

University of Nebraska - Lincoln

DigitalCommons@University of Nebraska - Lincoln

Agronomy & Horticulture -- Faculty Publications

Agronomy and Horticulture Department

7-17-2023

Genomes to Fields 2022 Maize genotype by Environment Prediction Competition

Dayane Cristina Lima

Jacob D. Washburn

José Ignacio Varela

Qiuyue Chen

Joseph L. Gage

See next page for additional authors

Follow this and additional works at: <https://digitalcommons.unl.edu/agronomyfacpub>



Part of the [Agricultural Science Commons](#), [Agriculture Commons](#), [Agronomy and Crop Sciences Commons](#), [Botany Commons](#), [Horticulture Commons](#), [Other Plant Sciences Commons](#), and the [Plant Biology Commons](#)

This Article is brought to you for free and open access by the Agronomy and Horticulture Department at DigitalCommons@University of Nebraska - Lincoln. It has been accepted for inclusion in Agronomy & Horticulture -- Faculty Publications by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.

Authors

Dayane Cristina Lima, Jacob D. Washburn, José Ignacio Varela, Qiuyue Chen, Joseph L. Gage, Maria Cinta Romay, James Holland, David Ertl, Marco Lopez-Cruz, Fernando M. Aguatero, Gustavo de los Campos, Shawn Kaeppler, Timothy Beissinger, Martin Bohn, Edward Buckler, Jode Edwards, Sherry Flint-Garcia, Michael A. Gore, Candice N. Hirsch, Joseph E. Knoll, John McKay, Richard Minyo, Seth C. Murray, Osler A. Ortez, James C. Schnable, Rajandeep S. Sekhon, Maninder P. Singh, Erin E. Sparks, Addie Thompson, Mitchell Tuinstra, Jason Wallace, Teclemariam Weldekidan, Wenwei Xu, and Natalia de Leon

DATA NOTE

Open Access



Genomes to Fields 2022 Maize genotype by Environment Prediction Competition

Dayane Cristina Lima^{1*}, Jacob D. Washburn², José Ignacio Varela¹, Qiuyue Chen³, Joseph L. Gage³, Maria Cinta Romay⁴, James Holland⁵, David Ertl⁶, Marco Lopez-Cruz⁷, Fernando M. Aguatero⁷, Gustavo de los Campos⁸, Shawn Kaeppler¹, Timothy Beissinger⁹, Martin Bohn¹⁰, Edward Buckler¹¹, Jode Edwards¹², Sherry Flint-Garcia², Michael A. Gore¹³, Candice N. Hirsch¹⁴, Joseph E. Knoll¹⁵, John McKay¹⁶, Richard Minyo¹⁷, Seth C. Murray¹⁸, Osler A. Ortiz¹⁹, James C. Schnable²⁰, Rajandeep S. Sekhon²¹, Maninder P. Singh²², Erin E. Sparks²³, Addie Thompson²², Mitchell Tuinstra²⁴, Jason Wallace²⁵, Teclamarium Weldekidan²³, Wenwei Xu²⁶ and Natalia de Leon¹

Abstract

Objectives The Genomes to Fields (G2F) 2022 Maize Genotype by Environment (GxE) Prediction Competition aimed to develop models for predicting grain yield for the 2022 Maize GxE project field trials, leveraging the datasets previously generated by this project and other publicly available data.

Data description This resource used data from the Maize GxE project within the G2F Initiative [1]. The dataset included phenotypic and genotypic data of the hybrids evaluated in 45 locations from 2014 to 2022. Also, soil, weather, environmental covariates data and metadata information for all environments (combination of year and location). Competitors also had access to ReadMe files which described all the files provided. The Maize GxE is a collaborative project and all the data generated becomes publicly available [2]. The dataset used in the 2022 Prediction Competition was curated and lightly filtered for quality and to ensure naming uniformity across years.

Keywords Grain yield, Maize, Root mean squared error

Objective

The Maize GxE project is a collaborative effort that involves researchers from diverse areas of study. The datasets collected by the project are some of the largest public data of their kind and are therefore of broad interest to communities from genetics to agronomy to computer science and beyond. The competition was organized to connect these communities and others with

interest in dissecting and exploring genotypic, environmental, and GxE information to predict hybrid maize performance in different environments across the US. The competition started on November 15, 2022, and ended on January 15, 2023. All the participants had access to the same curated data set, containing information collected on over 180,000 maize field plots and involving 4,683 hybrids. Participants were asked to create predictive models for maize grain yield for the 2022 Maize GxE project field trials, utilizing the existing Maize GxE project dataset and any other publicly available data. The trait of interest was grain yield, and the competitors were asked to submit absolute grain yield (Mg ha^{-1})

*Correspondence:
Dayane Cristina Lima
dclima@wisc.edu

Full list of author information is available at the end of the article



© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Table 1 Overview of Genomes to Fields 2022 Maize Genotype by Environment Prediction Competition data files

Label	Name of data file	File types (Extension)	Data repository and identifier
Data file 1	readme.txt	.txt	CyVerse (https://doi.org/10.25739/tq5e-ak26) [3]
Data file 2	COMPETITION_DATA_README.docx	.docx	CyVerse (https://doi.org/10.25739/tq5e-ak26) [3]
Data file 3	1_Training_Trait_Data_2014_2021.csv	.csv	CyVerse (https://doi.org/10.25739/tq5e-ak26) [3]
Data file 4	2_Training_Meta_Data_2014_2021.csv	.csv	CyVerse (https://doi.org/10.25739/tq5e-ak26) [3]
Data file 5	3_Training_Soil_Data_2015_2021.csv	.csv	CyVerse (https://doi.org/10.25739/tq5e-ak26) [3]
Data file 6	4_Training_Weather_Data_2014_2021.csv	.csv	CyVerse (https://doi.org/10.25739/tq5e-ak26) [3]
Data file 7	5_Genotype_Data_All_Years.vcf.zip	.vcf	CyVerse (https://doi.org/10.25739/tq5e-ak26) [3]
Data file 8	6_Training_EC_Data_2014_2021.csv	.csv	CyVerse (https://doi.org/10.25739/tq5e-ak26) [3]
Data file 9	All_hybrid_names_info.csv	.csv	CyVerse (https://doi.org/10.25739/tq5e-ak26) [3]
Data file 10	GenoDataSources.txt	.txt	CyVerse (https://doi.org/10.25739/tq5e-ak26) [3]
Data file 11	GenoDataSourcesWithUpdatedBioProject.txt	.txt	CyVerse (https://doi.org/10.25739/tq5e-ak26) [3]
Data file 12	1_Submission_Template_2022.csv	.csv	CyVerse (https://doi.org/10.25739/tq5e-ak26) [3]
Data file 13	2_Testing_Meta_Data_2022.csv	.csv	CyVerse (https://doi.org/10.25739/tq5e-ak26) [3]
Data file 14	3_Testing_Soil_Data_2022.csv	.csv	CyVerse (https://doi.org/10.25739/tq5e-ak26) [3]
Data file 15	4_Testing_Weather_Data_2022.csv	.csv	CyVerse (https://doi.org/10.25739/tq5e-ak26) [3]
Data file 16	6_Testing_EC_Data_2022.csv	.csv	CyVerse (https://doi.org/10.25739/tq5e-ak26) [3]
Data file 17	Test_Set_Observed_Values_ANSWER.csv	.csv	CyVerse (https://doi.org/10.25739/tq5e-ak26) [3]

adjusted to 15.5% moisture for each hybrid in each location where data had been collected during the 2022 field season. The winner of the competition was the model with the lowest average root mean squared error (RMSE) across locations when compared with the actual yield data obtained in 2022.

Data description

The Prediction Competition data are publicly available via CyVerse/iPlant. This dataset contains training and testing set data and has been structured according to the specifications outlined in Table 1.

- **Training data:** includes phenotypic, genotypic, soil, weather (downloaded from <https://power.larc.nasa.gov>), environmental covariate data, and metadata information from 2014 to 2021 for use in developing and training models.
- **Testing data:** includes genotypic, soil, weather, environmental covariate data, and metadata information for 2022 locations. Also, a submission template that contains the environments and hybrids that participants used to submit yield predictions.

Maize is cultivated as a hybrid crop, typically resulting from the cross of two inbred parents. Consequently, both the phenotypic data in the training and testing sets exhibit hybrid information. The genotypic data includes hybrid information generated in-silico from inbred genotypic data.

Limitations

These datasets contain missing data. When working with large agricultural datasets, missing data is a common occurrence due to various factors such as data collection limitations, measurement errors, plot losses,

and environmental events. The genotypic data provided contains hybrid information derived from inbred genotypic data, a common practice. However, depending on the study goals, this may pose limitations for specific types of analysis. In instances where precise GPS coordinates were not available for certain environments (i.e., a location in a particular year), field coordinates were estimated. Depending on the research objective, the unavailability of accurate GPS coordinates could impact the reliability of the results.

Abbreviations

G2F	Genomes to Fields
GxE	Genotype by Environment

Acknowledgements

We gratefully acknowledge contributions from National Corn Growers Association, Iowa Corn Promotion Board, and USDA-ARS. The weather data was obtained from the National Aeronautics and Space Administration (NASA) Langley Research Center (LaRC) Prediction of Worldwide Energy Resource (POWER) Project funded through the NASA Earth Science/Applied Science Program.

Author contributions

DCL, JDW, JIV, QC, JLG, MCR, JH, DE, MLC, FMA, GDLC, SK, TB, MB, EB, JE, SFG, MAG, CNH, JEK, JM, RM, SCM, OAO, JCS, RSS, MPS, EES, AT, MT, JW, TW, WX, NDL were responsible for advising on data collection methods, collecting the data, reviewing data collection and curation methods, and the resulting datasets for the 2022 season. DCL, JDW, JIV, QC, JLG, MCR, JH, DE, NDL organized the Genomes to Fields (G2F) 2022 Maize Genotype by Environment Prediction Competition.

Funding

We gratefully acknowledge support from: National Corn Growers Association, Iowa Corn Promotion Board, and USDA-ARS.

Data Availability

The data described in this Data note can be freely and openly accessed on CyVerse under <https://doi.org/10.25739/tq5e-ak26> [3]. Please see Table 1 for details and links to the data.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹Department of Agronomy, University of Wisconsin – Madison, Madison, WI 53706, USA

²USDA-ARS Plant Genetics Research Unit, 205 Curtis Hall, Columbia, MO 65211, USA

³Department of Crop and Soil Sciences, North Carolina State University, Raleigh, NC 27695, USA

⁴Institute for Genomic Diversity, Cornell University, Ithaca, NY 14853, USA

⁵USDA-ARS Plant Science Research Unit, Raleigh, NC 27606, USA

⁶Iowa Corn Promotion Board, Johnston, IA 50131, USA

⁷Department of Epidemiology and Biostatistics, Institute for Quantitative Health Science and Engineering, Michigan State University, East Lansing, MI 48824, USA

⁸Department of Plant, Soil and Microbial Sciences, Department of Epidemiology and Biostatistics, Institute for Quantitative Health Science and Engineering, Michigan State University, East Lansing, MI 48824, USA

⁹Department of Crop Science, Center for Integrated Breeding Research, University of Göttingen, Carl-Sprengel-Weg 1, 37075 Göttingen, Germany

¹⁰University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

¹¹USDA-ARS and Cornell University, Ithaca, NY 14853, USA

¹²USDA ARS CICGRU, 716 Farmhouse Ln, Ames, IA 50011-1051, USA

¹³Plant Breeding and Genetics Section, School of Integrative Plant Science, Cornell University, Ithaca, NY 14853, USA

¹⁴Department of Agronomy and Plant Genetics, University of Minnesota, St Paul, MN 55108, USA

¹⁵USDA-ARS Crop Genetics and Breeding Research Unit, Tifton, GA 31793, USA

¹⁶Department of Agricultural Biology, Colorado State University, Fort Collins, CO 80523, USA

¹⁷Department of Horticulture and Crop Science, College of Food, Agricultural, and Environmental Sciences, Ohio State University, Wooster, OH 44691, USA

¹⁸Department of Soil and Crop Sciences, Texas A&M University, College Station, TX 77843, USA

¹⁹Department of Horticulture and Crop Science, Ohio State University, Columbus, OH 43210, USA

²⁰Department of Agronomy and Horticulture, University of Nebraska-Lincoln, Lincoln, NE 68588, USA

²¹Department of Genetics and Biochemistry, Clemson University, Clemson, SC 29634, USA

²²Department of Plant, Soil and Microbial Sciences, Michigan State University, East Lansing, MI 48824, USA

²³Department of Plant and Soil Sciences, University of Delaware, Newark, DE 19716, USA

²⁴Department of Agronomy, Purdue University, West Lafayette, IN 47907, USA

²⁵Department of Crop & Soil Sciences, University of Georgia, Athens, GA 30602, USA

²⁶Texas A&M University, College Station, TX 77843, USA

Received: 23 May 2023 / Accepted: 28 June 2023

Published online: 17 July 2023

References

1. Genomes to Fields. 2023. <https://www.genomes2fields.org>.
2. Genomes to Fields resources. 2023. <https://www.genomes2fields.org/resources>.
3. G2F Consortium. Genomes to Fields 2022 Maize Genotype by Environment Prediction Competition. CyVerse Data Commons. 2023. <https://doi.org/10.25739/tq5e-ak26>.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.