Title: **A wild-derived antimutator drives germline mutation spectrum differences in a genetically diverse murine family**

Authors: Thomas A. Sasani[1], David G. Ashbrook[2], Lu Lu[2], Abraham A. Palmer[3,4], Robert W. Williams[2], Jonathan K. Pritchard[5,6], Kelley Harris[1*]

Affiliations:
[1]University of Washington, Department of Genome Sciences, Seattle, WA
[2]University of Tennessee Health Science Center, Department of Genetics, Genomics and Informatics, Memphis, TN
[3]University of California San Diego, Department of Psychiatry, La Jolla, CA
[4]University of California San Diego, Institute for Genomic Medicine, La Jolla, CA
[5]Stanford University, Department of Genetics, Stanford, CA
[6]Stanford University, Department of Biology, Stanford, CA
*To whom correspondence should be addressed

Abstract
Little is known about the impact of genetic variation on mutation rates within species. We conducted a QTL scan for alleles that influence germline mutagenesis using a panel of recombinant inbred mouse lines descended from the laboratory strains C57BL/6J (B6) and DBA/2J (D2). Mice inheriting *D* haplotypes at a locus on chromosome 4 accumulate C>A germline mutations at a 50% higher rate than those with *B* haplotypes. The QTL contains coding variation in *Mutyh,* a DNA repair gene that underlies C>A-dominated mutational signatures with similar sequence context biases in human cancers. Both *B* and *D Mutyh* alleles are present in wild populations of *Mus musculus domesticus*, and likely represent the first documented example of segregating variation that modulates germline mutation rates in mammals.

One-sentence summary: The rate of germline C>A mutations in the BXD family of mice is modulated by natural variation at a locus containing *Mutyh*, a component of the 8-oxoguanine DNA repair pathway.

Main Text

Although organisms generally maintain low mutation rates via intricate DNA repair and proofreading pathways, the precise fidelity of germline DNA replication varies by orders of magnitude across the tree of life (*1*). Evolutionary biologists have long debated why mutation rates vary so dramatically, citing among other factors the necessity of beneficial mutations for adaptation (*2*), the cost of DNA replication fidelity (*3*), and the inefficiency of selection on weak mutation rate modifiers (*1*). One reason for this long-standing debate is the paucity of data on the genetic basis of mutation rate variation. Although mutator alleles have been identified in bacteria (*2*) and candidate *cis* mutator loci have been discovered using patterns of human genetic variation (*4*), to our knowledge, no natural variants have been shown to impact the germline mutation rate of a mammalian species.

There exists abundant indirect evidence that DNA replication fidelity varies among humans. Some of the most compelling evidence is encoded in the "mutation spectrum," a summary of the relative abundances of specific base substitution types (C>A, C>T, A>G, etc.) in a particular collection of mutations. For example, germline variants private to European populations are enriched nearly 1.5-fold for TCC>TTC mutations compared to variants private to African and East Asian populations (*5*, *6*). Many other mutation types exhibit subtle differences in abundance between human continental groups (*7*), and there are pronounced differences between the mutation spectra of great ape species (*8*), suggesting that small-effect mutator alleles may be segregating in every hominin lineage. Direct measurements of mutation rates also vary among human families and between some ethnic groups (*9*–*11*); however, it is unclear how much variation in mutation rates and spectra is driven by genetics as opposed to the environment (*12*).

Previous attempts to either discover mutator alleles or estimate mutation rate heritability have been severely hampered by the noisy nature of mutation rate estimates. On average, fewer than 100 *de novo* mutations occur per generation in humans, and the exact number depends strongly on parental age (*9*, *13*, *14*). In this study, we avoided these complications by searching for mutator alleles in a large family of recombinant inbred mouse lines. Beginning in 1971, crosses of two inbred laboratory mice—C57BL/6J and DBA/2J—have been used to generate several cohorts of B-by-D (or BXD) recombinant inbred progeny (*15*). Each of the BXD lines is a unique linear mosaic of *B* and *D* haplotypes and has accumulated *de novo* germline mutations on a consistent genetic background throughout many generations of sibling inbreeding in a controlled environment. We recently sequenced the genomes of 153 BXDs in order to identify mutations unique to each BXD and absent from both parental lines.

Mutation spectrum differences between members of the extended C57 and DBA families have been documented in a prior study of *de novo* mutations in 29 inbred laboratory lines (*16*). We hypothesized that mutators underlying these differences might be detected using classical linkage analyses of *de novo* mutation rates and spectra in the BXD genomes.

Older BXD lineages contain high loads of *de novo* germline mutations that accumulated over many generations of inbreeding

The BXD family was generated via a series of six breeding epochs initiated between 1971 and 2014 (*15*) (**Figure S1)**. Each epoch comprises 7 to 49 recombinant inbred progeny lines; we included a total of 94 BXDs from five of these epochs in our analyses—all of which have been inbred for at least 20 generations (**Table S1**). To estimate germline mutation rates and spectra for each BXD family member, we first identified high quality singleton variants unique to each genome (**Materials and Methods**). After masking potentially error-prone repeats and segmental duplications, we discovered 63,914 high confidence autosomal homozygous singletons across all genomes (**Figure 1a**). Counts varied considerably among BXDs and as expected, were positively correlated with the number of generations of inbreeding (Poisson regression $p < 2.2 \times 10^{-16}$, **Fig. 1a**). We estimated each line's mutation rate by dividing its singleton count by the number of generations of inbreeding during which mutations could have occurred, and by the number of haploid base pairs accessible to singleton calling in the line's genome (**Materials and Methods**) (*17*). We calculated an average autosomal mutation rate of 5.6×10^{-9} per generation per base pair—similar to prior estimates of 3.9×10^{-9} and 5.4×10^{-9} from mouse pedigree sequencing and mutation accumulation experiments, respectively (*17*, *18*). As in (*16*), singletons were less depleted from conserved regions of the genome than older segregating variants (Kolmogorov-Smirnov test, $p < 2.2e-16$) **(Fig. S2**), suggesting that minimal purifying selection has acted to remove mutations that arose during BXD inbreeding.

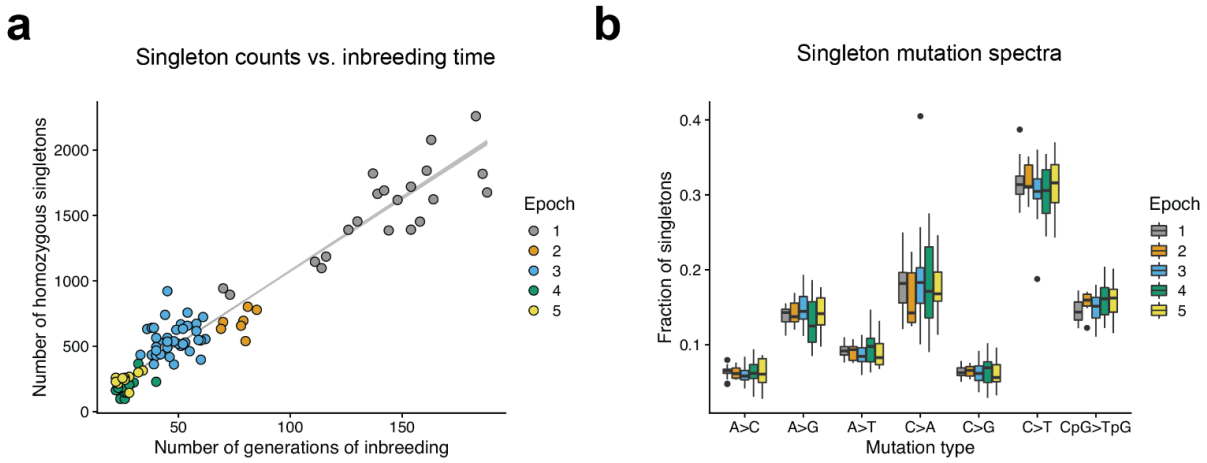

Figure 1: Homozygous singletons approximate recent *de novo* germline mutations

(a) Counts of autosomal homozygous singletons in 94 BXDs, correlated with the number of generations of inbreeding. Lower-numbered epochs are older and have been inbred for a longer period. Line is from a Poisson regression (identity link) with 95% confidence bands.

(b) Fractions of singletons that belong to each of seven mutation types, including the six possible transitions and transversions as well as CpG>TpG. Strand complements are collapsed.

Identification of a quantitative trait locus influencing the rate of germline C>A mutation

Overall, singleton mutation spectra were dominated by C>T transitions (**Fig. 1b**), and were similar to spectra previously inferred from *de novo* germline mutations in mice (*18*). We hypothesized that some of the mutation spectrum variation across the BXD family might be caused by one or more mutator loci, where a *B* allele has a different functional impact on DNA repair or replication fidelity than a *D* allele. Using the program R/qtl2 (*19*), we performed quantitative trait locus (QTL) scans for a total of fifteen mutagenesis-related phenotypes: the overall mutation rate in each strain, and the rates and fractions of the seven mutation types shown in **Fig. 1b** (**Table S2**).

After controlling for kinship among all BXD lines, the number of generations of intercrossing each BXD underwent prior to inbreeding, and each line's epoch membership, we did not find any QTL for the overall mutation rate (**Fig. S3a**). However, when we performed a QTL scan for the fraction of C>A singleton mutations, we discovered a single highly significant peak on chromosome 4 (**Fig. 2a-b**; maximum LOD of 18.1 at 116.8 Mbp; Bayes 95% credible interval = 114.8 –118.3 Mbp). BXD family members with *D* haplotypes at this locus have substantially higher fractions of C>A singleton mutations than those with *B* haplotypes (**Fig. 2c;** Welch's t-test $p < 2.2 \times 10^{-16}$). When we performed a QTL scan for the C>A mutation rate, we discovered a LOD peak at the same location on chromosome 4 (**Fig. 2a-b**; maximum LOD of 7.0 at 116.8 Mbp; Bayes 95% credible interval = 114.8 –118.8 Mbp). BXDs with the *D* haplotype

at the QTL accumulate C>A mutations at over 1.5 times the rate of strains with the *B* haplotype (1.22×10^{-9} and 7.32×10^{-10} per base pair per generation, respectively). None of the other mutagenesis-related QTL scans identified significant peaks (**Fig. S3b**). Since a higher C>A fraction distinguished the DBA mutation spectrum from C57 in a previous report (*16*), the observed QTL on chromosome 4 appears to fit the profile of a mutator haplotype responsible for the difference in mutation spectrum between the parental strains.

To further visualize differences between BXDs with either DBA/2J or C57BL/6J ancestry at the QTL on chromosome 4, we performed a principal components analysis of the mutation spectrum of singletons in each family member, as well as strain-private mutation spectra from DBA/2J and C57BL/6NJ (*16*) (**Fig. 2d**). The first principal component appears to separate BXDs with *D* haplotypes at the QTL from those with *B* haplotypes, due in large part to the differences in C>A mutation fractions between them (**Fig. 2d**).

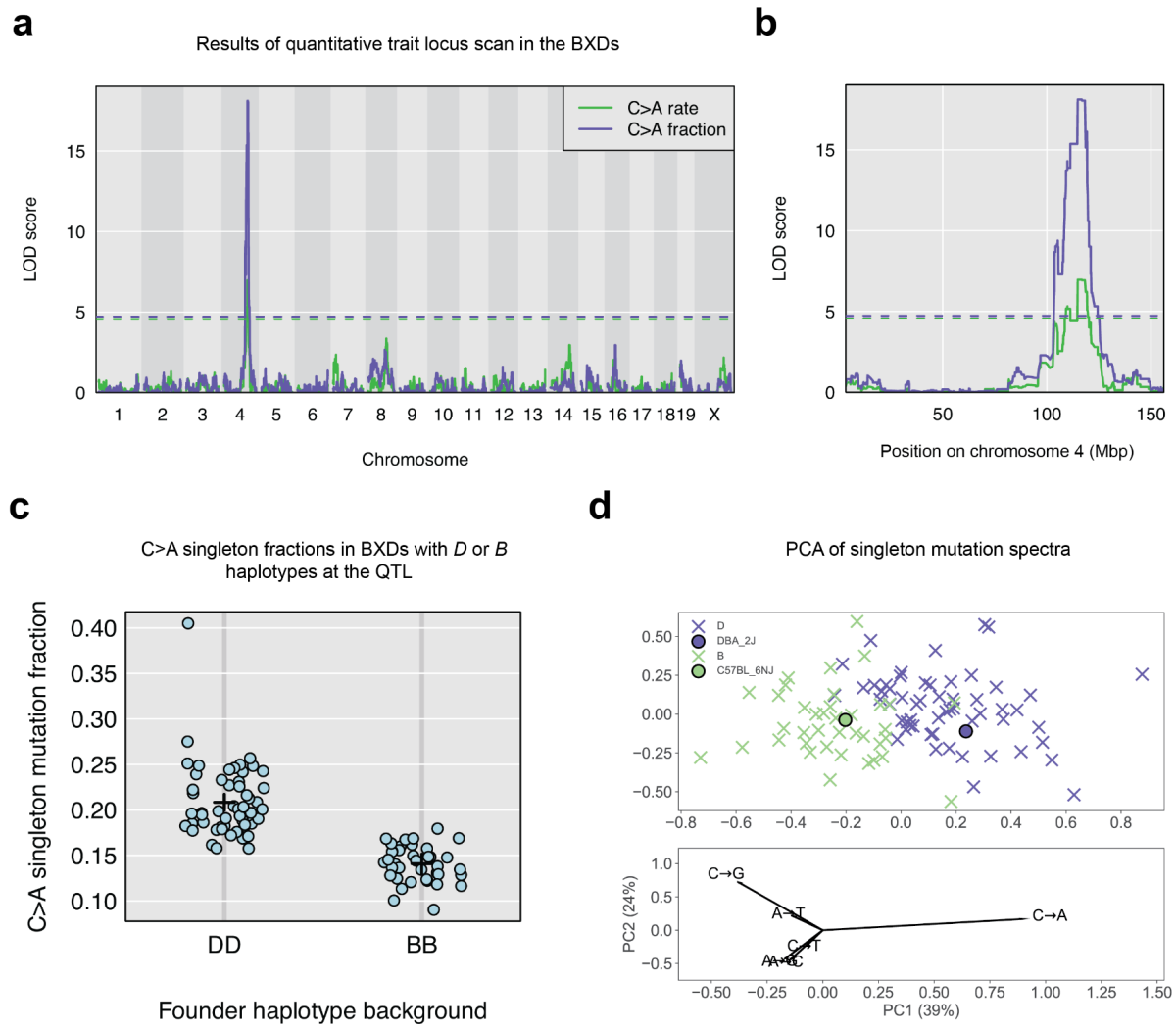

Figure 2: A quantitative trait locus on chromosome 4 for the rate of germline C>A mutation

(a) LOD scores for either the centered log-ratio transformed fraction or untransformed rate of C>A mutations. Blue and green dashed lines indicate genome-wide significance thresholds (using 1,000 permutations and alpha = 0.05/15) for the fraction and rate scans, respectively. C>A fraction and rate phenotypes are included in the GeneNetwork database as BXD_24430 and BXD_24437, respectively.

(b) Same as (a) but zoomed to only show chromosome 4.

(c) C>A singleton fractions in BXD strains that are homozygous for either the *D* or the *B* haplotype at the QTL on chromosome 4. Vertical bars indicate +/- 2 times the standard error of the mean; the mean is depicted as a horizontal bar. C>A fraction outlier is BXD68 and was not included in QTL scan.

(d) Principal component analysis plot of the 6-dimensional mutation spectrum of singletons in each BXD RIL (n = 94). BXDs are denoted as crosses and are colored by parental ancestry at the QTL on chromosome 4 (*D* or *B*). Strain-private mutation spectra for DBA/2J and C57BL/6NJ from (*16*) were included in PCA for reference, and those two strains are denoted with circles in the plot. Loadings are plotted for each mutation type below. Fractions of each mutation type were centered log-ratio transformed prior to PCA.

The DNA repair gene *Mutyh* resides in the C>A QTL interval and contains five missense differences between C57BL/6J and DBA/2J

The Bayes 95% credible interval surrounding the QTL spans approximately 4 Mbp on chromosome 4 and contains 76 protein-coding genes. We used SnpEff (*20*) to estimate the likely impact of variation within this interval on protein structure or function, and found that 21 of the genes in the QTL harbored sequence differences between C57BL/6J and DBA/2J annotated as having MODERATE or HIGH impact on any Ensembl transcripts (61 MODERATE-impact variants, 5 HIGH-impact variants). Of these genes, only one is annotated with either the "DNA repair" or "cellular response to DNA damage" Gene Ontology term: the mouse homolog of the mutY DNA glycosylase, *Mutyh*. We discovered 5 nonsynonymous differences between DBA/2J and C57BL/6J in the *Mutyh* gene, all of which were annotated as having MODERATE impact according to SnpEff (**Table 1**).

The MUTYH protein is required for base excision repair of 8-oxoguanine (8-oxoG) modified nucleotides. If left unrepaired, 8-oxoG are known to incorrectly pair with adenines, leading to G:C>T:A mutations during subsequent DNA replication (*21*). We note that three other genes within the QTL interval are annotated by the Gene Ontology database as being involved in "DNA repair" or the "cellular response to DNA damage" (*Plk3, Rad54L,* and *Dmap1*), and one gene in the interval, *Prdx1*, is involved in the "cellular response to oxidative stress." However, none of these genes harbors coding differences between DBA/2J and C57BL/6J.

To address the possibility that regulatory, rather than coding, variation might be responsible for the C>A QTL, we used GeneNetwork (*22*) to ask if the marker with the highest

LOD score at the QTL (rs52263933) was associated with the expression of protein-coding genes that might be related to the C>A mutator phenotype (**Materials and Methods**). We discovered significant correlations between rs52263933 genotypes and the expression of *Mutyh* in several tissues, including spleen, liver, and hematopoietic stem cells; in each case, the *B* allele was associated with higher expression of *Mutyh* (**Fig. S4**). We did not discover significant correlations between rs52263933 genotypes and *Mutyh* expression in other tissues, such as retina, kidney, and amygdala (**Fig. S4**), and we were not able to query BXD expression data from testis or ovary, the tissues where germline mutations occur. Nonetheless, it is possible that both regulatory and coding variation in the QTL interval interact to affect the rate of C>A germline mutation in the BXDs.

Table 1: *Mutyh* missense mutations in the BXD RILs
Amino acid positions are relative to Ensembl transcript ENSMUST00000102699.7.

Amino acid change	Coordinates in mm10	Fixed in DBA/2J?
p.Gln5Arg	Chr4:116814338	Yes
p.Arg24Cys	Chr4:116814394	Yes
p.Ser69Arg	Chr4:116815658	Yes
p.Thr312Pro	Chr4:116817416	Yes
p.Thr313Pro	Chr4:116817419	Yes
p.Arg153Gln	Chr4:116816476	No (private to BXD68)

Similar C>A-dominated mutational signatures occur in *Mutyh*-knockout mice and human colorectal cancers with impaired MUTYH function

Previously, germline mutation rates were measured in the TOY-KO mice, which possess triple knockouts of *Mutyh*, *Mth1*, and *Ogg1*, and therefore lack the machinery needed to repair 8-oxoguanine in the genome (*23*). The exonic germline mutation rate of the TOY-KO mice was elevated nearly 40-fold above the rate of wild-type mice, and approximately 99% of germline mutations in TOY-KO offspring were C:G>A:T (*23*). We discovered that *de novo* germline mutations observed in the TOY-KO mice occurred in similar 3-mer sequence contexts as the C>A mutations enriched in BXDs with *D* haplotypes at the QTL, suggesting that the mechanisms underlying germline mutations in these two groups are similar (**Fig. 3a-b**).

Two mutational signatures that exhibit strong C>A biases have also been observed in colorectal tumors occurring in human patients suffering from MUTYH-associated polyposis, an

autosomal recessive disorder caused by inherited MUTYH missense variants (*24–26*). These signatures, recorded in the COSMIC cancer mutational signature catalog (*27*) as SBS36 and SBS18, are dominated by CA>AA and CT>AT, the same mutation types most enriched in BXDs inheriting the *D* haplotype at the QTL on chromosome 4 (**Fig. S5**). These striking similarities provide further evidence that *Mutyh* variation is responsible for the C>A mutator phenotype in the BXD.

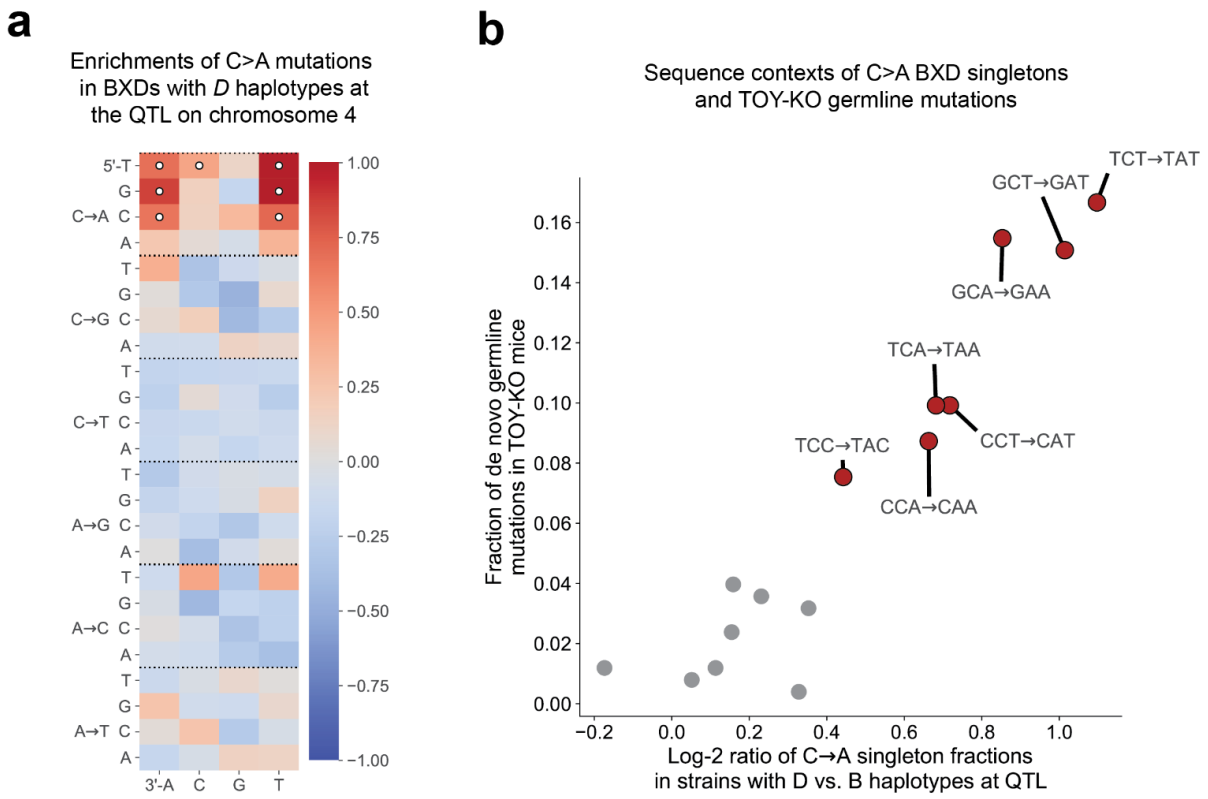

Figure 3: Particular C>A mutation types are more frequent in strains with the *D* haplotype at the C>A QTL

(a) Log-2 ratios of mutation fractions in strains with *D* haplotypes compared to strains with *B* haplotypes at the QTL on chromosome 4. Mutation types with Chi-square test of independence *p*-values < 0.05/96 are marked with white circles.

(b) The log-2 ratios of 3-mer C>A mutations in the heatmap in (a) are plotted against the relative abundances of each 3-mer C>A germline mutation in the *Mutyh*$^{-/-}$/*Ogg1*$^{-/-}$/*Mth1*$^{-/-}$ mice from (*23*) (*n* = 252 mutations). 3-mer mutation types that were significantly enriched in (a) are labeled and outlined in black.

The C>A-enriched outlier strain BXD68 harbors a deleterious singleton mutation in *Mutyh*

The C>A fraction of one outlier line, BXD68, is 5.6 standard deviations above the mean C>A fraction of all other lines with the *D* haplotype (**Fig. 2c)**. We searched the singleton mutations present in BXD68 for candidate *de novo* mutator alleles and discovered a unique

homozygous missense mutation in *Mutyh* (p.Arg153Gln). Notably, the BXD68 singleton affects an amino acid that is conserved between human and mouse (p.Arg179, relative to the human Ensembl transcript ENST00000372098.3). In fact, two missense mutations that affect the human p.Arg179 amino acid (rs747993448 and rs143353451) are both listed in the ClinVar database as being pathogenic or likely pathogenic (*28*).

The p.Arg153Gln amino acid change is also predicted by both PROVEAN (*29*) and SIFT (*30*) to be deleterious in the murine MUTYH peptide sequence. Finally, the most frequent mutation types observed in BXD68 are of the CA>AA and CT>AT types, much like the COSMIC signatures and TOY-KO germline mutations (**Fig. S6**). Based on this evidence, we hypothesize that p.Arg153Gln arose as a *de novo* germline mutation in BXD68, and severely impairs the 8-oxoguanine DNA damage response pathway in that strain.

The *B* and *D Mutyh* alleles are derived from wild mouse variation and appear to shape mutation accumulation in other laboratory strains

To determine whether the BXD mutator might be shaping genetic variation in other mouse populations, we analyzed genome sequencing data from wild populations of *Mus musculus domesticus, Mus musculus musculus*, *Mus musculus castaneus,* and *Mus spretus,* which were sampled from diverse geographical locations (*31*). We observed that all five candidate causal variants in *Mutyh* are segregating in wild *M.m.domesticus*, the subspecies from which the majority of DBA/2J and C57BL/6J ancestry is derived (*32*) (**Fig. 4a**). In contrast, the outgroup species *Mus spretus* appears to be fixed for the *D* allele at four of these sites and fixed for the *B* allele at the fifth site (p.Ser313Pro) (**Fig. 4a**). This suggests that DBA/2J, rather than the reference strain C57BL/6J, carries the ancestral allele at most of these sites, a conclusion supported by a multiple sequence alignment of additional vertebrates (**Fig. 4b**). If the C>A mutator phenotype is, in fact, caused by one of the four nonsynonymous variants at which the DBA/2J allele is ancestral, this would imply that C57BL/6J harbors a derived “antimutator” that increases the efficacy of 8-oxoguanine repair.

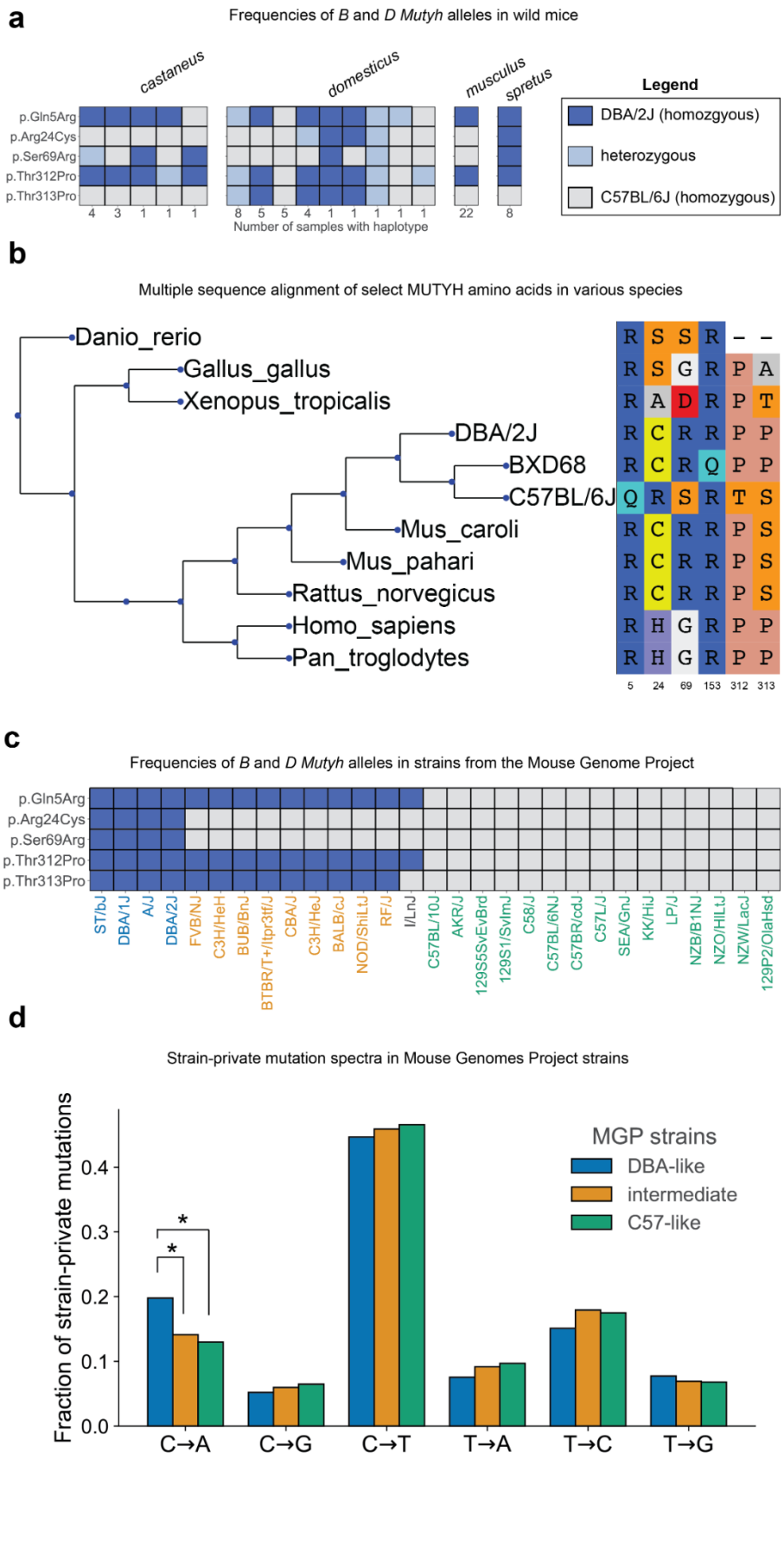
a
Frequencies of B and D Mutyh alleles in wild mice
castaneus
domesticus
musculus
spretus
p.Gln5Arg
p.Arg24Cys
p.Ser69Arg
p.Thr312Pro
p.Thr313Pro
4 3 1 1 1
8 5 5 4 1 1 1 1 1
22
8
Number of samples with haplotype
Legend
DBA/2J (homozgyous)
heterozygous
C57BL/6J (homozygous)
b
Multiple sequence alignment of select MUTYH amino acids in various species
Danio_rerio
Gallus_gallus
Xenopus_tropicalis
DBA/2J
BXD68
C57BL/6J
Mus_caroli
Mus_pahari
Rattus_norvegicus
Homo_sapiens
Pan_troglodytes
5 24 69 153 312 313
c
Frequencies of B and D Mutyh alleles in strains from the Mouse Genome Project
p.Gln5Arg
p.Arg24Cys
p.Ser69Arg
p.Thr312Pro
p.Thr313Pro
ST/bJ
DBA/1J
A/J
DBA/2J
FVB/NJ
C3H/HeH
BUB/BnJ
BTBR/T+/Itpr3tf/J
CBA/J
C3H/HeJ
BALB/cJ
NOD/ShiLtJ
RF/J
I/LnJ
C57BL/10J
AKR/J
129S5SvEvBrd
129S1/SvImJ
C58/J
C57BL/6NJ
C57BR/cdJ
C57L/J
SEA/GnJ
KK/HIJ
LP/J
NZB/B1NJ
NZO/HILtJ
NZW/LacJ
129P2/OlaHsd
d
Strain-private mutation spectra in Mouse Genomes Project strains
Fraction of strain-private mutations
0.0
0.1
0.2
0.3
0.4
*
*
C→A
C→G
C→T
T→A
T→C
T→G
MGP strains
DBA-like
intermediate
C57-like

Figure 4: *Mutyh* alleles segregate in both wild and inbred mouse strains and are ancestral in DBA/2J

(a) Presence of DBA/2J or C57BL/6J *Mutyh* alleles in 67 wild mice (*31*). Unique combinations of *Mutyh* alleles were identified in each *Mus* species or subspecies, and the numbers of mice with each combination are listed below each column.
(b) Multiple sequence alignment of MUTYH amino acids is subsetted to only show the six amino acids affected by missense mutations in the BXD. Positions of amino acids in the murine MUTYH peptide sequence (ENSMUST00000102699.7) are shown below each column.
(c) Presence of DBA/2J or C57BL/6J *Mutyh* alleles in Sanger Mouse Genomes Project strains that have associated strain-private singleton data from (*16*). Strain labels are colored according to whether their *Mutyh* genotypes match DBA, C57, or are intermediate between the two.
(d) Comparison of strain-private mutation spectra from (*16*) in MGP strains with DBA-like, intermediate, or C57-like genotypes at *Mutyh* missense mutations. *P*-value of Chi-square test comparing C>A fractions between DBA-like and intermediate strains: 3.3×10^{-7}; between DBA-like and C57-like strains: 1.4×10^{-10}; between intermediate and C57-like strains: 1.2×10^{-1}.

Of the 36 laboratory mouse strains sequenced by the Sanger Mouse Genomes Project (MGP) (*33*), strain-private mutations (which likely represent recent *de novo* germline mutations) were identified in 29 in a previous report (*16*). Of these 29, three match DBA/2J at all five nonsynonymous *Mutyh* sites, while 15 match C57BL/6J at all five sites (**Fig. 4c**). An additional 9 "intermediate" strains are homozygous for exactly three of the DBA/2J alleles at amino acids 5, 312, and 313 (**Fig. 4c**). After reanalyzing the spectra of strain-private mutations published in (*16*), we found that the four DBA-like strains had significantly higher C>A fractions than the intermediate and C57-like strains; the 3-mer mutation types most enriched in the four DBA-like strains were CA>AA and CT>AT, the same mutation types enriched in BXDs with *D* haplotypes (**Fig. 4d**, **Fig. S7**). However, we found no significant mutation spectrum differences between the intermediate and C57-like strains (**Fig. 4d**). These observations tentatively point to p.Arg24Cys and p.Ser69Arg as the variants most likely to cause a C>A mutator phenotype; however, the MGP strains differ at millions of other sites in their genomes, and we cannot rule out the possibility that the observed C>A fraction gradient is driven by variation outside *Mutyh*. At each of the nonsynonymous sites in *Mutyh*, DBA/2J shares a putatively ancestral allele with *Mus caroli, Mus pahari*, and *Rattus norvegicus* (**Fig. 4b**), suggesting that if any of these sites is responsible for the QTL, one or more mutations in C57BL/6J has managed to measurably improve the base excision repair function of *Mutyh*.

To situate the *Mutyh* signature in the broader landscape of natural mouse variation, we estimated the 3-mer mutation spectra of singletons private to wild *Mus spretus, M.m.domesticus*, *M.m.castaneus,* or *M.m.musculus* (**Materials and Methods**). The *Mutyh*-associated signature (comprising enrichments of CT>AT and CA>AA) appears somewhat enriched in *M.m.domesticus* compared to *M.m.musculus* and *M.m.castaneus,* but is only one of

many signatures that differ in dosage between wild mouse subspecies (**Fig. 5**). These results suggest that the BXD QTL is not the principal driver of mutation spectrum differences in wild populations, and is just one of many genetic or environmental mutators active in wild mice.

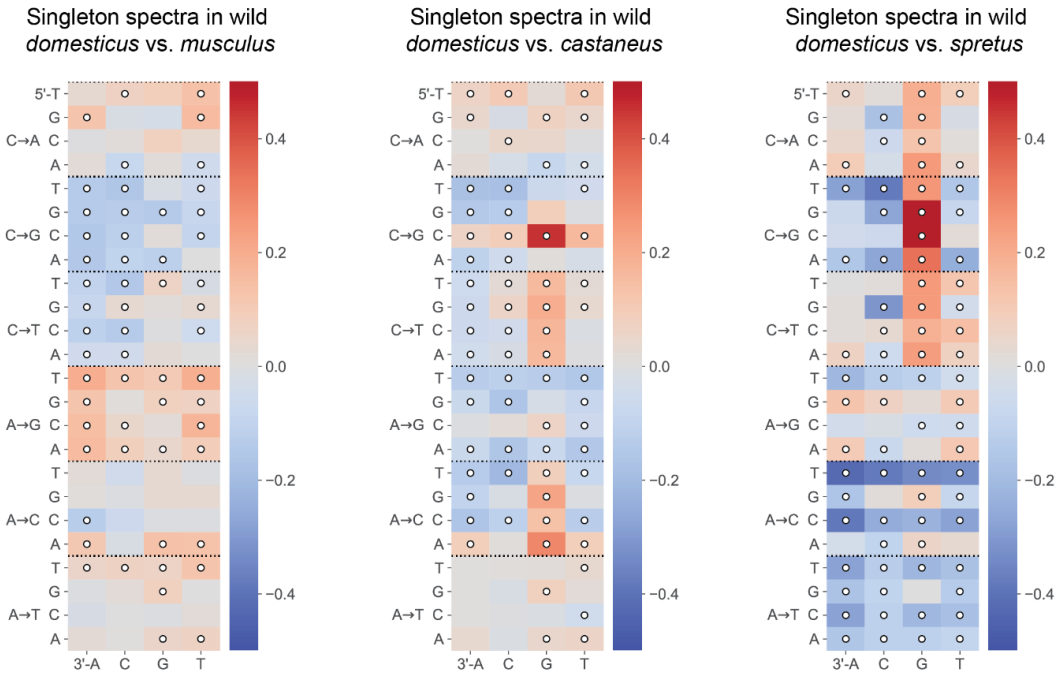

Figure 5: Comparisons of singleton spectra between wild *M.m.domesticus* and other wild species
Log-2 ratios of singleton fractions of each 3-mer mutation type in *Mus musculus domesticus*, compared to three other wild subspecies or species of *Mus*. Comparisons with Chi-square test of independence *p*-values < 0.05/96 are annotated with white circles.

Estimating the strength of selection against mutagenic *Mutyh* variation

Generally, mutators are expected to be disfavored in proportion to the strength of selection against the excess germline mutations they produce (*1*). Because all of the BXDs are inbred, we did not know whether the BXD mutator haplotype was dominant or recessive in this study; any fully recessive mutator or antimutator is expected to segregate neutrally until drifting to high enough frequency to appear in homozygous form. If we assume that the BXD mutator is rare and always heterozygous in wild populations, and that it generates an additional deleterious mutation load (ΔU) per haploid genome, each gamete containing the mutator incurs both an extra deleterious mutation load of ΔU, as well as about half of the additional mutation load produced in the previous generation. Each deleterious passenger mutation stays linked to the mutator for an average of two generations, producing a selective disadvantage of 2ΔU overall. Some additional deleterious effects might result from mutations occurring in the early

embryo throughout the diploid genome, which could slightly increase the selective effect of a mutator.

Using this framework, we conservatively assumed mutations occurring outside of exons to be neutral, and nonsynonymous mutations to have an expected selection coefficient of -5×10^{-3} (*34*). We calculated the mutation rate in BXD68 to be 1.1×10^{-8} per site per generation, about 2-fold higher than the average BXD mutation rate of 5.6×10^{-9}. Assuming a haploid genome size of 2.5×10^{9}, BXD68 is expected to accumulate about 13.5 excess mutations per genome per generation, including 1 extra coding mutation every four generations (assuming that 2% of mutations alter protein sequences). This implies a selection coefficient against the BXD68 mutator of approximately 2.5×10^{-3} (if additive) or 5×10^{-3} (if fully dominant), strong enough to be efficiently selected against in wild populations of *Mus musculus domesticus,* for which effective population sizes have been estimated to be between 5×10^{4} and 2×10^{5} (*35*, *36*). In comparison, the selective advantage of the *B* antimutator, which prevents about 1.5 mutations per haploid genome per generation, would be at most $2 * (1.5 * 0.02) * 5 \times 10^{-3} = 3 \times 10^{-4}$ (assuming additive selection). In theory, this should be advantageous enough to sweep to fixation in a mouse population of effective size $N = 5 \times 10^{4}$ (making $2Ns = 30$). The lack of evidence for such a sweep in wild *M.m.domesticus* may indicate the antimutator is recessive or otherwise suppressed by another allele or environmental factor absent from the BXD.

We also note that our estimates of selection on the BXD mutator and antimutator alleles are based on mutation rates from inbred mice in a highly controlled laboratory environment. In wild mouse populations, any number of environmental factors could affect the advantage or disadvantage of either allele. Moreover, given that the singleton mutation in BXD68 occurs at a position in *Mutyh* at which human variants cause polyposis and cancer susceptibility, it is possible that concomitant somatic phenotypes might reinforce selection against *Mutyh* variants and make them even more deleterious than our calculations predict.

Discussion

Our discovery of a QTL for the C>A mutation rate—to our knowledge, the first natural germline mutator identified in a mammalian species—provides empirical support for long-standing theoretical predictions about mutation rate evolution. Lynch's drift-barrier hypothesis postulates that evolution's ability to optimize the germline mutation rate is limited by the degree to which an antimutator decreases the genome-wide deleterious mutation rate (*U*) (*1*). If a mutation improves the efficacy of a DNA repair gene and lowers the mutation rate by a small amount, it might prevent a few future deleterious mutations; however, this fitness benefit may be

so insignificant that the antimutator will not be effectively favored by natural selection in the face of genetic drift. The C57BL/6J antimutator appears to fit this profile, as it prevents about 1.5 mutations per haploid genome per generation. Although we do not have conclusive proof that the antimutator phenotype is caused by one of the derived nonsynonymous *Mutyh* alleles in C57BL/6J, the drift-barrier hypothesis explains why it is plausible that *Mutyh* could have existed in a suboptimal state for millions of years of rodent evolution, before having its efficacy improved by a single mutation or small combination of mutations still segregating in *Mus musculus*.

Our findings add weight to the exciting possibility that natural mutator alleles underlie some of the species-specific and population-specific signatures previously observed in humans and other great apes (*6*, *8*). Our results also demonstrate that mutators may be mappable in model organisms using classical QTL analysis. Differences in mutation spectra observed across other mouse populations (*16*) suggest that the BXD mutator is just one of several active mutator alleles in mice, any of which might have been detected instead if the "right" parents had been selected to initiate a cross like the BXD. Mutator alleles have also recently been identified in organisms with shorter life cycles and smaller genomes, as demonstrated by recent work in *Saccharomyces cerevisiae (37, 38)*. We anticipate that mutator allele discovery will become increasingly feasible across the tree of life as sequencing costs continue to decline, providing long-awaited data needed to test theoretical predictions about selection on this fundamental phenotype.

Acknowledgments: We thank Beth Dumont, Melissa Gymrek, Milad Mortazavi, Andrew Clark, Daphna Rothschild, Uma Arora, Mikhail Maksimov, Hao Chen, and Jonathan Sebat for contributing helpful feedback as part of the BXD Genome Sequencing Consortium. We also thank Molly Przeworski for providing comments on a manuscript draft, Yu-Yu Ren for assistance with preliminary genotype calling, and members of the Harris and Pritchard labs for additional helpful discussions. We thank members of the staff of HudsonAlpha—Dr. Sean Levy and team—for DNA library preparation and sequencing, and for providing us with great support on data transfer. We thank staff at the UT ISAAC facility for storage and processing of all sequence-associated data files. We thank Arthur Centeno for assisting with the upload of phenotype data to GeneNetwork. Finally, we thank Dr. Cat Lutz, Alicia Valenzuela at The Jackson Laboratory and Jesse Ingels at UTHSC for DNA sample acquisition, handling, and assistance.

Competing interests: The authors declare that they have no competing interests.

Data and materials availability: All code used for data analysis and figure generation is deposited at https://github.com/harrispopgen/bxd_mutator_manuscript, along with mutation calls and other data files necessary to reproduce the manuscript. A VCF file containing all variant calls from the sequenced BXDs is being uploaded to the NCBI Sequence Read Archive.

Supplementary Materials:

Materials and Methods

Tables S1 to S2

Figures S1 to S7

References:

1. M. Lynch, M. S. Ackerman, J.-F. Gout, H. Long, W. Sung, W. K. Thomas, P. L. Foster, Genetic drift, selection and the evolution of the mutation rate. *Nat. Rev. Genet.* **17**, 704–714 (2016).

2. P. D. Sniegowski, P. J. Gerrish, R. E. Lenski, Evolution of high mutation rates in experimental populations of E. coli. *Nature*. **387**, 703–705 (1997).

3. K. J. Dawson, Evolutionarily stable mutation rates. *J. Theor. Biol.* **194**, 143–157 (1998).

4. C. Seoighe, A. Scally, Inference of Candidate Germline Mutator Loci in Humans from Genome-Wide Haplotype Data. *PLoS Genet.* **13**, e1006549 (2017).

5. K. Harris, Evidence for recent, population-specific evolution of the human mutation rate. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 3439–3444 (2015).

6. K. Harris, J. K. Pritchard, Rapid evolution of the human mutation spectrum. *Elife*. **6**, 415 (2017).

7. V. M. Narasimhan, R. Rahbari, A. Scally, A. Wuster, D. Mason, Y. Xue, J. Wright, R. C. Trembath, E. R. Maher, D. A. van Heel, A. Auton, M. E. Hurles, C. Tyler-Smith, R. Durbin, Estimating the human mutation rate from autozygous segments reveals population differences in human mutational processes. *Nat. Commun.* **8**, 303 (2017).

8. M. E. Goldberg, K. Harris, Great ape mutation spectra vary across the phylogeny and the genome due to distinct mutational processes that evolve at different rates. *Cold Spring Harbor Laboratory* (2019), p. 805598.

9. T. A. Sasani, B. S. Pedersen, Z. Gao, L. Baird, M. Przeworski, L. B. Jorde, A. R. Quinlan, Large, three-generation human families reveal post-zygotic mosaicism and variability in germline mutation accumulation. *Elife*. **8** (2019), doi:10.7554/eLife.46922.

10. R. Rahbari, A. Wuster, S. J. Lindsay, R. J. Hardwick, L. B. Alexandrov, S. A. Turki, A.

Dominiczak, A. Morris, D. Porteous, B. Smith, M. R. Stratton, UK10K Consortium, M. E. Hurles, Timing, rates and spectra of human germline mutation. *Nat. Genet.* **48**, 126–133 (2016).

11. M. D. Kessler, D. P. Loesch, J. A. Perry, N. L. Heard-Costa, D. Taliun, B. E. Cade, H. Wang, M. Daya, J. Ziniti, S. Datta, J. C. Celedón, M. E. Soto-Quiros, L. Avila, S. T. Weiss, K. Barnes, S. S. Redline, R. S. Vasan, A. D. Johnson, R. A. Mathias, R. Hernandez, J. G. Wilson, D. A. Nickerson, G. Abecasis, S. R. Browning, S. Zöllner, J. R. O'Connell, B. D. Mitchell, National Heart, Lung, and Blood Institute Trans-Omics for Precision Medicine (TOPMed) Consortium, TOPMed Population Genetics Working Group, T. D. O'Connor, De novo mutations across 1,465 diverse genomes reveal mutational insights and reductions in the Amish founder population. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 2560–2569 (2020).

12. I. Mathieson, D. Reich, Differences in the rare variant spectrum among human populations. *PLoS Genet.* **13**, e1006581 (2017).

13. H. Jónsson, P. Sulem, B. Kehr, S. Kristmundsdottir, F. Zink, E. Hjartarson, M. T. Hardarson, K. E. Hjorleifsson, H. P. Eggertsson, S. A. Gudjonsson, L. D. Ward, G. A. Arnadottir, E. A. Helgason, H. Helgason, A. Gylfason, A. Jonasdottir, A. Jonasdottir, T. Rafnar, M. Frigge, S. N. Stacey, O. Th Magnusson, U. Thorsteinsdottir, G. Masson, A. Kong, B. V. Halldorsson, A. Helgason, D. F. Gudbjartsson, K. Stefansson, Parental influence on human germline de novo mutations in 1,548 trios from Iceland. *Nature*. **549**, 519–522 (2017).

14. L. Ségurel, M. J. Wyman, M. Przeworski, Determinants of Mutation Rate Variation in the Human Germline. *Annu. Rev. Genomics Hum. Genet.* **15**, 47–70 (2014).

15. D. G. Ashbrook, D. Arends, P. Prins, M. K. Mulligan, S. Roy, E. G. Williams, C. M. Lutz, A. Valenzuela, C. J. Bohl, J. F. Ingels, M. S. McCarty, A. G. Centeno, R. Hager, J. Auwerx, L. Lu, R. W. Williams, A Platform for Experimental Precision Medicine: The Extended BXD Mouse Family. *Cell Systems* (2021), doi:10.1016/j.cels.2020.12.002.

16. B. L. Dumont, Significant Strain Variation in the Mutation Spectra of Inbred Laboratory Mice. *Mol. Biol. Evol.* **36**, 865–874 (2019).

17. A. Uchimura, M. Higuchi, Y. Minakuchi, M. Ohno, A. Toyoda, A. Fujiyama, I. Miura, S. Wakana, J. Nishino, T. Yagi, Germline mutation rates and the long-term phenotypic effects of mutation accumulation in wild-type laboratory mice and mutator mice. *Genome Res.* **25**, 1125–1134 (2015).

18. S. J. Lindsay, R. Rahbari, J. Kaplanis, T. Keane, M. E. Hurles, Similarities and differences in patterns of germline mutation between mice and humans. *Nat. Commun.* **10**, 4053 (2019).

19. K. W. Broman, D. M. Gatti, P. Simecek, N. A. Furlotte, P. Prins, Ś. Sen, B. S. Yandell, G. A. Churchill, R/qtl2: Software for Mapping Quantitative Trait Loci with High-Dimensional Data and Multiparent Populations. *Genetics*. **211**, 495–502 (2019).

20. P. Cingolani, A. Platts, L. L. Wang, M. Coon, T. Nguyen, L. Wang, S. J. Land, X. Lu, D. M. Ruden, A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* . **6**, 80–92 (2012).

21. S. S. David, V. L. O'Shea, S. Kundu, Base-excision repair of oxidative DNA damage. *Nature*. **447**, 941–950 (2007).

22. M. K. Mulligan, K. Mozhui, P. Prins, R. W. Williams, GeneNetwork: A Toolbox for Systems Genetics. *Methods Mol. Biol.* **1488**, 75–120 (2017).

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

24. N. Al-Tassan, N. H. Chmiel, J. Maynard, N. Fleming, A. L. Livingston, G. T. Williams, A. K. Hodges, D. Rhodri Davies, S. S. David, J. R. Sampson, J. P. Cheadle, Inherited variants of MYH associated with somatic G:C→T:A mutations in colorectal tumors. *Nat. Genet.* **30** (2002), doi:10.1038/ng828.

25. C. Pilati, J. Shinde, L. B. Alexandrov, G. Assié, T. André, Z. Hélias-Rodzewicz, R. Ducoudray, D. Le Corre, J. Zucman-Rossi, J.-F. Emile, J. Bertherat, E. Letouzé, P. Laurent-Puig, Mutational signature analysis identifies MUTYH deficiency in colorectal cancers and adrenocortical carcinomas. *J. Pathol.* **242**, 10–15 (2017).

26. A. Viel, A. Bruselles, E. Meccia, M. Fornasarig, M. Quaia, V. Canzonieri, E. Policicchio, E. D. Urso, M. Agostini, M. Genuardi, E. Lucci-Cordisco, T. Venesio, A. Martayan, M. G. Diodoro, L. Sanchez-Mete, V. Stigliano, F. Mazzei, F. Grasso, A. Giuliani, M. Baiocchi, R. Maestro, G. Giannini, M. Tartaglia, L. B. Alexandrov, M. Bignami, A Specific Mutational Signature Associated with DNA 8-Oxoguanine Persistence in MUTYH-defective Colorectal Cancer. *EBioMedicine*. **20**, 39–49 (2017).

27. J. G. Tate, S. Bamford, H. C. Jubb, Z. Sondka, D. M. Beare, N. Bindal, H. Boutselakis, C. G. Cole, C. Creatore, E. Dawson, P. Fish, B. Harsha, C. Hathaway, S. C. Jupe, C. Y. Kok, K. Noble, L. Ponting, C. C. Ramshaw, C. E. Rye, H. E. Speedy, R. Stefancsik, S. L. Thompson, S. Wang, S. Ward, P. J. Campbell, S. A. Forbes, COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res.* **47**, D941–D947 (2019).

28. M. J. Landrum, S. Chitipiralla, G. R. Brown, C. Chen, B. Gu, J. Hart, D. Hoffman, W. Jang, K. Kaur, C. Liu, V. Lyoshin, Z. Maddipatla, R. Maiti, J. Mitchell, N. O'Leary, G. R. Riley, W. Shi, G. Zhou, V. Schneider, D. Maglott, J. B. Holmes, B. L. Kattman, ClinVar: improvements to accessing data. *Nucleic Acids Res.* **48**, D835–D844 (2020).

29. Y. Choi, A. P. Chan, PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics*. **31**, 2745–2747 (2015).

30. P. C. Ng, S. Henikoff, SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **31**, 3812–3814 (2003).

31. B. Harr, E. Karakoc, R. Neme, M. Teschke, C. Pfeifle, Ž. Pezer, H. Babiker, M. Linnenbrink, I. Montero, R. Scavetta, M. R. Abai, M. P. Molins, M. Schlegel, R. G. Ulrich, J. Altmüller, M. Franitza, A. Büntge, S. Künzel, D. Tautz, Genomic resources for wild populations of the house mouse, Mus musculus and its close relative Mus spretus. *Sci Data*. **3** (2016), p. 160075.

32. H. Yang, J. R. Wang, J. P. Didion, R. J. Buus, T. A. Bell, C. E. Welsh, F. Bonhomme, A. H.-T. Yu, M. W. Nachman, J. Pialek, P. Tucker, P. Boursot, L. McMillan, G. A. Churchill, F. P.-

M. de Villena, Subspecific origin and haplotype diversity in the laboratory mouse. *Nat. Genet.* **43**, 648–655 (2011).

33. T. M. Keane, L. Goodstadt, P. Danecek, M. A. White, K. Wong, B. Yalcin, A. Heger, A. Agam, G. Slater, M. Goodson, N. A. Furlotte, E. Eskin, C. Nellåker, H. Whitley, J. Cleak, D. Janowitz, P. Hernandez-Pliego, A. Edwards, T. G. Belgard, P. L. Oliver, R. E. McIntyre, A. Bhomra, J. Nicod, X. Gan, W. Yuan, L. van der Weyden, C. A. Steward, S. Bala, J. Stalker, R. Mott, R. Durbin, I. J. Jackson, A. Czechanski, J. A. Guerra-Assunção, L. R. Donahue, L. G. Reinholdt, B. A. Payseur, C. P. Ponting, E. Birney, J. Flint, D. J. Adams, Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature*. **477**, 289–294 (2011).

34. C. D. Huber, B. Y. Kim, C. D. Marsden, K. E. Lohmueller, Determining the factors driving selective effects of new nonsynonymous mutations. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 4465–4470 (2017).

35. A. Geraldes, P. Basset, B. Gibson, K. L. Smith, B. Harr, H.-T. Yu, N. Bulatova, Y. Ziv, M. W. Nachman, Inferring the history of speciation in house mice from autosomal, X-linked, Y-linked and mitochondrial genes. *Mol. Ecol.* **17**, 5349–5363 (2008).

36. M. Phifer-Rixey, F. Bonhomme, P. Boursot, G. A. Churchill, J. Piálek, P. K. Tucker, M. W. Nachman, Adaptive evolution and effective population size in wild house mice. *Mol. Biol. Evol.* **29**, 2949–2955 (2012).

37. L. Gou, J. S. Bloom, L. Kruglyak, The Genetic Basis of Mutation Rate Variation in Yeast. *Genetics*. **211**, 731–740 (2019).

38. P. Jiang, A. R. Ollodart, V. Sudhesh, A. J. Herr, M. J. Dunham, K. Harris, A modified fluctuation assay reveals a natural mutator phenotype that drives mutation spectrum variation within Saccharomyces cerevisiae. *Cold Spring Harbor Laboratory* (2021), p. 2021.01.11.425955.

39. J. Köster, S. Rahmann, Snakemake--a scalable bioinformatics workflow engine. *Bioinformatics*. **28**, 2520–2522 (2012).

40. M. A. DePristo, E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, G. del Angel, M. A. Rivas, M. Hanna, A. McKenna, T. J. Fennell, A. M. Kernytsky, A. Y. Sivachenko, K. Cibulskis, S. B. Gabriel, D. Altshuler, M. J. Daly, A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).

41. X. Wang, R. Agarwala, J. A. Capra, Z. Chen, D. M. Church, D. C. Ciobanu, Z. Li, L. Lu, K. Mozhui, M. K. Mulligan, S. F. Nelson, K. S. Pollard, W. L. Taylor, D. B. Thomason, R. W. Williams, High-throughput sequencing of the DBA/2J mouse genome. *BMC Bioinformatics*. **11**, O7 (2010).

42. B. S. Pedersen, A. R. Quinlan, cyvcf2: fast, flexible variant analysis with Python. *Bioinformatics*. **33**, 1867–1869 (2017).

43. J. Schreiber, Pomegranate: fast and flexible probabilistic modeling in python. *J. Mach. Learn. Res.* **18**, 5992–5997 (2017).

44. W. S. DeWitt, mutyper: assigning and summarizing mutation types for analyzing germline

mutation spectra. *Cold Spring Harbor Laboratory* (2020), p. 2020.07.01.183392.

45. S. Neph, M. S. Kuehn, A. P. Reynolds, E. Haugen, R. E. Thurman, A. K. Johnson, E. Rynes, M. T. Maurano, J. Vierstra, S. Thomas, R. Sandstrom, R. Humbert, J. A. Stamatoyannopoulos, BEDOPS: high-performance genomic feature operations. *Bioinformatics*. **28**, 1919–1920 (2012).

46. A. R. Quinlan, BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Curr. Protoc. Bioinformatics*. **47**, 11.12.1–11.12.34 (2014).

47. B. S. Pedersen, A. R. Quinlan, Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics*. **34**, 867–868 (2018).

48. E. L. Green, *Genetics and Probability in Animal Breeding Experiments: A primer and reference book on probability, segregation, assortment, linkage and mating systems for biomedical scientists who breed and use genetically defined laboratory animals for research* (Macmillan International Higher Education, 1981).

49. M. F. Lyon, A. G. Searle, *Genetic variants and strains of the laboratory mouse* (Oxford University Press, 1989).