The Influence of *MUTYH* on Repair Deficiencies on Germline and Somatic Mutagenesis Across Mammalian Species

Candice L. Young

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2024

Reading Committee:

Kelley Harris, Chair

Brian Shirts

Rosana Risques

Program Authorized to Offer Degree:

Molecular & Cellular Biology

University of Washington

Abstract

The Influence of *MUTYH* on Germline and Somatic Mutagenesis Across Mammalian Species

Candice L. Young

Chair of the Supervisory Committee:

Kelley Harris

Department of Genome Sciences

The Base Excision Repair (BER) gene *MUTYH* is essential for maintaining genomic integrity by preventing C>A transversions induced by oxidative stress. Prior research has established that mutations in *MUTYH* are associated with increased cancer susceptibility in humans as well as elevated germline *mutation* rates in mice. However, the evolutionary dynamics of germline mutator alleles like MUTYH, and the extent to which their functional variations are permitted or constrained across different species and cellular contexts, remain poorly understood. Specifically, it remains unclear how *MUTYH* variants influence mutagenesis in the human germline compared to their well-established role in human somatic cells, and how they affect mutagenesis in mouse somatic cells relative to their known impact on mouse germline mutation rates. To explore this, we examined *MUTYH* alleles in a human pedigree, identifying an association between a pathogenic *MUTYH* genotype and elevated C>A de novo mutations in the

maternal germline. In parallel, we assessed murine *Mutyh* variation by sequencing colons and spleens from aged recombinantly inbred "BXD" mice, revealing aged-related increases in C>A somatic mutations in *Mutyh*-deficient strains as well as *Mutyh* associated single base substitution COSMIC cancer signatures in both tissues. Together, these findings advance our understanding of *MUTYH's* role in cancer, germline mutagenesis, and sex-specific mutation patterns across mammalian species.

Introduction

Many DNA repair deficiencies are linked with increased risk for cancer syndromes (Fearon 1997; Goode et al. 2002; Matullo et al. 2006; Randall et al. 2023). Pathogenic mutations leading to the loss of function in specific DNA repair pathways accelerate the accumulation of oncogenic variants. While each DNA repair defect often tends to cause cancers mainly in specific tissues, other tissues may also accumulate a higher mutation load than normal (Scarbrough et al. 2016; Dunlop et al. 1997; Aarnio et al. 1999). It is not well understood why accelerated mutagenesis only seems to lead to cancer in certain tissues, or whether somatic mutations that do not cause cancer might have other health impacts (Blokzijl et al. 2016; Elledge and Amon 2002; Chao and Lipkin 2006).

To address these uncertainties, researchers have increasingly turned their attention to the germline, where DNA repair deficiencies may have distinct impacts compared to somatic cells. Recent studies (Sherwood et al. 2023; Andrianova et al. 2023; Stendahl et al. 2023; Kaplanis et al. 2022) have concentrated on the impact of DNA repair deficiencies on the germline, recognizing that even modestly elevated germline mutation rates can influence congenital disease risk and the rate of evolution. Moreover, since germline mutations can be studied through relatively straightforward comparisons between relatives (Wei et al. 2015; Bergeron et al. 2022) and do not require the specialized technologies that are needed to detect low-frequency somatic variants (Kennedy et al. 2014; Ellis et al. 2021), germline mutator phenotypes have the potential to lead to discovery of new DNA repair defects that may be candidate drivers of novel cancer syndromes. For example, inherited variation was recently used to discover that a variant in the

murine Base Excision Repair (BER) DNA-glycosylase *Mutyh* gene acts as a germline mutator allele in inbred mouse strains (Sasani et al. 2022, 2024). Since impaired functioning of the human *MUTYH* protein is known to cause a colorectal cancer syndrome known as *MUTYH*-associated polyposis (MAP) (Smith et al. 2013a), this mutator allele is a promising candidate for exploring joint effects of DNA repair genes on the mammalian soma and germline.

The *MUTYH* gene plays a key role in BER, a DNA repair pathway that evolved to repair damage caused by reactive oxygen species (ROS), which are byproducts of aerobic metabolism (Banda et al. 2017). ROS can react with guanine to create the lesion 8-oxoguanine (8-OG), which has a propensity to mispair with adenine, resulting in G:C > T:A transversion mutations, often abbreviated as C > A mutations (David et al. 2007). BER DNA glycosylases have developed a specific mechanism to repair this mutagenic damage: OGG1 removes 8-OG from the compromised strand (Hayashi et al. 2002) while *MUTYH* excises the erroneously incorporated adenines opposite 8-OG (Woods et al. 2016; Krokan and Bjørås 2013). Due to *MUTYH's* role in this repair pathway, defects in this enzyme can cause excess accumulation of C>A mutations in tissues that are experiencing ROS damage (Pilati et al. 2017).

MAP follows a recessive inheritance pattern, affecting individuals who have inherited two sub-functional copies of the *MUTYH* gene (Morak et al. 2014). This condition is associated with intestinal adenomatous polyposis as well as an elevated risk for early-onset colorectal and duodenal malignancies (Nielsen et al. 2011; Al-Tassan et al. 2002). Individuals with "biallelic" *MUTYH* genotypes—either homozygous for a single pathogenic variant or compound heterozygotes with each allele carrying a different pathogenic mutation—are susceptible to

MAP. In contrast, individuals who have "monoallelic" genotypes (heterozygous for a single pathogenic *MUTYH* variant) generally do not develop intestinal polyposis and have at most a modestly elevated cancer risk (Barreiro et al. 2022). Notably, *MUTYH* is an example of a gene that plays a crucial role in genomic stability across all tissues affected by ROS damage, but mainly appears to modulate cancer risk in the colorectal epithelium (Nieuwenhuis et al. 2012; Hutchcraft et al. 2021). Despite the tissue specificity of MAP's cancer risk phenotype, recent evidence indicates that this condition also causes elevated somatic mutation rates in a wider variety of human cell types, including blood (Robinson et al. 2022), which might be why some studies have found *MUTYH* variants to be associated with increased risk of extracolonic cancers (Vogt et al. 2009; Win et al. 2016; Zhang et al. 2006; Beiner et al. 2009; Villy et al. 2022). These findings led us to hypothesize that *MUTYH*'s C>A mutator effect might extend to germline cells.

This dissertation examines the impact of DNA repair deficiencies on mutational patterns across the mammalian germline and somatic cells, presenting its findings in two main chapters. **Chapter 1** focuses on the role of *MUTYH* in human germline mutagenesis. We assessed whether pathogenic *MUTYH* genotypes induce a human germline mutator phenotype by sequencing fifteen genomes from an extended family affected by *MUTYH*-associated polyposis (MAP). Analysis of parent-child trios, where a parent has MAP, revealed elevated C>A mutation rates compared to trios with normal or monoallelic *MUTYH* genotypes. Additionally, our study builds upon previous findings reporting a C>A mutator effect in de novo mutations from trios with MAP-affected mothers (Sherwood et al. 2023). By analyzing a larger dataset that includes children of both mothers and fathers with pathogenic *MUTYH* genotypes and comparing results

to a null model based on parental age (Jónsson et al. 2017), we provide a comprehensive characterization of how pathogenic *MUTYH* variants affect the human germline.

While human studies provide valuable insights into the impact of DNA repair deficiencies on germline mutagenesis, complementary animal models are essential for mechanistic exploration and validation of observed mutational patterns. In this context, prior research by Sasani et al. (2022) identified a natural mutator allele in BXD mice that significantly influences germline mutation rates and spectra. Specifically, mice inheriting the "D" haplotype at a quantitative trait locus on chromosome 4 exhibited a 50% higher rate of C>A germline mutations compared to those with the "B" haplotype, primarily driven by the mutational signature SBS18. Additionally, an epistatic interaction between *Mutyh* alleles and a locus overlapping *Ogg1* on chromosome 6 further contributed to increased C>A mutations. These findings underscore the utility of BXD mouse models in dissecting the genetic factors underlying *MUTYH's* mutagenic effects and support the translational relevance of these models to human cancer syndromes.

Building on these insights, **Chapter 1** integrates human and mouse data to provide a comprehensive characterization of *MUTYH's* role in germline mutagenesis. By sequencing fifteen genomes from a family affected by MAP and analyzing parent-child trios, we assessed the induction of a germline mutator phenotype by pathogenic *MUTYH* genotypes. Our findings revealed elevated C>A mutation rates in trios with a MAP-affected parent, aligning with previous studies (Sherwood et al. 2023). Furthermore, by reanalyzing mutation rates in recombinant inbred BXD mouse strains carrying distinct *Mutyh* alleles, we demonstrated that *Mutyh* variation does not amplify the paternal age effect on C>A mutations. Instead, the

increased C>A mutation rate is primarily driven by early embryonic mutations rather than gamete aging. This comparative approach bridges human and murine studies, establishing a framework for investigating the effects of DNA repair deficiencies on germline mutagenesis and their broader implications for cancer syndromes and human evolution.

Having established the role of *MUTYH* in germline mutagenesis, **Chapter 2** extends this investigation to somatic tissues. This chapter examines how mutator alleles shape somatic mutation landscapes in murine models, focusing on the influence of repair deficiencies in non-cancerous tissues. By utilizing recombinant inbred BXD mouse strains, **Chapter 2** delves into the genetic modifiers of mutation rates, providing a complementary perspective to the germline-focused analyses of **Chapter 1**. Building on Sasani et al.'s (2022, 2024) findings in BXD mice, we examined whether the identified germline mutator alleles also affect somatic mutations in non-cancerous tissues, particularly within the context of aging. The mouse samples used in this study were sourced from recent BXD longevity studies (Roy et al., 2021), where colon samples were aged until natural death, and spleen samples were collected from mice maintained on either normal chow or high-fat diets. This approach allows us to analyze the interplay between aging, diet, genotype, and mutation accumulation within a controlled genetic background (Roy et al., 2021).

Through the analysis of somatic mutation burdens in aged BXD mice, we reveal how *Mutyh* genotype influences mutational processes in non-cancerous tissues, highlighting broader implications of DNA repair deficiencies beyond the germline. We observed a significant increase in C>A and C>T mutations specifically in colon tissues of aged D haplotype carriers.

Additionally, distinct mutational signatures associated with aging and oxidative damage were identified, underscoring the multifaceted role of *MUTYH* in maintaining genomic stability.

To contextualize these findings, we compared mutational signatures from spleen and colon tissues to published datasets from aged murine models (Cagan et al., 2022; Riva et al., 2020; Chin et al., 2021). Our analysis revealed that while some age-associated signatures overlapped, strain-specific variation played a dominant role in shaping mutation landscapes. Furthermore, we investigated clonal expansions in cancer-associated homolog genes such as *Trp53*, *Pik3ca*, and *Ctnnb1* in spleen tissues, finding frequent and expansive clones in both B and D allele carriers. Modeling mutation rates as a function of age and replication timing revealed replication-timing-dependent patterns and consistent age-related trends in substitution classes like C>A and C>T mutations, with D allele carriers exhibiting a higher mutation burden in colon tissues across all age groups.

These findings in **Chapter 2** provide insight into how *Mutyh* genotype influences mutational processes in non-cancerous tissues, based on data from two tissues in two mouse strains. Importantly, this work complements our investigations in **Chapter 1**, where we demonstrated that pathogenic *MUTYH* variants in humans lead to elevated germline mutation rates. Collectively, **Chapters 1** and **2** elucidate the dual impact of *MUTYH*-related DNA repair deficiencies on both germline and somatic mutagenesis. This consolidated approach underscores the influence of genetic variation on mutation accumulation across different biological contexts.

By integrating data from human familial studies and the well-characterized BXD mouse family, we underscore the significance of genetic background and age in shaping mutation landscapes across mammalian species. This comprehensive approach offers new insights into how intrinsic genetic factors contribute to genomic instability, disease risk, and evolutionary dynamics. Ultimately, this dissertation addresses the broader influence of DNA repair deficiencies on mutational processes in both germline and somatic tissues, with significant implications for cancer risk assessment and understanding the mechanisms underlying age-related diseases and cancer development. By probing the role of *MUTYH* mutations in compromising DNA integrity, we aim to deepen our understanding of how mutations in a putative mutator allele contributes to genomic instability across different biological contexts.

Chapter One:

A maternal germline mutator phenotype in a family affected by heritable colorectal cancer

Adapted from: “A maternal germline mutator phenotype in a family affected by heritable colorectal cancer”

Published in *Genetics* 2024 (iyae166) with co-lead Annabel Beichman as well as David Mas-Ponte, Shelby L Hemker, Luke Zhu, Jacob O Kitzman, Brian H Shirts, and Kelley Harris

Results

Whole genome sequencing of a family affected by a MUTYH genotype that shows reduced DNA repair efficiency in vitro

To investigate whether *MUTYH* genotype influences the accumulation of germline mutations in humans, we sequenced several parent-child trios from an extended family affected by a pathogenic *MUTYH* genotype associated with *MUTYH*-associated polyposis (MAP). This family provides a unique opportunity to study how specific *MUTYH* variants impact germline mutagenesis due to their known history of colorectal cancer and confirmed pathogenic genotypes.

We obtained saliva samples from three full siblings (labeled P1, P2, and P3 in **Figure 1**) who share a biallelic *MUTYH* genotype known as p.Y179C/V234M. (One gene copy has a tyrosine-to-cysteine substitution at amino acid position 179 and the other has a valine-to-methionine substitution at position 234. Both amino acid positions are indexed in the coordinates of *MUTYH* transcript NM_001128425; substitutions p.Y179C and p.V234M have DNA coordinates c.536A>G and c.700G>A, respectively.) We also sampled saliva from a fourth sibling (P4 in **Figure 1**) who is a monoallelic carrier of p.V234M. These siblings inherited p.Y179C from their father and inherited p.V234M from their mother. Two of the three biallelic siblings were previously diagnosed with colon cancer, and the third has a history of colon polyps (all three meet the diagnostic criterion for MAP by virtue of their *MUTYH* genotype). The siblings ' extended family is affected by a notably elevated level of colorectal cancer, including unsampled

relatives who are also pictured in **Figure 1**. All sampled individuals tested negative for Lynch syndrome variants and other variants known to cause heritable colorectal cancer.

While ClinVar classifies p.Y179C as pathogenic with evidence from many previous studies (Al-Tassan et al. 2002; Nielsen et al. 2005, 2009; Vogt et al. 2009), some laboratories consider p.V234M to be a variant of uncertain significance with mixed functional evidence (Peterlongo et al. 2006; Fleischmann et al. 2004; Yurgelun et al. 2015; Komine et al. 2015). To obtain more information about the pathogenicity of this genotype p.Y179C/V234M, we conducted functional assays in which mutant MUTYH expression is restored in human HEK293 *MUTYH* KO cells. Our approach uses a reporter construct engineered to contain an 8-oxoG:A lesion, such that proper repair corrects a premature stop codon in GFP and restores its expression. Notably, the p.Y179C allele exhibited severe loss of repair function, whereas the p.V234M variant displayed a partial loss of function with repair activity well below that of wild-type MUTYH (**Figure S2**). The deleterious effects observed for these two variants within the HEK293 cell context indicate that they likely have pathogenic effects and may result in elevated mutation accumulation across tissues in vivo.

We also used the same functional assay to measure the effects of c.461GT>AA p.R182Q (corresponding to the DNA substitution c.461 GT>AA on transcript NM_001128425), the human analog of the mutation found in an outlier mouse strain known as BXD68 that displayed a *Mutyh* hypermutator germline phenotype (Sasani et al. 2022). p.R182Q appears to be a total loss of function variant with a phenotype similar to that of p.Y179C (**Figure S2**).

We note that the previous study of *MUTYH's* germline mutator activity by Sherwood et al. (2023) analyzed families with a different compound heterozygous genotype known as p.Y179C/G368D. p.G368D (corresponding to the DNA substitution c. 1187 G>A on transcript NM_001128425) may be less deleterious than p.Y179C given its association with an older age at MAP diagnosis (Guarinos et al. 2014) and its less severe somatic mutator phenotype (Robinson et al. 2022). Note that Robinson, et al. refer to the variant p.G368D as p.G396D in the coordinates of a different reference *MUTYH* transcript.

To assess the impact of the *MUTYH* p.Y179C/V234M genotype on the germline mutation rate and spectrum, we performed 50X coverage whole genome sequencing on P1–P4 as well as nine of their adult children and two of their spouses (**Figure 1**). Individuals have been given labels according to which nuclear family they are a member of (1–4), and whether they are a *MUTYH* variant carrier parent (P), a spouse or partner of that parent (S), or a child of a *MUTYH* variant carrier (C). Colloquially, we refer to all parents as mothers and fathers if they conceived their children via oocytes and spermatocytes, respectively, although we recognize that these labels may not match parents ’individual gender identities.

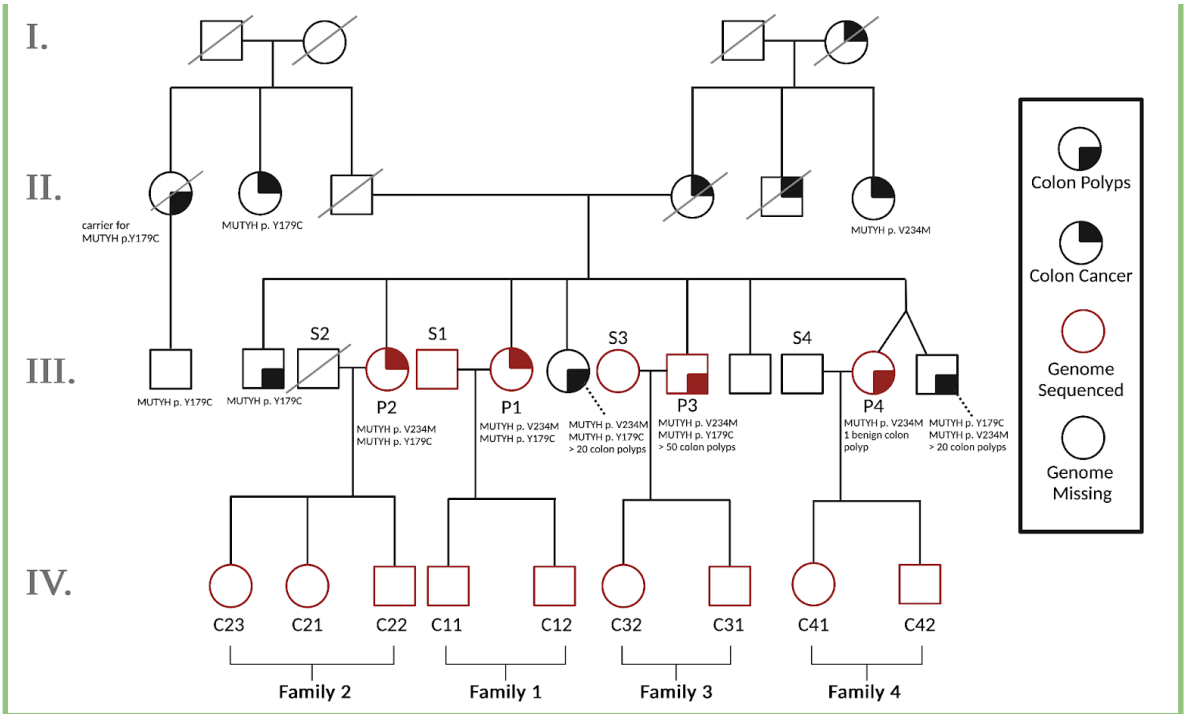


Figure 1. Pedigree of all sequenced individuals plus unsampled relatives. To measure the effects of biallelic *MUTYH* mutations (present in Generation III) on germline mutagenesis, we sequenced all offspring from Generation IV as well as all parents from Generation III from whom samples were available. Individuals have been given labels according to whether they are part of nuclear families 1–4, whether they are a *MUTYH* variant carrier parent (P), a spouse or partner of that parent (S), or a child (C). The sequenced individuals make up four nuclear families whose children are all first cousins: Families 1 and 2 include mothers with the biallelic genotype p.Y179C/V234M, while Family 3 includes a father with the same p.Y179C/V234M genotype. Family 4 includes a mother who is monoallelic for p.V234M. Parents of the Generation III individuals are each a carrier of one of these monoallelic mutations. Solid black quadrants indicate which individuals have been diagnosed with colon polyps (bottom right) or colon cancer (top right). *MUTYH* mutations and number of identified colon polyps are listed below individuals studied in this pedigree. Square = male; circle = female; red = genome sequenced; black = genome missing from sequenced pedigree trio.

Using siblings as "surrogate parents" for de novo mutation calling in incomplete nuclear families

DNMs are typically called by identifying sites that violate the principles of Mendelian inheritance. These are sites at which a child's genome contains a variant not observed in the genome of either of their parents. DNM calling normally requires the genomes of both parents to

be sequenced, and Families 1 and 3 are the only nuclear families in our dataset that meet this requirement (**Figure 1**). Families 2 and 4 were not suitable for standard mutation calling due to unavailability of the paternal genomes (one father is deceased and the other declined to participate). To maximize the power of this study, we developed a novel method that facilitated DNM calling in these families. Although our new method has slightly lower accuracy and precision than standard DNM calling, particularly in Family 4 which consists of just a mother and two children, it enables DNM analysis in previously inaccessible families within our study and potentially beyond.

Since each child in generations III and IV has at least one full sibling represented in our dataset, we were able to leverage the sharing of parental haplotypes among siblings to devise a "surrogate parent" method for estimating DNM rates and spectra. Instead of comparing each child's genome to their mother and father to identify variants that must have arisen de novo, we compared each child from Families 2 and 4 to their mother plus one or more siblings who inherited some of the same paternal DNA. If the child's genome contains a variant that is absent from both the mother and the sibling surrogate father, this implies that the unique variant arose de novo in the child (**Figure 2A-2B;** pipeline described in **Figure S1**). Jónsson, et al. (2018) previously used a similar procedure to identify mutations that arose early in parental embryonic development, which often have parental read support and are thus not detectable by standard parent/child trio mutation calling methodology.

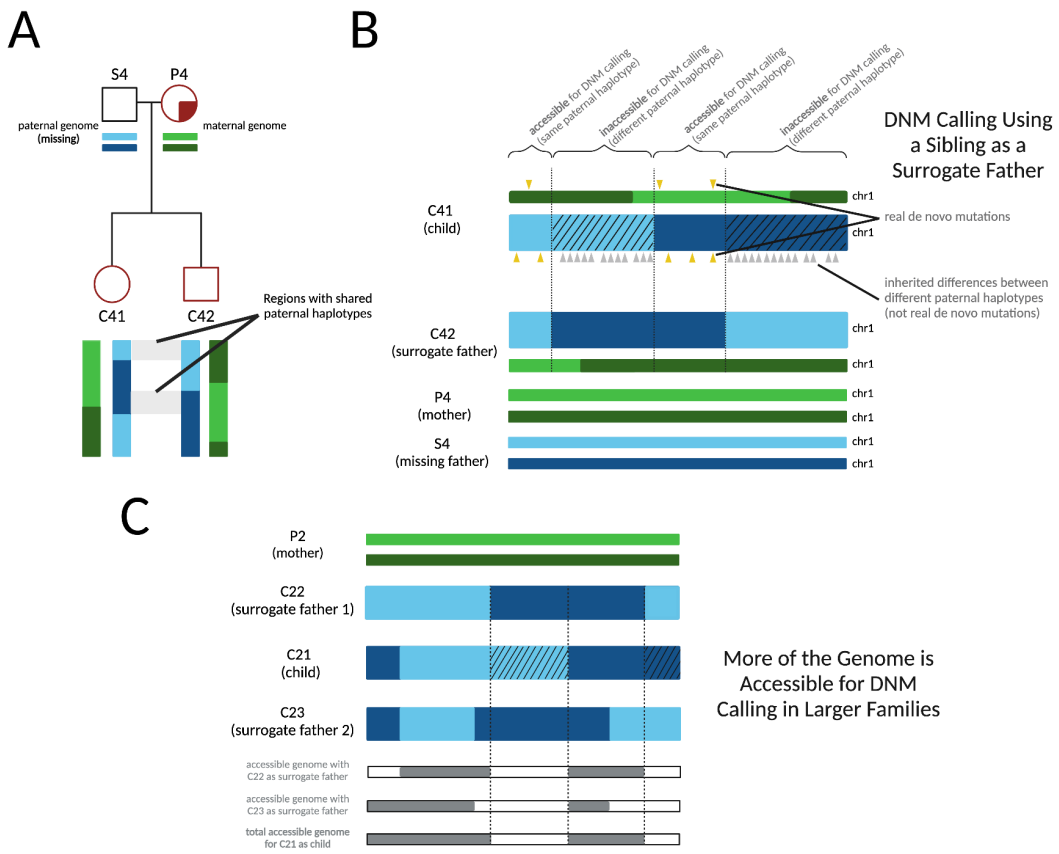


Figure 2. Calling DNMs using one parent and one surrogate parent. A) An illustration of the portions of an autosome with paternal haplotypes shared between two siblings. In an example chromosome from Family 4, DNMs can be called in regions where C41 and C42 share a paternal haplotype sequence with one another. **B)** An illustration of DNM calling using a sibling as surrogate father. In regions where the siblings inherited the same paternal haplotype, Mendelian violations (DNM calls, yellow triangles) are spaced far apart, but in regions where the siblings inherited different paternal haplotypes, Mendelian violations (gray triangles) are clustered close together, mostly stemming from polymorphic differences between the different paternal haplotypes inherited by the respective siblings. Hashed chromosome regions represent inaccessible regions of the genome, where DNMs cannot be called using the surrogate approach. **C)** An example of the surrogate method applied to Family 2, a three-child family where two different surrogate fathers can be used to call DNMs in each child. A set of partially overlapping candidate DNMs is generated from each sibling comparison, increasing the amount of accessible genome where mutations can be identified with more siblings used in this approach and allowing additional validation of calls in regions where accessible regions overlap.

As in standard DNM calling pipelines (Bergeron et al. 2022), we only call mutations at sites where both the real and surrogate parents are homozygous for the reference allele. This allows us to polarize mutation calls with confidence: if we are identifying DNMs in Sibling 1 using Sibling 2 as a surrogate parent, all DNM calls will occur at sites where Sibling 1 is heterozygous and

Sibling 2 is homozygous, which are not variants that could also be attributed to DNMs in Sibling 2.

Our use of the term "surrogate parent" is loosely based on the established use of relatives as surrogate parents for haplotype phasing (Kong et al. 2008). We also drew inspiration from several recent studies that successfully estimated mutation rates using tracts inherited identical-by-descent (IBD) between relatively distant relatives (Narasimhan et al. 2017; Tian et al. 2019, 2022). In contrast to these earlier studies, which estimated population-averaged mutation rates using mutations that likely occurred many generations ago, our method is designed to estimate an individual parent/child trio's mutation rate using mutations that arose within a single generation. This method enabled us to estimate germline mutation rates within five nuclear subfamilies of the pedigree and study how *MUTYH* genotype affected the germline mutation rate and spectrum.

To identify regions accessible for DNM calling using a mother and a surrogate father, we first used the software hap-IBD (Zhou et al. 2020) to identify long haplotypes shared identical by descent between each pair of siblings. We then filtered these regions to exclude maternally inherited haplotypes. In addition, we implemented a "SNP density filter" to exclude what appear to be false positive IBD calls: regions where the density of pairwise differences between the siblings is too high to be consistent with true sharing of paternal haplotypes (see "Filtering" in the supplementary methods). Within the remaining regions of paternal haplotype sharing, we called DNMs using a standard GATK-based pipeline, using the surrogate father as the father. In this setup, gene conversion between the two paternal haplotypes has the potential to create false

positive DNMs, as previously observed by Narasimhan et al. (2017). To minimize these errors, we filtered out putative DNMs that were present in the 1000 Genomes data or in two or more members of our pedigree (this should eliminate gene conversion errors at any loci where not all siblings inherited the same paternal haplotype). We note that the familial mutation sharing filter will cause us to miss the 1-2% of DNMs that are shared between siblings due to germline mosaicism (Jónsson et al. 2018). The 1000 Genomes filter may also cause us to exclude some true DNMs, particularly at CpG sites or other mutation hotspots, but it should effectively filter out many false positive DNMs that were actually inherited from a missing parent, except when those inherited variants are very rare in the population as a whole.

In a large family with a mother and n+1 siblings, the only regions inaccessible for surrogate-father DNM calling will be regions where n siblings inherited the same paternal haplotype and the remaining sibling inherited the other paternal haplotype. In this scenario, the sibling who inherited the unique paternal haplotype has no one to serve as a surrogate father, so that sibling's genome will be inaccessible for DNM calling while the n other siblings' genomes will be accessible. If we consider the paternal haplotype inherited at a specific locus in a specific sibling's genome, the probability that any given sibling inherited the same paternal haplotype is ½. Therefore, the probability that all n other siblings inherited the other paternal haplotype is 2^{-n}. If any one other sibling did inherit the same paternal haplotype, the first sibling's genome will be accessible for DNM calling at the locus we are considering. This implies that in our hypothetical family with one parent and n+1 children, a fraction $1 - 2^{-n}$ of all DNMs should be callable. For example, in Family 2, which consists of 3 children and their mother, ¾ of all DNMs should be callable.

Within each region that is accessible for DNM calling, most DNMs occurring on the proband's maternal and paternal haplotypes should be identifiable, with the exception of DNMs that are shared with the sibling surrogate parent. Neglecting these sib-shared mutations, the mutation rate can be estimated by dividing the mutation count by two times the length of the accessible genomic region spanned by paternal IBD tracts. Moreover, read-backed phasing tools that are designed for use in parent/child trios can be applied with the surrogate father substituted for the father (Belyeu et al. 2021). Read-backed phasing will deduce that a mutation arose on a paternally inherited chromosome if it can be phased to a haplotype shared between the siblings that is not shared with their mother. Similarly, we can deduce that a mutation occurred on a maternally inherited chromosome if it can be phased to a haplotype that is shared with the maternal genome.

Since full siblings share DNA inherited from both of their parents, they can serve as surrogate parents for DNM calling even when the mother and father are both deceased, as is the case for the Generation III siblings P1, P2, P3, and P4 (illustrated in **Figure 3A**; for more details, see "using surrogate parents for DNM calling" in the methods and supplementary methods). Since maternal and paternal chromosomes are passed down independently of one another, the fraction of DNMs accessible for calling in a family with n+1 children and no parents is expected to be $(1 - 2^{-n})^2$, slightly smaller than the fraction $1 - 2^{-n}$ that is callable when paternal or maternal DNA is available. To call DNMs in P1, P2, P3, and P4 using surrogate parents alone, we first used hap-IBD to identify tracts shared identical by descent between pairs of siblings. For each sibling trio (P_i, P_j; P_k), we then identified the set of regions where DNMs are callable in P_k

using P_i and P_j as surrogate parents: this is the set of regions where P_k shares one haplotype IBD with P_i and shares the other haplotype IBD with P_j (**Figure 3B**). As long as two surrogate parents share distinct overlapping IBD tracts with the proband, we are able to randomly designate one as the surrogate mother and the other as the surrogate father and proceed with DNM calling without determining which haplotypes are actually maternally and paternally inherited (though we cannot use read-backed phasing to deduce whether mutations occurred on maternal versus paternal haplotypes). We also identified regions of the genome where one sibling was able to serve simultaneously as surrogate mother and father to another sibling; this is possible wherever the two siblings inherited the same haplotype from each of their parents, which occurs across about 25% of the genome in full siblings (**Figure 3C**).

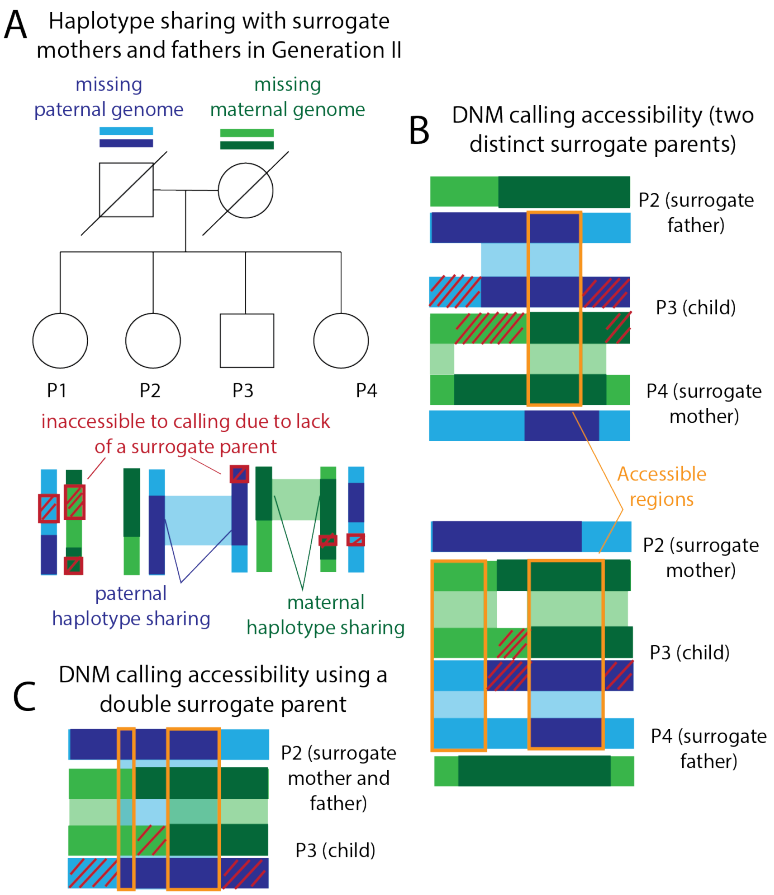


Figure 3. Calling DNMs using only surrogate parents. A) Generation III of our pedigree contains four full siblings (P1, P2, P3, P4) whose parents are deceased. We were able to call DNMs in these individuals using siblings as surrogate mothers and fathers. Example regions of paternal and maternal haplotype sharing are regions where P2 and P4 can act as surrogate father and mother to P3. Red crosshatches denote parental haplotype segments that were inherited by exactly one sibling–these regions are inaccessible to DNM calling due to lack of a surrogate parent. **B)** Illustration of maternal and paternal haplotype sharing in two example surrogate trios. DNMs are callable within orange regions due to haplotype sharing with both the surrogate mother and the surrogate father. **C)** Illustration of the regions that are accessible for calling in P3 using P2 as a double surrogate parent. This is possible when the surrogate and the proband share both maternal and paternal haplotypes.

To assess the quality of our surrogate-based DNM calls, we performed benchmarking using a family that we simulated as a composite of whole genome sequences from the 1000 Genomes project and mutation and recombination events from trios previously sequenced by Jónsson et. al (2017) (see "generation and analysis of simulated trio data for surrogate-method benchmarking" in the methods and supplementary methods). We simulated a family with 5 total children, randomly selected one child as the "proband," and called DNMs in this child using different subsets of the available real and surrogate parents. In both real and simulated families, DNM calling accessibility was similar to theoretical prediction (**Figure 4A**). As illustrated in **Figure 4B**, different sets of surrogate parents provided coverage of complementary genomic regions, and overlap between these tracts permitted error-correction of false positives that only appear when specific sets of surrogate parents are used. Overall inflation of the mutation rate by false positives is modest and appears to decrease with increasing family size (**Figure 4C-F**). Our simulation results suggest that caution is warranted when interpreting the mutation rates and spectra in Family 4, which contains just two siblings as well as their mother. Only half of the genome is accessible for DNM calling in this family (**Figure 4A, Figure S3**) and the ability to correct for paternal gene conversion will be limited, since none of the regions where DNMs are callable will contain both paternal haplotypes.

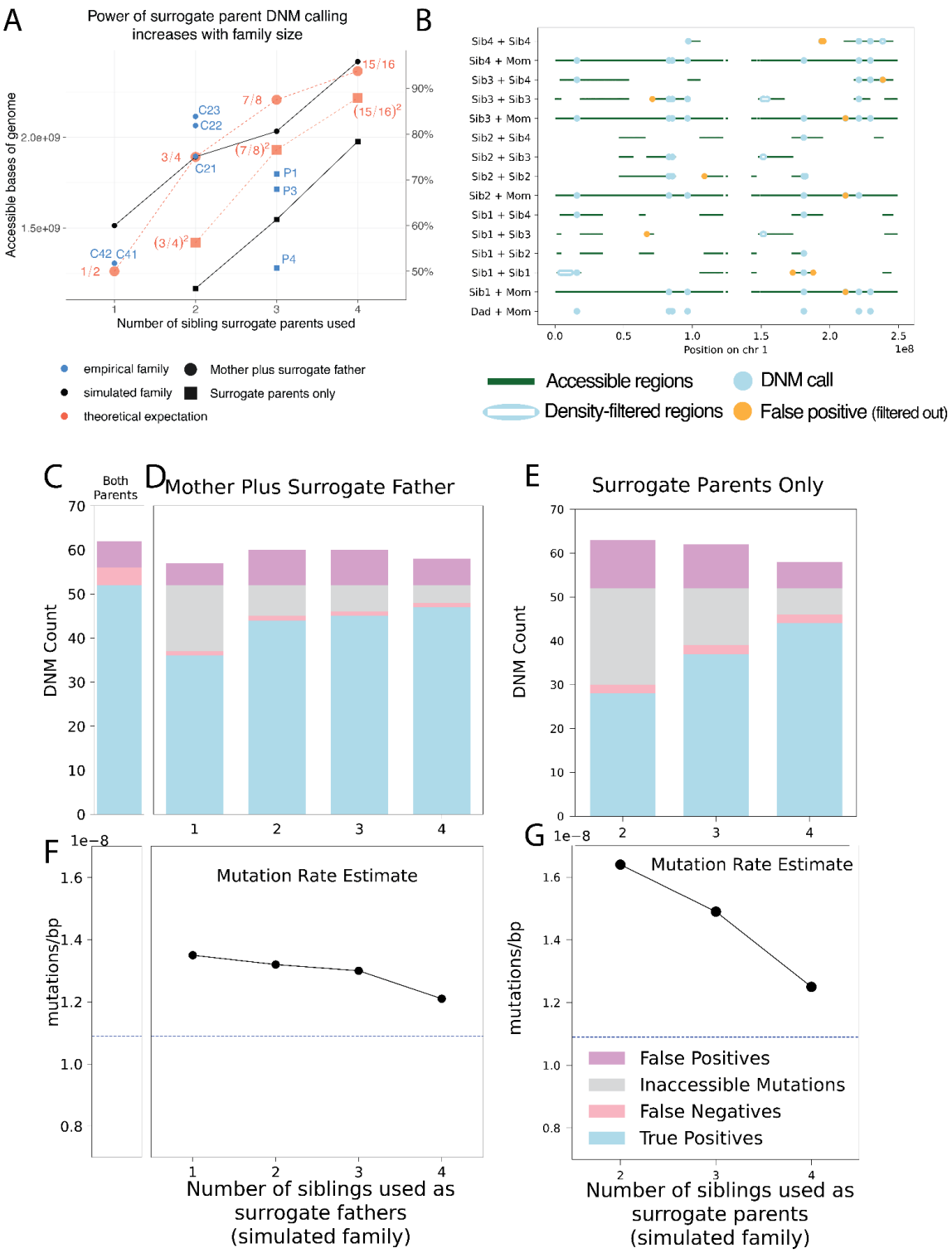


Figure 4: Precision and recall of surrogate-based DNM calling. **A)** The number of base pairs accessible for DNM calling in each individual from our pedigree (blue) depended on the number of real and surrogate parents available, broadly matching theoretical expectations (red dashed lines) as well as the proportion of accessible genome available for calling in our simulated pedigree (black) using similar configurations of real parents and siblings. Fractions in red indicate the proportion of the genome that is theoretically accessible. Families for which one real parent is available (circles) have a greater proportion of the genome

accessible than those with no parents available (squares), but both improve when additional siblings are available. Theoretical expectations reflect the size of the genome accessible for SNP calling in the simulated family, which is close but not identical to the callable genome sizes in the empirical family. Empirical individual P2 was excluded due to somatic mutation contamination which lowered the accessible genome size (**Figure S3**; see discussion below). **B)** Genomic accessibility, DNM calls, and density-filtered regions in chromosome 1 of the simulated pedigree. Note that error correction is facilitated by overlap between regions accessible to calling using different sets of surrogate parents. **C)** True positive, false positive, and false negative DNM calls made in a simulated proband with both mother and father's genomes available. **D)** As in C, but with a mother plus different numbers of siblings to act as surrogate fathers. Inaccessible mutations are DNMs that occur in regions of the genome without a suitable surrogate father. **E)** True positive, false positive, and false negative DNM calls made in a simulated proband with no parental genomes available, using different numbers of surrogate mothers and fathers. **F)** Mutation rates estimated with both simulated parents 'genomes available (dashed line). **G)** Mutation rates estimated using the mother's genome and surrogate fathers. **H)** Mutation rates estimated using surrogate mothers and fathers, with no parental genomes available.

Children of pathogenic MUTYH carriers have normal germline mutation rates

Previous studies have found that *MUTYH* variants specifically increase the C>A mutation rate in a variety of species and cell types (Sasani et al. 2022; Robinson et al. 2022). Since C>A comprises only about 10% of human DNMs, even relatively large perturbations of the C>A mutation rate are not necessarily enough to push the overall germline mutation rate significantly above its normal range, as previously seen in mice as well as humans (Sasani et al. 2022; Sherwood et al. 2023). In keeping with this expectation, we found most trios in this study to have normal mutation rates ranging from 8.23×10^{-9} to 2.14×10^{-8} mutations per base pair per generation (**Table S1**), comparable to the range between 7.9×10^{-9} to 1.9×10^{-8} expected in healthy individuals with parental ages between 15 and 50 based on a large previous study (Jónsson et al. 2017).

However, we called a much higher frequency of mutations (4.4×10^{-8} mutations per site per generation) in the genome of P2, the biallelic mother of Family 2. P2 was previously diagnosed

with colorectal cancer and underwent chemotherapy, which is known to induce somatic mutations in hematopoietic stem cells. Upon further examination, we found most of P2's mutations to have unusually low variant allele frequencies (VAFs), between 20% and 50%. All other individuals had mutation VAF distributions centered around 50%, as expected of germline mutations that arose on one of two parental haplotypes (**Figure S4**). P2's VAF skew suggests that most of their DNM calls are likely somatic mutations rather than germline mutations. Sherwood et al. (2023) previously noted a similar pattern in one of their biallelic *MUTYH* carriers who had undergone 5-fluorouracil chemotherapy for colorectal cancer, a treatment that can cause high-frequency mutations to emerge in the hematopoietic stem cell population. Due to this excess load of somatic mutations, which preclude estimation of an accurate germline mutation rate, we excluded P2 from further analysis and required a minimum VAF threshold of 30% for all mutations called in other individuals. Despite P2's high somatic mutation rate, we found that their mutation spectrum was dominated by normal background mutational signatures SBS1 and SBS5, with no sign of the MUTYH-associated signatures SBS18/SBS36 (**Figure S5**).

Testing the children of biallelic MUTYH carriers for skewed mutation spectra and parent-of-origin bias

Although we did not expect the overall mutation rate to be significantly elevated in trios with biallelic *MUTYH* carrier parents, we hypothesized that these families might have elevated proportions of C>A mutation types specifically, and/or a higher-than-expected proportion of C>A mutations inherited from the affected parent. To maximize our power to test for these effects, we calculated trio-specific expected C>A mutation counts and proportions using a model

fit to patterns of de novo mutations in 1,548 Icelandic trios with no known mutator phenotypes (Jónsson et al. 2017). Although this control dataset was generated separately from our study, we carried out similar filtering methodologies (**Figure S1A**), and all individuals in both studies are of European descent.

The Icelandic trio study by Jónsson et al. (2017) leveraged their data to predict the expected rate of each 1-mer mutation type per base pair per generation as a function of paternal and maternal age. Using this parental age model, we were able to calculate each trio's expected maternal and paternal 1-mer mutation burden as a function of the parents' ages at conception of the child (**Table S1**) and their accessible genome sizes (**Figure S6, Table S1**), following an approach recently used by Kaplanis, et al. (2022). For the most part, our empirical counts agreed with these expected counts (**Figure 5A**). In every child except for C42, the younger child with the abnormally high mutation rate in the family where we previously flagged DNM calling issues, the observed total mutation burden is within the upper one-tailed Poisson 95% confidence interval expected under the parental age model (**Figure S7**).

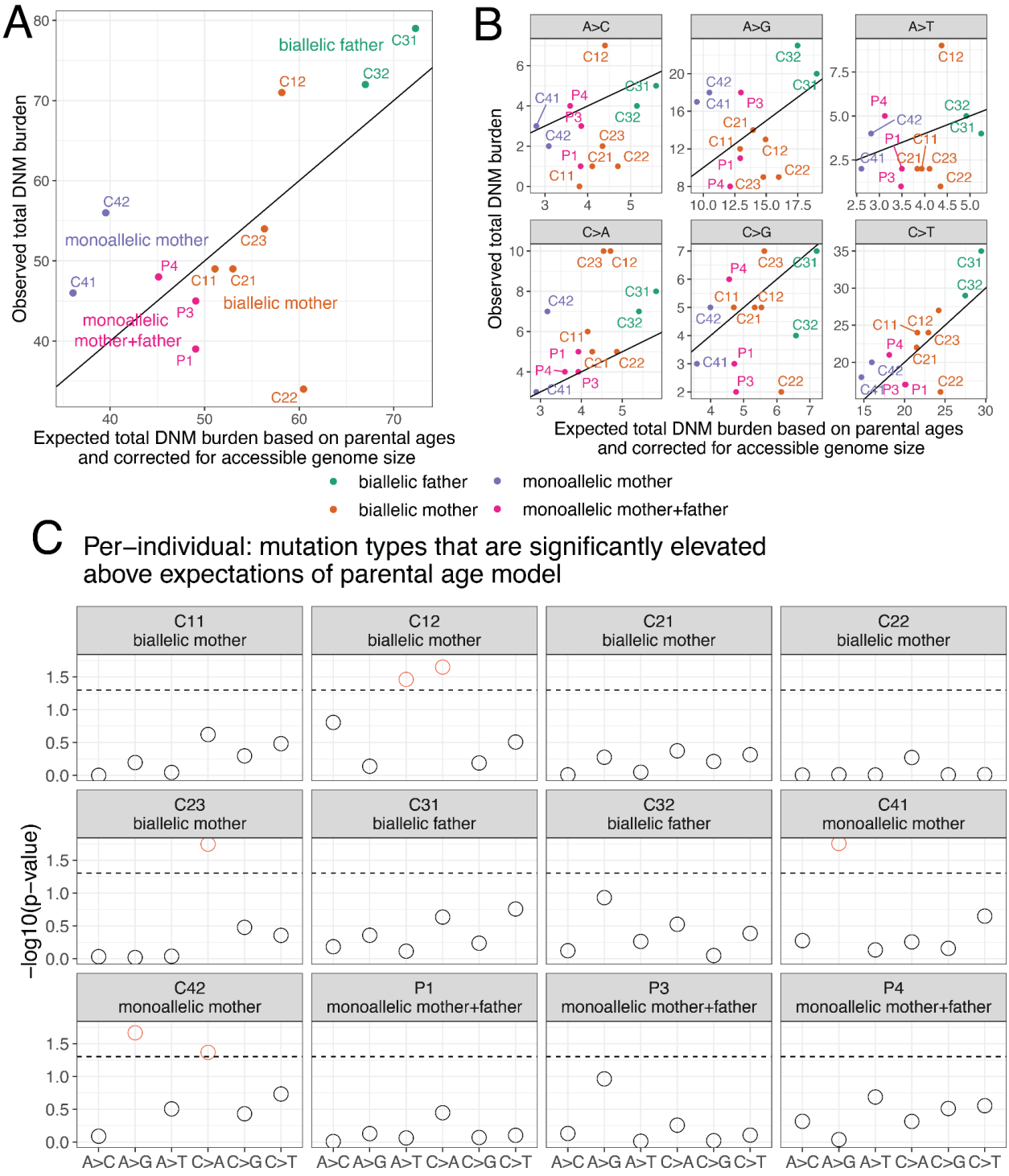


Figure 5. Observed and expected mutation counts. A) Comparison of observed DNM counts per child and the corresponding expected DNM counts under the parental age model (Jónsson et al. 2017), corrected for accessible genome size (**Figure S6**). P2 was excluded as discussed above due to evidence for somatic mutation contamination. Points are colored by the *MUTYH* carrier status of the child's parent(s). Each child except for C42 has an overall mutation count that is compatible with the Jónsson parental age model (**Figure S7**). See **Figure S8** for a comparison with the results of Sherwood et al. (2023). **B)** Observed and expected mutation counts, faceted by 1-mer mutation type. Note that C>A counts are above the $y = x$ line for nearly all

individuals. **C)** The probability of observing a mutation count of each of the six 1-mer mutation types under the parental age model that is greater than or equal to what we observed for each member of the pedigree. Points above the dashed line (red circles) fall below the upper one-tailed Poisson $p < 0.05$ significance threshold. C12 and C23, both children of biallelic mothers, show significant elevation of C>A DNM counts, as does C42 (child of a monoallelic mother).

When we categorized DNMs by 1-mer mutation type (**Table S2**), we found that individual trio mutation spectra were largely consistent with the parental age model (**Figure S9**), but that C>A is the mutation type whose observed counts were most consistently inflated above expected counts, exceeding the expected count in all 12 individuals (**Figure 5B-5C**). Across the remaining five 1-mer mutation types, the proportion of trios exceeding the mutation count predicted by the parental age model ranged from 3/12 trios (A>C mutations) to 9/12 trios (C>T mutations) (**Figure 5B**). Most of the elevated C>A counts fell within an upper one-tailed 95% Poisson confidence interval of the expected count, but three children's C>A burdens significantly exceeded the parental age model expectation (**Figure 5C**). These included C42, as well as C12 and C23, the children of two different biallelic mothers. The only non-C>A counts significantly exceeding the parental age model expectation were A>G mutations in C41 and C42 and A>T mutations in C12 (**Figure 5C**). However, the validation of false positive calls through our surrogate method was most challenging for C41 and C42 due to fewer available siblings as surrogate parents (**Figure 4A**), and thus the excess C>A and A>G mutations in these individuals may indicate a possible signal of inherited germline variant bleed-through resulting from the surrogate-calling method.

We then added up sibling mutation counts to estimate the aggregate C>A enrichment within each nuclear family and found that the two families with biallelic mothers (Families 1 and 2) were enriched for C>A mutations by 1.81-fold and 1.46-fold above the expectation of the parental age

model, respectively (**Figure 6A**). In Family 1, the total C>A mutation burden significantly exceeded the 95% upper 1-tailed confidence interval of the parental age model, while Family 2 falls 1 mutation below this significance threshold (**Figure 6B**). These C>A enrichments are comparable to the 1.57-fold-elevated C>A mutator phenotype recently identified in the mouse strain DBA/2J, but much less dramatic than the 6.04-fold enrichment phenotype identified in the mouse strain BXD68 caused by the homozygous loss of function R182Q-like mutation (**Figure 6A**). In contrast, neither child with a biallelic father had a significantly elevated C>A mutation load, and their family (Family 3) was only enriched 1.34-fold for C>A mutations overall (**Figure 6A-6B**). C>A mutation load was only 1.14-fold elevated (also nonsignificant) in the four parents P1–P4, whose own parents were all monoallelic and thus not expected to have a germline mutator phenotype. In Family 4, the family with a biallelic mother and bioinformatic limitations in our ability to accurately call DNMs, we observed a non-significant 1.65-fold C>A enrichment along with a significant 1.75-fold A>G enrichment (**Figure 6A- 6B**). A 1.65-fold C>A enrichment fails to reach significance in Family 4 because the siblings C41 and C42 each have a smaller callable genome proportion than individuals from families with a father or third sibling available for genotype calling. As in Sherwood et al. (2023), we then further summed up the mutation counts for children with the same carrier parent type, which in the case of our pedigree means summing up Families 1 and 2 which both have a biallelic mother as the carrier parent. This biallelic mother group showed significant enrichment of C>A DNMs (**Figure S10**).

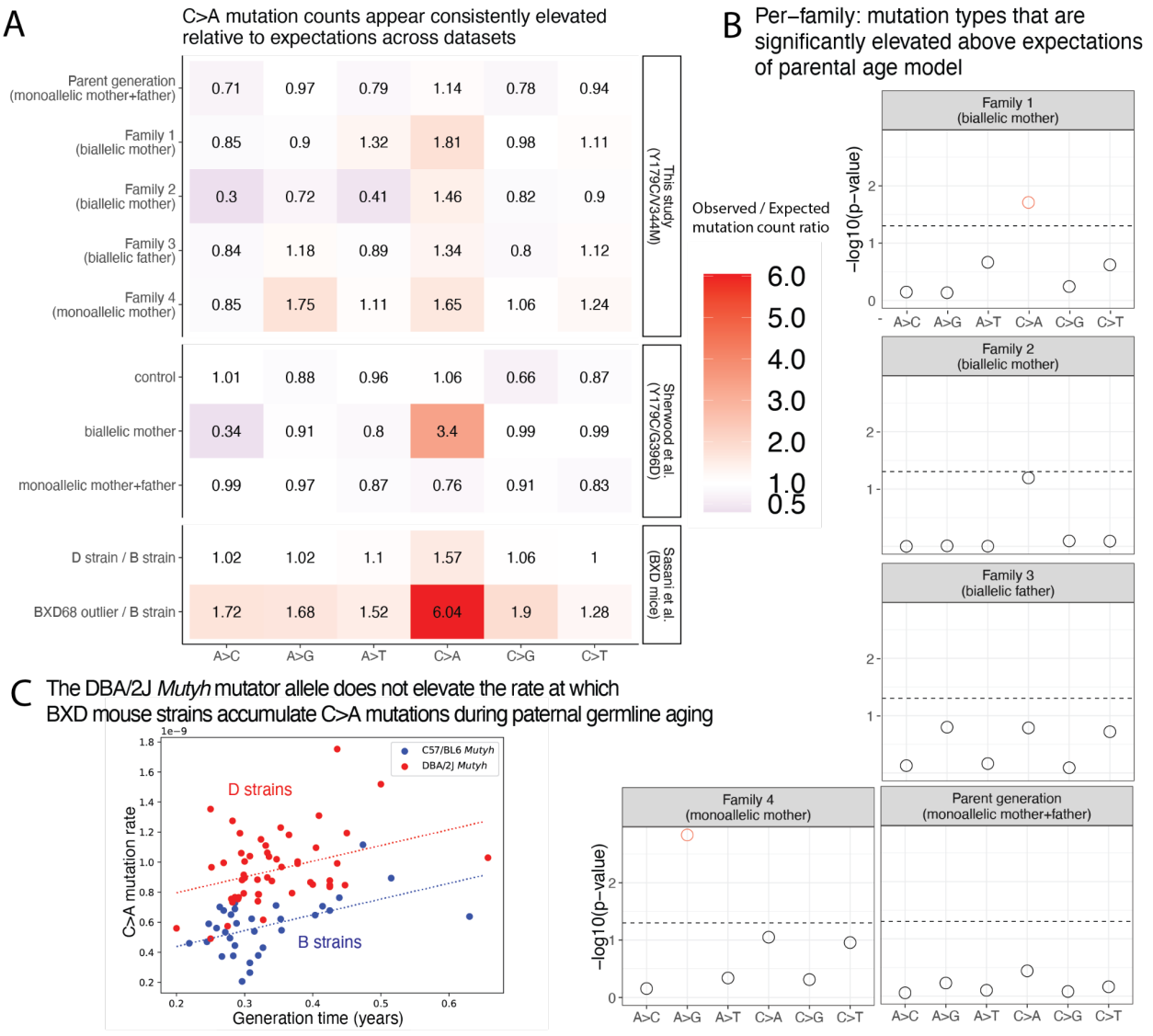


Figure 6. Children of mothers with biallelic *MUTYH* genotypes show significantly elevated C>A DNM counts. A) A heatmap showing the ratio of the observed / expected mutation counts per family (calculated by summing up the mutation counts per mutation type across all children within a family). These ratios are compared to the observed / expected ratio for the groups in Sherwood et al. (2023) (control group, individuals with a biallelic mother, and individuals with monoallelic parents), with expectations calculated using the parental age model. The bottom two rows show results from Sasani et al. (2022) for inbred BXD mouse strains: the "D" strain has an elevated mutation rate relative to the "B" strain, which has been linked to variation in *Mutyh*, and BXD68 is a mouse individual with an extreme outlier C>A mutator phenotype caused by a homozygous loss of function nonsynonymous mutation. The mouse ratios compare the per-generation rate of each mutation type between sets of inbred BXD mouse strains with different *Mutyh* genotypes. **B)** The probability of observing a mutation count of each of the six 1-mer mutation types under the parental age model that is greater than or equal to what we observed for each family in the pedigree. Points above the dashed line (red circles) fall below the upper one-tailed Poisson $p < 0.05$ significance threshold. Family 1 shows significant elevation for C>A DNM counts above what is expected under the parental age model, and Family 4 shows significant elevation of A>G mutations. **C)** Multilinear regression of C>A mutation rate per generation as a function of

generation time across BXD mouse strains. The *Mutyh* allele affects the regression intercept but not the slope, implying that mutator strains and non-mutator strains accumulate C>A mutations at the same rate during parental aging but accumulate these mutations at different rates during embryonic development.

We confirmed that our parental age model significance-testing framework was able to distill some of the main findings of Sherwood et al.'s (2023) study of germline mutator effects: in particular, the combined C>A burden of the children of the Sherwood et al. biallelic mother exceeded the 95% one-tailed confidence interval of the parental age model (**Figure S9**). In addition, all children of *POLE* and *POLD1* variant carriers in Sherwood et al. (genotypes which appear to have much more severe germline mutator effects than *MUTYH*) significantly exceeded the C>A and A>G mutation burdens predicted under the parental age model (**Figure S9**). We calculated a significant 3.4-fold enrichment of C>A mutations above the parental age model expectation in the family with a biallelic maternal *MUTYH genotype* sequenced by Sherwood et al. (**Figure 6A; Figure S11)**, suggesting that this family's p.Y179C/G368D genotype may have a more severe mutator phenotype than the p.Y179C/V234M genotype affecting our pedigree.

Unlike Sherwood et al. (2023), we did not detect a significant increase in overall DNMs phased to the haplotype of the carrier parent (**Figures S12-S13, Table S1**). However, we did detect a significant elevation in C>A mutations phased to the maternal haplotype in Family 1 (one of the two families in our pedigree where the mother is the biallelic *MUTYH* variant carrier) (**Figure S14, Table S3**), indicating that there may be a carrier-parent-specific elevation of C>A mutations in this family. We note that this result is based on very low sample sizes of phased de novo mutations: 3 C>A mutations phased to the maternal haplotype in Family 1, compared to an expectation of 0.79 mutations, and so may be largely driven by stochasticity. In their biallelic

families, Sherwood et al. were able to detect the activity of COSMIC mutational signature SBS18, a signature associated with defective *MUTYH* DNA repair (Alexandrov et al. 2020). However, mutational signature analysis of our DNM data did not identify any activity of either of the *MUTYH*-associated signatures SBS18 or SBS36. This likely reflects the small total sample size of C>A mutations in our data (**Figure S15**) and should not be interpreted as evidence of absence of SBS18/SBS36.

Estimating C>A mutator effect sizes in the maternal and paternal germline

One consistent feature of human germline mutagenesis is that only about 25% of mutations appear to arise in the maternal lineage (Wong et al., 2016; Gao et al., 2019). In a family where the mother's *MUTYH* genotype is pathogenic but the father's genotype is normal, any elevation of the C>A mutation rate observed in the children likely stems from either excess mutations that arose in the oocyte prior to conception or postzygotic mutations. Even if a child has inherited a normal *MUTYH* allele from their father, their postzygotic mutations may still be enriched for C>A if they arise prior to the maternal-zygotic transition, when the embryo first begins to express paternally inherited genes.

Given that almost 75% of germline mutations originate in the spermatocytes, the minimum maternal C>A enrichment required to explain our data is expected to be larger than the overall C>A enrichments recorded in **Figure 6**. To estimate the maternal germline C>A mutator effects that are required to explain the data, we started with the observed C>A mutation counts in Families 1 and 2 and subtracted the maternal and paternal C>A counts expected under the

parental age model (**Figure S16A**). We then added each excess C>A count to the expected maternal C>A count and computed the proportional inflation of this value above the expected maternal C>A count. Using this logic, we calculated that the 1.46-fold to 1.81-fold overall C>A rate elevations observed in Families 1 and 2 (**Figure 6A, S16A**) imply maternal C>A mutation rate elevations of 4.2-fold and 2.7-fold, respectively (**Figure S16B**). The maternal effect implied by the Sherwood et al.'s (2023) 3.4-fold increase in overall C>A count is even larger: this value translates to a 10.2-fold elevation of the maternal C>A mutation rate (**Figure S16B**).

Although Family 3's nonsignificant 1.34-fold C>A mutation rate elevation is not much lower than Family 2's C>A enrichment, it implies a much lower paternal C>A rate elevation of only 1.45-fold (**Figure S16B**). This is very different from the 4.2-fold and 2.7-fold maternal C>A rate elevations that we infer to affect the two mothers who share the same biallelic *MUTYH* genotype. To investigate the likelihood that we were simply underpowered to detect a male germline mutator effect in Family 3, we calculated a "mutator detection threshold" for each family, which is the minimum number of extra C>A mutations required to produce a significant deviation from the parental age model (horizontal black bars in **Figure S16A**).

We then calculated how much this minimum number of extra C>A mutations should inflate the germline rate in the parent with the biallelic *MUTYH* genotype: this is the minimum fold-elevation of the biallelic parent's C>A mutation rate that we have power to detect (horizontal black bars in **Figure S16B**). **Figure S16B** compares these minimum effect sizes to the effect sizes estimated using our empirical data (orange points). According to these calculations, we should have power to detect a paternal C>A mutator effect of 1.8-fold or greater, which is

notably smaller than the maternal effects supported by the data yet exceeds the level of paternal C>A enrichment that is supported by the data. We also carried this analysis on a per-individual level, rather than summed per family (**Figure S17**). Although this analysis is based on a limited sample size of individuals, it suggests that *MUTYH* variants may have a proportionally stronger effect on the maternal germline compared to the paternal germline.

Mutyh variation does not increase the strength of the paternal age effect in the BXD mice

To further investigate the etiology of *MUTYH*'s germline mutator effect, we turned our attention from the human genotype p.Y179C/V234M to the mutator allele affecting the murine homolog *Mutyh* in the mouse strain DBA/2J. This mutator allele, which we call *Mutyh-D*, occurs in half of the recombinant inbred mouse strains known as the BXDs, which are each descended from crosses of DBA/2J with the standard lab strain C57/BL6. Each BXD strain has been inbred for tens or hundreds of generations and was previously whole-genome sequenced, which allowed the average mutation rate over the inbreeding period to be measured with high precision (Sasani et al. 2022). These rates revealed that the "D strains," which have DBA/2J ancestry at *Mutyh*, have higher C>A mutation rates than the "B strains" which have C57/BL6 ancestry at this locus (Sasani et al. 2022). We were able to leverage these data to measure how the C>A mutation rate depends on parental age in the D strains as opposed to the B strains. Each BXD mouse strain has been inbred for a known number of generations spanning a known number of years, and we used these records to calculate each strain's average generation time. We found that these rates ranged from 0.2 years to 0.63 years, spanning more than half of the mouse reproductive lifespan.

We fit a multilinear regression model to infer the dependence of the C>A mutation rate on *Mutyh* genotype (B or D) jointly with parental age, letting the *y* intercept of the model be the C>A mutation rate at the minimum parental age of 0.2 years (**Figure 6C**). We inferred a parental age effect of 1.05×10^{-9} additional C>A mutations per site per year (ANCOVA $p < 0.001$), but found no significant interaction between this parental age effect and *Mutyh* genotype (ANCOVA $p > 0.82$), indicating that the rate of C>A mutations occurring during gamete aging does not appear to differ between the B and D strains. In contrast, the baseline C>A mutation rate at age 0.2 years differs significantly between the B and D strains (4.39×10^{-10} versus 7.96×10^{-10} mutations per site per generation; ANCOVA $p < 0.001$). This suggests that the elevated C>A mutation rate associated with the DBA/2J *Mutyh* allele is driven primarily by early embryonic mutations, not mutations that occur in the paternal (or maternal) gametes.

Materials and Methods

SNP calling and trio-based De Novo Mutation (DNM) calling

Variant detection and genotyping were performed using the GATK HaplotypeCaller (4.2.0.0) (Van der Auwera 2020). Variants were initially flagged using the filtration walker (GATK) to mark sites that were of lower quality [e.g., low quality scores (Q50), allelic imbalance (ABHet 0.75), long homopolymer runs (HRun> 4) and/or low quality by depth (QD < 5)]. Data QC included an assessment of: (1) mean coverage; (2) fraction of genome covered greater than 10X; (3) duplicate rate; (4) mean insert size; (5) contamination ratio; (6) mean Q20 base coverage; (7) Transition/Transversion ratio (Ti/Tv); (8) fingerprint concordance > 99%; (9) sample homozygosity and heterozygosity; and (10) sample contamination validation. Genome completion was defined as having > 95% of the target read at > 10X coverage and > 90% of the target at > 20X coverage.

Putative DNMs in parent-offspring and surrogate-offspring trios were identified using the GATK(v4.2.6.1) PossibleDeNovo tool, which uses the genotype information from individuals in family trios to identify possible de novo mutations and the sample(s) in which they occur.

Using surrogate parents for DNM calling

In children without two sequenced parents, we called DNMs using parental haplotypes shared with “surrogate parent” siblings. To identify parental haplotypes shared between relatives, we

began by phasing the full 15-genome dataset using Beagle (Browning et al. 2021). In order to improve phasing quality, we phased these genomes together with a panel of 3,202 genomes from the high-coverage 1000 Genomes Project (Byrska-Bishop et al. 2022). Since rare variants are generally uninformative for identity-by-descent (IBD) segments, and are prone to sequencing error and phasing error, we filtered for common variants that are found at minor allele frequency > 10% in a subset of 2,504 genomes and used them as the input to the program hap-IBD to infer shared tracts of IBD (Zhou et al. 2020). Additionally, the following hap-IBD parameter settings were used: min-seed=1.0, max-gap=1000, min-extend=0.2, min-output=2, min-markers=100. In this way, we were able to identify IBD segments that were shared between siblings but not present in any sequenced parent and then use these IBD segments as surrogates for missing paternal and maternal genome sequences. We noted that putative DNMs often occurred near the ends of our inferred surrogate parent tracts, and we hypothesized that these might be artifacts caused by inaccuracies in the boundaries of shared IBD tracts. To eliminate these artifacts, we implemented a density-based filter (see "Filtering" in the supplementary methods and **Figure S1B**).

In some probands, we called DNMs using a mother plus sibling surrogate fathers, restricting to regions where the proband and the sibling share paternal DNA. In other probands, we called DNMs using two surrogate parent siblings, restricting to regions where the proband shares each of their haplotypes with one or more siblings (for specific details on the specific surrogate parent configurations that were used, see "Using surrogate parents for DNM calling" in the supplementary methods). Within these accessible regions, DNMs were called using GATK PossibleDeNovo, substitution surrogate parents for one or both parents. All preliminary DNM

calls were then filtered as described below (see "Filtering" and "IGV inspection" in the supplementary methods). During the IGV inspection step, we eliminated any putative DNM shared between two or more individuals in the extended family, except when those individuals have a parent-child relationship, assuming that most such shared variants were in fact inherited from un-sequenced parents in regions erroneously identified as inherited IBD. To eliminate additional false positive calls from the data, we manually removed any set of mutations that were clustered (over 6 mutations within 50 bp of one another) nearby a putative DNM with an rsID annotation, as we inferred these sites were more likely to have been inherited from a missing parent rather than a multi-nucleotide germline mutation event.

Generation and analysis of simulated trio data for surrogate-method benchmarking

To benchmark the performance of the surrogate method, we simulated a realistic pattern of mutation accumulation in a 5-sibling family and then simulated sequencing reads consistent with the family's genotypes. In brief, we generated the family's inherited germline variation using the 1,000 Genomes high coverage phased dataset (Byrska-Bishop et al. 2022) and then generated *de novo* mutations based on the data from Jónsson et al. (2017), both mapping to the hg38 assembly. Only autosomes were simulated.

More specifically, we sampled one "parent" at random from the 1000 Genomes CEU population and sampled the other at random from the GBR population, NA11893 and HG00132. This sampling was designed to match the European ancestry of our pedigree but avoid sampling individuals who were too closely related. We randomly generated five "children" of these parents

by sampling recombination breakpoints from a chromosome map and adding the DNMs observed in a child sequenced by a previous study by Jonsson, et al. (2017). Finally, we generated a short read BAM file consistent with each simulated genome using DWGSIM (0.1.15, www.github.com/nh13/DWGSIM), a tool that simulates sequencing reads from a reference genome and can incorporate custom germline variants. For additional details, see "Generation and analysis of simulated trio data for surrogate-method benchmarking" in the supplementary methods.

Calculation of each family's expected mutation burden in the absence of a genetic mutator effect

Jónsson et al. (2017) carried out whole genome sequencing of Icelandic families and identified parental age impacts on the number and spectra of inherited de novo mutations. We used the Poisson regressions carried out in this study (listed in Table S9 of Jónsson et al. 2017) to predict expected de novo mutation burdens and spectra for each of the families in our study, based on parental ages.

In short, for each individual in our study, we plugged their parents ʾpaternal and maternal ages at the time of their birth into the following equations to get the expected count of each mutation type *c* (C>A, C>T, C>G, A>G, A>C, A>T):

$$y_{c,\text{mat}}(a_{\text{mat}}) = m_{c,\text{mat}} \cdot a_{\text{mat}} + b_{c,\text{mat}}$$

$$y_{c,\text{pat}}(a_{\text{pat}}) = m_{c,\text{pat}} \cdot a_{\text{pat}} + b_{c,\text{pat}}$$

In these equations, a_{mat} and a_{pat} are the maternal and paternal ages at the time of a child's birth, respectively, and $m_{c,mat}$ and $m_{c,pat}$ are the numbers of mutations of type c accumulated each year in the maternal and paternal germlines (linear regression slopes from Jónsson et al. (2017)'s Table S9), and $b_{c,mat}$ are $b_{c,pat}$ the numbers of mutations of type c that would theoretically be present in the maternal and paternal germlines at age zero (mutation-type-specific maternal and paternal linear regression y-intercepts). The resulting expected de novo mutation counts inherited from the mother and father add up to the expected burden of each type of de novo mutations in their child. For additional details on correcting the accessible genome size and application of the model to additional datasets, see "Calculation of each family's expected mutation burden in the absence of a genetic mutator effect" in the supplementary methods. For details on statistical analyses of deviation from this null model, see the additional supplementary methods sections "Comparing our observed mutation counts to the null parental age model of Jónsson et al. (2017)" and "Estimating the minimum mutator effect sizes that we have power to detect."

Cellular assay of MUTYH glycosylase function

Human HEK293 *MUTYH* KO cell lines were transduced with lentivirus containing *MUTYH* cDNAs, either WT or variant, each cloned into pCW57.1 (Addgene #41393; gift from Dr. David Root). Transduced cells were selected and stable MUTYH expression was induced as previously described (Jia et al. 2021). To measure *MUTYH* variant function, cells expressing each variant were then co-transfected with a GFP reporter containing an 8oxoG:A mispair (Raetz et al. 2012; Nagel et al. 2014) and an mCherry-expressing plasmid as a transfection control. After a ~72 hr incubation with the reporter, cells were analyzed via FACS with a BioRad Ze5. A function score

was calculated as the fraction of repair positive (mCherry+, GFP+) cells out of all transfected cells (mCherry+), divided by the same quantity for cells transduced with WT *MUTYH*, and scaled by a log2 transform, such that a score of 0 indicates WT-like repair function, and negative scores indicate deficient function.

Analysis of parental age dependence of the BXD mouse mutation rate

We downloaded data on the mutation rate per generation, B versus D haplotype status at the *Mutyh* locus, and the number of generations of inbreeding for each of the BXD mouse strains from the Github page associated with Sasani, et al. (2022) (https://github.com/tomsasani/bxd_mutator_manuscript). We then computed each strain's C>A mutation rate per site per generation (number of years of inbreeding divided by number of generations of inbreeding) and fit an ordinary least squares multilinear regression to explain this variable as a function of generation time minus the minimum observed generation time of 0.2 as well as the categorical *Mutyh* status variable.

Supplementary Materials and Methods

Genome sequencing and SNP calling

All sequencing was conducted at the University of Washington Northwest Genomics Center (NWGC). Samples had a detailed sample manifest (i.e., identification number/code, sex, DNA concentration, barcode, extraction method). Initial quality control (QC) entailed DNA quantification, sex typing, and molecular "fingerprinting" using a 63-SNP OpenArray assay derived from a custom exome SNP set. This "fingerprint" was used to identify potential sample handling errors and provided a unique genetic ID for each sample, which eliminated the possibility of sample assignment errors. Samples failed if: (1) the total amount, concentration, or integrity of DNA was too low; (2) the fingerprint assay produced poor genotype data; or (3) sex-typing was inconsistent with the sample manifest. No samples failed quality control at this stage.

Library construction was automated in 96-well plate format. At least 750 ng of genomic DNA was subjected to a series of library construction steps utilizing the KAPA Hyper Prep kit (KR0961 v1.14). All library construction steps were automated on the Perkin Elmer Janus platform. Libraries were validated using the Biorad CFX384 Real-Time System and KAPA Library Quantification Kit (KK4824). Barcoded genome libraries are pooled using liquid handling robotics prior to loading. Massively parallel sequencing-by-synthesis with fluorescently labeled, reversibly terminating nucleotides was carried out on the NovaSeq sequencer. Variant calling was carried out by the NWGC. Their variant calling pipeline combined a suite of Illumina software and other "industry standard" software packages (i.e., GenomeAnalysis ToolKit

[GATK], Picard, BWA, SAMTools, and in-house custom scripts) and consisted of (1) alignment to human reference genome GRCh38DH using BWA-MEM (v0.7.15) (Li and Durbin 2009), (2) local realignment, (3) PCR duplicate removal (Picard MarkDuplicates; v2.6.0), (4) base quality score recalibration (BQSR) (GATK BaseRecalibrator; v3.7), (5) data merging, (6) variant detection, (7) genotyping, and (8) annotation.

The SeattleSeq Annotation Server (http://gvs.gs.washington.edu/ SeattleSeqAnnotation/), an automated pipeline, was used for annotation of variants derived from genome data. This publicly accessible server returned annotations including dbSNP rsID (or whether the coding variant was novel), gene names and accession numbers, predicted functional effect (e.g., splice-site, nonsynonymous, missense, etc.), protein positions and amino-acid changes, PolyPhen predictions, conservation scores (e.g., PhastCons, GERP), ancestral allele, dbSNP allele frequencies, and known clinical associations.

Using surrogate parents for DNM calling

We used GATK PossibleDeNovo as in the previous section on each informative "trio" of a child, a real parent if available, and one or two surrogate parents. For children whose parents' genome sequences were both available (C11, C12, C31, and C32), we performed no surrogate DNM calling. For children whose mother's genome was available but whose father's genome was unavailable (C21, C22, C23, C41, C42), we called DNMs using the mother's sequence plus each available relative as surrogate father. This resulted in two overlapping DNM call sets for each of the three siblings C21, C22, and C23, but just a single call set for C41 and C42. To generate each

call set, we generated a positive mask file consisting of regions that we identified to be shared IBD between the child and the surrogate father, then called DNMs within the bounds of this positive mask minus the standard negative mask previously used to filter out low quality regions during standard DNM calling. We then merged together all call sets generated for the same child with different surrogate fathers.

To call mutations in each of the parents P1–P4, we ran PossibleDeNovo a total of nine times, each using a different combination of relatives as surrogate mother and father. Six of these runs involved a pair of two distinct relatives P*i* and P*j*, and the remaining three runs used the same sibling as both the surrogate mother and the surrogate father. For each run, a distinct positive mask was used to call mutations only in regions where the child shared two distinct parental haplotypes with its pair of surrogate parents. In the case where the same relative was used as both surrogate mother and surrogate father, this meant regions where the child shared two distinct IBD tracts with the same surrogate parent, because the two relatives had inherited the same chromosome from both their mother and their father. As before, DNM calls from all nine runs were merged to generate the total call set for each individual.

We generated additional mutation calls from P1–P4 by using each sibling P*i* as a "double surrogate parent" for each other sibling P*j*. We performed double surrogate calling within regions where hap-IBD found that P*i* and P*j* shared two overlapping IBD tracts, which indicates that they inherited the same maternal chromosome and also the same paternal chromosome. Since GATK PossibleDeNovo is designed for use with two distinct parental genomes, we called candidate

double-surrogate DNMs within these double-IBD regions by identifying sites where the child is heterozygous but the double surrogate parent is homozygous.

We filtered out all DNMs called at sites that appear in the accessible regions of multiple surrogate parent combinations but are not consistently called using all of those surrogate parent combinations. For example, if a putative DNM in C22's genome occurs at a locus that appears accessible for calling using either C21 or C23 as surrogate father, but that DNM is only called using C21 as surrogate father, it will be filtered out of the final call set.

Accessible Genome Size Estimation

Using both conventional Mendelian violation methods and our devised surrogate method, we derived the overall mutation rate for each offspring. Determining these rates required the computation of a denominator for each individual within the pedigree. This denominator represented the number of genomic sites where the read coverage was adequate (i.e., greater than 12 or less than 120) to ascertain a mutation, if present. Sites lacking confident inference of an individual's parental haplotype sequences were excluded.

For offspring without sequenced fathers, our focus shifted to chromosomal regions where the child had an identical paternal haplotype with at least one sibling. For example, in the offspring of P2 with three children, two children with adequate read coverage at a site were necessary to identify mutations at that locus for both. For the parent generation, mutation identification depended on factors such as sufficient read coverage, successful haplotype reconstruction, and

inheritance patterns. Using the surrogate method necessitated adjustments to the denominators based on the total length of shared parental haplotypes, leading to variable accessible base numbers for offspring in Families 2 and 4 and the parent generation (**Figure S3**).

Filtering

DNMs were subjected to a series of quality control steps to eliminate potential false positives (**Figure S1A**). Building on prior research findings (Bergeron et al. 2022), true germline DNMs are usually characterized by alternative allele read support, with a variant allele frequency (VAF) ranging from 30% to 70%, and lack reads from either parent. DNMs were only considered for further analysis if they adhered to these parameters:

- Displayed a read depth between 12 and 120 for all members of both full pedigree and surrogate pedigree trios.
- Were identified by GA TK PossibleDeNovo as being present in the child but not in either parent.
- Exhibited a V AF of 30-70% in the child.
- Had no reads supporting the variant in either parent.
- Genotypes filtered with GA TK recommended hard filters: QD > 2.0; FS < 60.0; MQRankSum > -12.5; ReadPosRankSum > -8.0; SOR < 3.0

DNMs located in centromeres, telomeres, and segmental duplications were further excluded. Only DNMs that appeared in unique, accessible regions of the genome were retained in the final

dataset. Additionally, any DNM that overlapped with variants having a minor allele frequency (MAF) of 1% or higher in the 1000 Genomes Phase 3 dataset was excluded. For DNMs identified using surrogate parents, a sliding window methodology was employed to pinpoint sparse mutations. The stipulated criteria for this was a maximum of 7 mutations within a 15MB sliding window, advancing in increments of 3MB.

IGV inspection

In order to verify the mutation calls from both the full trio sequences and the resulting variants from families with surrogate parental sequences, we performed visual inspection of the resulting calls by inspecting the raw reads around the called de novo mutations.

We queried the original mapped sequences (bam files) to obtain all reads within 10kb (5kb slop) all pre-called de novo mutations in each trio of samples. When a mutation was detected in one of the families with a missing paternal genome we included all other samples in that trio that were used as a surrogate-paternal sequence, thus including multiple bam files as parental sequences.

The reduced files were then processed to filter low quality reads by selecting unduplicated sequences (-F 1024) and requiring a mapping quality higher than 20 (--min-MQ 20). To select informative reads used by GATK for variant calling, the unfiltered reads were also used to re-call variants using GA TK HaplotypeCaller with the -bamout flag option that returns the informative reads for each call in bam format. The resulting variant files from this step were discarded and

not used in any of the analysis. Note that if the algorithm would not return a mutation in that position there would be no informative reads available.

For each trio or surrogate-parent trio we generated a IGV report using igv-reports (github.com/igvteam/igv-reports) that outputs a HTML file containing small snippets of all called variants from the original vcf files. Each variant has 3 extra tracks per sample: (1) the original mapped sequence (bam file used in the mutation calling pipeline), (2) the filtered bams without duplicated or lower quality mapped reads, and (3) the bams of 'informative reads' yielded from the re-run of GATK HaplotypeCaller.

These 3 tracks were included per sample in each trio, i.e. for a full trio a total of nine bam tracks will be included in the report while for a surrogate-parent trio the bams of all siblings and the available parents would be included. The reports included a 10Kb window around each variant and also included the allele count (AD, in each family) and the quality of the genotype (QD, in the original call).

Each variant in the IGV reports was then visually inspected to determine possible errors in the mutation dataset of each trio (**Figure S3**). The variants that failed our test were then classified according to their problematic features.

- Read evidence in the parental genomes, undetected due to indel realignment
- Read evidence in the parental genomes, undetected due to other reasons
- Unconventional or nuanced mapping

- Polymorphism evidence (as presence in dbSNP), for families with surrogate parents
- Polymorphism evidence (as presence in dbSNP), for families with surrogate parents
- A cluster of mutations (≥ 6 mutations per 50 bp), some or all of which have rsID annotations (indicative of misclassified germline mutations due to the surrogate-calling method)

This manual curation resulted in the number of DNMs being reduced by ~36% per individual.

Read-backed phasing

The tool Unfazed (v1.0.3) (Belyeu et al. 2021), a read-based phasing approach, was used to phase the de novo variants to maternal or paternal haplotypes. This approach required the existence of an "informative" inherited heterozygous variant that could be phased to a parent present on the same sequencing read as the DNM. This requirement resulted in 14-42% of DNMs being phased per individual (**Table S1**), a fraction typical for studies of phased de novo mutations.

Generation and analysis of simulated trio data for surrogate-method benchmarking

We randomly generated five "children" of these parents by generating recombination events using the recombination segments (S_i) from a chromosome map (M) which was downloaded from the Beagle (2021) resource page (see https://bochet.gcc.biostat.washington.edu/beagle/genetic_maps/). We compiled the code to

simulate the recombination events and combine the datasets from the multisample VCFs here, www.github.com/davidmasp/meiosim. Specifically, we simulated the number of crossings (x) using a Poisson distribution with λ equal to the length of the segment, measured in centimorgans (Δc), multiplied by a recombination rate (R) of 0.01 (crossings/cM). We then obtained x crossings (K_i) from a uniform distribution covering the positions of the recombination segment:

$$M = \{S_i\}$$

$$S_i = (p_0, p_1), (c_0, c_1)$$

$$x_i = Pois(\lambda = \Delta c \cdot R)$$

$$K_i = unif_{\{p_o, p_1\}}(n = x_i)$$

An initial haplotype was chosen at random from each parent and the haplotype was then swapped at each recombination breakpoint. SNPs from the parental haplotypes were then propagated to the children.

We added DNMs to each simulated child by selecting a proband uniformly at random without replacement from Jónsson et al. (2017) and editing the simulated child's genome to include these DNMs.

We generated a short read BAM file consistent with each simulated genome using DWGSIM (0.1.15, www.github.com/nh13/DWGSIM), a tool that simulates sequencing reads from a reference genome and can incorporate custom germline variants. We defined an error rate of 0.001, a read length of 151 bp, and a target coverage of 30X. All other parameters were left as default. Other software used in this process were bcftools (1.19) and samtools (1.19).

We mapped and processed all simulated reads using SAREK (3.3.2) (Garcia et al. 2023). In brief, Sarek checks for quality and trims raw reads using fastqc (0.11.9) and fastp (0.23.4) (Chen et al. 2018); thenmaps with BWA-MEM1 (0.7.17-r1188) and further process them with Mark Duplicates and Base Quality Score Recalibration from the GA TK suite (4.4.0.0) (McKenna et al. 2010). In addition to that, it measures the coverage using mosdepth (0.3.3) (Pedersen and Quinlan 2018) and integrates the quality reports with multiqc (1.15) (Ewels et al. 2016).

We then employed the same pipeline used for real families to process the jointly genotyped VCF file and simulated BAM files of the family. This involved applying the same filtering criteria for DNMs and utilizing identical genomic accessibility masks on the BAM files to calculate the final accessible genome denominators for each surrogate individual.

In simulating the surrogate method, we designated “sibling 0” as the child and performed DNM calling using all possible combinations of siblings 1-4 as surrogate parents, with scenarios including both the presence and absence of the real mother (HG00132) using GA TK’s PossibleDeNovo. We then generated IBD tracks using hap-IBD to create surrogate tracts for all sibling pairs, in order to identify all regions where siblings shared a maternal or paternal IBD tract with sibling 0.

To minimize the inclusion of likely false positives, we applied the density filter (as detailed in the "Filtering" section and depicted in **Figure S1B**) to retain only mutations in "sparse" genomic regions. While we did not curate these mutation calls via IGV , we did apply two stringent filters

to remove false positive calls: 1) excluding any DNMs called in sibling 0 if another sibling had the alternate allele at that site, and 2) excluding any DNMs not called in all surrogate parent combinations which had the appropriate IBD segments to make that region of the genome accessible. We additionally classified false positive calls by identifying regions where a mutation was called in one surrogate trio combination but not in another, despite being accessible for detection. This distinction helped to account for limitations in the surrogate method and differentiate between true de novo mutations and likely false positives.

To assess how the inclusion of additional siblings as surrogates affected the method's recall and precision, we conducted tests in two configurations:

- Group 1) one surrogate parent acting as the father with the real mother included, utilizing up to four siblings in various combinations (one sibling + mother, two siblings + mother, three siblings + mother, four siblings + mother)
- Group 2) two surrogate parents with no real parent, considering three potential combinations of sibling surrogates (two siblings used, three siblings used, four siblings used)

In each scenario, we subset the callset of true DNMs based on the accessible genome provided by these sibling combinations, thereby differentiating between false negatives and inaccessible mutations. We also applied a stepwise false positive filter by sites at which two or more siblings shared an alternate allele.

With each additional sibling included in the surrogate combinations, we assessed mutations that were accessible across multiple surrogate call sets but were called in fewer than expected, thus further reducing false positives. The mutation rate at each step was calculated similarly to the method described in the "Accessible Genome Size Estimation" section, with the number of true and false positives serving as the numerator and two times the size of the accessible genome as the denominator.

Calculation of each family's expected mutation burden in the absence of a genetic mutator effect

To correct for differences between Jónsson et al. (2017)'s accessible genome size (2.68×10^9 bp) and the accessible genome sizes of each individual in our study (which ranged from 1.28×10^9 bp to 2.67×10^9 bp), we multiplied each expected mutation count under the parental age model by g_i divided by g_J, the ratio of the accessible genome of individual *i* (g_i) to Jónsson et al. (2017)'s accessible genome size (g_J). When the accessible genome size of an individual is considerably smaller than that of Jónsson et al. (as is the case for the individuals whose DNMs were called using the surrogate method), this rescaling will reduce the count of each mutation type we expect to observe in the offspring (**Figure S6**).

In order to determine whether the families in Sherwood et al. (2023) are consistent with the model trained on the families sequenced by Jónsson et al. (2017), we repeated the above procedure for the families in that study. Sherwood et al. (2023) didn't report each individual's accessible genome size, but since they did not employ the surrogate-calling method, their accessible genome size should be comparable to that of Jónsson et al. (2017), and so we did not

carry out accessible genome size rescaling for these individuals. For each individual sequenced in our study and the Sherwood et al. (2023) study, we computed the ratio of observed to expected mutation counts for each mutation type.

When carrying out comparisons based on the subset of mutations we were able to phase to maternal and paternal haplotypes, we further downscaled the expected mutation counts by the phasing success rate per individual, which ranged from 14-40% (**Table S1**).

The above calculations yielded estimates of the relative rate of each mutation type in families with pathogenic human *MUTYH* genotypes relative to control families. To compare these effect sizes to the effect sizes of murine *Mutyh* mutator alleles, we computed analogous observed-over-expected ratios using mice with different *Mutyh* genotypes previously analyzed by Sasani et al. (2022). To compute the average mutation rate of each mutation type *c* in mice with a mutagenic *Mutyh* genotype known as the "D" genotype, we added up mutations of type *c* from all mice with the "D" genotype and divided this count by the total number of generations these mice were inbred, which is the total number of generations over which they had the opportunity to accumulate mutations. In the same way, we estimated a relative rate of mutations of type *c* in mice with the "B" *Mutyh* haplotype. Finally, we estimated the rate of mutations of type *c* in a single strain known as BXD68 affected by a unique *Mutyh* hypermutator phenotype. For the "D" allele and the BXD68 hypermutator allele, we divided the relative rate of each mutation type by the "B" allele rate to estimate the effect size of each of these *Mutyh* variants on mutagenesis in the mouse germline.

Comparing our observed mutation counts to the null parental age model of Jónsson et al. (2017)

We used the Poisson cumulative distribution function (CDF) to determine whether the overall and per-mutation type DNM counts we observe are consistent with the parental age model, or whether we see significant elevations of any mutation type, particularly the C>A type associated with a defective MUTYH protein.

For each individual, we calculated $P(X \geq k \mid \lambda)$: the probability that a Poisson random variable X will generate a value greater than or equal to our observed mutation count k, given that it has mean λ equal to the expected count calculated based on the parental age model regressions from Jónsson et al. (2017) (as described above). We used *R*'s ppois() Poisson CDF function to calculate this probability. The ppois() function with the "lower.tail = F" flag gives the probability $P(X > k \mid \lambda)$, and we calculated that $P(X \geq k \mid \lambda) = P(X > k\text{-}1 \mid \lambda)$, such that

$P(X \geq k \mid \lambda)$ = ppois(*q*= (ObservedMutationCount - 1), lambda = ExpectedMutationCount, lower.tail=F)

This approach was used to determine whether the total observed mutation counts per individual were significantly greater than what we'd expect under the null parental age model expectation. We separately carried out this analysis for each mutation type (C>A, C>G, C>T, A>G, A>T, A>C) per individual, per nuclear family, and for mutation counts phased to each parent (total counts and per-mutation type counts).

Estimating the minimum mutator effect sizes that we have power to detect

For each biallelic parent whose offspring might be affected by a C>A mutator phenotype, we calculated the minimum C>A mutator effect size that should be statistically detectable using the above one-tailed Poisson test (leading us to reject the parental age model from Jónsson et al. 2017). To calculate this minimum effect size, we used the qpois() function in *R* to calculate the number of C>A mutations that should yield a p-value < 0.05, with λ estimated from the parental age model:

qpois(p= 0.05, λ= parental age model expected C>A count, lower.tail = F).

We then added +1 to the mutation count given by qpois() to calculate the number of mutations needed to be observed (x) such that $P(X \geq x \mid \lambda) < 0.05$. We call this number of mutations the "mutator detection threshold." We calculated separate thresholds for each child of a biallelic parent (including C11, C12, C21, C22, C23, C31, C32) and also calculated a cumulative threshold for detecting an elevated C>A mutation rate in each family with a biallelic parent (Families 1, 2 and 3). The detection threshold varies slightly across individuals and families based on parental age, the sex of the biallelic parent, and the childrens' total accessible genome size.

To estimate the minimum biallelic *MUTYH* allele effect size we should be powered to detect, we assigned all excess C>A mutations above the parental age model's expectations to the carrier parent:

$$\hat{x}_{C>A,\,CP} = x_{C>A} - E_{C>A,\,NCP}$$

where $x_{C>A}$ is the mutator detection threshold (minimum number of mutations for which $P(X \geq x \mid \lambda) < 0.05$), $E_{C>A,\,NCP}$ is the C>A count expected for the non-carrier parent (NCP) under the parental age model, and $\hat{x}_{C>A,\,CP}$ is the contribution of C>A mutations from the carrier parent (CP) needed to reach the significance threshold x, assuming all excess C>A above the parental age model expectation are assigned to the carrier parent.

The minimum detectable effect size of the biallelic *MUTYH* genotype should then be:

$$\frac{\hat{x}_{C>A,\,CP}}{E_{C>A,\,CP}}$$

where $E_{C>A,\,CP}$ is the expected number of C>A mutations contributed by the carrier parent under the parental age model.

We can also use this framework to estimate the effect size of the C>A mutator phenotype in the germline of each biallelic parent, again making the assumption that all excess C>A mutation counts above the parental age model expectation can be assigned to the carrier parent:

$$\hat{O}_{C>A,\,CP} = O_{C>A,\,total} - E_{C>A,\,NCP}$$

where $O_{C>A,\,total}$ is the total observed C>A mutation count in an individual child or set of children of the same biallelic parent. As before, $E_{C>A,\,NCP}$ is the expected number of C>A mutations contributed by the non-carrier parent under the parental age model, and $\hat{O}_{C>A,\,CP}$ is the estimate of how many C>A mutations are contributed by the carrier parent, assuming all excess C>A

mutations are assigned to that parent. The *MUTYH* effect size required to yield this number of mutations is then:

$$\frac{\widehat{O}_{C>A,\,CP}}{E_{C>A,\,CP}}$$

where $E_{C>A,\,CP}$ is the expected number of C>A mutations contributed by the carrier parent under the parental age model.

Mutational Signature Analysis

Non-negative matrix (NMF) factorization was used to extract mutational signatures from the de novo 3-mer mutation spectra, either per-individual, or summed up per-family. *SigProfilerExtractorR* (v. 1.1.16), an R wrapper for *SigProfilerExtractor* (Islam et al. 2022), was used to carry out the analyses. The reference genome was set to "GRCh38" and 100 NMF replicates were used. A range of signature numbers were explored, ranging from 1-10 for the per-individual analysis, and 1-3 for the per-family analysis (above 3 there were too many signatures for the number of input samples when individuals were grouped per family). The optimal solution that maximizes stability while minimizing cosine similarity was chosen by the software: for each analysis (per-individual and per-family), one signature was chosen as the optimal solution.

The cosine similarity between the optimal reconstructed mutation spectra and the empirical data ranged from 0.563--0.821 in the per-individual analysis, from 0.803-0.882 in the per-family analysis.

The optimal single signature in each analysis was deconvoluted by SigProfilerExtractor into contributions from known COSMIC (Catalogue of Somatic Mutations in Cancer) signatures. In each case, the extracted signature was deconvoluted into signatures SBS1 and SBS5, two clock-like signatures that generally make up the bulk of mutations in both germline and somatic data. No contributions of SBS18 or SBS36, somatic mutational signatures associated with defective *MUTYH*, were detected.

Chapter Two:

Influence of DNA Repair Gene Variations on Somatic Mutagenesis in Murine Models

All analyses done in collaboration with David Mas-Ponte, Jeanne Fredrickson, Brendan Kohrn, Suheeta Roy, Robert Williams, David Ashbrook, Rosana Risques, and Kelley Harris.

Results

Somatic Mutation Burden Analysis in BXD Mice

After establishing that pathogenic *MUTYH* genotypes confer a germline mutator phenotype in humans, we next investigated whether mice carrying a different germline mutator allele of *Mutyh* exhibit a somatic mutator phenotype. Tissue samples were collected from a study on the effects of diet and genotype on murine longevity (Roy et al., 2021). Each mouse was necropsied after dying of natural causes, with half of the mice fed a high-fat diet and the remainder standard chow (applied only to spleen samples; all colon samples were from chow-fed mice). These experimental conditions were chosen to capture the combined effects of diet, genotype, and age on somatic mutation accumulation.

Colon and spleen tissues were chosen for their distinct biological properties to evaluate the mutational signature of *Mutyh* deficiency in aged BXD mice. Previous studies, such as Robinson et al. (2022), reported elevated C>A mutation loads in these tissues, highlighting them as putative candidate tissues to test in the mouse model. Further, colon tissue, with moderate epithelial turnover and significant exposure to dietary and microbiome-derived genotoxins including concentrated OG damage (Tudek and Speina 2012; Obtułowicz et al. 2010), offers a way to specifically hone in on *Mutyh*-deficient mutational processes. Unlike studies such as Cagan et al. (2022), which sequenced isolated colonic crypts from mice under uniform dietary conditions, our study reflects naturally aged mice under diverse dietary exposures, offering a complementary view of *Mutyh*-associated mutagenesis.

Mutation frequencies per sequenced base pair were compared using the Wilcoxon rank-sum test with Bonferroni correction for multiple testing (**Figure 1A**). Significant differences were observed for C>A, C>T, and T>C substitutions, with "D" allele carriers showing elevated rates across both tissues. For colon tissue, "D" mice exhibited significantly higher C>A ($p < 0.005$), C>T ($p < 0.005$), and T>C ($p < 0.001$) mutation frequencies. Similarly, spleen tissue in "D" mice showed significantly elevated C>A ($p < 0.001$) and C>T ($p < 0.005$) frequencies, while "B" mice had significantly higher C>G mutation frequencies ($p < 0.001$). Other mutation types, including T>A and T>G, showed no significant differences, potentially reflecting compensatory DNA repair mechanisms.

To further explore differences in mutation patterns across *Mutyh* genotypes, chi-squared tests were conducted to compare mutation frequencies between B and D haplotypes, stratified by "young" (<800 days old) and "old" (>800 days old) age groups and tissue type. In young mice, no significant differences in mutation frequencies were observed between B and D haplotypes ($\chi^2 = 2.27$, $p = 0.132$). However, in aged mice, D haplotype carriers exhibited significantly higher C>A mutation frequencies than B carriers ($\chi^2 = 239.90$, $p < 0.001$). In contrast, spleen tissues showed no significant differences between haplotypes ($\chi^2 = 0.02$, $p = 0.881$), indicating that factors such as variations in cellular turnover, immune activity, and exposure to environmental genotoxins may also modulate the somatic mutational effects of *Mutyh* deficiency, or that our study is underpowered to detect a significant difference in spleen tissues specifically.

.

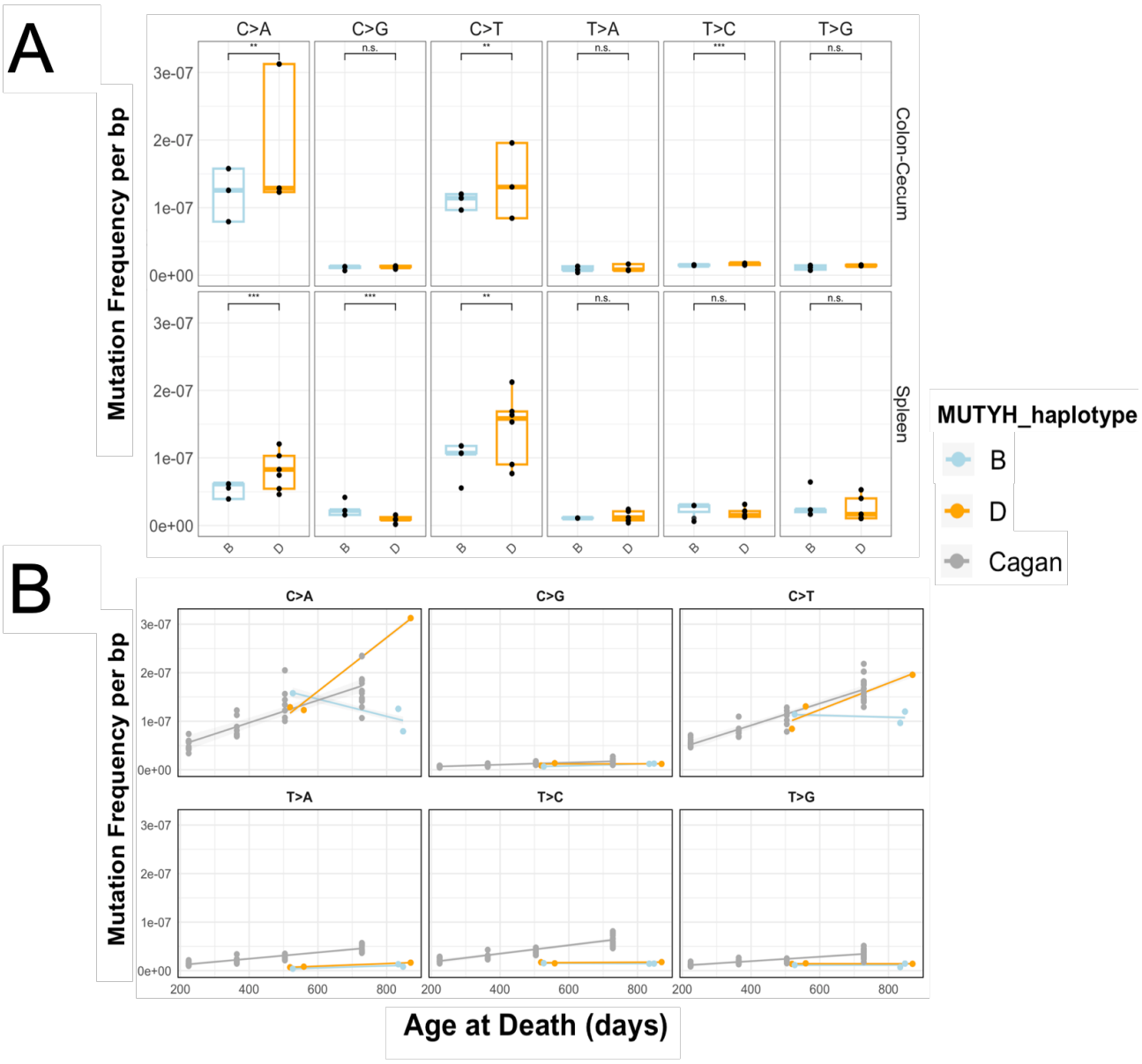


Figure 1. Mutation burden analysis across colon and spleen tissues in BXD mice with the "B" or "D" *Mutyh* allele. A) Mutation frequencies in colon and spleen tissues categorized into six pyrimidine-centered substitution classes (C>A, C>G, C>T, T>A, T>C, T>G) for mice carrying the "B" or "D" allele of the *Mutyh* gene. The x-axis indicates the substitution class, and the y-axis represents mutation frequency (mutations per base pair). Statistical significance between haplotypes for each substitution class was evaluated using a Wilcoxon rank-sum test with false discovery rate (FDR) correction to account for multiple testing. Significance levels are indicated as: *** ($p < 0.0001$), ** ($p < 0.001$), * ($p < 0.001$), and "n.s." (not significant). Orange bars indicate the "D" haplotype, while blue bars represent the "B" haplotype. One sample (BXD34-288) was excluded due to low-quality variant calls. **B)** Mutation frequencies in colon tissues as a function of mouse age at death (in days), stratified by substitution class and *Mutyh* haplotype. Mutation frequency per sequenced base pair (y-axis) is plotted against age at death (x-axis) for both "B" and "D" haplotypes. To calculate mutation frequencies, we aggregated depth-per-base values from duplex

sequencing data for each pyrimidine-centered context (e.g., all C>A sites), summed across the genome. We then divided the observed mutation counts by the total DP values for each sample, producing per-base-pair mutation frequencies (see more in Methods). Data for BXD mice carrying the "B" (blue) or "D" (orange) *Mutyh* allele are plotted alongside reference data (grey) from colon tissue in the Cagan et al. (2022) study.

While no significant differences in overall mutation burden were observed between "D" and "B" allele carriers (**Figure S1**), stratification by mutation type revealed consistently higher frequencies of C>A and C>T substitutions in "D" mice (**Figure 1A**). In colon tissues, this increase was largely driven by an outlier "D" mouse, strain BXD102, which exhibited markedly higher mutation frequencies compared to other samples (outlier indicated in **Figure S1**). Across tissues, colon samples showed slightly higher overall mutation burdens than spleen, irrespective of genotype. These findings suggest that the "D" allele broadly increases somatic mutation rates for C>A and C>T substitutions across tissues, independent of overall mutation burden or tissue type.

To place these findings in the context of published data, mutation frequencies were analyzed as a function of age at death and compared to colon tissue mutation frequencies from 43 C57BL/6-Ly5.1 laboratory mouse colonic crypts reported by Cagan et al. (2022) (12 individuals total; aged 0.5 to 2 years; **Figure 1B**; **Figure S2**). In contrast, our dataset included six bulk colon tissue samples from six BXD mice aged 519 to 870 days, with sampling occurring at the end of each mouse's natural lifespan. This difference in sampling strategy is notable, as mice that live longer might accumulate mutations at slower rates than those that die younger, potentially influencing mutation frequency trends.

To ensure comparability, mutation frequencies were normalized differently in the two datasets due to differences in sequencing methods. For our duplex sequencing data, mutation frequencies were calculated by dividing mutation counts by the total depth of sequencing for each pyrimidine-centered nucleotide context (e.g., all C>A sites) within each sample. This approach accounts for variations in sequencing coverage across samples. For the Cagan dataset, mutation frequencies were recalculated by dividing mutation counts by the number of accessible bases reported for whole-genome sequencing of colonic crypts. This slight methodological difference reflects the distinction between bulk tissue sequencing in our study and isolated crypt sequencing in the Cagan study.

In BXD colon tissues, "B" allele carriers showed no age-dependent increase in mutation frequencies, while "D" allele carriers exhibited age-related increases in somatic mutation loads for certain mutation types (**Figure 1B)**. Specifically, C>A and C>T substitutions in "D" mice followed trends consistent with the Cagan dataset, which reported age-dependent increases in somatic mutation frequencies. For other substitution types, mutation frequencies in BXD mice either declined or remained static with age, diverging from Cagan's findings. These differences could reflect the combined influence of sampling strategies, smaller sample sizes in our dataset (n = 3 per genotype group), and dietary differences, as Cagan et al. used a uniform diet while our study involved mice fed either a chow or high-fat diet. Despite these discrepancies, the overall mutation rates in BXD colon tissues were of the same order of magnitude as those reported by Cagan et al., indicating general consistency in age-related mutational burdens between datasets.

To further investigate deviations from the Cagan dataset, linear models were employed to estimate the effects of *Mutyh* genotypes on mutation frequencies while controlling for age at death (**Figure S3**). This analysis revealed significant deviations ($p < 0.01$, FDR-corrected) for most substitution types in both "B" and "D" haplotype carriers. Notably, these deviations were predominantly negative, indicating lower mutation frequencies in BXD mice compared to Cagan's colon tissue data. However, a unique and significant positive deviation was observed for C>A substitutions in "D" mice, which exhibited increased mutation frequencies relative to the Cagan dataset. This finding highlights an age-related increase in C>A colon mutations in "D" mice, deviating from the broader trend of reduced mutation frequencies observed for other substitution types.

In spleen tissues (**Figure S2**), mutation frequencies did not show significant age-dependent trends, likely due to the narrower age range of mice selected for sequencing (735–856 days old). We specifically prioritized the oldest available mice for sequencing, rather than sampling across the full range of ages in our cohort, to maximize mutation counts. This selection may have limited the detection of subtle age-related effects. Mutation burdens in spleen tissues were generally lower than those in colon tissues (**Figure S1)**, consistent with expectations given the biological differences between these tissues. However, the mutation rates in spleen remained broadly comparable in magnitude to those reported in the Cagan (2022) colon dataset and other studies of mouse spleen tissues. This consistency suggests that our measurements of spleen somatic mutation burdens fall within expected ranges for mouse tissues, though direct comparisons between spleen and colon tissues should be interpreted cautiously given their distinct biological contexts.

Inference of Mutational Signature Activities in BXD Tissues

To elucidate the distinct mutational processes in spleen and colon tissues of wild-type-like "B" and *Mutyh*-deficient "D" BXD mice, we performed mutational signature analysis using SigProfiler Assignment with the COSMIC mm10 reference profiles. SigProfiler Assignment was chosen for its robust capability to deconvolute complex mutational patterns and its applicability in murine models. We focused on signatures with established relevance in murine models to ensure biological plausibility (Sasani et al., 2022; Beal et al., 2020; Maura et al., 2023; Minko et al., 2024).

The overall mutation spectra showed distinct clustering of spleen and colon samples, with notable differences between 'B' and 'D' mice in spleen tissues, particularly in the composition of mutation types across pyrimidine-centered trinucleotide substitutions (**Figure 2A and Figure 2C**; individual spectra and PCA shown in **Figure S4 – S6**). Signature SBS5, associated with age-related mutagenesis, was predominant in both genotypes, contributing an average of 28.0% of mutations in "B" mice and 37.8% in "D" mice (**Figure 2B**). Notably, SBS30, a mutational signature linked to base excision repair (BER) deficiencies due to *NTHL1* inactivation in humans, comprised 5.06% of mutations in "B" mice and 25.0% in "D" mice. This finding aligns with observations from Sasani et al. (2024) and (2022), where SBS30 was identified in a majority of BXD lines, with no significant differences between "B" and "D" mice reported. SBS36, indicative of DNA repair defects specifically associated with *MUTYH* deficiency in human cancers, was identified in a single "D" mouse (BX070), accounting for 16.2% of its

mutational profile and 3.24% overall in “D” spleen tissues. Conversely, SBS9, reflecting polymerase η activity during lymphoid cell replication, was substantially higher in “B” mice (58.4%) compared to “D” mice (27.2%). Signature SBS1, associated with the spontaneous deamination of methylated cytosines and considered “clock-like,” showed similar contributions in both genotypes, accounting for 8.54% in “B” and 6.74% in “D” mice. Individual mutational signature contributions for spleen tissues are illustrated in **Figure S5**, displaying noticeable variability across samples both within and across the two genotypes.

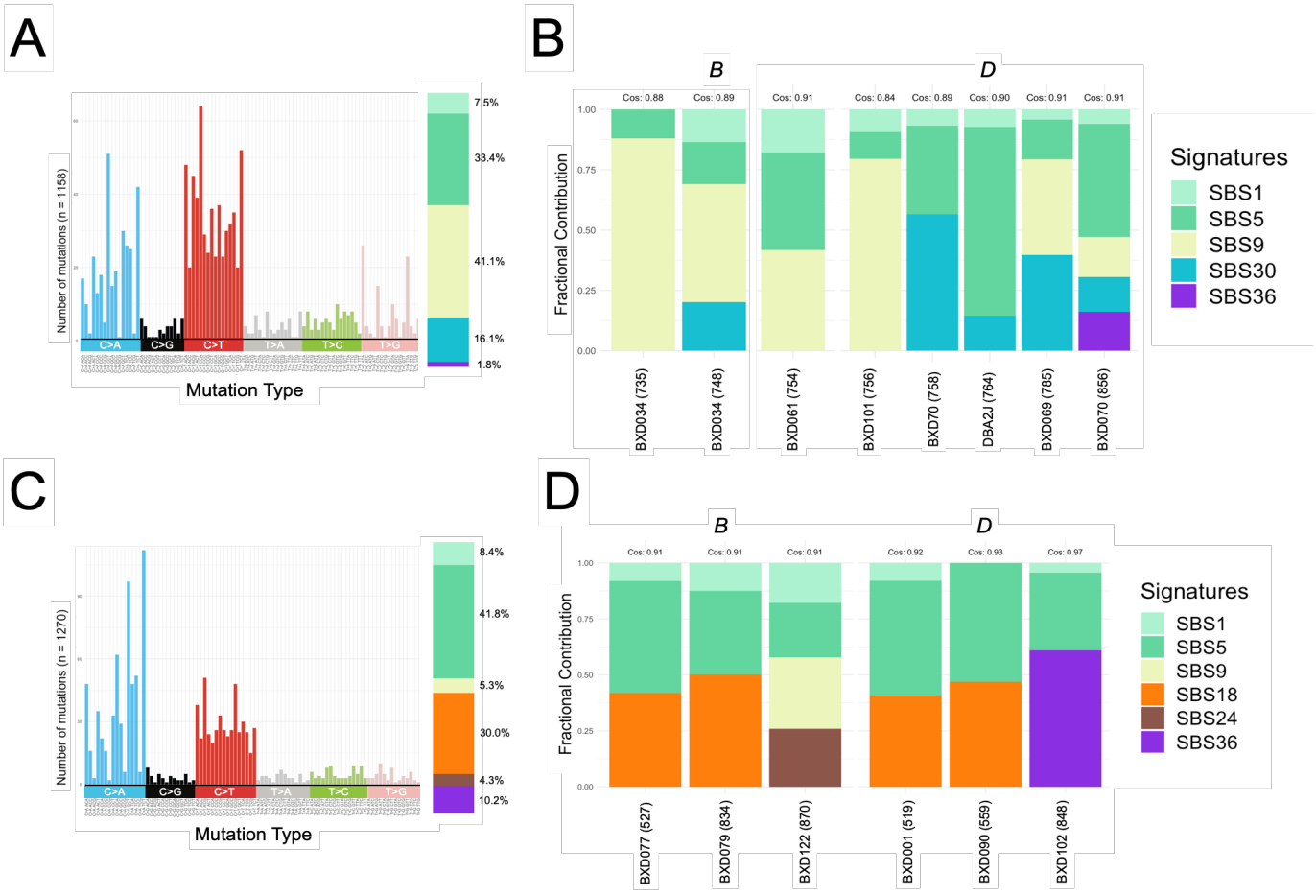


Figure 2. Mutational signature analysis in spleen and colon tissues of BXD mice carrying the "B" or "D" allele of the *Mutyh* locus. A) Mutational spectra in spleen tissues, showing the distribution of mutations across six pyrimidine-centered substitution classes. The x-axis represents the substitution class, while the y-axis shows the number of mutations observed. The

stacked bar on the right illustrates the proportional contributions of COSMIC mutational signatures (SBS1, SBS5, SBS9, SBS18, SBS24, SBS30, SBS36) to the overall mutation profile. Percentage contributions of each signature are labeled adjacent to their respective segments. **B)** Stacked bar plots depicting the fractional contributions of COSMIC mutational signatures in spleen tissues of "B" and "D" BXD mice. Each bar represents an individual sample, segmented by signatures as defined in the legend. The x-axis orders samples by haplotype and age at death (in days), while the y-axis denotes the fractional contribution of each signature. Cosine similarity values, indicating the quality of signature assignment, are annotated above each bar. **C)** Mutational spectra in colon tissues, displayed similarly to spleen tissues in panel A, showing mutation counts across substitution classes and the proportional contributions of the five COSMIC mutational signatures. **D)** Stacked bar plots illustrating the fractional contributions of COSMIC mutational signatures in colon tissues of "B" and "D" BXD mice. As in **B)**, the x-axis orders samples by haplotype and age at death, and the y-axis represents the fractional contributions of each signature. Cosine similarity values are annotated above each bar.

In colon tissues (**Figure 2C**), "D" mice exhibited higher contributions from SBS5 (46.3% versus 37.3%) and SBS36 (averaging 20.3%) compared to "B" mice (**Figure 2D**), suggesting enhanced age-related mutagenesis and specific DNA repair deficiencies associated with *Mutyh* loss. Conversely, "B" mice demonstrated higher levels of SBS1 (12.8% versus 4.08%) and uniquely harbored signatures SBS9 and SBS24 in a single individual (BXD122), where they contributed 31.9% and 26.0% of the mutational profile, respectively. SBS24 is associated with aflatoxin exposures commonly encountered in laboratory settings, such as mutagenic components present in mouse feed (Luzadder et al., 2024). The presence of SBS24 alongside SBS9 in an aged "B" mouse colon sample may reflect noise in mutational signature inference, or individual variability in DNA repair pathways or differential exposure to environmental mutagens.

Both genotypes showed similar contributions of SBS18, a signature linked to oxidative damage and defective BER, accounting for approximately 30% of mutations in both groups (30.7% in "B" and 29.2% in "D" mice). Interestingly, the elevated C>A mutation burden observed in the aged "D" mouse colons does not correspond to an increased load of SBS18. Instead, this difference appears to be driven by the significant C>A enrichment in one aged "D" mouse colon

sample, while elevated C>A mutations were absent in half of the aged "B" mouse colons sequenced. This suggests that the higher C>A load in the "D" samples may reflect stochastic variation or contributions from mutational processes not fully captured by SBS18. Further detail on mutational signature contributions for colon tissues at the individual level are presented in **Figure S6**.

Consistency of De Novo Mutational Signatures Across Tissues and Studies

To assess the consistency of mutational signatures across various tissues and studies, we compared our de novo extraction results with data from Cagan et al. (2022), Riva et al. (2020), and Chin et al. (2022). These studies were selected for their focus on diverse murine tissues and contexts, including intestinal crypts, multiple tumor types, and clonal hematopoiesis, providing a comprehensive framework for evaluating somatic mutations.

De novo extraction of mutational signatures using SigProfiler Extractor identified consistent profiles across colon samples, with mutational contributions stratified into four distinct signatures: SBS96A, SBS96B, SBS96C, and SBS96D (**Figure 3A**). Colon samples from this study demonstrated a dominant contribution of SBS96A, with smaller but notable contributions from SBS96B and SBS96C. This pattern aligns with Cagan et al. (2022), who reported similar signatures in colonic crypts of aged mice. Spleen samples exhibited a strong SBS96A presence along with SBS96C, mirroring the mutational profiles observed in lymphoid tissues from Chin et al. (2022) and Riva et al. (2020).

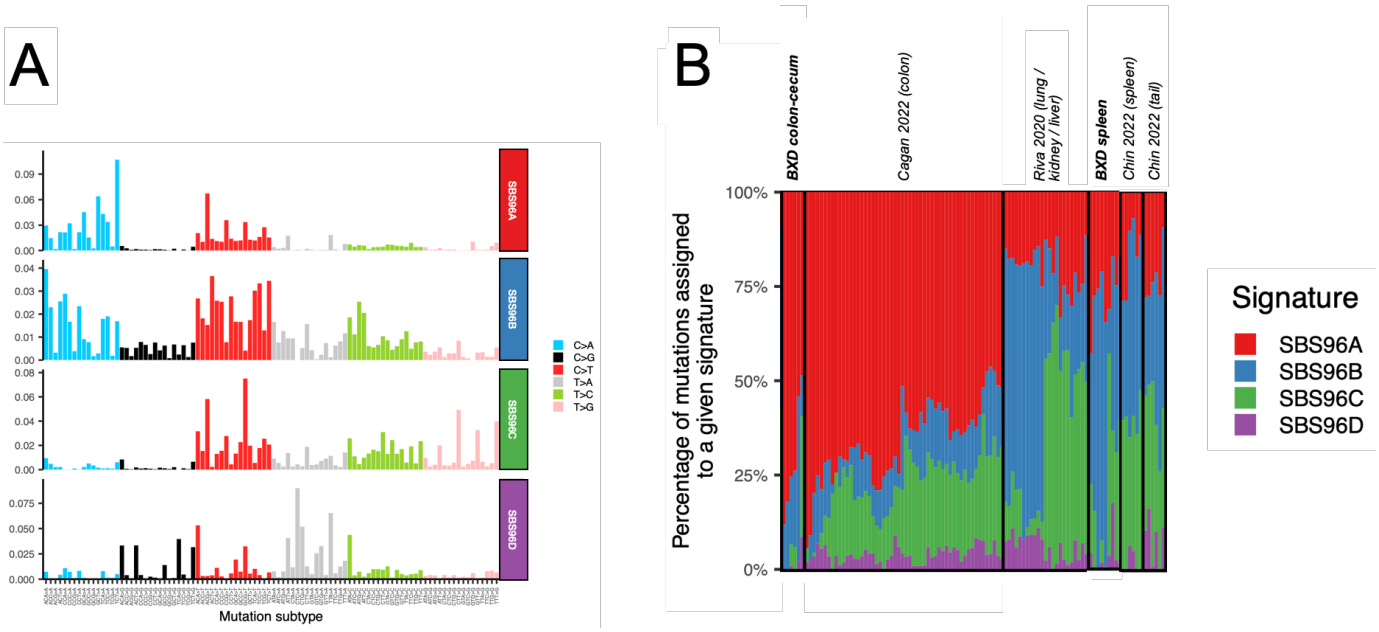


Figure 3. De novo extraction of mutational signatures and their contributions across tissues and datasets.
A) Mutational spectra of colon and spleen samples, stratified into four de novo extracted mutational signatures: SBS96A, SBS96B, SBS96C, and SBS96D. The x-axis represents mutation subtypes organized by trinucleotide context, while the y-axis indicates the proportion of mutations assigned to each subtype. Panels show the relative contributions of mutation subtypes for colon and spleen samples, highlighting tissue-specific patterns. **B)** Relative contributions of the four de novo mutational signatures (SBS96A–SBS96D) across samples from this study and previously published datasets. The x-axis displays individual samples grouped by study (e.g., Cagan et al., Chin et al., and Riva et al.), while the y-axis represents the percentage of mutations attributed to each signature. Distinct colors represent the contributions of SBS96A (red), SBS96B (blue), SBS96C (green), and SBS96D (purple), illustrating both tissue-specific and study-specific variations in mutational processes.

Figure 3B highlights the relative contributions of these signatures across our samples and the comparative studies. A decomposition of the data indicated that four mutational signatures provided the best fit for the overall dataset when whole-genome sequencing data from BXD mice was included. However, with the duplex sequencing data alone, the fit was more residual, reflecting the limitations of analyzing smaller-scale or targeted sequencing data. SBS96A, which shows high cosine similarity to SBS18 (add in exact number), was consistently prominent across colon samples from our study, Riva et al. (2020), and Cagan et al. (2022), reproducing earlier findings of higher SBS18 load in colon tissues compared to blood.

A separate de novo signature analysis conducted exclusively on the BXD mouse dataset (**Figure S7A**) identified two primary signatures: BXD-Sig1 and BXD-Sig2. BXD-Sig1 closely matched established signatures SBS18 and SBS5, demonstrating high similarity (cosine similarity of 0.925 and correlation of 0.864). BXD-Sig2 was predominantly composed of SBS40C with a minor SBS1 contribution, indicating moderate alignment (cosine similarity of 0.817 and correlation of 0.589). Notably, BXD-Sig1 was consistently present across colon, spleen, and other tissues, while BXD-Sig2 showed greater variation, highlighting the influence of tissue-specific factors on mutational processes.

Interestingly, sample BXD102, an exceptionally aged “D” mouse (848 days old), previously noted as an outlier for its elevated C>A mutation burden, also exhibited the highest overall contribution from SBS18 (**Figure S7B**). SBS18 is associated with oxidative damage, a process known to accumulate with aging, so the elevated SBS18 in BXD102 could reflect increased oxidative stress in this very aged individual. Additionally, SBS40C, a substitution signature of unknown etiology, was detected across nearly all samples, suggesting the presence of a ubiquitous, yet unidentified, mutational process in multiple mouse tissues.

High rates of Clonal Expansion in Cancer-Associated Genes in Spleen Tissues of "D" Allele Mice

Clonal expansions are defined as the outgrowth of genetically distinct cell populations within a tissue, often driven by mutations that confer a selective advantage. While clonal expansions are common in aging tissues and can reflect early stages of cancer, they frequently remain benign and do not progress to malignancy (Martincorena et al., 2015). Age-related clonal expansions have been well documented in human tissues, particularly in hematopoietic lineages, where mutations in genes involved in cell proliferation and genomic stability can promote expansion. Previous work has demonstrated the power of duplex sequencing to detect low-frequency clonal mutations in normal tissues, revealing that these expansions often accumulate with age and may serve as early markers of cancer risk (Salk et al., 2019; Kennedy et al., 2019).

Given that subfunctional human *MUTYH* alleles are strongly associated with increased cancer risk through defective base excision repair, we hypothesized that *Mutyh*-deficient "D" mice might exhibit higher rates of clonal expansions compared to "B" mice. This expectation is further supported by the role of the interdomain connector segment of *MUTYH* in the DNA damage response pathway (Raetz & David, 2020), suggesting that *Mutyh* mutations could compromise genomic stability beyond the germline.

To test this, we used duplex sequencing targeting a panel of cancer-associated homologs to detect clonal expansions in spleen tissues, focusing on mutations frequently implicated in cell proliferation and genomic stability. Using the homologs *Trp53*, *Pik3ca*, *Ctnnb1*, *Kras*, and *Nras*, duplex sequencing detected clonal expansions characterized by repeated mutations within hematopoietic lineages. These expansions appeared more frequent in "D" allele mice compared to "B" allele counterparts (**Figure 4**), though this difference was not statistically significant (χ^2

test, p = 0.214). Though *Trp53*-mutated clones were not significantly more frequent in "D" allele mice (χ^2 test, p = 0.73). with BXD69 displaying the highest burden, exceeding 1000 clones with mutations in *Trp53*, far surpassing any clonal activity observed in "B" allele samples.

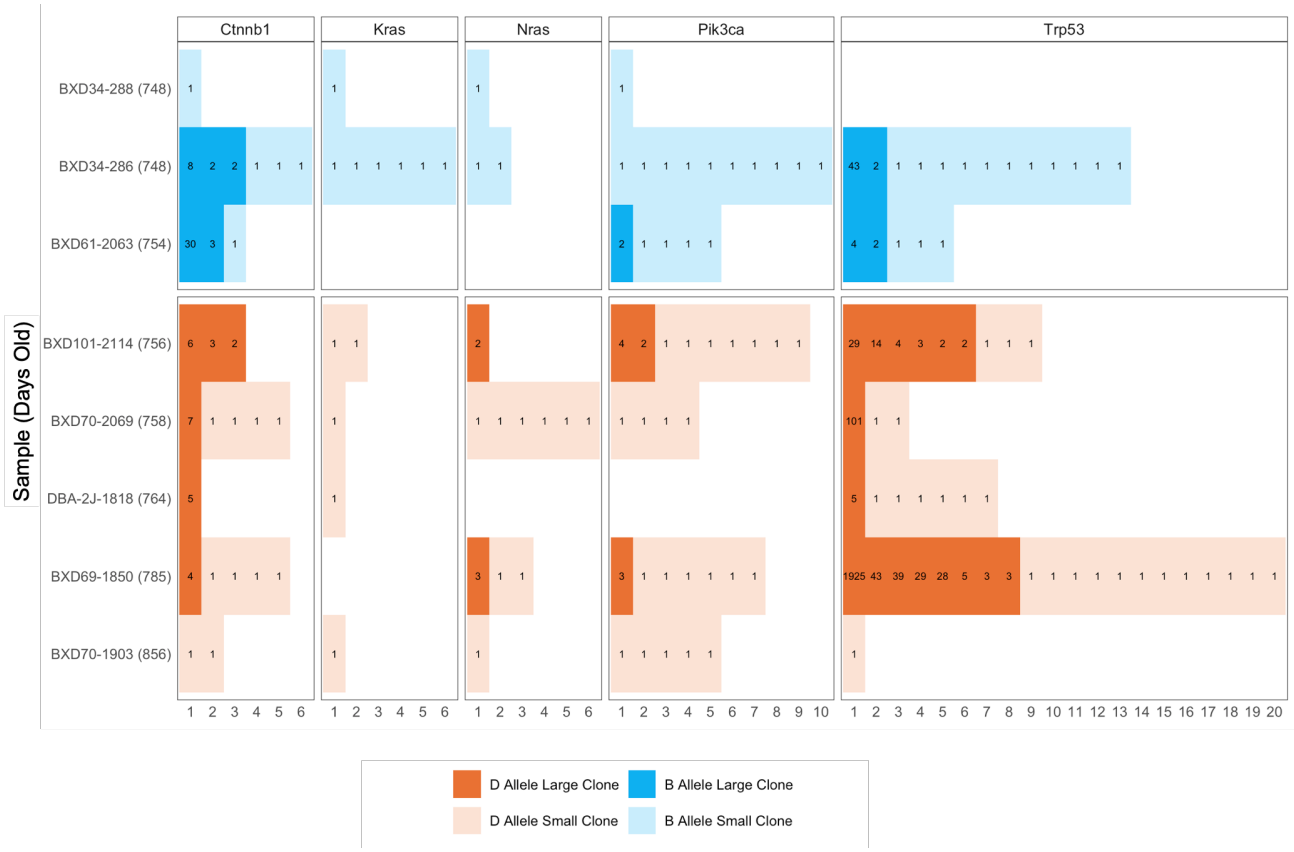


Figure 4. Clonal expansions in spleen tissues of BXD mice carrying the "B" or "D" allele of the *Mutyh* gene.
Clonal expansions were detected using a targeted panel of cancer-associated genes (*Ctnnb1*, *Kras*, *Nras*, *Pik3ca*, and *Trp53*) through duplex sequencing. Mutation counts across large and small clones for each gene. The x-axis represents individual genes, while the y-axis indicates the presence of clones in each strain categorized as large (bold shading) or small (light shading) for both "B" and "D" allele mice. The numbers inside of each rectangle indicate number of mutant duplex reads. Large clones are those with two or more mutant reads. Age of each mouse strain is reported in parentheses next to the strain name.

Mutations in *Ctnnb1*, a regulator of the *Wnt* signaling pathway involved in cell adhesion and transcriptional regulation, were identified in both large and small clones across "B" and "D" allele mice. One "B" allele mouse (BXD61) exhibited the largest expansion detected for this

gene, with 30 clones harboring mutations in *Ctnnb1*, while "D" allele mice exhibited fewer, smaller clones. Similarly, clonal expansions in the *Ras* family genes (*Kras* and *Nras*), which play critical roles in cell growth and differentiation signaling, were relatively rare and small across both genotypes. Slightly larger clones in these genes were observed in "D" allele mice compared to "B" allele mice. *Pik3ca*, a gene involved in the *Pi3k*/*Akt* signaling pathway that regulates cell growth and survival, exhibited more pronounced clonal expansions in "D" allele mice. In two samples (BXD69 and BXD101), multiple large clones with *Pik3ca* mutations were identified, while "B" allele mice exhibited only a few smaller clones for this gene.

No strong correlation was observed between age at death and clonal expansion size (**Figure S8**), likely due to the limited age range of the study cohort (735–856 days). While "D" allele mice were the only group to harbor clones exceeding 100 in count, this trend was not statistically significant (**Figure S9**). Additionally, no significant differences in clonal expansion counts were detected per individual gene between genotypes (**Figure S10**). These findings suggest that the observed differences in total clonal activity could reflect broader genomic instability in "D" allele mice or an increased propensity for certain mutations to drive cell proliferation, particularly in hematopoietic tissues. The interdomain connector segment of *MUTYH*, which plays a role in DNA damage signaling beyond base excision repair, may also contribute to this phenomenon by failing to suppress proliferation of damaged cells in *Mutyh*-deficient mice. Although the large clones in *Trp53* in "D" allele mice suggests a potential impact of *Mutyh* deficiency, additional sequenced samples are needed to determine the extent to which these differences are attributable to *Mutyh*-linked genomic instability, altered DNA damage responses, or other stochastic or biological factors.

Impact of Covariates on Mutation Burdens and Signatures

Mutational processes in mammalian genomes are known to be influenced by replication timing, with late-replicating regions consistently showing higher mutation rates across both germline and somatic tissues (Pope et al., 2014 ; Chen et al., 2017 ; Cornejo-Páramo et al., 2024). This relationship likely arises due to reduced DNA repair efficiency, lower chromatin accessibility, and increased mutational vulnerability in late-replicating regions. However, few studies have explored this association using duplex sequencing data, particularly in hematopoietic tissues such as the spleen. To fill this gap, we examined the relationship between replication timing and mutation rates in spleen samples from BXD mice, leveraging replication timing profiles derived from Repli-ChIP data of CH12.LX lymphoma cells (Pope et al., 2014).

We found that mutation rates were correlated with replication timing, with late-replicating regions exhibiting higher mutation frequencies compared to early-replicating regions across all substitution classes (**Figure 5**). This trend is consistent with findings from prior studies— for example, the genomic landscape of replication timing is broadly conserved across species, as shown in a recent study on great apes by Goldberg & Harris (2021), where late-replicating DNA was biased toward accumulating more C>A and A>T mutations. Similarly, Agarwal & Przeworski (2019) have demonstrated that late-replicating DNA in humans accumulates proportionally more C>A and A>T mutations compared to early-replicating regions, likely due to oxidative damage and repair inefficiencies in single-stranded DNA during replication.

In our data, derived from spleen samples and aligned with replication timing profiles from Repli-ChIP data of CH12.LX lymphoma cells (Pope et al., 2014), we observed this same trend across all mutation types, including C>A, C>G, C>T, T>A, T>C, and T>G substitutions. Notably, late-replicating regions were most enriched for mutations linked to spontaneous deamination of cytosines and oxidative damage, while early-replicating regions exhibited more uniform distributions of mutation types. The most prominent association was seen in C>T mutations, which showed a pronounced decrease in mutation frequency from late- to early-replicating regions. Similar, albeit weaker, associations were observed for other substitution types, such as C>A and T>C mutations, which showed a modest decline in frequency along the replication timing gradient.

To better characterize the relationship between replication timing and mutation frequencies, we further divided replication timing scores into quantile-based bins (early, mid-early, mid-late, and late). This binning analysis revealed a clear gradient of increasing mutation frequency as replication timing shifted from early to late bins. Notably, late-replicating bins exhibited a disproportionate enrichment of C>A and T>A mutations (**Figure S11**), consistent with previous observations of mutational biases linked to oxidative damage and single-stranded DNA repair inefficiencies. By leveraging this binning approach, we identified finer-scale patterns of mutation accumulation, demonstrating that the replication timing gradient modulates mutation burdens not only globally but also within specific genomic compartments.

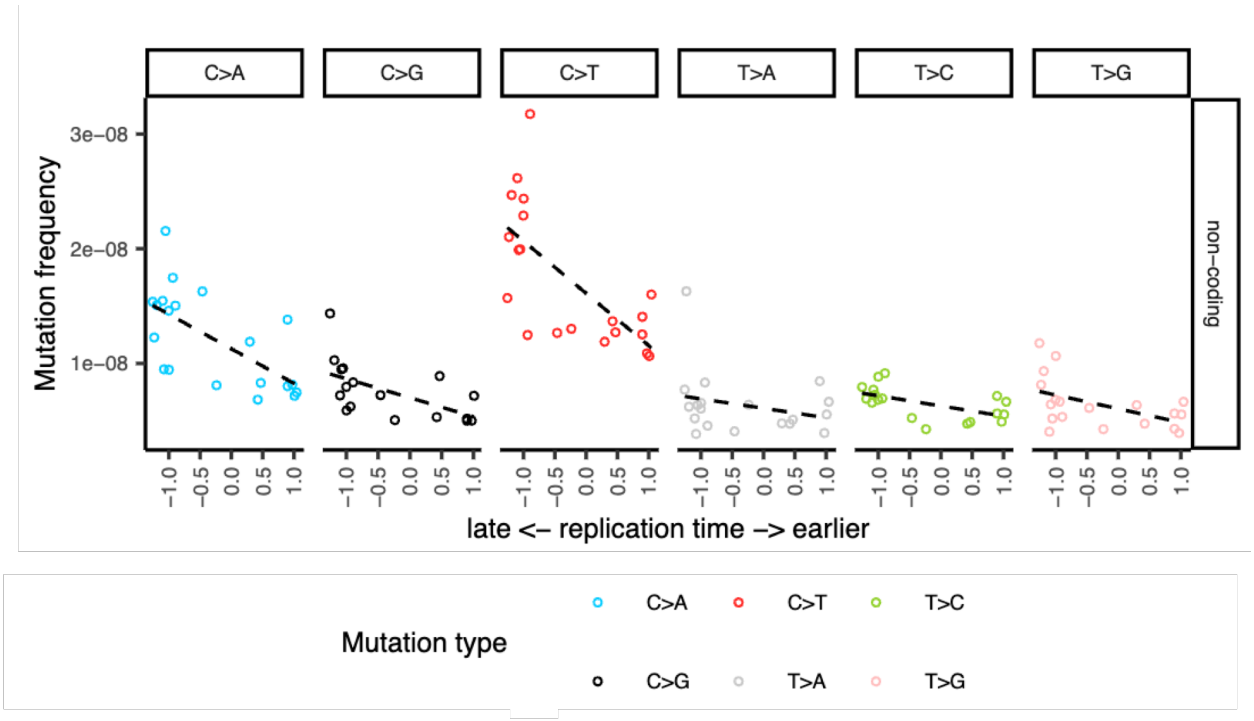


Figure 5. Correlation of mutation frequency with replication timing across substitution classes. Panels display mutation frequencies (y-axis) plotted against replication timing (x-axis), which ranges from late-replicating regions (left, -1.0) to early-replicating regions (right, 1.0). Each panel corresponds to one of the six pyrimidine-centered substitution classes. Mutation frequencies were derived from duplex sequencing data of spleen samples, while replication timing was determined using Repli-ChIP data from CH12.LX lymphoma cells. Dashed lines represent the linear trend for each substitution class, showing a general decrease in mutation frequency from late- to early-replicating regions**.**

Given that diet can influence cellular stress, inflammation, and metabolic processes linked to DNA damage and repair (Vermeij et al., 2016; Chen et al., 1999), we examined whether dietary differences could affect mutation burden and mutational signatures. In our study, some mice were fed a high-fat diet while others were maintained on a standard chow diet, reflecting the original design of the longevity study from which these samples were derived (Roy et al., 2021). We hypothesized that the high-fat diet might exacerbate oxidative stress, potentially leading to increased mutation burdens in metabolically active tissues such as the spleen. Additional analyses were conducted to assess the impact of diet on mutation burden and mutational signatures. Although the high-fat diet did not have a statistically significant effect on overall

mutation burden ($p = 0.256$; negative binomial test), a qualitative higher mutation burden associated with the high fat diet was observed in the spleen (**Figure S12**). However, only three mice on the high-fat diet exhibited clonal expansions in the sequenced gene panel, a frequency that was no higher than expected by chance based on chi-squared testing ($p = 0.453$). These findings warrant further investigation to clarify the interplay between diet, oxidative stress, and mutation accumulation in metabolically active tissues such as the spleen.

Materials and Methods

BXD Mouse Tissue Collection

Fifteen tissue samples were collected from aged BXD recombinant inbred mouse strains, a genetically diverse population well-suited for mutation analysis, at the University of Tennessee Health Sciences Center. Colon and spleen tissues were harvested to assess mutation rates across distinct somatic tissues. Spleens were excised in their entirety, promptly flash-frozen in liquid nitrogen, and stored at -80°C to preserve DNA integrity. Colon tissues were dissected to include the full length of the organ and were processed immediately in a fresh state. All sample handling adhered to standard preservation protocols, with tissues kept on ice during processing to maintain DNA quality for downstream analyses.

Colon Epithelial Cell Isolation and gDNA Extraction

Epithelial cells from the colon tissues of six BXD mice were isolated using an EDTA-based shake-off method, optimized to selectively detach epithelial cells while minimizing contamination from stromal and other cell types. Colon biopsies were incubated in a sterile 30 mM EDTA solution prepared in Hank's Balanced Salt Solution (HBSS) without divalent cations for approximately 35–45 minutes at room temperature, with gentle inversions to ensure even exposure. The epithelial layer was released from the underlying stroma through further gentle agitation and manually teased off using sterile instruments under a dissecting microscope.

Once isolated, epithelial sheets were microdissected into smaller fragments under sterile conditions using disposable needles treated with freezing media to prevent adhesion. Viable crypts were evaluated for quality, with optimal samples displaying well-formed crypt structures free from gelatinous stromal connections. Microdissected crypt fragments were transferred to microcentrifuge tubes prefilled with freezing media and stored on ice until genomic DNA (gDNA) extraction. The isolation and extraction protocols were rigorously controlled to ensure high yields of intact gDNA suitable for downstream library preparation and sequencing.

Spleens were collected from 15 BXD mice by excising the entire organ under sterile conditions immediately after dissection. Genomic DNA (gDNA) was later extracted from both spleen and colon tissues using the QIAGEN DNeasy Blood & Tissue Kit, following the manufacturer's protocol. Spleen tissues (up to 10 mg) and colon tissues were lysed in Buffer ATL with Proteinase K, with digestion performed at 56°C to ensure complete tissue breakdown. Following lysis, gDNA was purified through DNeasy Mini spin columns with sequential washes in Buffers AW1 and AW2. DNA was eluted in Buffer AE and stored at -20°C for downstream analyses. The DNA concentration was measured using a Qubit fluorometer to ensure adequate yield. DNA integrity was assessed using genomic TapeStation analysis to check the DNA Integrity Number (DIN). All steps were conducted under sterile conditions to preserve sample integrity.

Library Preparation and Sequencing

Library preparation for duplex sequencing was performed on DNA samples extracted from colon and spleen tissues. Starting with 1000 ng of genomic DNA input, we performed enzymatic

fragmentation prior to library preparation, following the standard TwinStrand Biosciences duplex sequencing protocol. Unique molecular identifiers (UMIs) were ligated to each DNA strand and its complement during library preparation to enable error correction through comparison of both DNA strands, resulting in highly accurate sequencing. DNA fragments were enzymatically end-repaired, adenylated, and ligated with UMIs, followed by amplification and purification.

To target DNA repair genes and mutation hotspots, a mutagenesis panel and homologous cancer gene probes provided by Dr. Scott Kennedy's lab were utilized (all regions captured are catalogued in **Table 1**). The mutagenesis panel is a hybrid selection assay designed for mice, covering approximately 48 kilobases of genomically representative sequences spread across nearly all autosomes. It targets intergenic regions optimized to exclude challenging sequences—such as repetitive elements, homopolymers, pseudogenes, and regions with extreme GC content—to facilitate accurate sequencing and alignment. By focusing on regions without known cancer-related loci, the panel minimizes potential biases due to positive or negative selective pressures, allowing for precise quantification of mutation frequency and spectrum. Using these mutagenesis and cancer probes, we performed hybridization capture on enriched colon and spleen DNA libraries to precisely target genomic regions of interest. The captured libraries were then amplified to produce sufficient material for sequencing.

Prepared libraries were sequenced on a high-throughput Illumina NovaSeq6000 platform to generate duplex reads. All steps were performed according to standard protocols to maintain fidelity and reproducibility.

BXD Variant Calling & Filtering Pipeline

The publicly available Duplex-Seq Pipeline (https://github.com/Kennedy-Lab-UW/Duplex-Seq-Pipeline), was used to process raw sequencing data into Variant Call Format (VCF) files. The pipeline was implemented using Snakemake for workflow management, with dependencies including Python3.6+, bwa (v0.7.17), and NCBI BLAST (v2.9.0). Genome preparation included downloading the mm10 reference genome from UCSC and indexing it with bwa, samtools, and Picard's CreateSequenceDictionary. Target regions for analysis were defined using BED files specific to a mutagenesis panel constructed by Twinstrand Biosciences (https://twinstrandbio.com/mutagenesis-assay/), and homologous murine cancer probes lent by Dr. Scott Kennedy's lab. A pre-configured contaminant database was used during data processing to identify and exclude non-target species reads.

Raw sequencing reads were preprocessed using Cutadapt to remove unique molecular identifiers (UMIs). Reads were then aligned to the reference genome using bwa mem, and low-quality alignments were filtered out based on mapping quality thresholds. Single-stranded consensus sequences (SSCS) were generated by grouping reads with matching UMIs and applying majority-rule base calling. Duplex consensus sequences (DCS) were subsequently constructed by comparing complementary SSCSs, enabling high-fidelity identification of true mutations.

Reads identified as originating from non-target species or with unresolved alignments were excluded. Filtered DCS reads were processed through variant calling using GATK tools and

custom scripts, with stringent quality filters, including depth thresholds of [insert final filter after verifying with Brendan], applied to distinguish true variants from sequencing artifacts.

The pipeline produced output files in multiple formats, including final DCS BAM files, mutation frequency summaries, and VCF files. Quality control metrics, such as insert size distributions, mutation rates per read cycle, and family size distributions, were also generated to validate data integrity. Comprehensive reports in both CSV and HTML formats summarized the results for all processed samples.

Following variant calling, an additional filtering pipeline was applied to isolate high-confidence somatic mutations. This pipeline involved filtering criteria based on base quality and allelic fraction thresholds, ensuring that only mutations with robust evidence were retained. Quality-filtered mutations were then subjected to downstream analyses to characterize the mutational landscape of the samples.

Mutational Signature Analysis

Filtered mutations from colon and spleen samples were analyzed to identify mutational signatures within 15 colon and spleen samples. Genomic mutation calls were loaded from pre-filtered variant datasets and processed to retain mutations within the targeted regions of the mutagenesis panel. Variants were annotated and categorized into trinucleotide mutation contexts by extracting the flanking bases and classifying mutations as single-base substitutions. Mutation frequencies were aggregated by trinucleotide context and sample.

External mutation datasets (Cagan et al., 2022; Chin et al. 2021) were processed in a similar manner. Single-nucleotide variant (SNV) data in tab-delimited format were standardized by aligning chromosome positions and reference/alternate alleles with the UCSC genome coordinate system. Variants were converted to genomic ranges for consistency across datasets. Mutation frequencies were then tabulated by trinucleotide context and merged with the duplex sequencing dataset.

All mutation datasets were integrated into a single table of mutation frequencies across 96 trinucleotide contexts. Mutation contexts were standardized, and data from multiple sources, including duplex sequencing and other variant-calling pipelines, were consolidated into a unified format suitable for mutational signature analysis. The final dataset was formatted and exported for use in signature fitting.

COSMIC reference signatures were used to fit the observed mutation data to known mutational profiles. Mutation frequencies across the 96 trinucleotide contexts were decomposed into contributions from specific mutational signatures using the SigProfiler framework. The resulting data were used for downstream analyses and comparisons between tissue types.

Comparative Analysis of Somatic Mutation Rates

To evaluate mutation burdens in BXD mice relative to published data, a workflow was implemented to calculate mutation rates and frequencies within the colon and spleen samples. This pipeline integrated genomic sequencing data, coverage metrics, and mutation counts to

normalize mutation frequencies per base pair, accounting for genome-wide coverage variations and nucleotide content. Mutation datasets were obtained from published studies on mouse mutation rates in the colon (Cagan et al. 2022) and blood (Chin et al. 2022). Each dataset included genomic annotations, analyzable genome sizes, coverage data, and sample ages. Variants were standardized to the UCSC genome coordinate system and converted into genomic ranges objects (VRanges). High-confidence single-nucleotide variants (SNVs) were retained for analysis. To ensure consistency, mutation data were cross-referenced with a precomputed callable genome dataset, excluding samples lacking coverage data.

For the BXD mouse samples, depth-of-coverage data were extracted from duplex consensus sequence (DCS) depth files generated during sequencing. Sample metadata—including strain, donor ID, age, and *Mutyh* genotype—were incorporated into the analysis. Depth data were filtered to exclude positions outside the targeted genomic regions, as defined by the mutagenesis panel BED file. Reference bases were simplified to pyrimidines (C and T) by complementing G to C and A to T, facilitating analysis focused on mutation types of interest.

Positions with low-quality reads, identified by a threshold number of ambiguous base calls (N bases), were removed to enhance data reliability. Total read depth and counts of ambiguous reads were aggregated per region, reference base, and sample. Mutation counts were determined by grouping variants according to trinucleotide context, mutation type, and sample. These counts were merged with corresponding coverage data to calculate mutation frequencies accurately.

Mutation frequencies were normalized by calculating the total number of nucleotides at risk (T and C bases) within the callable genome for each sample. The callable genome size was adjusted based on an assumed GC content of 40%, reflecting the proportion of cytosine and thymine bases. Normalization accounted for genome-wide coverage variations and the exclusion of ambiguous reads. Mutation frequencies were expressed as mutations per base pair, providing a standardized metric for comparison across samples. Confidence intervals for each mutation frequency were estimated using binomial exact tests.

Mutation counts and frequencies were categorized by mutation type (based on six pyrimidine-centered substitution classes), genomic region (coding versus non-coding), and sample attributes such as age and *Mutyh* genotype. Sample ages were standardized in days to facilitate temporal comparisons. Linear models were employed to assess relationships between mutation frequencies and variables such as age and replication timing. The replication timing data were integrated by associating genomic regions with replication timing scores, allowing for analysis of mutation rates in relation to early or late replication phases.

Statistical Modeling of Mutation Rates

Sample metadata, including age (standardized in days), tissue type (e.g., colon, spleen), diet, and *Mutyh* genotype, were integrated into the dataset. Outlier samples and those with insufficient data were excluded. Mutation rates per cell were calculated by scaling counts based on the diploid genome size of mice (approximately 2.7×10^9 base pairs).

Generalized linear models (GLMs) were employed to assess the effects of explanatory variables on mutation counts. The following models were considered:

- Poisson Regression: Used for counting data under the assumption of equidispersion, tested with both log and identity link functions.
- Negative Binomial Regression: Applied to account for overdispersion in count data when variance exceeds the mean.
- Linear Regression: Used when modeling mutation rates as continuous outcomes.

For each mutation type, models were specified with mutation counts as the response variable and age, tissue type, diet, and *Mutyh* genotype as predictors. An offset term based on the logarithm of the number of nucleotides at risk was included to adjust for varying coverage across samples. Interaction terms between predictors were explored but included only if they improved model fit.

Model performance was evaluated using:

- Akaike Information Criterion (AIC): To compare models, with lower AIC values indicating a better fit.
- Pseudo R-squared (Efron's R^2): To measure the proportion of variance explained by the model.
- Confidence Intervals: Calculated for model coefficients to assess statistical significance.
- Likelihood Ratio Tests and ANOVA: Performed to compare nested models and determine the contribution of additional predictors.

Diagnostic plots and residual analyses were conducted to assess model assumptions and identify potential outliers. Further, a simulation study was conducted to evaluate the properties of different statistical models. Synthetic datasets were generated by simulating mutation counts as Poisson-distributed random variables with rates dependent on age. Various models—including Poisson regression with log and identity link functions, negative binomial regression, and Gaussian models—were fitted to the simulated data. Model performance was compared using AIC and pseudo R-squared values to assess the impact of model choice.

Discussion

The findings in this section build on the investigations presented in **Chapters 1** and **2** of this dissertation. In **Chapter 1**, we demonstrated that pathogenic *MUTYH* variants in humans lead to elevated germline mutation rates, providing evidence for their significant impact on genomic stability in the germline. **Chapter 2** extended this work by examining the influence of the *Mutyh* genotype on mutational processes in non-cancerous tissues, using data from two tissues across 15 BXD mouse strains. Together, these chapters illustrate how *MUTYH*-related DNA repair deficiencies affect both germline and somatic mutation processes, highlighting the interplay between genetic background and cellular context in shaping mutation landscapes.

Further analyzing our findings, in **Chapter 1** we investigated the germline mutation rate and spectrum within a large extended family affected by a *MUTYH* genotype, p.Y179C/V234M, consisting of a relatively common pathogenic variant plus a rarer variant with conflicting interpretations. This family's history of colon cancer previously suggested that the p.Y179C/V234M genotype had a pathogenic effect, and we were able to use a cell-based *in vitro* functional assay to classify p.V234M as a partial loss of function variant. By calling de novo mutations in the children of two mothers with the p.Y179C/V234M genotype, we documented a modest but significant maternal mutator effect that appears weaker than the maternal germline mutator effect recently discovered in the children of two mothers with the more common MAP-associated genotype p.Y179C/G368D (Sherwood et al. 2023). A complementary analysis of parental age dependence of the BXD mouse *Mutyh* mutator effect confirms that the mutator is unlikely to act on the paternal germline and is most likely to increase the mutation rate during

embryonic development, similarly to a maternal mutator allele that Stendahl, et al. (2023) recently discovered in rhesus macaques.

Even in a pedigree as large as the one we study here, DNM data sparsity limits the power to estimate precise mutator effect sizes. Based on prior knowledge about the biology of *MUTYH*, we expected to see excess germline C>A mutations in the children of biallelic carriers, and though our data appear to support this hypothesis, the observed C>A enrichments are likely not extreme enough to survive a stringent Bonferroni correction for the number of distinct tests performed throughout the manuscript, let alone an agnostic scan for mutators affecting other mutation types. We did not attempt to formulate a less conservative multiple test correction by estimating the number of truly independent tests being performed, which would have been challenging to do given the nested nature of testing both individuals and larger nuclear families for the same mutator effect. To give readers an accurate sense of data heterogeneity and noise, we perform more tests than the minimum number required, computing C>A enrichments individual by individual and observing nominally significant enrichments in only a few children ($p<0.05$ in a one-tailed test without multiple testing correction).

To our knowledge, this study is the first to call DNMs in the children of a father with a biallelic *MUTYH* genotype. Since about three-fourths of human variation arises in the paternal germline, we expected to have more power to measure a germline mutator effect in this family compared to families with maternal *MUTYH* variation. We were thus quite surprised that this father was the only biallelic parent whose children did not have a significantly elevated C>A mutation load, suggesting that *MUTYH* variation has a proportionally weaker effect on the paternal germline.

This result should be interpreted with caution given our small sample sizes, but it could indicate that oxidative stress causes a smaller proportion of mutations in spermatocytes compared to oocytes or the developing embryo, or else that spermatocytes rely more on DNA repair pathways not involving *MUTYH*. Although it is possible that the one biallelic father in our dataset is an outlier, our mouse analysis also fails to detect an effect of *Mutyh* variation on the spermatocytes (**Figure 6C**), which would be expected to increase the parental age dependence of the C>A mutation rate, and points to the embryo as the most likely site of the mutator effect.

Maternal biallelic *MUTYH* genotypes could plausibly increase the rate of germline C>A mutations during the first few embryonic cell divisions prior to the maternal-zygotic transition; maternal DNA repair machinery is responsible for repairing embryonic DNA damage prior to the activation of zygotic transcription (Huang et al. 2014; Harland et al. 2017). Early embryonic mutations are enriched for C>A, possibly due to 8-oxoguanine damage that occurs during oocyte and spermatocyte maturation, and maternal *MUTYH* mutations may interfere with the repair of such damage (Ohno et al. 2014; Smith et al. 2013b; Gao et al. 2019). Advanced maternal age appears to increase the rate of C>A mutations occurring on the paternal haplotype of the embryo (Gao et al. 2019), and biallelic maternal *MUTYH* mutations might cause a similar attenuation of 8-oxoguanine repair during the earliest stage of development.

One possibility is that 8-oxoguanine lesions cause similar absolute numbers of mutations per generation in males and females, but that the excess male mutation load is caused by factors unrelated to oxidative stress, which would seemingly contradict the widespread assertion that oxidative stress is a major cause of DNA damage in aging sperm (Aitken et al. 2003; Aitken

2020; Aitken and Krausz 2001). Further study of germline mutagenesis in families with paternal *MUTYH* mutations may thus shed light on the etiology of germline mutagenesis in males with normal *MUTYH* genotypes, helping us better understand whether oxidative stress is truly to blame for age-related infertility and the genetic disorders associated with paternal age. Our results suggest that 8-oxoguanine lesions may beget a larger fraction of oocyte mutations, making oxidative stress a notable contributor to reproductive decline in the general female population.

Because germline mutator phenotypes appear to be rare, at least at the current limits of our ability to detect them, these phenotypes have often been measured in the offspring of just one carrier parent, leaving us no information about whether these phenotypes are sex-specific. The mutator phenotypes recently measured by Kaplanis et al. (2022) were mostly found to affect male parents, and a study of an extended family affected by a DNA polymerase delta mutator definitively measured a stronger effect in male carriers compared to female carriers (Andrianova et al. 2023). Though the *MBD4* mutator allele that was recently discovered in rhesus macaques clearly exerts a maternal effect, the absence of a male breeder with the same phenotype precluded estimation of the relative strength of the corresponding male mutator phenotype (Stendahl et al. 2023). Pedigree studies like ours and the work of Andrianova et al. (2023) will likely be instrumental for further study of possible sex differences affecting mutagenesis and DNA repair.

A technical innovation that improved the power of this study was new methodology for calling DNMs in incomplete nuclear families, with siblings acting as surrogate parents. Given our goal

of calling DNMs in the children of individuals with rare pathogenic *MUTYH* genotypes, we were able to maximize our pool of study subjects by relaxing the usual restriction to calling DNMs only in children whose parents 'genomes were both available for sequencing. Our surrogate parent approach does have drawbacks compared to DNM calling in complete nuclear families, most notably the restriction of the callable genome to regions where the proband inherited the same missing parental haplotypes as at least one sequenced sibling. Surrogate-based DNM calling is also susceptible to false positives caused by gene conversion and errors in IBD calling, and we note that these false positives will look cleanly mapped upon visual inspection of sequencing reads. The filters we employed to control these errors may have increased our false negative rates, particularly when using both a surrogate mother and a surrogate father. Conversely, our false positive rates appear to be elevated when only a single sibling is available as a surrogate parent and sib-sharing cannot be used to identify false positive DNMs that were actually inherited from the missing parent.

One new technology that will likely improve the performance of the surrogate method in the future is high-fidelity long read sequencing, which enables nearly all mutations to be phased to their maternal or paternal haplotype of origin (Noyes et al. 2022; Porubsky et al. 2024). When comparing the genomes of siblings who were sequenced using PacBio HiFi or a similar technology, it will become very straightforward to determine whether any read from the proband is derived from the same haplotype as a read present in a parent or surrogate parent, which will limit the ability of inherited variants to masquerade as DNMs. Surrogate-based mutation calling has the potential to make DNM analysis accessible to the many families where parents are

deceased or not in contact with their children, including families affected by rare or undiagnosed genetic diseases where DNM calling is likely to have the greatest scientific and clinical utility.

Our findings from **Chapter 1** suggest that the germline mutator effect of *MUTYH* predominantly operates in a recessive manner, paralleling its role in cancer predisposition. However, we note that all available data on C>A mutation rates in normal human cells is derived from individuals who have at least one loss of function allele (p.Y179C). Although our study and previous studies (Sherwood et al. 2023) find mutagenesis and cancer risk to be associated with biallelic genotypes that combine p.Y179C with a partial loss-of-function allele (p.V234M or p.G368D), we do not have similar data from biallelic genotypes that combine two partial loss of function alleles, and we still lack an estimate of the human germline effect of two complete loss of function alleles. As we move toward better quantification of partial loss-of-function genotypes, it will be important to consider how they interact epistatically with each other and additional genes–for example, variants that impair the function of *MUTYH* and *OGG1* appear to interact epistatically in both the germline and the soma (Robinson et al. 2022; Sasani et al. 2024).

The apparent effect size difference between p.Y179C/V234M and p.Y179C/G368D suggests that there may be utility in moving beyond the binary classification of *MUTYH* variants as simply pathogenic or non-pathogenic. Although data sparsity issues imply that this effect size difference should be interpreted with caution, recent studies of *MUTYH* mutator alleles in the mouse germline and the human soma have also found that some genotypes have more severe mutator phenotypes than others (Robinson et al., 2022; Sherwood et al., 2023). Previous somatic mutation data found an effect size difference between the common genotypes p.Y179C/G368D

and p.Y179C/Y179C that appeared concordant with an earlier age of polyposis onset in p.Y179C/Y179C carriers (Robinson et al. 2022). For a rare genotype like p.Y179C/V234M, epidemiological data can likely not predict variant effect severity, and sequencing of normal tissues obtained from carriers of this genotype may prove to be a more viable option for obtaining this information. In this way, the mutation load in healthy tissues like the germline or blood might eventually prove useful for predicting the severity of cancer risk likely to be associated with different pathogenic *MUTYH* genotypes, allowing clinicians to use whole genome sequencing to discern whether a family or an individual with a suspicious DNA repair variant is accumulating mutations in normal tissues faster than expected and might be at elevated risk of acquiring a mutation that transforms normal tissue into cancer.

To further explore these questions and address the limitations inherent in human studies, we complemented our findings of **Chapter 1** with a detailed analysis of somatic mutation burdens and mutational signatures in BXD mice carrying either the "B" or "D" allele of the *Mutyh* gene in **Chapter 2**. This mouse model allowed us to examine the mutational consequences of *Mutyh* deficiency in a controlled environment, revealing important genotype- and tissue-specific differences. While we observed a significant increase in C>A mutations in spleen tissues of "D" allele carriers, the differences in colon tissues were not statistically significant and were largely influenced by a single outlier mouse (BXD102). These findings suggest that the impact of *Mutyh* deficiency on somatic mutation rates in mice may vary between tissues or require larger sample sizes to detect subtle effects. Additionally, the variability underscores the complex interplay between genetic background and tissue-specific factors in modulating mutational outcomes.

Although we were able to sequence six fresh colon samples, several technical challenges constrained our capacity to obtain larger and higher-quality datasets for this tissue type. We were unable to sequence additional colon samples from the aged BXD mice because the remaining specimens were flash-frozen without preservative media such as EDTA. Upon thawing, the epithelial cells containing the colon crypts tended to burst, resulting in samples not enriched for the target epithelial tissue. This potentially diluted the signal of mutational processes specific to that cell population, as these epithelial cells are enriched for ROS-induced damage and are primary sites of *Mutyh* activity (Irrazabal et al., 2020). Moreover, working with fresh tissue is generally preferable for sequencing low-frequency somatic mutations with high fidelity. The freeze-thaw process can cause DNA damage due to ice crystal formation and cell lysis, potentially introducing artifacts or reducing DNA quality (Falk et al., 2018). These technical limitations underscore the importance of sample preservation methods and tissue handling in studies of somatic mutagenesis.

Similarly, while we had access to additional fresh blood samples, they were preserved with heparin, which interferes with downstream PCR reactions and sequencing protocols, precluding their use in our study (Yokota et al., 1999). This limitation prevented us from including blood as an additional tissue for assessing somatic mutations. In contrast, flash-frozen spleen tissue was available and relatively easier to work with for bulk sequencing and extraction of potentially clonally expanded hematopoietic stem cells. However, the possibility of interference from expanded immune cells, which may proliferate stochastically under immune-stress-related conditions (Muralidharan & Mandrekar 2013; Tsyglakova et al., 2019), could affect the observed mutation spectra. Laboratory mice can exhibit immune cell expansions due to subclinical

infections or stress, potentially introducing variability in mutation rates and spectra that are unrelated to *Mutyh* deficiency.

Despite these constraints and the inherent variability of mutation rates among different BXD mouse strains, our study provided valuable insights into the mutational processes associated with *Mutyh* deficiency. The observed dominance of SBS18 in colon tissues, alongside the elevated contributions of SBS5, SBS30, and SBS36 in spleen tissues, underscores tissue-specific differences in mutational processes. These differences likely reflect distinct cellular contexts and interactions with other DNA repair pathways. For instance, spleen tissue, rich in proliferating lymphocytes, may have different oxidative stress levels and DNA repair activities compared to colon epithelial cells (Crane et al., 2021; Ma et al., 2008).

The advanced ages of the mice at the time of sampling (735–856 days) raise questions about when mutations occurred, as some may have accumulated during early development while others likely arose later in life. Mutations detected at low frequencies in only a few duplex reads were likely acquired post-embryonically, as early embryonic mutations would typically be present at much higher allele frequencies due to clonal expansion.

When this study began, we hypothesized that *Mutyh* deficiency would primarily increase mutation rates during tissue aging. However, our **Chapter 1** findings suggest that the mutator effect may be strongest during early embryonic development, potentially driving higher mutation burdens in adult tissues due to persistent mutations from that early period. This raises the possibility that sampling younger mice could reveal larger genotype-specific differences, as a

greater proportion of their mutations would have arisen during development, when *Mutyh*-mediated BER might play a more critical role.

While studying older mice allowed us to capture the lifelong accumulation of mutations, our sampling design was not well-powered to explore how *Mutyh* deficiency influences mutation accumulation across different stages of the lifespan. Future studies should sequence both younger and older mice to better resolve whether the mutator effect is limited to early development or continues into adulthood, and to assess whether mutation accumulation accelerates with age or occurs at a steady rate.

Analyses of diet as a potential covariate impacting somatic mutation rates did not reveal statistically significant effects on overall mutation burden or interactions with *Mutyh* allele status. However, dietary factors are known to influence oxidative stress levels and could potentially affect ROS-induced DNA damage (Aleksandrova et al., 2021). One possible explanation for the lack of observable differences is that the germline mutator effect of *Mutyh* deficiency appears to occur primarily during embryonic development, rather than through continuous mutation accumulation with age. If a similar mechanism operates in somatic tissues, *Mutyh*-deficient mice might acquire most of their SBS18-associated mutations early in development, limiting the extent to which age-related mutation accumulation could differentiate "B" and "D" allele carriers later in life. Alternatively, the absence of significant findings could be due to limited sample sizes, insufficient diet-induced oxidative stress differences within our cohort, or tissue-specific factors such as compensatory DNA repair pathways that reduce mutation burdens in certain tissues. Future studies should explore how *Mutyh* deficiency

interacts with environmental exposures, tissue context, and life stage to better define its role in shaping somatic mutational landscapes.

We also uncovered insights into replication timing-dependent mutation patterns across substitution classes. Late-replicating regions exhibited higher mutation frequencies, particularly for C>T mutations, which may reflect increased exposure to mutagenic processes due to prolonged single-stranded DNA during replication stress (Tomkova et al., 2018). This pattern aligns with the notion that late-replicating regions are more prone to replication errors and DNA damage (Briu et al., 2021). Advanced techniques such as single-cell sequencing could provide higher-resolution insights into the effects of replication timing on mutation accumulation. Single-cell sequencing would enable the identification of mutations in individual cells, allowing for precise mapping of replication timing-dependent mutational profiles without averaging effects across heterogeneous cell populations. Incorporating such methods in future studies would enhance our capacity to detect low-frequency mutations, track their accumulation over time, and better understand the interplay between DNA replication timing and *Mutyh*-mediated repair in the murine context.

The clonal expansions observed in cancer-associated homologs underscore the potential oncogenic consequences of *Mutyh* deficiency. These findings are consistent with reports that mutations in these genes contribute to clonal hematopoiesis and increase the risk of hematological malignancies (Pourebrahim et al., 2024; Pich et al., 2022). Our observations suggest that even in non-cancerous tissues, *Mutyh* deficiency predisposes cells to increased genomic instability, elevating the risk of cancer development. Expanding these analyses to

additional tissues and incorporating mouse models of cancer could provide a broader understanding of how *Mutyh* mutations influence cancer pathways. Furthermore, investigating epistatic interactions between *Mutyh* and other base excision repair (BER) genes, such as *Ogg1*, may reveal how combined deficiencies synergistically exacerbate mutagenesis and genomic instability (Sasani 2024).

By addressing these technical challenges and building upon our findings, future research can further elucidate the complex relationship between DNA repair deficiencies, mutational processes, and disease risk. The findings presented in **Chapter 2** contribute to a nuanced understanding of the role of *Mutyh* deficiency in somatic mutagenesis across tissues and throughout aging. They emphasize the importance of tissue-specific contexts and the interplay of multiple factors in shaping mutational landscapes.

Recognizing the significance of both positive and negative results provides a more comprehensive perspective on how *Mutyh* mutations influence genomic stability, critical for evaluating cancer risk and designing targeted interventions for populations with DNA repair deficiencies. Investigating low-frequency mutations in normal tissues could identify early mutational events that signify increased cancer risk (Fiala 2020). Developing and applying sensitive techniques for detecting these mutations will be crucial for early detection and risk assessment. Importantly, examining the effects of more deleterious *Mutyh* mutations and considering epistatic interactions, such as concurrent impairments in *Mutyh* and *Ogg1*, may offer deeper insights into how combinations of DNA repair deficiencies drive mutagenesis and cancer risk. Our mouse model provides a valuable system to study these interactions in a controlled

environment, advancing our understanding of the mechanisms underlying DNA repair deficiencies and their contribution to genomic instability.

Conclusion

Our combined human and mouse studies underscore the complex interplay between *MUTYH* genotype, tissue context, and developmental timing in shaping both somatic and germline mutation landscapes. By addressing the limitations of our current studies—such as small sample sizes and variability in genetic backgrounds—and by developing advanced techniques to detect low-frequency somatic mutations in normal tissues, we can achieve a more comprehensive understanding of *MUTYH*-mediated mutagenesis. This enhanced understanding will not only elucidate the mechanisms underlying genomic instability in normal aging and disease contexts but also aid in identifying specific genomic mutational patterns associated with increased cancer risk. Ultimately, such insights may inform strategies for early detection, cancer risk assessment, and the development of preventive interventions.

Data and Code Availability

All genomic data are available for controlled access via dbGaP, accession number phs003554.v1.p1. Per-individual de novo mutation counts and mutation spectra are available in Tables S1-S3. Custom scripts necessary for reproducing our analyses are available on GitHub at https://github.com/harrispopgen/mutyh_human_pedigree.

A human reference panel of phased VCF files from the high coverage 1000 Genomes project (Byrska-Bishop et al. 2022) was used to phase the data and infer shared haplotype tracts between relatives. These data can be found at https://www.internationalgenome.org/data-portal/data-collection/30x-grch38.

Poisson regression coefficients used for the parental age model can be found in Jónsson et al. (2017)'s Table S9. Sherwood et al. (2023)'s de novo mutation counts and mutation spectra are found in Table 1 and Table S2 of that study, respectively.

Acknowledgements

We thank all the study participants for their time and engagement with our research. We thank Martha Horike-Pyne for her assistance drafting consent forms and applying for Institutional Review Board approval, and we thank Jailanie Kaganovsky and Vidha Sudhesh for their assistance mailing DNA collection kits to the study participants. We also thank the editor and three anonymous reviewers for providing feedback that helped us improve the manuscript. We also benefited from helpful discussions with Rosana Risques, Lea Starita, and members of the Harris Lab, whose expertise and camaraderie have greatly enriched this work. In particular, we thank Annabel Beichman, David Mas Ponte, and Georgia Tsambos for their critical contributions and constant encouragement throughout this project. Their mentorship, collaboration, and feedback were instrumental in shaping the ideas and analyses presented here.

Lastly, we extend our heartfelt gratitude to our friends and family, whose unwavering support and understanding have been a source of strength and motivation throughout this work.

Funding

The collection and sequencing of all human subjects data was funded by a Searle Scholarship to K.H. We acknowledge additional financial support from NIH/NIGMS grant R35GM133428, a Burroughs Wellcome Fund Career Award at the Scientific Interface, a Pew Scholarship, the Allen Discovery Center for Cell Lineage Tracing, and a Sloan Fellowship, all to K.H. A.C.B. received additional support from the NIH Biological Mechanisms of Healthy Aging training grant T32 AG066574, and C.Y. received support from the NIH Cellular and Molecular Biology training grant T32 GM007270. S.H. and J.K. were supported by NIH/NIGMS R01 GM129123. B.S. received support from the Damon Runyon-Rachleff Innovation Award (DRR-33-15) and the Brotman Baty Institute for Precision Medicine.

Conflict of Interest

B.S. consults for the company Constantiam Biosciences. J.O.K serves as a scientific advisor to the company MyOme. R.A.R. is a consultant and equity holder at TwinStrand Biosciences Inc. and an equity holder at NanoString Technologies Inc. R.A.R. is named inventor on patents owned by the University of Washington and licensed to TwinStrand Biosciences Inc. The authors declare no other competing interests.

Chapter 1: Supplemental Figures

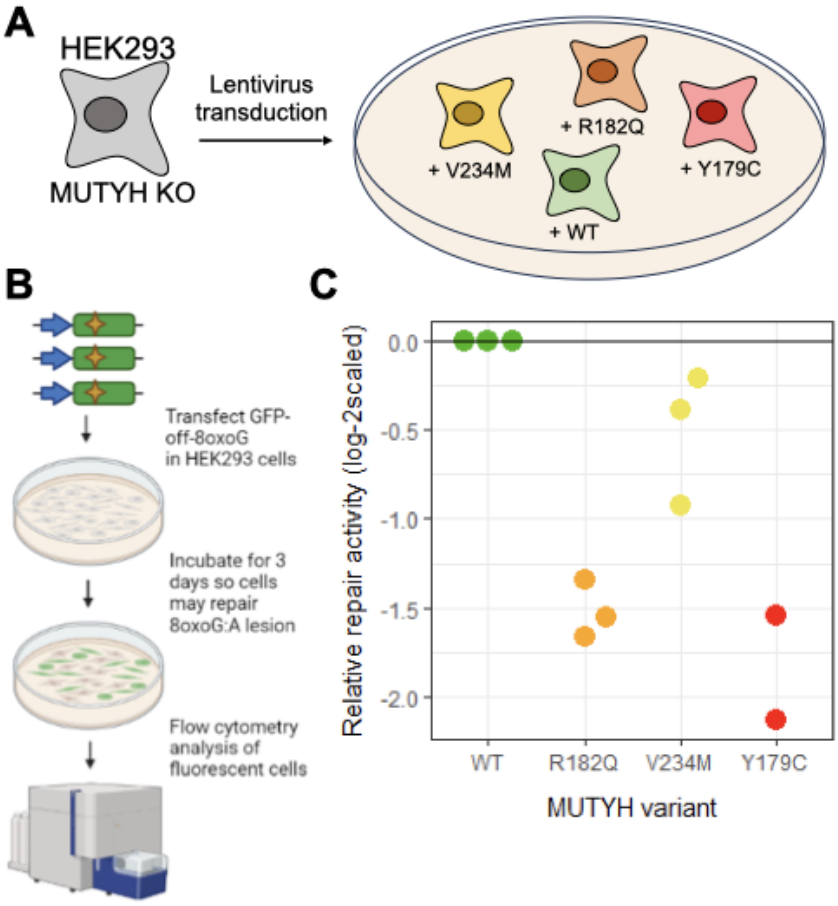


Figure S1. Cell-based *in vitro* assay of MUTYH 8-oxoG:A repair function. **A)** We first generated knock-in HEK293 cells expressing different MUTYH variants. **B)** We then transfected in a GFP reporter containing an 8-oxoG:A mispair, which turns cells green when the A is replaced with a C. Flow cytometry was used to sort cells based on GFP fluorescence. (Panel was generated using biorender.com) **C)** Results of the GFP-off assay, with the relative repair activity measured as the frequency of each variant in the GFP+ fraction compared to the frequency before sorting.

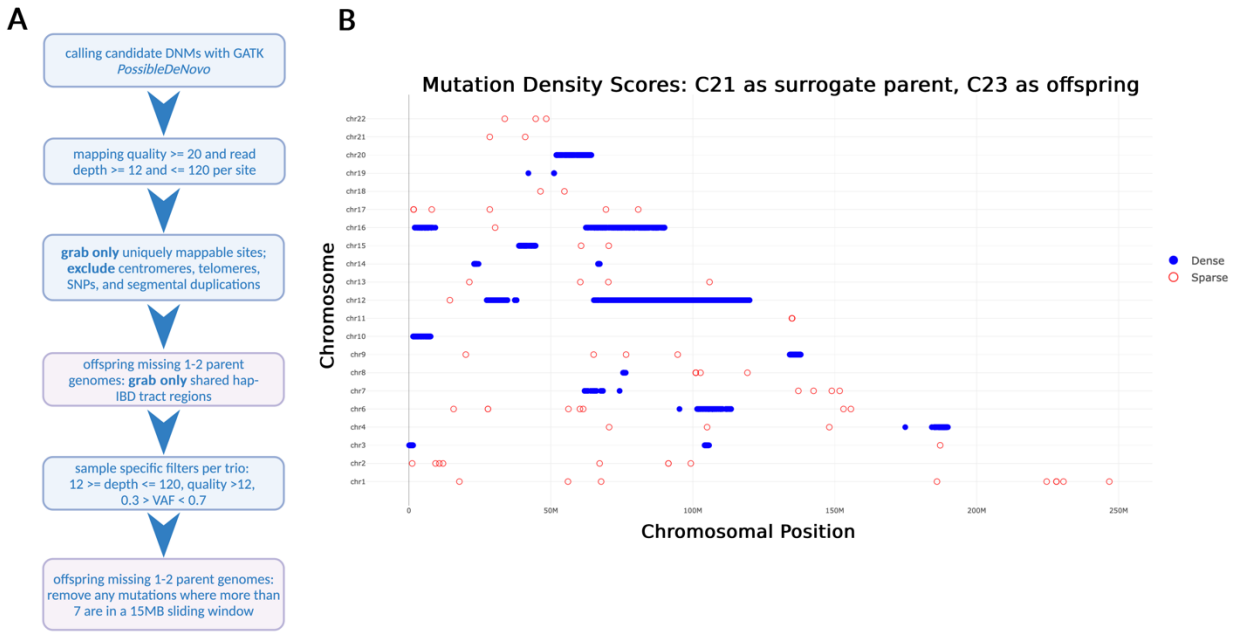


Figure S2. DNM calling and assessment. **A)** DNM calling and filtering workflow diagram. **B)** An example of the density filtering method used in the sibling-as-surrogate-parent calling method applied to C22 as the offspring of surrogate parent C21. Candidate DNMs are "sparse," meaning they are in regions where no more than 7 mutations were identified in sliding window sizes of 15MB (with step size of 3MB). Sparse mutations are outlined in red, while "dense" mutations (that did not pass this density filter) are outlined in blue, and likely represent regions of the genome where the two siblings did not share the same paternal haplotype.

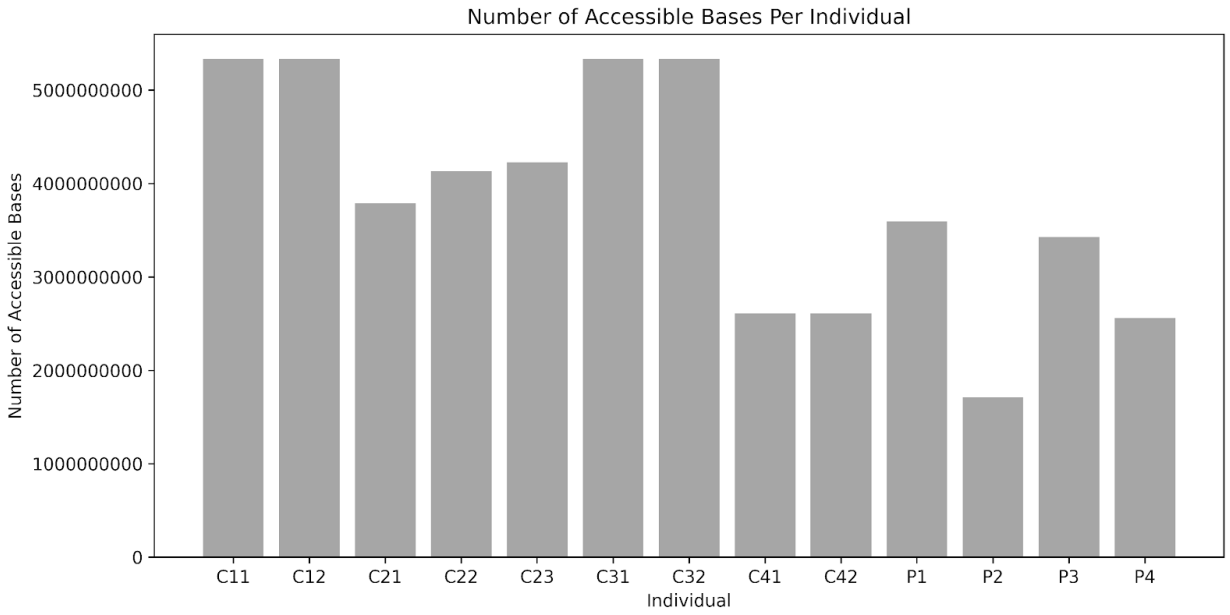


Figure S3. Number of Accessible Bases Calculated per Individual. The number of accessible bases identified per each sequenced individual (excluding S1 and S3). Individuals that did not require the surrogate DNM calling approach (C11, C12, C31, C32) share the same maximum number of accessible bases. Individuals in Families 2 and 4 and the parent generation all have a lower amount of accessible bases, which is dependent in each case on the number of shared paternal haplotype bases seen in each surrogate parent sibling combination. As there is only one combination for offspring of Family 4, the number of accessible bases is lowesst in these individuals.

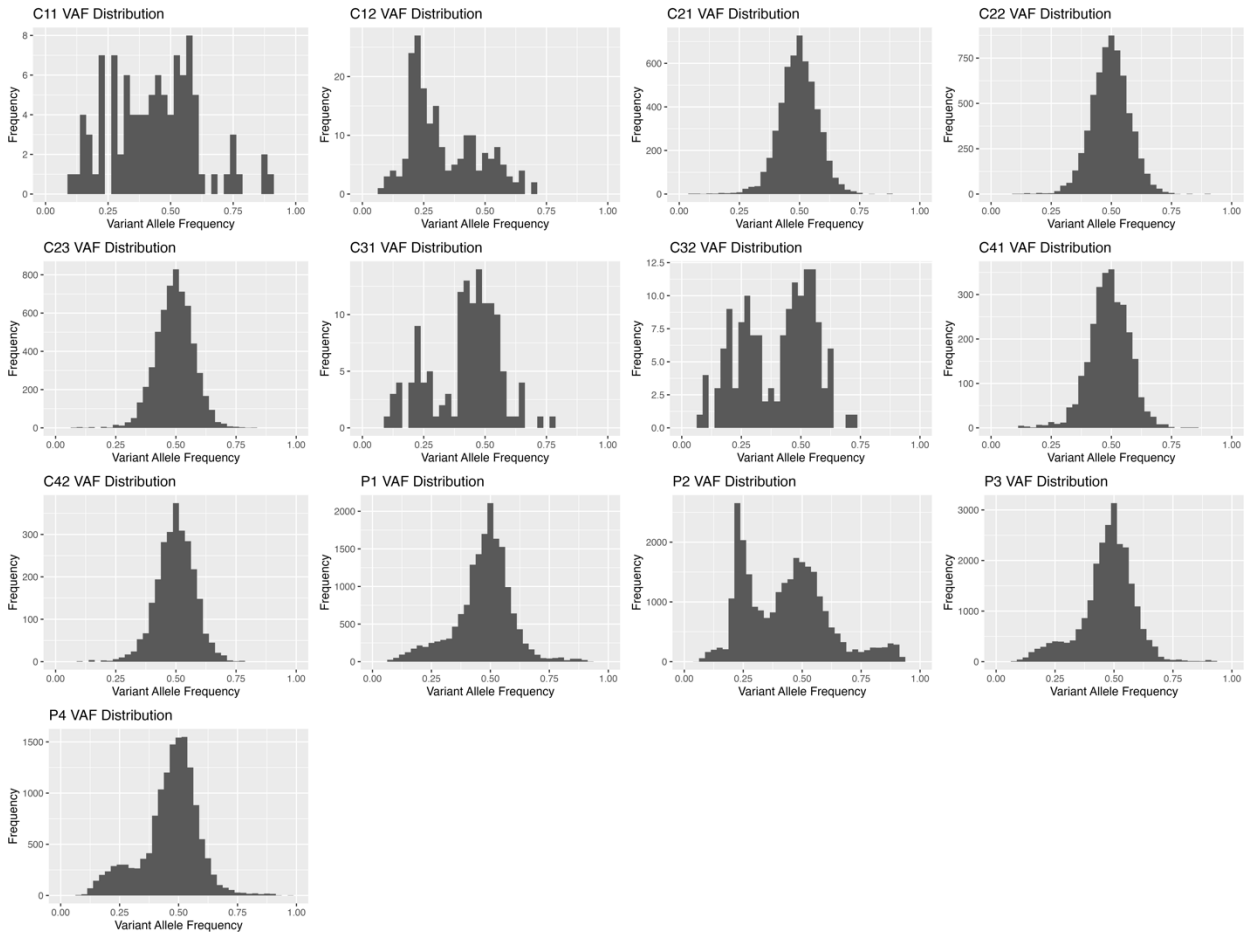


Figure S4. VAF distributions of all candidate DNMs per individual. Distributions of variant allele frequency (VAF) scores are displayed for all identified DNMs of the 13 studied individuals. Two modes are frequently observed across individuals, typically centering below and above 0.40- these likely reflect clonal somatic mutations and germline mutations, respectively. Note that P2 is a clear outlier, with a VAF distribution heavily skewed to the left with relatively few mutations in the candidate germline part of the distribution.

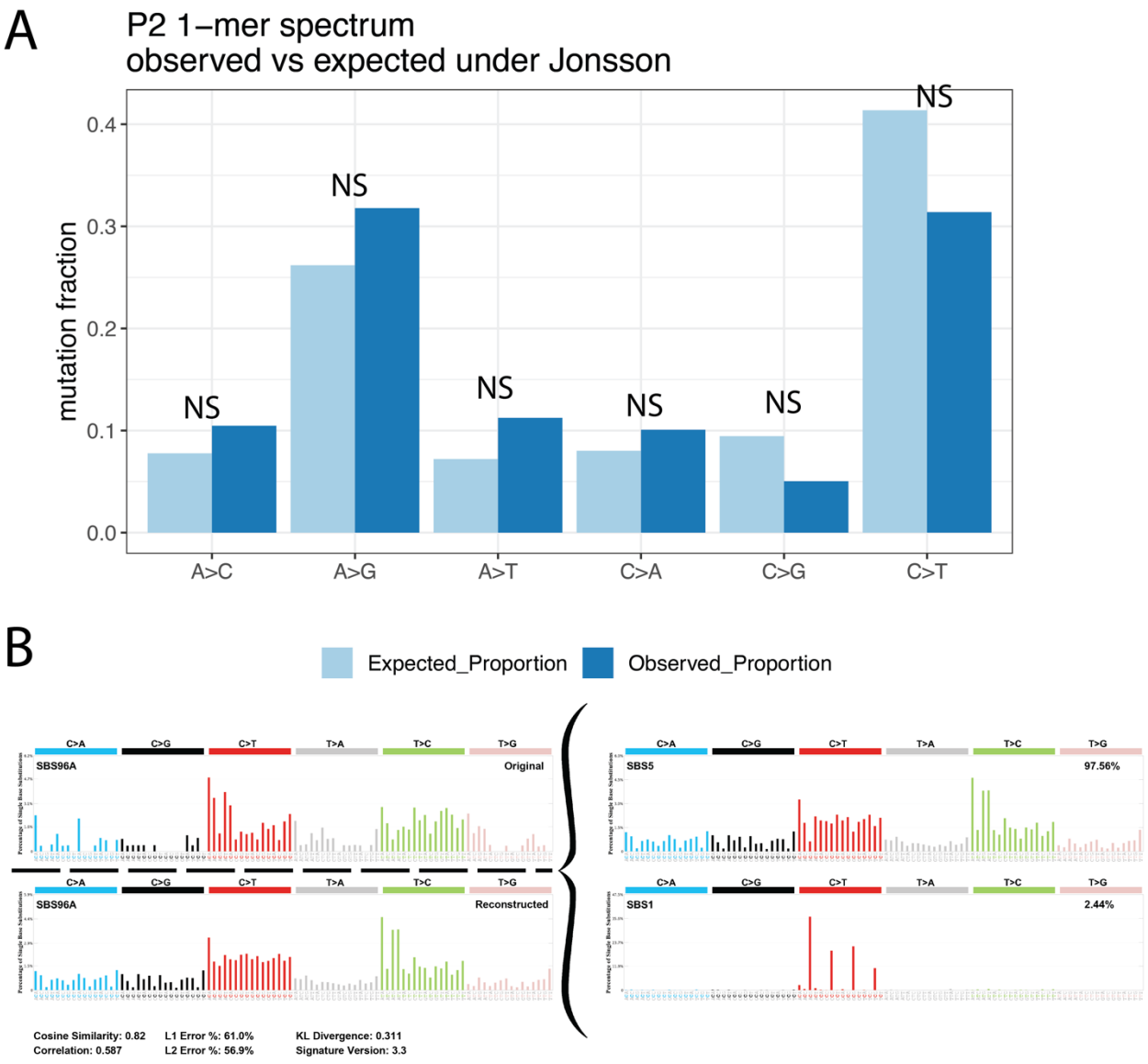


Figure S5. Analysis of spectra in the individual (P2) with elevated number of mutations. **A)** Comparison of the 1-mer de novo mutation spectrum for individual P2, who showed signs of somatic mutation contamination. The proportional 1-mer mutation spectrum was compared to that expected based on the ages of P2's parents at the time of their birth under the parental aging model. Each mutation type's count relative to the sum of the other five types was compared to model expectations using a 2x2 Chi-Squared test and Fisher's Exact test. No comparison resulted in a p-value < 0.05 (labeled as NS for non-significant), indicating that the spectrum is not

significantly different from expectation. Note that these mutations did not go through extensive filtering or manual inspection due to P2 being excluded from further analysis early in the pipeline. The Chi-Squared test was used instead of the Poisson-based test for other individuals because P2's mutation counts were so highly elevated above the counts expected under the parental aging model, so comparing relative proportions of mutation types was more suitable. **B)** We carried out mutation signature decomposition on the 3-mer spectrum of P2 using SigProfilerExtractor and both the COSMIC v3.3 and v2 catalogs (results using v3.3 shown). The mutation signature was decomposed into SBS signatures 1 and 5; signatures SBS18 and SBS36 did not appear.

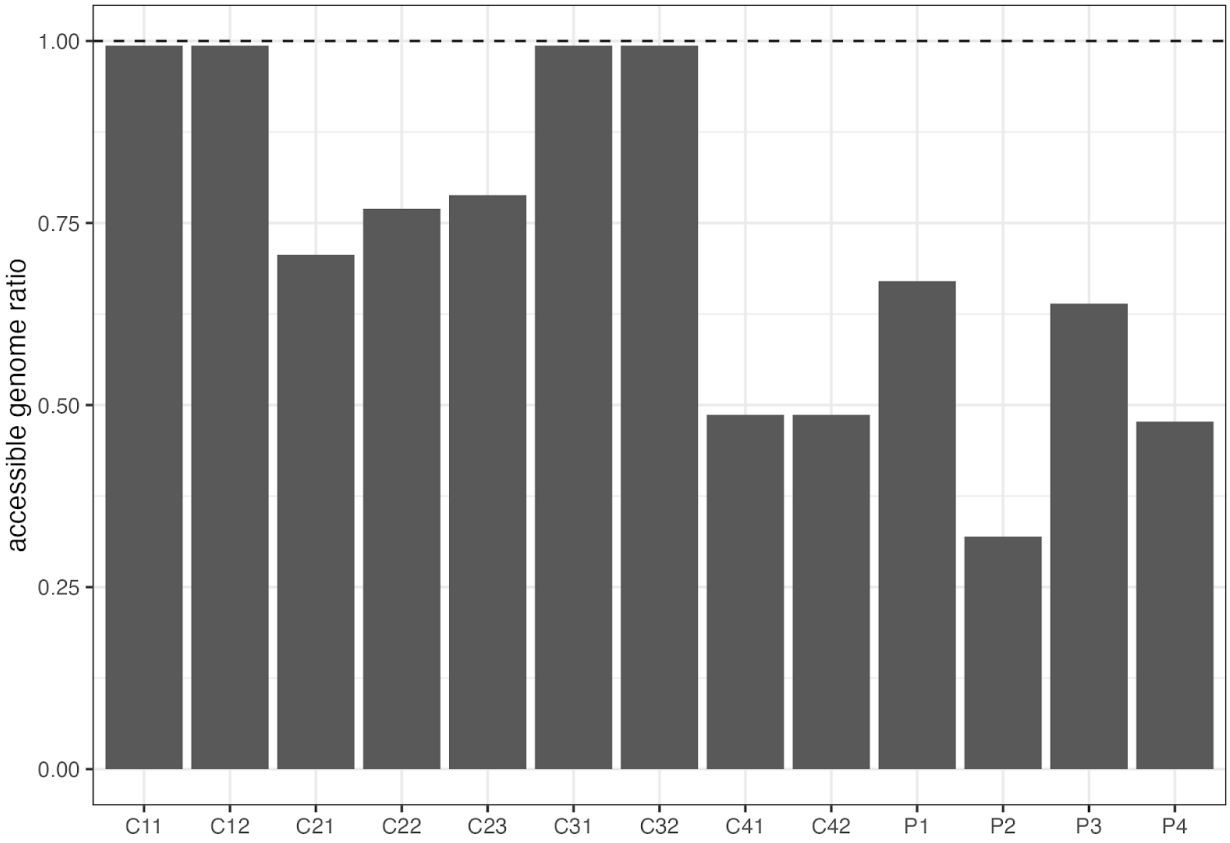


Figure S6. Ratio of accessible genome size for the individuals in this study over the average accessible genome size reported in Jonsson et al. (2017) (2,682,890,000 base pairs). Individuals C21, C22, C23, C41, C42, P1, P2, P3, and P4 all have accessible genome ratios substantially <1 due to the use of the surrogate-parent DNM calling method.

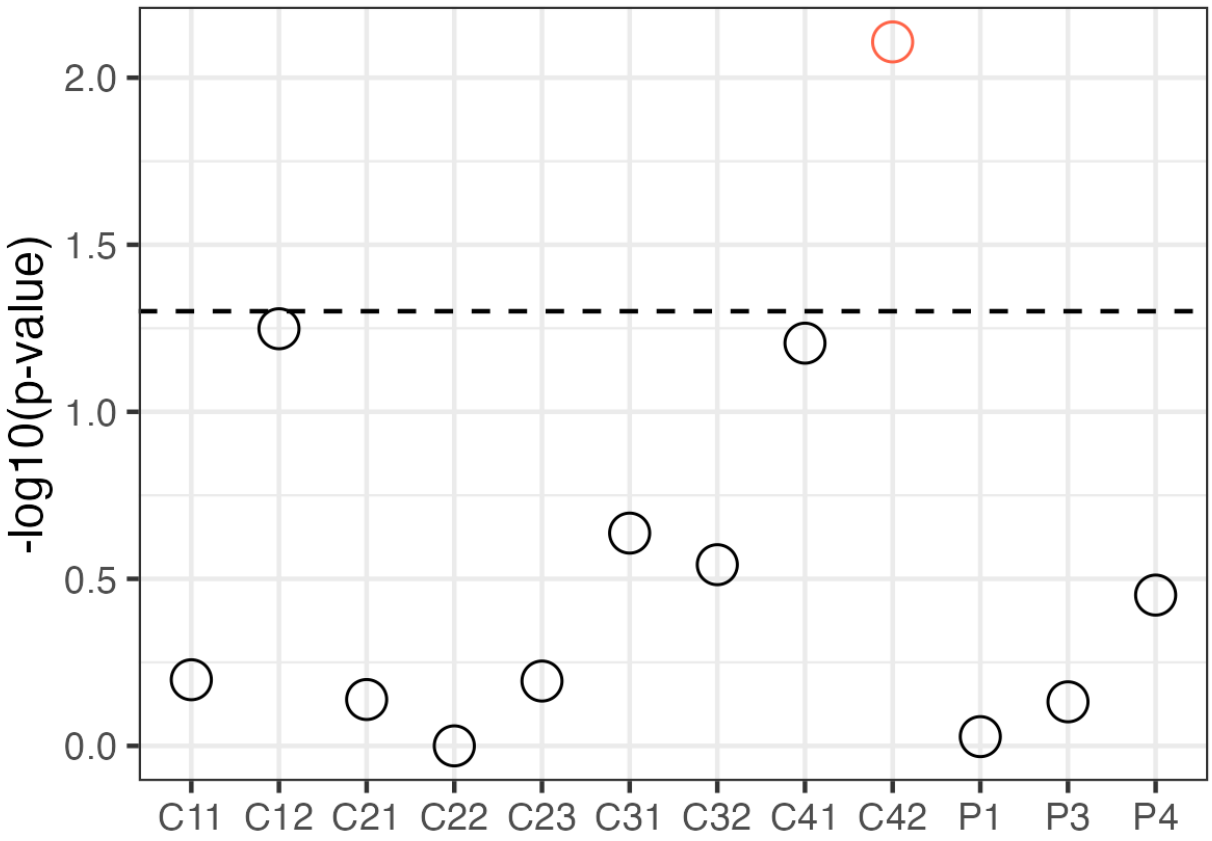


Figure S7. The probability (from the Poisson cumulative distribution) under the parental age model (Jonsson et al. 2017) of observing an overall mutation count greater than or equal to what we observe. The dashed line indicates the p-value threshold of 0.05 (significant points colored in red). All individuals other than C42 have overall DNM counts that are compatible with the parental age model. See the Methods for more detail on how probabilities are calculated.

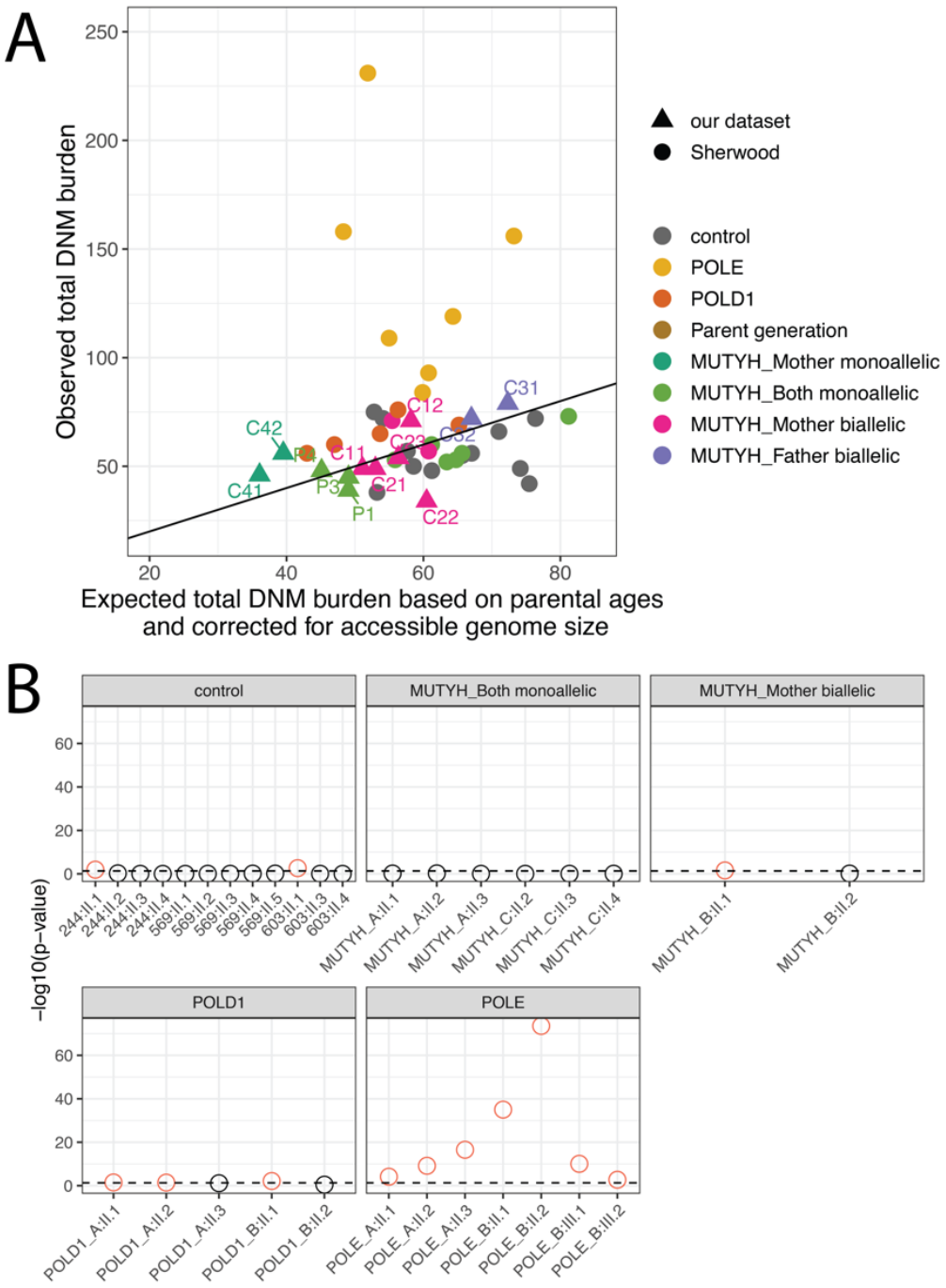


Figure S8. Deviations from the parental age model in Sherwood et al. (2023). **A)** Comparing observed and expected DNM burdens from individuals in this study (triangles) and individuals from Sherwood et al. (2023) (circles). Expected DNM burdens are based on parental age and accessible genome size for the individuals in this study, and based on parental age only for

Sherwood et al. (2023), since accessible genome size was not reported (and was therefore assumed to be ~equivalent to the accessible genome size used to generate the parental age model in Jonsson et al.). "*POLE*" and "*POLD1*" refer to individuals in Sherwood et al.'s dataset that have variation in those polymerase genes and show an extreme effect on the germline mutation rate. The $y = x$ line is shown in black. As in Sherwood et al., individual *MUTYH*_C:II.1 was excluded due to high levels of somatic variant bleed-through. **B)** The probability (from the Poisson cumulative distribution) under the parental age model of observing an overall mutation count greater than or equal to what Sherwood et al. observed. The dashed line indicates the p-value threshold of 0.05 (significant points colored in red). Most of Sherwood et al.'s control individuals (except for two) have overall mutation counts consistent with the parental age model, as do all their individuals with monoallelic *MUTYH* parents. However, one of their individuals with a biallelic *MUTYH* mother has a significantly elevated DNM burden, and the majority of individuals with the more severe variants in *POLE* and *POLD1* show extremely significant elevations of overall mutation count.

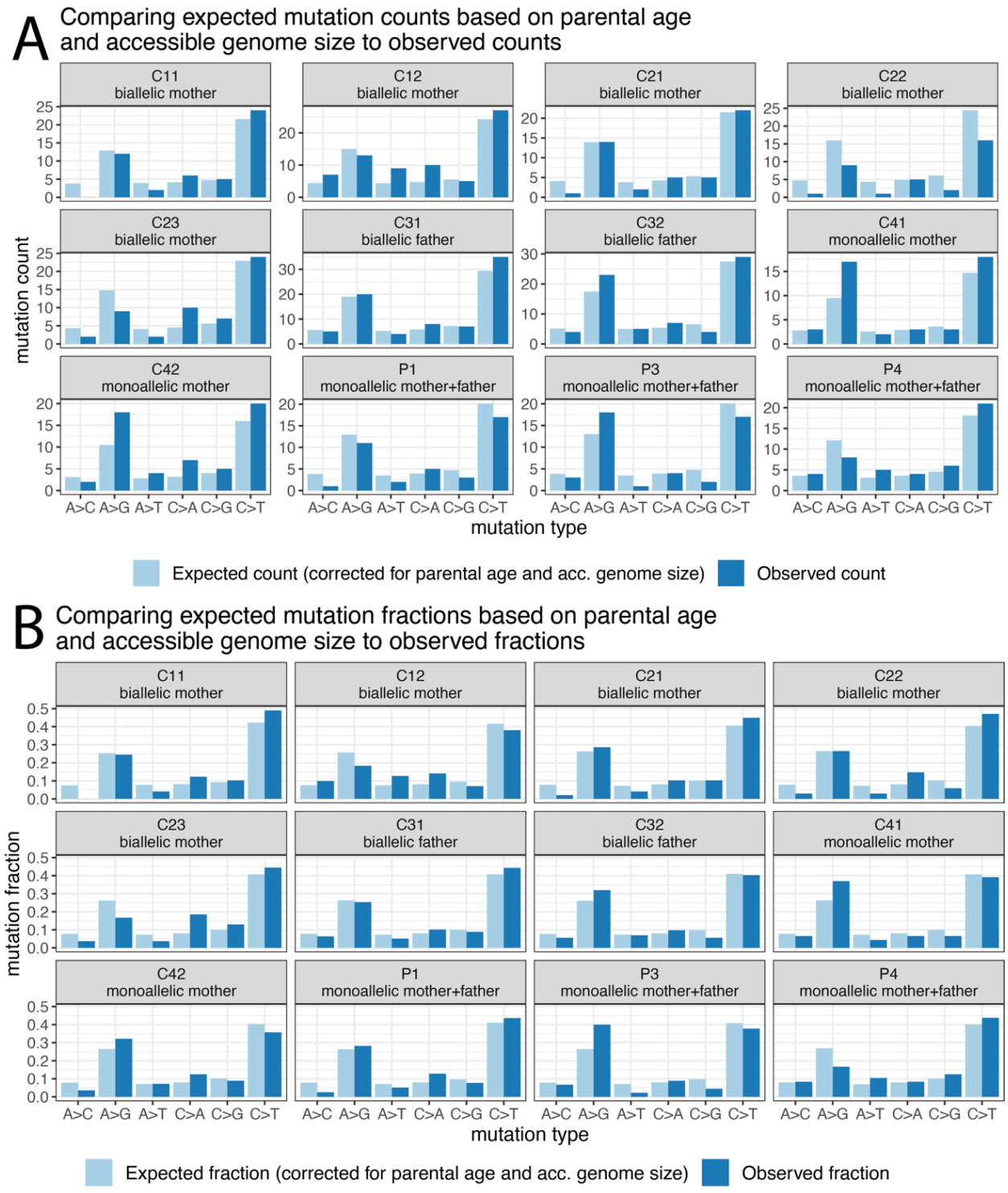


Figure S9. Comparing observed and expected (under the parental age model) mutation spectra. **A)** Comparing observed (dark blue) mutation counts for each 1-mer mutation type to

expectations (light blue) under the parental age model. **B)** Comparing observed and expected mutation fractions (proportion of the total mutations) across mutation types.

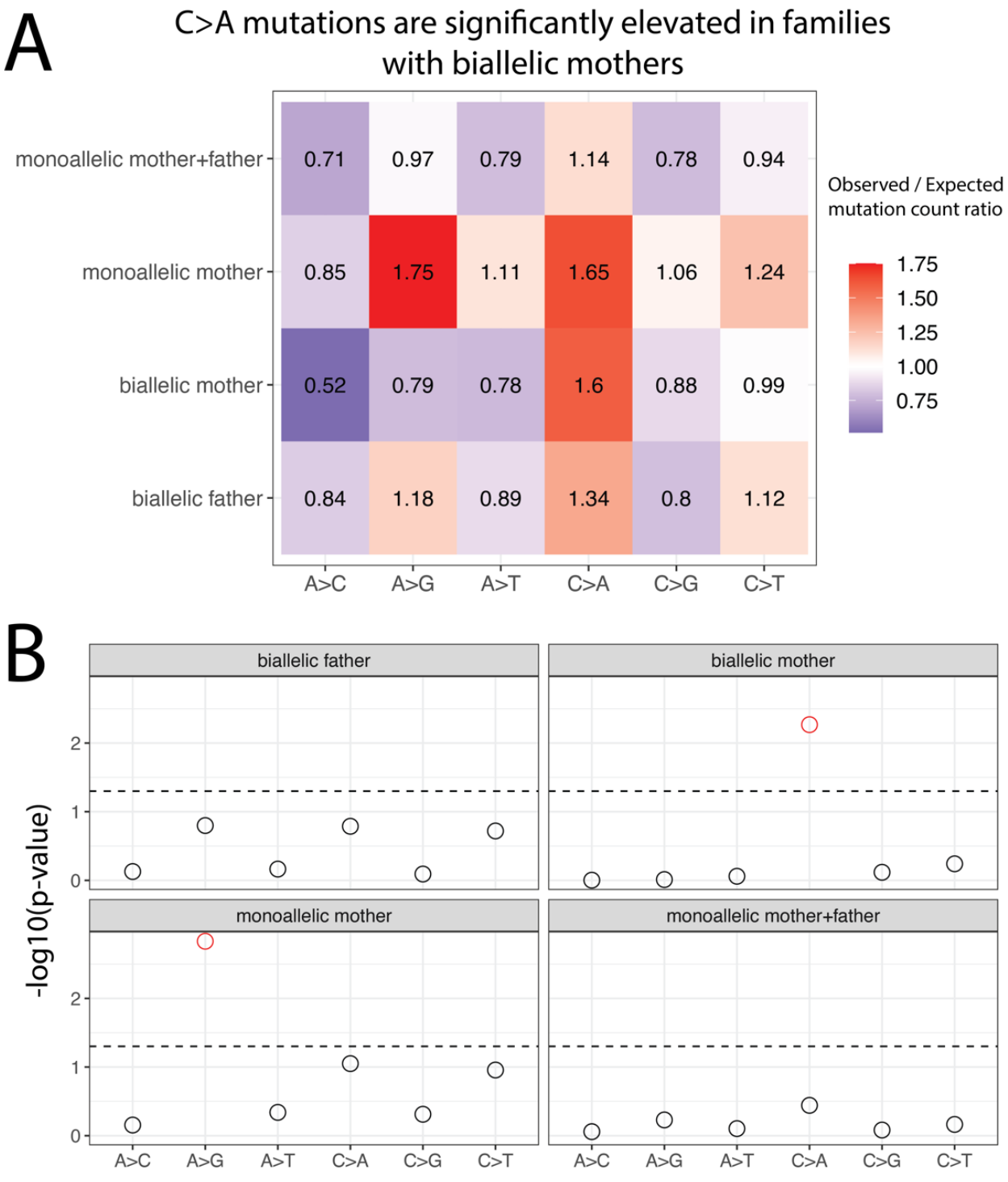


Figure S10. Families with biallelic mothers show significant C>A elevation. **A)** A heatmap showing the ratio of the observed / expected mutation counts per group of children in our dataset that have parents in the category listed on the y-axis (calculated by summing up the mutation

counts per mutation type across all children with a parent of the types listed on the y-axis, as in Sherwood et al.). Only our ‘biallelic mother ’category contains more than one family (Families 1 and 2). The ‘monoallelic mother+father ’represents the parent generation individuals, the ‘monoallelic mother ’group is Family 4, and the biallelic father group is Family 3. **B)** The probability of observing a mutation count of each of the six 1-mer mutation types under the parental age model that is greater than or equal to what we observed for each group. Points above the dashed line (red circles) fall below the upper one-tailed Poisson $p < 0.05$ significance threshold. The biallelic mother group (Families 1 and 2) shows significant elevation for C>A DNM counts above what is expected under the parental age model, and the monoallelic mother group (Family 4) shows significant elevation of A>G mutation types above expectations.

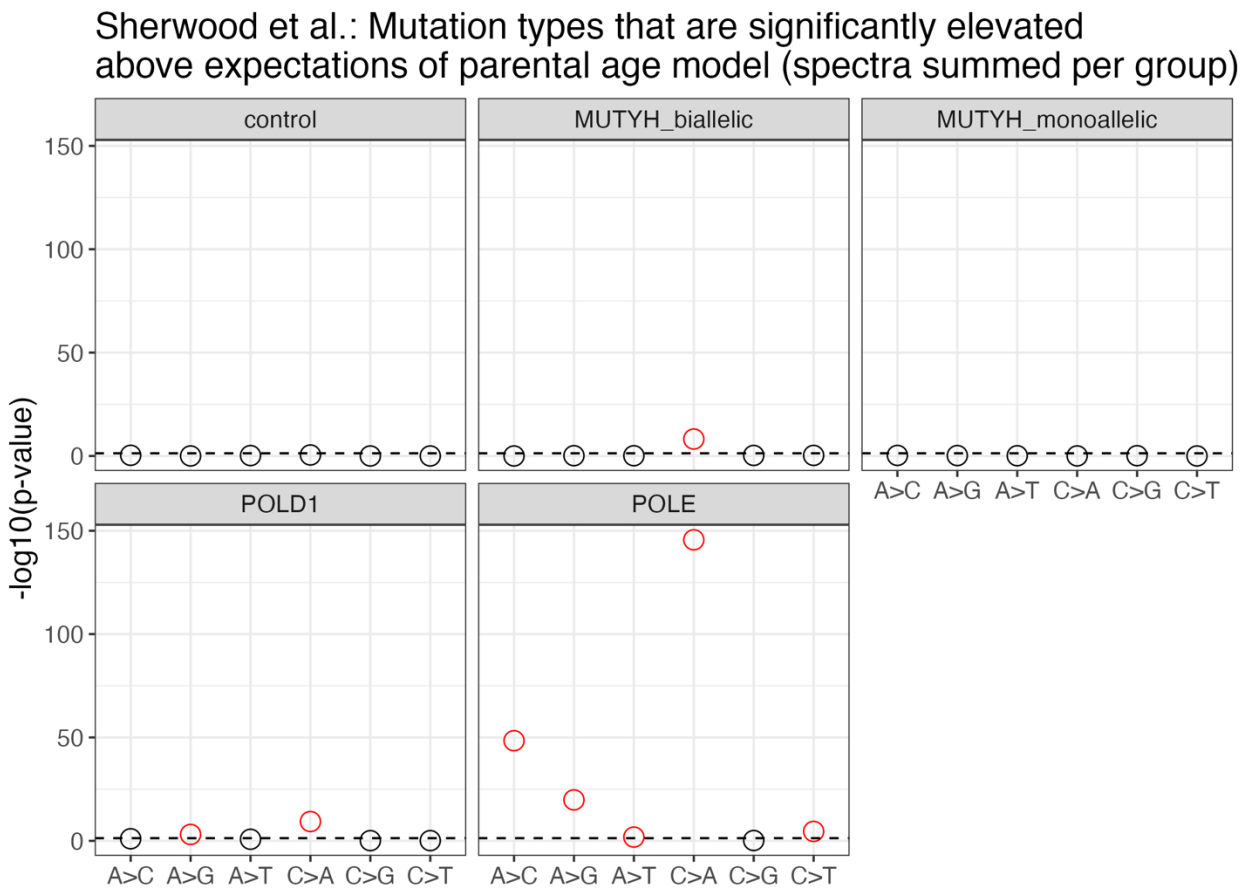
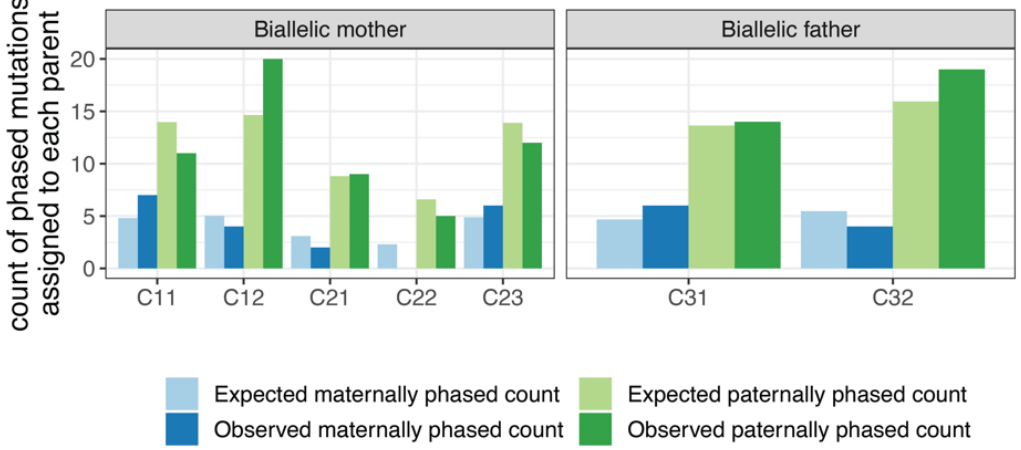


Figure S11. Elevated C>A counts in biallelic *MUTYH*, *POLD1* and *POLE* groups from Sherwood et al. (2023). Under our significance testing framework, the mutation spectra summed per group from Sherwood et al. (2023) (summarized in the heatmap in **Figure 4A)** show a significantly elevated C>A count for the biallelic *MUTYH* family, as well as the groups of individuals with more severe *POLD1* and *POLE* variants.

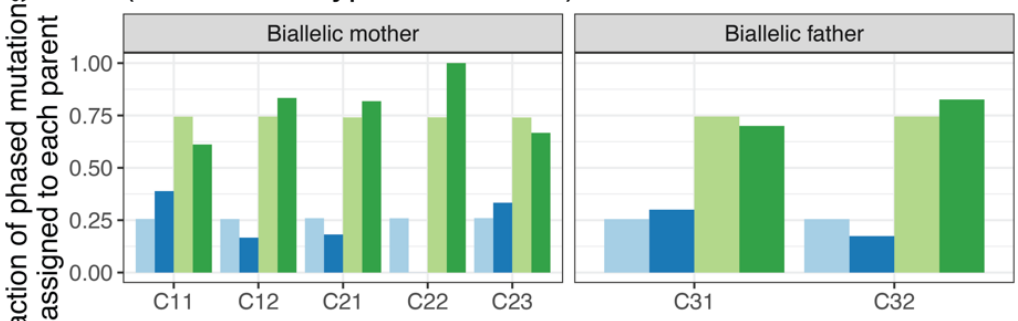
A

Comparing observed and expected counts of mutations phased to each parent. Expectations from parental age model * phasing rate (all mutation types combined)



B

Comparing observed and expected proportions of mutations phased to each parent. Expectations from parental age model * phasing rate (all mutation types combined)



C

Testing for significant deviations from expectations in the number of mutations phased to each parent (all mutation types)

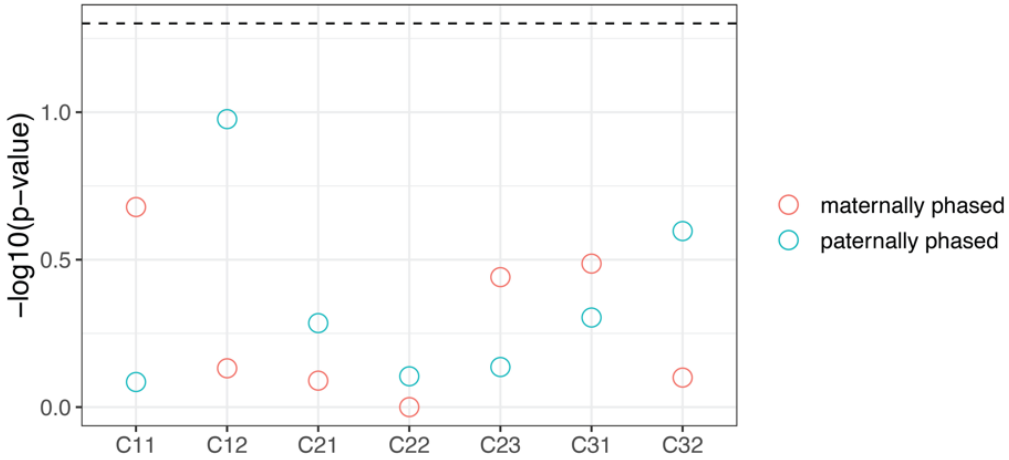


Figure S12. No significant differences in the amount of mutations phased to the carrier parents are observed. Counts **(A)** and relative fractions **(B)** of phased mutations assigned to either the maternal (dark blue) or paternal (dark green) haplotypes. Expectations (light blue and light green) are based on the number of mutations expected to come from each parent under the parental aging model (corrected for accessible genome size and individual phasing success rate). **C)** The probability (from the Poisson cumulative distribution) of observing greater than or equal to the number of mutations phased to each parent under the parental age model (corrected for accessible genome size and individual phasing success rate). Dashed line indicates $p < 0.05$ threshold. No individual shows significantly more mutations phased to either parent than in expectation. See the Methods for more details on how probabilities are calculated.

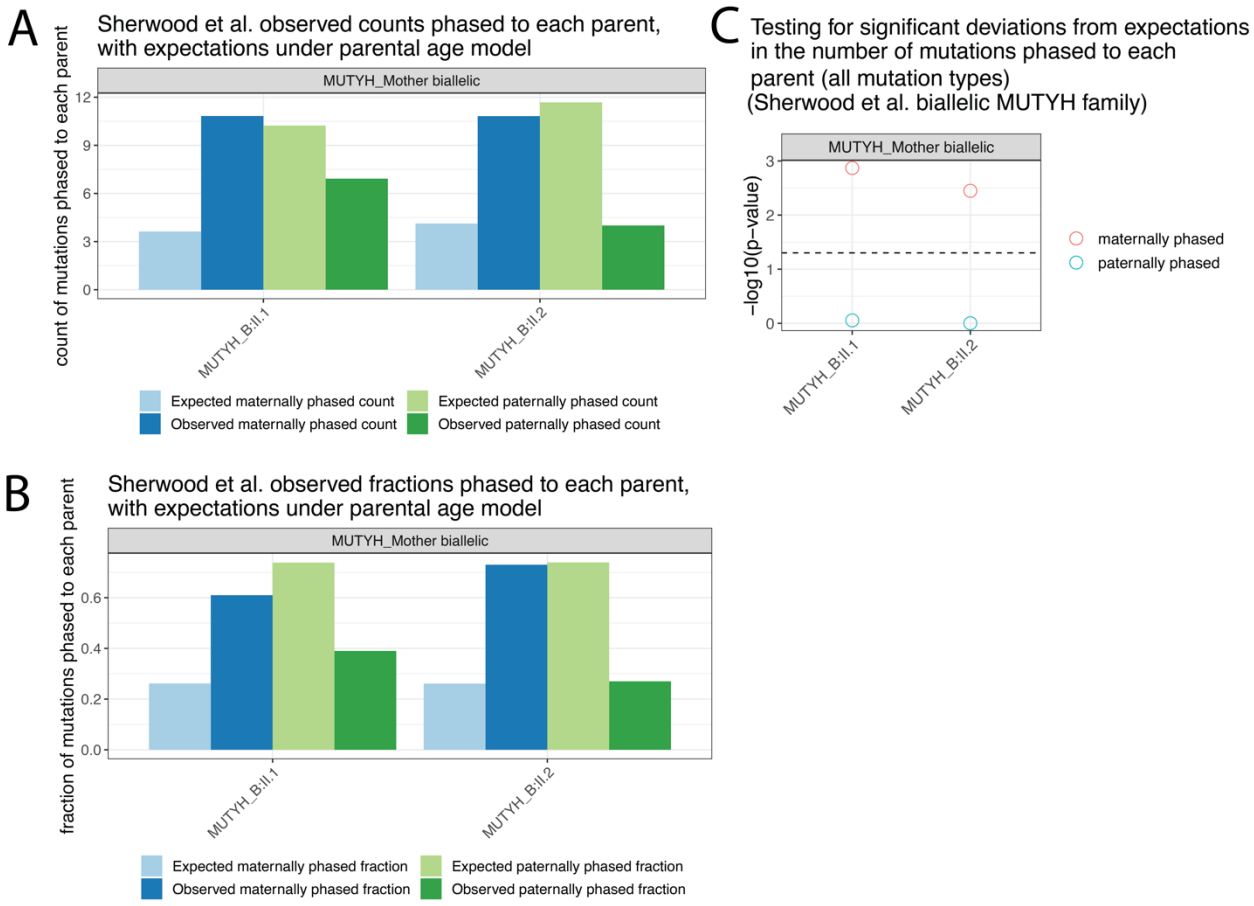


Figure S13. Significantly more DNMs were phased to the biallelic *MUTYH* mother in Sherwood et al. (2023) in both of her children (*MUTYH*_B.II.1 and *MUTYH*_BII.2). **A)** The counts of mutations phased to maternal (dark blue) and paternal (dark green) haplotypes reported by Sherwood et al. (2023), with expectations from the parental age model in light blue and light green. **B)** As in **(A)**, but showing the fraction of phased mutations phased to each parent. Note the substantial elevations of maternally-phased mutations that they report. **C)** Under our significance testing threshold, both children of the biallelic *MUTYH* mother from Sherwood et al. show a significant elevation of overall mutations phased to the maternal haplotype than what is expected under the parental age model.

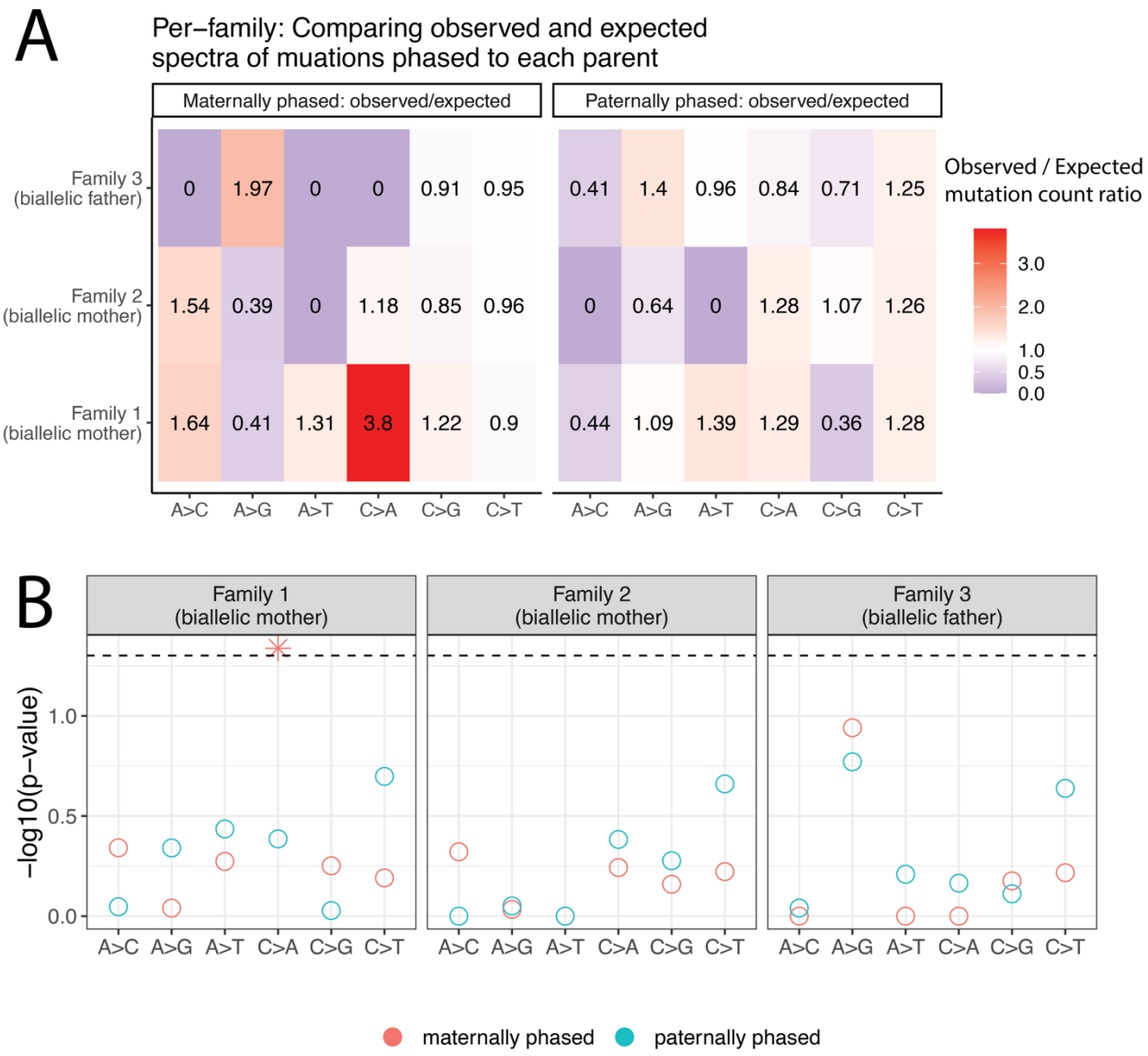


Figure S14. Significantly more C>A mutations phased to the children of Family 1 (biallelic mother) than expected. **A)** The ratio of observed/expected (under the parental age model (corrected for accessible genome size and individual phasing success rates) mutations phased to maternal and paternal haplotypes across the three biallelic *MUTYH* families in this study. **B)** The probability of observing greater than or equal to the number of mutations phased to each parent under the parental age model (corrected for accessible genome size and individual phasing success rates). Only Family 1 shows a significant result for C>A mutations phased to the carrier parent (biallelic mother).

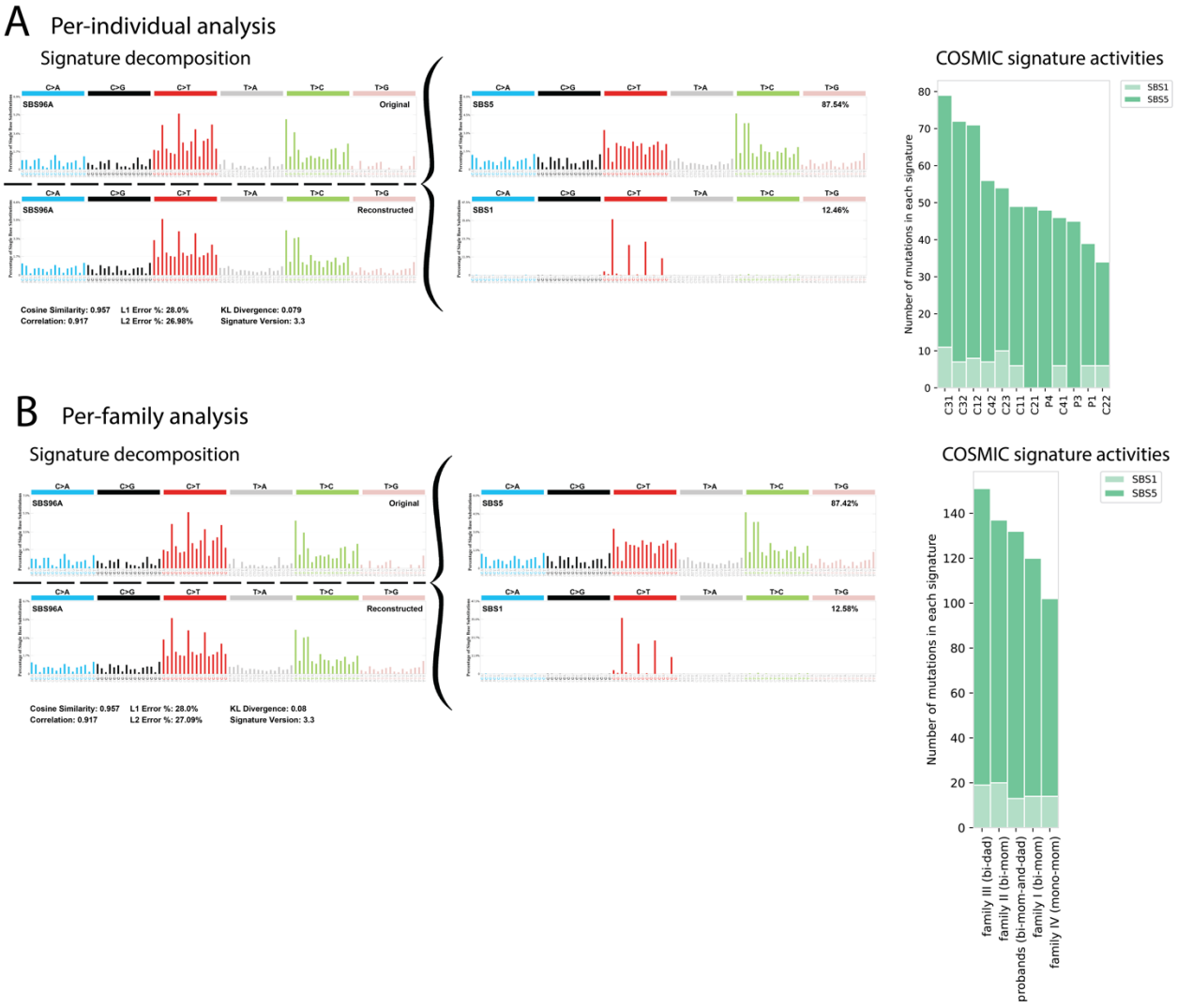


Figure S15. Mutational signature extraction does not find activity of SBS18 or SBS36. Results of SigProfilerExtractor, which extracts novel mutation signatures from the 3-mer mutation spectrum either at the per-individual level **(A)**, or across spectra summed across siblings within the same family **(B)**. The novel signature is then deconvoluted into known COSMIC signatures. The novel signature is shown in each row as "SBS96A (original)", and its reconstruction based on known COSMIC signatures is shown below ("SBS96A (reconstructed)"). The COSMIC signatures used to reconstruct the signatures are shown to the right of the brackets (SBS1 + SBS5). The cosine similarity reported is between the original and reconstructed signatures, indicating how well COSMIC signatures can be used to reconstruct the signatures extracted from the empirical data. The inferred activities (numbers of mutations contributed) of each signature is shown in the "COSMIC signature activities" plots. Analysis was repeated using COSMIC v2 in

which signatures SBS18 and SBS36 are not separated, but did not yield qualitatively different results.

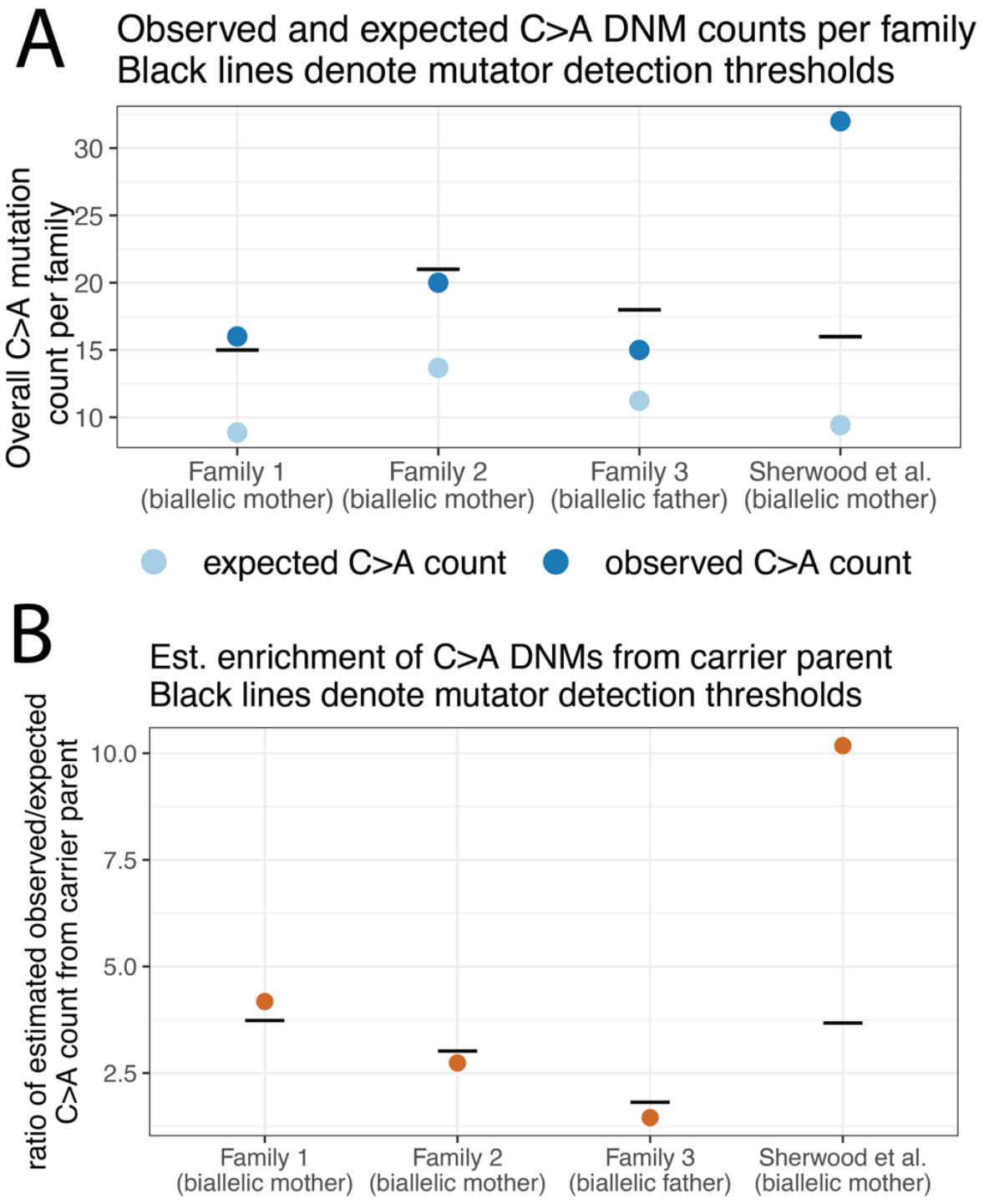


Figure S16. Estimating the minimum *MUTYH* effect sizes that we have power to detect in the male and female germlines. **A)** Observed (dark blue) and expected (light blue) C>A mutation counts in the children of each family with a biallelic parent. Horizontal black lines show the minimum number of mutations needed to reject the null parental age model ("mutator detection

threshold"). Of the two biallelic mother families, Family 1 exceeds the threshold, implying a significant C>A mutation rate elevation while Family 2 falls just short, and Family 3 (biallelic father) falls further short. The family with a biallelic *MUTYH* mother from Sherwood et al. (2023) is included, and has a much more elevated C>A count than the families in this study. **B)** Estimates of the effect size of *MUTYH* on the number of C>A mutations transmitted by the carrier parent relative to expectations under the parental age model. Orange points indicate an estimate based on observed mutation counts in the children of each family, assuming all excess C>A mutations beyond the parental age expectations were inherited from the carrier parent. The horizontal black lines show the minimum effect size that exceeds a one-tailed 95% confidence interval above the Jónsson (2017) parental age model expectation (corresponding to the mutation counts denoted by the horizontal lines in **(A)**). These effect sizes represent estimates of the overall effect of *MUTYH* variants across gametes from the biallelic carrier parent. The minimal detectable effect size is much lower for Family 3 (biallelic father) than for Families 1 or 2 (biallelic mothers), as fathers transmit much higher numbers of mutations to their offspring, which makes it surprising that we detect significantly elevated C>A rates in Families 1 and 2 but not Family 3. This result suggests that *MUTYH* variation may exert a proportionally stronger effect on the female germline compared to the male germline. The large elevation of C>A mutations in the biallelic mother family from Sherwood et al. (2023) implies a higher effect size in the carrier parent than any seen in the families in this study. See **Figure S17** for this analysis based on per-individual mutation counts.

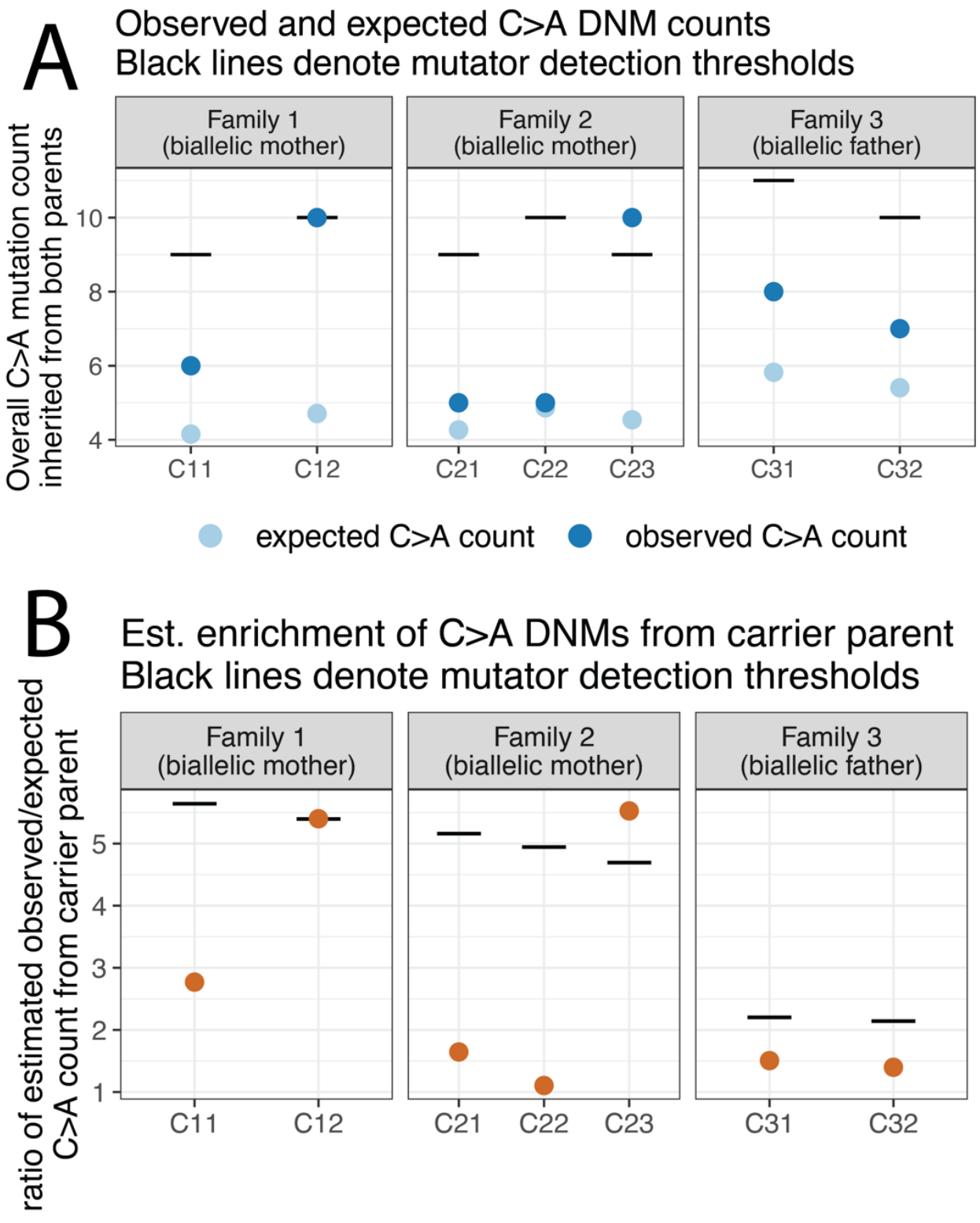


Figure S17. Estimating the minimum per-individual *MUTYH* effect sizes that we have power to detect in the male and female germlines. As in main text **Figure S16**, but based on individuals '

C>A DNM counts rather than C>A DNM counts summed per family. **A)** Observed (dark blue) and expected (light blue) C>A mutation counts per individual in biallelic carrier parent families. Horizontal black lines show the number of mutations needed to reject the null parental age model ("mutator detection threshold"). As can be seen in Figure 3C, among the children of biallelic carriers, only C12 and C23 reach that threshold. **B)** Estimates of the effect size of *MUTYH* on the number of C>A mutations transmitted by the carrier parent relative to expectations under the parental age model. Orange points indicate an estimate based on observed mutation counts, if all excess C>A mutations beyond the parental age expectations are assigned to the carrier parent. The horizontal black lines show the minimum effect size that exceeds a one-tailed 95% confidence interval above the Jónsson (2017) parental age model expectation (corresponding to the mutation counts denoted by the horizontal lines in **(A)**). Note that the minimum detectable effect size in the children of the biallelic father is much lower than that of the children of biallelic mothers, as fathers transmit much higher numbers of mutations to their offspring. Despite this lower threshold, we observe no significant elevation of the C>A count in the biallelic father family.

Chapter 2: Tables and Supplemental Figures

Chromosome	Start Position	End Position	Region Name	Gene Symbol	Transcript ID	Panel
chr1	69,304,217	69,306,617	region01	N/A	N/A	Mutagenesis Panel
chr1	155,235,938	155,238,338	region02	N/A	N/A	Mutagenesis Panel
chr2	50,833,175	50,835,575	region03	N/A	N/A	Mutagenesis Panel
chr3	32,397,570	32,466,207	PIK3CA_ENSMUST00000108243.7	PIK3CA	ENSMUST00000108243.7	Homologous Cancer Gene Panel
chr3	103,058,184	103,068,014	NRAS_ENSMUST00000029445.12	NRAS	ENSMUST00000029445.12	Homologous Cancer Gene Panel
chr3	109,633,160	109,635,560	region04	N/A	N/A	Mutagenesis Panel
chr4	96,825,280	96,827,680	region05	N/A	N/A	Mutagenesis Panel
chr5	18,210,612	18,213,012	region06	N/A	N/A	Mutagenesis Panel
chr6	119,170,706	119,173,106	region07	N/A	N/A	Mutagenesis Panel
chr6	145,216,598	145,250,339	KRAS_ENSMUST00000032399.11	KRAS	ENSMUST00000032399.11	Homologous Cancer Gene Panel
chr7	142,683,053	142,685,453	region08	N/A	N/A	Mutagenesis Panel
chr8	43,954,521	43,956,921	region09	N/A	N/A	Mutagenesis Panel
chr9	28,648,072	28,650,472	region10	N/A	N/A	Mutagenesis Panel
chr9	120,933,299	120,960,607	CTNNB1_ENSMUST00000007130.14	CTNNB1	ENSMUST00000007130.14	Homologous Cancer Gene Panel
chr10	21,442,014	21,444,414	region11	N/A	N/A	Mutagenesis Panel
chr11	37,934,364	37,936,764	region12	N/A	N/A	Mutagenesis Panel
chr11	69,580,258	69,591,972	TRP53_ENSMUST00000108658.9	TP53	ENSMUST00000108658.9	Homologous Cancer Gene Panel

Chromosome	Start Position	End Position	Region Name	Gene Symbol	Transcript ID	Panel
chr12	80,601,542	80,603,942	region13	N/A	N/A	Mutagenesis Panel
chr13	74,030,071	74,032,471	region14	N/A	N/A	Mutagenesis Panel
chr14	13,076,171	13,078,571	region15	N/A	N/A	Mutagenesis Panel
chr15	66,779,762	66,782,162	region16	N/A	N/A	Mutagenesis Panel
chr16	72,381,580	72,383,980	region17	N/A	N/A	Mutagenesis Panel
chr17	94,009,028	94,011,428	region18	N/A	N/A	Mutagenesis Panel
chr18	81,262,078	81,264,478	region19	N/A	N/A	Mutagenesis Panel
chr19	4,618,813	4,621,213	region20	N/A	N/A	Mutagenesis Panel

Table 1. Probe regions and associated information used in the study. The Mutagenesis Panel probes (regions labeled "region01" to "region20") were purchased from TwinStrand Biosciences® as part of their Mutagenesis Assay kit. The Mutagenesis Panel includes synthetic regions designed to assess mutagenesis rates. The Homologous Cancer Gene Panel probes, targeting genes such as *Pik3ca*, *Nras*, *Kras*, *Ctnnb1*, and *Trp53*, were generously provided by Dr. Scott Kennedy from the Department of Laboratory Medicine & Pathology at the University of Washington. The Homologous Cancer Gene Panel focuses on genes commonly associated with cancer in humans, facilitating the detection of relevant mutations. Chromosome positions are based on the mm10 reference genome assembly.

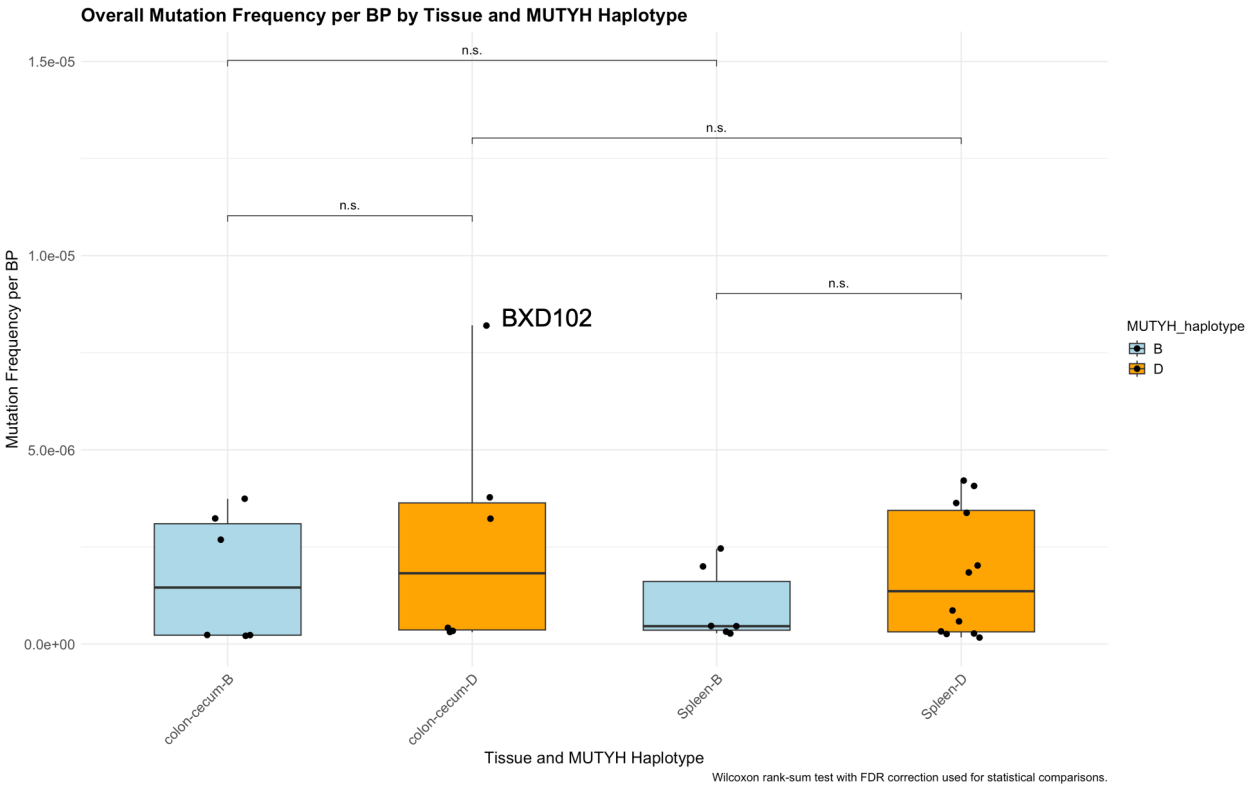


Figure S1. Overall mutation frequency per base pair by tissue type (colon-cecum and spleen) and *Mutyh* haplotype (B and D). Mutation frequencies were calculated from duplex sequencing data and normalized by the total number of base pairs sequenced per sample. Mutation burdens were generally higher in colon tissues compared to spleen tissues, with an outlier "D" mouse, strain BXD102, exhibiting markedly higher mutation frequencies in the colon-cecum group. This outlier is annotated in the plot. Statistical comparisons between groups were performed using the Wilcoxon rank-sum test with FDR correction, with "n.s." indicating non-significant differences. Error bars represent the interquartile range, and individual data points are plotted to highlight sample-level variation.

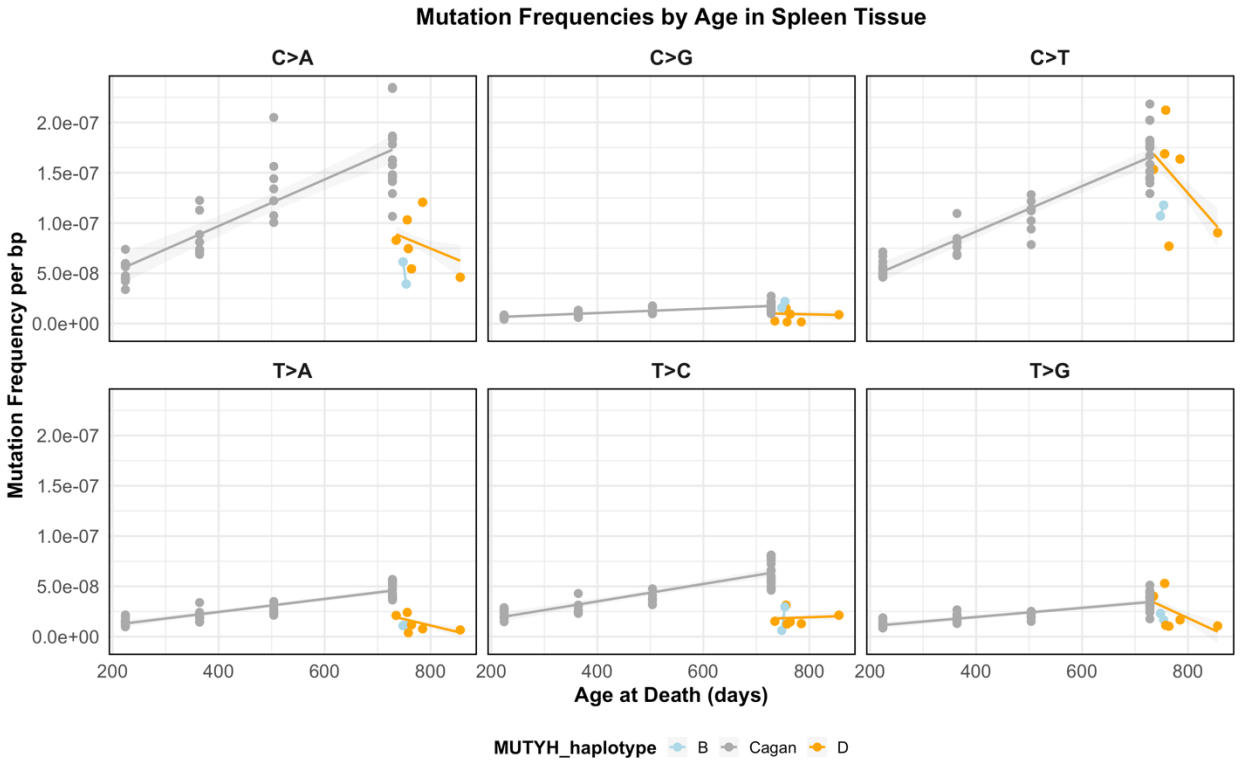


Figure S2. Mutation frequencies per base pair by age at death in spleen tissue, grouped by *Mutyh* haplotype (B, D) and compared with published colon tissue data from Cagan et al. (2022). Each panel corresponds to one of six pyrimidine-centered substitution classes. Trend lines represent linear regressions for each group, with shaded regions indicating the confidence intervals. Mutation frequencies in spleen tissues showed no significant age-dependent trends, likely due to the narrower age range of mice sequenced (735–856 days old). Data from Cagan et al. are included as a reference, covering a broader age range (180–800 days), and are shown in gray. This comparison highlights tissue-specific and age-related patterns of mutation accumulation.

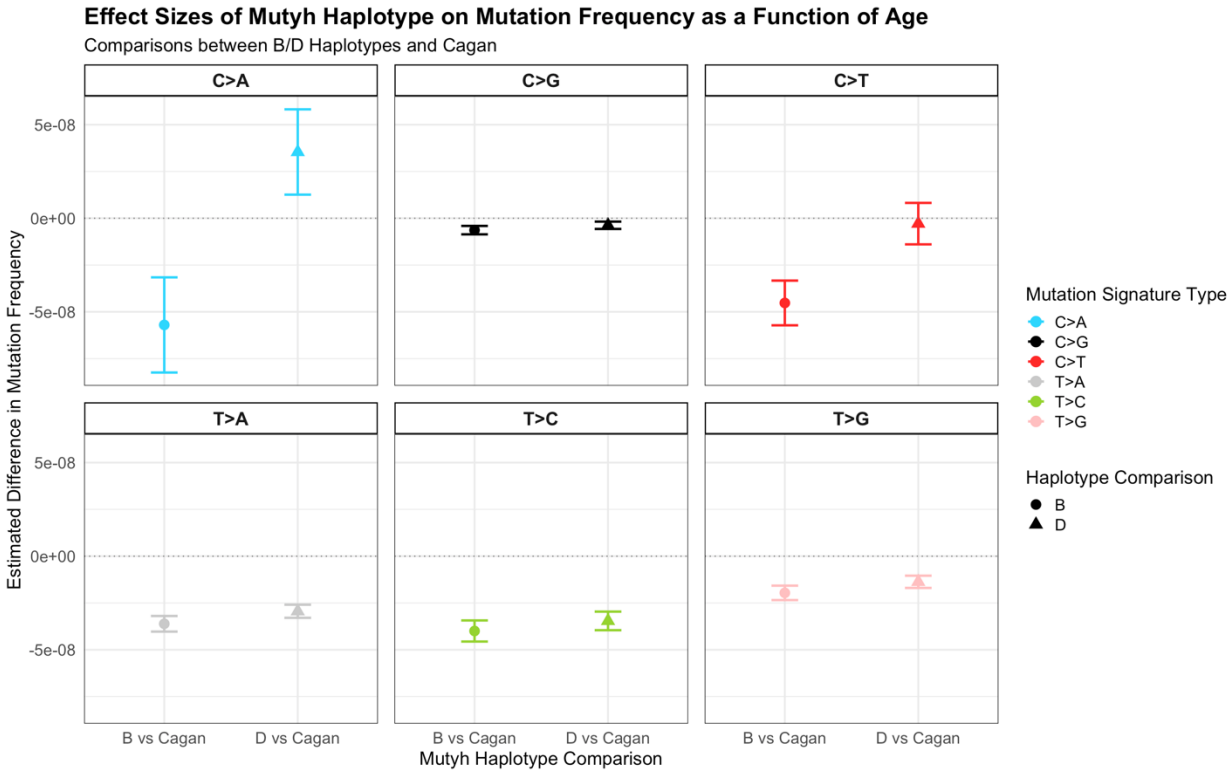


Figure S3. Effect sizes of *Mutyh* haplotype on mutation frequencies as a function of age, comparing B and D haplotypes in BXD mice to colonic crypt mutation data from Cagan et al. (2022). Each panel corresponds to a pyrimidine-centered substitution class. Points represent estimated differences in mutation frequency relative to the Cagan dataset, with error bars indicating 95% confidence intervals. A dotted horizontal line at y=0 represents no difference from the reference dataset. Significant deviations ($p < 0.01$, FDR-corrected) were observed for most substitution types, with predominantly negative effect sizes indicating lower mutation frequencies in BXD mice compared to Cagan's colon tissue data. Notably, C>A substitutions in "D" haplotype mice showed a significant positive effect size, reflecting increased mutation frequencies relative to the reference. Shapes indicate the haplotype comparison group (solid circles for "B vs Cagan" and triangles for "D vs Cagan").

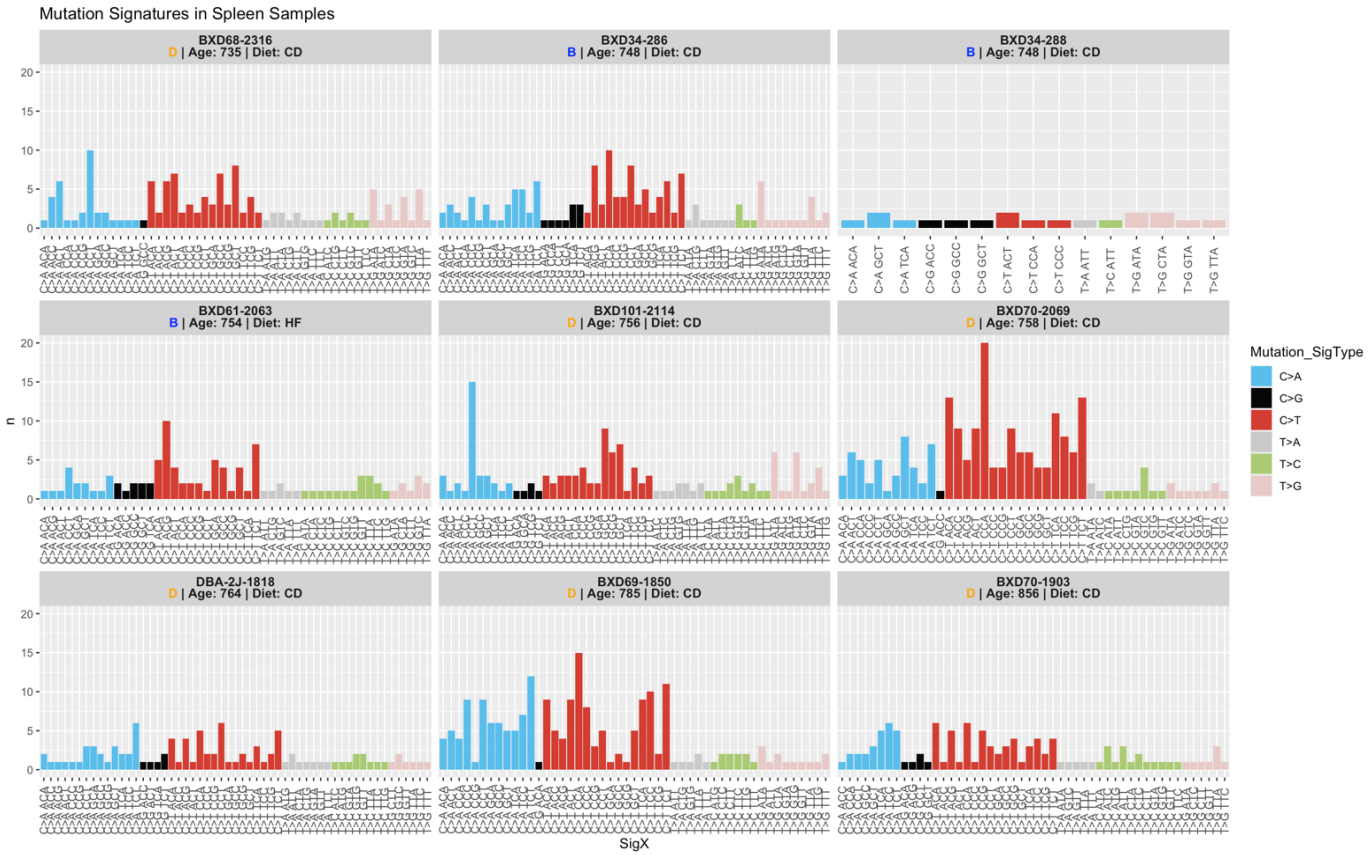


Figure S4. Mutation signatures in spleen samples from individual BXD mice, categorized by pyrimidine-centered trinucleotide substitutions and grouped by *Mutyh* haplotype (B, D), age at death (days), and diet (CD: control diet, HF: high-fat diet). Each panel represents a single spleen sample, showing the frequency of mutations for specific trinucleotide contexts (SigX) across six mutation signature types. Overall, mutation spectra demonstrated notable differences between "B" and "D" haplotype mice, reflecting distinct patterns of mutation accumulation in spleen tissues. This analysis complements PCA clustering results shown in **Figure S6**, which indicate distinct mutation profiles between spleen and colon tissues and between *Mutyh* haplotypes.

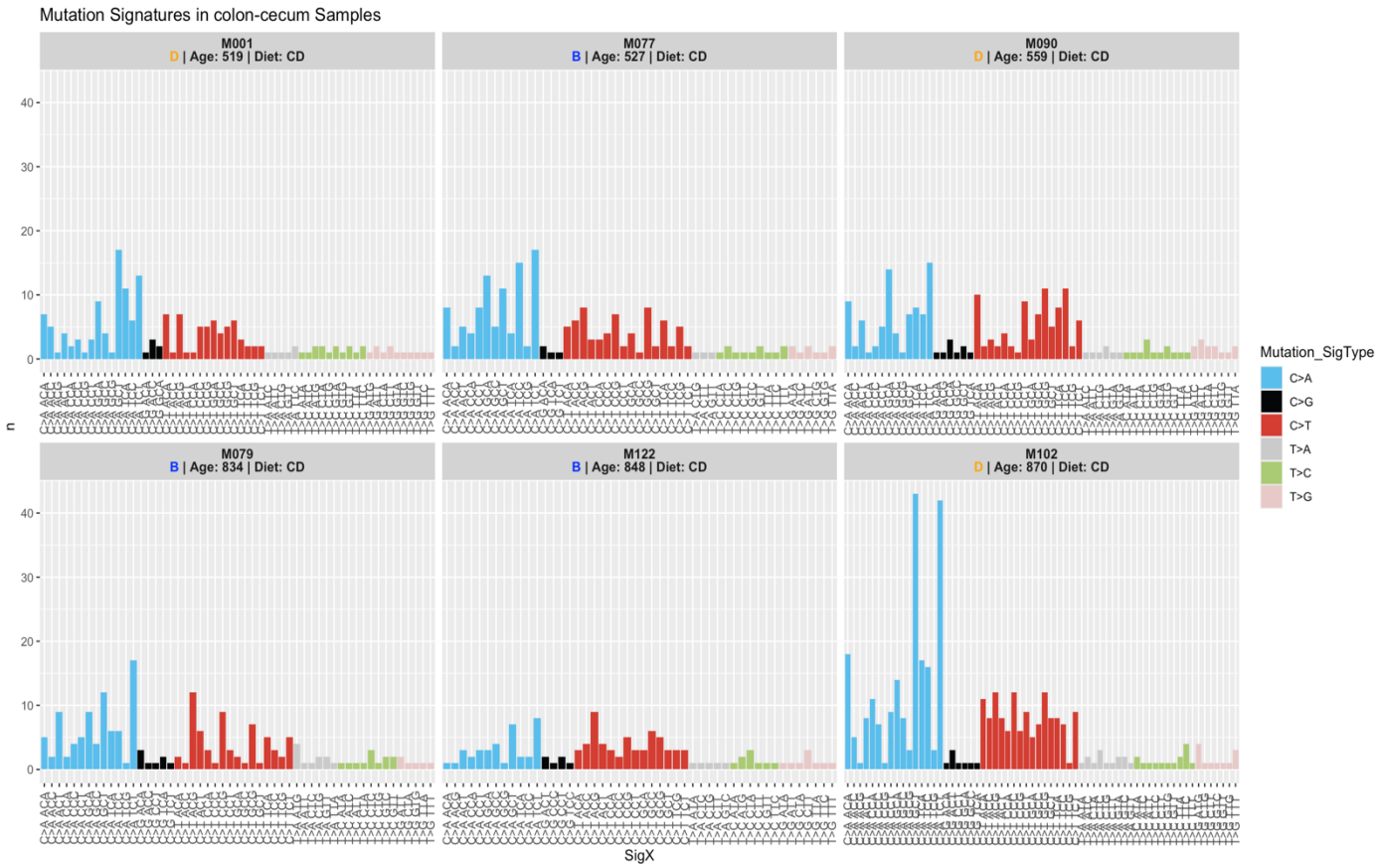


Figure S5. Mutation signatures in colon samples from individual BXD mice, categorized by pyrimidine-centered trinucleotide substitutions and grouped by *Mutyh* haplotype (B, D), age at death (days), and diet (CD: control diet, HF: high-fat diet). Each panel represents a single colon sample, showing the frequency of mutations for specific trinucleotide contexts (SigX) across six mutation signature types. This analysis complements PCA clustering results shown in **Figure S6**, which indicate distinct mutation profiles between spleen and colon tissues and between *Mutyh* haplotypes.

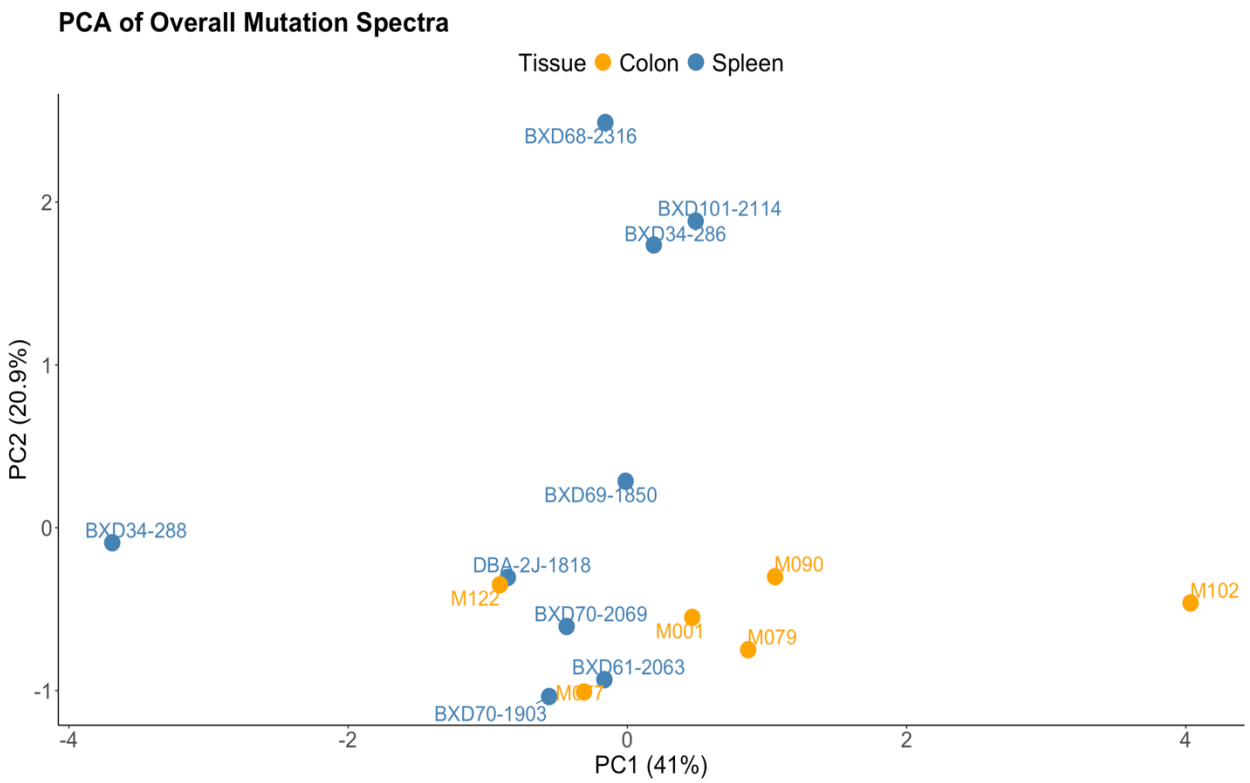


Figure S6. PCA of overall mutation spectra, comparing spleen and colon samples across BXD and *Mutyh* haplotype groups. Each point represents an individual sample, with PC1 (21.5% variance explained) and PC2 (12.2% variance explained) summarizing the dominant patterns of variation in mutation spectra. Spleen samples (blue) and colon samples (orange) cluster distinctly, reflecting tissue-specific differences in mutation composition.

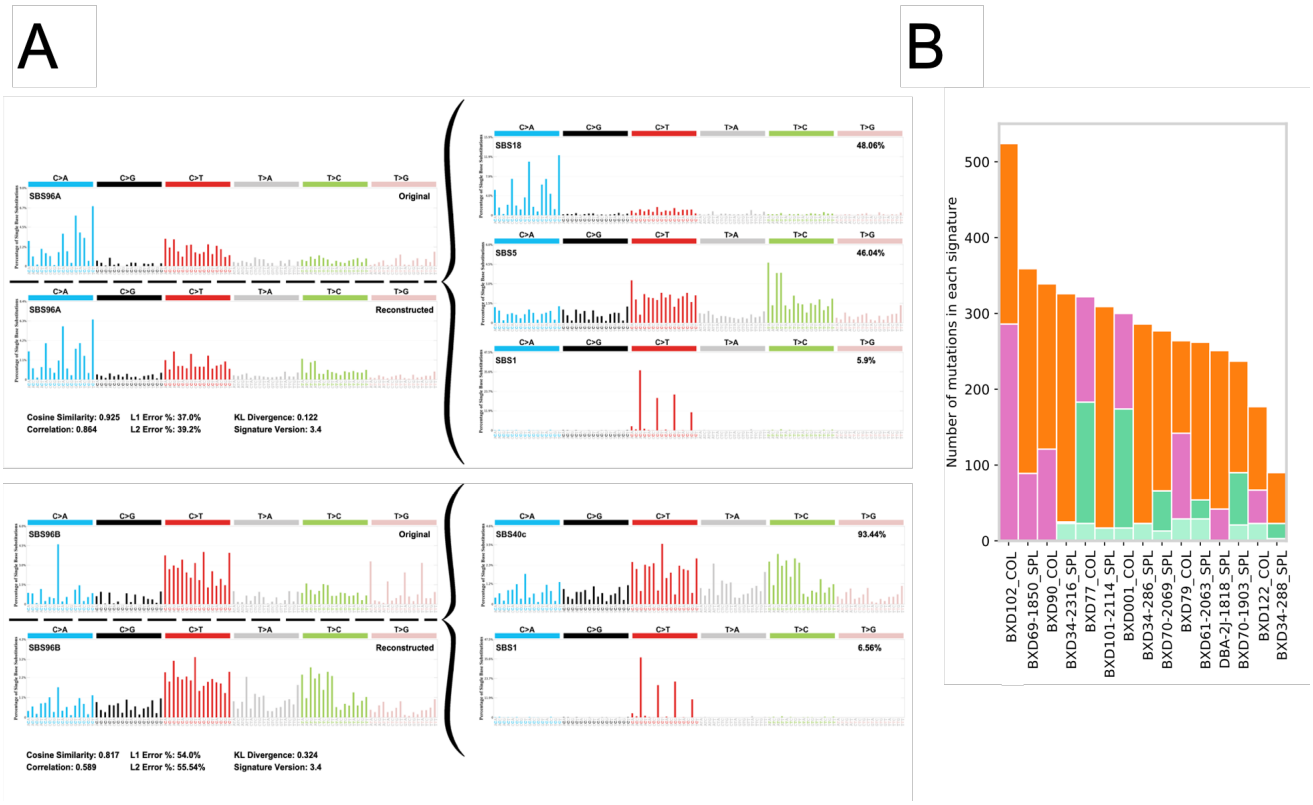


Figure S7. De novo mutational signature analysis in BXD mouse tissues. **(A)** Reconstruction of mutational spectra using de novo extracted signatures from BXD mouse data. Two dominant signatures were identified: BXD-Sig1 and BXD-Sig2. BXD-Sig1 closely aligns with COSMIC signatures SBS18 and SBS5, as shown by high cosine similarity (0.925) and correlation (0.864). BXD-Sig2 shows moderate similarity to SBS40C with a minor contribution from SBS1, with a cosine similarity of 0.817 and a correlation of 0.589. BXD-Sig1 is consistently detected across all tissues analyzed, while BXD-Sig2 exhibits greater variation, reflecting tissue-specific influences on mutational processes. **(B)** Stacked bar plot showing the number of mutations attributed to COSMIC signatures in individual samples. Sample BXD102, an aged "D" haplotype mouse (848 days old), displays the highest contribution from SBS18, associated with oxidative damage, consistent with its elevated C>A mutation burden. The widespread presence of SBS40C suggests an underlying ubiquitous mutational process across various tissues.

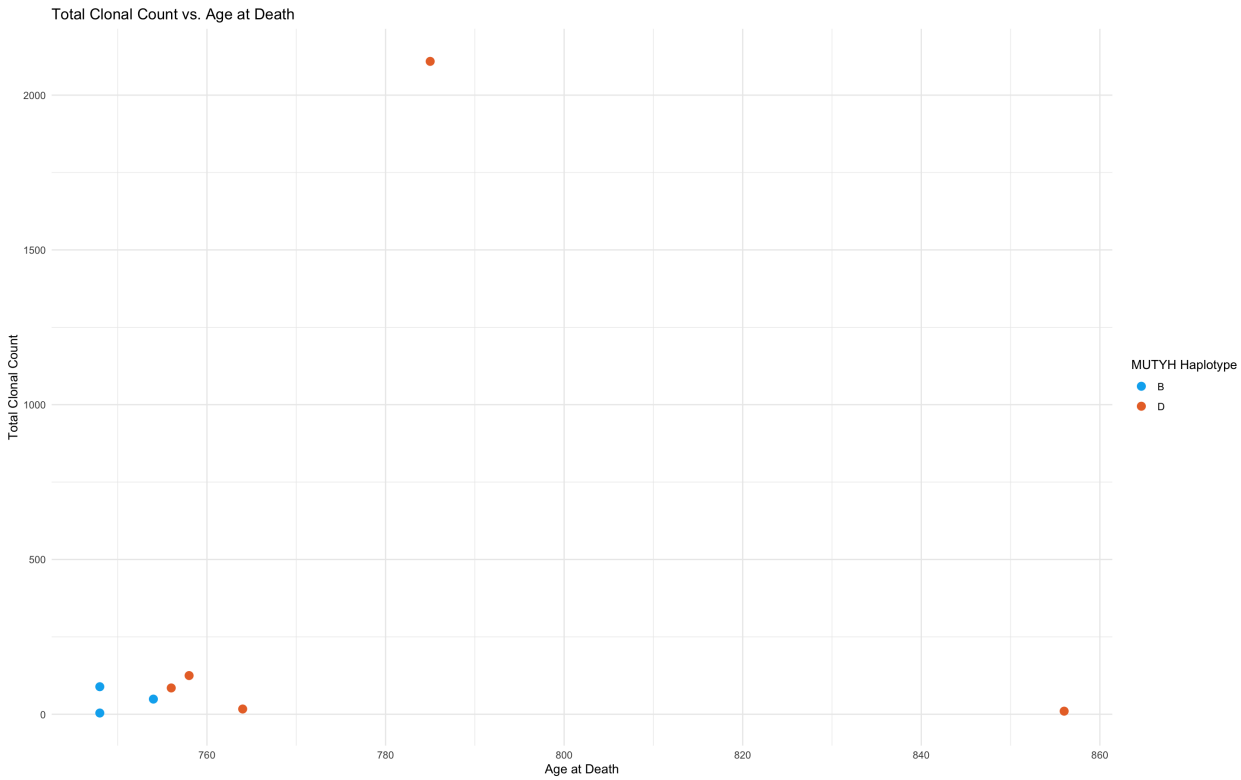


Figure S8. Total clonal counts as a function of age at death in BXD mice. The scatter plot depicts the total clonal counts (y-axis) against the age at death (x-axis) for individual samples, categorized by *Mutyh* haplotype ("B" shown in blue and "D" shown in orange). No significant correlation was observed between age at death and total clonal counts across the cohort, likely due to the narrow age range of the study cohort (735–856 days).

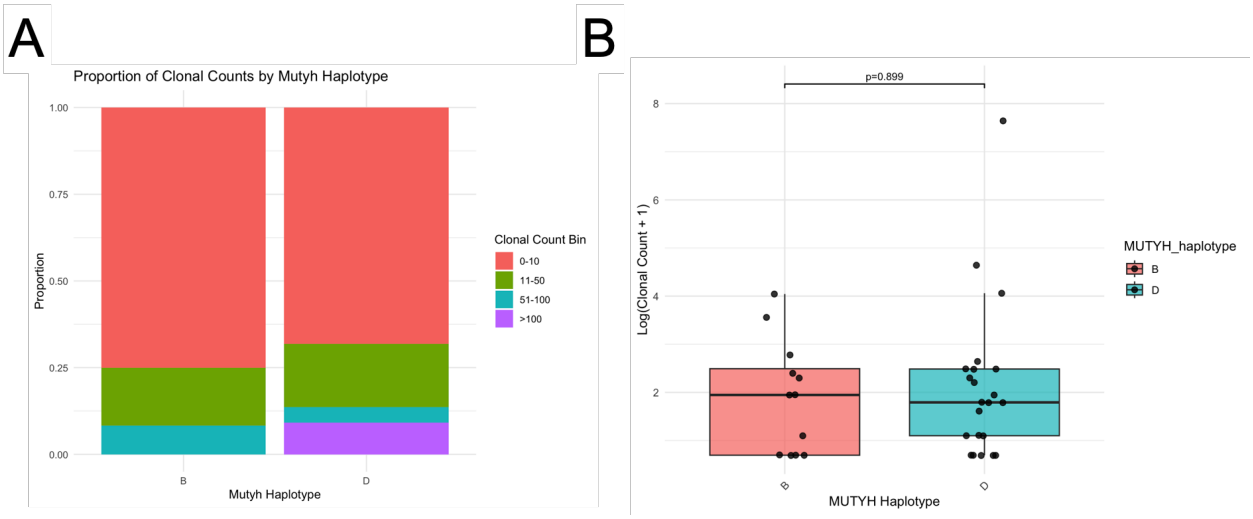


Figure S9. Comparison of clonal counts between *Mutyh* haplotypes. **(A)** Proportion of clonal counts stratified by clonal size bins (0–10, 11–50, 51–100, and >100) across B (blue) and D (orange) haplotypes. While "D" allele mice were the only group to harbor clones exceeding 100 in count, these differences were not statistically significant. **(B)** Log-transformed clonal counts by *Mutyh* haplotype. The boxplot shows no significant difference in log-transformed clonal counts between B and D haplotype groups (Wilcoxon rank-sum test, p = 0.899). Outlier clones, such as those exceeding 100 in size, did not drive overall differences, suggesting similar distributions of clonal expansion sizes across haplotypes.

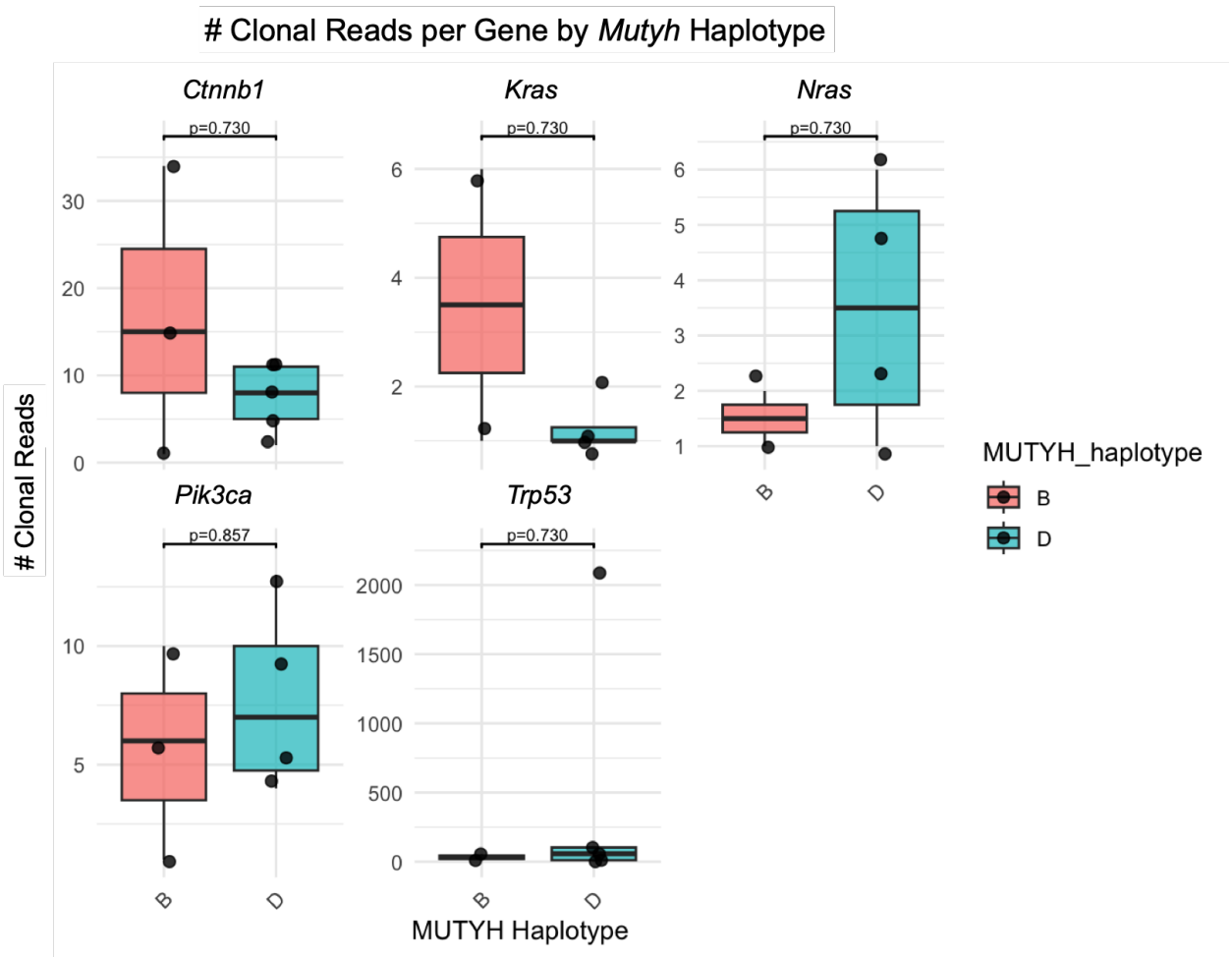


Figure S10. Clonal read counts per gene stratified by *MUTYH* haplotype. Boxplots depict the distribution of clonal reads for five homologs (*Ctnnb1*, *Kras*, *Nras*, *Pik3ca*, *Trp53*) across B (red) and D (blue) haplotype groups. No statistically significant differences were observed in clonal read counts between haplotypes for any individual gene (Wilcoxon rank-sum test, all $p > 0.7$).

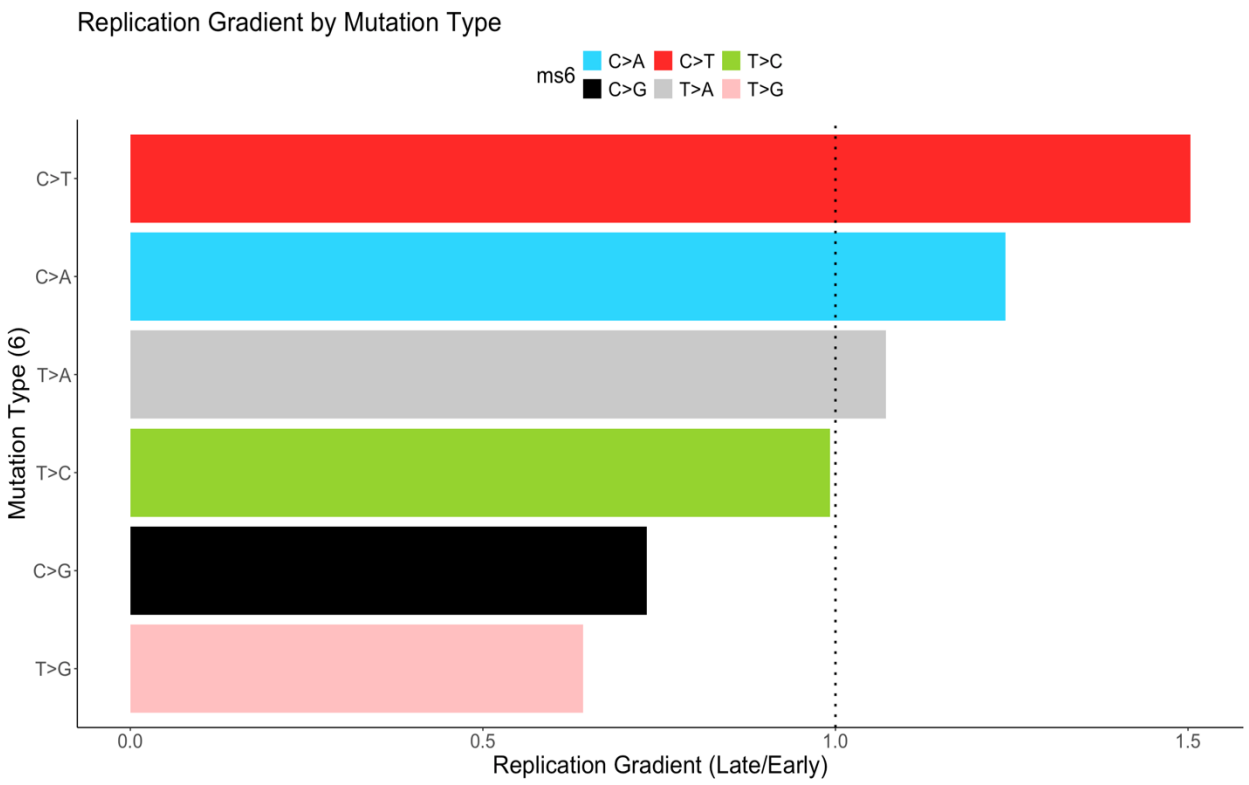


Figure S11. Replication gradient across mutation types, showing the ratio of mutation frequencies in late- versus early-replicating bins for six substitution types. Late-replicating regions demonstrated a notable enrichment of C>A and T>A mutations, consistent with prior studies linking these substitutions to oxidative damage and inefficient repair of single-stranded DNA. The dotted line at a replication gradient value of 1 indicates equal mutation frequencies halfway between late and early replicating regions. No slopes in replication timing gradients were found to be significantly deviated from the other six mutation types.

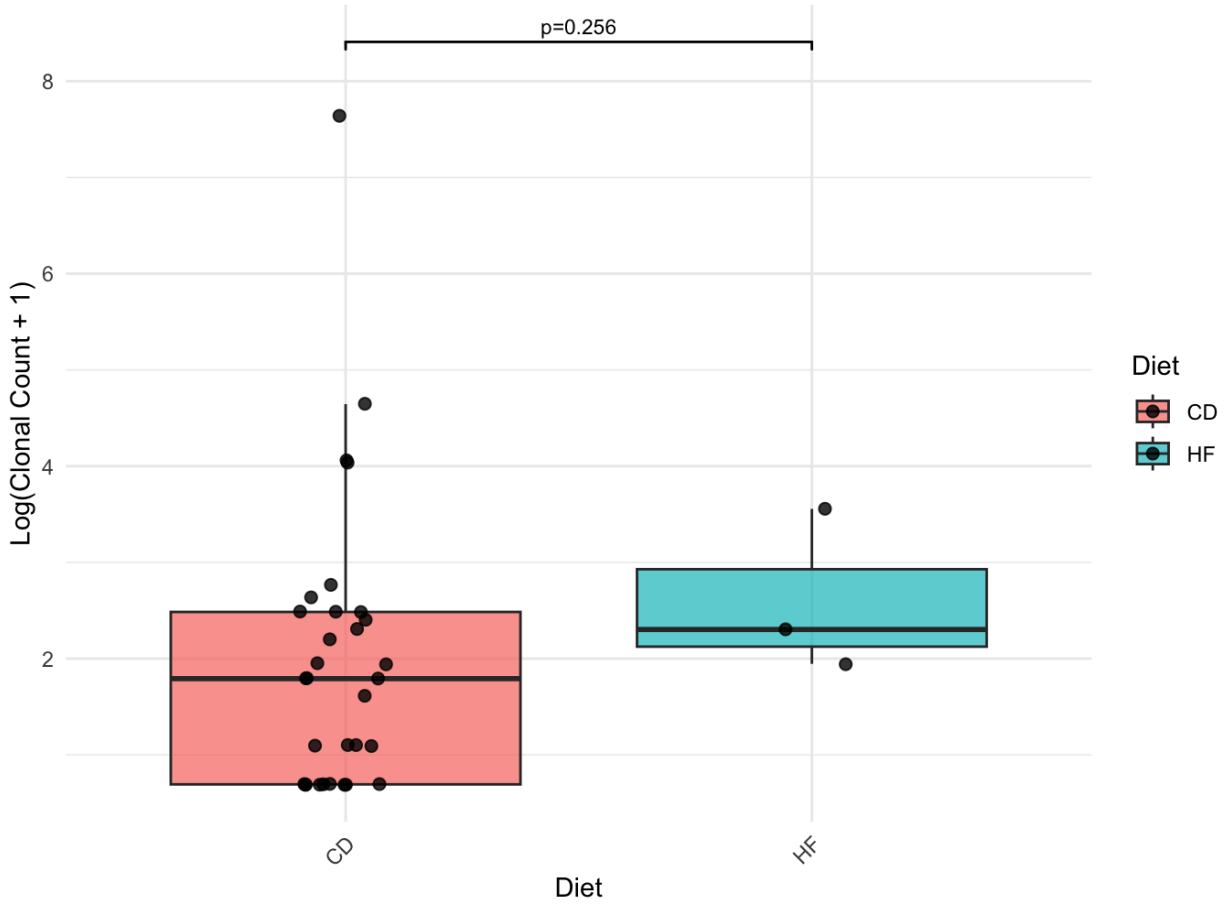


Figure S12. Clonal expansion counts by diet in mice. Boxplots illustrate the distribution of log-transformed clonal counts *(Log(Clonal Count + 1))* across two diet groups: standard chow diet (CD, red) and high-fat diet (HF, blue). Statistical analysis using the Wilcoxon rank-sum test did not reveal a significant difference between diets ($p = 0.256$). These results suggest that diet does not have a significant impact on clonal expansion size in the observed cohort. Data points represent individual mouse clones, with outliers excluded from the boxplot whiskers but plotted as individual points.

References

Aarnio M, Sankila R, Pukkala E, Salovaara R, Aaltonen LA, de la Chapelle A, Peltomäki P, Mecklin J-P, Järvinen HJ. 1999. Cancer risk in mutation carriers of DNA-mismatch-repair genes. *Int J Cancer* **81**: 214–218.

Agarwal I, Przeworski M. 2019. Signatures of replication timing, recombination, and sex in the spectrum of rare variants on the human X chromosome and autosomes. *Proc Natl Acad Sci USA* **116:** 17916–17924.

Aitken RJ. 2020. Impact of oxidative stress on male and female germ cells: implications for fertility. *Reproduction* **159**: R189–R201.

Aitken RJ, Baker MA, Sawyer D. 2003. Oxidative stress in the male germ line and its role in the aetiology of male infertility and genetic disease. *Reprod Biomed Online* **7**: 65–70.

Aitken RJ, Krausz C. 2001. Oxidative stress, DNA damage and the Y chromosome. *Reprod Camb Engl* **122**: 497–506.

Aleksandrova K, Koelman L, Rodrigues CE. 2021. Dietary patterns and biomarkers of oxidative stress and inflammation: A systematic review of observational and intervention studies. *Redox Biol* **42**: 101869.

Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, Wu Y, Boot A, Covington KR, Gordenin DA, Bergstrom EN, et al. 2020. The repertoire of mutational signatures in human cancer. *Nature* **578**: 94–101.

Al-Tassan N, Chmiel NH, Maynard J, Fleming N, Livingston AL, Williams GT, Hodges AK, Davies DR, David SS, Sampson JR, et al. 2002. Inherited variants of *MYH* associated with somatic G:C-->T:A mutations in colorectal tumors. *Nat Genet* **30**: 227–232.

Andrianova MA, Seplyarskiy VB, Terradas M, Sánchez-Heras AB, Mur P, Soto JL, Aiza G,

Kondrashov FA, Kondrashov AS, Bazykin GA, et al. 2023. Extended family with an inherited pathogenic variant in polymerase delta provides strong evidence for recessive effect of proofreading deficiency in human cells. *BioRxiv Prepr Serv Biol* 2022.07.20.500591.

Banda DM, Nuñez NN, Burnside MA, Bradshaw KM, David SS. 2017. Repair of 8-oxoG:A mismatches by the MUTYH glycosylase: Mechanism, metals and medicine. *Free Radic Biol Med* **107**: 202–215.

Barreiro RAS, Sabbaga J, Rossi BM, Achatz MIW, Bettoni F, Camargo AA, Asprino PF, A F Galante P. 2022. Monoallelic deleterious germline variants as a driver for tumorigenesis. *J Pathol* **256**: 214–222.

Beal MA, Meier MJ, LeBlanc DP, et al. 2020. Chemically induced mutations in a MutaMouse reporter gene inform mechanisms underlying human cancer mutational signatures. *Commun Biol* **3:** 438.

Beiner ME, Zhang WW, Zhang S, Gallinger S, Sun P, Narod SA. 2009. Mutations of the MYH gene do not substantially contribute to the risk of breast cancer. *Breast Cancer Res Treat* **114**: 575–578.

Belyeu JR, Sasani TA, Pedersen BS, Quinlan AR. 2021. Unfazed: parent-of-origin detection for large and small *de novo* variants. *Bioinforma Oxf Engl* **37**: 4860–4861.

Bergeron LA, Besenbacher S, Turner T, Versoza CJ, Wang RJ, Price AL, Armstrong E, Riera M, Carlson J, Chen H, et al. 2022. The Mutationathon highlights the importance of reaching standardization in estimates of pedigree-based germline mutation rates. *eLife* **11**: e73577.

Blokzijl F, de Ligt J, Jager M, Sasselli V, Roerink S, Sasaki N, Huch M, Boymans S, Kuijk E, Prins P, et al. 2016. Tissue-specific mutation accumulation in human adult stem cells

during life. *Nature* **538**: 260–264.

Briu L-M, Maric C, Cadoret J-C. 2021. Replication stress, genomic instability, and replication timing: a complex relationship. Int J Mol Sci 22: 4764.

Browning BL, Tian X, Zhou Y, Browning SR. 2021. Fast two-stage phasing of large-scale sequence data. *Am J Hum Genet* **108**: 1880–1890.

Byrska-Bishop M, Evani US, Zhao X, Basile AO, Abel HJ, Regier AA, Corvelo A, Clarke WE, Musunuri R, Nagulapalli K, et al. 2022. High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell* **185**: 3426-3440.e19.

Chao EC, Lipkin SM. 2006. Molecular models for the tissue specificity of DNA mismatch repair-deficient carcinogenesis. *Nucleic Acids Res* **34**: 840–852.

Chen C, Qi H, Shen Y, Pickrell J, Przeworski M. 2017. Contrasting determinants of mutation rates in germline and soma. *Genetics* **207**: 255–267.

Chen L, Bowen PE, Berzy D, Aryee F, Stacewicz-Sapuntzakis M, Riley RE. 1999. Diet modification affects DNA oxidative damage in healthy humans. *Free Radic Biol Med* **26**: 695–703.

Chen S, Zhou Y, Chen Y, Gu J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**: i884–i890.

Chin DWL, Yoshizato T, Virding Culleton S, Grasso F, Barbachowska M, Ogawa S, et al. 2022. Aged healthy mice acquire clonal hematopoiesis mutations. *Blood* **139**: 629–634.

Cornejo-Páramo P, Petrova V, Zhang X, Young RS, Wong ES. 2024. Emergence of enhancers at late DNA replicating regions. *Nat Commun* **15**: 3451

Crane GM, Liu Y-C, Chadburn A. 2021. Spleen: Development, anatomy and reactive lymphoid proliferations. *Semin Diagn Pathol* **38**: 112–124.

David SS, O'Shea VL, Kundu S. 2007. Base-excision repair of oxidative DNA damage. *Nature* **447**: 941–950.

Dunlop MG, Farrington SM, Carothers AD, Wyllie AH, Sharp L, Burn J, Liu B, Kinzler KW, Vogelstein B. 1997. Cancer Risk Associated with Germline DNA Mismatch Repair Gene Mutations. *Hum Mol Genet* **6**: 105–110.

Elledge SJ, Amon A. 2002. The BRCA1 suppressor hypothesis: An explanation for the tissue-specific tumor development in BRCA1 patients. *Cancer Cell* **1**: 129–132.

Ellis P, Moore L, Sanders MA, Butler TM, Brunner SF, Lee-Six H, Osborne R, Farr B, Coorens THH, Lawson ARJ, et al. 2021. Reliable detection of somatic mutations in solid tissues by laser-capture microdissection and low-input DNA sequencing. *Nat Protoc* **16**: 841–871.

Ewels P, Magnusson M, Lundin S, Käller M. 2016. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinforma Oxf Engl* **32**: 3047–3048.

Fearon ER. 1997. Human cancer syndromes: clues to the origin and nature of cancer. *Science* **278**: 1043–1050.

Fiala C, Diamandis EP. 2020. Can a broad molecular screen based on circulating tumor DNA aid in early cancer detection? *J Appl Lab Med* **5**: 1372–1377.

Fleischmann C, Peto J, Cheadle J, Shah B, Sampson J, Houlston RS. 2004. Comprehensive analysis of the contribution of germline *MYH* variation to early-onset colorectal cancer. *Int J Cancer* **109**: 554–558.

Gao Z, Moorjani P, Sasani TA, Pedersen BS, Quinlan AR, Jorde LB, Amster G, Przeworski M. 2019. Overlooked roles of DNA damage and maternal age in generating human germline mutations. *Proc Natl Acad Sci* **116**: 9491–9500.

Garcia MU, Hanssen F, Pedersen AS, Gabernet G, WackerO, SusiJo, Talbot A, James C, Ergüner B, Peltzer A, et al. 2023. nf-core/sarek: Sarek 3.4.0 - Pårtetjåkko. https://zenodo.org/records/10138031 (Accessed August 29, 2024).

Goldberg ME, Harris K. 2022. Mutational signatures of replication timing and epigenetic modification persist through the global divergence of mutation spectra across the great ape phylogeny. *Genome Biol Evol* **14**(1).

Goode EL, Ulrich CM, Potter JD. 2002. Polymorphisms in DNA Repair Genes and Associations with Cancer Risk. *Cancer Epidemiol Biomarkers Prev* **11**: 1513–1530.

Guarinos C, Juárez M, Egoavil C, Rodríguez-Soler M, Pérez-Carbonell L, Salas R, Cubiella J, Rodríguez-Moranta F, de-Castro L, Bujanda L, et al. 2014. Prevalence and characteristics of *MUTYH*-associated polyposis in patients with multiple adenomatous and serrated polyps. *Clin Cancer Res Off J Am Assoc Cancer Res* **20**: 1158–1168.

Harland C, Charlier C, Karim L, Cambisano N, Deckers M, Mni M, et al. 2016. Frequency of mosaicism points towards mutation-prone early cleavage cell divisions in cattle. *BioRxiv Prepr Serv Biol* 2016.079863.

Hayashi H, Tominaga Y, Hirano S, McKenna AE, Nakabeppu Y, Matsumoto Y. 2002. Replication-associated repair of adenine:8-oxoguanine mispairs by MYH. *Curr Biol CB* **12**: 335–339.

Huang AY, Xu X, Ye AY, Wu Q, Yan L, Zhao B, Yang X, He Y, Wang S, Zhang Z, et al. 2014. Postzygotic single-nucleotide mosaicisms in whole-genome sequences of clinically unremarkable individuals. *Cell Res* **24**: 1311–1327.

Hutchcraft ML, Gallion HH, Kolesar JM. 2021. *MUTYH* as an Emerging Predictive Biomarker in Ovarian Cancer. *Diagn Basel Switz* **11**: 84.

Islam SMA, Díaz-Gay M, Wu Y, Barnes M, Vangara R, Bergstrom EN, He Y, Vella M, Wang J, Teague JW, et al. 2022. Uncovering novel mutational signatures by *de novo* extraction with SigProfilerExtractor. *Cell Genomics* **2**: None.

Jia X, Burugula BB, Chen V, Lemons RM, Jayakody S, Maksutova M, Kitzman JO. 2021. Massively parallel functional testing of *MSH2* missense variants conferring Lynch syndrome risk. *Am J Hum Genet* **108**: 163–175.

Jónsson H, Sulem P, Arnadottir GA, Pálsson G, Eggertsson HP, Kristmundsdottir S, Zink F, Kehr B, Hjorleifsson KE, Jensson BÖ, et al. 2018. Multiple transmissions of de novo mutations in families. *Nat Genet* **50**: 1674–1680.

Jónsson H, Sulem P, Kehr B, Kristmundsdottir S, Zink F, Hjartarson E, Hardarson MT, Hjorleifsson KE, Eggertsson HP, Gudjonsson SA, et al. 2017. Parental influence on human germline *de novo* mutations in 1,548 trios from Iceland. *Nature* **549**: 519–522.

Kaplanis J, Ide B, Sanghvi R, Neville M, Danecek P, Coorens T, Prigmore E, Short P, Gallone G, McRae J, et al. 2022. Genetic and chemotherapeutic influences on germline hypermutation. *Nature* **605**: 503–508.

Kennedy SR, Schmitt MW, Fox EJ, Kohrn BF, Salk JJ, Ahn EH, Prindle MJ, Kuong KJ, Shen J-C, Risques R-A, et al. 2014. Detecting ultralow-frequency mutations by Duplex Sequencing. *Nat Protoc* **9**: 2586–2606.

Kennedy SR, Zhang Y, Risques RA. 2019. Cancer-associated mutations but no cancer: Insights into the early steps of carcinogenesis and implications for early cancer detection. *Trends Cancer* **5**: 531–540.

Komine K, Shimodaira H, Takao M, Soeda H, Zhang X, Takahashi M, Ishioka C. 2015. Functional Complementation Assay for 47 *MUTYH* Variants in a *MutY*-Disrupted

Escherichia coli Strain. *Hum Mutat* **36**: 704–711.

Kong A, Masson G, Frigge ML, Gylfason A, Zusmanovich P, Thorleifsson G, Olason PI, Ingason A, Steinberg S, Rafnar T, et al. 2008. Detection of sharing by descent, long-range phasing and haplotype imputation. *Nat Genet* **40**: 1068–1075.

Krokan HE, Bjørås M. 2013. Base excision repair. *Cold Spring Harb Perspect Biol* **5**: a012583.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinforma Oxf Engl* **25**: 1754–1760.

Luzadder MM, Minko IG, Vartanian VL, Davenport M, Fedorov LM, McCullough AK, et al. 2024. The distinct roles of NEIL1 and XPA in limiting aflatoxin B_1-induced mutagenesis in mice. *Mol Cancer Res* **22**: OF1–13.

Ma H, Wang J, Abdel-Rahman SZ, Boor PJ, Khan MF. 2008. Oxidative DNA damage and its repair in rat spleen following subchronic exposure to aniline. *Toxicol Appl Pharmacol* **233**: 247–253.

Martincorena I, Roshan A, Gerstung M, Ellis P, Van Loo P, McLaren S, et al. 2015. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348**: 880–886.

Matullo G, Dunning AM, Guarrera S, Baynes C, Polidoro S, Garte S, Autrup H, Malaveille C, Peluso M, Airoldi L, et al. 2006. DNA repair polymorphisms and cancer risk in non-smokers in a cohort study. *Carcinogenesis* **27**: 997–1007.

Maura F, Coffey DG, Stein CK, Braggio E, Ziccheddu B, Sharik ME, et al. 2023. The Vk*MYC mouse model recapitulates human multiple myeloma evolution and genomic diversity. *bioRxiv Prepr Serv Biol* 2023.07.25.550482.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K,

Altshuler D, Gabriel S, Daly M, et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**: 1297–1303.

Minko IG, Luzadder MM, Vartanian VL, Rice SPM, Nguyen MM, Sanchez-Contreras M, et al. 2024. Frequencies and spectra of aflatoxin B1-induced mutations in liver genomes of NEIL1-deficient mice as revealed by duplex sequencing. *NAR Mol Med* **1**: ugae006.

Morak M, Heidenreich B, Keller G, Hampel H, Laner A, de la Chapelle A, Holinski-Feder E. 2014. Biallelic *MUTYH* mutations can mimic Lynch syndrome. *Eur J Hum Genet* **22**: 1334–1337.

Muralidharan S, Mandrekar P. 2013. Cellular stress response and innate immune signaling: integrating pathways in host defense and inflammation. *J Leukoc Biol* **94**: 1167–1184.

Nagel ZD, Margulies CM, Chaim IA, McRee SK, Mazzucato P, Ahmad A, Abo RP, Butty VL, Forget AL, Samson LD. 2014. Multiplexed DNA repair assays for multiple lesions and multiple doses via transcription inhibition and transcriptional mutagenesis. *Proc Natl Acad Sci* **111**: E1823–E1832.

Narasimhan VM, Rahbari R, Scally A, Wuster A, Mason D, Xue Y, Wright J, Trembath RC, Maher ER, van Heel DA, et al. 2017. Estimating the human mutation rate from autozygous segments reveals population differences in human mutational processes. *Nat Commun* **8**: 303.

Nielsen M, Franken PF, Reinards THCM, Weiss MM, Wagner A, van der Klift H, Kloosterman S, Houwing-Duistermaat JJ, Aalfs CM, Ausems MGEM, et al. 2005. Multiplicity in polyp count and extracolonic manifestations in 40 Dutch patients with *MYH* associated polyposis coli (MAP). *J Med Genet* **42**: e54.

Nielsen M, Joerink-van de Beld MC, Jones N, Vogt S, Tops CM, Vasen HFA, Sampson JR, Aretz S, Hes FJ. 2009. Analysis of *MUTYH* genotypes and colorectal phenotypes in patients With MUTYH-associated polyposis. *Gastroenterology* **136**: 471–476.

Nielsen M, Morreau H, Vasen HFA, Hes FJ. 2011. MUTYH-associated polyposis (MAP). *Crit Rev Oncol Hematol* **79**: 1–16.

Nieuwenhuis MH, Vogt S, Jones N, Nielsen M, Hes FJ, Sampson JR, Aretz S, Vasen HFA. 2012. Evidence for accelerated colorectal adenoma-carcinoma progression in *MUTYH*-associated polyposis? *Gut* **61**: 734–738.

Noyes MD, Harvey WT, Porubsky D, Sulovari A, Li R, Rose NR, Audano PA, Munson KM, Lewis AP, Hoekzema K, et al. 2022. Familial long-read sequencing increases yield of de novo mutations. *Am J Hum Genet* **109**: 631–646.

Obtulowicz T, Swoboda M, Speina E, Gackowski D, Rozalski R, Siomek A, et al. 2010. Oxidative stress and 8-oxoguanine repair are enhanced in colon adenoma and carcinoma patients. *Mutagenesis* **25**: 463–471.

Ohno M, Sakumi K, Fukumura R, Furuichi M, Iwasaki Y, Hokama M, Ikemura T, Tsuzuki T, Gondo Y, Nakabeppu Y. 2014. 8-oxoguanine causes spontaneous de novo germline mutations in mice. *Sci Rep* **4**: 4689.

Pedersen BS, Quinlan AR. 2018. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics* **34**: 867–868.

Peterlongo P, Mitra N, Sanchez de Abajo A, de la Hoya M, Bassi C, Bertario L, Radice P, Glogowski E, Nafa K, Caldes T, et al. 2006. Increased frequency of disease-causing *MYH* mutations in colon cancer families. *Carcinogenesis* **27**: 2243–2249.

Pich O, Reyes-Salazar I, Gonzalez-Perez A, Lopez-Bigas N. 2022. Discovering the drivers of

clonal hematopoiesis. *Nat Commun* **13**: 4267.

Pilati C, Shinde J, Alexandrov LB, Assié G, André T, Hélias-Rodzewicz Z, Ducoudray R, Le Corre D, Zucman-Rossi J, Emile J-F, et al. 2017. Mutational signature analysis identifies *MUTYH* deficiency in colorectal cancers and adrenocortical carcinomas. *J Pathol* **242**: 10–15.

Pope BD, Ryba T, Dileep V, Yue F, Wu W, Denas O, et al. 2014. Topologically associating domains are stable units of replication-timing regulation. *Nature* **515**: 402–405.

Porubsky D, Dashnow H, Sasani TA, Logsdon GA, Hallast P, Noyes MD, et al. 2024. A familial, telomere-to-telomere reference for human de novo mutation and recombination from a four-generation pedigree. *bioRxiv Prepr Serv Biol* 2024.08.05.606142.

Pourebrahim R, Montoya RH, Akiyama H, Ostermann L, Khazaei S, Muftuoglu M, et al. 2024. Age-specific induction of mutant p53 drives clonal hematopoiesis and acute myeloid leukemia in adult mice. *Cell Rep Med* **5**: 101558.

Raetz AG, Xie Y, Kundu S, Brinkmeyer MK, Chang C, David SS. 2012. Cancer-associated variants and a common polymorphism of *MUTYH* exhibit reduced repair of oxidative DNA damage using a GFP-based assay in mammalian cells. *Carcinogenesis* **33**: 2301–2309.

Raetz AG, David SS. 2019. When you're strange: Unusual features of the MUTYH glycosylase and implications in cancer. *DNA Repair (Amst)* **80**: 16–25.

Randall MP, Egolf LE, Vaksman Z, Samanta M, Tsang M, Groff D, Evans JP, Rokita JL, Layeghifard M, Shlien A, et al. 2023. *BARD1* germline variants induce haploinsufficiency and DNA repair defects in neuroblastoma. *BioRxiv Prepr Serv Biol* 2023.01.31.525066.

Riva L, Pandiri AR, Li YR, Droop A, Hewinson J, Quail MA, et al. 2020. The mutational signature profile of known and suspected human carcinogens in mice. *Nat Genet* **52**: 1189–1197.

Robinson PS, Thomas LE, Abascal F, Jung H, Harvey LMR, West HD, Olafsson S, Lee BCH, Coorens THH, Lee-Six H, et al. 2022. Inherited *MUTYH* mutations cause elevated somatic mutation rates and distinctive mutational signatures in normal human cells. *Nat Commun* **13**: 3949.

Roy S, Sleiman MB, Jha P, Ingels JF, Chapman CJ, McCarty MS, et al. 2021. Gene-by-environment modulation of lifespan and weight gain in the murine BXD family. *Nat Metab* **3**: 1217–1227.

Sasani TA, Ashbrook DG, Beichman AC, Lu L, Palmer AA, Williams RW, Pritchard JK, Harris K. 2022. A natural mutator allele shapes mutation spectrum variation in mice. *Nature* **605**: 497–502.

Sasani TA, Quinlan AR, Harris K. 2024. Epistasis between mutator alleles contributes to germline mutation rate variability in laboratory mice. *eLife* **12**. https://elifesciences.org/reviewed-preprints/89096.

Salk JJ, Loubet-Senear K, Maritschnegg E, Valentine CC, Williams LN, Higgins JE, et al. 2019. Ultra-sensitive *TP53* sequencing for cancer detection reveals progressive clonal selection in normal tissue over a century of human lifespan. *Cell Rep* **28**: 132–144.e3.

Scarbrough PM, Weber RP, Iversen ES, Brhane Y, Amos CI, Kraft P, Hung RJ, Sellers TA, Witte JS, Pharoah P, et al. 2016. A Cross-Cancer Genetic Association Analysis of the DNA Repair and DNA Damage Signaling Pathways for Lung, Ovary, Prostate, Breast, and Colorectal Cancer. *Cancer Epidemiol Biomarkers Prev* **25**: 193–200.

Sherwood K, Ward JC, Soriano I, Martin L, Campbell A, Rahbari R, Kafetzopoulos I, Sproul D, Green A, Sampson JR, et al. 2023. Germline de novo mutations in families with Mendelian cancer syndromes caused by defects in DNA repair. *Nat Commun* **14**: 3636.

Smith CG, West H, Harris R, Idziaszczyk S, Maughan TS, Kaplan R, Richman S, Quirke P, Seymour M, Moskvina V, et al. 2013a. Role of the Oxidative DNA Damage Repair Gene *OGG1* in Colorectal Tumorigenesis. *JNCI J Natl Cancer Inst* **105**: 1249–1253.

Smith TB, Dun MD, Smith ND, Curry BJ, Connaughton HS, Aitken RJ. 2013b. The presence of a truncated base excision repair pathway in human spermatozoa that is mediated by OGG1. *J Cell Sci* **126**: 1488–1497.

Stendahl AM, Sanghvi R, Peterson S, Ray K, Lima AC, Rahbari R, et al. 2023. A naturally occurring variant of MBD4 causes maternal germline hypermutation in primates. *Genome Res* **33**: 2053–2059.

Tian X, Browning BL, Browning SR. 2019. Estimating the Genome-wide Mutation Rate with Three-Way Identity by Descent. *Am J Hum Genet* **105**: 883–893.

Tian X, Cai R, Browning SR. 2022. Estimating the genome-wide mutation rate from thousands of unrelated individuals. *Am J Hum Genet* **109**: 2178–2184.

Tomkova M, Tomek J, Kriaucionis S, Schuster-Böckler B. 2018. Mutational signature distribution varies with DNA replication timing and strand asymmetry. *Genome Biol* **19**: 129.

Tsyglakova M, McDaniel D, Hodes GE. 2019. Immune mechanisms of stress susceptibility and resilience: Lessons from animal models. *Front Neuroendocrinol* **54**: 100771.

Tudek B, Speina E. 2012. Oxidatively damaged DNA and its repair in colon carcinogenesis. *Mutat Res* **736**: 82–92.

Van der Auwera GA. 2020. *Genomics in the cloud: using Docker, GATK, and WDL in Terra*. First edition. O'Reilly Media, Sebastopol, CA.

Vermeij WP, Dollé MET, Reiling E, Jaarsma D, Payan-Gomez C, Bombardieri CR, et al. 2016. Restricted diet delays accelerated ageing and genomic stress in DNA-repair-deficient mice. *Nature* **537**: 427–431.

Villy M-C, Masliah-Planchon J, Buecher B, Beaulaton C, Vincent-Salomon A, Stoppa-Lyonnet D, Colas C. 2022. Endometrial cancer may be part of the *MUTYH*-associated polyposis cancer spectrum. *Eur J Med Genet* **65**: 104385.

Vogt S, Jones N, Christian D, Engel C, Nielsen M, Kaufmann A, Steinke V, Vasen HF, Propping P, Sampson JR, et al. 2009. Expanded extracolonic tumor spectrum in *MUTYH*-associated polyposis. *Gastroenterology* **137**: 1976-1985.e1–10.

Wei Q, Zhan X, Zhong X, Liu Y, Han Y, Chen W, Li B. 2015. A Bayesian framework for *de novo* mutation calling in parents-offspring trios. *Bioinformatics* **31**: 1375–1381.

Win AK, Reece JC, Dowty JG, Buchanan DD, Clendenning M, Rosty C, Southey MC, Young JP, Cleary SP, Kim H, et al. 2016. Risk of extracolonic cancers for people with biallelic and monoallelic mutations in *MUTYH*. *Int J Cancer* **139**: 1557–1563.

Wong WSW, Solomon BD, Bodian DL, Kothiyal P, Eley G, Huddleston KC, et al. 2016. New observations on maternal age effect on germline de novo mutations. *Nat Commun* **7**: 10486.

Woods RD, O'Shea VL, Chu A, Cao S, Richards JL, Horvath MP, David SS. 2016. Structure and stereochemistry of the base excision repair glycosylase MutY reveal a mechanism similar to retaining glycosidases. *Nucleic Acids Res* **44**: 801–810.

Yokota M, Tatsumi N, Nathalang O, Yamada T, Tsuda I. 1999. Effects of heparin on polymerase

chain reaction for blood white cells. *J Clin Lab Anal* **13**: 133–140.

Yurgelun MB, Allen B, Kaldate RR, Bowles KR, Judkins T, Kaushik P, Roa BB, Wenstrup RJ, Hartman A-R, Syngal S. 2015. Identification of a Variety of Mutations in Cancer Predisposition Genes in Patients With Suspected Lynch Syndrome. *Gastroenterology* **149**: 604-613.e20.

Zhang Y, Newcomb PA, Egan KM, Titus-Ernstoff L, Chanock S, Welch R, Brinton LA, Lissowska J, Bardin-Mikolajczak A, Peplonska B, et al. 2006. Genetic polymorphisms in base-excision repair pathway genes and risk of breast cancer. *Cancer Epidemiol Biomark Prev Publ Am Assoc Cancer Res Cosponsored Am Soc Prev Oncol* **15**: 353–358.

Zhou Y, Browning SR, Browning BL. 2020. A Fast and Simple Method for Detecting Identity-by-Descent Segments in Large-Scale Data. *Am J Hum Genet* **106**: 426–437.