Manuscript Title

This manuscript (permalink) was automatically generated from quinlan-lab/mutator-epistasis-manuscript@500f931 on February 9, 2023.

Authors

- **John Doe**
  XXXX-XXXX-XXXX-XXXX · johndoe · johndoe · @johndoe@mastodon.social
  Department of Something, University of Whatever · Funded by Grant XXXXXXXX

- **Jane Roe** ✉
  XXXX-XXXX-XXXX-XXXX · janeroe
  Department of Something, University of Whatever; Department of Whatever, University of Something

✉ — Correspondence possible via GitHub Issues or email to Jane Roe <jane.roe@whatever.edu>.

Abstract

Lorem ipsum.

Introduction

Maintaining genome integrity in the mammalian germline is 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 even low-effect mutator alleles to be efficiently selected against. The selection coefficient on a mutator allele is 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 mamalian genomes and the strength of selection on a potential mutator allele, 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 carry 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.

Previously, 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 linear mosaic of the parental haplotypes that was almostly completely homozygous. 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 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.

Results

Summary of *de novo* germline mutation data 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 intercross lines (AILs), and subsequently inbred by brother-sister mating for up to 180 generations. Previously, we analyzed whole-genome sequencing data from the BXDs and identified *de novo* germline mutations in each line. A detailed description of the methods used for variant processing and filtering, as well as the characteristics of the high-quality *de novo* mutations, are available in a previous manuscript [8].

Briefly, we identified variants in each BXD that were absent from all other RILs, as well as from the C57BL/6J and DBA/2J parents [Methods]. We required each private variant to be homozygous for the alternate allele, but included heterozygous variants for which at least 90% of sequencing reads supported the alternate allele. Counts of homozygous private mutations were positively correlated with duration of inbreeding, and as expected for *de novo* germline mutations, were enriched in conserved regions of the genome [8].

A new approach to discover germline mutator alleles

Using this 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]. We assume that a collection of haplotypes have been genotyped at informative markers, and that each haplotype carries its own private *de novo* germline mutations.

To detect loci that influence the germline mutation spectrum, 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 $2 \times 3 \times 4^{k-1}$ 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 compute 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.

To assess the significance of cosine distances at particular loci, we use a permutation test to establish genome-wide distance thresholds. In each of N permutation trials, we randomly shuffle the individual 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 then compute the $1 - p$ percentile of the maximum distance distribution, and use that percentile value as a genome-wide significance threshold.

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

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

We first applied our inter-haplotype distance method to 94 BXD RILs [Methods] using the previously described *de novo* germline mutation data and approximately 7,500 genotype markers. Reassuringly, we observed a large peak in cosine distance at a locus on chromosome 4 (maximum distance of X at marker ID rsYYYYY; position 116.8 Mbp in mm10 coordinates). 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. 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 our method recovered the previously-discovered mutator locus overlapping *Mutyh*, we then asked if the inter-haplotype distance approach could identify additional mutator loci in the BXD. To account for the effects of the C>A germline mutator locus near *Mutyh*, we further 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 re-ran a genome-wide distance scan using each group separately.

Using only the BXDs with *B* haplotypes at the *Mutyh* mutator locus, we did not observe any additional genome-wide significant distance peaks. However, using the BXDs with *D* haplotypes at the same locus, we identified an additional peak in cosine distance on chromosome 6 (maximum distance of X at marker rsYYYYY; position 133.2 Mbp in mm10 coordinates).

Evidence of epistasis between germline mutator alleles

Strikingly, BXDs with *D* alleles at both mutator loci exhibit even higher C>A germline mutation rates than those with *D* alleles at only one of the two loci. 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.

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