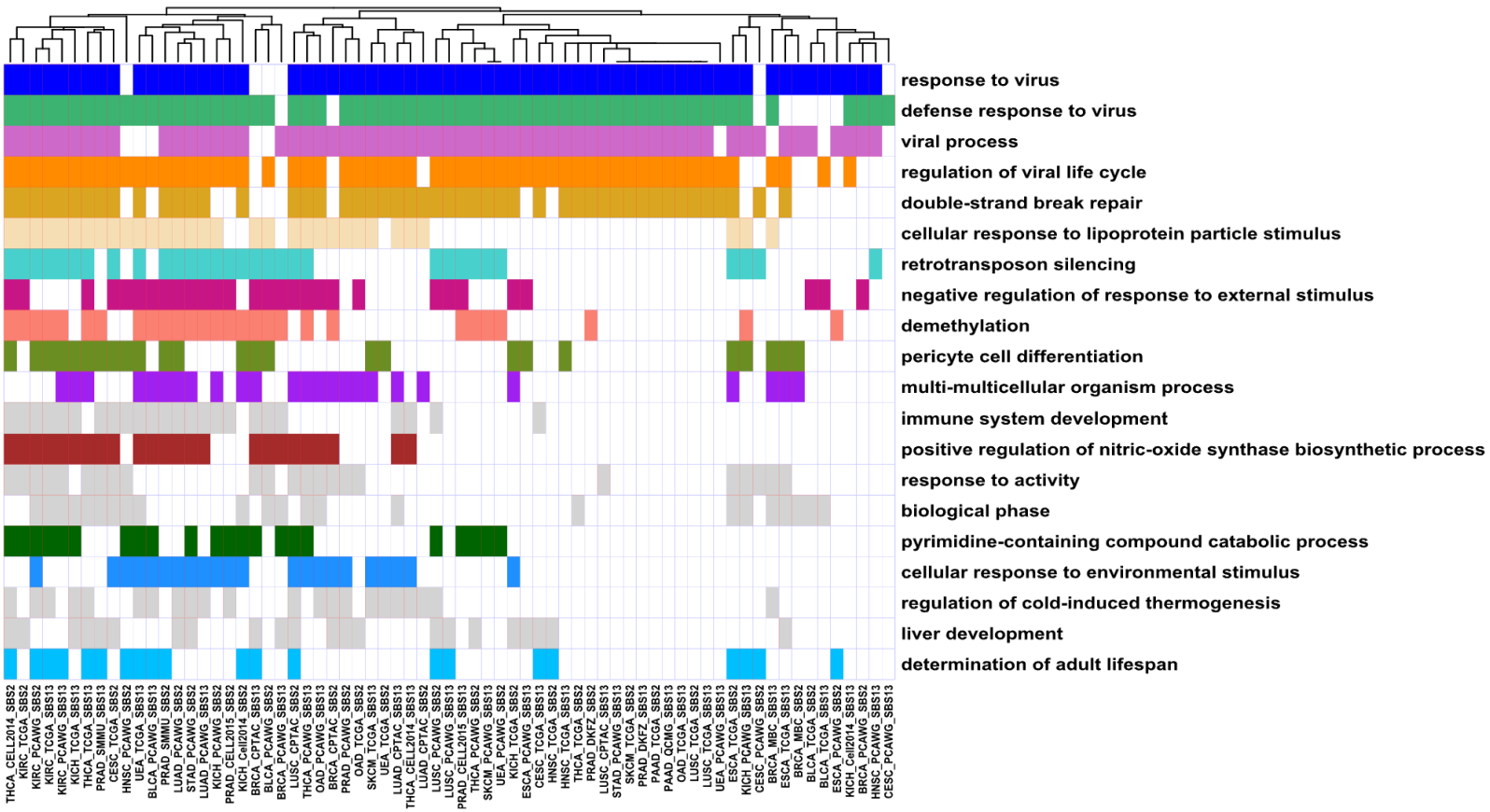
All Datasets

SBS1/5



All Datasets

SBS2/13

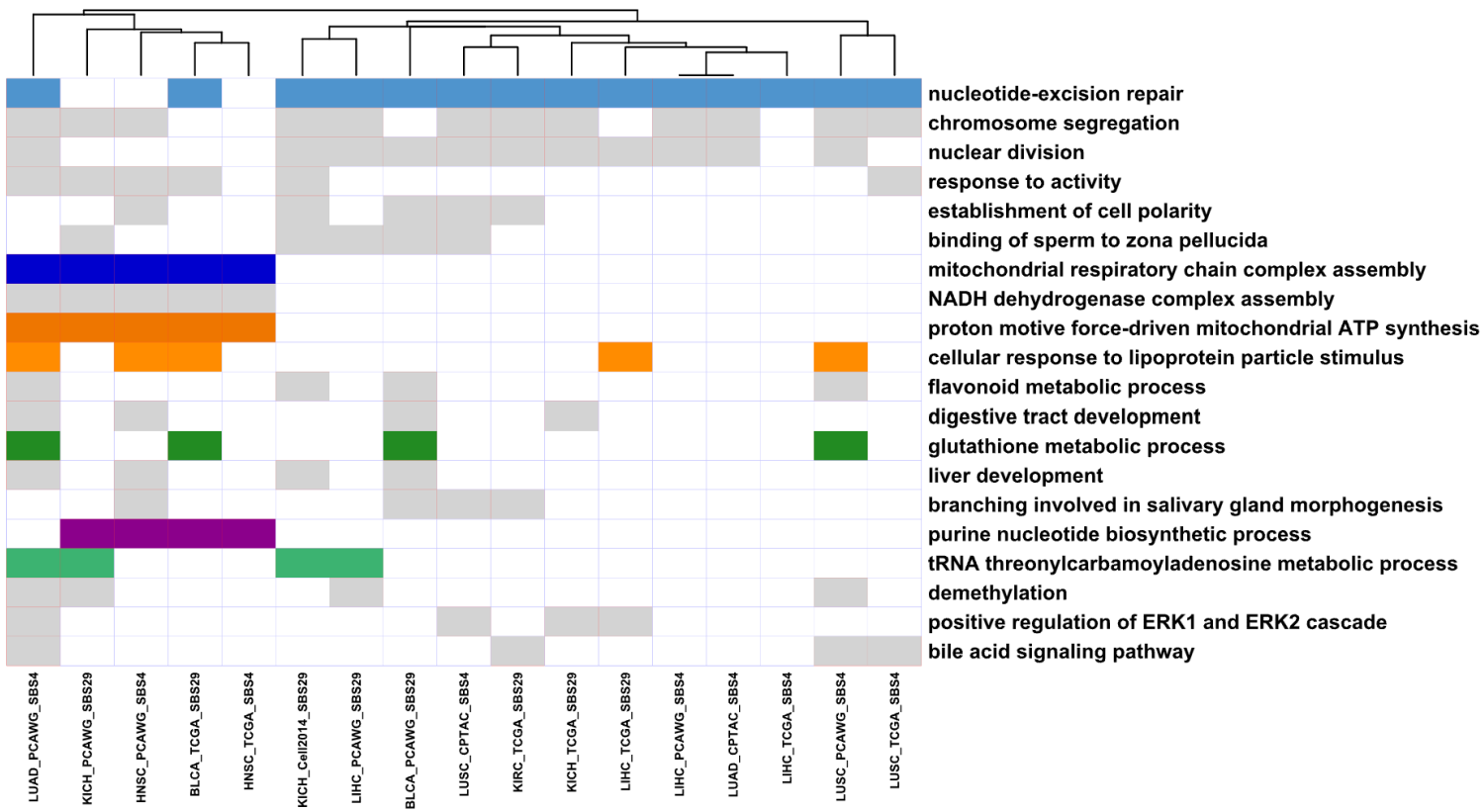


All Datasets

SBS3

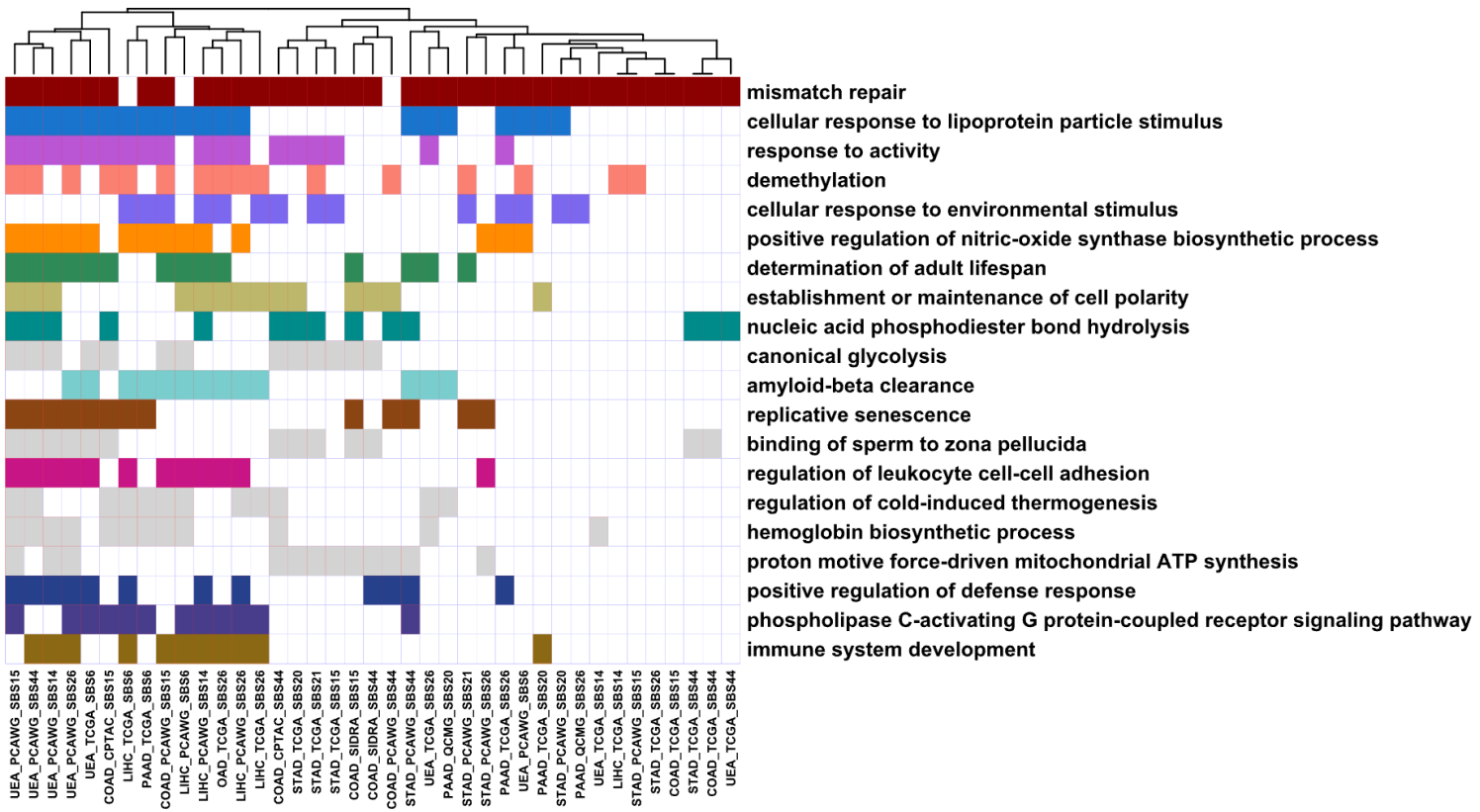
All Datasets

SBS4/29



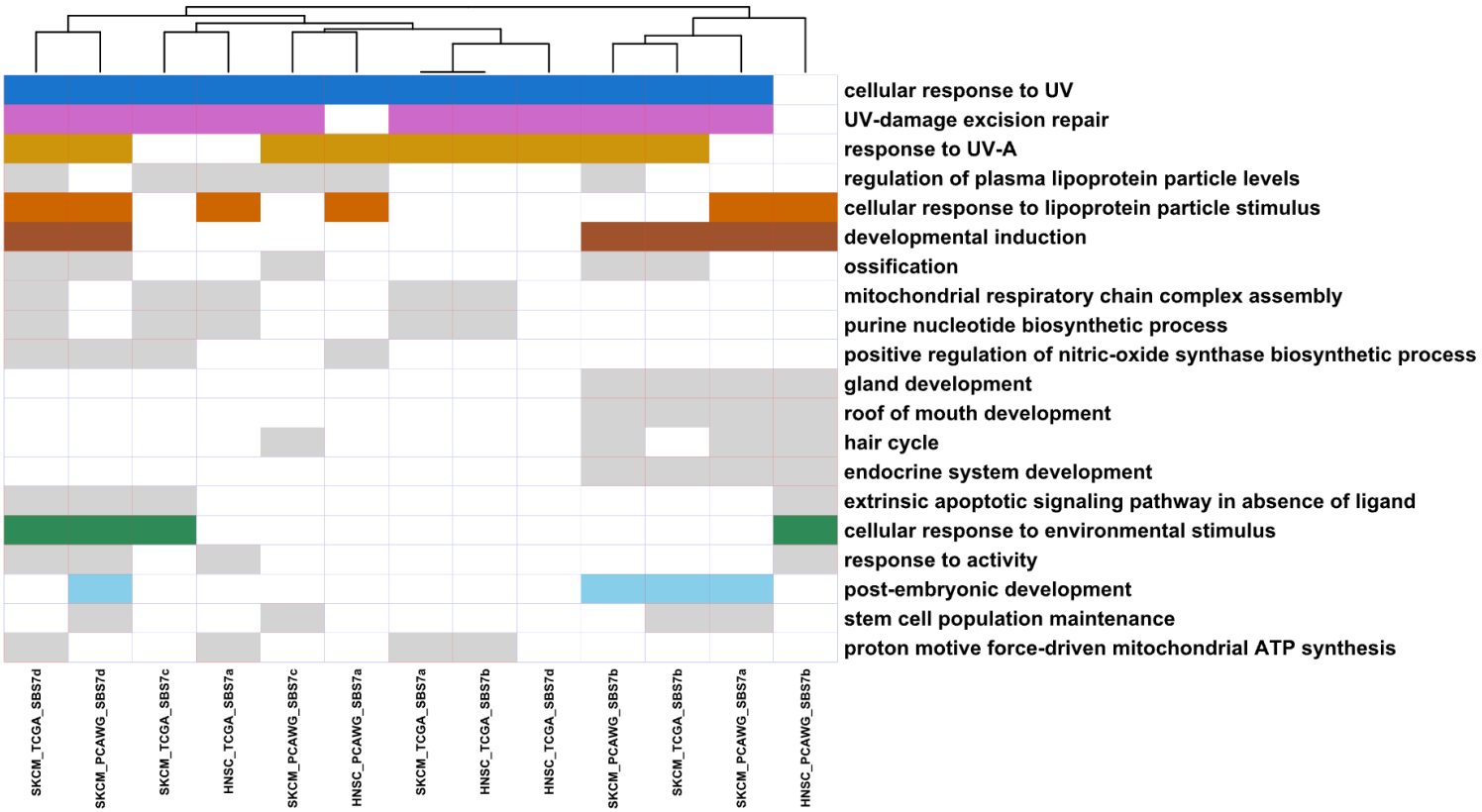
All Datasets

SBS6/14/15/20/21/26/44



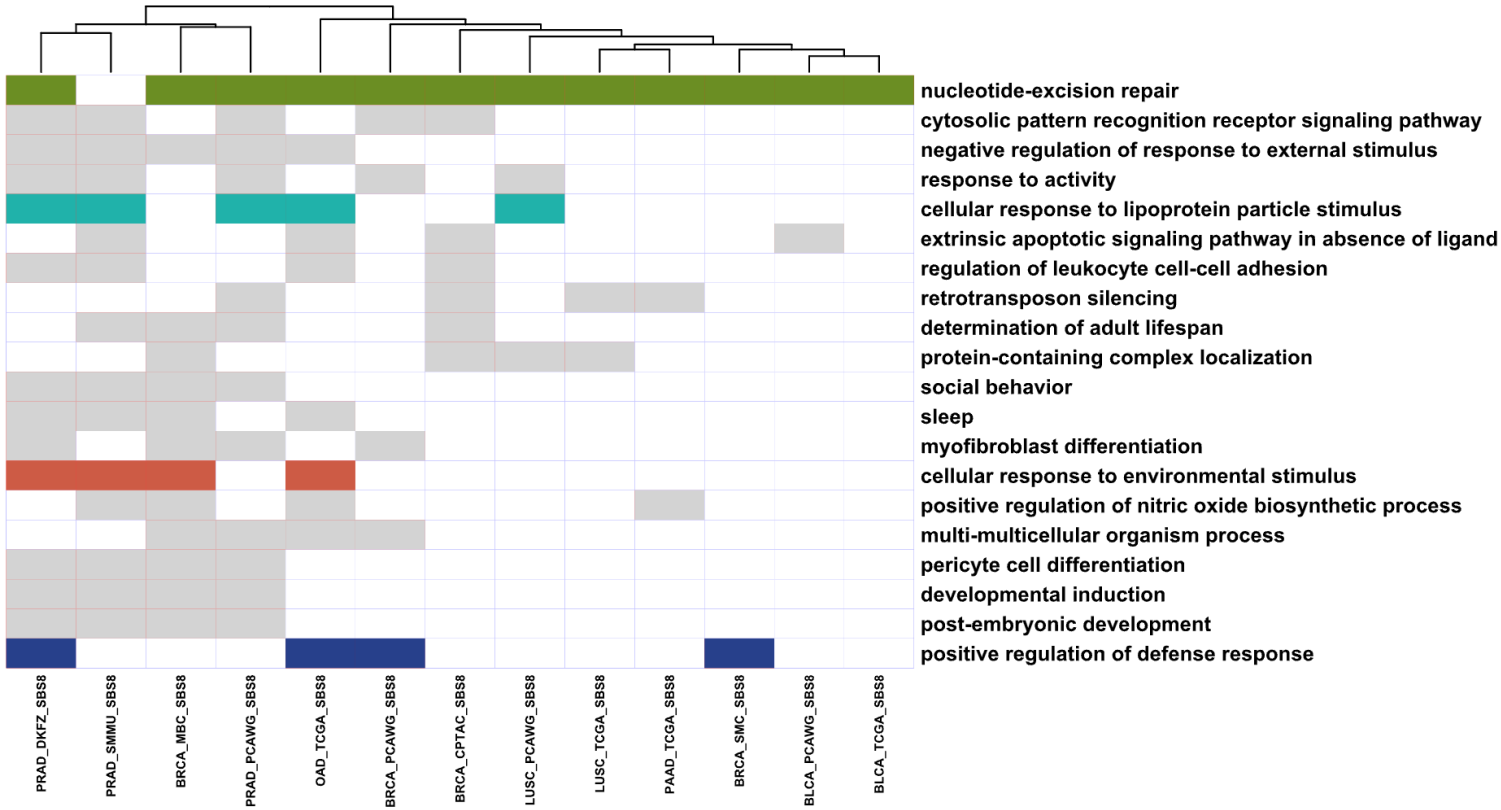
All Datasets

SBS7a/7b/7c/7d



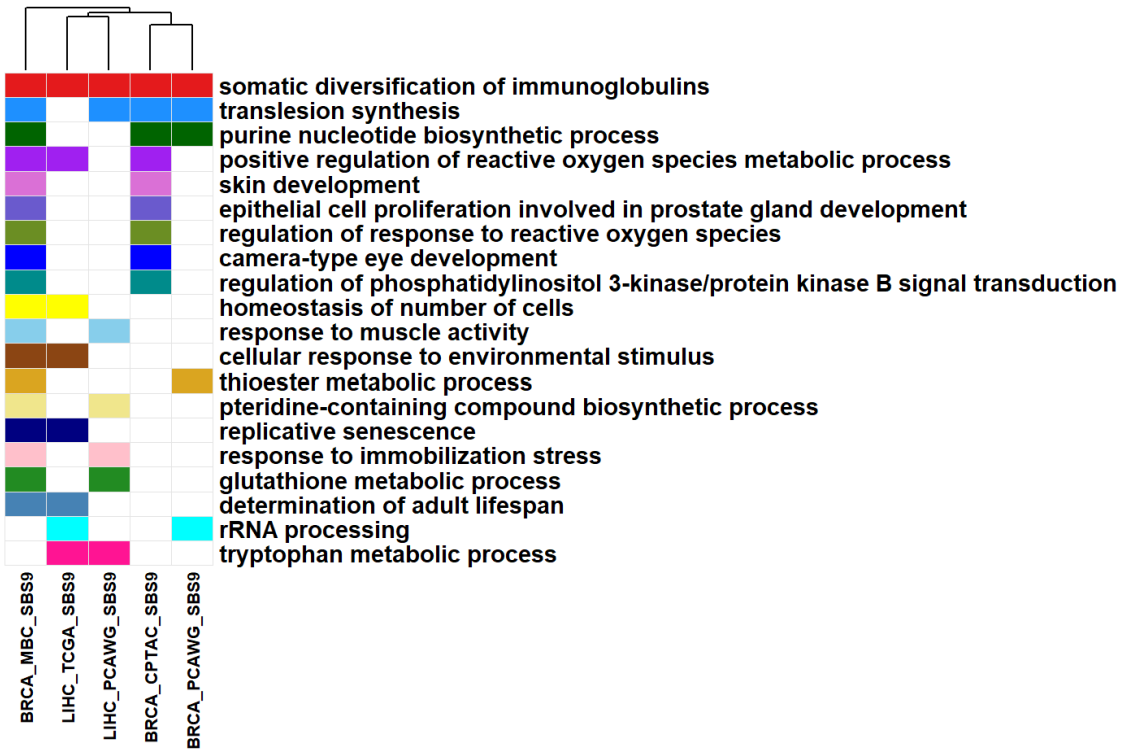
All Datasets

SBS8



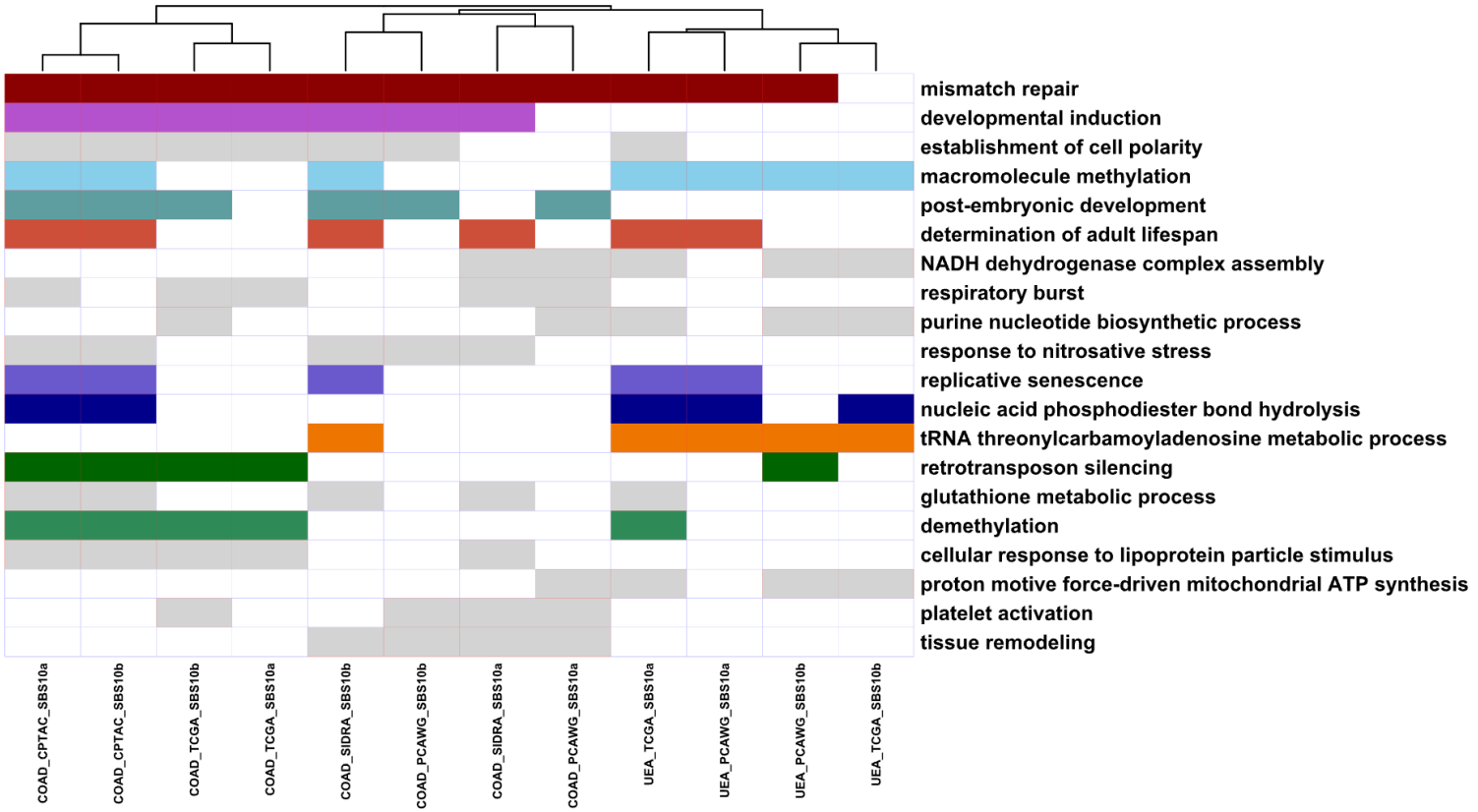
All Datasets

SBS9



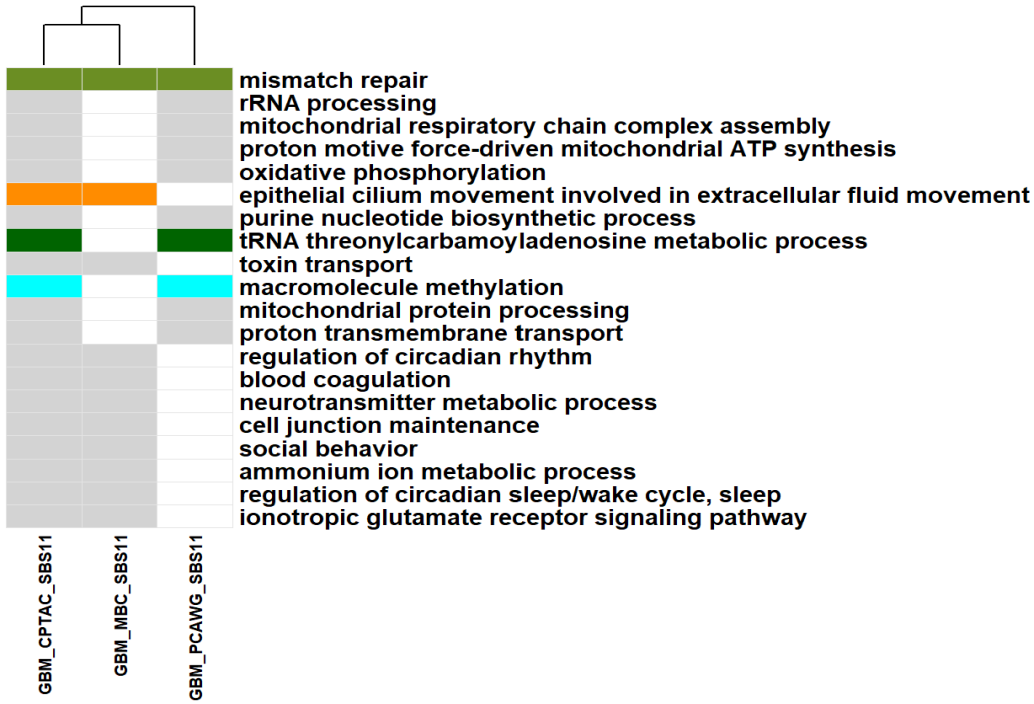
All Datasets

SBS10a/10b



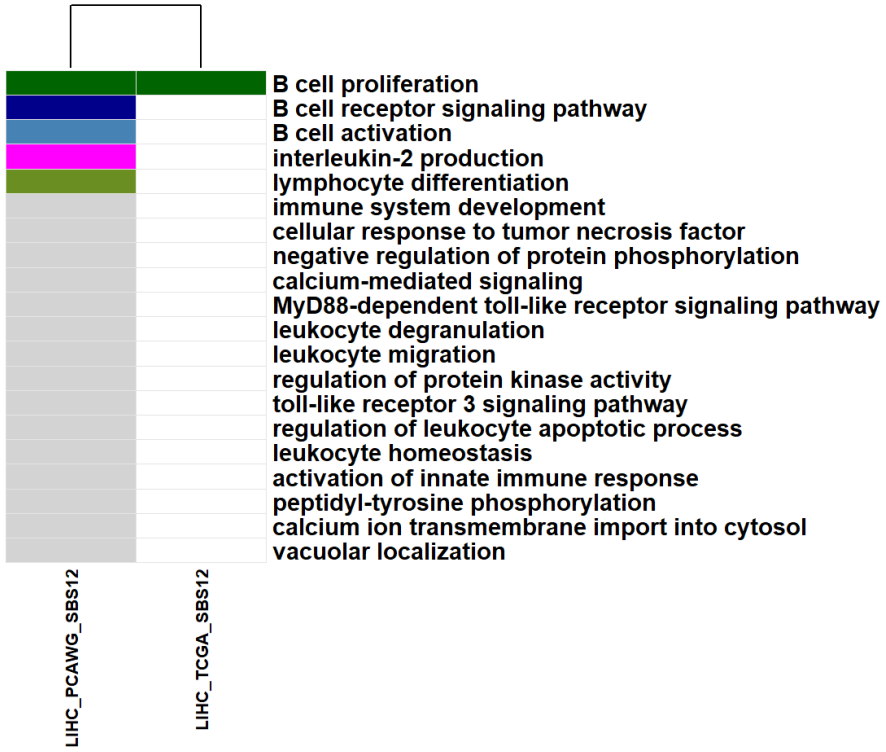
All Datasets

SBS11



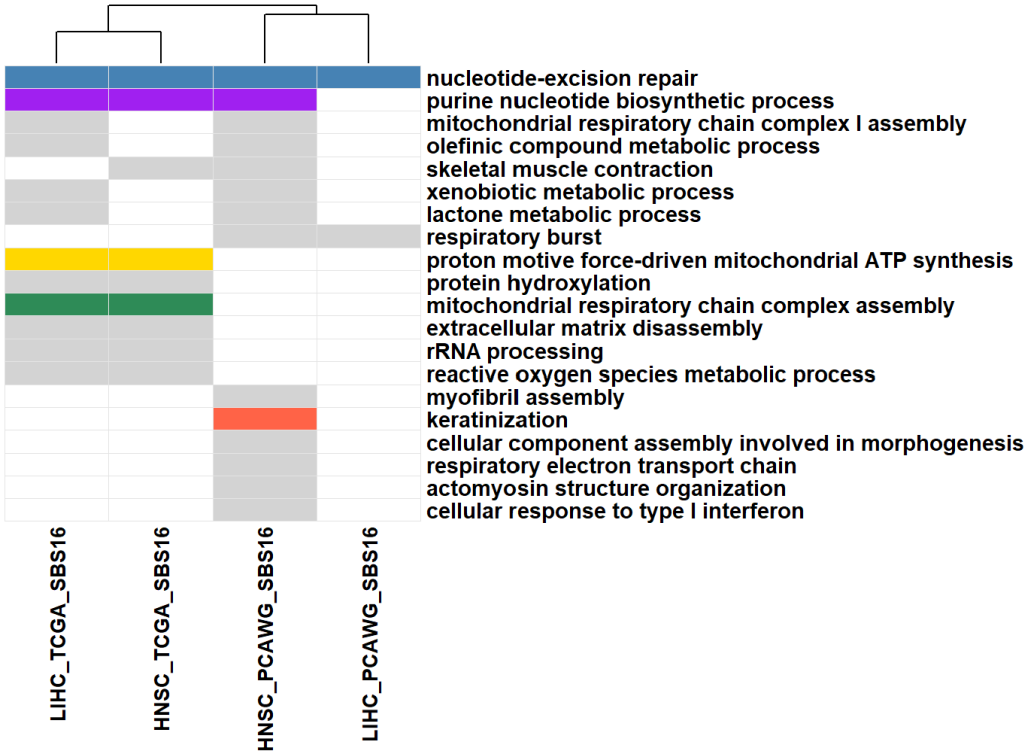
All Datasets

SBS12



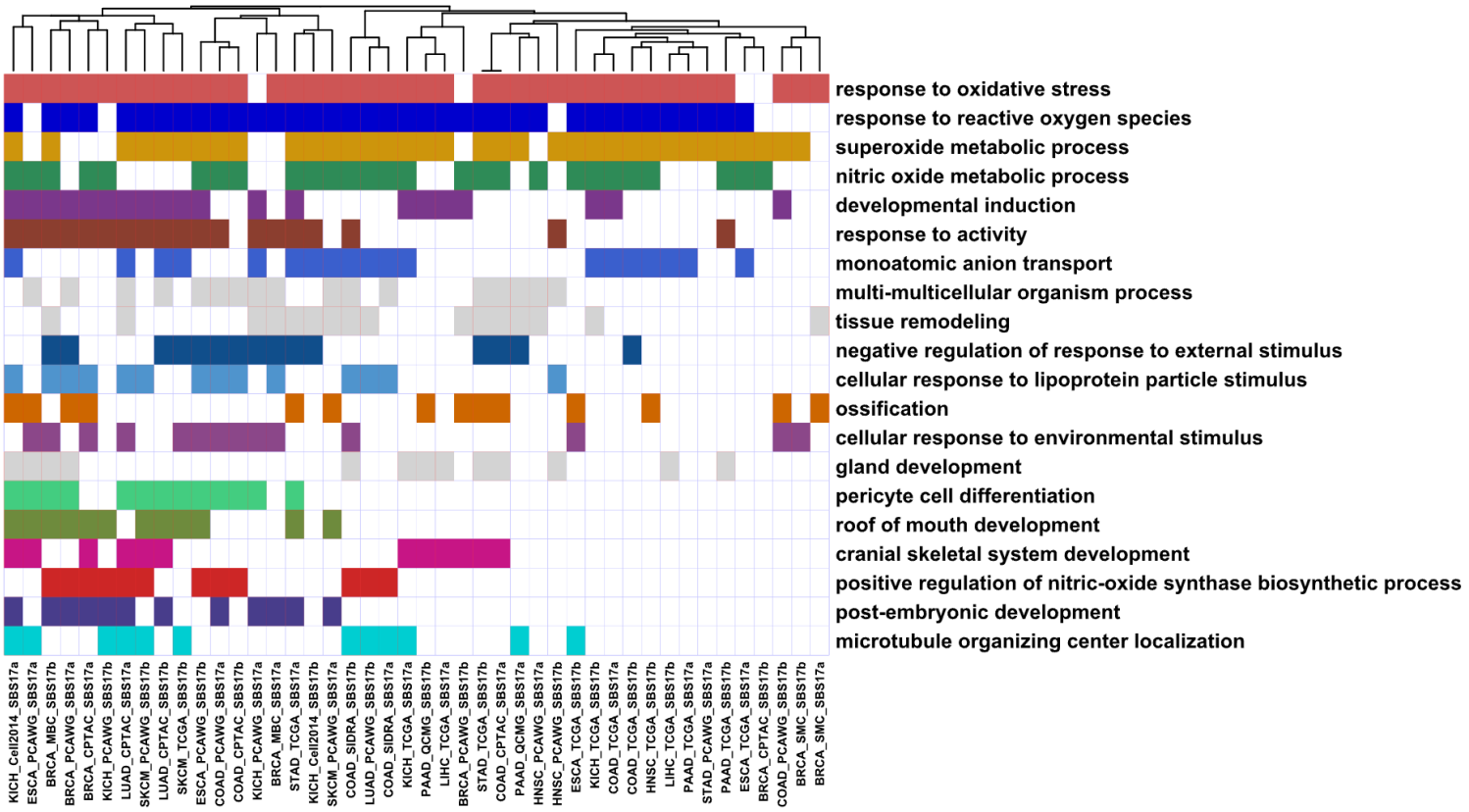
All Datasets

SBS16



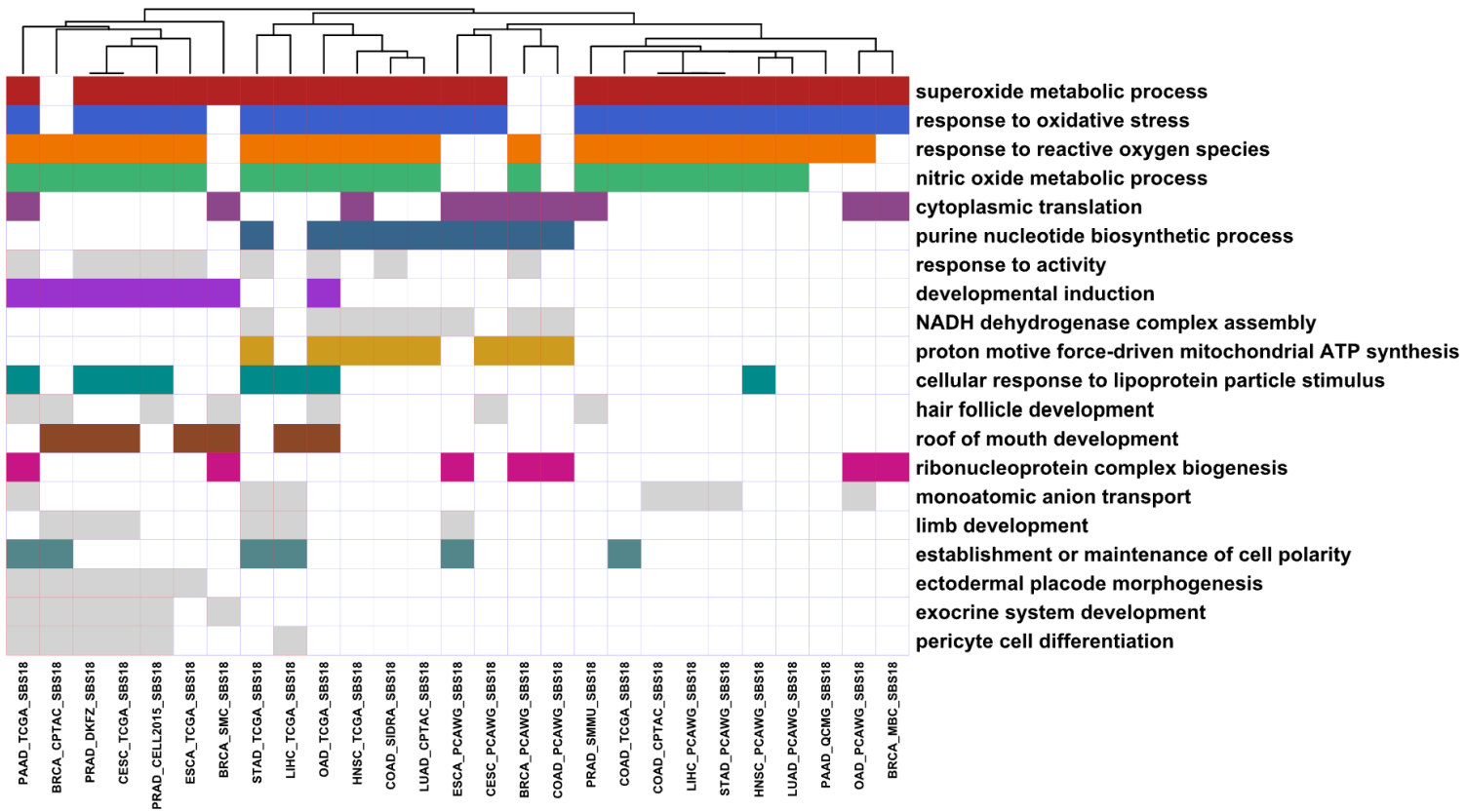
All Datasets

SBS17a/17b



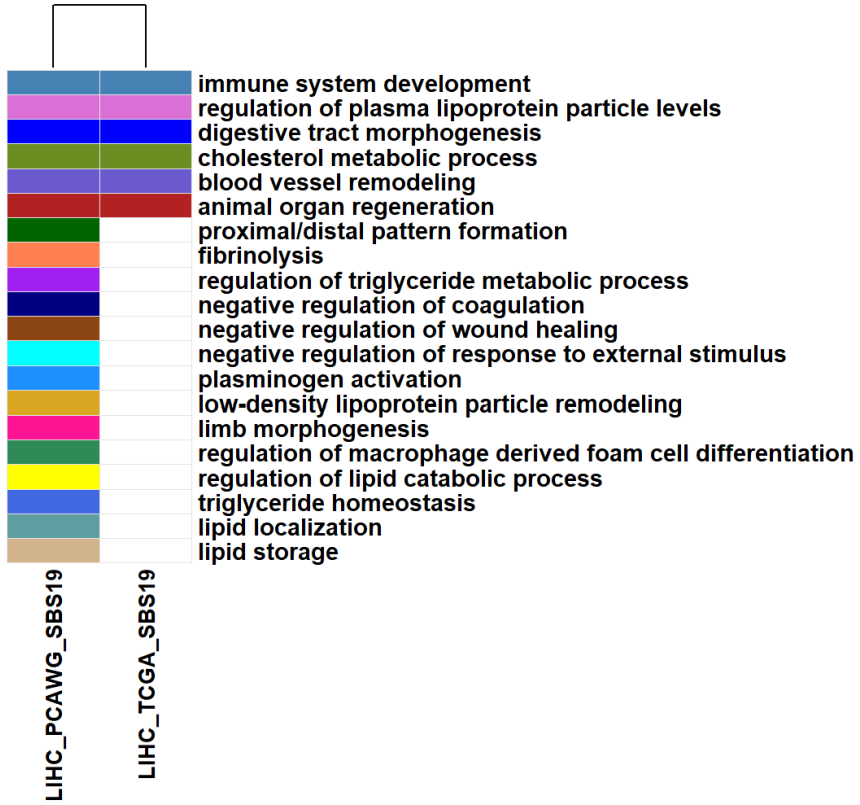
All Datasets

SBS18



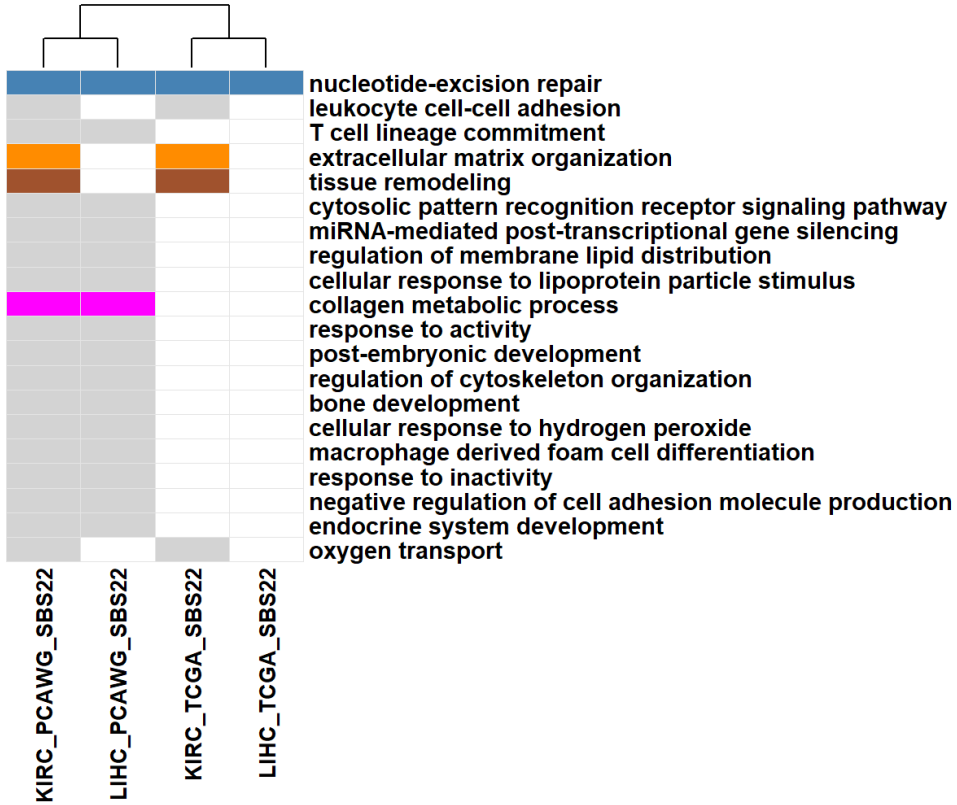
All Datasets

SBS19



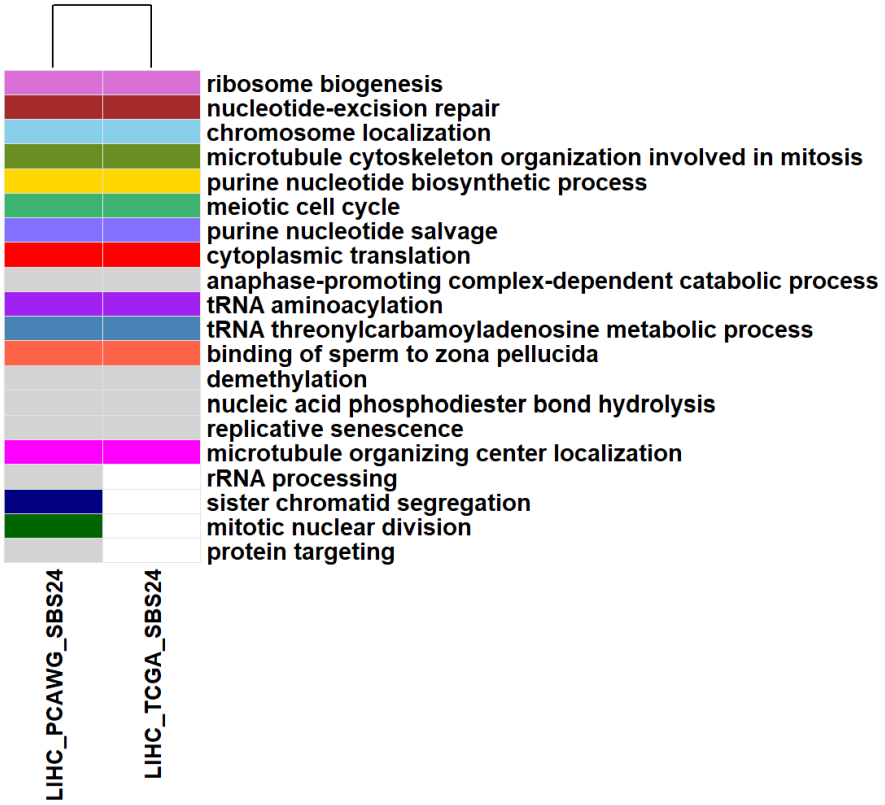
All Datasets

SBS22



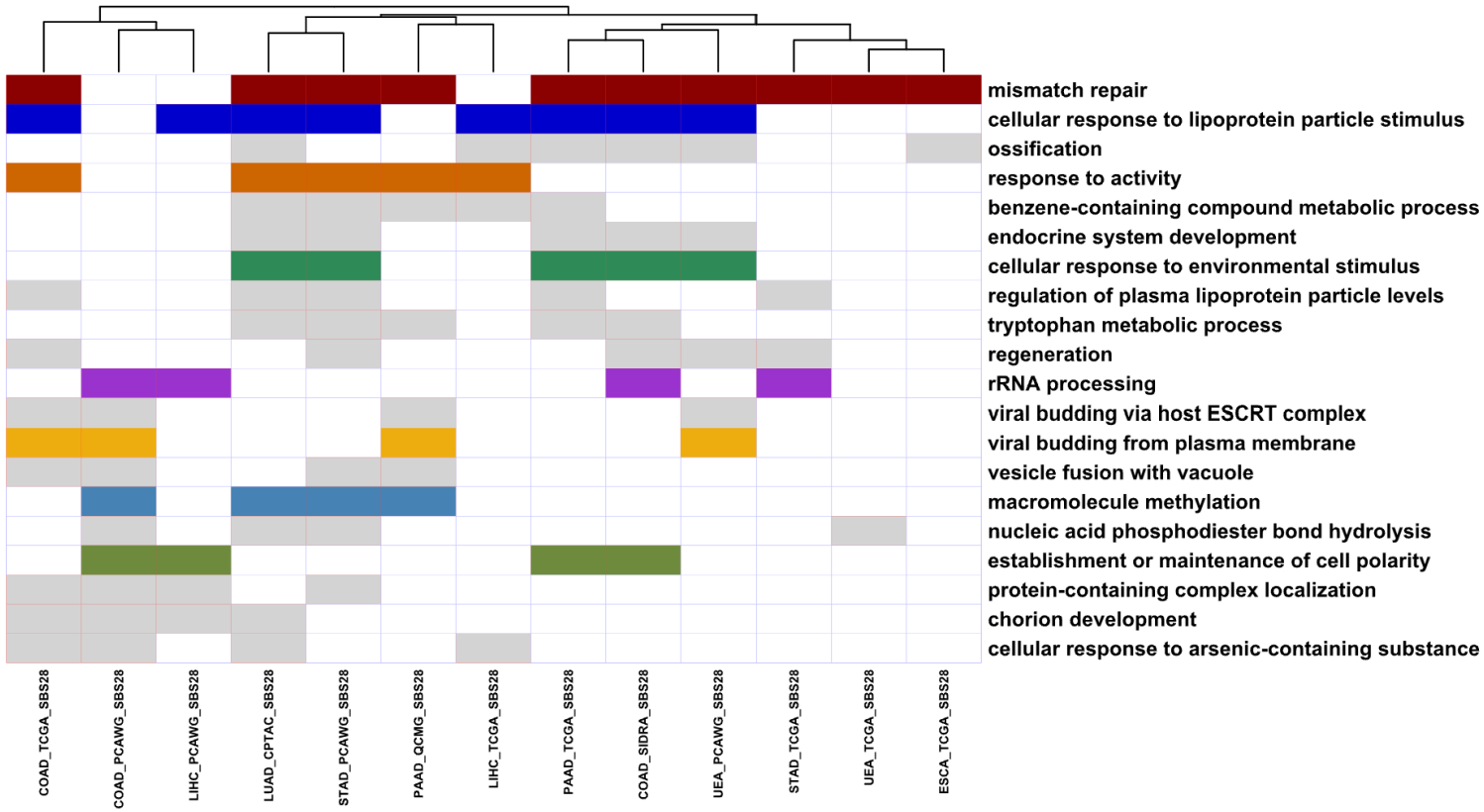
All Datasets

SBS24



All Datasets

SBS28

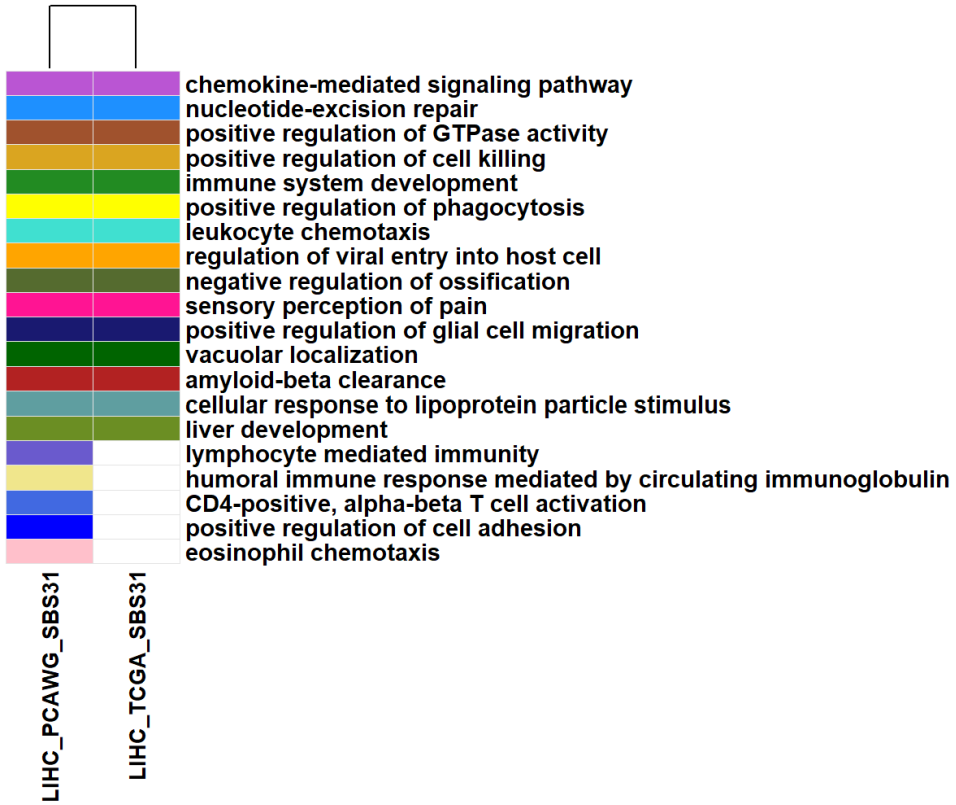


All Datasets

SBS30

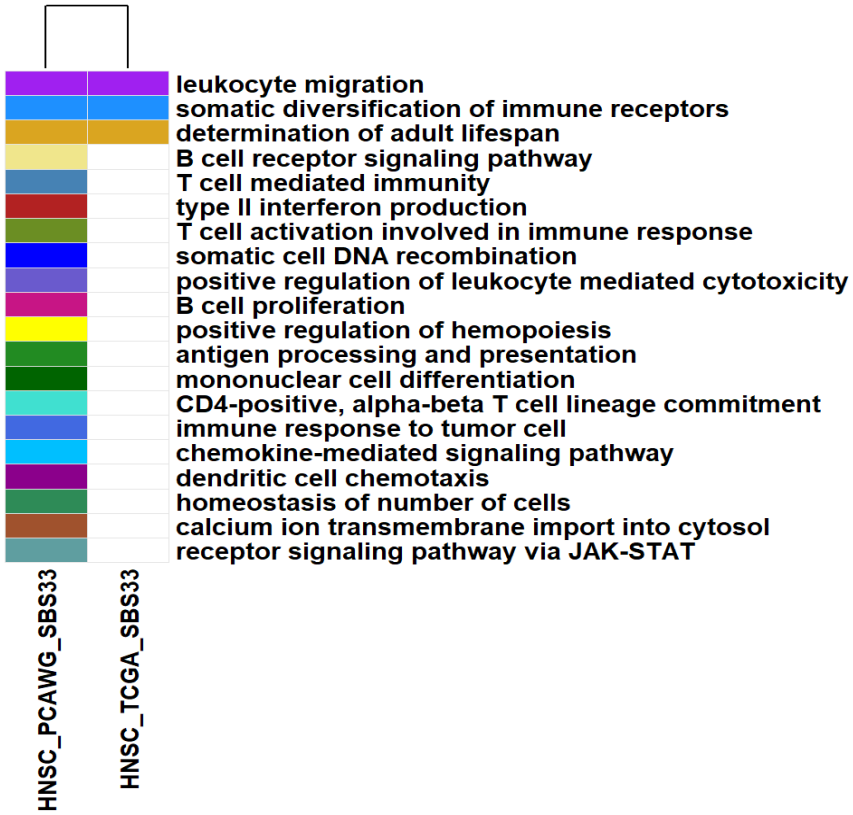
All Datasets

SBS31



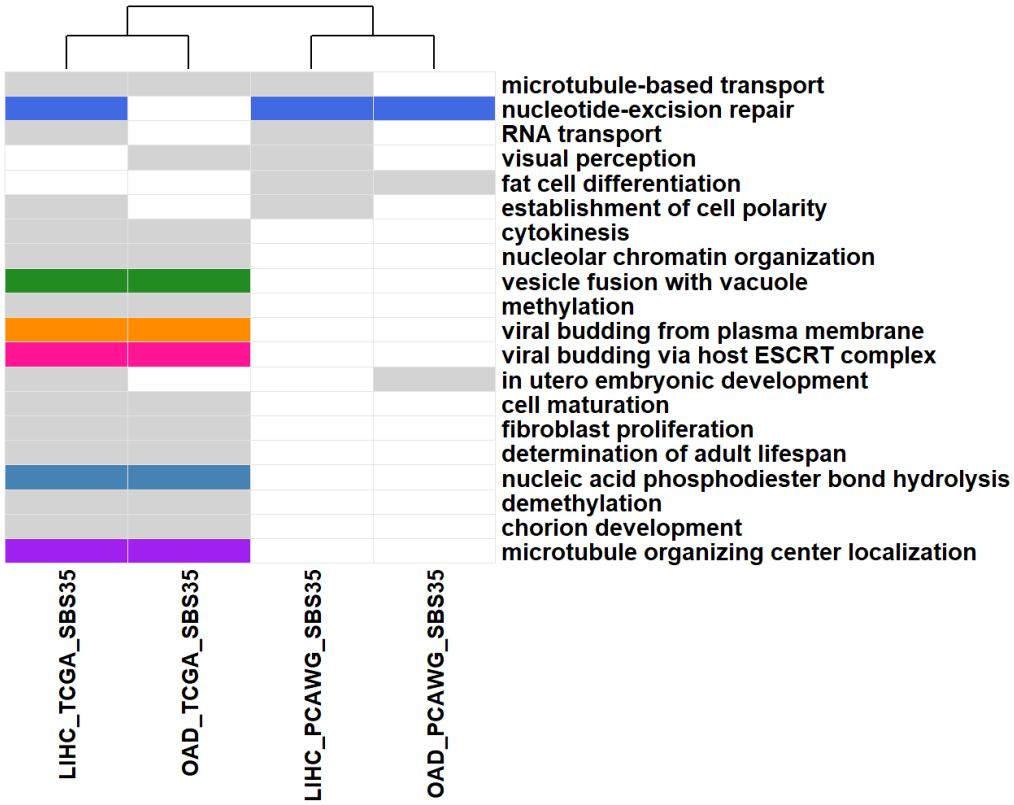
All Datasets

SBS33



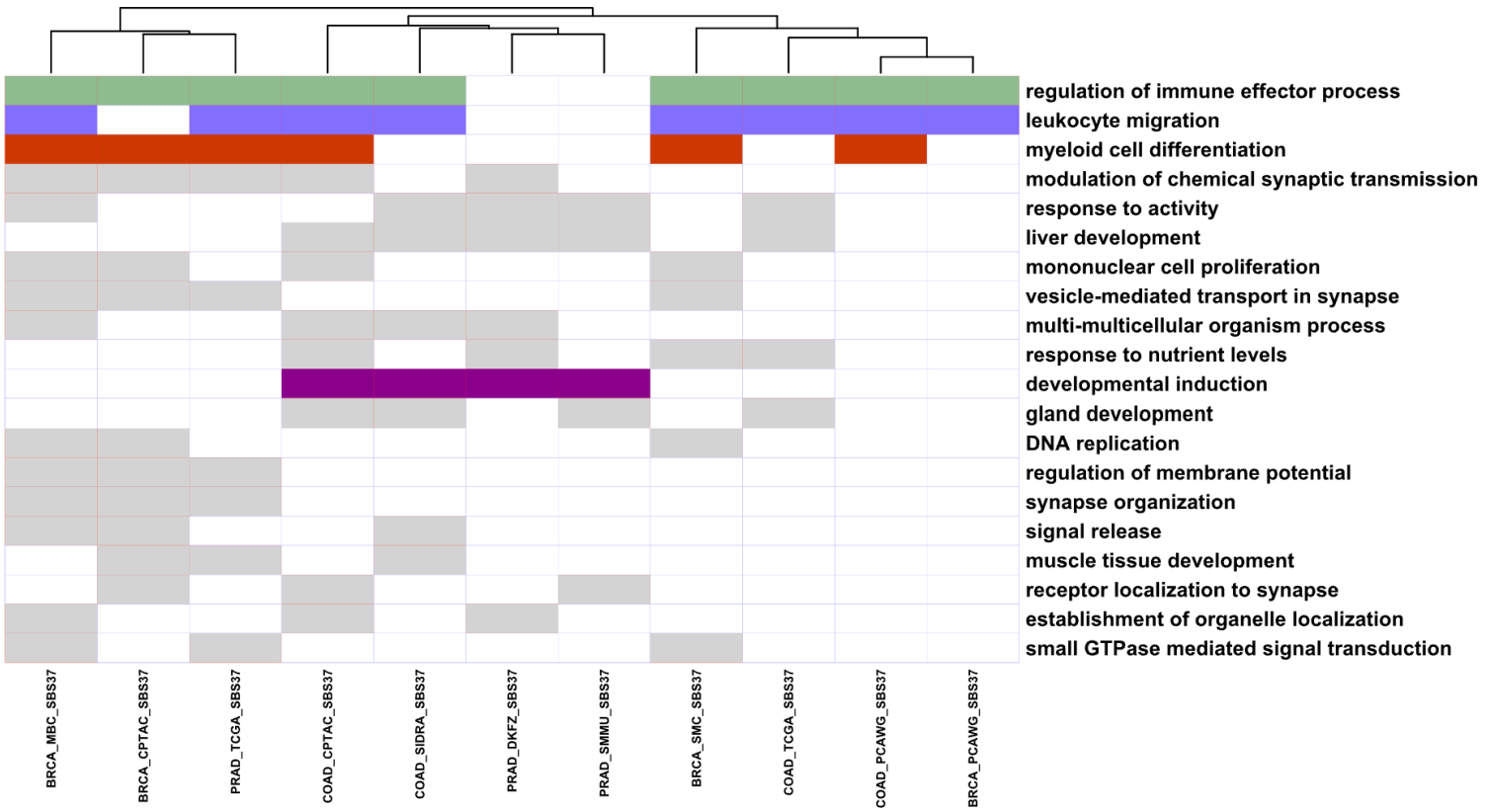
All Datasets

SBS35



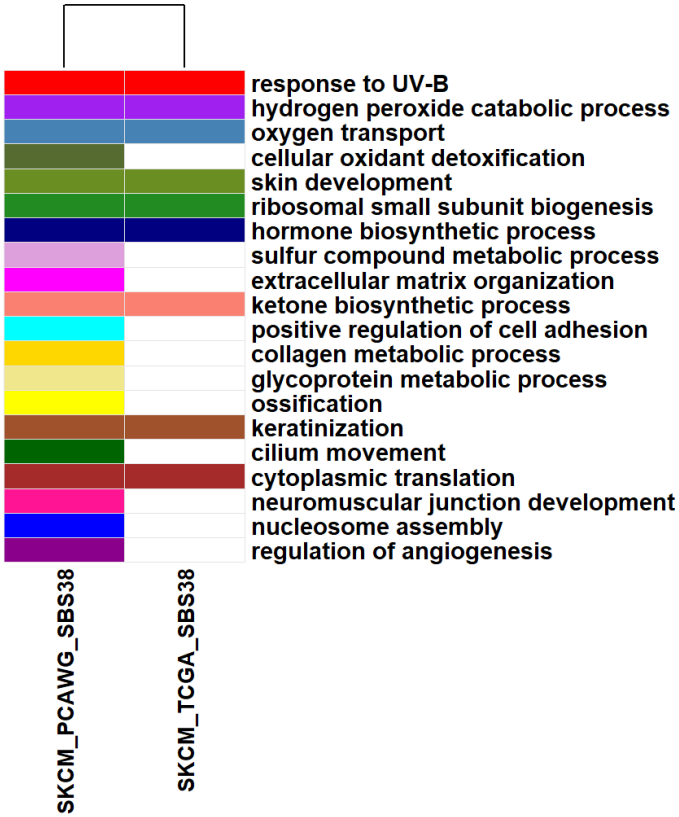
All Datasets

SBS37



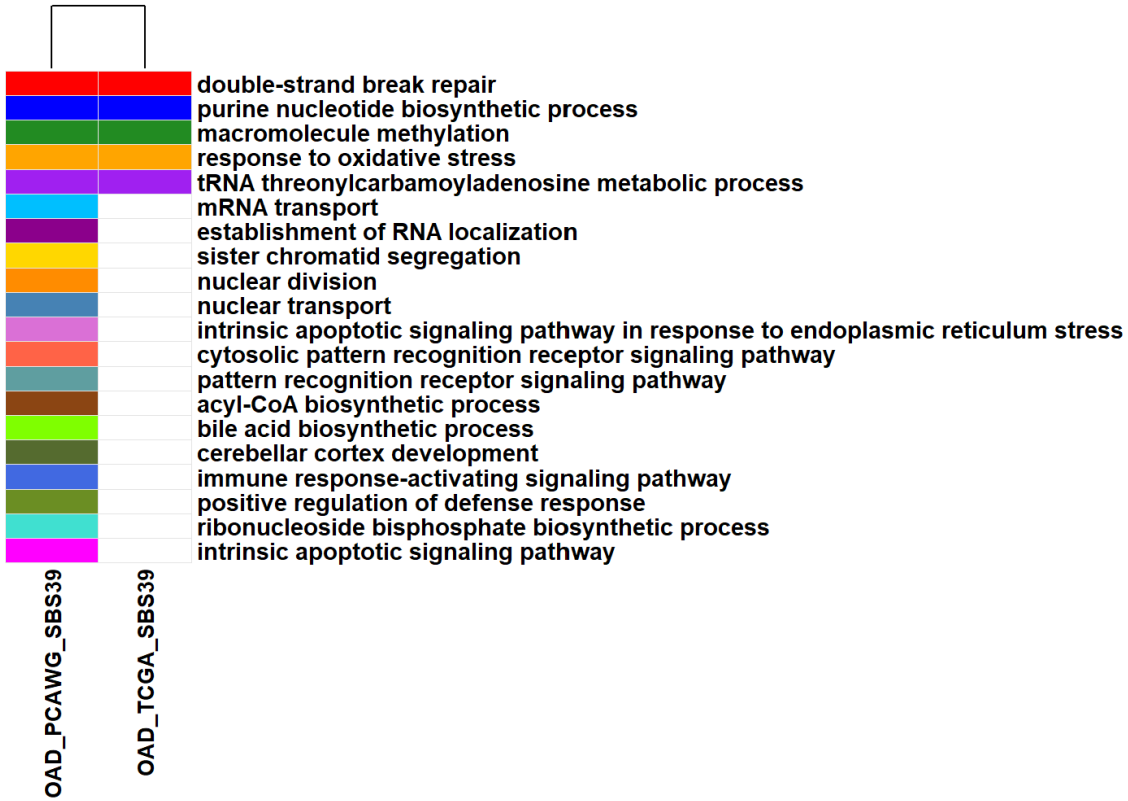
All Datasets

SBS38



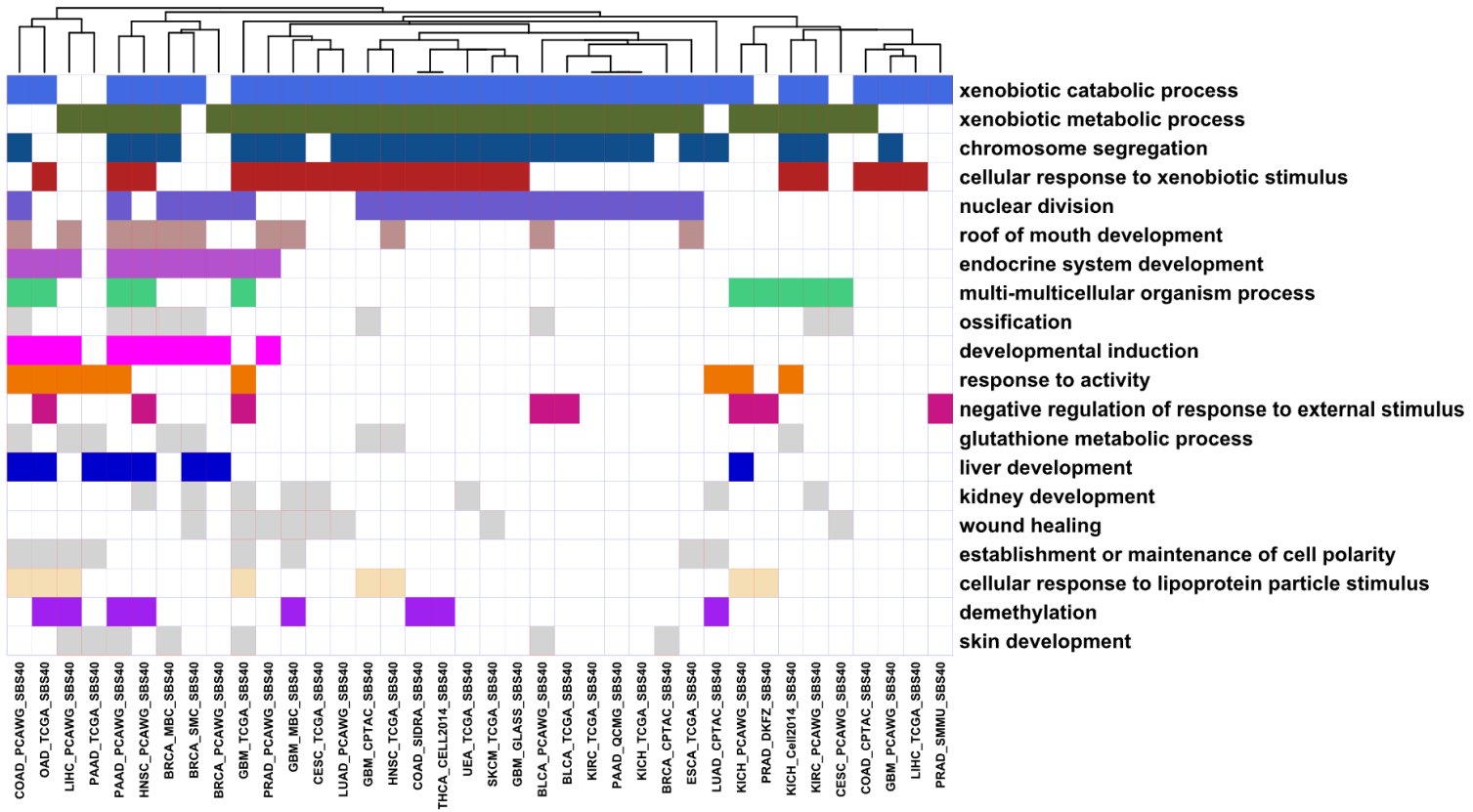
All Datasets

SBS39



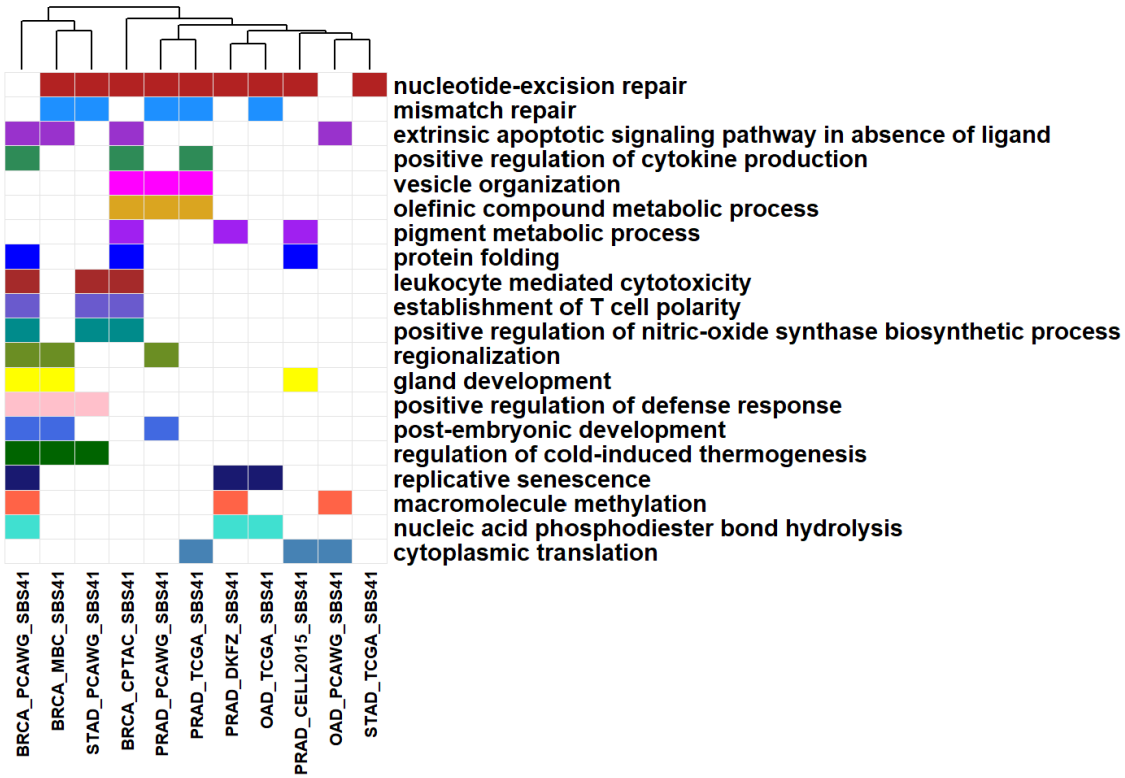
All Datasets

SBS40



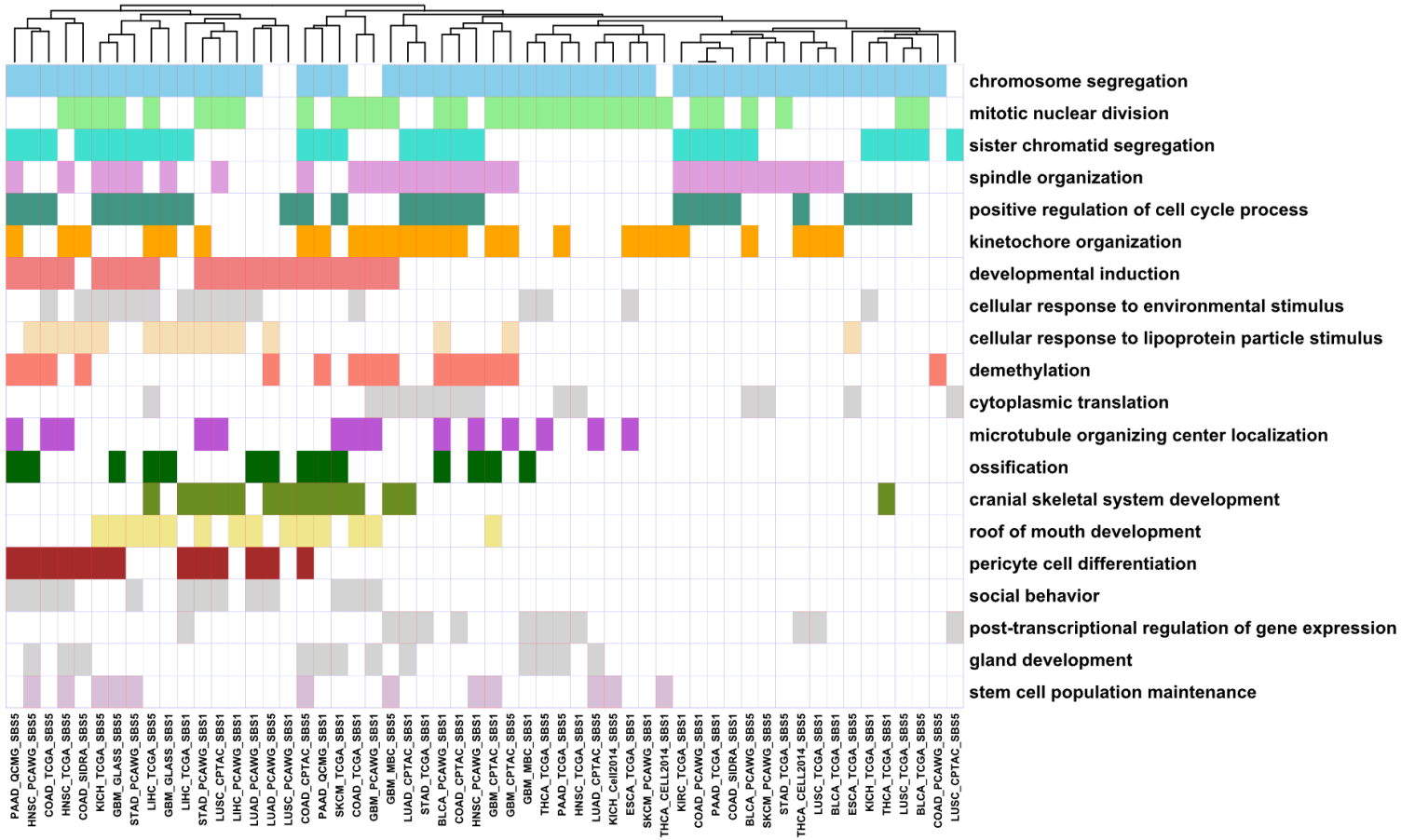
All Datasets

SBS41



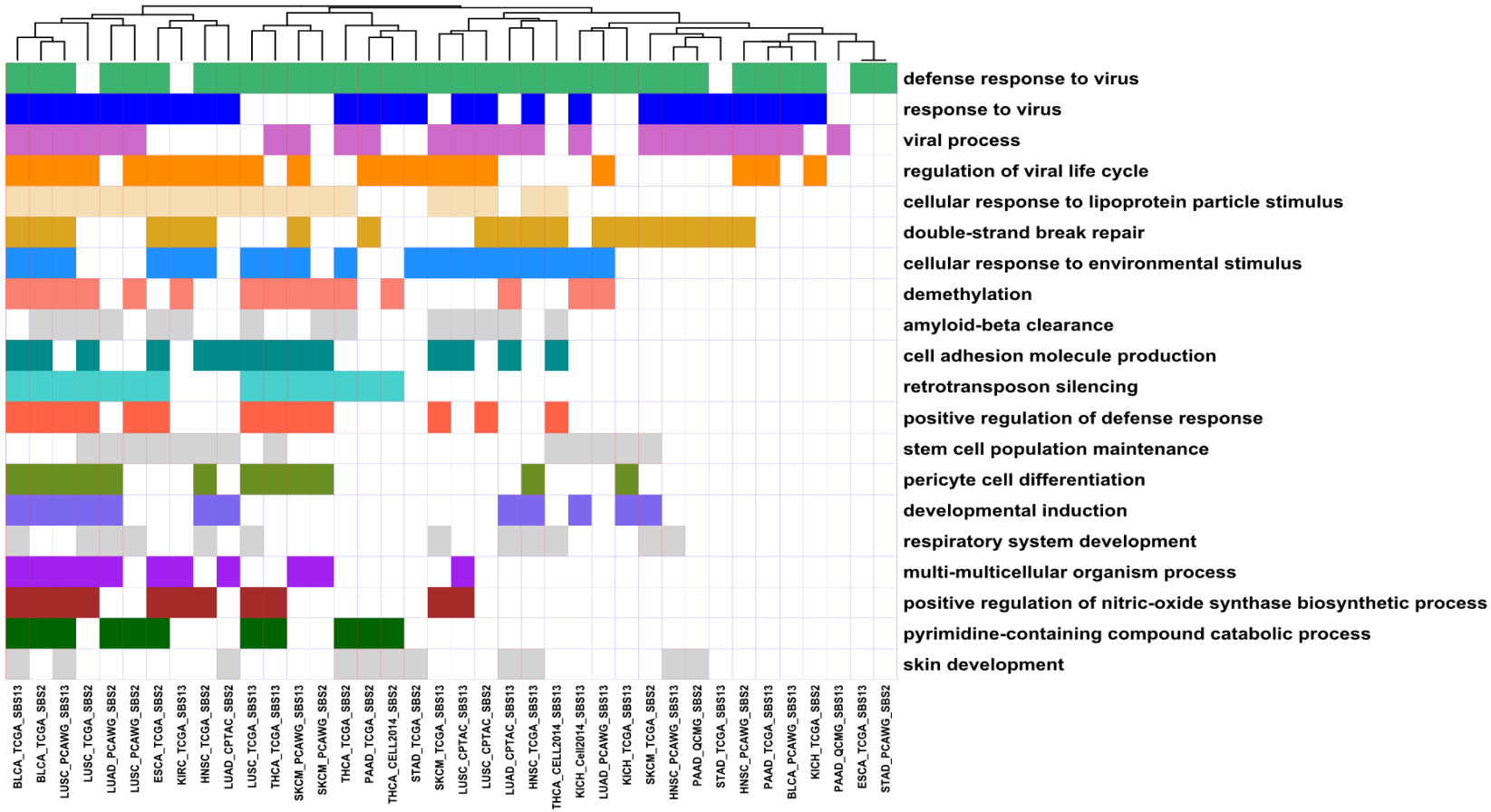
Female

SBS1/5



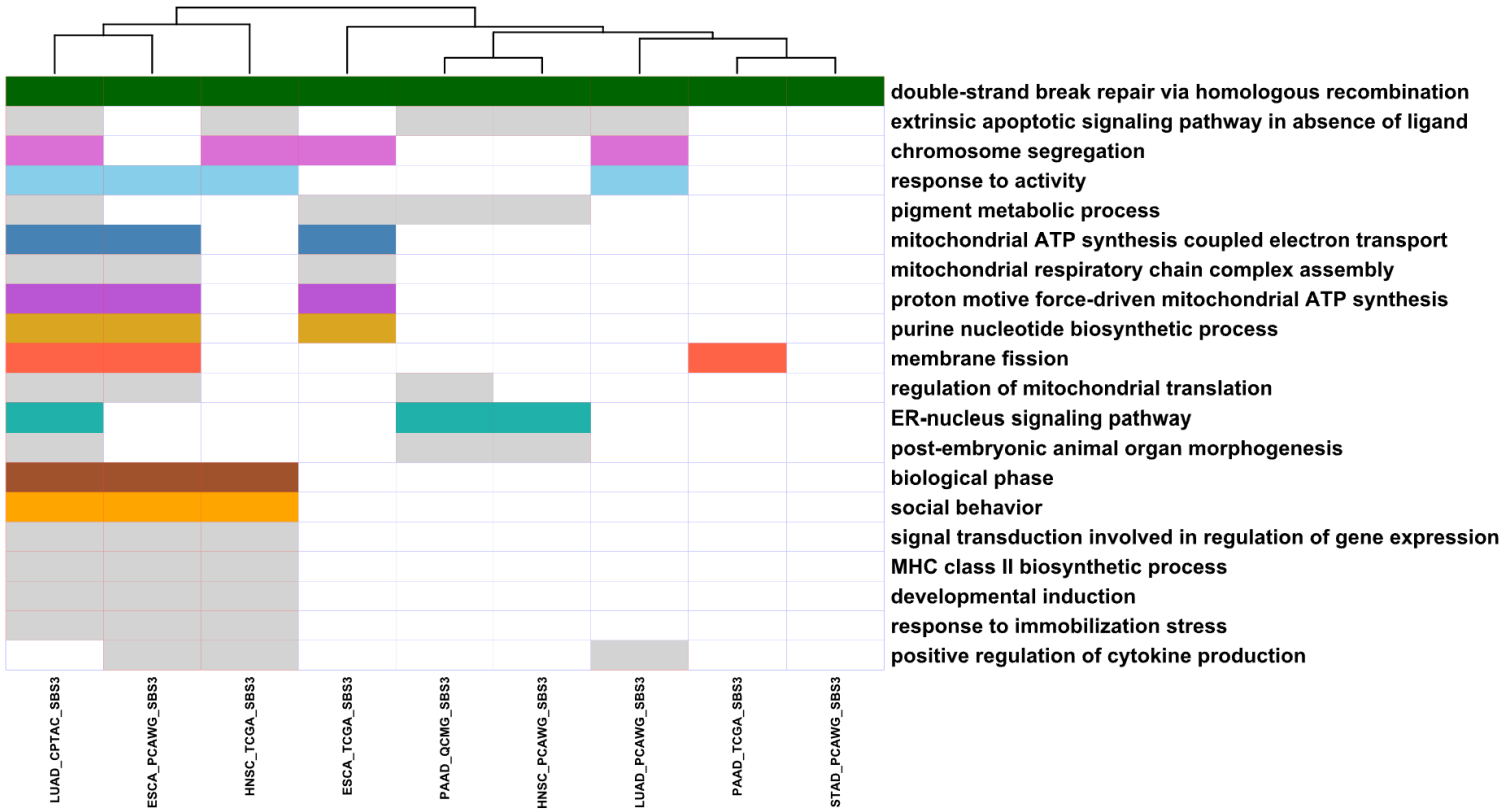
Female

SBS2/13



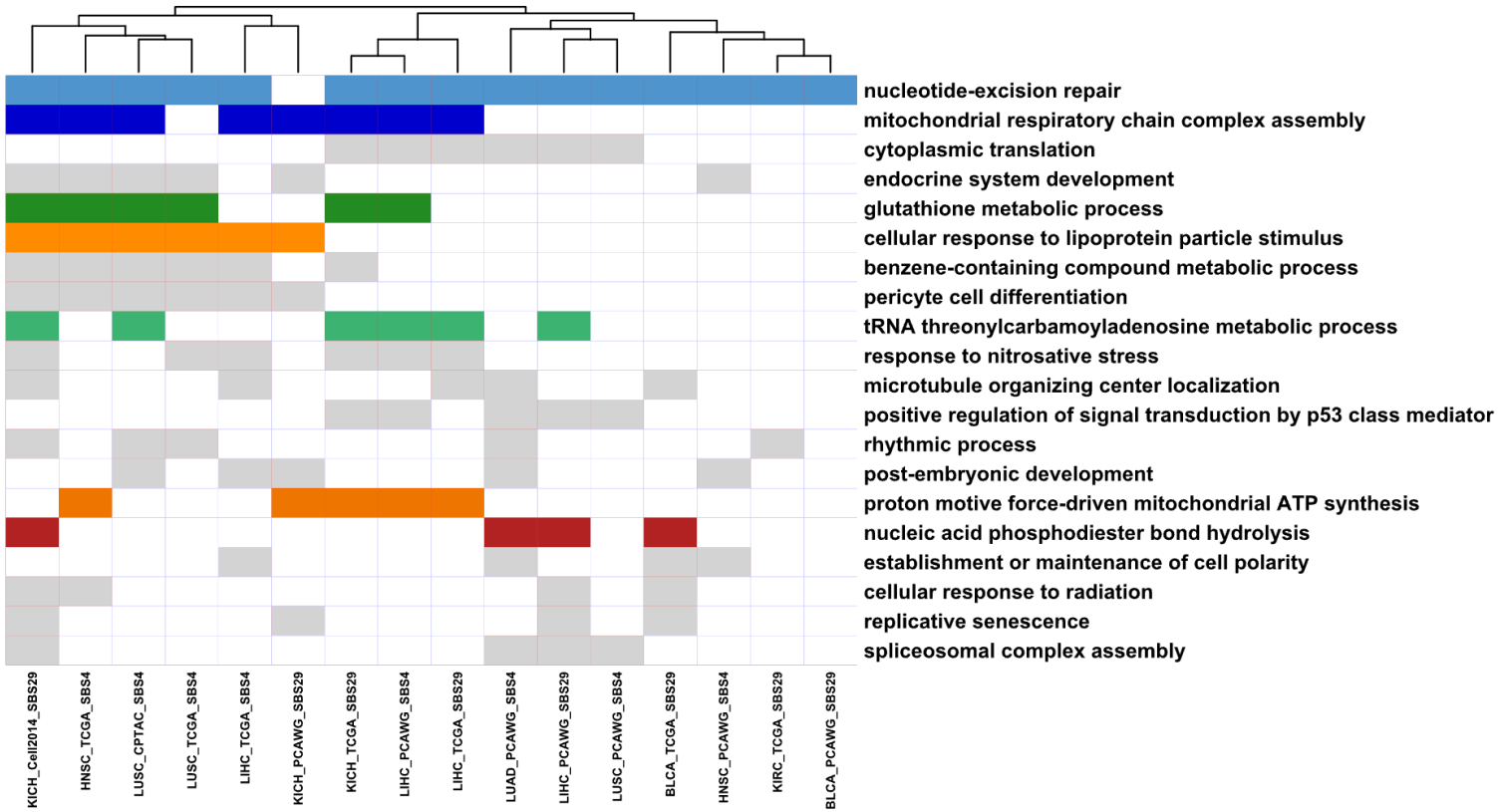
Female

SBS3


double-strand break repair via homologous recombination
extrinsic apoptotic signaling pathway in absence of ligand
chromosome segregation
response to activity
pigment metabolic process
mitochondrial ATP synthesis coupled electron transport
mitochondrial respiratory chain complex assembly
proton motive force-driven mitochondrial ATP synthesis
purine nucleotide biosynthetic process
membrane fission
regulation of mitochondrial translation
ER-nucleus signaling pathway
post-embryonic animal organ morphogenesis
biological phase
social behavior
signal transduction involved in regulation of gene expression
MHC class II biosynthetic process
developmental induction
response to immobilization stress
positive regulation of cytokine production
LUAD_CPTAC_SBS3
ESCA_PCAWG_SBS3
HNSC_TCGA_SBS3
ESCA_TCGA_SBS3
PAAD_QCMG_SBS3
HNSC_PCAWG_SBS3
LUAD_PCAWG_SBS3
PAAD_TCGA_SBS3
STAD_PCAWG_SBS3

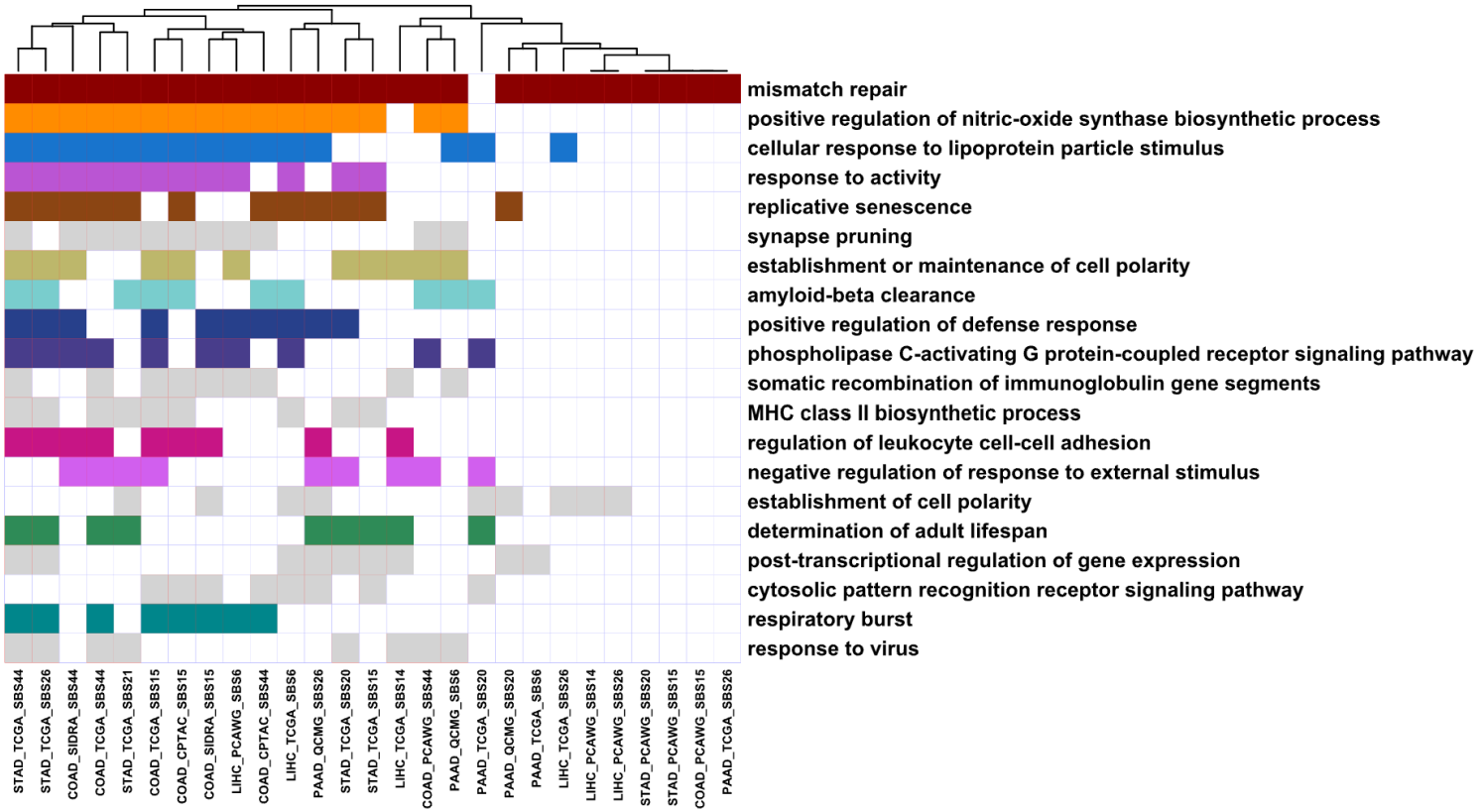
Female

SBS4/29



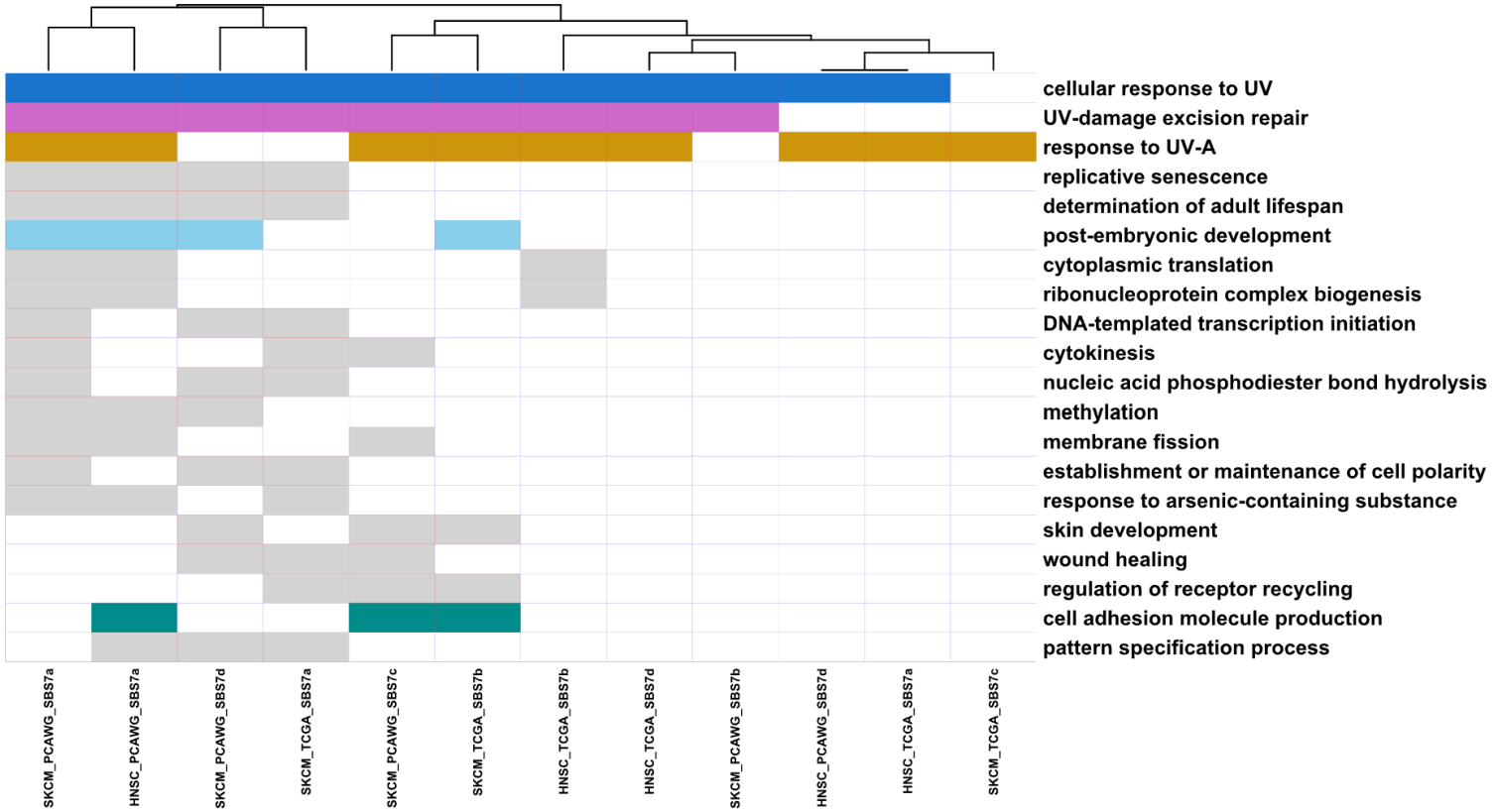
Female

SBS6/14/15/20/21/26/44



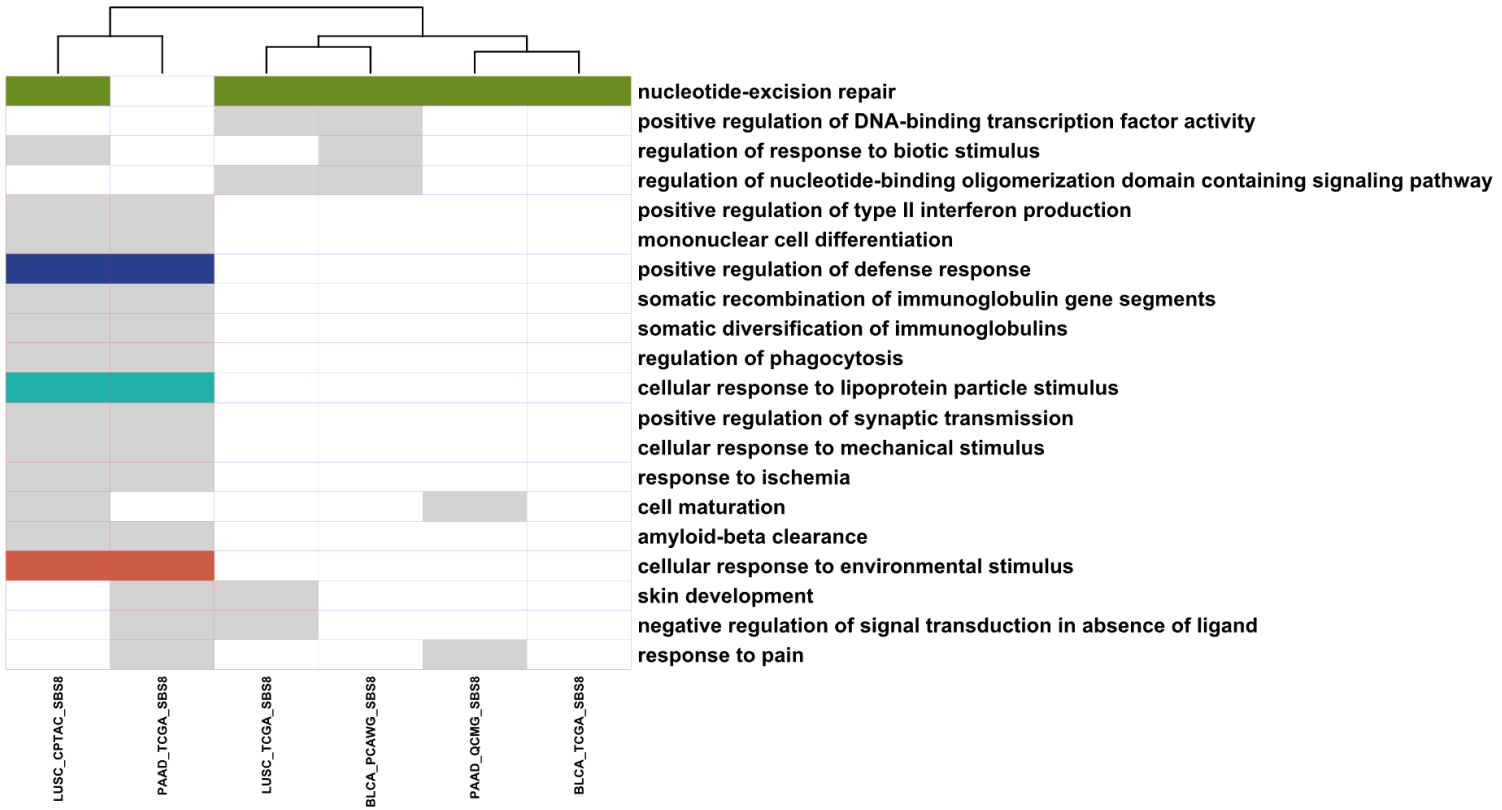
Female

SBS7a/7b/7c/7d



Female

SBS8

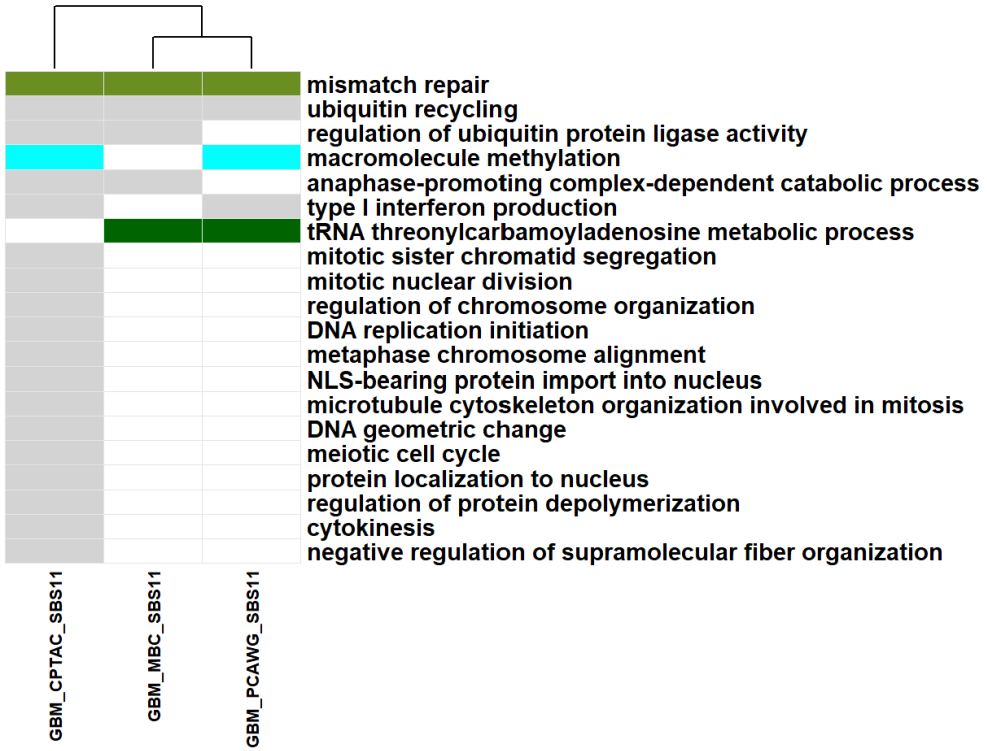

nucleotide-excision repair
positive regulation of DNA-binding transcription factor activity
regulation of response to biotic stimulus
regulation of nucleotide-binding oligomerization domain containing signaling pathway
positive regulation of type II interferon production
mononuclear cell differentiation
positive regulation of defense response
somatic recombination of immunoglobulin gene segments
somatic diversification of immunoglobulins
regulation of phagocytosis
cellular response to lipoprotein particle stimulus
positive regulation of synaptic transmission
cellular response to mechanical stimulus
response to ischemia
cell maturation
amyloid-beta clearance
cellular response to environmental stimulus
skin development
negative regulation of signal transduction in absence of ligand
response to pain
LUSC_CPTAC_SBS8
PAAD_TCGA_SBS8
LUSC_TCGA_SBS8
BLCA_PCAWG_SBS8
PAAD_QCMG_SBS8
BLCA_TCGA_SBS8

Female

SBS10a/10b

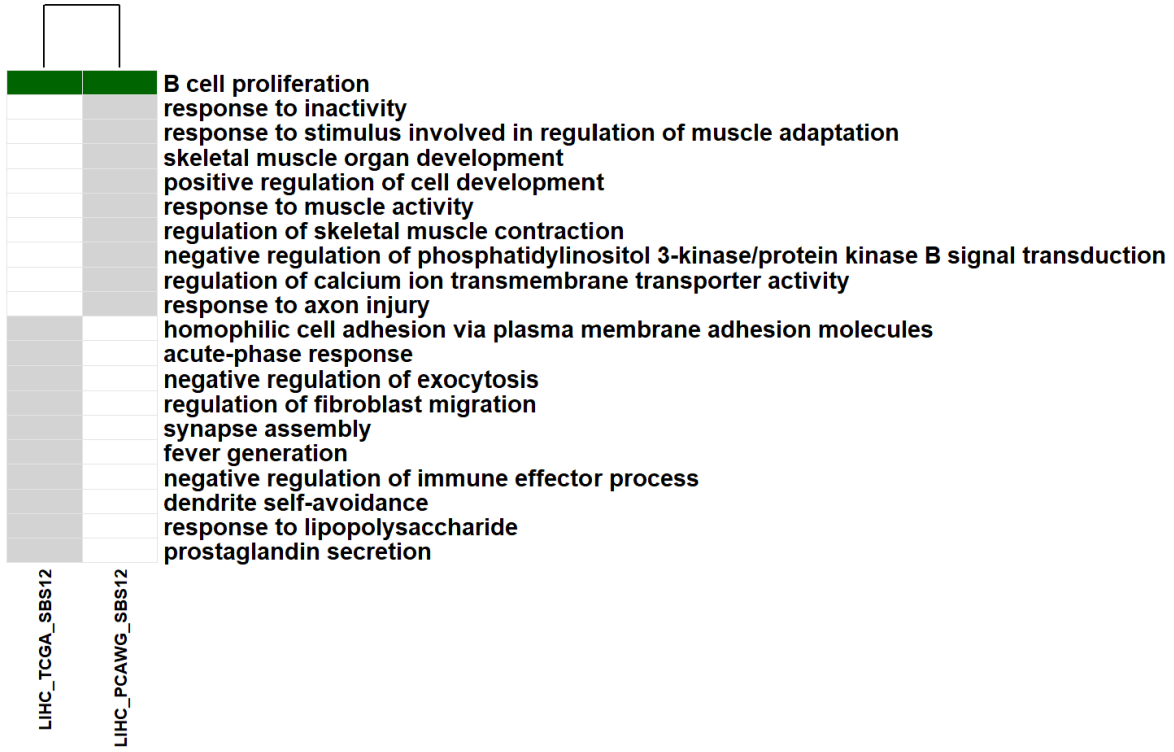
Female

SBS11



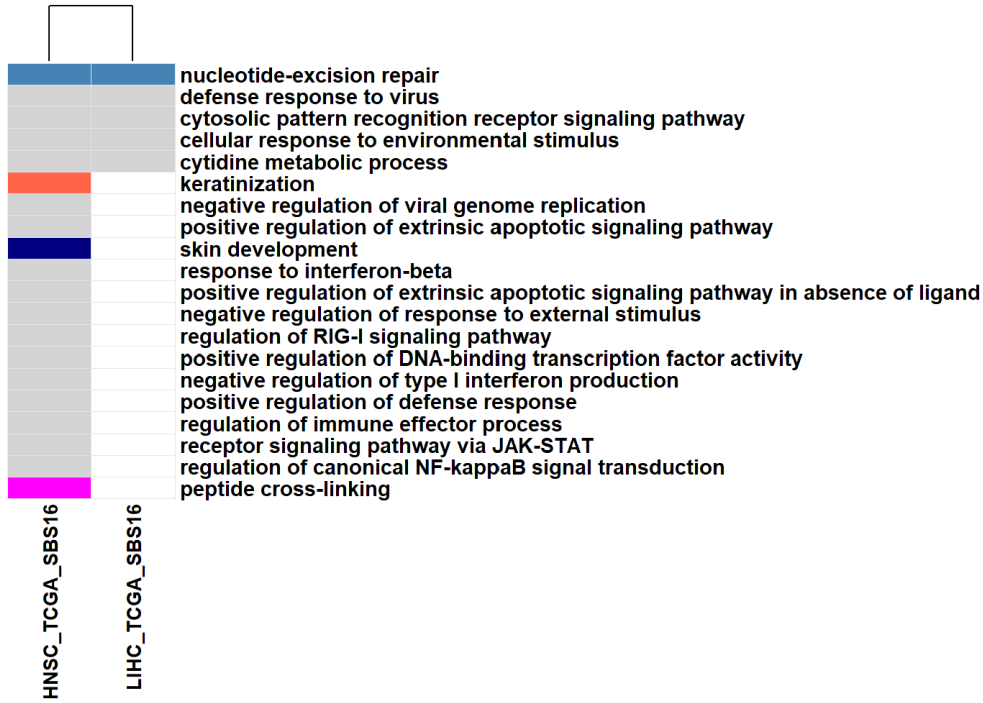
Female

SBS12



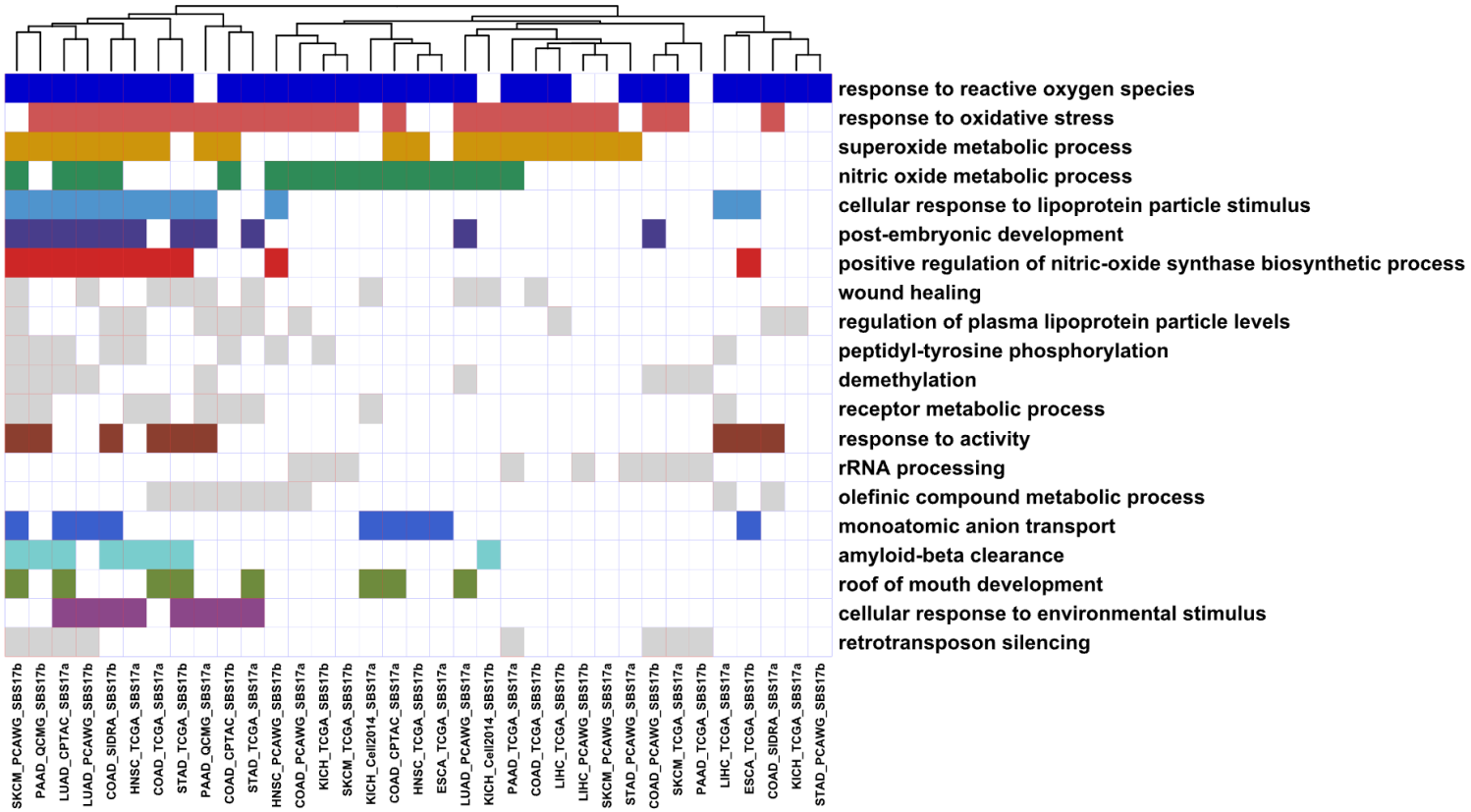
Female

SBS16



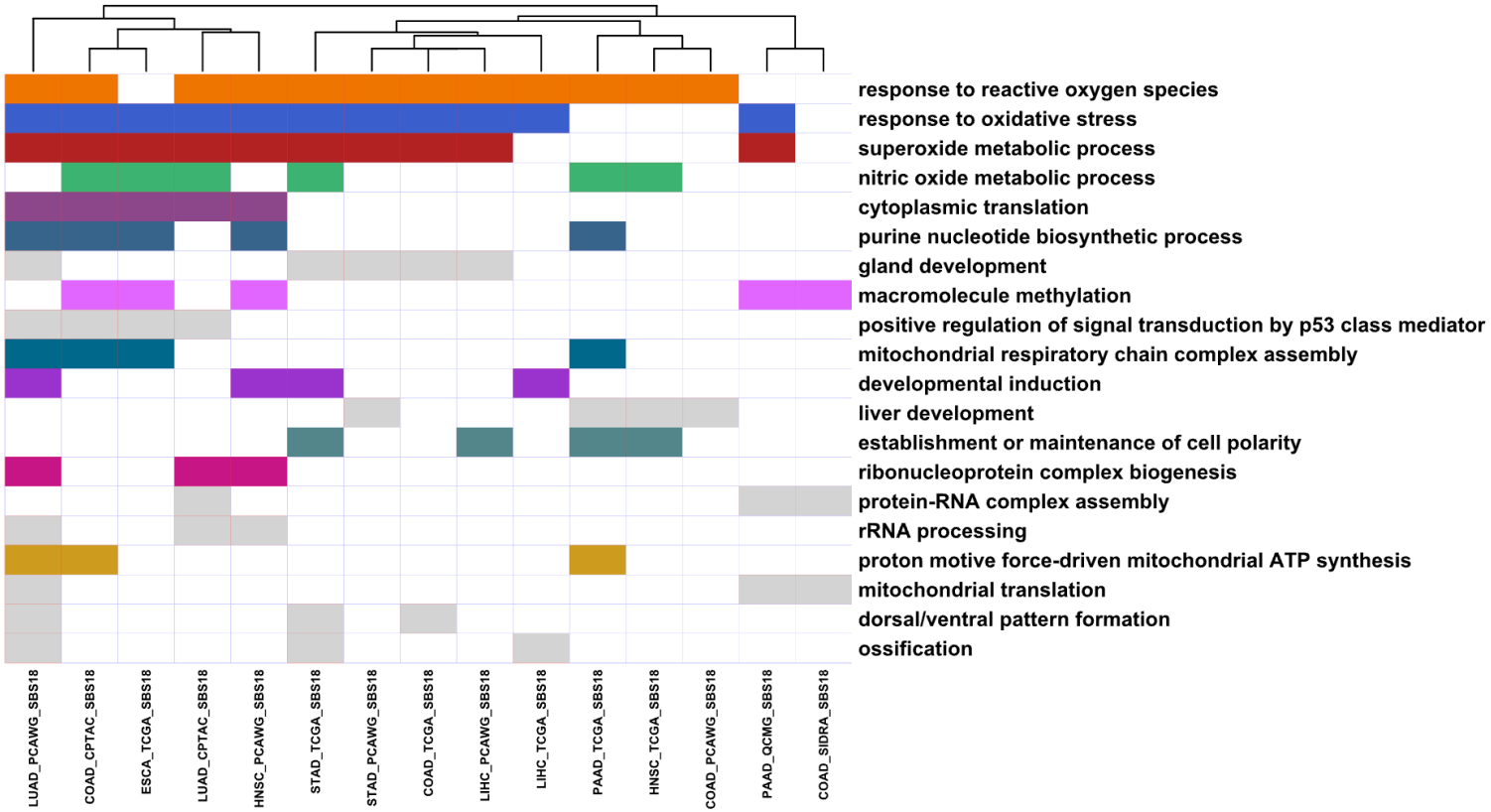
Female

SBS17a/17b



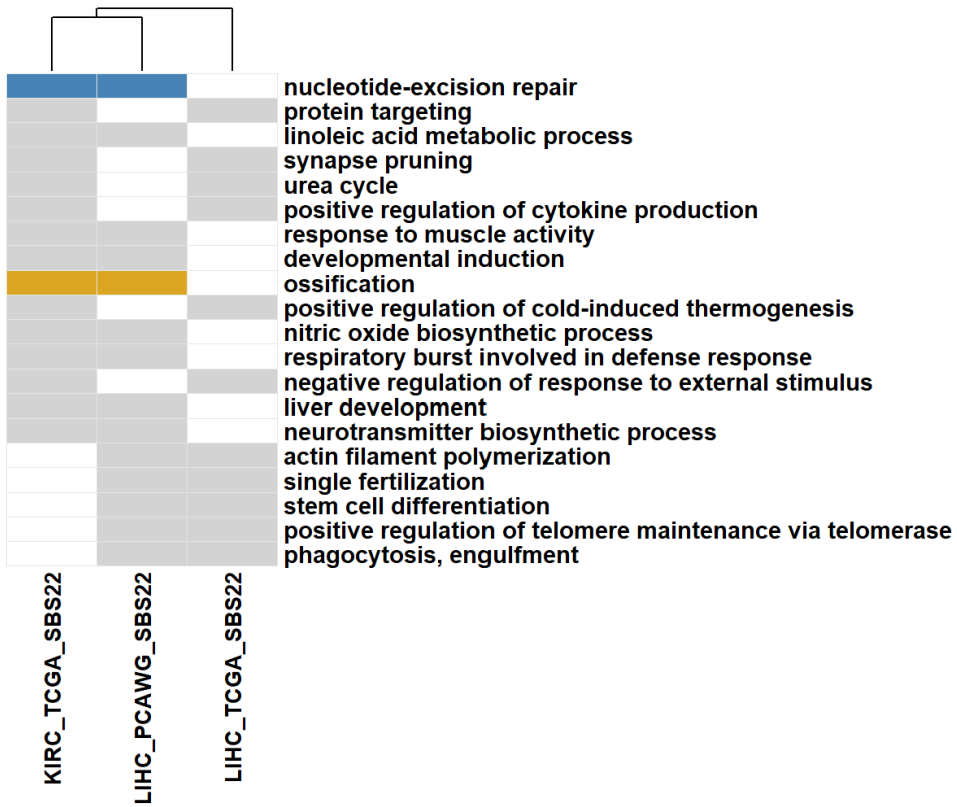
Female

SBS18



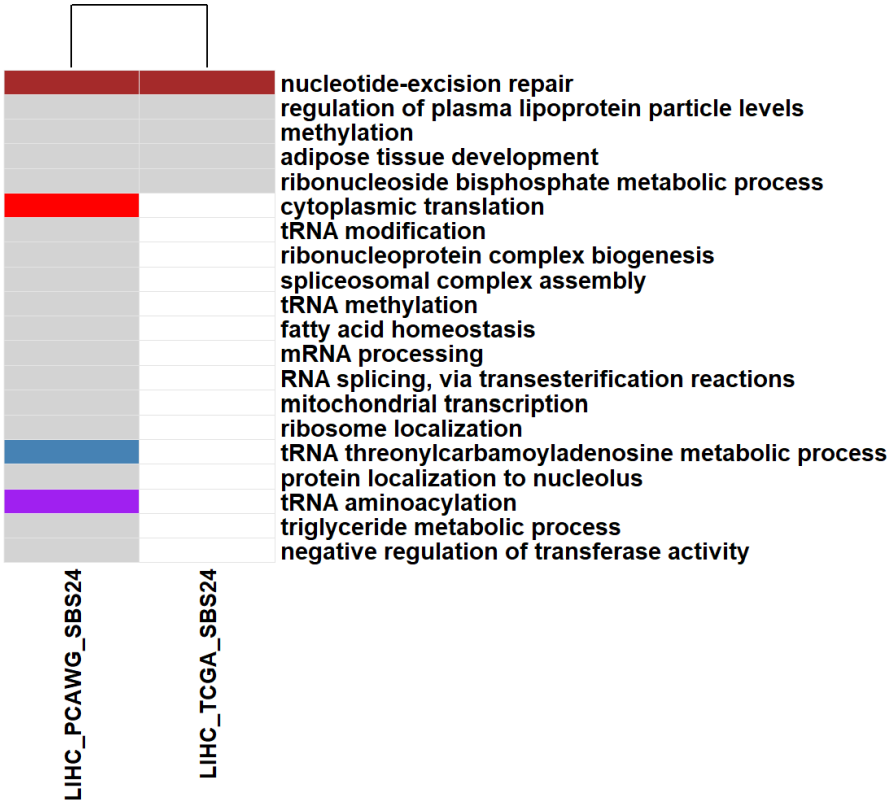
Female

SBS22



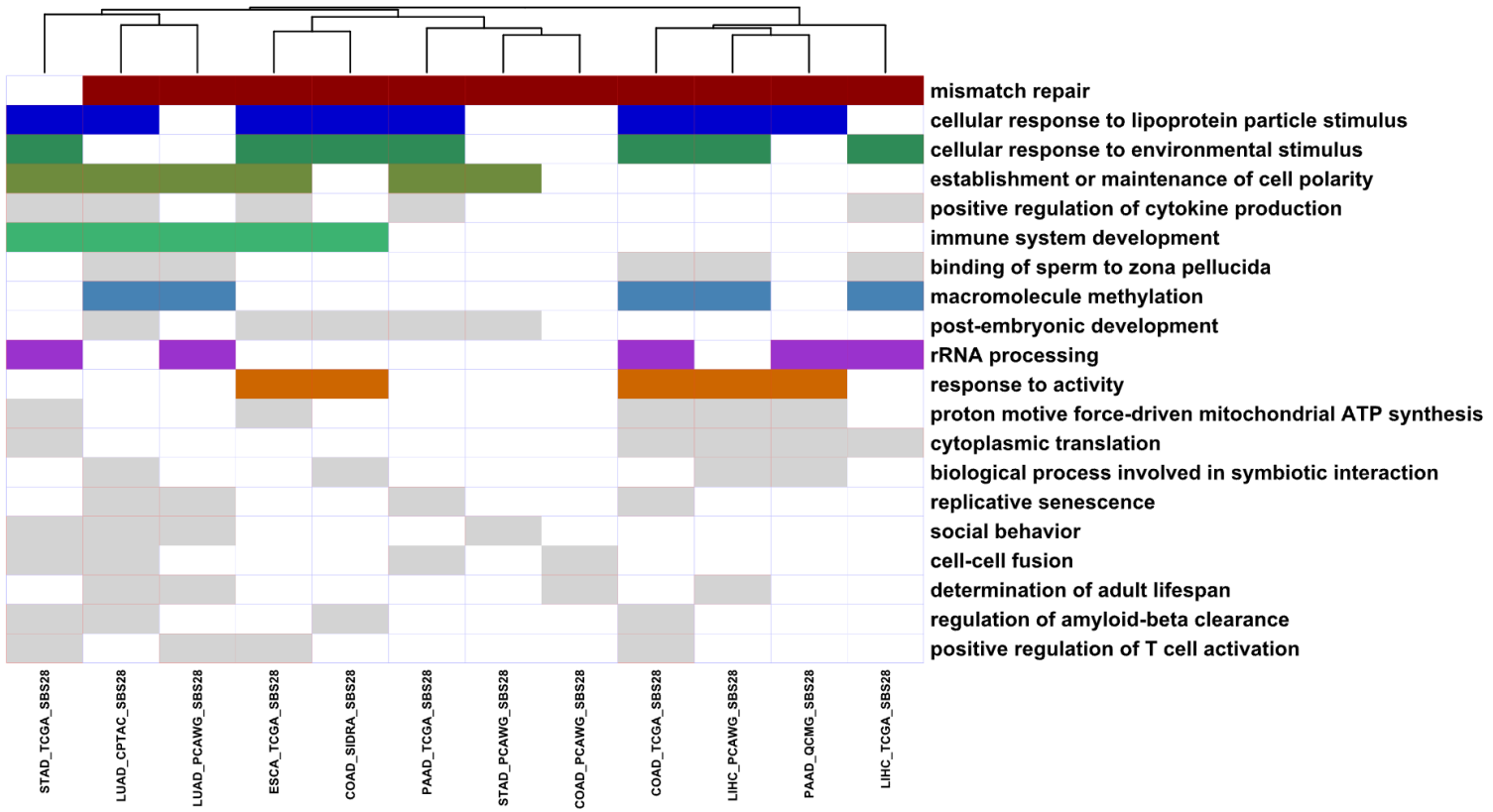
Female

SBS24



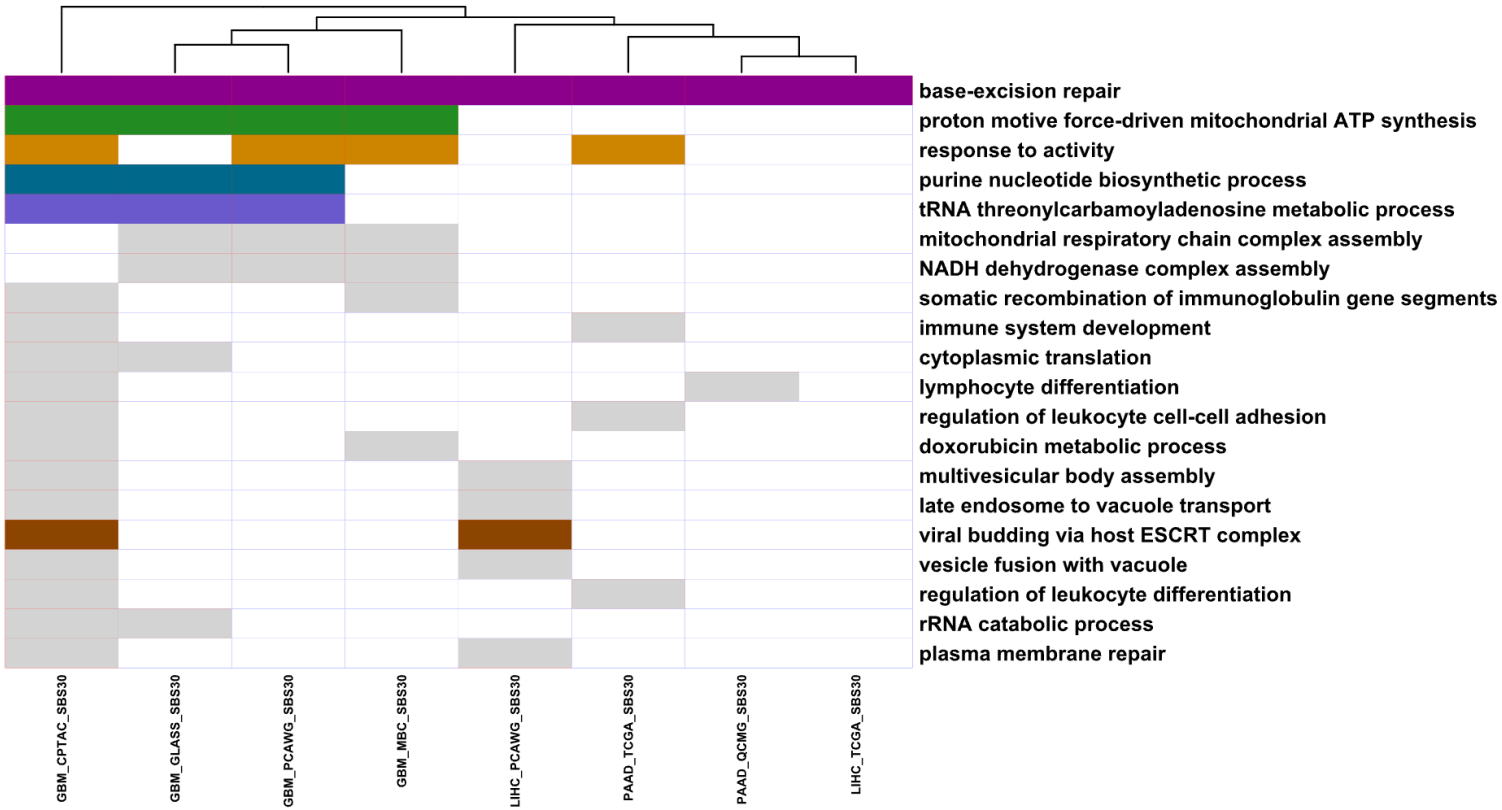
Female

SBS28



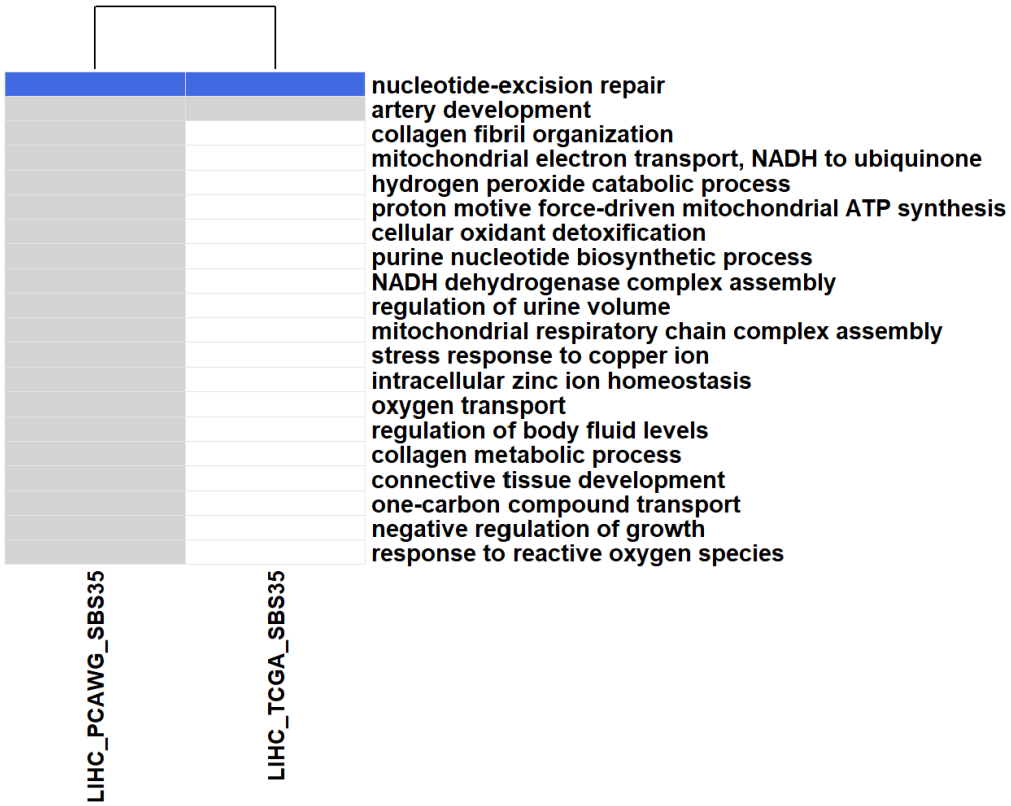
Female

SBS30



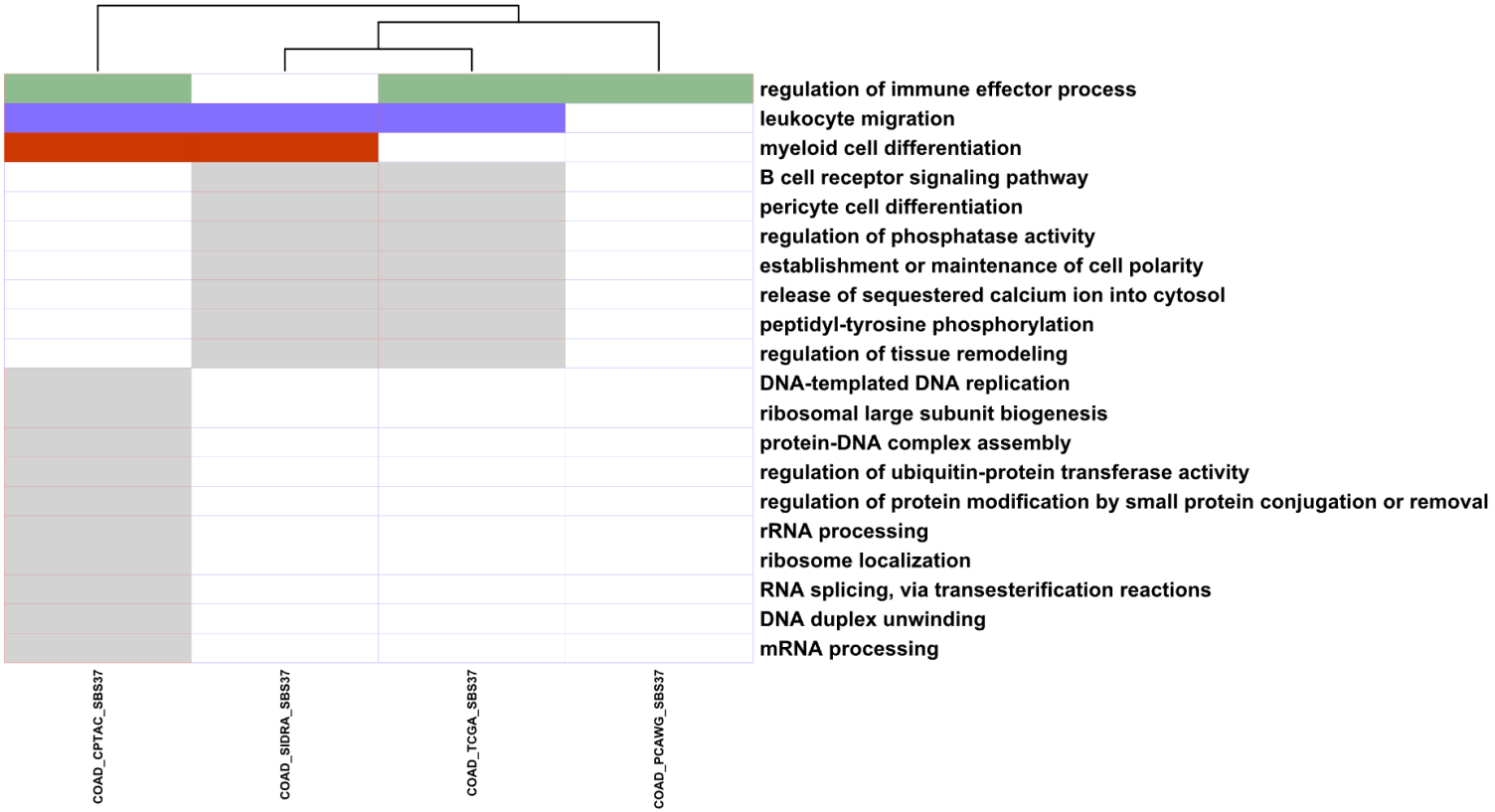
Female

SBS35



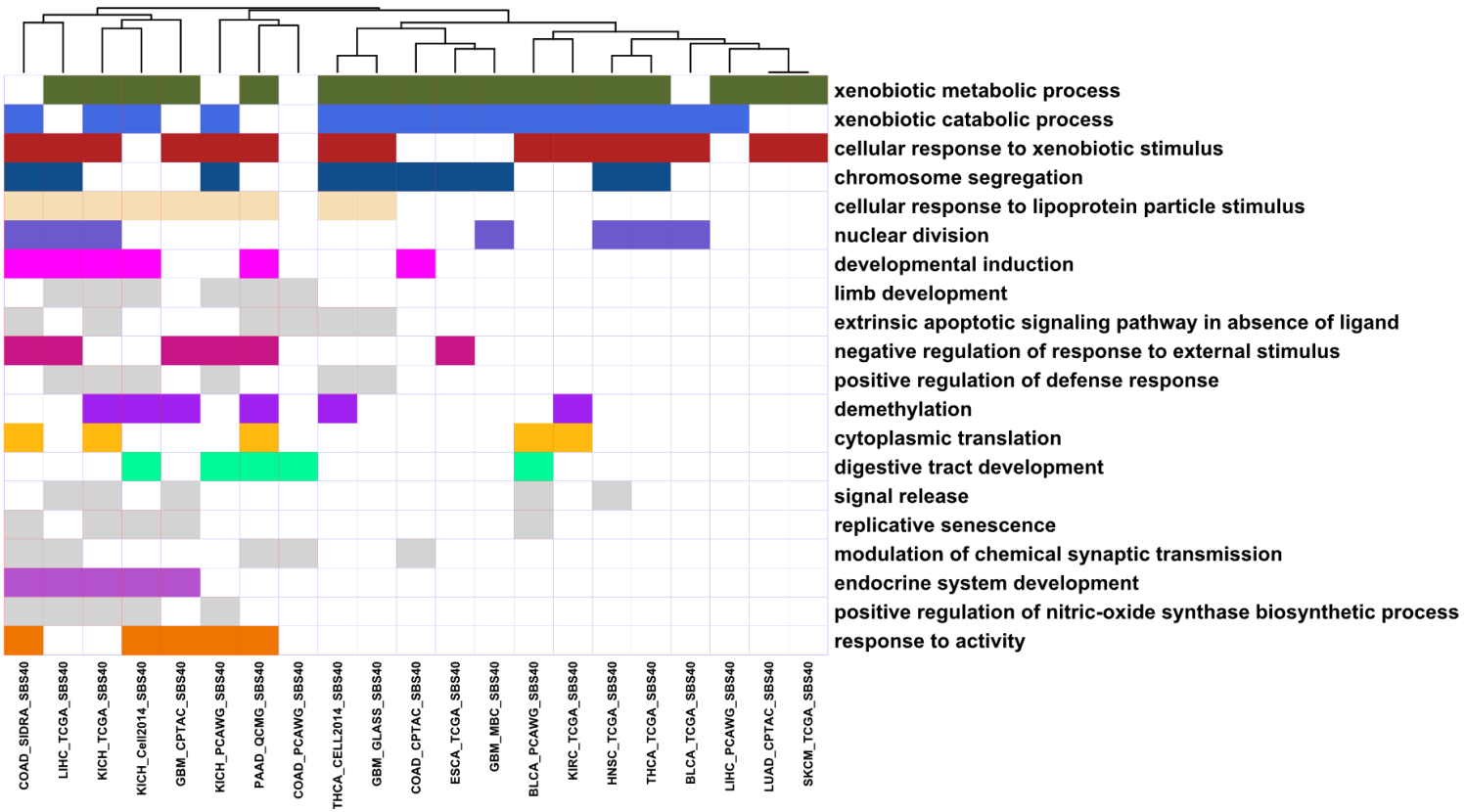
Female

SBS37



Female

SBS40



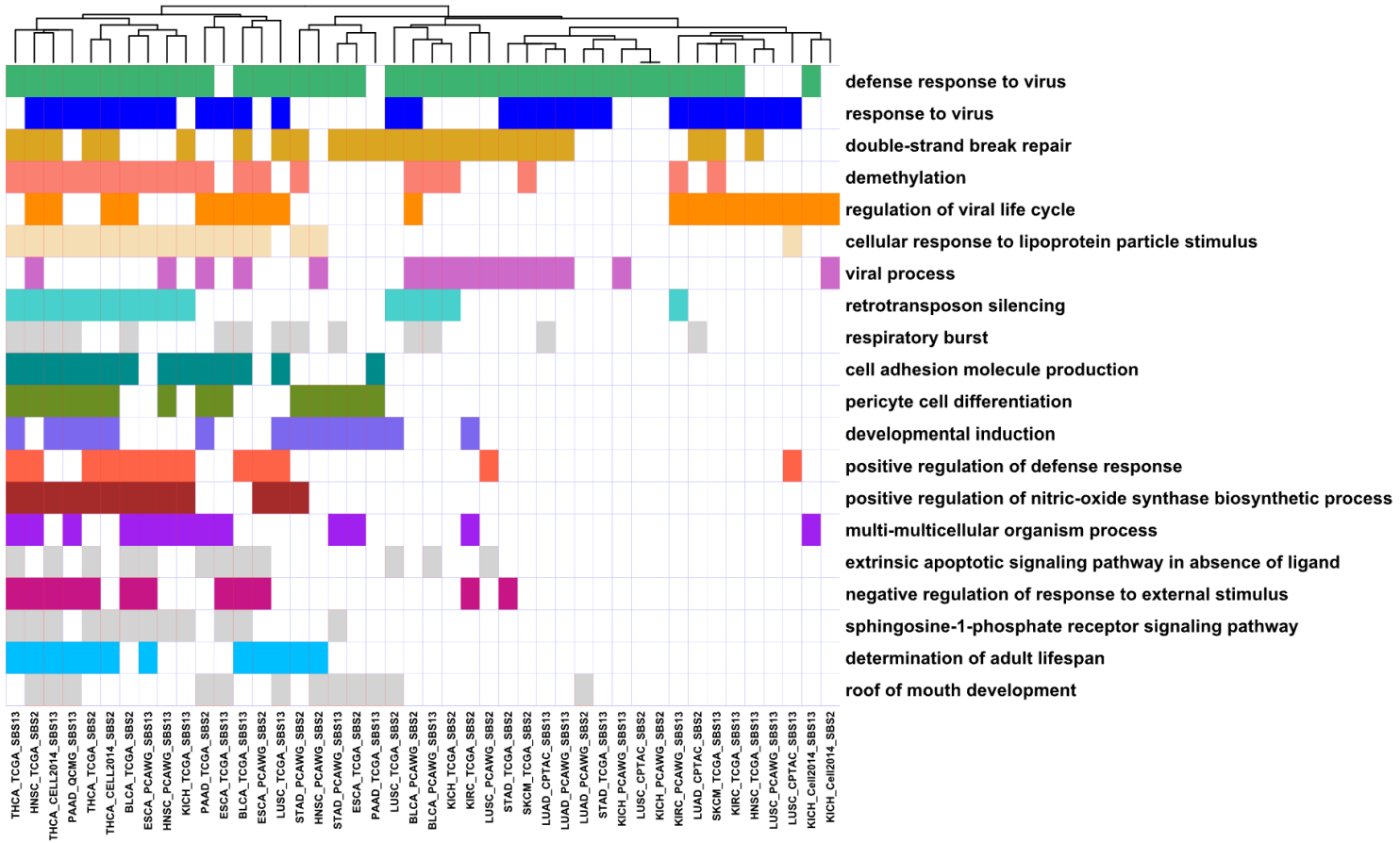
Male

SBS1/5


chromosome segregation
mitotic nuclear division
spindle organization
kinetochore organization
positive regulation of cell cycle process
developmental induction
sister chromatid segregation
signal transduction involved in regulation of gene expression
ossification
cellular response to lipoprotein particle stimulus
liver development
reproductive structure development
demethylation
stem cell population maintenance
post-embryonic development
roof of mouth development
respiratory system development
response to activity
pericyte cell differentiation
microtubule organizing center localization
GBM_PCAWG_SBS5
HNSC_TCGA_SBS5
KIRC_TCGA_SBS5
BLCA_PCAWG_SBS5
ESCA_TCGA_SBS5
KICH_TCGA_SBS5
STAD_PCAWG_SBS1
STAD_PCAWG_SBS5
COAD_PCAWG_SBS1
ESCA_PCAWG_SBS1
HNSC_PCAWG_SBS5
GBM_CPTAC_SBS5
LUSC_CPTAC_SBS1
LUSC_PCAWG_SBS5
LIHC_PCAWG_SBS5
GBM_GLASS_SBS1
GBM_GLASS_SBS5
KICH_Cell2014_SBS5
COAD_TCGA_SBS5
HNSC_TCGA_SBS1
COAD_CPTAC_SBS1
SKCM_TCGA_SBS1
LUAD_PCAWG_SBS5
KICH_Cell2014_SBS1
KIRC_TCGA_SBS1
THCA_CELL2014_SBS1
LUSC_TCGA_SBS1
PAAD_QCMG_SBS5
THCA_TCGA_SBS1
COAD_TCGA_SBS1
LUSC_TCGA_SBS5
GBM_TCGAX_SBS5
COAD_SIDRA_SBS5
GBM_CPTAC_SBS1
LUAD_CPTAC_SBS1
SKCM_TCGA_SBS5
ESCA_TCGA_SBS1
LUAD_PCAWG_SBS1
KICH_TCGA_SBS1
LIHC_PCAWG_SBS1
SKCM_PCAWG_SBS5
KICH_PCAWG_SBS1
LIHC_TCGA_SBS5
PAAD_QCMG_SBS1
LUAD_CPTAC_SBS5
SKCM_PCAWG_SBS1
KIRC_PCAWG_SBS5
COAD_SIDRA_SBS1
LUSC_PCAWG_SBS1
BLCA_PCAWG_SBS1
THCA_PCAWG_SBS1
COAD_CPTAC_SBS5
BLCA_TCGA_SBS1
GBM_TCGAX_SBS1
PAAD_TCGA_SBS1
BLCA_TCGA_SBS5
STAD_TCGA_SBS1
LIHC_TCGA_SBS1
PAAD_TCGA_SBS5
GBM_PCAWG_SBS1
THCA_CELL2014_SBS5
KICH_PCAWG_SBS5
THCA_TCGA_SBS5

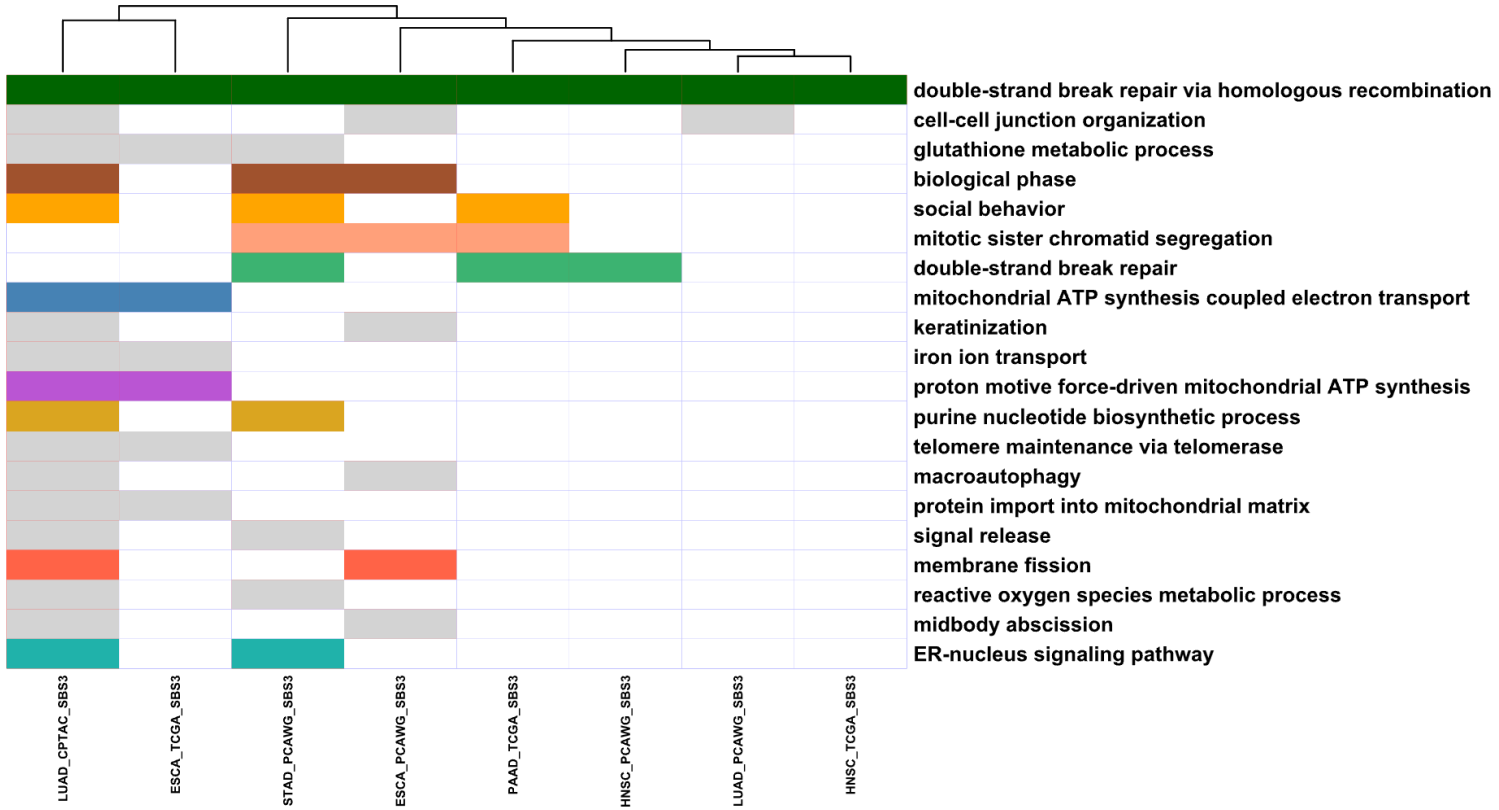
Male

SBS2/13


defense response to virus
response to virus
double-strand break repair
demethylation
regulation of viral life cycle
cellular response to lipoprotein particle stimulus
viral process
retrotransposon silencing
respiratory burst
cell adhesion molecule production
pericyte cell differentiation
developmental induction
positive regulation of defense response
positive regulation of nitric-oxide synthase biosynthetic process
multi-multicellular organism process
extrinsic apoptotic signaling pathway in absence of ligand
negative regulation of response to external stimulus
sphingosine-1-phosphate receptor signaling pathway
determination of adult lifespan
roof of mouth development
THCA_TCGA_SBS13
HNSC_TCGA_SBS2
THCA_CELL2014_SBS13
PAAD_QCMG_SBS13
THCA_TCGA_SBS2
THCA_CELL2014_SBS2
BLCA_TCGA_SBS2
ESCA_PCAWG_SBS13
HNSC_PCAWG_SBS13
KICH_TCGA_SBS13
PAAD_TCGA_SBS2
ESCA_TCGA_SBS13
BLCA_TCGA_SBS13
ESCA_PCAWG_SBS2
LUSC_TCGA_SBS13
STAD_PCAWG_SBS2
HNSC_PCAWG_SBS2
STAD_PCAWG_SBS13
ESCA_TCGA_SBS2
PAAD_TCGA_SBS13
LUSC_TCGA_SBS2
BLCA_PCAWG_SBS2
BLCA_PCAWG_SBS13
KICH_TCGA_SBS2
KIRC_TCGA_SBS2
LUSC_PCAWG_SBS2
STAD_TCGA_SBS2
SKCM_TCGA_SBS2
LUAD_CPTAC_SBS13
LUAD_PCAWG_SBS13
LUAD_PCAWG_SBS2
STAD_TCGA_SBS13
KICH_PCAWG_SBS13
LUSC_CPTAC_SBS2
KICH_PCAWG_SBS2
KIRC_PCAWG_SBS13
LUAD_CPTAC_SBS2
SKCM_TCGA_SBS13
KIRC_TCGA_SBS13
HNSC_TCGA_SBS13
LUSC_PCAWG_SBS13
LUSC_CPTAC_SBS13
KICH_Cell2014_SBS13
KICH_Cell2014_SBS2

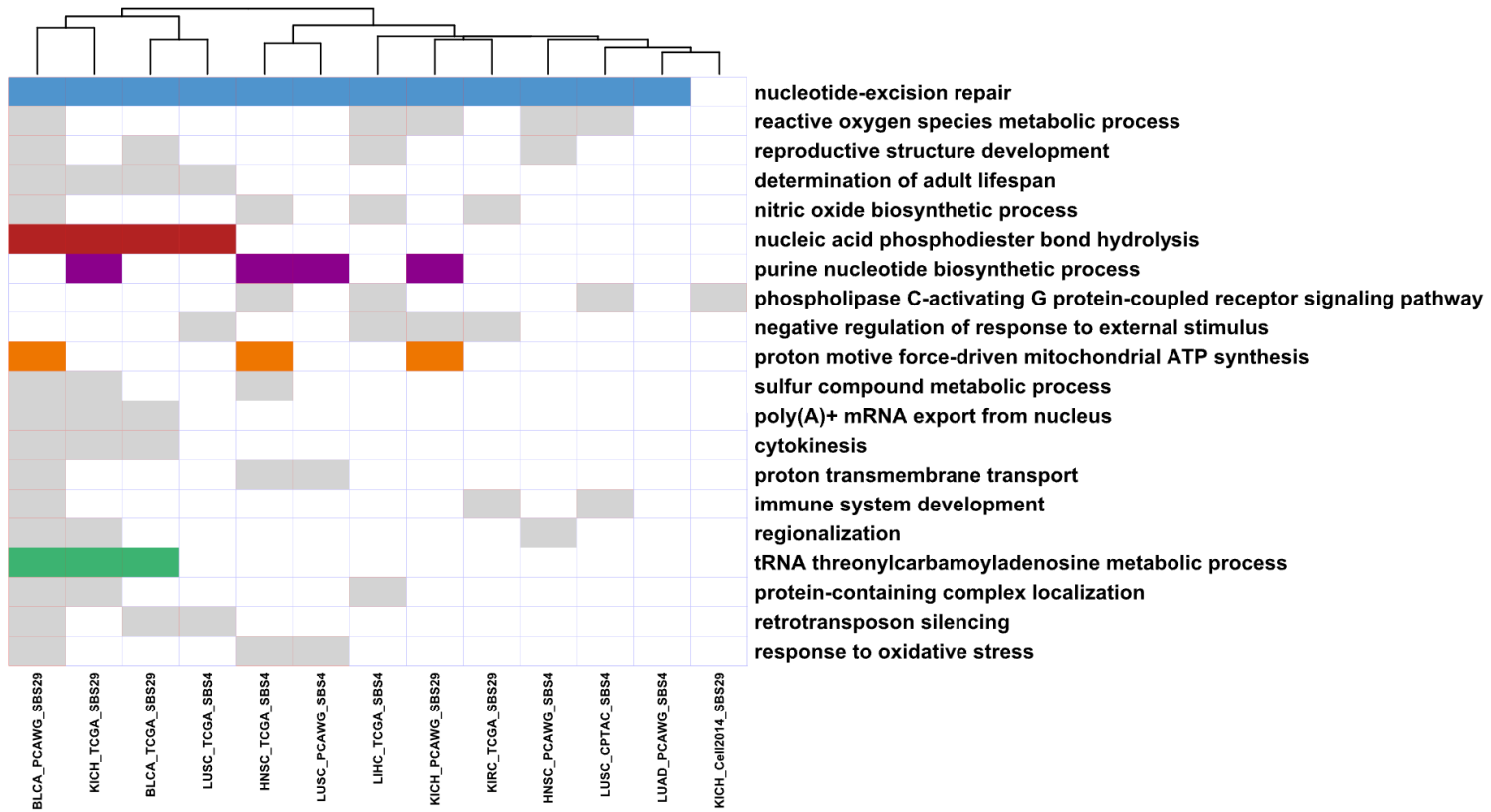
Male

SBS3


double-strand break repair via homologous recombination
cell-cell junction organization
glutathione metabolic process
biological phase
social behavior
mitotic sister chromatid segregation
double-strand break repair
mitochondrial ATP synthesis coupled electron transport
keratinization
iron ion transport
proton motive force-driven mitochondrial ATP synthesis
purine nucleotide biosynthetic process
telomere maintenance via telomerase
macroautophagy
protein import into mitochondrial matrix
signal release
membrane fission
reactive oxygen species metabolic process
midbody abscission
ER-nucleus signaling pathway
LUAD_CPTAC_SBS3
ESCA_TCGA_SBS3
STAD_PCAWG_SBS3
ESCA_PCAWG_SBS3
PAAD_TCGA_SBS3
HNSC_PCAWG_SBS3
LUAD_PCAWG_SBS3
HNSC_TCGA_SBS3

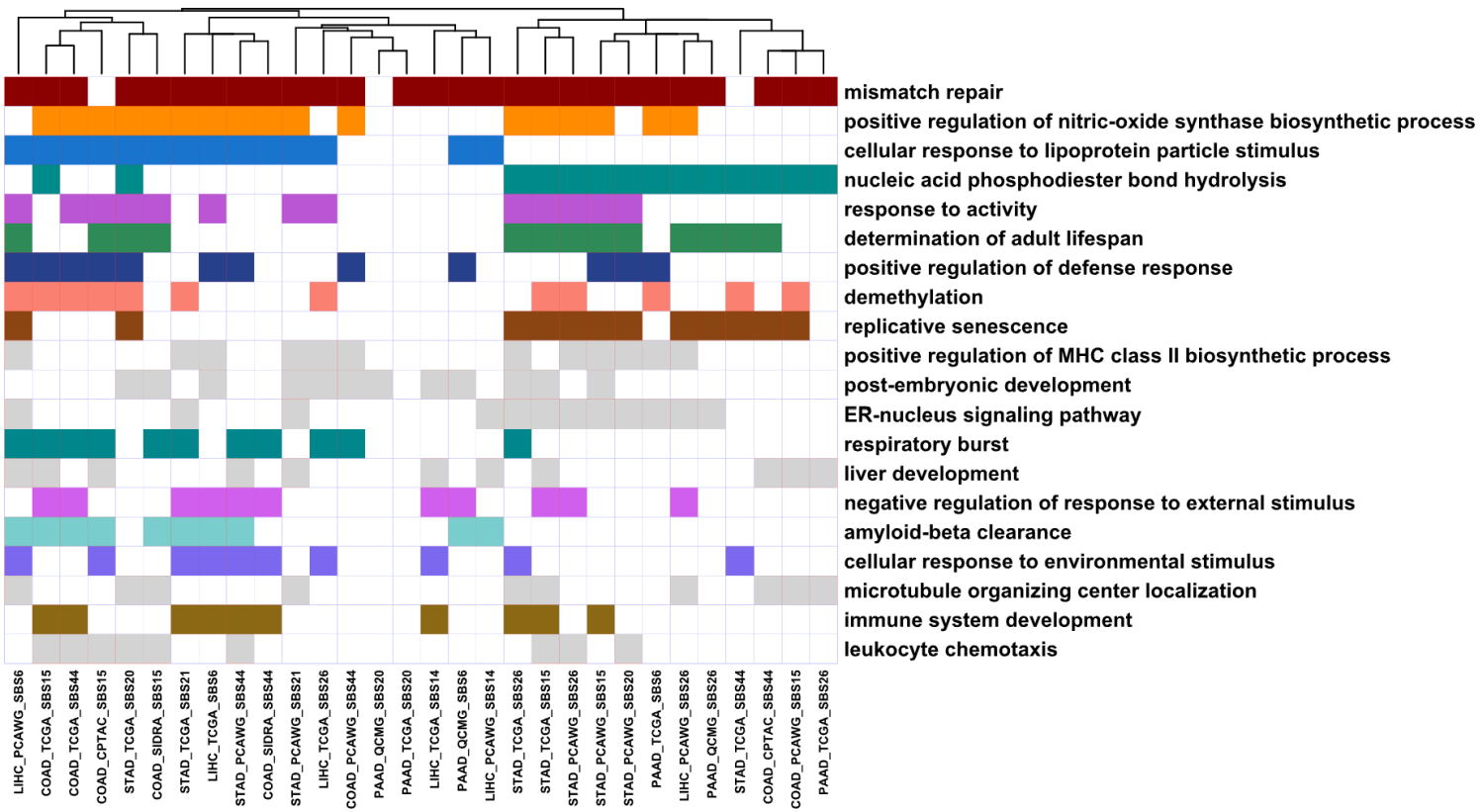
Male

SBS4/29


nucleotide-excision repair
reactive oxygen species metabolic process
reproductive structure development
determination of adult lifespan
nitric oxide biosynthetic process
nucleic acid phosphodiester bond hydrolysis
purine nucleotide biosynthetic process
phospholipase C-activating G protein-coupled receptor signaling pathway
negative regulation of response to external stimulus
proton motive force-driven mitochondrial ATP synthesis
sulfur compound metabolic process
poly(A)+ mRNA export from nucleus
cytokinesis
proton transmembrane transport
immune system development
regionalization
tRNA threonylcarbamoyladenosine metabolic process
protein-containing complex localization
retrotransposon silencing
response to oxidative stress
BLCA_PCAWG_SBS29
KICH_TCGA_SBS29
BLCA_TCGA_SBS29
LUSC_TCGA_SBS4
HNSC_TCGA_SBS4
LUSC_PCAWG_SBS4
LIHC_TCGA_SBS4
KICH_PCAWG_SBS29
KIRC_TCGA_SBS29
HNSC_PCAWG_SBS4
LUSC_CPTAC_SBS4
LUAD_PCAWG_SBS4
KICH_Cell2014_SBS29

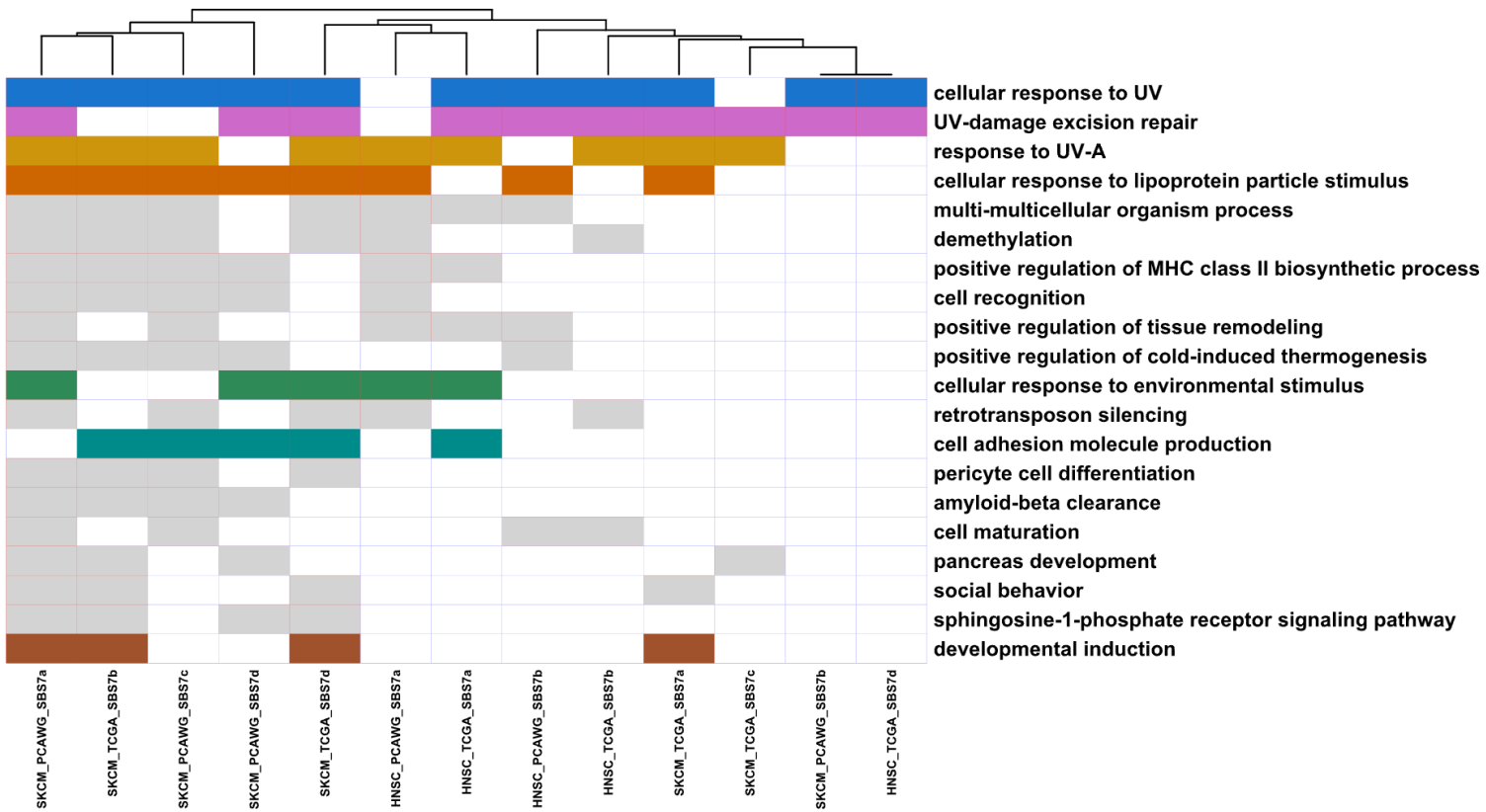
Male

SBS6/14/15/20/21/26/44



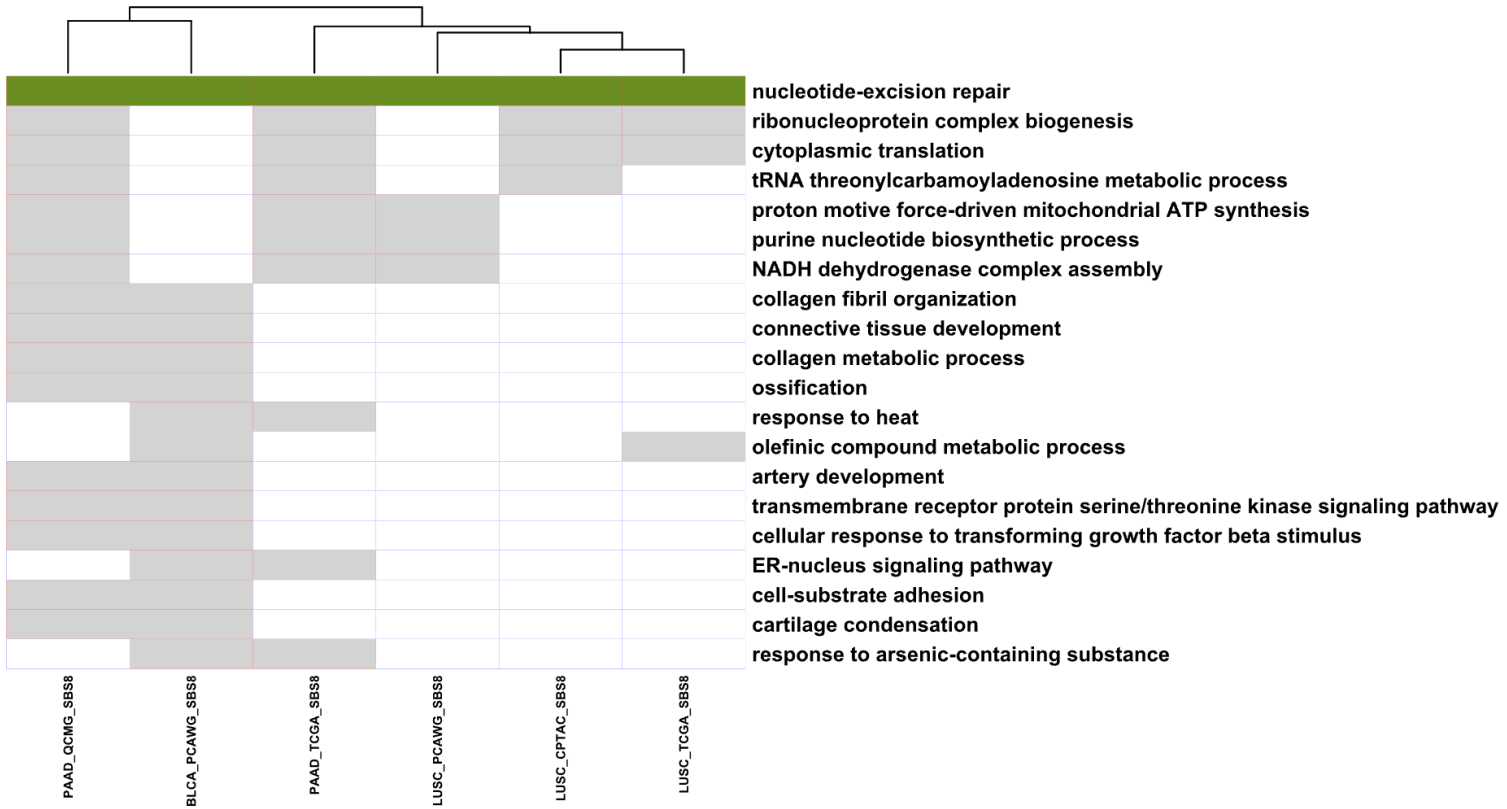
Male

SBS7a/7b/7c/7d



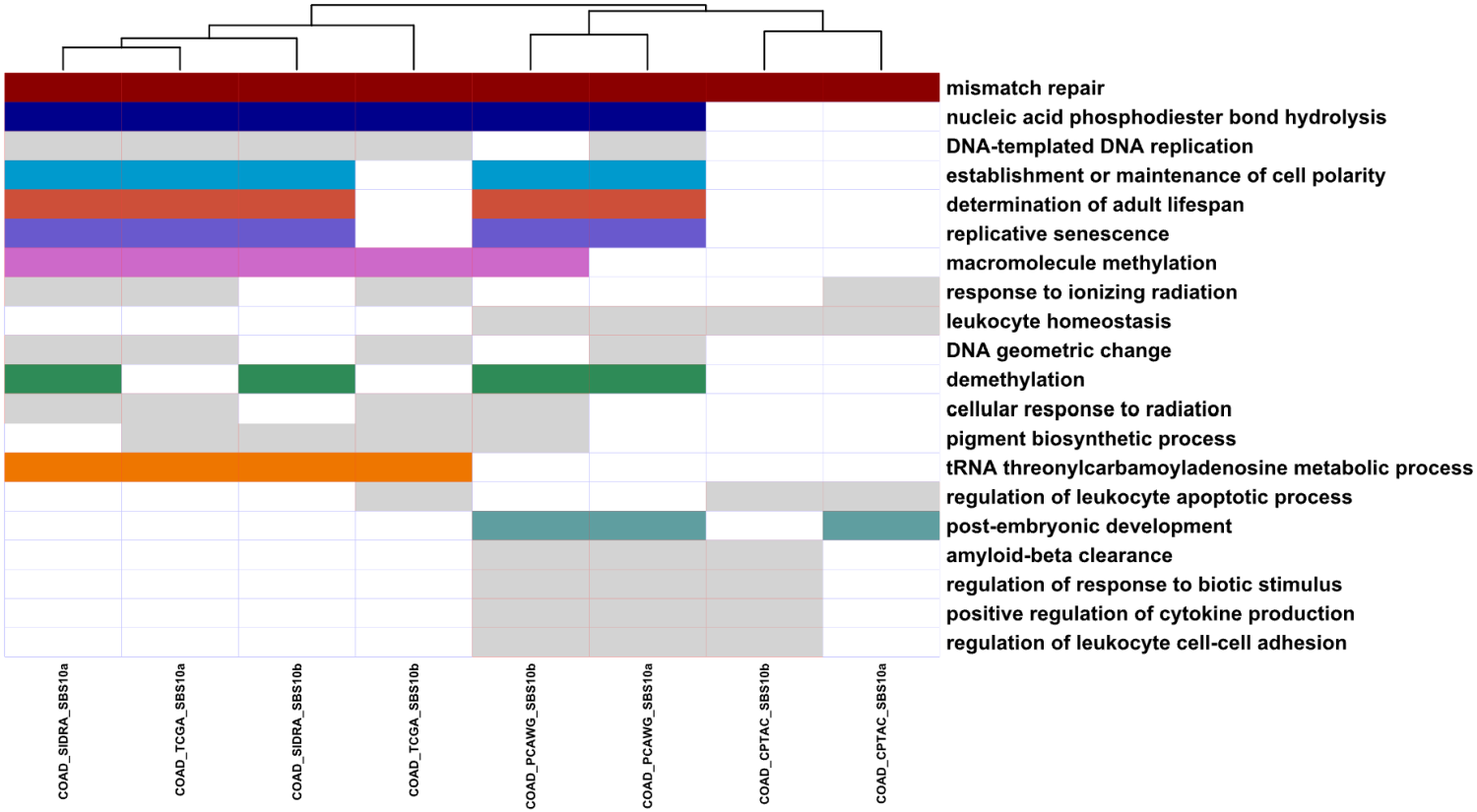
Male

SBS8


nucleotide-excision repair
ribonucleoprotein complex biogenesis
cytoplasmic translation
tRNA threonylcarbamoyladenosine metabolic process
proton motive force-driven mitochondrial ATP synthesis
purine nucleotide biosynthetic process
NADH dehydrogenase complex assembly
collagen fibril organization
connective tissue development
collagen metabolic process
ossification
response to heat
olefinic compound metabolic process
artery development
transmembrane receptor protein serine/threonine kinase signaling pathway
cellular response to transforming growth factor beta stimulus
ER-nucleus signaling pathway
cell-substrate adhesion
cartilage condensation
response to arsenic-containing substance
PAAD_QCMG_SBS8
BLCA_PCAWG_SBS8
PAAD_TCGA_SBS8
LUSC_PCAWG_SBS8
LUSC_CPTAC_SBS8
LUSC_TCGA_SBS8

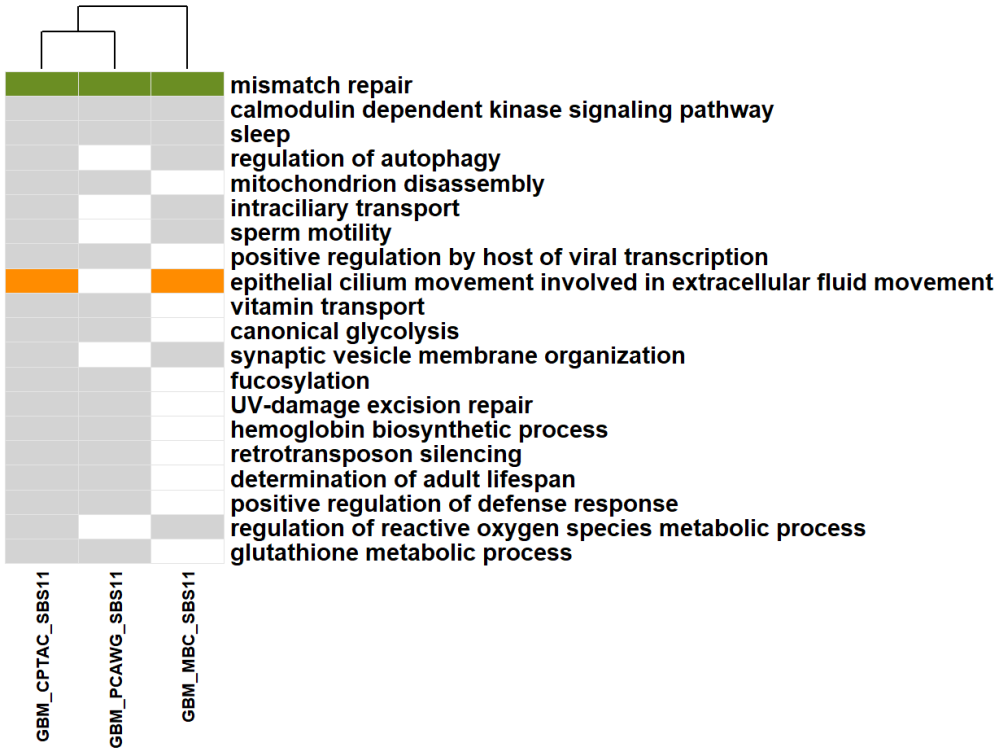
Male

SBS10a/10b



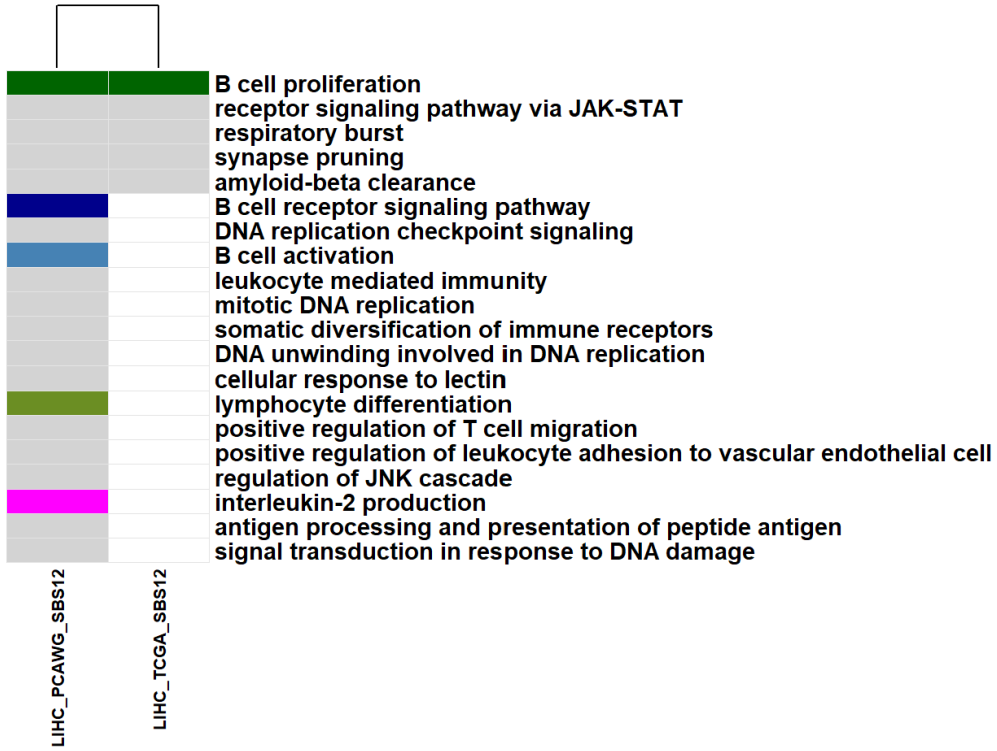
Male

SBS11


mismatch repair
calmodulin dependent kinase signaling pathway
sleep
regulation of autophagy
mitochondrion disassembly
intraciliary transport
sperm motility
positive regulation by host of viral transcription
epithelial cilium movement involved in extracellular fluid movement
vitamin transport
canonical glycolysis
synaptic vesicle membrane organization
fucosylation
UV-damage excision repair
hemoglobin biosynthetic process
retrotransposon silencing
determination of adult lifespan
positive regulation of defense response
regulation of reactive oxygen species metabolic process
glutathione metabolic process
GBM_CPTAC_SBS11
GBM_PCAWG_SBS11
GBM_MBC_SBS11

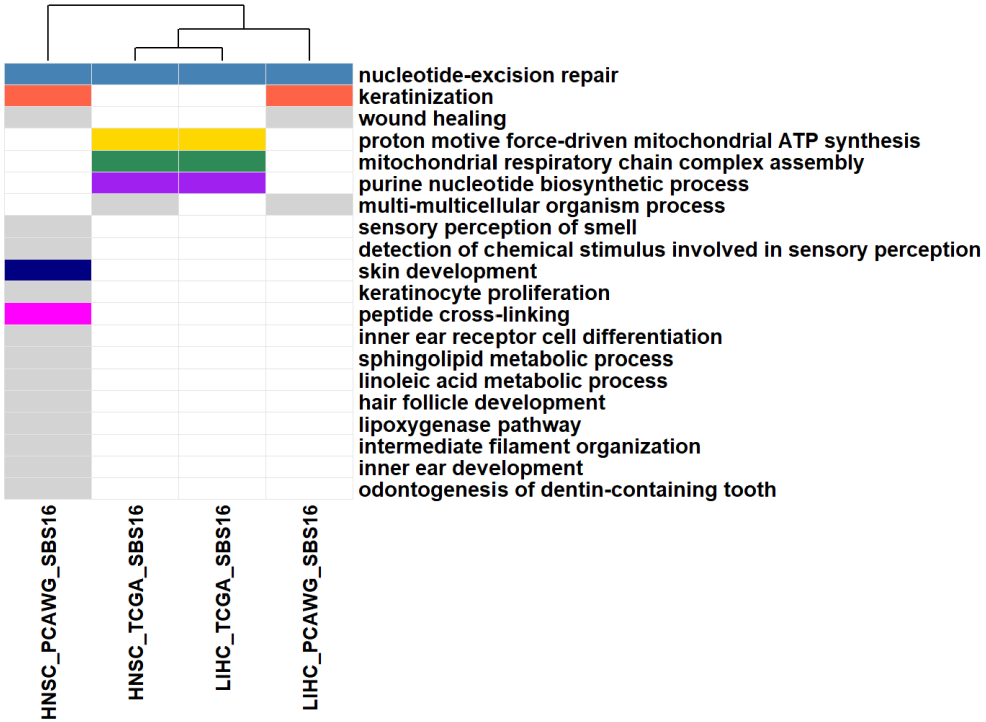
Male

SBS12


B cell proliferation
receptor signaling pathway via JAK-STAT
respiratory burst
synapse pruning
amyloid-beta clearance
B cell receptor signaling pathway
DNA replication checkpoint signaling
B cell activation
leukocyte mediated immunity
mitotic DNA replication
somatic diversification of immune receptors
DNA unwinding involved in DNA replication
cellular response to lectin
lymphocyte differentiation
positive regulation of T cell migration
positive regulation of leukocyte adhesion to vascular endothelial cell
regulation of JNK cascade
interleukin-2 production
antigen processing and presentation of peptide antigen
signal transduction in response to DNA damage
LIHC_PCAWG_SBS12
LIHC_TCGA_SBS12

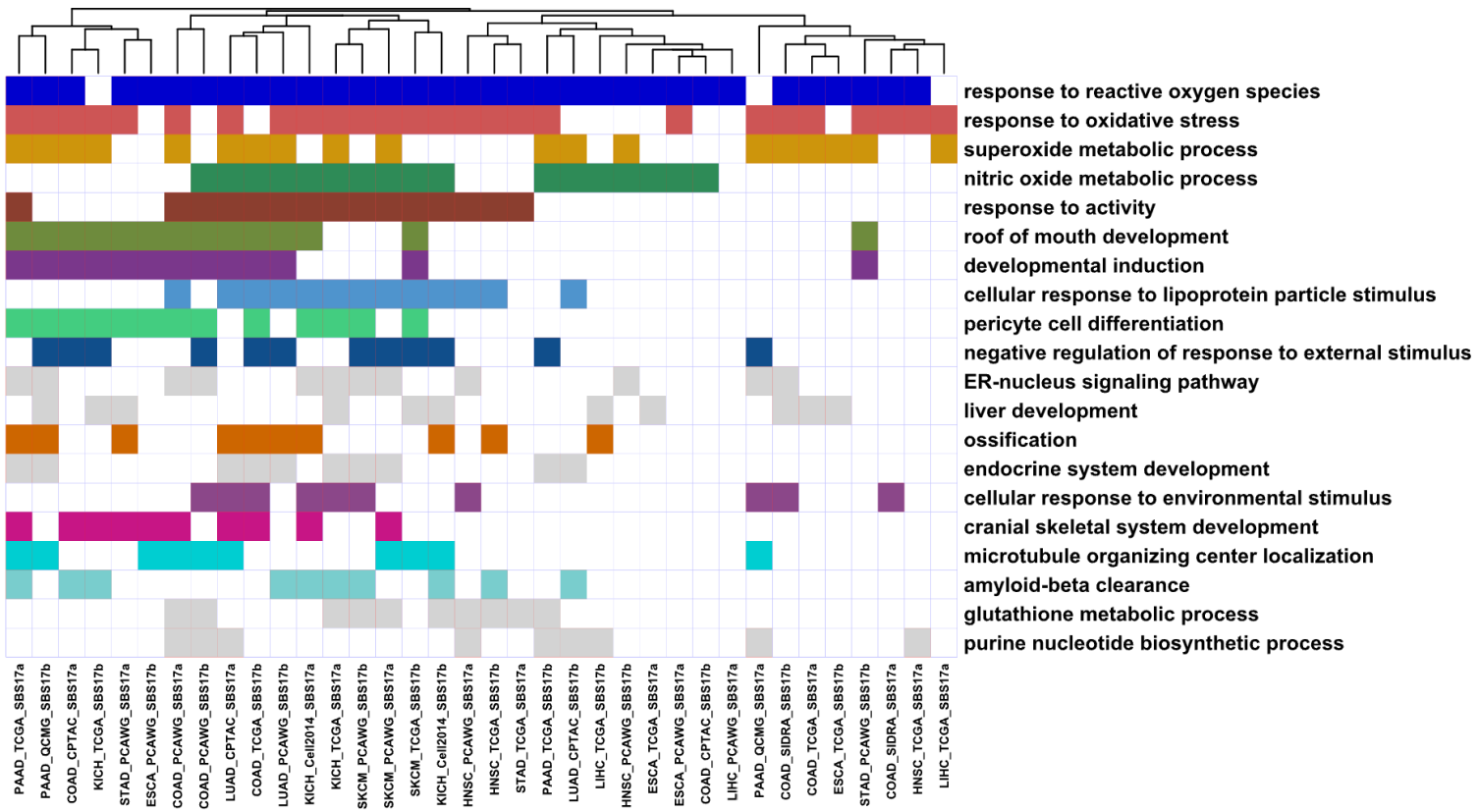
Male

SBS16


nucleotide-excision repair
keratinization
wound healing
proton motive force-driven mitochondrial ATP synthesis
mitochondrial respiratory chain complex assembly
purine nucleotide biosynthetic process
multi-multicellular organism process
sensory perception of smell
detection of chemical stimulus involved in sensory perception
skin development
keratinocyte proliferation
peptide cross-linking
inner ear receptor cell differentiation
sphingolipid metabolic process
linoleic acid metabolic process
hair follicle development
lipoxygenase pathway
intermediate filament organization
inner ear development
odontogenesis of dentin-containing tooth
HNSC_PCAWG_SBS16
HNSC_TCGA_SBS16
LIHC_TCGA_SBS16
LIHC_PCAWG_SBS16

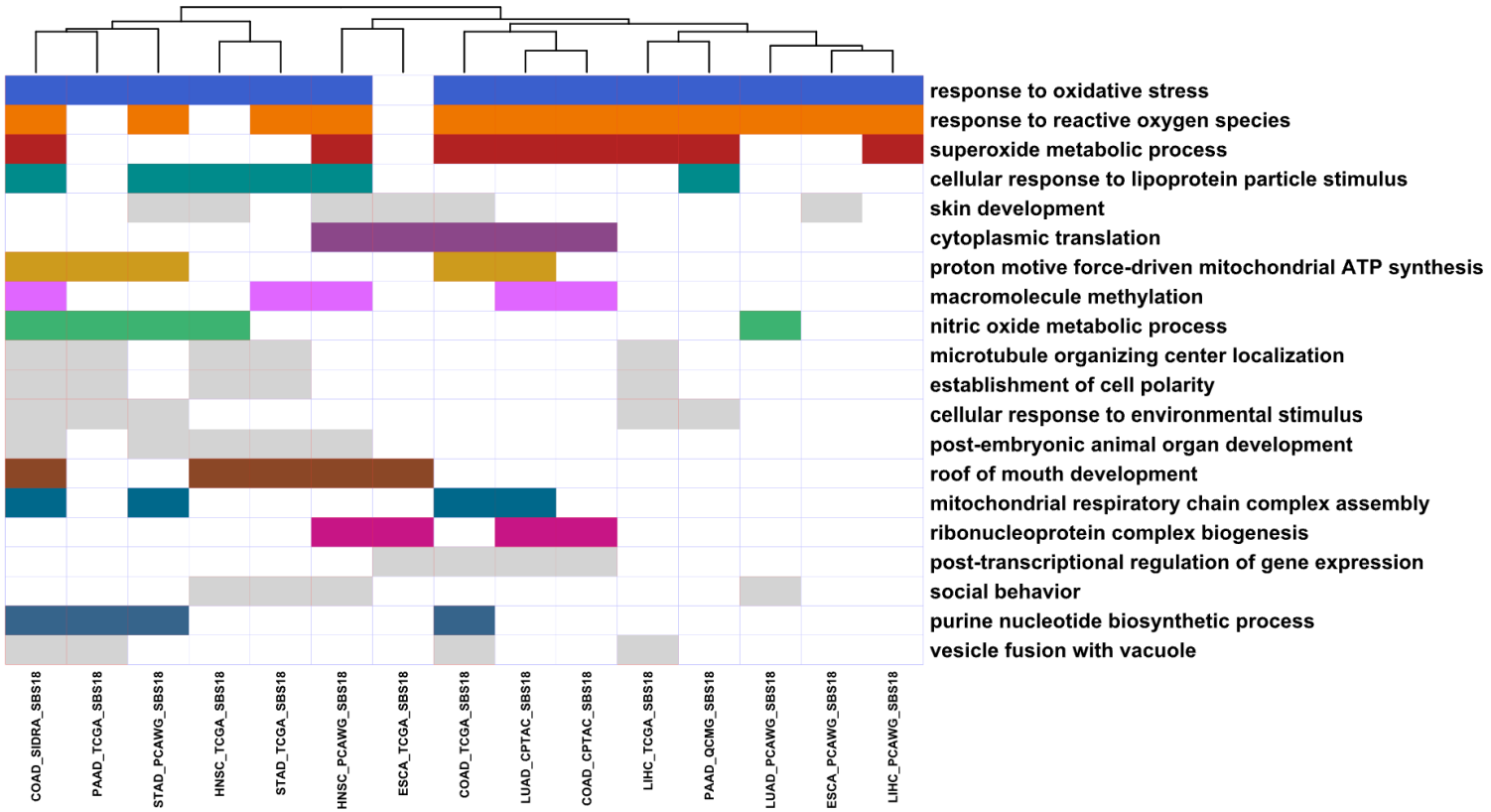
Male

SBS17a/17b



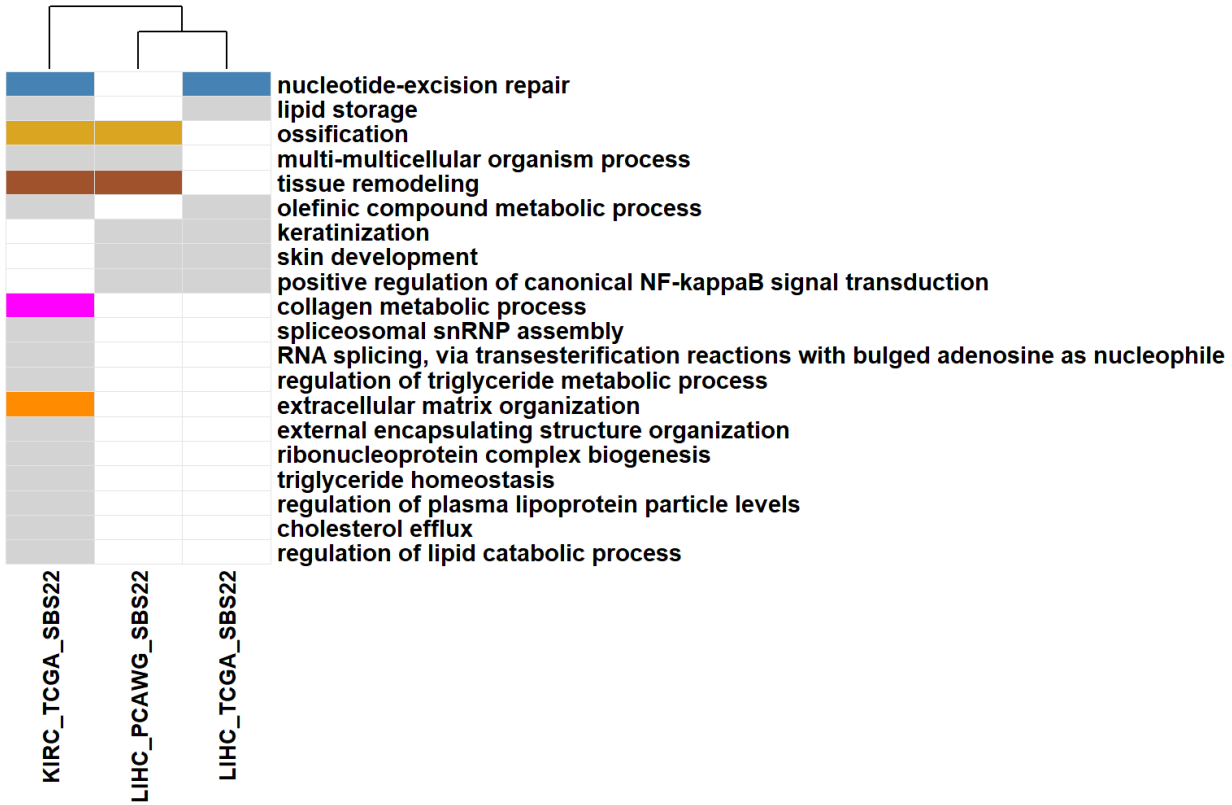
Male

SBS18


response to oxidative stress
response to reactive oxygen species
superoxide metabolic process
cellular response to lipoprotein particle stimulus
skin development
cytoplasmic translation
proton motive force-driven mitochondrial ATP synthesis
macromolecule methylation
nitric oxide metabolic process
microtubule organizing center localization
establishment of cell polarity
cellular response to environmental stimulus
post-embryonic animal organ development
roof of mouth development
mitochondrial respiratory chain complex assembly
ribonucleoprotein complex biogenesis
post-transcriptional regulation of gene expression
social behavior
purine nucleotide biosynthetic process
vesicle fusion with vacuole
COAD_SIDRA_SBS18
PAAD_TCGA_SBS18
STAD_PCAWG_SBS18
HNSC_TCGA_SBS18
STAD_TCGA_SBS18
HNSC_PCAWG_SBS18
ESCA_TCGA_SBS18
COAD_TCGA_SBS18
LUAD_CPTAC_SBS18
COAD_CPTAC_SBS18
LIHC_TCGA_SBS18
PAAD_QCMG_SBS18
LUAD_PCAWG_SBS18
ESCA_PCAWG_SBS18
LIHC_PCAWG_SBS18

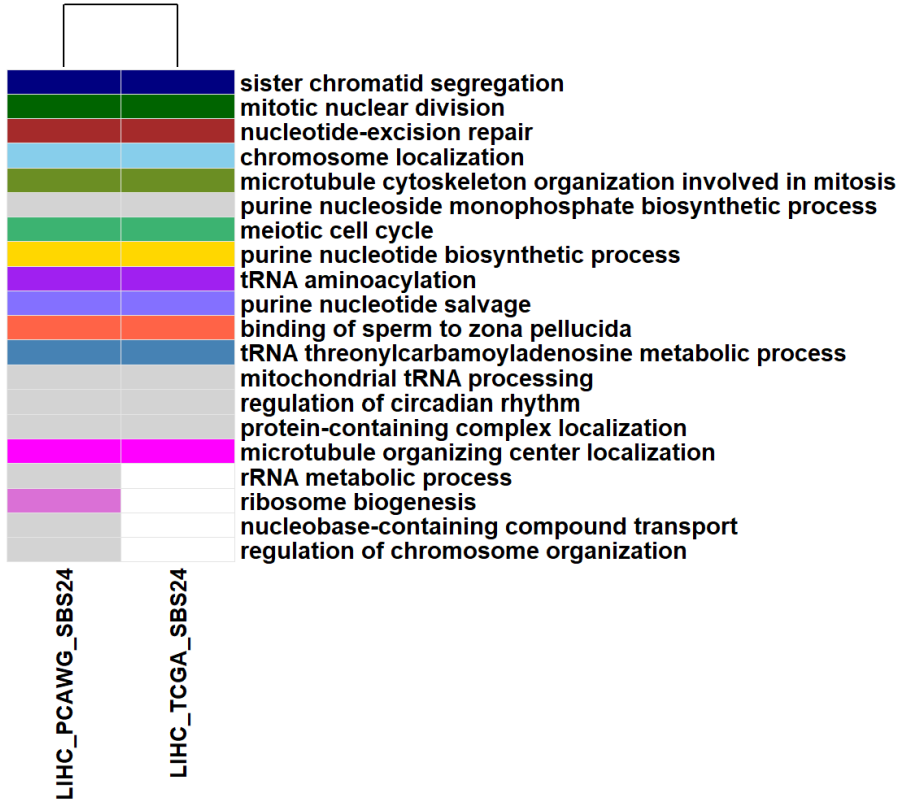
Male

SBS22


nucleotide-excision repair
lipid storage
ossification
multi-multicellular organism process
tissue remodeling
olefinic compound metabolic process
keratinization
skin development
positive regulation of canonical NF-kappaB signal transduction
collagen metabolic process
spliceosomal snRNP assembly
RNA splicing, via transesterification reactions with bulged adenosine as nucleophile
regulation of triglyceride metabolic process
extracellular matrix organization
external encapsulating structure organization
ribonucleoprotein complex biogenesis
triglyceride homeostasis
regulation of plasma lipoprotein particle levels
cholesterol efflux
regulation of lipid catabolic process
KIRC_TCGA_SBS22
LIHC_PCAWG_SBS22
LIHC_TCGA_SBS22

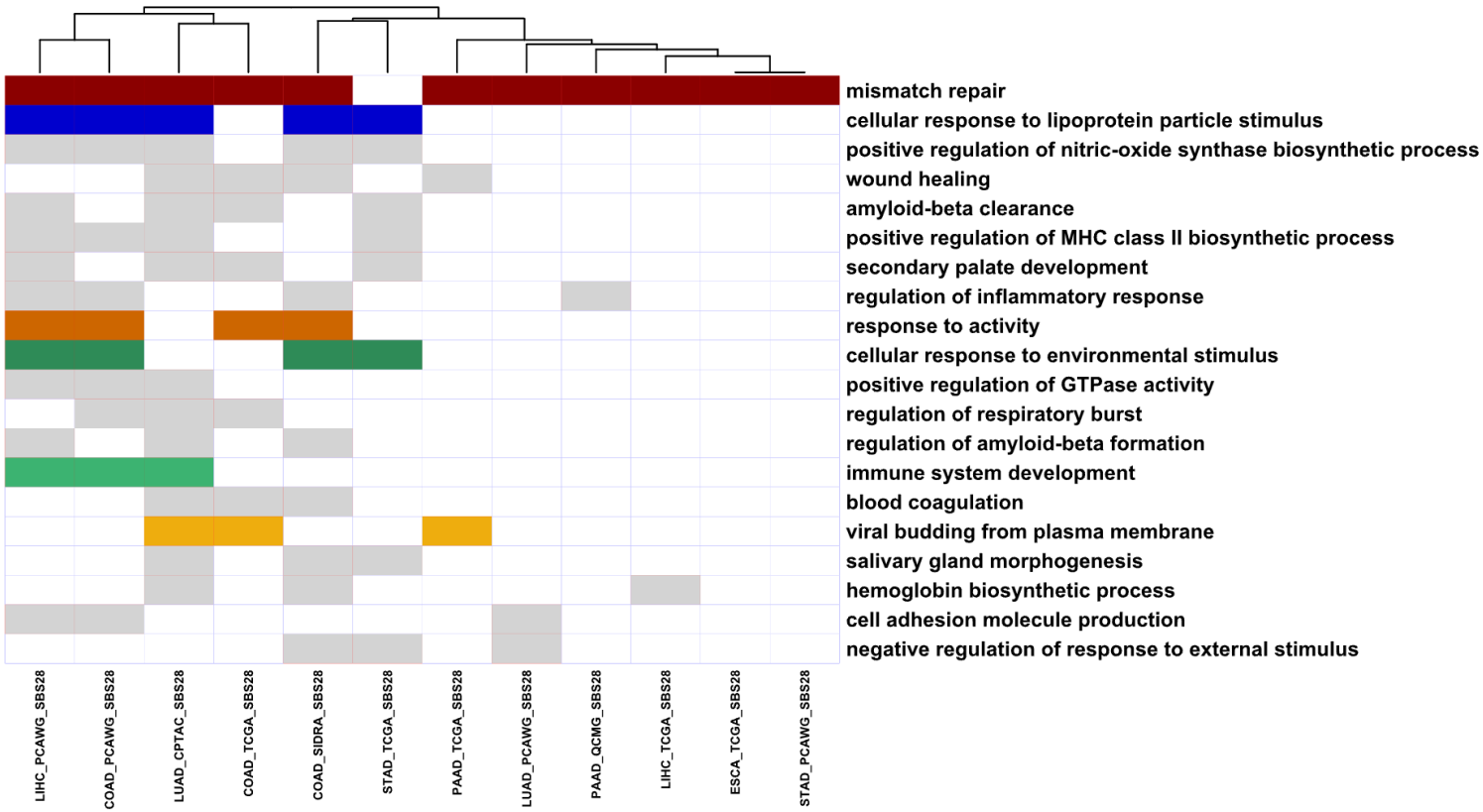
Male

SBS24


sister chromatid segregation
mitotic nuclear division
nucleotide-excision repair
chromosome localization
microtubule cytoskeleton organization involved in mitosis
purine nucleoside monophosphate biosynthetic process
meiotic cell cycle
purine nucleotide biosynthetic process
tRNA aminoacylation
purine nucleotide salvage
binding of sperm to zona pellucida
tRNA threonylcarbamoyladenosine metabolic process
mitochondrial tRNA processing
regulation of circadian rhythm
protein-containing complex localization
microtubule organizing center localization
rRNA metabolic process
ribosome biogenesis
nucleobase-containing compound transport
regulation of chromosome organization
LIHC_PCAWG_SBS24
LIHC_TCGA_SBS24

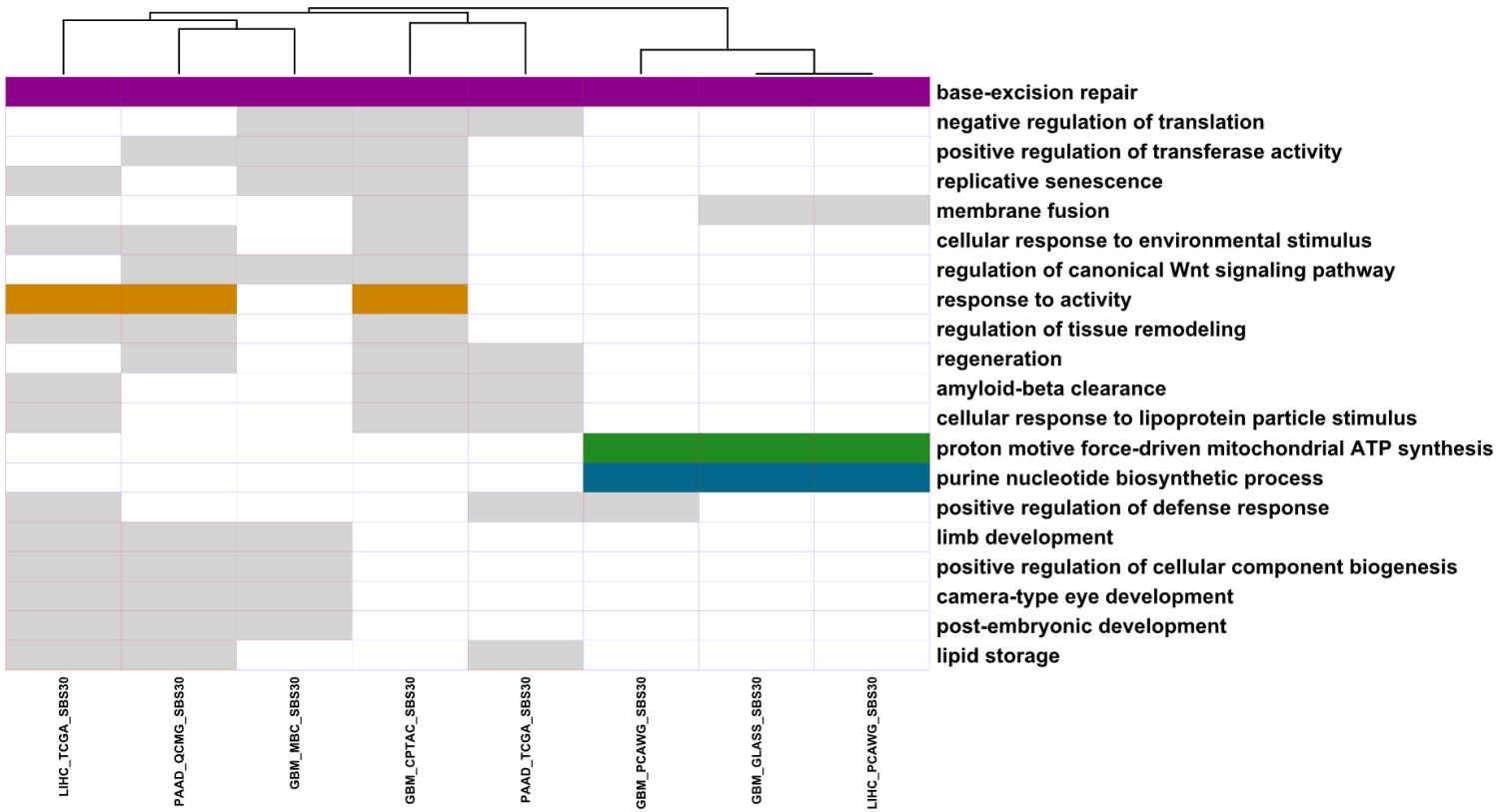
Male

SBS28


mismatch repair
cellular response to lipoprotein particle stimulus
positive regulation of nitric-oxide synthase biosynthetic process
wound healing
amyloid-beta clearance
positive regulation of MHC class II biosynthetic process
secondary palate development
regulation of inflammatory response
response to activity
cellular response to environmental stimulus
positive regulation of GTPase activity
regulation of respiratory burst
regulation of amyloid-beta formation
immune system development
blood coagulation
viral budding from plasma membrane
salivary gland morphogenesis
hemoglobin biosynthetic process
cell adhesion molecule production
negative regulation of response to external stimulus
LIHC_PCAWG_SBS28
COAD_PCAWG_SBS28
LUAD_CPTAC_SBS28
COAD_TCGA_SBS28
COAD_SIDRA_SBS28
STAD_TCGA_SBS28
PAAD_TCGA_SBS28
LUAD_PCAWG_SBS28
PAAD_QCMG_SBS28
LIHC_TCGA_SBS28
ESCA_TCGA_SBS28
STAD_PCAWG_SBS28

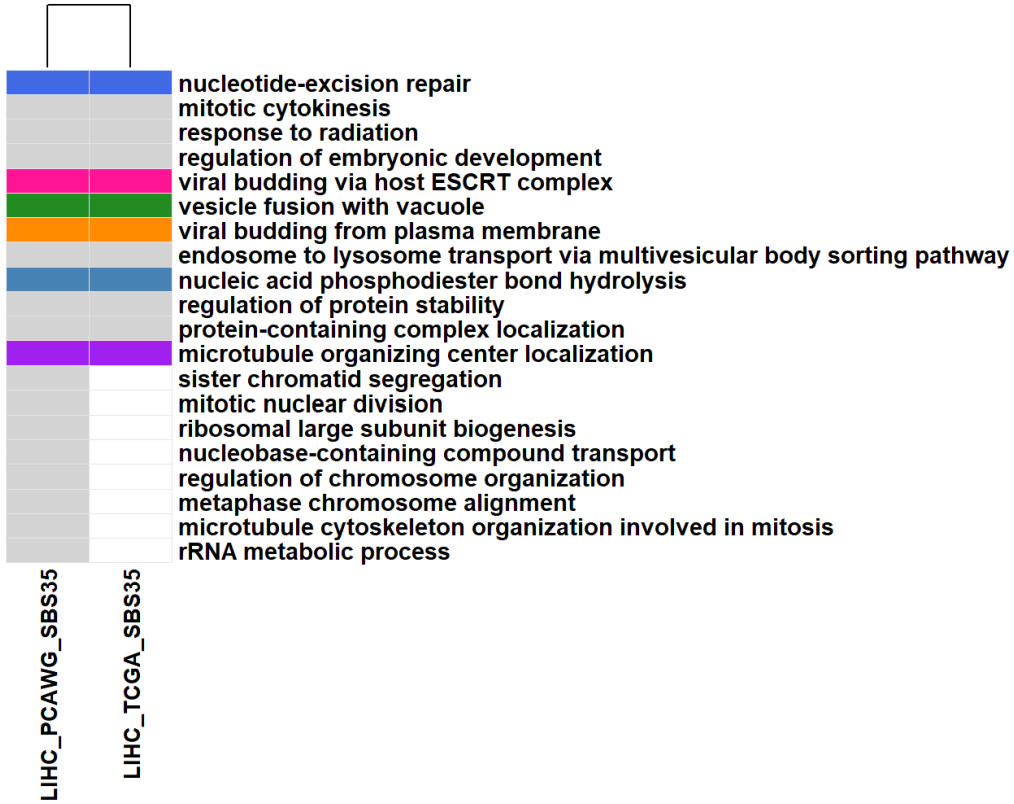
Male

SBS30


base-excision repair
negative regulation of translation
positive regulation of transferase activity
replicative senescence
membrane fusion
cellular response to environmental stimulus
regulation of canonical Wnt signaling pathway
response to activity
regulation of tissue remodeling
regeneration
amyloid-beta clearance
cellular response to lipoprotein particle stimulus
proton motive force-driven mitochondrial ATP synthesis
purine nucleotide biosynthetic process
positive regulation of defense response
limb development
positive regulation of cellular component biogenesis
camera-type eye development
post-embryonic development
lipid storage
LIHC_TCGA_SBS30
PAAD_QCMG_SBS30
GBM_MBC_SBS30
GBM_CPTAC_SBS30
PAAD_TCGA_SBS30
GBM_PCAWG_SBS30
GBM_GLASS_SBS30
LIHC_PCAWG_SBS30

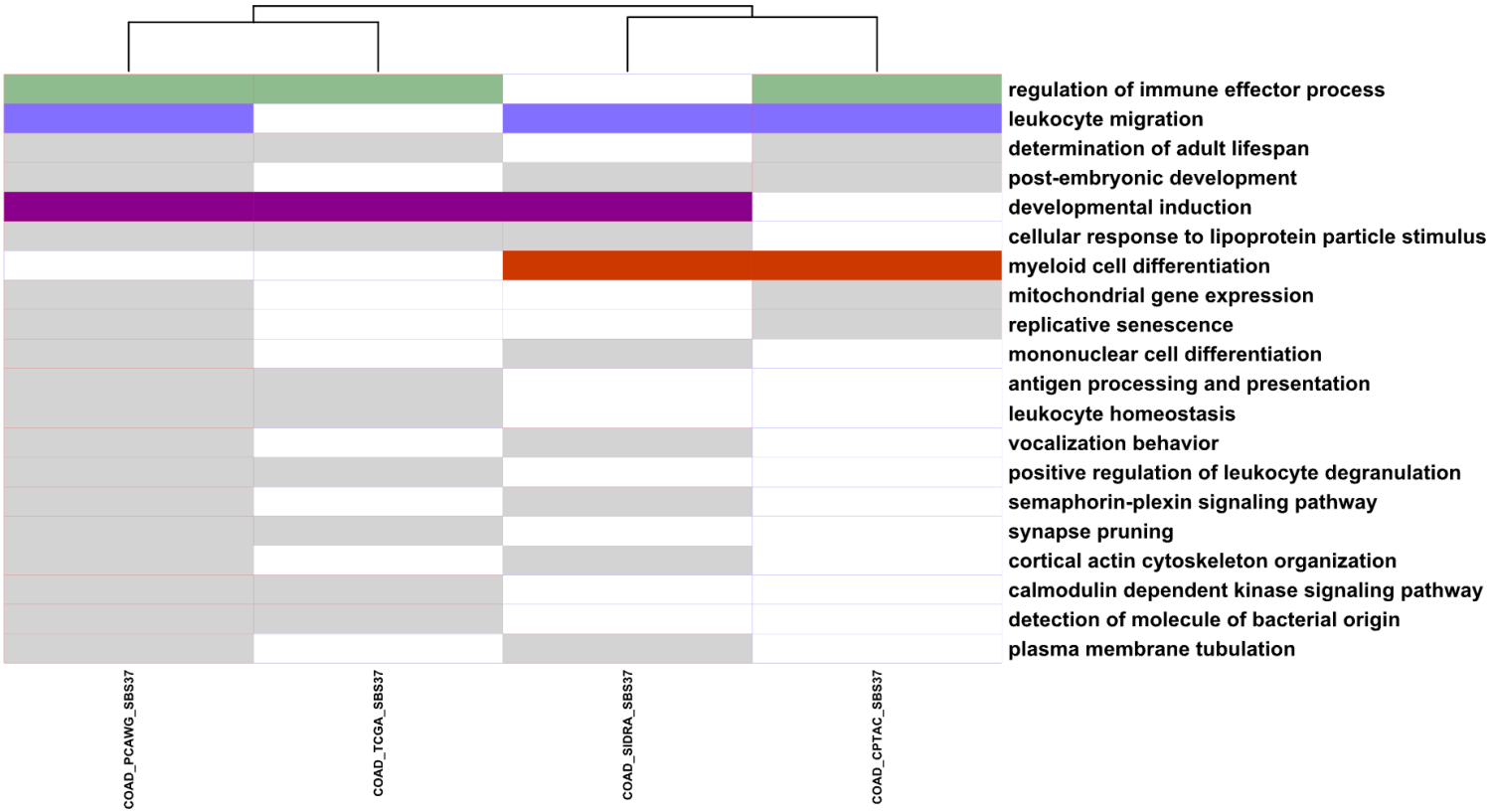
Male

SBS35


nucleotide-excision repair
mitotic cytokinesis
response to radiation
regulation of embryonic development
viral budding via host ESCRT complex
vesicle fusion with vacuole
viral budding from plasma membrane
endosome to lysosome transport via multivesicular body sorting pathway
nucleic acid phosphodiester bond hydrolysis
regulation of protein stability
protein-containing complex localization
microtubule organizing center localization
sister chromatid segregation
mitotic nuclear division
ribosomal large subunit biogenesis
nucleobase-containing compound transport
regulation of chromosome organization
metaphase chromosome alignment
microtubule cytoskeleton organization involved in mitosis
rRNA metabolic process
LIHC_PCAWG_SBS35
LIHC_TCGA_SBS35

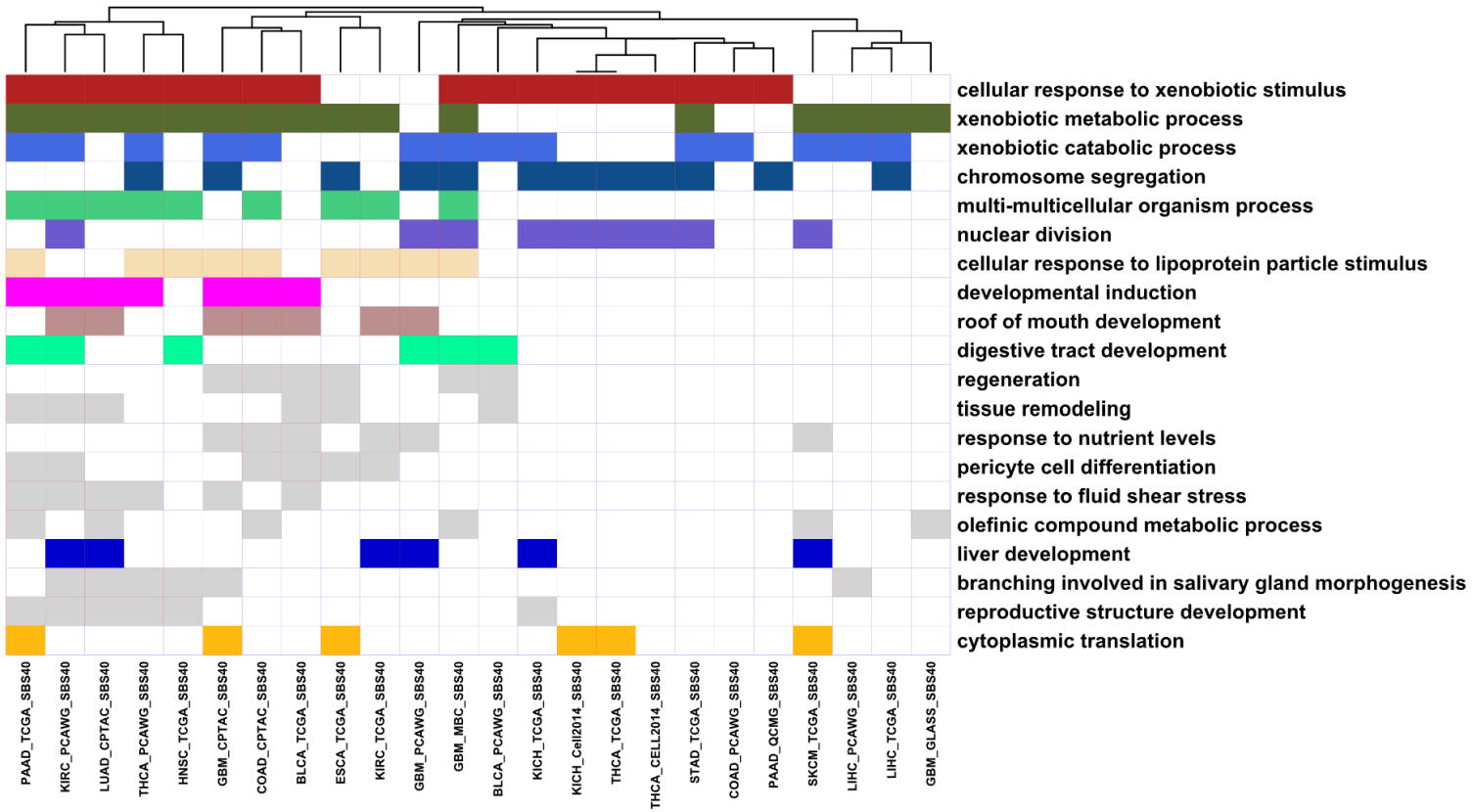
Male

SBS37


regulation of immune effector process
leukocyte migration
determination of adult lifespan
post-embryonic development
developmental induction
cellular response to lipoprotein particle stimulus
myeloid cell differentiation
mitochondrial gene expression
replicative senescence
mononuclear cell differentiation
antigen processing and presentation
leukocyte homeostasis
vocalization behavior
positive regulation of leukocyte degranulation
semaphorin-plexin signaling pathway
synapse pruning
cortical actin cytoskeleton organization
calmodulin dependent kinase signaling pathway
detection of molecule of bacterial origin
plasma membrane tubulation
COAD_PCAWG_SBS37
COAD_TCGA_SBS37
COAD_SIDRA_SBS37
COAD_CPTAC_SBS37

Male

SBS40


cellular response to xenobiotic stimulus
xenobiotic metabolic process
xenobiotic catabolic process
chromosome segregation
multi-multicellular organism process
nuclear division
cellular response to lipoprotein particle stimulus
developmental induction
roof of mouth development
digestive tract development
regeneration
tissue remodeling
response to nutrient levels
pericyte cell differentiation
response to fluid shear stress
olefinic compound metabolic process
liver development
branching involved in salivary gland morphogenesis
reproductive structure development
cytoplasmic translation
PAAD_TCGA_SBS40
KIRC_PCAWG_SBS40
LUAD_CPTAC_SBS40
THCA_PCAWG_SBS40
HNSC_TCGA_SBS40
GBM_CPTAC_SBS40
COAD_CPTAC_SBS40
BLCA_TCGA_SBS40
ESCA_TCGA_SBS40
KIRC_TCGA_SBS40
GBM_PCAWG_SBS40
GBM_MBC_SBS40
BLCA_PCAWG_SBS40
KICH_TCGA_SBS40
KICH_Cell2014_SBS40
THCA_TCGA_SBS40
THCA_CELL2014_SBS40
STAD_TCGA_SBS40
COAD_PCAWG_SBS40
PAAD_QCMG_SBS40
SKCM_TCGA_SBS40
LIHC_PCAWG_SBS40
LIHC_TCGA_SBS40
GBM_GLASS_SBS40