

# Epistasis between mutator alleles contributes to germline mutation rate variability in laboratory mice

This manuscript ([permalink](#)) was automatically generated from [quinlan-lab/mutator-epistasis-manuscript@92c6e65](mailto:quinlan-lab/mutator-epistasis-manuscript@92c6e65) on April 18, 2023.

## Authors

---

- **Thomas A. Sasani**  
 [0000-0003-2317-1374](#) ·  [tomsasani](#) ·  [tomsasani](#)  
Department of Human Genetics, University of Utah
- **Aaron R. Quinlan**    
 [0000-0003-1756-0859](#) ·  [aaronquinlan](#)  
Department of Human Genetics, University of Utah; Department of Biomedical Informatics, University of Utah · Funded by G
- **Kelley Harris**    
 [0000-0003-0302-2523](#) ·  [Kelley\\_Harris](#)  
Department of Genome Sciences, University of Washington · Funded by NIH/NIGMS R35GM133428; Burroughs Wellcome Career Award at the Scientific Interface; Searle Scholarship; Pew Scholarship; Sloan Fellowship; Allen Discovery Center for Cell Lineage Tracing

✉ — Correspondence possible via [GitHub Issues](#) or email to Aaron R. Quinlan <[aquinlan@genetics.utah.edu](mailto:aquinlan@genetics.utah.edu)>, Kelley Harris <[harriske@uw.edu](mailto:harriske@uw.edu)>.

# Abstract

---

Maintaining germline genome integrity is essential and enormously complex. Hundreds of proteins comprise pathways involved in DNA proofreading, 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 — that is, 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 mapping to discover a mutator allele overlapping the DNA repair gene *Mutyh* that increased 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 approach to mutation data from the BXDs, we confirmed the activity of the germline mutator locus near *Mutyh* and discovered an additional C>A mutator locus on chromosome 6 that overlaps *Ogg1* and *Mbd4*, two DNA glycosylases involved in base-excision repair [4,5]. Strikingly, BXDs with mutator alleles on chromosome 6 only exhibit elevated rates of C>A germline mutation if they also possess mutator alleles near *Mutyh*, and BXDs with both alleles have even higher C>A mutation rates than those with either one alone.

Our new methods for analyzing mutation spectra reveal evidence of epistasis between eukaryotic germline mutator alleles, and may be applicable to mutation data from humans and other model organisms.

## Introduction

---

Germline mutation rates reflect the complex interplay between DNA proofreading and repair pathways, exogenous sources of DNA damage, and life-history traits. For example, parental age is an important determinant of mutation rate variability; in many mammalian species, the number of germline *de novo* mutations observed in offspring increases as a function of paternal and maternal age [6,7,8,9,10]. Rates of germline mutation accumulation are also variable across human families [7,11], potentially due to genetic variation or differences in environmental exposures. Nonetheless, genetic variants that increase germline mutation rates, known as *mutator alleles*, have proven difficult to discover in mammals.

The dearth of observed germline mutators in mammalian genomes is not necessarily surprising, since alleles that lead to elevated germline mutation rates would likely have deleterious consequences and be purged by negative selection [12]. Moreover, germline mutation rates are relatively low, and direct mutation rate measurements require whole-genome sequencing data from both parents and their offspring. As a result, large-scale association studies — which have been used to map the contributions of common genetic variants to many complex traits — are not currently well-powered to investigate the polygenic architecture of germline mutation rates.

Despite these challenges, less traditional strategies have been used to identify a small number of mutator alleles in both humans and mice. By focusing on families with rare genetic diseases, a recent study identified two mutator alleles that lead to significantly elevated rates of *de novo* germline mutation in human genomes [13]. Another group discovered mutator phenotypes in the sperm and somatic tissues of adults who possess cancer-predisposing inherited mutations in the POLE/POLD1

exonucleases [14]. Candidate mutator loci were also discovered by identifying human haplotypes with excess counts of derived alleles in the Thousand Genomes Project, though these loci could not be replicated using *de novo* germline mutation data from pedigrees [15].

In mice, a germline mutator allele was recently discovered by sequencing a large collection of 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 [16]. The BXD RILs were maintained via brother-sister mating for up to 180 generations, and each line therefore accumulated hundreds or thousands of germline mutations on a nearly-homozygous linear mosaic of parental B and D haplotypes. Due to their husbandry in a controlled laboratory setting, the BXDs were largely free from confounding by environmental heterogeneity, and the effects of selection on *de novo* mutations were attenuated by strict inbreeding [17].

In this previous study, whole-genome sequencing data from the BXD strains were used to map a quantitative trait locus (QTL) for the C>A mutation rate [2]. Germline C>A mutation rates were nearly 50% higher in mice with D haplotypes at the QTL, likely due to genetic variation in the DNA glycosylase *MutYh* that reduced the efficacy of oxidative DNA damage repair. Importantly, the QTL was undetectable in a genome-wide scan for variation in overall germline mutation rates, which were only modestly higher in BXDs with D alleles, demonstrating the utility of *mutation spectrum* analysis for mutator allele discovery. Close examination of the mutation spectrum is likely to be broadly useful for detecting mutator alleles, as genes involved in DNA proofreading and repair often recognize particular sequence motifs or excise specific types of DNA lesions [18]. Mutation spectra are usually defined in terms of  $k$ -mer nucleotide context [18]; the 1-mer mutation spectrum, for example, consists of 6 mutation types after collapsing by strand complement (C>T, C>A, C>G, A>T, A>C, A>G), while the 3-mer mutation spectrum contains 96 (each of the 1-mer mutations partitioned by trinucleotide context).

Although mutation spectrum analysis can enable the discovery of mutator alleles that affect the rates of specific mutation types, early implementations of this strategy have suffered from a few drawbacks. For example, performing association tests on the rates or fractions of every  $k$ -mer mutation type can quickly incur a substantial multiple testing burden. Since germline mutation rates are generally quite low, estimates of  $k$ -mer mutation type frequencies from individual samples may also be noisy and imprecise. We were therefore motivated to develop a statistical method that could overcome the sparsity of *de novo* mutation spectra, eliminate the need to test each  $k$ -mer mutation type separately, and enable sensitive detection of alleles that influence the germline mutation spectrum.

Here, we present a new mutation spectrum association test that minimizes multiple testing burdens and mitigates the challenges of sparsity in *de novo* mutation datasets. We leverage this method to re-analyze germline mutation data from the BXD cohort and find compelling evidence for the existence of a second mutator allele that was undetectable using previous approaches. The new allele appears to interact epistatically with the mutator that was previously discovered in the BXDs, further augmenting the C>A germline mutation rate in a subset of RILs. Our observation of epistasis suggests that mild DNA repair deficiencies may compound one another, as mutator alleles chip away at the redundant systems that collectively maintain germline integrity.

## Results

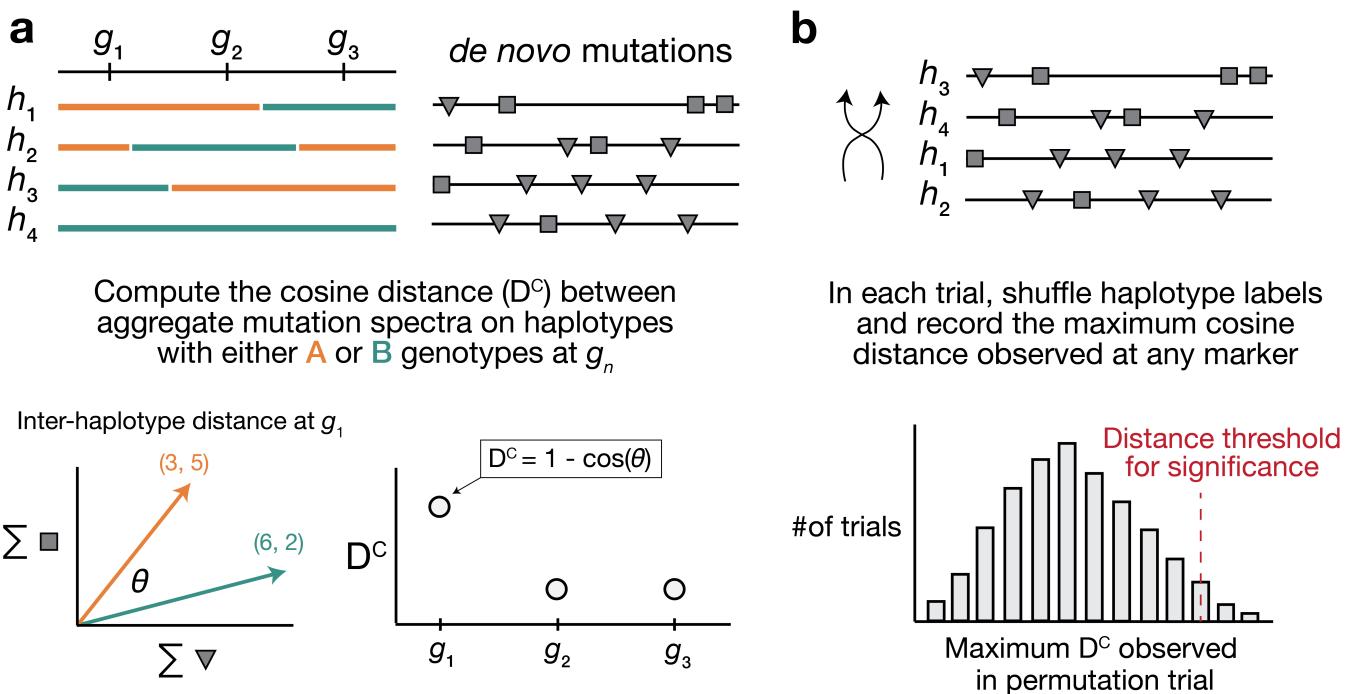
---

### A novel method for detecting mutator alleles

We developed a statistical method, termed “inter-haplotype distance” (IHD), to detect loci that are associated with mutation spectrum variation in biparental RILs (Figure 1). Our approach leverages the fact that mutator alleles often leave behind distinct and detectable impressions on the *mutation*

spectrum, even if they increase the overall mutation rate by a relatively small amount (*Materials and Methods*). Given a population of haplotypes, we assume that a) each has been genotyped at the same collection of biallelic loci and b) each harbors *de novo* mutations that have been partitioned by  $k$ -mer context (Figure 1). At every locus, we calculate a cosine distance between the aggregate mutation spectra of haplotypes that inherited either parental allele. Using permutation tests, we then identify loci at which those distances are larger than what we'd expect by random chance.

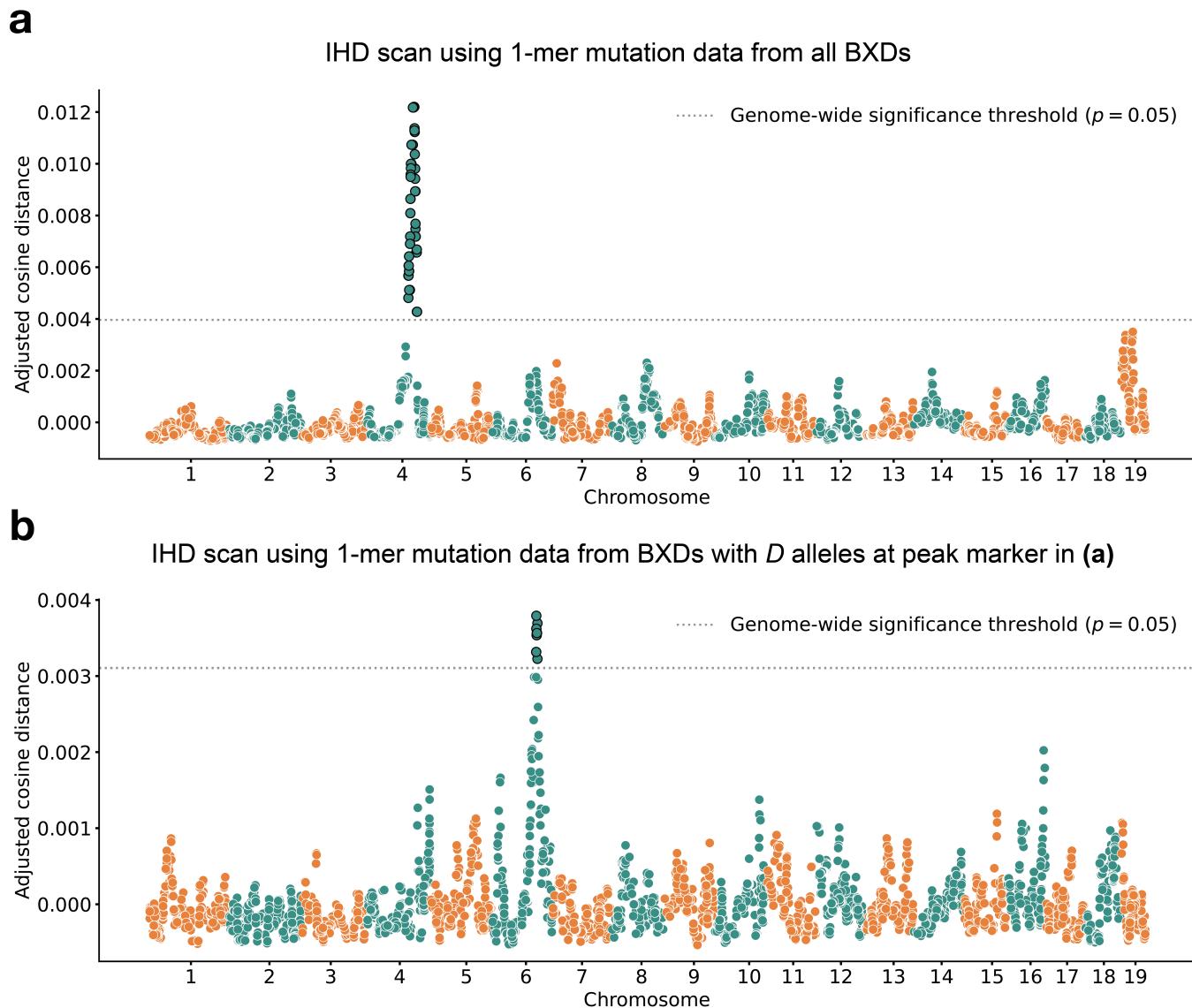
The method's power is primarily limited by the initial mutation rate of the  $k$ -mer mutation type affected by a mutator allele and the total number of *de novo* mutations used to detect it (Figure 1-[figure supplement 1](#)). Given 100 haplotypes with an average of 500 *de novo* germline mutations each, IHD has approximately 90% power to detect a mutator allele that increases the C>A *de novo* mutation rate by as little as 20%. However, the approach has less than 20% power to detect a mutator of identical effect size that augments the C>G mutation rate, since C>G mutations are expected to make up a smaller fraction of all *de novo* germline mutations to begin with. Simulations also demonstrate that our approach 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-[figure supplement 1](#)). Both IHD and traditional quantitative trait locus (QTL) mapping have similar power to detect alleles that augment the rates of individual 1-mer mutation types (Figure 1-[figure supplement 2](#)), but IHD has a number of potential advantages for mutator allele discovery; for a more detailed comparison of the methods, see the *Discussion*.



**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 ( $g_1$  through  $g_3$ ); each haplotype also harbors unique *de novo* germline mutations. In practice, *de novo* mutations are partitioned by  $k$ -mer context; for simplicity in this toy example, *de novo* mutations are simply classified into two possible mutation types (grey squares represent C>(A/T/G) mutations, while grey triangles represent A>(C/T/G) mutations). At each informative marker  $g_n$ , we calculate the total number of each mutation type observed on haplotypes that carry either parental allele (i.e., the aggregate mutation spectrum) using all genome-wide *de novo* mutations. We then calculate the cosine distance between the two aggregate mutation spectra, which we call the “inter-haplotype distance.” Cosine distance can be defined as  $1 - \cos(\theta)$ , where  $\theta$  is the angle between two vectors; in this case, the two vectors are the two aggregate spectra. We repeat this process for every informative marker  $g_n$ . **b)** To assess the significance of any distance peaks in a), we perform permutation tests. In each of  $N$  permutations, we shuffle the haplotype labels associated with the *de novo* mutation data, run a genome-wide distance scan, and record the maximum cosine distance encountered at any locus in the scan. Finally, we calculate the  $1 - p$  percentile of the distribution of those maximum distances to obtain a genome-wide cosine distance threshold at the specified value of  $p$ .

## 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]. Using mutation data that were partitioned by 1-mer nucleotide context, we discovered a locus on chromosome 4 that was significantly associated with mutation spectrum variation (Figure 2a; maximum adjusted cosine distance of 1.22e-2 at marker ID rs52263933 ; position 116.75 Mbp in GRCm38/mm10 coordinates).



**Figure 2: Results of inter-haplotype distance scans in the BXD RILs.** **a)** Adjusted cosine distances 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 7,321 informative markers. Cosine distance threshold at  $p = 0.05$  was calculated by performing 10,000 permutations of the BXD mutation data, and is shown as a dotted grey line. **b)** Adjusted cosine distances 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 7,278 informative markers. Cosine distance threshold at  $p = 0.05$  was calculated by performing 10,000 permutations of the BXD mutation data, and is shown as a dotted grey line.

Using quantitative trait locus (QTL) mapping, we previously identified a nearly-identical locus on chromosome 4 that was significantly associated with the C>A germline mutation rate in the BXDs [2]. This locus overlapped 21 protein-coding genes that were annotated by the Gene Ontology as being involved in “DNA repair,” but only one of these genes contained 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 [4,19,20]. C>A germline mutation rates are nearly 50% higher in BXDs that inherit *D* genotypes at marker ID rs52263933 (the marker at which we observed the highest adjusted cosine distance on chromosome 4) than in those that inherit *B* genotypes (Figure 3) [2].

## An additional germline mutator allele on chromosome 6

After confirming that IHD could recover the mutator locus overlapping *Mutyh*, we asked if our approach could identify additional mutator loci in the BXDs. In particular, we were interested in discovering epistatic interactions between alleles at the chromosome 4 locus and mutator alleles elsewhere in the genome. We hypothesized that such interactions could be detectable by first “conditioning” on the presence of *B* or *D* alleles at the mutator locus on chromosome 4, and then running another genome-wide scan for loci associated with mutation spectrum variation. To account for the effects of the large-effect mutator locus near *Mutyh*, we divided the BXD RILs into those with either *D* (n = 55) or *B* (n = 38) genotypes at rs52263933 and ran an inter-haplotype distance scan using each group separately (Figure 2b and [2-figure supplement 1](#)).

Using the BXDs with *D* genotypes at rs52263933, we identified a locus on chromosome 6 that was significantly associated with mutation spectrum variation (Figure 2b; maximum adjusted cosine distance of 3.69e-3 at marker rs31001331; position 114.05 Mbp in GRCm38/mm10 coordinates). This signal was specific to BXDs with *D* genotypes at the rs52263933 locus, as we did not observe any new mutator loci (other than some residual signal from the *Mutyh* locus) after performing an IHD scan using BXDs with *B* genotypes at rs52263933 ([Figure 2-figure supplement 1](#)). We also performed QTL scans for the rates of each 1-mer mutation type using the same mutation data, but none produced a genome-wide significant log-odds score at any locus ([Figure 2-figure supplement 2](#); *Materials and Methods*).

We queried the region surrounding the locus on chromosome 6 (+/- 5 Mbp) and discovered 15 protein-coding genes that harbored nonsynonymous differences between the parental C57BL/6J and DBA/2J strains. Two of these genes were also annotated with the Gene Ontology term “DNA repair”: *Ogg1* and *Mbd4*. *Ogg1* encodes a key member of the base-excision repair response to oxidative DNA damage (a pathway that also includes *Mutyh*), and *Mbd4* encodes a protein that is involved in the repair of G:T mismatches caused by spontaneous deamination of methylated CpGs. Each of these genes harbors a single fixed nonsynonymous difference between the C57BL/6J and DBA/2J parental strains (Table 1).

**Table 1:** Nonsynonymous mutations in DNA repair genes near the chr6 peak

Gene name	Ensembl transcript name	Nucleotide change	Amino acid change	Position in GRCm38/mm10 coordinates	PhyloP conservation score	SIFT prediction
<i>Ogg1</i>	ENSMUST00000032406	A>G	p.Thr95Ala	chr6:113,328,510	-0.016	0.84 (tolerant/benign)
<i>Mbd4</i>	ENSMUST00000032469	C>T	p.Asp129Asn	chr6:115,849,644	2.28	0.02 (intolerant/damaging)

We also considered the possibility that expression quantitative trait loci (eQTLs), rather than nonsynonymous mutations, could contribute to the C>A mutator phenotype associated with the locus on chromosome 6. Using GeneNetwork [21] we mapped cis-eQTLs for *Ogg1* and *Mbd4* in a number of tissues, including hematopoietic stem cells, kidney, and spleen; we did not have access to expression data from germline tissues. BXD genotypes near the cosine distance 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 [supplement 1](#)). We discovered one significant cis-eQTL for *Mbd4* in spleen at which *D* alleles were associated with increased expression. We also queried a previously published collection of eQTLs derived from Diversity Outbred (DO) mouse embryonic stem cell expression data [22], but did not find any significant eQTLs for either *Ogg1* or *Mbd4*.

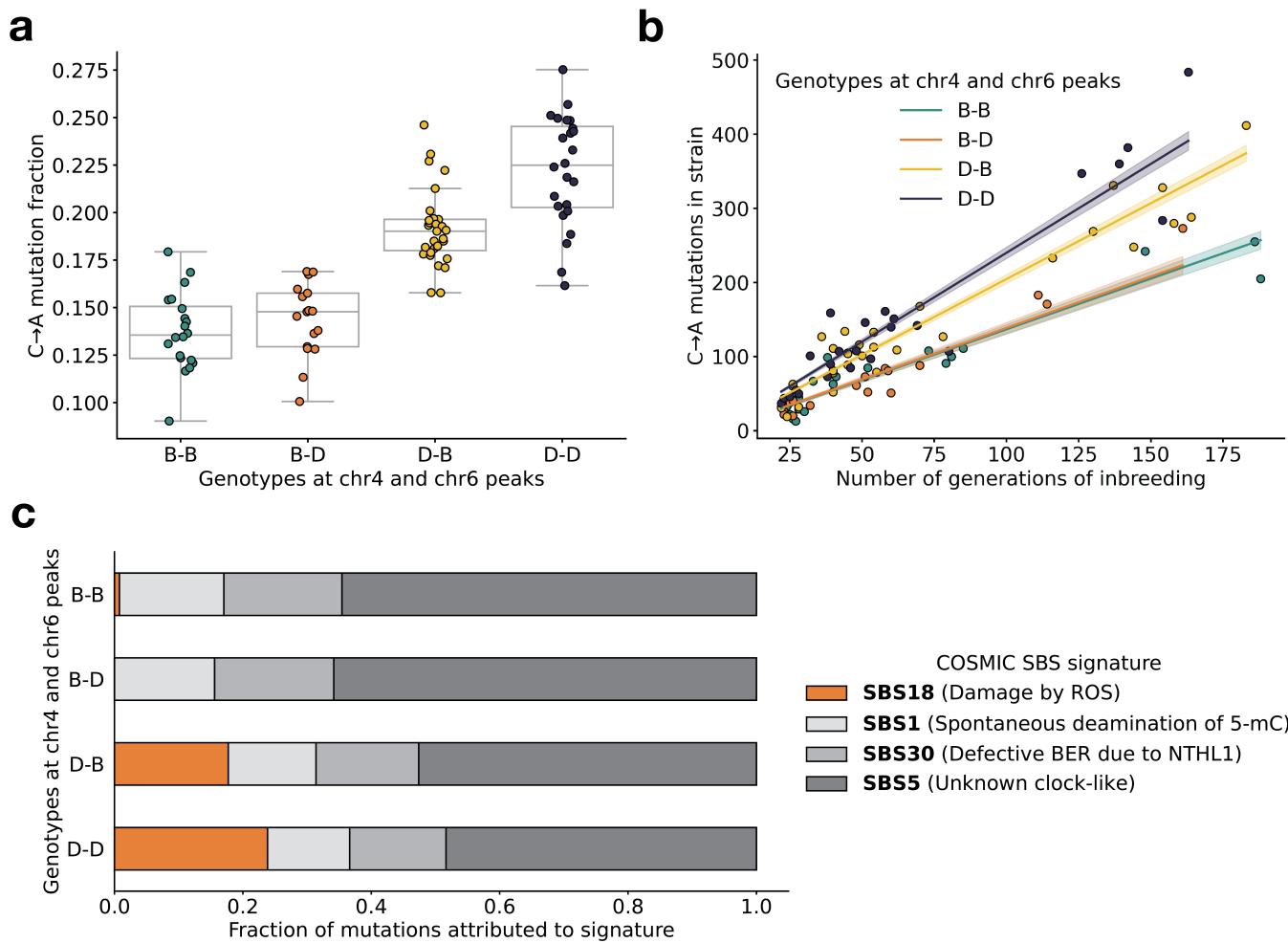
Finally, we queried a dataset of structural variants (SVs) identified via high-quality, long-read assembly of inbred laboratory mouse strains [23] and found 148 large insertions or deletions within 5 Mbp of the cosine distance peak on chromosome 6. Of these, seven overlapped the exonic sequences of protein-coding genes (Table [supplement 2](#)), though none of the genes has a previously annotated role in DNA binding, repair or replication, or in a pathway that would likely affect germline mutation rates. Two protein-coding genes that are involved in DNA repair (*Mbd4* and *Rad18*) harbored intronic insertions or deletions (Table [supplement 2](#)); however, it is challenging to interpret the functional impact of these SVs without additional experimental evidence.

## Evidence of epistasis between germline mutator alleles

Next, we more precisely characterized the effects of the chromosome 4 and 6 mutator alleles on mutation spectra in the BXDs. On average, C>A germline mutation fractions were higher in BXDs with *D* alleles at both mutator loci than in BXDs with *D* alleles at either locus alone (Figure 3a and [3-figure supplement 1](#)). The effects of *D* alleles at the chromosome 6 locus also depended on the presence of *D* alleles at *Mutyh*; among BXDs with *B* alleles at the locus overlapping *Mutyh*, those with *D* alleles on chromosome 6 did not exhibit higher C>A mutation fractions (Figure 3a). Regardless of how long the RILs were inbred, counts of C>A *de novo* mutations were always highest in BXDs with *D* alleles at both mutator loci (Figure 3b).

We also used SigProfilerExtractor [24] to assign the germline mutations in each BXD to single-base substitution (SBS) mutation signatures from the COSMIC catalog [25]. Mutation signatures often reflect specific exogenous or endogenous sources of DNA damage, and the fraction of mutations attributable to a particular SBS signature can suggest a genetic or environmental etiology. The SBS1, SBS5, and SBS30 mutation signatures were active in nearly all BXDs, regardless of genotypes at the chromosome 4 and 6 mutator loci (Figure 3). However, the SBS18 signature, which is dominated by C>A mutations and likely reflects unrepaired DNA damage from reactive oxygen species, was almost exclusively active in mice with *D* alleles at the chromosome 4 locus; the highest SBS18 activity was observed in mice with *D* alleles at both mutator loci (Figure 3). SBS18 activity was absent from mice with *D* alleles at the chromosome 6 mutator locus alone (Figure 3), further demonstrating that *D* alleles at this locus are not sufficient to cause a mutator phenotype.

To more formally test for epistasis, we fit a linear model predicting counts of C>A mutations in each RIL as a function of genotypes at [rs52263933](#) and [rs31001331](#) (the peak markers at the two mutator loci) (*Materials and Methods*). A model that included an interaction term between genotypes at the two markers fit the data significantly better than a model including only additive effects ( $p = 8.8e-4$ ; *Materials and Methods*). Genotypes at the chromosome 4 locus explain approximately 61% of the variance in BXD C>A mutation fractions, while genotypes at both loci together explain nearly 70%.



**Figure 3: BXD mutation spectra are affected by alleles at both mutator loci.** **a)** C<sub>></sub>A *de novo* germline mutation fractions in BXDs with either *D* or *B* genotypes at markers rs52263933 (chr4 peak) and rs31001331 (chr6 peak). **b)** Counts of C<sub>></sub>A *de novo* germline mutations in each BXD RIL plotted against the number of generations for which it was inbred. Lines represent predicted C<sub>></sub>A counts in each haplotype group from a Poisson regression (identity link), and shading around each line represents the 95% confidence interval. **c)** Germline mutations in each BXD were assigned to COSMIC SBS mutation signatures using SigProfilerExtractor [24]. After grouping BXDs by their genotypes at rs52263933 and rs31001331, we calculated the fraction of mutations in each group that was attributed to each signature. The proposed etiologies of each mutation signature are: SBS1 (spontaneous deamination of methylated cytosine nucleotides at CpG contexts), SBS5 (unknown, clock-like signature), SBS18 (damage by reactive oxygen species, possibly loss-of-function mutations in MUTYH), and SBS30 (defective base-excision repair due to NTHL1 mutations).

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 [26]. Dumont [16] previously identified germline mutations that were private to each of the 29 MGP strains; these private variants likely represent recent *de novo* mutations (Figure 3—figure supplement 2). 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<sub>></sub>A mutation counts as a function of genotypes at rs52263933 and rs31001331. A model incorporating an interaction term between genotypes at these loci did not fit the MGP data significantly better than a model with additive effects alone ( $p = 0.806$ ), so we are unable to confirm the signal of epistasis; however, this may be due to the smaller number of MGP strains with *de novo* germline mutation data.

## Only one of the candidate mutator alleles is present in wild mice

To determine whether the candidate mutator alleles on chromosome 6 were segregating in natural populations, we queried previously published sequencing data generated from 67 wild-derived mice

[27]. These data include three subspecies of *Mus musculus*, as well as the outgroup *Mus spretus*. We found that the *Ogg1 D* allele was segregating at an allele frequency of 0.259 in *Mus musculus domesticus*, the species from which C57BL/6J and DBA/2J derive the majority of their genomes [28], and was fixed in *Mus musculus musculus*, *Mus musculus castaneus*, and the outgroup *Mus spretus*. The *Mbd4 D* allele was not present in any of the wild mice.

## Discussion

---

### Epistasis between germline mutator alleles

We have identified a locus on chromosome 6 that amplifies a C>A germline mutator phenotype in the BXDs, a population of recombinant inbred lines (RILs) derived from the laboratory mice DBA/2J and C57BL/6J. DBA/2J (*D*) alleles at this locus have no detectable effect on C>A mutation rates in mice that also harbor “wild-type” C57BL/6J (*B*) alleles at a previously discovered mutator locus on chromosome 4 [2]. However, mice with *D* alleles at *both* loci have even higher mutation rates than those with *D* alleles at the chromosome 4 mutator locus alone (Figure 3). Epistatic interactions between mutator alleles have been previously documented in yeast [29] and in human cell lines [30], but never to our knowledge in a whole-animal context.

Importantly, we discovered this germline mutator allele epistasis in an unnatural population of model organisms that have been inbred by brother-sister mating in a highly controlled laboratory environment [3]. This breeding setup has likely attenuated the effects of natural selection on all but the most deleterious alleles [17], and may have facilitated the fixation of large-effect mutator alleles that would be less common in wild, outbreeding populations.

We have not conclusively fine-mapped the chromosome 6 mutator locus to a causal variant, but we argue that nonsynonymous mutations in the DNA glycosylases *Ogg1* and *Mbd4* are the best candidates. Our results demonstrate that multiple mutator alleles have spontaneously arisen during the evolutionary history of inbred laboratory mice, supporting the hypothesis that purifying selection is required to keep mutation rates low. Our finding suggests that mutational pressure can cause mutation rates to rise in just a few generations of relaxed selection in captivity. This is corroborated by the finding of a large-effect mutator allele in a rhesus macaque research colony [31], as well as the observation that domesticated animals tend to have higher mutation rates than wild animals [32].

### Possible causal alleles underlying the chromosome 6 mutator locus

Two DNA repair genes overlap the C>A mutator locus on chromosome 6 and also contain nonsynonymous differences between the C57BL/6J and DBA/2J founder strains: *Ogg1*, a glycosylase that excises the oxidative DNA lesion 8-oxoguanine (8-oxoG) [4], and *Mbd4*, a glycosylase that removes thymine nucleotides at G:T mispairs following spontaneous deamination of methylated CpGs. Both genes have been previously linked to somatic mutator phenotypes and carcinogenesis. Missense mutations and loss-of-heterozygosity in *Ogg1* have been associated with increased risk of human cancer [33,34], and copy-number losses of either *Ogg1* or *MutYH* are linked to elevated rates of spontaneous C>A mutation in human neuroblastoma [35]. Biallelic loss-of-function (LOF) mutations in human *MBD4* underlie a neoplastic syndrome that mimics forms of familial adenomatous polyposis caused by LOF mutations in *MUTYH* [36]. Similarly, biallelic loss of *MBD4* function was recently found to cause a hypermutator phenotype in the maternal germline of rhesus macaques [31]. This macaque germline mutator phenotype primarily comprises C>T mutations at CpG and CpA sites, a mutational signature that has previously been associated with LOF mutations in *Mbd4* [37].

Although we are unable to conclusively determine that either *Mbd4* or *Ogg1* harbors the causal variant underlying the observed C>A mutator phenotype, both genes are plausible candidates.

## An *Ogg1* mutator allele might impair the excision of 8-oxoguanine lesions

*Ogg1* is a member of the same base-excision repair (BER) pathway as *Mutyh*, the protein-coding gene we previously implicated as harboring mutator alleles at the locus on chromosome 4 [4]. Each of these genes has a distinct role in the BER response to oxidative DNA damage, and thereby the prevention of C>A mutations [38,39]. Following damage by reactive oxygen species, *Ogg1* is able to recognize and remove 8-oxoguanine lesions that are base-paired with cytosine nucleotides; once 8-oxoG is excised, other members of the BER pathway are mobilized to restore a proper G:C base pair at the site. If an 8-oxoG lesion is not removed before the cell enters S-phase, adenine can be mis-incorporated opposite 8-oxoG during DNA replication [38]. If this occurs, *Mutyh* can excise the mispaired adenine, leaving a one-nucleotide gap that is processed and filled with a cytosine by other BER proteins. The resulting C:8-oxoG base pair can then be “returned” to *Ogg1* for excision and repair. Defects in the BER response to oxidative damage lead to significantly elevated rates of C>A mutation. For example, triple-knockout (KO) mice lacking *Ogg1*, *Mutyh*, and *Mth1* (which encodes an enzyme that prevents 8-oxo-dGTP from being incorporated during DNA synthesis [40]) accumulate 100-fold excess 8-oxoG in their gonadal cells [39]. Almost 99% of *de novo* germline mutations in the *Ogg1/Mutyh/Mth1* triple KO mice are C>A transversions, demonstrating the clear role of 8-oxoG repair in preventing C>A mutation.

The p.Thr95Ala *Ogg1* missense variant is not predicted to be deleterious by *in silico* tools like SIFT [41], and does not occur at a nucleotide that is well-conserved across mammalian species (Table 1). We also observe that the *D* allele at p.Thr95Ala is segregating at an allele frequency of nearly 26% among wild-derived *Mus musculus domesticus* animals, and is fixed in other wild populations of *Mus musculus musculus*, *Mus musculus castaneus*, and *Mus spretus*. Although we would expect *a priori* that *Ogg1* deficiency would lead to increased 8-oxoG accumulation and elevated C>A mutation rates, these lines of evidence suggest that p.Thr95Ala is not highly deleterious on its own, and might only exert a detectable effect on the BER pathway when *Mutyh* function is also impaired. Overall, it is challenging to predict the functional consequences of the *Ogg1* p.Thr95Ala variant, and as we discuss below, the p.Asp129Asn missense mutation in *Mbd4* appears to be a more compelling candidate mutator allele based on conservation and variant effect prediction.

## *Mbd4* may buffer the effects of *Mutyh* mutator alleles by triggering apoptosis

Unlike the *Ogg1* p.Thr95Ala mutation, the p.Asp129Asn variant in *Mbd4* resides within an annotated protein domain (the *Mbd4* methyl-CpG binding domain), occurs at a nucleotide and amino acid residue that are both well-conserved, and is predicted to be deleterious by SIFT [41] (Table 1). A missense mutation that affects the homologous amino acid in humans (p.Asp142Gly in GRCh38/hg38) is also present on a single haplotype in the Genome Aggregation Database (gnomAD) [42] and is predicted by SIFT and Polyphen [43] to be “deleterious” and “probably\_damaging” in human genomes, respectively.

One puzzle is that loss-of-function mutations in *Mbd4* are not typically associated with C>A mutator phenotypes. Instead, *Mbd4* deficiency is usually implicated in C>T mutagenesis at CpG sites, and we did not detect an excess of C>T mutations in BXDs with *D* alleles at the chromosome 6 mutator locus (Figure 3-figure supplement 1). However, loss of function mutations in *Mbd4* are also known to exacerbate the effects of exogenous DNA damage agents. For example, mouse embryonic fibroblasts that harbor homozygous *Mbd4* knockouts fail to undergo apoptosis following treatment with a number of therapeutics and mutagenic compounds [44]. Most of these exogenous mutagens cause DNA damage that is normally repaired by mismatch repair (MMR) machinery, but murine intestinal cells with biallelic *Mbd4* LOF mutations also showed a reduced apoptotic response to gamma irradiation, which is repaired independently of the MMR gene *Mlh1* [45]. Homozygous loss of *Mbd4* function also leads to accelerated intestinal tumor formation in mice that harbor an *Apc* allele

that predisposes them to intestinal neoplasia [37], and mice with biallelic truncations of the *Mbd4* coding sequence exhibit modestly increased mutation rates in colon cancer cell lines, including increased C>A mutation rates in certain lines [46].

Together, these lines of evidence suggest that *Mbd4* can modulate sensitivity to many types of exogenous mutagens, potentially through its role in determining whether cells harboring DNA damage should undergo apoptosis [44,45]. We speculate that in mice with deficient 8-oxoguanine repair — caused by a mutator allele in *Mutyh*, for example — reactive oxygen species (ROS) could cause accumulation of DNA damage in the germline. If those germ cells harbor fully functional copies of *Mbd4*, they might be able to trigger apoptosis and mitigate the effects of a *Mutyh* mutator allele. However, mice with reduced activity of both *Mbd4* and *Mutyh* may have a reduced ability to initiate cell death in response to DNA damage; as a result, their germ cells may accumulate even higher levels of ROS-mediated damage, leading to substantially elevated germline C>A mutation rates.

Our model makes the testable prediction that mice with *D* haplotypes at *Mbd4* should have elevated sensitivity to a variety of mutagenic agents. We thus anticipate that future experimental work will be able to more conclusively establish a mechanistic explanation for the epistatic interaction between the mutator loci described in this paper.

## No indication of structural variation or mobile element insertions near the chromosome 6 mutator locus

Although we argue above that *Mbd4* and *Ogg1* are the best candidate genes to explain the augmented C>A mutator phenotype in a subset of BXDs, we cannot conclusively determine that either the p.Asp129Asn or p.Thr95Ala missense mutation is a causal allele. We previously hypothesized that *Mutyh* missense mutations on *D* haplotypes were responsible for the large-effect C>A mutator phenotype we observed in the BXDs [2]. However, subsequent long-read assemblies of several inbred laboratory mouse strains revealed that this mutator phenotype might be caused by a ~5 kbp mobile element insertion (MEI) within the first intron of *Mutyh* [23], which is associated with significantly reduced expression of *Mutyh* in embryonic stem cells. We queried the new high-quality assemblies for evidence of mobile elements or other large structural variants (SVs) in the region surrounding the mutator locus on chromosome 6, but found no similarly compelling evidence that either SVs or MEIs might underlie the mutator phenotype described in this study.

## Strengths and limitations of the inter-haplotype distance approach

Our inter-haplotype distance (IHD) approach was able to identify a mutator allele that escaped notice using quantitative trait locus (QTL) mapping. To more systematically compare the power of IHD and QTL mapping, we performed simulations under a variety of possible parameter regimes. Overall, we found that IHD and QTL mapping have similar power to detect mutator alleles in biallelic recombinant inbred lines (RILs) that each harbor tens or hundreds of *de novo* germline mutations (Figure [1-figure supplement 2](#)). Nonetheless, only IHD was able to discover the mutator locus on chromosome 6 in the BXDs, demonstrating that it outperforms QTL mapping in certain experimental systems where different strains have been inbred for different lengths of time and harbor very different numbers of mutations that can be leveraged for mutator QTL mapping. Because the BXDs were generated in six breeding epochs over a period of nearly 40 years, the oldest lines have accumulated orders of magnitude more mutations than the youngest lines; these younger RILs also have much noisier mutation spectra as a result. While approaches for QTL mapping typically weight the phenotypic measurements of each sample equally, IHD compares the *aggregate* mutation spectra of haplotypes at every locus, a property that likely increased its power to detect mutators in the BXD dataset.

Another benefit of the IHD approach is that it obviates the need to perform separate association tests for every possible  $k$ -mer mutation type, and therefore the need to adjust significance thresholds for multiple tests. Since IHD compares the complete mutation spectrum between haplotypes that carry either allele at a site, it would also be well-powered to detect a mutator allele that exerted a coordinated effect on multiple  $k$ -mer mutation types (e.g., increased the rates of both C>T and C>A mutations).

However, the IHD method suffers a handful of drawbacks when compared to QTL mapping. Popular QTL mapping methods (such as R/qtL2 [47]) use linear models to test associations between genotypes and phenotypes, enabling the inclusion of additive and interactive covariates, as well as kinship matrices, in QTL scans. Although we have developed simple methods to account for inter-sample relatedness in the IHD approach, they are not as flexible as similar methods in QTL mapping software. Additionally, the IHD method assumes that mutator alleles affect a subset of  $k$ -mer mutation types; if a mutator allele increased the rates of all mutation types equally on haplotypes that carried it, IHD would be unable to detect it.

## Discovering mutator alleles in other experimental systems

Our discovery of a second BXD mutator allele underscores the power of recombinant inbred lines (RILs) as a resource for dissecting the genetic architecture of germline mutation rates. Large populations of RILs exist for many model organisms, and we anticipate that as whole-genome sequencing becomes cheaper and cheaper, the IHD method could be useful for future mutator allele discovery outside of the BXDs. At the same time, RILs are a finite resource that require enormous investments of time and labor to construct. If germline mutator alleles are only detectable in these highly unusual experimental populations, we are unlikely to discover more than a small fraction of the mutator alleles that may exist in nature.

Fortunately, the statistical approach introduced in this paper is readily adaptable to datasets beyond RILs. Thousands of human pedigrees have been sequenced in an effort to precisely estimate the rate of human *de novo* germline mutation [6,7,48], and as family sequencing has become a more common step in the diagnosis of many congenital disorders, these datasets are growing on a daily basis. Large cohorts of two- or three-generation families are an example of a regime in which IHD could enjoy high power; by pooling sparse mutation counts across many individuals who share the same candidate mutator allele, even a subtle mutator signal could rise above the noise of *de novo* germline mutation rate estimates.

Selection on germline mutator alleles will likely prevent large-effect mutators from reaching high allele frequencies; however, a subset may be detectable by sequencing a sufficient number of human trios [49]. Previous estimates of power to detect germline mutators in human pedigrees have assumed that mutators affect all mutation types equally, and that methods for mutator discovery will rely on identifying haplotypes with excess total mutation counts [49]. Our results in the BXDs suggest that power analysis based on total mutation counts have likely underestimated the power of pedigrees to map mutator alleles—since germline mutators seem to often exert their effects on a small number of mutation types, mutation spectrum analyses may often have the power to detect the genes that underlie heritable variation of the mutation rate, even if each gene only has a modest effect on the overall mutation rate per generation.

## 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

We developed a new approach to discover loci that affect the germline *de novo* mutation spectrum in biparental RILs (Figure 1).

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 a 6-element vector that contains the frequencies of C>T, C>G, C>A, A>G, A>T, and A>C mutations. Since C>T transitions at CpG nucleotides are often caused by a distinct mechanism (spontaneous deamination of methylated cytosine), we expand the 1-mer mutation spectrum to include a separate category for CpG>TpG mutations [50].

At each marker, we then compute the cosine distance between the two aggregate spectra. The cosine distance between two vectors **A** and **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 it adjusts for the magnitude of the two input vectors, cosine distance 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.

Inspired by methods from QTL mapping [47,51], 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 scan as described above, and record the maximum cosine distance observed at any locus. After  $N$  permutations (usually 10,000), we compute the  $1 - p$  percentile of the distribution of maximum statistics, 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-state 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  $D$  alleles are more genetically dissimilar from the RILs that inherited  $B$  alleles (considering all loci throughout the genome in our measurement of genetic similarity), we might expect the aggregate mutation spectra in the two groups to also be more dissimilar.

We implemented a simple approach to account for these potential issues of relatedness. At each marker  $g_i$ , 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 cosine distance between the two aggregate spectra ( $D_i^C$ ). 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.

Put another way, at every marker  $g_i$  along the genome, we divide BXD haplotypes into two groups and compute two metrics:  $D_i^C$  (the cosine distance 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 cosine distances, we regress  $(D_1^C, D_2^C, \dots, D_n^C)$  on  $(r_1, r_2, \dots, r_n)$  for all  $n$  markers using an ordinary least-squares model. We then use the residuals from the fitted model as the “adjusted” cosine distance values for each marker. If genome-wide genetic similarity between haplotypes perfectly predicts cosine distances 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 the observed cosine distance at a single marker and the marker-wide mean of cosine distances.

## 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` [52,53,54,55,56,57,58].

The code underlying IHD, as well as documentation of the method, is available on GitHub (<https://github.com/quinlan-lab/proj-mutator-mapping>). We have also deposited a reproducible Snakemake [59] workflow for running reproducing all analyses and figures presented in the manuscript.

## 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.

### Simulating genotypes

First, we simulate genotypes on a population of haplotypes at a collection of sites. We define a matrix  $G$  of size  $(s, h)$ , where  $s$  is the number of sites and  $h$  is the number of haplotypes. We assume that every site is biallelic, and that the minor allele frequency at every site is 0.5. For every entry  $G_{i,j}$ , we take a single draw from a uniform distribution in the interval  $[0.0, 1.0]$ . If the value of that draw is less than 0.5, we assign the value of  $G_{i,j}$  to be 1. Otherwise, we assign the value of  $G_{i,j}$  to be 0.

### Defining expected mutation type probabilities

Next, we define a vector of 1-mer mutation probabilities:

$$P = (0.29, 0.17, 0.12, 0.075, 0.1, 0.075, 0.17)$$

These probabilities sum to 1 and roughly correspond to the expected frequencies of C>T, CpG>TpG, C>A, C>G, A>T, A>C, and A>G *de novo* germline mutations in mice, respectively [10]. 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  $\frac{P_c}{16} = \frac{0.46}{16} = 0.02875$ . We then generate a vector of lambda values by scaling the mutation probabilities by the number of mutations we wish to simulate ( $m$ ):

$$\lambda = Pm$$

We also create a second vector of lambda values ( $\lambda'$ ), in which we multiply the  $\lambda$  value of a single mutation type by the mutator effect size  $e$ .

In our simulations, we assume that genotypes at a single site (the “mutator locus”) are associated with variation in the mutation spectrum. That is, at a single site  $s_i$ , all of the haplotypes with 1 alleles should have elevated rates of a particular mutation type and draw their mutation counts from  $\lambda'$ , while all of the haplotypes with 0 alleles should have “wild-type” rates of that mutation type and draw their mutation counts from  $\lambda$ . We therefore pick a random site  $s_i$  to be the “mutator locus,” and identify the indices of haplotypes in  $G$  that were assigned 1 alleles at  $s_i$ . We call these indices  $h_{mut}$ .

### Simulating mutation spectra

To simulate the mutation spectrum on our toy population of haplotypes, we define a matrix  $C$  of size  $(h, n)$ , where  $n = 6 \times 4^{k-1}$  (or if  $k = 1$  and we include CpG>TpG mutations,  $6 \times 4^{k-1} + 1$ ).

Then, we populate the matrix  $C$  separately for *mutator* and *wild-type* haplotypes. For every row  $i$  in the matrix (i.e., for every haplotype), we first ask if  $i$  is in  $h_{mut}$  (that is, if the haplotype at index  $i$  was assigned a 1 allele at the “mutator locus”). If so, we set the values of  $C_i$  to be the results of a single Poisson draw from  $\lambda'$ . If row  $i$  is not in  $h_{mut}$ , we set the values of  $C_i$  to be the results of a single Poisson draw from  $\lambda$ .

## Assessing power to detect a simulated mutator allele using IHD

For each combination of parameters (number of simulated haplotypes, number of simulated markers, mutator effect size, etc.), we run 100 independent trials. In each trial, we simulate the genotype matrix  $G$  and the mutation counts  $C$ . We calculate a “focal” cosine distance as the cosine distance between the aggregate mutation spectra of haplotypes with either genotype at  $s_i$  (the site at which we artificially simulated an association between genotypes and mutation spectrum variation). We then perform an inter-haplotype distance scan using  $N = 1,000$  permutations. If fewer than 5% of the  $N$  permutations produced a cosine distance greater than or equal to the focal distance, we say that the approach successfully identified the mutator allele in that trial.

## Assessing power to detect a simulated mutator allele using quantitative trait locus (QTL) mapping

Using simulated data, we also assessed the power of traditional quantitative trait locus (QTL) mapping to detect a locus associated with mutation spectrum variation. As described above, we simulated both genotype and mutation spectra for a population of haplotypes under various conditions (number of mutations per haplotype, mutator effect size, etc.). Using those simulated data, we used R/qtl2 [47] to perform a genome scan for significant QTL as follows; we assume that the simulated genotype markers are evenly spaced (in physical Mbp coordinates) on a single chromosome. First, we calculate the fraction of each haplotype’s *de novo* mutations that belong to each of the  $6 \times 4^{k-1}$  possible  $k$ -mer mutation types. We then convert the simulated genotypes at each marker to genotype probabilities using the `calc_genoprob` function in R/qtl2, with `map_function = "c-f"` and `error_prob = 0`. For every  $k$ -mer mutation type, we use genotype probabilities and per-haplotype mutation fractions to perform a scan for QTL with the `scan1` function; to make the results more comparable to those from the IHD method, we do not include any covariates or kinship matrices in these QTL scans. We then use the `scan1perm` function to perform 1,000 permutations of the per-haplotype mutation fractions and calculate log-odds (LOD) thresholds for significance. We consider the QTL scan to be “successful” if it produces a LOD score above the significance threshold (defined using  $\alpha = \frac{0.05}{7}$ ) for the marker at which we simulated an association with mutation spectrum variation.

**Note:** In our simulations, we augment the mutation rate of a single  $k$ -mer mutation type on haplotypes carrying the simulated mutator allele. However, in an experimental setting, we would not expect to have *a priori* knowledge of the mutation type affected by the mutator. Thus, by using an alpha threshold of 0.05, we would likely over-estimate the power of QTL mapping for detecting the mutator. Since we would need to perform 7 separate QTL scans (one for each 1-mer mutation type plus CpG>TpG) in an experimental setting, we calculate QTL LOD thresholds at a Bonferroni-corrected alpha value of  $\alpha = \frac{0.05}{7}$ .

## 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 [2,60,61], and downloaded a CSV file of BXD genotypes at ~7,300 informative markers from GeneNetwork [21,62].

We also downloaded relevant metadata about each BXD RIL from the manuscript describing the updated BXD resource [3]. These files are included in the GitHub repository associated with this manuscript.

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 [63]. 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 [64]. 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 93 BXD RILs in our genome-wide scans.

## Identifying candidate single-nucleotide 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 [65]) using `bcftools` [66]. We defined the candidate interval surrounding the cosine distance peak on chromosome 6 as +/- 5 Mbp from the genotype marker with the largest adjusted cosine distance 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` [67] command:

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

and used `cyvcf2` [68] to iterate over the annotated VCF file in order to identify nonsynonymous fixed differences between the parental C57BL/6J and DBA/2J strains.

## Identifying candidate structural variant alleles overlapping the chromosome 6 peak

We downloaded summary VCFs containing insertion, deletion and inversion structural variants (identified via high-quality, long-read assembly of inbred laboratory mouse strains [23]) from the Zenodo link associated with the Ferraj et al. manuscript: <https://doi.org/10.5281/zenodo.7644286>.

We then downloaded a TSV file containing RefSeq gene predictions in GRCm39/mm39 from the UCSC Table Browser [69], and used the `bx-python` library [70] to intersect the interval spanned by each structural variant with the intervals spanned by the `txStart` and `txEnd` of every RefSeq entry.

We queried all structural variants within a region +/- 5 Mbp from the adjusted cosine distance peak on chromosome 6 at marker `rs31001331`.

## Extracting mutation signatures

We used SigProfilerExtractor (v.1.1.21) [25] to extract mutation signatures from the BXD mutation data. After converting the BXD mutation data to the “matrix” input format expected by SigProfilerExtractor, we ran the `sigProfilerExtractor` method as follows:

```
# install the mm10 mouse reference data
genInstall.install('mm10')

# run mutation signature extraction
sig.sigProfilerExtractor(
  'matrix',
  /path/to/output/directory,
  /path/to/input/mutations,
  maximum_signatures=10,
  nmf_replicates=100,
  opportunity_genome="mm10",
)
```

## Comparing mutation spectra between Mouse Genomes Project strains

We downloaded mutation data from a previously published analysis [16] (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 [26]. 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 eQTLs at the mutator locus

We used the online GeneNetwork resource [21], which contains array- and RNA-seq-derived expression measurements in a wide variety of tissues, 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 2. In the **Get Any** text box, we then entered the listed gene name and clicked **Search**. After selecting the appropriate trait ID on the next page, we used the **Mapping Tools** dropdown to run Haley-Knott regression [71] 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

Tissue name	Complete name of GeneNetwork expression data	GeneNetwork trait ID
Hematopoietic stem cells	UMCG Stem Cells ILM6v1.1 (Apr09) transformed	ILM1940279
Hematopoietic 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 frequencies of the *Ogg1* p.Thr95Ala and *Mbd4* p.Asp129Asn mutations in other populations of mice, we queried a VCF file containing genome-wide variation in 67 wild-derived mice from four species of *Mus* [27]. 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 on chromosome 4 and chromosome 6, we modeled C>A mutation rates in the BXDs as a function of genotypes at either locus. Specifically, we tested for statistical interaction between 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/guanine nucleotides in each RIL (i.e., the total number of cytosines/guanines 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 cosine distances on chromosomes 4 and 6 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)) + Genotype_A + Genotype_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 [16]. 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 RIL 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 [72] in the R statistical language, we fit a mixed effects model predicting centered log ratio (CLR)-transformed 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(CLR_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` [47] 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, ]

# calculate QTL genotype probabilities
pr <- calc_genoprob(bxd, bxd$pmap, error_prob = 0, map_function = "c-f")

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

kinship_matrix = as.matrix(k)

```

The interaction term (between genotypes at the chromosome 4 and 6 loci) in the `lmekin` model returned a p-value of 0.029.

## Supplementary information

---

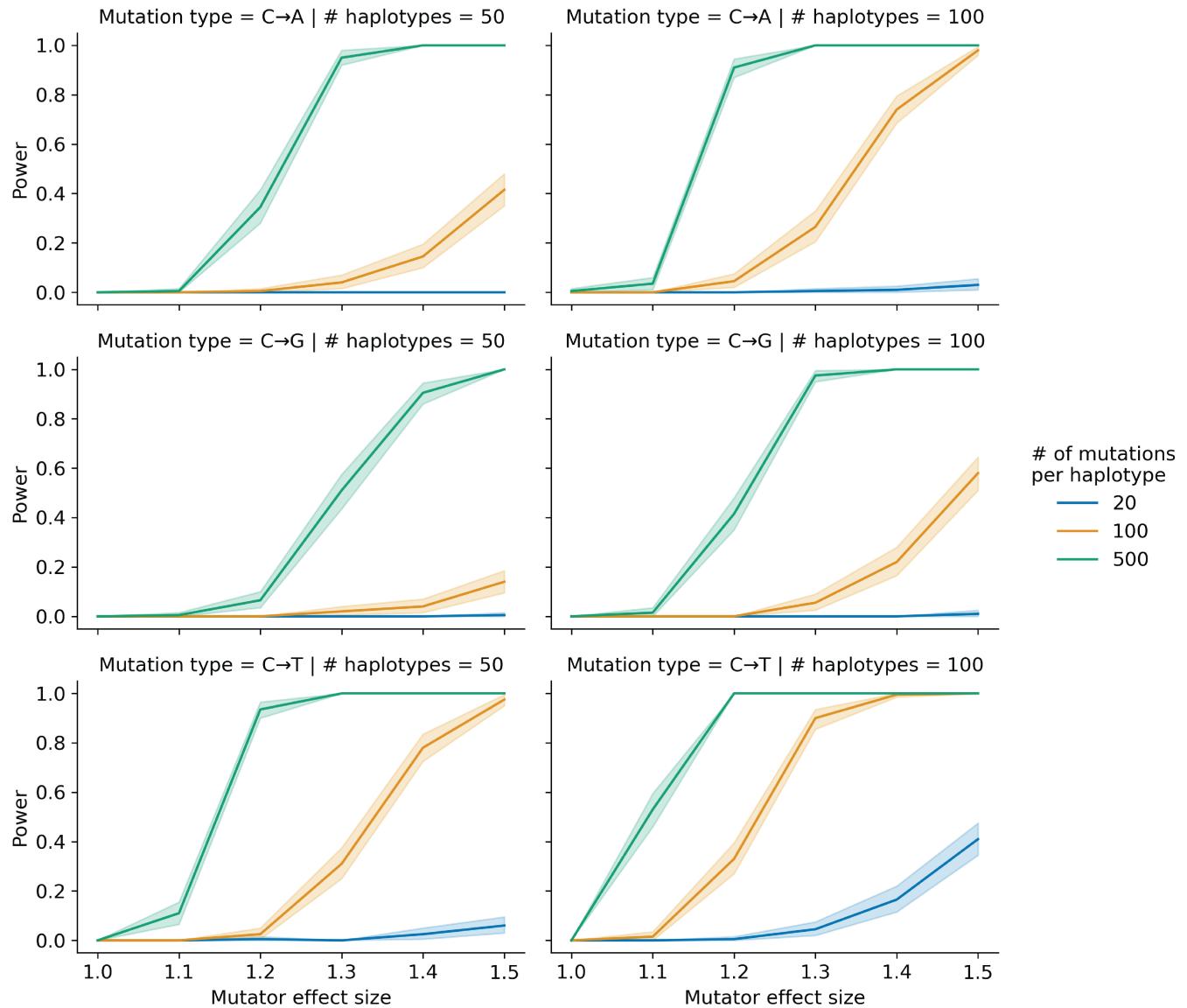
### Chi-square statistics as an alternative to cosine distance in the inter-haplotype distance method

Although the current implementation of inter-haplotype distance (IHD) uses the cosine distance metric to compare aggregate mutation spectra at biallelic markers, we also explored the use of chi-square statistics to compare spectra.

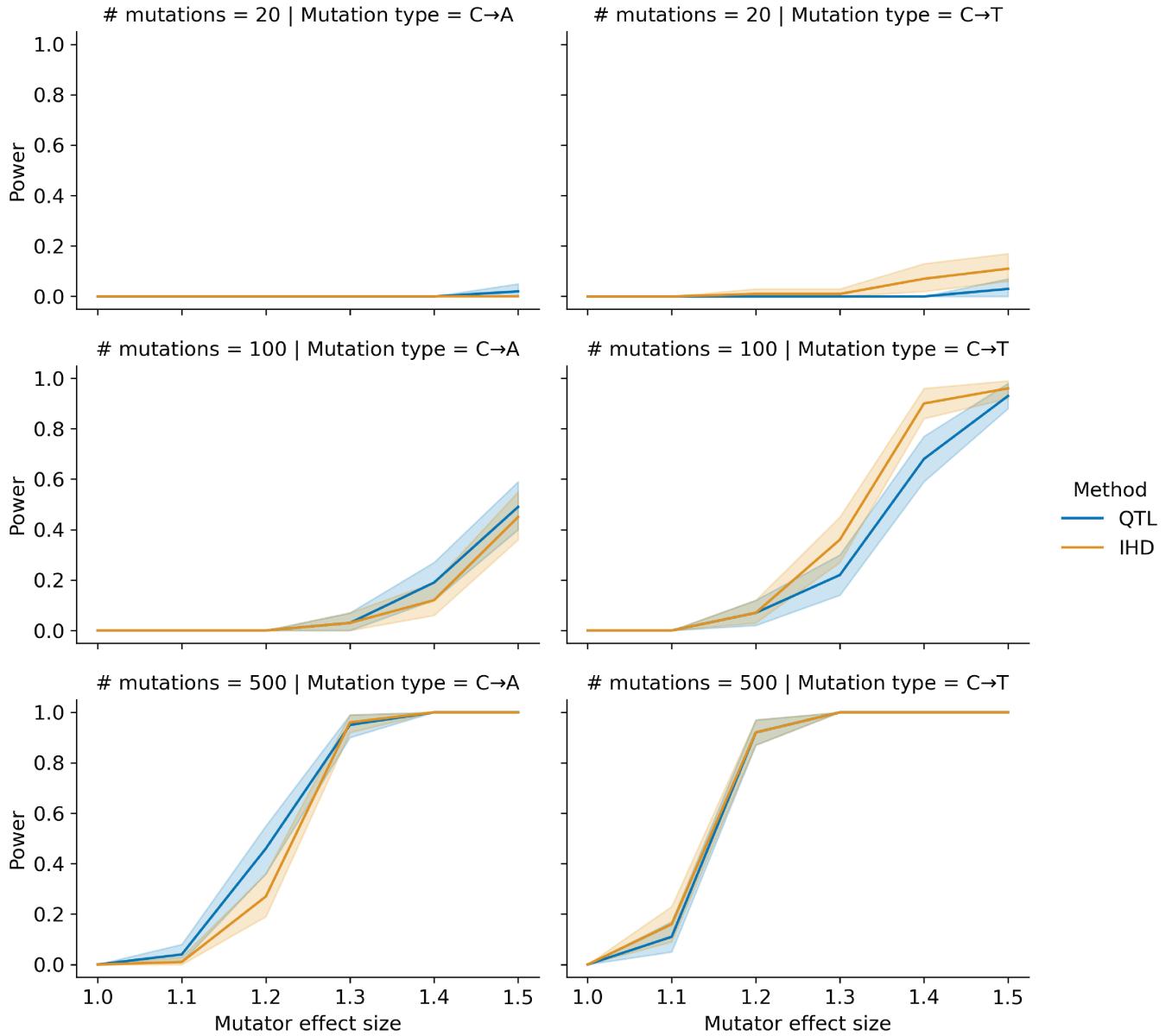
As described in the *Materials and Methods*, we compute aggregate mutation spectra on haplotypes that inherited either parental allele at each biallelic marker. Rather than compute the cosine distance between those spectra, we can create a contingency table of size  $(2, 6 \times 4^{k-1})$  (or size  $(2, 6 \times 4^{k-1} + 1)$  if  $k = 1$  and we're including CpG>TpG) that contains the aggregate  $k$ -mer mutation spectrum on haplotypes that inherited either parental allele. Using the contingency table, we then calculate a  $\chi^2$  test statistic; larger values of the  $\chi^2$  statistic suggest more “distance” between the aggregate mutation spectra on the two haplotypes.

We discovered the BXD mutator loci on chromosome 4 and 6 using either cosine distances or chi-square statistics to compare mutation spectra; however, using simulated data we found that the cosine distance metric was more sensitive for mutator allele detection across a broad range of parameter choices (Figure [1-figure supplement 3](#)).

## Supplementary Figures

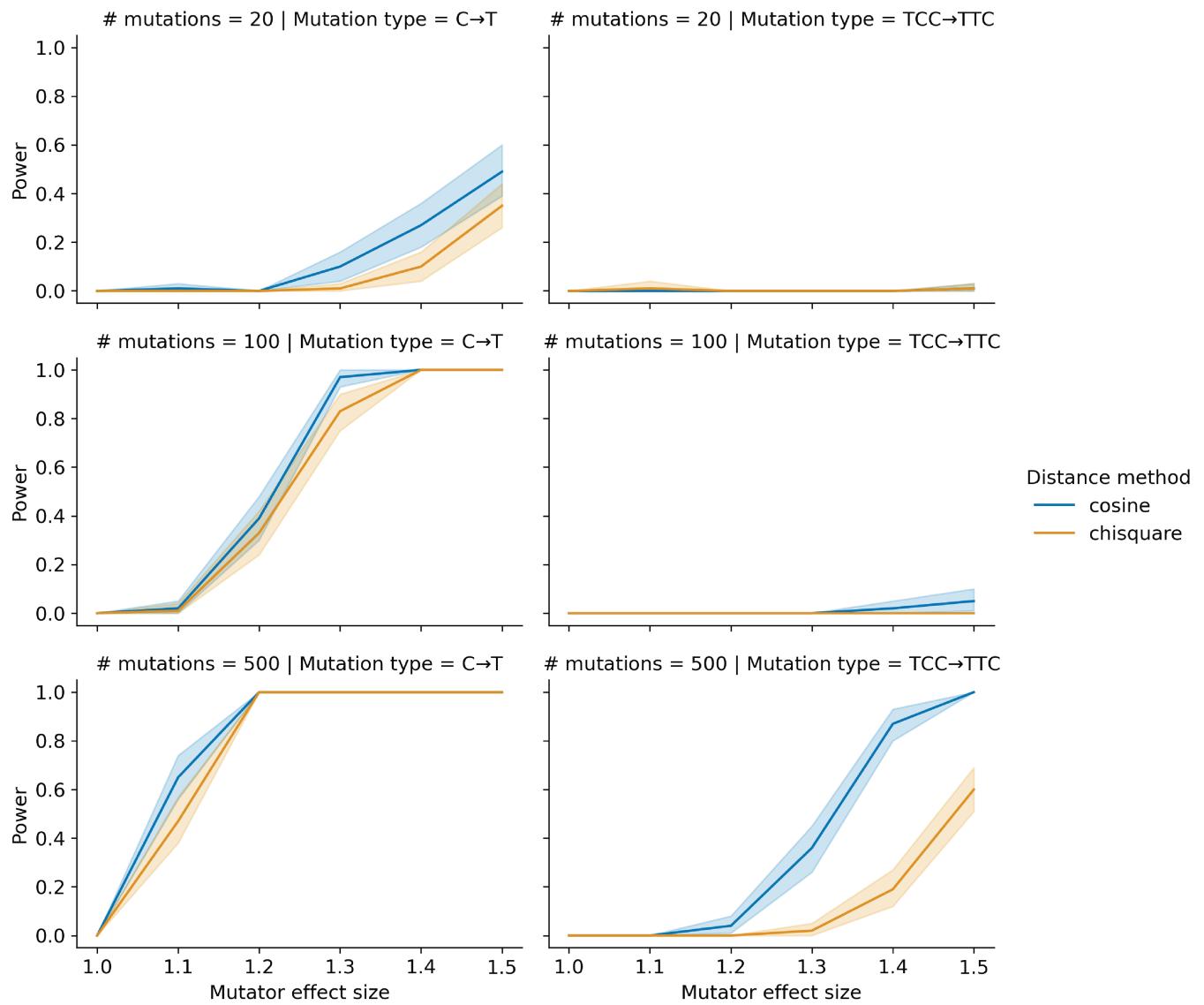


**Figure 1-figure supplement 1: Simulations to assess the power of the inter-haplotype distance method.** In each of 100 trials, we simulated genotypes at 1,000 biallelic loci on a toy population of either 50 or 100 haplotypes as follows. At every locus on every haplotype, we drew a single floating point value from a uniform distribution  $[0, 1]$ . If that value was less than or equal to 0.5, we set the allele to be "A"; otherwise, we set the allele to be "B". In each trial, we also simulated *de novo* germline mutations on the population of haplotypes, such that at a single locus  $g_i$ , we augmented the mutation rate of a particular  $k$ -mer by the specified effect size (an effect size of 1.5 indicates a 50% increase in the mutation rate) on haplotypes carrying "A" alleles. We then applied the inter-haplotype distance method to these simulated data and asked if the adjusted cosine distance at locus  $g_i$  was greater than expected by chance. Given a specific combination of parameters, the y-axis denotes the fraction of 100 trials in which the simulated mutator allele could be detected at a significance threshold of  $p = 0.05$ . Shaded areas indicate the standard deviation of that fraction across 100 simulations.

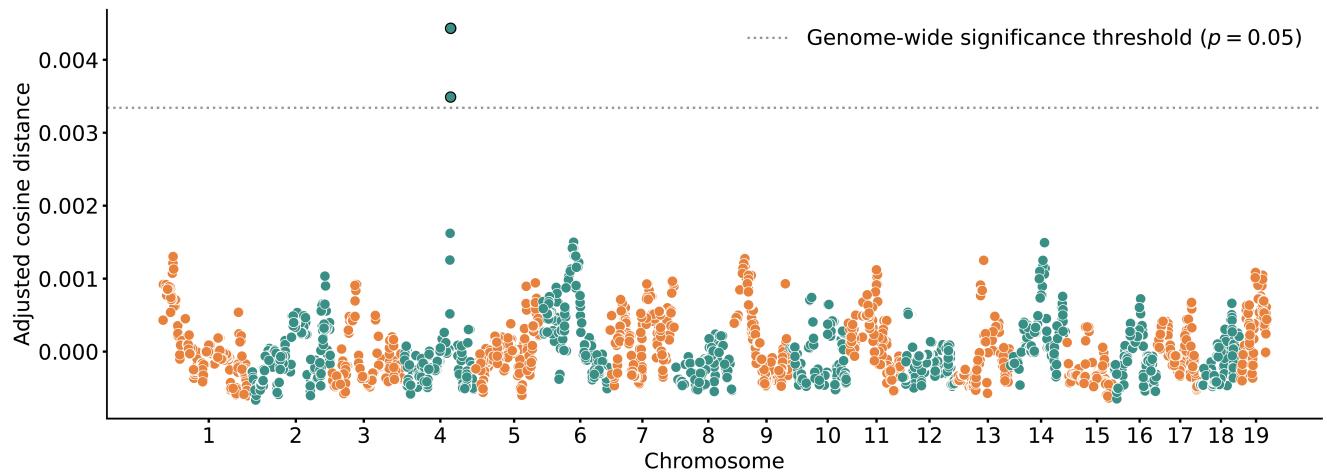


**Figure 1-figure supplement 2: Comparing power between the inter-haplotype distance method and QTL mapping.**

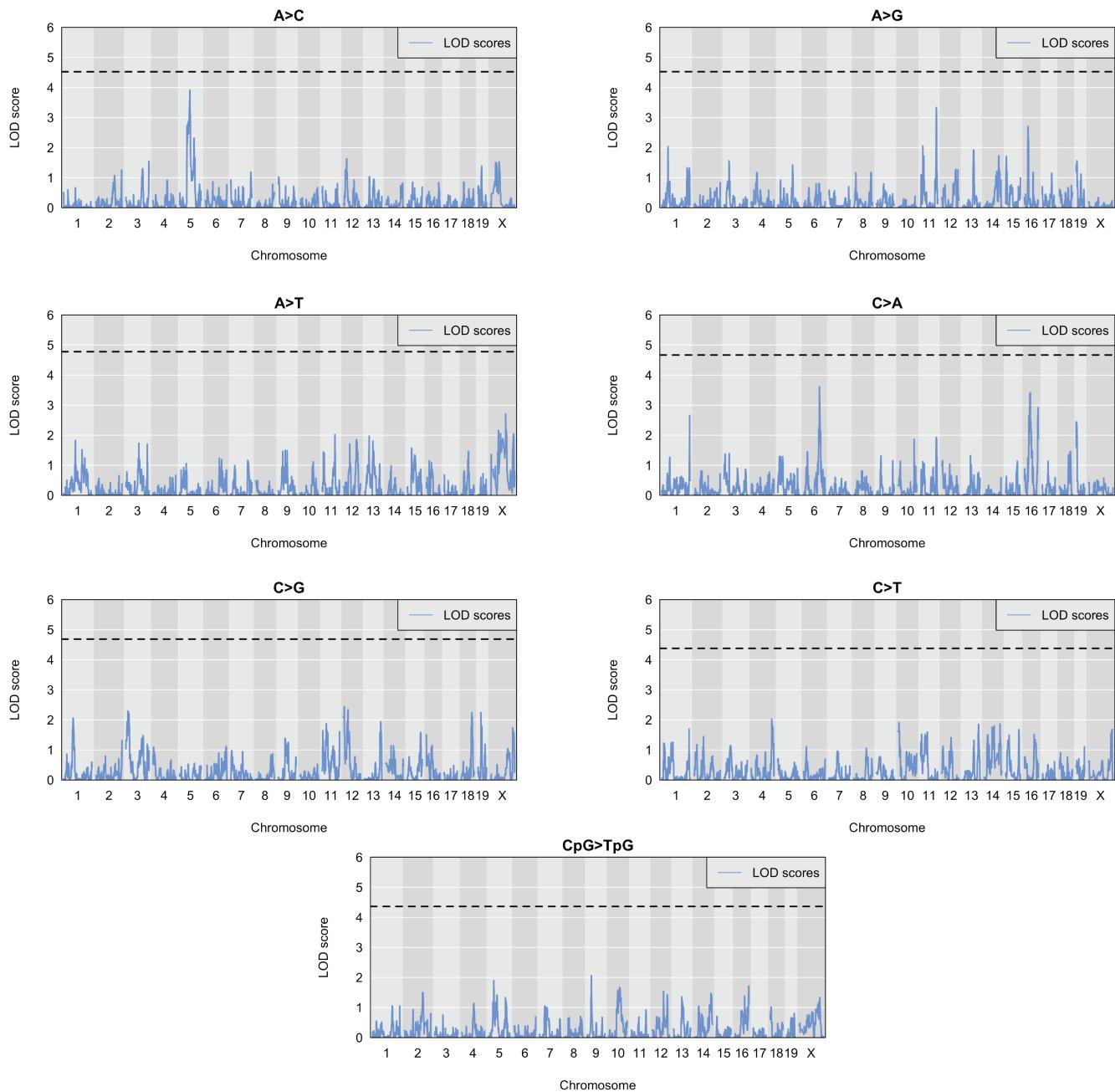
**mapping.** In each of 100 trials, we simulated genotypes at 1,000 biallelic loci on a toy population of 50 haplotypes as follows. At every locus on every haplotype, we drew a single floating point value from a uniform distribution  $[0, 1]$ . If that value was less than or equal to 0.5, we set the allele to be "A"; otherwise, we set the allele to be "B". In each trial, we also simulated *de novo* germline mutations on the population of haplotypes, such that at a single locus  $g_i$ , we augmented the rate of the specified mutation type by the specified effect size (an effect size of 1.5 indicates a 50% increase in the mutation rate) on haplotypes carrying "A" alleles. We then applied the inter-haplotype distance method to these simulated data and asked if the adjusted cosine distance at locus  $g_i$  was greater than expected by chance. Similarly, in each trial, we used R/qtL2 to perform a genome scan for QTL and asked if the log-odds score at  $g_i$  was greater than expected by chance. Given a specific combination of parameters, the y-axis denotes the fraction of 100 trials in which the simulated mutator allele could be detected at a significance threshold of  $p = 0.05$  (for IHD) or at an alpha of  $\frac{0.05}{7}$  (for QTL mapping). Shaded areas indicate the standard deviation of that fraction across 100 simulations.



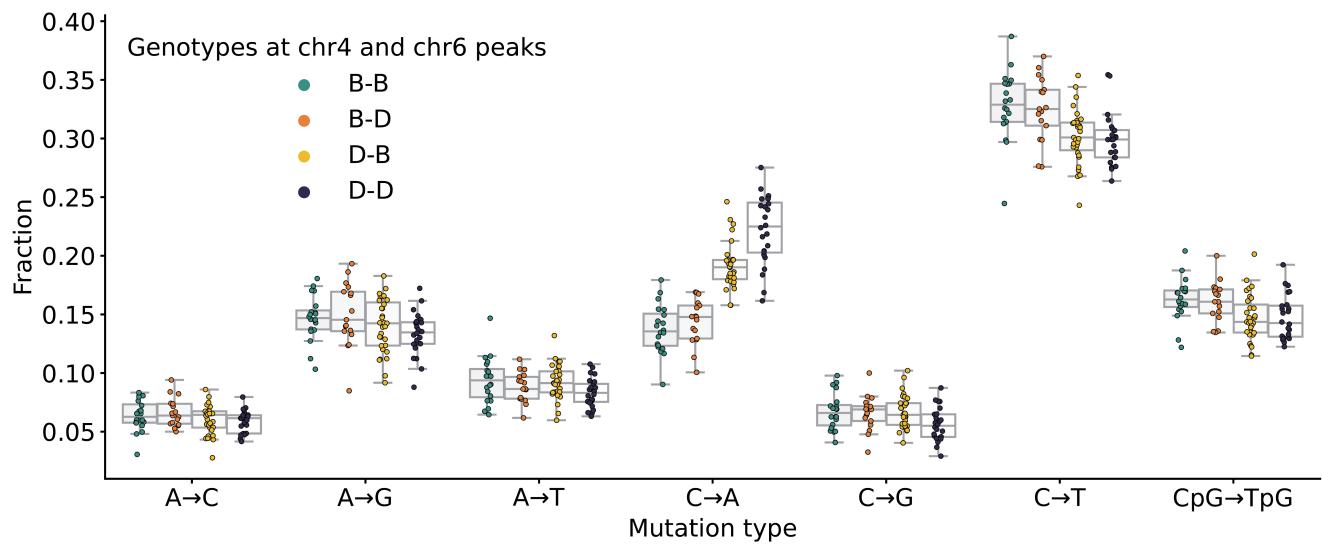
**Figure 1-figure supplement 3: Comparing the power of cosine distance and chi-square statistics in the inter-haplotype distance method.** In each of 100 trials, we simulated genotypes at 1,000 biallelic loci on a toy population of 50 haplotypes as follows. At every locus on every haplotype, we drew a single floating point value from a uniform distribution  $[0, 1]$ . If that value was less than or equal to 0.5, we set the allele to be "A"; otherwise, we set the allele to be "B". In each trial, we also simulated *de novo* germline mutations on the population of haplotypes, such that at a single locus  $g_i$ , we augmented the rate of the specified mutation type by the specified effect size (an effect size of 1.5 indicates a 50% increase in the mutation rate) on haplotypes carrying "A" alleles. We then applied the inter-haplotype distance method, using either cosine distance or chi-square statistics to compare aggregate mutation spectra at each marker. In each trial, we asked if the adjusted cosine distance or chi-square statistic at locus  $g_i$  was greater than expected by chance. Given a specific combination of parameters, the y-axis denotes the fraction of 100 trials in which the simulated mutator allele could be detected at a significance threshold of  $p = 0.05$ . Shaded areas indicate the standard deviation of that fraction across 100 simulations.



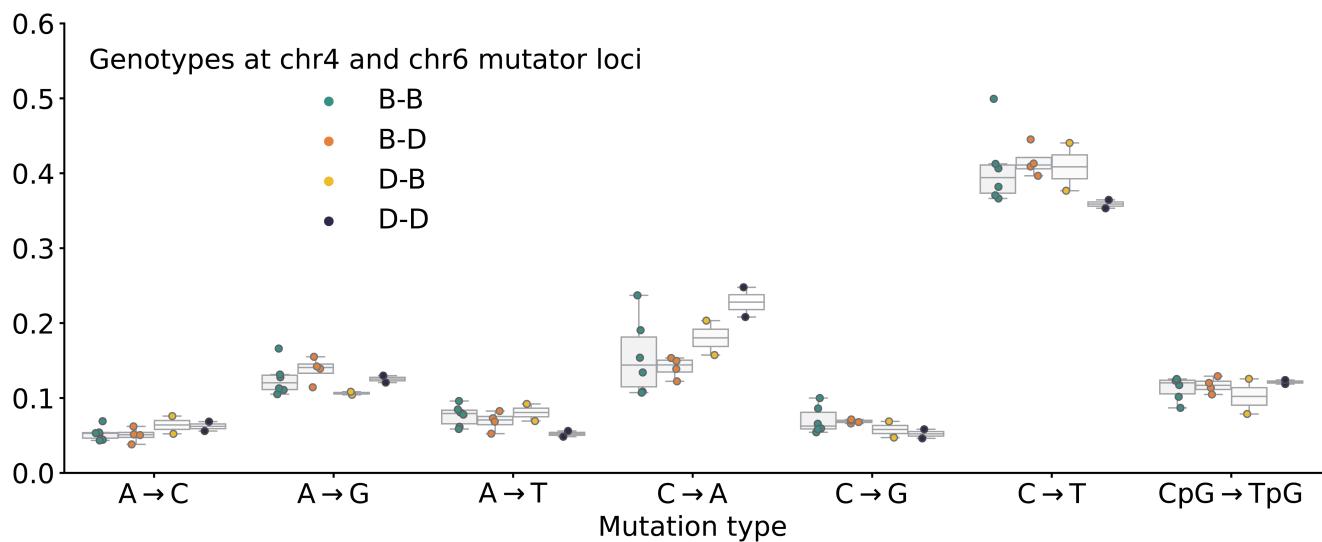
**Figure 2-figure supplement 1: Results of inter-haplotype distance scans using BXDs with *B* alleles at the chromosome 4 locus.** Adjusted cosine distances between aggregate 1-mer *de novo* mutation spectra on BXD haplotypes with *B* alleles at rs52263933 (n = 38 haplotypes; 22,080 total mutations) and either *D* or *B* alleles at 7,278 informative markers. Cosine 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.



**Figure 2-figure supplement 2: Quantitative trait locus scans for mutation spectrum phenotypes.** Using the BXDs with D genotypes at rs52263933 (the marker with the highest cosine distance on chromosome 4; n = 55 RILs, 40,913 total mutations), we used R/qtl2 to perform QTL scans for the centered log-ratio (CLR) transformed fraction of each 1-mer mutation type. QTL scans also included a kinship matrix (that contained the pairwise genetic similarity between each pair of BXDs, calculated using the leave-one-chromosome-out method) as a random effect term using the `kinship` keyword argument in the `scan1` function. Plots show the log-odds (LOD) score at every genotyped marker in blue; the dotted black line represents the genome-wide LOD significance threshold (established using 1,000 permutations at an alpha of  $\frac{0.05}{7}$  to account for the fact that 7 separate association tests were performed.)



**Figure 3-figure supplement 1: Mutation spectra comparison in BXD strains.** Fractions of *de novo* germline mutations in BXDs with either *D* or *B* genotypes at markers rs52263933 and rs31001331, stratified by mutation type.



**Figure 3—figure supplement 2: Mutation spectra comparison in Sanger Mouse Genomes Project strains.** Fractions of *de novo* germline mutations in Sanger MGP strains with either *D* or *B* haplotypes at the chromosome 4 and chromosome 6 mutator loci, stratified by mutation type.

## Supplementary Tables

**Table supplement 1:** Presence or absence of cis-eQTLs for *Ogg1* and *Mbd4* in various tissues identified using GeneNetwork.

Gene name	Tissue name	# BXDs with expression data	Top significant marker	LRS at top significant marker	Significant LRS threshold
<i>Ogg1</i>	Kidney	53	rsm100000041 88	52.25	17.82
<i>Ogg1</i>	Gastrointestinal	46	rsm100000034 41	23.39	16.09

Gene name	Tissue name	# BXDs with expression data	Top significant marker	LRS at top significant marker	Significant LRS threshold
<i>Ogg1</i>	Hematopoietic stem cells	22	-	-	16.45
<i>Ogg1</i>	Hematopoietic progenitor cells	23	-	-	18.52
<i>Ogg1</i>	Spleen	79	rsm100000034 18	-	17.51
<i>Ogg1</i>	Liver	50	rsm100000041 88	53.54	18.77
<i>Ogg1</i>	Heart	73	-	-	16.22
<i>Ogg1</i>	Eye	87	rsm100000041 94	23.05	16.96
<i>Mbd4</i>	Kidney	53	-	-	18.48
<i>Mbd4</i>	Gastrointestinal	46	-	-	15.97
<i>Mbd4</i>	Hematopoietic stem cells	22	-	-	18.12
<i>Mbd4</i>	Hematopoietic progenitor cells	23	-	-	18.55
<i>Mbd4</i>	Spleen	79	rsm100000041 99	21.42	16.99
<i>Mbd4</i>	Liver	50	-	-	16.15
<i>Mbd4</i>	Heart	73	-	-	17.17
<i>Mbd4</i>	Eye	87	-	-	16.63

**Table supplement 2:** Large structural variants overlapping protein-coding genes in the mutator locus on chromosome 6 All coordinates are with respect to GRCm39/mm39.

SV start	SV end	SV type	Gene name(s)	Overlaps exon?
108,437,600	108,437,652	DEL	<i>Gm35165, Itpr1</i>	Yes
110,617,754	110,618,347	DEL	<i>Gm20387</i>	Yes
113,110,214	113,110,215	INS	<i>Setd5</i>	Yes
116,217,366	116,217,367	INS	<i>Gm52873, Washc2</i>	Yes
116,323,201	116,323,202	INS	<i>Marchf8</i>	Yes
116,662,107	116,668,530	DEL	<i>Tmem72</i>	Yes
116,976,917	116,976,984	DEL	<i>LOC118567448</i>	Yes
112,605,594	112,605,595	INS	<i>Rad18</i>	No
112,629,618	112,636,619	DEL	<i>Rad18</i>	No
115,823,860	115,824,076	DEL	<i>Mbd4</i>	No

# 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](https://doi.org/10.1002/em.22087) · PMID: [28485537](https://pubmed.ncbi.nlm.nih.gov/28485537/) · PMCID: [PMC5474181](https://pmc.ncbi.nlm.nih.gov/pmc/articles/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](https://doi.org/10.1038/s41586-022-04701-5) · PMID: [35545679](https://pubmed.ncbi.nlm.nih.gov/35545679/) · PMCID: [PMC9272728](https://pmc.ncbi.nlm.nih.gov/pmc/articles/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](https://doi.org/10.1016/j.cels.2020.12.002) · PMID: [33472028](https://pubmed.ncbi.nlm.nih.gov/33472028/) · PMCID: [PMC7979527](https://pmc.ncbi.nlm.nih.gov/pmc/articles/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](https://doi.org/10.1038/nature05978) · PMID: [17581577](https://pubmed.ncbi.nlm.nih.gov/17581577/) · PMCID: [PMC2896554](https://pmc.ncbi.nlm.nih.gov/pmc/articles/PMC2896554/)
5. **MED1, a novel human methyl-CpG-binding endonuclease, interacts with DNA mismatch repair protein MLH1.**  
A Bellacosa, L Cicchillitti, F Schepis, A Riccio, AT Yeung, Y Matsumoto, EA Golemis, M Genuardi, G Neri  
*Proceedings of the National Academy of Sciences of the United States of America* (1999-03-30)  
<https://www.ncbi.nlm.nih.gov/pubmed/10097147>  
DOI: [10.1073/pnas.96.7.3969](https://doi.org/10.1073/pnas.96.7.3969) · PMID: [10097147](https://pubmed.ncbi.nlm.nih.gov/10097147/) · PMCID: [PMC22404](https://pmc.ncbi.nlm.nih.gov/pmc/articles/PMC22404/)
6. **Parental influence on human germline de novo mutations in 1,548 trios from Iceland.**  
Hákon Jónsson, Patrick Sulem, Birte Kehr, Snaedis Kristmundsdóttir, Florian Zink, Eiríkur Hjartarson, Marteinn T Hardarson, Kristjan E Hjorleifsson, Hannes P Eggertsson, Sigurður Axel Gudjonsson, ... Kari Stefansson  
*Nature* (2017-09-20) <https://www.ncbi.nlm.nih.gov/pubmed/28959963>  
DOI: [10.1038/nature24018](https://doi.org/10.1038/nature24018) · PMID: [28959963](https://pubmed.ncbi.nlm.nih.gov/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](https://doi.org/10.7554/elife.46922) · PMID: [31549960](https://pubmed.ncbi.nlm.nih.gov/31549960/) · PMCID: [PMC6759356](https://pmc.ncbi.nlm.nih.gov/pmc/articles/PMC6759356/)
8. **De novo Mutations in Domestic Cat are Consistent with an Effect of Reproductive Longevity on Both the Rate and Spectrum of Mutations.**  
Richard J Wang, Muthuswamy Raveendran, RAlan Harris, William J Murphy, Leslie A Lyons, Jeffrey Rogers, Matthew W Hahn  
*Molecular biology and evolution* (2022-07-02) <https://www.ncbi.nlm.nih.gov/pubmed/35771663>  
DOI: [10.1093/molbev/msac147](https://doi.org/10.1093/molbev/msac147) · PMID: [35771663](https://pubmed.ncbi.nlm.nih.gov/35771663/) · PMCID: [PMC9290555](https://pmc.ncbi.nlm.nih.gov/pmc/articles/PMC9290555/)

9. **A comparison of humans and baboons suggests germline mutation rates do not track cell divisions.**  
Felix L Wu, Alva I Strand, Laura A Cox, Carole Ober, Jeffrey D Wall, Priya Moorjani, Molly Przeworski  
*PLoS biology* (2020-08-17) <https://www.ncbi.nlm.nih.gov/pubmed/32804933>  
DOI: [10.1371/journal.pbio.3000838](https://doi.org/10.1371/journal.pbio.3000838) · PMID: [32804933](#) · PMCID: [PMC7467331](#)
10. **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](https://doi.org/10.1038/s41467-019-12023-w) · PMID: [31492841](#) · PMCID: [PMC6731245](#)
11. **Timing, rates and spectra of human germline mutation.**  
Raheleh Rahbari, Arthur Wuster, Sarah J Lindsay, Robert J Hardwick, Ludmil B Alexandrov, Saeed Al Turki, Anna Dominiczak, Andrew Morris, David Porteous, Blair Smith, ... Matthew E Hurles  
*Nature genetics* (2015-12-14) <https://www.ncbi.nlm.nih.gov/pubmed/26656846>  
DOI: [10.1038/ng.3469](https://doi.org/10.1038/ng.3469) · PMID: [26656846](#) · PMCID: [PMC4731925](#)
12. **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](https://doi.org/10.1038/nrg.2016.104) · PMID: [27739533](#)
13. **Genetic and chemotherapeutic influences on germline hypermutation.**  
Joanna Kaplanis, Benjamin Ide, Rakesh Sanghvi, Matthew Neville, Petr Danecek, Tim Coorens, Elena Prigmore, Patrick Short, Giuseppe Gallone, Jeremy McRae, ... Matthew Hurles  
*Nature* (2022-05-11) <https://www.ncbi.nlm.nih.gov/pubmed/35545669>  
DOI: [10.1038/s41586-022-04712-2](https://doi.org/10.1038/s41586-022-04712-2) · PMID: [35545669](#) · PMCID: [PMC9117138](#)
14. **Increased somatic mutation burdens in normal human cells due to defective DNA polymerases.**  
Philip S Robinson, Tim HH Coorens, Claire Palles, Emily Mitchell, Federico Abascal, Sigurgeir Olafsson, Bernard CH Lee, Andrew RJ Lawson, Henry Lee-Six, Luiza Moore, ... Michael R Stratton  
*Nature genetics* (2021-09-30) <https://www.ncbi.nlm.nih.gov/pubmed/34594041>  
DOI: [10.1038/s41588-021-00930-y](https://doi.org/10.1038/s41588-021-00930-y) · PMID: [34594041](#) · PMCID: [PMC8492474](#)
15. **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](https://doi.org/10.1371/journal.pgen.1006549) · PMID: [28095480](#) · PMCID: [PMC5283766](#)
16. **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](https://doi.org/10.1093/molbev/msz026) · PMID: [30753674](#) · PMCID: [PMC6501876](#)
17. **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](https://doi.org/10.1146/annurev.ecolsys.39.110707.173437)

18. **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](https://doi.org/10.1016/j.gde.2020.05.024) · PMID: [32619789](https://pubmed.ncbi.nlm.nih.gov/32619789/) · PMCID: [PMC7646088](https://pubmed.ncbi.nlm.nih.gov/PMC7646088/)
19. **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](https://doi.org/10.1016/j.ebiom.2017.04.022) · PMID: [28551381](https://pubmed.ncbi.nlm.nih.gov/28551381/) · PMCID: [PMC5478212](https://pubmed.ncbi.nlm.nih.gov/PMC5478212/)
20. **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](https://doi.org/10.1002/path.4880) · PMID: [28127763](https://pubmed.ncbi.nlm.nih.gov/28127763/)
21. **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](https://doi.org/10.1007/978-1-4939-6427-7_4) · PMID: [27933521](https://pubmed.ncbi.nlm.nih.gov/27933521/) · PMCID: [PMC7495243](https://pubmed.ncbi.nlm.nih.gov/PMC7495243/)
22. **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](https://doi.org/10.1016/j.stem.2020.07.005) · PMID: [32795400](https://pubmed.ncbi.nlm.nih.gov/32795400/) · PMCID: [PMC7484384](https://pubmed.ncbi.nlm.nih.gov/PMC7484384/)
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  
*Cell Genomics* (2023-04) <https://doi.org/gr3t7b>  
DOI: [10.1016/j.xgen.2023.100291](https://doi.org/10.1016/j.xgen.2023.100291)
24. **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](https://doi.org/10.1016/j.xgen.2022.100179) · PMID: [36388765](https://pubmed.ncbi.nlm.nih.gov/36388765/) · PMCID: [PMC9646490](https://pubmed.ncbi.nlm.nih.gov/PMC9646490/)
25. **COSMIC: the Catalogue Of Somatic Mutations In Cancer.**  
John G Tate, Sally Bamford, Harry C Jubb, Zbyslaw Sondka, David M Beare, Nidhi Bindal, Harry Boutselakis, Charlotte G Cole, Celestino Creatore, Elisabeth Dawson, ... Simon A Forbes  
*Nucleic acids research* (2019-01-08) <https://www.ncbi.nlm.nih.gov/pubmed/30371878>  
DOI: [10.1093/nar/gky1015](https://doi.org/10.1093/nar/gky1015) · PMID: [30371878](https://pubmed.ncbi.nlm.nih.gov/30371878/) · PMCID: [PMC6323903](https://pubmed.ncbi.nlm.nih.gov/PMC6323903/)

26. **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](https://doi.org/10.1038/nature10413) · PMID: [21921910](https://pubmed.ncbi.nlm.nih.gov/21921910/) · PMCID: [PMC3276836](https://pmc.ncbi.nlm.nih.gov/pmcid/PMC3276836/)
27. **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](https://doi.org/10.1038/sdata.2016.75) · PMID: [27622383](https://pubmed.ncbi.nlm.nih.gov/27622383/) · PMCID: [PMC5020872](https://pmc.ncbi.nlm.nih.gov/pmcid/PMC5020872/)
28. **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](https://doi.org/10.1038/ng2087) · PMID: [17660819](https://pubmed.ncbi.nlm.nih.gov/17660819/)
29. **Negative epistasis between natural variants of the *Saccharomyces cerevisiae* MLH1 and PMS1 genes results in a defect in mismatch repair.**  
Julie Akiko Heck, Juan Lucas Argueso, Zekeriyya Gemici, Richard Guy Reeves, Ann Bernard, Charles F Aquadro, Eric Alani  
*Proceedings of the National Academy of Sciences of the United States of America* (2006-02-21) <https://www.ncbi.nlm.nih.gov/pubmed/16492773>  
DOI: [10.1073/pnas.0510998103](https://doi.org/10.1073/pnas.0510998103) · PMID: [16492773](https://pubmed.ncbi.nlm.nih.gov/16492773/) · PMCID: [PMC1413905](https://pmc.ncbi.nlm.nih.gov/pmcid/PMC1413905/)
30. **Mechanisms of APOBEC3 mutagenesis in human cancer cells.**  
Mia Petljak, Alexandra Dananberg, Kevan Chu, Erik N Bergstrom, Josefine Striepen, Patrick von Morgen, Yanyang Chen, Hina Shah, Julian E Sale, Ludmil B Alexandrov, ... John Maciejowski  
*Nature* (2022-07-20) <https://www.ncbi.nlm.nih.gov/pubmed/35859169>  
DOI: [10.1038/s41586-022-04972-y](https://doi.org/10.1038/s41586-022-04972-y) · PMID: [35859169](https://pubmed.ncbi.nlm.nih.gov/35859169/) · PMCID: [PMC9329121](https://pmc.ncbi.nlm.nih.gov/pmcid/PMC9329121/)
31. **A naturally occurring variant of <math>\text{MBD4}</math> causes maternal germline hypermutation in primates**  
Alexandra M Stendahl, Rashed Sanghvi, Samuel Peterson, Karina Ray, Ana C Lima, Raheleh Rahbari, Donald F Conrad  
*Cold Spring Harbor Laboratory* (2023-03-29) <https://doi.org/gr3ghz>  
DOI: [10.1101/2023.03.27.534460](https://doi.org/10.1101/2023.03.27.534460)
32. **Evolution of the germline mutation rate across vertebrates.**  
Lucie A Bergeron, Søren Besenbacher, Jiao Zheng, Panyi Li, Mads Frost Bertelsen, Benoit Quintard, Joseph I Hoffman, Zhipeng Li, Judy St Leger, Changwei Shao, ... Guojie Zhang  
*Nature* (2023-03-01) <https://www.ncbi.nlm.nih.gov/pubmed/36859541>  
DOI: [10.1038/s41586-023-05752-y](https://doi.org/10.1038/s41586-023-05752-y) · PMID: [36859541](https://pubmed.ncbi.nlm.nih.gov/36859541/) · PMCID: [PMC9995274](https://pmc.ncbi.nlm.nih.gov/pmcid/PMC9995274/)
33. **Novel mutations of OGG1 base excision repair pathway gene in laryngeal cancer patients.**  
Ishrat Mahjabeen, Nosheen Masood, Ruqia Mahmood 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](https://doi.org/10.1007/s10689-012-9554-2) · PMID: [22829015](https://pubmed.ncbi.nlm.nih.gov/22829015/)
34. **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](https://doi.org/10.1038/sj.onc.1202096) · PMID: [9662341](#)

35. **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](https://doi.org/10.1073/pnas.2007898118) · PMID: [34479993](#) · PMCID: [PMC8433536](#)

36. **Germline MBD4 deficiency causes a multi-tumor predisposition syndrome.**

Claire Palles, Hannah D West, Edward Chew, Sara Galavotti, Christoffer Flensburg, Judith E Grolleman, Erik AM Jansen, Helen Curley, Laura Chegwidden, Edward H Arbe-Barnes, ... Richarda M de Voer

*American journal of human genetics* (2022-04-22)  
<https://www.ncbi.nlm.nih.gov/pubmed/35460607>  
DOI: [10.1016/j.ajhg.2022.03.018](https://doi.org/10.1016/j.ajhg.2022.03.018) · PMID: [35460607](#) · PMCID: [PMC9118112](#)

37. **Enhanced CpG mutability and tumorigenesis in MBD4-deficient mice.**

Catherine B Millar, Jacky Guy, Owen J Sansom, Jim Selfridge, Eilidh MacDougall, Brian Hendrich, Peter D Keightley, Stefan M Bishop, Alan R Clarke, Adrian Bird  
*Science (New York, N.Y.)* (2002-07-19) <https://www.ncbi.nlm.nih.gov/pubmed/12130785>  
DOI: [10.1126/science.1073354](https://doi.org/10.1126/science.1073354) · PMID: [12130785](#)

38. **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](https://doi.org/10.1016/j.dnarep.2017.09.007) · PMID: [28963982](#)

39. **8-oxoguanine causes spontaneous de novo germline mutations in mice.**

Mizuki Ohno, Kunihiko Sakumi, Ryutaro Fukumura, Masato Furuichi, Yuki Iwasaki, Masaaki Hokama, Toshimichi Ikemura, Teruhisa Tsuzuki, Yoichi Gondo, Yusaku Nakabeppu  
*Scientific reports* (2014-04-15) <https://www.ncbi.nlm.nih.gov/pubmed/24732879>  
DOI: [10.1038/srep04689](https://doi.org/10.1038/srep04689) · PMID: [24732879](#) · PMCID: [PMC3986730](#)

40. **Cloning and expression of cDNA for a human enzyme that hydrolyzes 8-oxo-dGTP, a mutagenic substrate for DNA synthesis.**

K Sakumi, M Furuichi, T Tsuzuki, T Kakuma, S Kawabata, H Maki, M Sekiguchi  
*The Journal of biological chemistry* (1993-11-05)  
<https://www.ncbi.nlm.nih.gov/pubmed/8226881>  
PMID: [8226881](#)

41. **SIFT: Predicting amino acid changes that affect protein function.**

Pauline C Ng, Steven Henikoff  
*Nucleic acids research* (2003-07-01) <https://www.ncbi.nlm.nih.gov/pubmed/12824425>  
DOI: [10.1093/nar/gkg509](https://doi.org/10.1093/nar/gkg509) · PMID: [12824425](#) · PMCID: [PMC168916](#)

42. **The mutational constraint spectrum quantified from variation in 141,456 humans.**

Konrad J Karczewski, Laurent C Francioli, Grace Tiao, Beryl B Cummings, Jessica Alföldi, Qingbo Wang, Ryan L Collins, Kristen M Laricchia, Andrea Ganna, Daniel P Birnbaum, ... Daniel G MacArthur  
*Nature* (2020-05-27) <https://www.ncbi.nlm.nih.gov/pubmed/32461654>  
DOI: [10.1038/s41586-020-2308-7](https://doi.org/10.1038/s41586-020-2308-7) · PMID: [32461654](#) · PMCID: [PMC7334197](#)

43. **A method and server for predicting damaging missense mutations.**  
Ivan A Adzhubei, Steffen Schmidt, Leonid Peshkin, Vasily E Ramensky, Anna Gerasimova, Peer Bork, Alexey S Kondrashov, Shamil R Sunyaev  
*Nature methods* (2010-04) <https://www.ncbi.nlm.nih.gov/pubmed/20354512>  
DOI: [10.1038/nmeth0410-248](https://doi.org/10.1038/nmeth0410-248) · PMID: [20354512](https://pubmed.ncbi.nlm.nih.gov/20354512/) · PMCID: [PMC2855889](https://pmc.ncbi.nlm.nih.gov/pmc/articles/PMC2855889/)
44. **The base excision repair enzyme MED1 mediates DNA damage response to antitumor drugs and is associated with mismatch repair system integrity.**  
Salvatore Cortellino, David Turner, Valeria Masciullo, Filippo Schepis, Domenico Albino, Rene Daniel, Anna Marie Skalka, Neal J Meropol, Christophe Alberti, Lionel Larue, Alfonso Bellacosa  
*Proceedings of the National Academy of Sciences of the United States of America* (2003-11-12) <https://www.ncbi.nlm.nih.gov/pubmed/14614141>  
DOI: [10.1073/pnas.2334585100](https://doi.org/10.1073/pnas.2334585100) · PMID: [14614141](https://pubmed.ncbi.nlm.nih.gov/14614141/) · PMCID: [PMC299910](https://pmc.ncbi.nlm.nih.gov/pmc/articles/PMC299910/)
45. **MBD4 deficiency reduces the apoptotic response to DNA-damaging agents in the murine small intestine.**  
Owen James Sansom, Joanna Zabkiewicz, Stefan Mark Bishop, Jackie Guy, Adrian Bird, Alan Richard Clarke  
*Oncogene* (2003-10-16) <https://www.ncbi.nlm.nih.gov/pubmed/14562041>  
DOI: [10.1038/sj.onc.1206850](https://doi.org/10.1038/sj.onc.1206850) · PMID: [14562041](https://pubmed.ncbi.nlm.nih.gov/14562041/)
46. **A human cancer-associated truncation of MBD4 causes dominant negative impairment of DNA repair in colon cancer cells.**  
SA Bader, M Walker, DJ Harrison  
*British journal of cancer* (2007-02-06) <https://www.ncbi.nlm.nih.gov/pubmed/17285135>  
DOI: [10.1038/sj.bjc.6603592](https://doi.org/10.1038/sj.bjc.6603592) · PMID: [17285135](https://pubmed.ncbi.nlm.nih.gov/17285135/) · PMCID: [PMC2360052](https://pmc.ncbi.nlm.nih.gov/pmc/articles/PMC2360052/)
47. **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, Šaunak 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](https://doi.org/10.1534/genetics.118.301595) · PMID: [30591514](https://pubmed.ncbi.nlm.nih.gov/30591514/) · PMCID: [PMC6366910](https://pmc.ncbi.nlm.nih.gov/pmc/articles/PMC6366910/)
48. **An analytical framework for whole-genome sequence association studies and its implications for autism spectrum disorder.**  
Donna M Werling, Harrison Brand, Joon-Yong An, Matthew R Stone, Lingxue Zhu, Joseph T Glessner, Ryan L Collins, Shan Dong, Ryan M Layer, Irene Markenscoff-Papadimitriou, ... Stephan J Sanders  
*Nature genetics* (2018-04-26) <https://www.ncbi.nlm.nih.gov/pubmed/29700473>  
DOI: [10.1038/s41588-018-0107-y](https://doi.org/10.1038/s41588-018-0107-y) · PMID: [29700473](https://pubmed.ncbi.nlm.nih.gov/29700473/) · PMCID: [PMC5961723](https://pmc.ncbi.nlm.nih.gov/pmc/articles/PMC5961723/)
49. **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/iyc087](https://doi.org/10.1093/genetics/iyc087) · PMID: [35666194](https://pubmed.ncbi.nlm.nih.gov/35666194/) · PMCID: [PMC9339295](https://pmc.ncbi.nlm.nih.gov/pmc/articles/PMC9339295/)
50. **Understanding what determines the frequency and pattern of human germline mutations.**  
Norman Arnheim, Peter Calabrese  
*Nature reviews. Genetics* (2009-07) <https://www.ncbi.nlm.nih.gov/pubmed/19488047>  
DOI: [10.1038/nrg2529](https://doi.org/10.1038/nrg2529) · PMID: [19488047](https://pubmed.ncbi.nlm.nih.gov/19488047/) · PMCID: [PMC2744436](https://pmc.ncbi.nlm.nih.gov/pmc/articles/PMC2744436/)
51. **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](https://doi.org/10.1093/genetics/138.3.963) · PMID: [7851788](https://pubmed.ncbi.nlm.nih.gov/7851788/) · PMCID: [PMC1206241](https://pubmed.ncbi.nlm.nih.gov/PMC1206241/)
52. **Array programming with NumPy**  
Charles R Harris, Kjarrrod 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/ghbfz2>  
DOI: [10.1038/s41586-020-2649-2](https://doi.org/10.1038/s41586-020-2649-2) · PMID: [32939066](https://pubmed.ncbi.nlm.nih.gov/32939066/) · PMCID: [PMC7759461](https://pubmed.ncbi.nlm.nih.gov/PMC7759461/)
53. **pandas-dev/pandas: Pandas**  
The Pandas Development Team  
*Zenodo* (2023-04-03) <https://doi.org/ggt8bh>  
DOI: [10.5281/zenodo.3509134](https://zenodo.org/record/3509134)
54. **Matplotlib: A 2D Graphics Environment**  
John D Hunter  
*Computing in Science & Engineering* (2007) <https://doi.org/drjhg>  
DOI: [10.1109/mcse.2007.55](https://doi.org/10.1109/mcse.2007.55)
55. **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>
56. **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](https://doi.org/10.25080/majora-342d178e-010)
57. **seaborn: statistical data visualization**  
Michael Waskom  
*Journal of Open Source Software* (2021-04-06) <https://doi.org/gjqn3g>  
DOI: [10.21105/joss.03021](https://doi.org/10.21105/joss.03021)
58. **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](https://doi.org/10.1145/2833157.2833162)
59. **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](https://doi.org/10.12688/f1000research.29032.1) · PMID: [34035898](https://pubmed.ncbi.nlm.nih.gov/34035898/) · PMCID: [PMC8114187](https://pubmed.ncbi.nlm.nih.gov/PMC8114187/)
60. **A natural mutator allele shapes mutation spectrum variation in mice**  
Tom Sasani  
(2023-01-24) [https://github.com/tomsasani/bxd\\_mutator\\_manuscript](https://github.com/tomsasani/bxd_mutator_manuscript)
61. **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](https://doi.org/10.5281/zenodo.5941048)
62. **BXD Genotype / WebQTL** <https://gn1.genenetwork.org/dbdoc/BXDGeno.html>
63. **Genetics and Probability in Animal Breeding Experiments**  
<https://link.springer.com/book/10.1007/978-1-349-04904-2>
64. **MGI-Guidelines for Nomenclature of Mouse and Rat Strains**  
<http://www.informatics.jax.org/mgihome/nomen/strains.shtml>
65. **ENA Browser** <https://www.ebi.ac.uk/ena/browser/view/PRJEB45429>
66. **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](https://doi.org/10.1093/gigascience/giab008) · PMID: [33590861](https://pubmed.ncbi.nlm.nih.gov/33590861/) · PMCID: [PMC7931819](https://pmcid.ncbi.nlm.nih.gov/pmcid/PMC7931819/)
67. **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](https://doi.org/10.4161/fly.19695) · PMID: [22728672](https://pubmed.ncbi.nlm.nih.gov/22728672/) · PMCID: [PMC3679285](https://pmcid.ncbi.nlm.nih.gov/pmcid/PMC3679285/)
68. **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](https://doi.org/10.1093/bioinformatics/btx057) · PMID: [28165109](https://pubmed.ncbi.nlm.nih.gov/28165109/) · PMCID: [PMC5870853](https://pmcid.ncbi.nlm.nih.gov/pmcid/PMC5870853/)
69. **The UCSC Table Browser data retrieval tool.**  
Donna Karolchik, Angela S Hinrichs, Terrence S Furey, Krishna M Roskin, Charles W Sugnet, David Haussler, WJames Kent  
*Nucleic acids research* (2004-01-01) <https://www.ncbi.nlm.nih.gov/pubmed/14681465>  
DOI: [10.1093/nar/gkh103](https://doi.org/10.1093/nar/gkh103) · PMID: [14681465](https://pubmed.ncbi.nlm.nih.gov/14681465/) · PMCID: [PMC308837](https://pmcid.ncbi.nlm.nih.gov/pmcid/PMC308837/)
70. **bx-python**  
Taylor Lab at Johns Hopkins University  
(2023-03-16) <https://github.com/bxlab/bx-python>
71. **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](https://doi.org/10.1038/hdy.1992.131) · PMID: [16718932](https://pubmed.ncbi.nlm.nih.gov/16718932/)
72. **coxme: Mixed Effects Cox Models**  
Terry M Therneau  
(2022-10-03) <https://CRAN.R-project.org/package=coxme>