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

Candice L. Young[1,2*], Annabel C. Beichman[1*], David Mas-Ponte[1], Shelby L. Hemker[3], Luke Zhu[4], Jacob O. Kitzman[3], Brian Shirts[5], and Kelley Harris[1,6**]

[1]Department of Genome Sciences, University of Washington

[2]Department of Molecular and Cellular Biology, University of Washington

[3]Department of Human Genetics, University of Michigan

[4]Department of Bioengineering, University of Washington

[5]Department of Laboratory Medicine and Pathology, University of Washington

[6]Computational Biology Division, Fred Hutchinson Cancer Center

[*]These authors contributed equally to this work

[**]Corresponding author

Kelley Harris (harriske@uw.edu)
Department of Genome Sciences
William H. Foege Hall
3720 15th Ave NE
Seattle, WA 98195

Running head: A human maternal mutator phenotype linked to *MUTYH*

Abstract

Variation in DNA repair genes can increase cancer risk by elevating the rate of oncogenic mutation. Defects in one such gene, *MUTYH*, are known to elevate the incidence of colorectal cancer in a recessive Mendelian manner, and some evidence has also linked *MUTYH* to elevated incidence of other cancers as well as elevated mutation rates in normal somatic and germline cells. Here, we use whole genome sequencing to measure germline de novo mutation rates in a large extended family affected by pathogenic *MUTYH* variation and a history of colorectal cancer. Although this family's genotype, p.Y179C/V234M (c.536A>G/700G>A on transcript NM_001128425), contains a variant with conflicting functional interpretations, we use an *in vitro* cell line assay to determine that it partially attenuates MUTYH's function. In the children of mothers affected by the Y179C/V234M genotype, we identify an elevation of the C>A mutation rate that is weaker than mutator effects previously reported to be caused by other pathogenic *MUTYH* genotypes, suggesting that mutation rates in normal tissues may be useful for classifying cancer-associated variation along a continuum of severity. Surprisingly, we detect no significant elevation of the C>A mutation rate in children born to a father with the same biallelic *MUTYH* genotype, despite calculating that we should have adequate power to detect such a mutator effect. This suggests that the oxidative stress repaired by *MUTYH* may contribute more to female reproductive aging than male reproductive aging in the general population.

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).

Some recent studies (Sherwood et al. 2023; Andrianova et al. 2023; Stendahl et al. 2023; Kaplanis et al. 2022) have paid particular attention to the impact of DNA repair deficiencies on the germline because even modestly elevated germline mutation rates can impact 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, 2023). 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. 2013), 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 base excision repair (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).

MUTYH-associated polyposis (MAP) follows a recessive inheritance pattern, occurring in "biallelic" individuals who have inherited two sub-functional copies of the *MUTYH* gene (Morak et al. 2014). Individuals affected by this phenotype exhibit intestinal adenomatous polyposis and have an elevated risk for early-onset colorectal and duodenal malignancies (Nielsen et al. 2011; Al-Tassan et al. 2002). 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.

To test whether pathogenic *MUTYH* mutations might cause a germline mutator phenotype, we sequenced fifteen genomes from a large extended family affected by MAP. We used these sequences to measure germline de novo mutation (DNM) rates and spectra in seven children of three biallelic *MUTYH* variant carriers as well as six children of monoallelic variant carriers. Recently, another group published de novo mutation data from two children born to a mother with biallelic *MUTYH* mutations (Sherwood et al. 2023), obtaining evidence for a C>A mutator effect that we were able to further investigate in our larger dataset, which includes children of both male and female carriers of biallelic *MUTYH* genotypes. We then further contextualized our results through a comparison to a null model of mutation rate as function of parental age that was previously constructed from thousands of control trios (Jónsson et al. 2017). In this way, we were able to characterize how pathogenic *MUTYH* variants affect the human germline, using an analysis framework that is broadly appropriate for investigating the effects of other cancer syndromes on germline mutagenesis and human evolution.

Results

Sequencing whole genomes from a large extended family affected by pathogenic MUTYH mutations

We performed 50X-coverage whole-genome sequencing on saliva samples from three individuals who are compound heterozygotes for two *MUTYH* variants known as c.536A>G p.Y179C (NM_001128425) and c.700G>A V234M (NM_001128425), as well as a fourth individual who is a monoallelic carrier of p.V234M. Two of the three biallelic individuals were previously diagnosed with colon cancer, while the other two had histories of colon polyps. While ClinVar classifies Y179C as pathogenic with evidence from many previous studies (Al-Tassan et al. 2002; Nielsen et al. 2005, 2009; Vogt et al. 2009) and some laboratories consider 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), this family is affected by a notably elevated level of colorectal cancer, including in family members not sequenced as part of our study.

To assess the impact of the *MUTYH* Y179C/V234M genotype on the germline mutation rate and spectrum, we sequenced a total of nine adult children of these four individuals, as well as 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 (C). Colloquially, we refer to all parents as mothers and fathers if they conceived their children via oocytes and spermatocytes, respectively, recognizing that in some cases these labels may not match parents' social gender identities.

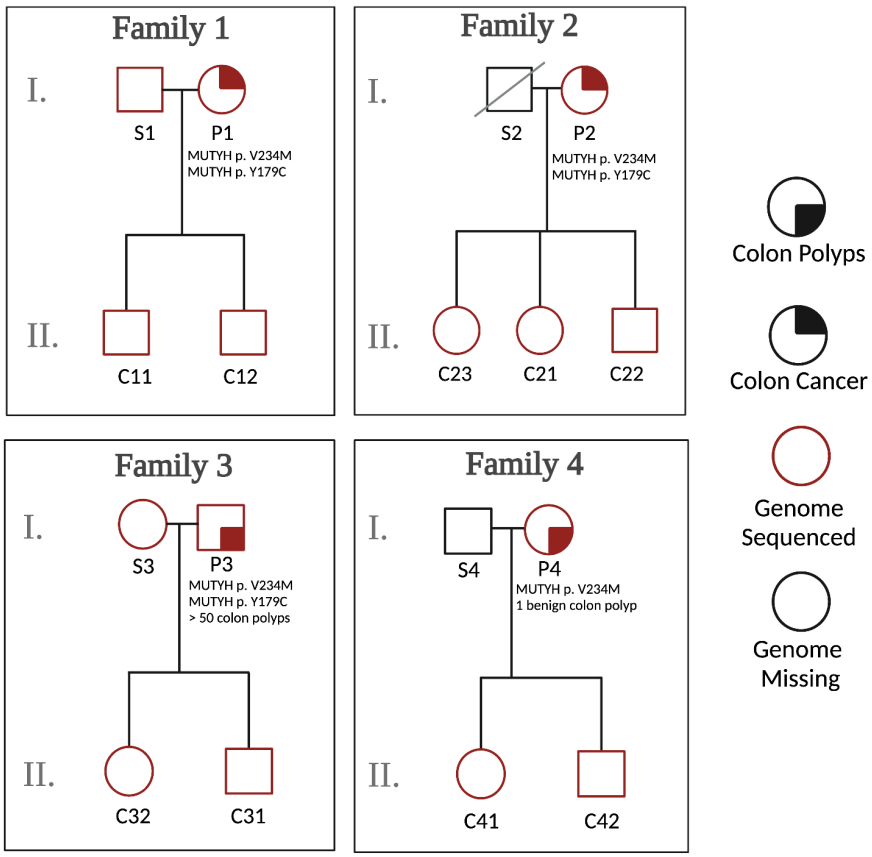

Figure 1. Sequencing four families of *MUTYH* variant carriers. To measure the effects of biallelic *MUTYH* mutations on germline mutagenesis, we sequenced three individuals with the same pathogenic *MUTYH* genotype, as well as one related monoallelic *MUTYH* variant carrier, along with their children and partners. Individuals have been given labels that indicate which nuclear family they are part of (1–4), whether they are a *MUTYH* variant carrier parent (P), a spouse or partner of that parent (S), or a child (C). Families 1 and 2 include mothers with the biallelic genotype Y179C/V234M, while Family 3 includes a father with the same Y179C/V234M genotype. Family 4 includes a mother with the monoallelic mutation V234M. Shaded quadrants indicate which individuals have been diagnosed with colon polyps (bottom right) or colon cancer (top right). *MUTYH* mutations and age at cancer diagnosis / number of identified colon polyps are listed below individuals for which this information is known. Square = male; circle = female; red = genome sequenced; black = genome not sequenced.

Cellular functional scan of MUTYH variant effects

The *MUTYH* genotype Y179C/V234M contains one variant annotated as pathogenic in ClinVar and one variant with conflicting interpretations (including pathogenic, likely pathogenic, and uncertain significance). To obtain more information about the pathogenicity of this genotype and compare it to the *MUTYH* genotype that was found by Sherwood et al. (2023) to have a slight mutagenic effect, 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 Y179C allele exhibited severe loss of repair function, whereas the V234M variant displayed a partial loss of

function with repair activity well below that of wild-type MUTYH (**Figure S1**). 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*. The genotype recently studied by Sherwood et al. (2023) contained Y179C along with c.1187G>A p.G368D (NM_001128425), a second common pathogenic *MUTYH* variant that may be less deleterious than 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 G368D as G396D in the coordinates of a different reference *MUTYH* transcript). We also used the same functional assay to measure the effects of c.461GT>AA p.R182Q (NM_001128425), the human analog of the mutation found in an outlier mouse strain known as BXD68 that displayed a *Mutyh* hypermutator germline phenotype. R182Q appears to be a total loss of function variant with a phenotype similar to that of Y179C.

De novo mutation calling in complete and 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, requiring the genomes of both parents to be sequenced. Two of the nuclear families in this study (1 and 3) had the genomes of both parents sequenced, while the other two families (2 and 4) only had the genome of the carrier parent sequenced. We called DNMs in the four children of the two complete nuclear families by identifying variants that violated the principles of Mendelian inheritance, followed by extensive filtering and visual curation (**Figure S2**).

We were unable to sequence the genomes of the fathers of the five children in Families 2 and 4 (one was deceased and one declined to participate). However, since each of these children had at least one full sibling represented in our dataset, we were able to leverage the sharing of paternal haplotypes among siblings to devise a "surrogate parent" method for estimating DNM rates and spectra in incomplete nuclear families, which is loosely based on the established use of relatives a surrogate parents for haplotype phasing (Kong et al. 2008). This method enabled us to estimate germline mutation rates for all members of the extended family except for the spouses who had no close relatives other than their children.

In each nuclear family where the mother's genome sequence was available but the father's genome sequence was missing, we were able to call DNMs in the subset of the genome where two siblings had inherited the same haplotype from their father (**Figure 2A-2B; Figure S2-S3**). In each of these regions, if

one sibling's genome contained an allele that was not present in either their mother's or their sibling's genome (acting as a proxy for their father's genome), we were able to deduce that the unique variant arose as a DNM. This implies that in a family with a mother and two siblings, about half of the siblings' genomes should be accessible for DNM calling. In a family with a mother and three siblings, about three-fourths of each sibling's genome is available for mutation calling, which is the expected proportion of the genome where each sibling inherited the same paternal haplotype as at least one other sibling.

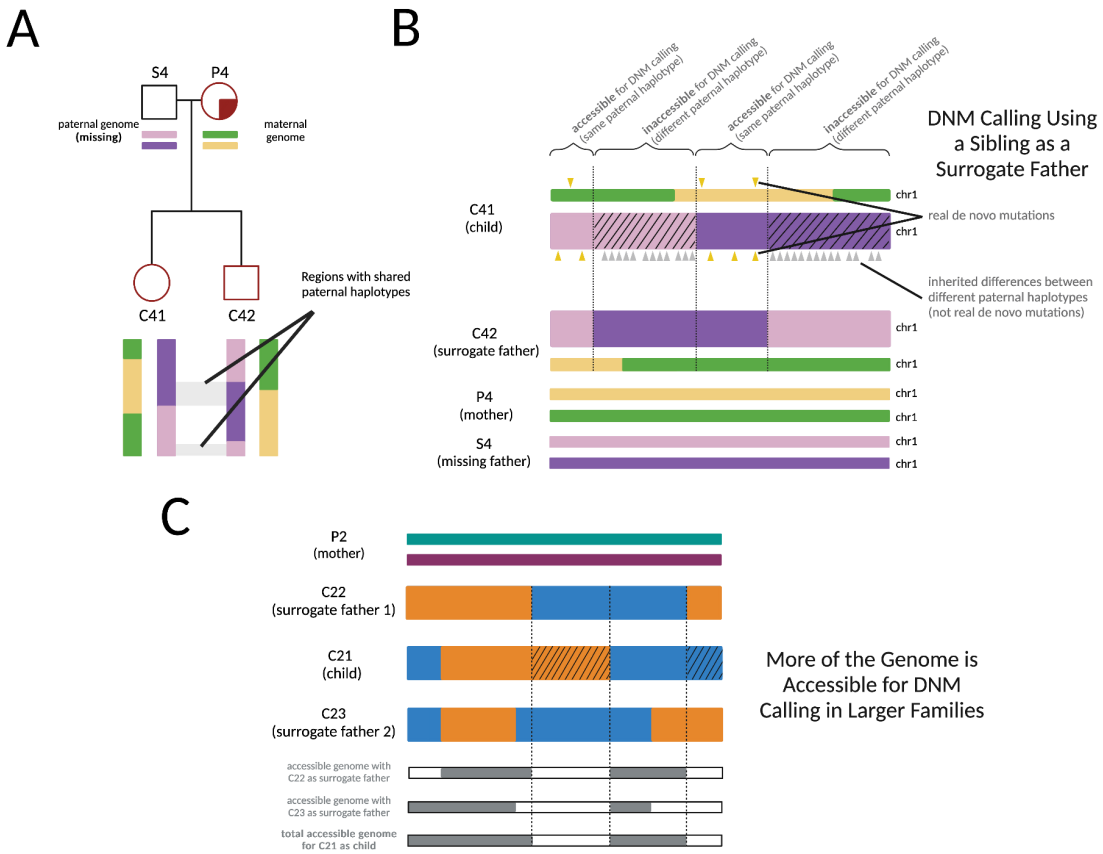

Figure 2. Using siblings as surrogate parents to identify DNMs. A) An illustration of the portions of an autosome with paternal haplotypes shared between two siblings. In an example chromosome from Family 4, DNMs can be called in regions where C41 and C42 share a paternal haplotype sequence with one 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.

To call mutations in the four relatives P1-P4, whose parents' genomes were all unavailable, we used these individuals as surrogate parents for one another in lieu of both maternal and paternal genomes. For example, in a region where P1 shares one IBD tract with P2 and a distinct IBD tract with P3, we were able to call DNMs in the genome of P1 using P2 and P3 as surrogate mother and father. It is also possible for P1 and P2 to have inherited the same haplotype from the same common ancestor in some genomic regions, in which case the genome of P2 can be used twice in lieu of both the maternal and paternal genomes. Though we were not able to determine which shared haplotypes were maternally versus paternally inherited, this information is not required for DNM calling. In practice, 49-76% of these genomes were callable using the IBD segments we were able to empirically infer.

We found that surrogate families with three or more children (Family 2 and the set of four parents P1–P4) allowed for better performance of the surrogate calling method. Larger family sizes increased the proportion of each individual's genome shared identity by descent (IBD) tracts with another sequenced relative, leading to a greater amount of accessible genome where DNMs could be identified using the surrogate approach (**Figure 2C**). More choices of surrogate parents also led to better elimination of false-positive DNM calls, as putative DNMs identified in one child when using a given relative as a parent could be screened for presence in additional relatives. Although about 1% of DNMs are expected to be shared between siblings as a result of parental gamete mosaicism (Jónsson et al. 2018), we initially noted in some preliminary call data that surrogate parent calling resulted a higher proportion of shared DNMs, leading us to conclude that the majority of DNMs shared between such siblings were likely to be mis-identified germline variants inherited from a missing parent. Both of these factors increased our uncertainty about the mutation rate and spectrum we estimated in Family 4, where we attempted to call DNMs using only the genomes of two siblings and their mother. In this family, only half the genome of each sibling was accessible to DNM calling (**Figure S4**) and we could not filter DNM calls using sharing between siblings.

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 individuals in this study to have normal mutation rates ranging from 9.92×10^{-9} to 2.26×10^{-8}

mutations per base pair per generation (**Table S1**), comparable to the range between 7.9×10^{-9} to 1.9×10^{-8} expected of healthy individuals with parental ages between 15 and 50 based on a large previous study (Jónsson et al. 2017). However, we found P2, the biallelic mother of Family 2, to have a much higher mutation rate (~ 6.3×10^{-8} mutations per site per generation). 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 S5**). 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 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.

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

Although we did not expect to find elevated mutation rates in the children of biallelic *MUTYH* variant carriers, we hypothesized that we might see an elevated proportion of C>A mutations 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 individual-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 S2A**), 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 individual's expected maternal and paternal 1-mer mutation burden as a function of their parents' ages (**Table S1**) and their accessible genome size (**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 3A**). For every individual 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**).

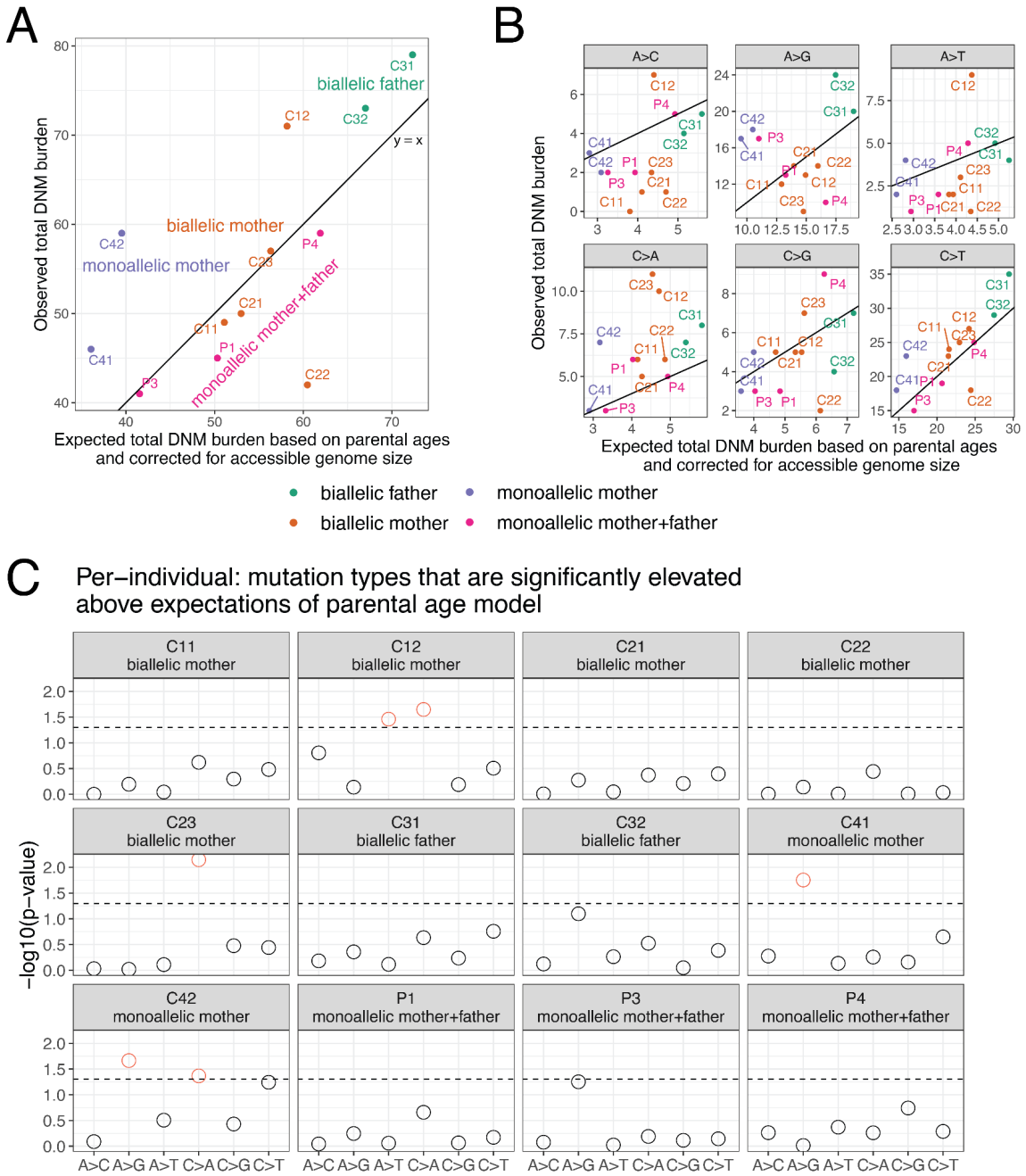

Figure 3. Observed and expected mutation counts. A) Comparison of observed DNM counts per-individual and the corresponding expected DNM counts under the parental age model (Jónsson et al. 2017), corrected for accessible genome size (**Figure S6**). P2 was excluded as discussed above due to evidence for somatic mutation contamination. Points are colored by the *MUTYH* carrier status of their parent(s). Each individual 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 *MUTYH* 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 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 (**Figure 3B-3C**). For 11/12 individuals (the exception being P3, biallelic father of Family 3), the observed C>A mutation count exceeded the expected C>A mutation count from the parental age model (**Figure 3B**). Across the remaining five 1-mer mutation types, the proportion of individuals exceeding the parental age model expectation ranged from 3/12 individuals (A>C mutations) to 8/12 individuals (C>T mutations) (**Figure 3B**). Most of the elevated C>A counts fell within an upper one-tailed 95% Poisson confidence interval of the expected count, but three individuals' C>A burdens significantly exceeded the parental age model expectation (**Figure 3C**). These included C42 (one of our bioinformatic outliers), but also included 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 (a possible signal of inherited germline variant bleed-through due to the surrogate-calling method) and A>T mutations in C12 (**Figure 3C**).

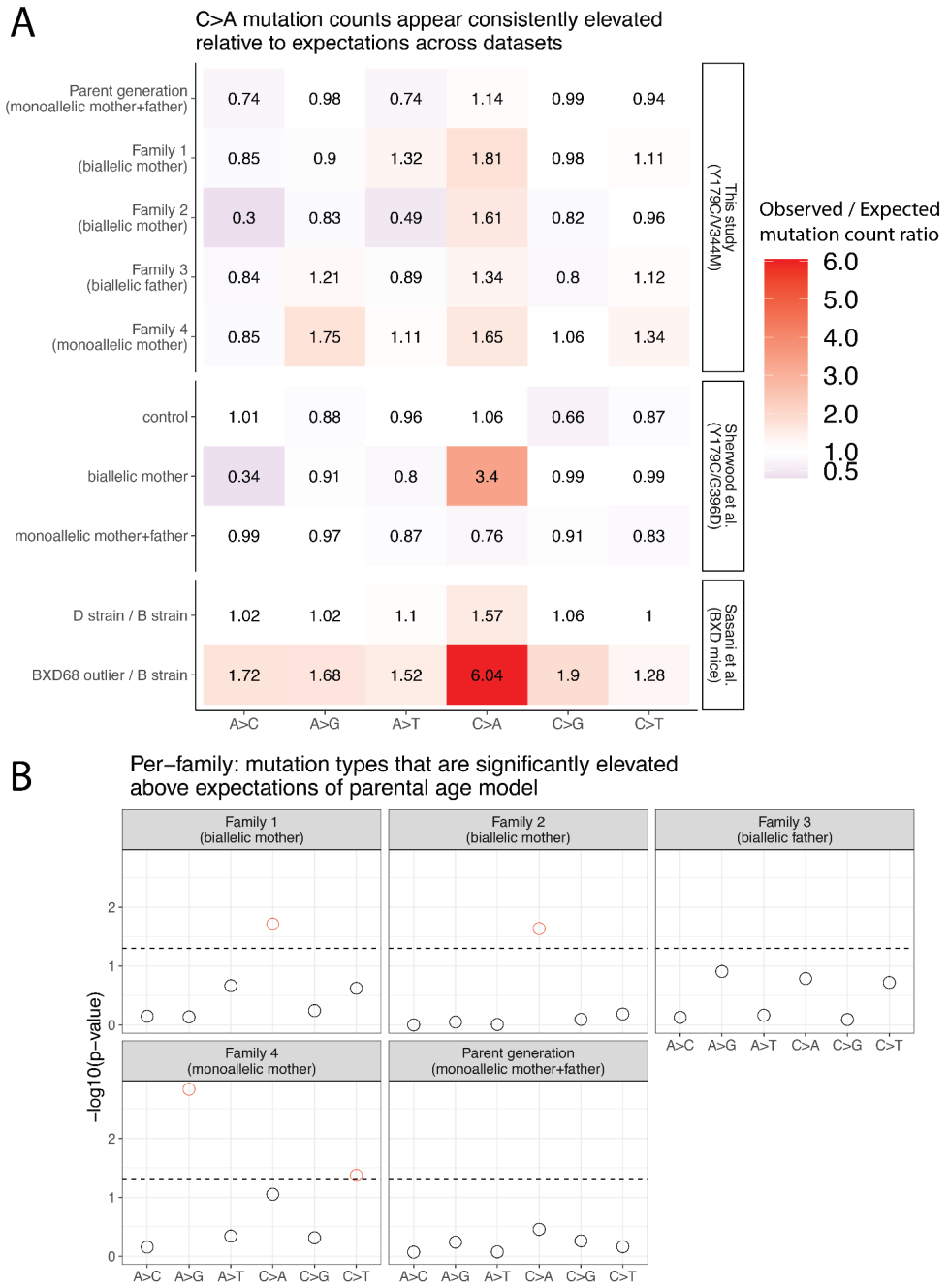

Figure 4. Children of mothers with biallelic *MUTYH* genotypes show significantly elevated C>A DNM counts. A) A heatmap showing the ratio of the observed / expected mutation counts per family (calculated by summing up the mutation counts per mutation type across all children within a family). These ratios are compared to the observed / expected ratio for the groups in Sherwood et al. (2023) (control group, individuals with a biallelic *MUTYH* mother, and individuals with monoallelic *MUTYH* 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. Families 1 and 2 show significant elevation for C>A DNM counts above what is expected under the parental age model, and Family 4 shows significant elevation of C>T and A>G mutation types above expectations.

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 1.61-fold above the expectation of the parental age model, respectively (**Figure 4A**). In each of these families, the total C>A mutation burden significantly exceeded the 95% upper 1-tailed confidence interval of the parental age model (**Figure 4B**). 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 4A**). In contrast, the family with a biallelic father (Family 3) was only enriched 1.34-fold for C>A mutations, a burden falling within the upper parental age 95% confidence interval (**Figure 4A-4B**). C>A enrichment was only 1.14-fold (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 significant bioinformatic obstacles to accurate DNM calling, we observed a non-significant 1.65-fold C>A enrichment along with a significant 1.75-fold A>G and significant 1.34-fold C>T enrichment (**Figure 4A- 4B**). 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.

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 *MUTYH* mother sequenced by Sherwood et al. (**Figure 4A; Figure S10)**, suggesting that this family's Y179C/G368D genotype may have a more severe mutator phenotype than the Y179C/V234M genotype affecting our pedigree.

Unlike Sherwood et al., we did not detect a significant increase in overall DNMs phased to the haplotype of the carrier parent (**Figures S11-S12, 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 S13, 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 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. Sherwood et al. were able to detect the activity of COSMIC mutational signature SBS18 in their biallelic *MUTYH* dataset, 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, perhaps reflecting the small total sample size of C>A mutations in our data (**Figure S14**).

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. 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 arises due to excess mutations that arose in the oocyte prior to conception, or possibly as postzygotic mutations. If the child has inherited one normal copy of *MUTYH* from their father, postzygotic mutations are unlikely to be enriched for C>A unless they arose prior to the maternal-zygotic transition, when the embryo first begins to express paternally inherited genes. In humans, paternal gene expression has already begun by the 2-cell stage (Li et al. 2013), which leads us to infer that if most of the excess C>A mutations in Families 1 and 2 are due to *MUTYH* variants, they likely arose in the maternal germline. Given that the maternal germline normally contributes only 25% of all germline mutations, the 1.34-fold to 3.4-fold C>A rate elevations observed in children of biallelic mothers imply much more drastic elevation of the C>A rate within the oocyte itself.

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 5A**). 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 1.61-fold to 1.81-fold overall C>A rate elevations observed in Families 1 and 2 (**Figure 4A, 5A**) imply maternal C>A mutation rate elevations of

3.3-fold and 4.2-fold, respectively (**Figure 5B**). 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 5B**).

We further estimated that Family 3's nonsignificant 1.34-fold C>A mutation rate elevation implies a C>A rate elevation of only 1.45-fold in that family's biallelic father (**Figure 5B**). This is surprisingly low compared to the 3.3-fold and 4.2-fold maternal C>A rate elevations that we infer to affect the two mothers who share the same biallelic *MUTYH* genotype. This might imply that *MUTYH* variants affect the female germline more severely than they affect the male germline, but in principle, the differences between these small numbers of families might also be driven by stochasticity. 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 5A**).

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 5B**). **Figure 5B** 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 1.7-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. Although this analysis is based on a limited sample size of individuals and neglects any potential effect of *MUTYH* on postzygotic mutations, it suggests that *MUTYH* variants may have a proportionally stronger effect on the maternal germline compared to the paternal germline.

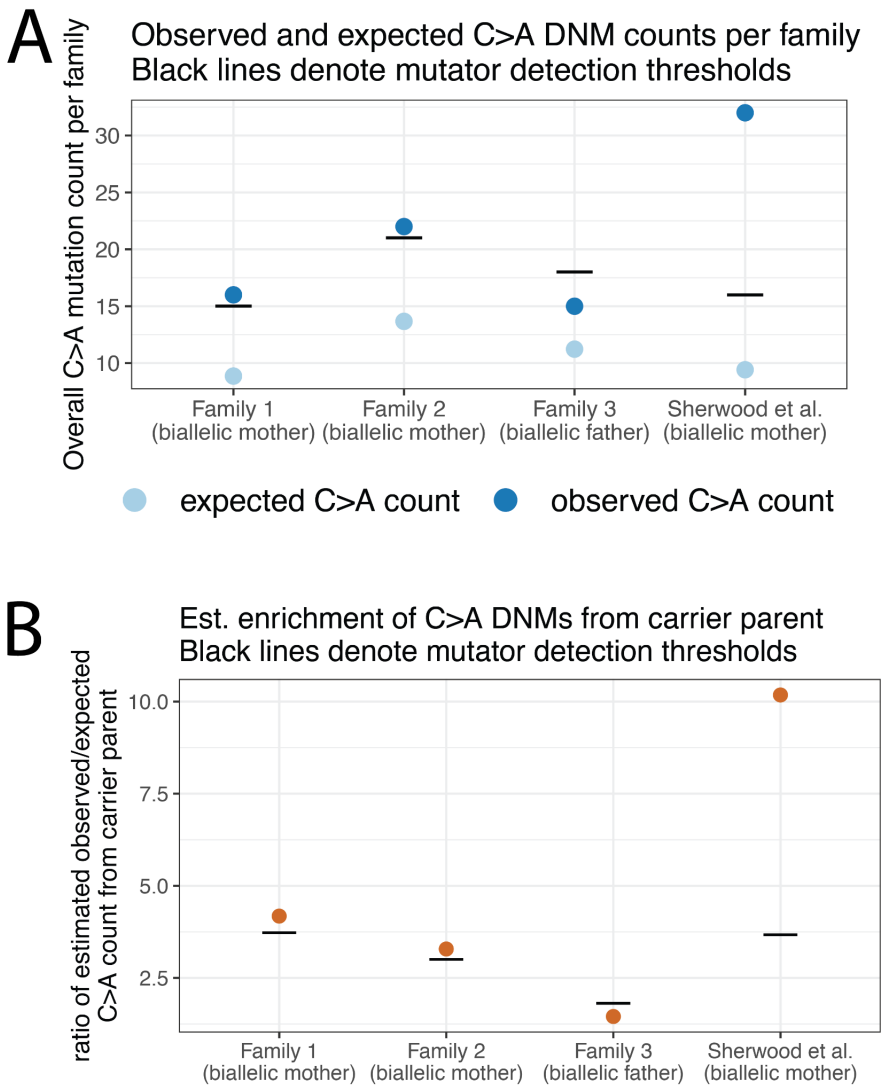

Figure 5. Estimating the minimum *MUTYH* effect sizes that we have power to detect in the male and female germlines. A) Observed (dark blue) and expected (light blue) C>A mutation counts in the children of each family with a biallelic parent. Horizontal black lines show the minimum number of mutations needed to reject the null parental age model ("mutator detection threshold"). Families 1 and 2 (biallelic mothers) exceed the threshold, implying a significant C>A mutation rate elevation, but Family 3 (biallelic father) does not. 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 S15** for this analysis based on per-individual mutation counts.

Discussion

We have investigated the germline mutation rate and spectrum within a large extended family affected by a *MUTYH* genotype, 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 Y179C/V234M genotype had a pathogenic effect, and we were able to use a cell-based *in vitro* functional assay to classify V234M as a partial loss of function variant that may impair protein function. By calling de novo mutations in the children of two mothers who carry the Y179C/V234M genotype, we documented a modest but significant maternal mutator effect that appears weaker than the maternal germline mutator effect recently discovered in children of mothers with the more common MAP-associated genotype Y179C/G368D (Sherwood et al. 2023).

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). However, C>A enrichment is less noisy and more interpretable when summed across multiple children within nuclear families, which reveals a consistent elevation of the C>A load in the children of biallelic mothers, both in our study and in the family studied by Sherwood et al (2023).

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 else that spermatocytes rely more on DNA repair pathways not involving *MUTYH*.

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 are so 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). To our knowledge, our work presents the strongest known evidence of a female-biased germline mutator allele–a recent macaque study documented a strong female mutator phenotype but contained no information on 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. Although we found that DNM calling using surrogate parents was most reliable in families with three or more children, this method will be particularly useful for opening up more families for multigenerational DNM analysis.

Our data suggests 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 (Y179C). Although our study and previous studies (Sherwood et al. 2023) find mutagenesis and cancer risk to be associated with biallelic genotypes that combine Y179C with a partial loss-of-function allele (V234M or 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. 2023).

The apparent effect size difference between Y179C/V234M and 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. Previous somatic mutation data found an effect size difference between the common genotypes Y179C/G368D and Y179C/Y179C that appeared concordant with an earlier age of polyposis onset in Y179C/Y179C carriers (Robinson et al. 2022). For a rare genotype like 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 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.

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. 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. We also benefited from helpful discussions with Rosana Risques, Lea Starita, and members of the Harris Lab.

Author Contributions

K.H. conceived the study. B.S. identified a suitable family for study, collected this family's genotype and phenotype data, and led the design of the human subject engagement and biological sample collection protocol. K.H. contacted and consented the study participants. C.Y. coordinated the sample processing and genomic data generation. A.C.B., D.M.P., and K.H. designed the study's computational analysis framework. Computational analyses were carried out by C.Y., A.C.B., and D.M.P. with technical assistance from L.Z. The cell line variant assay was designed by J.O.K. and S.L.H. and executed by S.H. C.Y., A.C.B, and S.H. generated main text and SI figures. C.Y., A.C.B., and K.H. drafted the manuscript, and D.M.P., S.L.H., L.Z., J.O.K., and B.S. contributed manuscript edits.

Competing interest statement

B.S. consults for the company Constantiam Biosciences. J.O.K serves as a scientific advisor to the company MyOme. The authors declare no other competing interests.

Data and Code Availability

All genomic data will be made available for controlled access via dbGaP. 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. Some details on exact relationships among members of this extended family are omitted from the manuscript to protect participants' privacy–for more details, contact the corresponding author (harriske@uw.edu).

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.

Methods

Recruitment and consenting of study subjects for biospecimen collection

The design of this study received prior approval from the University of Washington Institutional Review Board. After study participants gave written informed consent, they were each mailed an OGR-500 Oragene saliva collection kit. DNA was extracted from the Oragene kits at the Fred Hutchinson Cancer Center specimen processing core facility using the recommended standard protocol.

Genome sequencing, SNP calling, and DNM 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., Genome Analysis 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.

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.

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.

Putative de novo mutations 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.

<u>Surrogate method</u>

To identify haplotypes shared between relatives, we began by phasing the full 15-genome dataset using Beagle (Browning and Browning 2016). 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" and **Figure S2B**).

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 Pi and Pj, 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.

Raw DNM calls from PossibleDeNovo were then filtered as described below (see "Filtering" and "IGV inspection." During the IGV inspection step, we eliminated any putative DNM shared between two or more siblings, assuming that most of these variants were in fact inherited from un-sequenced parents in regions erroneously identified as inherited IBD.

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 S4**).

Filtering

DNMs were subjected to a series of quality control steps to eliminate potential false positives (**Figure S2A**). 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 GATK PossibleDeNovo as being present in the child but not in either parent.
- Exhibited a VAF of 30-70% in the child.
- Had no reads supporting the variant in either parent.
- Genotypes filtered with GATK 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 GATK 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

This manual curation resulted in the number of DNMs being reduced by ~30%.

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 14-40% of DNMs being phased per individual (**Table S1**), a fraction typical for studies of phased de novo mutations.

Comparison to Jónsson model, 'mouse model', downstream statistics

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' paternal 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):

$$y_{c,mat}(a_{mat}) = m_{c,mat} * a_{mat} + b_{c,mat}$$

$$y_{c,pat}(a_{pat}) = m_{c,pat} * a_{pat} + b_{c,pat}$$

In these equations, a_{mat} and a_{pat} are the maternal and paternal ages at the time of a child's birth, respectively, $m_{c,mat}$ 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 $b_{c,mat}$ and $b_{c,pat}$ are 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.

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 1.31×10^9 bp to 2.67×10^9 bp), we multiplied each expected mutation count under the parental age model by $\frac{g_i}{g_J}$, 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.

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

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 $P(X \geq k \mid \lambda) = P(X > k - 1 \mid \lambda)$, 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).

<u>Estimating the minimum mutator effect sizes that we have power to detect</u>

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 *MUTYH*, 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

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

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

<u>Mutational Signature Analysis</u>

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 0.662-0.849 in the per-individual analysis, from 0.871-0.913 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.

Cellular assay of MUTYH 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.

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.

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.
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. 2022.07.20.500591. https://www.biorxiv.org/content/10.1101/2022.07.20.500591v2.
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.
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.
Browning BL, Browning SR. 2016. Genotype Imputation with Millions of Reference Samples. *Am J Hum Genet* **98**: 116–126.
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.
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.
Fearon ER. 1997. Human cancer syndromes: clues to the origin and nature of cancer. *Science* **278**: 1043–1050.
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.
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.
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.
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.
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.
Li L, Lu X, Dean J. 2013. The Maternal to Zygotic Transition in Mammals. *Mol Aspects Med* **34**: 919–938.
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.
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.
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.
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.
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.
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.
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.
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.
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.
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. 2023. Epistasis between mutator alleles contributes to germline mutation rate variability in laboratory mice. *eLife* **12**. https://elifesciences.org/reviewed-preprints/89096.
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. 2013. Role of the Oxidative DNA Damage Repair Gene *OGG1* in Colorectal Tumorigenesis. *JNCI J Natl Cancer Inst* **105**: 1249–1253.
Stendahl AM, Sanghvi R, Peterson S, Ray K, Lima AC, Rahbari R, Conrad DF. 2023. A naturally occurring variant of *MBD4* causes maternal germline hypermutation in primates. *Genome Res* gr.277977.123.
Van der Auwera GA. 2020. *Genomics in the cloud: using Docker, GATK, and WDL in Terra*. First edition. O'Reilly Media, Sebastopol, CA.
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.
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.
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.