


DATA NOTE

Open Access



Maize genomes to fields (G2F): 2014–2017 field seasons: genotype, phenotype, climatic, soil, and inbred ear image datasets

Bridget A. McFarland^{1†}, Naser AlKhalifah^{1†}, Martin Bohn², Jessica Bubert², Edward S. Buckler^{3,4}, Ignacio Ciampitti⁵, Jode Edwards^{4,6}, David Ertl⁷, Joseph L. Gage³, Celeste M. Falcon¹, Sherry Flint-Garcia^{4,8}, Michael A. Gore³, Christopher Graham⁹, Candice N. Hirsch¹⁰, James B. Holland^{4,11}, Elizabeth Hood¹², David Hooker¹³, Diego Jarquin¹⁴, Shawn M. Kaeppler¹, Joseph Knoll⁴, Greg Kruger¹⁴, Nick Lauter^{4,6}, Elizabeth C. Lee¹³, Dayane C. Lima¹, Aaron Lorenz¹⁰, Jonathan P. Lynch¹⁵, John McKay¹⁶, Nathan D. Miller¹, Stephen P. Moose², Seth C. Murray¹⁷, Rebecca Nelson³, Christina Poudyal¹⁰, Torbert Rocheford¹⁸, Oscar Rodriguez¹⁴, Maria Cinta Romay³, James C. Schnable¹⁴, Patrick S. Schnable⁶, Brian Scully^{4,19}, Rajandeep Sekhon²⁰, Kevin Silverstein¹⁰, Maninder Singh²¹, Margaret Smith³, Edgar P. Spalding¹, Nathan Springer¹⁰, Kurt Thelen²¹, Peter Thomison²², Mitchell Tuinstra¹⁸, Jason Wallace²³, Ramona Walls²⁴, David Wills⁸, Randall J. Wisser²⁵, Wenwei Xu¹⁷, Cheng-Ting Yeh⁶ and Natalia de Leon^{1*} 

Abstract

Objectives: Advanced tools and resources are needed to efficiently and sustainably produce food for an increasing world population in the context of variable environmental conditions. The maize genomes to fields (G2F) initiative is a multi-institutional initiative effort that seeks to approach this challenge by developing a flexible and distributed infrastructure addressing emerging problems. G2F has generated large-scale phenotypic, genotypic, and environmental datasets using publicly available inbred lines and hybrids evaluated through a network of collaborators that are part of the G2F's genotype-by-environment (G × E) project. This report covers the public release of datasets for 2014–2017.

Data description: Datasets include inbred genotypic information; phenotypic, climatic, and soil measurements and metadata information for each testing location across years. For a subset of inbreds in 2014 and 2015, yield component phenotypes were quantified by image analysis. Data released are accompanied by README descriptions. For genotypic and phenotypic data, both raw data and a version without outliers are reported. For climatic data, a version calibrated to the nearest airport weather station and a version without outliers are reported. The 2014 and 2015 datasets are updated versions from the previously released files [1] while 2016 and 2017 datasets are newly available to the public.

Keywords: Maize, Genome, Genotype, GBS, G × E, Hybrid, Inbred, Phenotype, Environment, Field metadata

Objective

Genomes to fields (G2F) is a multi-institutional, public collaborative to develop information and tools that support the translation of maize (*Zea mays* L.) genomic information into relevant phenotypes for the benefit of

*Correspondence: ndeleongatti@wisc.edu

†Bridget A. McFarland and Naser AlKhalifah are joint first authors

¹ University of Wisconsin, Madison, WI 53706, USA

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



growers, consumers, and society. Building on existing maize genome sequence resources, the project focuses on developing approaches to improve phenomic predictability and facilitate the development and deployment of tools and resources that help address fundamental problems of sustainable agricultural productivity. Specific projects within G2F involve collaboration from research fields such as genetics, genomics, plant physiology, agronomy, climatology and crop modeling, computational sciences, statistics, and engineering.

As part of this effort, the G2F $G \times E$ project has collected, utilized, and shared multi-year, large-scale genotypic, phenotypic, environmental, and metadata datasets. The datasets described here were generated using standard formats between 2014 and 2017. For each of the testing locations, metadata and soil characterization are also included. During these four growing seasons, over 55,000 plots across 68 unique locations were used to evaluate inbred and hybrid plants. The resulting datasets are unique as they represent, to our knowledge, the most extensive publicly available datasets of their kind in maize, reporting a consistent set of traits across common sets of fully genotyped germplasm across many locations, along with relevant information reported down to the level of specific plots. Making these datasets publicly available is expected to enable researchers to conduct novel data analyses and develop tools using the curated and organized data described here. The 2014 and 2015 datasets are recently updated versions from previously released files (AlKhalifah et al. in *BMC Res Notes* 11:452, 2018) while 2016 and 2017 datasets are newly available to the public.

Data description

Online forms were developed for logging field site coordinates, field management metadata, and other site-specific information. Datasets include:

- Genotypic information for inbreds (with and without imputation): This includes single nucleotide polymorphism (SNP) information generated using a genotyping-by-sequence (GBS) method [2] for the inbreds used to produce the hybrids tested across all locations. Data is formatted to be readily analyzed using the TASSEL software [3].
- Phenotypic measurements for inbreds and hybrids: A handbook of instructions for making traditional phenotypic measurements (reviewed in [4]) is available via the G2F website [5]. Standard traits include stand count, stalk lodging, root lodging, days to anthesis, days to silking, ear height, plant height, plot weight, grain moisture, test weight, and estimated grain yield. Datatypes reported as both raw files and files with outliers removed are described in README

files. Additionally, a set of ear, cob, and kernel measurements was made using flatbed scanners and a machine vision platform to quantify components of yield [6]. These data are reported in millimeters with shape descriptors reported as principal components of contour data points. Cob color was reported as RGB (red/green/blue) pixel values. Kernel row number, counted manually, is reported as an integer.

- Environmental data: Data was collected using WatchDog 2700 weather stations (Spectrum Technologies) measuring at 30-min intervals from planting through harvest at each location. Collected information includes wind speed, direction, and gust; air temperature, dewpoint, and relative humidity; rainfall; and photoperiod. Data are reported based on calibration derived from nearby National Weather Service (NWS) Automated Surface Observing Systems (ASOS) airport weather stations and cleaned by removing obvious artifacts from the calibrated dataset.
- Soil characterizations: Information was first collected in 2015. Measurements include plow depth, pH, buffered pH, organic matter, texture and nitrogen, phosphorous, potassium, sulfur, and sodium levels (in parts per million).
- The previously released 2014 and 2015 datasets have been updated through additional quality control of the phenotypic and environmental datasets, the addition of missing site-specific field information and an update of the genotypic data to version 4 of the B73 reference genome.

The 2014–2017 datasets are publicly available via CyVerse/iPlant [7] with files and access links as shown in Table 1.

As the number of collaborators, plots evaluated and research questions across this project grows, it is anticipated that the variety and depth of data collected will also increase. Several projects have utilized aspects of these datasets [13–16], and more are in preparation. The potential scope of application for these data is broad and is anticipated to impact the field simply by being the first public dataset of its scale that has been collected and reported in a crop sciences using standardized protocols and formats, thus defining standards for data collection, formatting, and access for maize and other species.

Limitations

These datasets contain missing data. In the phenotypic and genotypic datasets, missing data is left blank instead of indicated by ‘null’ or zero to not interfere with software compatibility and interpretation. The only exception is for traits extracted from 2014 and 2015 ear imaging data, which are demarcated with ‘NA’.

Table 1 Overview of data file/data set

Label	Name of data file/data set	File types (Extension)	Data repository and identifier		
2014 Planting season	_readme.txt	.txt	CyVerse [8] (https://doi.org/10.25739/9wjm-ec41)		
	/a_2014_hybrid_phenotypic_data	directory			
	g2f_2014_hybrid_data_clean.csv	.csv			
	g2f_2014_hybrid_raw.csv	.csv			
	/b_2014_weather_data	directory			
	g2f_2014_weather.csv	.csv			
	/c_2014_inbred_phenotypic_data	directory			
	g2f_2014_hybrid_data_clean.csv	.csv			
	g2f_2014_hybrid_raw.csv	.csv			
	/z_2014_supplemental_info	directory			
	g2f_2014_field_characteristics.csv	.csv			
	2015 Planting season	_readme.txt		.txt	CyVerse [9] (https://doi.org/10.25739/kjsn-dz84)
		/a_2015_hybrid_phenotypic_data		directory	
g2f_2015_hybrid_data_clean.csv		.csv			
g2f_2015_hybrid_raw.csv		.csv			
/b_2015_weather_data		directory			
g2f_2015_weather.csv		.csv			
/c_2015_inbred_phenotypic_data		directory			
g2f_2015_hybrid_data_clean.csv		.csv			
g2f_2015_hybrid_raw.csv		.csv			
/d_2015_soil_data		directory			
g2f_2016_soil_data.txt		.txt			
g2f_2016_soil_data.csv		.csv			
z_2015_supplemental_info		directory			
g2f_2015_cooperator_list.csv		.csv			
g2f_2015_field_irrigation.csv		.csv			
g2f_2015_field_metadata.csv	.csv				
g2f_2015_supplemental_information.txt	.csv				
2016 Planting season	_readme.txt	.txt	CyVerse [10] (https://doi.org/10.25739/yjnh-kt21)		
	/a_2016_hybrid_phenotypic_data	directory			
	g2f_2016_hybrid_data_clean.csv	.csv			
	g2f_2016_hybrid_raw.csv	.csv			
	/c_2016_weather_data	directory			
	g2f_2016_weather.csv	.csv			
	/c_2016_soil_data	directory			
	g2f_2016_soil_data.txt	.txt			
	g2f_2016_soil_data_clean.csv	.csv			
	g2f_2016_soil_data_raw.csv	.csv			
	/z_2016_supplemental_info	directory			
	g2f_2016_supplemental_information.txt	.txt			
	g2f_2016_agronomic_information.csv	.csv			
	g2f_2016_cooperators_list.csv	.csv			
	g2f_2016_field_metadata.csv	.csv			
2017 Planting season	_readme.txt	txt	CyVerse [11] (https://doi.org/10.25739/w560-2114)		
	/a_2017_hybrid_phenotypic_data	directory			
	g2f_2017_hybrid_data_clean.csv	.csv			
	g2f_2017_hybrid_data_raw.csv	.csv			
	/b_2017_weather_data	directory			
	g2f_2017_weather_data.csv	.txt			

Table 1 (continued)

Label	Name of data file/data set	File types (Extension)	Data repository and identifier
2014 and 2015 Inbred ear imaging	/c_2017_soil_data	directory	CyVerse [12] (https://doi.org/10.7946/P2C34P)
	g2f_2017_soil_data.txt	.txt	
	g2f_2017_soil_data_clean.csv	.csv	
	g2f_2017_soil_data_raw.csv	.csv	
	/d_2017_genotypic_data	directory	
	g2f_2017_gbs_hybrid_codes.xlsx	.xlsx	
	g2f_2017_ZeaGBSv27_Imputed_ABpV4.h5	.h5	
	g2f_2017_ZeaGBSv27_Imputed_ABpV4.h5.zip	.zip	
	g2f_2017_ZeaGBSv27_Raw_ABpV4.h5	.h5	
	g2f_2017_ZeaGBSv27_Raw_ABpV4.h5.zip	.zip	
	/z_2017_supplemental_info	directory	
	g2f_2017_supplemental_information.txt	.txt	
	g2f_2017_agronomic_information.csv	.csv	
	g2f_2017_cooperators_list.csv	.csv	
	g2f_2017_field_metadata.csv	.csv	
	_readme.txt	.txt	
	2014_2015_compiledData.tar.gz	.tar.gz	
	2014_gxe_compiledDataAndFileNames.csv	.csv	
	2014_gxe_compiledDataAndFileNames_Raw.csv	.csv	
	2015_gxe_compiledDataAndFileNames.csv	.csv	
	2015_gxe_compiledDataAndFileNames_Raw.csv	.csv	
	CEK_Data_Files.tar.gz	.tar.gz	
	/cob	directory	
	_cob.txt	txt	
	cob.tar.gz	.tar.gz	
	cob_01of05.tar.gz	.tar.gz	
	cob_02of05.tar.gz	.tar.gz	
	cob_03of05.tar.gz	.tar.gz	
	cob_04of05.tar.gz	.tar.gz	
	cob_05of05.tar.gz	.tar.gz	
	/ear	directory	
	_ear.txt	.txt	
	ear.tar.gz	tar.gz	
	ear_01of08.tar.gz	tar.gz	
	ear_02of08.tar.gz	tar.gz	
	ear_03of08.tar.gz	tar.gz	
	ear_04of08.tar.gz	tar.gz	
	ear_05of08.tar.gz	tar.gz	
	ear_06of08.tar.gz	tar.gz	
	ear_07of08.tar.gz	tar.gz	
ear_08of08.tar.gz	tar.gz		
/kernel	directory		
_kernel.txt	.txt		
kernel.tar.gz	tar.gz		
kernel_01of05.tar.gz	tar.gz		
kernel_02of05.tar.gz	tar.gz		
kernel_03of05.tar.gz	tar.gz		
kernel_04of05.tar.gz	tar.gz		
kernel_05of05.tar.gz	tar.gz		

For weather datasets, raw files reported by sensors are not provided because machine data were calibrated based on information from nearby weather stations to ensure accuracy (e.g., if the wind vane was set improperly, a calibration correction was required). Instead, only the cleaned version of the file is reported to reduce misinterpretation.

The geographic locations of field locations are not identical across years due to crop rotation management practices. Along with the field location code, the GPS coordinates are reported. While the germplasm used in the experiments is publicly accessible, it was not generated directly by national public genebanks. Seed access and availability are handled by the G2F collaborators directly.

Abbreviations

G2F: Genomes to fields; G × E: Genotype-by-environment; GBS: Genotyping-by-sequencing; RGB: Red/green/blue; DOI: Digital Object Identifier.

Acknowledgements

We gratefully acknowledge the data management training and transition contributions from Darwin A. Campbell, Jack M. Gardiner, Carolyn Lawrence-Dill and Renee Walton. We also acknowledge contributions from many field managers and data collectors including: Lisa Coffey (P. Schnable lab); Dustin Eilert, Marina Borsecnik, Rachel Perry, Emily Rothfusz, and Jane Petzoldt (de Leon/Kaeppeler labs); Nick Lepak, Josh Budka, Nicholas Kaczmar, and Judy Kolkman (Cornell University); Miriam Lopez, Grace Kuehne, and Sarah Weirich (Lauter lab); Teclemariam Weldekidan (Wisser lab); Christine Smith (J. Schnable lab); Jacob Garfin, Amanda Gilbert and Thomas Hoverstad (Hirsch lab); Pete Hermanson (Springer lab); Nicole Yana (Bohn lab); Jacob Pekar (Texas A&M University); Susan Melia-Hancock (USDA-ARS, Columbia, MO); and Bill Widdicombe (Michigan State University). We also benefitted from data management discussions with Nicole Hopkins and Jeremy DeBarry (formerly with CyVerse); Kate Dreher, Clarissa Pimental, Julian Pietragalla, Jean-Marcel Ribaut, and Sarah Hearne (CIMMYT); Jan Erik Backlund and Kelly Robbins (Cornell University); and Matthew Berrigan (LeafNode).

Authors' contributions

BAM, NAK, JE, CMF, JLG, DJ, DCL, NDM, CP, MCR, KS, RW, CTY: data management team; MB, JB, ESB, IC, JE, SFG, MAG, CG, CH, JBH, EH, DH, SMK, JK, GK, NL, ECL, AL, JPL, JM, SPM, SCM, RN, TR, OR, JCS, BS, RS, MS, MS, EPS, NS, KT, PT, MT, JW, DW, RJW, WX, NDL: data contributors; DE, PSS, NDL: communication. The data management team aggregated, curated, and made available data resources. Contributors advised on data collection methods, collected the data, and reviewed data collection and curation methods as well as datasets. Communicating authors wrote the manuscript and guided data collection, curation, and distribution. All authors reviewed the manuscript. All authors read and approved the final manuscript.

Funding

We gratefully acknowledge support from: USDA Hatch program funds to multiple PIs in this project; the USDA Agricultural Research Service; the Arkansas Corn and Grain Sorghum Board; the Clemson University, the Colorado Corn Administrative Committee; the Georgia Agricultural Commodity Commission for Corn; the Corn Marketing Program of Michigan; the Illinois Corn Marketing Board; the Iowa Corn Promotion Board; the Iowa State University Plant Sciences Institute; the Kansas Corn Commission; the Minnesota Corn Research and Promotion Council; National Corn Growers Association; Nebraska Corn Board; the Ohio Corn Marketing Program; the Ontario Ministry of Agriculture, Food, and Rural Affairs; the Texas Corn Producers Board and the Wisconsin Corn Promotion Board. We also acknowledge funding from the National Science Foundation under Grant Numbers #DBI-0735191 and #DBI-1265383 to support CyVerse (<http://www.cyverse.org>), #IOS-1339362 to support

phenotyping by JC and SPM, and USDA-NIFA 2011-67003-30342 to RJW, SFG, JH, NL, SM, WX, and NDL. The funders had no role in the design and conduct of the study, data collection, and writing of the manuscript.

Availability of data materials

The data described in this Data Note can be freely and openly accessed at CyVerse via the following Digital Object Identifiers (DOIs): <https://www.doi.org/10.25739/frmv-wj25>, <https://www.doi.org/10.25739/9wjm-eq41>, <https://www.doi.org/10.25739/kjsn-dz84>, <https://www.doi.org/10.25739/yjnh-kt21>, <https://www.doi.org/10.25739/w560-2114> and <https://doi.org/10.7946/P2C34>. See Table 1 and reference list for details and links to the data.

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

¹ University of Wisconsin, Madison, WI 53706, USA. ² University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA. ³ Cornell University, Ithaca, NY 14853, USA. ⁴ USDA-ARS, Beltsville, MD, USA. ⁵ Kansas State University, Manhattan, KS 66502, USA. ⁶ Iowa State University, Ames, IA 50011, USA. ⁷ Iowa Corn Growers Association, Johnston, IA 50131, USA. ⁸ University of Missouri, Columbia, MO 65211, USA. ⁹ South Dakota State University, Rapid City, SD 57702, USA. ¹⁰ University of Minnesota, St. Paul, MN 55108, USA. ¹¹ North Carolina State University, Raleigh, NC 27695, USA. ¹² Arkansas State University, Jonesboro, AR 72401, USA. ¹³ University of Guelph, Ridgetown, ON, Canada. ¹⁴ University of Nebraska, Lincoln, NE 68583, USA. ¹⁵ Pennsylvania State University, State College, PA 16802, USA. ¹⁶ Colorado State University, Fort Collins, CO 80523, USA. ¹⁷ Texas A&M University, College Station, TX 77843, USA. ¹⁸ Purdue University, West Lafayette, IN 47907, USA. ¹⁹ University of Florida, Gainesville, FL 32611, USA. ²⁰ Clemson University, Clemson, SC 29634, USA. ²¹ Michigan State University, East Lansing, MI 48824, USA. ²² Ohio State University, Columbus, OH 43210, USA. ²³ University of Georgia, Athens, GA 30602, USA. ²⁴ University of Arizona, Tucson, AZ 85721, USA. ²⁵ University of Delaware, Newark, DE 19716, USA.

Received: 31 October 2019 Accepted: 27 January 2020

Published online: 12 February 2020

References

1. Alkhalifah N, et al. Maize genomes to fields: 2014 and 2015 field season genotype, phenotype, environment, and inbred ear image datasets. *BMC Res Notes*. 2018;11:452.
2. Elshire RJ, et al. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE*. 2011;6(5):e19379.
3. Bradbury PJ, et al. TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics*. 2007;23:2633–5.
4. Pauli D, et al. The quest for understanding phenotypic variation via integrated approaches in the field environment. *Plant Physiol*. 2016;172:622–34.
5. Genomes to fields. phenotyping handbook <https://www.genomes2fields.org/about/project-overview/#standards-and-methods>. Accessed 30 Aug 2019.
6. Miller ND, et al. A robust, high-throughput method for computing maize ear, cob, and kernel attributes automatically from images. *Plant J*. 2017;89:169–78.
7. Merchant N, et al. The iPlant collaborative: cyberinfrastructure for enabling data to discovery for the life sciences. *PLoS Biol*. 2016;14:e1002342.
8. G2F Consortium. G2F planting season 2014. CyVerse Data Commons. 2019. <https://doi.org/10.25739/9wjm-eq41>.
9. G2F Consortium. G2F planting season 2015. CyVerse Data Commons. 2019. <https://www.doi.org/10.25739/kjsn-dz84>.
10. G2F Consortium. G2F planting season 2016. CyVerse Data Commons. 2019. <https://www.doi.org/10.25739/yjnh-kt21>.

11. G2F Consortium. G2F planting season 2017. CyVerse Data Commons. 2019. <https://www.doi.org/10.25739/w560-2114>.
12. Spalding E. Genomes to fields inbred ear imaging 2017. CyVerse Data Commons. 2017. <https://doi.org/10.7946/p2c34p>.
13. Gage JL, et al. The effect of artificial selection on phenotypic plasticity in maize. *Nat Commun.* 2017;8:1348.
14. Lawrence-Dill C, et al. Idea factory: the maize genomes to fields initiative. *Crop Sci.* 2019;59(4):1406–10.
15. Anderson SL, et al. Prediction of maize grain yield before maturity using improved temporal height estimates of unmanned aerial systems. *Plant Phenome J.* 2019;2:190004.
16. Falcon CM, Kaeppeler SM, Spalding EP, et al. Relative utility of agronomic, phenological, and morphological traits for assessing genotype-by-environment interaction in maize inbreds. *Crop Sci.* 2020. <https://doi.org/10.1002/csc2.20035>.

Publisher's Note

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

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

