

Discovering epistasis between germline mutator alleles in mice

This manuscript ([permalink](#)) was automatically generated from quinlan-lab/mutator-epistasis-manuscript@9f3f419 on March 13, 2023.

Authors

- **Thomas A. Sasani**

 [0000-0003-2317-1374](#) ·  [tomsasani](#) ·  [tomsasani](#)

Department of Human Genetics, University of Utah · Funded by Grant XXXXXXXX

- **Aaron R. Quinlan** 

 [0000-0003-1756-0859](#) ·  [aaronquinlan](#)

Department of Human Genetics, University of Utah; Department of Biomedical Informatics, University of Utah

- **Kelley Harris** 

 [0000-0003-0302-2523](#) ·  [Kelley_Harris](#)

Department of Genome Sciences, University of Washington

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

Abstract

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

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

In this study, we developed a new method to detect alleles that affect the mutation spectrum in biparental RILs. By applying this method to mutation data from the BXDs, we confirmed the activity of the germline mutator locus near *Mutyh* and discovered an additional C>A germline mutator locus on chromosome 6 that overlaps *Ogg1* 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 exhibit even higher C>A mutation rates than those with either one alone.

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

Introduction

The germline mutation rate is a fundamental parameter in population genetics, and reflects the complex interplay between DNA replication and repair pathways, exogenous sources of DNA damage, and life-history traits. *Mutator alleles* may explain some of the within- and between-species variation in germline mutation rates [6], but have proven challenging to identify in mammalian genomes.

Germline mutator alleles are difficult to detect for a number of reasons, including the fidelity of germline genome replication and the effects of selection on mutators. On average, humans are born with 70 to 100 single-nucleotide *de novo* germline mutations per diploid genome [7,8]; in mice, that number is closer to 20 or 30 [9]. Due to the low baseline germline mutation rate in many mammals, it can be challenging to ascertain sequencing data from enough haplotypes to reliably detect those with significantly elevated *de novo* mutation counts. Moreover, in a population of sufficiently large N_e (effective population size), large-effect mutator alleles will likely be efficiently purged by negative selection. The estimated selection coefficient on a mutator allele is approximately $2s\Delta U$ [10], where s is the mean selective coefficient on a new deleterious mutation and ΔU is the excess number of deleterious mutations caused by the mutator allele; the product of s and ΔU is multiplied by 2 to account for the average number of generations for which mutator is linked to the excess mutations it causes.

Compared to haplotypes that harbor wild-type alleles at a particular locus, those with mutator alleles will likely carry an excess of total germline mutations. Indeed, candidate germline mutator loci have been discovered in human genomes by identifying haplotypes with significantly more derived alleles

than the population mean [11]. However, protein-coding genes involved in DNA replication and repair often recognize particular sequence motifs or excise lesions at specific nucleotides [6]. Thus, we might also expect the spectrum of de novo mutations — that is, the frequency of each individual mutation type — to differ between genomes that carry either a mutator or wild-type allele at a given locus. Mutation spectra are usually defined in terms of k -mer nucleotide context [6]; 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).

To discover alleles that affect elements of the mutation spectrum, a straightforward approach would involve comparing the rates of individual k -mer mutation types between groups of haplotypes at a set of candidate loci (e.g., via quantitative trait locus mapping). However, this approach can quickly incur a substantial multiple testing burden, and since germline mutation rates are generally quite low, estimates of k -mer mutation type frequencies on each haplotype may 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.

To maximize our power to detect germline mutator alleles, we analyzed whole-genome sequencing data derived from 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 [12]. Since the BXD RILs were maintained via brother-sister mating for up to 180 generations and housed in a controlled laboratory environment, they represent an ideal population for mutator allele discovery. Each line accumulated hundreds or thousands of germline mutations on a nearly-homozygous linear mosaic of parental B and D haplotypes, while the effects of negative selection on new and standing variation were attenuated by strict inbreeding [13].

In a previous study, we used quantitative trait locus (QTL) mapping to identify a locus on chromosome 4 that was strongly associated with the C>A germline mutation rate in the BXDs [2]. The QTL overlapped *Mutyh*, which encodes a protein that normally prevents C>A mutations by repairing oxidative DNA damage [4], and we hypothesized that missense mutations in *Mutyh* were responsible for a 50% increase in the C>A mutation rate between BXDs with either parental haplotype at the QTL [2]. Using QTL mapping, however, lacked power to detect more subtle germline mutation rate modifiers in the BXDs.

Here, we demonstrate the power of our new approach to detect candidate mutator alleles, use it to re-identify the mutator near *Mutyh*, and discover compelling evidence of epistasis between two germline mutator alleles that augment the C>A germline mutation rate in the BXD RILs.

Results

Benchmarking the inter-haplotype distance method using simulations

We developed a statistical method, termed “inter-haplotype distance” (IHD), to detect loci that are associated with the mutation spectrum in biparental RILs (Materials and Methods). Using simulated data (Materials and Methods), we find that the method’s power is primarily limited by the initial mutation rate of the k -mer mutation type affected by the mutator allele and the total number of *de novo* germline mutations used to detect it (that is, the product of the number of haplotypes and the mean number of mutations per haplotype) (Figure 1). For example, given 50 haplotypes with an average of 500 *de novo* germline mutations each, the IHD approach has approximately 60% power to detect a mutator allele that increases the C>T *de novo* mutation rate by as little as 10%. However, the

method only has about 30% power to detect a mutator of identical effect size that affects the C>G mutation rate, since C>G mutations are expected to make up a smaller fraction of all *de novo* germline mutations to begin with. These 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).

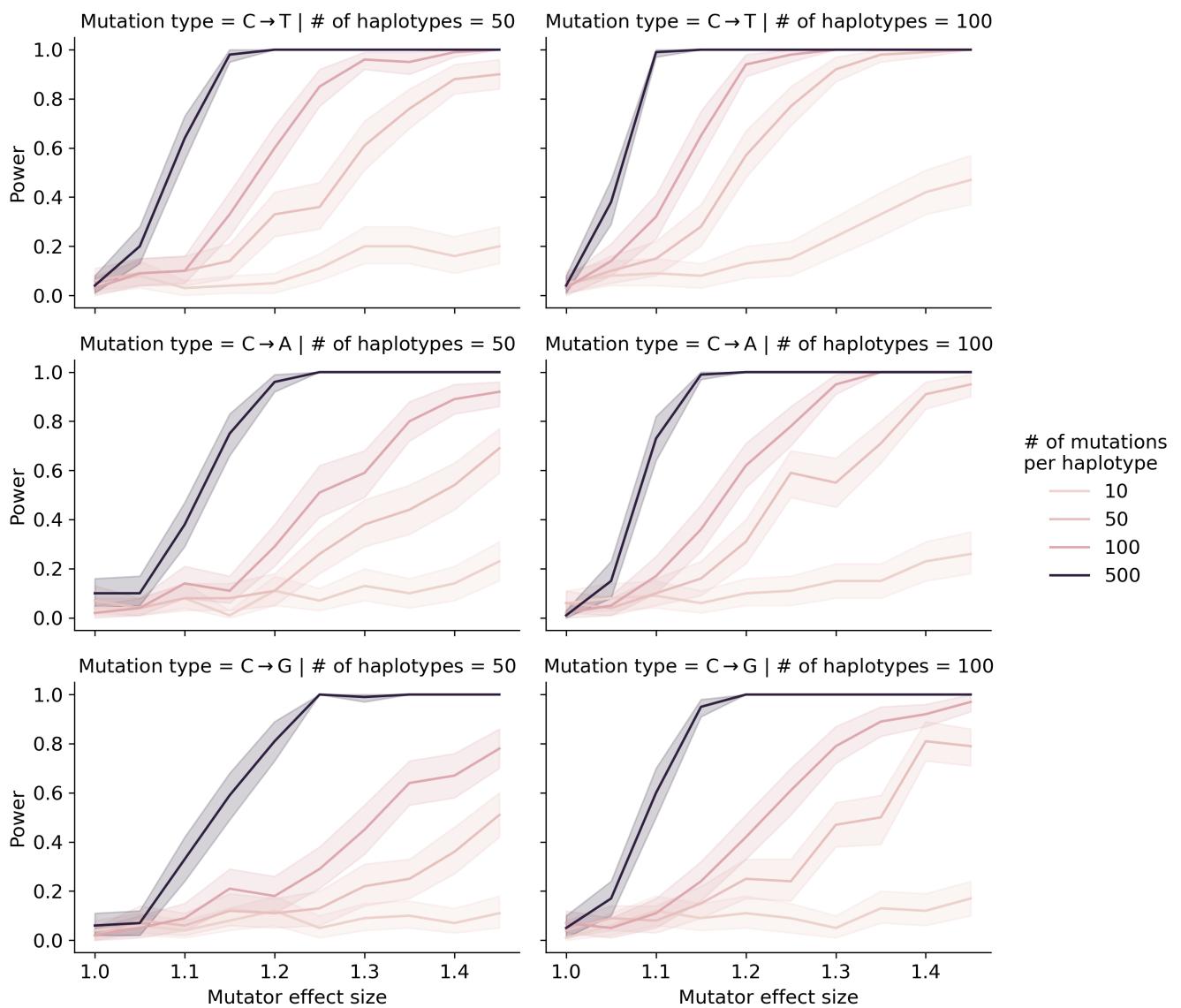


Figure 1: Simulations to assess the power of the inter-haplotype distance method. We simulated *de novo* germline mutations on a toy population of haplotypes such that 50% of haplotypes were affected by a mutator allele that increased the mutation rate of the specified k -mer by the specified effect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

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

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

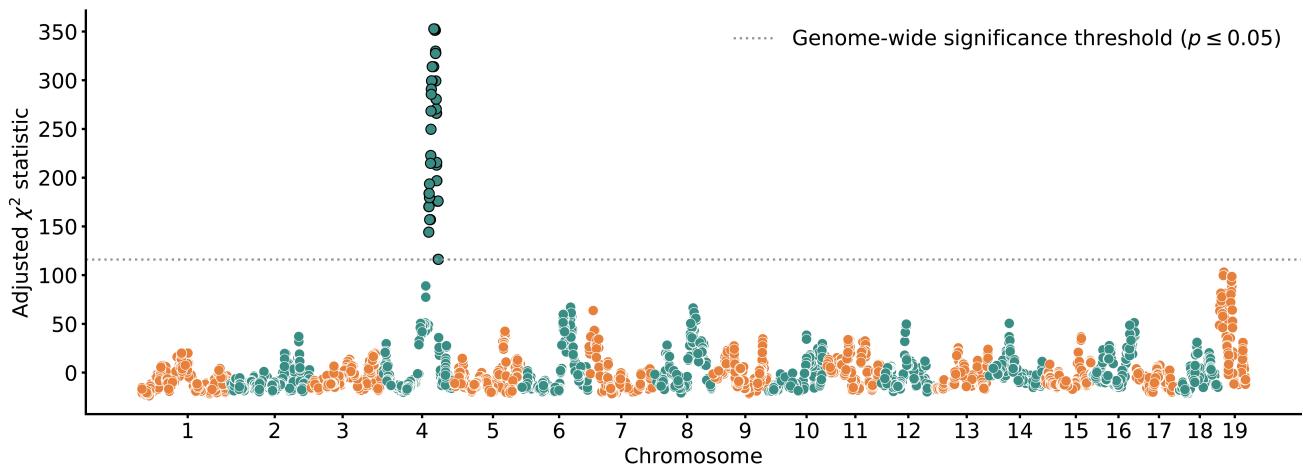
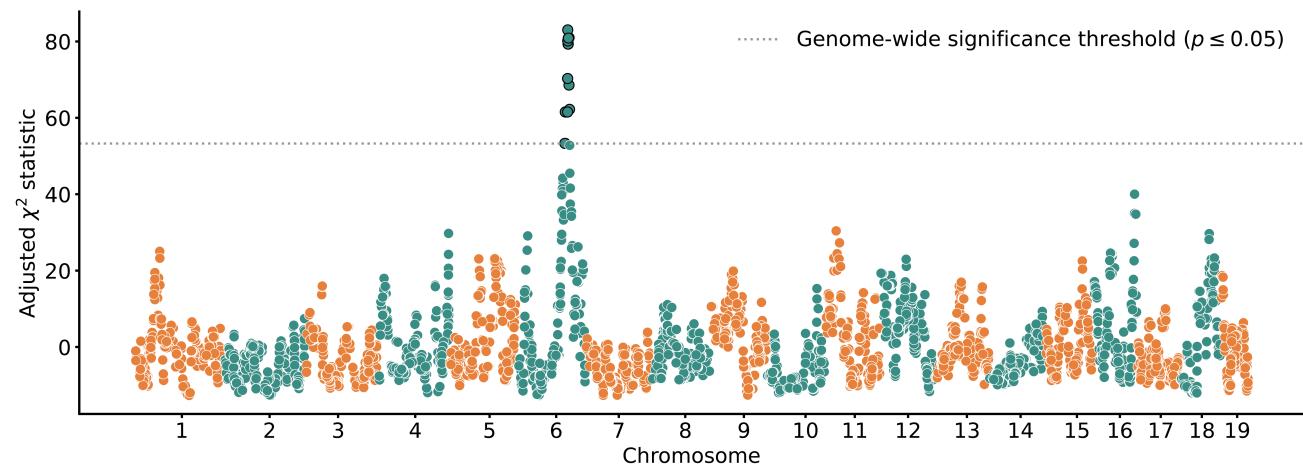
a**b**

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

In a previous analysis, we used quantitative trait locus (QTL) mapping to identify a nearly identical locus on chromosome 4 that was significantly associated with the C>A germline mutation rate in the BXDs [2]. This locus overlapped 21 protein-coding genes that are 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 [4,14,15]. C>A germline mutation rates are nearly 50% higher in BXDs that inherit *D* genotypes at marker ID rs52263933 than in those that inherit *B* genotypes (Figure 3) [2].

An additional germline mutator allele on chromosome 6

After confirming that the inter-haplotype distance method could recover the mutator locus overlapping *Mutyh*, we asked if our approach could identify additional mutator loci in the BXD. To account for the effects of the large-effect C>A germline mutator locus near *Mutyh*, we divided the BXD RILs into those with either *D* ($n = 55$) or *B* ($n = 38$) genotypes at rs52263933 (the marker at which we

observed the highest adjusted χ^2 statistic on chromosome 4), and ran a genome-wide distance scan using each group separately (Figure 2B).

Using only the BXDs with *B* genotypes at the *Mutyh* mutator locus, we did not observe any genome-wide significant peaks. But using the BXDs with *D* genotypes at the same locus, we identified a χ^2 statistic peak on chromosome 6 (Figure 2B and [2—figure supplement 1](#); maximum adjusted χ^2 statistic of 81.0 at marker rs31001331 ; position 114.05 Mbp in GRCm38/mm10 coordinates). We queried the region underneath this peak (+/- 5 Mbp) and discovered 16 protein-coding genes that harbored nonsynonymous differences between C57BL/6J and DBA/2J. 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 at methylated CpG sites that have undergone spontaneous deamination. 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 | Amino acid change | Position in GRCm38/mm10 coordinates | SIFT prediction |
|-------------|-------------------------|-------------------|-------------------------------------|----------------------------------|
| <i>Ogg1</i> | ENSMUST00000032406 | p.Thr95Ala | chr6:113,328,510 | 0.84 (tolerant/benign) |
| <i>Mbd4</i> | ENSMUST00000032469 | p.Asp129Asn | chr6:115,849,644 | 0.02 (intolerant/deleterious) |

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 [16] 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 χ^2 statistic peak on chromosome 6 were significantly associated with *Ogg1* expression in some (but not all) tissues, and *D* genotypes were nearly always associated with decreased gene expression (Table [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 [17], but did not find any significant eQTLs for either *Ogg1* or *Mbd4*.

Finally, we queried the structural variants identified in a population of inbred laboratory strains sequenced by the Sanger Mouse Genomes Project (MGP) [18] and found 93 deletions within the implicated interval on chromosome 6. Of these, five overlapped the exonic sequences of three protein-coding genes: *Tmem72*, *Tmcc1*, and *Gm20387*. None of these genes has a previously annotated role in DNA binding, repair or replication, or in a pathway that would likely affect germline mutation rates.

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. We observed that C>A germline mutation fractions in BXDs with *D* alleles at both mutator loci were higher than in BXDs with *D* alleles at either locus alone; GCT>GAT and GCA>GAA mutations were specifically enriched in the former (Figure 3). However, compared to BXDs with *B* alleles at the chromosome 6 mutator locus, those with *D* alleles did not exhibit higher C>A

mutation fractions, indicating that the effects of the chromosome 6 mutator locus depend on the presence of a *D* allele at the chromosome 4 locus (Figure 3). To more formally test for epistasis, we fit a linear model predicting counts of C>A mutations in each strain 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 of the two markers ($p = 9.8\text{e-}4$).

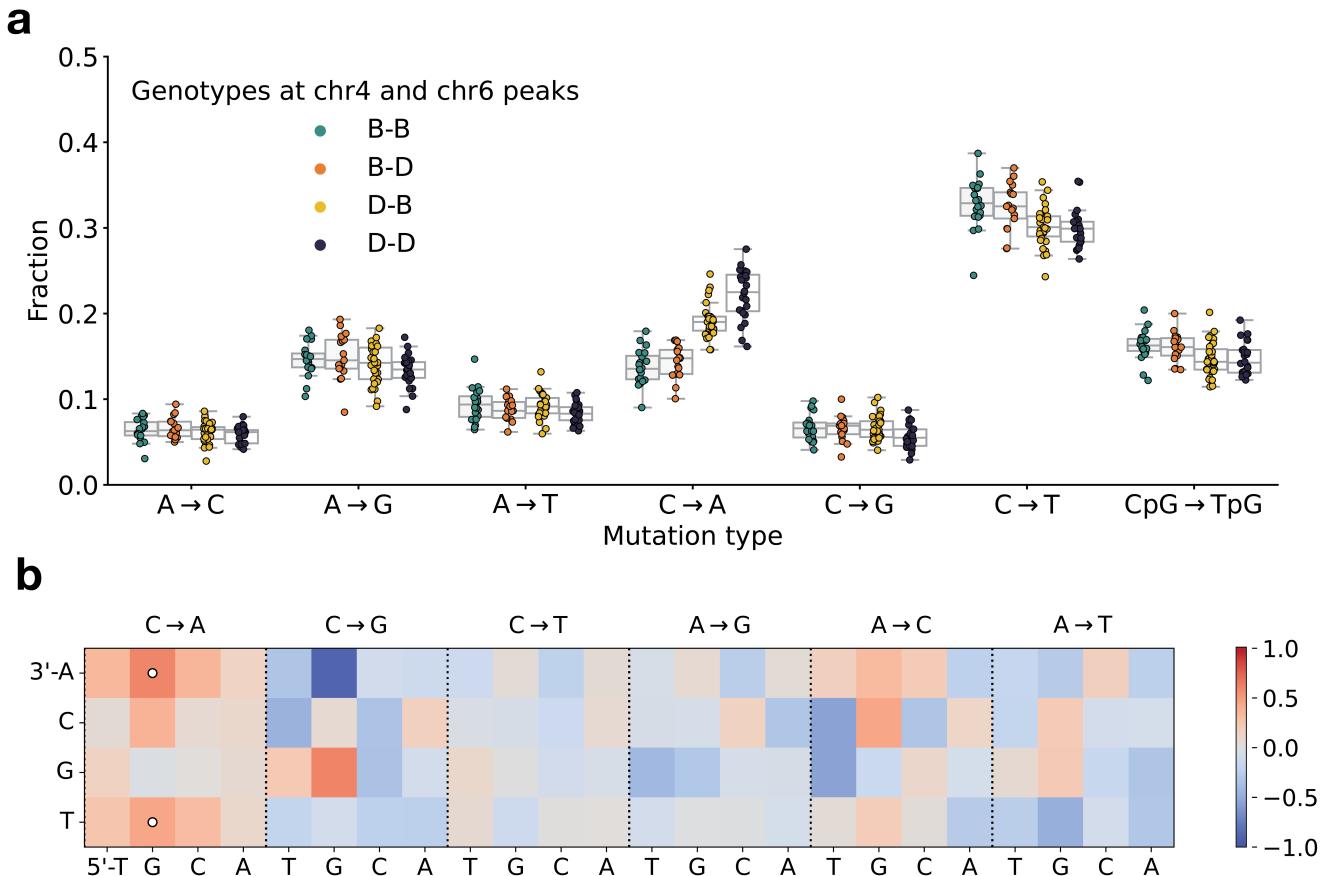


Figure 3: BXD mutation spectra are affected by alleles at both mutator loci. **a)** Fractions of *de novo* germline mutations in BXDs with either *D* or *B* genotypes at markers rs52263933 and rs31001331, stratified by mutation type. **b)** \log_2 -transformed ratios of 3-mer mutation fractions in BXDs with *D* genotypes at rs52263933 and either *D* or *B* haplotypes at rs31001331. Comparisons for which a χ^2 test of independence returned $p < \frac{0.05}{96}$ (Bonferroni corrected $p < 0.05$) are annotated with a white dot.

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 [18]. Dumont [12] previously identified germline mutations that were private to each of the 29 MGP strains; these private variants likely represent recent *de novo* germline mutations (Figure 3—figure supplement 1). Only two of the MGP strains possess *D* genotypes at both the chromosome 4 and chromosome 6 mutator loci: DBA/1J and DBA/2J. As before, we tested for epistasis in the MGP strains by fitting two linear models predicting C>A mutation 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 ($p = 0.174$). Thus, we are unable to confirm the signal of epistasis observed in the BXDs, but this may be due to the smaller number of MGP strains with *de novo* germline mutation data.

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 [19]. These data include three subspecies of *Mus musculus*, as well as the outgroup *Mus spretus*. We found that the *D* allele of *Ogg1* 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 [20], and was fixed in *Mus musculus musculus*, *Mus musculus castaneus*, and the outgroup *Mus spretus*. However, the *D* allele of *Mbd4* was not present in any of the wild mice.

Discussion

Epistasis between germline mutator alleles

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

Importantly, we note that we observed epistasis between germline mutator alleles in an unnatural population; the BXDs were inbred by brother-sister mating in a highly controlled laboratory environment that attenuated the effects of natural selection on all but the most deleterious alleles [13]. Large-effect mutator alleles (and epistasis between them) may be much less common in natural, outbreeding mammalian populations. However, our results demonstrate that germline mutation rates in recombinant inbred populations are highly plastic, and that RILs represent a powerful system in which to discover germline mutators.

Causal variants underlying the mutator allele

Two protein-coding DNA repair genes overlap the C>A mutator locus on chromosome 6 and also contain nonsynonymous fixed 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 can bind to methylated CpG sites and remove mispaired thymine nucleotides opposite spontaneously deaminated CpGs.

Missense mutations and loss-of-heterozygosity in *Ogg1* have been associated with increased risk of human cancer [21,22], and copy-number losses of either *Ogg1* or *Mutyh* are linked to elevated rates of spontaneous C>A mutation in human neuroblastoma [23]. Although *Ogg1* is a member of the same base-excision repair pathway as *Mutyh* (the protein-coding gene we previously implicated as harboring mutator alleles at the locus on chromosome 4), a number of lines of evidence suggest that the p.Asp129Asn missense mutation in *Mbd4* is the more compelling candidate mutator allele on chromosome 6. Unlike the *Ogg1* p.Thr95Ala mutation, p.Asp129Asn occurs at an amino acid residue within the *Mbd4* methyl-CpG binding domain that is well-conserved across mammalian species and is predicted to be deleterious by *in silico* tools like SIFT [24] (Table 1). Additionally, a missense mutation that affects the homologous amino acid in humans (p.Asp142Gly in GRCh38/hg38) is present on a single haplotype in the Genome Aggregation Database (gnomAD) [25] and is predicted by SIFT and Polyphen [26] to be “deleterious” and “probably_damaging” in human genomes, respectively.

One of the key roles of *Mbd4* is to excise thymine nucleotides at G:T mispairs, which arise due to the spontaneous deamination of methylated CpGs. As a result, loss-of-function (LOF) mutations in *Mbd4* often lead to increased rates of CpG>TpG mutation [27]. Although LOF mutations in *Mbd4* are not

known to cause C>A mutator phenotypes in mammalian cells, *Mbd4* is involved in a number of DNA repair processes that may contribute to the mutator phenotype we observed in the BXDs. For example, bi-allelic LOF mutations in human *MBD4* underlie a neoplastic syndrome that closely mimics forms of familial adenomatous polyposis caused by LOF mutations in *MUTYH* [28]. Perhaps most intriguingly, LOF mutations in *Mbd4* can exacerbate the effects of exogenous DNA damage agents. Mouse embryonic fibroblasts that harbor homozygous LOF mutations in *Mbd4* fail to undergo apoptosis following treatment with a number of chemotherapeutics and mutagenic compounds [29]. Most of these exogenous mutagens cause DNA damage that is normally repaired by mismatch repair (MMR) machinery, but murine intestinal cells with biallelic LOF mutations in *Mbd4* also show a reduced apoptotic response to gamma irradiation, which is repaired independently of the MMR gene *Mlh1* [30]. Homozygous LOF mutations in *Mbd4* lead to accelerated intestinal tumor formation in mice that harbor an *Apc* allele that predisposes them to intestinal neoplasia [27], 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 [31].

Potential roles of structural variation and mobile elements as mutator alleles

Given the evidence outlined above, we believe that *Mbd4* is the most likely candidate gene to explain the additional C>A mutator phenotype in the BXDs, but we are unable to conclusively determine that the p.Asp129Asn missense mutation is the 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]. Using high-quality long-read assemblies of inbred laboratory strains, another group recently identified a ~5 kbp mobile element insertion (MEI) within the first intron of *Mutyh* [32] that is present on *D* haplotypes and absent from *B* haplotypes. This MEI is associated with significantly reduced expression of *Mutyh* in embryonic stem cells from laboratory strains, and may in fact underlie the previous C>A germline mutator phenotype in the BXDs. Although we did not find compelling evidence that structural variants or mobile element insertions were responsible for the novel C>A mutator phenotype observed in this study, it remains plausible that large SVs or MEIs contribute to mutation spectrum evolution in mammalian genomes.

Mechanisms of epistasis between mutator alleles

Although both *Mutyh* and *Mbd4* are DNA glycosylases that participate in base-excision repair, they operate on distinct DNA lesions: *Mutyh* excises adenines paired with 8-oxoguanine, while *Mbd4* excises thymines opposite spontaneously deaminated CpGs. Given *Mbd4*'s role in suppressing DNA damage-induced apoptosis [29,30], we hypothesize that *D* alleles in *Mutyh* and *Mbd4* exhibit epistasis through the following mechanism in the BXD RILs.

In the absence of other defects in the DNA repair response, *D* alleles at *Mbd4* appear to have little or no detectable effect on *de novo* mutation rates (Figure 3). Since loss-of-function mutations in *Mbd4* are expected to increase CpG>TpG mutation rates [27], it is possible that the p.Asp129Asn missense mutation is not sufficient to cause such a phenotype in the BXDs. As we demonstrated in this and a previous manuscript, *D* alleles in *Mutyh* alone lead to significantly increased C>A mutation rates (Figure 3) [2]. We hypothesize that in response to an accumulation of C>A mutations, a fraction of spermatagonial stem cells with *D* alleles at *Mutyh* will undergo apoptosis to prevent further unrepaired DNA damage. However, if those germline cells harbor *D* alleles at both *Mutyh* and *Mbd4*, they may not be able to arrest the cell cycle and complete apoptosis, allowing the effects of *D* alleles in *Mutyh* to exacerbate C>A mutation rates even further.

Discovering mutator alleles in other systems

Mutator alleles likely contribute to variation in mutation rates and spectra across the tree of life. In two natural isolates of *Saccharomyces cerevisiae*, nonsynonymous variation in *OGG1* causes a substantial increase in the C>A *de novo* mutation rate [33]. Recent analyses have suggested that mutator alleles and/or environmental mutagens have shaped mutation rate evolution both in human genomes [34] and more broadly during great ape evolution [35]. The heritability of paternal *de novo* mutation counts in the human germline has also been estimated to be between 10 and 20%, demonstrating a contribution of genetic factors to germline mutation rates [36]. However, mutator discovery remains challenging in mammalian genomes.

Recently, thousands of human pedigrees have been sequenced in an effort to precisely estimate the rate of human *de novo* germline mutation [7,8]. Selection on germline mutator alleles will likely prevent large-effect mutators from reaching high allele frequencies; as the number of active mutators increases, however, it becomes much more likely that a subset will be detectable by sequencing a sufficient number of human trios [37]. Current estimates of power to detect germline mutators in human pedigrees generally assume that mutators affect all mutation types equally, and that methods for mutator discovery will rely on identifying haplotypes with excess total mutation counts [37]. Our results in the BXD suggest that germline mutators often exert their effects on a small number of *k*-mer mutation types, and may be far more amenable to detection by analyzing mutation spectra instead.

Materials and Methods

Identifying *de novo* germline mutations in the BXD RILs

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

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

- genotyped as either homozygous or heterozygous for the alternate allele, with at least 90% of sequencing reads supporting the alternate allele
- supported by at least 10 sequencing reads
- Phred-scaled genotype quality of at least 20
- must not overlap regions of the genome annotated as segmental duplications or simple repeats in GRCm38/mm10
- must occur on a parental haplotype that was inherited by at least one other BXD at the same locus; these other BXDs must be homozygous for the reference allele at the variant site

A new approach to discover germline mutator alleles

Calculating inter-haplotype distance

Using the existing catalog of *de novo* germline mutations in the BXDs, we developed a new approach to discover loci that affect the germline *de novo* mutation spectrum in biparental RILs (Figure 4).

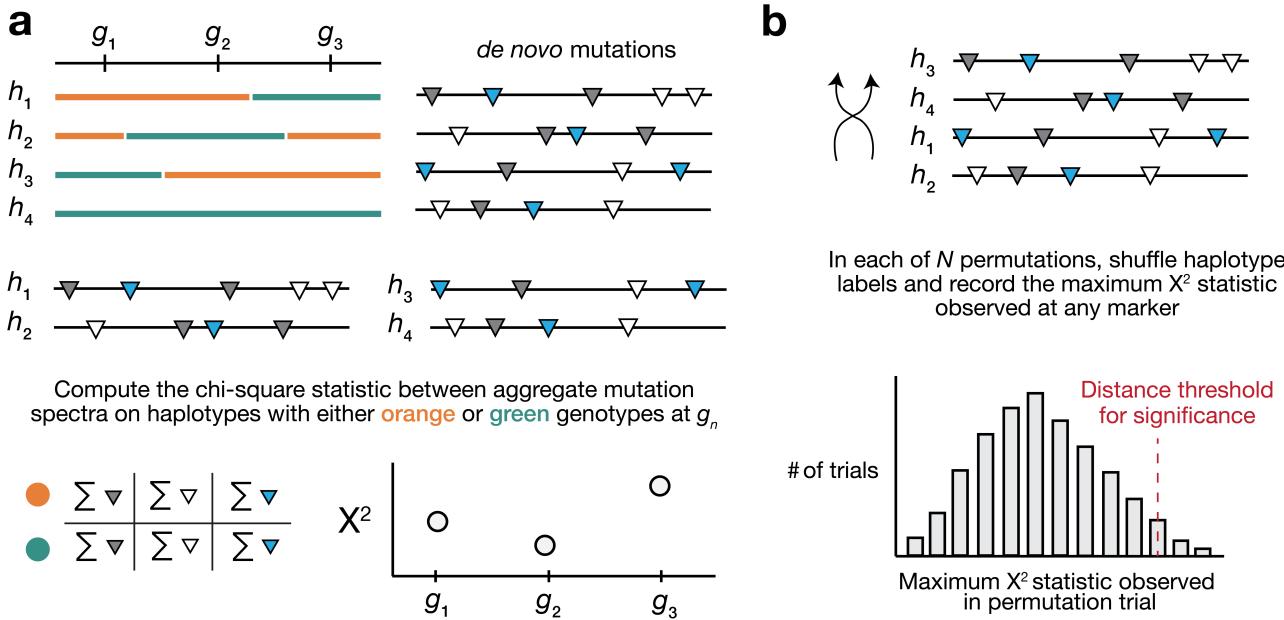


Figure 4: Overview of inter-haplotype distance method for discovering mutator alleles. **a)** A population of four haplotypes has been genotyped at three informative markers (g_1 through g_3); each haplotype also harbors private *de novo* germline mutations, which are denoted as triangles. *De novo* mutations are further classified into three possible mutation types, indicated by the color of each triangle. 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). We then calculate a χ^2 statistic between the two aggregate mutation spectra. We repeat this process for every informative marker g_n . **b)** To assess the significance of any χ^2 statistic peaks in a), we perform a permutation test by shuffling the haplotype labels associated with the *de novo* mutation data. In each of N permutations, we record the maximum χ^2 statistic encountered at any locus in the distance scan. Finally, we calculate the $1 - p$ percentile of the distribution of those maximum statistics to obtain a genome-wide χ^2 statistic threshold at the specified value of p .

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

At each informative marker, we divide haplotypes into two groups based on the parental allele that they inherited. We then compute a k -mer mutation spectrum using the aggregate mutation counts in each haplotype group. The k -mer mutation spectrum contains the frequency of every possible k -mer mutation type in a collection of mutations, and can be represented as a vector of size $6 \times 4^{k-1}$ after collapsing by strand complement. For example, the 1-mer mutation spectrum is 6-element vector that contains the frequencies of C>T, C>G, C>A, A>G, A>T, and A>C mutations. 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 [38].

At each marker, we then create a contingency table of size $(2, 6 \times 4^{k-1})$ (or when $k = 1$, size $(2, 6 \times 4^{k-1} + 1)$) if we include CpG>TpG that contains the aggregate k -mer mutation spectrum in haplotypes that inherited either parental allele. Using the contingency table, we then calculate a χ^2 test statistic; larger values of the χ^2 statistic suggest more “distance” between the aggregate mutation spectra on the two haplotypes.

Inspired by methods from QTL mapping [39,40], we use permutation tests to establish genome-wide χ^2 statistic thresholds. In each of N permutation trials, we randomly shuffle the per-haplotype mutation data such that haplotype labels no longer correspond to the correct mutation counts. Using the shuffled mutation data, we perform a genome-wide scan as described above, and record the maximum χ^2 statistic 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-descent (IBD) at a fraction of genotyped markers. Pairs of more genetically similar BXDs may also have more similar mutation spectra, potentially due to shared polygenic effects on the mutation process. Therefore, at a given marker, if the BXD RILs that inherited *D* haplotypes are more genetically dissimilar from the RILs that inherited *B* haplotypes (considering all loci throughout the genome), we might expect the aggregate mutation spectra in the two groups to also be 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 χ^2 statistic between the two aggregate spectra (χ_i^2). Then, within each group of haplotypes, we calculate the allele frequency of the *D* allele at every marker along the genome to obtain a vector of length n , where n is the number of genotyped markers. To quantify the genetic similarity between the two groups of haplotypes, we calculate the Pearson correlation coefficient r_i between the two vectors of marker-wide *D* allele frequencies.

Put another way, at every marker g_i along the genome, we divide BXD haplotypes into two groups and compute two metrics: χ_i^2 (the χ^2 statistic between the two groups' aggregate spectra) and r_i (the correlation between genome-wide *D* allele frequencies in the two groups). To control for the potential effects of genetic similarity on χ^2 statistics, we regress $(\chi_1^2, \chi_2^2, \dots, \chi_n^2)$ 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" χ^2 statistic values for each marker. If genome-wide genetic similarity between haplotypes perfectly predicts χ^2 statistics at each marker, these residuals will all be 0 (or very close to 0). If genome-wide genetic similarity has no predictive power, the residuals will simply represent the difference between the observed χ^2 statistic at a single marker and the marker-wide mean of χ^2 statistics.

Implementation and source code

The inter-haplotype distance method was implemented in Python, and relies heavily on the following Python libraries: `numpy`, `pandas`, `matplotlib`, `scikit-learn`, `pandera`, `seaborn`, and `numba` [41,42,43,44,45,46,47].

Additional documentation is available on GitHub [48], along with a reproducible Snakemake [49] workflow for running the method from start to finish using the BXDs (including downloading the mutation data, downloading genotypes, and running a genome-wide distance scan).

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

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

First, we simulate the k -mer mutation spectrum in a population of h haplotypes. We assume that exactly $\frac{h}{2}$ of the haplotypes are under the effects of a mutator allele that increases the mutation rate of a particular mutation type(s) by an effect size e . We also assume that at any arbitrary marker, haplotypes can possess one of two alleles (allele A or allele B), and that exactly half of the genotypes possess genotype A at each marker. We simulate m mutations on each haplotype as follows:

We first 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 [9].

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 $\frac{P_c}{16} = \frac{0.46}{16} = 0.02875$.

To simulate the mutation spectrum on the *wild-type* haplotypes, we define a matrix C of size $(\frac{h}{2}, n)$, where $n = 6 \times 4^{k-1}$ (or if $k = 1$ and we include CpG>TpG mutations, $6 \times 4^{k-1} + 1$). First, we generate a vector of lambda values by scaling the mutation probabilities by the number of mutations we wish to simulate:

$$\lambda = Pm$$

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

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

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

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

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

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

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

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

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

Applying the inter-haplotype distance method to the BXDs

We downloaded previously-generated BXD *de novo* germline mutation data from the GitHub repository associated with our previous manuscript, which was also archived at Zenodo [2,50,51], and downloaded a CSV file of BXD genotypes at 7,320 informative markers from GeneNetwork [16,52]. 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 [48].

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 [53]. 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 [54]. 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.

We used Snakemake [55] to write a reproducible workflow for running the inter-haplotype distance method on the BXD dataset, which has been deposited in the GitHub repository associated with this manuscript [48].

Identifying candidate mutator alleles overlapping the chromosome 6 peak

We investigated the region implicated by our inter-haplotype distance approach on chromosome 6 by subsetting the joint-genotyped BXD VCF file (European Nucleotide Archive accession **PRJEB45429** [56]) using `bcftools` [57]. We defined the candidate interval surrounding the χ^2 statistic peak on chromosome 6 as +/- 5 Mbp from the genotype marker with the largest adjusted χ^2 statistic value (rs31001331). To predict the functional impacts of both single-nucleotide variants and indels on splicing, protein structure, etc., we annotated variants in the BXD VCF using the following `snpEff` [58] command:

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

and used `cyvcf2` [59] to iterate over the annotated VCF file in order to identify nonsynonymous fixed differences between the parental C57BL/6J and DBA/2J strains.

Comparing mutation spectra between Mouse Genomes Project strains

We downloaded mutation data from a previously published analysis [12] (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 [18]. 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 [16], 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 χ^2 statistic peak on chromosome 6. On the GeneNetwork homepage (genenetwork.org), we selected the “BXD Family” **Group** and used the **Type** dropdown menu to select each of the specific expression datasets described in Table 2. In the **Get Any** text box, we then entered the 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 [60] 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 |
| 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 |

| Tissue name | Complete name of GeneNetwork expression data | GeneNetwork trait ID |
|-------------|--|----------------------|
| 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* [19]. 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 χ^2 statistics 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 [12]. 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 [61] in the R statistical language, we fit a mixed effects model predicting C>A mutation fractions as a function of genotypes at both `rs52263933` and `rs31001331`, and included a pairwise kinship matrix as a random effect.

```

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

```

The rows and columns of the kinship matrix were labeled with the `sample` name of each BXD, such that the `(1|sample)` term in the model captured the random effect of kinship. We calculated the `kinship_matrix` using the `calc_kinship` method from `R/qt12` [40] as follows:

```

# read in the JSON-formatted file that directs R/qt12 to sample
# genotypes, phenotypes, and covariates if applicable
bxd <- read_cross2("path/to/bxd.json")

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

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

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

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

kinship_matrix = as.matrix(k)

```

Supplementary information

Using cosine distance instead of chi-square statistics for comparing mutation spectra

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

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

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

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

Supplementary Figures

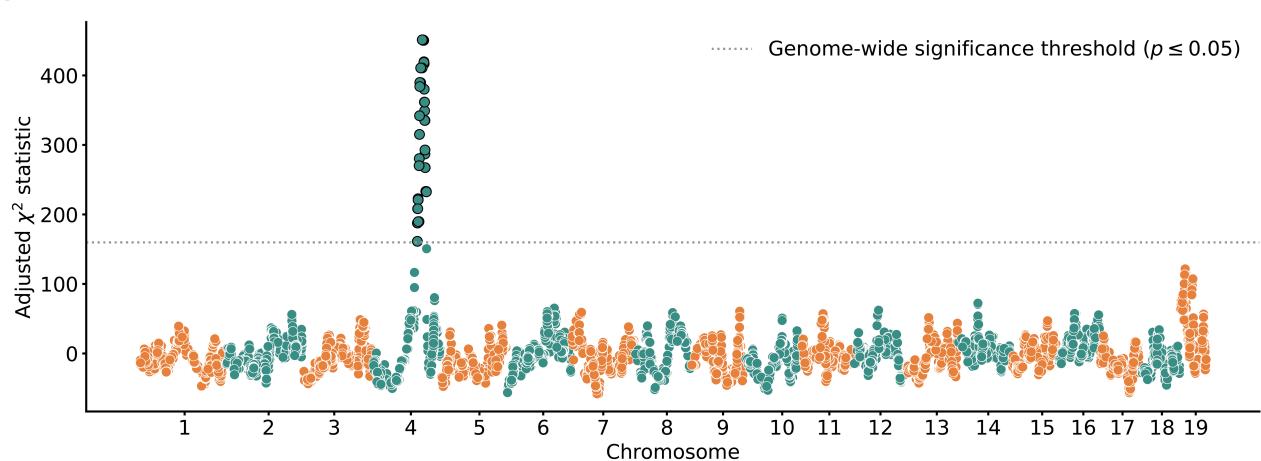
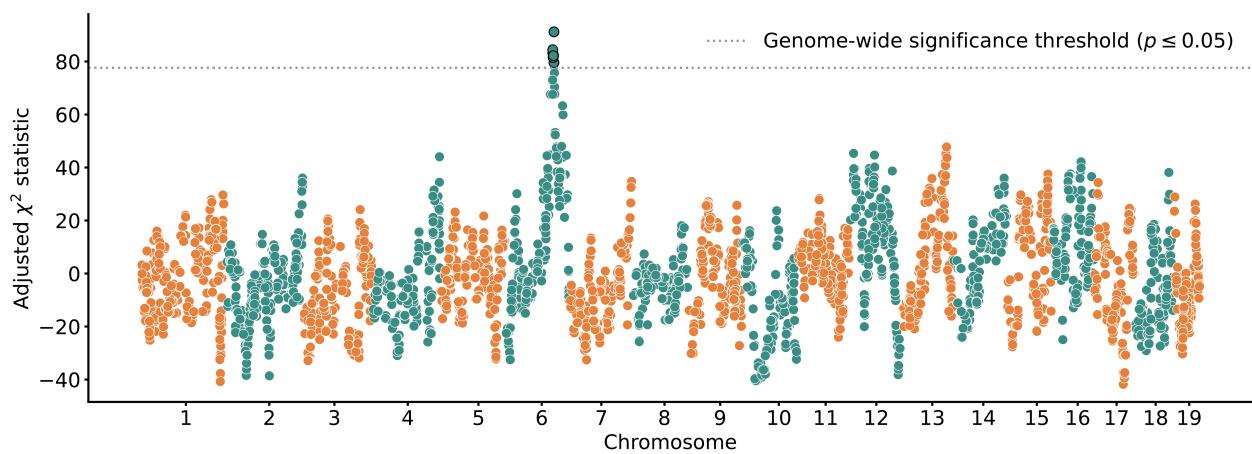
a**b**

Figure 2—figure supplement 1: Results of inter-haplotype distance scans in the BXD RILs using 3-mer mutation spectra. **a)** Adjusted χ^2 statistics between aggregate 3-mer *de novo* mutation spectra on BXD haplotypes ($n = 93$ haplotypes; 62,993 total mutations) with either *D* or *B* alleles at 7,320 informative markers. Distance threshold at $p = 0.05$ was calculated by performing 10,000 permutations of the BXD haplotype mutation data, and is shown as a dotted grey line. **b)** Adjusted χ^2 statistics between aggregate 3-mer *de novo* mutation spectra on BXD haplotypes with *D* alleles at rs52263933 ($n = 55$ haplotypes; 40,913 total mutations) and either *D* or *B* alleles at 7,320 informative markers. Distance threshold at $p = 0.05$ was calculated by performing 10,000 permutations of the BXD haplotype mutation data, and is shown as a dotted grey line.

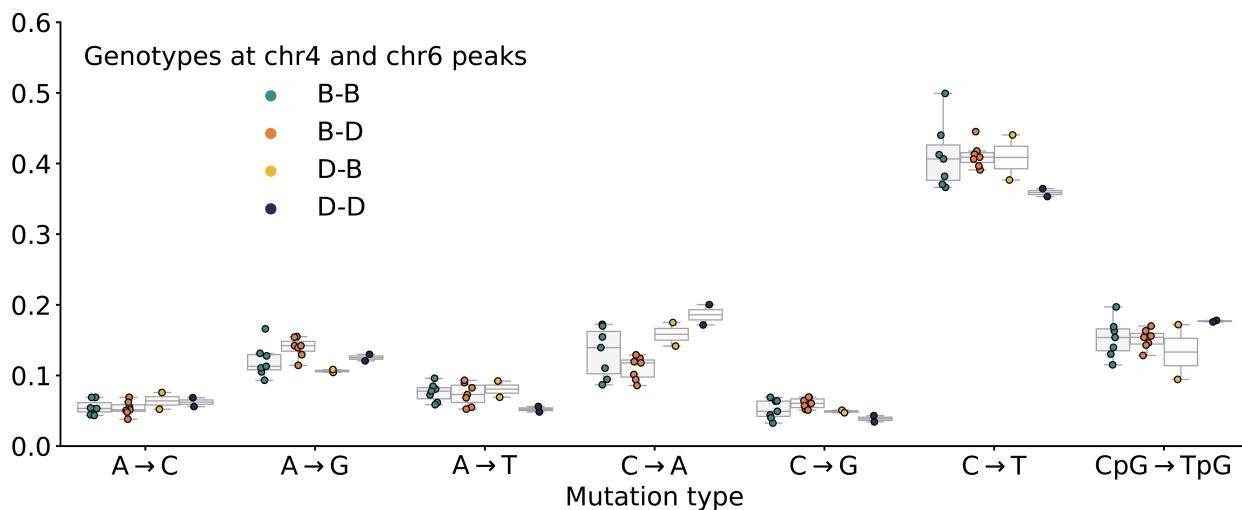


Figure 3—figure supplement 1: Mutation spectra comparison in Sanger Mouse Genomes Project strains.

Fractions of *de novo* germline mutations in Sanger MGP strains with either *D* or *B* genotypes at markers rs52263933 and rs31001331, 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 |
| <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 |

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://pubmed.ncbi.nlm.nih.gov/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://pubmed.ncbi.nlm.nih.gov/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://pubmed.ncbi.nlm.nih.gov/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://pubmed.ncbi.nlm.nih.gov/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://pubmed.ncbi.nlm.nih.gov/PMC22404/)
6. **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/)
7. **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/)
8. **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://pubmed.ncbi.nlm.nih.gov/PMC6759356/)

9. **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](https://pubmed.ncbi.nlm.nih.gov/31492841/) · PMCID: [PMC6731245](https://pmc.ncbi.nlm.nih.gov/pmc/articles/PMC6731245/)
10. **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](https://pubmed.ncbi.nlm.nih.gov/27739533/)
11. **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](https://pubmed.ncbi.nlm.nih.gov/28095480/) · PMCID: [PMC5283766](https://pmc.ncbi.nlm.nih.gov/pmc/articles/PMC5283766/)
12. **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](https://pubmed.ncbi.nlm.nih.gov/30753674/) · PMCID: [PMC6501876](https://pmc.ncbi.nlm.nih.gov/pmc/articles/PMC6501876/)
13. **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)
14. **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://pmc.ncbi.nlm.nih.gov/pmc/articles/PMC5478212/)
15. **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/)
16. **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://pmc.ncbi.nlm.nih.gov/pmc/articles/PMC7495243/)
17. **Mapping the Effects of Genetic Variation on Chromatin State and Gene Expression Reveals Loci That Control Ground State Pluripotency.**
Daniel A Skelly, Anne Czechanski, Candice Byers, Selcan Aydin, Catrina Spruce, Chris Olivier, Kwangbom Choi, Daniel M Gatti, Narayanan Raghupathy, Gregory R Keele, ... Laura G Reinholdt
Cell stem cell (2020-08-13) <https://www.ncbi.nlm.nih.gov/pubmed/32795400>
DOI: [10.1016/j.stem.2020.07.005](https://doi.org/10.1016/j.stem.2020.07.005) · PMID: [32795400](https://pubmed.ncbi.nlm.nih.gov/32795400/) · PMCID: [PMC7484384](https://pmc.ncbi.nlm.nih.gov/pmc/articles/PMC7484384/)

18. **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/)
19. **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/)
20. **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/)
21. **Novel mutations of OGG1 base excision repair pathway gene in laryngeal cancer patients.**
Ishrat Mahjabeen, Nosheen Masood, Ruqia Mehmood Baig, Maimoona Sabir, Uzma Inayat, Faraz Arshad Malik, Mahmood Akhtar Kayani
Familial cancer (2012-12) <https://www.ncbi.nlm.nih.gov/pubmed/22829015>
DOI: [10.1007/s10689-012-9554-2](https://doi.org/10.1007/s10689-012-9554-2) · PMID: [22829015](https://pubmed.ncbi.nlm.nih.gov/22829015/)
22. **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](https://pubmed.ncbi.nlm.nih.gov/9662341/)
23. **Defects in 8-oxo-guanine repair pathway cause high frequency of C > A substitutions in neuroblastoma**
Marlinde L van den Boogaard, Rurika Oka, Anne Hakkert, Linda Schild, Marli E Ebus, Michael R van Gerven, Danny A Zwijnenburg, Piet Molenaar, Lieke L Hoyng, MEmmy M Dolman, ... Jan J Molenaar
Proceedings of the National Academy of Sciences (2021-09-03) <https://doi.org/grtcs9>
DOI: [10.1073/pnas.2007898118](https://doi.org/10.1073/pnas.2007898118) · PMID: [34479993](https://pubmed.ncbi.nlm.nih.gov/34479993/) · PMCID: [PMC8433536](https://pmc.ncbi.nlm.nih.gov/pmcid/PMC8433536/)
24. **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](https://pubmed.ncbi.nlm.nih.gov/12824425/) · PMCID: [PMC168916](https://pmc.ncbi.nlm.nih.gov/pmcid/PMC168916/)
25. **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](https://pubmed.ncbi.nlm.nih.gov/32461654/) · PMCID: [PMC7334197](https://pmc.ncbi.nlm.nih.gov/pmcid/PMC7334197/)
26. **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>

27. **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](https://pubmed.ncbi.nlm.nih.gov/12130785/)
28. **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](https://pubmed.ncbi.nlm.nih.gov/35460607/) · PMCID: [PMC9118112](https://pubmed.ncbi.nlm.nih.gov/PMC9118112/)
29. **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://pubmed.ncbi.nlm.nih.gov/PMC299910/)
30. **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/)
31. **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://pubmed.ncbi.nlm.nih.gov/PMC2360052/)
32. **Resolution of structural variation in diverse mouse genomes reveals chromatin remodeling due to transposable elements**
Ardian Ferraj, Peter A Audano, Parithi Balachandran, Anne Czechanski, Jacob I Flores, Alexander A Radecki, Varun Mosur, David S Gordon, Isha A Walawalkar, Evan E Eichler, ... Christine R Beck
Cold Spring Harbor Laboratory (2022-09-27) <https://doi.org/grtctb>
DOI: [10.1101/2022.09.26.509577](https://doi.org/10.1101/2022.09.26.509577)
33. **A modified fluctuation assay reveals a natural mutator phenotype that drives mutation spectrum variation within**
Pengyao Jiang, Anja R Ollodart, Vidha Sudhesh, Alan J Herr, Maitreya J Dunham, Kelley Harris
eLife (2021-09-15) <https://www.ncbi.nlm.nih.gov/pubmed/34523420>
DOI: [10.7554/elife.68285](https://doi.org/10.7554/elife.68285) · PMID: [34523420](https://pubmed.ncbi.nlm.nih.gov/34523420/) · PMCID: [PMC8497059](https://pubmed.ncbi.nlm.nih.gov/PMC8497059/)
34. **Limited role of generation time changes in driving the evolution of mutation spectrum in humans**
Ziyue Gao, Yulin Zhang, Nathan Cramer, Molly Przeworski, Priya Moorjani
bioRxiv (2023-01-13) <https://doi.org/grr525>
DOI: <https://doi.org/10.1101/2022.06.17.496622>

35. **Mutational Signatures of Replication Timing and Epigenetic Modification Persist through the Global Divergence of Mutation Spectra across the Great Ape Phylogeny.**
Michael E Goldberg, Kelley Harris
Genome biology and evolution (2022-01-04) <https://www.ncbi.nlm.nih.gov/pubmed/33983415>
DOI: [10.1093/gbe/evab104](https://doi.org/10.1093/gbe/evab104) · PMID: [33983415](#) · PMCID: [PMC8743035](#)
36. **Heritability of de novo germline mutation reveals a contribution from paternal but not maternal genetic factors**
Seongwon Hwang, Matthew DC Neville, Genomics England Research Consortium, Felix R Day, Aylwyn Scally
bioRxiv (2022-12-17) <https://doi.org/grr526>
DOI: <https://doi.org/10.1101/2022.12.17.520885>
37. **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](#) · PMCID: [PMC9339295](#)
38. **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](#) · PMCID: [PMC2744436](#)
39. **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](#) · PMCID: [PMC1206241](#)
40. **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](#) · PMCID: [PMC6366910](#)
41. **Array programming with NumPy**
Charles R Harris, Kjarrod Millman, Stéfan J van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J Smith, ... Travis E Oliphant
Nature (2020-09-16) <https://doi.org/ghbf2>
DOI: [10.1038/s41586-020-2649-2](https://doi.org/10.1038/s41586-020-2649-2) · PMID: [32939066](#) · PMCID: [PMC7759461](#)
42. **pandas-dev/pandas: Pandas**
The Pandas Development Team
Zenodo (2023-02-20) <https://doi.org/ggt8bh>
DOI: [10.5281/zenodo.3509134](https://doi.org/10.5281/zenodo.3509134)
43. **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)
44. **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>

45. **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)

46. **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)

47. **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)

48. <https://github.com/quinlan-lab/proj-mutator-mapping>

49. **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](#) · PMCID: [PMC8114187](#)

50. **A natural mutator allele shapes mutation spectrum variation in mice**

Tom Sasani

(2023-01-24) https://github.com/tomsasani/bxd_mutator_manuscript

51. **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)

52. **BXD Genotype / WebQTL** <https://gn1.genenetwork.org/dbdoc/BXDGeno.html>

53. **Genetics and Probability in Animal Breeding Experiments**

<https://link.springer.com/book/10.1007/978-1-349-04904-2>

54. **MGI-Guidelines for Nomenclature of Mouse and Rat Strains**

<http://www.informatics.jax.org/mgihome/nomen/strains.shtml>

55. **Sustainable data analysis with Snakemake.**

Felix Mölder, Kim Philipp Jablonski, Brice Letcher, Michael B Hall, Christopher H Tomkins-Tinch, Vanessa Sochat, Jan Forster, Soohyun Lee, Sven O Twardziok, Alexander Kanitz, ... Johannes Köster

F1000Research (2021-01-18) <https://www.ncbi.nlm.nih.gov/pubmed/34035898>

DOI: [10.12688/f1000research.29032.2](https://doi.org/10.12688/f1000research.29032.2) · PMID: [34035898](#) · PMCID: [PMC8114187](#)

56. **ENA Browser** <https://www.ebi.ac.uk/ena/browser/view/PRJEB45429>

57. **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://pmc.ncbi.nlm.nih.gov/pmcid/PMC7931819/)
58. **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://pmc.ncbi.nlm.nih.gov/pmcid/PMC3679285/)
59. **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://pmc.ncbi.nlm.nih.gov/pmcid/PMC5870853/)
60. **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/)
61. **coxme: Mixed Effects Cox Models**
Terry M Therneau
(2022-10-03) <https://CRAN.R-project.org/package=coxme>