Discovering epistasis between germline mutator alleles in mice

This manuscript (permalink) was automatically generated from quinlan-lab/mutator-epistasis-manuscript@7261c41 on February 22, 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 [PMID:28485537?]. 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 [PMID:35545679?,PMID:33472028?].

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 [PMID:17581577?]. 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 [PMID:32619789?], 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 [PMID:28959963?,PMID:31549960?]; in mice, that number is closer to 10 or 15 [PMID:31492841?]. 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 $2s\Delta U$ [PMID:27739533?], 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 [PMID:28095480?]. However, protein-coding genes involved in DNA replication and repair often recognize particular sequence motifs or excise lesions at specific nucleotides [PMID:32619789?]. 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 [PMID:33472028?], 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 [PMID:30753674?]. 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 [1]. 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 [PMID:35545679?]. The QTL overlapped *Mutyh*, which encodes a protein that normally prevents C>A mutations by repairing oxidative DNA damage [PMID:17581577?], 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 [PMID:35545679?].

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.

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 [PMID:33472028?]. Previously, we analyzed whole-genome sequencing data from the BXDs and identified candidate *de novo* germline mutations in each line [PMID:35545679?]. 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 [PMID:35545679?].

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

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 1).

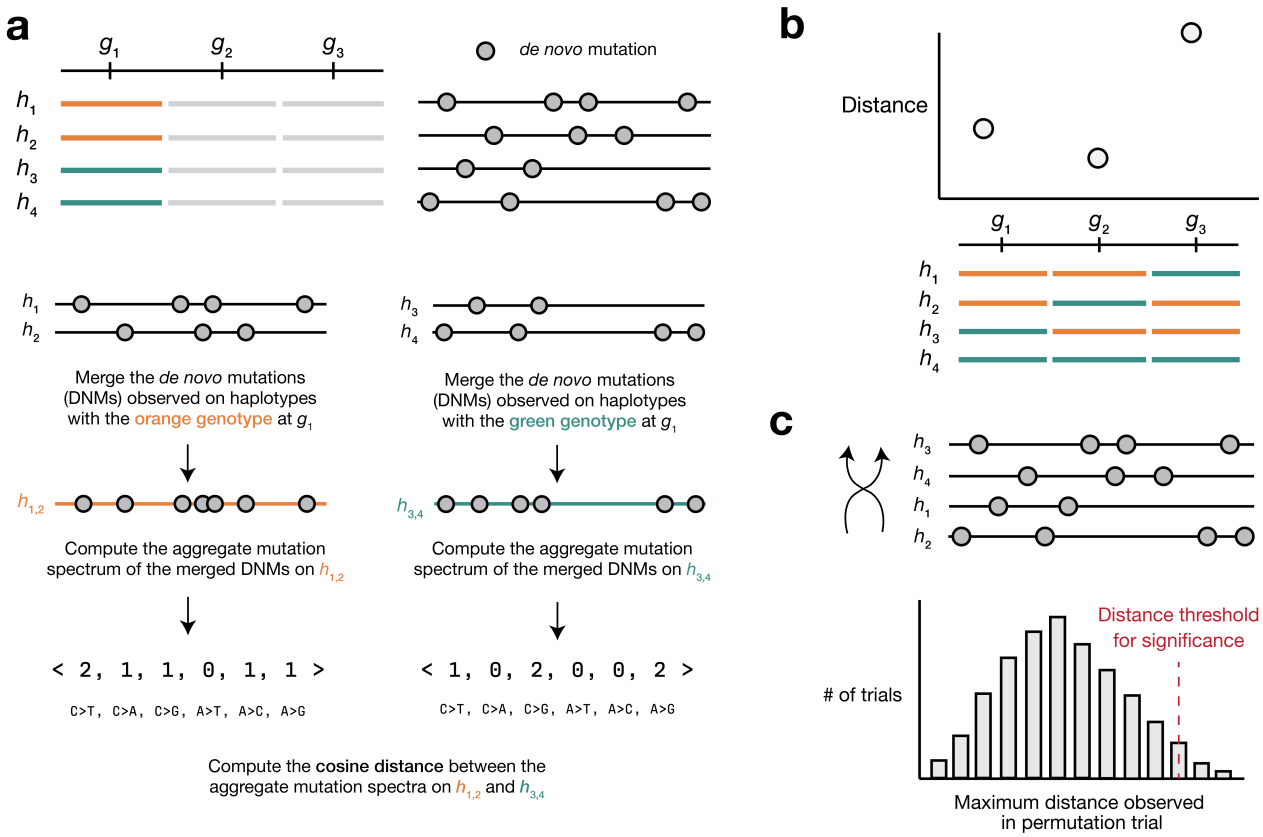

Figure 1: Overview of inter-haplotype distance method for discovering mutator alleles. a) A population of four haplotypes has been genotyped at three informative markers; each haplotype also harbors private *de novo* germline mutations. At each informative marker, we compute an aggregate *de novo* germline mutation spectrum in the haplotypes that carry either parental allele, and calculate the cosine distance 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 cosine distance 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 cosine distance 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 cosine distance between the aggregate mutation spectra of haplotypes with either parental allele. 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.

Similar to existing methods for quantitative trait locus mapping [**PMID:30591514?**], we use permutation tests to establish genome-wide cosine distance 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 $p = 0.05$).

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` [2,3,4,5,6,7,8].

Additional documentation is available on GitHub [9], along with a reproducible Snakemake [10] 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 [**PMID:31492841?**].

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 N$\underline{\text{C}}$N>N$\underline{\text{T}}$N 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 the 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 "base" 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 cosine distance 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 cosine distance between the aggregate spectra of the wild-type and mutator haplotypes. If fewer than 5% of the N permutations produces a cosine distance greater than or equal to D_f, we say that the approach succesfully 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 [11,12,**PMID:35545679?**], and downloaded a CSV file of BXD genotypes at 7,320 informative markers from GeneNetwork [13,**PMID:27933521?**]. We also downloaded relevant metadata about each BXD RIL from the manuscript describing the updated BXD resource [**PMID:33472028?**].

As in our previous manuscript [**PMID:35545679?**], 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 ~99.8% homozygous [14]. 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. 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* [**PMID:35545679?**]. In total, we included 94 BXD RILs in our genome-wide scans.

We used Snakemake [15] 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 [9].

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** [16]) using `bcftools` [**PMID:33590861?**]. 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` [**PMID:22728672?**] command:

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

and used `cyvcf2` [**PMID:28165109?**] 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 [**PMID:30753674?**] (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 [**PMID:21921910?**]. 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 [**PMID:35545679?**].

Querying GeneNetwork for evidence of eQTLs at the mutator locus

We used the online GeneNetwork resource [PMID:27933521?], 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 cosine distance 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 3. In the **Get Any** text box, we then entered the specified gene name and clicked **Search**. After selecting the appropriate data record on the next page, we used the **Mapping Tools** dropdown to run Haley-Knott regression [PMID:16718932?] with the following parameters: WGS-based marker genotypes, 1,000 permutations for LOD threshold calculations, and controlling for BXD genotypes at the `rsm10000007390` marker.

The exact names of the expression datasets we used for each tissue are shown in Table 1 below:

Table 1: 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

For each of the three candidate nonsynonymous mutations we identified under the cosine distance peak on chromosome 6 (Table 2), we queried a VCF file containing genome-wide variation in 67 wild-derived mice from four species of *Mus* [PMID:27622383?]. 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

Results

Benchmarking the inter-haplotype distance method using simulations

We first tested the inter-haplotype cosine 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 2). 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.

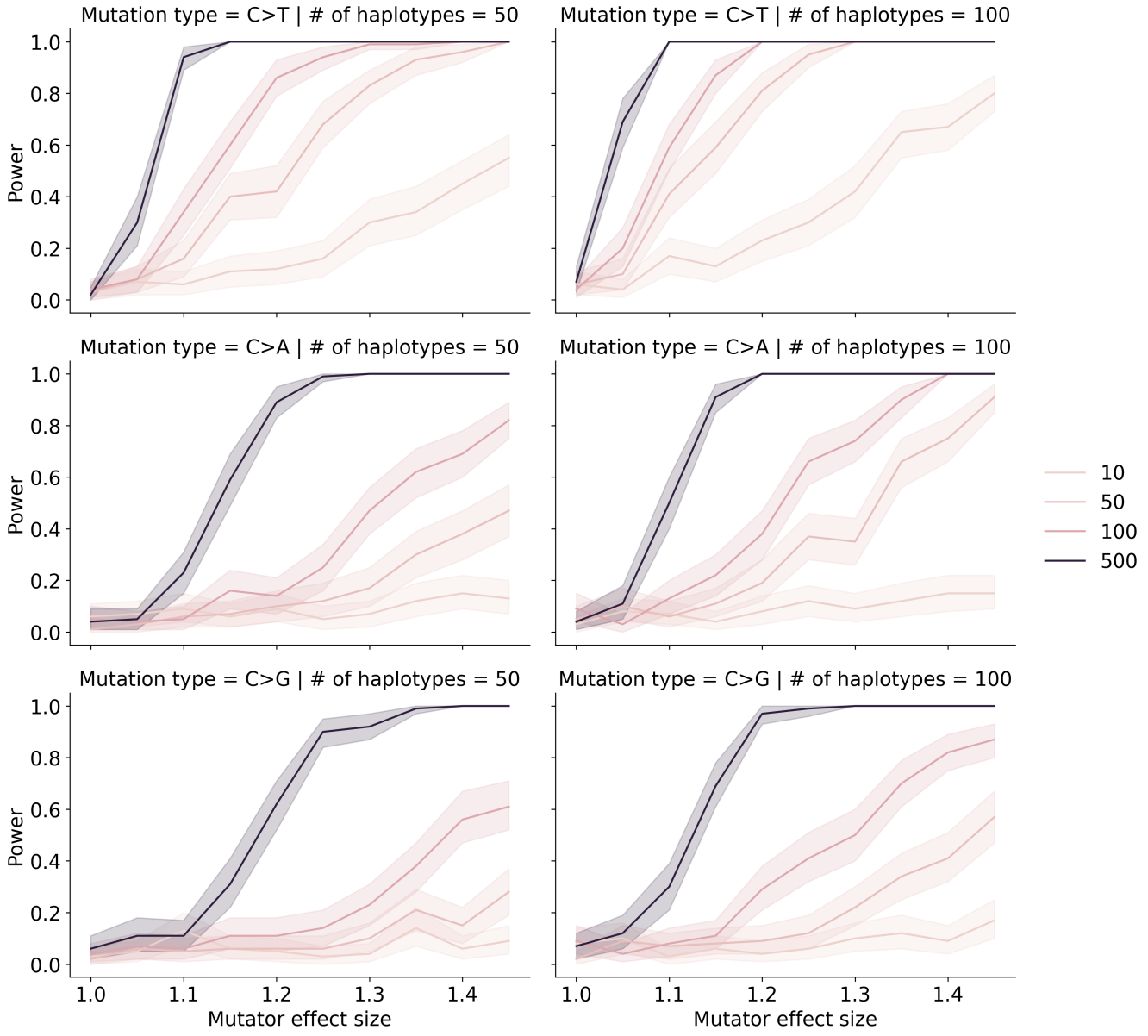

Figure 2: Simulations to assess the power of the inter-haplotype distance method. We simulated *de novo* germline mutations on the specified number of haplotypes, such that 50% of haplotypes were affected by a mutator allele that increased the mutation rate of the specified k-mer by the specified effect 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 94 BXD RILs (Materials and Methods) with a total of 63,914 *de novo* germline mutations [**PMID:35545679?**]. Reassuringly, we observed a large peak in cosine distance at a locus on chromosome 4 (Figure 3A; maximum distance of X at marker ID rsYYYYY; position 116.8 Mbp in mm10 coordinates).

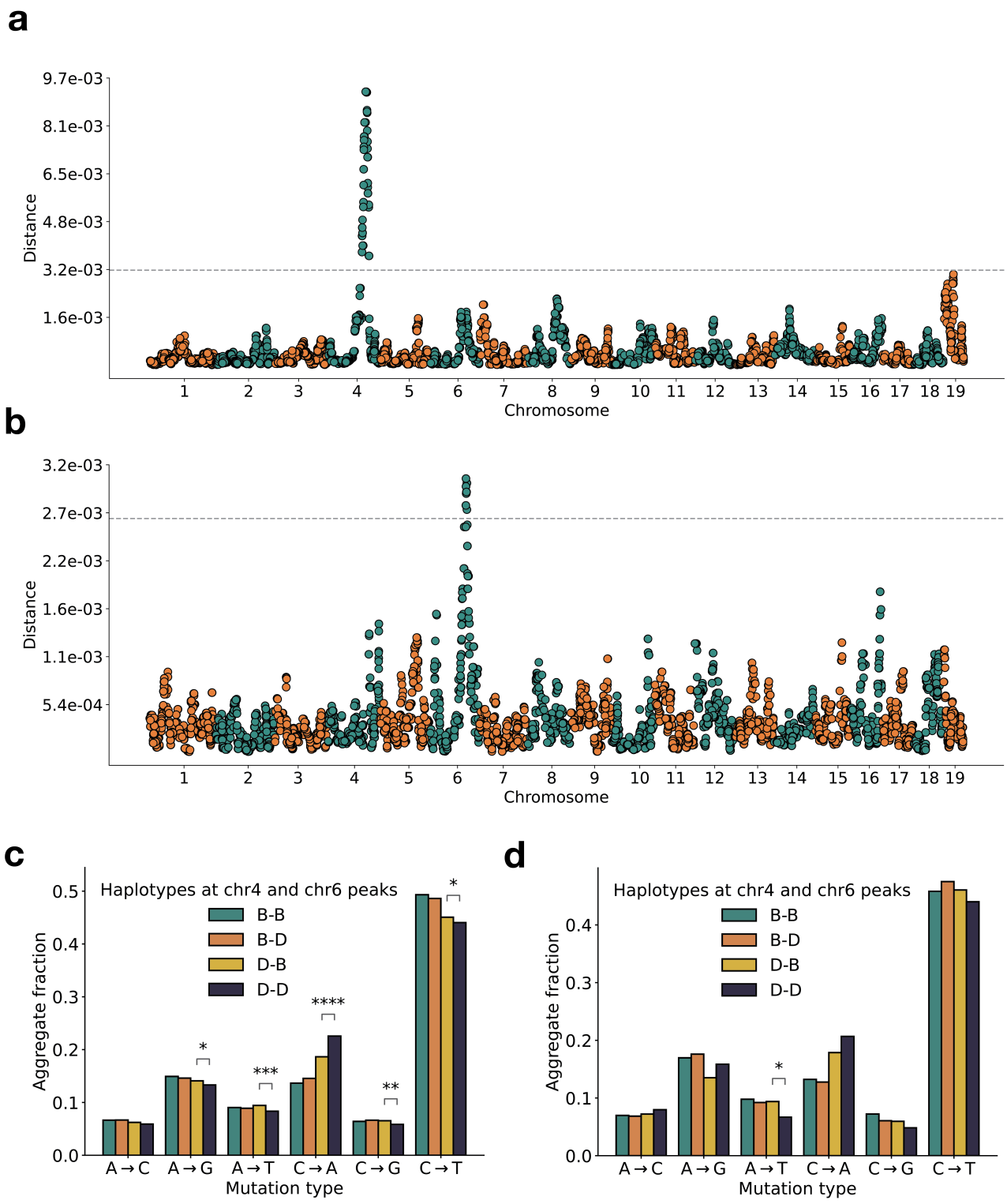

Figure 3: Results of inter-haplotype distance scans in the BXD RILs. a) Cosine distances between aggregate *de novo* mutation spectra on BXD haplotypes (n = 94)with either *D* or *B* alleles at approximately 7,500 informative markers. Distance threshold at $p = 0.05$ was calculated by performing 10,000 permutations of the BXD haplotype mutation data, and is shown as a dotted grey line. One outlier RIL with an extremely high C>A mutation rate (BXD68) was removed prior to running the distance scan. **b)** Cosine distances between aggregate *de novo* mutation spectra on BXD haplotypes with

either *D* or *B* alleles at approximately 7,500 informative markers. Only BXDs with *D* haplotypes at marker rsYYYY on chromosome 4 were included in the scan. 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 rsXXXX and rsYYYY. Total counts of each mutation type were aggregated in groups of BXDs with either of the four possible haplotype combinations, and fractions of each mutation type were calculated in each group separately. Chi-square tests of independence were used to compare counts of individual mutation types between the D-D and D-B groups of BXDs, as well as between the B-D and B-B groups. **d)** Fractions of *de novo* germline mutations in Sanger Mouse Genome Project (MGP) strains with either *D* or *B* haplotypes at markers rsXXXX and rsYYYY. Total counts of each mutation type were aggregated in groups of MGP strains with either of the four possible haplotype combinations, and fractions of each mutation type were calculated in each group separately. Chi-square tests of independence were used to compare counts of individual mutation types between the D-D and D-B groups of BXDs, as well as between the B-D and B-B groups.

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 [PMID:35545679?]. 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 [PMID:28551381?,PMID:28127763?,PMID:17581577?]. C>A germline mutation rates are nearly 50% higher in BXDs that inherited *D* haplotypes at marker ID rsYYYY than in those that inherited *B* haplotypes [PMID:35545679?].

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 = X) or *B* (n = Y) genotypes at rsYYYYY (the marker at which we observed the highest inter-haplotype cosine distance on chromosome 4), and ran a genome-wide distance scan using each group separately (Figure 3B.

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 cosine distance peak on chromosome 6 (Figure 3B; maximum distance of X at marker rsYYYYY; position 133.2 Mbp in mm10 coordinates). We queried the region underneath this peak and discovered two genes annotated with the Gene Ontology term "DNA repair": *Setmar*, a protein with histone methyltransferase and transposase activity, and remarkably, *Ogg1*. The latter encodes a key member of the base-exision repair response to oxidative DNA damage, a pathway that also includes *Mutyh* and a related gene, *Mth1*. Both *Setmar* and *Ogg1* harbor fixed nonsynonymous differences between the C57BL/6J and DBA/2J parental strains (Table 2).

Table 2: Summary of nonsynonymous differences between C57BL/6J and DBA/2J haplotypes in DNA repair genes near the mutator locus on chromosome 6.

Gene name	Amino acid change	Position in GRCm38/mm10 coordinates
Ogg1	p.Ala95Thr	chr6:
Setmar	p.Leu103Phe	chr6:
Setmar	p.Ser273Arg	chr6:

We also considered the possibility that expression quantitative trait loci (eQTLs), rather than nonsynonymous mutations, were responsible for the C>A mutator phenotype linked to the locus on chromosome 6. Using GeneNetwork [PMID:27933521?], we mapped cis-eQTLs for *Ogg1* in a number of tissues, including hematopoetic stem cells, kidney, and spleen. BXD genotypes at the cosine

distance peak on chromosome 6 were significantly associated with *Ogg1* expression in many tissues, and *D* genotypes were nearly always associated with decreased gene expression (Table 3). We also queried a previously published collection of eQTLs derived from Diversity Outbred (DO) mouse embryonic stem cell (mESCs) expression data [17], but did not find any significant eQTLs for *Ogg1*.

Table 3: Presence of 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`	**56.41**	17.80	-0.18
Gastrointestinal	46	`rsm10000003441`	**30.10**	16.21	-0.081
Hematopoetic stem cells	22	-	-	16.43	-
Hematopoetic progenitor cells	23	-	-	18.27	-
Spleen	79	`rsm10000003418`	**17.72**	17.49	-0.056
Liver	50	`rsm10000004188`	**52.27**	18.81	-0.155
Heart	73	-	-	16.12	-
Eye	87	`rsm10000004194`	**22.66**	17.20	0.087

Evidence of epistasis between germline mutator alleles

We next compared the mutation spectra of BXDs with either *B* or *D* genotypes at the mutator loci on chromosomes 4 and 6. 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 3C). 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 3C).

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 [**PMID:30753674?**] previously identified private germline mutations in 29 inbred laboratory strains; these private variants likely represent recent *de novo* germline mutations. Only two of the MGP strains possess *D* genotypes at both the chromosome 4 and chromosome 6 mutator loci: DBA/1J and DBA/2J. Compared to strains with *B* alleles at both mutator loci, those with *D* alleles at both mutator loci exhibit significantly higher C>A germline mutation fractions ($p = 3.5\text{e}{-8}$, Figure 3D). MGP strains with *D* alleles at both mutator loci appear to have higher C>A mutation fractions than those with *D* alleles at either locus alone (Figure 3D), but this difference is not significant ($p = 0.16$). Therefore, given the smaller number of MGP strains with *de novo* germline mutation data, we are unable to confirm the signal of epistasis observed in the BXDs.

The candidate mutator alleles are present in wild mice

To determine whether the candidate mutator alleles on chromosome 6 were segregating in natural populations of mice, we queried previously published sequencing data generated from 67 wild-derived mice [PMID:27622383?]. These data include three subspecies of *Mus musculus*, as well as the outgroup *Mus spretus*. We found that every *D* allele in *Setmar* or *Ogg1* was segregating in at least one population of *Mus* (Figure 4). Notably, the nonsynonymous p.Ala95Thr mutation was present at approximately 25% allele frequency in *Mus musculus domesticus*, and fixed in all other subspecies.

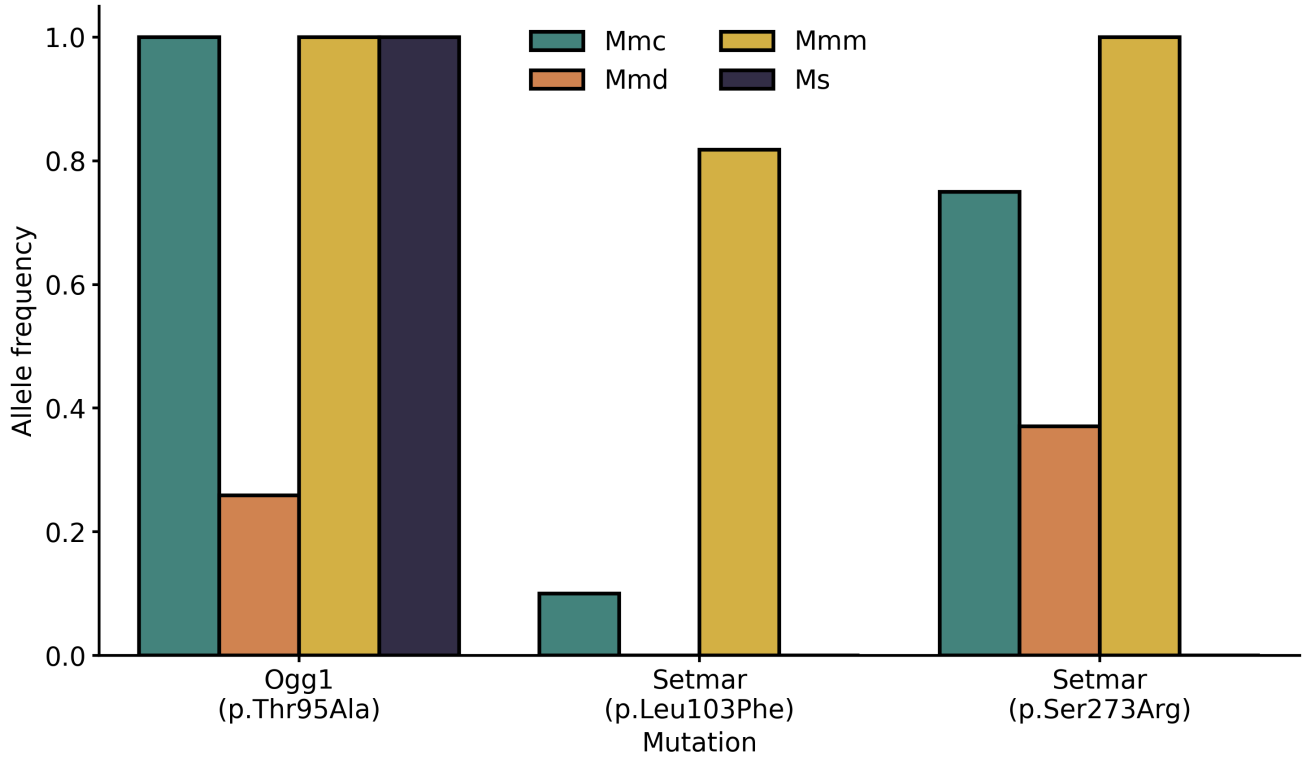

Figure 4: Frequency of candidate mutator alleles in wild mice. We queried a VCF file containing variant calls from 67 wild-derived *Mus* samples for each of the three fixed nonsynonymous differences between C57BL/6J and DBA/2J that were observed in DNA repair genes near the mutator locus on chromosome 6. The species abbreviations are: Ms: *Mus spretus*; Mmd: *Mus musculus domesticus*; Mmc: *Mus musculus castaneus*; Mmm: *Mus musculus musculus*

Discussion

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 [PMID:28440220?] 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 [PMID:34016747?]).

Within somatic tissues, mutation spectra can also be used to uncover the mutational processes active in particular populations of cells [PMID:23945592?]. New computational methods have been developed to extract “mutational signatures” from large databases of somatic mutations in cancer [18]. 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 [PMID:23945592?,PMID:31740835?,PMID:27811275?].

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 3). 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.

Epistasis between germline mutator alleles

Our results also 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 [**PMID:35545679?**] 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 3C). 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 [1]. Howver, when we queried wild mouse genomes for for the three nonsynonymous DNA repair mutations overlapping the chromosome 6 mutator locus, we found all three at high frequency in *Mus musculus domesticus*, the strain from which C57BL/6J and DBA/2J derive most of their genomes [**PMID:17660819?**]. 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

We discovered three nonsynonymous fixed differences in DNA repair genes between C57BL/6J and DBA/2J under the C>A mutator locus on chromosome 6 (Table 2). One of these mutations affects *Ogg1*, a protein-coding gene that participates in the base-excision repair of 8-oxoguanine (8-oxoG), a DNA lesion produced by oxidative DNA damage [**PMID:17581577?**]. Both missense mutations and loss-of-heterozygosity in *Ogg1* have been associated with initiation and progression of various types of human cancer [**PMID:22829015?**,**PMID:10987279?**,**PMID:9662341?**]. 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 [19]. 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.Ala95Thr missense mutation (Table 2) is the causal allele. We hypothesized that missense mutations in *Mutyh* were responsible for the large-effect C>A mutator phenotype we previously observed in the BXDs [**PMID:35545679?**]. 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* [20] that is present on *D* haplotypes and absent from *B* haplotypes. The MEI is associated with significantly reduced expression of *Mutyh* in laboratory strains, and may 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* are responsible for the C>A mutator phenotype we observed in this study (Table 3).

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 [**PMID:34523420?**]. A recent analysis suggested that mutator alleles and/or environmental mutagens have shaped mutation rate evolution in human genomes [21], and more generally, the mutation spectrum has evolved rapidly during great ape evolution, potentially due to the effects of *trans*-acting mutation rate modifiers [**PMID:33983415?**]. 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 [22]). 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 [**PMID:31549960?**,**PMID:28959963?**]. 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 [**PMID:35666194?**]. 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 [**PMID:35666194?**]. 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.

References

1. **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

2. **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) https://doi.org/ghbzf2
DOI: https://doi.org/10.1038/s41586-020-2649-2

3. **pandas-dev/pandas: Pandas**
The pandas development team
Zenodo (2023-02-20) https://doi.org/ggt8bh
DOI: https://doi.org/10.5281/zenodo.3509134

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

5. **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

6. **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

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

8. **Numba: a LLVM-based Python JIT compiler**
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: https://doi.org/10.1145/2833157.2833162 · ISBN: 9781450340052

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

10. **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: https://doi.org/10.12688/f1000research.29032.1

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

12. **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

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

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

15. **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

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

17. **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

18. **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

19. **Defects in 8-oxo-guanine repair pathway cause high frequency of C > 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

20. **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

21. **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

22. **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