Discovering epistasis between germline mutator alleles in mice

Authors

- **Thomas A. Sasani**
  ⓘD [0000-0003-2317-1374](#) · ○ [tomsasani](#) · 🐦 [tomsasani](#)
  Department of Human Genetics, University of Utah · Funded by Grant XXXXXXXX

- **Kelley Harris**
  ⓘD [0000-0003-0302-2523](#) · 🐦 [Kelley__Harris](#)
  Department of Genome Sciences, University of Washington

- **Aaron R. Quinlan** ✉
  ⓘD [0000-0003-1756-0859](#) · 🐦 [aaronquinlan](#)
  Department of Human Genetics, University of Utah; Department of Biomedical Informatics, University of Utah

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

Abstract

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

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

In this study, we developed a new method to detect alleles that affect the mutation spectrum in two-parent RILs. By applying this method to mutation data from the BXDs, we confirmed the activity of the germline mutator allele near *Mutyh*, and discovered an additional C>A germline mutator locus on chromosome 6 that overlaps *Ogg1*, a key partner of *Mutyh* in base-excision repair of oxidative DNA damage.

Strikingly, BXDs with the mutator allele near *Ogg1* do not exhibit elevated rates of C>A germline mutation unless they also possess the mutator allele near *Mutyh*. However, BXDs with both alleles exhibit even higher C>A mutation rates than those with either one alone.

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

Introduction

Maintaining genome integrity in the mammalian germline is essential and enormously complex. Hundreds of protein-coding genes contribute to pathways involved in DNA replication, and hundreds more are mobilized in response to damage by exogenous and endogenous mutagens [1]. Despite this abundance of potential targets, *mutator alleles* that augment the germline mutation rate have largely eluded detection in mammals.

Germline mutator alleles are difficult to detect for a number of reasons, including the fidelity of germline genome replication and the effects of selection on mutators. On average, humans are born with about 70 to 100 *de novo* germline mutations per diploid genome [2,3]; in mice, the number is closer to 20 or 30 [4]. Moreover, in a population of sufficiently large N_e, we would also expect many mutator alleles to be efficiently selected against. The selection coefficient on a mutator allele is estimated to be approximately $2s\Delta U$ [5], where s is the mean selective coefficient on a new deleterious mutation and ΔU is the excess number of new deleterious mutations caused by the mutator allele; the product of s and ΔU is multiplied by 2 to account for the expected number of generations for which mutator will be linked to the excess mutations it causes. Given the low germline *de novo* mutation rate in mammalian genomes and the potential strength of selection on a mutator, we would likely require a very large number of offspring, as well as an environment that attenuates the effects of selection, in order to detect the effects of a germline mutator allele.

In general, we would expect haplotypes that harbor mutator alleles at a particular locus to carry an excess of total germline mutations, compared to those that harbor wild-type alleles. However, protein-coding genes involved in DNA replication and repair often recognize particular sequence motifs or excise lesions at specific nucleotides [1]. Thus, we might also expect that the spectrum of de novo mutations – i.e, the frequency of each individual mutation type (C>T, A>G, etc.) – will differ between genomes that harbor either a mutator or wild-type allele at a given locus.

In 2022, we discovered a germline mutator allele in mice by analyzing whole-genome sequencing data from 152 recombinant inbred lines (RILs). These RILs, known as the **B**X**D**s [6], were derived from C57**B**L/6J and **D**BA/2J, two laboratory strains that exhibit significant differences in their germline mutation spectra [7]. Following either F2 or advanced intercrosses of the parental strains, the BXDs were inbred by brother-sister mating for up to 180 generations, attenuating the effects of natural selection on both standing and new variation. Over the course of inbreeding, each BXD therefore accumulated hundreds or thousands of germline de novo mutations on a nearly-homozygous linear mosaic of the parental haplotypes. Previously, we identified up to 2,000 germline de novo mutations in each line and used quantitative trait locus (QTL) mapping to identify a locus on chromosome 4 that was strongly associated with the C>A germline mutation rate [8]. The QTL overlapped *Mutyh*, which encodes a protein that normally prevents C>A mutations by repairing oxidative DNA damage [9], and we hypothesized that missense mutations in *Mutyh* were responsible for a 50% increase in the C>A mutation rate between BXDs with either parental haplotype at the QTL.

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

Materials and Methods

Identifying *de novo* germline mutations in the BXD RILs

The BXD resource currently comprises a total of 152 recombinant inbred lines (RILs). RILs were derived from either F2 or advanced intercrosses, and subsequently inbred by brother-sister mating for up to 180 generations [6]. Previously, we analyzed whole-genome sequencing data from the BXDs and identified *de novo* germline mutations in each line [8]. A detailed description of the methods used for DNA extraction, sequencing, alignment, and variant processing, as well as the characteristics of the high-quality *de novo* mutations, are available in the previous manuscript [8].

Briefly, we aligned paired-end Illumina sequencing data to the GRCm38/mm10 reference genome using the 10X LongRanger software (v2.1.6; Lariat approach), called single-nucleotide variants using GATK version (v3.8-1-0-gf15c1c3ef) [10], and identified mutations in each BXD that were absent from all other RILs, as well as from the C57BL/6J and DBA/2J parents. We required each private variant to be meet the following criteria:

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

- supported by at least 10 sequencing reads

- Phred-scaled genotype quality of at least 20

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

- must occur on a parental haplotype that was inherited by at least one other BXD at the same locus, and must be homozygous for the reference allele in those BXDs

A new approach to discover germline mutator alleles

Using the existing catalog of *de novo* germline mutations in the BXDs, we developed a new approach to discover loci that affect the germline *de novo* mutation spectrum in biparental RILs (Figure 1).
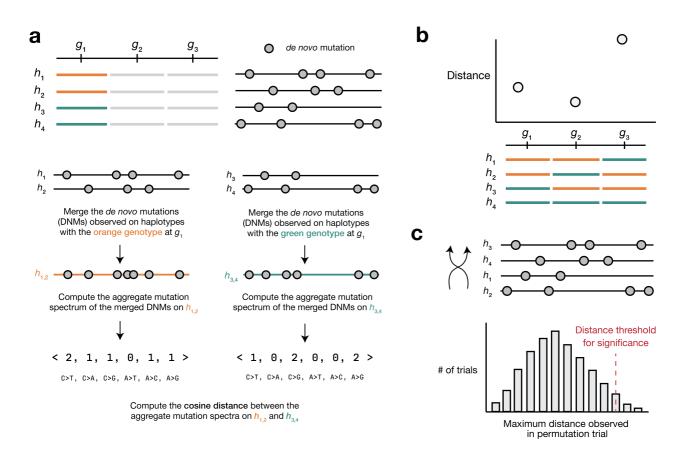

Figure 1: **Overview of inter-haplotype distance method for discovering mutator alleles. a)** A population of four haplotypes have been genotyped at three informative markers, and each harbors a set of private *de novo* germline mutations. At each informative marker, we compute an aggregate *de novo*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

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

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

At each locus, we then calculate the cosine distance between the aggregate mutation spectra of haplotypes with either parental allele. The cosine distance between two vectors $\mathbf{A}$ and $\mathbf{B}$ is defined

as

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

where $||\mathbf{A}||$ and $||\mathbf{B}||$ are the L_2 norms of $\mathbf{A}$ and $\mathbf{B}$, respectively. The cosine distance metric has a number of favorable properties for comparing mutation spectra. Since cosine distance does not take the magnitude of vectors into account, it can be used to compare two spectra with unequal total mutation counts. 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.

We use permutation tests to establish genome-wide cosine distance thresholds. In each of N permutation trials, we randomly shuffle the per-haplotype mutation data such that haplotype labels no longer correspond to the correct mutation counts. Using the shuffled mutation data, we perform a genome-wide distance scan as described above, and record the maximum distance observed at any locus. After N permutations (usually 1,000 or 10,000), we compute the $1 - p$ percentile of the maximum distance distribution, and use that percentile value as a genome-wide significance threshold (for example, at $p = 0.05$).

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

We performed a series of simple simulations to estimate our power to detect alleles that affect the germline mutation spectrum in biparental RILs.

First, we simulate the k-mer mutation spectrum in a population of H haplotypes. We assume that 50% of the haplotypes are under the effects of a mutator allele that increases the mutation rate of a particular mutation type(s) by an effect size E. We simulate M mutations on each haplotype by taking draws from a Poisson distribution as follows:

We first define a vector of mutation probabilities:

$$\mathbf{P} = (0.4,\ 0.1,\ 0.075,\ 0.075,\ 0.075,\ 0.275)$$

These probabilities sum to 1 and correspond to the expected frequencies of C>T, C>A, C>G, A>T, A>C, and A>G *de novo* germline mutations in mice, respectively [4].

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

To simulate the mutation spectrum on each wild-type haplotype, we define a vector of lambda values by scaling the mutation probabilities by the number of mutations we wish to simulate:

$$\lambda = \mathbf{P}M$$

and take a Poisson draw from this vector of lambda values.

To simulate the mutation spectrum on each mutator haplotype, we multiply the mutation probability of a particular mutation type (or multiple mutation types) $\mathbf{P}_n$ by the mutator effect size E. When $k = 1$, we only augment the effect size of one mutation type at a time, but when $k = 3$, we augment a fraction (25%, 50%, or 100%) of the 3-mer mutation types associated with a single "base" mutation type. Then, we define the vector of lambda values by scaling the mutation probabilities by M and take a Poisson draw from that vector on each haplotype.

After generating "mutator" and "wild-type" haplotypes, we randomly shuffle the simulated haplotypes $N = 10,000$ times and compute the cosine distance between wild-type and mutator haplotypes each time. If fewer than 5% of the N permutations produces a cosine distance greater than or equal to the cosine distance between the "true" wild-type and mutator haplotypes, we say that the approach succesfully identified the mutator allele. For every combination of simulation parameters (H, M, E, and so on) we perform 100 trials and record the number of trials in which we successfully identify the mutator allele.

Results

Benchmarking the inter-haplotype distance method using simulations

We first tested the inter-haplotype cosine distance approach using simulated data (Materials and Methods). Perhaps unsurprisingly, we find that the method's power is mostly limited by a) the initial mutation rate of the k-mer mutation type affected by the mutator allele and b) the total number of *de novo* germline mutations in the dataset (that is, the product of the number of haplotypes and the number of mutations per haplotype) (Figure {fig:figure-two}). For example, given 100 haplotypes with an average of 500 *de novo* germline mutations each, our method has nearly 100% power detect a mutator allele that increases the C>T *de novo* mutation rate by 10%. However, under the same conditions, the method only has about 10% power to detect a mutator that increases the C>G mutation rate, since C>G mutations are expected to make up a much smaller fraction of all *de novo* germline mutations to begin with. These simulations also demonstrate that our method is well-powered to detect large-effect mutator alleles (e.g., those that increase the mutation rate of a specific k-mer by 50%), even with a relatively smaller number of total mutations in the dataset.
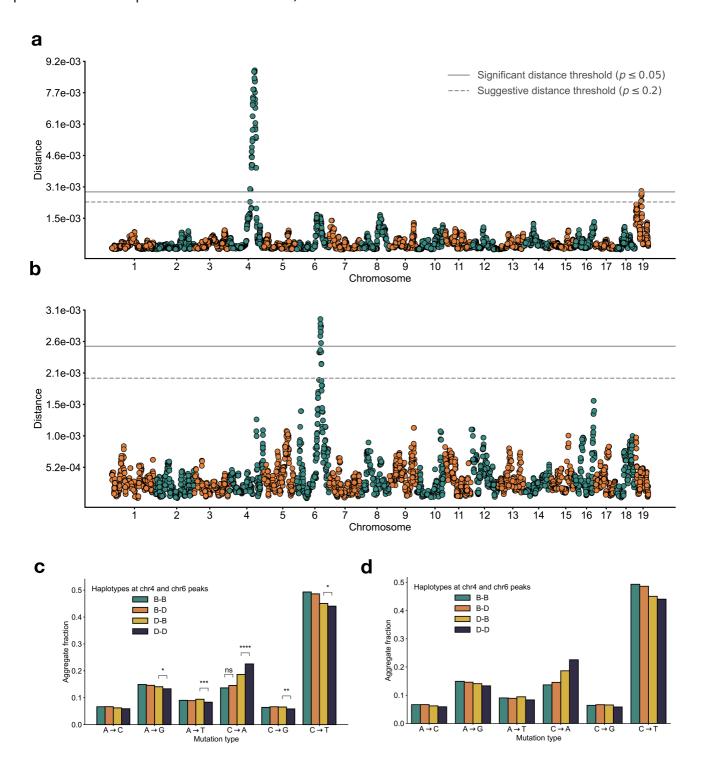

Figure 2: Simulations to assess the power of the inter-haplotype distance method. We simulated de novo germline mutations on the specified number of haplotypes, such that 50% of haplotypes were affected by a mutator allele that increased the mutation rate of the specified k-mer by the specified effect size (an effect size of 1.5 indicates a 50% increase in the mutation rate). The colors of the lines indicate the number of simulated mutations on each haplotype (before augmenting the mutation rate with a mutator allele). Given a specific combination of parameters, the y-axis denotes the fraction of 100 simulations in which the simulated mutator allele could be detected at a p-value of 0.05. Shaded areas indicate the standard deviation of that fraction.

Figure 2: Simulations to assess the power of the inter-haplotype distance method. We simulated *de novo* germline mutations on the specified number of haplotypes, such that 50% of haplotypes were affected by a mutator allele that increased the mutation rate of the specified k-mer by the specified effect size (an effect size of 1.5 indicates a 50% increase in the mutation rate). The colors of the lines indicate the number of simulated mutations on each haplotype (before augmenting the mutation rate with a mutator allele). Given a specific combination of parameters, the y-axis denotes the fraction of 100 simulations in which the simulated mutator allele could be detected at a p-value of 0.05. Shaded areas indicate the standard deviation of that fraction.

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

We applied our inter-haplotype distance method to 94 BXD RILs [Methods] using the previously described *de novo* germline mutation data [8]. Reassuringly, we observed a large peak in cosine

distance at a locus on chromosome 4 (Figure 3A; maximum distance of X at marker ID rsYYYYY; position 116.8 Mbp in mm10 coordinates).

Figure 3: Results of inter-haplotype distance scans in the BXD RILs. a) Cosine distances between aggregate *de novo* mutation spectra on BXD haplotypes with either *D* (n = X) or *B* (n = Y) alleles at approximately 7,500 informative markers. Distance thresholds at $p = 0.2$ and $p = 0.05$ were calculated by performing 10,000 permutations of the BXD haplotype mutation data. **b)** Cosine distances between aggregate *de novo* mutation spectra on BXD haplotypes with either *D* or *B* alleles at approximately 7,500 informative markers. Only BXDs with *D* haplotypes at marker rsYYYY on chromosome 4 were included in the scan. Distance thresholds at $p = 0.2$ and $p = 0.05$ were calculated by performing 10,000 permutations of the BXD haplotype mutation data. **c)** Fractions of *de novo* germline mutations in BXDs with either *D* or *B* haplotypes at markers rsXXXX and rsYYYY. Total counts of each mutation type were aggregated in groups of BXDs with either of the four possible haplotype combinations, and fractions of each mutation type were calculated in each group separately. Chi-square tests of independence were used to compare counts of individual mutation types between the D-D and D-B groups of BXDs, as well as between the B-D and B-B groups. **d)** Fractions of *de novo* germline mutations in

Sanger Mouse Genome Project (MGP) strains with either *D* or *B* haplotypes at markers rsXXXX and rsYYYY. Total counts of each mutation type were aggregated in groups of MGP strains with either of the four possible haplotype combinations, and fractions of each mutation type were calculated in each group separately. Chi-square tests of independence were used to compare counts of individual mutation types between the D-D and D-B groups of BXDs, as well as between the B-D and B-B groups.

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

An additional germline mutator allele on chromosome 6

After confirming that the inter-haplotype distance method could recover the mutator locus overlapping *Mutyh*, we asked if our approach could identify additional mutator loci in the BXD. To account for the effects of the C>A germline mutator locus near *Mutyh*, we divided the BXD RILs into those with either *D* (n = X) or *B* (n = Y) genotypes at rsYYYYY (the peak marker on chromosome 4), and ran a genome-wide distance scan using each group separately (Figure 3B.

Using only the BXDs with *B* haplotypes at the *Mutyh* mutator locus, we did not observe any genome-wide significant peaks. But using the BXDs with *D* haplotypes at the same locus, we identified a cosine distance peak on chromosome 6 (Figure 3B; maximum distance of X at marker rsYYYYY; position 133.2 Mbp in mm10 coordinates). We queried the region underneath this peak and discovered two genes annotated with the Gene Ontology term "DNA repair": *Setmar*, a protein with histone methyltransferase and transposase activity, and remarkably, *Ogg1*. The latter encodes a key member of the base-exision repair response to oxidative DNA damage, a pathway that also includes *Mutyh* and a related gene, *Mth1*. Both *Setmar* and *Ogg1* harbor nonsynonymous differences between *D* and *B* haplotypes (Table ??).

Summary of nonsynonymous differences between *D* and *B* haplotypes in DNA repair genes at the mutator locus on chromosome 6.{#tbl:table-one}

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

Evidence of epistasis between germline mutator alleles

We next compared the mutation spectra of BXDs with either *B* or *D* alleles at the mutator loci on chromosomes 4 and 6. Strikingly, we observed that BXDs with *D* alleles at both mutator loci exhibited even higher C>A germline mutation rates than those with *D* alleles at either locus alone (Figure {fig:figure-three}C). However, BXDs with *D* alleles at the mutator locus on chromosome 6 alone do not exhibit elevate elevated C>A mutation rates, suggesting that the effects of the chromosome 6 mutator locus depend on the presence of a *D* allele at the chromosome 4 locus (Figure {fig:figure-three}C). This result suggests the presence of epistasis between *D* alleles at the two mutator loci in the BXD.

To explore the effects of the two mutator loci in other inbred laboratory mice, we also compared the mutation spectra of Sanger Mouse Genomes Project (MGP) strains. Dumont [7] previously identified private germline mutations in 29 inbred laboratory strains; these private variants likely represent recent *de novo* germline mutations. Compared to MGP strains with *B* alleles at both mutator loci, those with *D* alleles at both mutator loci exhibit significantly higher C>A germline mutation rates (Figure 3D). MGP strains with *D* alleles at both mutator loci also appear to have higher C>A mutation rates than those with *D* alleles at either locus alone (Figure 3D), but given the smaller number of strains with *de novo* germline mutation data, we are unable to confirm the signal of epistasis using the MGP mutation data.

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. **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

3. **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

4. **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

5. **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

6. **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

7. **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

8. **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

9. **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

10. **A framework for variation discovery and genotyping using next-generation DNA sequencing data.**
Mark A DePristo, Eric Banks, Ryan Poplin, Kiran V Garimella, Jared R Maguire, Christopher Hartl, Anthony A Philippakis, Guillermo del Angel, Manuel A Rivas, Matt Hanna, ... Mark J Daly
Nature genetics (2011-04-10) https://www.ncbi.nlm.nih.gov/pubmed/21478889
DOI: 10.1038/ng.806 · PMID: 21478889 · PMCID: PMC3083463