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@26fc714 on April 25, 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 NIH/NHGRI R01HG012252

- **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>, Kelley Harris <harriske@uw.edu>.

Abstract

Maintaining germline genome integrity is essential and enormously complex. Hundreds of proteins are involved in DNA replication and 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 candidate mutator alleles in the DNA repair gene *Mutyh* that increased the C>A germline mutation rate in a family of inbred mice known as the BXDs [2,3].

In this study we developed a new method, called "inter-haplotype distance," to detect alleles associated with mutation spectrum variation. By applying this approach to mutation data from the BXDs, we confirmed the presence 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]. The effect of a chromosome 6 mutator allele depended on the presence of a mutator allele near *Mutyh*, and BXDs with mutator alleles at both loci had even greater numbers of C>A mutations than those with mutator alleles at either locus alone.

Our new methods for analyzing mutation spectra reveal evidence of epistasis between 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], likely due to either genetic variation or differences in environmental exposures. Although numerous protein-coding genes contribute to the maintenance of genome integrity, 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 [13].

Despite these challenges, less traditional strategies have been used to identify a small number of mutator alleles in humans, macaques, and mice [14]. By focusing on families with rare genetic diseases, a recent study discovered two mutator alleles that led to significantly elevated rates of *de novo* germline mutation in human genomes [15]. Another group observed mutator phenotypes in the sperm and somatic tissues of adults who carry cancer-predisposing inherited mutations in the POLE/POLD1 exonucleases [16]. Candidate mutator loci were also found by identifying human

haplotypes from the Thousand Genomes Project with excess counts of derived alleles in genomic windows [17].

In mice, a germline mutator allele was recently discovered by sequencing a large family of inbred mice [2]. Commonly known as the BXDs, these recombinant inbred lines (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 [3,18]. The BXDs were maintained via brother-sister mating for up to 180 generations, and each BXD 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 [19].

In this previous study, whole-genome sequencing data from the BXD family 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 did not reach genome-wide significance in a 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 [20]. Mutation spectra are usually defined in terms of k-mer nucleotide context; 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 can 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, called “inter-haplotype distance,” 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 family and find compelling evidence for a second mutator allele that was not detected 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 inbred mice. Our observation of epistasis suggests that mild DNA repair deficiencies can 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 recombinant inbred lines (RILs) (Figure 1; *Materials and Methods*). 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. Given a population of haplotypes, we assume that each has been genotyped at the same collection of biallelic loci and that each harbors *de novo* mutations which 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. To account for polygenic effects on the mutation process that might be shared between BXDs, we also regress the cosine distance at each marker against the genetic similarity between haplotype groups, and assess significance using the fitted residuals (which we call the "adjusted" cosine distances) (*Materials and Methods*).

Using simulated data, we find that our 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*.

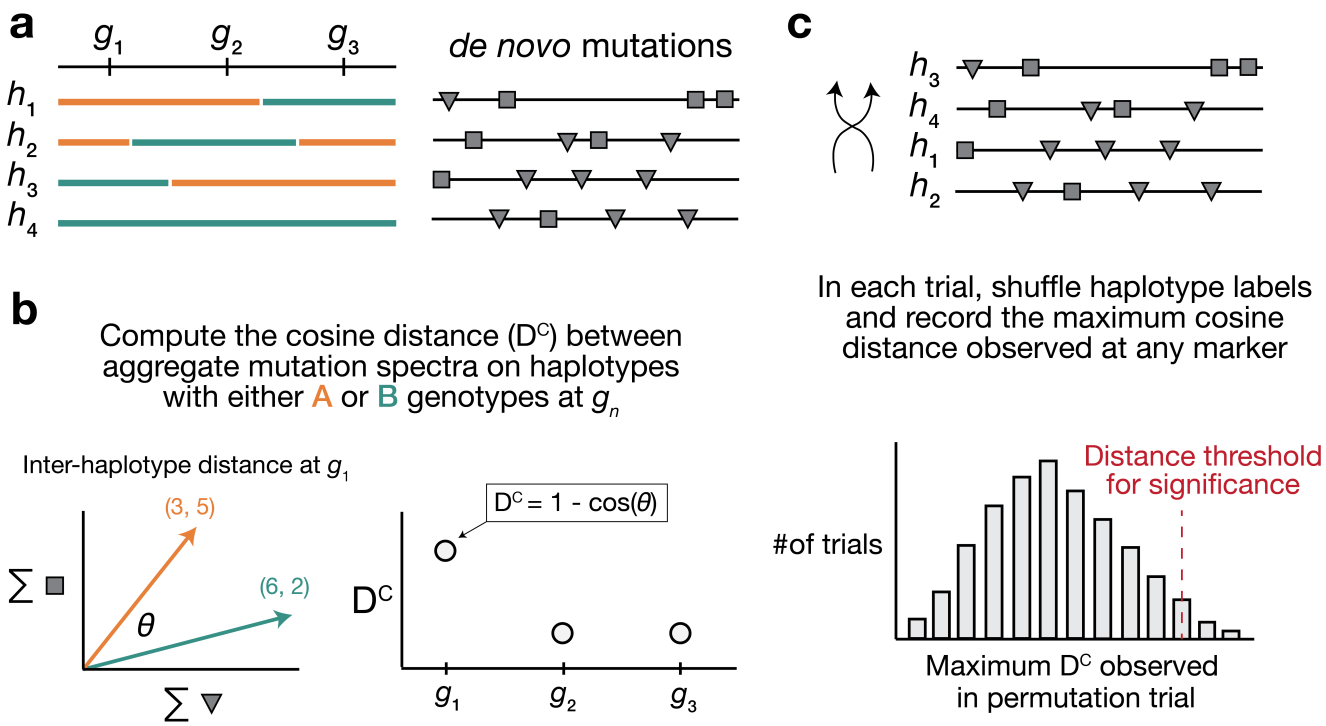

Figure 1: Overview of inter-haplotype distance method for discovering mutator alleles. a) A population of four haplotypes has been genotyped at three informative markers (g_1 through g_3); each haplotype also harbors unique *de novo* germline mutations. In practice, *de novo* mutations are partitioned 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). **b)** 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. For example, haplotypes with *A* (orange) genotypes at g_1 carry a total of three "triangle" mutations and five "square" mutations, and haplotypes with *B* (green) genotypes carry a total of six triangle and two square 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 θ 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. **c)** To assess the significance of any distance peaks in b), 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 a mutator allele on chromosome 4 in the BXDs

We applied our inter-haplotype distance method to 117 BXDs (*Materials and Methods*) with a total of 65,552 *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 1.20e-2 at marker ID `rs27509845`; position 118.28 Mbp in GRCm38/mm10 coordinates).

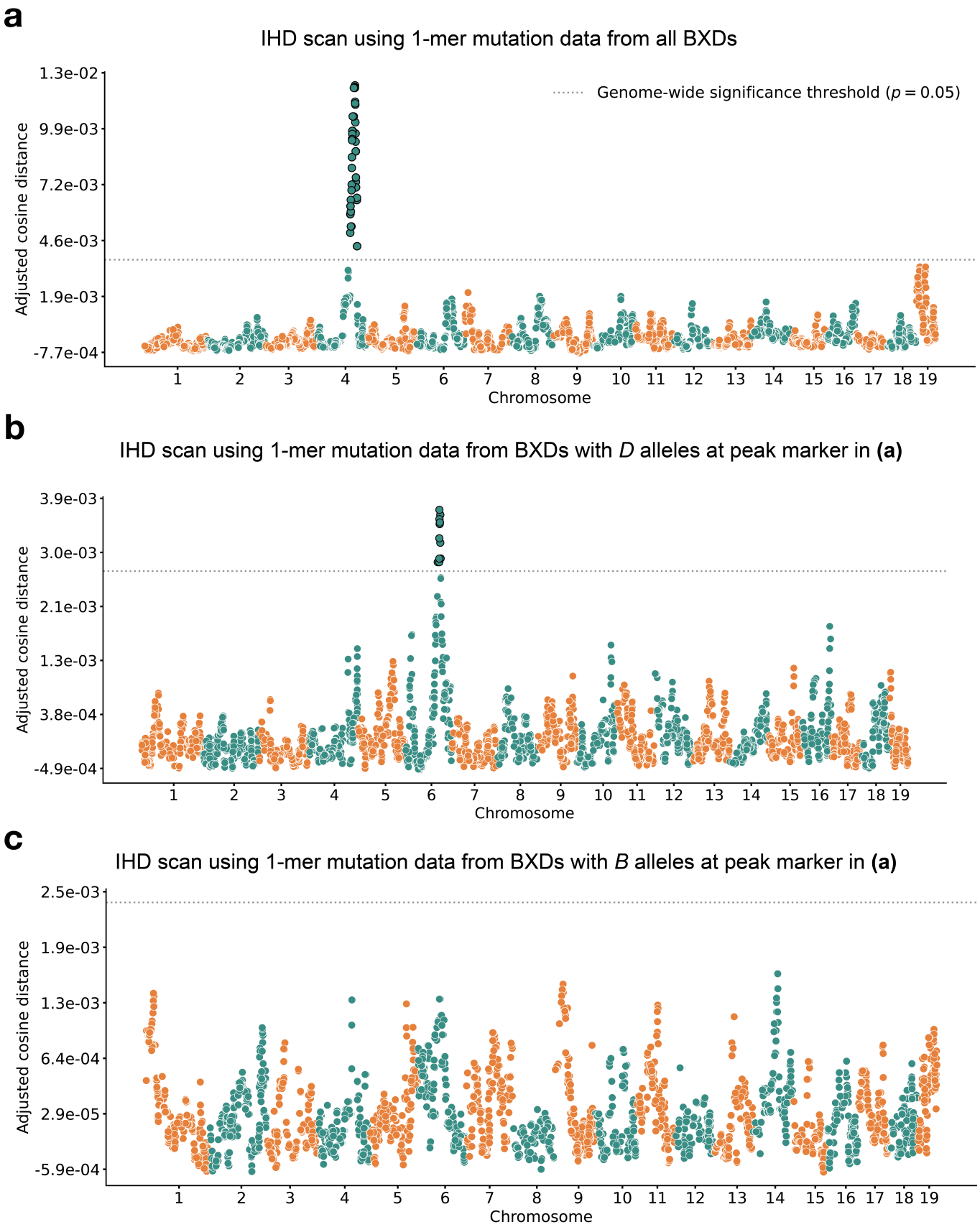

Figure 2: Results of inter-haplotype distance scans in the BXDs. a) Adjusted cosine distances between aggregate 1-mer *de novo* mutation spectra on BXD haplotypes (n = 117 haplotypes; 65,552 total mutations) with either *D* or *B* alleles at 7,321 informative markers. Cosine distance threshold at p = 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 `rs27509845` (n = 66 haplotypes; 42,171 total mutations) and either *D* or *B* alleles at 7,276 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. **c)** Adjusted cosine distances between aggregate 1-mer *de novo* mutation spectra on BXD haplotypes with *B* alleles at `rs27509845` (n = 44

haplotypes; 22,645 total mutations) and either *D* or *B* alleles at 7,273 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 those genes contained nonsynonymous 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 [4,21,22]. C>A germline mutation fractions are nearly 50% higher in BXDs that inherit *D* genotypes at marker ID `rs27509845` (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 BXDs into those with either *D* (n = 66) or *B* (n = 44) genotypes at `rs27509845` (n = 7 BXDs were heterozygous) and ran an inter-haplotype distance scan using each group separately (Figure 2b-c).

Using the BXDs with *D* genotypes at `rs27509845`, we identified a locus on chromosome 6 that was significantly associated with mutation spectrum variation (Figure 2b; maximum adjusted cosine distance of 3.68e-3 at marker `rs46276051`; position 111.27 Mbp in GRCm38/mm10 coordinates). This signal was specific to BXDs with *D* genotypes at the `rs27509845` locus, as we did not observe any new mutator loci after performing an IHD scan using BXDs with *B* genotypes at `rs27509845` (Figure 2c). We also performed QTL scans for the fractions 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 1; *Materials and Methods*).

We queried the region surrounding the top marker on chromosome 6 (+/- 5 Mbp) and discovered 63 protein-coding genes, of which five were annotated with a Gene Ontology [23,24] term related to "DNA repair": *Fancd2*, *Mbd4*, *Ogg1*, *Rad18*, and *Setmar*. Of these, only three harbored nonsynonymous differences between the parental C57BL/6J and DBA/2J strains (Table 1). *Ogg1* encodes a key member of the base-excision repair response to oxidative DNA damage (a pathway that also includes *Mutyh*), *Mbd4* encodes a protein that is involved in the repair of G:T mismatches caused by spontaneous deamination of methylated CpGs, and in mice *Setmar* encodes a SET domain-containing histone methyltransferase.

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
Setmar	ENSMUST00000049246	C>T	p.Leu103Phe	chr6:108,075,853	0.422	0.0 (intolerant/deleterious)

Gene name	Ensembl transcript name	Nucleotide change	Amino acid change	Position in GRCm38/mm10 coordinates	PhyloP conservation score	SIFT prediction
Setmar	ENSMUST00000049246	T>G	p.Ser273Arg	chr6:108,076,365	-0.355	0.3 (tolerant/benign)
Ogg1	ENSMUST00000032406	A>G	p.Thr95Ala	chr6:113,328,510	-0.016	0.84 (tolerant/benign)
Mbd4	ENSMUST00000032469	C>T	p.Asp129Asn	chr6:115,849,644	2.28	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 [25] we mapped cis-eQTLs for the five aforementioned DNA repair genes in a number of tissues, though we did not have access to expression data from germline cells. *D* alleles near the cosine distance peak on chromosome 6 were significantly associated with decreased *Ogg1* expression in kidney, liver, and gastrointestinal tissues; *D* alleles were also associated with lower *Rad18* expression in the retina (Table supplement 1). The cis-eQTLs that lead to decreased *Ogg1* or *Rad18* expression are challenging to interpret, however, given their tissue specificity and our lack of access to germline expression data.

Finally, we queried a dataset of structural variants (SVs) identified via high-quality, long-read assembly of inbred laboratory mouse strains [26] and found 136 large insertions or deletions within 5 Mbp of the cosine distance peak on chromosome 6. Of these, four 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, additional experimental evidence will be needed to probe the functional impact of these SVs.

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 significantly 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). Among BXDs with *B* alleles at the locus overlapping *Mutyh*, those with *D* alleles on chromosome 6 did not exhibit significantly elevated C>A mutation fractions (Figure 3a). After controlling for inbreeding duration, we observed that C>A *de novo* mutation counts were always highest in BXDs with *D* alleles at both mutator loci (Figure 3b). After 100 generations of inbreeding, BXDs with *D* alleles at both mutator loci were predicted to have 238.9 C>A mutations (95% CI: 231.4 - 246.4), about 20% more than the 199.0 mutations (95% CI: 193.3 - 204.7) predicted in those with *D* and *B* alleles at the chromosome 4 and chromosome 6 loci, respectively (Figure 3b).

We also used SigProfilerExtractor [27] to assign the germline mutations in each BXD to single-base substitution (SBS) mutation signatures from the COSMIC catalog [28]. Mutation signatures often reflect specific exogenous or endogenous sources of DNA damage, and the proportions of mutations attributable to particular SBS signatures 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 3c). 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 3c). SBS18 activity was lowest in mice with *D* alleles at the chromosome 6 mutator locus alone (Figure 3c), further demonstrating that *D* alleles at this locus are not sufficient to cause a mutator phenotype.

To more formally test for statistical epistasis, we fit a generalized (Poisson) linear model predicting counts of C>A mutations in each BXD as a function of genotypes at `rs27509845` and `rs46276051` (the markers with the largest adjusted cosine distance at the two mutator loci); the model also accounted for differences in inbreeding duration and sequencing coverage between the BXDs (*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 = 7.92e-7; *Materials and Methods*).

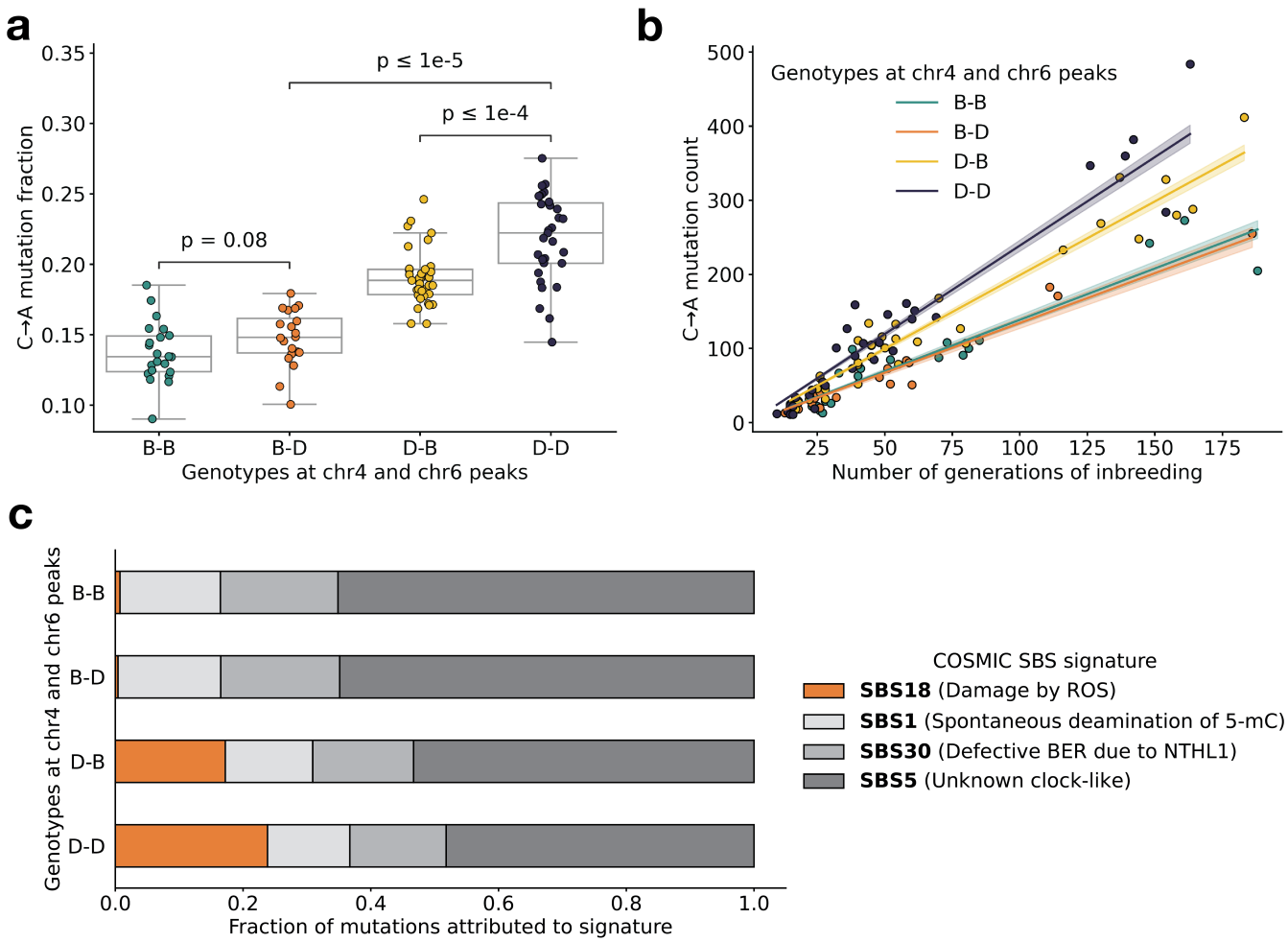

Figure 3: BXD mutation spectra are affected by alleles at both mutator loci. a) C>A *de novo* germline mutation fractions in BXDs with either *D* or *B* genotypes at markers `rs27509845` (chr4 peak) and `rs46276051` (chr6 peak). Distributions of C>A mutation fractions were compared with two-sided Mann-Whitney U-tests; annotated p-values are uncorrected. `B−B` vs. `B−D` comparison: U-statistic = 149.0, p = 7.58e-2; `B−D` vs `D−D` comparison: U-statistic = 21.0, p = 2.61e-8; `D−B` vs `D−D` comparison: U-statistic = 232.5, p = 6.99e-5. **b)** The count of C>A *de novo* germline mutations in each BXD was plotted against the number of generations for which it was inbred. Lines represent predicted C>A counts in each haplotype group from a generalized linear model (Poisson family, 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 [27]. After grouping BXDs by their genotypes at `rs27509845` and `rs46276051`, 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, related to SBS36 and defective base-excision repair due to 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 [29]. Dumont [18] 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>A mutation counts as a function of genotypes at the two mutator loci. A model incorporating an interaction term 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.

Some of the candidate mutator alleles are segregating 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 [30]. 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 [31], and was fixed in *Mus musculus musculus*, *Mus musculus castaneus*, and the outgroup *Mus spretus* (Figure 3-figure supplement 3). The *Setmar* p.Ser273Arg *D* allele was also present at an allele frequency of 0.37 in *Mus musculus domesticus*, while *D* alleles at the *Setmar* p.Leu103Phe variant were not observed in any wild *Mus musculus domesticus* animals. Notably, *D* alleles at the *Mbd4* p.Asp129Asn variant were absent from all wild mouse populations (Figure 3-figure supplement 3).

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 family of inbred mice derived from the laboratory strains 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 [32] and in human cell lines [33], but never to our knowledge in a whole-animal context.

Importantly, we discovered epistasis between germline mutator alleles 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 [19], 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 also suggests that mutational pressure can cause mutation rates to rise in just a few generations of relaxed selection in captivity. This hypothesis is corroborated by the recent discovery of a large-effect mutator allele in a rhesus macaque research colony [14], as well as the observation that domesticated animals tend to have higher mutation rates than those in the wild [34].

Possible causal alleles underlying the chromosome 6 mutator locus

Three protein-coding genes involved in DNA repair 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], *Mbd4*, a glycosylase that removes thymine nucleotides at G:T mispairs following spontaneous deamination of methylated CpGs, and *Setmar*, a histone methyltransferase involved in non-homologous end joining (NHEJ) of double-stranded breaks (DSBs) [35,36].

Both *Ogg1* and *Mbd4* 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 [37,38], and copy-number losses of either *Ogg1* or *Mutyh* are linked to elevated rates of spontaneous C>A mutation in human neuroblastoma [39]. 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* [40]. Similarly, biallelic loss of *MBD4* function was recently found to cause a hypermutator phenotype in the maternal germline of rhesus macaques [14]; the macaque germline mutator phenotype primarily comprised C>T mutations at CpG and CpA sites, a mutational signature that has previously been associated with LOF mutations in *Mbd4* [41].

Although we are unable to conclusively determine that either *Setmar*, *Ogg1*, or *Mbd4* harbors the causal variant underlying the observed C>A mutator phenotype, we believe that *Ogg1* and *Mbd4* are the most plausible candidates. The primate *SETMAR* ortholog is involved in NHEJ of double-strand breaks, but its role in DNA repair appears to depend on the function of both a SET methyltransferase domain and a *Mariner*-family transposase domain [36,42,43]. Since the murine *Setmar* ortholog lacks the latter element, we believe it is unlikely to underlie the epistatic interaction between the chromosome 4 and 6 mutator loci in the BXDs (*Supplementary Information*).

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 [44,45]. 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 [44]. 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 [46]) accumulate a 100-fold excess of 8-oxoG in their gonadal cells [45]. 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 the *in silico* tool SIFT [47], and occurs at a nucleotide that is not particularly well-conserved across mammalian species (Table 1). We also observe that the *D* allele at p.Thr95Ala is segregating at an allele frequency of approximately 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 should 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 gene network when *Mutyh* function is also impaired. It is also possible that *D* alleles at *Ogg1* lead to a very subtle increase in C>A mutation rates, and we are simply underpowered to detect such a small mutation rate effect in the BXDs. 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* may be an even more compelling candidate mutator allele.

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 [47] (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) [48] and is predicted by SIFT and Polyphen [49] to be “deleterious” and “probably_damaging” in human genomes, respectively.

One puzzling observation 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* have also been shown 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 chemotherapeutics and mutagenic compounds [50]. 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* [51]. 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 [41], 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 [52].

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 [50,51]. 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 partially 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 mutator loci described in this paper.

No indication of causal 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* [26], 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 on haplotypes 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 (for example, in which RILs have been inbred for different lengths of time and carry varying numbers of mutations that can be leveraged for mutator 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 BXDs 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 [53]) 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 methods to account for inter-sample relatedness in the IHD approach (*Materials and Methods*), 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 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 [6,7,54], 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 potentially 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, but a subset may be detectable by sequencing a sufficient number of human trios [55]. Since germline mutators often seem to exert their effects on a small number of mutation types, mutation spectrum analyses may have greater power to detect the genes that underlie heritable mutation rate variation, even if each gene has only a modest effect on the overall mutation rate per generation.

Materials and Methods

Identifying *de novo* germline mutations in the BXDs

The BXD resource currently comprises a total of 152 recombinant inbred lines (RILs). BXDs 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 BXDs, 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 [56].

At each marker, we then compute the cosine distance between the two aggregate spectra. The cosine distance between two vectors $\mathbf{A}$ and $\mathbf{B}$ is defined as

$$D^C = 1 - \frac{\mathbf{A} \cdot \mathbf{B}}{||\mathbf{A}||\ ||\mathbf{B}||}$$

where $||\mathbf{A}||$ and $||\mathbf{B}||$ are the L^2 (or Euclidean) norms of $\mathbf{A}$ and $\mathbf{B}$, respectively. The cosine distance metric has a number of favorable properties for comparing mutation spectra. Since 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 [53,57], 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 to derive approximately 50% of its genome from C57BL/6J and 50% from DBA/2J. As a result, every pair of BXDs will likely have identical genotypes at a fraction of 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 BXDs that inherited *D* alleles are more genetically dissimilar from those 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 $\left(D_1^C, D_2^C, \ldots D_n^C\right)$ on $(r_1, r_2, \ldots r_n)$ 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.

Accounting for BXD population structure due to breeding epochs

The current BXD family was generated in six breeding "epochs." As discussed previously, each epoch was initiated with a distinct cross of C57BL/6J and DBA/2J parents; BXDs in four of the epochs were generated following F2 crosses of C57BL/6J and DBA/2J, and BXDs in the other two were generated following advanced intercrosses. Due to this breeding approach the BXD epochs differ from each other in a few important ways. For example, BXDs derived in epochs 3 and 5 (i.e., from advanced intercross) harbor larger numbers of fixed recombination breakpoints than those from epochs 1, 2, 4, and 6 [3]. Although the C57BL/6J and DBA/2J parents used to initialize each epoch were completely inbred, they each possessed a small number unique *de novo* germline mutations that were subsequently inherited by many of their offspring. A number of these "epoch-specific" variants have also been linked to phenotypic variation observed between BXDs from different epochs [3,58,59,60].

To account for potential population structure, as well as these epoch-specific effects, we introduced the ability to perform stratified permutation tests in the inter-haplotype distance approach. Normally, in each of *N* permutations we shuffle the per-haplotype mutation spectrum data such that haplotype labels no longer correspond to the correct mutation spectra (i.e., shuffle mutation spectra *across* epochs). In the stratified approach, we instead shuffle per-haplotype mutation data *within* epochs, preserving epoch structure while still enabling mutation spectra permutations.

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` [61,62,63,64,65,66,67].

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 [68] 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 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 N<u>C</u>N>N<u>T</u>N 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 (λ'), in which we multiply the λ 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 λ', while all of the haplotypes with 0 alleles should have "wild-type" rates of that mutation type and draw their mutation counts from λ. 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 λ'. 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 λ.

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 [53] 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 in our simulations, 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 [2,69,70], and downloaded a CSV file of BXD genotypes at ~7,300 informative markers from GeneNetwork [25,71]. We also downloaded relevant metadata about each BXD 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. Specifically, we removed BXDs that were backcrossed to a C57BL/6J or DBA/2J parent at any point during the inbreeding process (usually, in order to rescue that BXD from inbreeding depression [3]). We also removed BXD68 from our genome-wide scans, since we previously discovered a hyper-mutator phenotype in that line; 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 our previous manuscript, 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 [72]. 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. In other words, the *de novo* mutations present in the genome of a "young" BXD (i.e., a BXD that was inbred for fewer than 20 generations) would not reflect a mutator allele's activity as strongly as the mutations present in the genome of a much older BXD. This presented a challenge when we used quantitative trait locus mapping to discover mutator alleles in our previous manuscript, since the phenotypes (i.e., C>A mutation rates) of young and old BXDs were weighted equally; thus, we simply removed the younger BXDs from our analysis to avoid using their especially noisy mutation spectra. Since IHD computes an *aggregate* mutation spectrum using all BXDs that inherited a particular allele at a locus, and can overcome the sparsity and noise of individual mutation spectra, we chose to include these younger BXDs in our genome-wide scans in this study.

In total, we included 117 BXDs 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 [73]) using `bcftools` [74]. 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 (`rs46276051`). 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` [75] command:

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

and used `cyvcf2` [76] 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 [26]) 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 [77], and used the `bx-python` library [78] 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 `rs46276051`.

Extracting mutation signatures

We used SigProfilerExtractor (v.1.1.21) [28] 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 [18] (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 [29]. 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 [25], 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 [79] 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
Kidney	`Mouse kidney M430v2 Sex Balanced (Aug06) RMA`
Gastrointestinal	`UTHSC Mouse BXD Gastrointestinal Affy MoGene 1.0 ST Gene Level (Apr14) RMA`
Hematopoetic stem cells	`UMCG Stem Cells ILM6v1.1 (Apr09) transformed`

Tissue name	Complete name of GeneNetwork expression data
Hematopoetic progenitor cells	`UMCG Progenitor Cells ILM6v1.1 (Apr09) transformed`
Spleen	`UTHSC Affy MoGene 1.0 ST Spleen (Dec10) RMA`
Liver	`UTHSC BXD Liver RNA-Seq Avg (Oct19) TPM Log2`
Heart	`NHLBI BXD All Ages Heart RNA-Seq (Nov20) TMP Log2 **`
Eye	`UTHSC BXD All Ages Eye RNA-Seq (Nov20) TPM Log2 **`

Calculating the frequencies of candidate mutator alleles in wild mice

To determine the frequencies of the *Ogg1*, *Mbd4*, and *Setmar* nonsynonymous 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* [30]. 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 statistical 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. `ADJ_AGE` is the product of the number of "callable" cytosine/guanine nucleotides in each BXD (i.e., the total number of cytosines/guanines covered by at least 10 sequencing reads) and the number of generations for which the BXD was inbred. We included the logarithm of `ADJ_AGE` as an "offset" in order to model the response variable as a rate (expressed per base-pair, per generation) 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 BXD. The `Genotype_A` and `Genotype_B` terms represent the genotypes of BXDs at markers `rs27509845` and `rs46276051` (the markers with peak cosine distances on chromosomes 4 and 6 in the two inter-haplotype distance scans). We limited our analysis to the n = 108 BXDs that were homozygous at both sites, allowing us to model genotypes at either locus as binary variables ("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 [18]. 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 the chromosome 4 and chromosome 6 mutator loci, respectively. We compared the two models using ANOVA as described above.

Acknowledgments

We thank Robert W. Williams (University of Tennessee Health Sciences Center) and Don F. Conrad (Oregon Health & Science University) for very helpful comments and feedback on a draft of this manuscript.

References

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

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

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

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

5. **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 · PMID: 10097147 · PMCID: PMC22404

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

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

8. **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 · PMID: 35771663 · PMCID: 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 · 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 · 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 · 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 · PMID: 27739533

13. **De novo mutations across 1,465 diverse genomes reveal mutational insights and reductions in the Amish founder population.**
Michael D Kessler, Douglas P Loesch, James A Perry, Nancy L Heard-Costa, Daniel Taliun, Brian E Cade, Heming Wang, Michelle Daya, John Ziniti, Soma Datta, ... Timothy D O'Connor
Proceedings of the National Academy of Sciences of the United States of America (2020-01-21) https://www.ncbi.nlm.nih.gov/pubmed/31964835
DOI: 10.1073/pnas.1902766117 · PMID: 31964835 · PMCID: PMC7007577

14. **A naturally occurring variant ofMBD4causes maternal germline hypermutation in primates**
Alexandra M Stendahl, Rashesh 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

15. **Genetic and chemotherapeutic influences on germline hypermutation.**
Joanna Kaplanis, Benjamin Ide, Rashesh 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 · PMID: 35545669 · PMCID: PMC9117138

16. **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 · PMID: 34594041 · PMCID: PMC8492474

17. **Inference of Candidate Germline Mutator Loci in Humans from Genome-Wide Haplotype Data.**

Cathal Seoighe, Aylwyn Scally
PLoS genetics (2017-01-17) https://www.ncbi.nlm.nih.gov/pubmed/28095480
DOI: 10.1371/journal.pgen.1006549 · PMID: 28095480 · PMCID: PMC5283766

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

19. **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

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

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

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

23. **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.**
M Ashburner, CA Ball, JA Blake, D Botstein, H Butler, JM Cherry, AP Davis, K Dolinski, SS Dwight, JT Eppig, … G Sherlock
Nature genetics (2000-05) https://www.ncbi.nlm.nih.gov/pubmed/10802651
DOI: 10.1038/75556 · PMID: 10802651 · PMCID: PMC3037419

24. **The Gene Ontology resource: enriching a GOld mine.** *Nucleic acids research* (2021-01-08)
https://www.ncbi.nlm.nih.gov/pubmed/33290552
DOI: 10.1093/nar/gkaa1113 · PMID: 33290552 · PMCID: PMC7779012

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

26. **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

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

28. **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 · PMID: 30371878 · PMCID: PMC6323903

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

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

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

32. **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 · PMID: 16492773 · PMCID: PMC1413905

33. **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 · PMID: 35859169 · PMCID: PMC9329121

34. **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 · PMID: 36859541 · PMCID: PMC9995274

35. **Methylation of histone H3 lysine 36 enhances DNA repair by nonhomologous end-joining.**
Sheema Fnu, Elizabeth A Williamson, Leyma P De Haro, Mark Brenneman, Justin Wray, Montaser Shaheen, Krishnan Radhakrishnan, Suk-Hee Lee, Jac A Nickoloff, Robert Hromas
Proceedings of the National Academy of Sciences of the United States of America (2010-12-27) https://www.ncbi.nlm.nih.gov/pubmed/21187428
DOI: 10.1073/pnas.1013571108 · PMID: 21187428 · PMCID: PMC3021059

36. **The SET domain protein Metnase mediates foreign DNA integration and links integration to nonhomologous end-joining repair.**
Suk-Hee Lee, Masahiko Oshige, Stephen T Durant, Kanwaldeep Kaur Rasila, Elizabeth A Williamson, Heather Ramsey, Lori Kwan, Jac A Nickoloff, Robert Hromas
Proceedings of the National Academy of Sciences of the United States of America (2005-12-06) https://www.ncbi.nlm.nih.gov/pubmed/16332963
DOI: 10.1073/pnas.0503676102 · PMID: 16332963 · PMCID: PMC1312370

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

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

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

40. **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 · PMID: 35460607 · PMCID: PMC9118112

41. **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 · PMID: 12130785

42. **The DDN catalytic motif is required for Metnase functions in non-homologous end joining (NHEJ) repair and replication restart.**
Hyun-Suk Kim, Qiujia Chen, Sung-Kyung Kim, Jac A Nickoloff, Robert Hromas, Millie M Georgiadis, Suk-Hee Lee

The Journal of biological chemistry (2014-02-25)
https://www.ncbi.nlm.nih.gov/pubmed/24573677
DOI: 10.1074/jbc.m113.533216 · PMID: 24573677 · PMCID: PMC4036204

43. **Biochemical characterization of metnase's endonuclease activity and its role in NHEJ repair.**
Brian D Beck, Sung-Sook Lee, Elizabeth Williamson, Robert A Hromas, Suk-Hee Lee
Biochemistry (2011-04-27) https://www.ncbi.nlm.nih.gov/pubmed/21491884
DOI: 10.1021/bi200333k · PMID: 21491884 · PMCID: PMC3388547

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

45. **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 · PMID: 24732879 · PMCID: PMC3986730

46. **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

47. **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 · PMID: 12824425 · PMCID: PMC168916

48. **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 · PMID: 32461654 · PMCID: PMC7334197

49. **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 · PMID: 20354512 · PMCID: PMC2855889

50. **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 · PMID: 14614141 · PMCID: PMC299910

51. **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 · PMID: 14562041

52. **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 · PMID: 17285135 · PMCID: PMC2360052

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

54. **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, Eirene 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 · PMID: 29700473 · PMCID: PMC5961723

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

56. **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 · PMID: 19488047 · PMCID: PMC2744436

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

58. Alexandra M Stafford, Cheryl Reed, Harue Baba, Nicole Ar Walter, John Rk Mootz, Robert W Williams, Kim A Neve, Lev M Fedorov, Aaron J Janowsky, Tamara J Phillips
eLife (2019-07-09) https://www.ncbi.nlm.nih.gov/pubmed/31274109
DOI: 10.7554/elife.46472 · PMID: 31274109 · PMCID: PMC6682400

59. **Identification of a Functional Non-coding Variant in the GABA**
Megan K Mulligan, Timothy Abreo, Sarah M Neuner, Cory Parks, Christine E Watkins, MTrevor Houseal, Thomas M Shapaker, Michael Hook, Haiyan Tan, Xusheng Wang, … Robert W Williams
Frontiers in genetics (2019-03-29) https://www.ncbi.nlm.nih.gov/pubmed/30984232
DOI: 10.3389/fgene.2019.00188 · PMID: 30984232 · PMCID: PMC6449455

60. **Investigating a downstream gene of**
Ye Lu, Diana Zhou, Hong Lu, Fuyi Xu, Junming Yue, Jianping Tong, Lu Lu
Molecular vision (2019-04-23) https://www.ncbi.nlm.nih.gov/pubmed/31057322
PMID: 31057322 · PMCID: PMC6478243

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

62. **pandas-dev/pandas: Pandas**
The Pandas Development Team
Zenodo (2023-04-03) https://doi.org/ggt8bh
DOI: 10.5281/zenodo.3509134

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

64. **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

65. **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

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

67. **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

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

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

70. **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

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

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

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

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

75. **A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3.**
Pablo Cingolani, Adrian Platts, Le Lily Wang, Melissa Coon, Tung Nguyen, Luan Wang, Susan J Land, Xiangyi Lu, Douglas M Ruden
Fly (2012) https://www.ncbi.nlm.nih.gov/pubmed/22728672
DOI: 10.4161/fly.19695 · PMID: 22728672 · PMCID: PMC3679285

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

77. **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 · PMID: 14681465 · PMCID: PMC308837

78. **bx-python**
Taylor Lab at Johns Hopkins University
(2023-04-25) https://github.com/bxlab/bx-python

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

80. **Birth of a chimeric primate gene by capture of the transposase gene from a mobile element.**
Richard Cordaux, Swalpa Udit, Mark A Batzer, Cédric Feschotte
Proceedings of the National Academy of Sciences of the United States of America (2006-05-03)
https://www.ncbi.nlm.nih.gov/pubmed/16672366
DOI: 10.1073/pnas.0601161103 · PMID: 16672366 · PMCID: PMC1472436

81. **The roles of the human SETMAR (Metnase) protein in illegitimate DNA recombination and non-homologous end joining repair.**
Michael Tellier, Ronald Chalmers
DNA repair (2019-06-19) https://www.ncbi.nlm.nih.gov/pubmed/31238295
DOI: 10.1016/j.dnarep.2019.06.006 · PMID: 31238295 · PMCID: PMC6715855