Aguate, F., Watson, M., Neelam, H., Deng, Y., Dekkers, J., Steibel, J., Cheng, H., de los Campos, G. (2026) G2P Datasets: a hub for genomic datasets for predictive modeling in plants and animals, *Proceedings 13th World Congress on Genetics Applied to Livestock Production*: 24148 https://doi.org/10.31274/wcgalp.24148 ©2026 *Aguate, et al.* CC BY-NC-ND

G2P Datasets: a hub for genomic datasets for predictive modeling in plants and animals

F. Aguate[1*], M. Watson[2], H. Neelam[1], Y. Deng[2], J. Dekkers[3], J.P. Steibel[3], H. Cheng[2], and G. de los Campos[1]

[1]*Michigan State University, East Lansing, MI, USA;* [2]*University of California, Davis, CA, USA;* [3]*Iowa State University, Ames, IA, USA;* [*]*Corresponding author: matiasfe@msu.edu*

Introduction

Every year, thousands of studies involving quantitative genetics analysis of plant and animal data are published. Increasingly, scientific journals encourage -and some mandate- making the data used in those studies publicly available. Moreover, many funding agencies and journals require research data to be made public following the FAIR (Findable, Accessible, Interoperable, and Reusable) principles (Wilkinson et al., 2016).

Although a vast number of datasets containing genomic data have been made available through publications, several issues often hinder their reuse. The datasets are often presented in a diverse range of formats and are stored in different locations, ranging from software packages and public repositories to supplementary files in publications. To facilitate data accessibility, reproducibility, and collaborative research, we developed an extensive, open-access repository of meta-data to access genomic datasets, with scripts to read the data, unify format, and run genomic models that are commonly used in plant and animal breeding. The vision for this project was to enhance access to existing genomic resources, with a primary -though not exclusive- focus on plant and animal breeding.

Currently, G2P Datasets references more than 100 datasets comprising single nucleotide polymorphism (SNP) genotypes, phenotypic data, and genetic maps for more than 60 animal and plant species. Meta-data from diverse publications were meticulously collected, encompassing a wide range of traits and species, to provide researchers with a broad and representative resource for genomic predictive modeling. While other projects have compiled genomic datasets (Chen et al., 2024; Morales et al., 2022), to our knowledge, this is the first attempt to consolidate such an extensive collection of genomic datasets across plant and animal breeding applications, and to provide scripts to read and format datasets that can be readily used in genomics research.

G2P Datasets is accessible through a GitHub repository (https://github.com/QuantGen/G2P-Datasets) and through a Shiny web app (https://mtwatson.shinyapps.io/G2P-datasets/) featuring a searchable database. This interactive tool enables researchers to search for datasets and learn about their features before downloading them, thereby saving time and effort.

To illustrate some of the utilities that G2P Datasets can offer, we present a benchmark of commonly used genomic prediction models using a subset of hosted datasets (selected based on the data included and sample size). For these datasets, in addition to the meta-data, we also offer analysis scripts and links to download the curated datasets directly from the repository and through Kaggle (https://www.kaggle.com/organizations/ag2p-disc/datasets).

Materials & Methods

We populated the database by searching on the web for publicly available genomic datasets suitable for genomic prediction modeling in plants and animals. Our search focused on plant and animal data and excluded fungi, bacteria, protist, and human data. We searched online

repositories, supplementary materials of published articles, and data archives, and we filtered out data archives that require submitting a data request or a subscription, in an effort to maximize data accessibility. The search was performed using a combination of keywords related to genomic prediction, GWAS, plant breeding, animal breeding, SNP data, and phenotypic data. We included in the repository datasets that: (i) were from plants and animals of agricultural or veterinary importance, (ii) included DNA genotypes linked to phenotypes, and (iii) were publicly available and downloadable without requiring registration or consent.

Data types. The datasets listed in the repository may contain a variety of data types to accommodate diverse research needs in genomic predictive modeling. The primary data types include: 1) DNA genotypes stored in formats such as .vcf, .bed, .rds, .txt, .csv, or .ped extensions. 2) Phenotypic data (either quantitative or qualitative), including single or multiple individual phenotypes, as well means (possibly adjusted by experimental factors). In some cases, the phenotype file also included information about the experiments from which the data was collected. Phenotype data are typically stored in tables in ASCII (.txt or .cvs extensions), .xls, or using software-specific binary formats (e.g., .rds). 3) Many datasets also include genetic maps for the SNPs included in the genotype files, while some datasets offer pedigree and/or genomic relationships instead of or along with the individual genotype data. 4) Likewise, some datasets include meta-data describing management or environmental conditions under which the phenotypic data were collected.

Data formatting. To streamline analyses, for each dataset, we developed R-scripts that, starting from the file format provided by the source, produce R-objects (data-frames for phenotypes, matrices for genotype data) that are readily useable within an R-environment. These scripts can be obtained from the Web application. To maintain the integrity of the original data and allow researchers flexibility to perform their own quality control and filtering steps, the scripts that we provide preform minimal data pre-processing.

Web Application. We developed a Web Application interface for the G2P Datasets repository, built using the R package Shiny (https://shiny.posit.co/) and is available at https://mtwatson.shinyapps.io/G2P-datasets/. This application facilitates user interactions with the project through two main modules: Browse Datasets and Add Dataset.
The Browse Datasets module displays a searchable table of datasets that can be filtered based on meta-data features, a short title, number of individuals that were genotyped, number of DNA markers genotyped, whether there is phenotypic data available, and links to articles and datasets. The search tool will also filter by information contained in the abstract, title, scientific name, and/or common name.
The Add dataset module includes a web-form that users can complete to submit new datasets, with clear instructions in the "About the project" section of the Web Application. The form collects information about the publication (including a DOI to the data), sample size, and the number of DNA markers, along with an entry where scripts to read the data can be uploaded. To secure a proper use of the application, we check each submission before it's available online.

Benchmarks. To showcase an application using large and diverse genomic data, we present a genomic prediction benchmark involving 42 datasets that are currently available in G2P Datasets. We benchmarked the cross-validation prediction performance of four genomic prediction models, including three linear models, GBLUP (Van Raden, 2008), BayesB (Meuwissen et al., 2001), tree boosting model, XGBT (Chen & Guestrin, 2016), and a Reproducing Kernel Hilbert Spaces regression using three Gaussian kernels, GKAD (de los

Campos et al., 2010). Following Couto-Alves et al. (2019), for the GKAD, we first computed a distance matrix based on a (linear) genomic relationship matrix and then derived from it three Gaussian kernels using bandwidth parameters of 0.5, 1, and 1.5. Thus, the GKAD model included three random effects, each with a different kernel matrix. To evaluate prediction performance, for each dataset and trait we conducted 100 training-testing partitions, each time randomly assigning 80% of the data to training and 20% to testing. Thus, for each dataset, trait, and model we had 100 estimates of the correlation between predictions and observed phenotypes in the testing data. The benchmark results are not shown due to space limitations.

Results & Discussion

Figure 1 illustrates the G2P Datasets project, including a GitHub repository (hosting meta-data for each of the data sets), a Web Application, and a subset of the data sets (in binary R-format) in Kaggle.

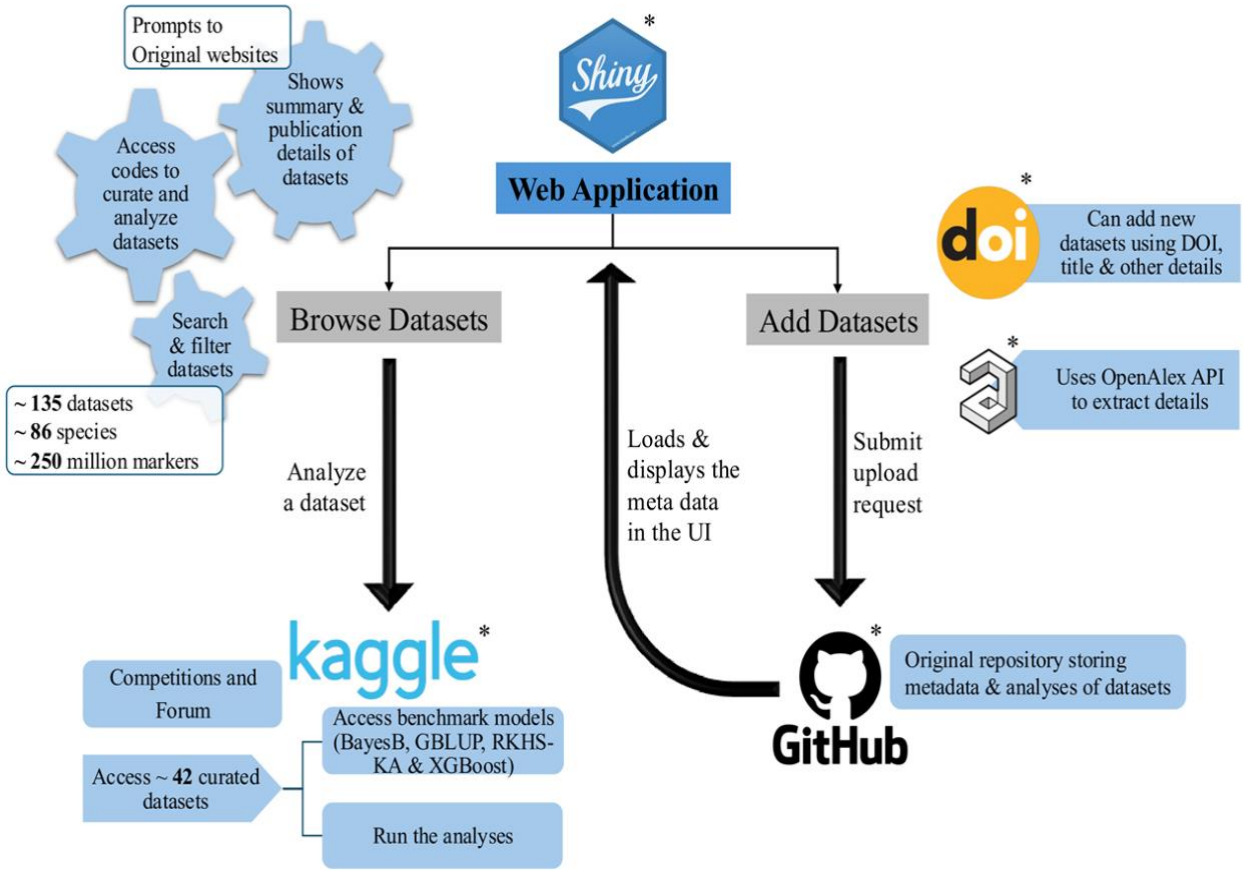

Figure 1. The G2P Datasets project. The core of the project (where meta-data is stored) is a GitHub repository. A Web Application provides a searchable database (Browse Datasets) and an electronic form to contribute new data sets (Add Datasets). A subset of the datasets included in the database are also available (as R binary objects) in Kaggle. * Outside entities.

The Web Application reads the meta-data from the repository in GitHub, provides a searchable database through the Browse Dataset module, and facilitates contribution of new datasets to the project through the Add Datasets module. The Web Application is the main User Interface (UI) to browse through all the datasets and R-scripts.

Within the Web Application, the Add Dataset module can be used to submit a new dataset to the repository. In this module, a project contributor must provide and validate meta-data for a

dataset that is not currently in the repository. Instructions to download a raw dataset can be found within the R-Scripts sub-panel in the Web Application, which also contains copy-to-clipboard buttons to facilitate copying the scripts into any Integrated Development Environment (IDE) for R programming. Scripts to run the different predictive models are also provided within this sub-panel.
We used 42 datasets for the benchmarks; these datasets can be accessed (as binary R-objects) either from Kaggle (https://www.kaggle.com/organizations/ag2p-disc) or by downloading directly from the GitHub repository using a script like the one in Figure 2.

```
repo<-   'https://github.com/QuantGen/G2P-Datasets/raw/refs/heads/main/Datasets/'
folder<- '00012_AmericanChestnutAncestry/'
file<-   'curated_geno_pheno_map.rdata'
fname<-  paste0(repo,folder,file)

load(url(fname)); rm(folder,repo,file,fname)
ls()
```

```
## [1] "geno"  "map"   "pheno"
```

Figure 2. An example R-script to load a dataset from the GitHub repository into an R-environment.

Conclusions

Data sharing and collaboration are essential to advance open science; this is especially important in genomic prediction modelling. In the last two decades, a large number of genomic datasets have been developed, and many of these were shared in the public domain, usually linked to scientific publications. The G2P Datasets initiative aims to facilitate access to the large collection of datasets for investigations and benchmarking of genomic prediction models. The repository is populated with more than 100 datasets and provides a platform for scientists to contribute with more datasets. We invite scientists to contribute with new datasets to the repository and to use the existing ones in their methods and applied research.

Acknowledgments

We thank all the scientists and institutions that made the genomic datasets available. The development of G2P Datasets was funded by the USDA Agriculture and Food Research Initiative, 11 National Institute of Food and Agriculture Competitive Grant 2023-67015-39564, and 2023-12 70412-41054.

References

Chen J, Tan C, Zhu M, Zhang C, et al. 2024. Nucleic Acids Res. 52(D1):D1519–D1529.
Chen T, Guestrin C. 2016. Proc 22nd ACM SIGKDD Int Conf Knowl Discov Data Min. p. 785–794.
Couto-Alves F, Granato ÍSC, Galli G, Lyra DH, Fritsche-Neto R, de los Campos G. 2019. Plant Methods. 15:1.
de los Campos G, Gianola D, Rosa GJM, Weigel KA, Crossa J. 2010. Genet Res. 92(4):295–308.
Meuwissen THE, Hayes BJ, Goddard ME. 2001. Genetics. 157(4):1819–1829.
Morales N, Ogbonna AC, Ellerbrock BJ, Bauchet GJ, Tantikanjana T, Tecle IY, Powell AF, Lyon D, Menda N,
R Core Team. 2024. R: a language and environment for statistical computing. https://www.R-project.org/
Van Raden PM. 2008. J Dairy Sci. 91(11):4414–4423.
Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, et al. 2016. Sci Data. 3(1):160018. https://doi.org/10.1038/sdata.2016.18