AlKhalifah *et al. BMC Res Notes (2018) 11:452*
https://doi.org/10.1186/s13104-018-3508-1

BMC Research Notes

DATA NOTE

Open Access

Maize Genomes to Fields: 2014 and 2015 field season genotype, phenotype, environment, and inbred ear image datasets

Naser AlKhalifah[1,23†], Darwin A. Campbell[1†], Celeste M. Falcon[2†], Jack M. Gardiner[1,24†], Nathan D. Miller[2†], Maria Cinta Romay[3†], Ramona Walls[4†], Renee Walton[1†], Cheng-Ting Yeh[1†], Martin Bohn[5], Jessica Bubert[5], Edward S. Buckler[3,6], Ignacio Ciampitti[7], Sherry Flint-Garcia[6,8], Michael A. Gore[3], Christopher Graham[9], Candice Hirsch[10], James B. Holland[6,11], David Hooker[12], Shawn Kaeppler[2], Joseph Knoll[6], Nick Lauter[1,6], Elizabeth C. Lee[13], Aaron Lorenz[14,25], Jonathan P. Lynch[15], Stephen P. Moose[5], Seth C. Murray[16], Rebecca Nelson[3], Torbert Rocheford[17], Oscar Rodriguez[14], James C. Schnable[14], Brian Scully[6,18], Margaret Smith[3], Nathan Springer[10], Peter Thomison[19], Mitchell Tuinstra[17], Randall J. Wisser[20], Wenwei Xu[21], David Ertl[22*], Patrick S. Schnable[1*], Natalia De Leon[2*], Edgar P. Spalding[2*], Jode Edwards[1,6*] and Carolyn J. Lawrence-Dill[1*]

Abstract

Objectives: Crop improvement relies on analysis of phenotypic, genotypic, and environmental data. Given large, well-integrated, multi-year datasets, diverse queries can be made: Which lines perform best in hot, dry environments? Which alleles of specific genes are required for optimal performance in each environment? Such datasets also can be leveraged to *predict* cultivar performance, even in uncharacterized environments. The maize Genomes to Fields (G2F) Initiative is a multi-institutional organization of scientists working to generate and analyze such datasets from existing, publicly available inbred lines and hybrids. G2F's genotype by environment project has released 2014 and 2015 datasets to the public, with 2016 and 2017 collected and soon to be made available.

Data description: Datasets include DNA sequences; traditional phenotype descriptions, as well as detailed ear, cob, and kernel phenotypes quantified by image analysis; weather station measurements; and soil characterizations by site. Data are released as comma separated value spreadsheets accompanied by extensive README text descriptions. For genotypic and phenotypic data, both raw data and a version with outliers removed are reported. For weather data, two versions are reported: a full dataset calibrated against nearby National Weather Service sites and a second calibrated set with outliers and apparent artifacts removed.

Keywords: Maize, Genome, Genotype, Environment, Breeding, Phenotype, Prediction, Soil, Inbred, Hybrid

*Correspondence: dertl@iowacorn.org; schnable@iastate.edu; ndeleongatti@wisc.edu; spalding@wisc.edu; jode.edwards@ars.usda.gov; triffid@iastate.edu
†Naser AlKhalifah, Darwin A. Campbell, Celeste M. Falcon, Jack M. Gardiner, Nathan D. Miller, Maria Cinta Romay, Ramona Walls, Renee Walton, Cheng-Ting Yeh are joint first authors
[1] Iowa State University, Ames, IA 50011, USA
[2] University of Wisconsin, Madison, WI 53706, USA
[6] USDA-ARS, Beltsville, MD, USA
[22] Iowa Corn Growers Association, Johnston, IA 50131, USA
Full list of author information is available at the end of the article

© The Author(s) 2018. This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. 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.

Objective

G2F is a multi-institutional, collaborative initiative to develop tools that efficiently predict performance of diverse maize (*Zea mays* ssp. *mays*) varieties across multiple growing conditions. G2F projects aim to collect, share, and analyze multi-year, large-scale genomic, phenotypic, and environmental datasets. The project builds on existing maize genome sequence resources by developing approaches to understand the functions of genes and specific alleles based on their expression in typical field conditions. There are many dimensions to the goal of understanding genotype-by-environment (G × E) interactions, including which genes impact which traits and trait components, how genes interact among themselves, the relevance of specific genes under different growing conditions, and how genes influence plant growth during various stages of development.

G2F projects foster integration of diverse research disciplines, including (but not limited to) genetics, genomics, plant physiology, agronomy, climatology, and crop modeling as well as analytical perspectives and tools derived from computational sciences, statistics, and engineering. Under the umbrella of G2F are enterprises such as the G × E project that began in 2014. The G × E project aims to document and measure genotypes, phenotypes, and environmental data in standard formats across more than twenty distributed field locations in North America annually. The resulting dataset is unique as it represents, to our knowledge, the most extensive publicly available dataset of its kind, reporting a consistent set of traits across common sets of fully genotyped germplasm not only across many locations, but also with relevant information reported down to the level of specific plots. Making these datasets publicly available enables researchers from many different disciplines to tackle the daunting analyses necessary to make useful predictions of crop performance. Novel data analysis approaches and tools are expected to result from the curated and organized data described here.

Data description

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

- DNA sequences of inbreds (with and without imputation), including those inbreds used to produce featured hybrids. The process for creating files and metadata pertaining to the genotype by sequencing (GBS) process [1] is described. Data are most readily analyzed using TASSEL software [2]. Raw sequence reads generated are accessible via the Sequence Read Archive [3].
- Phenotype measurements for inbreds and hybrids. A handbook of instructions for making traditional phenotype measurements (reviewed in [4]) is available via the G2F website [5]. Traditional traits include stand count, stalk lodging, root lodging, days to anthesis, days to silking, ear height, plant height, plot weight, grain moisture, and test weight. Datatypes reported as both raw files and files with outliers removed are described in README files. Additionally, a large set of ear, cob, and kernel measurements was made with a non-traditional machine vision platform to quantify the 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 collected by WatchDog 2700 weather stations (Spectrum Technologies) at 30-min intervals from planting through harvest. Collected information includes wind speed, direction, and gust; air temperature, dewpoint, and relative humidity; rainfall; and solar radiation. Data are reported as a calibrated set (based on calibration derived from nearby National Weather Service stations) and "clean" (based on removing obvious artifacts from the calibrated dataset).
- Soil characterizations by site (first taken in 2015) including plow depth, pH, buffered pH, organic matter, phosphorus levels (in parts per million), and potassium levels (in parts per million).

Data collected in year n are released to project members in spring of the following year ($n+1$), and released to the public the subsequent year ($n+2$). The 2014 and 2015 datasets are publicly available via the NCBI SRA [7] and CyVerse/iPlant [8] with files and access links shown in Table 1.

As technologies develop and the number of researchers involved in the project grows, it is anticipated that increasingly diverse datatypes will be documented. An example of the use of these data has been reported [12]. In that study, phenotypic plasticity was found to be disproportionately controlled by regulatory regions. Because these datasets support lines of inquiry limited only by the questions researchers pose, the potential scope of application for these data is broad. The dataset is anticipated to additionally impact the field simply by being the first public dataset of its scale that has been collected and reported using standardized protocols and

Table 1 Overview of data files and data sets

Label	Name of data file/data set	File types (extension)	Data repository and identifier
DNA Sequences of Inbreds	GBS sequencing Maize G2F (G × E) inbreds	Sequence reads	NCBI SRA PRJNA385022 [3] (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA385022)
2014 Field Season Phenotypic and Genotypic Data	_readme.txt	.txt	CyVerse [9] (https://doi.org/10.7946/P2V888)
	/a._2014_hybrid_phenotypic_data	directory	
	_g2f_2014_hybrid_data_description.txt	.txt	
	g2f_2014_hybrid_no_outliers.csv	.csv	
	g2f_2014_hybrid_raw.csv	.csv	
	/b._2014_gbs_data	directory	
	_g2f_2014_gbs_data_description.txt	.txt	
	g2f_2014_gbs_data.csv	.csv	
	g2f_2014_zeagbsv27.imp.h5	.h5	
	g2f_2014_zeagbsv27.imp.h5.gz	.gz	
	g2f_2014_zeagbsv27.raw.h5	.h5	
	g2f_2014_zeagbsv27.raw.h5.gz	.gz	
	g2f_2014_zeagbsv27impv5hmp.txt.gz	.gz	
	g2f_2014_zeagbsv27v5hmp.txt.gz	.gz	
	/c._2014_weather_data	directory	
	_g2f_2014_weather_data_description.txt	.txt	
	g2f_2014_weather_calibrated.csv	.csv	
	g2f_2014_weather_clean.csv	.csv	
	/d._2014_inbred_phenotypic_data	directory	
	_g2f_2014_inbred_data_description.txt	.txt	
	g2f_2014_inbred_no_outliers.csv	.csv	
	g2f_2014_inbred_raw.csv	.csv	
	/z._2014_supplemental_info	directory	
	g2f_2014_field_characteristics.csv	.csv	
2015 Field Season Phenotypic and Genotypic Data	_readme.txt	.txt	CyVerse [10] (https://doi.org/10.7946/P24S31)
	/a._2015_hybrid_phenotypic_data	directory	
	_g2f_2015_hybrid_data_description.txt	.txt	
	g2f_2015_hybrid_no_outliers.csv	.csv	
	g2f_2015_hybrid_raw.csv	.csv	
	/b._2015_gbs_data	directory	
	_g2f_2014_gbs_data_description.txt	.txt	
	/c._2015_weather_data	directory	
	_g2f_2015_weather_data_description.txt	.txt	
	g2f_2015_weather_calibrated.csv	.csv	
	g2f_2015_weather_clean.csv	.csv	
	/d._2015_inbred_phenotypic_data	directory	
	_g2f_2015_inbred_data_description.txt	.txt	
	g2f_2015_inbred_raw.csv	directory	
	/e._2015_soils	directory	
	_g2f_2015_soil_data.txt	.txt	
	g2f_2015_soil_data.csv	.csv	
	/z._2015_supplemental_info	directory	
	_g2f_2015_supplemental_information.txt	.txt	
	g2f_2015_cooperator_list.csv	.csv	
	g2f_2015_field_irrigation.csv	.csv	
	g2f_2015_field_metadata.csv	.csv	

Table 1 (continued)

Label	Name of data file/data set	File types (extension)	Data repository and identifier
2014 and 2015 Inbred Ear Imaging	_readme.txt	txt	CyVerse [11] (https://doi.org/10.7946/P2C34P)
	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	

formats, respectively, thus defining standards for data collection, formatting, and access.

Limitations

Missing data occurs in most datasets. For genotypic and phenotypic datasets, missing data are left blank rather than zero or 'null' representation because some measured data report zero values and some software will only accept numeric values (not strings). The exception is for traits extracted from inbred ear, cob, and kernel image data, which are demarcated with 'NA'.

In some instances, reported data were maintained rather than editing for consistency. These decisions were made to minimize misinterpretation that could lead to incorrect documentation or measurements.

For weather data, 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).

Field locations are not always identical year-to-year, primarily due to crop rotation management practices. Each field's GPS coordinates are reported annually to enable data aggregation in keeping with specific research objectives.

Germplasm used and reported are specific to the project and are held by researchers involved in the project. They do not derive directly from national public genebanks. Seed access is granted in keeping with seed availability from cooperating researchers directly.

Abbreviations

G2F: Genomes to Fields; G × E: genotype by environment interaction; GBS: genotyping by sequencing; RGB: red/green/blue; DOI: Digital Object Identifier.

Authors' contributions

NA, DAC, CMF, JMG, NDM, MCR, RW, RW, CTY: data management team; MB, JB, ESB, IC, SFG, MAG, CG, CH, JBH, DH, SK, JK, NL, ECL, AL, JPL, SPM, SCM, RN, TR, OR, JCS, BS, MS, NS, PT, MT, RJW, WX: data contributors; DE, PSS, NL, EPS, JE, CJLD: 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.

Author details

[1] Iowa State University, Ames, IA 50011, USA. [2] University of Wisconsin, Madison, WI 53706, USA. [3] Cornell University, Ithaca, NY 14853, USA. [4] University of Arizona, Tucson, AZ 85721, USA. [5] University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA. [6] USDA-ARS, Beltsville, MD, USA. [7] Kansas State University, Manhattan, KS 66502, 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] University of Guelph, Ridgetown, ON, Canada. [13] University of Guelph, Guelph, ON, Canada. [14] University of Nebraska, Lincoln, NE 68583, USA. [15] Pennsylvania State University, University Park, PA 16802, USA. [16] Texas A&M University, College Station, TX 77843, USA. [17] Purdue University, West Lafayette, IN 47907, USA. [18] University of Florida, Gainesville, FL 32611, USA. [19] Ohio State University, Columbus, OH 43210, USA. [20] University of Delaware, Newark, DE 19716, USA. [21] Texas A&M AgriLife Research, Lubbock, TX 79403, USA. [22] Iowa Corn Growers Association, Johnston, IA 50131, USA. [23] Present Address: University of Wisconsin, Madison, WI 53706, USA. [24] Present Address: University of Missouri, Columbia, MO 65211, USA. [25] Present Address: University of Minnesota, St. Paul, MN 55108, USA.

Acknowledgements

We gratefully acknowledge contributions from many field managers and data collectors including: Lisa Coffey (Schnable lab); Dustin Eilert, Marina Borsecnik, Emily Rothfusz, and Jane Petzoldt (De Leon lab); Nick Lepak, Josh Budka, and Nicholas Kaczmar (Cornell University); Miriam Lopez, Grace Kuehne, and Sarah Weirich (Lauter lab); Teclemariam Weldekidan (Wisser lab); Jacob Garfin and Amanda Gilbert (Hirsch lab), Pete Hermanson (Springer lab); Jacob Pekar (Texas A&M University); and Susan Melia-Hancock (USDA-ARS, Columbia, MO). 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).

Competing interests

The authors declare that they have no competing interests.

Availability of data materials

The data described in this Data Note can be freely and openly accessed at the NCBI Sequence Read Archive via the identifier PRJNA385022 and at CyVerse via the following Digital Object Identifiers (DOIs): https://doi.org/10.7946/p2v888, https://doi.org/10.7946/p24s31, and https://doi.org/10.7946/p2c34p. See Table 1 and reference list for details and links to the data.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

Funding

We gratefully acknowledge support from: USDA Hatch program funds to multiple PIs in this project; the USDA Agricultural Research Service; the Iowa State University Plant Sciences Institute; the Ontario Ministry of Agriculture, Food, and Rural Affairs; the Illinois Corn Marketing Board; the Iowa Corn Promotion Board; the Kansas Corn Commission; the Minnesota Corn Research and Promotion Council; the Nebraska Corn Board; the Ohio Corn Marketing Program; the Texas Corn Producers Board; and the National Corn Growers Association. We also acknowledge funding from the National Science Foundation under Grant Numbers #DBI-0735191 and #DBI-1265383 to support CyVerse (http://www.cyverse.org) and USDA-NIFA 2011-67003-30342 to SFG, JH, NL, SM, RW, WX, and NDL.

Publisher's Note

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

Received: 14 February 2018 Accepted: 18 June 2018
Published online: 09 July 2018

References

1. Elshire RJ, et al. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS ONE. 2011;6(5):e19379.
2. Bradbury PJ, et al. TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics. 2007;23:2633–5.
3. Sornapudi T, Nayak R, Uppada V, Guthikonda PK, Kethavath S, Yellaboina S, Pasupulati AK, Kurukuti S. 2018: NCBI Sequence Read Archive. PRJNA385022.
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 1 Mar 2018.
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. Leinonen R, Sugawara H. Shumway. The sequence read archive. Nucleic Acids Res. 2011;39(Database issue):D19–21.
8. Merchant N, et al. The iPlant collaborative: cyberinfrastructure for enabling data to discovery for the life sciences. PLoS Biol. 2016;14:e1002342.
9. Lawrence-Dill C. Genomes To Fields 2014. CyVerse Data Commons; 2016. https://doi.org/10.7946/p2v888.
10. Lawrence-Dill C. Genomes To Fields 2015. CyVerse Data Commons; 2017. https://doi.org/10.7946/p24s31.
11. Spalding E. Genomes to fields inbred ear imaging 2017. CyVerse Data Commons; 2017. https://doi.org/10.7946/p2c34p.
12. Gage JL, et al. The effect of artificial selection on phenotypic plasticity in maize. Nat Commun. 2017;8:1348.