Discovering epistasis between germline mutator alleles in mice

This manuscript (permalink) was automatically generated from quinlan-lab/mutator-epistasis-manuscript@d2bc0c3 on March 1, 2023.

Authors

- **Thomas A. Sasani**
  0000-0003-2317-1374 · tomsasani · tomsasani
  Department of Human Genetics, University of Utah · Funded by Grant XXXXXXXX

- **Aaron R. Quinlan** ✉
  0000-0003-1756-0859 · aaronquinlan
  Department of Human Genetics, University of Utah; Department of Biomedical Informatics, University of Utah

- **Kelley Harris** ✉
  0000-0003-0302-2523 · Kelley__Harris
  Department of Genome Sciences, University of Washington

✉ — Correspondence possible via GitHub Issues or email to Aaron R. Quinlan <aquinlan@genetics.utah.edu>, Kelley Harris <harriske@uw.edu>.

Abstract

Maintaining genome integrity in the mammalian germline is essential and enormously complex. Hundreds of proteins comprise pathways involved in DNA replication, and hundreds more are mobilized to repair DNA damage [1]. While loss-of-function mutations in any of the genes encoding these proteins might lead to elevated mutation rates, *mutator alleles* have largely eluded detection in mammals.

DNA replication and repair proteins often recognize particular sequence motifs or excise lesions at specific nucleotides. Thus, we might expect that the spectrum of de novo mutations — i.e, the frequency of each individual mutation type (C>T, A>G, etc.) — will differ between genomes that harbor either a mutator or wild-type allele at a given locus. Previously, we used quantitative trait locus (QTL) mapping to discover a mutator allele near the DNA repair gene *Mutyh* that increases the rate of *de novo* C>A germline mutation in a collection of recombinant inbred lines (RILs) known as the BXDs [2,3].

In this study, we developed a new method to detect alleles that affect the mutation spectrum in biparental RILs. By applying this method to mutation data from the BXDs, we confirmed the activity of the germline mutator locus near *Mutyh*, and discovered an additional C>A germline mutator locus on chromosome 6 that overlaps *Ogg1*, a key partner of *Mutyh* in base-excision repair of oxidative DNA damage [4]. Strikingly, BXDs with the mutator allele near *Ogg1* do not exhibit elevated rates of C>A germline mutation unless they also possess the mutator allele near *Mutyh*, but BXDs with both alleles exhibit even higher C>A mutation rates than those with either one alone.

To our knowledge, these new methods for analyzing mutation spectra reveal the first evidence of epistasis between mammalian germline mutator alleles, and may be applicable to mutation data from humans and other model organisms.

Introduction

The germline mutation rate is a fundamental parameter in population genetics, and reflects the complex interplay between DNA replication and repair pathways, exogenous sources of DNA damage, and life-history traits. *Mutator alleles* may explain some of the within- and between-species variation in germline mutation rates [5], but have proven challenging to identify in mammalian genomes.

Germline mutator alleles are difficult to detect for a number of reasons, including the fidelity of germline genome replication and the effects of selection on mutators. On average, humans are born with 30 to 50 single-nucleotide *de novo* germline mutations per haploid genome [6,7]; in mice, that number is closer to 10 or 15 [8]. Due to the low baseline germline mutation rate in many mammals, it can be challenging to ascertain sequencing data from enough haplotypes to reliably detect those with significantly elevated *de novo* mutation counts. Moreover, in a population of sufficiently large N_e (effective population size), large-effect mutator alleles will likely be efficiently purged by negative selection. The estimated selection coefficient on a mutator allele is approximately $2s\Delta U$ [9], where s is the mean selective coefficient on a new deleterious mutation and ΔU is the excess number of new deleterious mutations caused by the mutator allele; the product of s and ΔU is multiplied by 2 to account for the average number of generations for which mutator is linked to the excess mutations it causes.

Compared to haplotypes that harbor wild-type alleles at a particular locus, those harboring mutator alleles will likely carry an excess of total germline mutations. Indeed, candidate germline mutator loci have been discovered in human genomes by identifying haplotypes with significantly more derived alleles than the population mean [10]. However, protein-coding genes involved in DNA replication and repair often recognize particular sequence motifs or excise lesions at specific nucleotides [5]. Thus, we might also expect the spectrum of de novo mutations — that is, the frequency of each individual mutation type (C>T, A>G, etc.) — to differ between genomes that carry either a mutator or wild-type allele at a given locus.

In 2022, we discovered a germline mutator allele in mice by analyzing whole-genome sequencing data from 152 recombinant inbred lines (RILs). Commonly known as the BXDs [3], these RILs were derived from either F2 or advanced intercrosses of C57BL/6J and DBA/2J, two laboratory strains that exhibit significant differences in their germline mutation spectra [11]. Since the BXD RILs were maintained via brother-sister mating for up to 180 generations and housed in a controlled laboratory environment, they were an ideal population for mutator allele discovery. Namely, each line accumulated hundreds or thousands of germline mutations on a nearly-homozygous linear mosaic of parental B and D haplotypes, while the effects of negative selection on new and standing variation were attenuated by strict inbreeding [12]. We used quantitative trait locus (QTL) mapping to identify a locus on chromosome 4 that was strongly associated with the C>A germline mutation rate in the BXDs [2]. The QTL overlapped *Mutyh*, which encodes a protein that normally prevents C>A mutations by repairing oxidative DNA damage [4], and we hypothesized that missense mutations in *Mutyh* were responsible for a 50% increase in the C>A mutation rate between BXDs with either parental haplotype at the QTL [2].

In this study, we developed a new method to detect alleles that affect the mutation spectrum in biparental RILs, and applied it to *de novo* germline mutation data from the BXDs. We assessed its power to detect candidate mutator alleles, re-identified the mutator near *Mutyh*, and discovered compelling evidence of epistasis between two germline mutator alleles that augment the C>A germline mutation rate.

Results

Benchmarking the inter-haplotype distance method using simulations

We first tested the inter-haplotype distance approach using simulated data (Materials and Methods). We find that the method's power is mostly limited by the initial mutation rate of the k-mer mutation type affected by the mutator allele and the total number of *de novo* germline mutations in the dataset (that is, the product of the number of haplotypes and the mean number of mutations per haplotype) (Figure 1). For example, given 50 haplotypes with an average of 500 *de novo* germline mutations each, our method has nearly 90% power detect a mutator allele that increases the C>T *de novo* mutation rate by 10%. However, the method only has about 10% power to detect a mutator of identical effect size that affects the C>G mutation rate, since C>G mutations are expected to make up a much smaller fraction of all *de novo* germline mutations to begin with. These simulations also demonstrate that our method is well-powered to detect large-effect mutator alleles (e.g., those that increase the mutation rate of a specific k-mer by 50%), even with a relatively small number of mutations per haplotype.

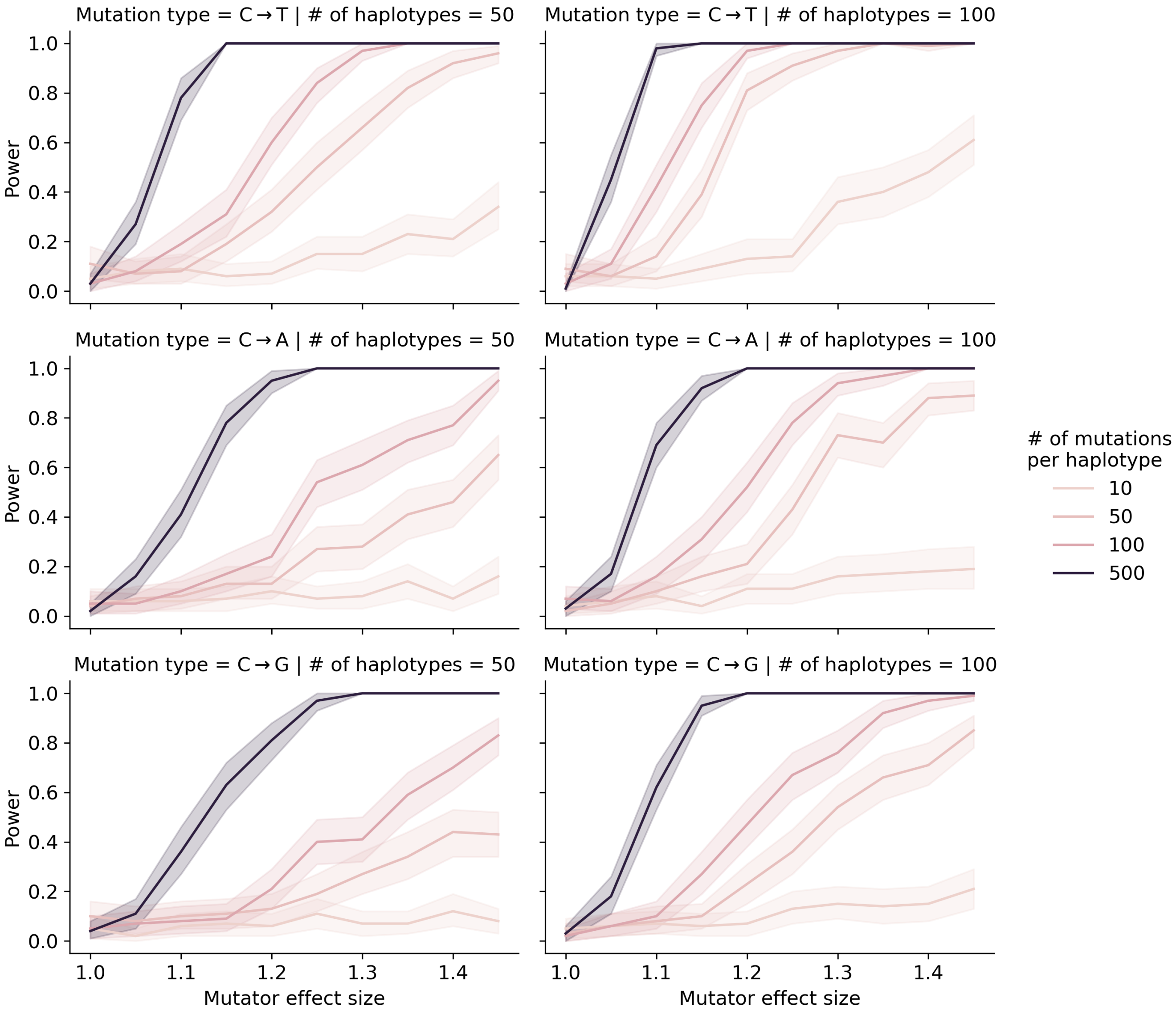

Figure 1: Simulations to assess the power of the inter-haplotype distance method. We simulated *de novo* germline mutations on the specified number of haplotypes, such that 50% of haplotypes were affected by a mutator allele that increased the mutation rate of the specified k-mer by the specified effect size (an effect size of 1.5 indicates a 50% increase in the mutation rate). The colors of the lines indicate the number of simulated mutations on each haplotype (before augmenting the mutation rate with a mutator allele). Given a specific combination of parameters, the y-axis denotes the fraction of 100 simulations in which the simulated mutator allele could be detected at a p-value of 0.05. Shaded areas indicate the standard deviation of that fraction.

Re-identifying the mutator allele on chromosome 4 in the BXDs

We applied our inter-haplotype distance method to 93 BXD RILs (Materials and Methods) with a total of 62,993 *de novo* germline mutations [2]. Reassuringly, using 1-mer mutation spectra, we observed a large χ^2 statistic peak at a locus on chromosome 4 (Figure 2A; maximum χ^2 statistic of 352.7 at marker ID `rs52263933`; position 116.75 Mbp in GRCm38/mm10 coordinates). We observed the same peak on chromosome 4 using the $3-mer$ mutation spectrum, as well (Figure 4).

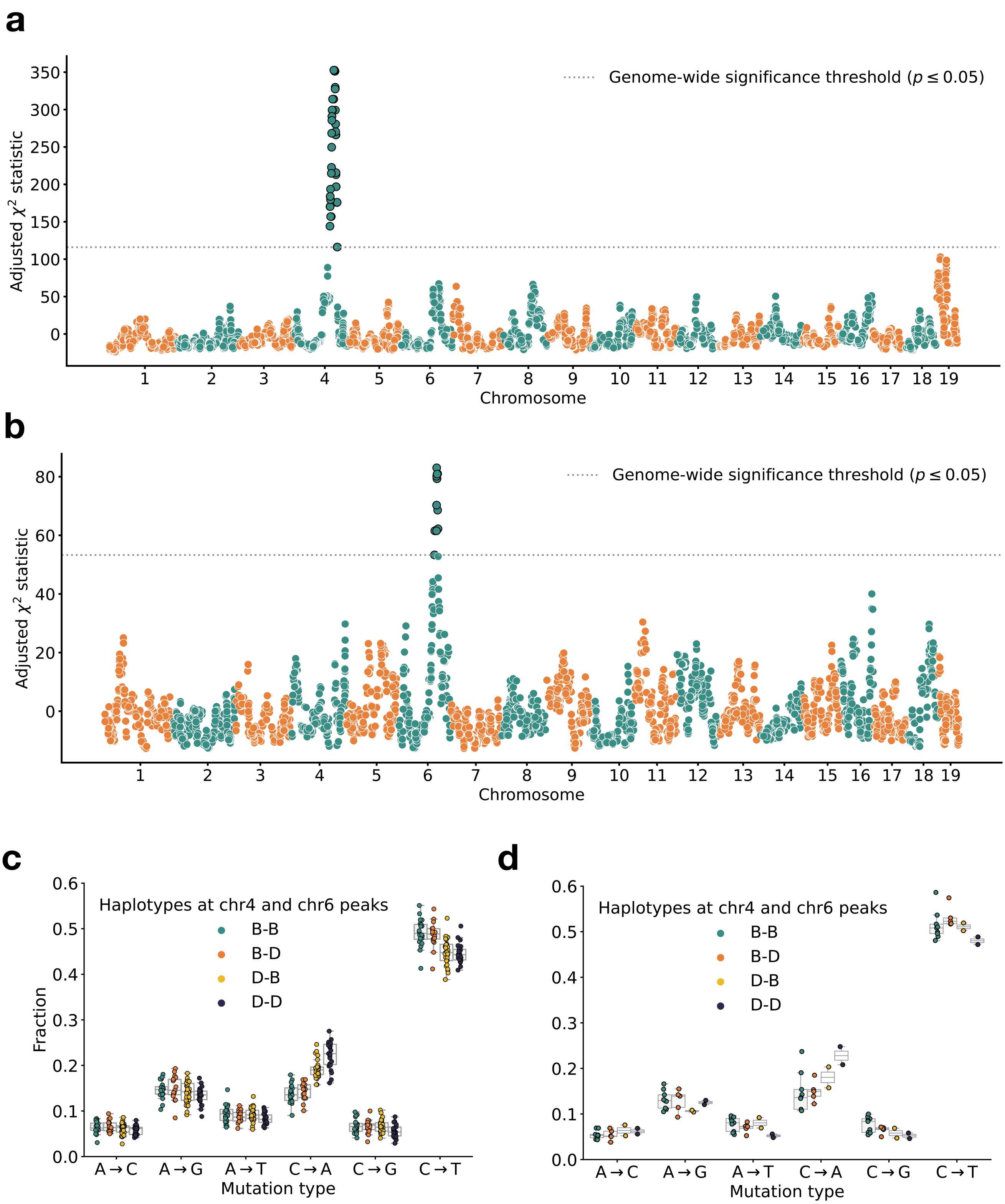

Figure 2: Results of inter-haplotype distance scans in the BXD RILs. a) χ^2 statistics between aggregate 1-mer *de novo* mutation spectra on BXD haplotypes (n = 93 haplotypes; 62,993 total mutations) with either *D* or *B* alleles at 7,320 informative markers. Distance threshold at $p = 0.05$ was calculated by performing 10,000 permutations of the BXD haplotype mutation data, and is shown as a dotted grey line. **b)** χ^2 statistics between aggregate $1 - mer$ *de novo* mutation spectra on BXD haplotypes with *D* alleles at `rs52263933` (n = 55 haplotypes; 40,913 total mutations) and either *D* or *B* alleles at 7,320 informative markers. Distance threshold at $p = 0.05$ was calculated by performing 10,000 permutations of the BXD haplotype mutation data, and is shown as a dotted grey line. **c)** Fractions of *de novo* germline mutations in BXDs with either *D* or *B* haplotypes at markers `rs52263933` and `rs31001331`, stratified by mutation type. **d)** Fractions of *de novo* germline mutations in Sanger Mouse Genome Project (MGP) strains with either *D* or *B* haplotypes at markers `rs52263933` and `rs31001331`, stratified by mutation type.

In a previous analysis, we used quantitative trait locus (QTL) mapping to identify a nearly identical locus on chromosome 4 that was significantly associated with the C>A germline mutation rate in the BXDs [2]. This locus overlaps 21 protein-coding genes that are annotated by the Gene Ontology as being involved in "DNA repair," but only one of these genes contains non-synonymous differences between the two parental strains: *Mutyh*. *Mutyh* encodes a protein involved in the base-excision repair of 8-oxoguanine (8-oxoG), a DNA lesion caused by oxidative damage, and prevents the accumulation of C>A mutations [4,13,14]. C>A germline mutation rates are nearly 50% higher in BXDs that inherited *D* haplotypes at marker ID `rs52263933` than in those that inherited *B* haplotypes [2].

An additional germline mutator allele on chromosome 6

After confirming that the inter-haplotype distance method could recover the mutator locus overlapping *Mutyh*, we asked if our approach could identify additional mutator loci in the BXD. To account for the effects of the large-effect C>A germline mutator locus near *Mutyh*, we divided the BXD RILs into those with either *D* (n = 55) or *B* (n = 38) genotypes at `rs52263933` (the marker at which we observed the highest inter-haplotype χ^2 statistic on chromosome 4), and ran a genome-wide distance scan using each group separately (Figure 2B.

Using only the BXDs with *B* genotypes at the *Mutyh* mutator locus, we did not observe any genome-wide significant peaks. But using the BXDs with *D* genotypes at the same locus, we identified a χ^2 statistic peak on chromosome 6 (Figure 2B; maximum χ^2 statistic of 81.0 at marker `rs31001331`; position 114.05 Mbp in GRCm38/mm10 coordinates). We identified the same peak using the $3 - mer$ mutation spectra, as well (Figure 4). We queried the region underneath this peak (+/- 5 Mbp) and discovered 87 protein-coding genes. Remarkably, only one was both annotated with the Gene Ontology term "DNA repair" and contained nonsynonymous differences between C57BL/6J and DBA/2J: *Ogg1*. *Ogg1* encodes a key member of the base-excision repair response to oxidative DNA damage, a pathway that also includes *Mutyh*. *Ogg1* harbors a single fixed nonsynonymous differences between the C57BL/6J and DBA/2J parental strains: p.Thr95Ala, at position 113,328,510 on chromosome 6 in GRCm38/mm10 coordinates.

We also considered the possibility that expression quantitative trait loci (eQTLs), rather than nonsynonymous mutations, could contribute to the C>A mutator phenotype associated to the locus on chromosome 6. Using GeneNetwork [15], we mapped cis-eQTLs for *Ogg1* in a number of tissues, including hematopoetic stem cells, kidney, and spleen. BXD genotypes near the χ^2 statistic peak on chromosome 6 were significantly associated with *Ogg1* expression in some (but not all) tissues, and *D* genotypes were nearly always associated with decreased gene expression (Table 1). We also queried a previously published collection of eQTLs derived from Diversity Outbred (DO) mouse embryonic stem cell (mESC) expression data [16], but did not find any significant eQTLs for *Ogg1*.

Table 1: Presence or absence of cis-eQTLs for *Ogg1* in various tissues identified using GeneNetwork.

Tissue name	# BXDs with expression data	Top significant marker	LRS at top significant marker	Significant LRS threshold	Additive effect of D allele on expression
Kidney	53	`rsm10000004188`	52.25	17.82	-0.186
Gastrointestinal	46	`rsm10000003441`	23.39	16.09	-0.074
Hematopoetic stem cells	22	-	-	16.45	-
Hematopoetic progenitor cells	23	-	-	18.52	-

Tissue name	# BXDs with expression data	Top significant marker	LRS at top significant marker	Significant LRS threshold	Additive effect of D allele on expression
Spleen	79	`rsm10000003418`	-	17.51	-
Liver	50	`rsm10000004188`	53.54	18.77	-0.156
Heart	73	-	-	16.22	-
Eye	87	`rsm10000004194`	23.05	16.96	0.088

Evidence of epistasis between germline mutator alleles

Next, we more precisely characterized the effects of the *Mutyh* and *Ogg1* mutator alleles on mutation spectra in the BXDs. We observed that C>A germline mutation fractions in BXDs with *D* alleles at both mutator loci were significantly higher than C>A fractions in BXDs with *D* alleles at either locus alone (Figure 2C). However, compared to BXDs with *B* alleles at the chromosome 6 mutator locus, those with *D* alleles did not exhibit significantly higher C>A mutation fractions, indicating that the effects of the chromosome 6 mutator locus depend on the presence of a *D* allele at the chromosome 4 locus (Figure 2C). To more formally test for epistasis, we fit a linear model predicting C>A mutation rates as a function of genotypes at `rs52263933` and `rs31001331` (the peak markers at the chr4 and chr6 mutator loci, respectively) (Materials and Methods). A model that included an interaction term between genotypes at the two markers fit the data significantly better (p = 0.0048) than a model including only additive effects of the two markers.

To explore the effects of the two mutator loci in other inbred laboratory mice, we also compared the germline mutation spectra of Sanger Mouse Genomes Project (MGP) strains. Dumont [11] previously identified private germline mutations in 29 inbred laboratory strains; these private variants likely represent recent *de novo* germline mutations (Figure 2D). Only two of the MGP strains possess *D* genotypes at both the chromosome 4 and chromosome 6 mutator loci: DBA/1J and DBA/2J. As before, we tested for epistasis in the MGP strains by fitting two linear models predicting C>A mutation rates as a function of genotypes at `rs52263933` and `rs31001331`. A model incorporating an interaction term between genotypes at these loci did not fit the data significantly better than a model with additive effects alone (p = 0.474). Thus, we are unable to confirm the signal of epistasis observed in the BXDs, but this may be due to the smaller number of MGP strains with *de novo* germline mutation data.

The candidate *Ogg1* mutator allele is present in wild mice

To determine whether the candidate mutator allele on chromosome 6 was segregating in natural populations, we queried previously published sequencing data generated from 67 wild-derived mice [17]. These data include three subspecies of *Mus musculus*, as well as the outgroup *Mus spretus*. We found that the *D* allele in *Ogg1* was segregating at approximately 25% frequency in *Mus musculus domesticus*, the species from which C57BL/6J and DBA/2J derive the majority of their genomes [18], and was fixed in *Mus musculus musculus*, *Mus musculus castaneus*, and the outgroup *Mus spretus*.

Discussion

Epistasis between germline mutator alleles

To our knowledge, these results reveal evidence of epistasis between mammalian germline mutator alleles for the first time. BXDs with *D* alleles at the mutator locus on chromosome 6 only exhibit elevated C>A mutation rates if they also carry *D* alleles at the previously-identified [2] mutator locus on chromosome 4. And BXDs with *D* alleles at both loci have significantly higher C>A germline mutation rates than lines with *D* alleles at only one mutator locus alone (Figure 2C). This raises the exciting possibility that epistasis between mutator alleles has contributed to the evolution of germline mutation rates and spectra in mammalian genomes.

Importantly, we note that we observed epistasis between germline mutator alleles in an unnatural population; the BXDs were inbred by brother-sister mating in a highly controlled laboratory environment that attenuated the effects of natural selection on all but the most deleterious alleles [12]. However, we found the *D* allele in *Ogg1* to be at nearly 25% frequency in *Mus musculus domesticus*, the strain from which C57BL/6J and DBA/2J derive most of their genomes [18]. Since the *D* mutator haplotype on chromosome 6 does not appear to increase the C>A germline mutation rate on its own (even in a homozygous state), we hypothesize that similar alleles may be at intermediate or high frequency in other natural populations.

Causal variants underlying the mutator allele

Only one DNA repair gene overlapping the C>A mutator locus on chromosome 6 also contained nonsynonymous fixed differences between the C57BL/6J and DBA/2J founder strains: *Ogg1*, a protein-coding gene that participates in base-excision repair of the oxidative DNA lesion 8-oxoguanine (8-oxoG) [4]. Both missense mutations and loss-of-heterozygosity in *Ogg1* have been associated with initiation and progression of various types of human cancer [19,20,21]. Unrepaired 8-oxoG lesions can also lead to C>A mutations, and copy-number losses of either *Ogg1* or *Mutyh* are linked to elevated rates of spontaneous C>A mutation in human neuroblastoma [22]. Given these various lines of evidence, we believe that *Ogg1* is the most likely candidate gene to explain the additional C>A mutator phenotype in the BXDs, but it remains unclear whether the p.Thr95Ala missense mutation is the causal allele. We hypothesized that *Mutyh* missense mutations on *D* haplotypes were responsible for the large-effect C>A mutator phenotype we previously observed in the BXDs [2]. However, using high-quality long-read assemblies of inbred laboratory strains, another group recently identified a ~5 kbp mobile element insertion (MEI) within the first intron of *Mutyh* [23] that is present on *D* haplotypes and absent from *B* haplotypes. The MEI is associated with significantly reduced expression of *Mutyh* in embryonic stem cells from laboratory strains, and may therefore underlie the previous C>A germline mutator phenotype in the BXDs. In light of this new evidence, we cannot discount the possibility that eQTLs associated with decreased expression of *Ogg1* (Table 1) are responsible for the C>A mutator phenotype we observed in this study.

Mechanism of epistasis between *Mutyh* and *Ogg1* mutator alleles

Mutyh and *Ogg1* are key members of the base-excision repair (BER) response to 7,8-dihydro-8oxo-deoxyguanine (8-oxoG), one of the most common products of DNA damage by reactive oxygen species [4,24]. *Mutyh* and *Ogg1* fulfill two distinct roles in the BER response to 8-oxoG. If a cell is not actively undergoing DNA replication, *Ogg1* can excise the 8-oxoG lesion, leaving behind an unpaired cytosine on the opposite strand [24]. Then, additional BER proteins can incorporate an unmodified guanine and restore the appropriate G:C base-pair. However, if the 8-oxoG lesion is not repaired prior to a single round of DNA replication, DNA polymerases will often misincorporate adenines (via Hoogsteen base-pairing) opposite the 8-oxoG lesion. In this case, *Mutyh* can excise the mispaired adenine and enable *Ogg1* (along with other members of the BER pathway) to correct the lesion [22].

Given the distinct roles of *Mutyh* and *Ogg1* in the BER pathway, it seems plausible that mutator alleles in both protein-coding genes could together exhibit non-additive effects on C>A mutation rates.

However, even in cells with functional copies of *Mutyh*, repair of 8-oxoG likely requires the activity of *Ogg1*. It is therefore somewhat surprising that the *D* haplotype at *Ogg1*, which further augments the effects of the *Mutyh* mutator haplotype on C>A mutation rates, has no apparent effect on its own. One possibility is that *D* haplotype overlapping *Ogg1* does, in fact, lead to elevated C>A mutation rates, but that we are underpowered to detect its effect in this study. Notably, copy-number-loss of *Ogg1* in human neuroblastoma causes a much more modest increase in C>A mutation rates than copy-number-loss of *Mutyh* [22].

Discovering mutator alleles in other systems

Numerous lines of evidence suggest that mutator alleles contribute to variation in mutation rates and spectra across the tree of life. In two natural isolates of *Saccharomyces cerevisiae*, nonsynonymous variation in *OGG1* causes a substantial increase in the C>A *de novo* mutation rate [25]. Recent analyses have suggested that mutator alleles and/or environmental mutagens have shaped mutation rate evolution both in human genomes [26] and more broadly during great ape evolution [27]. The heritability of paternal *de novo* mutation counts in the human germline has also been estimated to be between 10 and 20%, demonstrating a contribution of genetic factors to germline mutation rates [28]). However, mutator discovery remains challenging in mammalian genomes.

What conditions must be met in order to detect a germline mutator allele? Presumably, one must have access to many haplotypes, each with a reasonably large number of *de novo* germline mutations that remain linked to the mutator allele(s) that caused them. Recently, thousands of human pedigrees have been sequenced in an effort to precisely estimate the rate of human *de novo* germline mutation [6,7]. Selection on germline mutator alleles will likely prevent large-effect mutators from reaching high allele frequencies; however, if multiple mutators are active in a particular population, it becomes much more likely that a subset will be detectable by sequencing human trios [29]. Current estimates of power to detect germline mutators in human pedigrees generally assume that mutators affect all mutation types equally, and that methods for mutator discovery will rely on identifying haplotypes with excess total mutation counts [29]. However, our results in the BXD suggest that germline mutators often exert their effects on a small number of k-mer mutation types, and may be far more amenable to detection by analyzing mutation spectra instead.

Using germline mutation spectra to identify mutator alleles

Germline mutation spectra are a rich source of information about the demographic history of populations, as well as the activity of both exogenous and endogenous sources of mutation throughout time. For example, by analyzing the 3-mer mutation spectrum in a collection of human genomes, Harris and Pritchard [30] discovered a “pulse” of TCC>TTC mutation activity in European populations that likely occurred between 15,000 and 2,000 years ago, and perhaps began even earlier [31]).

Within somatic tissues, mutation spectra can also be used to uncover the mutational processes active in particular populations of cells [32]. New computational methods have been developed to extract “mutational signatures” from large databases of somatic mutations in cancer [33]. These signatures, which describe the relative frequency of each 3-mer mutation type, can often be precisely attributed to chemotherapeutic agents, exposures to environmental mutagens, or loss-of-function mutations in genes encoding DNA repair or replication proteins [32,34,35].

Although a germline mutator allele should increase the absolute count of mutations on a linked haplotype, our results demonstrate that its effects can be more easily detectable by examining mutation *spectra* instead. For example, *D* alleles at the mutator locus on chromosome 6 augment the C>A mutation rate by a factor of approximately 1.2 (Figure 2). Since C>A mutations comprise approximately 10% of all germline mutations to begin with, *D* alleles only increase the overall germline mutation rate by about 2%. Given the depth of information that can be encoded in the mutation spectrum, we expect that mutation spectra can be further exploited to discover genetic modifiers of the mutation rate in other study systems, as well.

Materials and Methods

Identifying *de novo* germline mutations in the BXD RILs

The BXD resource currently comprises a total of 152 recombinant inbred lines (RILs). RILs were derived from either F2 or advanced intercrosses, and subsequently inbred by brother-sister mating for up to 180 generations [3]. BXDs were generated in distinct breeding "epochs," which were each initiated with a distinct cross of C57BL/6J and DBA/2J parents; epochs 1, 2, 4, and 6 were derived from F2 crosses, while epochs 3 and 5 were derived from advanced intercrosses [3]. Previously, we analyzed whole-genome sequencing data from the BXDs and identified candidate *de novo* germline mutations in each line [2]. A detailed description of the methods used for DNA extraction, sequencing, alignment, and variant processing, as well as the characteristics of the *de novo* mutations, are available in a previous manuscript [2].

Briefly, we identified private single-nucleotide mutations in each BXD that were absent from all other RILs, as well as from the C57BL/6J and DBA/2J parents. We required each private variant to be meet the following criteria:

- genotyped as either homozygous or heterozygous for the alternate allele, with at least 90% of sequencing reads supporting the alternate allele

- supported by at least 10 sequencing reads

- Phred-scaled genotype quality of at least 20

- must not overlap regions of the genome annotated as segmental duplications or simple repeats in GRCm38/mm10

- must occur on a parental haplotype that was inherited by at least one other BXD at the same locus; these other BXDs must be homozygous for the reference allele at the variant site

A new approach to discover germline mutator alleles

Calculating inter-haplotype distance

Using the existing catalog of *de novo* germline mutations in the BXDs, we developed a new approach to discover loci that affect the germline *de novo* mutation spectrum in biparental RILs (Figure 3).

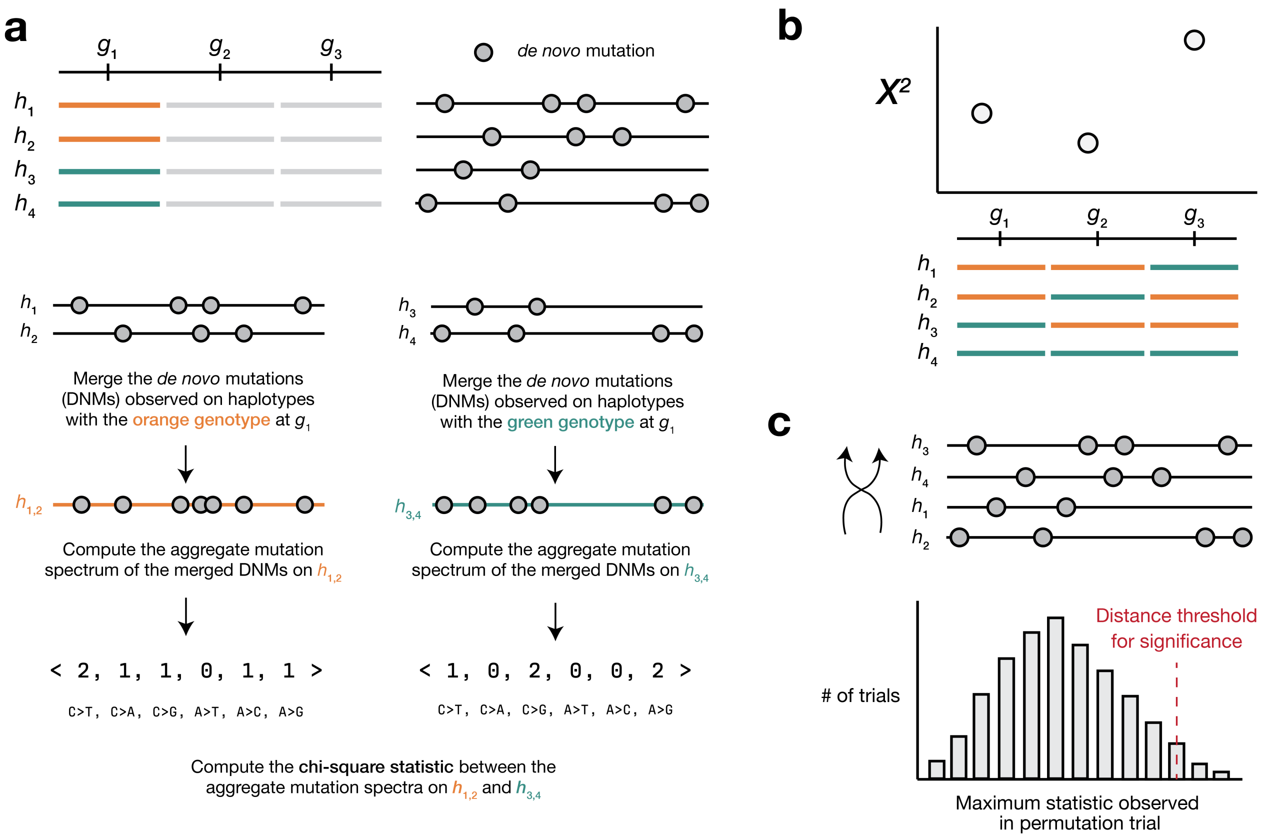
a
g_1
g_2
g_3
h_1
h_2
h_3
h_4
de novo mutation
h_1
h_2
h_3
h_4
Merge the de novo mutations (DNMs) observed on haplotypes with the orange genotype at g_1
Merge the de novo mutations (DNMs) observed on haplotypes with the green genotype at g_1
$h_{1,2}$
$h_{3,4}$
Compute the aggregate mutation spectrum of the merged DNMs on $h_{1,2}$
Compute the aggregate mutation spectrum of the merged DNMs on $h_{3,4}$
< 2, 1, 1, 0, 1, 1 >
C>T, C>A, C>G, A>T, A>C, A>G
< 1, 0, 2, 0, 0, 2 >
C>T, C>A, C>G, A>T, A>C, A>G
Compute the chi-square statistic between the aggregate mutation spectra on $h_{1,2}$ and $h_{3,4}$
b
X^2
g_1
g_2
g_3
h_1
h_2
h_3
h_4
c
h_3
h_4
h_1
h_2
Distance threshold for significance
# of trials
Maximum statistic observed in permutation trial

Figure 3: Overview of inter-haplotype distance method for discovering mutator alleles. a) A population of four haplotypes has been genotyped at three informative markers; each haplotype also harbors private *de novo* germline mutations. At each informative marker, we compute an aggregate *de novo* germline mutation spectrum in the haplotypes that carry either parental allele, and calculate the χ^2 statistic between the two aggregate spectra. **b)** We repeat the process outlined in a) for every informative marker along the genome. **c)** To assess the significance of any χ^2 statistic peaks in b), we perform a permutation test by shuffling the labels associated with each haplotype's mutation data and running a genome-wide distance scan. In each of N permutations, we record the maximum distance encountered at any locus in the distance scan. Finally, we calculate the $1 - p$ percentile of that maximum distance distribution to obtain a genome-wide χ^2 statistic threshold at the specified value of p.

We assume that a collection of haplotypes has been genotyped at informative markers, and that *de novo* germline mutations have been identified on each haplotype.

At each informative marker, we divide haplotypes into two groups based on the parental allele that they inherited. We then compute a k-mer mutation spectrum using the aggregate mutation counts in each haplotype group. The k-mer mutation spectrum contains the frequency of every possible k-mer mutation type in a collection of mutations, and can be represented as a vector of size $6 \times 4^{k-1}$ after collapsing by strand complement. For example, the 1-mer mutation spectrum is 6-element vector that contains the frequencies of C>T, C>G, C>A, A>G, A>T, and A>C mutations.

At each marker, we then calculate the χ^2 statistic between the aggregate mutation spectra of haplotypes with either parental allele. Larger values of the χ^2 statistic suggests more dissimilarity between the two aggregate mutation spectra.

Inspired by methods from QTL mapping [36], we use permutation tests to establish genome-wide χ^2 statistic thresholds. In each of N permutation trials, we randomly shuffle the per-haplotype mutation data such that haplotype labels no longer correspond to the correct mutation counts. Using the shuffled mutation data, we perform a genome-wide distance scan as described above, and record the maximum distance observed at any locus. After N permutations (usually 10,000), we compute the

$1-p$ percentile of the maximum distance distribution, and use that percentile value as a genome-wide significance threshold (for example, at $p = 0.05$).

Accounting for relatedness between strains

We expect each BXD RIL to derive approximately 50% of its genome from C57BL/6J and 50% from DBA/2J. As a result, every pair of RILs will likely be identical-by-descent (IBD) at a fraction of genotyped markers. Pairs of more genetically similar BXDs may also have more similar mutation spectra, potentially due to shared polygenic effects on the mutation process. Therefore, at a given marker, if the BXD RILs that inherited *D* haplotypes are genetically dissimilar from the RILs that inherited *B* haplotypes (considering all loci throughout the genome), we might expect the aggregate mutation spectra in the two groups to also be dissimilar.

We therefore implemented a simple approach to account for these potential issues of relatedness. At each marker i, we divide BXD haplotypes into two groups based on the parental allele they inherited. As before, we first compute the aggregate mutation spectrum in each group of haplotypes and calculate the χ^2 statistic between the two aggregate spectra (χ_i^2). Then, within each group of haplotypes, we calculate the allele frequency of the *D* allele at every marker along the genome to obtain a vector of length n, where n is the number of genotyped markers. To quantify the genetic similarity between the two groups of haplotypes, we calculate the Pearson correlation coefficient r_i between the two vectors of marker-wide *D* allele frequencies.

Thus, at every marker i along the genome, we divide BXD haplotypes into two groups and compute two metrics: χ_i^2 (the χ^2 statistic between the two groups' aggregate spectra) and r_i (the correlation between genome-wide *D* allele frequencies in the two groups). To control for the potential effects of genetic similarity on χ^2 statistics, we regress $\left(\chi_1^2, \chi_2^2, \ldots \chi_n^2\right)$ on $(r_1, r_2, \ldots r_n)$ for all n markers and fit an ordinary least-squares regression model. We then use the residuals from the fitted model as the "adjusted" χ^2 statistic values for each marker. If genome-wide genetic similarity between haplotypes perfectly predicts χ^2 statistics at each marker, these residuals will all be 0 (or very close to 0). If genome-wide genetic similarity has no predictive power, the residuals will simply represent the difference between an observed χ^2 statistic and the marker-wide mean of χ^2 statistics.

Implementation and source code

The inter-haplotype distance method was implemented in Python, and relies heavily on the following Python libraries: `numpy`, `pandas`, `matplotlib`, `scikit-learn`, `pandera`, `seaborn`, and `numba` [37,38,39,40,41,42,43].

Additional documentation is available on GitHub [44], along with a reproducible Snakemake [45] workflow for running the method from start to finish using the BXDs (including downloading the mutation data, downloading genotypes, and running a genome-wide distance scan).

Simulations to assess the power of the inter-haplotype distance approach

We performed a series of simple simulations to estimate our power to detect alleles that affect the germline mutation spectrum in biparental RILs using the inter-haplotype distance method.

First, we simulate the k-mer mutation spectrum in a population of h haplotypes. We assume that exactly $\frac{h}{2}$ of the haplotypes are under the effects of a mutator allele that increases the mutation rate of a particular mutation type(s) by an effect size e. We simulate m mutations on each haplotype as follows:

We first define a vector of 1-mer mutation probabilities:

$$P = (0.4,\ 0.1,\ 0.075,\ 0.075,\ 0.075,\ 0.275)$$

These probabilities sum to 1 and correspond to the expected frequencies of C>T, C>A, C>G, A>T, A>C, and A>G *de novo* germline mutations in mice, respectively [8].

If we are simulating the 3-mer mutation spectrum, we modify the vector of mutation probabilities P to be length 96, and assign every 3-mer mutation type a value of $\frac{P_c}{16}$, where P_c is the probability of the "central" mutation type associated with the 3-mer mutation type. In other words, each of the 16 possible NCN>NTN 3-mer mutation types would be assigned a mutation probability of $\frac{P_c}{16} = \frac{0.4}{16} = 0.025$.

To simulate the mutation spectrum on the *wild-type* haplotypes, we define a matrix C of size $(\frac{h}{2}, n)$, where $n = 6 \times 4^{k-1}$ (i.e., the number of k-mer mutation types being simulated). First, we generate a vector of lambda values by scaling the mutation probabilities by the number of mutations we wish to simulate:

$$\lambda = Pm$$

Then, we populate the matrix C by taking a single Poisson draw from the vector of λ values for each mutation type on each haplotype. Thus, for every row i in the matrix (i.e., for every haplotype), we perform the following for mutation type j:

$$C_{i,j} = \text{Pois}(\lambda_j)$$

To simulate the mutation spectrum on the $\frac{h}{2}$ *mutator* haplotypes, we define a new matrix C' of size $(\frac{h}{2}, n)$ as defined above. We then multiply the lambda value of a particular mutation type (or multiple mutation types) by the mutator effect size e to obtain λ'. Then, for every row i in the matrix:

$$C'_{i,j} = \text{Pois}(\lambda'_j)$$

When $k = 1$, we only augment the effect size of one mutation type at a time, but when $k = 3$, we augment a fraction (25%, 50%, or 100%) of the 3-mer mutation types associated with a single "central" mutation type.

After generating mutator and wild-type haplotypes, we compute the aggregate mutation spectrum in either group by summing the columns of C and C'. We then calculate the χ^2 statistic between the two aggregate spectra, which we call the "focal" distance D_f. To determine whether D_f is greater than what we'd expect by chance, we perform a permutation test.

First, we concatenate the matrices of wild-type and mutator haplotype spectra:

$$A = \begin{bmatrix} C \\ C' \end{bmatrix}$$

Then, in each of $N = 1,000$ trials, we randomly permute the rows of A. In every permutation, we consider the row indices from $\left[0, \frac{h}{2}\right)$ to correspond to the wild-type haplotypes, and the row indices from $\left[\frac{h}{2}, h\right)$ to correspond to the mutator haplotypes. We then compute the χ^2 statistic between the aggregate spectra of the wild-type and mutator haplotypes. If fewer than 5% of the N permutations produces a χ^2 statistic greater than or equal to D_f, we say that the approach successfully identified the mutator allele. For every combination of simulation parameters (h, m, e, and so on) we perform 100 trials and record the number of trials in which we successfully identify the mutator allele.

Applying the inter-haplotype distance method to the BXDs

We downloaded previously-generated BXD *de novo* germline mutation data from the GitHub repository associated with our previous manuscript, which was also archived at Zenodo [2,46,47], and downloaded a CSV file of BXD genotypes at 7,320 informative markers from GeneNetwork [15,48]. We also downloaded relevant metadata about each BXD RIL from the manuscript describing the updated BXD resource [3].

As in our previous manuscript [2], we included mutation data from a subset of the 152 BXDs in our inter-haplotype distance scans. We removed any BXDs that had been inbred for fewer than 20 generations, as it takes approximately 20 generations of strict brother-sister mating for an RIL genome to become >98% homozygous [49]. As a result, any potential mutator allele would almost certainly be either fixed or lost after 20 generations. If fixed, the allele would remain linked to any excess mutations it causes for the duration of subsequent inbreeding, and its effects would be detectable using our methods. Additionally, a strain only meets the canonical definition of “inbred” if it has been subject to brother-sister mating for at least 20 generations [50]. is one We also removed the BXD68 RIL from our genome-wide scans, since we previously discovered a hyper-mutator phenotype in that strain; the C>A germline mutation rate in BXD68 is over 5 times the population mean, likely due to a private deleterious nonsynonymous mutation in *Mutyh* [2]. In total, we included 94 BXD RILs in our genome-wide scans.

We used Snakemake [51] to write a reproducible workflow for running the inter-haplotype distance method on the BXD dataset, which has been deposited in the GitHub repository associated with this manuscript [44].

Identifying candidate mutator alleles overlapping the chromosome 6 peak

We investigated the region implicated by our inter-haplotype distance approach on chromosome 6 by subsetting the joint-genotyped BXD VCF file (European Nucleotide Archive accession **PRJEB45429** [52]) using `bcftools` [53]. We defined the candidate interval surrounding the χ^2 statistic peak on chromosome 6 as +/- 5 Mbp from the genotype marker with the largest χ^2 statistic value (`rs31001331`). To predict the functional impacts of both single-nucleotide variants and indels on splicing, protein structure, etc., we annotated variants in the BXD VCF using the following `snpEff` [54] command:

```
 java -Xmx16g -jar /path/to/snpeff/jarfile GRCm38.75 /path/to/bxd/vcf >
/path/to/uncompressed/output/vcf
```

and used `cyvcf2` [55] to iterate over the annotated VCF file in order to identify nonsynonymous fixed differences between the parental C57BL/6J and DBA/2J strains.

Comparing mutation spectra between Mouse Genomes Project strains

We downloaded mutation data from a previously published analysis [11] (Supplementary File 1, Excel Table S3) that identified strain-private mutations in 29 strains that were originally whole-genome sequenced as part of the Sanger Mouse Genomes (MGP) project [56]. When comparing counts of each mutation type between MGP strains that harbored either *D* or *B* alleles at the chromosome 4 or chromosome 6 mutator loci, we adjusted mutation counts by the number of callable A, T, C, or G nucleotides in each strain as described previously [2].

Querying GeneNetwork for evidence of eQTLs at the mutator locus

We used the online GeneNetwork resource [15], which contains array- and RNA-seq-derived expression measurements in a wide variety of tissues from numerous datasets, to find *cis*-eQTLs for the DNA repair genes we implicated under the χ^2 statistic peak on chromosome 6. On the GeneNetwork homepage (genenetwork.org), we selected the "BXD Family" **Group** and used the **Type** dropdown menu to select each of the specific expression datasets described in Table 2. In the **Get Any** text box, we then entered the gene name (*Ogg1*) and clicked **Search**. After selecting the appropriate trait ID on the next page, we used the **Mapping Tools** dropdown to run Haley-Knott regression [57] with the following parameters: WGS-based marker genotypes, 1,000 permutations for LOD threshold calculations, and controlling for BXD genotypes at the `rs32497085` marker.

The exact names of the expression datasets we used for each tissue are shown in Table 2 below:

Table 2: Names of gene expression datasets used for each tissue type on GeneNetwork

Tissue name	Complete name of GeneNetwork expression data	GeneNetwork trait ID
Kidney	`Mouse kidney M430v2 Sex Balanced (Aug06) RMA`	`1448815_at`
Gastrointestinal	`UTHSC Mouse BXD Gastrointestinal Affy MoGene 1.0 ST Gene Level (Apr14) RMA`	`10540639`
Hematopoetic stem cells	`UMCG Stem Cells ILM6v1.1 (Apr09) transformed`	`ILM1940279`
Hematopoetic progenitor cells	`UMCG Progenitor Cells ILM6v1.1 (Apr09) transformed`	`ILM1940279`
Spleen	`UTHSC Affy MoGene 1.0 ST Spleen (Dec10) RMA`	`10540639`
Liver	`UTHSC BXD Liver RNA-Seq Avg (Oct19) TPM Log2`	`ENSMUST00000032406`
Heart	`NHLBI BXD All Ages Heart RNA-Seq (Nov20) TMP Log2 **`	`ENSMUSG00000030271`
Eye	`UTHSC BXD All Ages Eye RNA-Seq (Nov20) TPM Log2 **`	`ENSMUSG00000030271`

Calculating the frequencies of candidate mutator alleles in wild mice

To determine the frequency of the *Ogg1* p.Thr95Ala mutation in other populations of mice, we queried a VCF file containing genome-wide variation in 67 wild-derived mice from four species of *Mus* [17]. We calculated the allele frequency of each nonsynonymous mutation in each of the four species or subspecies (*Mus musculus domesticus*, *Mus musculus musculus*, *Mus musculus castaneus*, and *Mus spretus*), including genotypes that met the following criteria:

- supported by at least 10 sequencing reads
- Phred-scaled genotype quality of at least 20

Testing for epistasis between the two mutator loci

To test for the presence of epistasis between the mutator loci near *Mutyh* and *Ogg1*, we modeled C>A mutation rates in the BXDs as a function of genotypes at either locus. Specifically, we tested for statistical interaction between *Mutyh* and *Ogg1* genotypes by fitting a generalized linear model in the R statistical language as follows:

```
m1 <- glm(Count ~ offset(log(ADJ_AGE)) + Genotype_A * Genotype_B, data =
        data, family=poisson())
```

In this model, `Count` is the count of C>A *de novo* mutations observed in each BXD RIL. `ADJ_AGE` is the product of the number of “callable” cytosine nucleotides in each RIL (i.e., the total number of cytosines covered by at least 10 sequencing reads in the RIL) and the number of generations for which the RIL was inbred. We included the logarithm of `ADJ_AGE` as an “offset” in order to model the response variable as a rate rather than an absolute count; the BXDs differ in both their durations of inbreeding and the proportions of their genomes that were sequenced to sufficient depth, which influences the number of mutations we observe in each RIL. The `Genotype_A` and `Genotype_B` terms represent the genotypes of BXDs at markers `rs52263933` and `rs31001331` (the markers with peak χ^2 statistic near *Mutyh* and *Ogg1* in the two inter-haplotype distance scans). Since each BXD is inbred for at least 20 generations, we considered genotypes at either locus to be binary (“B” or “D”). Using analysis of variance (ANOVA), we then compared the model including an interaction effect to a model including only additive effects:

```
m2 <- glm(Count ~ offset(log(ADJ_AGE)) + Haplotype_A + Haplotype_B, data =
        data, family=poisson())
```

```
anova(m1, m2, test="Chisq")
```

We tested for epistasis in the Sanger Mouse Genomes Project (MGP) strains using a nearly-identical approach. In this analysis, we fit two models as follows:

```
m1 <- glm(Count ~ offset(log(CALLABLE_C)) + Genotype_A * Genotype_B, data =
        data, family=poisson())

m2 <- glm(Count ~ offset(log(CALLABLE_C)) + Genotype_A + Genotype_B, data =
        data, family=poisson())
```

where `Count` is the count of strain-private C>A mutations observed in each MGP strain [11]. The `CALLABLE_C` term represents the total number of cytosine and guanine nucleotides that were accessible for mutation calling in each strain, and the `Genotype_A` and `Genotype_B` terms represent MGP genotypes at `rs52263933` and `rs31001331`. We compared the two models using ANOVA as described above.

Since each BXD derives approximately 50% of its genome from C57BL/6J and 50% from DBA/2J, we performed an additional test for epistasis that accounted for kinship between the BXD RILs. Using the `lmekin` method from the `coxme` package [58] in the R statistical language, we fit a mixed effects model predicting C>A mutation fractions as a function of genotypes at both `rs52263933` and `rs31001331`, and included a pairwise kinship matrix as a random effect.

```
m = lmekin(Fraction ~ Genotype_A * Genotype_B + (1|sample), data = data,
        varlist = kinship_matrix)
```

The rows and columns of the kinship matrix were labeled with the `sample` name of each BXD, such that the `(1|sample)` term in the model captured the random effect of kinship. We calculated the `kinship_matrix` using the `calc_kinship` method from `R/qtl2` [59] as follows:

```
# read in the JSON-formatted file that directs R/qtl2 to sample
# genotypes, phenotypes, and covariates if applicable
bxd <- read_cross2("path/to/bxd.json")

# subset cross2 object to BXDs with C>A fractions in `data`
bxd <- bxd[data$sample, ]

# insert pseudomarkers into the genotype map
gmap <- insert_pseudomarkers(bxd$gmap, step = 0.2, stepwidth = "max")

# calculate QTL genotype probabilities
pr <- calc_genoprob(bxd, gmap, error_prob = 0.002, map_function = "c-f")

# calculate kinship between strains using all chromosomes
k <- calc_kinship(pr, "overall")

kinship_matrix = as.matrix(k)
```

Supplementary information

Using cosine distance instead of χ^2 statistics for comparing mutation spectra

We also explored the use of cosine distance as an alternative to the χ^2 statistic for comparing mutation spectra. The cosine distance between two vectors $\mathbf{A}$ and $\mathbf{B}$ is defined as

$$D_C = 1 - \frac{\mathbf{A} \cdot \mathbf{B}}{||\mathbf{A}||\ ||\mathbf{B}||}$$

where $||\mathbf{A}||$ and $||\mathbf{B}||$ are the L^2 (or Euclidean) norms of $\mathbf{A}$ and $\mathbf{B}$, respectively. The cosine distance metric has a number of favorable properties for comparing mutation spectra. Since cosine distance does not take the magnitude of vectors into account, it can be used to compare two spectra with unequal total mutation counts (even if those total counts are relatively small). Additionally, by calculating the cosine distance between mutation *spectra*, we avoid the need to perform separate comparisons of mutation counts at each individual k-mer mutation type.

In practice, we found that the χ^2 statistic was more sensitive for detecting loci associated with differences in mutation spectra. However, we provide the ability to use cosine distance in our method, as well, since the χ^2 statistic may not behave as expected in certain situations (e.g., if the counts of mutations in each k-mer type are small).

Using 3-mer mutation spectra to perform the inter-haplotype distance scans

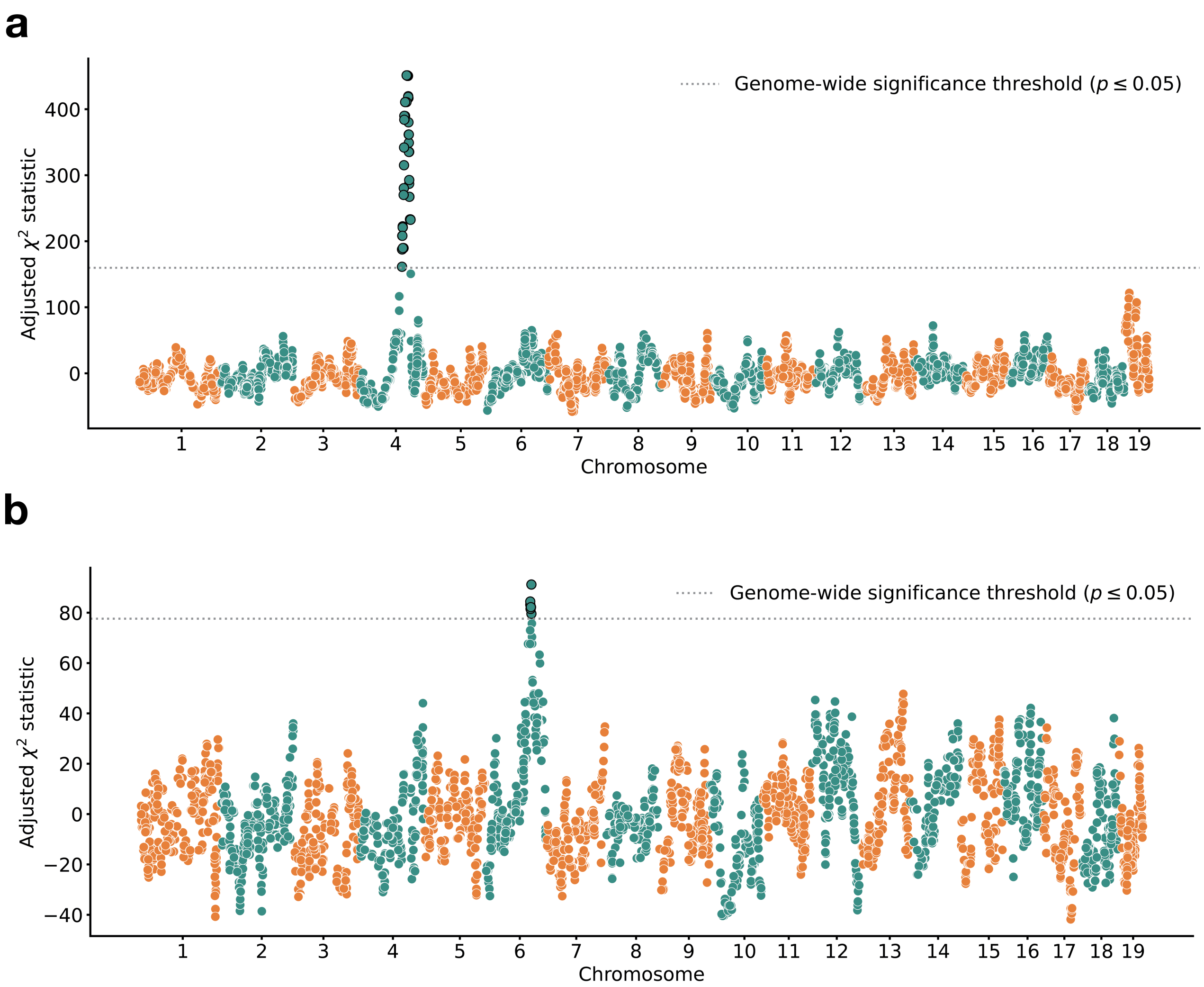

Figure 4: Results of inter-haplotype distance scans in the BXD RILs using 3-mer mutation spectra. a) χ^2 statistics between aggregate 3-mer *de novo* mutation spectra on BXD haplotypes (n = 93 haplotypes; 62,993 total mutations) with either *D* or *B* alleles at 7,320 informative markers. Distance threshold at $p = 0.05$ was calculated by performing 10,000 permutations of the BXD haplotype mutation data, and is shown as a dotted grey line. **b)** χ^2 statistics between aggregate $3 - mer$ *de novo* mutation spectra on BXD haplotypes with *D* alleles at `rs52263933` (n = 55 haplotypes;

40,913 total mutations) and either *D* or *B* alleles at 7,320 informative markers. Distance threshold at $p = 0.05$ was calculated by performing 10,000 permutations of the BXD haplotype mutation data, and is shown as a dotted grey line.

References

1. **Mechanisms of DNA damage, repair, and mutagenesis.**
Nimrat Chatterjee, Graham C Walker
Environmental and molecular mutagenesis (2017-05-09)
https://www.ncbi.nlm.nih.gov/pubmed/28485537
DOI: 10.1002/em.22087 · PMID: 28485537 · PMCID: PMC5474181

2. **A natural mutator allele shapes mutation spectrum variation in mice.**
Thomas A Sasani, David G Ashbrook, Annabel C Beichman, Lu Lu, Abraham A Palmer, Robert W Williams, Jonathan K Pritchard, Kelley Harris
Nature (2022-05-11) https://www.ncbi.nlm.nih.gov/pubmed/35545679
DOI: 10.1038/s41586-022-04701-5 · PMID: 35545679 · PMCID: PMC9272728

3. **A platform for experimental precision medicine: The extended BXD mouse family.**
David G Ashbrook, Danny Arends, Pjotr Prins, Megan K Mulligan, Suheeta Roy, Evan G Williams, Cathleen M Lutz, Alicia Valenzuela, Casey J Bohl, Jesse F Ingels, … Robert W Williams
Cell systems (2021-01-19) https://www.ncbi.nlm.nih.gov/pubmed/33472028
DOI: 10.1016/j.cels.2020.12.002 · PMID: 33472028 · PMCID: PMC7979527

4. **Base-excision repair of oxidative DNA damage.**
Sheila S David, Valerie L O'Shea, Sucharita Kundu
Nature (2007-06-21) https://www.ncbi.nlm.nih.gov/pubmed/17581577
DOI: 10.1038/nature05978 · PMID: 17581577 · PMCID: PMC2896554

5. **Inferring evolutionary dynamics of mutation rates through the lens of mutation spectrum variation.**
Jedidiah Carlson, William S DeWitt, Kelley Harris
Current opinion in genetics & development (2020-06-30)
https://www.ncbi.nlm.nih.gov/pubmed/32619789
DOI: 10.1016/j.gde.2020.05.024 · PMID: 32619789 · PMCID: PMC7646088

6. **Parental influence on human germline de novo mutations in 1,548 trios from Iceland.**
Hákon Jónsson, Patrick Sulem, Birte Kehr, Snaedis Kristmundsdottir, Florian Zink, Eirikur Hjartarson, Marteinn T Hardarson, Kristjan E Hjorleifsson, Hannes P Eggertsson, Sigurjon Axel Gudjonsson, … Kari Stefansson
Nature (2017-09-20) https://www.ncbi.nlm.nih.gov/pubmed/28959963
DOI: 10.1038/nature24018 · PMID: 28959963

7. **Large, three-generation human families reveal post-zygotic mosaicism and variability in germline mutation accumulation.**
Thomas A Sasani, Brent S Pedersen, Ziyue Gao, Lisa Baird, Molly Przeworski, Lynn B Jorde, Aaron R Quinlan
eLife (2019-09-24) https://www.ncbi.nlm.nih.gov/pubmed/31549960
DOI: 10.7554/elife.46922 · PMID: 31549960 · PMCID: PMC6759356

8. **Similarities and differences in patterns of germline mutation between mice and humans.**
Sarah J Lindsay, Raheleh Rahbari, Joanna Kaplanis, Thomas Keane, Matthew E Hurles
Nature communications (2019-09-06) https://www.ncbi.nlm.nih.gov/pubmed/31492841
DOI: 10.1038/s41467-019-12023-w · PMID: 31492841 · PMCID: PMC6731245

9. **Genetic drift, selection and the evolution of the mutation rate.**
Michael Lynch, Matthew S Ackerman, Jean-Francois Gout, Hongan Long, Way Sung, WKelley Thomas, Patricia L Foster

Nature reviews. Genetics (2016-10-14) https://www.ncbi.nlm.nih.gov/pubmed/27739533
DOI: 10.1038/nrg.2016.104 · PMID: 27739533

10. **Inference of Candidate Germline Mutator Loci in Humans from Genome-Wide Haplotype Data.**
Cathal Seoighe, Aylwyn Scally
PLoS genetics (2017-01-17) https://www.ncbi.nlm.nih.gov/pubmed/28095480
DOI: 10.1371/journal.pgen.1006549 · PMID: 28095480 · PMCID: PMC5283766

11. **Significant Strain Variation in the Mutation Spectra of Inbred Laboratory Mice.**
Beth L Dumont
Molecular biology and evolution (2019-05-01) https://www.ncbi.nlm.nih.gov/pubmed/30753674
DOI: 10.1093/molbev/msz026 · PMID: 30753674 · PMCID: PMC6501876

12. **Spontaneous Mutation Accumulation Studies in Evolutionary Genetics**
Daniel L Halligan, Peter D Keightley
Annual Review of Ecology, Evolution, and Systematics (2009-12-01) https://doi.org/dvrjz8
DOI: 10.1146/annurev.ecolsys.39.110707.173437

13. **A Specific Mutational Signature Associated with DNA 8-Oxoguanine Persistence in MUTYH-defective Colorectal Cancer.**
Alessandra Viel, Alessandro Bruselles, Ettore Meccia, Mara Fornasarig, Michele Quaia, Vincenzo Canzonieri, Eleonora Policicchio, Emanuele Damiano Urso, Marco Agostini, Maurizio Genuardi, … Margherita Bignami
EBioMedicine (2017-04-13) https://www.ncbi.nlm.nih.gov/pubmed/28551381
DOI: 10.1016/j.ebiom.2017.04.022 · PMID: 28551381 · PMCID: PMC5478212

14. **Mutational signature analysis identifies MUTYH deficiency in colorectal cancers and adrenocortical carcinomas.**
Camilla Pilati, Jayendra Shinde, Ludmil B Alexandrov, Guillaume Assié, Thierry André, Zofia Hélias-Rodzewicz, Romain Ducoudray, Delphine Le Corre, Jessica Zucman-Rossi, Jean-François Emile, … Pierre Laurent-Puig
The Journal of pathology (2017-03-29) https://www.ncbi.nlm.nih.gov/pubmed/28127763
DOI: 10.1002/path.4880 · PMID: 28127763

15. **GeneNetwork: A Toolbox for Systems Genetics.**
Megan K Mulligan, Khyobeni Mozhui, Pjotr Prins, Robert W Williams
Methods in molecular biology (Clifton, N.J.) (2017)
https://www.ncbi.nlm.nih.gov/pubmed/27933521
DOI: 10.1007/978-1-4939-6427-7_4 · PMID: 27933521 · PMCID: PMC7495243

16. **Mapping the Effects of Genetic Variation on Chromatin State and Gene Expression Reveals Loci That Control Ground State Pluripotency.**
Daniel A Skelly, Anne Czechanski, Candice Byers, Selcan Aydin, Catrina Spruce, Chris Olivier, Kwangbom Choi, Daniel M Gatti, Narayanan Raghupathy, Gregory R Keele, … Laura G Reinholdt
Cell stem cell (2020-08-13) https://www.ncbi.nlm.nih.gov/pubmed/32795400
DOI: 10.1016/j.stem.2020.07.005 · PMID: 32795400 · PMCID: PMC7484384

17. **Genomic resources for wild populations of the house mouse, Mus musculus and its close relative Mus spretus.**
Bettina Harr, Emre Karakoc, Rafik Neme, Meike Teschke, Christine Pfeifle, Željka Pezer, Hiba Babiker, Miriam Linnenbrink, Inka Montero, Rick Scavetta, … Diethard Tautz
Scientific data (2016-09-13) https://www.ncbi.nlm.nih.gov/pubmed/27622383
DOI: 10.1038/sdata.2016.75 · PMID: 27622383 · PMCID: PMC5020872

18. **On the subspecific origin of the laboratory mouse.**
Hyuna Yang, Timothy A Bell, Gary A Churchill, Fernando Pardo-Manuel de Villena
Nature genetics (2007-07-29) https://www.ncbi.nlm.nih.gov/pubmed/17660819
DOI: 10.1038/ng2087 · PMID: 17660819

19. **Novel mutations of OGG1 base excision repair pathway gene in laryngeal cancer patients.**
Ishrat Mahjabeen, Nosheen Masood, Ruqia Mehmood Baig, Maimoona Sabir, Uzma Inayat, Faraz Arshad Malik, Mahmood Akhtar Kayani
Familial cancer (2012-12) https://www.ncbi.nlm.nih.gov/pubmed/22829015
DOI: 10.1007/s10689-012-9554-2 · PMID: 22829015

20. **Alterations of the DNA repair gene OGG1 in human clear cell carcinomas of the kidney.**
M Audebert, S Chevillard, C Levalois, G Gyapay, A Vieillefond, J Klijanienko, P Vielh, AK El Naggar, S Oudard, S Boiteux, JP Radicella
Cancer research (2000-09-01) https://www.ncbi.nlm.nih.gov/pubmed/10987279
PMID: 10987279

21. **Mutations in OGG1, a gene involved in the repair of oxidative DNA damage, are found in human lung and kidney tumours.**
S Chevillard, JP Radicella, C Levalois, J Lebeau, MF Poupon, S Oudard, B Dutrillaux, S Boiteux
Oncogene (1998-06-11) https://www.ncbi.nlm.nih.gov/pubmed/9662341
DOI: 10.1038/sj.onc.1202096 · PMID: 9662341

22. **Defects in 8-oxo-guanine repair pathway cause high frequency of C &gt; A substitutions in neuroblastoma**
Marlinde L van den Boogaard, Rurika Oka, Anne Hakkert, Linda Schild, Marli E Ebus, Michael R van Gerven, Danny A Zwijnenburg, Piet Molenaar, Lieke L Hoyng, MEmmy M Dolman, … Jan J Molenaar
Proceedings of the National Academy of Sciences (2021-09-03) https://doi.org/grtcs9
DOI: 10.1073/pnas.2007898118 · PMID: 34479993 · PMCID: PMC8433536

23. **Resolution of structural variation in diverse mouse genomes reveals chromatin remodeling due to transposable elements**
Ardian Ferraj, Peter A Audano, Parithi Balachandran, Anne Czechanski, Jacob I Flores, Alexander A Radecki, Varun Mosur, David S Gordon, Isha A Walawalkar, Evan E Eichler, … Christine R Beck
Cold Spring Harbor Laboratory (2022-09-27) https://doi.org/grtctb
DOI: 10.1101/2022.09.26.509577

24. **Not breathing is not an option: How to deal with oxidative DNA damage.**
Enni Markkanen
DNA repair (2017-09-22) https://www.ncbi.nlm.nih.gov/pubmed/28963982
DOI: 10.1016/j.dnarep.2017.09.007 · PMID: 28963982

25. **A modified fluctuation assay reveals a natural mutator phenotype that drives mutation spectrum variation within**
Pengyao Jiang, Anja R Ollodart, Vidha Sudhesh, Alan J Herr, Maitreya J Dunham, Kelley Harris
eLife (2021-09-15) https://www.ncbi.nlm.nih.gov/pubmed/34523420
DOI: 10.7554/elife.68285 · PMID: 34523420 · PMCID: PMC8497059

26. **Limited role of generation time changes in driving the evolution of mutation spectrum in humans**
Ziyue Gao, Yulin Zhang, Nathan Cramer, Molly Przeworski, Priya Moorjani
bioRxiv (2023-01-13) https://doi.org/grr525
DOI: https://doi.org/10.1101/2022.06.17.496622

27. **Mutational Signatures of Replication Timing and Epigenetic Modification Persist through the Global Divergence of Mutation Spectra across the Great Ape Phylogeny.**
Michael E Goldberg, Kelley Harris
Genome biology and evolution (2022-01-04) https://www.ncbi.nlm.nih.gov/pubmed/33983415
DOI: 10.1093/gbe/evab104 · PMID: 33983415 · PMCID: PMC8743035

28. **Heritability of de novo germline mutation reveals a contribution from paternal but not maternal genetic factors**
Seongwon Hwang, Matthew DC Neville, Genomics England Research Consortium, Felix R Day, Aylwyn Scally
bioRxiv (2022-12-17) https://doi.org/grr526
DOI: https://doi.org/10.1101/2022.12.17.520885

29. **The impact of genetic modifiers on variation in germline mutation rates within and among human populations.**
William R Milligan, Guy Amster, Guy Sella
Genetics (2022-07-30) https://www.ncbi.nlm.nih.gov/pubmed/35666194
DOI: 10.1093/genetics/iyac087 · PMID: 35666194 · PMCID: PMC9339295

30. **Rapid evolution of the human mutation spectrum.**
Kelley Harris, Jonathan K Pritchard
eLife (2017-04-25) https://www.ncbi.nlm.nih.gov/pubmed/28440220
DOI: 10.7554/elife.24284 · PMID: 28440220 · PMCID: PMC5435464

31. **Nonparametric coalescent inference of mutation spectrum history and demography.**
William S DeWitt, Kameron Decker Harris, Aaron P Ragsdale, Kelley Harris
Proceedings of the National Academy of Sciences of the United States of America (2021-05-25) https://www.ncbi.nlm.nih.gov/pubmed/34016747
DOI: 10.1073/pnas.2013798118 · PMID: 34016747 · PMCID: PMC8166128

32. **Signatures of mutational processes in human cancer.**
Ludmil B Alexandrov, Serena Nik-Zainal, David C Wedge, Samuel AJR Aparicio, Sam Behjati, Andrew V Biankin, Graham R Bignell, Niccolò Bolli, Ake Borg, Anne-Lise Børresen-Dale, … Michael R Stratton
Nature (2013-08-14) https://www.ncbi.nlm.nih.gov/pubmed/23945592
DOI: 10.1038/nature12477 · PMID: 23945592 · PMCID: PMC3776390

33. **Uncovering novel mutational signatures by**
SMAshiqul Islam, Marcos Díaz-Gay, Yang Wu, Mark Barnes, Raviteja Vangara, Erik N Bergstrom, Yudou He, Mike Vella, Jingwei Wang, Jon W Teague, … Ludmil B Alexandrov
Cell genomics (2022-11-09) https://www.ncbi.nlm.nih.gov/pubmed/36388765
DOI: 10.1016/j.xgen.2022.100179 · PMID: 36388765 · PMCID: PMC9646490

34. **The mutational footprints of cancer therapies.**
Oriol Pich, Ferran Muiños, Martijn Paul Lolkema, Neeltje Steeghs, Abel Gonzalez-Perez, Nuria Lopez-Bigas
Nature genetics (2019-11-18) https://www.ncbi.nlm.nih.gov/pubmed/31740835
DOI: 10.1038/s41588-019-0525-5 · PMID: 31740835 · PMCID: PMC6887544

35. **Mutational signatures associated with tobacco smoking in human cancer.**
Ludmil B Alexandrov, Young Seok Ju, Kerstin Haase, Peter Van Loo, Iñigo Martincorena, Serena Nik-Zainal, Yasushi Totoki, Akihiro Fujimoto, Hidewaki Nakagawa, Tatsuhiro Shibata, … Michael R Stratton
Science (New York, N.Y.) (2016-11-04) https://www.ncbi.nlm.nih.gov/pubmed/27811275
DOI: 10.1126/science.aag0299 · PMID: 27811275 · PMCID: PMC6141049

36. **Empirical threshold values for quantitative trait mapping.**
GA Churchill, RW Doerge
Genetics (1994-11) https://www.ncbi.nlm.nih.gov/pubmed/7851788
DOI: 10.1093/genetics/138.3.963 · PMID: 7851788 · PMCID: PMC1206241

37. **Array programming with NumPy**
Charles R Harris, KJarrod Millman, Stéfan J van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J Smith, … Travis E Oliphant
Nature (2020-09-16) https://doi.org/ghbzf2
DOI: 10.1038/s41586-020-2649-2 · PMID: 32939066 · PMCID: PMC7759461

38. **pandas-dev/pandas: Pandas**
The Pandas Development Team
Zenodo (2023-02-20) https://doi.org/ggt8bh
DOI: 10.5281/zenodo.3509134

39. **Matplotlib: A 2D Graphics Environment**
John D Hunter
Computing in Science &amp; Engineering (2007) https://doi.org/drbjhg
DOI: 10.1109/mcse.2007.55

40. **Scikit-learn: Machine Learning in Python**
Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, … Édouard Duchesnay
Journal of Machine Learning Research (2011) http://jmlr.org/papers/v12/pedregosa11a.html

41. **pandera: Statistical Data Validation of Pandas Dataframes**
Niels Bantilan
Proceedings of the Python in Science Conference (2020) https://doi.org/grr54q
DOI: 10.25080/majora-342d178e-010

42. **seaborn: statistical data visualization**
Michael Waskom
Journal of Open Source Software (2021-04-06) https://doi.org/gjqn3g
DOI: 10.21105/joss.03021

43. **Numba**
Siu Kwan Lam, Antoine Pitrou, Stanley Seibert
Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC (2015-11-15) https://doi.org/gf3nks
DOI: 10.1145/2833157.2833162

44. https://github.com/quinlan-lab/proj-mutator-mapping

45. **Sustainable data analysis with Snakemake**
Felix Mölder, Kim Philipp Jablonski, Brice Letcher, Michael B Hall, Christopher H Tomkins-Tinch, Vanessa Sochat, Jan Forster, Soohyun Lee, Sven O Twardziok, Alexander Kanitz, … Johannes Köster
F1000Research (2021-01-18) https://doi.org/gjjkwv
DOI: 10.12688/f1000research.29032.1 · PMID: 34035898 · PMCID: PMC8114187

46. **A natural mutator allele shapes mutation spectrum variation in mice**
Tom Sasani
(2023-01-24) https://github.com/tomsasani/bxd_mutator_manuscript

47. **tomsasani/bxd_mutator_manuscript: Final figure generation updates prior to publication**
Tom Sasani
Zenodo (2022-02-01) https://doi.org/grrwv8
DOI: 10.5281/zenodo.5941048

48. **BXD Genotype / WebQTL** https://gn1.genenetwork.org/dbdoc/BXDGeno.html

49. **Genetics and Probability in Animal Breeding Experiments**
https://link.springer.com/book/10.1007/978-1-349-04904-2

50. **MGI-Guidelines for Nomenclature of Mouse and Rat Strains**
http://www.informatics.jax.org/mgihome/nomen/strains.shtml

51. **Sustainable data analysis with Snakemake.**
Felix Mölder, Kim Philipp Jablonski, Brice Letcher, Michael B Hall, Christopher H Tomkins-Tinch, Vanessa Sochat, Jan Forster, Soohyun Lee, Sven O Twardziok, Alexander Kanitz, … Johannes Köster
F1000Research (2021-01-18) https://www.ncbi.nlm.nih.gov/pubmed/34035898
DOI: 10.12688/f1000research.29032.2 · PMID: 34035898 · PMCID: PMC8114187

52. **ENA Browser** https://www.ebi.ac.uk/ena/browser/view/PRJEB45429

53. **Twelve years of SAMtools and BCFtools.**
Petr Danecek, James K Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O Pollard, Andrew Whitwham, Thomas Keane, Shane A McCarthy, Robert M Davies, Heng Li
GigaScience (2021-02-16) https://www.ncbi.nlm.nih.gov/pubmed/33590861
DOI: 10.1093/gigascience/giab008 · PMID: 33590861 · PMCID: PMC7931819

54. **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.**
Pablo Cingolani, Adrian Platts, Le Lily Wang, Melissa Coon, Tung Nguyen, Luan Wang, Susan J Land, Xiangyi Lu, Douglas M Ruden
Fly (2012) https://www.ncbi.nlm.nih.gov/pubmed/22728672
DOI: 10.4161/fly.19695 · PMID: 22728672 · PMCID: PMC3679285

55. **cyvcf2: fast, flexible variant analysis with Python.**
Brent S Pedersen, Aaron R Quinlan
Bioinformatics (Oxford, England) (2017-06-15) https://www.ncbi.nlm.nih.gov/pubmed/28165109
DOI: 10.1093/bioinformatics/btx057 · PMID: 28165109 · PMCID: PMC5870853

56. **Mouse genomic variation and its effect on phenotypes and gene regulation.**
Thomas M Keane, Leo Goodstadt, Petr Danecek, Michael A White, Kim Wong, Binnaz Yalcin, Andreas Heger, Avigail Agam, Guy Slater, Martin Goodson, … David J Adams
Nature (2011-09-14) https://www.ncbi.nlm.nih.gov/pubmed/21921910
DOI: 10.1038/nature10413 · PMID: 21921910 · PMCID: PMC3276836

57. **A simple regression method for mapping quantitative trait loci in line crosses using flanking markers.**
CS Haley, SA Knott
Heredity (1992-10) https://www.ncbi.nlm.nih.gov/pubmed/16718932
DOI: 10.1038/hdy.1992.131 · PMID: 16718932

58. **coxme: Mixed Effects Cox Models**
Terry M Therneau
(2022-10-03) https://CRAN.R-project.org/package=coxme

59. **R/qtl2: Software for Mapping Quantitative Trait Loci with High-Dimensional Data and Multiparent Populations.**
Karl W Broman, Daniel M Gatti, Petr Simecek, Nicholas A Furlotte, Pjotr Prins, Śaunak Sen, Brian S Yandell, Gary A Churchill
Genetics (2018-12-27) https://www.ncbi.nlm.nih.gov/pubmed/30591514
DOI: 10.1534/genetics.118.301595 · PMID: 30591514 · PMCID: PMC6366910