Discovering epistasis between germline mutator alleles in mice

This manuscript (permalink) was automatically generated from quinlan-lab/mutator-epistasis-manuscript@65cff54 on February 14, 2023.

Authors

- **Thomas A. Sasani**
  0000-0003-2317-1374 · tomsasani · tomsasani
  Department of Human Genetics, University of Utah · Funded by Grant XXXXXXXX

- **Aaron R. Quinlan** ✉
  0000-0003-1756-0859 · aaronquinlan
  Department of Human Genetics, University of Utah; Department of Biomedical Informatics, University of Utah

- **Kelley Harris** ✉
  0000-0003-0302-2523 · Kelley__Harris
  Department of Genome Sciences, University of Washington

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

Abstract

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

DNA replication and repair proteins often recognize particular sequence motifs or excise lesions at specific nucleotides. Thus, we might expect that the spectrum of de novo mutations – i.e, the frequency of each individual mutation type (C>T, A>G, etc.) – will differ between genomes that harbor either a mutator or wild-type allele at a given locus. Previously, we used quantitative trait locus (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 [2,3].

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

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

Introduction

The germline mutation rate is a fundamental parameter in population genetics, and reflects the interplay between complex DNA replication and repair pathways, exogenous sources of DNA damage, and life-history traits. *Mutator alleles* may explain both within- and between-species variation in germline mutation rates [5], but have proven challenging to identify in mammalian genomes.

Germline mutator alleles are difficult to detect for a number of reasons, including the fidelity of germline genome replication and the effects of selection on mutators. On average, humans are born with 70 to 100 single-nucleotide *de novo* germline mutations per diploid genome [6,7]; in mice, that number is closer to 20 or 30 [8]. Moreover, in a population of sufficiently large effective size 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 $2s\Delta U$ [9], where s is the mean selective coefficient on a new deleterious mutation and ΔU is the excess number of new deleterious mutations caused by the mutator allele; the product of s and ΔU is multiplied by 2 to account for the average number of generations for which mutator is linked to the excess mutations it causes.

Compared to those that harbor wild-type alleles, we expect haplotypes harboring mutator alleles at a particular locus to carry an excess of total germline mutations. Indeed, candidate germline mutator loci have been discovered in human genomes by identifying haplotypes with excess derived allele counts [10]. 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 the

spectrum of de novo mutations – i.e, the frequency of each individual mutation type (C>T, A>G, etc.) – to differ between genomes that carry either a mutator or wild-type allele at a given locus.

In 2022, we discovered a germline mutator allele in mice by analyzing whole-genome sequencing data from 152 recombinant inbred lines (RILs). Commonly known as the BXDs [3], these RILs were derived from either F2 or advanced intercrosses of C57BL/6J and DBA/2J, two laboratory strains that exhibit significant differences in their germline mutation spectra [11]. Since the BXD RILs were maintained via brother-sister mating for up to 180 generations and housed in a controlled laboratory environment, they presented an ideal population for mutator allele discovery. Namely, each line accumulated hundreds or thousands of germline *de novo* mutations on a nearly-homozygous linear mosaic of the parental haplotypes, while the effects of negative selection on new and standing variation were attenuated [12]. We used quantitative trait locus (QTL) mapping to identify a locus on chromosome 4 that was strongly associated with the C>A germline mutation rate in the BXDs [2]. The QTL overlapped *Mutyh*, which encodes a protein that normally prevents C>A mutations by repairing oxidative DNA damage [4], and we hypothesized that missense mutations in *Mutyh* were responsible for a 50% increase in the C>A mutation rate between BXDs with either parental haplotype at the QTL.

In this study, we developed a new method to detect alleles that affect the mutation spectrum in two-parent RILs, and applied it to the *de novo* germline mutation data from the BXDs. We assessed its power to detect candidate mutator alleles, re-discovered the mutator near *Mutyh*, and identified 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 [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 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) [13], 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 other lines

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).

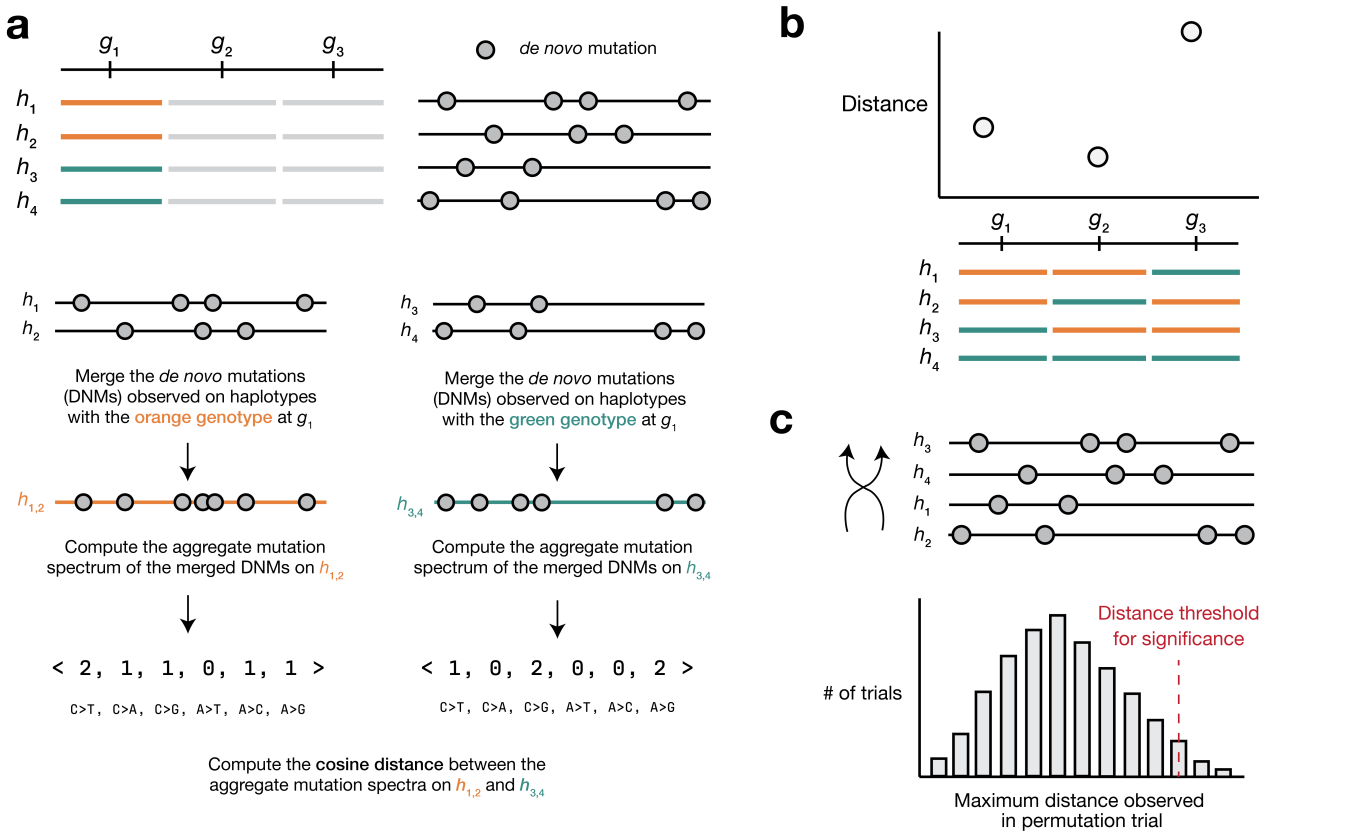

Figure 1: Overview of inter-haplotype distance method for discovering mutator alleles. a) A population of four haplotypes have been genotyped at three informative markers, and each harbors a set of private *de novo* germline mutations. At each informative marker, we compute an aggregate *de novo* germline 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 has been genotyped at informative markers, and that *de novo* germline mutations have been identified on each haplotype.

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

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

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

where $||\mathbf{A}||$ and $||\mathbf{B}||$ are the L_2 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 10,000), we compute the $1 - p$ percentile of the maximum distance distribution, and use that percentile value as a genome-wide significance threshold (for example, at $p = 0.05$).

The inter-haplotype distance method was implemented in Python, and relies heavily on the following Python libraries: `numpy`, `pandas`, `matplotlib`, `scikit-learn`, `pandera`, `seaborn`, and `numba` [14,15,16,17,18,19,20].

Additional documentation is available on GitHub [21], along with a reproducible Snakemake [22] workflow for running the method on the BXDs or other similar datasets.

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 [8].

If we are simulating the 3-mer mutation spectrum, we instead define a vector of mutation probabilities $\mathbf{P}$ 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 16 possible 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) 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.

After generating mutator and wild-type haplotypes, we compute the aggregate mutation spectrum in either group and calculate the cosine distance between the two, which we call the "focal" distance D_f. To determine whether D_f is greater than what we'd expect by chance, we concatenate the mutator and wild-type haplotypes, randomly shuffle the 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 D_f, we say that the approach succesfully identified the mutator allele. For every combination of simulation parameters (h, m, e, and so on) we perform 100 trials and record the number of trials in which we successfully identify the mutator allele.

Applying the inter-haplotype distance method to the BXDs

We downloaded previously-generated BXD *de novo* germline mutation data from the GitHub repository associated with the previous manuscript, which was also archived at Zenodo [2,23,24], and downloaded a CSV file of BXD genotypes at each of 7,320 informative markers from GeneNetwork [25,26]. We also downloaded relevant metadata about each BXD RIL from the manuscript describing the updated BXD resource [3].

We used Snakemake [27] to write a reproducible workflow for running the inter-haplotype distance method on the BXD dataset, which has been deposited in the GitHub repository associated with this manuscript [21].

Identifying candidate mutator alleles overlapping the chromosome 6 peak

We investigated the region implicated by our inter-haplotype distance approach on chromosome 6 by querying the joint-genotyped BXD VCF file (European Nucleotide Archive accession **PRJEB45429** [28]). We annotated the BXD VCF using the following `snpEff` [29] command:

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

and used `cyvcf2` [30] to iterate over the annotated VCF file in order to identify nonsynonymous differences between the parental C57BL/6J and DBA/2J strains.

Comparing mutation spectra between Mouse Genomes Project strains

We downloaded mutation data from a previous analysis [11] (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 [31]. 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 evidence of eQTLs at the mutator locus

We used the online GeneNetwork resource [26], which contains array- and RNA-seq-derived expression measurements in a wide variety of tissues from numerous datasets, to find *cis*-eQTLs for the DNA repair genes we implicated under the cosine distance peak on chromosome 6. On the GeneNetwork homepage (genenetwork.org), we selected the "BXD Family" **Group** and used the **Type** dropdown menu to select each of the specific expression datasets described in Table 2. In the **Get Any** text box, we then entered the specified gene name and clicked **Search**. After selecting the appropriate data record on the next page, we used the **Mapping Tools** dropdown to run genome-wide efficient mixed model association (GEMMA) [32] with the WGS-based marker genotypes and the leave-one-chromosome-out (LOCO) method to account for kinship between RILs.

Results

Benchmarking the inter-haplotype distance method using simulations

We first tested the inter-haplotype cosine distance approach using simulated data (Materials and Methods). We find that the method's power is mostly limited by the initial mutation rate of the k-mer mutation type affected by the mutator allele and the total number of *de novo* germline mutations in the dataset (that is, the product of the number of haplotypes and the mean number of mutations per haplotype) (Figure 2). For example, given 50 haplotypes with an average of 500 *de novo* germline mutations each, our method has nearly 90% power detect a mutator allele that increases the C>T *de novo* mutation rate by 10%. However, the method only has about 10% power to detect a mutator of identical effect size that affects the C>G mutation rate, since C>G mutations are expected to make up a much smaller fraction of all *de novo* germline mutations to begin with. These simulations also demonstrate that our method is well-powered to detect large-effect mutator alleles (e.g., those that increase the mutation rate of a specific k-mer by 50%), even with a relatively small number of mutations per haplotype.

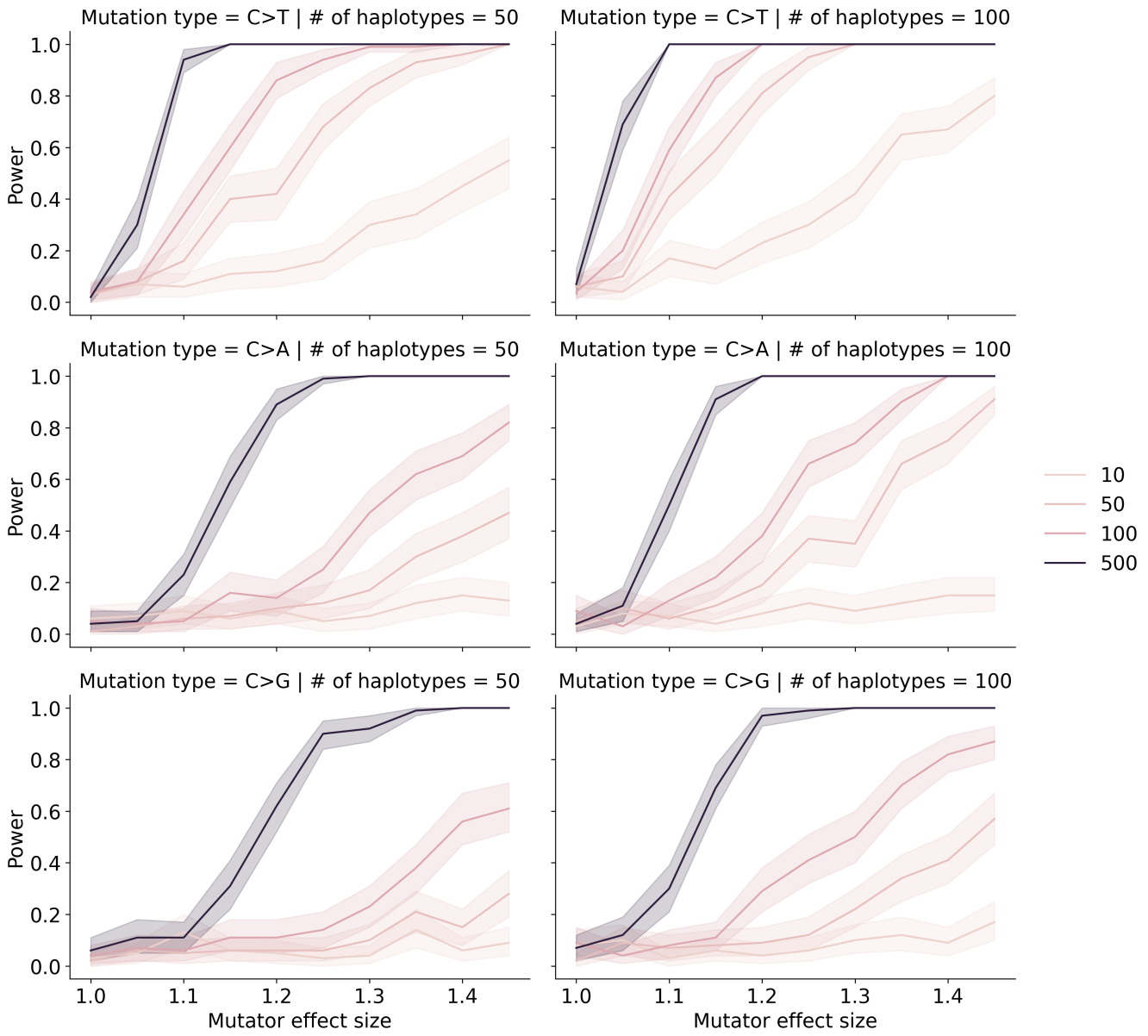

Figure 2: Simulations to assess the power of the inter-haplotype distance method. We simulated *de novo* germline mutations on the specified number of haplotypes, such that 50% of haplotypes were affected by a mutator allele that increased the mutation rate of the specified k-mer by the specified effect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

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

We applied our inter-haplotype distance method to 94 BXD RILs (Materials and Methods) using the previously described *de novo* germline mutation data [2]. Reassuringly, we observed a large peak in cosine distance at a locus on chromosome 4 (Figure 3A; maximum distance of X at marker ID rsYYYYY; position 116.8 Mbp in mm10 coordinates).

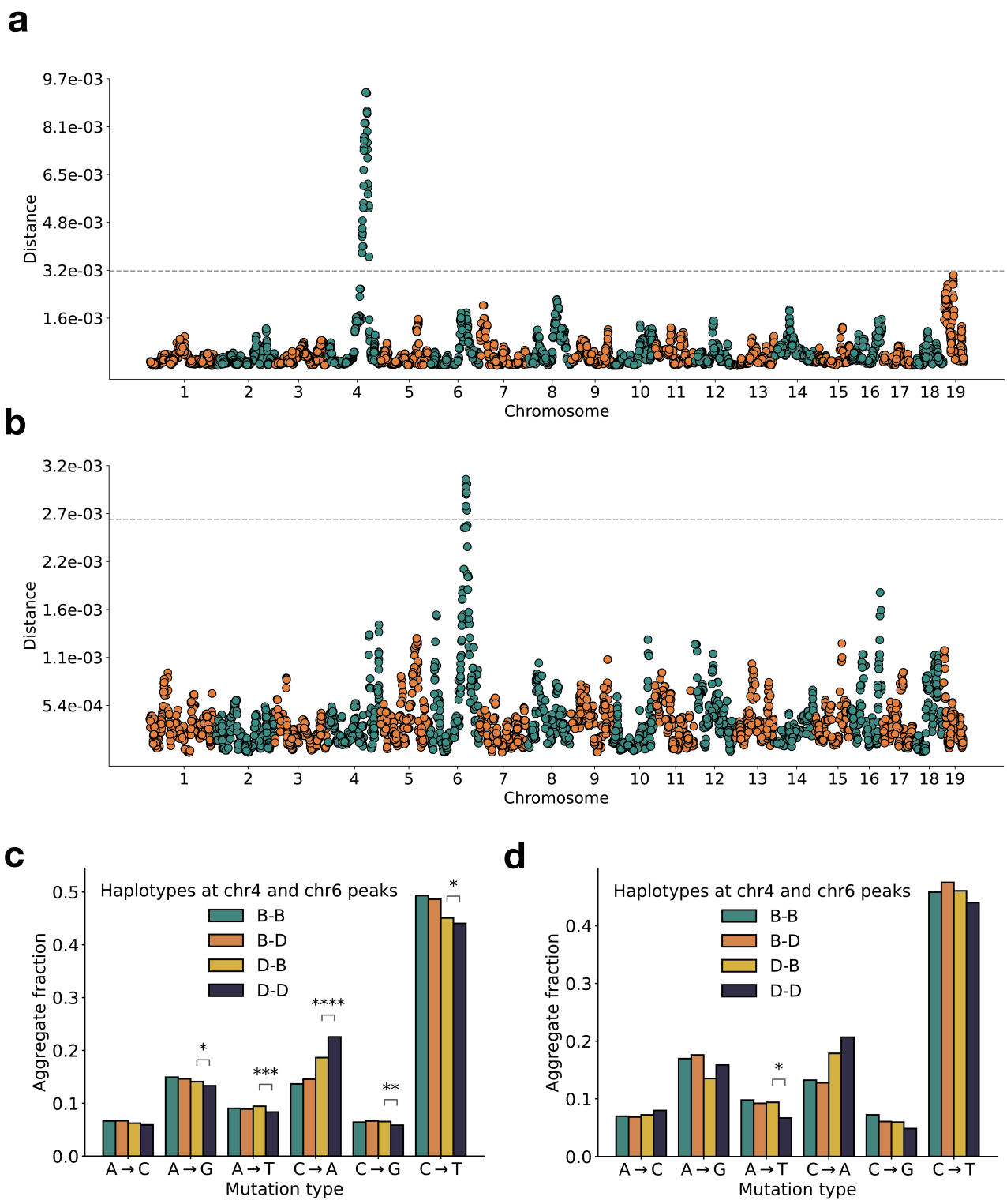

Figure 3: Results of inter-haplotype distance scans in the BXD RILs. a) Cosine distances between aggregate *de novo* mutation spectra on BXD haplotypes (n = 94)with either *D* or *B* alleles at approximately 7,500 informative markers. Distance thresholds at $p = 0.2$ and $p = 0.05$ were calculated by performing 10,000 permutations of the BXD haplotype mutation data. One outlier RIL with an extremely high C>A mutation rate (BXD68) was removed prior to running the distance scan. **b)** Cosine distances between aggregate *de novo* mutation spectra on BXD haplotypes with either *D* or *B* alleles at approximately 7,500 informative markers. Only BXDs with *D* haplotypes at marker rsYYYY on chromosome 4 were included in the scan. Distance 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 [2]. 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 [4,33,34]. 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 [2].

An additional germline mutator allele on chromosome 6

After confirming that the inter-haplotype distance method could recover the mutator locus overlapping *Mutyh*, we asked if our approach could identify additional mutator loci in the BXD. To account for the effects of the large-effect C>A germline mutator locus near *Mutyh*, we divided the BXD RILs into those with either *D* (n = 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 1).

Table 1: Summary of nonsynonymous differences between *D* and *B* haplotypes in DNA repair genes at the mutator locus on chromosome 6.

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

We also considered the possibility that expression quantitative trait loci (eQTLs), rather than nonsynonymous mutations, were responsible for the C>A mutator phenotype linked to the locus on chromosome 6. Using GeneNetwork [26], we mapped eQTLs for *Ogg1* in a number of tissues, including hematopoetic stem cells, kidney, and spleen. Overall, BXD genotypes at the cosine distance peak on chromosome 6 were significantly associated with *Ogg1* expression in many tissues, and *D* genotypes were nearly always associated with decreased gene expression (Table 2). Although we did not have access to expression data from germline tissues, such as testis or ovary, these results suggest that the expression of DNA repair genes may contribute to the observed C>A mutator phenotype, as well.

Table 2: Presence of absence of eQTLs for *Ogg1* in various tissues identified using GeneNetwork.

Tissue name	Expression data provenance	# BXDs with expression data	cis-eQTL near peak?	-log10(p) of top marker	Additive effect of D allele on expression
Kidney	`Mouse kidney M430v2 Sex Balanced (Aug06) RMA`	53	Yes	12.89	-0.18
Gastrointestinal	`UTHSC Mouse BXD Gastrointestinal Affy MoGene 1.0 ST Gene Level (Apr14) RMA`	46	Yes	5.15	-0.076
Stem cells	`UMCG Stem Cells ILM6v1.1 (Apr09) transformed`	22	No	-	-
Stem cells	`UMCG Progenitor Cells ILM6v1.1 (Apr09) transformed`	23	No	-	-
Spleen	`UTHSC Affy MoGene 1.0 ST Spleen (Dec10) RMA`	79	Yes	4.73	-0.056
Liver	`UTHSC BXD Liver RNA-Seq Avg (Oct19) TPM Log2`	50	Yes	13.57	-0.155
Heart	`NHLBI BXD All Ages Heart RNA-Seq (Nov20) TMP Log2 **`	73	No	-	-
Eye	`UTHSC BXD All Ages Eye RNA-Seq (Nov20) TPM Log2 **`	87	Yes	5.58	0.085
Lung	`HZI Lung M430v2 (Apr08) RMA`	47	Yes	14.52	-0.335

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 3C). 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 3C). 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 [11] 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.

Discussion

Using germline mutation spectra to identify mutator alleles

Germline mutation spectra are a rich source of information about the demographic history of populations, as well as the activity of both exogenous and endogenous sources of mutation throughout time. By analyzing the 3-mer mutation spectrum in a collection of human genomes, Harris and Pritchard [35] discovered a "pulse" of TCC>TTC mutation activity in European populations that likely occurred between 15,000 and 2,000 years ago, and perhaps began even earlier [36]). The mutation spectrum has also evolved rapidly during great ape evolution, potentially due to the effects of *trans*-acting mutation rate modifiers [37].

Within somatic tissues, mutation spectra can also be used to uncover the mutational processes active in particular populations of cells [38]. New computational methods have been developed to extract "mutational signatures" from large databases of somatic mutations in cancer [39]. These signatures, which describe the relative frequency of each 3-mer mutation type, can often be precisely attributed to chemotherapeutic agents, exposures to environmental mutagens, or loss-of-function mutations in genes encoding DNA repair or replication proteins [38,40,41].

Although we would expect a germline mutator allele to increase the absolute count of mutations on a haplotype, its effects may be more easily detectable by examining mutation *spectra* instead. For example, *D* alleles at the locus on chromosome 6 augment the C>A mutation rate by a factor of approximately 1.2 (Figure 3). Since C>A mutations comprise approximately 10% of all germline mutations to begin with, *D* alleles only increase the overall germline mutation rate by about 2%. Our results demonstrate that by examining mutation spectra rather than total mutation counts, we can detect germline mutator alleles with greater power. Given the depth of information that can be encoded in the mutation spectrum, we expect that these spectra can be further exploited to discover genetic modifiers of the mutation rate in other organisms, as well.

Epistasis between germline mutator alleles

Our results also reveal evidence of epistasis between mammalian germline mutator alleles for the first time. BXDs with *D* alleles at the mutator locus on chromosome 6 only exhibit elevated C>A mutation rates if they also carry *D* alleles at the previously-identified [2] mutator locus on chromosome 4. And BXDs with *D* alleles at both loci have significantly higher C>A germline mutation rates than lines with *D* alleles at only one mutator locus alone (Figure 3C). This raises the exciting possibility that epistasis between mutator alleles has contributed to the evolution of germline mutation rates and spectra in mammalian genomes.

Importantly, though, we observed epistasis between germline mutator alleles in an unnatural population; the BXDs were inbred by brother-sister mating in a highly controlled laboratory environment that attenuated the effects of natural selection on all but the most deleterious alleles

[12]. When we queried wild mouse genomes for for the three nonsynonymous DNA repair mutations overlapping the chromosome 6 mutator locus, we found all three at high frequency in *Mus mus domesticus*, the strain from which C57BL/6J and DBA/2J derive most of their genomes. Since the *D* mutator haplotype on chromosome 6 does not appear to increase the C>A germline mutation rate on its own (even in a homozygous state), we hypothesize that similar alleles may be at intermediate or high frequency in other natural populations. However, it remains unclear if our observations of epistasis in the BXD RILs reflect similar genomic processes that are active outside of the laboratory.

Discovering mutator alleles in other systems

Numerous lines of evidence suggest that mutator alleles contribute to variation in mutation rates and spectra in various organisms. In two natural isolates of *Saccharomyces cerevisiae*, nonsynonymous variation in *OGG1* causes a substantial increase in the C>A *de novo* mutation rate [42]. A recent analysis also suggests that mutator alleles and/or environmental mutagens have shaped mutation rate evolution in human genomes [43], and the heritability of paternal *de novo* mutation counts in the human germline has been estimated to be between 10 and 20%, indicating a contribution of genetic factors to germline mutation rates [44]). However, mutator discovery remains challenging in mammalian genomes.

What conditions must be met in order to detect a germline mutator allele? Presumably, one must have access to many haplotypes, each with a reasonably large number of *de novo* germline mutations that remain linked to the mutator allele(s) that caused them. Recently, thousands of human pedigrees have been sequenced in an effort to precisely estimate the rate of human *de novo* germline mutation [6,7]. Selection on germline mutator alleles will likely prevent large-effect mutators from reaching high allele frequencies; however, if multiple mutators are active in a particular population, it becomes much more likely that a subset will be detectable by sequencing human trios [45]. Current estimates of power to detect germline mutators in human pedigrees generally assume that mutators affect all mutation types equally, and that methods for mutator discovery will rely on identifying haplotypes with excess total mutation counts [45]. However, our results in the BXD suggest that germline mutators often exert their effects on a small number of k-mer mutation types, and may only be detectable by analyzing mutation spectra.

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. **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

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. **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

9. **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

10. **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

11. **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

12. **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

13. **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

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

15. **pandas-dev/pandas: Pandas**
The pandas development team
Zenodo (2023-01-19) https://doi.org/ggt8bh
DOI: https://doi.org/10.5281/zenodo.3509134

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

17. **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

18. **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

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

DOI: https://doi.org/10.21105/joss.03021

20. **Numba: a LLVM-based Python JIT compiler**
Siu Kwan Lam, Antoine Pitrou, Stanley Seibert
Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC (2015-11-15) https://doi.org/gf3nks
DOI: https://doi.org/10.1145/2833157.2833162 · ISBN: 9781450340052

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

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

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

24. **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

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

26. **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

27. **Sustainable data analysis with Snakemake.**
Felix Mölder, Kim Philipp Jablonski, Brice Letcher, Michael B Hall, Christopher H Tomkins-Tinch, Vanessa Sochat, Jan Forster, Soohyun Lee, Sven O Twardziok, Alexander Kanitz, … Johannes Köster
F1000Research (2021-01-18) https://www.ncbi.nlm.nih.gov/pubmed/34035898
DOI: 10.12688/f1000research.29032.2 · PMID: 34035898 · PMCID: PMC8114187

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

29. **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

30. **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

31. **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

32. **Study on a new chromogenic substrate for the prothrombin time determination.**
A Girolami, A Santarossa, S Martinelli, MT Sartori, I Visentin
Folia haematologica (Leipzig, Germany : 1928) (1987)
https://www.ncbi.nlm.nih.gov/pubmed/2453419
PMID: 2453419

33. **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

34. **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

35. **Rapid evolution of the human mutation spectrum.**
Kelley Harris, Jonathan K Pritchard
eLife (2017-04-25) https://www.ncbi.nlm.nih.gov/pubmed/28440220
DOI: 10.7554/elife.24284 · PMID: 28440220 · PMCID: PMC5435464

36. **Nonparametric coalescent inference of mutation spectrum history and demography.**
William S DeWitt, Kameron Decker Harris, Aaron P Ragsdale, Kelley Harris
Proceedings of the National Academy of Sciences of the United States of America (2021-05-25)
https://www.ncbi.nlm.nih.gov/pubmed/34016747
DOI: 10.1073/pnas.2013798118 · PMID: 34016747 · PMCID: PMC8166128

37. **Mutational Signatures of Replication Timing and Epigenetic Modification Persist through the Global Divergence of Mutation Spectra across the Great Ape Phylogeny.**
Michael E Goldberg, Kelley Harris
Genome biology and evolution (2022-01-04) https://www.ncbi.nlm.nih.gov/pubmed/33983415
DOI: 10.1093/gbe/evab104 · PMID: 33983415 · PMCID: PMC8743035

38. **Signatures of mutational processes in human cancer.**
Ludmil B Alexandrov, Serena Nik-Zainal, David C Wedge, Samuel AJR Aparicio, Sam Behjati, Andrew V Biankin, Graham R Bignell, Niccolò Bolli, Ake Borg, Anne-Lise Børresen-Dale, … Michael R Stratton
Nature (2013-08-14) https://www.ncbi.nlm.nih.gov/pubmed/23945592
DOI: 10.1038/nature12477 · PMID: 23945592 · PMCID: PMC3776390

39. **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

40. **The mutational footprints of cancer therapies.**
Oriol Pich, Ferran Muiños, Martijn Paul Lolkema, Neeltje Steeghs, Abel Gonzalez-Perez, Nuria Lopez-Bigas
Nature genetics (2019-11-18) https://www.ncbi.nlm.nih.gov/pubmed/31740835
DOI: 10.1038/s41588-019-0525-5 · PMID: 31740835 · PMCID: PMC6887544

41. **Mutational signatures associated with tobacco smoking in human cancer.**
Ludmil B Alexandrov, Young Seok Ju, Kerstin Haase, Peter Van Loo, Iñigo Martincorena, Serena Nik-Zainal, Yasushi Totoki, Akihiro Fujimoto, Hidewaki Nakagawa, Tatsuhiro Shibata, … Michael R Stratton
Science (New York, N.Y.) (2016-11-04) https://www.ncbi.nlm.nih.gov/pubmed/27811275
DOI: 10.1126/science.aag0299 · PMID: 27811275 · PMCID: PMC6141049

42. **A modified fluctuation assay reveals a natural mutator phenotype that drives mutation spectrum variation within**
Pengyao Jiang, Anja R Ollodart, Vidha Sudhesh, Alan J Herr, Maitreya J Dunham, Kelley Harris
eLife (2021-09-15) https://www.ncbi.nlm.nih.gov/pubmed/34523420
DOI: 10.7554/elife.68285 · PMID: 34523420 · PMCID: PMC8497059

43. **Limited role of generation time changes in driving the evolution of mutation spectrum in humans**
Ziyue Gao, Yulin Zhang, Nathan Cramer, Molly Przeworski, Priya Moorjani
bioRxiv (2023-01-13) https://doi.org/grr525
DOI: https://doi.org/10.1101/2022.06.17.496622

44. **Heritability of de novo germline mutation reveals a contribution from paternal but not maternal genetic factors**
Seongwon Hwang, Matthew DC Neville, Genomics England Research Consortium, Felix R Day, Aylwyn Scally
bioRxiv (2022-12-17) https://doi.org/grr526
DOI: https://doi.org/10.1101/2022.12.17.520885

45. **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