Data Use Policy for the Wild Mouse Genomes Project

January 28, 2025

This policy is modeled on that developed for the Vertebrate Genomes Project (VGP) and posted on the VGP website: https://vertebrategenomesproject.org/data-use-policies

The Wild Mouse Genomes Project is a collaborative initiative to sequence and analyze the genomes of over 1000 wild-caught house mice (*Mus musculus*). Our goals are to catalog the breadth of wild mouse diversity, unlock the evolutionary mechanisms that have enabled house mice to successfully colonize diverse habitats across the globe, and understand how the rich diversity in wild mice can be harnessed to improve existing biomedical mouse models for human disease.

The WMGP was borne out of a series of informal discussions among a small working group in early 2024. These early discussions focused on aggregating existing wild mouse genome sequences (both published and unpublished) and conceptualizing analysis possibilities for a large-scale wild mouse population genomic resource. Throughout 2024, the project organically grew to fold in the expertise and resource contributions of other investigators. In October 2024, the WMGP secured a 2-year grant from The Jackson Laboratory to support generation of new whole-genome sequences from wild-caught mice and a course of genomic analyses aiming to (1) provide greater clarity on the origins of laboratory inbred strains and (2) aid in the functionalization of variants in the lab.

To support fair and productive use of WMGP data, our group has developed the following data use policy. This policy is consistent with those of widely used public annotation and genome databases (Ensembl, NCBI, and UCSC), has been enforced by journal editors (e.g. Nature, Science), and follows standards in genomics. We ask all users of WMGP data to respect and follow this data use policy. We anticipate that this policy will evolve over time as we accumulate data and refine our project goals.

WMGP Data Use Policy

Many of the genomes included in WMGP analyses are already available in public archives. There are no restrictions on the use of those data (unless otherwise indicated on the associated BioProject pages).

Genome sequences generated using designated WMGP funds: Before publication, the WMGP releases raw reads and assembled genomes as a service to the research community. These data releases occur through the public archives, such as SRA at NCBI, European Nucleotide Archive (ENA) at EMBL-EBI, and the GenomeArk, a public Amazon Web Services (AWS) S3 Bucket dedicated as a working space and home for high-quality reference genomes.

The WMGP project team encourages others to use these data but expects them to respect our right to first presentation (including journal publications, pre-prints such as in bioRxiv, public conference talks, and press releases) of genome-wide analyses. For those not responsible for producing, sponsoring, or analyzing the data, exceptions to the policy during the embargo period (see below) will be considered on a case-by-case basis. Possible exceptions include, but are not limited to:

- Analyses of either a single locus or a single gene family
- Use of a reference for mapping reads from independent studies

- Proposed analyses that fall outside the expertise or research interests of WMGP team members.

Genome sequences generated using other funds: Unpublished genome sequences contributed to the WMGP by individual WMGP group members will follow, at a minimum, the data release terms of the funding agency that supported data generation.

Timeline of Embargo

Genome sequences generated using designated WMGP funds: The timeline of the embargo period on newly generated data for the WMGP is 2 years from the release date of the raw sequence reads and/or genome assembly in the relevant public archive, or immediately upon WMGP sponsored publication on the data, whichever comes first. The specific embargo or (lack of embargo) will be stated along with the metadata provided for each genome. At the time that the embargo is released due to publication, the genome assembly will also be updated with the citation in the public archives. In all cases of WMGP data use, the relevant persons or publications responsible for the genome(s) should be cited.

Genome sequences generated using other funds: Unpublished genome sequences contributed to the WMGP by individual WMGP group members will follow, at a minimum, the data release timeline specified by the supporting funding agency. Investigators interested in using these data in advance of public data release should reach out to the donating investigator.

Data Sharing by WMGP members

Successfully achieving the goals of the WMGP will require that all members share their data as widely and effectively as possible with the following considerations:

- **Access**– Primary data, processed files, and final assemblies should be deposited in accessible repositories or file sharing platforms.
    - Large data files will be shared through Globus, FileSender, or as embargoed data deposited on appropriate public archives (e.g., ENA, SRA)
    - Processed data files will be shared on GitHub
- **Rights of Data Providers**- Researchers should be appropriately credited for their contribution to data generation and analyses.

To support fair and productive use of this resource within the consortium:

1. All WMGP members have pre-publication access to all genomes generated by and/or for the WGMP, regardless of source of the data. All WMGP members are encouraged to participate collaboratively in analyses whenever possible.
2. Each WMGP member is expected to follow the embargo policy.

Planned studies led by the WMGP

The WMGP intends to use the genomic data that it produces for multiple studies. This list will be periodically updated.

1. Structural diversity of wild mouse genomes
2. Creation of a wild mouse pangenome
3. Analysis of centromere diversity and Robertsonian fusions in wild mice
4. Comparative genomic analyses of wild-caught mice and their derivative inbred strains
5. Comparative genomics of wild-derived inbred strains
6. Demographic history of wild mice from across the globe
7. Genomic signatures of selection
8. Genetic diversity across immune genes and gene families
9. Contrasting genomic diversity on the X versus autosomes
10. Transposable element diversity and mobilization

11. Y chromosome diversity
12. Geographic and subspecies ancestry inference of lab strains
13. Assessing the rare variant load of inbred strains
14. Genetic contrasts between island and mainland mice

WMGP Data Use Contact Inquires

For general inquiries, including on this Data Use Policy, referencing WMGP data, or joining the WMGP, contact Beth Dumont, beth.dumont@jax.org. For inquiries limited to data contributed to the WMGP from specific group members, please contact the individual responsible for the genome(s) or questions of interest.

Members of the WMGP can be found at https://www.wildmousegenomes.com/team.