Tenure Package Executive Summary: James C. Schnable

Position Summary

Initial appointment: Assistant Professor at the University of Nebraska-Lincoln. Start date: May 1st, 2014. Nine-month appointment responsibilities: 80% research, 18% teaching, 2% service.

Unit Membership: Tenure home in the Department of Agronomy and Horticulture. Hired as part of a cluster for the Quantitative Life Sciences Initiative. Appointment to the Center for Plant Science Innovation after my initial hire.

Research Metrics Summary (80%)

Funding: 5 federal grants (2 USDA; 2 NSF; 1 ARPA-E), two as lead PI (1 NSF; 1 USDA). Sixth grant (NSF) currently recommended for funding. Current support from the North Central Sun Grant program, the Nebraska Corn Board, the Wheat Innovation Foundation, and the Water for Food Global Institute and the Midwest Big Data Hub. Prior support from a USAID Linkages Grant with ICRISAT (Hyderabad, India), ConAgra's Popcorn program, and Iowa Corn Board. $2.42M in funding over four years (See page 2). $1.37M in funding currently "recommended" for support by NSF. $4.78M in proposals currently under under review at the National Science Foundation, Department of Energy, and Foundation for Food and Agriculture (See page 51).

Publications: 56 total peer reviewed publications; 38 since appointment as an assistant professor at UNL; 28 resulting from work conducted at UNL; 14 papers where a member of my lab (or I) was the first or co-first author (See page 6). These papers have been cited a total of 3,166 times, or 1,930 times if two marker papers for the release of genome sequences for new grass species are excluded. H-index of 21.

Invited Presentations: 50 total invited talks or seminars; 37 since I was hired as an assistant professor at the University of Nebraska-Lincoln; 29 excluding seminars and conferences affiliated with the University of Nebraska-Lincoln. Three additional invited speaking engagements scheduled before the deadline for the final tenure package (December 1st) (See page 15)

Notable Recognition: Junior Faculty Research Award (2016); M. Rhoades Early Career Maize Genetics Award (2018) (see page 51)

Teaching and Mentoring Metrics Summary (18%)

Mentoring: 4 Postdoctoral Scholars; 1 Visiting Scientist; 3 PhD students; 1 co-advised PhD student; 2 visiting PhD students; 1 MS student; 2 co-advised MS students; 13 undergraduates – 4 REU (Research Experience for Undergraduates) students; 2 UCARE (Undergraduate Creative Activities and Research Experience) students; and 7 undergraduate students supported by regular research funding – and 2 high school students. (See page 4)

Teaching: Developed and taught a new graduate level course: Professional Development for 1st year graduate students (taught Fall '15 '16 '17 and '18. Taught "Big Question in Complex Biosystems" Fall '17 and '18. (See page 24)

Service Metrics Summary (2%)

University Service: Currently an active member of 6 university committees. Prior service on seven additional committees (4 search committees, 3 organizing committees for meetings). (See page 14)

Professional Service: Associate Editor for Molecular Plant; MaizeGDB Advisory Committee; Grant Reviewer: NSF, USDA, JGI, Genome British Columbia. Peer reviewer for 21 journals (includes Science, Nature Plants, PNAS, and Plant Cell). (See page 15)

Contents

Curriculum Vitae

James C. Schnable

Quantitative Life Sciences Initiative
Center for Plant Science Innovation
Nebraska Food for Health Center
Department of Agronomy & Horticulture
University of Nebraska-Lincoln

Office: E207 Beadle Center
Phone: (402) 472-3192
Email: schnable@unl.edu
Web: schnablelab.org

Employment

Assistant Professor 2014-present
Department of Agronomy and Horticulture, University of Nebraska-Lincoln
Start Date: May 1st. Appointment: 80% Research, 18% Teaching and Mentoring 2% Service.

NSF PGRP Fellowship Supported Visiting Scholar 2014
Chinese Academy of Agricultural Sciences

NSF PGRP Fellowship Supported Postdoctoral Researcher 2013
Donald Danforth Plant Science Center

Education

PhD Plant Biology (with Michael Freeling) 2008-2012
University of California-Berkeley

BA Biology 2004-2008
Cornell University

Honors and Awards

Marcus Rhoades Early Career Award 2018

Junior Faculty Excellence in Research Award, University of Nebraska-Lincoln 2016

Faculty Fellow, Robert B. Dougherty Water for Food Institute 2016-Present

Research Support

$2.42M in funding over four years. An additional $1.37M in funding currently recommended for support by NSF.
$4.78M in proposals currently under under review at the National Science Foundation, Department of Energy, and Foundation for Food and Agriculture (see page 51)

Current

1. "RoL: FELS: EAGER: Genetic Constraints on the Increase of Organismal Complexity Over Time"
   National Science Foundation - Rules of Life
   Schnable JC (PI)
   Award Period: 08/01/2018 - 07/31/2020
   Award Amount: $299,801

2. "Identifying mechanisms conferring low temperature tolerance in maize, sorghum, and frost tolerant relatives."
   US Department of Agriculture - National Institute of Food and Agriculture - Agriculture and Food Research Initiative
   Schnable JC (PI), Roston RL (Co-PI) *(Department of Biochemistry, UNL)*
   Award Period: 12/15/2015 - 12/14/2019
   Award Amount: $455,000

3. "In-plant and in-soil microsensors enabled high-throughput phenotyping of root nitrogen uptake and nitrogen use efficiency.'
   ARPA-E ROOTS
   Dong, L (PI) *(Department of Electrical and Computer Engineering, Iowa State)*, **Schnable JC (co-PI)**, Castellano, M (co-PI) *(Department of Agronomy, Iowa State)*
   Award Period: 06/12/2017 - 06/11/2019
   Award Amount: $1,100,000. Award Amount to UNL: $334,169

4. "Center for Root and Rhizobiome Innovation." (Investigator & Management Team Member)
   National Science Foundation - EPSCoR
   Choobineh F (PI), Cahoon E (co-PI), Alfano J (co-PI). *JCS: wrote one of four objectives in the original grant. Manages that the team of 6 PIs working on that objective. Serves on the management team for the overall project.*
   Award Period: 06/15/2016 - 05/31/2021
   Award Amount: $20M total. Awarded to UNL: $10M. Funding directly and specifically to the Schnable Lab: $649,170

5. "PAPM EAGER: Transitioning to the next generation plant phenotyping robots." (co-PI)
   USDA/NSF Joint Program
   Ge, Y (PI) *(Department of Biological Systems Engineering (BSE), UNL)*, **Schnable JC (co-PI)**, Pitla S (co-PI) *(BSE, UNL)*
   Award Period: 11/15/2016 - 11/14/2018
   Award Amount: $285,000.

6. Nebraska Corn Board "Genomes to Fields (G2F) - Predicting Final Yield Performance in Variable Environments."
   Nebraska Corn Board *(Three separate and sequential competitively awarded grants of one year each)*
   Award Period: 07/01/2016 - 06/30/2017 Award Amount: $42,620 Team: **Schnable JC (PI)**, Ge, Y (co-PI) (BSE, UNL), Rodriguez, O (co-PI) (Agronomy and Horticulture, UNL)
   Award Period: 07/01/2017 - 06/30/2018 Award Amount: $47,945 Team: **Schnable JC (PI)**, Ge, Y (co-PI) (BSE, UNL), Rodriguez, O (co-PI) (Agronomy and Horticulture, UNL)
   Award Period: 07/01/2018 - 06/30/2019 Award Amount: $51,054 Team: **Schnable JC (PI)**, Ge, Y (co-PI) (BSE, UNL), Shi, Y (co-PI) (BSE, UNL)

7. "Optimizing the Water Use Efficiency of C4 Grain Crops Using Comparative Phenomics and Crop Models to Guide Breeding Targets."
   Daugherty Water for Food Global Institute – University of Nebraska Foundation *(Original Award and Accomplishment Based Renewal)*
   Award Period: 07/01/2017 - 06/30/2018 Award Amount: $17,658 Team: **Schnable JC (PI)**
   Award Period: 07/01/2018 - 06/30/2019 Award Amount: $12,000 Team: **Schnable JC (PI)**

8. "High through put phenotyping to accelerate biomass sorghum improvement." (co-PI)
   Sun Grant Program - North Central Region *(Original Award and Competitive Renewal)*
   Award Period: 09/01/2016 - 12/31/2018 Award Amount: $149,633 Team: Ge, Y (PI) **Schnable JC (co-PI)**, Sibel Irmak (co-PI) (BSE, UNL), Jin, J (Agricultural & Biological Engineering, Purdue)
   Award Period: 07/01/2018 - 06/30/2019 Award Amount: $44,000 Team: Ge, Y (PI) **Schnable JC (co-PI)**

9. "A Low-Cost, High-Throughput Cold Stress Perception Assay for Sorghum Breeding."
   Wheat Innovation Fund – University of Nebraska Foundation
   Roston RL (PI) (Biochemisty, UNL), **Schnable JC (co-PI)**
   Award Period: 1/1/2019 - 12/31/2021
   Award Amount: $205,000

10. "Automatic feature extraction pipeline development for high-throughput plant phenotyping"
    National Science Foundation Big Data Hub - Competitive Subaward
    Xu, Z (PI) (Statistics, UNL), Cui J (co-PI) (Computer Science and Engineering (CSE), UNL), Ge, Y (co-PI) (BSE, UNL), Qiu Y (co-PI) (Statistics, UNL), **Schnable JC (co-PI)**
    Award Period: 10/01/2017 - 09/30/2018
    Award Amount: $5,000

Completed

11. "Application of tGBS And Genomic Selection to a Hybrid Pearl Millet Breeding Program."
    USAID Linkages Grant with ICRISAT (A CGAR Center)
    Schnable JC (PI), Gupta SK (co-PI) (Pearl Millet Breeding Lead, ICRISAT), Schnable PS (co-PI) (Agronomy, Iowa State)
    Award Period: 7/1/15 - 9/30/17
    Award Amount: $45,000 Awarded to UNL: $23,000

12. "Field Deployable Cameras to Quantify Dynamic Whole Plant Phenotypes in the Field."
    Iowa Corn Board
    Schnable JC (PI)
    Award Period: 7/1/14 - 6/30/16
    Award Amount: $45,395

13. "Marker Discovery & Genetic Diversity. (In Popcorn)"
    ConAgra Foods
    Lorenz A (original PI), **Schnable JC (replacement PI)**
    Award Period: 01/01/2014 - 12/31/2017 Award Amount: $162,284

14. "A High Throughput Phenotyping Reference Dataset for GWAS in Sorghum"
    Agricultural Research Division - Internal UNL Funding
    Schnable JC (PI), Ge Y (co-PI) (BSE, UNL), Qiu Y (co-PI) (Statistics, UNL), Samal A (co-PI) (CSE, UNL), Sigmon B (Agronomy & Horticulture, UNL)
    Award Period: 07/01/2016 - 6/30/2018
    Award Amount: $99,159

Recommended for Funding

15. "RII Track-2 FEC: Functional analysis of nitrogen responsive networks in Sorghum"
    National Science Foundation - EPSCoR
    Schmutz J (PI) (Hudson Alpha), Lamb N (co-PI) (Hudson Alpha), **Schnable JC (co-PI)**, Swaminathan K (co-PI) (Hudson Alpha), Clemente T (co-PI) (Agronomy and Horticulture, UNL)

Award Period: 01/01/2019 - 12/31/2022
Award Amount: $4M total. Awarded to UNL: $1,337,633. Funding directly and specifically to the Schnable Lab: $648,710

Economic Development

Co-Founder, EnGeniousAg LLC 2017-Present
Designs, manufactures, and deploys low-cost, instant readout, high-performance, field-based nutrient sensors for crops, soil, and water, improving agronomic management practices, increasing grower profitability and reducing the environmental footprint of agriculture.

Founder, Dryland Genetics LLC 2014-Present
Using high throughput quantitative genetics and field phenotyping techologies to develop and commericialize higher yielding cultivars of crops already naturally adapted to using little water and growing arid regions where conventional agriculture fails in the absence of irrigation.

Co-Founder, Data2Bio LLC (USA) & DATA生物科技（北京）有限公司 (China) 2010-Present
Providing patented tGBS genotyping and genomic selection services to public and private sector plant and animal breeders in the USA and China.

Scientific Advisory Council, GeneSeek, Inc 2017-Present

External Advisor to the Scientific Advisory Board, Indigo Agriculture 2017

External Advisor to the Scientific Advisory Board, Syngenta AG 2016

Mentoring

Table 1: **Graduate Students Mentored**

Student	**Degree**	**Program**	**Years**
Zhikai Liang	PhD	Agronomy & Horticulture	2015-Present
Daniel Carvalho	PhD	Agronomy & Horticulture	2015-Present
Chenyong Miao	PhD	Agronomy & Horticulture	2016-Present
Preston Hurst	MS	Agronomy & Horticulture	2016-Present
Nate Korth	PhD (co-advised)	Food Science and Technology	2017-Present
Xiuru Dai	PhD (CSC student)	Shandong Agriculture University	2017-Present
Xianjun Lai	PhD (CSC student)	Sichuan Agriculture University	2015-2017
Bhushit Agarwall	MS (co-advised)	Computer Science & Engineering	2016-2017
Srinidhi Bashyan	MS (co-advised)	Computer Science & Engineering	2016-2017

Table 2: **Undergraduates and High School Students Engaged In Research**

Student	**Program**	**Years**
Daniel Ngu	Startup and Grant Funds	2014-2017
Kyle Johnson	Bioenergy REU Program	Summer 2016
Taylor Horn	QLSI REU Program	Summer 2016
Logan Olson	Startup and Grant Funds	2016-2017
Xiaoyang Ye	Startup and Grant Funds	2016-2017
Holly Podliska	UCARE Program	2016-2017
Nicole Hollander	HS Student (Young Nebraska Scientist Program)	Summer 2017
Connor Pedersen	Grant Funds	2016-2018
Tom Hoban	Startup and Grant Funds	2016-Present
Isabel Sigmon	HS Student (Grant Funds)	Summer 2018
Christian Butera	Bioenergy REU Program	Summer 2018
Ashley Foltz	CRRI REU Program	Summer 2018
Alex Enersen	Grant Funds	2018-Present
Alexandra Bradley	Grant Funds	2018-Present
Alejandro Pages	UCARE Program	2018-Present

Table 3: **Postdoctoral Scholars and Visiting Scientists**

Name	**Years**	**Present Position**
Yang Zhang (Postdoc)	2014-2017	Scientist, St. Jude's Children Hospital
Lang Yan (Visiting Scholar)	2016-2017	Deputy Director, Potato Functional Genomics, XiChang College
Jinliang Yang (Postdoc)	2016-2107	Assistant Professor, University of Nebraska-Lincoln
Sunil Kumar (Postdoc)	2017-2018	Postdoc, Niederhuth Lab, Michigan State
Guangchao Sun (Postdoc)	2017-Present	Schnable Lab
Xiaoxi Meng (Postdoc)	2018-Present	Schnable Lab

Publications

H-Index: 21 Google Scholar
*Lab members in bold, *equal contribution, ‡undergraduate, §corresponding*

The quality, importance, and impact of scientific articles can be assessed in numerous ways. Some methods, such as individual citation counts, are considered to have greater reliability, but take longer to accumulate. Based on department guidelines, I am providing the Impact Factors for each journal in which I have published a manuscript since beginning my position at the University of Nebraska-Lincoln.

In addition, I am providing the number of citations for each article retrieved from Google Scholar, and each article's Altmetric score, an aggregate estimate of the amount of attention an article receives in the press and social media following publication. Altmetric scores have been shown to exhibit a statistically significant, albeit imperfect, correlation with future citation counts, and provide a non-impact factor based estimator of individual article significance which can be quantified for rapidly that citation counts which often require several years to accumulate.

Preprints

Miao C, **Yang, J**, **Schnable JC**.§ Optimizing the identification of causal variants across varying genetic architectures in crops. BIORXIV doi: 10.1101/310391

Yan L, **Raju SKK**, **Lai X**, **Zhang Y**, **Dai X**, Rodriguez O, Mahboub S, Roston RL, **Schnable JC**.§ Parallels between artificial selection in temperate maize and natural selection in the cold-adapted crop-wild relative *Tripsacum*. BIORXIV doi: 10.1101/187575

Other Manuscripts in Review

Zou C, Miki D, Li D, Tang Q, Xiao L, Rajput S, Deng P, Peng L, Huang R, Zhang M, Sun Y, Hu J, Fu X, Schnable P, Li F, Zhang H, Feng B, Zhu X, Liu R, **Schnable JC**, Zhu JK, Zhang H.§ The genome of broomcorn millet (*Panicum miliaceum* L.) *(In Review)*

Faculty Publications

56. **Lai X**, **Yan L**, Lu Y, **Schnable JC**§ (2018) Largely unlinked gene sets targeted by selection for domestication syndrome phenotypes in maize and sorghum. THE PLANT JOURNAL doi: 10.1111/tpj.13806 BIORXIV doi: 10.1101/184424
ALTMETRIC SCORE: 22 (87th percentile for papers of similar age (+/- 6 weeks) published in this journal).
TIMES CITED TO DATE: 1
JOURNAL IMPACT FACTOR (2017): 5.8
SCHNABLE LAB CONTRIBUTION: *All analyses and writing conducted by lab members.*

55. **Liang Z**, Gupta SK, Yeh CT, **Zhang Y**, **Ngu DW**,‡ Kumar R, Patil HT, Mungra KD, Yadav DV, Rathore A, Srivastava RK, Gupkta R, **Yang J**, Varshney RK, Schnable PS, **Schnable JC**§ (2018) Phenotypic data from inbred parents can improve genomic prediction in pearl millet hybrids. G3: GENES GENOMES GENETICS doi: 10.1534/g3.118.200242
ALTMETRIC SCORE: 26 (97th percentile for papers of similar age (+/- 6 weeks) published in this journal).
TIMES CITED TO DATE: 0
JOURNAL IMPACT FACTOR (2017): 2.7
SCHNABLE LAB CONTRIBUTION: *Built the libraries, analyzed the SNP data, conducted the GS tests, wrote the paper. Field data and extracted DNA contributed by ICRISAT collaborators. Sequencing and SNP calling contributed by ISU collaborators.*

54. **Miao C**, Fang J, Li D, Liang P, Zhang X, **Yang J**, **Schnable JC**, Tang H§ (2018) Genotype-Corrector: improved genotype calls for genetic mapping. Scientific Reports doi: 10.1038/s41598-018-28294-0
Altmetric Score: 29 (91st percentile for papers of similar age (+/- 6 weeks) published in this journal).
Times Cited to Date: 0
Journal Impact Factor (2017): 4.1
Schnable Lab Contribution: *Improved and documented the core algorithm. Conducted tests of how much the core algorithm improved genotype call accuracy in a RIL and F2 population when using sub-optimal sequencing depth. Wrote the paper collaboratively with Haibao Tang.*

53. **Raju SKK**, Barnes A, **Schnable JC**, Roston RL§ (2018) Low-temperature tolerance in land plants: Are transcript and membrane responses conserved? Plant Science doi: 10.1016/j.plantsci.2018.08.002
Altmetric Score: 13 (97th percentile for papers of similar age (+/- 6 weeks) published in this journal).
Times Cited to Date: 0
Journal Impact Factor (2017): 3.7
Schnable Lab Contribution: *Sunil and I wrote the portions of this review focused on conserved patterns transcriptional responses to cold stress across diverse plants, and worked collaboratively with the Roston lab on the combined transcript/lipid analyses.*

52. **Carvalho DS**, **Schnable JC**, Almeida AMR§ (2018) Integrating phylogenetic and network approaches to study gene family evolution: the case of the AGAMOUS family of floral genes. Evolutionary Bioinformatics doi: 10.1177/1176934318764683 bioRxiv doi: 10.1101/195669
Altmetric Score: 6 (89th percentile for papers published in this journal). *Insufficient papers of similar age to generate a percentile ranking.*
Times Cited to Date: 0
Journal Impact Factor (2017): 1.9
Schnable Lab Contribution: *Conducted the majority of the analyses. Wrote the paper.*

51. Xu Y, Qiu Y, **Schnable JC**§ (2018) Functional modeling of plant growth dynamics. The Plant Phenome doi: 10.2135/tppj2017.09.0007 bioRxiv doi: 10.1101/190967
Altmetric Score: 13 (Altmetric score from preprint. 67th percentile for preprints of similar age (+/- 6 weeks) on bioRxiv).
Times Cited to Date: 0
Journal Impact Factor (2017): Not Yet Assigned
Schnable Lab Contribution: *Conceived of the experiment to test subsampling on different days. Wrote the paper*

50. Ott A, **Schnable JC**, Yeh CT, Wu L, Liu C, Hu HC, Dolgard CL, Sarkar S, Schnable PS.§ (2018) Linked read technology for assembling large complex and polyploid genomes. BMC Genomics *(Accepted)*
Altmetric Score: NA.
Times Cited to Date: NA
Journal Impact Factor (2017): 3.7
Schnable Lab Contribution: *Conducted an analysis of a linked read genome assembly of proso millet, a previously unsequenced allotetraploid grass to assess the accuracy with which separate subgenomes were assembled and resolved using this new linked-reads technique.*

49. Liu S,* **Schnable JC**,* Ott A,* Yeh CT, Springer NM, Yu J, Meuhbauer G, Timmermans MCP, Scanlon MJ, Schnable PS§ (2018) Intragenic Meiotic Crossovers Generate Novel Alleles with Transgressive Expression Levels. Molecular Biology and Evolution *(Accepted)*
Altmetric Score: NA.
Times Cited to Date: NA
Journal Impact Factor (2017): 10.2
Schnable Lab Contribution: *Analysis of relative correlation between recombination frequency per megabase and the relative density of either syntenic or nonsyntenic genes separately was conducted in the Schnable Lab@UNL.*

48. Alkhalifah N, Campbell DA, Falcon CM, ... **Schnable JC** (31 of 44 authors) ... Spalding EP, Edwards J, Lawrence-Dill CJ[§] (2018) Maize Genomes to Fields: 2014 and 2015 field season genotype, phenotype, environment, and inbred ear image datasets. BMC Research Notes doi: 10.1186/s13104-018-3508-1
Altmetric Score: 12 (89th percentile for papers of similar age (+/- 6 weeks) published in this journal).
Times Cited to Date: 0
Journal Impact Factor (2017): Not Yet Assigned
Schnable Lab Contribution: *Data collection from grow outs of Genomes to Fields hybrids at Nebraska field sites, assisted in writing the manuscript itself.*

47. **Zhang Y**, **Ngu DW**,[‡] **Carvalho D**, **Liang Z**, Qiu Y, Roston RL, **Schnable JC**[§] (2017) Differentially regulated orthologs in sorghum and the subgenomes of maize. The Plant Cell doi: 10.1105/tpc.17.00354
Selected as an Editor's Choice by MaizeGDB Editorial Board August 2017
Altmetric Score: 34 (97th percentile for papers published in this journal). *Insufficient papers of similar age to generate a percentile ranking.*
Times Cited to Date: 6
Journal Impact Factor (2017): 8.2
Schnable Lab Contribution: *All analyses and writing conducted by lab members. Qiu lab assisted in developing new analytical approaches to comparing gene expression across species. Roston lab assisted in interpreting biological responses to plant cold stress.*

46. **Lai X**,[*] Behera S,[*] **Liang Z**, Lu Y, Deogun JS, **Schnable JC**[§] (2017) STAG-CNS: An order-aware conserved noncoding sequence discovery tool for arbitrary numbers of species. Molecular Plant. doi: 10.1016/j.molp.2017.05.010
Altmetric Score: 8 (62nd percentile for papers of similar age (+/- 6 weeks) published in this journal).
Times Cited to Date: 3
Journal Impact Factor (2017): 9.3
Schnable Lab Contribution: *I defined the problem, Sairam Behera created the algorithm, Zhikai Liang and Xianjun Lai, both from my group, conducted multiple rounds of biological validation and provided feedback to Sairam, improving the core algorithm in an iterative process. My lab wrote the paper.*

45. **Liang Z**, Pandey P, Stoerger V, Xu Y, Qiu Y, Ge Y, **Schnable JC**[§] (2017) Conventional and hyperspectral time-series imaging of maize lines widely used in field trials. GigaScience doi: 10.1093/gigascience/gix117 bioRxiv doi: 10.1101/169045
Altmetric Score: 14 (61st percentile for papers of similar age (+/- 6 weeks) published in this journal).
Times Cited to Date: 3
Journal Impact Factor (2017): 7.3
Schnable Lab Contribution: *All analyses and writing conducted by lab members. Ge, Qiu, and Xu labs each assisted in developing new analytical approaches. Vincent Stoerger assisted with data generation.*

44. **Liang Z**, **Schnable JC**[§] (2017) Functional divergence between subgenomes and gene pairs after whole genome duplications. Molecular Plant doi: 10.1016/j.molp.2017.12.010
Altmetric Score: 8 (65th percentile for papers of similar age (+/- 6 weeks) published in this journal).
Times Cited to Date: 2
Journal Impact Factor (2017): 9.3
Schnable Lab Contribution: *All analyses writing conducted by lab members.*

43. Pandey P, Ge Y[§], Stoerger V, **Schnable JC** (2017) High throughput in vivo analysis of plant leaf chemical properties using hyperspectral imaging. Frontiers in Plant Science doi 10.3389/fpls.2017.01348
Altmetric Score: 15 (96th percentile for papers of similar age (+/- 6 weeks) published in this journal).
Times Cited to Date: 20
Journal Impact Factor (2017): 3.7

Schnable Lab Contribution: *Designed a different cross validation technique which was implemented by the first author. Drafted portions of the introduction and discussion and revised the manuscript.*

42. Gage J, Jarquin D, Romay M, ... **Schnable JC** (29th of 40 authors) .. Yu J, de Leon N[§] (2017) The effect of artificial selection on phenotypic plasticity in maize. Nature Communications doi: 10.1038/s41467-017-01450-2
Selected as an Editor's Choice by MaizeGDB Editorial Board December 2017
Altmetric Score: 85 (83st percentile for papers of similar age (+/- 6 weeks) published in this journal).
Times Cited to Date: 5
Journal Impact Factor (2017): 12.4
Schnable Lab Contribution: *Generated and contributed yield and field phenotyping data from Nebraska field sites of Genomes to Fields project.*

41. Washburn JD, **Schnable JC**, Brutnell TP, Shao Y, **Zhang Y**, Ludwig M, Davidse G, Pires JC[§] (2017) Genome-guided phylo-transcriptomic methods and the nuclear phylogentic tree of the paniceae grasses. Scientific Reports doi: 10.1038/s41598-017-13236-z
Altmetric Score: 1 (36th percentile for papers of similar age (+/- 6 weeks) published in this journal).
Times Cited to Date: 1
Journal Impact Factor (2017): 4.1
Schnable Lab Contribution: *Grew plants, extracted RNA, built and sequenced libraries and shared data. Consulted with the lead author on the syntenic gene analysis.*

40. Ott A,* Liu S,* **Schnable JC**, Yeh CT, Wang C, Schnable PS[§] (2017) Tunable Genotyping-By-Sequencing (tGBS®) enables reliable genotyping of heterozygous loci. Nucleic Acids Research doi: 10.1093/nar/gkx853
Altmetric Score: 15 (87th percentile for papers of similar age (+/- 6 weeks) published in this journal).
Times Cited to Date: 4
Journal Impact Factor (2017): 11.6
Schnable Lab Contribution: *Wrote portions of the manuscript, designed additional analyses to validate datasets which were executed by Alina Ott.*

39. **Lai X**, **Schnable JC**, Liao Z, Xu J, Zhang G, Li C, Hu E, Rong T, Xu Y, Lu Y[§] (2017) Genome-wide characterization of non-reference transposable elements insertion polymorphisms reveals genetic diversity in tropical and temperate maize. BMC Genomics doi: 10.1186/s12864-017-4103-x
Altmetric Score: 0
Times Cited to Date: 3
Journal Impact Factor (2017): 3.7

textscSchnable Lab Contribution: textitThe majority of this paper was written by Xianjun Lai during his time in the Schnable lab. I redesigned several analyses for him to carry out and helped to re-write the paper.

38. Mei W, Boatwright L, Feng G, **Schnable JC**, Barbazuk WB[§] (2017) Evolutionarily conserved alternative splicing across monocots. Genetics doi: 10.1534/genetics.117.300189
Cover Article October 2017 Issue
Altmetric Score: 24 (88th percentile for papers of similar age (+/- 6 weeks) published in this journal).
Times Cited to Date: 6
Journal Impact Factor (2017): 4.1
Schnable Lab Contribution: *Conceived and designed a new approach to identifying orthologous plant exons based on a directed acyclic graph which was robust to the insertion or deletion of entire introns.*

37. Mei W, Liu S, **Schnable JC**, Yeh C, Springer NM, Schnable PS, Barbazuk WB[§] (2017) A comprehensive analysis of alternative splicing in paleopolyploid maize. Frontiers in Plant Science doi:

10.3389/fpls.2017.00694
Altmetric Score: 9 (90th percentile for papers of similar age (+/- 6 weeks) published in this journal).
Times Cited to Date: 14
Journal Impact Factor (2017): 3.7
Schnable Lab Contribution: *Developed approached to identifying orthologous exons across both maize subgenomes and co-orthologous genes in sorghum (an earlier iteration of the algorithm later used for paper # 42). Consulted with the lead author on the best ways to make comparisons across subgenomes.*

36. Walley JW,* Sartor RC,* Shen Z, Schmitz RJ, Wu KJ, Urich MA, Nery JR, Smith LG, **Schnable JC**, Ecker JR, Briggs SP§ (2016) Integration of omic networks in a developmental atlas of maize. Science doi: 10.1126/science.aag1125
Selected as an Editor's Choice by MaizeGDB Editorial Board September 2016
Altmetric Score: 82 (77th percentile for papers of similar age (+/- 6 weeks) published in this journal).
Times Cited to Date: 48
Journal Impact Factor (2017): 41.1
Schnable Lab Contribution: *Suggested and provided the data which enabled the separate analysis of maize syntenic and non-syntenic genes. This analysis lead to the discovery than non-syntenic maize genes are much less likely to be translated enough protein, even when they are transcribed into mRNAs than genes conserved at syntenic locations across multiple grass species. See Figure 1 in the final paper.*

35. Ge Y§, Bai G, Stoerger V, **Schnable JC** (2016) Temporal dynamics of maize plant growth, water use, and plant water content using automated high throughput RGB and hyperspectral imaging. Computers and Electronics in Agriculture doi: 10.1016/j.compag.2016.07.028
Altmetric Score: 29 (99th percentile for papers published in this journal). *Single highest altmetric score recorded for this journal.*
Times Cited to Date: 35
Journal Impact Factor (2017): 2.4
Schnable Lab Contribution: *Provided plant material. Interpreted a portion of the resulting trait datasets. Wrote portions of the manuscript particularly those focused on the biological relevance of the measured traits.*

34. **Liang Z**, **Schnable JC**§ (2016) RNA-seq based analysis of population structure within the maize inbred B73. PLoS One doi: 10.1371/journal.pone.0157942
Altmetric Score: 16 (90th percentile for papers of similar age (+/- 6 weeks) published in this journal).
Times Cited to Date: 4
Journal Impact Factor (2017): 2.8
Schnable Lab Contribution: *All analyses and writing conducted by lab members.*

33. Rajput SG, Santra DK§, **Schnable JC** (2016) Mapping QTLs for morpho-agronomic traits in proso millet (*Panicum miliaceum* L.). Molecular Breeding doi: 10.1007/s11032-016-0460-4
Altmetric Score: 7 (87th percentile for published in this journal). Times Cited to Date: 4
Journal Impact Factor (2017): 2.1
Schnable Lab Contribution: *Taught Santosh Rajput how to analyze GBS data. Conducted analyses to generate a filtered set of dominant markers from homeologous loci collapsed across subgenomes which were ultimately used to generate the genetic map created in this paper.*

32. Joyce BL, Huag-Baltzell A, Davey S, Bomhoff M, **Schnable JC**, Lyons E§ (2016) FractBias: a graphical tool for assessing fractionation bias after whole genome duplications. Bioinformatics doi: 10.1093/bioinformatics/btw666
Altmetric Score: 4 (72nd percentile for papers of similar age (+/- 6 weeks) published in this journal).
Times Cited to Date: 2
Journal Impact Factor (2017): 5.5
Schnable Lab Contribution: *Generated semi-manual subgenome assignments which are used as the basis*

for evaluating the accuracy of the automated assignments made by FractBias. Consulted with the lead author on the best ways to make comparisons across subgenomes.

31. Chao S, Wu J, Liang J, **Schnable JC**, Yang W, Cheng F, Wang X[§] (2016) Impacts of whole genome triplication on MIRNA evolution in *Brassica rapa*. Genome Biology and Evolution doi: 10.1093/gbe/evv206
Altmetric Score: 6 (43rd percentile for papers of similar age (+/- 6 weeks) published in this journal).
Times Cited to Date: 10
Journal Impact Factor (2017): 3.9
Schnable Lab Contribution: *Contributed to the design of comparisons across the Brassica rapa subgenomes. Assisted in drafting portions of the paper.*

30. Tang H, Bomhoff MD, Briones E, **Schnabe JC**, Lyons E[§] (2015) SynFind: compiling syntenic regions across any set of genomes on demand. Genome Biology and Evolution doi: 10.1093/gbe/evv219
Altmetric Score: 12 (86th percentile for papers of similar age (+/- 6 weeks) published in this journal).
Times Cited to Date: 22
Journal Impact Factor (2017): 3.9
Schnable Lab Contribution: *Tested/validated the orthology assignments of the core algorithms and provided feedback to the author to improve the algorithm in an iterative process. Drafted portions of the manuscript related to validation of the core algorithm.*

29. Washburn JD, **Schnable JC**, Davidse G, Pires JC[§] (2015) Phylogeny and photosynthesis of the grass tribe Paniceae. American Journal of Botany doi: 10.3732/ajb.1500222
Altmetric Score: 5 (58th percentile for papers of similar age (+/- 6 weeks) published in this journal).
Times Cited to Date: 22
Journal Impact Factor (2017): 2.8
Schnable Lab Contribution: *Collected germplasm, grew plants and contributed tissue for plastid sequencing. Revised the research questions to be addressed in the manuscript with the first and last authors.*

28. Tang H, Zhang X, **Miao C**, Zhang J, Ming R, **Schnable JC**, Schnable PS, Lyons E, Lu J[§] (2015) ALLMAPS: robust scaffold ordering based on multiple maps. Genome Biology doi: 10.1186/s13059-014-0573-1
Altmetric Score: 4 (11th percentile for papers of similar age (+/- 6 weeks) published in this journal).
Times Cited to Date: 48
Journal Impact Factor (2017): 13.2
Schnable Lab Contribution: *Designed analyses for validating the accuracy of consensus ordering produced by the ALLMAPS algorithm. Contributed data for validation of the algorithm. Created visualizations for figures. Wrote or re-wrote portions of the manuscript text.*

27. **Schnable JC**[§] (2015) Genome evolution in maize: from genomes back to genes. Annual Review of Plant Biology doi: 10.1146/annurev-arplant-043014-115604
Altmetric Score: 15 (92nd percentile for papers of similar age (+/- 6 weeks) published in this journal).
Times Cited to Date: 18
Journal Impact Factor (2017): 18.7
Schnable Lab Contribution: *Wrote the manuscript.*

26. Paschold A, Larson NB, Marcon C, **Schnable JC**, Yeh C, Lanz C, Nettleton D, Piepho H, Schnable PS, Hochholdinger F[§] (2014) Nonsyntenic genes drive highly dynamic complementation of gene expression in maize hybrids. Plant Cell doi: 10.1105/tpc.114.130948
Altmetric Score: 18 (93rd percentile for papers of similar age (+/- 6 weeks) published in this journal).
Times Cited to Date: 27

Journal Impact Factor (2017): 8.2
Schnable Lab Contribution: *Suggested a key analyses to the remaining authors – the separation of maize genes into syntenically conserved and non-syntenic classes to look for different patterns of gene expression in the F1 hybrid – and provided the datasets and analytical approaches necessary to conduct this analysis.*

Postdoctoral Publications

25. Nani TF, **Schnable JC**, Washburn JD, Albert P, Pereira WA, Sobrinho FS, Birchler JA, Techia VH[§] (2018). Location of low copy genes in chromosomes of *Brachiaria* spp. Molecular Biology Reports doi: 10.1007/s11033-018-4144-5
Times Cited to Date: 0
Journal Impact Factor (2017): 1.9

24. Studer AJ[*], **Schnable JC**[*], Weissmann S, Kolbe AR, McKain MR, Shao Y, Cousins AB, Kellogg EA, Brutnell TP[§] (2016) The draft genome of *Dichanthelium oligosanthes*: A C3 panicoid grass species. Genome Biology doi: 10.1186/s13059-016-1080-3
Times Cited to Date: 8
Journal Impact Factor (2017): 13.2

23. Huang P, Studer AJ, **Schnable JC**, Kellogg EA, Brutnell TP[§] (2016) Cross species selection scans identify components of C4 photosynthesis in the grasses. Journal of Experimental Botany doi: 10.1093/jxb/erw256
"Insight" highlighting this article by PA Christin also published in JXB doi: 10.1093/jxb/erw390
Times Cited to Date: 17
Journal Impact Factor (2017): 5.4

22. Liu X, Tang S, Jia G, **Schnable JC**, Su X, Tang C, Zhi H, Diao X[§] (2016) The C-terminal motif of SiAGO1b is required for the regulation of growth, development and stress responses in foxtail millet [*Setaria italica* (L.) P. Beauv]. Journal of Experimental Botany doi: 10.1093/jxb/erw135
Times Cited to Date: 13
Journal Impact Factor (2017): 5.4

21. Jia G, Liu X, **Schnable JC**, Niu Z, Wang C, Li Y, Wang Sh, Wang Su, Liu J, Gou E, Diao X[§] (2015) Microsatellite variations of elite Setaria varieties released during last six decades in China. PLoS One doi: 10.1371/journal.pone.0125688
Times Cited to Date: 11
Journal Impact Factor (2017): 2.8

20. Qie L, Jia G, Zhang W, **Schnable JC**, Shang Z, Li W, Liu B, Li M, Chai, Y, Zhi H, Diao X[§] (2014) Mapping of quantitative trait loci (QTLs) that contribute to germination and early seedling drought tolerance in the interspecific cross *Setaria italica* x *Setaria viridis*. PLoS One doi: 10.1371/journal.pone.0101868
Times Cited to Date: 33
Journal Impact Factor (2017): 2.8

19. Diao X[§], **Schnable JC**, Bennetzen JL, Li J[§] (2014) Initiation of Setaria as a model plant. Frontiers of Agricultural Science and Engineering doi: 10.15302/J-FASE-2014011
Times Cited to Date: 48
Journal Impact Factor (2017): Impact Factor Not Yet Assigned

Graduate Publications

18. Cheng F, Sun C, Wu J, **Schnable JC**, Woodhouse MR, Liang J, Cai C, Freeling M,[§] Wang X[§] (2016) Epigenetic regulation of subgenome dominance following whole genome triplication in *Brassica rapa*. New Phytologist doi: 10.1111/nph.13884

Times Cited to Date: 19
Journal Impact Factor (2017): 7.4

17. Almeida AMR, Yockteng R, **Schnable JC**, Alvarez-Buylla ER, Freeling M, Specht CD[§] (2014) Co-option of the polarity gene network shapes filament morphology in angiosperms. Scientific Reports doi: 10.1038/srep06194
Times Cited to Date: 8
Journal Impact Factor (2017): 41

16. Martin JA, Johnson NV, Gross SM, **Schnable JC**, Meng X, Wang M, Coleman-Derr D, Lindquist E, Wei C, Kaeppler S, Chen F, Wang Z[§] (2014) A near complete snapshot of the *Zea mays* seedling transcriptome revealed from ultra-deep sequencing. Scientific Reports doi: 10.1038/srep04519
Selected as an Editor's Choice by MaizeGDB Editorial Board May 2014
Times Cited to Date: 21
Journal Impact Factor (2017): 4.1

15. Garsmeur O,[*] **Schnable JC**,[*] Almeida A, Jourda C, D'Hont A,[§] Freeling M[§] (2014) Two evolutionarily distinct classes of paleopolyploidy. Molecular Biology and Evolution doi: 10.1093/molbev/mst230
Times Cited to Date: 84
Journal Impact Factor (2017): 10.2

14. Turco G, **Schnable JC**, Pedersen B, Freeling M[§] (2013) Automated conserved noncoding sequence (CNS) discovery reveals differences in gene content and promoter evolution among the grasses. Frontiers in Plant Sciences doi: 10.3389/fpls.2013.00170
Times Cited to Date: 21
Journal Impact Factor (2017): 3.7

13. **Schnable JC**, Wang X, Pires JC, Freeling M[§] (2012) Escape from preferential retention following repeated whole genome duplication in plants. Frontiers in Plant Science doi: 10.3389/fpls.2012.00094
Times Cited to Date: 48
Journal Impact Factor (2017): 3.7

12. Freeling M[§], Woodhouse MR, Subramaniam S, Turco G, Lisch D, **Schnable JC** (2012) Fractionation mutagenesis and similar consequences of mechanisms removing dispensable or less-expressed DNA in plants. Current Opinion in Plant Biology doi: 10.1016/j.pbi.2012.01.015
Times Cited to Date: 105
Journal Impact Factor (2017): 7.3

11. Tang H[§], Woodhouse MR, Cheng F, **Schnable JC**, Pedersen BS, Conant GC, Wang X, Freeling M, Pires JC (2012) Altered patterns of fractionation and exon deletions in Brassica rapa support a two-step model of paleohexaploidy. Genetics doi: 10.1534/genetics.111.137349
Times Cited to Date: 111
Journal Impact Factor (2017): 4.1

10. **Schnable JC**, Freeling M, Lyons E[§] (2012) Genome-wide analysis of syntenic gene deletion in the grasses. Genome Biology and Evolution doi: 10.1093/gbe/evs009
Selected as an Editor's Choice by MaizeGDB Editorial Board Dec 2012
Times Cited to Date: 106
Journal Impact Factor (2017): 3.9

9. Zhang W, Wu Y, **Schnable JC,** Zeng Z, Freeling M, Crawford GE, and Jiang J[§] (2012) High-resolution mapping of open chromatin in the rice genome. Genome Research doi: 10.1101/gr.131342.111
Times Cited to Date: 110
Journal Impact Factor (2017): 4.1

8. Eichten SR,* Swanson-Wagner RA,* **Schnable JC,** Waters AJ, Hermanson PJ, Liu S, Yeh C, Jia Y, Gendler K, Freeling M, Schnable PS, Vaughn MW, Springer NM§ (2011) Heritable epigenetic variation among maize inbreds. PLoS Genetics doi: 10.1371/journal.pgen.1002372
Selected as an Editor's Choice by MaizeGDB Editorial Board Jan 2012
Times Cited to Date: 115
Journal Impact Factor (2017): 10.1

7. **Schnable JC**, Lyons E§ (2011) Comparative genomics with maize and other grasses: from genes to genomes. Maydica 56(1763) 77-93 Link directly to PDF
Times Cited to Date: 11
Journal Impact Factor (2017): 0.2

6. Tang H, Lyons E, Pedersen B, **Schnable JC**, Paterson AH, Freeling M. (2011) Screening synteny blocks in pairwise genome comparisons through integer programming. BMC Bioinformatics doi: 10.1186/1471-2105-12-102
Times Cited to Date: 71
Journal Impact Factor (2017): 2.2

5. **Schnable JC**, Pedersen BS, Subramaniam S, Freeling M§ (2011) Dose-sensitivity, conserved noncoding sequences and duplicate gene retention through multiple tetraploidies in the grasses. Frontiers in Plant Science doi: 10.3389/fpls.2011.00002
Commentary by Birchlier and Veitia also published in Frontiers in Plant Science doi: 10.3389/fpls.2011.00064
Times Cited to Date: 29
Journal Impact Factor (2017): 3.7

4. **Schnable JC**§, Freeling M (2011) Genes identifed by visible mutant phenotypes show increased bias towards one of two maize subgenomes. PLoS One doi: 10.1371/journal.pone.0017855
Times Cited to Date: 102
Journal Impact Factor (2017): 2.8

3. **Schnable JC**, Springer NM, Freeling M§ (2011) Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. Proceedings of the National Academy of Sciences doi: 10.1073/pnas.1101368108
Selected as an Editor's Choice by MaizeGDB Editorial Board May 2011
Times Cited to Date: 327
Journal Impact Factor (2017): 9.5

2. Woodhouse MR,* **Schnable JC**,* Pedersen BS, Lyons E, Lisch D, Subramaniam S, Freeling M§ (2010) Following tetraploidy in maize, a short deletion mechanism removed genes preferentially from one of the two homeologs. PLoS Biology doi: 10.1371/journal.pbio.1000409
Selected as an Editor's Choice by MaizeGDB Editorial Board August 2010
PLoS Biology Cover Article
Times Cited to Date: 187
Journal Impact Factor (2017): 9.1

1. The International Brachypodium Initiative (2010) Genome sequencing and analysis of the model grass Brachypodium distachyon. Nature doi: 10.1038/nature08747
Times Cited to Date: 1,226
Journal Impact Factor (2017): 41.6

Service: selected, 2014-Present

University

Consortium for Integrated Translational Biology (CITB) 2014-Present

UNL Faculty Greenhouse Committee	2015-Present
Department of Agronomy and Horticulture Peer Evaluation Committee	2016-Present
Biotech Seminar Series Committee	2017-Present
Agronomy and Horticulture Faculty Advisory Committee	2017-Present
Nebraska Food for Health Center Faculty Advisory Committee	2017-Present
Organizing Committee "International Millet Symposium 2018"	2018
Organizing Committee "Predictive Crop Design, Genome to Phenome"	2017
Search Committee, Director of Phenomic Sciences	2017
Search Committee, Agricultural Research Division	2016
Search Committee, Quantitative Life Sciences Initiative	2016
Search Committee, Department of Agronomy and Horticulture	2016
Organizing Committee "Plant Phenomics: from pixels to traits"	2015

Professional

Associate Editor: Molecular Plant	2014-Present
Data Management Subcommittee, Maize Genetics Research Collaboration Network	2018-Present
MaizeGDB Advisory Committee	2018-Present

Grant Reviewer: NSF (panel & ad hoc), USDA (panel), JGI (panel), Genome British Columbia (ad hoc).

Peer Reviewer (selected, recent): Bioinformatics, BMC Plant Biology, G3: Genes | Genomes | Genetics, Genome Biology & Evolution, Heredity, Journal of Experimental Botany, Molecular Biology and Evolution, Molecular Plant, Nature Communications, Nature Plants, New Phytologist, PeerJ, Photosynthesis Research, Plant Cell, Plant Cell & Environment, The Plant Genome, The Plant Journal, Plant Methods, Plant Physiology, PLoS Genetics, Science

Invited Talks:

I have delivered a total of 50 invited talks or seminars, including 37 since I was hired as an assistant professor at the University of Nebraska-Lincoln. 29 excluding seminars and conferences affiliated with the University of Nebraska-Lincoln.

Talks in italics are scheduled for Fall '18 or Spring of '19 but have not yet been delivered and are not including in the count of delivered presentations.

External at Institutions

Invited presentations only. Excludes presentations selected based on abstracts or applications.
42 total and 29 during my time at UNL

University of Massachusetts-Amhert, MA, USA	*(Feb. 2019)*
Research Triangle, Raleigh, NC, USA	*(Oct. 2018)*

Plant Energy Biology Annual Forum, Perth, Australia *(Sept. 2018)*
Expenses covered by invitation.

Washington State University, Pullman, WA, USA *(Sept. 2018)*

42. The Plant Phenome Journal Webinar Series 2018
41. University of Delaware, Newark, DE, USA 2018
40. Entrepreneurship Panel, USDA FACT: Genomes to Fields, Ames, IA, USA 2018
39. Plant Phenotyping Session, Plant and Animal Genome, San Deigo, CA, USA 2018
38. Chinese Academy of Agricultural Sciences, Beijing, China 2017
37. Beijing Academy of Agricultural and Forestry Sciences, Beijing, China 2017
36. University of Minnesota, St. Paul, MN 2017
35. Plant Genome Evolution, Sitges, Spain 2017
34. Iowa State University, Ames, IA, USA 2017
33. Purdue Plant Science Symposium (Student Organized), West Lafayette, IN, USA 2017
32. P²IRC Annual Symposium, Saskatoon, Saskatchewan, Canada 2017
31. University of Missouri-Columbia, Columbia, MO, USA 2017
30. Maize Tools and Resources (Maize Genetics Conference pre-meeting), St. Louis, MO, USA 2017
29. Phenome, Tucson, AZ, USA 2017
28. Kansas State University, Manhattan, KS, USA 2016
27. University of Georgia-Athens, Athens, GA, USA 2016
26. University of California-San Diego, San Diego, CA, USA 2016
25. Corn Breeding Research Meeting, Jacksonville, FL, USA 2016
24. Chinese Academy of Agricultural Sciences, Beijing, China 2015
23. Beijing Academy of Agricultural and Forestry Sciences, Beijing, China 2015
22. Molecular Plant Symposium: From Model Species to Crops, Shanghai, China 2015
21. Sichuan Agricultural University, Chengdu, China 2015
20. Huazhong Agricultural University, Wuhan, China 2015
19. Shandong Agricultural University, Tai'an, China 2015
18. Monsanto, St. Louis, MO, USA 2015
17. Corn Breeding Research Meeting, St. Charles, IL, USA 2015
16. Life Technologies Session, Plant and Animal Genome, San Diego, CA, USA 2015
15. Maize Session, Plant and Animal Genome, San Diego, CA, USA 2015
14. Millet as Crop: Past and Future, Aohan, Inner Mongolia, China 2014

(May 1st, 2014. Start of my time at UNL)

13. Henan Agricultural University, Zhengzhou, China 2014
12. Chinese Academy of Tropical Agriculture, Haikou, China 2014
11. Cornell University, Ithaca, NY, USA 2014
10. Interdisciplinary Plant Group Seminar Series, University of Missouri, Columbia, MO, USA 2012
9. Donald Danforth Plant Science Center, St. Louis, MO, USA 2012
8. Plant Genomes in China Meeting, Tai'an, China 2012
7. China Agricultural University, Beijing, China 2012
6. Chinese Academy of Agricultural Sciences, Beijing, China 2012
5. American Society of Plant Biology, Austin, TX, USA 2012
4. MaizeGDB, Ames, IA, USA 2012
3. Polyploidy Session, Plant and Animal Genome, San Deigo, CA, USA 2012
2. CSSA Translational Genomics Session, Plant and Animal Genome, San Diego, CA, USA 2012
1. University of Arizona, Tucson, AZ, USA 2011

Internal

8 total and 8 during my time at UNL

8. UNL Plant Phenomics Symposium 2018
7. NeDA 2017: 2nd Nebraska Data Analytics Workshop, UNL 2017
6. Water for Food Global Conference, UNL 2017
5. Complex Biosystems Seminar Series, UNL 2017
4. Food Science Departmental Seminar Series, UNL 2016
3. Animal Science Departmental Seminar Series, UNL 2016
2. Agronomy & Horticulture Departmental Seminar Series, UNL 2015
1. Plant Science Retreat, UNL 2014

Candidate Statement

Program Overview

My research program at the University of Nebraska-Lincoln has had two – sometimes overlapping but logically distinct – themes. The first, which I was originally hired to pursue, is to employ comparative genomic approaches, including functional genomic data, to share information across maize, sorghum, and allied domesticated and wild species to develop new ways to link genes to functions and phenotypes. As an extension of this theme, research in my lab has also focused on the ability to distinguish between genes which are likely to play functional roles in determining the phenotype of a plant, and genomic sequences annotated as genes which are extremely unlikely to ever be identified as playing a consequential role in determining any plant trait. The second research theme within my group is one I was asked to take on by the UNL administration a little over a year after I started at the university: developing and deploying new approaches to utilize high throughput phenotyping in the context of quantitative genetics and crop improvement efforts. This effort currently focuses primarily on maize and sorghum and includes both controlled environment and field phenotyping efforts. Both programs have been productive bringing in grant dollars and each are produced both publications lead by my own research group with external collaborators as co-authors and publications by other research groups with members of my own research group as co-authors. A third, recently developed research focus involves the use to plant quantitative genetics to identify new specialized plant metabolites with the potential the perturb the human gut microbiome, and ultimately use this knowledge to develop new functional foods.

During my time at the University of Nebraska I have also been active in economic development. I have founded or co-founded three startups, two during my time at UNL. The second of the three, Dryland Genetics, has also invested over $100,000 funding work in the lab of one of my colleagues in the department (Dipak Santra). I serve on the scientific advisory council of a local genotyping company (GeneSeek) which is a notable employer of UNL graduates, and I have been a guest member of scientific advisory board meetings for Syngenta and Indigo Agriculture.

My appointment at the University of Nebraska-Lincoln coincided with the launch of the new Complex Biosystems graduate program. Complex Biosystems is one of the first, if not the first, interdepartmental graduate programs at UNL. I was involved in the original design of this requirements for this program and volunteered to teach Professional Development (LIFE 843) to first year graduate students entering the program in its first year (Fall 2015). I have continued to teach this course in each subsequent year, and in 2017 picked up Big Questions in Complex Biosystems (LIFE 841), a second required course for first year graduate students in Complex Biosystems, when the original instructor was hired away to another university. This year I am teaching Big Questions for the second time and Professional Development for the fourth time. However, I believe my greatest educations contributions have come in the context of involving undergraduates in the research process. In the past four years, 12 undergraduates and 2 high school students have had a chance to work on research projects in my lab, supported by a combination of Research Experience for Undergraduate (NSF), Undergraduate Creative Activities & Research Experiences (UNL), and startup or grant research funds (see details below).

Both my research focuses necessitate working in team-based environments with statisticians, engineers, computer scientists, and applied plant breeders. I have been extremely pleased with the opportunity this provides for both the undergraduate and graduate students working in my lab to experience working as part of an interdisciplinary team and to develop the necessary skills and vocabulary to communicate across disciplinary silos. This experience, more than any of the specific skills I can teach them – with the possible exception of learning how to apply the highly successful hypothesis-testing mindset of geneticists to genomics and big data research questions – is the way I can best prepare students in my research group to be successful in private industry, in academic research, or as entrepreneurs.

Research Contributions

Introduction

The grasses are one of the single most evolutionarily, ecologically, and economically successful clades of plants on the planet today. This group of species has reshaped whole ecosystems and adapted to grow and thrive everywhere from salt soaked tropical beaches to frozen tundras. Multiple grass clades have made the jump to using C_4 photosynthesis and have evolved multiple distinct enzymatic pathways to carry out the C_4 pathway. More than half of all calories consumed by humans around the world come, directly or indirectly, from only three highly productive domesticated grasses: maize, wheat, and rice. Yet, in addition to these three species, at least 27 other grass species have also been domesticated by humans as grain crops (see Glemin & Bataillon 2009 doi: 10.1111/j.1469-8137.2009.02884.x). Yet, at the genomic level, the genomes of grasses remain surprisingly similar with many of the same genes, conserved in the same order, present across all of these species. **Research in my group focuses on ways to generate genomic and phenotypic data from sets of related grass species and both developing and using comparative, functional, and quantitative genomic approaches to link genes to functions.** In many cases we are able to leverage the fact that most phenotypic and evolutionary transitions observed in the grasses have occurred multiple times independently in related lineages.

Since my hire, I have invested a substantial fraction of my total time and effort in building and sustaining research teams which cross disciplinary expertise. I believe that my success in that endeavor is reflected both in the breadth of disciplines represented by my co-PIs (Computer Science, Statistics, Engineering, Biochemistry, Food Science, Genetics, and Applied Plant Breeding), but more importantly in the fact that many of my co-PIs are found on multiple projects, including both those I have lead, and those I have contributed to as a team member. Putting teams together is, in some ways, a lot easier than managing and contributing to teams in such a way that the people continue to want to work together in the future. Similarly, you will see a number of names who both show up as middle authors on papers lead by my lab, and who show up as the lead or anchor authors of papers where I and my lab members are middle authors in turn.

The sections below describe some of the most productive areas of research in my lab over the past four years. Through these and other projects I have published a total of 38 peer reviewed papers since my appointment as an assistant professor at the University of Nebraska-Lincoln (56 total peer reviewed publications), including 28 resulting from work conducted here at the University. I have been fortunate enough to be quite successful at securing external funding, including two federal grants as PI (one from the USDA and one from NSF), three federal grants as a co-PI (USDA, ARPA-E, and NSF) (see page 44 for the cover page summaries of each funded federal grant) and a large number of nonfederal grants as both a PI and co-PI.

Research Highlights

Emphasis I: Comparative Genomics of Maize, Sorghum, and Allied Species

The comparative genomics work in my lab focuses primarily on the panicoid grasses, a group of species which includes maize, sorghum, and sugar cane, miscanthus, switchgrass, and many of the domesticated species collectively referred to as millets. As part of an effort to improve the resources for comparative genomics in this group of species, I have worked with JGI and HudsonAlpha to sequence the genomes of additional grass species within this clade, including *Dichanthelium oligosanthes* (published), *Paspalum vaginatum* (completed), and *Urochloa fusca* (initial draft). All the comparative genomics work performed in my lab depends on a set of high quality syntenic ortholog calls across different grass species which is generated and updated in house. We have distributed updated versions of these lists of syntenic orthologs across different grass genomes online using FigShare prior to their use in lab publications, and these datasets have been widely downloaded and employed in published research by other groups across the USA, EU, and China.

and using examples of parallel evolution to generate hypotheses about the functions of specific genomic sequences.

Emphasis IA: Separating Functional and Functionless or Function Mimicking Parts of the Genome

A combination of homology-based, mRNA-sequencing based, and *ab initio* prediction based approaches now make it straightforward to identify protein coding exons within any existing or newly sequenced plant genome. However, two significant challenges remain. The first is the identification of the regulatory sequences – which are predominantly noncoding – and control where, when, and in response to what stimuli those exonic sequences are transcribed into mRNA remain far more challenging to identify *in silico*. The second is that recent work, including from my own research group, is beginning to demonstrate that a nontrivial number of gene models present in plant genomes may not be true genes according to the classic requirement that a gene "contribute in any detected way to plant morphology, physiology or development" (Bennetzen et al. 2004 doi: 10.1016/j.pbi.2004.09.003). Comparing orthologous regions of the genome across related species can be used to address both of these questions, at least in part.

Identifying regulatory regions in noncoding DNA: One successful approach for identifying regulatory sequences has been to use comparisons between the noncoding sequence surrounding orthologous genes in related species to identify islands of conserved sequence evolving more slowly than the surrounding noncoding DNA. This slower evolutionary rate is seen as a marker for functional constraint, and regions identified as conserved noncoding sequences have been experimentally validated as regulators of the expression pattern of adjacent protein coding genes. However, previous approaches employed for identifying these functionally constrained functional regulatory regions within noncoding sequence were based on pairwise comparisons between related genes and as a result of multiple testing – analogous to the birthday problem – could not confidently identify sequences shorter than 15 base pairs long as exhibiting statistically significant functional constraint. This represented a major limitation of existing methods as many transcription factor binding sites can be as short as 4-8 base pairs long. Alignments of more than two sequences have the potential to reduce sequence matches resulting from coincidental matches rather than conserved homology, however conventional multiple sequence alignment algorithms cannot be effectively employed to compare plant promoters as the high rates of sequence divergence, insertion, and deletion mean that the majority of intergenic sequence associated with orthologous genes is truly non-homologous, and much of the remainder has diverged to the point where homology is no longer detectable through sequence similarity. Working with Sairam Behera, a computer science student in Jitender Deogun's lab I and two of my students – Xianjun Lai and Zhikai Liang – developed a new approach to identifying small conserved sequences in the noncoding sequence associated with orthologous genes in multiple species. Unlike previous approaches, this algorithm – STAG-CNS – is inherently scalable to incorporate data from orthologous genes in >2 species at once. Using data from genes in six different grass species, we demonstrated that unlike previous pairwise approaches, STAG-CNS could confidently identify sequences as short as 9 base pairs long as showing functional constraint, and that these sequences showed greater overlap with chromatin marks known to be associated with regulatory sequences – such as DNase1 hypersensitive sites – than conserved noncoding sequences identified through pairwise comparisons. As additional grass genome have become available since our original paper was published in 2017, STAG-CNS has continued to scale, providing greater and greater resolution in identifying functionally constrained regulatory sequences (Lai et al. 2017 doi: 10.1016/j.molp.2017.05.010).

Separating genes and gene mimics: Maize is rapidly becoming one of the best plant models to study what is, in fact, a gene. This is because of both its long history of forward genetic investigation which provides a significant set of verified "true positive" genes to use as ground truth, as well as growing evidence that different maize haplotypes vary significantly in their number of gene-like sequences, including the number of transcribed sequences with valid open reading frames which show homology to genes in other plant species. These sequences present a major challenge to efforts to annotated the genome, and, in fact, the first version of the maize genome included multiple sets of gene model annotations, based on differing degrees of supporting evidence which identified between 32,000 and 110,000 putative genes. The conservation of a gene at the same position in the genome between related species (synteny), rather than solely the conservation of gene sequence itself, appears to be a far better mark for function and contribution to plant morphology, physiology or development. Genes conserved at syntenic locations are 9x as likely to be identified as the causal loci responsible for phenotypic variation mapped

through forward genetics, and 4x as likely to be harbor SNPs identified in GWAS studies (as reviewed Schnable 2015 doi: 10.1146/annurev-arplant-043014-115604). Since joining UNL I have worked to uncover the reason for this dramatic difference in the functional relevance of nonsyntenic and syntenically conserved genes. Working with a group in Germany interested in allele specific expression in hybrids, we demonstrated that nonsyntenic genes show much less conservation of expression between alleles in maize hybrids (Paschold et al, published in The Plant Cell in 2014). Working with researchers at Iowa State and Kansas State, we have demonstrated at, at least in maize, the long known correlation between gene density and recombination frequency across the genome is entirely explained by syntenic gene density, and that nonsyntenic gene density shows no correlation with recombination rate (Liu et al 2018, Molecular Biology and Evolution). Working with scientists at the University of California-San Diego, we demonstrated that not only on non-syntenic genes less likely to be transcribed into mRNA, but when mRNAs are detected from nonsyntenic genes these mRNAs are less likely to be translated into proteins, suggesting potential mechanistic explanation for why nonsyntenic genes are much less likely to play a role in determining plant traits (Walley et al. published in Science in 2016).

Emphasis IB: Using parallel evolution to identify genes involved in complex changes in phenotype

As mentioned, the grasses have been successful enough as a clade that many complex changes have happened multiple times in parallel in different lineages. These cases of parallel evolution provide an opportunity to address several questions. Firstly, do parallel phenotypic changes in related lineages result from changes in the function of the orthologous genes in each lineage? Secondly, if the answer to the first question is yes, can we use cases of parallel evolution to identify the specific genetic loci involved in complex changes in plant traits?

Domestication Syndrome: In order to test the first question, my lab first worked with parallel artificial selection for "domestication syndrome" traits in maize and sorghum. Domestication of different grain crops from different wild grasses involved conscious or unconscious selection for an array of traits including a loss of seed dormancy and shattering, decreasing tillering, increases in seed size, decreases in shade avoidance responses, etc. Reanalyzing published data from separate studies on maize and sorghum and their respective wild relatives using a common pipeline members of my lab discovered that there was no statistically significant overlap in the relatively large sets of genes which show population genetic signatures of selection between wild and domesticated accessions. However, at the same time, we found that genes with specific and known phenotypic roles in producing domestication syndrome traits in one species were disproportionately likely to show population genetic signatures of selection in the other (Lai et al. published in The Plant Journal in 2018).

Low temperature tolerance and temperate latitude adaptation: The first federally funded project in my research group – a USDA NIFA AFRI grant with Rebecca Roston, a collaborator from the Biochemistry department with a background in cold and freezing stress biology as a co-PI – sought to employ the parallel adaption of different groups of panicoid grass species to temperate climates to identify genes and pathways involved in conferring low temperature tolerance. Both maize and sorghum were domesticated from wild species native to tropical latitudes and are extremely sensitive to cold and freezing temperatures. As a first step in this project, we set out to develop robust methods for comparing patterns of transcriptional responses to stimuli across related species using time course gene expression data collected from maize and sorghum exposed to cold stress and grown in a common experiment. Working with a collaborator from the statistics department – Yumou Qiu – we validated an approach to specifically compare the pattern of transcriptional response to stimulus across orthologous genes in related species which may exhibit different baseline levels of expression under control conditions. In the specific maize-sorghum comparison, we demonstrated that genes with conserved patterns of expression in response to cold stress across the two species experienced stronger purifying selection and were enriched in genes with plausible mechanistic links to cold acclimation/tolerance while genes with dissimilar patterns of response to cold across the two species were more similar to a random sample of expressed genes (Zhang et al. Published in The Plant Cell in 2017).

This finding lead to proposing a model that gene regulation, like noncoding sequence as a

whole, is a somewhat fast evolving trait, with many transcriptional responses being selectively neutral or nearly neutral. This model is also consistent with allele specific expression studies in maize which indicated that transcriptional responses to cold stress are frequently not conserved between different alleles of the same gene. If ultimately proven to be correct, one conclusion is that parallel whole genome transcriptional studies in several related species can provide a way to separate the signal of functionally constrained transcriptional responses from the noise of rapidly evolving but largely selectively neutral responses which are currently confounded in many studies. A review paper which summarized some of the potential implications of this model, and performed a meta analysis of published cold stress experiments across multiple species was written and published with our collaborators in the Roston Lab and came out earlier this year (Raju et al 2018, published in Plant Science).

Leveraging support from our $20M NSF EPSCoR grant (The Center for Root and Rhizobiome Innovation), members of my lab have also generated a library of full length cDNA sequences from *Tripsacum dactyloides* using PacBio IsoSeq sequencing, a member of a genus sister to Zea which is indigenous to Nebraska and much of the lower 48 United States east of the rocky mountains. Using an analytical method which identified genes with elevated ratios of non-synonymous substitutions to synonymous substitutions in either *T. dactyloides* or maize relative to the ratios observed for orthologs of that specific gene in other related grasses. Using maize as a control, we found that genes with elevated rates of protein sequence evolution specifically in *T. dactyloides* were clustered in a lipid biosynthesis pathway known to be involved in conferring freezing tolerance in eudicts, and identified a statistically significant overlap between genes showing elevated rates of protein sequence evolution in *T. dactyloides* and genes showing signatures of artificial selection between tropical and temperate latitude adapted maize accessions (Yan et al, Preprint).

Emphasis II: Phenomics for Breeding and Quantitative Genetics of Maize, Sorghum, and Allied Species

My interest in high throughput phenotyping of plants dates back to my postdoc at the Danforth Center where I used cameras connected to a raspberry pi computer to measure leaf rolling and leaf dropping of different maize genotypes in response to drought stress. I did not originally anticipate continuing to work on plant phenotyping after my hire at UNL. However about a year into my time at the University I was asked by several administrators to resume this program in addition to my multi-species genomics work. Ultimately I envision being able to connect these two areas of emphasis through the development of phenotyping methodologies which can be applied across maize, sorghum, foxtail millet and pearl millet for effective identification of homologous traits and cross species quantitative genomics. In the interregnum, this research emphasis has a strong component of service to my institution, as well as benefiting student training by ensuring members of my lab regularly come into contact with and collaborate with faculty and students from statistics, computer science, and engineering, disciplines that plant biology students often lack sufficient exposure to.

Controlled Environment Phenotyping: Employing high throughput phenotyping methodologies in the context of a breeding program or quantitative genetic research requires two very different types of data analysis. The first is simply to convert raw sensor or image data into some sort of biologically meaningful numerical measurement. The second is to take those sets of numerical data and convert them into biological insight. The latter problem can be addressed with existing tools for genomic prediction and QTL mapping/GWAS, although potentially intriguing alternative methods may also become practical as high throughput phenotyping datasets tends to have much higher dimensionality than conventional phenotyping datasets. However, addressing the first problem absolutely requires the development and deployment of new algorithms, and early work in the lab focused on collaborating with computer scientists, statisticians, and engineers simply to develop algorithms to extract meaningful measurements of plant traits from RGB and hyperspectral images. This work, supported by startup funds, internal grants, and an USDA/NSF joint EAGER award with my colleague Yufeng Ge as lead PI, resulted in a number of publications, including Ge et al 2016 doi: 10.1016/j.compag.2016.07.028 (published in Computers and Science in Agriculture and already cited 35 times), Pandey et al 2017 doi: 10.3389/fpls.2017.01348 (published in Frontiers in Plant Science), and Liang et al 2018 doi: 10.1093/gigascience/gix117 (published in Gigascience). One key finding of Liang et al 2018 was that the amount error in measurements of biological

traits (such as biomass) using image data was, itself, subject to genetic control. After presenting this work at the AGU conference in December of 2017, A colleague has contacted me to tell me that, after seeing this work presented at the AGU conference in December, they have begun running separate GWAS for error between manual and high throughput measurements of the same traits and are able to identify specific genetic loci involved in controlling error. My own lab is just starting to do similar work, using multiple replicates of the the Sorghum Association Panel which were grown and phenotyped through reproductive maturity on the greenhouse system (completed grant #14) and multiple replicates of the Buckler Goodman 292 maize association panel which are currently being grown and phenotyped on the system as part of the NSF CRRI project (active grant #4). Going forward I anticipate we have hit an inflection point where papers published by the lab in this area will start to shift back to addressing biological questions, rather than focusing primarily on methods development and validation.

Field Phenotyping: Field phenotyping efforts in the lab have been enabled by my participation in Genomes to Fields Initiative (https://www.genomes2fields.org/). This participation has in turn been supported by research funding from Nebraska Corn Growers – one of the key stakeholders of my own research program – who have now supported my G2F and field phenotyping work for three years running, each year evaluating a new research proposal as part of their competitive funding process. The overall goal of Genomes to Fields is to enable plant science (particularly in quantitative genetics and high throughput phenotyping) on a geographic scale beyond what any individual academic research group (or small company) could manage individually. This in turn should create datasets that make it possible to train students with the big data skillsets that big ag companies are having trouble hiring, catalyze the development and validation of new plant phenotyping technologies, and provide a platform for startup to medium sized ag companies to conduct geographically broad experiments on a scale that would only otherwise be feasible for big ag companies. In addition to these initiative-wide goals, I have found that simply having plots in the field, combined with the commitment to manually collect ground truth data from our plots, and access to data on the performance of the same lines at dozens of sites in states across the corn belt and beyond is excellent collaborator bait, particularly for statisticians and engineers interested in new ways to analyze and collect plant phenotypic data respectively. My participation in Genomes to Fields has resulted in three publications thus far: Gage et al 2017 doi: 10.1038/s41467-017-01450-2 (published in Nature Communications), Liang et al 2018 (also references in the previous section), and Alkhalifah et al 2018 doi: 10.1186/s13104-018-3508-1 (published in BMC Research Notes).

Emphasis III: The Nebraska Food for Health Center

Several years ago I was recruited to join a group of six faculty members who created the original concept for the Nebraska Food for Health Center (NFHC). This team pitched the concept of a center which could unite plant quantitative genetics with studies of the human microbiome to identify new dietary molecules which can perturb the gut microbiome to the Bill and Melinda Gates Foundation and the Raikes Foundation, resulting in a $5M charitable give to the University of Nebraska from these two foundations to establish the center, and additional fundraising by the University of Nebraska Foundation with a target of $40M in total investment over coming years. The Nebraska Food for Health center seeks to develop new approaches to perturb the human gut microbiome – both as a tool for basic research and to improve human health – through the application of plant quantitative genetics. Essentially we will identify sets of specialized plant metabolites present in food with microbially active properties by conducting GWAS to identify genetic loci in plants which are associated with changes in the population structure and composition of human gut microbiomes feed grain derived from specific plant accessions. Because these microbially active compounds are, by definition, already produced by existing food crops in varying concentrates, conventional breeding work could be used to develop new varieties of functional foods enriched in compounds with beneficial effects or depleted in compounds with detrimental effects, potentially providing new valued-added crop variety options to Nebraska farms – key stakeholders of my research program and the Department of Agronomy and Horticulture as a whole. Work in this area commenced in earnest only in 2017 with the recruitment of the lab's most recent PhD student, supported by the NFHC graduate fellowship program and co-mentored with Andrew Benson in Food Science and has, as of yet, not resulted in any peer reviewed publications.

Mentoring and Teaching Contributions

My goal as a teacher and mentor is to train students who are equally comfortable in the field, at the lab bench, or working at the command line at both the undergraduate and graduate level. Within my lab, I design each graduate student's project to require significant direct interaction with at least one faculty member in statistics, computer science, or engineering, as well as at least one UNL faculty member (besides me) in either plant breeding, genetics, or biochemistry. I have also been working to make sure as many of the graduate and undergraduate students in my lab have interactions with private sector companies prior to graduation. In recent years this has included presenting to groups of visiting scientists from Pioneer Hi-Bred and directly interacting with employees at Indigo Agriculture as part of a collaborative project.

Mentoring: I am currently advising three PhD students and one masters student from the Agronomy and Horticulture department. In addition I serve as the co-advisor for one PhD student in food science and one visiting PhD student from Shandong Agriculture University. During my time at the University of Nebraska-Lincoln I have also served as the co-advisor for a second visiting PhD student from Sichuan Agriculture University, and two masters students from the Department of Computer Science and Engineering (see page 4. Students in the lab are making excellent progress towards graduation. My first PhD student has published six papers during his three years in the lab (four as first author), the second has published three (one as first author) in three years, and the third has published one (as first author) in two years, with a second first author paper in review. Xianjun Lai, a CSC supported graduate student published three first author papers – plus one perspective/review piece – in the two years he was part of the lab, allowing him to be hired as an Associate Professor when he returned to China.

I also currently mentor two postdoctoral scholars, and have previously mentored three postdocs and one visiting scholar during my time at UNL. Of those four lab alumni, one is now an assistant professor, a second is a staff scientist, and a third the deputy director of a functional genomics center in China, while the fourth did a short term postdoc in my lab while waiting for his spouse to graduate and has now transferred to a new institution for his primary postdoc project. For a complete list of nonundergraduate lab alumni and their present positions please see the mentoring and teaching appendix (page 26).

Teaching: I have placed a high priority on involving undergraduates in the academic research process since the start of my time at UNL. Over the past four years my lab has hosted 13 undergraduate researchers and 2 high school interns. Four students were supported through the Research Experience for Undergraduates summer program, two as part of an internal UNL program called UCARE (Undergraduate Creative Activities and Research Experience), one high school intern through the Young Nebraska Scientist program, and the remaining eight were supported by startup funds, and when those were exhausted, from research grant funding (see page 4). In recent years undergraduate and high school student research projects in my research group include a statistical reconstruction of ancestral character states for grass phenotypes, identifying computer vision phenotypes which correlate with field performance in ex-PVP maize lines and training neural networks to count corn leaves using computer generated corn plants. The preceding examples were all projects conducted by biology or agronomy students, not students with backgrounds in Computer Science or Statistics. For a list of poster presentations including those by undergraduate researchers see page 26. The first undergraduate in my lab – Daniel Ngu – has already been a co-author on two published papers, and several more recent undergraduates are co-authors on projects currently being written up for submissions. A sample of written feedback from former mentees within the lab is provided as part of Appendix A (see page 28).

I joined UNL as a new Complex Biosystems graduate major was being developed and deployed. Complex Biosystems is one of the first, if not the first, interdepartmental graduate program in biology at the University of Nebraska. After being involved in the initial curriculum and program design, I began teaching one of the required first year graduate courses for the new program (LIFE 843), which I am currently teaching for the fourth time this fall. This course incorporates training on both oral and written scientific communication, scientific misconduct, and professional development. The most successful aspect of the course has been a module in which students draft a research statement following the guidelines for the National Science Foundation Graduate Research Fellowship Program. Two years ago I updated this module so that, prior to the drafting their own research statements, students are divided into groups and conduct a mock peer review and stack rank of research statements from successful applicants

in prior years, which drew both positive feedback from students and a notable improvement in the quality of the final research statements turned in by students. In addition, after another of the original founding faculty for the Complex Biosystems program left the University, I have taken over coordinating the Fall semester of Life 841, which is another required first year course for students in the Complex Biosystems program. Life 841 focuses on the big open questions across a number of fields including quantitative genetics, plant biology, and microbiome host interactions. Example syllabi for LIFE 843 and LIFE 841 (formerly LIFE 891) are provided in Appendix A (see page 36) I am currently involved in an effort lead by George Graef to revise and revitalize the Plant Breeding curriculum within the Department of Agronomy and Horticulture and hopefully to revive the teaching of Plant Breeding at the undergraduate level. In addition to formal teaching I have been active in conducting outreach to both early learners (Sunday with a Scientist) and high school students (Fascination with Plants Day), as well as to the broader scientific and research community (both through twitter and podcast interviews). Details of outreach activities are provided as part of Appendix A (see page 34).

Service Contributions

Scholarly Service
I serve as a member of the advisory committee for MaizeGDB, the USDA funded genetics and genomics database for maize research. For folks more familiar with Arabidopsis, MaizeGDB is the equivalent of TAIR. In addition, I was recently recruited to become a member of the Data Management Subcommittee of the NSF-funded Maize Genetics Research Collaboration Network. I am also conscious of the fact that high publication frequency brings with it a responsibility to be active in the peer review process. I have reviewed an average of 2-3 manuscripts per month during my time as an assistant professor for journals ranging from Science to G3. Finally, I have served as an associate editor for Molecular Plant since 2014.
University and Department Service
During my time as the University of Nebraska I have served on four search committees and the organizing committees for two conferences based at the University of Nebraska (Predictive Crop Design: Genome to Phenome, Plant Phenomics: From Pixels to Traits) and one based at a neighboring school (International Millet Symposium at Colorado State University). I have also served on a number of internal committees with varying degrees of both time commitment required and exposure to political consequences and fallout. The most controversial committee I serve on is the UNL Faculty Greenhouse Committee. On one notable occasion, my faculty mentor within the department fired me as a mentee because she was unhappy that I had brought a concern of hers to the greenhouse committee, but was unable to sway the committee as a whole to adopt the remedy she had requested I seek.

I was elected by my peers within the Department of Agronomy and Horticulture to serve a two year on the Faculty Advisory Committee in 2017. Over this period the department has had two chairs, and a third is expected to assume the office before the end of my current term on the committee so, in addition to providing a service back to the department, this role as served as a fascinating learning experience into differences in management style and the challenges facing any faculty member who finds themselves in charge of a department including 67 faculty members with home bases spread across four hundred miles and two time zones. I also serve as a member of the faculty advisory committee for the Raikes and Gates Foundation funded Nebraska Food for Health Center – the only assistant professor to be asked to serve in this role. Within the department I also service on the Peer Evaluation Committee which reviews the annual progress reports submitted by faculty members and makes recommendations and provides text to the department chair for use in the annual evaluation process.

Appendix A: Supporting Evidence for Mentoring Activity and Outcomes

Present Employment of Lab Alumni

Name	Schnable Lab Position	Tenure	Current Position
Yang Zhang	Postdoctoral Scholar	2014-2017	Research Scientist, St. Jude Children's Research Hospital
Jinliang Yang	Postdoctoral Scholar	2016-2017	Assistant Professor, University of Nebraska-Lincoln
Sunil Kumar	Postdoctoral Scholar	2017-2018	Postdoc, Niederhuth Lab, Michigan State University
Lang Yan	Visiting Scientist	2016-2017	Deputy Director, Potato Functional Genomics, Xi-Chang College
Xianjun Lai	CSC PhD Student	2015-2017	Associate Professor, XiChang College
Bhushit Agarwal	Masters Student	2015-2016	Software Engineer, Mode.ai
Srinidhi Bashyam	Masters Student	2015-2016	Systems Software Developer, University of Nebraska-Lincoln

Partial List of Poster Presentations With Undergraduate Authors Highlighted

Lab Members in **bold**. Undergraduates from the Schnable lab in **red**

Carvalho DS, **Liang Z**, **Butera C**, Stoerger V, **Schnable JC**. (2018) High-throughput imaging and phenotyping of panicoid grain crops. 3rd International Millet Symposium. Fort Collins, Colorado.

Butera C, **Carvalho DS**, **Liang Z**, **Miao C**, **Sun G**, **Schnable JC**. (2018) Automated phenotyping of maize and pearl millet growth patterns and drought stress responses. Summer Research Fair - University of Nebraska-Lincoln. Lincoln, Nebraska.

Pages AD, **Miao C**, Clarke J, **Schnable JC**. (2018) Automated trait extraction from images of Sorghum. Summer Research Fair - University of Nebraska-Lincoln. Lincoln, Nebraska.

Foltz A, **Sun G**, **Schnable JC**. (2018) Differences in the responses to nutrient stress of the root systems of maize and its domesticated and wild relatives. Summer Research Fair - University of Nebraska-Lincoln. Lincoln, Nebraska.

Miao C, Pandey P, **Liang Z**, ... **Schnable JC**. (2018) Analysis of sorghum time-series phenotype data using nonparametric curve fitting and machine learning. Phenome 2018. Tucson, Arizona.

Pedersen C, **Schnable JC, Liang Z**. (2018) Analyzing phenotypic correlations in large scale studies combining field and greenhouse datasets. Nebraska Plant Breeding Symposium. Lincoln, Nebraska
Connor Pedersen was awarded the first prize in the undergraduate poster competition at the Nebraska Plant Breeding Symposium.

Hoban T, **Schnable JC**, **Miao C**, Xu Z, **Liang Z**. (2018) Using machine learning to count leaves in maize. Nebraska Plant Breeding Symposium. Lincoln, Nebraska

Liang Z, Pandey P, Stoerger V, Xu Y, Qiu Y, Ge Y, **Schnable JC**. (2018) High-throughput imaging of maize lines from public and private sectors employed in field trials. Supercomputing and Life Sciences Symposium. Lincoln, Nebraska.

Podliska H, **Schnable JC**, **Carvalho DS**. (2018) Cold tolerance in PACMAD grasses. Spring Research Fair - University of Nebraska-Lincoln. Lincoln, Nebraska.

Miao C, Yang J, Schnable JC. (2018) Large-scale simulation studies enabled by HPC reveal the powers of GWAS approaches in dissecting highly polygenic traits in crop species. Supercomputing and Life Sciences Symposium. Lincoln, Nebraska.

Miao C, Pandey P, **Liang Z, Carvalho DS**, Ye X, Stoerger V, Xu Y, Ge Y, **Schnable JC**. (2018) Analysis of sorghum time-series phenotype data using functional ANOVA and machine learning. Phenome 2018. Tuscon, Arizona.

Liang Z, Bai G, Ge Y, Rodriguez O, **Schnable JC**. (2018) Field phenotype prediction on maize using novel phenomic tools and environmental information. 2018 NIFA FACT G2F Workshop. Ames, Iowa.

Schnable JC, Pandey P, Ge Y, Xu Y, Qiu Y, **Liang Z**. (2017) Lessons From Paired Data From exPVP Maize Lines in Agronomic Field Trials and RGB And Hyperspectral Time-Series Imaging In Controlled Environments. AGU 2017 Fall Meeting. New Orleans, Louisiana.

Shi Y, Veeranampalayam-Sivakumar AN, Li J, Ge Y, **Schnable JC**, Rodriguez O, **Liang Z, Miao C**. (2017) Breeding for Increased Water Use Efficiency in Corn (Maize) Using a Low-altitude Unmanned Aircraft System. AGU Fall Meeting. New Orleans, Louisiana.

Hoban T, **Liang Z**, **Schnable JC**. (2017) Identifying sorghum root hair mutants as a first step in comparative genetic analysis of maize and sorghum. Spring Research Fair - University of Nebraska-Lincoln. Lincoln, Nebraska.

Carvalho DS, Zhang Y, Schnable JC. (2017) Identifying common and unique enzymatic changesassociated with three C4 biochemical pathways in related grasses. Predictive Crop Design: Genome-to-Phenome. Lincoln, Nebraska.

Zhang Y, **Ngu DW**, **Carvalho DS, Liang Z**, Qiu Y, Roston RL, **Schnable JC**. (2017) Statistical approaches to identifying differentially regulated orthologs (DROs) across related grass species. 59th Maize Genetics Conference. St. Louis, Missouri.

Miao C, Yang J, Schnable JC. (2017) Comparative GWAS in *Sorghum bicolor* and *Setaria italica*. 59th Maize Genetics Conference. St. Louis, Missouri.

Yan L, Lai X, Rodriguez O, **Schnable JC**. (2017) Developing transcriptomic resources for Tripsacum to study the adaptation of a maize relative to temperate climates. 59th Maize Genetics Conference. St. Louis, Missouri.

Lai X, Yan L, Lu Y, **Schnable JC**. (2017) Searching for parallel signatures of selection during domestication in maize and sorghum. 59th Maize Genetics Conference. St. Louis, Missouri.

Liang Z, Bai G, Ge Y, Rodriguez O, **Schnable JC**. (2017) Field phenotype prediction on maize using novel phenomic tools and environmental information. 59th Maize Genetics Conference. St. Louis, Missouri.

Liang Z, Pandey P, Stoerger V, Xu Y, Qiu Y, Ge Y, **Schnable JC**. (2017) Conventional and hyperspectral time-series image data sets of maize inbred lines widely used in North American field trials. Nebraska EPSCoR RII Track 1 Grant External Review Panel Visit. Lincoln, Nebraska.

Liang Z, Bashyam S, Agarwal B, Samal A, Bai G, Chaudhury SD, Rodriguez O, Qiu Y, Ge Y, **Schnable JC**. (2016) Maize Phenomap1 and Phenomap2 datasets: Integration with genomes to fields. 4th International Plant Phenotyping Symposium. Texcoco, Mexico.

Liang Z, Bashyam S, Samal A, Choudhury SD, Geng B, Ge Y, Rodriguez O, **Schnable JC**. (2016) Computer vision based phenotyping of panicoid crops. 2016 Purdue Plant Science Symposium. West Lafayette, Indiana.

Johnson K, **Liang Z**, **Schnable JC**. (2016) Maize association studies with high throughput image based phenotyping. Summer Research Fair - University of Nebraska-Lincoln. Lincoln, Nebraska.

Horn T, **Zhang Y**, **Schnable JC**. (2016) Cold sensitivity and genetic regulatory responses in panicoid grasses. Summer Research Fair - University of Nebraska-Lincoln. Lincoln, Nebraska.

Liang Z, Schnable JC. (2016) B73 maize population structure analysis by RNA-seq data. 2016 UNL Plant Breeding and Genetics Symposium. Lincoln, Nebraska.

Zhang Y, **Ngu DW**, Mahboub S, Qiu Y, Roston RL, **Schnable JC**. (2016) Conservation and divergence of synthetic gene regulation in response to stress in maize and relatives. 58th Maize Genetics Conference. Jacksonville, Florida

Ngu DW, **Zhang Y, Schnable JC.** (2016). Updates to qTeller: A tool for visualizing published gene expression data. 58th Maize Genetics Conference. Jacksonville, Florida

Carvalho DS, Zhang Y, Schnable JC. (2016) Comparative analysis of C4 photosynthesis genes in two independent origins of C4 in grasses. 58th Maize Genetics Conference. Jacksonville, Florida.

Lai XJ, Bendix C, **Zhang Y**, **Ngu DW**, Lu YL, Harmon FG, **Schnable JC**. (2016) Conserved and lineage-specific alternative splicing of orthologous genes in maize, sorghum, and setaria. 58th Maize Genetics Conference. Jacksonville, Florida.

Carvalho DS, Zhang Y, Schnable JC. (2015) Comparative transcriptomic analysis of *Danthoniopsis dinteri*, a novel C4 grass species. Plant Phenomics Symposium. Lincoln, Nebraska.

Zhang Y, **Ngu DW**, Roston RL, **Schnable JC**. (2015) Core cold responsive genes in the panicoid grasses. Plant Science Symposium "Plant Phenomics: from pixels to traits". Lincoln, Nebraska

Letters from Former Mentees:

Figure 1: Ashley Foltz was an REU student from the University of Wyoming who worked on comparisons of how the root morphology and gene expression of eight different grasses (maize, sorghum, setaria and wild relatives) responded to different nutrient deficits as part of the CRRI EPSCoR project.

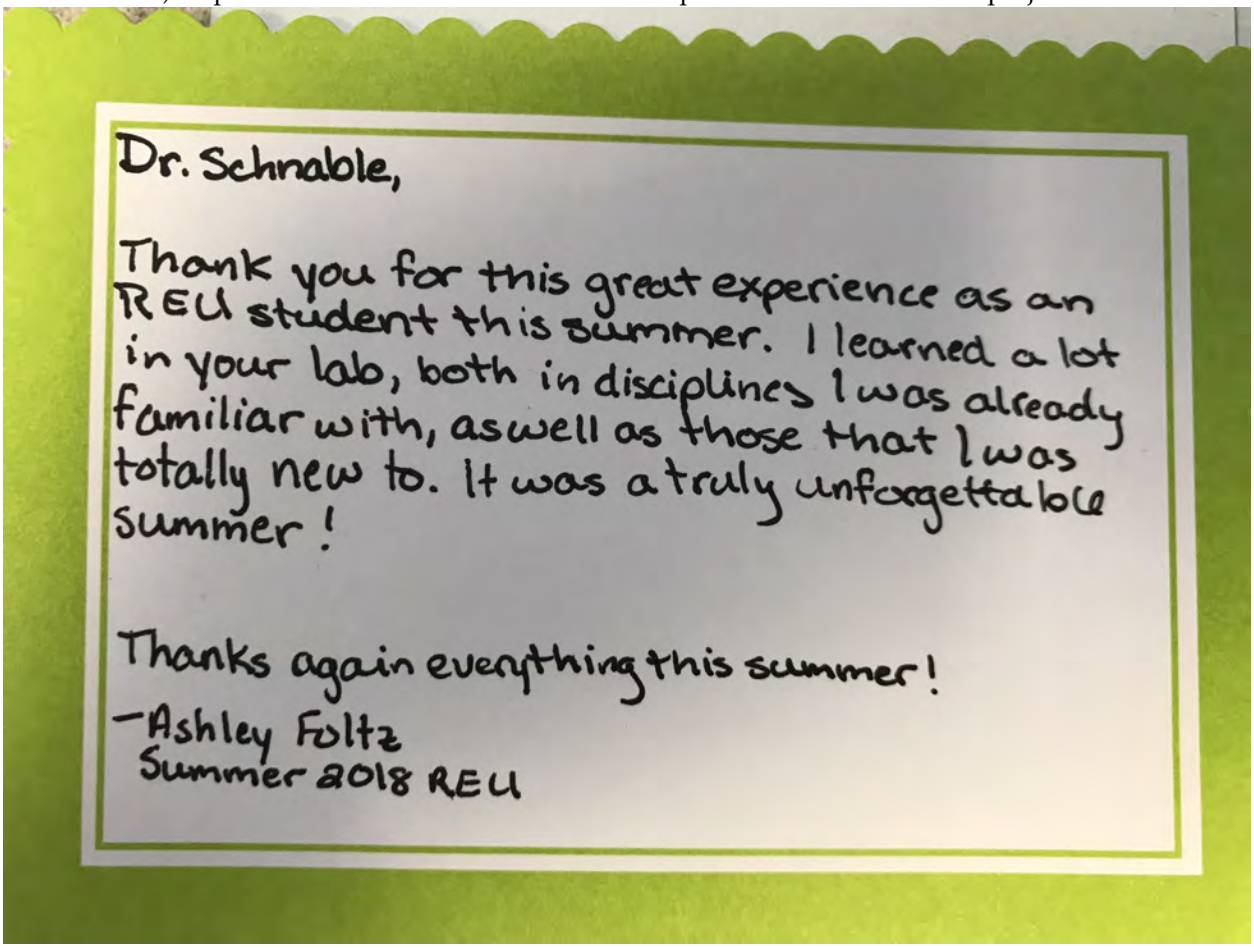


Figure 2: Taylor Horn was an REU student from Baylor University who worked on a project linking variation in the cold stress tolerance of wild grasses to their native ranges, conducting phenotypic screening, and mining data from the Global Biodiversity Information Facility (GBIF) for a GIS based analysis that she learned to conduct herself using python.

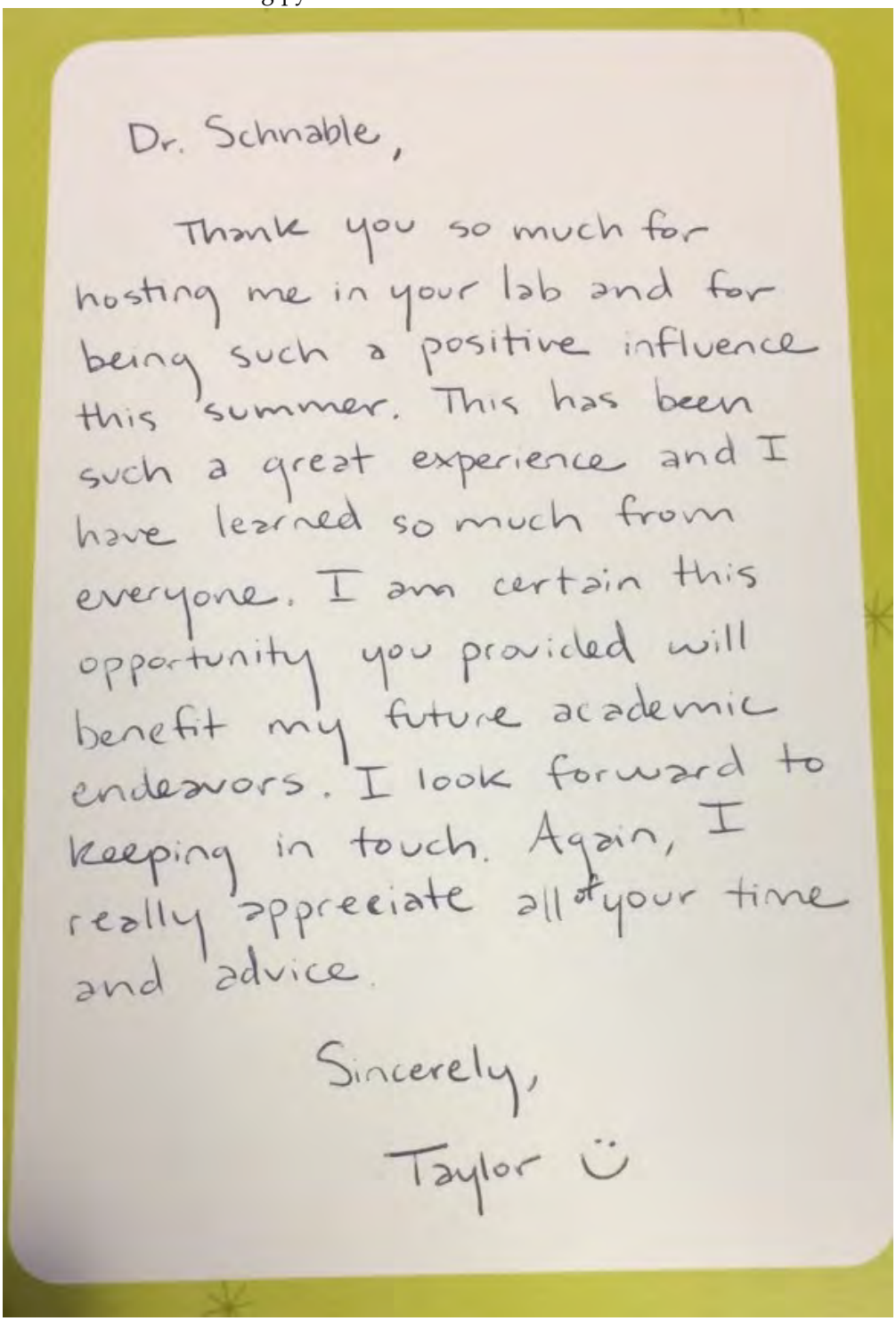


Figure 3: Kyle Johnson was an REU student from BYU who worked on kernel phenotyping project, imaging kernels from the maize Buckler/Goodman 282 association panel and conducting GWAS for traits measured from those images using computer vision techniques related to size and shape.

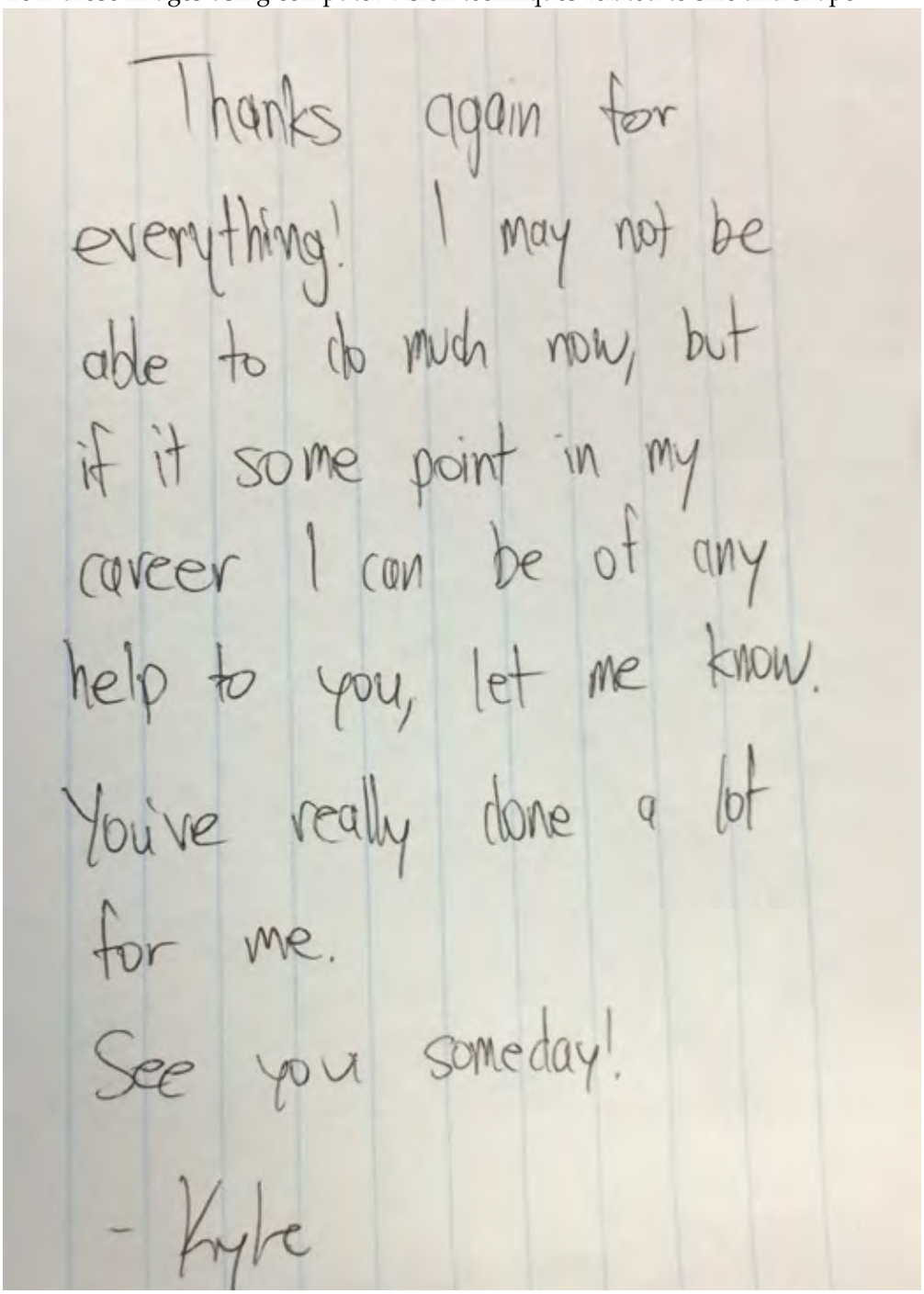
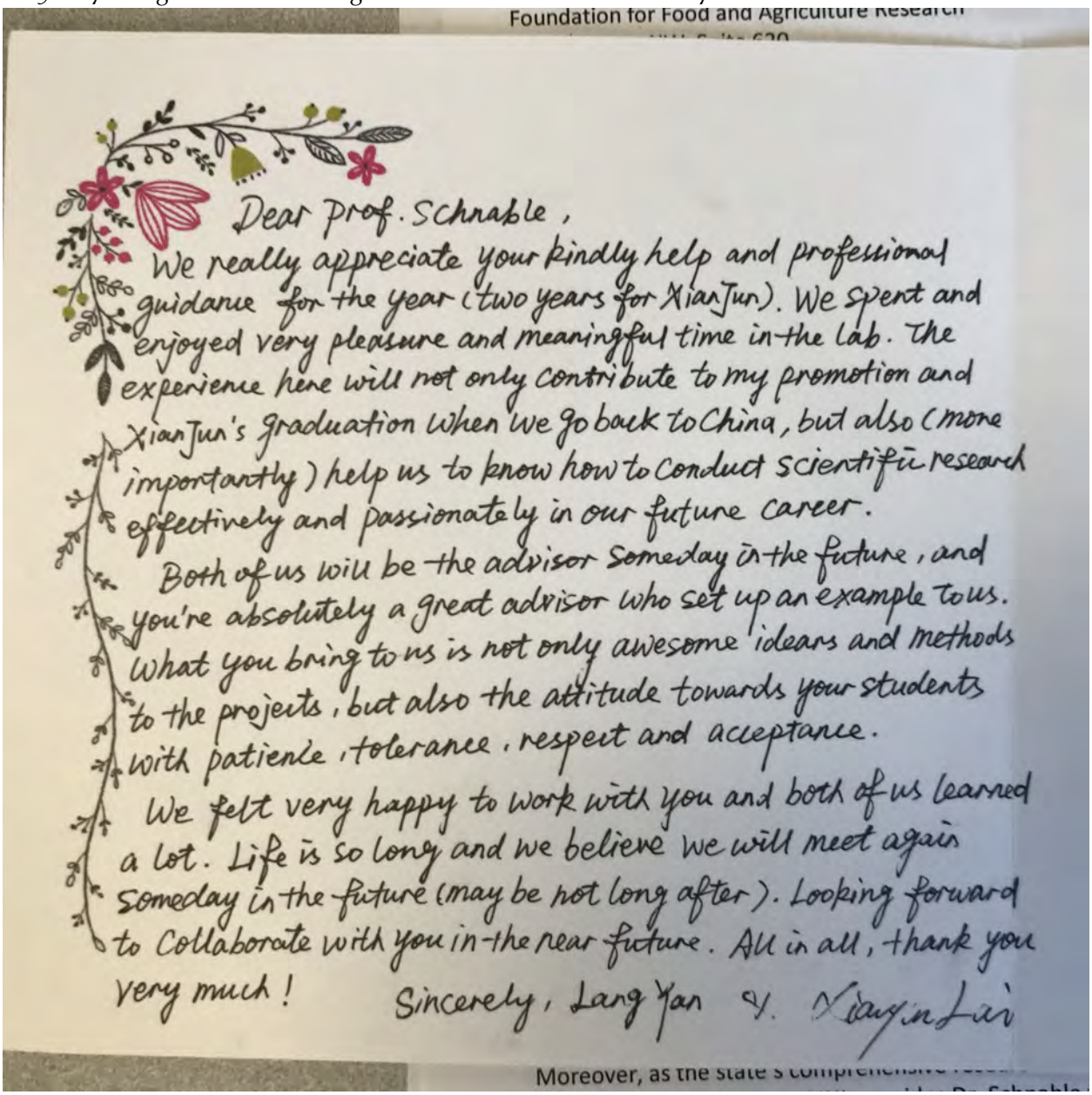


Figure 4: Xianjun Lai was a "sandwich" PhD student from Sichuan Agricultural University who was supported for two years of his graduate research in my lab by the Chinese Scholarship Council from 2015-2017. Lang Yan was a visiting scholar in the lab from 2016-2017.



Foundation for Food and Agriculture Research

Dear Prof. Schnable,

We really appreciate your kindly help and professional guidance for the year (two years for XianJun). We spent and enjoyed very pleasure and meaningful time in the lab. The experience here will not only contribute to my promotion and XianJun's graduation when we go back to China, but also (more importantly) help us to know how to conduct scientific research effectively and passionately in our future career.

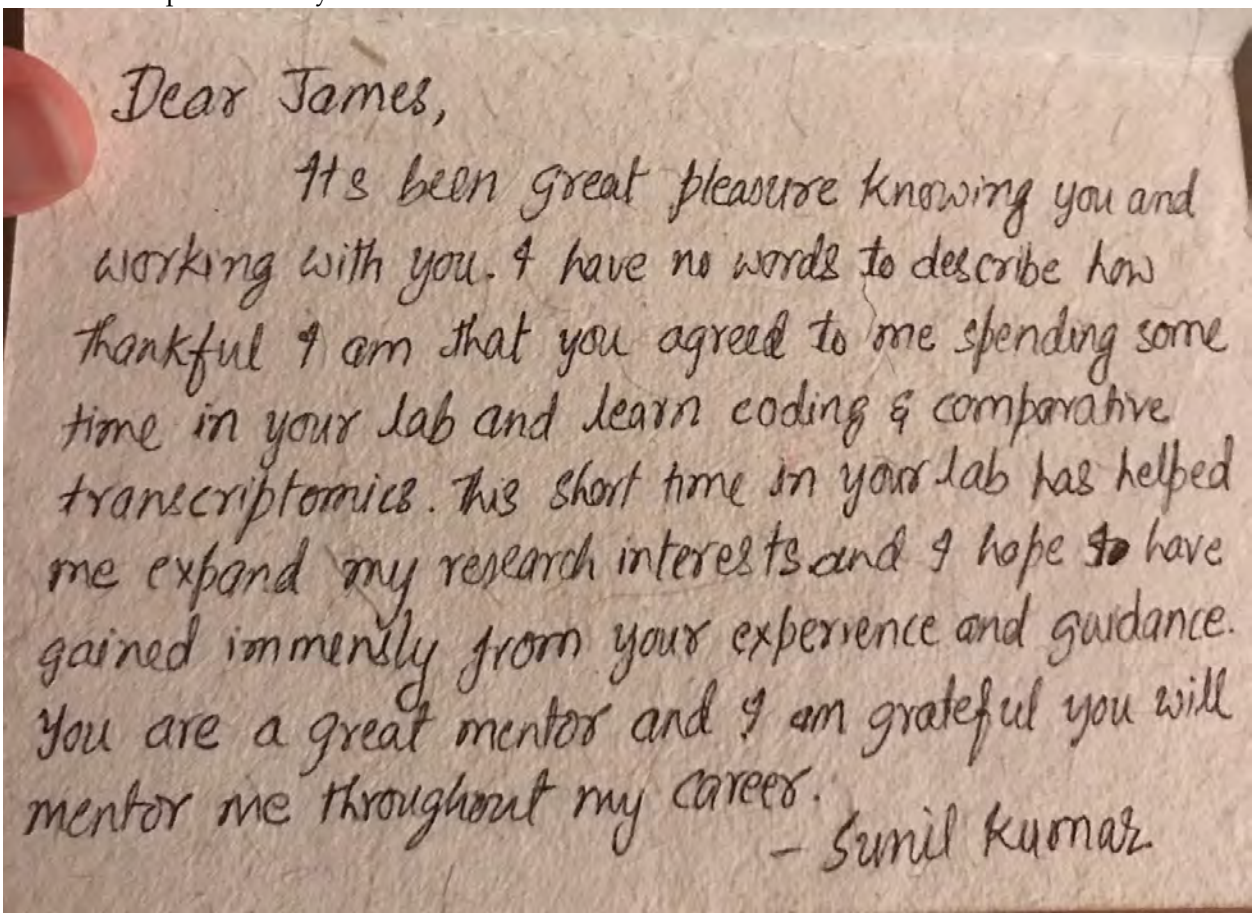
Both of us will be the advisor someday in the future, and you're absolutely a great advisor who set up an example to us. What you bring to us is not only awesome ideas and methods to the projects, but also the attitude towards your students with patience, tolerance, respect and acceptance.

We felt very happy to work with you and both of us learned a lot. Life is so long and we believe we will meet again someday in the future (may be not long after). Looking forward to collaborate with you in the near future. All in all, thank you very much!

Sincerely, Lang Yan & Xianjun Lai

Moreover, as the state's comprehensive

Figure 5: Sunil Kumar came to the lab as a molecular biologist, having just finished a PhD at UNL, and spent a semester as a postdoc learning both bioinformatic and comparative genomic techniques prior to starting his primary postdoc at Michigan State where he is now working on comparative epigenomics in maize and Tripsacum dactyloides.



Outreach Activities

Sunday with a Scientist: Working with members of both the Herr lab (Plant Pathology) and the Roston Lab (Biochemistry), I and other members of my lab designed and delivered in a module for the successful "Sunday with a Scientist" program titled "Why Plants Don't Wear Sweaters" which is run by the Nebraska State Museum and held in Morrill Hall on the third Sunday of each month.

http://museum.unl.edu/sundaywithascientist/may2016.html

Figure 6: Schnable and Roston lab members at a Sunday with a Scientist event held in 2016. Total attendance at the 2016 "Why Plants Don't Wear Sweaters" Sunday with a Scientist event was 110 individuals: Adults 61, Youth 45 and UNL students 4.

Fascination of Plants: As part of the 2017 celebration of "Fascination of Plants" day, the Schnable Lab designed and taught a module to introduce high school students from a local magnet school to comparative genomics techniques used the web interface/database CoGe. As a result of limited dry-lab space available to the Schnable Lab, enrollment in this module was capped at two shifts of 20 students each and every spot was filled.

Figure 7: One shift of 20 high school students crowded into the dry lab space of the Schnable lab to learn and practice basic comparative genomics techniques.

Peggy Smedley Interview: Invited guest interviewed on the Peggy Smedley show – a weekly podcast focused on the Internet of Things – about developments in plant phenotyping. Reported audience size of 100,000 downloads per week.

https://peggysmedleyshow.com/portfolio-items/05-22-18-episode-564-segment-2-iot-measures-and-manages-plants/

Twitter: In order to for the work I do to have impact, it is important to work to ensure people actually hear about it. As the number of scientific papers published each year continues to accelerate at a rate of 5-10% year, depending on which estimates you believe, it is less and less practical to simply publish papers in journals and sit back and wait for impact to happen – whether that impact is measured in terms of changes in real world outcomes, economic activity, changes in the broad scientific consensus within a field, or simply citations to the paper in question. Some of this necessary evil of self promotion can be achieved through one on one conversations and attendance at scientific conferences.

However, in addition to these approaches, I have developed and maintained a twitter account which I use to disseminate and promote the research and achievements of lab members and scientific collaborators. The account currently has 1,832 unique followers and generates 100,000-150,000 unique impressions per month. Generally it achieves 1,000-5,000 per tweet, although my record is >54,000 impressions for a twitter showcasing time lapse imaging of corn plant development. A modified version of this small time lapse dataset was later incorporated into the wikipedia page on "Heterosis" by a third party, with proper citation and attribution back to my lab and the student involved.

Water of Food: Interview by Daniel Carvalho (a PhD student in the lab) about his work studying the genetic basis of the convert evolution of C4 photosynthesis as part of the mission of the Daugherty Water for Food Global Institute.

https://www.youtube.com/watch?v=uFo2kuE4Qno

Example Syllabi

Biosystems Research I
Big Questions
LIFE891: 3 credits, Fall 2017

Instructional coordinator:
Dr. James Schnable
E207 Beadle Center
Phone: 472-3192
Email: schnable@unl.edu
Office hours: Normally, I will be available for office hours the hour after class. If you need to talk with me outside that time, please email for an appointment time. Additional office hours will be set by each team of instructors.

Meeting place and time: Tuesday/Thursday, 9:30-10:45 AM, 110 Othmer Hall.

Course Objectives: The course will provide an overview of major research questions in the life sciences focused on understanding complex biological systems. Emphasis will be on developing confidence in reading and critically evaluating seminal and current primary literature in a broad range of research on living systems. This will include an introduction to key techniques in biochemical, molecular, organismal, quantitative and computational biology relevant to the life sciences. Students will develop the following set of skills:

1) Conversant in major concepts and issues underlying significant ongoing research questions in the life sciences, including the state of our understanding, and the array of approaches that have been used to address the questions
2) Familiar with contemporary research methods in the life sciences
3) Able to articulate the strengths and limitations of discovery-based versus hypothesis-driven research, understand how use of methodologies differs with these objectives, and appreciate what research questions are best addressed by which approaches
4) Competent in critically reading and interpreting primary literature in diverse systems
5) Strong written communication and scientific presentation in life sciences

---

Prerequisites: No formal prerequisites, but basic knowledge of math, biology, biochemistry, and/or chemistry will be assumed.

***Recommended* Reference Text:** You do not need to purchase a textbook exclusively for this course. Instructor teams will provide recommendations for resources to consult for reference.

Primary literature: The overall schedule of topics and due dates is [will be] attached to this syllabus. Reading assignments for each class period's lecture topic will typically consist of a review article and/or 1-2 primary papers. These will be posted in pdf format on the Blackboard site, typically no less than a week before the material will be discussed in class. It will be essential to complete reading assignments prior to the class period so that you are prepared for the discussion. You may also need to consult reference texts or look up additional review materials as needed for background.

Big questions in current research will be examined in five modules. Examples of such questions include the following:

1. How do computational approaches advance the pace of life sciences research?
2. Cancer: why don't we have a cure?
3. How does the gut microbiome impact obesity and disease susceptibility?
4. Feeding the world: can water use be optimized with drought tolerant crops?
5. What are we doing to the soil (or water): impact of industry on environment?

Assignments:

The course will be assessed on the basis of class readings, discussions, one written project or collection of assignments per modular topic, and an integrative final. Dates and instructors for each module are in the schedule attached. Projects/assignments and respective due dates will be defined by each team. No late submission of written assignments will receive credit.

Grades will be based on the following criteria: grading scale

Five written projects Project 1 Project 2 Project 3 Project 4 Project 5	50% total 10% 10% 10% 10% 10%
Overall participation and preparation*	30%
Final	20%

*Attendance is a component of class participation and preparation, and is MANDATORY. If you will be absent, you must notify the instructor of the current topic *before* the class meeting with a detailed reason for absence and the date of absence (acceptable reasons for absence: at scientific meeting; surgery or similar significant medical issue (not just a cold); major experiment that cannot be done another day and time). This is for your benefit.

Academic Dishonesty: Academic dishonesty includes fabrication, falsification and plagiarism. Acts of academic dishonesty are not acceptable or tolerated by the scientific community. Thus, these acts should not be tolerated by students. Falsification of research data or its deliberate misinterpretation is a serious offense with consequences ranging from reprimand to dismissal to professional ruin. Plagiarism is more complex and may be the result of carelessness or ignorance, rather than an intentional attempt to deceive. Plagiarism is defined as passing off someone else's ideas, words or writings as your own. Inclusion of a sentence in a paper that is copied from any source without quotation marks and citation is an obvious example of plagiarism. **However, paraphrasing another person's writing or organization of ideas is also regarded as plagiarism.** Anything you present that is under your name should be entirely your own work unless so indicated and appropriately cited. This includes ideas, artwork and figures, as well as specific sentence, paragraph or word use. Note that in scientific writing, quotations are typically not used and it is therefore expected that all phraseology is your own, ***whether cited or not***.

This is the University's policy. We uphold it implicitly and we are committed to making sure you understand the reprehensible practice of plagiarism in all its nuances.

It is that important to your career, starting now.

If you do not fully understand what constitutes plagiarism, *please* ask for clarification immediately.

We expect and encourage you to discuss assignments and work together as you process the material, but that your individual compositions will be your own original writing. Assignments will be submitted electronically and automatically subjected to plagiarism screening by academic software. If you plagiarize, you will receive an F for the course. UNL's policies concerning grade appeals will be followed.

Students with disabilities are encouraged to contact me for a confidential discussion of their

individual needs for academic accommodation. It is the policy of the University of Nebraska to provide flexible and individualized accommodation to students with documented disabilities that may affect their ability to fully participate in course activities or to meet course requirements. To receive accommodation services, students must be registered with the Services for Students with Disabilities (SSD) office, 132 Canfield Administration, 472-3787 voice or TTY.

Grades of Incomplete: According to UNL policy, an Incomplete will be given only in the event of acute illness, military service, hardship, or death in the immediate family (i.e.; parents, children, spouses or siblings). Students are eligible for an Incomplete only if coursework is substantially completed. For this course, at least 50% of the class discussions and coursework must have been completed with a minimum overall grade of C to be considered for the Incomplete option. Resolution of the Incomplete grade would entail retaking the entire course within two years.

Classroom Emergency Preparedness and Response Information

- **Fire Alarm (or other evacuation):** In the event of a fire alarm: Gather belongings (Purse, keys, cellphone, N-Card, etc.) and use the nearest exit to leave the building. Do not use the elevators. After exiting notify emergency personnel of the location of persons unable to exit the building. Do not return to building unless told to do so by emergency personnel.
- **Tornado Warning:** When sirens sound, move to the lowest interior area of building or designated shelter**.** Stay away from windows and stay near an inside wall when possible.
- **Active Shooter**
  - **Evacuate:** if there is a safe escape path, leave belongings behind, keep hands visible and follow police officer instructions.
  - **Hide out:** If evacuation is impossible secure yourself in your space by turning out lights, closing blinds and barricading doors if possible.
  - **Take action**: As a last resort, and only when your life is in imminent danger, attempt to disrupt and/or incapacitate the active shooter.

UNL Alert: Notifications about serious incidents on campus are sent via text message, email, unl.edu website, and social media. For more information go to: http://unlalert.unl.edu.

Additional Emergency Procedures can be found here:
http://emergency.unl.edu/doc/Emergency_Procedures_Quicklist.pdf

JClarke 7/30/17

Schedule of lectures and assignments

SA	Tues	Aug 22	Overview and history	TBD
	Thurs	Aug 24	Systems analysis	Jean-Jack Riethoven
	Tues	Aug 29		
	Thurs	Aug 31		
PBS	Tues	Sept 5	Pathobiology and	Rodrigo Franco Cruz
	Thurs	Sept 7	Biomedical science	
	Tues	Sept 12		
	Thurs	Sept 14		
	Tues	Sept 19		
	Thurs	Sept 21		
MI/IPB	Tues	Sept 26	Microbiome and Metagenomics	Joshua Herr
	Thurs	Sept 28		
	Tues	Oct 3		
MI	Thurs	Oct 5	Microbial interactions	Amanda Ramer-Tait, Andrew Benson, Jacques Izard
	Tues	Oct 10	Gut microbiome	
	Thurs	Oct 12		
	Tues	Oct 17	*Fall break, no class*	
	Thurs	Oct 19		
	Tues	Oct 24		
	Thurs	Oct 26		
IPB	Tues	Oct 31	Soil microbes and interactions with roots	Daniel Schachtman
	Thurs	Nov 2	TBD	
	Tues	Nov 7	Integrative plant biology	Tom Clemente
COB EE	Thurs	Nov 9	Computational organismal	Colin Meiklejohn
	Tues	Nov 14	Biology, Ecology, Evolution	TBD
	Thurs	Nov 16		Drew Tyre
	Tues	Nov 21		Clay Cressler
	Thurs	Nov 23	*Thanksgiving break, no class*	
	Tues	Nov 28		John DeLong and Kristi Montooth
	Thurs	Nov 30		Steven Thomas
SA	Tues	Dec 5	Systems analysis	James Schnable
	Thurs	Dec 7		
	Wed	Dec 13	10am-Noon official final time	

SA, Systems Analysis; **PBS**, Pathobiology and Biomedical Science; **MI**, Microbial Interactions; **IPB**, Integrative Plant Biology; **COBEE**, Computational Organismal Biology, Ecology and Evolution

LIFE 843 – Professional Development

1 Credit

Instructor: James Schnable
E207 Beadle Center
City Campus
schnable@unl.edu

No required text book

Course Information:
This course will meet once per week and is intended for first year graduate students in the life sciences.
Course Objectives:
At the conclusion of this course, students should be able to:
1. Present scientific information in both oral and written formats in ways accessible to others with scientific backgrounds outside of their own specialization
2. Effectively read the scientific literature, and for individual papers be capable of assessing what was known previously, the hypotheses put forward, and critically assess the whether the results put forward by the authors represent an effective test by the authors.
3. Understand what constitutes scientific misconduct, how to avoid committing misconduct yourself, and how to deal with observing others committing apparent misconduct
4. Describe their career goals post graduate school and describe how they are selecting a lab and research project which will enable them to work towards those goals.
Assignments and Assessment:
Throughout this course you will give two presentations, write one two page research statement (and associated preliminary documents), and participate in a mock peer review, as well as discussions of scientific misconduct.
Five minute presentation (10% of final grade): This presentation should include no more than three slides, take no more than four minutes (leaving one minute for questions). The first slide should introduce the topic for those with no background in the topic (what does the audience need to know to understand your presentation, why is it relevant to their lives), the second slide should present more detailed information, and the third should explain the significance of the information presented on the first two.

Presentations can be on any topic other than your past or current research. Previous topics have included: Strategies for conserving water, the origins and future threat of killer bees, and how to cook thai curry.

To receive full credit your presentation cannot exceed 240 seconds and must be intelligible to both the instructor and other students.
Academic paper presentation (30% of final grade): This presentation should be approximately 15 minutes in length. At a minimum you should include a title slide, a slide listing the hypothesis you believe the authors tested in this paper, a bullet point list summarizing the introduction, slides for each figure in the paper you are presenting (supplementary figures can be optionally included if you think they help describe the paper), a slide summarizing the conclusions the authors drew from their results, and a final slide describing your assessment of whether the authors results A) support their conclusions B) represent a valid test of their hypothesis.
When you present each figure you say what it is, what it is showing, and why you think the

authors included in their manuscript.
Two page research statement (40% of final grade): This research statement will track the form of the statement required by the NSF GRFP. Students will submit a list of 3-5 potential research topics, a research statement outline, and rough draft, and a final draft. All documents should be submitted electronically and are due the midnight before class on the week the assignment is due. Point can be lost for failure to turn in research topics, outline, and rough draft on time, however quality of each student's work will be assessed only from the final paper.

Class participation (Mock peer review 10% Academic misconduct discussion 10%).
Attendance and Participation:
Attendance is mandatory. If you need to miss a class for an outside event, illness, or other event, contact me prior to the start of class.
Weekly schedule:
Week 1 (August 21nd): Introduction to the course. Anatomy of a good scientific presentation. *Sign up for presentation dates.*
Week 2 (August 28th): Five minute/three slide presentations for all students
Week 3 (September 4th): Receive and read three example research statements. Conduct mock peer review and stack rank during class.
Week 4 (September 11th): No class*. Students should submit 3-5 ideas for research projects (as little as one sentence each) by e-mail.*
Week 5 (September 18nd): Three student presentations on academic papers
Week 6 (September 25th): Three student presentations on academic papers. One-on-one meetings to go over research proposal ideas should be scheduled for this week.
Research proposal outlines due
Week 7 (October 2th): Draft research statements due. Peer review of draft research statements. Class will divide into two groups, evaluate anonymous research statements, and provide a summaries of strengths and weaknesses of each proposal to the instructor.
Students should submit draft research statements electronically to the instructor no later than midnight the night before class.
Week 8 (October 9th): No class. Final research statements, including “response to reviewers” due.
October 16th: Fall break, no class
Students will received instructor feedback and anonymized peer feedback on final research statements.
(Reminder for those planning to submit NSF-GRFP proposals that they will be due by October 22nd (Monday).
Week 9 (October 23th): Class discussion of examples of recently retracted scientific papers taken from “Retraction Watch”
Week 10 (October 30th): TBD.
Week 11 (November 6th): Group discussions of academic misconduct take home examples;
Short responses to academic misconduct examples due
Week 12 (November 13th): Plagiarism and dual publication.
Week 13: (November 20nd) Academic Life Skills: Choosing an advisor, choosing a research project.
Week 14: (November 27th) Academic Life Skills: Life after grad school (industry research, academic research, academic teaching, science related non-research fields).
Week 15: (December 4th): Open discussion
Additional required information:

Fire Alarm (or other evacuation): In the event of a fire alarm: Gather belongings (Purse, keys, cellphone, N-Card, etc.) and use the nearest exit to leave the building. Do not use the elevators. After exiting notify emergency personnel of the location of persons unable to exit the building. Do not return to building unless told to do so by emergency personnel.
Tornado Warning: When sirens sound, move to the lowest interior area of building or designated shelter. Stay away from windows and stay near an inside wall when possible.
Active Shooter
o **Evacuate:** if there is a safe escape path, leave belongings behind, keep hands visible and follow police officer instructions.
o **Hide out**: If evacuation is impossible secure yourself in your space by turning out lights, closing blinds and barricading doors if possible.
o **Take action:** As a last resort, and only when your life is in imminent danger, attempt to disrupt and/or incapacitate the active shooter.
UNL Alert: Notifications about serious incidents on campus are sent via text message, email, unl.edu website, and social media. For more information go to:http://unlalert.unl.edu.
Additional Emergency Procedures can be found here:
http://emergency.unl.edu/doc/Emergency_Procedures_Quicklist.pdf
Students with disabilities are encouraged to contact the instructor for a confidential discussion of their individual needs for academic accommodation. It is the policy of the University of Nebraska-Lincoln to provide flexible and individualized accommodation to students with documented disabilities that may affect their ability to fully participate in course activities or to meet course requirements. To receive accommodation services, students must be registered with the Services for Students with Disabilities (SSD) office, 132 Canfield Administration, 472-3787 voice or TTY.
Academic Integrity:
Academic integrity is an essential indicator of the student's ethical standards. For this reason students are expected to adhere to guidelines concerning academic honesty outlined in Section 4.2 of University's Student Code of Conduct which can be found at http://stuafs.unl.edu/ja/code/three.shtml. Students are encouraged to contact the instructor to seek clarification of these guidelines whenever they have questions and/or potential concerns.
a. Breaches of academic integrity and their consequences vary considerably, so it is not possible to outline one set of absolute chain of consequences for every situation
b. Each instructor may impose a consequence(s) for a breach of academic integrity in his/her own course, consistent with the magnitude of the breach. The consequences may range from reduced credit for a test or assignment to failure in the course.
c. If the student feels that the consequence(s) imposed are inappropriate, the student should discuss the matter first with the instructor within 7 days of the incident.
d. If the student is still dissatisfied with the consequences imposed, he/she may appeal to the Department Head or his/her designee within 14 days of the incident.
e. If the student is dissatisfied with the results of his/her appeal to the Department Head, then he/she may appeal to the Dean of the College of Agricultural Sciences and Natural Resources within 21 days of the incident.
f. Further appeal may be pursued with the University Judicial Officer as described in http://stuafs.unl.edu/ja/code/three.shtml.
g. The course instructor will inform the student's academic advisor of the final disposition of the breach of academic integrity immediately after the final decision

Appendix B: Supporting Evidence for Research Activity and Outcomes

Cover Page Summaries of Funded Grants

Figure 8: RoL: FELS: EAGER: Genetic Constraints on the Increase of Organismal Complexity Over Time. (Schnable is PI)

Overview:
The research included as part of this project is aimed at testing a proposed rule of life: that increases in organismal complexity are constrained by the availability of certain types of genes which are recalcitrant to most forms of gene duplication, but can be duplicated as part of polyploidy (whole genome duplication) (Freeling and Thomas, 2006; Freeling, 2009). Most apparent links between whole genome duplication and an increase in the number of separately defined body parts – for example the evolution of early tetrapods or the emergence of flowering plants – occurred tens or hundreds of millions of years ago. However, we will use a much more recent model to test this hypothesis. The model system *Zea mays* (maize; corn) produces two specialized types of inflorescences for male and female reproduction which have been shown to be controlled by distinct genetic architectures, while all the other genera in the grass tribe to which it belongs – Andropogoneae – produce only a single type of inflorescence. This proposal seeks to test the link between a whole genome duplication in the maize lineage and the evolution of its developmentally distinct inflorescences. These tests will be conducted using both conventional comparative genomics and novel comparative genetics techniques enabled by reverse genetics resources in both maize, and a closely related species *Sorghum bicolor*, which lacks both the maize whole genome duplication and the specialized and genetically differentiated inflorescences found in maize.

Intellectual Merit:
In both plants and animals, the emergence of new specialized body parts – for example the development of floral organs in plants or the multiple specialized types of teeth found in heterodont animals – is a rare process. Generally these specialized organs appear to originate as specialized versions of existing organs, yet their specialization requires a divergence in regulation between different copies of the same organ. This EAGER proposal seeks to test the hypothesis that, in the case of the specialized reproductive organs of maize, the separate regulation and evolution of what were, initially, duplicate copies of the same organs were enabled by a whole genome duplication, which created duplicate copies of many transcription factors which rarely duplicate through other processes. In addition, through the generation and phenotypic characterization of knockouts of syntenic orthologous genes in both maize and sorghum, this proposal provides one of the first systematic tests of the Ortholog Conjecture in plants (Koonin, 2005; Nehrt et al., 2011; Chen and Zhang, 2012).

Broader Impacts:
If successful, this proposal will have defined a potential rule of life that may also explain why the emergence of new specialized body parts is so rare, can predict when the emergence of new specialized body parts is more likely, and provide initial insights into how specialized body parts could be engineered in future synthetic biology efforts. Research in this area also has the potential to guide the development of new engineered varieties of crop plants with multiple specialized leaf types deployed in different parts of their canopies, okafor example – engineering top layer leaves to have fewer chloroplasts, allowing more light to penetrate deeper into the canopy where lower wind speeds and higher humidity reduce the transpirational water cost of photosynthesis (Ort et al., 2011, 2015) or engineering bottom leaves to express modified forms of chlorophyll such as chlorophylls D & F which can harvest energy from abundant far red light present lower in crop canopies (Chen et al., 2010; Croce and Van Amerongen, 2014). In addition this proposal will provide valuable training on conducting science at the intersection of genomics, genetics, and phenomics to one postdoc, and has the potential to provide spin off projects characterizing individual orthologous gene triplets in more detail to multiple undergraduates, supported by UNL's internally funded UCARE program as well as an existing REU program in which Schnable participates at UNL.

Figure 9: USDA NIFA AFRI Foundational: Identifying mechanisms conferring low temperature tolerance in maize, sorghum, and frost tolerant relatives. (Schnable is PI)

PROJECT SUMMARY

Instructions:
The summary is limited to 250 words. The names and affiliated organizations of all Project Directors/Principal Investigators (PD/PI) should be listed in addition to the title of the project. The summary should be a self-contained, specific description of the activity to be undertaken and should focus on: overall project goal(s) and supporting objectives; plans to accomplish project goal(s); and relevance of the project to the goals of the program. The importance of a concise, informative Project Summary cannot be overemphasized.

Title: Identifying mechanisms conferring low temperature tolerance in maize, sorghum, and frost tolerant relatives.

PD: Schnable, James, C	**Institution:** University of Nebraska-Lincoln
CO-PD: Roston, Rebecca	**Institution:** University of Nebraska-Lincoln
CO-PD:	**Institution:**
CO-PD:	**Institution:**
CO-PD:	**Institution:**
CO-PD:	**Institution:**
CO-PD:	**Institution:**

Key crops in the US – maize and sorghum – are quite sensitive to both cold and freezing temperatures resulting in crop losses from unexpected cold snaps, and limiting growing seasons at more northern lattitudes, resulting in lower yields. Maize and sorghum show extremely limited genetic variation for survival under freezing temperatures, limiting the effectiveness of conventional intraspecific quantitative genetic approaches. The central premise of this proposal is that interspecific comparisons between low temperate tolerant and low temperature sensitive species within the same grass subfamily will make it possible to link changes in gene regulation to changes in membrane lipid composition and metabolite accumulation known to be two central components freeze and cold tolerances. To this end, changes in membrane lipid composition and gene expression in response to cold and subsequent freezing stress will be profiled for 10 panicoid grasses in objective one. Synteny-based comparative genomic approaches will be employed to enable the comparisons of the transcriptional response of the same genes across multiple species to the same changes in environmental condition. A broader set of 180 grass species will be assayed for cold and freezing tolerance in objective two. And in objective three patterns of metabolic change in response to cold stress will be assayed in maize, sorghum, and two close relatives in which freeze tolerance has evolved independently. Integrating multiple types of data from multiple species in response to different degrees of low temperature stress will permit the identification of mechanisms which convey low temperature tolerance in panicoid grasses.

Figure 10: ARPA-E Roots: In-plant and in-soil microsensors enabled high-throughput phenotyping of root nitrogen uptake and nitrogen use efficiency. (Schnable is co-PI)

1. INNOVATION AND IMPACT

1.1 Overall Description.

Nitrogen (N) is a key nutrient for crop plants, and improved nitrogen use efficiency (NUE) can significantly reduce fertilizer applications, increase crop yields, and reduce the environmental footprint of agriculture. By examining plant N uptake and NUE on appropriate populations, the genetic control of physiological phenotypes can be defined. Although progress has been made in collecting in-field trait data, high-throughput and high-accuracy measurements of below-ground phenotypes and *in planta* phenotyping of N status have so far not been possible. Sensing of plant-soil processes that control NUE is currently limited to methods that are time-intensive, laborious, and destructive and have low information content with respect to spatiotemporal characteristics of N soil N supply and plant N uptake.

We propose to develop Micro-Electro-Mechanical Systems (MEMS)-based *in planta* and soil N sensors that will extend rapidly measureable plant phenotypes from yield, number of leaves, and flowering time to deep physiological traits directly related to root N uptake and NUE. The *in planta* and soil nitrate measurements will be coupled to create a direct measurement of fertilizer NUE that is *not* currently possible, and thereby enable rapid identification of genotypes with N-uptake-proficient root systems. The core concept is to create a low cost sensors-enable high-throughput, high-accuracy, large-scale NUE phenotyping platform combining minimally invasive, *in planta* nitrate sensors and microfluidic soil nitrate sensors, to enable integrated sensing of plant and soil processes that influence NUE.

The unique *in planta* sensors, in the form of a microscale needle, are inserted into multiple sites of the plant to prove frequent and accurate monitoring of nitrate uptake. The sensors will create an unprecedented ability to describe the spatial distribution and temporal variations of *in planta* nitrate status. In addition, we will develop microfluidic soil nitrate sensors capable of monitoring soil nitrate concentrations with full automation from sampling to quantification to signal processing. We will validate the proposed sensors and NUE phenotyping platform, by deploying hundreds of low-cost nitrate *in planta* and soil sensors within yield trials of corn with known genotypes across multiple, well-defined environments.

Together, *in planta* and soil nitrate sensors will create tremendous synergy from which the integrated data could efficiently identify interactions between genotype and environment that drive N-uptake-proficient root systems, potentially reducing the need for costly and laborious root phenotyping. At present, a plant's ability to access soil N requires destructive, labor-intensive measurements. With coupled *in planta* and soil nitrate sensing, the amount of nitrate uptake per amount of soil nitrate can be directly measured, creating a continuous measurement of N uptake per amount of available soil N.

Objectives

This proposal has two objectives that will enable next-generation field-based, high-throughput phenotyping to accurately assess NUE.

(1) Develop, calibrate and optimize low cost, efficient MEMS-based *in planta* sensors and integrated microfluidic soil sensors for accurate measurements of plant and soil nitrate levels.

(2) Establish and validate a field-based, high-throughput phenotyping platform with coupled *in planta* and soil sensors to generate data from yield trials of known genotypes in multiple, well-defined environments.

Figure 11: NSF EPSCoR Track I: Center for Root and Rhizobiome Innovation. (Schnable is a member of the management)

PROJECT SUMMARY

Overview:

The Center for Root and Rhizobiome Innovation (CRRI) will be established to develop tools and technologies for more rapid, precise, and predictable crop genetic improvement that complement and transcend methods currently used by biotechnologists and plant breeders. These innovations are needed because of the urgency and enormity of challenges facing global agriculture, including the need to feed a rapidly growing population in the face of extreme climate variations and limitations in water and soil vitality. CRRI research will be structured around a systems and synthetic biology core to generate and iteratively improve network models of plant metabolism for predictable outcomes from genetic modifications. CRRI's systems and synthetic biology research will be applied to the study of root metabolism and its influence on root-interactions with soil microbes for improved plant health. Research will focus on root metabolism in maize, a plant genetic model and important crop species, but findings will be broadly applicable to other plants and crop species. CRRI will develop and use fundamental knowledge to create translational products with far-reaching impact on plant and microbial biology and global agriculture.

Intellectual Merit :

CRRI research will be based on generation of large omics datasets from analyses and measurements of root gene expression and metabolites, soil microbiota, and plant phenotypes. Computational innovations for extraction and integration of these datasets will drive the creation of algorithms for predictive model construction and databases for more efficient data mining. CRRI researchers will also devise next-generation synthetic biology tools for precise delivery and expression of large numbers of trait genes in crop plants. By combining computationally-derived models and synthetic biology tools, CRRI will advance the predictive engineering of plant metabolism to accelerate crop improvement. Systems and synthetic biology-enabled research on root-microbiome interactions also lead to innovative solutions for agroecosystem sustainability. Furthermore, collaborations among CRRI engineers and biologists will result in new fiber optic technology for real-time, non-destructive sampling of root-associated microbes and root exuded chemicals. This technology will advance rhizosphere and root-microbiome studies, emerging research areas of agricultural and ecological significance.

Broader Impacts :

Innovations arising from CRRI will enable strategies for increasing the rate and precision of crop genetic improvement and will generate maize germplasm with altered root metabolism and soil microbe interactions for improved tolerance to drought and low soil fertility. These products will be important for the sustainability of Nebraska's agriculture-based economy and for meeting the global grand challenges to agricultural production. CRRI will implement an innovative and comprehensive research-based STEM portfolio of education and workforce development activities along the continuum of the STEM pipeline. CRRI will implement: 1) secondary education programs to build the pipeline of students choosing STEM studies and careers; 2) undergraduate and graduate training programs that prepare the next generation of researchers and industry leaders and a rigorous postdoctoral scientist mentoring program; 3) faculty development programs including early career faculty and small colleges that increase research capabilities and provide new opportunities for students; and 4) STEM events that educate the public in the science and ethics of new technologies for informed consumer choices. Over the entire award period, CRRI will provide 55 person-years of postdoctoral scientist training and will support 35 graduate students and 120 undergraduates (cumulative) as participants in research-based activities. CRRI will further impact 15 small college faculty, 15 small college undergraduates, 950 undergraduates in CRRI enhanced courses, 20 students in internships, and 4,680 grade 7-12 students (cumulative). The CRRI diversity plan, through targeted strategies and support of successful programs, will accelerate the pace of diversification among students, postdoctoral scientists, and faculty while facilitating the participation of a broadly-defined group of individuals in CRRI activities.

Figure 12: NSF EPSCoR Track-2 FEC: Functional analysis of nitrogen responsive networks in Sorghum. (Schnable is co-PI)

Project Summary

Overview
We propose to establish a partnership for cutting edge plant genomic research between two EPSCoR regions of Alabama and Nebraska. The team at University of Nebraska-Lincoln will contribute their expertise in plant transformation and automated phenotyping using their new state-of-the-art LemnaTeC high-throughput system for imaging large plants. The team at HudsonAlpha Institute for Biotechnology in Huntsville, Alabama will contribute dedicated outreach for agricultural biotechnology education and genomic and molecular analysis of plant networks. We will combine these advanced tools to better understand the regulation of a complex agronomic trait of agricultural, economic, and environmental importance: how nitrogen affects plant growth and development.

Intellectual Merit
We will collect baseline information of plant response to nitrogen levels through combined transcriptomics, automated phenotyping, and molecular function methods in the widely used grain and biomass crop sorghum. We will modify the function and regulation of key hubs and transcription factors using CRISPR/Cas9 transformation methods targeting the promoter regions of these genes. We will then characterize plants with the resulting modified genotypes that show nitrogen response deploying new nitrogen uptake sensors, automated phenotyping through the full cycle of seasonal plant development, and genomics of expression based network reconstruction to identify key modifiers of nitrogen use efficiency. We anticipate that the model and pipelines established through the combined efforts of these two institutions will be extendable to understanding the biology underlying other complex agronomic traits in addition to nitrogen use efficiency.

Broader Impacts
In addition to a deep understanding of the molecular basis of plant nitrogen uptake throughout the growing season and the creation of many sorghum lines that can be used for additional nitrogen experiments, this proposal also seeks to train and motivate students to become involved in genetic and biotechnology based research for agriculture. As part of the proposed research, two graduate students will develop methods for automated phenotyping image analysis, a postdoctoral researcher will learn new skills and develop pipelines to go from transcriptomic analysis to empirical determination of gene function, and an early career faculty member will further develop her program in crop functional genomics. HudsonAlpha will develop and deploy a 3-week summer course, 'AgriGenomics Academy', for advanced high school students that will be offered for 3 years to 16-18 students each year to excite students to go on to genomics based research efforts in plants. Additional impacts include the recruitment of three undergraduate students who will complete summer internships at both HudsonAlpha and University of Nebraska-Lincoln to learn advanced techniques, and support for the Launching Aspiring Biotechnology Students (LABS), which introduces low and moderate-income students to biotechnology.

Figure 13: USDA/NSF Joint Program PAPM EAGER: Transitioning to the next generation plant phenotyping robots. (Schnable is co-PI)

1 Project Overview, Goal, and Objectives

One of the grand challenges facing agriculture today is to produce enough food and energy for a world population likely exceeding 9.7 billion by 2050. To achieve this goal, the overall production of all major food and energy crops will have to double, which has to occur in the context of climate change (Tilman et al., 2011). Crop yield improvement in the past few decades has been mainly attributed to green revolution, inorganic fertilizer and mechanization; but these technologies are unlikely to sustain the needed yield increase for another 35 years (Ray et al., 2013). Plant phenomics, the use of holistic large scale approaches to collect plant phenotypic information, has the potential to spark a new green revolution (Houle et al., 2010). It would fill the gap between the low cost of generating large scale datasets of plant genotypes and the time consuming and expensive processes of collecting large scale phenotypic datasets. Advancement in plant phenomics would enable more effective utilization of genetic data, and ultimately lead to novel gene discovery and improved crop yield in the field (Furbank and Tester 2011).

In recent years there has been a rapid expansion of high throughput plant phenotyping using digital image analysis (Golzarian et al., 2011; Chen et al., 2014). As the current state of the art, digital imaging has proven to be very useful in obtaining plant traits such as size and growth (Neilson et al., 2015; Neumann et al., 2015). An intrinsic limitation of image-based phenotyping, however, lies in the fact that images are indirect measurements. The results are in terms of pixel count or pixel intensity, which by its own convey little information regarding the traits of biological importance. This is particularly true when plant chemical and physiological traits such as water content and photosynthesis are to be measured. To maximize the use of images, researchers are required to collect ground truth trait measurements to establish correlations between the ground truth data and images. Because ground truthing is performed by humans, it is slow and expensive, and represents a major limiting factor for image-based plant phenotyping.

There are many specialized plant sensors designed to measure a wide array of plant physiological or chemical traits. Several examples are the fiber-optic sensing head coupled with a NIR spectrometer for leaf reflectance, handheld leaf porometers for stomatal conductance and gas exchange, anthocyanin meters, and portable photosynthesis system for leaf fluorescence and Photosystem II analysis (Figure 1). All these sensors are designed to be operated by human for *in vivo* plant sensing. Conceivably, robotic systems can be developed to integrate these sensors for autonomous trait measurements.

Figure 1. Specialized plant leaf sensors for *in vivo* plant sensing

The goal of this project is to develop automated robotic systems that can realize *in vivo*, human-like plant phenotyping in the greenhouse. Our central hypothesis is that the throughput, capacity and accuracy of phenotyping by the automated robots will be much better than human, whereas the cost will be substantially lower. Toward that end, there are three specific research aims listed in Research Tasks section (see section 4).

Pending Grants

1. "Hybrid Pearl Millet as a Biomass Crop for the Arid Great Plains"
   Department of Energy - Affordable and Sustainable Energy Crops
   Schnable JC (PI), Ge, Yufeng (co-PI) (Biological and Systems Engineering, UNL), Dweikat, I (Agronomy and Horticulture, UNL), Wilkins M (Biological and Systems Engineering, UNL), Yang J (Agronomy and Horticulture, UNL), Cheng X (Mathematics, University of Nebraska-Omaha), Serba, D (Agricultural Research Center, Kansas State University) Award Period: 7/1/2019 - 6/30/2024
   Award Amount: $3.9M total.

2. "Crops in silico: Increasing crop production by connecting models from the microscale to the macroscale"
   Foundation for Food and Agricultural Research
   Amy Marshal-Colon (University of Illinois- Urbana Champaign, **Schnable JC (co-PI)** (One of eight total co-PIs, including faculty at UIUC, Purdue, and Penn State)
   Award Period: 12/1/2018 - 12/31/2022
   Award Amount: $5M total. Funding directly and specifically to the Schnable Lab: $493,823

3. "Ultra-Low Power Sensor Network"
   ARPA-E - Open Call
   Full proposal submitted after encouragement from ARPA-E based on evaluation of a preproposal.
   Kim H (PI) (Department of Electrical and Computer Engineering, University of Utah), **Schnable JC (co-PI)** (one of four total co-PIs, sole plant biologist)
   Award Period: 1/1/2019 - 12/31/2020
   Award Amount: $1.1M total. Funding directly and specifically to the Schnable Lab: $198,845

4. "BTT EAGER: A wearable plant sensor for real-time monitoring of sap flow and stem diameter to accelerate breeding for water use efficiency"
   National Science Foundation - Breakthrough Technology Call
   Full proposal submitted based on an invitation from NSF after evaluation of a preproposal.
   Schnable JC (PI), Dong L (co-PI) (Electrical and Computer Engineering, ISU), Castellano M (co-PI) (Agronomy, ISU), Schnable P (co-PI) (Agronomy, ISU)
   Award Period: 1/1/2019 - 12/31/2020
   Award Amount: $300k total. Funding directly and specifically to the Schnable Lab: $99,299

5. "EAGER-SitS: Ultra-Low-Power, Event-based-Wake-Up, Wireless Chemical Sensor Networks for Long-Term Underground Soil Monitoring"
   Full proposal submitted based on an invitation from NSF after evaluation of a preproposal.
   National Science Foundation - Signals in the Soil
   Kim H (PI) (Department of Electrical and Computer Engineering, University of Utah, **Schnable JC (co-PI)** (one of two total co-PIs)
   Award Period: 9/1/2018 - 8/31/2020
   Award Amount: $300k total. Funding directly and specifically to the Schnable Lab: $100,000

6. "EAGER SitS: High-resolution Measurement of N Dynamics Using a Miniature in-Soil Lab"
   Full proposal submitted based on an invitation from NSF after evaluation of a preproposal.
   National Science Foundation - Signals in the Soil
   Dong L (PI) (Electrical and Computer Engineering, ISU) **Schnable JC (co-PI)** (one of four total co-PIs
   Award Period: 1/1/2019 - 12/31/2020
   Award Amount: $300k total. Funding directly and specifically to the Schnable Lab: $15,273

Letters From UNL Administrators

University of Nebraska Lincoln

INSTITUTE OF AGRICULTURE AND NATURAL RESOURCES
AGRICULTURAL RESEARCH DIVISION
Office of the Dean

November 7, 2016

James Schnable, Ph.D.
Assistant Professor
Department of Agronomy and Horticulture
BEAD E207
Lincoln, NE 68583-0660

Dear Dr. Schnable:

The Agricultural Research Division (ARD) in conjunction with the ARD Advisory Council recently reviewed eight nominations for the 2016 Junior Faculty Excellence in Research award. This year's nominees were of exceptional high quality.

Congratulations! It is my pleasure to inform you of your selection as a recipient of the Junior Faculty Excellence in Research award presented by ARD. This honor attests to the excellence of your research program and to the potential that you have to make outstanding contributions in the future. As part of this recognition, you will receive $3,000 from ARD to use for research or for professional development activities. Staff in the ARD office will work with staff in your Business Center to process the paperwork for this award.

ARD will host a reception to honor you as one of the recipients of the Junior Faculty Excellence in Research award. Staff in the ARD office will contact you to discuss potential dates and times prior to making the final arrangements for the reception.

Best wishes for continued success with your research program at the University of Nebraska-Lincoln!

Sincerely,

Archie C. Clutter
Dean, Agricultural Research Division
Director, Nebraska Agricultural Experiment Station

cc: Roch Gaussoin
Ed Cahoon
Sue Walker
Interim Vice Chancellor Ron Yoder

University of Nebraska Lincoln

OFFICE OF THE CHANCELLOR

April 30, 2018

Dr. James Schnable, Assistant Professor
Department of Agronomy and Horticulture
University of Nebraska-Lincoln
E207 Beadle Center
1901 Vine St.
Lincoln, NE 68588-0665

Dear James,

Congratulations on your recent Marcus Rhoades Early Career Maize Genetics Award from Maize Genetics! This accomplishment demonstrates the significant impact of your research and the potential of your career moving forward. I wanted to let you know how much our university community recognizes and values this important distinction.

Indeed, we all appreciate how important national and international societies are to scholarship and the academy—this award from Maize Genetics exemplifies your commitment to cutting-edge research and discovery. And, while external recognitions are important to you personally and professionally, they also enhance the profile of the Department of Agronomy and Horticulture, the College of Agricultural Sciences and Natural Resources, and the university. Again, congratulations on this well-deserved award.

James, we are so pleased to have distinguished faculty like you on our campus. Thank you for your commitment to excellence and your contributions to the University of Nebraska-Lincoln.

Sincerely,

Ronnie D. Green

Ronnie D. Green, Ph.D.
Chancellor

James-
You are rocking it!
RDG

c: Michael Boehm, Ph.D., Harlan Vice Chancellor, Institute of Agriculture and Natural Resources
Tiffany Heng-Moss, Ph.D., Interim Dean, College of Agricultural Sciences and Natural Resources
Richard Ferguson, Ph.D., Interim Department Head, Department of Agronomy and Horticulture

Press Releases and News Articles

UNIVERSITY OF NEBRASKA–LINCOLN

NEBRASKA TODAY

September 1, 2016

Study in contrasts: System advances analysis of corn

by Scott Schrage | University Communication (/written-by/scott-schrage-university-communication/)

Greenhouse Innovation Center

The prospect of a higher-yielding Corn Belt could rest – or advance – on a conveyer belt monitored by cameras that boast superhuman sight, according to new research from the University of Nebraska-Lincoln.

Known as a high-throughput phenotyping system, the automated set-up resides at the Greenhouse Innovation Center (http://innovate.unl.edu/greenhouse-innovation-center) on Nebraska Innovation Campus. The system can rapidly measure and compare the physical traits, or phenotypes, of different crop varieties by transporting plants through several 360-degree imaging chambers.

Researchers Yufeng Ge and James Schnable are investigating how the phenotyping system, one of just a few in the United States, can be used to estimate certain properties of corn. The crop's unwieldy size and complex anatomy have

left it mostly ignored by previous automated phenotyping work, the researchers said.

In a recent study, Ge and Schnable demonstrated that images taken by the system's hyperspectral camera – a technology that detects a much wider range of the electromagnetic spectrum than the human eye – can help quantify the amount of water in a corn plant. Whereas a conventional camera detects wavelengths of only visible light, the system's hyperspectral camera can capture 240 slivers of wavelengths from both the visible and near-infrared portions of the spectrum.

"In a lot of previous studies, phenotyping was just trying to quantify the size and growth of the plants," said Ge, assistant professor of biological systems engineering. "But we were also trying to answer the question of whether we can use a hyperspectral imaging system to predict water content, which is one of the most important (traits) for plant physiology and breeding. We were fairly successful in doing that."

The researchers already knew that water-filled plant tissues absorb different wavelengths of light – and absorb the same wavelengths differently – than do their drier counterparts. Using this knowledge, they applied statistical methods to connect changes in various wavelengths with known changes in the water content of corn plants. This allowed them to build a mathematical model that hewed closely to measurements of actual water content, the study reported.

"This is essentially the same technology we use to analyze the atmospheres of planets in other solar systems," said Schnable, assistant professor of agronomy and horticulture. "You know the intensity of all the light coming from the star, so when a planet comes in front of the star, you look at what wavelengths (disappear). Similarly, we know the wavelengths of all the lights pointing at the plant, so we can look at which ones come back from the plant and which ones don't. That lets us see what's (being absorbed).

"The exciting thing here is that there are now so many things that we could potentially measure. So we have this whole new challenge. What are the measurements that are going to be the most informative? We don't even know in a lot of cases. Before automated phenotyping technology, we picked the measurement that was easy to make. Now there are thousands of measurements that are, in principle, equally easy to make. It's a very good problem to have."

Second sight

Ge and Schnable also showed that conventional RGB imagery from the phenotyping system can be used to estimate the daily growth of corn plants – and how efficiently they use water to stimulate that growth – during their first few weeks of development.

"There are probably other studies that have looked at corn seedlings," Schnable said. "But I don't think anyone has been able to take corn to the advanced stage of development while doing this type of imaging because it's a big plant that wouldn't fit into the smaller imaging chambers used in other automated phenotyping systems."

The researchers began by feeding daily images of each plant from two perpendicular angles into a program capable of distinguishing plant from background. Mathematical software averaged the two images into one, approximating a plant's total surface area by counting the number of plant-covered pixels in the composite image.

Ge and Schnable found that the software's estimates of plant size correlated strongly with their own measurements of plant weight, leaf area and water use efficiency. The methods required to establish those baseline values help illustrate why the RGB and hyperspectral imaging techniques should prove so useful, the researchers said.

The phenotyping system does automatically weigh and water the plants at regular intervals, allowing the team to periodically measure water consumption of sampled plants. But teasing apart a plant's water weight from its new biomass growth – and subsequently determining how efficiently each plant turned water into new tissue – required multiple steps that ultimately destroyed the plant. The researchers previously had to remove a plant from soil, weigh

it, then dehydrate it in an oven before weighing it again. They also employed a scanning instrument to individually measure the surface area of leaves, a step that required cutting each leaf off the plant.

In killing the plant, these methods kept the researchers from observing how it would have grown and developed afterward. Hyperspectral and RGB imaging not only address this issue, Ge said, but should also further speed the process of simultaneously comparing multiple traits among plant varieties.

"How to capture all of these dynamic traits is really a challenging task without these high-throughput systems," Ge said. "Now we can take daily images and put them together. We can analyze the growth rate and look at the changes over time at different developmental stages. I think that's the beauty of this phenotyping that wasn't really possible in the past."

Greenhouse to green acres

The team is also working with colleagues from the Department of Computer Science and Engineering to refine its image-analysis program. The hope is that it can eventually distinguish among components of a corn plant – individual leaves, stem segments, ears and more. Achieving that level of specificity might allow it to recognize and track those same components across changes in appearance and location, an important consideration for a plant that develops as quickly and dynamically as corn.

And in an effort to ensure their work will be useful to farmers, Ge and Schnable recently finished growing 140 corn hybrids from major seed companies in both the greenhouse and a research field that simulates the agricultural conditions of Nebraska farms. The researchers are currently analyzing their greenhouse data and comparing it with that from the field, aiming to glean insights on how well the former translates to the latter.

"That's a very important connection we need to make," Schnable said. "We don't want to just figure out how plants grow in a greenhouse. This data has to be relevant to field conditions. Hopefully we can build that into our models to some extent, so that we can make testable predictions in the greenhouse about what's going to happen in the field."

Ge and Schnable reported their recent findings in the journal Computers and Electronics in Agriculture (http://www.sciencedirect.com/science/article/pii/S0168169916305464). They authored the study with Geng "Frank" Bai, a postdoctoral researcher in biological systems engineering, and Vincent Stoerger, the plant phenotyping facilities manager at the Greenhouse Innovation Center.

The researchers received support from the Agricultural Research Division, housed within the university's Institute of Agriculture and Natural Resources (http://ianr.unl.edu/).

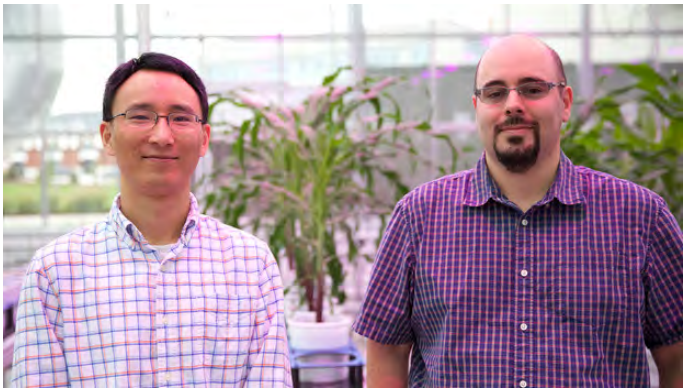
RELATED LINKS:

- Greenhouse Innovation Center (http://innovate.unl.edu/greenhouse-innovation-center)
- Read the study (http://www.sciencedirect.com/science/article/pii/S0168169916305464)
- Schnable Lab (http://www.schnablelab.org/)
- Yufeng Ge (http://engineering.unl.edu/bse/faculty/yufeng-ge-1/)

TAGS:

agriculture (/free-tags/agriculture/) phenotyping (/free-tags/phenotyping/) corn (/free-tags/corn/) hyperspectral (/free-tags/hyperspectral/) Yufeng Ge (/free-tags/yufeng-ge/) James Schnable (/free-tags/james-schnable/) Geng Bai (/free-tags/geng-bai/) Vincent Stoerger (/free-tags/vincent-stoerger/) Greenhouse Innovation Center (/free-tags/greenhouse-innovation-center/) Nebraska Innovation Campus (/free-tags/nebraska-innovation-campus/) Agronomy and Horticulture (/free-tags/agronomy-and-horticulture/)

Institute of Agriculture and Natural Resources (/free-tags/institute-agriculture-and-natural-resources/) biological systems engineering (/free-tags/biological-systems-engineering/) Computer Science and Engineering (/free-tags/computer-science-and-engineering/) Engineering (/free-tags/engineering/)



(https://news.unl.edu/sites/default/files/media/20160901-geandschnable.jpg)
Yufeng Ge (left) and James Schnable

Greenhouse Innovation Center

(https://news.unl.edu/sites/default/files/media/20160901-phenotyping.gif)
Images taken by a phenotyping system at the Greenhouse Innovation Center show the growth of a corn plant over the course of a month. (Click to play.)

UNIVERSITY OF NEBRASKA–LINCOLN

Institute of Agriculture and Natural Resources (http://ianr.unl.edu)

IANR NEWS

Schnable receives early career award in maize genetics

James Schnable

April 16, 2018

Lincoln, Neb. — James Schnable, an assistant professor in the Department of Agronomy and Horticulture and Center for Plant Science Innovation at the University of Nebraska–Lincoln, received the Marcus Rhoades Early Career Award in maize genetics at the 60th annual Maize Genetics Conference held in France on March 24.

“It is an honor to receive this award from the maize genetics community,” Schnable said. “It’s also quite special to be given an award named in honor of my academic great-grandmentor, Marcus Rhoades.” Rhoades was an American cytogeneticist who started his career in maize genetics working alongside future Noble Prize winners Barbara

McClintock and George Beadle. His research on maize spanned both basic genetics and advances in applied plant breeding.

Schnable was honored for his work delineating the functionality district subgenomes of maize and separate patterns of selection across during the domestication of maize, sorghum and foxtail millet.

At the same the conference, Michael Scanlon of Cornell University received the Lewis Stadler Mid-Career award and Robert Martienssen from the Howard Hughes Medical Institute was awarded the Barbara McClintock Prize for Plant Genetics and Genome Studies.

James Schnable
Assistant Professor
Agronomy & Horticulture
Center for Plant Science Innovation
402-472-3192
schnable@unl.edu (mailto:schnable@unl.edu)

Tags: People in the News (/tags/people-news)

Related Links

- Op-Eds (/op-eds)
- Jasa honored with excellence in conservation award (/jasa-honored-excellence-conservation-award)
- Volunteers needed for State Fair ice cream social (/volunteers-needed-state-fair-ice-cream-social)
- Williams receives national achievement award (/williams-receives-national-achievement-award)
- Husker entomology student receives FFAR fellow award (/husker-entomology-student-receives-ffar-fellow-award)

Interview with Yufeng Ge and James Schnable
By Breanna Jakubowski

SCIENTISTS' COLLABORATION

Developing technologies that can eventually 'alter the way we live.'

If you've ever driven through Nebraska in a dry year, you've seen the leaves of corn rolling up from too little moisture.

But how much water does corn actually need at each stage of growth, and can cultivars be developed that can thrive with less?

Researchers at the University of Nebraska–Lincoln are learning more about the growth rate and the percentage of water in the plant through phenotyping and through research facilities that carefully monitor growth, water and environment, recording it all with specialized photography and sensors.

Plant phenotyping research may make it possible to develop locally adapted crop varieties that produce higher yields. That could be especially important in geographic areas where weather or other conditions traditionally produce lower yields.

Scientists Yufeng Ge and James Schnable collaborate on research that measures and compares physical traits of corn, using a high-throughput phenotyping system in the Greenhouse Innovation Center at Nebraska Innovation Campus. On the surface, the collaboration seems ordinary, but it is complex – on several levels.

Ge is an assistant professor in the university's Department of Biological Systems Engineering; Schnable is an assistant professor in the Department of Agronomy and Horticulture. Ge, the engineer, works with technologies such as instrumentation, cameras and collection of informative images, while Schnable, the agronomist, works with the plant traits and genetics. Together, they are making progress toward higher, more consistent crop yields.

"These are the first steps in accelerating development of technologies that can alter the way we live," Schnable said. The work right now is making first advances so phenotyping equipment

becomes less expensive and easier to integrate into every farm in America.

Technological advances have increased speed and precision in terms of fertilizer applications on the farm and selecting plants for genetic potential, but until recently, the actual measurement of the plants slowed the plant breeding process. Now, high-throughput phenotyping offers the potential to make everything go faster, Schnable added.

The high-throughput phenotyping system is one of only a few in the United States. It uses sensors, robotics and computer technology, Ge said, rapidly measuring and comparing the physical traits of individual plants and capturing it all in high-content images.

The scientists conduct research on the university's test fields, located both right on campus and around Nebraska, but also in the Greenhouse Innovation Center. Climate can be controlled in each of the greenhouse's two sections, so a scientist can conduct research on plants' growth in a specific temperature or relative humidity. The greenhouse has 672 plant pots; each can hold one plant. There are three automated watering stations that measure how much water each plant used since it was last watered and can add individually specified amounts of water to each plant each day during a research experiment.

The greenhouse phenotyping facility has the capacity to capture tens of thousands of images a day in different forms, Ge said, with five kinds of sophisticated cameras: a consumer-grade camera with higher resolution; a fluorescent camera that measures the chlorophyll in the plant; a thermal infrared camera that measures the temperature of plant leaves and stems; a near-infrared camera that looks at plants' water content; and a hyperspectral camera that captures the reflective spectrum of a plant leaf. The hyperspectral camera also quantifies differences in the chemical composition of plants, including variation in water, pigment and cell wall composition.

The system stores all of those images in the computer and database.

'SEA OF DATA'

Thousands of photos are taken every day, representing an immense amount of data that is analyzed for conclusions about plant growth and water consumption.

"It is the responsibility of the researcher to go into that big sea of data, trying to figure out what you are looking for," Ge said.

Ge and Schnable collaborate with computer scientists and statisticians, who work with the data after the cameras capture the images. Even a simple experiment, Schnable said, can generate hundreds of gigabytes of data, with the possibility of growing to terabytes. The team of researchers is working on challenges of data transmission so others can easily access the data. Another challenge, Schnable said, is capturing the metadata, such as the corn plant's variety, genotype, growth environment and treatment.

After the image collection, they face the challenge of turning the pictures into numbers, which statisticians need to work with the data.

"After the engineering, we have lots of images of individual plants or plots, taken many times throughout the growing season," Schnable said. These images are added to the data collected with non-high-throughput phenotyping, such as bushels per acre in yield; height; flowering; climb and more, which all are numerical values.

"How you take an image with a cellphone camera or something more complex, like the hyperspectral photos, and reduce them to numerical values that are informative about the health or stress tolerance or yield potential of a corn, soybean or sorghum plant – that's one of the real analysis challenges," Schnable said.

INTERDISCIPLINARY COLLABORATION

The collaboration between the two scientists is rewarding, Ge said. It also is the research trend, driven by funding agencies that suggest a variety of perspectives when solving a challenge.

"I trained as an engineer and I always interacted with engineers, and we would talk about things like tractors and irrigation pivots and precision agriculture, remote sensing and instrumentation. But it wasn't until I joined th university and started working on the high-throughput phenotyping research that I realized it opened up a whole new arena," he said.

Schnable said that on a fundamental level, it means that "I bring the plants and he brings the sensors and computers" to the collaboration.

"I have an experiment where I'm growing plants in Mead (at the university's Agricultural Research and Development Center). I can go to Yufeng and he already has developed this wonderful sensor platform that he can roll through the field," Schnable said. "I can say, 'well, if you're developing this anyway, can you image these plants for me?'"

The university invests not only in the high-throughput phenotyping greenhouse at Nebraska Innovation Campus, but also in the phenotyping field facility at the ARDC in Mead, Ge said.

"Everyone understands just how vital this work is, and that is very powerful – both in terms of the resources the state invests in it, but also in the mindset of students you meet, the people you work with on campus. They understand just how important agriculture is," Schnable said.

TIMELY RESEARCH

Ge said the research team wants to find out whether hyperspectral imaging can non-invasively measure biochemical traits in the plant leaf, such as nitrogen, phosphorus and potassium. If it can be done, he said, farmers will benefit by having to apply only the amount of nutrients the plant needs, and no more.

Schnable said that in the next decade, there will be more pressure to apply only the specific amount of nitrogen or other fertilizers that are needed. If research can quantify plants' requirements more precisely, farmers can adopt the techniques that will save them money, he added.

"The grand challenge that we are facing here is that we wanted to produce sufficient food, fuel and fiber for the global population that is projected to exceed 9.7 or 10 billion by the year 2050. Everything we do fits into that grand challenge," Ge said. Using the phenotyping method, scientists are able to look at not only Nebraska, but around the United States and potentially to developing countries like Africa, India or China.

"That is the key to phenotyping; you want to develop crop cultivars that can survive in different environments," Ge said.

UNIVERSITY OF NEBRASKA–LINCOLN

Institute of Agriculture and Natural Resources (http://ianr.unl.edu)

IANR NEWS

Husker research to explore the emergence of specialized body parts, plants

Tripsacum, a member of the most closely related genus to maize produces only one type of flower structure, with male flowers at the top and female flowers at the base. (Lang Yan, Schnable Lab)

August 21, 2018

Lincoln, Neb. — New research conducted by the University of Nebraska–Lincoln's James Schnable will use corn to test the idea that the emergence of specialized body parts occurs through whole genome duplication.

A gene consists of enough DNA to code one protein, and a genome is the sum total of an organism's DNA. Schnable, an assistant professor in the Department of Agronomy and Horticulture and Center for Plant Science Innovation, has

earned a 2-year, $299,801 grant from the National Science Foundation for a project to prove that certain genes are not available for individual duplication, and can only be replicated through whole genome duplication.

“Transcription factors are genes which control when other genes are turned off and on. Because cells are very sensitive to quantities of these transcription factors present, duplicating a single transcription factor gene can throw things out of balance, usually with negative consequences for the plant or animal in question. In a whole genome duplication all the transcription factors and all the genes they regulate are duplicated at the same time which avoids many of the problems caused by single gene duplication.”

In both plants and animals, the emergence of new specialized body parts, such as floral organs in plants or the multiple specialized types of teeth found in heterodont animals, is a rare process. Generally these specialized organs appear to originate as specialized versions of existing organs, yet their specialization requires a separation in regulation between different copies of the same organ.

The research could explain why the emergence of new, specialized body parts is so rare. Most apparent links between whole genome duplication and an increase in separately defined body parts, such as the 4-footed animal or flowering plants, occurred hundreds of millions of years ago.

Schnable’s team has focused this research on corn, as it produces two specialized flower heads for male and female reproduction, while all other related plants only produce a single kind of head. Researchers will test the link between a whole genome duplication in the corn lineage and the evolution of its distinct flowering heads. Reverse genetic techniques will be applied to both corn and sorghum, which lacks two flower heads,

to prove that this emergence required whole genome duplication, which created duplicate copies of many genes which rarely duplicate through other processes.

“We could potentially predict when the emergence of new specialized body parts is more likely, and provide initial insights into how specialized body parts could be engineered in future synthetic biology efforts,” Schnable said.

Research in this area also has the potential to guide the development of new engineered varieties of crop plants with multiple specialized leaf types. Engineering leaves high in the canopy to have fewer chloroplasts, would allow light to penetrate deeper leaves where lower wind speeds and higher humidity reduce the transpirational water cost of photosynthesis.

The research project will provide valuable training on conducting science at the intersection of genomics, genetics and phenomics to multiple graduate and undergraduate students.

To learn more about Schnable’s research, visit http://schnablelab.org/ (http://schnablelab.org/).

James Schnable
Assistant Professor
Center for Plant Science Innovation
Department of Agronomy and Horticulture
402-472-3192
schnable@unl.edu (mailto:schnable@unl.edu)

Tags: ARD (/tags/ard)
Crops (/crops)

Related Links

- Husker research to explore the emergence of specialized body parts, plants (/husker-research-explore-emergence-specialized-body-parts-plants)
- Nebraska partners to fight antimicrobial resistance (/nebraska-partners-fight-antimicrobial-resistance)
- Husker-led research team to examine irrigation's role in precipitation (/husker-led-research-team-examine-irrigations-role-precipitation)
- Study: Food fraud spoils value for all (/study-food-fraud-spoils-value-all)

- 115 years of data reveal longer U.S. growing season, temp trends (/115-years-data-reveal-longer-us-growing-season-temp-trends)

Five Recent and Significant Manuscripts

CellPress
PARTNER JOURNAL

Molecular Plant
Resource Article

STAG-CNS: An Order-Aware Conserved Noncoding Sequences Discovery Tool for Arbitrary Numbers of Species

Xianjun Lai[1,3,4], Sairam Behera[2,4], Zhikai Liang[1], Yanli Lu[3], Jitender S. Deogun[2,*] and James C. Schnable[1,*]

[1]Department of Agronomy and Horticulture, Center for Plant Science Innovation, University of Nebraska-Lincoln, Lincoln, NE 68588, USA

[2]Department of Computer Science and Engineering, University of Nebraska-Lincoln, Lincoln, NE 68588, USA

[3]Maize Research Institute, Sichuan Agricultural University, Chengdu 611130, China

[4]These authors contributed equally to this article.

***Correspondence: Jitender S. Deogun (deogun@cse.unl.edu), James C. Schnable (schnable@unl.edu)**

http://dx.doi.org/10.1016/j.molp.2017.05.010

ABSTRACT

One method for identifying noncoding regulatory regions of a genome is to quantify rates of divergence between related species, as functional sequence will generally diverge more slowly. Most approaches to identifying these conserved noncoding sequences (CNSs) based on alignment have had relatively large minimum sequence lengths ($\geq$15 bp) compared with the average length of known transcription factor binding sites. To circumvent this constraint, STAG-CNS that can simultaneously integrate the data from the promoters of conserved orthologous genes in three or more species was developed. Using the data from up to six grass species made it possible to identify conserved sequences as short as 9 bp with false discovery rate $\leq$0.05. These CNSs exhibit greater overlap with open chromatin regions identified using DNase I hypersensitivity assays, and are enriched in the promoters of genes involved in transcriptional regulation. STAG-CNS was further employed to characterize loss of conserved noncoding sequences associated with retained duplicate genes from the ancient maize polyploidy. Genes with fewer retained CNSs show lower overall expression, although this bias is more apparent in samples of complex organ systems containing many cell types, suggesting that CNS loss may correspond to a reduced number of expression contexts rather than lower expression levels across the entire ancestral expression domain.

Key words: conserved noncoding sequence, comparative genomics, suffix tree, longest path algorithm, grain crops

Lai X., Behera S., Liang Z., Lu Y., Deogun J.S., and Schnable J.C. (2017). STAG-CNS: An Order-Aware Conserved Noncoding Sequences Discovery Tool for Arbitrary Numbers of Species. Mol. Plant. **10**, 990–999.

INTRODUCTION

Mutations accumulate in different parts of the genome at different rates. Protein-coding exons tend to have significantly lower rates of nucleotide substitutions than introns or intergenic sequences. The majority of possible substitutions in protein-coding sequences usually change the amino acid sequences of the resulting protein. Many such changes in the amino acid sequence will have negative effects on the protein. Therefore, many mutations which occur in the protein-coding sequence are purged from the genome by purifying selection. Protein-coding sequences can be said to be functionally constrained, resulting in nucleotide substations accumulating more slowly. In both animals and plants, there are islands of noncoding sequences which also exhibit low rates of nucleotide substitution, indicating that these regions are also subject to functional constraint (Hardison et al., 1997; Levy et al., 2001; Kaplinsky et al., 2002; Guo and Moose, 2003). There regions are referred to as conserved noncoding sequences (CNSs) (Hardison et al., 1997), or sometimes in the animal literature as conserved noncoding elements (CNEs) (Shin et al., 2005). In both animals and plants, CNSs have been shown to confer extremely specific spatiotemporal patterns of transcriptional regulation (Shin et al., 2005; Visel et al., 2008; Raatz et al., 2011).

Identification of CNSs in animals and plants presents very different challenges. In animals, many CNSs are large ($\geq$100 bp) (Stephen

Published by the Molecular Plant Shanghai Editorial Office in association with Cell Press, an imprint of Elsevier Inc., on behalf of CSPB and IPPE, SIBS, CAS.

et al., 2008), while different analyses in plants have primarily identified smaller (15–50 bp) CNSs (Thomas et al., 2007; Baxter et al., 2012; Haudry et al., 2013; Turco et al., 2013). Various algorithmic approaches are currently employed to identify CNSs in plants including both manual (Thomas et al., 2007) and automated (Turco et al., 2013) curation of BLASTN results, global alignments of sliding windows of promoter regions (Baxter et al., 2012), the use of whole-genome alignment algorithms (Haudry et al., 2013), alignment-independent detection of enriched IUPAC motifs (De Witte et al., 2015), and searches for biologically defined transcription factor binding sites across orthologous genes (Van de Velde et al., 2016). The majority of these approaches are either based on two-at-a-time sequence alignments (Baxter et al., 2012; Haudry et al., 2013; Turco et al., 2013) and/or do not retain information on conserved microsynteny within the promoter (De Witte et al., 2015; Van de Velde et al., 2016).

Here we describe a new method for CNS detection using suffix tree and maximum-flow algorithms (see Methods), which identifies sets of sequences conserved in the same order in the promoters of arbitrarily large numbers of orthologous or paralogous genes. This approach is inherently non-pairwise, allowing researchers to adjust the number of genes compared and significance cutoffs for CNS discovery based on the tolerance of their research program to either false positives or false negatives. The analysis described herein demonstrates that as greater numbers of species are used in the comparison, CNSs with smaller sizes can be identified while retaining equivalent or better false positive discovery rates.

RESULTS

A New Approach for Fast and Accurate CNS Identification

A comparative genomics approach combining data on conserved sequences and conserved order was employed identify CNSs within the promoters of syntenic orthologous genes drawn from two to six grass species. This approach identifies conserved regions, i.e., maximal exact matches (MEMs), present in all sequences being analyzed, and then searches for and locates the optimal path including the greatest number of conserved sequences without violations of microsynteny using a weighted acyclic graph (Figure 1A). This approach provides flexibility to identify CNS using data from different numbers of species, or multiple paralogous gene copies from a single family, and is runtime efficient, only taking several seconds to complete the computation for one group of orthologous or paralogous genes and a few hours to complete the computation for all syntenic orthologous genes present within a group of species. The software tool (STAG-CNS: Suffix Tree Arbitrary Gene number: Conserved Noncoding Sequence) and its source code are now publicly available at https://github.com/srbehera11/stag-cns.

Estimating the Accuracy and Sensitivity of STAG-CNS

STAG-CNS has a configurable minimum CNS length that allows the users to balance the trade-off between increasing sensitivity to small conserved sequences and controlling false positive discovery rates. The probability of the same short sequence existing in multiple sequences by chance alone is dependent on the number of sequences being compared, the length of each comparator sequence, and the minimum length of the matching sequence. Assuming a random ordering of four nucleotides at equal frequencies, this probability can be approximated by the formula: $1 - ((1 - {}^1/_4{}^n)^{L-n+1})^{L-n+1}$, where L represents the length of DNA sequence and n is the length of small sequence fragment (Figure 2A). However, the assumptions given above are violated by most genomes. Among 239 monocot species, GC content was found to vary between 33.6% and 49.9% (Šmarda et al., 2014). In grasses, individual genes exhibit binomial distributions of gene content (Tatarinova et al., 2010). In addition, the frequency of individual short sequences is nonrandom within a given genome, with certain strings never or rarely occurring, and others, particularly those found in MITEs and other transposons, occurring at high frequency. Therefore, rather than approximating the false positive discovery rate for STAG-CNS using the formula described above, it was instead estimated using permutation testing, whereby the number of putative conserved noncoding sequences identified in comparisons of nonorthologous genes was assayed following the method described by Baxter et al. (2012). This method allows the minimum number of CNSs that can be detected at an acceptable false positive rate to be determined empirically. A set of 200 genes (Supplemental Table 1) conserved as syntenic orthologous located in sorghum, rice, setaria, brachypodium, oropetium, and dichanthelium were randomly selected from a previously published syntenic gene list (Schnable et al., 2016). Conserved noncoding sequences were identified between syntenic orthologous genes in three species, sorghum, rice, and setaria, using minimum CNS lengths between 8 and 22 bp. Average false positive discovery rates per gene were estimated using 100 random permutations of the dataset (Figure 2B). As expected, both true positives and false positives declined as minimum CNS lengths increased, while the percentage of all identified CNSs predicted to be true positives increased. Controlling the false positive discovery rate at 5% of total discovered CNSs indicated that a CNS as short as 12 bp can be identified with high confidence using data from three species.

To determine how minimum CNS length responded to variation in the number of species employed in the comparison, CNSs were identified between the same set of 200 genes employed above (Supplemental Table 1) using data from syntenic gene copies in 2, 4, 5, or 6 species, with minimum CNS lengths from 8 to 22 bp (Figure 3A). The corresponding false positive discovery rates are displayed in Figure 3B. As expected, increasing the number of species included in the analysis improves the proportion of true positive CNSs identified at any given minimum CNS length, and with six species it was possible to identify conserved sequences as short as 9 bp while maintaining a false discovery rate (FDR) of less than 5%.

Comparison of STAG-CNS and CDP

The CNS Discovery Pipeline (CDP) is one of the tools previously used to identify conserved noncoding sequences among different grass species (Turco et al., 2013). Unlike STAG-CNS, the CDP works by performing pairwise comparisons based on BLASTN, and identifies CNSs present in three or more species through overlap with a single common reference. The CNSs

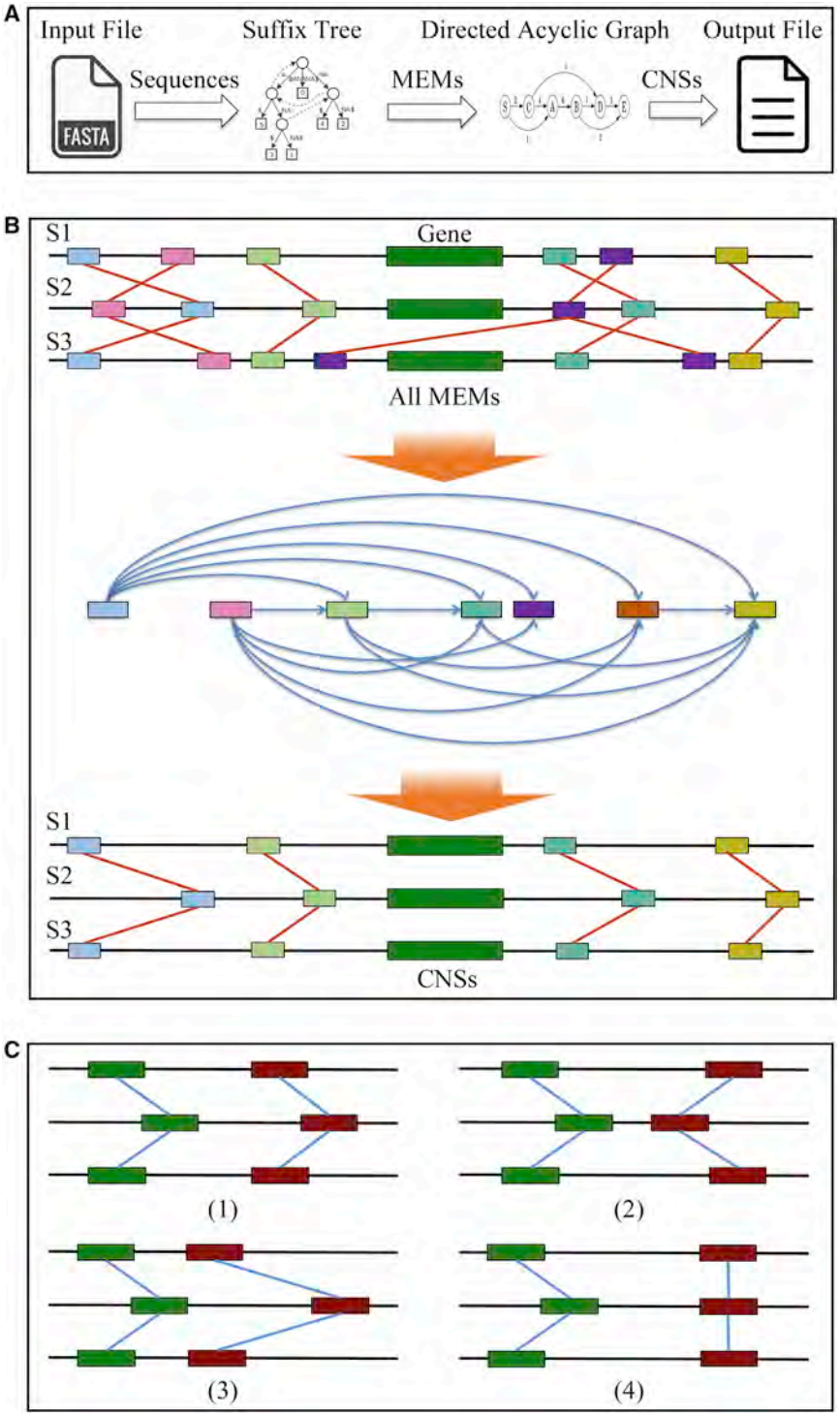


Figure 1. The Work Flow for Identifying CNSs across Grass Species Using STAG-CNS.

(A) CNS discovery strategy using suffix tree and maximum-flow algorithms.

(B) CNS of three species and directed acyclic graph using maximal exact matches (MEMs).

(C) Breaking ties: choose the one with highest rank. The ranking of the scenario above is (1) > (3) > (4) > (2) (see Methods).

were identified in 200 genes conserved in sorghum, setaria, and rice using the CDP and STAG-CNS for direct comparisons of the results produced by both methods. A set of genes previously shown to be CNS rich were selected to maximize the number of informative comparisons (Supplemental Table 3). The CDP was run with the default minimum CNS length of 15 bp for sorghum–rice and sorghum–setaria comparisons. A CNS was considered to be present across all three species if there was at least a 12-bp overlap in sorghum between a sorghum–rice CNS and sorghum–setaria CNS. The number and total length of three-species CNSs identified for each syntenic orthologous gene group was compared between the two methods and found to be moderately correlated. The Pearson correlation coefficient between the results from the CDP and STAG-CNS was 0.415 (p value = 1.01e-09) for the number of CNSs identified per gene and 0.525 (p value = 1.332e-15) for the total length of CNS sequence identified per gene (Supplemental Figure 1).

Subsequently, all of the 17 996 orthologous syntenic genes of three species (sorghum, setaria, and rice) (Supplemental Table 2) were used to identify the CNS using STAG-CNS and CDP (sorghum and rice, sorghum and setaria, and the pan-grass species). The CNS information from both methods was summarized (Table 1) and the orthologous CNSs of the syntenic genes of three species listed (Supplemental Table 4).

The CDP uses nucleotide–nucleotide BLAST (blastn) as its core aligner, allowing it to identify longer sequences with multiple mismatches or gaps in CNSs, while STAG-CNS currently requires exact matches. The mismatch rates for each pairwise CNS identified by the CDP were 9.3% between sorghum and rice and 11.6% between sorghum and setaria, suggesting one mismatch for each 10-bp sequence on average using the CDP method. As a result, STAG-CNS identified fewer CNSs per gene in this three-way comparison. Across the 17 996 syntenic gene triplets employed in this analysis, STAG-CNS identified an average of 0.58 CNS per gene while the CDP identified an average of 1.25 CNSs per gene. Both the average length and median length of CNSs identified by STAG-CNS were smaller than those calculated by CDP (Table 1). Overall, the total length of the CNS of syntenic genes identified by CDP and STAG-CNS were 774.2 kb and 162.3 kb, respectively.

The CNSs identified by each method were manually proofed for a number of individual syntenic genes. One example

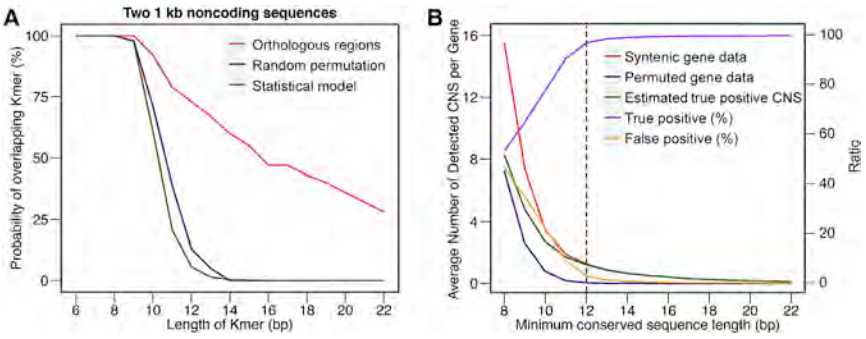


Figure 2. Relationship of Minimum Conserved Sequence Length and Statistical Power.

(A) The number of overlapping sequences of a given length between two noncoding sequences based on either a statistical model assuming random sequence and equal frequencies of all four nucleotides, random regions extracted from actual grass genomes, or syntenic orthologous noncoding regions extracted from grass genomes.

(B) Relationship between the minimum length of shared subsequence before it is considered a CNS, number of CNSs discovered, and false discovery percentage.

(Sobic.004G325100) is shown in Figure 4. The CDP identified a total of 12 CNSs between sorghum and rice and eight CNSs between sorghum and setaria. Comparing all three species simultaneously using STAG-CNS identified 13 short CNSs mostly representing more precise regions within larger CNSS identified in pairwise species analysis. Including data from syntenic orthologs in all six grass species employed in this analysis increased the number of CNSs identified by STAG-CNS to 18. As an additional validation step, sequences identified by STAG-CNS were compared with a functionally characterized noncoding regulatory sequence associated with the classical maize gene *knotted1* (Greene et al., 1994; Inada et al., 2003). Using a minimum CNS length of 12 bp and comparison across sorghum, setaria, and rice, six CNSs were identified in the third intron of *knotted1*, which largely corresponded to the functional regulator region defined through characterization of transposon insertion alleles (Supplemental Figure 2) (Greene et al., 1994).

Association of CNSs with DNase I Hypersensitive Sites

Regulatory sequences are often correlated with regions of open or accessible chromatin (Tsompana and Buck, 2014; Rodgers-Melnick et al., 2016). Open chromatin can be assayed in a whole-genome fashion using a range of techniques including FAIRE-seq, MNase-seq, and DNase I hypersensitivity-seq (Zhang et al., 2012; Rodgers-Melnick et al., 2016). The overlap between CNS identified by STAG-CNS and open chromatin regions was tested using a pre-existing set of DNase I hypersensitive sites (DH sites) generated from rice seedling and callus tissue (Zhang et al., 2012). A total of 8934 CNSs identified from the syntenic genes in sorghum, rice, and setaria with minimum CNS of 12 bp were compared with the DH sites (Figure 5 and Supplemental Table 5).

Of these CNSs, 34.0% (3037) and 58.1% (5190) overlapped with DH sites identified in seeding and callus tissues, respectively, significantly more than would be expected from random sequence (binomial test p value < 1.0e-06 in both cases). For example, in callus tissue, total length of accumulative DH sites accounts for 8% of rice genome and the overlap between STAG-CNS and DH sites was enriched ~7.26-fold, more than expected. This observation is potentially confounding by the enrichment of both open chromatin and conserved promoter sequences around transcription start sites. After excluding DH sites and CNSs within 1000 bp of annotated transcription start sites, 5591 CNSs remained. Of these sequences 1130 and 2289 overlapped with DH sites identified in seedling and callus tissues, respectively, which continued to show a significant enrichment (binomial test p value < 1.0e-06 in both cases). The overall percentage of the remaining portion of the rice genome and DH sites in callus tissue drops from 8% to 5.8%. However, the enrichment of overlap between STAG-CNS and DH sites actually decreases slightly from ~7.26-fold enrichment to ~7-fold enrichment. These overlaps are also significantly higher than those previously observed when comparing CNSs identified by the CDP between rice and sorghum with the same rice open chromatin datasets (25.7% and 41.6% for seedling and callus tissue, respectively) (Zhang et al., 2012).

A set of 1873 rice genes with conserved syntenic orthologs across all six species used in this analysis were used to test how the relative overlap between STAG-CNS-identified sequences and DH open chromatin regions responded to variation in the number of species employed to identify CNS and minimum CNS length. The overlap between potential regulatory sites identified using the STAG-CNS and potential regulatory sites identified using DNase1 hypersensitivity-seq increases either when the number of species used in the STAG-CNS analysis is increased or when the minimum length of the CNS is increased (Figure 3C and 3D).

Functional Enrichments among CNS-Rich Genes

Previous studies have shown that genes with regulatory functions tend to be associated with greater numbers of CNSs (Freeling and Subramaniam, 2009; Turco et al., 2013) and that genes with more complex regulatory patterns tend to have larger promoters in both animals (Nelson et al., 2004) and plants (Sun et al., 2010). Here, genes were grouped based on the number of CNS for each syntenic gene (0, 1, 2, or ≥3). Gene ontology (GO) enrichment analysis was performed on each group of syntenic genes independently (Supplemental Table 6). For the genes present in the group with no CNS, nine GO terms related to "metabolic process," "catalytic activity," "single-organism metabolic," "oxidoreductase activity," etc. were identified as enriched (Bonferroni test p value < 0.05, FDR < 0.05). In the group of genes associated with one CNS, 20 GO terms were significantly enriched (Bonferroni test p value < 0.05, FDR < 0.05); among these terms 17 were related to "regulation of," two were related to "transcription factor activity," and one was related to "biological regulation." Similarly, 22 GO terms were significantly enriched in the group of genes with two CNSs (Bonferroni test p value < 0.05, FDR < 0.05), including all 20 GO terms identified as enriched among genes with one CNS and

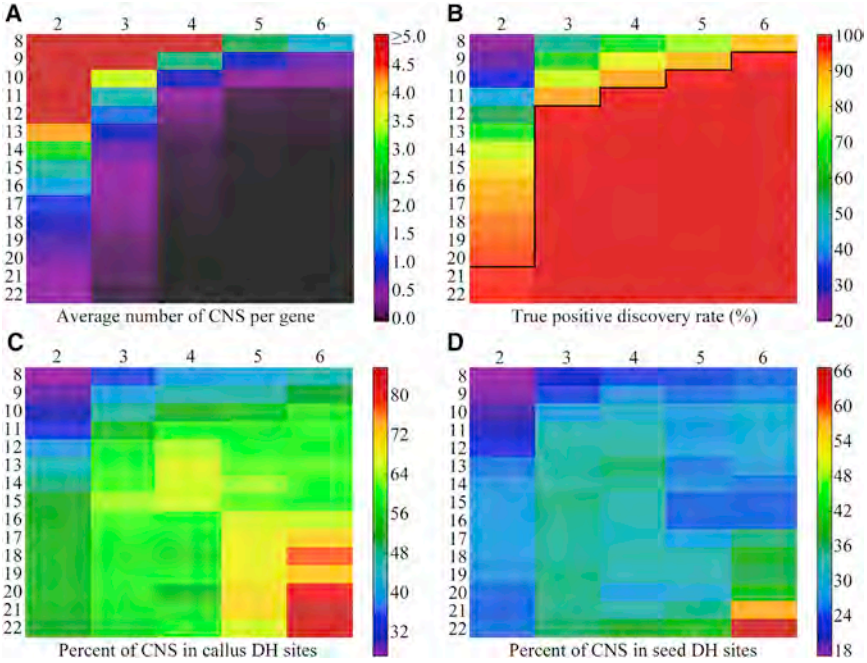


Figure 3. Analysis Showing that Both Increasing the Number of Species and Increasing the Minimum Length Can Increase the Power of STAG-CNS to Detect the CNS.

For each sub-figure the *x*-axis shows the species number analyzed and the *y*-axis shows the minimum length used to identify CNSs.

(A) The number of CNSs identified in a different number of species and the minimum length of CNSs.

(B) True positive discovery rate of CNSs across a different number of species. The true positive rates below the black line indicate reliable minimum length of CNSs (true positive rates ≥95%).

(C and **D)** Overlap rate of CNS and DNase sites in callus **(C)** and seed **(D)**.

two others related to "DNA binding" and "nucleic acid binding." Among the genes with at least three CNSs, a total of 32 GO terms were significantly enriched (Bonferroni test p value < 0.05, FDR < 0.05). Among these GO terms, 18 were related to "regulation" and 10 to "binding." Overall these findings are consistent with previous reports using CNSs identified by other methods.

A significant proportion of syntenic genes were not associated with any GO term annotations (6121 genes). The distribution of these unannotated genes across the four categories of CNS richness described above was tested using a "dummy" GO term not associated with any genes. Significantly fewer genes with no GO annotations were present in the 0 CNS/gene category than expected, and significantly more genes without any GO terms were present in the 1 CNS/gene category. No significant difference from the null model was observed for these genes in the 2 CNSs/gene category and ≥3 CNSs/gene category. However, these were also the two categories with the fewest total genes, reducing statistical power to detect significant enrichment or purification of these genes.

Biased Regulatory Sequence Loss Tracks Biased Gene Loss between Maize Subgenomes

STAG-CNS can also be used to identify differences in the loss or retention of conserved noncoding sequences between duplicated genes. Maize experienced a whole-genome duplication (WGD) within the last 5–12 million years, after its divergence from the lineage leading to sorghum (Swigoňová et al., 2004) forming two subgenomes (maize I and maize II) (Schnable et al., 2011). Following the WGD, duplicate copies of many genes were deleted from one or the other subgenome; however, 4000–7000 duplicate gene pairs are retained on both maize subgenomes today. Gene copies were more likely to be lost from the maize II subgenome, while gene copies on the maize I subgenome tend to be expressed at higher levels than their retained duplicates on the maize II subgenome (Schnable et al., 2011). Unlike other forms of gene duplication, both copies of a WGD-derived gene pair are initially associated with equivalent sets of regulatory sequences, although these sequences can be lost from one or the other gene copy, a process known as fractionation mutagenesis (Freeling et al., 2012).

A set of 6156 syntenic gene groups including gene copies in setaria, sorghum, maize I, and maize II were used in the following analysis. Using the same sorghum and setaria gene, CNSs shared by both retained duplicate maize genes, as well as sequences conserved in both outgroup species but retained by only one maize gene copy, were identified using STAG-CNS (Supplemental Figure 3A and Supplemental Table 8). In approximately half the cases (2925 gene groups), no CNSs were found to be associated with either maize gene copy. In an additional 297 cases, equal numbers of CNSs were retained by each gene copy. However, of the remaining 2934 cases, approximately two-thirds (1925 gene groups) showed greater amounts of conserved noncoding sequence associated the maize I copy of a retained duplicate gene pair, and one-third (1010 gene groups) showed greater amounts of conserved noncoding sequence associated with the maize II gene copy. Bias in CNS loss correlated with bias in expression, with greater bias toward expression of the maize I gene copy in cases where the maize I gene copy retained more CNSs, and less or no bias toward greater expression of the maize I gene copy when the maize II gene copy retained more CNSs. The strength of this effect was quantified in a range of gene expression datasets generated by multiple research groups (Wang et al., 2009; Li et al., 2010; Davidson et al., 2011; Waters et al., 2011; Bolduc et al., 2012; Chang et al., 2012; Chettoor et al., 2014; Zhang et al., 2017). Supplemental Figure 3B shows the difference in the magnitude of biased expression toward maize I between the group of genes where maize I retained more CNSs and the group of genes where maize II retained more CNSs. Pollen and anthers showed some of the smallest effects of relative CNS number on bias toward maize I gene copy expression, while whole leaf/whole seedling/whole root samples showed some the greatest

	Software		
	CDP	CDP	STAG-CNS
CNS data	Os-Sb (15 bp)	Pan-grass (12 bp)	Sb-Si-Os (12 bp)
Total no. of orthologous CNSs	46 005	22 510	10 498
Percent of syntenic genes with at least 1 CNS	72.47% (13 041)	41.67% (7499)	27.61% (4968)
Average no. of CNSs	2.56 CNS/gene	1.25 CNS/gene	0.58 CNS/gene
Mean length of CNS (bp)	32.77	34.39	32.66
Median length of CNS (bp)	24.00	25.00	18.00
Total quantity of conserved noncoding sequence (bp)	1 507 726	774 177	162 250

Table 1. Summary of CNS Distributions in 17 996 Syntenic Orthologous Genes.

effects of relative CNS number on bias toward maize I gene copy expression.

DISCUSSION

We developed an algorithm and software implementation, STAG-CNS, for the identification of conserved noncoding sequences distributed in the noncoding regions flanking conserved syntenic orthologous genes across multiple species. The proposed approach employs the suffix tree and maximum-flow algorithm to identify CNSs using both sequence conservation constraints and conserved microsynteny constraints in each species. Unlike previous CNS identification tools based on sequence alignment, STAG-CNS can directly align the promoters from three or more species simultaneously. This flexibility makes it possible to identify shorter, but highly conserved, sequences that could not be identified at acceptable FDRs using two-at-a-time approaches to CNS discovery (Figure 3B). Effective use of STAG-CNS requires tuning the minimum CNS length. This will vary based on a number of factors including the number of species included in the analysis, the phylogenetic distance between these species, and the total amount of flanking sequence included on either side of target genes. For these reasons, the optimal solution for any gene group of species may be to test the false discovery proportions at different minimum CNS lengths as described above, before choosing a cutoff for the final analysis. However, many individual users may not have the time or computational resources to perform optimization for individual projects. In the absence of tuning, our implementation of the STAG-CNS algorithm will use default cutoffs based on the threshold values for different numbers of species being compared, defined in Figure 3B. Because transcription factor binding sites are small, STAG-CNS depends on the presence of some type of anchor sequence, such as a conserved syntenic orthologous gene group, to define workably small genome regions to compare between species. However, the use of this tool is not confined only to promixal promoters, but can also be employed to compare intron sequences across multiple species (Supplemental Figure 2) or to compare deep intergenic regions if other methods are used to identify specific regions of the genome for comparison across species.

Known transcription factor binding sites range in length from 6 to 15 bp with an average length of 10 bp (Stewart et al., 2012; Tuğrul et al., 2015; Yu et al., 2016), so the advance from identifying CNSs $\geq$15 bp to identify conserved sequences as short as 9 bp with acceptable ($\leq$0.05) FDRs means that it should be possible to identify most of the conserved binding sites of a wider range of transcription factors using CNS analysis. As the number of species with high-quality genome sequences in groups such as the grasses, crucifers, and legumes increases, it is anticipated that it will be possible to employ STAG-CNS to identify even smaller regions of conserved sequence within gene promoters with acceptable false discovery proportions.

Many CNSs identified with previous methods contained mismatches and indels (about 10% of base pair positions for CDP CNSs). It is likely that many of these mismatches do not disrupt the function of the conserved noncoding region. Currently STAG-CNS can only identify exact match sequences. However, there are precedents in the literature for adapting suffix tree-based algorithms to identify conserved sequences containing up to one mismatch without unmanageable increases in runtime or memory requirements. This can be achieved by extending the longest common substring with *k* mismatch problem for more than two sequences. Crochemore et al. (2006) used suffix tree and reverse suffix tree, i.e., suffix tree of sequences in reverse, to find the longest repeats with *k* consecutive mismatches among two strings. A modified version of Crochemore's algorithm was developed by Flouri et al. (2015) for finding the longest common substring with one mismatch. This method could be extended for more than two sequences, as employed in STAG-CNS, by using a generalized suffix tree and a reverse generalized suffix tree.

Finally, the ability to analyze the promoter sequence surrounding three or more genes at once makes it possible to more accurately study the differential loss of conserved sequences from duplicate genes following WGD. Here STAG-CNS was used to demonstrate that the maize subgenome which lost more genes following WGD also has less conserved regulatory sequence associated with retained copies of duplicate gene pairs. Several recent reports have suggested that duplicate maize genes have experienced significant regulatory sub- or neofunctionalization (Hughes et al., 2014; Pophaly and Tellier, 2015; Li et al., 2016). The findings here and elsewhere suggest that genes from the maize II subgenome may have subfunctionalized into more specialized expression domains while maize I gene copies retained broader patterns of expression. This model would be broadly consistent with the findings of Pophaly and Tellier (2015) that a large population of maize WGD gene pairs exhibit bidirectional expression divergence when

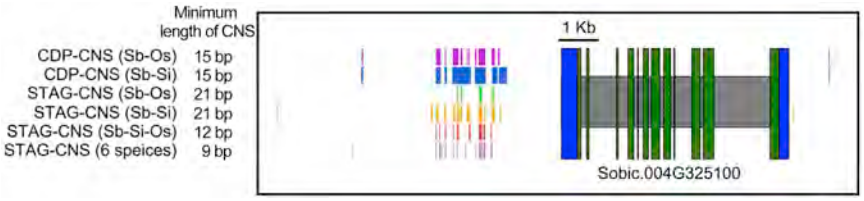


Figure 4. Visual Comparison of CNSs Identified for a Given CNS-Rich Gene Using the CDP or STAG-CNS.

CNSs between Sb-Os and Sb-Si were identified using CDP with default parameters (minimum length of CNS: 15 bp). CNSs in Sb-Si-Os were identified using STAG-CNS with a minimum CNS length of 12 bp. CNSs in six species (Sb-Si-Os-Or-Br-Do) were identified using STAG-CNS with minimum CNS length of 9 bp. The blocks in blue, green, and gray correspond to untranslated regions, exons, and introns, respectively.

examining expression patterns across specific tissue types, and that these genes are enriched in transcription factors, a class of gene that tends to have large promoters containing many conserved regulatory elements.

METHODS

STAG-CNS Algorithm and Implementation

The STAG-CNS uses suffix tree and maximum-flow algorithm to discover the CNSs in multiple grass species. Repeated subsequences, i.e., MEMs, are identified using suffix tree and then the microsyntenic path through the target sequences including the greatest number of base pairs of repeated subsequences is identified using maximum-flow algorithm. Figure 1A shows the overview of the algorithm to detect CNSs. The repeated subsequences, i.e., MEMs produced from suffix tree of sequences, are used to construct a weighted directed acyclic graph (DAG) and then the maximum-flow algorithm is used on DAG to generate the CNSs. The input for STAG-CNS is a file containing fasta sequences of the surrounding regions, i.e., 10 kb upstream and downstream of each gene to be compared, with additional information such as gene name, start and end position of the gene, chromosome name, direction ("+" or "−"), and the actual start and end positions in the chromosome. All of this information is extracted from the annotation files of corresponding species. If the direction is "−", i.e., the gene of interest is on the reverse strand, the reverse complement of sequence is constructed prior to analysis. The STAG-CNS can be used to find CNS located in nonpromoter regions by modifying the input file. For example, to find the CNSs in the intron region, the start and end positions of exons should be provided in the input file. The generalized suffix tree of all sequences is constructed by using Ukkonen's linear time online algorithm (Ukkonen, 1995). The generalized suffix tree is a tree data structure that stores all suffixes of the sequences in a compressed manner. The suffix of a string is a substring which ends at the last position of the string. The leaf nodes of the generalized suffix tree are labeled with the index of the sequence and the start position of the suffix in that sequence. This data structure has been used in a variety of applications in computational biology including search applications, exact match searches, subsequence composition searches, homology searches, single-sequence analysis applications, and multiple-sequence analysis applications (Bieganski et al., 1994). The MEM is a repeat subsequence that cannot be extended at both the ends. The generalized suffix tree is used to find the MEMs efficiently by traversing the tree and using some additional information associated with suffix tree data structure such as suffix links. The suffix tree was first studied by Gusfield (1997) and was also used for multiple-sequence alignment in the MGA tool (Höhl et al., 2002). Our algorithm uses the similar approach to find all MEMs with length greater than or equal to the given minimum length (input parameter) using the generalized suffix tree. The implementation of this part of the algorithm (finding MEMs from generalized suffix tree) is similar to splitMEM (Marcus et al., 2014). The first figure in Figure 1B shows an example of MEMs present in three sequences. The blocks of same color represent the matching subsequences. Instead of obtaining all MEMs, the program only finds the MEMs which are (a) present in all sequences, (b) not present in genomic regions, (c) present in the same side of the gene. A weighted DAG is constructed using the MEMs obtained in the previous stage. The middle part of Figure 1B shows the DAG obtained from the MEMs. The MEM with light-blue color intersects with the MEM with the pink color, so there is no edge between them. The MEM with light-blue color does not intersect with all other MEMs, so there are directed edges from that MEM to all other MEMs except the pink one. Each MEM becomes a node in the graph and the weight of the node is the length of that MEM. If two MEMs are not overlapping or intersecting with each other, a directed edge is constructed between them. For each node, the weight of each incoming edge is equal to the node weight. The CNSs are extracted from the graph using the approach similar to the longest path algorithm (Ma and Deogun, 2010). Here we use a maximum-flow algorithm instead of the longest path for the weighted DAG to find the path with maximum weight. The algorithm finds the optimal path consisting of CNSs that gives the maximum cumulative score (shown in the bottom part of Figure 1B). The ties are broken based on the distance between consecutive CNSs. If two or more CNS sets have same maximum score, the set with minimum intra-CNS distance is chosen. In Figure 1C, the first set of CNSs is chosen among the four CNS sets having equal score. The CNS that is selected among four have min $\{|d_1 - d_2| + |d_1 - d_3| + |d_2 - d_3|\}$, where d_1, d_2, and d_3 are the distances between MEM with the green color and MEM with maroon color in the sequences 1, 2, and 3, respectively. Basically, the CNSs with less variation in the distance between two of its consecutive MEMs is ranked higher.

The current implementation of the STAG-CNS algorithm has been tested using between 3 and 15 sequences and for sequence lengths between 2 kb and 50 kb. It takes little less than an hour to generate the CNS of 10 sequences, each 25 kb long, using 10 Gb of memory. The memory usage may increase when the total length of the sequences becomes more than 500 kb. The output of STAG-CNS is a file containing the list of CNSs with their start and end positions, and length and name of the chromosome. It also produces the visualization file that can be used in the Gobe visualization tool (Pedersen et al., 2011). The source code for STAG-CNS is freely available at https://github.com/srbehera11/stag-cns.

Genomes and Orthologous Syntenic Gene Sets

The genomes and annotations of seven related species were downloaded from Phytozome (https://phytozome.jgi.doe.gov/pz/portal.html): sorghum (*Sorghum bicolor* v3.1) (McCormick et al., 2017), rice (*Oryza sativa* v7) (Ouyang et al., 2007), setaria (*Setaria italica* v2.2) (Bennetzen et al., 2012), brachypodium (*Brachypodium distachyon* v3.1) (Vogel et al., 2010), maize (*Zea mays*) (Schnable et al., 2009); and CoGe (https://genomevolution.org/coge/): oropetium (*Oropetium thomaeum* v1.0) (VanBuren et al., 2015) and dichanthelium (*Dichanthelium oligosanthes* v1.001) (Studer et al., 2016). A pan-grass orthologous gene list including all of these seven species generated on the GEvo panel of CoGe is available (https://genomevolution.org/coge/Gevo.pl) (Schnable et al., 2016). A set of 200 orthologous syntenic genes of six species (excluding maize) were randomly selected from the pan-grass syntenic gene list, which were used to test the optimal minimum length of CNS when comparing

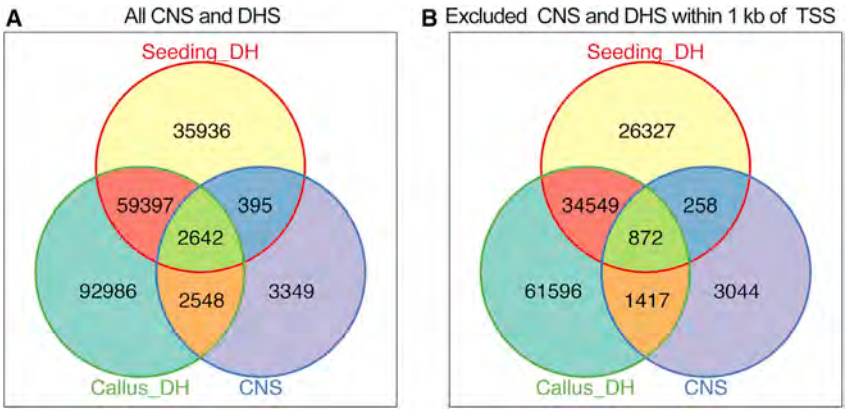


Figure 5. Overlap of Rice, Sorghum, and Setaria CNSs Identified Using STAG-CNS with Regions of Open Chromatin Identified using DNase1 Hypersensitivity in Rice Seedlings and Rice Callus.

As both CNS and DHS are commonly found near the start of transcription, this analysis was conducted separately for all CNS and DHS **(A)** and only for those CNS and DHS more than 1 kb distant from an annotated TSS **(B)**.

TSS, transcription start site.

orthologous genes from different numbers of species (Supplemental Table 1). Another set consisted of 17 996 orthologous syntenic genes present in three species (sorghum, setaria, and rice) were also extracted from the pan-grass syntenic gene list (Supplemental Table 2). A subset of 200 syntenic genes, which had the greatest number of CNSs between sorghum and rice based on CDP, was extracted (Supplemental Table 3). These syntenic gene sets were used for the comparison of STAG-CNS and CDP. A set of 6156 maize duplicated genes from two maize subgenome and corresponding syntenic genes in sorghum and setaria were obtained from the pan-grass syntenic genes list (Supplemental Table 7).

Extracting Noncoding Regions Associated with Genes and Permutation Testing of CNS False Discovery Rates

The region used for CNS discovery started 10 kb upstream and ended 10 kb downstream of a given syntenic gene. However, the sequence was truncated at the next conserved syntenic gene if this gene was less than 10 kb away from the start or end of the target gene. CNS may be identified in the nonsyntenic gene if one such gene is included in the flanking sequences of the given syntenic gene at an extremely small probability. For extremely high-confidence removal of sequences with coding potential or transcribed sequences, STAG-CNS can be combined with downstream filtering using ORF discovery and/or BAM files including RNA-seq reads (Turco et al., 2013). For permutation testing, these sequences were grouped randomly across species for each permutation. The average number of CNS identified across 100 permutations was compared with the original set of CNSs identified without shuffling syntenic relationships to estimate the FDR. The difference between the average number of CNSs from the permutation test and the number of CNSs from the original set (without shuffling) is considered to be false positive. The estimated true positive number of CNSs was the number of CNSs in syntenic genes minus the false discovery number of CNSs. The true positive rate is ratio of true positive number of CNS and all numbers of CNSs in syntenic genes. This analysis was repeated separately for minimum CNS lengths between 8 bp and 22 bp and for genes extracted from between two species and six species. The species compositions of each group were: two species (sorghum and rice), three species (sorghum, rice, and setaria), four species (sorghum, rice, setaria, and oropetium), five species (sorghum, rice, setaria, oropetium, and brachypodium), and six species (sorghum, rice, setaria, oropetium, brachypodium, and dichanthelium).

Comparing STAG-CNS and CDP Using Sorghum, Rice, and Setaria

The approach described in the previous subsection was also applied to extract the noncoding sequences 10 kb upstream and downstream of the genes. A total of 17 996 syntenic genes of sorghum, setaria, and rice were used to find the CNS with optimal minimum length of CNS at 12 bp.

At the same time, the CNSs in syntenic genes of sorghum–rice and sorghum–setaria were identified using CDP with default parameters (Turco et al., 2013) but with the modification that syntenic gene pairs were provided externally from the same list used by STAG-CNS rather than identifying syntenic gene pairs automatically (the default of the CDP). Then the CNSs of sorghum–rice and sorghum–setaria were combined based on positions on the genome of sorghum. The overlapped region ($\geq$12 bp) of CNSs between two pairwise species was regarded as the CNSs of pan-grass syntenic genes. Summary statistics were computed with custom perl scripts and basic functions in R software.

GO Term Enrichment Analyses

The enrichment analyses were performed using the goatools (version 0.5.9) (Tang et al., 2015), a Python script to find enrichment of GO terms. The genes in the GO annotation file of sorghum were used as association data. A total of 17 996 sorghum syntenic genes were used as a reference population. Nonannotated genes were temporarily annotated as a GO term that was not present in the analyses to test whether these unannotated genes were enriched in any groups. The genes with 0, 1, 2, and $\geq$3 CNSs were grouped respectively into four subsets. The Bonferroni correction and FDR implementation using resampling method were applied with a significance p value of <0.05. The GO enrichment data for sorghum are shown in Supplemental Table 6.

Association Analyses between CNSs and DNase I Sites

The DNase I sites of seeding and callus were downloaded from the NCBI GEO (http://www.ncbi.nlm.nih.gov/geo/), which have been generated by Zhang et al. (2012). The overlapping positions between DNase I sites and CNSs identified from 17 996 syntenic genes in sorghum, rice (v5), and setaria were calculated using custom perl scripts. A total number of 1873 syntenic genes in six species (sorghum, rice, setaria, oropetium, brachypodium, and dichanthelium) were extracted from the syntenic gene list. The CNSs with minimum length from 8 bp to 22 bp were extracted from different numbers (2–6) of species to identify the CNSs and compared with the DNase sites. The percentages of overlap of CNS and DNase sites for each case were calculated and shown in the heatmap (Figure 3C and 3D).

Comparison of CNSs in Maize Subgenomes

The set of 6156 syntenic genes form sorghum, setaria, and duplicated maize genes (Supplemental Table 1) were divided into two groups (sorghum-setaria-maize I and sorghum-setaria-maize II) that were used to identify the CNSs in the noncoding sequences 10 kb upstream and downstream of these syntenic genes. The numbers of CNSs were counted for each syntenic gene of the two groups and the differences between the duplicated genes, maize I and maize II, computed using in-house perl scripts. The expression values of these duplicated genes in multiple tissues were calculated from several datasets (Wang et al., 2009; Li et al., 2010; Davidson et al., 2011; Waters et al., 2011; Bolduc et al., 2012; Chang et al., 2012; Chettoor et al., 2014; Zhang et al., 2017) to

compare the expression level between maize I and maize II. For case 1, maize I genes having more CNS than maize II, we count the number if maize I genes have higher/lower expression than maize II and calculate the ratio between them to measure the bias. For case 2, maize I genes having less CNS than maize II, we also count the number if maize II genes have higher/lower expression than maize I and calculate the ratio. We use the values of the ratio of case 1 minus the ratio of case 2 to show the gene expression bias between the two cases in different tissues.

SUPPLEMENTAL INFORMATION

Supplemental Information is available at *Molecular Plant Online*.

FUNDING

This work was supported by internal funding to J.C.S. and J.S.D., and by a China Scholarship Council fellowship awarded to X.L.

AUTHOR CONTRIBUTIONS

J.C.S., J.S.D. and Y.L. designed the research. S.B., J.S.D., and X.L. conducted the software development. X.L., S.B., and Z.L. tested the software and analyzed the data. X.L. and S.B. wrote the initial draft. All authors reviewed and edited the paper.

ACKNOWLEDGMENTS

No conflict of interest declared.

Received: March 28, 2017
Revised: May 24, 2017
Accepted: May 30, 2017
Published: June 5, 2017

REFERENCES

Baxter, L., Jironkin, A., Hickman, R., Moore, J., Barrington, C., Krusche, P., Dyer, N.P., Buchanan-Wollaston, V., Tiskin, A., Beynon, J., et al. (2012). Conserved noncoding sequences highlight shared components of regulatory networks in dicotyledonous plants. Plant Cell **24**:3949–3965.

Bennetzen, J.L., Schmutz, J., Wang, H., Percifield, R., Hawkins, J., Pontaroli, A.C., Estep, M., Feng, L., Vaughn, J.N., Grimwood, J., et al. (2012). Reference genome sequence of the model plant *Setaria*. Nat. Biotechnol. **30**:555–561.

Bieganski, P., Riedl, J., Cartis, J.V., and Retzel, E.F. (1994). Generalized suffix trees for biological sequence data: applications and implementation. IEEE **5**:35–44.

Bolduc, N., Yilmaz, A., Mejia-Guerra, M.K., Morohashi, K., O'Connor, D., Grotewold, E., and Hake, S. (2012). Unraveling the KNOTTED1 regulatory network in maize meristems. Genes Dev. **26**:1685–1690.

Chang, Y.-M., Liu, W.-Y., Shih, A.C.-C., Shen, M.-N., Lu, C.-H., Lu, M.-Y.J., Yang, H.-W., Wang, T.-Y., Chen, S.C.-C., Chen, S.M., et al. (2012). Characterizing regulatory and functional differentiation between maize mesophyll and bundle sheath cells by transcriptomic analysis. Plant Physiol. **160**:165–177.

Chettoor, A.M., Givan, S.A., Cole, R.A., Coker, C.T., Unger-Wallace, E., Vejlupkova, Z., Vollbrecht, E., Fowler, J.E., and Evans, M. (2014). Discovery of novel transcripts and gametophytic functions via RNA-seq analysis of maize gametophytic transcriptomes. Genome Biol. **15**:414.

Crochemore, M., Iliopoulos, C.S., Mohamed, M., and Sagot, M.-F. (2006). Longest repeats with a block of k don't cares. Theor. Computer Sci. **362**:248–254.

Davidson, R.M., Hansey, C.N., Gowda, M., Childs, K.L., Lin, H., Vaillancourt, B., Sekhon, R.S., de Leon, N., Kaeppler, S.M., and Jiang, N. (2011). Utility of RNA sequencing for analysis of maize reproductive transcriptomes. Plant Genome **4**:191–203.

De Witte, D., Van de Velde, J., Decap, D., Van Bel, M., Audenaert, P., Demeester, P., Dhoedt, B., Vandepoele, K., and Fostier, J. (2015). BLSSpeller: exhaustive comparative discovery of conserved cis-regulatory elements. Bioinformatics **31**:3758–3766.

Flouri, T., Giaquinta, E., Kobert, K., and Ukkonen, E. (2015). Longest common substrings with k mismatches. Inf. Process. Lett. **115**:643–647.

Freeling, M., and Subramaniam, S. (2009). Conserved noncoding sequences (CNSs) in higher plants. Curr. Opin. Plant Biol. **12**:126–132.

Freeling, M., Woodhouse, M.R., Subramaniam, S., Turco, G., Lisch, D., and Schnable, J.C. (2012). Fractionation mutagenesis and similar consequences of mechanisms removing dispensable or less-expressed DNA in plants. Curr. Opin. Plant Biol. **15**:131–139.

Greene, B., Walko, R., and Hake, S. (1994). Mutator insertions in an intron of the maize knotted1 gene result in dominant suppressible mutations. Genetics **138**:1275–1285.

Guo, H., and Moose, S.P. (2003). Conserved noncoding sequences among cultivated cereal genomes identify candidate regulatory sequence elements and patterns of promoter evolution. Plant Cell **15**:1143–1158.

Gusfield, D. (1997). Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology (New York, NY, USA: Cambridge University Press).

Höhl, M., Kurtz, S., and Ohlebusch, E. (2002). Efficient multiple genome alignment. Bioinformatics **18**:S312–S320.

Hardison, R.C., Oeltjen, J., and Miller, W. (1997). Long human-mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome. Genome Res. **7**:959–966.

Haudry, A., Platts, A.E., Vello, E., Hoen, D.R., Leclercq, M., Williamson, R.J., Forczek, E., Joly-Lopez, Z., Steffen, J.G., Hazzouri, K.M., et al. (2013). An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. Nat. Genet. **45**:891–898.

Hughes, T.E., Langdale, J.A., and Kelly, S. (2014). The impact of widespread regulatory neofunctionalization on homolog gene evolution following whole-genome duplication in maize. Genome Res. **24**:1348–1355.

Inada, D.C., Bashir, A., Lee, C., Thomas, B.C., Ko, C., Goff, S.A., and Freeling, M. (2003). Conserved noncoding sequences in the grasses4. Genome Res. **13**:2030–2041.

Kaplinsky, N.J., Braun, D.M., Penterman, J., Goff, S.A., and Freeling, M. (2002). Utility and distribution of conserved noncoding sequences in the grasses. Proc. Natl. Acad. Sci. USA **99**:6147–6151.

Levy, S., Hannenhalli, S., and Workman, C. (2001). Enrichment of regulatory signals in conserved non-coding genomic sequence. Bioinformatics **17**:871–877.

Li, P., Ponnala, L., Gandotra, N., Wang, L., Si, Y., Tausta, S.L., Kebrom, T.H., Provart, N., Patel, R., Myers, C.R., et al. (2010). The developmental dynamics of the maize leaf transcriptome. Nat. Genet. **42**:1060–1067.

Li, L., Briskine, R., Schaefer, R., Schnable, P.S., Myers, C.L., Flagel, L.E., Springer, N.M., and Muehlbauer, G.J. (2016). Co-expression network analysis of duplicate genes in maize (*Zea mays* L.) reveals no subgenome bias. BMC Genomics **17**:875.

Ma, F., and Deogun, J.S. (2010). Multiple genome alignment based on longest path in directed acyclic graphs. Int. J. Bioinformatics Res. Appl. **6**:366–383.

Marcus, S., Lee, H., and Schatz, M.C. (2014). SplitMEM: a graphical algorithm for pan-genome analysis with suffix skips. Bioinformatics **30**:3476–3483.

McCormick, R.F., Truong, S.K., Sreedasyam, A., Jenkins, J., Shu, S., Sims, D., Kennedy, M., Amirebrahimi, M., Weers, B., McKinley, B., et al. (2017). The Sorghum bicolor reference genome: improved assembly and annotations, a transcriptome atlas, and signatures of genome organization. bioRxiv, 110593. http://dx.doi.org/10.1101/110593.

Nelson, C.E., Hersh, B.M., and Carroll, S.B. (2004). The regulatory content of intergenic DNA shapes genome architecture. Genome Biol. **5**:R25.

Ouyang, S., Zhu, W., Hamilton, J., Lin, H., Campbell, M., Childs, K., Thibaud-Nissen, F., Malek, R.L., Lee, Y., Zheng, L., et al. (2007). The TIGR rice genome annotation resource: improvements and new features. Nucleic Acids Res. **35**:D883–D887.

Pedersen, B.S., Tang, H., and Freeling, M. (2011). Gobe: an interactive, web-based tool for comparative genomic visualization. Bioinformatics **27**:1015–1016.

Pophaly, S.D., and Tellier, A. (2015). Population level purifying selection and gene expression shape subgenome evolution in maize. Mol. Biol. Evol. **32**:3226–3235.

Raatz, B., Eicker, A., Schmitz, G., Fuss, E., Müller, D., Rossmann, S., and Theres, K. (2011). Specific expression of LATERAL SUPPRESSOR is controlled by an evolutionarily conserved 3′ enhancer. Plant J. **68**:400–412.

Rodgers-Melnick, E., Vera, D.L., Bass, H.W., and Buckler, E.S. (2016). Open chromatin reveals the functional maize genome. Proc. Natl. Acad. Sci. USA **113**:E3177–E3184.

Schnable, P.S., Ware, D., Fulton, R.S., Stein, J.C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T.A., et al. (2009). The B73 maize genome: complexity, diversity, and dynamics. Science **326**:1112–1115.

Schnable, J.C., Springer, N.M., and Freeling, M. (2011). Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. Proc. Natl. Acad. Sci. USA **108**:4069–4074.

Schnable, J., **Zang, Y.**, and **D.W.C. Ngu**. (2016). Pan-Grass Syntenic Gene Set (Sorghum Referenced). Figshare. Available online at: https://dx.doi.org/10.6084/m6089.figshare.3113488.v3113481.

Shin, J.T., Priest, J.R., Ovcharenko, I., Ronco, A., Moore, R.K., Burns, C.G., and MacRae, C.A. (2005). Human-zebrafish non-coding conserved elements act in vivo to regulate transcription. Nucleic Acids Res. **33**:5437–5445.

Šmarda, P., Bureš, P., Horová, L., Leitch, I.J., Mucina, L., Pacini, E., Tichʾy, L., Grulich, V., and Rotreklová, O. (2014). Ecological and evolutionary significance of genomic GC content diversity in monocots. Proc. Natl. Acad. Sci. USA **111**:E4096–E4102.

Stephen, S., Pheasant, M., Makunin, I.V., and Mattick, J.S. (2008). Large-scale appearance of ultraconserved elements in tetrapod genomes and slowdown of the molecular clock. Mol. Biol. Evol. **25**:402–408.

Stewart, A.J., Hannenhalli, S., and Plotkin, J.B. (2012). Why transcription factor binding sites are ten nucleotides long. Genetics **192**:973–985.

Studer, A.J., Schnable, J.C., Weissmann, S., Kolbe, A.R., McKain, M.R., Shao, Y., Cousins, A.B., Kellogg, E.A., and Brutnell, T.P. (2016). The draft genome of the C 3 panicoid grass species *Dichanthelium oligosanthes*. Genome Biol. **17**:223.

Sun, X., Zou, Y., Nikiforova, V., Kurths, J., and Walther, D. (2010). The complexity of gene expression dynamics revealed by permutation entropy. BMC Bioinformatics **11**:607.

Swigoňová, Z., Lai, J., Ma, J., Ramakrishna, W., Llaca, V., Bennetzen, J.L., and Messing, J. (2004). Close split of sorghum and maize genome progenitors. Genome Res. **14**:1916–1923.

Tang, H., **Klopfenstein, D.**, **Pedersen, B.**, **Flick, P.**, **Sato, K.**, **Ramirez, F.**, **Yunes, J.**, and **Mungall, C**. (2015). GOATOOLS: Tools for Gene Ontology. Available online at: https://zenodo.org/record/31628#.WTkXHNR97Gg.10.5281/zenodo.31628.

Tatarinova, T.V., Alexandrov, N.N., Bouck, J.B., and Feldmann, K.A. (2010). GC 3 biology in corn, rice, sorghum and other grasses. BMC Genomics **11**:308.

Thomas, B.C., Rapaka, L., Lyons, E., Pedersen, B., and Freeling, M. (2007). Arabidopsis intragenomic conserved noncoding sequence. Proc. Natl. Acad. Sci. USA **104**:3348–3353.

Tsompana, M., and Buck, M.J. (2014). Chromatin accessibility: a window into the genome. Epigenetics & Chromatin **7**:33.

Tuğrul, M., Paixão, T., Barton, N.H., and Tkačik, G. (2015). Dynamics of transcription factor binding site evolution. PLoS Genet. **11**:e1005639.

Turco, G., Schnable, J.C., Pedersen, B., and Freeling, M. (2013). Automated conserved non-coding sequence (CNS) discovery reveals differences in gene content and promoter evolution among grasses. Front. Plant Sci. **4**:170.

Ukkonen, E. (1995). On-line construction of suffix trees. Algorithmica **14**:249–260.

Van de Velde, J., Van Bel, M., Van Eechoutte, D., and Vandepoele, K. (2016). A collection of conserved non-coding sequences to study gene regulation in flowering plants. Plant Physiol. **171**:2586–2598.

VanBuren, R., Bryant, D., Edger, P.P., Tang, H., Burgess, D., Challabathula, D., Spittle, K., Hall, R., Gu, J., Lyons, E., et al. (2015). Single-molecule sequencing of the desiccation-tolerant grass *Oropetium thomaeum*. Nature **527**:508–511.

Visel, A., Prabhakar, S., Akiyama, J.A., Shoukry, M., Lewis, K.D., Holt, A., Plajzer-Frick, I., Afzal, V., Rubin, E.M., and Pennacchio, L.A. (2008). Ultraconservation identifies a small subset of extremely constrained developmental enhancers. Nat. Genet. **40**:158–160.

Vogel, J.P., Garvin, D.F., Mockler, T.C., Schmutz, J., Rokhsar, D., Bevan, M.W., Barry, K., Lucas, S., Harmon-Smith, M., Lail, K., et al. (2010). Genome sequencing and analysis of the model grass Brachypodium distachyon. Nature **463**:763–768.

Wang, X., Elling, A.A., Li, X., Li, N., Peng, Z., He, G., Sun, H., Qi, Y., Liu, X.S., and Deng, X.W. (2009). Genome-wide and organ-specific landscapes of epigenetic modifications and their relationships to mRNA and small RNA transcriptomes in maize. Plant Cell **21**:1053–1069.

Waters, A.J., Makarevitch, I., Eichten, S.R., Swanson-Wagner, R.A., Yeh, C.-T., Xu, W., Schnable, P.S., Vaughn, M.W., Gehring, M., and Springer, N.M. (2011). Parent-of-origin effects on gene expression and DNA methylation in the maize endosperm. Plant Cell **23**:4221–4233.

Yu, C.-P., Lin, J.-J., and Li, W.-H. (2016). Positional distribution of transcription factor binding sites in *Arabidopsis thaliana*. Scientific Rep. **6**:25164.

Zhang, W., Wu, Y., Schnable, J.C., Zeng, Z., Freeling, M., Crawford, G.E., and Jiang, J. (2012). High-resolution mapping of open chromatin in the rice genome. Genome Res. **22**:151–162.

Zhang, Y., Ngu, D.W., Carvalho, D., Liang, Z., Qiu, Y., Roston, R.L., and Schnable, J.C. (2017). Differentially regulated ortholog analysis demonstrates that early transcriptional responses to cold are more conserved in Andropogoneae. bioRxiv, 120303. http://dx.doi.org/10.1101/120303.

The Plant Cell, Vol. 29: 1938–1951, August 2017, www.plantcell.org © 2017 ASPB.

Differentially Regulated Orthologs in Sorghum and the Subgenomes of Maize[OPEN]

Yang Zhang,[a,b] Daniel W. Ngu,[a,b] Daniel Carvalho,[a,b] Zhikai Liang,[a,b] Yumou Qiu,[c] Rebecca L. Roston,[a,d] and James C. Schnable[a,b,1]

[a] Center for Plant Science Innovation, University of Nebraska-Lincoln, Lincoln, Nebraska 68588
[b] Department of Agronomy and Horticulture, University of Nebraska-Lincoln, Lincoln, Nebraska 68588
[c] Department of Statistics, University of Nebraska-Lincoln, Lincoln, Nebraska 68588
[d] Department of Biochemistry, University of Nebraska-Lincoln, Lincoln, Nebraska 68588

ORCID IDs: 0000-0003-1712-7211 (Y.Z.); 0000-0002-7448-3629 (D.W.N.); 0000-0003-4846-1263 (Y.Q.); 0000-0002-3063-5002 (R.L.R.); 0000-0001-6739-5527 (J.C.S.)

Identifying interspecies changes in gene regulation, one of the two primary sources of phenotypic variation, is challenging on a genome-wide scale. The use of paired time-course data on cold-responsive gene expression in maize (*Zea mays*) and sorghum (*Sorghum bicolor*) allowed us to identify differentially regulated orthologs. While the majority of cold-responsive transcriptional regulation of conserved gene pairs is species specific, the initial transcriptional responses to cold appear to be more conserved than later responses. In maize, the promoters of genes with conserved transcriptional responses to cold tend to contain more micrococcal nuclease hypersensitive sites in their promoters, a proxy for open chromatin. Genes with conserved patterns of transcriptional regulation between the two species show lower ratios of nonsynonymous to synonymous substitutions. Genes involved in lipid metabolism, known to be involved in cold acclimation, tended to show consistent regulation in both species. Genes with species-specific cold responses did not cluster in particular pathways nor were they enriched in particular functional categories. We propose that cold-responsive transcriptional regulation in individual species may not be a reliable marker for function, while a core set of genes involved in perceiving and responding to cold stress are subject to functionally constrained cold-responsive regulation across the grass tribe Andropogoneae.

INTRODUCTION

The grasses are a clade of more than 10,000 species, which exhibit conserved morphology and genome architecture (Bennetzen and Freeling, 1993). Grasses have adapted to grow in a wide range of climates and ecologies across the globe, with 20% of total land area covered by ecosystems dominated by grasses (Shantz, 1954). As a result, the range of tolerance to abiotic stresses present in the grass family (Poaceae) far exceeds that present within any single grass species. However, to date, studies attempting to identify determinants of abiotic stress tolerance at a genetic or genomic level have predominantly focused on individual species (Chopra et al., 2017; Priest et al., 2014; Revilla et al., 2016; Tiwari et al., 2016; Waters et al., 2017). The majority of genetic changes with phenotypic effects can be broadly classified into two categories: those that alter protein-coding sequence and those that alter the regulation of gene expression.

DNA sequence changes that alter protein-coding sequences can be identified in a straightforward fashion. The probability that a given polymorphism in a protein-coding sequence will have a phenotypic effect can also often be estimated. At a basic level, this involves classification as synonymous, missense, and nonsense mutations. Information on the overall level of evolutionary conservation for a given amino acid residue can also be used to increase the accuracy of these predictions (Cooper et al., 2005; Ng and Henikoff, 2001; Reva et al., 2011). Cross-species comparisons of the protein-coding sequences from genes co-opted into new functional roles in C4 photosynthesis have been able to identify protein changes linked to changes in function at a resolution of individual amino acid residues (Christin et al., 2007).

Identifying changes in gene regulation across related species is more challenging, and the associated methods are far less advanced. For extremely close relatives, such as *Arabidopsis thaliana* and *Arabidopsis arenosa*, RNA-seq reads from both species can be mapped to a common reference genome (Burkart-Waco et al., 2015). For species with greater levels of sequence divergence in transcribed regions, this approach becomes impractical. Recent work in Sophophora (formerly Drosophila) described some of the many challenges present in comparing changes in baseline expression levels across closely related species with independently sequenced and assembled reference genomes (Torres-Oliva et al., 2016). However, this approach is limited to identifying changes in baselines expression in the same treatment rather than examining patterns of regulation across multiple treatments. Within the grasses, several research groups have employed clustering-based methods to identify genes with conserved patterns of regulation during either reproductive or photosynthetic development (Davidson et al., 2012; Wang et al., 2014). Among other results, one of these studies concluded that orthologous genes conserved at syntenic locations

[1] Address correspondence to schnable@unl.edu.
The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantcell.org) is: James C. Schnable (schnable@unl.edu).
[OPEN]Articles can be viewed without a subscription.
www.plantcell.org/cgi/doi/10.1105/tpc.17.00354

are more likely to share correlated expression patterns across multiple species than genes classified as orthologs based on phylogenetic analysis but located at nonsyntenic locations (Davidson et al., 2012). Clustering-based methods can identify genes with conserved patterns of regulation across multiple species, but they have high false positive rates when used to identify genes with changes in regulatory pattern.

In even closely related species, the baseline expression levels of orthologous genes can diverge significantly (Hollister and Gaut, 2009; Hollister et al., 2011). Testing for conserved or divergent patterns of regulation across different genotypes or different species when baseline expression levels have diverged creates a statistical challenge. Modeling of multiple environmental or genotype level effects can be combined either additively or multiplicatively. The model selected will determine which set of genes will be classified as differentially regulated between species. While few attempts have been made to identify differential patterns of gene regulation across species, attempts to do so between subspecies or diverse accessions have largely used either only a multiplicative model (Lovell et al., 2016), an additive model, or additive and multiplicative models separately (Waters et al., 2017) but have not made comparisons between the suitability of the two models.

Here, we sought to develop effective methods for comparing gene regulatory patterns between syntenic orthologous genes in closely related species. For initial cross-species comparisons, data on changes in the transcriptional responses to cold stress in maize (*Zea mays*) and sorghum (*Sorghum bicolor*) were employed. Cold was selected as a stress that could be delivered in a consistent fashion and time frame. Maize and sorghum were selected based on their close evolutionary relationship (Swigonová et al., 2004), high-quality sequenced genomes (Paterson et al., 2009; Schnable et al., 2009), and common susceptibility to cold stress (Chinnusamy et al., 2007; Hetherington et al., 1989; Wendorf et al., 1992). In addition, maize is a mesotetraploid species that experienced a whole-genome duplication ~12 million years ago after its divergence from sorghum (Swigonová et al., 2004), producing two functionally distinct maize subgenomes, maize1 and maize2 (Schnable et al., 2011). Approximately 3000 to 5000 pairs of genes are retained on both maize subgenomes (Schnable et al., 2009, 2011, 2012). Unlike other types of gene duplication, whole-genome duplicates initially retain almost all the same associated conserved regulatory sequences (Freeling et al., 2012). Comparing the expression patterns of duplicated genes exposed to the same *trans*-regulatory factors provides a bridge to comparing the expression patterns of orthologous genes in closely related species with similar phenotypes. These two systems provide a useful platform for developing and testing approaches to comparative gene regulatory analysis. However, one goal of cross-species comparisons of transcriptional regulation must ultimately be to link changes in regulation to changes in phenotype, which in the case of low-temperature stress will require conducting comparisons between species with differing, rather than similar, tolerance to cold.

RESULTS

A set of 15,231 syntenic orthologous gene pairs conserved between the maize1 subgenome and sorghum and 9553 syntenic gene pairs conserved between the maize2 subgenome was employed in this study (Figure 1A). The sequence identity in coding regions of syntenic genes between sorghum and either maize subgenome or between maize subgenomes is ~90% (Supplemental Figure 1), which is a level of divergence that makes alignment to a common reference sequence impractical. We conducted parallel expression analyses of the set of syntenic orthologous gene pairs conserved between the maize1 subgenome and sorghum and the smaller set of syntenic gene pairs conserved between the maize2 subgenome and sorghum.

Syntenic orthologs exhibited reasonably well-correlated patterns of absolute gene expression levels between sorghum and either subgenome of maize based on expression data generated from whole seedlings under control conditions (Spearman's rho = 0.79–0.84, Pearson r = 0.67–0.85, Kendall rank correlation 0.67–0.63; Figure 1B). This observation is consistent with previous reports about the analysis of expression across reproductive tissues in three grass species (Davidson et al., 2012). However, it should be noted that these correlations were significantly lower than those observed between biological replicates (see Methods for a detailed explanation of what constituted a biological replicate in this study) of the same species (Spearman's rho = 0.88–0.98, Pearson r = 0.89–0.99, Kendall rank correlation 0.78–0.91), and many individual genes have large divergence in baseline

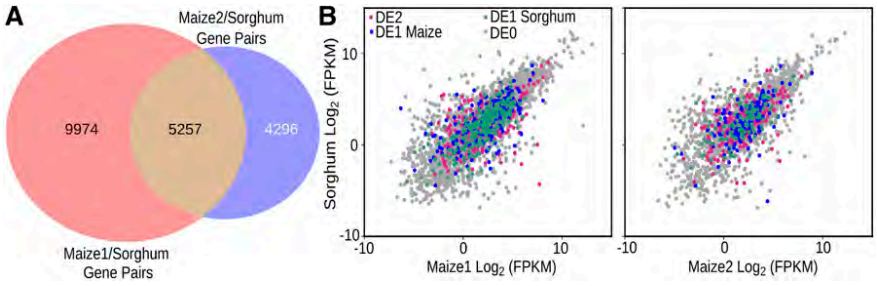


Figure 1. Gene Level and Expression Level Conservation between Sorghum, Maize1, and Maize2.

(A) The overlap between syntenic orthologous gene pairs conserved between maize1/sorghum and maize2/sorghum.

(B) Comparison of average control condition expression levels ($\log_2$ transformed FPKM) for either maize1/sorghum or maize2/sorghum gene pairs. (To improve readability, a random sample of 1/3 of all gene pairs is displayed for each category.)

expression levels between the two species, creating divergence between the predictions of additive and multiplicative statistical models of gene regulation, as described above.

We visually confirmed the lethal effect of prolonged cold stress on maize and sorghum (Ercoli et al., 2004; Hetherington et al., 1989; Olsen et al., 1993; Sánchez et al., 2014; Shaykewich, 1995) following prolonged cold treatment (Figures 2A to 2C; Supplemental Figure 2; see Methods). We employed measurements of impairment of CO_2 assimilation rates after recovery from a controlled length cold stress to provide more quantitative measures of cold stress and to assess the suitability of the level of cold stress employed to distinguish differing degrees of cold stress sensitivity or cold stress tolerance among maize, sorghum, and several related panicoid grass species. Data were generated from a total of six panicoid grasses, including the relatively cold tolerant paspalum (*Paspalum vaginatum*) and the extremely cold sensitive proso millet (*Panicum miliaceum*) (Figure 2D). After 1 d of cold stress, the species could be broadly classified as either cold stress insensitive or cold stress sensitive, with both maize and sorghum in the cold stress sensitive category. A longer period of cold stress (3 d) revealed greater impairment of CO_2 assimilation rates in sorghum than in maize, consistent with previous reports on the relative cold sensitivity of these two species (Chinnusamy et al., 2007; Chopra et al., 2017; Fiedler et al., 2016; Hetherington et al., 1989; Wendorf et al., 1992) and separated the six species into three broad categories of cold tolerant, moderately cold sensitive and extremely cold sensitive. Based on these data, we selected one day of cold stress, when maize and sorghum still exhibit comparable levels of CO_2 assimilation impairment (Figure 2D), for downstream expression analysis.

Conventional Differentially Expressed Gene Analysis

We identified differentially expressed genes in each species by comparing gene expression data in control seedlings to those subjected to one day of cold stress (Supplemental Data Set 1). Among maize1/sorghum syntenic gene pairs, 1686 (11.1%, 1686 out of 15,231) and 2343 (15.4%, 2343 out of 15,231) genes were classified as differentially expressed genes (DEGs), respectively (Figure 3A; see Methods). For maize2/sorghum syntenic gene pairs, these values were 968 (10.1%, 968 out of 9553) and 1446 (15.1%, 1446 out of 9553) genes, respectively. Only 836 (5.5%, 836 out of 15,231) of maize1/sorghum syntenic genes were classified as showing differential regulation in response to cold in both species (Figure 3A). In addition, there were 29 and 16 genes pairs in the maize1/sorghum and maize2/sorghum gene pairs, respectively, where both genes were classified as differentially expressed but in opposite directions (Figure 3B).

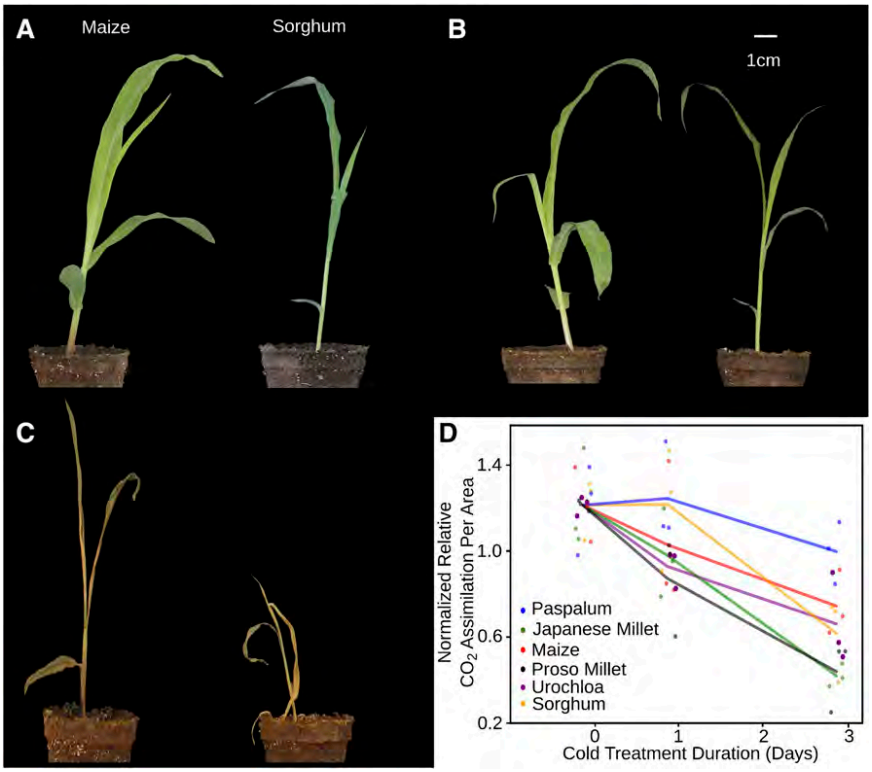


Figure 2. Effects of Cold Stress on Maize, Sorghum, and Related Species.

(A) to **(C)** Representative seedling phenotypes for maize and sorghum. Control conditions **(A)**, 24 h of stress at 6°C **(B)**, and 14 d at 6°C and 2 d recovery under greenhouse conditions **(C)**.

(D) Normalized relative CO_2 assimilation rates for six panicoid grass species with differing degrees of sensitivity or tolerance to cold stress. Individual data points were jittered (adding random noise to data in order to prevent overplotting in statistical graphs) on the *x* axis to avoid overlap and improve readability.

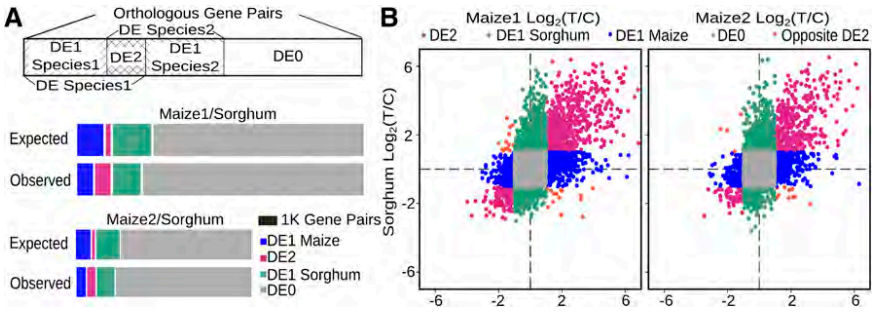


Figure 3. Combined DEG Analysis of Maize and Sorghum.

(A) An illustration of the DEG-based gene pair classification model and a comparison of expected and observed values for gene pairs classified as differentially expressed in response to cold in zero, one, or both species. Expected distributions were calculated based on a null hypothesis of no correlation in gene regulation between maize and sorghum (see Methods). DE0, gene pairs classified as differentially expressed in response to cold in neither species; DE1, gene pairs classified as differentially expressed in response to cold in one species but not the other; DE2, gene pairs classified as differentially expressed in response to cold in both species. Observed number of gene pairs in maize1/sorghum: DE1 maize = 850, DE2 = 836, DE1 sorghum = 1507, DE0 = 12,038. Observed number of gene pairs in maize2/sorghum: DE1 maize = 508, DE2 = 460, DE1 sorghum = 986, DE0 = 7599. Expected number of gene pairs in maize1/sorghum: DE1 maize = 1427, DE2 = 259, DE1 sorghum = 2084, DE0 = 11,461. Expected number of gene pairs in maize2/sorghum: DE1 maize = 822, DE2 = 146, DE1 sorghum = 1300, and DE0 = 7285.

(B) Comparison of fold change in gene expression between the treatment and control groups for pairs of orthologous genes in maize and sorghum. Log_2-transformed treatment/control expression ratios are shown.

The 836 observed syntenic gene pairs is ~3.2 times higher than the 259 genes pairs that should have been identified if cold-responsive gene regulation were not correlated between the two species (see legend of Figure 3 for a detailed breakdown of how this value was calculated). With these two values, the maximum number of genes responding to cold in the same fashion as a result of common descent from an ancestrally cold-responsive gene in the common ancestor of maize and sorghum can be calculated using the formula ((observed number of shared DEGs) − (expected number of shared DEGs))/(observed number of shared DEGs). In this case, a maximum of approximately two-thirds (69.0%, 577 out of 836) of genes identified as responding to cold in both species are likely to do so as a result of common descent. However, this may in fact be an overestimate if some of the same changes in cold-responsive gene regulation have been selected for in parallel in both lineages. Extending this calculation to the set of gene pairs that responded transcriptionally to cold in either maize or sorghum or both, only 18.1% (577 out of 3193) of gene pairs responding to cold in either species are likely to have retained a conserved pattern of cold-responsive gene expression since the divergence of maize and sorghum from a common ancestor 12 million years ago (Swigonová et al., 2004).

One potential explanation for this observation is that low statistical power to detect differentially expressed genes may create a false impression that differential expression is not conserved between related species. Prior estimates from real biological data in yeast (*Saccharomyces cerevisiae*) suggest that, given the number of replicates and minimum cutoff for differential expression employed here, the power of DESeq2 to identify differentially expressed genes should be between 0.65 and 0.90 (Schurch et al., 2016). In addition, a simulation study using observed expression values and variances in the maize data set generated here indicated that the power to detect differential gene expression ranged from 0.63 for genes with a change in expression exactly at the minimum cutoff to 0.961 for genes with larger changes in expression value (Supplemental Data Set 2). The expected proportion of genes classified as differentially expressed in either species that are classified as differentially expressed in both species is given by the formula $power^2/1-(1-power)^2$. Given the worst-case assumption (power = 0.628), this value would be 46% if gene regulation were perfectly conserved between maize and sorghum, which is higher than the observed value of 25%.

Results for maize2/sorghum gene pairs were largely comparable. However, the proportion of genes classified as not differentially expressed in either species was greater for maize2/sorghum gene pairs (Figure 3A), likely because maize2 genes tend to have lower overall levels of expression (Schnable et al., 2011). In total, 766 nonsyntenic maize genes were classified as differentially expressed in response to cold (2.0% of all nonsyntenic genes in maize, 766 out of 38,664), while 1333 (9.1%, 1333 out of 14,683) of nonsyntenic genes in sorghum were classified as differentially expressed in response to cold. The absolute numbers of differentially expressed nonsyntenic genes are more similar to each other than the proportions, as the current set of maize gene model annotations includes many lower confidence genes, which are generally nonsyntenic and often show little or no detectable expression (Schnable, 2015), than the current set of sorghum gene model annotations.

Maize and sorghum share a close relationship (Swigonová et al., 2004), and both originated from tropical latitudes (De Wet, 1978; van Heerwaarden et al., 2011). The two species even have a high degree of promoter conservation in abiotic stress-responsive genes (Freeling et al., 2007). Therefore, the apparent low degree of conservation in cold stress-responsive regulation is unexpected. However, this result is also consistent with studies that have found significant divergence in abiotic stress responses between different haplotypes in maize (Waters et al., 2017).

One potential explanation is that the same cold stress pathways are being induced in maize and sorghum, but these pathways are

induced more rapidly in one crop than the other when exposed to equivalent cold stresses. To test this hypothesis, we used data from a more detailed time course to compare the expression levels between matched pairs of cold stressed and control plants of each species at six time points distributed over 24 h (see Methods; Supplemental Data Set 1). The number of gene pairs classified as differentially expressed at different time points ranged from 60 to 2199 for maize1/sorghum gene pairs and 29 to 1235 for maize2/sorghum gene pairs. Comparing the number of genes identified as differentially expressed in each of all 36 possible pairwise combinations of time points between the two species showed that the greatest proportion of shared differentially expressed gene pairs was identified when identical time points were compared between the two species and that the overall number of shared differentially expressed gene pairs increases at later time points (Figure 4A). Overall, genes tended to remain in the same categories, with a general trend toward more DE0 genes moving into all three cold-responsive expression categories as the length of cold stress increased (Figure 4A). Because the proportion of all genes classified as differentially expressed increases at later time points, the expected number of gene pairs classified at DE2 under the null model described above also increases. Therefore, considering only the absolute number of gene pairs classified as DEGs in both species (DE2) at each time point can be misleading. After controlling for the expected number of DE2 genes, early time points show significantly higher proportions of true positives than later time points (Figure 4B).

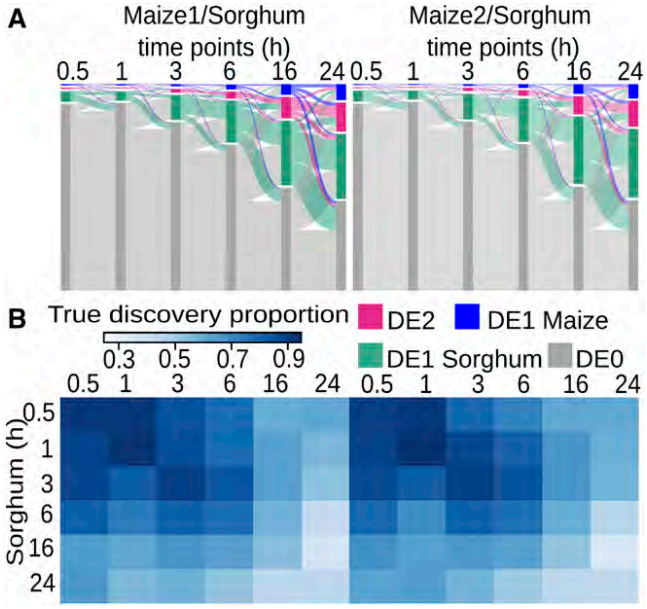


Figure 4. Patterns of Gene Expression across a Cold-Stress Time Series in Maize and Sorghum.

(A) Changes in classification of individual gene pairs as DE0, DE1 maize, DE1 sorghum, and DE2 across adjacent time points.

(B) The proportion of genes identified as differentially expressed in both species in excess of the number of gene pairs expected in this category in the absence of either conservation of gene regulation or parallel evolution of gene regulation. True discovery proportion is defined as (observed positives − estimated false positives)/observed positives. The expected number false positive DE2 gene pairs was calculated from the proportion of all genes classified as DEGs in maize and sorghum using the null model described in Figure 3A.

Differentially Regulated Ortholog Analysis

Another potential explanation for the finding that relatively few shared differentially expressed genes were identified between maize and sorghum is that differential gene expression analysis may not be testing the correct null hypothesis for between-species comparisons (Paschold et al., 2014). The null hypothesis of conventional DEG analysis is that the expression values observed for a given gene under control and stress conditions are drawn from the same underlying distribution. This approach is perfectly suitable for single-species analysis. In a two-species analyses, such as those conducted above, a DEG approach divides gene pairs into three categories: genes pairs classified as differentially expressed in neither species (DE0), in one species but not the other (DE1), and in both species (DE2; Figure 3A).

As shown in Figure 5A, in principle, each of those three categories (DE0, DE1, and DE2) can include gene pairs without significant differences in the pattern of regulation between species (comparably regulated orthologs [CROs]), as well as gene pairs that do show significant differences in regulation between the two species (differentially regulated orthologs [DROs]). All six theoretical cases from Figure 5A were observed in the RNA-seq expression data set generated above (Supplemental Figure 3A). DROs and CROs were both observed in all the DEG groups (Supplemental Figure 3B). Distinguishing between DROs and CROs requires testing a different null hypothesis: that the change in expression for a given gene between two treatments is equivalent to the change in expression for an ortholog of that same gene, in a different species, across the same two treatments. Another way of describing this same experimental approach is testing for a statistically significant treatment by species interaction effect. Several existing statistical packages incorporate the ability to test for significant interactions between different treatments (Love et al., 2014; Ritchie et al., 2015; Robinson et al., 2010) by including species as an effect in the model. However, comparing across species under different conditions, including testing for interaction effects to cross species comparisons, requires us to define an accurate model for what the same change in gene regulation looks like starting from different baseline levels of expression. Testing this null hypothesis across species in turn requires us to define an accurate model of what the same pattern of gene expression looks like when starting from different baseline levels of expression.

For an orthologous gene pair where gene copies are expressed at different baseline levels in two species, two different models can be used to compare a change in expression between treatment and control conditions: additive and multiplicative (Figure 5B). When expression under control conditions is equivalent between the two species, these models yield the same predicted expression under stressed conditions. However, when control condition expression is different between the two species, the models produce different expected expression values under stress conditions. Using simulated data on additive and multiplicative models, an

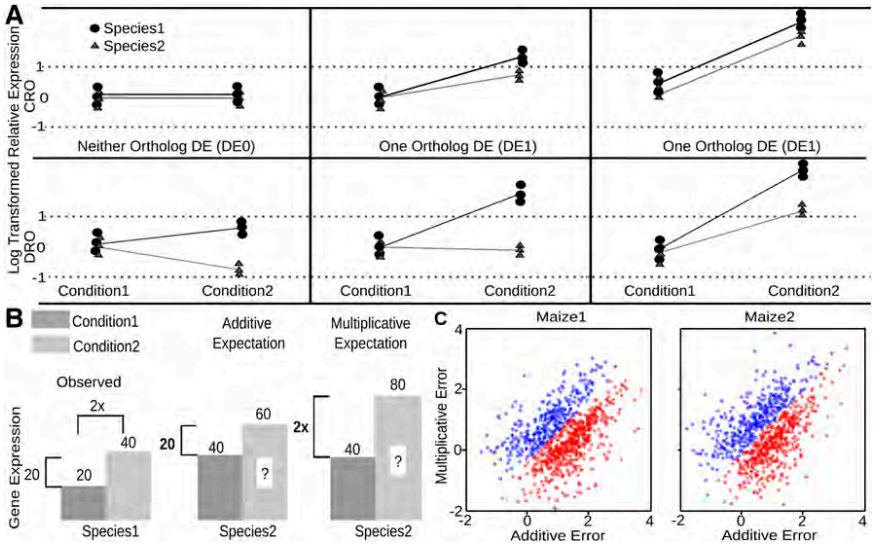


Figure 5. Conceptual Illustration of the Differentially Regulated Ortholog Model.

(A) Illustration of the different classification outcomes that can be produced for a given gene pair using both a DEG-based analysis (testing whether the expression pattern of each gene changes significantly between conditions) and a DRO-based analysis (testing whether the pattern across the two conditions is significantly different between copies of the same gene in both species).

(B) Two models, additive and multiplicative, for predicting what a conserved pattern of gene regulation should look like when the underlying level of expression changes.

(C) Relationship between prediction error (log_{10} transformed) for expression under cold stress using a multiplicative model to predict expression between maize1/maize2 gene pairs or an additive model to predict expression between maize1/maize2 gene pairs. Maize1: Predictions for the expression pattern of maize2 genes using data from their maize1 homoeologs. Maize2: Predictions for the expression pattern of maize1 genes using data from their maize2 homoeologs. Blue dots mark cases where the additive model was the better predictor; red dots mark cases where the multiplicative model was the better predictor.

ANOVA-based test classified genes with different baseline expression levels but the same pattern of expression (as simulated by a multiplicative model) as significantly differentially regulated between species, while the generalized linear model-based DESeq2 classified genes with different baseline expression levels but the same pattern of expression (as simulated by a additive model) as significantly differentially regulated between species (Supplemental Data Set 3).

To test which of these models is a better representation of how cold-responsive gene regulation actually operates, we used a set of 5257 gene pairs retained from the maize whole-genome duplication (WGD) (Schnable et al., 2011). The maize WGD created two copies of each gene in the genome, each associated with the same chromatin environments and regulatory sequences. RNA-seq-based measurements of expression for duplicate genes can be unreliable when gene copies are similar enough that reads cannot be unambiguously mapped to individual copies. Maize WGD-derived duplicate gene pairs show ~93% sequence similarity in exon regions (Supplemental Figure 1). This is equivalent to 4.5 mismatches per 50-bp sequence read, significantly reducing the risk of ambiguous or incorrect read mapping. The expression level of each gene copy in a WGD gene pairs in the maize genome in the same samples results from the exact same *trans*-factors acting in the exact same tissue and cell types. Therefore, divergence in the regulation of these genes should start out with the same *cis*-regulatory sequence prior to their divergence from their most recent common ancestor (whether at the time of WDG for autopolyploids or at the time of speciation prior to WGD for allopolyploids) (Freeling et al., 2012).

To test the additive and multiplicative null models, we used the expression pattern of one maize gene copy between control and cold stress conditions to predict the expression pattern of the other maize gene copy using each null model from Figure 5B. We conducted the analysis in parallel at each of the six time points in maize using maize1/maize2 gene pairs where at least one copy was identified as differentially expressed at that time point. Gene pairs were omitted from the analysis if the predictions of both models were more similar to each other than either was to the observed value.

The multiplicative model was more accurate at predicting cold-responsive expression patterns between maize WGD duplicates than the additive model at all time points (P = 0.004-2.4*10-15, paired two tailed *t* test) (Supplemental Data Set 4). Requiring the difference between the predictions of the two models to be at least twice as large as the difference between the better model and the observed expression pattern produced similar results (Figure 5C; Supplemental Data Set 4). The set of genes where the additive model produced better predictions was examined for differences in expression, selection (Ka/Ks ratio) (Supplemental Figure 4), or Gene Ontology (GO) annotation. No significant markers for which genes could be best predicted with which model were identified. Therefore, going forward, we employed the multiplicative model for conserved gene regulation across species,

as implemented in DESeq2's test for multiple factors (Love et al., 2014) (see Methods).

Figure 6A shows the proportion of gene pairs classified as DROs among all gene pairs in the DE0, DE1, and DE2 groups at each of the six time points. Comparing the same time points for maize and sorghum identifies fewer differentially regulated orthologs than comparisons between nonequivalent time points in the two species. Fewer differentially regulated orthologs were identified at earlier cold treatment time points than at later time points. This is consistent with the results of DEG analysis described above, which suggested early cold stress responses were more conserved across sorghum and maize than later cold stress responses.

Functional Differences between Genes with Conserved or Lineage-Specific Regulatory Patterns

Genes classified as responding to cold stress in both species (DE2) tended to have significantly lower ratios of nonsynonymous nucleotide changes to synonymous nucleotide changes (Ka/Ks ratio) than genes that responded to cold stress in only one species or in neither species. This suggests genes with conserved patterns of cold-responsive regulation experience stronger purifying selection than genes with lineage-specific patterns of cold-responsive regulation (Figures 6B and 6C). GO enrichment analysis identified genes differentially regulated in both species as enriched in transcription factor-related GO terms, such as GO:0006355 "regulation of transcription, DNA-templated." This enrichment was further confirmed in a separate test for enrichment of genes annotated as transcription factors in the GRASSIUS database (Yilmaz et al., 2009). No non-transcription factor-related GO term showed significant enrichment when compared with the population of gene pairs that were syntenically conserved between both species. Comparison to the total population of annotated genes in maize or sorghum showed many additional enrichments; however, this approach can produce misleading results, as nonsyntenic genes are enriched among genes without any functional annotation (Schnable et al., 2012). We used MapMan (Usadel et al., 2009) to visualize the patterns of expression within particular functional categories among DE2 genes as well as DE1 maize and DE1 sorghum genes. As expected, genes related to cell wall growth, a marker for plant growth, were downregulated in both species in the cold, including xyloglucosyl transferase (Sobic.001g538000 and GRMZM2G388684) and leucine-rich repeat family protein (Sobic.003g205600 and GRMZM2G333811) genes (Cui et al., 2005; Pearce, 2001; Tenhaken, 2014). Genes involved in lipid metabolism were upregulated in both species, including glycerol-3-phosphate acyltransferase 8 (Sobic.009g162000 and GRMZM2G166176), diacylglycerol kinase (Sobic.006g230400 and GRMZM2G106578), choline-phosphate cytidylyltransferase (Sobic.001g282900 and GRMZM2G132898), MGDG synthase (Sobic.004g334000 and GRMZM2G178892, Sobic.007g211900 and GRMZM2G141320), glycerophosphodiester phosphodiesterase (Sobic.007g190700 and GRMZM2G064962, Sobic.004g157300 and GRMZM2G018820), and fatty acid elongation acyl-CoA ligase (Sobic.004g015400 and GRMZM2G120539) genes. This observation is consistent with the reported role of changes in membrane composition to avoid stiffening in the cold as an adaptive response to cold (Quinn, 1988; Singer and Nicolson, 1972). No consistent expression patterns of genes in particular metabolic processes (up- or downregulated) were observed among the DE1 maize or DE1 sorghum gene pairs.

The previously defined binding site for DREB/CBF transcription factors, which are induced in response to drought and cold stress (Muiño et al., 2016), showed significant enrichment in the proximal promoters of gene pairs in the DE2 category, as well as significant purification in the proximal promoters of gene pairs in the DE0 category (Supplemental Figure 5). As transcription factors are often associated with larger quantities of conserved noncoding sequences (CNSs) (Freeling et al., 2007; Turco et al., 2013), we also investigated the number and quantity of conserved noncoding sequence associated with different classes of genes; however, no strong patterns were observed (Figure 6D). The use of conserved noncoding sequence data to identify regulatory sequence requires that the regulatory sequence be conserved between species. Given that many of the genes identified as responding to cold in either maize or sorghum appear to do so in a lineage-specific fashion, this requirement may not be satisfied in many cases. Various measurements of open chromatin have been shown to be good predictors of where regulatory sequences will be identified using CNS-based methods (Lai et al., 2017; Vera et al., 2014; Zhang et al., 2012), and unlike CNS-based methods, chromatin structure-based methods do not require that the same regulatory sequence be conserved across multiple species. We therefore examined the chromatin states in the promoters of genes with different patterns of cold-responsive regulation using a published data set of MNase hypersensitive sites (HSs) generated from maize seedlings grown under nonstressed conditions (Rodgers-Melnick et al., 2016). Comparisons were made for maize DE0, maize DE1, sorghum DE1, DE2, and nonsyntenic genes at each of the six cold stress time points. Many nonsyntenic genes responded to cold; however, nonsyntenic genes as a whole showed little or no open chromatin (as defined by MNase HS) associated with their TSSs (transcriptional start sites) or proximal promoters. Previous studies of other epigenetic marks have also concluded that the chromatin signatures of nonsyntenic genes in maize are more similar to those of intergenic sequences versus syntenic genes (Eichten et al., 2011). All categories of syntenic genes tended to have a peak of MNase sensitivity associated with their TSS and more open chromatin in their proximal promoters than nonsyntenic genes. Genes with conserved cold-responsive regulation (DE2) appear to have the greatest amount of open chromatin in their proximal promoters (Figure 7). Intriguingly, the maize copies of maize DE1 gene pairs exhibited stronger open chromatin signals that the maize copies of sorghum DE1 gene pairs, even though data on MNase hypersensitive sites came from seedlings grown under control conditions. The patterns reported above remained apparent when genes were divided into nine categories based on their relative expression level and Ka/Ks ratio, although statistical significance was reduced substantially as a result of the smaller number of genes included in each analysis (Supplemental Figure 6).

DISCUSSION

The above results indicate that there are roughly equivalent numbers of genes differentially expressed in response to cold compared with those reported from separate studies in each

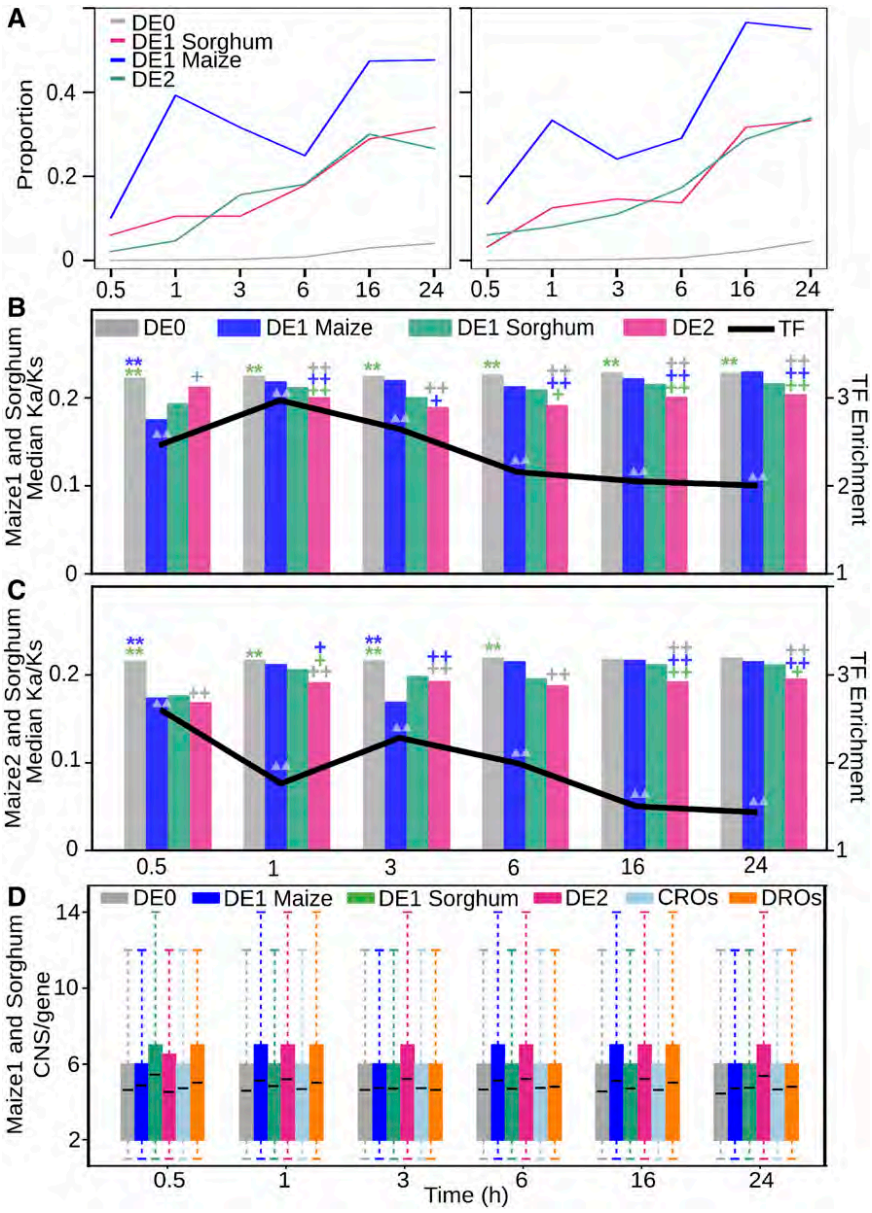


Figure 6. Characteristics of Genes in Different DEG Groups at Different Time Points.

(A) The proportion of gene pairs classified as DROs between maize and sorghum in different DEG groups at each of the six time points examined.
(B) and **(C)** Median ratios of nonsynonymous substitutions to synonymous substations in coding sequences for maize and sorghum for gene pairs classified as DE0, DE1, or DE2 at each of six time points. Time points where there is a statistically significant difference in Ka/Ks ratio between DE2 and any of the other three categories are marked with either + (if $P < 0.05$) or ++ (if $P < 0.01$). Color of the + indicates the category to which DE2 is being compared. Time points where there is a statistically significant difference in Ka/Ks ratio between DE0 and either DE1 maize or DE1 sorghum categories are marked with either * (if $P < 0.05$) or ** (if $P < 0.01$). Color of the asterisk indicates the category to which DE0 is being compared. Enrichment of genes annotated as transcription factor genes among DE2 gene pairs relative to all syntenic gene pairs indicated by the black line and the right-hand axis. Double white triangles mark time points where the enrichment is statistically significant ($P < 0.01$).
(D) Frequency of CNS within the promoters of genes classified as DE0, DE1 maize, DE1 sorghum, DE2, DRO, or CRO at each of the six time points. Black lines within the box plot mark the average number of CNS per gene for each category.

species (Chopra et al., 2015; Makarevitch et al., 2015). However, cross-species comparisons of the transcriptional regulation of the same genes in these two different species reveals that many cold-responsive patterns of regulation are not conserved between the two species. Correcting for the expected overlap across conserved genes based solely on the absolute genes number exhibiting cold-responsive transcriptional changes in each species further reduced the expected number of gene pairs where shared regulation resulted from the conservation of an ancestral pattern of cold-responsive transcriptional regulation. These data imply that

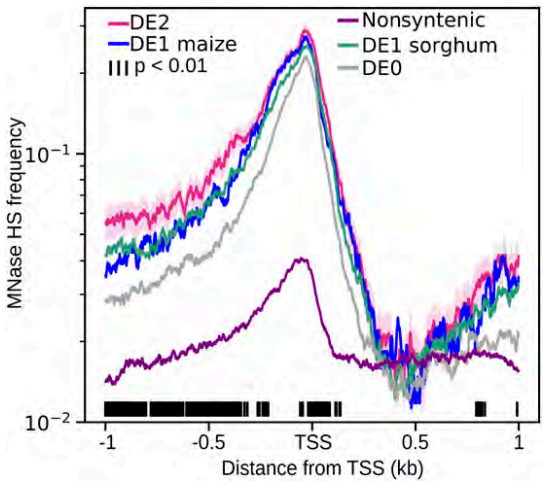


Figure 7. Chromatin Patterns Associated with Different Groups of Genes in Maize and Sorghum.

Patterns of MNase HS regions around the transcriptional start sites of genes classified based on their pattern of gene regulation in the 24-h stress time point. Maize1 sorghum gene pairs and maize2 sorghum gene pairs were aggregated to increase statistical power. The lighter band around the DE2 line indicates a 2 SD confidence interval. Black bars at the bottom of the graph indicate individual base pair positions where the amount of open chromatin associated with DE2 genes is significantly different from that of each of the other four categories displayed with a P value < 0.01 for each comparison. Pairwise comparisons were performed using Fisher's exact test.

gains or losses of cold-responsive regulation are relatively frequent in the grass tribe Andropogoneae. Genes that respond to cold in only a single lineage experience lower levels of purifying selection and are less likely to be annotated as transcription factor genes than genes that are cold-responsive in both lineages. It should be noted that these results are based on data from a single accession of maize (B73) and a single accession of sorghum (BTx623). Evidence suggests that lower, but still significant, levels of divergence in transcriptional regulation in response to cold are present in different accessions of a single species (Makarevitch et al., 2015; Waters et al., 2017).

It appears that a relatively small core set of genes exhibit conserved responses to cold across the two species in this initial analysis, and functional analysis suggests that these genes are more likely to be present in pathways with logical links to cold stress (decreases in growth and cell wall biosynthesis, increases in lipid metabolism). Thus, we propose a model where a small core set of genes involved in the mechanisms by which panicoid grasses perceive and respond to cold stress are under functionally constrained cold-responsive transcriptional regulation, while a much larger set of genes can gain or lose cold-responsive transcriptional regulation in a neutral fashion or potentially as a result stabilizing selection, potentially through transposon-mediated mechanisms (Makarevitch et al., 2015; Naito et al., 2009). Consistent with this model, the genes with conserved cold-responsive gene regulation exhibited lower ratios of nonsynonymous-to-synonymous coding sequence substitutions than the other genes, which would imply their coding sequence is also subject to greater functional constraint. This model would also be consistent with the relatively high proportion of maize cold-responsive genes that exhibit variation in cold-responsive regulation across alleles (Waters et al., 2017).

We evaluated two different models for predicting conserved regulation across different expression levels and found that the multiplicative model was more effective at predicting orthologous gene pair expression than the additive model (Figure 5C; Supplemental Data Set 4). However, while this different was statistically significant, the additive model remained the better predictor for many gene pairs. While no obvious markers that distinguish genes where one model is the better predictor than the other were identified in this study, further study may identify additional molecular traits measured from the genome that can forecast which model is more appropriate for testing the expression pattern of a given gene across multiple related species.

The Challenge of Linking Genes to Functions Based on Expression Evidence

The model above would predict that the observation of stress-responsive changes in transcript abundance in a single species is not strong evidence that the associated gene plays a role in the response to that particular stress. While sequencing genomes and identifying genes are becoming more straightforward tasks, confidently assigning functional roles to newly identified genes remains challenging. Many genes in maize (35.1%) and sorghum (16.2%) are not associated with any GO annotations in the current release of Phytozome (v12). Many genes that do possess GO annotations are associated with only extremely broad annotation categories, such as protein binding or catalytic activity. "Guilt by association" studies using coexpression analysis are an intriguing method for assigning putative functional roles to some orphan or poorly annotated genes (Li et al., 2016; Schaefer et al., 2014). However, the use of these methods in a single species may also produce false positive annotations in the case of selectively neutral or stabilizing changes in gene regulation. It may prove to be the case that functionally constrained transcriptional responses are an effective method for identifying these links. Collecting parallel expression data sets in multiple species can be time consuming and costly. We therefore tested a number of alternative approaches to identifying functionally constrained cold-responsive transcriptional regulation. Early transcriptional responses to cold (30 min to 3 h) appeared to show greater conservation across species than later transcriptional responses. Regions of open chromatin detected through MNase HS (Rodgers-Melnick et al., 2016; Vera et al., 2014) were preferentially associated with genes that responded transcriptionally to cold stress in maize; however, this association was observed for genes with either conserved or lineage-specific patterns of cold-responsive regulation.

Importance of Developing Methods for Cross-Species Comparisons of Transcriptional Regulation

Both modeling (Orr, 1998, 1999) and empirical studies (Chan et al., 2010; Studer et al., 2011) have found that genetic variants responsible for large, sudden changes in natural or artificial selection tend to have large, pleiotropic effects. In maize, distinct genetic architectures underlie traits that have been subjected to selection

during domestication (one large-effect quantitative trait locus and many small modifiers) and traits that were not selected on during domestication (many small-effect quantitative trait loci) (Wallace et al., 2014). This model was supported by recent work with an intersubspecies cross of maize and its wild progenitor teosinte (*Z. mays* ssp *parviglumis*). Looking at tassel morphology, distinctly genetic architectures were reported for traits believed to have been under selection during domestication compared with those traits that were not (Xu et al., 2017). Developing effective approaches for comparing transcriptional regulation of conserved syntenic genes across related grass species has the potential to identify large-effect polymorphisms responsible for interspecies phenotypic variation in traits such as abiotic stress tolerance where substantial phenotypic variation exists between species (Figure 2D).

Here, we have shown that by using synteny to identify pairs of conserved orthologs across related species, it is possible to identify species by treatment interactions, which signify changes in gene regulation across species (DROs), using a multiplicative model of gene regulation. The use of a multiplicative model was in turn supported by analysis of the regulation of duplicated maize genes within the same sample. By increasing the number of species sampled, it may soon be possible to define a consistent core set of genes subjected to functionally constrained regulation in response to cold across the grasses. Changes in the regulation of these core genes in specific lineages with different cold stress-response phenotypes would be useful candidates for the type of large-effect changes predicted to produce between-species phenotypic variation. However, the interpretation of such data must take into account that, unlike within-species studies of allelic variation in cold-responsive regulation, between-species analysis cannot distinguish *cis*-regulatory from *trans*-regulatory sources of variation in transcriptional responses.

METHODS

Plant Growth and Cold Treatment

For maize (*Zea mays*) and sorghum (*Sorghum bicolor*), the reference genotypes used for genome sequencing and assembly were B73 and BTx623, respectively. SNP calling using RNA-seq data from B73 was used to verify that the plants used in this study came from the USA South clade of B73 accessions, i.e., those closest to the original reference genome (Liang and Schnable, 2016). Under the growing conditions employed, maize developed more quickly than sorghum, and sorghum seedlings 12 d after planting were selected as being roughly developmentally equivalent to maize seedlings 10 d after planting based on leaf number and morphology (Figure 2A). Planting dates were staggered so that all species reached this developmental time point simultaneously. For the original RNA-seq presented in Figure 2A, seeds were planted in MetroMix 200 and grown in greenhouse conditions under 13 h daylength in greenhouses at University of Nebraska- Lincoln's Beadle Center, with target conditions of 320 mol m^{-2} s^{-1}, high-pressure sodium bulb, 13 h/11 h 29°C /23°C day/night, and 60% relative humidity. Control plants were harvested directly from the greenhouse three hours before lights on. Plants subjected to cold stress treatment were moved to a cold treatment growth chamber, with 33 mol m^{-2} s^{-1}, metal halide grow bulb, 12 h/12 h 6°C /6°C day/night. Cold-stressed plants were harvested 3 h before lights on. Each sample consisted of pooled aboveground tissue from at least three seedlings. Each biological replicate was harvested from plants that were planted, grown, and harvested at a distinct and separate time from each other biological replicate. A total of three independent biological replicates where generated for this experiment. For the time course RNA-seq data presented in Figure 4 and onward in the study, maize and sorghum were planted as above and grown in a Percival growth chamber (Percival model E-41L2) with target conditions of 111 mol m^{-2} s^{-1} light levels, 60% relative humidity, a 12 h/12 h day night cycle with a target temperature of 29°C during the day and 23°C at night. The onset of cold stress treatment was immediately before the end of daylight illumination, at which point half of the plants were moved to a second growth chamber with equivalent settings with the exception of a target temperature of 6°C both during the day and at night. Each sample represents a pool of all aboveground tissue from at least three seedlings. Samples were harvested from both the paired control and cold stress treatments at 0.5, 1, 3, 6, 16, and 24 h after the onset of cold stress. Biological replicates included both maize and sorghum plants that were offset in planting but stressed and harvested at the same time in the same growth chambers. A total of three independent biological replicates were generated for this experiment.

Definition of Samples and Biological Replicates

Sample: Each sample consists of RNA extracted from the pooled tissue of no less than three and no more than five separate plants planted and harvested on the same date and grown in the same growth chamber. All aboveground tissue was harvested from each plant included in a pool. All aboveground tissue: At the stage plants were harvested, all aboveground tissue included leaf blades, ligules, and leaf sheaths, but not apical meristems, stems, or roots. Biological replicate: each biological replicate consists of RNA extracted from pooled tissue harvested from plants of the same genotype planted and harvested on separate dates from any other biological replicate. Paired replicate: biological replicates were paired across species, with tissue harvested on the same day from plants of each species growth in the same growth chamber.

CO_2 Assimilation Rate Measurements

Plants were grown and cold treated as above, with the modification that in the case of sorghum, small plastic caps were placed over the seedlings to prevent the plants from becoming too tall to fit into the LiCor measurement chamber (~2 inches). After 0, 1, or 3 d of cold treatment, the plants were allowed to recover in the greenhouse overnight. The following morning, CO_2 assimilation rates were measured using the Li-6400 portable photosystem unit under the following conditions: PAR 200 mol mol^{-1}, CO_2 at 400 mol mol^{-1} with flow at 400 mol mol^{-1}, and humidity at greenhouse conditions. Whole-plant readings were measured for sorghum, paspalum, Japanese millet (*Echincloa esculenta*), proso millet, and urochloa (*Urochloa fusca*) after covering their pots with clay and using the LiCor Arabidopsis chamber. Maize was measured using the leaf clamp attachment, which was consistently placed on the second leaf at a position 3 cm above the ligule. Leaf area was measured using the Li-3100c Area meter (Li-Cor). The accessions used for each species presented in Figure 1D included the following: paspalum, USDA PI 509022; Japanese millet, USDA PI 647850; proso millet, earlybird USDA PI 578073; urochloa, LBJWC-52; sorghum, BTx623; and maize, B73.

Identifying Syntenic Orthologs

Coding sequence data for primary transcripts of each annotated gene in the genome assemblies of eight grass species, including maize and sorghum used in the analysis, were obtained from Phytozome 10.2. Similar sequences were identified using LASTZ (Harris, 2007), requiring an alignment spanning at least 50% of total sequence length and 70% sequence identity. In addition, the arguments -ambiguous=iupac, -notransition, and

-seed=match12 were all set in each run. LASTZ output was converted to QuotaAlign's "RAW" format using a version of the blast to raw.py script that had been modified to take into account differences in output format between BLAST and LASTZ. The additional parameters -tandem Nmax=10 and -cscore=0.5 were specified when running this script.

RAW formatted data were processed using the core QuotaAlign algorithm with the parameters -merge, and -Dm=20. -quota was set to 1:2 in comparisons to maize and 1:1 in all other comparisons. Pure QuotaAlign pan-grass syntenic gene sets were constructed using this data set directly. Polished QuotaAlign pan-grass syntenic gene sets were constructed by first predicting the expected location for a given query gene in the target genome and then selecting the gene showing the greatest sequence similarity (as determined by lastz alignment score) within the window from 20 genes downstream of the predicted location to 20 genes upstream of the predicted location.

RNA-Seq Data Generation

RNA isolation and library construction followed the protocol described by Zhang et al. (2015). The number of reads generated per library is summarized in Supplemental Data Set 1. Sequencing was conducted at Illumina Sequencing Genomics Resources Core Facility at Weill Cornell Medical College. Raw sequencing data are available through the NCBI (http://www.ncbi.nlm.nih.gov/bioproject) under accession numbers PRJNA343268 and PRJNA344653. Adapters were removed from raw sequence reads using cutadapt version 1.6 (Martin, 2011). RNA-seq reads were mapped to genome assemblies downloaded from Phytozome: RefGen v3 (*Z. mays*) and v3.1 (*S. bicolor*). RNA-seq reads from each species were aligned using GSNAP version 2014-12-29 (Wu and Nacu, 2010; Wu and Watanabe, 2005). Per-gene read counts were obtained using HTSeq version 0.6.1 (Anders et al., 2015).

Identifying DEGs

DEGs were identified using count data generated as described above and DESeq2 (version 1.14.0) (Love et al., 2014) based on a comparison of the treatment and control with adjusted P value ≤ 0.05, meaning absolute $\log_2$ of fold change of between treatment and control value ≥ 1. All expressed syntenic orthologous genes were classified into one of three categories. The three categories include genes that were classified as responding transcriptionally to cold in at least one species (DE1) (Figure 3A). The remaining category includes all expressed syntenic orthologous genes that were not classified as cold-responsive in either of the two species (DE0). The number of shared genes identified as differentially expressed in the two species (DE2) was tested relative to the expected overlap if there was no correlation in gene regulation across species. For the time course RNA-seq, analysis was conducted as above for all 36 possible pairwise comparisons of the six sorghum time points and six maize time points.

When estimating the true discovery proportion in analyses of DE2 genes (see Figures 3A and 4B), it was necessary to calculate the number of DE2 genes expected under a null hypothesis of no conservation of gene regulation. This expected number of DE2 genes was calculated using the formula (percentage of gene pairs DE in species 1)*(percentage of gene pairs DE in species 2)*(total number of gene pairs analyzed was used). Total number of gene pairs was fixed at 15,232 syntenic orthologous gene pairs for maize1/sorghum comparisons and 9554 for maize2/sorghum comparisons.

Estimating the Power of DESeq2 in This Data Set Using Simulated Data

One thousand genes were randomly sampled from the maize1/sorghum syntenic gene list in each repetition of the simulation. These selected genes included three replicates from both normal growth conditions (control) and 1-d cold treatment (treatment). The geometric mean of each gene was calculated (adding 1 to the data to avoid 0 readings). A random sample from the uniform distribution on (5, 50) was used as the estimate of the true dispersion parameter. The simulated data for the non-differentially expressed genes were generated from a negative binomial distribution with the calculated geometric mean from the actual data and the sampled dispersion parameter. To generate the list of differentially expressed genes, the first 100 genes out of the 1000 sampled genes were selected with a treatment mean value equal to the geometric mean from the original data, whereas the mean value of the control was a multiple of the geometric mean (multiples of 2, 2.5, and 3 are reported). The calculated false discovery rate (ratio of number of false positives over total number of discoveries) and the power (ratio of true positives over the true number of differentially expressed genes) of the DESeq2 procedure are reported in Supplemental Data Set 2.

Evaluating the Additive and Multiplicative Models of Gene Regulation

From the 5257 duplicate genes retained from the maize WGD (Schnable et al., 2011) in each of the six time points in maize, gene pairs where both copies were classified as differentially expressed in response to cold were used to test both models. The expression pattern of the maize1 gene under control and cold stress conditions plus the expression of the maize2 gene under control conditions was used to predict the expression of the maize2 gene under cold stress using both the additive and multiplicative models defined in Figure 5B. The distance between the prediction from the additive model and the observed value was defined as "a," the distance between the prediction from the multiplicative model and the observed value was defined as "b," and the predictions between the two models were defined as "c." In the relaxed case, gene pairs where the two models produced predictions that were closer to each other than either was to the observed expression value of the maize2 gene under cold stress were excluded. That is, if $c<a$ and $c<b$, the multiplicative model works better than the additive model, while if $b<a$ and $b<c$, the additive model works better than the other model. In the most stringent case, gene pairs where the two models produced predictions that were less than twice as large as the difference between the better model and the observed value were excluded (Supplemental Data Set 4). In other words, if $b>2a$ and $b>c$, the multiplicative model was considered to be the better model; if $c>2a$ and $c>b$, the additive model was considered to be the better model. Analyses were also conducted reciprocally using data from control and cold stress conditions in maize2 plus data from maize1 under control conditions to predict the expression of the maize1 gene under cold stress conditions.

Identifying DROs

DROs were identified using count data generated as described above and an interaction term for species (maize or sorghum) and treatment (cold or control) in DESeq2 (Love et al., 2014). Species (maize and sorghum) and condition (cold and control) were considered to be two factors for design in this analysis. Simulated data for CROs generated using additive and multiplicative models were used to confirm that this approach did not classify simulated CROs based on the multiplicative model as having significant species-by-treatment interactions. The formula used was as follows: design _ condition + genotype + condition: genotype. Maize sorghum gene pairs with an interaction adjusted P value ≤ 0.001 were classified as DROs, those with interaction adjusted P value ≥ 0.05 were classified as CROs, and those with intermediate P values were disregarded (Yoav and Yosef, 1995). The decision was made to retain an ambiguous case of gene pairs with interaction P values too high to be classified as DROs but too significant to be classified as CROs rather than increase the number of classification errors by forcing all gene pairs to be assigned to one category or the other.

Calculating Ka/Ks Values

"Primary transcript only" coding sequences for maize (v6a), sorghum (v3.1), and setaria (v2.2) were retrieved from Phytozome version 12.0. The gene model annotations v6a for maize were annotated onto the B73 RefGen v3 pseudomolecules. Coding sequences were translated to protein sequences and aligned using Kalign version 2.04 (Lassmann and Sonnhammer, 2005). The protein alignment was used as a guide to create a codon level alignment of coding sequences. The codon alignment was supplied to PAML (version 4.09) (Yang, 2007). Synonymous and nonsynonymous substitution rates were calculated independently for each branch of the tree. When both maize1 and maize2 gene copies were present for the same syntenic gene group, alignment and substitution rate calculations were conducted separately for the maize1 gene and its syntenic orthologs in sorghum and setaria and for the maize2 and the same syntenic orthologous genes. To eliminate genes with extreme Ka/Ks ratios resulting from very low numbers of synonymous substitutions, only Ka/Ks ratios from genes with an estimated synonymous substitution rate greater than or equal to 0.05 (~1/2 the median Ks ratio observed between maize and the most common recent ancestor of maize and sorghum) were considered.

MNase HS Analysis

Intervals defined as MNase HSs were taken from Rodgers-Melnick et al. (2016). The same TSS was used for MNase and RNA-seq analysis. Average coverage of MNase HS was calculated on a per-base basis from 1 kb upstream of the annotated TSS to 1 kb downstream of the TSS. When multiple transcripts with different TSS were present, the transcript with the earliest TSS was selected for analysis.

Identifying CNSs

CNSs were identified using the CNS Discovery Pipeline 3.0 (CDP) (Turco et al., 2013) with some modifications. Specifically, the built-in syntenic gene identification pipeline from the CDP was replaced with the previously defined syntenic gene list described above. Functions for finding local duplicates and comparing CNSs to Arabidopsis proteins and RNA were omitted. CNSs were identified between the region 12 kb upstream and 12 kb downstream using a word size of 15 bp. CNSs with bit scores for each gene pair < 29.5 were removed following the same scoring parameter settings outlined in the original software pipeline.

Transcription Factor Enrichment Calculation

Transcription factor enrichment was calculated using the maize transcription factor list from GRASSIUS (Yilmaz et al., 2009).

GO Enrichment Analysis

GO analysis was performed using GOATOOLS (Haibao et al., 2015) and functional additions associated with the sorghum v3.1 sorghum gene model and maize RefGen-v3 maize gene model annotations.

Pathway Analysis

Pathway analysis was conducted using the MapMan software package (http://mapman.gabipd.org/web/guest) (Usadel et al., 2009).

Accession Numbers

Gene IDs for all syntenic gene sets and the final syntenic gene list used in this study are posted at figShare (http://dx.doi.org/10.6084/m9.figshare.3113488.v1). Adapter sequences used for library construction and for adapter trimming are those provided in Illumina TruSeq Library Prep Pooling Guide, with sequences reported on page 5 of the user manual.

Supplemental Data

Supplemental Figure 1. Coding sequence similarity among syntenic genes in sorghum, maize1, and maize2.

Supplemental Figure 2. Representative sample of cold stressed seedling phenotypes.

Supplemental Figure 3. Individual examples of genes in each of six possible DRO/DEG classification categories.

Supplemental Figure 4. Comparison of Ka/Ks ratio and expression level for genes grouped based on expression classification model.

Supplemental Figure 5. Frequency of known CBF binding motifs within the 1-kb proximal promoters of maize and sorghum.

Supplemental Figure 6. Relationship between gene pair expression pattern in maize and sorghum after subdividing genes based on Ka/Ks ratio and expression tertile.

Supplemental Data Set 1. Number of sequenced and aligned reads per library.

Supplemental Data Set 2. Estimates of power and FDR for DESeq2.

Supplemental Data Set 3. ANOVA and DESeq2 tests for DROs using simulated data.

Supplemental Data Set 4. Accuracy of additive and multiplicative expression models across maize duplicate gene pairs.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Institute of Food and Agriculture, U.S. Department of Agriculture under Award 16-67013-24613 to R.L.R. and J.C.S. This material is based on work supported by the National Science Foundation under Grant OIA-1557417. In addition, this work was supported by start-up funding from the University of Nebraska-Lincoln to R.L.R., Y.Q., and J.C.S.

AUTHOR CONTRIBUTIONS

J.C.S. and R.L.R. conceived the project and designed the studies. Y.Z., D.W.N., D.C., and Z.L. performed the research. Y.Z. and Y.Q. analyzed the data. Y.Z., J.C.S., and R.L.R. wrote the article. All authors reviewed the manuscript.

Received May 5, 2017; revised July 5, 2017; accepted July 18, 2017; published July 21, 2017.

REFERENCES

Anders, S., Pyl, P.T., and Huber, W. (2015). HTSeq–a Python framework to work with high-throughput sequencing data. Bioinformatics **31:** 166–169.

Bennetzen, J.L., and Freeling, M. (1993). Grasses as a single genetic system: genome composition, collinearity and compatibility. Trends Genet. **9:** 259–261.

Burkart-Waco, D., Ngo, K., Lieberman, M., and Comai, L. (2015). Perturbation of parentally biased gene expression during interspecific hybridization. PLoS One **10:** e0117293.

Chan, Y.F., et al. (2010). Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 enhancer. Science **327:** 302–305.

Chinnusamy, V., Zhu, J., and Zhu, J.K. (2007). Cold stress regulation of gene expression in plants. Trends Plant Sci. **12:** 444–451.

Chopra, R., Burow, G., Hayes, C., Emendack, Y., Xin, Z., and Burke, J. (2015). Transcriptome profiling and validation of gene based single nucleotide polymorphisms (SNPs) in sorghum genotypes with contrasting responses to cold stress. BMC Genomics **16:** 1040.

Chopra, R., Burow, G., Burke, J.J., Gladman, N., and Xin, Z. (2017). Genome-wide association analysis of seedling traits in diverse Sorghum germplasm under thermal stress. BMC Plant Biol. **17:** 12.

Christin, P.A., Salamin, N., Savolainen, V., Duvall, M.R., and Besnard, G. (2007). C_4 Photosynthesis evolved in grasses via parallel adaptive genetic changes. Curr. Biol. **17:** 1241–1247.

Cooper, G.M., Stone, E.A., Asimenos, G., Green, E.D., Batzoglou, S., and Sidow, A.; NISC Comparative Sequencing Program (2005). Distribution and intensity of constraint in mammalian genomic sequence. Genome Res. **15:** 901–913.

Cui, S., Huang, F., Wang, J., Ma, X., Cheng, Y., and Liu, J. (2005). A proteomic analysis of cold stress responses in rice seedlings. Proteomics **5:** 3162–3172.

Davidson, R.M., Gowda, M., Moghe, G., Lin, H., Vaillancourt, B., Shiu, S.-H., Jiang, N., and Robin Buell, C. (2012). Comparative transcriptomics of three Poaceae species reveals patterns of gene expression evolution. Plant J. **71:** 492–502.

De Wet, J. (1978). Systematics and evolution of sorghum sect. sorghum (gramineae). Am. J. Bot. **65:** 477–484.

Eichten, S.R., et al. (2011). Heritable epigenetic variation among maize inbreds. PLoS Genet. **7:** e1002372.

Ercoli, L., Mariotti, M., Masoni, A., and Arduini, I. (2004). Growth responses of sorghum plants to chilling temperature and duration of exposure. Eur. J. Agron. **21:** 93–103.

Fiedler, K., Bekele, W.A., Matschegewski, C., Snowdon, R., Wieckhorst, S., Zacharias, A., and Uptmoor, R. (2016). Cold tolerance during juvenile development in sorghum: a comparative analysis by genomewide association and linkage mapping. Plant Breed. **135:** 598–606.

Freeling, M., Rapaka, L., Lyons, E., Pedersen, B., and Thomas, B.C. (2007). G-boxes, bigfoot genes, and environmental response: characterization of intragenomic conserved noncoding sequences in Arabidopsis. Plant Cell **19:** 1441–1457.

Freeling, M., Woodhouse, M.R., Subramaniam, S., Turco, G., Lisch, D., and Schnable, J.C. (2012). Fractionation mutagenesis and similar consequences of mechanisms removing dispensable or less-expressed DNA in plants. Curr. Opin. Plant Biol. **15:** 131–139.

Haibao, T., Klopfenstein, D., Pedersen, B., Flick, P., Sato, K., Ramirez, F., Yunes, J., and Mungall, C. (2015). Goatools: Tools for Gene Ontology. http://dx.doi.org/10.5281/zenodo.31628.

Harris, R.S. (2007). Improved Pairwise Alignment of Genomic DNA. PhD dissertation (State College, PA: Pennsylvania State University).

Hetherington, S.E., He, J., and Smillie, R.M. (1989). Photoinhibition at low temperature in chilling-sensitive and -resistant plants. Plant Physiol. **90:** 1609–1615.

Hollister, J.D., and Gaut, B.S. (2009). Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression. Genome Res. **19:** 1419–1428.

Hollister, J.D., Smith, L.M., Guo, Y.-L., Ott, F., Weigel, D., and Gaut, B.S. (2011). Transposable elements and small RNAs contribute to gene expression divergence between *Arabidopsis thaliana* and *Arabidopsis lyrata.* Proc. Natl. Acad. Sci. USA **108:** 2322–2327.

Lai, X., Behera, S., Liang, Z., Lu, Y., Deogun, J.S., and Schnable, J.C. (2017). Stag-CNS: An order-aware conserved non-coding sequences discovery tool for arbitrary numbers of species. Mol. Plant. **10:** 990–999.

Lassmann, T., and Sonnhammer, E.L. (2005). Kalign–an accurate and fast multiple sequence alignment algorithm. BMC Bioinformatics **6:** 298.

Li, L., Briskine, R., Schaefer, R., Schnable, P.S., Myers, C.L., Flagel, L.E., Springer, N.M., and Muehlbauer, G.J. (2016). Co-expression network analysis of duplicate genes in maize (*Zea mays L.*) reveals no subgenome bias. BMC Genomics **17:** 875.

Liang, Z., and Schnable, J.C. (2016). RNA-seq based analysis of population structure within the maize inbred B73. PLoS One **11:** e0157942.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. **15:** 550.

Lovell, J.T., et al. (2016). Drought responsive gene expression regulatory divergence between upland and lowland ecotypes of a perennial C4 grass. Genome Res. **26:** 510–518.

Makarevitch, I., Waters, A.J., West, P.T., Stitzer, M., Hirsch, C.N., Ross-Ibarra, J., and Springer, N.M. (2015). Transposable elements contribute to activation of maize genes in response to abiotic stress. PLoS Genet. **11:** e1004915.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal **17:**10–12.

Muiño, J.M., de Bruijn, S., Pajoro, A., Geuten, K., Vingron, M., Angenent, G.C., and Kaufmann, K. (2016). Evolution of dna-binding sites of a floral master regulatory transcription factor. Mol. Biol. Evol. **33:** 185–200.

Naito, K., Zhang, F., Tsukiyama, T., Saito, H., Hancock, C.N., Richardson, A.O., Okumoto, Y., Tanisaka, T., and Wessler, S.R. (2009). Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. Nature **461:** 1130–1134.

Ng, P.C., and Henikoff, S. (2001). Predicting deleterious amino acid substitutions. Genome Res. **11:** 863–874.

Olsen, J., McMahon, C., and Hammer, G. (1993). Prediction of sweet corn phenology in subtropical environments. Agron. J. **85:** 410–415.

Orr, H.A. (1998). The population genetics of adaptation: the distribution of factors fixed during adaptive evolution. Evolution **52:** 935–949.

Orr, H.A. (1999). The evolutionary genetics of adaptation: a simulation study. Genet. Res. **74:** 207–214.

Paschold, A., Larson, N.B., Marcon, C., Schnable, J.C., Yeh, C.-T., Lanz, C., Nettleton, D., Piepho, H.-P., Schnable, P.S., and Hochholdinger, F. (2014). Nonsyntenic genes drive highly dynamic complementation of gene expression in maize hybrids. Plant Cell **26:** 3939–3948.

Paterson, A.H., et al. (2009). The *Sorghum bicolor* genome and the diversification of grasses. Nature **457:** 551–556.

Pearce, R.S. (2001). Plant freezing and damage. Ann. Bot. (Lond.) **87:** 417–424.

Priest, H.D., Fox, S.E., Rowley, E.R., Murray, J.R., Michael, T.P., and Mockler, T.C. (2014). Analysis of global gene expression in *Brachypodium distachyon* reveals extensive network plasticity in response to abiotic stress. PLoS One **9:** e87499.

Quinn, P.J. (1988). Effects of temperature on cell membranes. Symp. Soc. Exp. Biol. **42:** 237–258.

Reva, B., Antipin, Y., and Sander, C. (2011). Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res. **39:** e118.

Revilla, P., et al. (2016). Association mapping for cold tolerance in two large maize inbred panels. BMC Plant Biol. **16:** 127.

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. **43:** e47.

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics **26:** 139–140.

Rodgers-Melnick, E., Vera, D.L., Bass, H.W., and Buckler, E.S. (2016). Open chromatin reveals the functional maize genome. Proc. Natl. Acad. Sci. USA **113:** E3177–E3184.

Sánchez, B., Rasmussen, A., and Porter, J.R. (2014). Temperatures and the growth and development of maize and rice: a review. Glob. Change Biol. **20:** 408–417.

Schaefer, R.J., Briskine, R., Springer, N.M., and Myers, C.L. (2014). Discovering functional modules across diverse maize transcriptomes using COB, the Co-expression Browser. PLoS One **9:** e99193.

Schnable, J.C. (2015). Genome evolution in maize: from genomes back to genes. Annu. Rev. Plant Biol. **66:** 329–343.

Schnable, J.C., Springer, N.M., and Freeling, M. (2011). Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. Proc. Natl. Acad. Sci. USA **108:** 4069–4074.

Schnable, J.C., Freeling, M., and Lyons, E. (2012). Genome-wide analysis of syntenic gene deletion in the grasses. Genome Biol. Evol. **4:** 265–277.

Schnable, P.S., et al. (2009). The B73 maize genome: complexity, diversity, and dynamics. Science **326:** 1112–1115.

Schurch, N.J., Schofield, P., Gierliński, M., Cole, C., Sherstnev, A., Singh, V., Wrobel, N., Gharbi, K., Simpson, G.G., Owen-Hughes, T., Blaxter, M., and Barton, G.J. (2016). How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? RNA **22:** 839–851.

Shantz, H. (1954). The place of grasslands in the earth's cover. Ecology **35:** 143–145.

Shaykewich, C. (1995). An appraisal of cereal crop phenology modelling. Can. J. Plant Sci. **75:** 329–341.

Singer, S.J., and Nicolson, G.L. (1972). The fluid mosaic model of the structure of cell membranes. Science **175:** 720–731.

Studer, A., Zhao, Q., Ross-Ibarra, J., and Doebley, J. (2011). Identification of a functional transposon insertion in the maize domestication gene tb1. Nat. Genet. **43:** 1160–1163.

Swigonová, Z., Lai, J., Ma, J., Ramakrishna, W., Llaca, V., Bennetzen, J.L., and Messing, J. (2004). Close split of sorghum and maize genome progenitors. Genome Res. **14:** 1916–1923.

Tenhaken, R. (2014). Cell wall remodeling under abiotic stress. Front. Plant Sci. **5:** 771.

Tiwari, S., Sl, K., Kumar, V., Singh, B., Rao, A.R., Mithra Sv, A., Rai, V., Singh, A.K., and Singh, N.K. (2016). Mapping qtls for salt tolerance in rice (*Oryza sativa* L.) by bulked segregant analysis of recombinant inbred lines using 50k snp chip. PLoS One **11:** e0153610.

Torres-Oliva, M., Almudi, I., McGregor, A.P., and Posnien, N. (2016). A robust (re-)annotation approach to generate unbiased mapping references for RNA-seq-based analyses of differential expression across closely related species. BMC Genomics **17:** 392.

Turco, G., Schnable, J.C., Pedersen, B., and Freeling, M. (2013). Automated conserved non-coding sequence (CNS) discovery reveals differences in gene content and promoter evolution among grasses. Front. Plant Sci. **4:** 170.

Usadel, B., Poree, F., Nagel, A., Lohse, M., Czedik-Eysenberg, A., and Stitt, M. (2009). A guide to using MapMan to visualize and compare Omics data in plants: a case study in the crop species, Maize. Plant Cell Environ. **32:** 1211–1229.

van Heerwaarden, J., Doebley, J., Briggs, W.H., Glaubitz, J.C., Goodman, M.M., de Jesus Sanchez Gonzalez, J., and Ross-Ibarra, J. (2011). Genetic signals of origin, spread, and introgression in a large sample of maize landraces. Proc. Natl. Acad. Sci. USA **108:** 1088–1092.

Vera, D.L., Madzima, T.F., Labonne, J.D., Alam, M.P., Hoffman, G.G., Girimurugan, S.B., Zhang, J., McGinnis, K.M., Dennis, J.H., and Bass, H.W. (2014). Differential nuclease sensitivity profiling of chromatin reveals biochemical footprints coupled to gene expression and functional DNA elements in maize. Plant Cell **26:** 3883–3893.

Wallace, J.G., Larsson, S.J., and Buckler, E.S. (2014). Entering the second century of maize quantitative genetics. Heredity (Edinb) **112:** 30–38.

Wang, L., et al. (2014). Comparative analyses of C_4 and C_3 photosynthesis in developing leaves of maize and rice. Nat. Biotechnol. **32:** 1158–1165.

Waters, A.J., Makarevitch, I., Noshay, J., Burghardt, L.T., Hirsch, C.N., Hirsch, C.D., and Springer, N.M. (2017). Natural variation for gene expression responses to abiotic stress in maize. Plant J. **89:** 706–717.

Wendorf, F., Close, A.E., Schild, R., Wasylikowa, K., Housley, R.A., Harlan, J.R., and Krolik, H. (1992). Saharan exploitation of plants 8,000 years bp. Nature **359:** 721–724.

Wu, T.D., and Nacu, S. (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics **26:** 873–881.

Wu, T.D., and Watanabe, C.K. (2005). GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics **21:** 1859–1875.

Xu, G., Wang, X., Huang, C., Xu, D., Li, D., Tian, J., Chen, Q., Wang, C., Liang, Y., Wu, Y., Yang, X., and Tian, F. (2017). Complex genetic architecture underlies maize tassel domestication. New Phytol. **214:** 852–864.

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. **24:** 1586–1591.

Yilmaz, A., Nishiyama, M.Y., Jr., Fuentes, B.G., Souza, G.M., Janies, D., Gray, J., and Grotewold, E. (2009). GRASSIUS: a platform for comparative regulatory genomics across the grasses. Plant Physiol. **149:** 171–180.

Yoav, B.Y., and Yosef, H. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. B **57:** 289–300.

Zhang, W., Wu, Y., Schnable, J.C., Zeng, Z., Freeling, M., Crawford, G.E., and Jiang, J. (2012). High-resolution mapping of open chromatin in the rice genome. Genome Res. **22:** 151–162.

Zhang, Y., Ding, Z., Ma, F., Chauhan, R.D., Allen, D.K., Brutnell, T.P., Wang, W., Peng, M., and Li, P. (2015). Transcriptional response to petiole heat girdling in cassava. Sci. Rep. **5:** 8414.

the plant journal

SEB Society for Experimental Biology

The Plant Journal (2018) **93,** 843–855 doi: 10.1111/tpj.13806

Largely unlinked gene sets targeted by selection for domestication syndrome phenotypes in maize and sorghum

Xianjun Lai[1,2], Lang Yan[1,3,4], Yanli Lu[2] and James C. Schnable[1,*]

[1]*Center for Plant Science Innovation and Department of Agronomy and Horticulture, University of Nebraska-Lincoln, NE 68588, USA,*

[2]*Maize Research Institute, Sichuan Agricultural University, Chengdu 611130, China,*

[3]*Laboratory of Functional Genome and Application of Potato, Xichang College, Liangshan 615000, China, and*

[4]*College of Life Sciences, Sichuan University, Chengdu 610065, China*

Received 4 September 2017; revised 27 November 2017; accepted 4 December 2017; published online 19 December 2017.

*For correspondence (e-mail schnable@unl.edu).

Xianjun Lai and Lang Yan are the authors contributed equally to this work.

SUMMARY

The domestication of diverse grain crops from wild grasses was a result of artificial selection for a suite of overlapping traits producing changes referred to in aggregate as 'domestication syndrome'. Parallel phenotypic change can be accomplished by either selection on orthologous genes or selection on non-orthologous genes with parallel phenotypic effects. To determine how often artificial selection for domestication traits in the grasses targeted orthologous genes, we employed resequencing data from wild and domesticated accessions of *Zea* (maize) and *Sorghum* (sorghum). Many 'classic' domestication genes identified through quantitative trait locus mapping in populations resulting from wild/domesticated crosses indeed show signatures of parallel selection in both maize and sorghum. However, the overall number of genes showing signatures of parallel selection in both species is not significantly different from that expected by chance. This suggests that while a small number of genes will extremely large phenotypic effects have been targeted repeatedly by artificial selection during domestication, the optimization part of domestication targeted small and largely non-overlapping subsets of all possible genes which could produce equivalent phenotypic alterations.

Keywords: grasses, domestication, selection, parallel evolution, *Zea mays*, *Sorghum bicolor*.

INTRODUCTION

The characteristics of modern crops are the result of thousands of years of artificial selection applied consciously or unconsciously by farmers and plant breeders. An estimated 2500 plant species have experienced some degree of artificial selection, with approximately 10% of these being domesticated to the point that the species depends on humans for survival (Dirzo and Raven, 2003). Of the hundreds of crops domesticated by human civilizations, three species – rice, wheat and maize – provide more than half of all calories consumed around the world. These three crops all belong to the same family Poaceae (the grasses), a clade that has contributed a total of at least 48 domesticated crop species to human civilization, including at least 30 species domesticated as sources of grain (Glémin and Bataillon, 2009). Artificial selection for grain production produced a suite of shared phenotypic changes in grain crops referred to as 'the domestication syndrome' (Harlan *et al.*, 1973). Grain crop domestication syndrome includes loss of seed shattering, increased apical dominance, more uniform maturity across inflorescences and across tillers, increase in size and/or number of inflorescences, larger seeds, greater carbohydrate content and lower protein content per seed, and reduction or loss of seed dormancy (Harlan *et al.*, 1973). The genes involved in producing domestication syndrome phenotypes can be identified through two broad types of studies which are sometimes referred to as 'top-down' and 'bottom-up' approaches (Ross-Ibarra *et al.*, 2007).

Top-down approaches utilize quantitative genetic studies to identify large-effect genes involved in producing the changes associated with the domestication syndrome in crop species that are interfertile with their wild progenitors. In maize, an estimated five loci have large enough effects on domestication traits to be mapped using conventional quantitative trait locus (QTL) analysis (Doebley and Stec,

1993). Of these loci, several have been mapped including *teosinte branched1* (*tb1*), where the allele selected for significantly reduces the development of tillers (Clark *et al.*, 2004), *teosinte glume architecture1* (*tga1*), where the allele selected for abolishes the stony fruitcase surrounding teosinte seeds (Dorweiler *et al.*, 1993), and *grassy tillers1* (*gt1*), where the allele selected for results in many fewer ears per plant during domestication (Whipple *et al.*, 2011; Wills *et al.*, 2013). In rice, many functionally characterized genes that underlie phenotypic changes during the domestication process have also been identified through QTL mapping, such as *Seed dormancy 4* (*Sdr4*), where the allele selected for produces a reduction in seed dormancy (Sugimoto *et al.*, 2010), *Tiller Angle Control1* (*TAC1*), where the allele selected for decreases tiller angle, producing more photosynthetically efficient canopy architecture (Yu *et al.*, 2007), and *betaine aldehyde dehydrogenase2* (*BADH2*), a loss-of-function allele selected for during the domestication process that results in the accumulation of 2-acetyl-1-pyrroline in fragrant rice (Bradbury *et al.*, 2005).

In contrast, bottom-up approaches use changes in the diversity and frequency of haplotypes at particular regions of the genome between populations of a crop species and populations of wild relatives to identify loci which were targets of artificial selection. Notably, while quantitative genetic evaluation of recombinant populations generally identifies relatively small numbers of large-effect loci responsible for many of the differences observed between domesticated grain crops and their wild relatives, genome-wide population genetic approaches generally identify hundreds to thousands of loci as targets of selection during domestication in the same species. A similar orders of magnitude difference between candidate gene (top-down) and selection scans (bottom-up) has been noted in other systems, including studies of positive selection in humans (Akey, 2009). Hufford and co-workers used resequencing data from a set of 75 teosinte and maize lines to identify 484 regions of the maize genome likely to have experienced selection during transition from wild teosintes to maize landraces and another 695 regions likely to have experienced selection during transition from largely tropical landraces to largely temperate elite lines (Hufford *et al.*, 2012). Huang and co-workers used genome resequencing data of 1083 cultivated rice and 446 wild rice accessions to identify 55 selection regions encompassed 2547 candidate artificially selected genes during domestication from wild to cultivated rice (Huang *et al.*, 2012). In sorghum, a set of 725 genes which were likely the targets of artificial selection during domestication and/or crop improvement were identified from resequencing of 44 lines of sorghum and wild relatives (Mace *et al.*, 2013). However, one critical limitation of bottom-up approaches is that candidate genes identified using these techniques will not initially be linked to any specific phenotypic trait (Ross-Ibarra *et al.*, 2007).

Parallel phenotypic changes which are part of the domestication syndrome in grain crops could result from parallel or lineage-specific changes at a molecular level. A recently published study demonstrated the loss of seed shattering resulted from disruption of the same gene in maize, sorghum and rice (Lin *et al.*, 2012). *Heading Date1* is a major QTL controlling flowering time which shows evidence of being under parallel artificial selection during the process of domestication for sorghum, setaria and rice (Liu *et al.*, 2015). A flowering time QTL identified in a population of wild × domesticated *Setaria* lines co-localizes with a flowering time QTL identified at syntenic orthologous locations in the genomes of maize and sorghum (Mauro-Herrera *et al.*, 2013). A significant number of candidate genes associated with seed size exhibited signals of parallel selection during domestication in maize, rice and sorghum (Tao *et al.*, 2017). However, not all parallel phenotypic changes produced by artificial selection result from parallel evolution at the molecular level. Artificial selection for adaption to high altitudes in different maize populations targeted largely unrelated sets of genes in Mexican and Andean highland populations (Takuno *et al.*, 2015).

Here we focus on two grain crops, maize (*Zea mays* ssp. *mays*) and sorghum (*Sorghum bicolor* ssp. *bicolor*). Maize was domesticated from Balsas teosinte (*Zea mays* ssp *parviglumis*) in Mesoamerica, with a center of origin in the lowlands of southwestern Mexico (Van Heerwaarden *et al.*, 2011). Sorghum is believed to have first been domesticated from broomcorn (*Sorghum bicolor* ssp. *verticilliflorum*) in Ethiopia (Wendorf *et al.*, 1992), with a potential second independent domestication in west Africa (Sagnard *et al.*, 2011; Mace *et al.*, 2013). The wild ancestors of these two crops diverged approximately 12 million years ago (Swigoˇová *et al.*, 2004). Subsequent to the divergence of these two lineages, maize experienced a whole-genome duplication creating two functionally distinct subgenomes – maize1 and maize2 – each of which is, in principle orthologous to the entire genome of sorghum (Schnable *et al.*, 2011). In some cases orthologs of a single sorghum gene from both subgenomes are still present, creating a pair of maize genes which are co-orthologous to a single sorghum gene. In other cases the maize1 or maize2 gene copy was lost from the genome after the whole-genome duplication, restoring a 1:1 orthologous relationship between the two species. Maize2 gene copies have been lost from the genome more frequently than maize1 gene copies, tend to be expressed to lower mRNA levels and, on average, tend to explain less phenotypic variation than maize1 gene copies (Schnable and Freeling, 2011; Renny-Byfield *et al.*, 2017). The parallel set of phenotypic changes during domestication in these two species (Harlan *et al.*, 1973), the high degrees of conserved collinearity across grass genomes (Bennetzin and Freeling, 1993; Moore *et al.*, 1995) and the

bias towards genes with detectable phenotypic effects being conserved at syntenic locations across grass genomes (Schnable, 2015) offer an opportunity to test the hypothesis that the parallel phenotypic changes in maize and sorghum resulted from artificial selection acting on orthologous genes in both species.

We found that genes conserved at syntenic orthologous locations in maize and sorghum were significantly more likely to be targets of selection during domestication than non-syntenic genes unique to one species. In maize, domestication preferentially targeted genes on the dominant maize1 subgenome rather than their retained duplicates on the maize2 subgenome. Consistent with a much earlier study of maize and sorghum QTLs controlling domestication phenotypes (Paterson *et al.*, 1995), genes identified through quantitative genetic studies of domestication traits in one species were likely to show signatures of selection in the other species. However, the overall overlap between genes identified using population genetic methods in both species was only marginally greater than expected by chance.

RESULTS

Population genetic datasets for both species

Maize and sorghum accessions were sampled from published datasets (Chia *et al.*, 2012; Mace *et al.*, 2013; Luo *et al.*, 2016) (see Experimental Procedures and Table S1 in the Supporting Information). After quality filtering to remove low-quality single nucleotide polymorphisms (SNPs) and those potentially representing alignments of paralogous sequence elsewhere in the genome (see Experimental Procedures), a total of 10.3 million segregating SNPs in 56 maize accessions and 3.3 million segregating SNPs in 42 sorghum accessions remained. These proportions roughly correspond to the difference in genome size between the two species (approximately 2.0 Gb for maize and 700 Mb for sorghum); however, as much of the maize genome is repetitive and cannot be uniquely identified using short sequence reads, this is consistent with higher overall levels of nucleotide diversity in maize relative to sorghum. Also consistent with previous reports, wild relatives had higher levels of nucleotide diversity ($\pi = 0.00377$ in maize and $\pi = 0.00381$ in sorghum), than both landraces ($\pi = 0.00338$ in maize and $\pi = 0.00242$ in sorghum) and improved inbreds ($\pi = 0.00334$ in maize and $\pi = 0.00226$ in sorghum) (Table S3) (Hufford *et al.*, 2012; Mace *et al.*, 2013).

Maize accessions were primarily collected in the Western Hemisphere and sorghum accessions primarily from the Eastern Hemisphere, with some exceptions in both cases (Figure 1a). Wild relatives were primarily collected near the centers of domestication: southwestern Mexico for maize and central east Africa for sorghum. The sorghum dataset also included data for a forage sudangrass line (Greenleaf: sweet sorghum × Sudan grass) from North America. Among the maize lines, modern elite lines and wild relatives each formed distinct clades (Figure 1b). In sorghum, wild relatives formed a distinct clade; however, lines reported to be elite or landrace lines were intercalated, potentially as a result of distinct sorghum breeding efforts developing lines for different agroclimactic zones around the world (Figure 1c).

Genetic maps for both species were sourced from public datasets. For maize, a genetic map was employed that included 10 085 markers genotyped in a set of 232 recombinant inbred lines (RILs) from the maize IBM population using tGBS (Zou *et al.*, 2012; Ott *et al.*, 2017) while for sorghum a genetic map was employed which was constructed from a set of 3418 markers genotyped in a set of 244 RILs from a grain sorghum × sweet sorghum cross using resequencing (Zou *et al.*, 2012).

Genomic signals of selection in maize and sorghum

In each species, the identification of regions under selection was performed for three separate pairwise comparisons: landraces versus wild relatives (domestication), improved lines versus landraces (improvement), and improved lines versus wild relatives. The genome was scanned with XP-CLR using a window size of 0.05 cM and a step size of 1 kb (see Experimental Procedures) and each gene was assigned the XP-CLR score of the highest scoring bin that overlapped with the gene. Genes above the 90th percentile of XP-CLR scores for a given pairwise comparison were considered as candidate 'under selection' genes. The set of gene annotations employed in this analysis included 63 480 maize gene models and 34 027 sorghum gene models. Thus, for each of the three possible pairwise comparisons, 6348 genes in maize and 3403 genes in sorghum were identified as candidates for selection (Figure S1). In both species, estimated selection coefficients were higher during domestication (mean $s = 0.06$ in maize and 0.047 in sorghum) than improvement (mean $s = 0.045$ in maize and 0.024 in sorghum).

Since 10% of genes were identified as candidates for selection during domestication and 10% of genes were identified as candidates for selection during improvement, the overlap expected if these datasets are unrelated is 1%. In fact, approximately 1% of all annotated maize genes (620 genes) were in the top 10% in both comparisons, and approximately 1% of all annotated sorghum genes (345 genes) were in the top 10% in both comparisons (Figure 2a-b). As expected, the set of candidate genes identified in the comparison of wild relatives and improved lines showed significant overlap with both the domestication and crop improvement candidate gene sets (Figure 2a). Among a set of 112 classical maize mutants cloned using forward genetics (Schnable and Freeling, 2011), 23 were

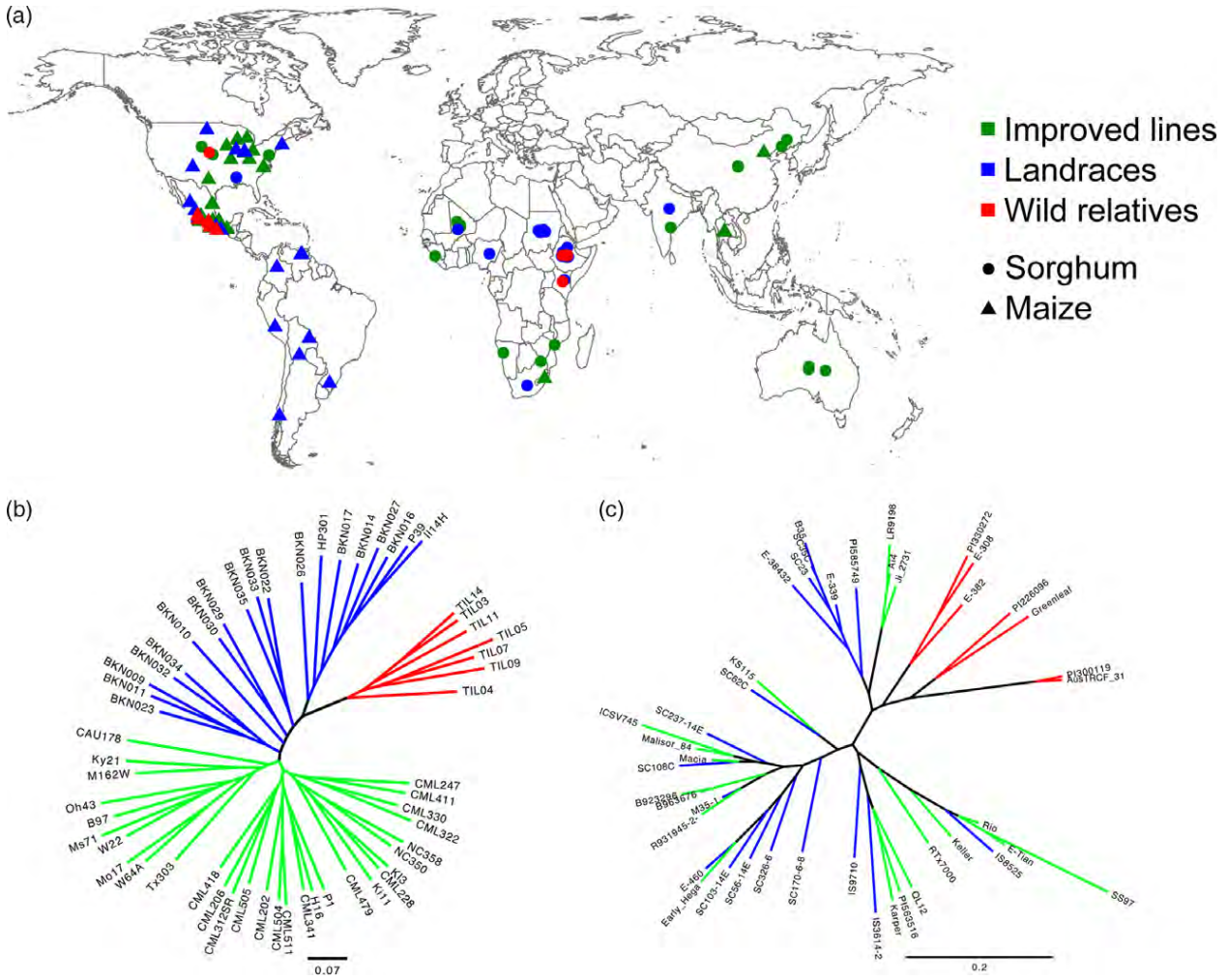


Figure 1. Geographic distribution and phylogenetic relationship of maize and sorghum accessions employed in this study.
(a) The geographic distribution of maize and sorghum accessions. Dots and triangles represent sorghum and maize, respectively. Colors represent membership in different populations: improved lines (green), landraces (blue) and wild relatives (red).
(b), (c) Neighbor-joining (NJ) trees for maize accessions (b) and sorghum accessions (c). Taxa in the NJ tree are color coded using the same system as in (a).

identified as candidate genes under selection during domestication and/or improvement (*P*-value = 0.0007, Fisher exact test).

The maize and sorghum genetic maps are largely collinear (Moore *et al.*, 1995); however, maize has experienced a whole-genome duplication relative to sorghum, producing two functionally distinct subgenomes with differing levels of gene expression, purifying selection and loss rates for both genes and conserved regulatory sequences (Schnable and Freeling, 2011; Pophaly and Tellier, 2015; Lai *et al.*, 2017). In a dataset of 14 433 genes conserved at syntenic locations in sorghum and maize, 7041 genes are conserved between sorghum and maize1 but lost from maize2 and 3031 genes are conserved between sorghum and maize2 but lost from maize1; the balance (4361 genes) are present in all three genomes (Table S2) (Schnable *et al.*, 2016). The balance of the gene complement of each species consists of non-syntenic genes, which are the majority (57.6% and 70.4% for sorghum and maize, respectively) of all annotated genes in both species. Genes with known mutant phenotypes tend to be syntenic (Schnable and Freeling, 2011), while non-syntenic genes tend to exhibit greater allelic variation in gene regulation, and it has been speculated that they may contribute to phenotypic plasticity and heterosis (Paschold *et al.*, 2014; Baldauf *et al.*, 2016; Li *et al.*, 2016).

In sorghum, genes conserved at syntenic orthologous locations between maize and sorghum were significantly more likely to be identified as targets of selection in all three pairwise comparisons (binomial testing *P*-value $< 1.0 \times 10^{-6}$ for the domestication and improvement, *P*-value = 0.0058 for the wild–improved comparison). In maize, syntenic genes were significantly more likely to be identified as targets of selection during domestication and

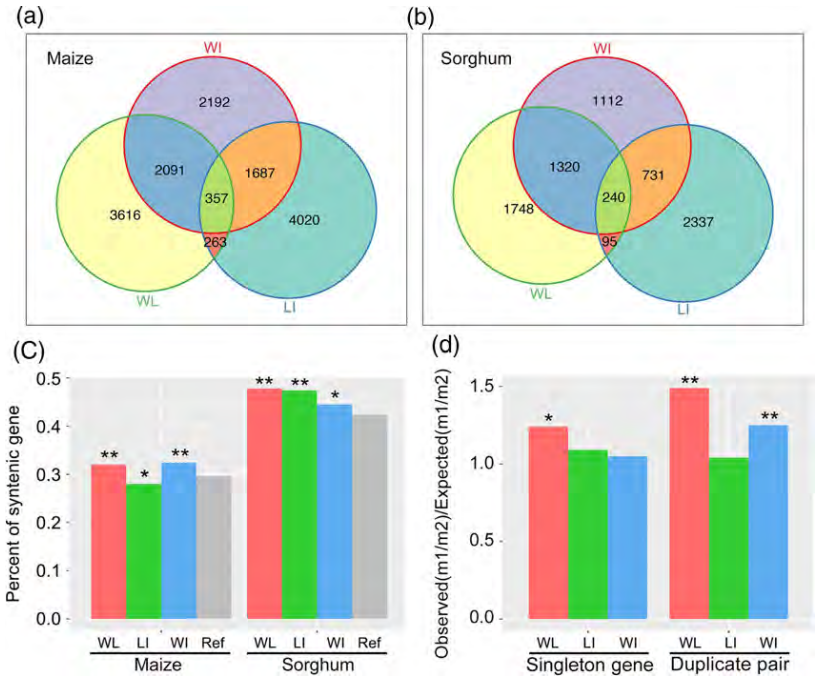


Figure 2. Summary information for candidate genes under selection.
(a), (b) The number of candidate genes shared among the three pairwise comparisons of populations in maize (a) and sorghum (b), respectively.
(c) The proportion of non-syntenic genes and syntenic genes under selection in pairwise comparison in maize and sorghum.
(d) Ratio of maize1:maize2 genes among genes identified as selection candidates.
In (c) and (d) analysis was conducted separately for singleton genes and duplicate genes. One asterisk denotes cases which are significantly different at a threshold of $P < 0.01$ and two asterisks denote cases which were significantly different at a threshold of $P < 0.001$. I, improved lines; L, landraces; W, wild relatives. [Colour figure can be viewed at wileyonlinelibrary.com].

in the wild relative improved line comparison (*P*-value = 0.00024 and *P*-value = 1.0×10^{-6}, respectively); however, genes identified as targets of selection during the crop improvement process were significantly more likely to be non-syntenic genes (*P*-value = 0.0026) (Figure 2c).

In maize, selection candidates were also unevenly distributed between subgenomes. Maize1 genes were more likely to be identified as candidates for selection during domestication, both among genes retained as duplicate pairs (1.49×) and genes which fractionated to single copy (1.24×) (*P*-value = 0.00013 and *P*-value < 1.0×10^{-6}, respectively, binomial test) (Figure 2d). Fewer singleton genes than duplicate gene pairs were identified as likely to be under selection during domestication, while the opposite pattern was observed for genes identified as likely to be under selection during improvement; however, these differences were not statistically significant (Figure S2).

Testing for parallel selection during domestication

In order to control for differences in gene content and biases towards syntenic genes, new sets of candidate genes were selected consisting of only genes above the 90th percentile for XP-CLR scores of syntenically conserved genes in each species. Based on the percentage of genes classified as domestication candidates in each species, in the absence of parallel selection on the same genes during domestication 189 gene pairs would be expected to be identified as gene candidates in both species. Among domestication candidate genes 196 gene pairs were identified independently in both species, slightly more than the 189 gene pairs expected in the absence of parallel selection (determined via permutation testing). This difference of seven genes was not statistically significant [false discovery rate (FDR) < 0.27, permutations] (Figure 3a). The gene exhibiting the strongest combined selection signal across the two species *GRMZM2G026024*/*Sobic.004G272100* encodes a phosphoribulokinase, an enzyme that catalyzes a key step in carbon fixation as part of the Calvin cycle.

Comparison of genes under apparent selection in the landrace versus elite comparison identified fewer pairs of syntenic genes under parallel selection than during the domestication (Figure 3b). This may be linked to a lower proportion of the syntenic genes being under selection in maize during the improvement process (Figure 2c). In the case of genes under selection during crop improvement, a total of 186 overlapping genes were expected but 174 were observed (FDR < 0.85, permutations) (Figure 3b). In the wild relatives versus elite comparison, 195 overlapping gene pairs were identified and 188 were expected (FDR < 0.31, permutations) (Figure S3c).

The cut-off of genes in the 90th percentile of XP-CLR scores was chosen somewhat arbitrarily. In order to test whether the lack of a greater than expected overlap between genes identified in maize and those identified in sorghum was an artifact of the threshold score employed, the analysis above was repeated using a range of percentile-based score thresholds from the 85th percentile to the 99th percentile. None of these thresholds identified a significant enrichment of gene pairs under selection in both species relative to the expectations of the null hypothesis (Figure S4a, b).

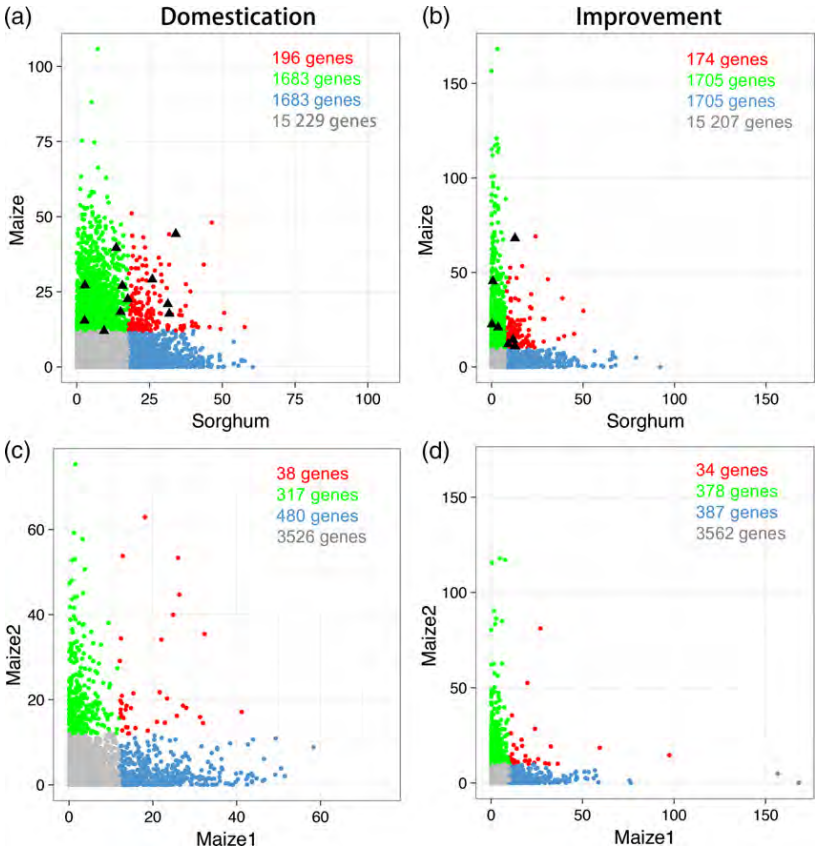


Figure 3. Comparison of scores for syntenic orthologous gene pairs. Comparison of scores for syntenic orthologous gene pairs in the wild relatives/landrace (a) and landrace/improved (b) lines XP-CLR analyses of maize and sorghum as well as the selection in duplicated maize genes during domestication (c) and improvement (d). Red, blue, orange and black dots (in (a) and (b)) mark gene pairs identified as putative selection candidates in both maize and sorghum, only in sorghum, only in maize, or in neither species, respectively. The triangles in (a) and (b) show the classic domestication genes of maize listed in Table 1. Red, blue, orange and black dots (in c and d) mark gene pairs identified as putative selection candidates in both maize1 and maize2, only in maize1, only in maize2, or in neither subgenome, respectively.

Another potential explanation is that the analysis above was partially confounded as a result of the partially paired data structure, with some sorghum genes paired with a single maize syntenic ortholog and others paired with two syntenic orthologs on opposite maize subgenomes. Separate permutation tests were conducted using only maize1–sorghum or maize2–sorghum gene pairs. Slightly more gene pairs were identified as likely under selection during improvement between maize1 and sorghum than expected under the null hypothesis, and slightly fewer gene pairs than expected were identified in the maize2 sorghum comparison. However, this bias was not large and was not replicated in the comparison of genes identified as likely under selection during domestication (Figure S4c-f).

While many phenotypic changes during domestication appear to be shared between sorghum and maize – the domestication syndrome referred to above – domestication probably also involved selection on some traits only in one species or the other. Therefore we also searched for signatures of parallel selection between homeologous gene pairs retained between the two subgenomes of maize, as these genes experienced identical whole-plant-level artificial selection during the domestication of maize. Thirty-eight duplicate pairs under selection during domestication were identified from 4362 pairs of retained maize duplicates tested (Figure 3c) and only 34 duplicate pairs under parallel selection during improvement were identified (Figure 3d). There were fewer duplicate genes under parallel selection during the improvement because of a smaller proportion of syntenic genes under selection during this period. In both cases, the number of gene pairs identified as likely under parallel selection was lower than the expectation for unlinked genes, but not by a statistically significant amount (FDR < 0.88 and 0.80, permutations, respectively).

Another potential explanation for the absence of significant overlap between genes appearing to have experienced selection in maize and sorghum, or between the two maize subgenomes, is simply that the dataset used or the analysis method employed was invalid in some way. To test this concern, we employed a positive control set of 16 maize genes with known and functionally validated links to domestication phenotypes in maize described above (Table 1). All 16 of these genes were indeed included in the set of maize gene candidates identified through XP-CLR analysis. In nine cases the sorghum orthologs of these target genes were also identified as likely targets of artificial selection (P-value < 1.0×10^{-6}, binomial test).

A set of 16 genes shown to exhibit functional variation between maize and teosinte or between maize landraces and improved lines for traits linked to domestication based

Table 1 Well-characterized domestication genes in maize and their orthologs/homeologs

Gene symbol	Gene name	Gene ID in sorghum	Gene ID in maize1	Gene ID in maize2
sh1	Seed shattering1	Sobic.001G152901	GRMZM2G085873	GRMZM2G074124
gt1	Grassy tillers1	Sobic.001G468400	GRMZM2G005624	NoGene
tga1	Teosinte glume architecture1	Sobic.007G193500	NoGene	GRMZM2G101511
tb1	Teosinte branched1	Sobic.001G121600	AC233950.1_FG002	AC190734.2_FG003
ra1	Ramosa1	Sobic.002G197700	GRMZM2G003927	NoGene
ELF4	Early flowering 4	Sobic.002G193000	GRMZM2G025646	NoGene
CCT	Flowering time related	Sobic.002G275100	GRMZM2G179024	NoGene
ohp2	Opaque2 zein storage protein synthesis	Sobic.001G056700	GRMZM2G007063	NoGene
yab14	Yabby14	Sobic.006G160800	GRMZM2G054795	GRMZM2G005353
GI	GIGANTEA	Sobic.003G040900	GRMZM5G844173	GRMZM2G107101
zag2	Zea agamous2	Sobic.008G072900	GRMZM2G010669	GRMZM2G160687
bif2	Barren inflorescence2	Sobic.008G170500	GRMZM2G171822	NoGene
zfl2	Zea floricaula/leafy2	Sobic.006G201600	GRMZM2G180190	GRMZM2G098813
gln2	glutamine synthetase2	Sobic.001G116400	GRMZM2G024104	NoGene
sbe3	starch branching enzyme3	Sobic.006G066800	GRMZM2G073054	NoGene
c2	colorless2	Sobic.005G136200	GRMZM2G422750	GRMZM2G151227

Genes in red, cyan and black represent the genes identified as likely under selection during domestication or improvement phases, or not showing evidence of selection in either process, respectively.

on single-gene or single-gene-family studies was assembled (Table 1). Characterized genes showing signatures of parallel selection in both maize and sorghum include the previously reported *sh1* gene involved in the loss of seed shattering (Lin *et al.*, 2012), genes involved in reshaping plant architecture such as *gt1*, identified as a controller of ear number in maize (Wills *et al.*, 2013), and genes involved in regulation of flowering time such as *ELF4* and *GI (Bendix et al., 2015)*, as well as two important genes in starch synthesis pathway, *ss1* and *sbe3 (Whitt et al., 2002; Campbell et al., 2016)*. The *tb1* gene, which is involved in the repression of axillary branching in both maize (Doebley *et al.*, 1997) and sorghum (Kebrom *et al.*, 2006), showed signatures of parallel selection, and was identified as a selection candidate in both maize and sorghum. However, *tb1* was not one of the strongest signals of selection, and was not even found to be a candidate locus for selection when *mexicana* teosinte lines were included as part of the wild population (Hufford *et al.*, 2012). A second TCP transcription factor, belonging to the same gene family as *tb1*, was identified as under parallel selection in sorghum, maize1 and maize2 (Figure 4).

Functional roles of genes selected in parallel

A total of 1014 maize/sorghum syntenic gene pairs were identified as under parallel selection, including both genes under parallel selection in the same comparison (i.e. wild versus landrance or landrace versus improved) or in opposite comparisons in different species. These genes were enriched in transcription factors relative to all syntenic gene pairs in both sorghum (P-value = 8.70×10^{-4}, Fisher exact test) and maize (P-value = 3.50×10^{-5}, Fisher exact test), although the absolute enrichment is modest (1.36× and 1.46× for maize and sorghum, respectively) (Table S4).

To test whether genes identified as targets of selection in both maize and sorghum exhibited parallel expression patterns, we utilized a set of RNA-seq data generated from homologous tissues in maize and sorghum, with a specific focus on reproductive tissues (Davidson *et al.*, 2011, 2012). A total of 44 genes under apparent parallel selection exhibited conserved patterns of reproductive tissue-specific expression, with anther, embryo and endosperm being represented at the highest frequency (Table S5). One gene which showed strong parallel selective signals in sorghum (*Sobic.009G203900*) and both maize subgenomes – maize1 (*GRMZM2G074361*) and maize2 (*GRMZM2G109842*) – exhibited identical and highly specific expression patterns in the anthers of both species (Figure S5). This gene is annotated as the *profilin 1* (*PRF1*) gene which encodes a core cell-wall structural protein. Overall, no strong biases towards either expression in specific reproductive tissues or greater conservation of tissue-specific expression among genes with statistical signatures of parallel selection were detected.

A total of 237 domestication candidate genes in sorghum were involved in 112 annotated biochemical pathways, and 270 domestication candidate genes from the 18 794 syntenic genes of maize were involved in 113 annotated biochemical pathways. A total of 69 pathways overlapped between these two datasets, which was similar to the 71 pathways predicted to overlap based on permutations of orthologous relationships (FDR < 0.606,

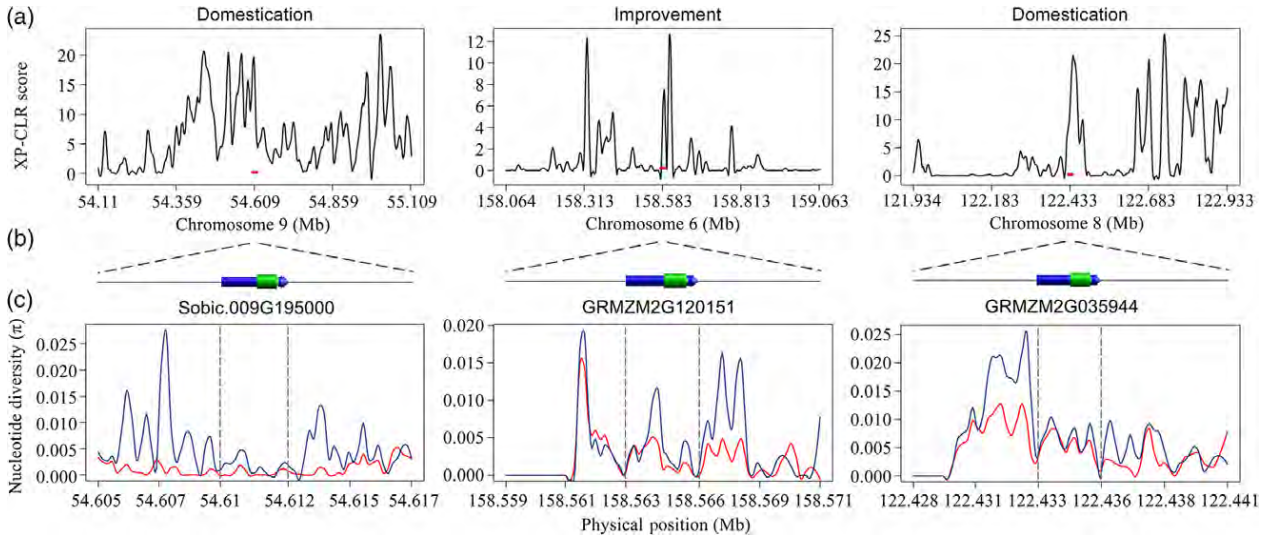


Figure 4. An example of evidence of selection on a *TCP* gene in maize and sorghum.
(a) Cross-population composite likelihood ratio test plot showing selection on partial regions of genomes in sorghum, maize1 and maize2. The red line represents the location of an orthologous *TCP* gene on the chromosomes.
(b) Gene model of syntenic *TCP* genes in sorghum, maize1 and maize2. The blue and green boxes represent the untranslated regions and exons.
(c) Level of nucleotide diversity (π) in the objective population (red) and background population (blue). [Colour figure can be viewed at wileyonlinelibrary.com].

permutations). For genes identified as candidates in the landrace–improved line comparison, a total of 231 candidate genes from sorghum were annotated as being part of 128 pathways and 258 selection candidate genes from maize were annotated as being part of 129 pathways. A total of 80 pathways had at least one maize gene and one sorghum gene under selection, which is moderately statistically significant compared with the expected number of overlapping pathways (FDR < 0.006, permutations). However, when analyzing gene pairs identified as under parallel selection, a set of 141 gene pairs were annotated as encoding enzymes which were involved in 89 different metabolic pathways (Table S4). This was a significantly larger number of metabolic pathways than would be expected given the overall number of gene pairs under parallel selection (expectation = 78 pathways, FDR < 0.028, permutations).

DISCUSSION

The high collinearity of genetic maps and gene content among related grasses (Moore *et al.*, 1995) makes it feasible to employ multiple grass species as a single genomic system. Here we sought to test whether the parallel phenotypic changes produced by artificial selection as part of the domestication syndrome in maize and sorghum resulted from parallel molecular changes targeting orthologous genes. However, as shown above, the number of genes showing parallel signatures of selection during domestication in maize and sorghum was not significantly different from the amount of overlap expected among random gene sets. This result stands in contrast to reports on individual large-effect genes where the same gene often appears to have been a target during independent domestication in different species (Lin *et al.*, 2012; Liu *et al.*, 2015), as well as the finding here that genes with validated links to domestication from single-gene studies in maize were significantly more likely to also be targets of selection in sorghum (Table 1).

A recent report that compared published results from analyses of domestication in maize and rice reached a similar result, with only 65 orthologous gene pairs being shared between a set 969 genes identified as candidate targets of selection in maize and 1526 gene pairs identified as candidate targets of selection in rice (Gaut, 2015). However, that study also identified a number of limitations in their analysis, including the relatively high linkage disequilibrium (LD) in rice, and the potential for selection for different traits during domestication in the two species given the large differences in growth habit between modern rice and maize cultivars. It was also noted that this analysis used candidate gene sets identified by different research groups using different sets of parameters, and even within the same species different scans for positive selection can identify different sets of candidate genes (Akey, 2009). Here we employed data from two more closely related species with similar, and low, levels of LD and greater similarities of plant architecture and growth habit as well as conducting a reanalysis starting from raw SNP calls in order to ensure balanced and equivalent approaches to identifying candidate genes in both species. However, we also found an absence of parallel selection on orthologous genes at a whole-genome level between maize and sorghum.

However, as with the previous comparison between maize and rice, the analyses presented above come with a number of important caveats. The first caveat comes from the differences in the domestication process between the two species. Maize was domesticated from teosinte approximately 9000 years ago in what appears to have been a single event (Piperno *et al.*, 2009; Van Heerwaarden *et al.*, 2011). Improved maize lines used in this study were largely drawn from temperate elite varieties adapted to North America. In contrast, sorghum appears to have been domesticated independently at least twice (Mace *et al.*, 2013), and the improved sorghum lines used here were drawn from separate breeding efforts aimed at developing improved cultivars for African, Australian and North American climates. Parallel domestication and independent crop improvement efforts in sorghum are likely to reduce the statistical power to identify genes which are targets of selection, both because different haplotypes of the same genes may have been targets of selection during distinct domestication or crop improvement efforts and because different domestication and crop improvement efforts may have targeted different genetic loci. While the statistical approach employed by XP-CLR to identify signatures of selective sweeps is a significant advance over previous approaches in that it can detect both 'hard' sweeps, where a single beneficial haplotype at a given locus rapidly increases in frequency as a result of selection, and 'soft' sweeps, where multiple pre-existing haplotypes provide the same fitness advantage, either as a result of recombination or independent origins (Hermisson and Pennings, 2005; Przeworski *et al.*, 2005), the statistical signatures of soft sweeps remain more difficult to detect. Sweeps which reflect selection in only a subset of a group of germplasm, for example genes under selection in elite North American sorghum lines but not in elite Australian germplasm, are also likely to be missed by the current analysis. The reproductive habits of the two species may also have altered the relative contributions of standing genetic variation, likely to contribute to soft sweeps, and novel mutations, likely to contribute to hard sweeps, to the domestication syndrome in maize and sorghum. Under field conditions outcrossing rates for sorghum have been reported to be in the range of 7–18% (Djè *et al.*, 2004; Barnaud *et al.*, 2008), while teosinte outcrossing rates can reach ~97% (Hufford *et al.*, 2011). The high outcrossing rate of wild teosinte may have allowed tolerance of alleles with a wider range of phenotypic consequences, producing a deeper pool of standing functional genetic variation than would have been present in more inbred wild sorghum plants.

The observation that in maize, but not in sorghum, non-syntenic genes were enriched among genes targeted for selection during crop improvement was unexpected. One potential explanation is the relative importance of hybrid breeding and heterosis for these two crops. The absolute size of the effect of heterosis for maize yield has remained constant while the inbred yield values have increased (Schnell, 1974). The effect of heterosis on yield in early maize single-cross hybrids was 300% in 1930s crosses, and in relative terms has slowly decreased to 100% in relative terms (Duvick, 2005a). In contrast, in sorghum, the increase in yield resulting from heterosis is generally of the order of 40% (Duvick, 1999; Mindaye *et al.*, 2016). Non-syntenic genes are more likely to exhibit presence–absence variation across different maize lines (Swanson-Wagner *et al.*, 2010; Schnable *et al.*, 2011) and to display non-additive expression (Paschold *et al.*, 2014). Both presence–absence variation and non-additive expression have been speculated to contribute to heterosis. The greater emphasis on heterosis and contribution to yield of heterosis in maize relative to sorghum may therefore have resulted in a greater proportion of artificial selection targeting non-syntenic genes in maize.

One final assumption in the analyses presented here is that selection during domestication truly did target the same phenotypic traits in both maize and sorghum. While this should generally be the case, selection during crop improvement has differed between these two species in at least one key phenotype: selection for higher yield in maize has resulted in indirect selection for decreased tassel size (Duvick, 2005b), while selection for decreased head size in sorghum would presumably be detrimental to yield. As a partial control for the potential explanation, namely that the lack of identified overlap between maize and sorghum resulted from selection for different traits during domestication in these two species, we also compared patterns of selection between conserved homeologous genes in the different maize subgenomes. These genes started out with equivalent functional roles prior to the maize whole-genome duplication, and experienced the same selective pressures during domestication and crop improvement. However, we also failed to identify any statistically significant correlation between genes identified as targets of selection between the two subgenomes. Another explanation, that selection targeted different genes in the same pathways, also failed to find support in this study; however, it should be noted that improved annotations of biochemical and transcriptional pathways may produce a different result in the future. Finally, the re-identification of many genes from a set of positive control genes previously identified through top-down approaches as playing a role in domestication provided a validation that the statistical methods, software implementations and genomic datasets employed do indeed have the power to identify genes which were targets of selection during domestication and crop improvement.

The observation that the few maize domestication genes characterized in conventional single-gene genetic studies were much more likely to also be identified as

domestication candidates in sorghum (Table 1) suggests that the genes involved in domestication may fall into a two-tier system. A few large-effect genes appear to have been repeatedly targeted to create the domestication syndrome in multiple grain crops (Lin *et al.*, 2012; Liu *et al.*, 2015). As described previously, strong and relatively recent selection should rapidly identify and fix a small number of large-effect alleles which may have pleiotropic consequences, while a range of different smaller-effect alleles at other genetic loci are then selected to fine tune the effect size and mitigate any negative pleiotropic effects of the initial large-effect alleles (Orr, 1998). Studies of the genetic architecture of different traits inferred to be under selection or largely neutral characters during domestication and crop improvement in maize and maize–teosinte RIL populations have produced findings consistent with this model (Wallace *et al.*, 2014; Xu *et al.*, 2017).

Here we propose that the initial, large-effect alleles selected for during selection for domestication syndrome traits in grain crops are drawn from a constrained pool of genes, and therefore orthologous genes are more likely to be selected for in parallel across multiple grain crops. In contrast, the set of small-effect genes which fine-tune domestication rates and mitigate potentially deleterious pleiotropic effects of large-effect alleles may be drawn from a much larger pool and would thus exhibit little repeat sampling of the same orthologous genes across different domesticated grasses. Alternatively, these fine-tuning genetic changes may more frequently be drawn from standing genetic variation, resulting in more soft sweeps or incomplete sweeps, reducing the statistical power to consistently identify these genes in repeated statistical trials across different species. Indeed, simulation studies suggest that small-effect loci are more likely to become fixed during selection if they originate from standing genetic variation than from novel mutations (Hermisson and Pennings, 2005). An additional potential confounding variable is that the mutational target space – defined as the number of potential mutations at a given locus which produce alleles with the same phenotypic outcome – may well vary between genes with larger and smaller phenotypic effects. A larger mutational target space at a given locus increases the probability that the response to artificial selection during domestication will result from a multiple-origin soft sweep and reduces the potential to identify the locus as a target of selection using bottom-up population genetic approaches (Hermisson and Pennings, 2017). The decreasing cost of whole-genome sequencing and resequencing, and the large number of different grain crops that have experienced parallel selection for domestication syndrome phenotypes, should enable more rigorous tests of this model in the near future incorporating data from syntenic orthologous genes across many different species.

EXPERIMENTAL PROCEDURES

Data collection and preliminary polishing

The maize and sorghum whole-genome resequencing data used in this study were taken from Hapmap2 (Chia *et al.*, 2012) and SorGSD (Luo *et al.*, 2016), respectively. SNPs that scored as heterozygous in 3.0% of accessions in the maize Hapmap2 and sorghum dataset were removed prior to analysis. A subset of maize accessions was selected and separated into three groups: 30 improved lines, 19 landraces and 7 wild relatives (Table S1). Data from a total of 42 sorghum accessions were obtained, including 17 improved lines, 18 landraces and 7 wild relatives (Table S1). SNPs with missing rates of >50% in either species, or with heterozygous calls in any of the remaining samples, were discarded, resulting in a final dataset consisting of 10.3 million SNPs in maize and 3.3 million SNPs in sorghum.

Population genetics analysis

The genetic distance between individuals was first calculated using a 0.1% subset of the total SNP set constructed by sampling every 1000th SNP position along each chromosome for a total of 10 286 SNPs in maize and using a 0.2% subset of the total SNP set constructed by sampling every 500th SNP position along each chromosome for a total of 6719 SNPs in sorghum.

Neighbor-joining trees were constructed for the accessions of each species using Phase (Jow *et al.*, 2002) and Phylip v3.696 (Felsenstein, 1981) with default parameters. The resulting phylogenetic trees were visualized using Figtree v1.4.3 (http://tree.bio.ed.ac.uk/software/figtree/). Nucleotide diversity (π) values were calculated for each species with non-overlapping windows of 10 kb using an in-house Perl script (https://figshare.com/articles/Tajima_D_pl/5544484). Reported π values are the average of all genomic windows.

Syntenic gene identification

A pan-grass syntenic gene set using the sorghum genes as reference was download from figshare (Schnable *et al.*, 2016). When multiple sorghum genes were identified as syntenic orthologs of the same gene in maize – a result that can be produced by tandem duplication events in sorghum – the tie was broken using a separate dataset of syntenic orthologous genes using the *Setaria italica* genome as a reference. This resulted in a final set of 14 433 sorghum genes paired with a syntenic ortholog in either the maize1 subgenome (11 402 gene pairs) and/or the maize2 subgenome (7392 gene pairs), including 4361 sorghum genes with syntenic co-orthologs on both maize subgenomes (Table S2).

Genome-wide scan for selection

To identify genes affected by selection during domestication in maize and sorghum, genome-wide scans for signals of selection were conducted using a cross-population composite likelihood approach (XP-CLR) [Chen *et al.*, 2010; updated by Hufford *et al.* (2012) to incorporate missing data], based on the allele frequency differentiation between target and reference populations. This approach was employed in three separate pair wise comparisons: wild relatives versus landraces, landraces versus improved lines and wild relatives versus improved lines.

Recombination rates in maize and sorghum were measured using high-density genetic maps constructed using RILs from biparental crosses in maize (Ott *et al.*, 2017) and sorghum (Zou *et al.*, 2012). Genetic maps were transferred to the more recent

versions of the maize and sorghum genomes used in this analysis (B73 RefGen v3 and v3.1, respectively). The transfer was performed using the two genes flanking the marker (when the marker was in a non-coding region) or the single gene the marker was located in. For each pseudomolecule in maize, a ninth-order polynomial curve was fitted to the genetic and physical coordinates of all markers presented on chromosomes, and genetic positions for each marker were reassigned based on the value predicted for the genetic and physical position of the marker and the polynomial formula.

The same parameters were employed for XP-CLR analysis for maize and sorghum. A 0.05-cM sliding window with 1000-bp steps across the whole-genome scan was used for scanning. Individual SNPs were assigned a position along the genetic map based on the polynomial fitting curves in maize and by assuming uniform recombination between pairs of genetic markers in sorghum. The number of SNPs assayed in each window was fixed at 100 and pairs of SNPs in high LD ($r^2 > 0.75$) were down-weighted.

To obtain XP-CLR scores for each gene, each gene was assigned a window starting 5 kb upstream of its annotated transcription start site and extending to 5 kb downstream of its annotated transcription stop site. The maximum XL-CLR score among all the XL-CLR intervals within this window was assigned to the gene.

Testing for enrichment of genes selected in parallel

Gene pairs were considered to be under selection if a sorghum gene and at least one maize syntenic ortholog were both identified as being under selection. To determine the optimal cut-off for testing the enrichment of syntenic genes under parallel selection, a series of cut-offs from 85% to 99% were used in the analysis. At each cut-off the number of gene pairs under selection in both species were recorded and compared with the number of gene pairs identified when orthologous relationships between maize and sorghum were shuffled using a permutation test repeated 100 times.

Gene annotation and enrichment analysis

Maize and sorghum GO annotations were retrieved from Phytozome (https://phytozome.jgi.doe.gov/). The maize transcription factor (TF) lists were downloaded from Grassius (http://grassius.org/grasstfdb.html). Metabolic pathway lists were downloaded from the Gramene (ftp://ftp.gramene.org/pub/gramene/pathways/). Annotated enzyme name and the corresponding pathways for these genes were obtained by searching the pathway list.

A set of 1000 permutations was used to calculate the expected number of pathways in the same number of random genes in the R package to examine whether maize and sorghum genes under selection were significantly more likely to be present in the same pathways than expected if selection was unlinked.

ACKNOWLEDGEMENTS

We thank Professor Edward Buckler (Cornell University) for advice on the Hapmap dataset and Professor Jeffrey Ross-Ibarra (UC Davis) for advice and access to an updated version of the XP-CLR software. This work was supported by a China Scholarship Council fellowship awarded to XL and a Science Foundation of Xichang College awarded to LY.

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Figure S1. Genome-wide scan for signatures of selection in maize and sorghum.

Figure S2. Ratio of singleton genes: duplicate pairs among genes identified as selection candidates.

Figure S3. Comparison of scores for syntenic orthologous gene pairs in the wild relative/improved line XP-CLR analyses of maize and sorghum.

Figure S4. Testing for enrichment of candidate selection genes for different proportions of genes with top XP-CLR scores.

Figure S5. Gene model and gene expression level visualization of *profilin1* in sorghum and maize.

Table S1. List of maize and sorghum accessions employed in this study, their data sources and their classifications as wild relative, landrace or improved line datasets.

Table S2. Set of high-confidence syntenic orthologous maize–sorghum gene pairs employed in this study.

Table S3. Population nucleotide diversity statistics for maize and sorghum.

Table S4. A list of the genes identified as likely under selection during the domestication and/or improvement process in maize and sorghum and functional annotations of each.

Table S5. Tissue-specific expression of genes under parallel selection between maize and sorghum.

REFERENCES

Akey, J.M. (2009) Constructing genomic maps of positive selection in humans: where do we go from here? *Genome Res.*, **19**, 711–722.

Baldauf, J.A., Marcon, C., Paschold, A. and Hochholdinger, F. (2016) Nonsyntenic genes drive tissue-specific dynamics of differential, nonadditive, and allelic expression patterns in maize hybrids. *Plant Physiol.*, **171**, 1144–1155.

Barnaud, A., Trigueros, G., McKey, D. and Joly, H.I. (2008) High outcrossing rates in fields with mixed sorghum landraces: how are landraces maintained? *Heredity*, **101**, 445.

Bendix, C., Marshall, C.M. and Harmon, F.G. (2015) Circadian clock genes universally control key agricultural traits. *Molecular plant*, **8**, 1135–1152.

Bennetzin, J.L. and Freeling, M. (1993) Grasses as a single genetic system: genome composition, collinearity and compatibility. *Trends Genet.*, **9**, 259–261.

Bradbury, L.M., Fitzgerald, T.L., Henry, R.J., Jin, Q. and Waters, D.L. (2005) The gene for fragrance in rice. *Plant Biotechnol. J.*, **3**, 363–370.

Campbell, B.C., Gilding, E.K., Mace, E.S., Tai, S., Tao, Y., Prentis, P.J., Thomelin, P., Jordan, D.R. and Godwin, I.D. (2016) Domestication and the storage starch biosynthesis pathway: signatures of selection from a whole sorghum genome sequencing strategy. *Plant Biotechnol. J.*, **14**, 2240–2253.

Chen, H., Patterson, N. and Reich, D. (2010) Population differentiation as a test for selective sweeps. *Genome Res.*, **20**, 393–402.

Chia, J.-M., Song, C., Bradbury, P.J. *et al.* (2012) Maize HapMap2 identifies extant variation from a genome in flux. *Nat. Genet.*, **44**, 803–807.

Clark, R.M., Linton, E., Messing, J. and Doebley, J.F. (2004) Pattern of diversity in the genomic region near the maize domestication gene tb1. *Proc. Natl Acad. Sci. USA*, **101**, 700–707.

Davidson, R.M., Hansey, C.N., Gowda, M. *et al.* (2011) Utility of RNA sequencing for analysis of maize reproductive transcriptomes. *The Plant Genome*, **4**, 191–203.

Davidson, R.M., Gowda, M., Moghe, G., Lin, H., Vaillancourt, B., Shiu, S.-H., Jiang, N. and Robin Buell, C. (2012) Comparative transcriptomics of three Poaceae species reveals patterns of gene expression evolution. *Plant J.*, **71**, 492–502.

Dirzo, R. and Raven, P.H. (2003) Global state of biodiversity and loss. *Annu. Rev. Environ. Resour.*, **28**, 137–167.

Djè, Y., Heuertz, M., Ater, M., Lefèbvre, C. and Vekemans, X. (2004) In situ estimation of outcrossing rate in sorghum landraces using microsatellite markers. *Euphytica*, **138**, 205–212.

Doebley, J. and Stec, A. (1993) Inheritance of the morphological differences between maize and teosinte: comparison of results for two F2 populations. *Genetics*, **134**, 559–570.

Doebley, J., Stec, A. and Hubbard, L. (1997) The evolution of apical dominance in maize. *Nature*, **386**, 485.

Dorweiler, J., Stec, A., Kermicle, J. and Doebley, J. (1993) Teosinte glume architecture 1: a genetic locus controlling a key step in maize evolution. *Science*, **262**, 233–233.

Duvick, D.N. (1999) Heterosis: Feeding people and protecting natural resources. In *The Genetics and Exploitation of Heterosis in Crops* (Coors, J.G. and Pandey, S. eds). ASA, CSSA: Madison, WI, pp. 19–29.

Duvick, D.N. (2005a) The contribution of breeding to yield advances in maize (Zea mays L.). *Adv. Agron.*, **86**, 83–145.

Duvick, D.N. (2005b) Genetic progress in yield of United States maize (Zea mays L.). *Maydica*, **50**, 193.

Felsenstein, J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.*, **17**, 368–376.

Gaut, B.S. (2015) Evolution Is an Experiment: assessing Parallelism in Crop Domestication and Experimental Evolution. *Mol. Biol. Evol.*, **32**, 1661–1671.

Glémin, S. and Bataillon, T. (2009) A comparative view of the evolution of grasses under domestication. *New Phytol.*, **183**, 273–290.

Harlan, J.R., Wet, J.D. and Price, E.G. (1973) Comparative evolution of cereals. *Evolution*, **27**, 311–325.

Hermisson, J. and Pennings, P.S. (2005) Soft sweeps. *Genetics*, **169**, 2335–2352.

Hermisson, J. and Pennings, P.S. (2017) Soft sweeps and beyond: understanding the patterns and probabilities of selection footprints under rapid adaptation. *Methods Ecol. Evol.*, **8**, 700–716.

Huang, X., Kurata, N., Wang, Z.-X. *et al.* (2012) A map of rice genome variation reveals the origin of cultivated rice. *Nature*, **490**, 497.

Hufford, M.B., Gepts, P. and ROSS-IBARRA, J. (2011) Influence of cryptic population structure on observed mating patterns in the wild progenitor of maize (Zea mays ssp. parviglumis). *Mol. Ecol.*, **20**, 46–55.

Hufford, M.B., Xu, X., van Heerwaarden, J. *et al.* (2012) Comparative population genomics of maize domestication and improvement. *Nat. Genet.*, **44**, 808–811.

Jow, H., Hudelot, C., Rattray, M. and Higgs, P.G. (2002) Bayesian phylogenetics using an RNA substitution model applied to early mammalian evolution. *Mol. Biol. Evol.*, **19**, 1591–1601.

Kebrom, T.H., Burson, B.L. and Finlayson, S.A. (2006) Phytochrome B represses Teosinte Branched1 expression and induces sorghum axillary bud outgrowth in response to light signals. *Plant Physiol.*, **140**, 1109–1117.

Lai, X., Behera, S., Liang, Z., Lu, Y., Deogun, J.S. and Schnable, J.C. (2017) STAG-CNS: an Order-Aware Conserved Non-coding Sequences Discovery Tool For Arbitrary Numbers of Species. *Molecular Plant*, **10**, 990–999.

Li, L., Briskine, R., Schaefer, R., Schnable, P.S., Myers, C.L., Flagel, L.E., Springer, N.M. and Muehlbauer, G.J. (2016) Co-expression network analysis of duplicate genes in maize (Zea mays L.) reveals no subgenome bias. *BMC Genom.*, **17**, 875.

Lin, Z., Li, X., Shannon, L.M. *et al.* (2012) Parallel domestication of the Shattering1 genes in cereals. *Nat. Genet.*, **44**, 720–724.

Liu, H., Liu, H., Zhou, L., Zhang, Z., Zhang, X., Wang, M., Li, H. and Lin, Z. (2015) Parallel domestication of the heading date 1 gene in cereals. *Mol. Biol. Evol.*, **32**, 2726–2737.

Luo, H., Zhao, W., Wang, Y. *et al.* (2016) SorGSD: a sorghum genome SNP database. *Biotechnol. Biofuels*, **9**, 1.

Mace, E.S., Tai, S., Gilding, E.K. *et al.* (2013) Whole-genome sequencing reveals untapped genetic potential in Africa's indigenous cereal crop sorghum. *Nat. Commun.*, **4**, 2320–2320.

Mauro-Herrera, M., Wang, X., Barbier, H., Brutnell, T.P., Devos, K.M. and Doust, A.N. (2013) Genetic control and comparative genomic analysis of flowering time in Setaria (Poaceae). *G3: Genes. Genomes, Genetics*, **3**, 283–295.

Mindaye, T.T., Mace, E.S., Godwin, I.D. and Jordan, D.R. (2016) Heterosis in locally adapted sorghum genotypes and potential of hybrids for increased productivity in contrasting environments in Ethiopia. *The Crop Journal*, **4**, 479–489.

Moore, G., Devos, K.M., Wang, Z. and Gale, M.D. (1995) Cereal genome evolution: grasses, line up and form a circle. *Curr. Biol.*, **5**, 737–739.

Orr, H.A. (1998) The population genetics of adaptation: the distribution of factors fixed during adaptive evolution. *Evolution*, **52**, 935–949.

Ott, A., Liu, S., Schnable, J. C., Yeh, C., Wang, K. and Schnable, P.S. (2017) tGBS® genotyping-by-sequencing enables reliable genotyping of heterozygous loci. *Nucleic Acids Res.* **45**, e178.

Paschold, A., Larson, N.B., Marcon, C., Schnable, J.C., Yeh, C.-T., Lanz, C., Nettleton, D., Piepho, H.-P., Schnable, P.S. and Hochholdinger, F. (2014) Nonsyntenic genes drive highly dynamic complementation of gene expression in maize hybrids. *Plant Cell*, **26**, 3939–3948.

Paterson, A.H., Lin, Y.-R., Li, Z., Schertz, K.F., Doebley, J.F., Pinson, S.R., Liu, S.-C., Stansel, J.W. and Irvine, J.E. (1995) Convergent domestication of cereal crops by independent mutations at corresponding genetic loci. *Science*, **269**, 1714–1718.

Piperno, D.R., Ranere, A.J., Holst, I., Iriarte, J. and Dickau, R. (2009) Starch grain and phytolith evidence for early ninth millennium BP maize from the Central Balsas River Valley, Mexico. *Proc. Natl Acad. Sci.*, **106**, 5019–5024.

Pophaly, S.D. and Tellier, A. (2015) Population level purifying selection and gene expression shape subgenome evolution in maize. *Mol. Biol. Evol.*, **32**, 3226–3235.

Przeworski, M., Coop, G. and Wall, J.D. (2005) The signature of positive selection on standing genetic variation. *Evolution*, **59**, 2312–2323.

Renny-Byfield, S., Rodgers-Melnick, E. and Ross-Ibarra, J. (2017) Gene fractionation and function in the ancient subgenomes of maize. *Mol. Biol. Evol.*, **34**, 1825–1832.

Ross-Ibarra, J., Morrell, P.L. and Gaut, B.S. (2007) Plant domestication, a unique opportunity to identify the genetic basis of adaptation. *Proc. Natl Acad. Sci.*, **104**, 8641–8648.

Sagnard, F., Deu, M., Dembélé, D. *et al.* (2011) Genetic diversity, structure, gene flow and evolutionary relationships within the Sorghum bicolor wild?weedy?crop complex in a western African region. *Theoretical and applied genetics*, **123**, 1231.

Schnable, J.C. (2015) Genome evolution in maize: from genomes back to genes. *Annu. Rev. Plant Biol.*, **66**, 329–343.

Schnable, J.C. and Freeling, M. (2011) Genes identified by visible mutant phenotypes show increased bias toward one of two subgenomes of maize. *PLoS ONE*, **6**, 17855.

Schnable, J.C., Springer, N.M. and Freeling, M. (2011) Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proc. Natl Acad. Sci.*, **108**, 4069–4074.

Schnable, J., Zang, Y. and Ngu, W.C.D. (2016) Pan-Grass Syntenic Gene Set (sorghum referenced). *Figshare*, https://doi.org/10.6084/m9.figshare.3113488.v1.

Schnell, F.W. (1974) *Trends and problems in breeding methods for hybrid corn*. Birmingham, England: Proc. of the British Poultry Breeders Roundtable, 16th. pp. 86–98.

Sugimoto, K., Takeuchi, Y., Ebana, K. *et al.* (2010) Molecular cloning of Sdr4, a regulator involved in seed dormancy and domestication of rice. *Proc. Natl Acad. Sci.*, **107**, 5792–5797.

Swanson-Wagner, R.A., Eichten, S.R., Kumari, S., Tiffin, P., Stein, J.C., Ware, D. and Springer, N.M. (2010) Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor. *Genome Res.*, **20**, 1689–1699.

Swigo˘ová, Z., Lai, J., Ma, J., Ramakrishna, W., Llaca, V., Bennetzen, J.L. and Messing, J. (2004) Close split of sorghum and maize genome progenitors. *Genome Res.*, **14**, 1916–1923.

Takuno, S., Ralph, P., Swarts, K., Elshire, R.J., Glaubitz, J.C., Buckler, E.S., Hufford, M.B. and Ross-Ibarra, J. (2015) Independent molecular basis of convergent highland adaptation in maize. *Genetics*, **200**, 1297–1312.

Tao, Y., Mace, E.S., Tai, S., Cruickshank, A., Campbell, B.C., Zhao, X., van Oosterom, E.J., Godwin, I.D., Botella, J.R. and Jordan, D.R. (2017) Whole-genome analysis of candidate genes associated with seed size and weight in Sorghum bicolor reveals signatures of artificial selection and insights into parallel domestication in cereal crops. *Frontiers in Plant Science*, **8**, 1237.

Van Heerwaarden, J., Doebley, J., Briggs, W.H., Glaubitz, J.C., Goodman, M.M., Gonzalez, J.d.J.S. and Ross-Ibarra, J. (2011) Genetic signals of

origin, spread, and introgression in a large sample of maize landraces. *Proc. Natl Acad. Sci.*, **108**, 1088–1092.

Wallace, J.G., Bradbury, P.J., Zhang, N., Gibon, Y., Stitt, M. and Buckler, E.S. (2014) Association mapping across numerous traits reveals patterns of functional variation in maize. *PLoS Genet.*, **10**, 1004845.

Wendorf, F., Close, A.E., Schild, R., Wasylikowa, K., Housley, R.A., Harlan, J.R. and Królik, H. (1992) Saharan exploitation of plants 8,000 years BP. *Nature*, **359**, 721–724.

Whipple, C.J., Kebrom, T.H., Weber, A.L., Yang, F., Hall, D., Meeley, R., Schmidt, R., Doebley, J., Brutnell, T.P. and Jackson, D.P. (2011) Grassy tillers 1 promotes apical dominance in maize and responds to shade signals in the grasses. *Proc. Natl Acad. Sci.*, **108**, 506.

Whitt, S.R., Wilson, L.M., Tenaillon, M.I., Gaut, B.S. and Buckler, E.S. (2002) Genetic diversity and selection in the maize starch pathway. *Proc. Natl Acad. Sci.*, **99**, 12959–12962.

Wills, D.M., Whipple, C.J., Takuno, S., Kursel, L.E., Shannon, L.M., Ross-Ibarra, J. and Doebley, J.F. (2013) From many, one: genetic control of prolificacy during maize domestication. *PLoS Genet.*, **9**, 1003604.

Xu, G., Wang, X., Huang, C. *et al.* (2017) Complex genetic architecture underlies maize tassel domestication. *New Phytol.*, **214**, 852–864.

Yu, B., Lin, Z., Li, H. *et al.* (2007) TAC1, a major quantitative trait locus controlling tiller angle in rice. *Plant J.*, **52**, 891–898.

Zou, G., Zhai, G., Feng, Q. *et al.* (2012) Identification of QTLs for eight agronomically important traits using an ultra-high-density map based on SNPs generated from high-throughput sequencing in sorghum under contrasting photoperiods. *J. Exp. Bot.*, **63**, 5451–5462.

G3 Genes | Genomes | Genetics

GENOMIC SELECTION

Phenotypic Data from Inbred Parents Can Improve Genomic Prediction in Pearl Millet Hybrids

Zhikai Liang,* Shashi K. Gupta,† Cheng-Ting Yeh,‡ Yang Zhang,*,1 Daniel W. Ngu,* Ramesh Kumar,§ Hemant T. Patil, Kanulal D. Mungra,†† Dev Vart Yadav,§ Abhishek Rathore,† Rakesh K. Srivastava,† Rajeev Gupta,† Jinliang Yang,* Rajeev K. Varshney,† Patrick S. Schnable,‡ and James C. Schnable*,2**

*University of Nebraska-Lincoln, Lincoln, NE †International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Hyderabad, Telangana State, India, ‡Iowa State University, Ames, IA §Chaudhary Charan Singh Haryana Agricultural University, Hisar, Haryana, India, **Mahatma Phule Krishi Vidyapeeth, Dhule, Maharashtra, India, and ††Junagadh Agricultural University, Jamnagar, Gujarat, India

ORCID IDs: 0000-0002-9963-8631 (Z.L.); 0000-0003-1712-7211 (Y.Z.); 0000-0001-6887-4095 (A.R.); 0000-0002-0999-3518 (J.Y.); 0000-0002-4562-9131 (R.K.V.); 0000-0001-9169-5204 (P.S.S.); 0000-0001-6739-5527 (J.C.S.)

ABSTRACT Pearl millet is a non-model grain and fodder crop adapted to extremely hot and dry environments globally. In India, a great deal of public and private sectors' investment has focused on developing pearl millet single cross hybrids based on the cytoplasmic-genetic male sterility (CMS) system, while in Africa most pearl millet production relies on open pollinated varieties. Pearl millet lines were phenotyped for both the inbred parents and hybrids stage. Many breeding efforts focus on phenotypic selection of inbred parents to generate improved parental lines and hybrids. This study evaluated two genotyping techniques and four genomic selection schemes in pearl millet. Despite the fact that 6× more sequencing data were generated per sample for RAD-seq than for tGBS, tGBS yielded more than 2× as many informative SNPs (defined as those having MAF $>$ 0.05) than RAD-seq. A genomic prediction scheme utilizing only data from hybrids generated prediction accuracies (median) ranging from 0.73-0.74 (1000-grain weight), 0.87-0.89 (days to flowering time), 0.48-0.51 (grain yield) and 0.72-0.73 (plant height). For traits with little to no heterosis, hybrid only and hybrid/inbred prediction schemes performed almost equivalently. For traits with significant mid-parent heterosis, the direct inclusion of phenotypic data from inbred lines significantly ($P < 0.05$) reduced prediction accuracy when all lines were analyzed together. However, when inbreds and hybrid trait values were both scored relative to the mean trait values for the respective populations, the inclusion of inbred phenotypic datasets moderately improved genomic predictions of the hybrid genomic estimated breeding values. Here we show that modern approaches to genotyping by sequencing can enable genomic selection in pearl millet. While historical pearl millet breeding records include a wealth of phenotypic data from inbred lines, we demonstrate that the naive incorporation of this data into a hybrid breeding program can reduce prediction accuracy, while controlling for the effects of heterosis *per se* allowed inbred genotype and trait data to improve the accuracy of genomic estimated breeding values for pearl millet hybrids.

KEYWORDS
pearl millet
Genomic Selection
hybrid breeding
genotyping
GenPred
Shared Data Resources

Pearl millet [*Cenchrus americanus* (L.) Morrone; Syn. *Pennisetum glaucum* (L.) R. Br.] is able to grow on infertile and marginal soils under limiting soil moisture conditions and high soil temperatures. It is a climate resilient species, and is one of the most widely grown millets globally Ramya *et al.* (2017). Pearl millet can thrive in arid environments, and successfully set seed at temperatures above 40°, which would kill the pollen/stigmas of many other grain crops Gupta *et al.* (2015). Pearl millet can also tolerate infertile and marginal soils, limited soil moisture, and high soil temperatures. While most pearl millet production in Africa utilizes open pollinated varieties, Indian pearl millet production now makes extensive use of hybrid seed generated using three line cytoplasmic-genetic male sterility systems (CMS) Hanna (1989). Three line CMS systems employ female lines which carry male sterile cytoplasm and non-restoring nuclear gene(s) (A-lines), maintainer lines carry an identical nuclear genome to each A-line in a compatible fertile cytoplasm, resulting in male fertile plants (B-lines)

and are able to maintain the male sterility of A-line, and pollinator/male lines which carry dominant nuclear restorer of fertility gene(s) (R-lines).

Plant breeding for hybrid crops requires generating and testing large numbers of hybrids under different field conditions. Performing crosses to generate F1 hybrids is a labor intensive process. Top-crossing between B- and R-lines can reduce the amount of labor required per cross, but only in crossing schemes where many female lines are being crossed to one or a few male lines. Evaluating each new hybrid across field trials with several environments also requires significant time and resources. As a result, methods for selecting parental inbred lines and determining which crosses are likely to yield the best hybrids is a critical part of crop improvement. In pearl millet, a widespread approach has been used to evaluate the phenotypes of new potential inbred parents as a first pass screen of their potential in hybrid breeding programs.

Traditionally, mid-parental values have been a common way to predict performance of hybrids on the basis of inbred values, combined with estimates of General Combining Ability (GCA) in cases where phenotypes cannot be scored in parental lines or individuals directly Gowda *et al.* (2013); Xing *et al.* (2014); Mühleisen *et al.* (2015). However, for traits where significant heterosis exists, the phenotypes of hybrids can vary significantly from what would be predicted through the use of mid-parent values and estimated GCA. In these cases, it can be necessary to estimate Specific Combining Ability (SCA) values for each potential cross. The incorporation of genetic markers can improve the accuracy with which both GCA and SCA can be predicted by enabling the sharing of data across multiple tested lines carrying common haplotypes Schrag *et al.* (2007). When applied to sets of genetic markers across the whole genome, this process is referred to as genomic prediction (GP), which can be used to implement breeding programs based on estimated breeding values from genome wide sets of markers, a process know as genomic selection (GS).

Approximately 90-100 pearl millet hybrids are currently cultivated on about 5 million hectares in India Yadav *et al.* (2016). Both public and private sector organizations, including 30-40 seed companies, perform thousands of test-crosses each year. Resulting hybrids are then evaluated over multiple years and multiple locations to identify small numbers of new hybrids with superior performance which can be marketed/released for cultivation. The high investment of time and resources into initial hybrid evaluation would benefit significantly from the use of GP/GS to exclude many potential test crosses which can be discarded as unlikely to outperform existing hybrids prior to field evaluation, reducing the vast number of crosses which must be performed and evaluated.

The use of GP/GS to obtain estimated breeding values have been widely evaluated and employed in inbreeding crops such as wheat Poland *et al.* (2012), barley Zhong *et al.* (2009), rice Spindel *et al.* (2015). In crops where production is based upon hybrids, genomic prediction for single-cross hybrid performance are only starting to appear in the public sector literature Technow *et al.* (2014); Kadam *et al.* (2016), although genomic predictions for hybrid performance across populations all crossed to a single common tester are more common Windhausen *et al.* (2012); Albrecht *et al.* (2014). Pearl millet presents an intriguing opportunity in that both hybrid and open pollinated production systems are widely employed, and phenotypic data are thus available from both hybrids and inbred R- and B- lines. A-lines, being male sterile, do not produce grain when grown in isolation.

Here we evaluated two potential genotyping strategies – RAD-seq Miller *et al.* (2007) and tGBS Ott *et al.* (2017) to characterize a set of inbred pearl millet lines developed by ICRISAT in Hyderabad, India, and then evaluated the utility of GP/GS to predict optimal hybrid combinations among possible combinations of these inbreds using a scheme trained using phenotypic data collected from hybrid trials alone, inbred trials alone, or both.

MATERIALS AND METHODS

Field Traits

Field trials were conducted at four locations in India, spanning two agro-ecological zones (A- and B- zone, having rainfall of >400mm/annum) of pearl millet cultivation. The Hisar and Jamnagar sites fall within the A zone of pearl millet cultivation in northwest India, while Dhule and Patancheru are located in the B zone of pearl millet cultivation in southern (peninsular) India Gupta *et al.* (2013). While pearl millet is also grown in the A1 zone - highly drought prone areas with less than 400 mm of rainfall per year - the majority of hybrid pearl millet is confined to the A and B agroecological zones. Data were collected from 320 hybrids and 37 inbreds at field trials in four locations in 2015 in India (Dhule: N20.90°,E74.77°; Patancheru: N17.53°,E78.27°; Jamnagar: N22.47°,E70.06° and Hisar: N29.10°,E75.46°). In CMS system, A-lines are sterile and hence will not produce grain when grown in isolation. Therefore, genotyping and phenotyping were conducted on non-sterile B-lines carrying the same nuclear genome as A-lines in a compatible cytoplasm, rendering them male fertile. Lines in plots were grown in an alpha lattice design with two replicates and 28 15-plot blocks in each location. Each block included two common control lines/hybrids (ICMH 356 and 9444) and 13 experimental lines. Hybrid plots were randomly assigned to the first 25 blocks of each replicate, and inbred plots were randomly assigned to the last three blocks of each replicate (Experimental design, plot distribution and recorded phenotypes was provided in FigShare https://figshare.com/articles/pearl_millet_genomic_selection_field_layout/5969230).

Phenotype measurement

Phenotypic traits scored include days to 50% flowering (days), plant height (centimeters), grain yield (kilograms/hectare), and 1000-grain weight (grams). The criteria used to measure each of these four traits were as follows. 1) Plant height (centimeters): Plant height for a given plant was measured from where the main stem meets to soil to the tip of the panicle of the primary tiller at the time of harvest. For each plot, five random plants were randomly selected for height measurements and the mean value of these five measurements was reported; 2) Days to 50% flowering (days): Days to flowering was measured as the time between the planting date and the date at which at least 50% of plants within a given plot exhibited the initiation of stigma emergence on the panicle of their primary tiller; 3) 1000 seed weight (grams): 200 seeds were counted out from the pooled grain collected from a given research plot, weighed, and multiplied by a factor of 5 to determine 1000 seed weight (grams); 4) Grain yield (kilograms/hectare): For each entry, all panicles within a

doi: https://doi.org/10.1534/g3.118.200242
Manuscript received March 14, 2018; accepted for publication May 21, 2018; published Early Online May 24, 2018.

Supplemental material available at Figshare: doi: https://doi.org/10.6084/m9.figshare.5969230; doi: https://doi.org/10.6084/m9.figshare.5566843.
[1]Present Address: St. Jude Children's Research Hospital, Memphis, TN.
[2]Corresponding author: Beadle Center E207, Department of Agronomy and Horticulture & Plant Science Innovation Center, University of Nebraska-Lincoln, Lincoln, NE 68583. E-mail: schnable@unl.edu

given a plot were harvested at physiological maturity, and these panicles were sun dried for 10 to 15 days and then threshed for grain yield. Planting density and plot size varied across locations, but for each location the total yield was multiplied by the number of plots per hectare to estimate the final yield per hectare.

DNA extraction and library construction

Thirty to thirty five seeds from each inbred line were sown in a four inch pot in a darkroom at ICRISAT°s Patancheru. The pots were maintained at a temperature between 18° and 25°. Etiolated leaf tissues were harvested eight days after planting. Pooled leaf tissue from 20 to 25 seedlings per line was collected for DNA extraction. DNA was extracted using a modified DNA extraction method described by Mace *et al.* (2003). The DNA was stained by 5 ng/μl of ethidium bromide and checked using 0.8% (w/v) agarose gel electrophoresis in Tris-acetate-EDTA (TAE) buffer for 1 h at 90 V with visualization under ultraviolet (UV) light. tGBS sequencing libraries for 192 B-lines and 192 R-lines were prepared following the protocol outlined in Ott *et al.* (2017). RAD-seq libraries for a set of inbreds including all but 12 of the lines genotyped using tGBS were constructed as described in Varshney *et al.* (2017). RAD-seq libraries were sequenced using an HiSeq 2000 and tGBS libraries were sequenced using an Ion Proton (Table 1).

SNP calling and filtering

Raw sequence data obtained from both genotyping strategies was analyzed using the same analytical pipeline to enable accurate comparisons between the two. Raw reads were aligned to the pearl millet reference genome (v1.1 Varshney *et al.* (2017)) using default settings of GSNAP Wu and Nacu (2010).

After alignment, SNPs were called using the software package 123SNP Yu *et al.* (2012) ignoring the first and end 3bp of aligned reads. After ignoring the first and last 3 bp of each read, polymorphic sites were determined using the following criteria: 1) 5 aligned reads covering the position in the genome; 2) PHRED quality greater than 20. Genotype calls for individual samples were determined in the following fashion. The genotype for a given SNP marker in a given sample was determined to be homozygous if the site was covered by 5 aligned reads from that individual samples and one allele had a frequency >0.9. The genotype for a given SNP marker in a given sample was determined to be heterozygous if the site was covered by 5 reads, at least 90% of the reads support the two most frequent alleles, at least two reads supported the two most frequent alleles, and both alleles had a frequency >0.2. Any cases which did not satisfy the conditions for either a homozygous or heterozygous SNP call were treated as missing data.

The initial set of SNPs was filtered to exclude any SNP site more than two alleles were identified, sites where only one genotype call was present, sites where more than 10% of samples with genotype calls heterozygous, sites where the minor allele was not identified in at least 5 samples, and sites where <20% of individuals had a genotype call for the site. These sets of filtered SNPs were used to calculate missing data rate (# of samples with missing data / total sample), heterozygosity (# of samples with heterozygous genotype calls / (# samples with homozygous genotype calls + # of samples with heterozygous genotype calls)) and minor allele frequency ((2×# of samples with homozygous minor allele genotype calls) + # of samples with heterozygous genotype calls) / 2 × total sample without missing sample). Then they were imputed using Beagle (Version: 16-06-2016). The filtered but unimputed and imputed SNP sets used in this paper have been uploaded to FigShare (https://doi.org/10.6084/m9.figshare.5566843.v1). Genetic markers with low MAFs (Minor Allele Frequency) are frequently removed prior to quantitative genetic analysis Tabangin *et al.* (2009). In downstream analysis, only SNPs with MAF larger than 0.05 were employed.

Projecting Hybrid Genotypes

Hybrid genotypes for each possible combination of an A/B-line and an R-line were derived from genotypes of the corresponding parental inbred lines. If both parental lines were homozygous for the same allele at a given marker, the F1 progeny received the same genotype call at that marker. If the parental lines were homozygous for opposite alleles at a given marker, the F1 progeny received a heterozygous genotype call at that marker. If either parent was genotyped as heterozygous at a given marker was treated as having a genotype of (parent 1 genotype + parent 2 genotype)/2 on a scale of 0 to 2, where 0 is a genotype call of homozygous reference allele, and 2 is a genotype call of homozygous non-reference allele.

Phenotype calculation

A linear mixed model was used to estimate the best linear unbiased prediction (BLUP) for the phenotypic traits of inbred and hybrid lines. In the model, genotype (G), location (L), genotype by location interaction (G×L), replication (R) and block (B) were treated as random effects.

In addition, the calculated variance of factors were used to estimate broad-sense heritability (H_2) using the following formula (from Holland *et al.* (2003)):

$$H_2 = \frac{V_G}{V_G + V_{G\times L}/N_L + V_\epsilon/(N_R \times N_L)} \quad (1)$$

Table 1 Comparison between RAD-seq and tGBS genotyping technologies

	RAD-seq	tGBS
Total number of samples genotyped	372	384
Sequencing platform	Paired-end Illumina HiSeq 2000	Single-end Ion Proton
Average (Median) Reads/Sample after QC	12,221,976 (12,097,256)	1,793,300 (1,365,265)
Average (Median) Sequence/Sample after QC	965,295,176 (955,561,340)	195,057,311 (146,026,776)
Average (Median) missing rate / SNP	41.39% (41.67%)	58.65% (63.02%)
Average (Median) Proportion Het Calls / SNP before imputation	2.05% (0.42%)	4.12% (3.82%)
Average (Median) Proportion Het Calls / SNP after imputation	1.63% (0.53%)	4.72% (2.86%)
Average (Median) MAF / SNP before imputation	1.89% (1.18%)	11.69% (5.43%)
Average (Median) MAF / SNP after imputation	1.24% (0.67%)	10.37% (3.26%)
Total SNPs	649,067	73,291
SNPs with MAF >0.05 after imputation	15,306	32,463

where N_L is the number of locations and N_R is the number of replicates. V_G, $V_{G\times L}$ and V_ϵ represent the variance of the studied phenotypes controlled by genotype (G), interaction between genotype and environment (G×E) and residual factors.

In the analysis presented in Figure 2A, variance attributed to residual includes both V_ϵ as well as block and replicate variance. A customized R package (available at: https://jyanglab.github.io/g3tools/) was used to perform the above analyses.

The mid-parent heterosis for each trait was calculated from the BLUP values using the following formula:

$$\frac{Y_{hybrid} - \left(Y_{female} + Y_{male}\right)/2}{\left(Y_{female} + Y_{male}\right)/2} \quad (2)$$

Genomic selection and cross-validation

BLUP values for lines with genotypic information – either direct genotyping for inbred lines or projected genotypes for hybrids – were used as training data for genomic prediction. For each approach described in results, predictions were conducted using the implementation of RR-BLUP (Ridge Regression Best Linear Unbiased Prediction) in the R-package rrBLUP Endelman (2011). The genomic prediction model was represented as:

$$y = \mu + \sum_{i=1}^{k} x_i g_i + e \quad (3)$$

where y is the matrix of BLUPs of all individuals, μ is the overall mean, k is the total number of SNPs, x_i is the ith SNP genotype, g_i is the effect for ith SNP and e is the residual.

For each trait, randomly selected subsets of SNPs ranging from 64 (2^6) to 16,384 (2^{14}) plus total projected hybrid SNPs were tested. For each subsampled set of SNPs, the individuals were divided into 5 groups, and five separate genomic prediction analyses were conducted, using four of the five groups as training data and the remaining group as testing data. The mean correlation coefficient across these five sub-predictions was treated as a single estimate of the accuracy of the prediction model for estimates of accuracy and standard deviation.

The creation of the five individual sub-predictions varied somewhat across the four different prediction schemes described below (Figure 1). For each set of parameters with each scheme, a total of 20 sets of fivefold of cross-validation were performed. Thus, for each number of SNPs for each trait, a total of 20 × 5 = 100 sets of predictions were made. In scheme 1 (M1), the total set of genotyped and phenotyped inbreds was divided into five equal groups. Each sub-prediction used four of these five groups to predict phenotypic values for all genotyped and phenotyped hybrids. Scheme 2 (M2) utilized conventional five fold cross validation where the total set of genotyped and phenotyped hybrids was divided into five equal groups, and each sub-prediction used training data from four of the five groups to predict the remaining 20% of the data. The other two schemes utilized the same system as scheme 2, with the addition of all genotyped and phenotyped inbreds to the training dataset for all five subpredictions which either used BLUPs calculated across all individuals (M3A) or BLUPs calculated separately for inbred and hybrid populations (M3B).

To assess the accuracy of predictions for hybrids where one or more parents are completely unobserved, one B-line and one R-line were selected as "hold out" parents, and all hybrids which had either of these lines as a parent were excluded prior to the division of the remaining data into five groups. Each sub-prediction consisted of training the model using the hybrids four of these the five groups, and then predicting the genomic estimated breeding values of the hybrids with a "hold out" parent. Relative to the analysis without hold-out parents, prediction accuracy in this scenario decreased modestly and variance in prediction accuracy increased dramatically (Figure S2). Finally, all hybrids with genotype and genotype data were used to train a model that then produced genomic estimated breeding values for all 36,864 possible hybrid (Figure S3).

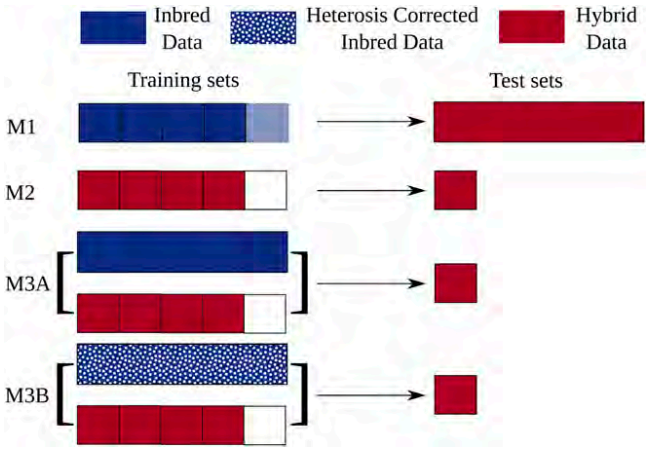


Figure 1 Four approaches taken to training and testing genomic prediction schemes. Scheme 1 (M1) uses different sets of 4/5s of the inbred phenotypic data to build a model which is tested by comparing predicted and measured traits for all hybrids. Scheme 2 (M2) is conventional fivefold cross validation, where the hybrids tested are divided into five equal parts, and the genomic estimated breeding values for hybrids in each are predicted using a model trained with the other four parts of the dataset. Scheme 3A (M3A), follows the same strategy outlined for M2, with the the training set extended to include the phenotypic and genotypic data for the inbred lines from M1. Scheme 3B (M3B) follows the same strategy as M3A but normalizes for the separate mean trait values of the inbred and hybrid populations prior to combining them into the training dataset.

Data availability

The authors affirm that all data necessary for confirming the conclusions of this article are represented fully within the article and its tables and figures. Supplemental material available at Figshare: doi: https://doi.org/10.6084/m9.figshare.5969230; doi: https://doi.org/10.6084/m9.figshare.5566843.

RESULTS

Phenotype analysis

Phenotypic variance was partitioned into four components: genotype (G), environment (E), interaction between genotype and environment (G×E), residual (R). For each trait, the pattern of relative contribution of each of these factors was roughly similar between inbred and hybrid pearl millet populations (Figure 2A). Plant height was the trait with the greatest proportion of variance explained by purely genetic factors, while flowering time had the great proportion of variance explained by environments. As expected, grain yield had the highest residual value, making this critical trait the most difficult to predict accurately using quantitative genetic models. Broad sense heritability – *i.e.*, the proportion of total variance in trait values explained by genetic factors – for grain yield, plant height, flowering time and 1000-grain weight were estimated to be 0.60, 0.86, 0.88 and 0.74 respectively for pearl millet

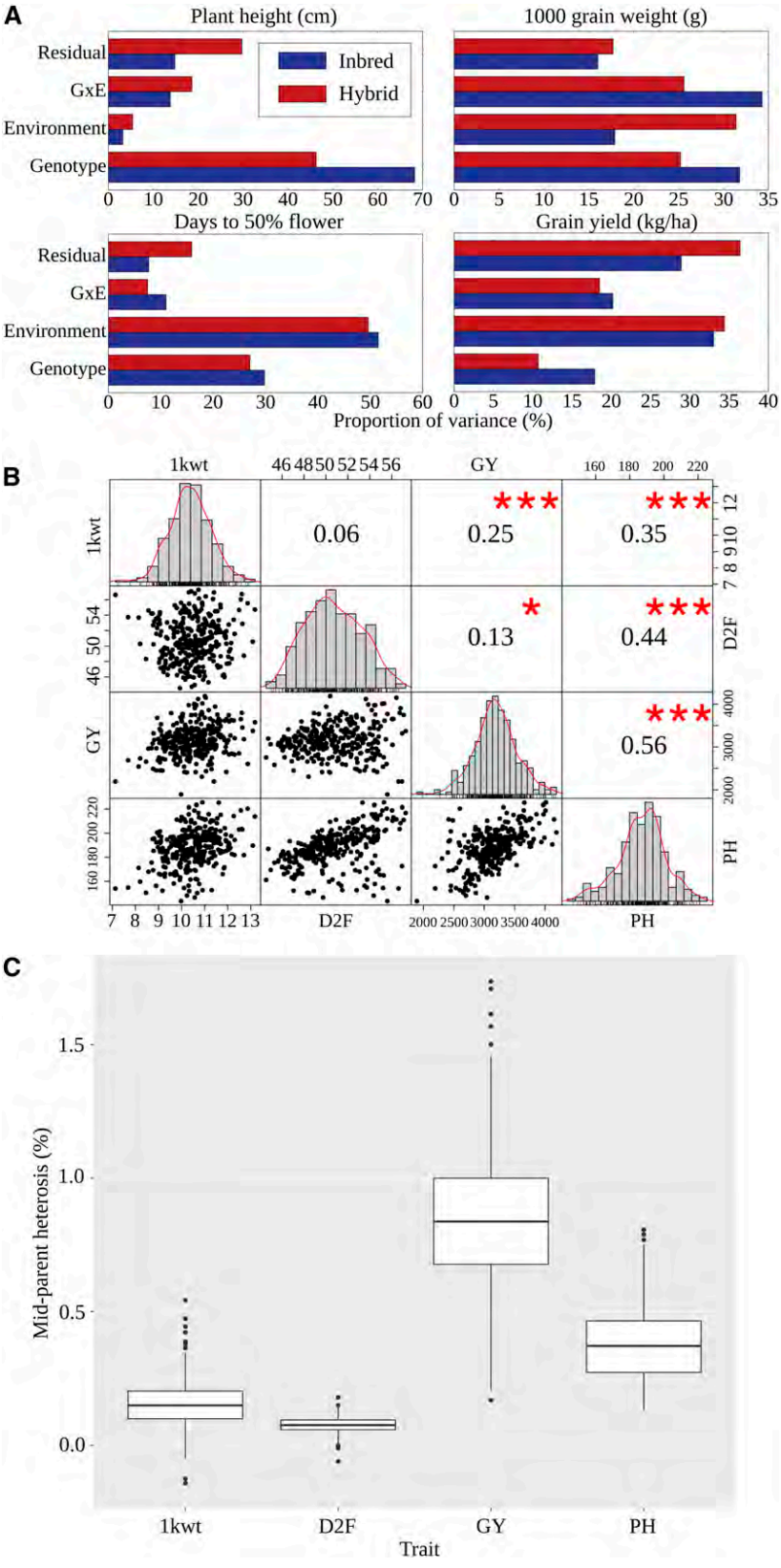


Figure 2 (A) Proportion of phenotypic variance explained by genotype, location (considered as a environmental factor), genotype by location (GxE) interaction for either inbred pearl millet lines or hybrid pearl millet lines. (B) Phenotype investigation of four studied traits in pearl millet population. *** p value of the significance of this correlation is ≤ 0.001, ** p value of the significance of this correlation is ≤ 0.01 and * p value of the significance of this correlation is ≤ 0.05; (C) Distribution of observed mid-parent heterosis for each of the four traits scored in this study.

hybrids and 0.72, 0.92, 0.88 and 0.74 for inbreds. However, caution should be used in interpreting differences between the inbred and hybrid heritability values, given the large difference in the number of individuals between the two populations. BLUP values were calculated for these four traits (see methods). Each trait exhibited an approximately normal distribution (Figure 2B). Heterosis can be defined in

different ways but the definition employed here is mid-parent heterosis which is the degree to which the measured trait values of hybrids tend to exceed the average of measured values for the same trait in both parents. Mid-parent heterosis was observed for all four traits with median values of 8% (flowering time), 15% (1000-grain weight), 37% (plant height) and 84% (grain yield) (Figure 2C). Note that the direct of effect for heterosis was reversed for flowering time. This is consistent with studies in maize which indicate that hybrid tend to flower earlier than their parents Dickert and Tracy (2002).

Characteristics of the tGBS and RAD-seq datasets

A total of 4,550 million barcoded RAD-seq reads were generated on an Illumina HiSeq 2000, for an average of 12.2 million reads per sample. tGBS libraries were sequenced using seven Ion Proton runs, generating a total of 810 million raw reads included 584 million of barcoded reads, for an average of 2.1 million reads per sample.

After aligning to the pearl millet reference genome and quality filtering (See Methods), 649,067 polymorphic SNPs were identified using RAD-seq data and 73,291 SNPs were identified using tGBS data. As expected given the different subsets of the genome targeted by these two technologies Miller *et al.* (2007); Ott *et al.* (2017), there was only minimal overlap between the two methods with only 439 SNPs identified and scored by both technologies. The missing data rates for RAD-seq genotypes exhibited a bimodal distribution while tGBS genotypes exhibited a unimodal distribution skewed toward high missing data rates. RAD-seq genotyping was much less likely to genotype sites as heterozygous, which may reflect a difference in the technologies, or may be explained by the observation that many SNPs identified by RAD-seq had low minor allele frequencies, while tGBS SNPs, tended to have higher minor allele frequencies (Figure 3). A more detailed comparison of the outcomes of RAD-seq and tGBS genotyping is provided in Table 1. Downstream analyses utilized only those SNPs with MAF >0.05 from each dataset (15,306 RAD-seq SNPs and 32,463 tGBS SNPs).

Evaluating the accuracy of genomic prediction

The ability of genomic prediction using projected hybrid genotypes from both genotyping methods was then assessed for each phenotype using cross validation. For each tested set of SNPs, 20 random rounds of fivefold cross validation were performed. The median correlation coefficient of 1000-grain weight, days to flowering time, grain yield and plant height using all available SNPs was 0.73, 0.89, 0.51 and 0.72 for RAD-seq and 0.74, 0.87, 0.48 and 0.73 for tGBS (Figure S1). The differences in prediction accuracies observed for the two methods, either utilizing random sub-sampling of equal numbers of SNPs for each dataset or all SNPs obtained using each genotyping method were not statistically significant (student's *t*-test).

Comparison of prediction models

Four different approaches (see Methods) to genomic prediction were evaluated to test whether inbred phenotypic data can add value to genomic prediction as part of a hybrid breeding program. Scheme M1, which used trait trait from inbreds to predict genomic estimated breeding values of hybrids performed the worst of all four approaches for all four phenotypic traits tested. Notably, the rank of traits by mid-parent heterosis had a perfect negative correlation with the rank of the traits by phenotypic prediction accuracy using inbred parent training data. Scheme 2 (M2) produced a significant increase ($P < 0.05$) in accuracy relative to the scheme 1 (M1) for all four phenotypes, although, consistent with its high residual values when fitting the original

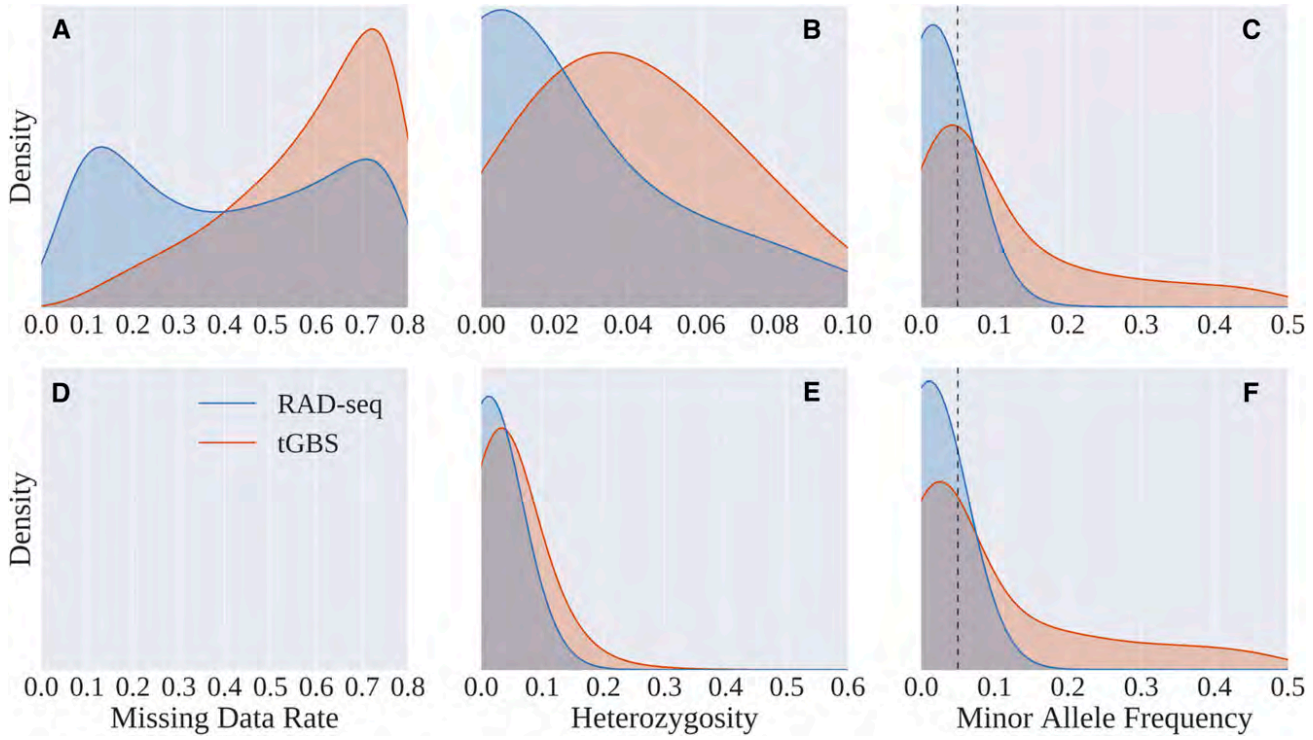


Figure 3 Distribution of missing data rates (A, D), heterozygosity (B, E), and minor allele frequency (C, F) for SNPs identified and scored in either the RAD-seq or tGBS dataset. A-C summarize raw SNP data prior to imputation. D-F show densities for the same characteristics subsequent to imputation. However, no missing sites were left after imputation, hence panel D is blank. Dashed line in C & F indicates the cut off of MAF = 0.05 for SNPs which were utilized in downstream genomic prediction.

BLUP, the accuracy of prediction for yield was the lowest of the four phenotypes (Figure 4). Scheme 3A (M3A), which merged phenotypes from both inbred parents and hybrid trials to predict hybrid trials performed equivalently to scheme 2 (M2) for 1000-grain weight (1kwt) and flowering time (D2F), the two traits with the lowest degree of mid-parent heterosis. However, for traits with significant amounts of mid-parent heterosis (grain yield and plant height), adding inbred parent phenotype data to the prediction model either provided no benefit (plant height, $P = 0.17$) or significant *decreases* in prediction accuracy (grain yield, $P = 0.09e\text{-}2$) compared to a purely hybrid phenotype data training set. Scheme 3B (M3B), which, instead of employing absolute trait values for inbreds and hybrids, summarized phenotypes as the differences between the predicted trait value for a given inbred or hybrid line and the mean trait value for either all hybrids or all inbreds (Figure 5D), performed approximately equal to, or sometimes marginally better than M2 (hybrid only scheme). An additional 1,000 permutations of fivefold cross validation were conducted for scheme 2 (M2) and scheme 3B (M3B) using the "All SNPs" dataset. The increase in prediction accuracy in M3B relative to M2 was statistically significant for two out of four traits tested: flowering time ($P = 6.00e\text{-}4$), and grain yield ($P = 5.03e\text{-}9$).

Increases in prediction accuracy coming from increasing numbers of markers tended to plateau at smaller total marker numbers for scheme 1 (M1 which is inbred only) than for scheme 2 (M2 which is hybrid only), with scheme 3 (M3 which is inbreds plus hybrids) was intermediate between the two. However, even 64 (2^6) random SNPs provided significant ($P < 0.05$) predictive ability for all traits and all schemes tested. The relatively small set of inbred parents used to create the set of hybrids tested as part of this analysis may have resulted in inflated apparent prediction accuracies for each trait. When using a hold-two-parents out approach to segregating hybrids with common parentage between the training and testing datasets (see Methods), accuracy was lower and standard deviations of prediction accuracy were higher (Figure S2), indicating that our estimates of prediction accuracy are likely to be optimistic relative to the ability to predictions for hybrids where one or both parents have not previously served as parents for tested hybrids.

Finally, grain yield and time to flowering were predicted for every possible F1 hybrid between a genotyped A/B-line and a genotyped R-line in the dataset. Within this prediction space, the highest yielding potential hybrids tend to be associated with somewhat longer flowering times defining a production possibility frontier for the trade off between growing season length and yield among the pearl millet hybrids which could be generated using the inbred germplasm genotyped as part of this study (Figure S3).

DISCUSSION

Here we found that the naive integration of trait data collected from inbred lines into genomic prediction for a hybrid breeding program can actually reduce prediction accuracy, particularly for traits exhibiting significant heterosis. However, controlling for the effect of heterosis *per se* by calculating BLUPs separated for inbred and hybrid lines eliminated this negative impact on prediction accuracy and could in fact

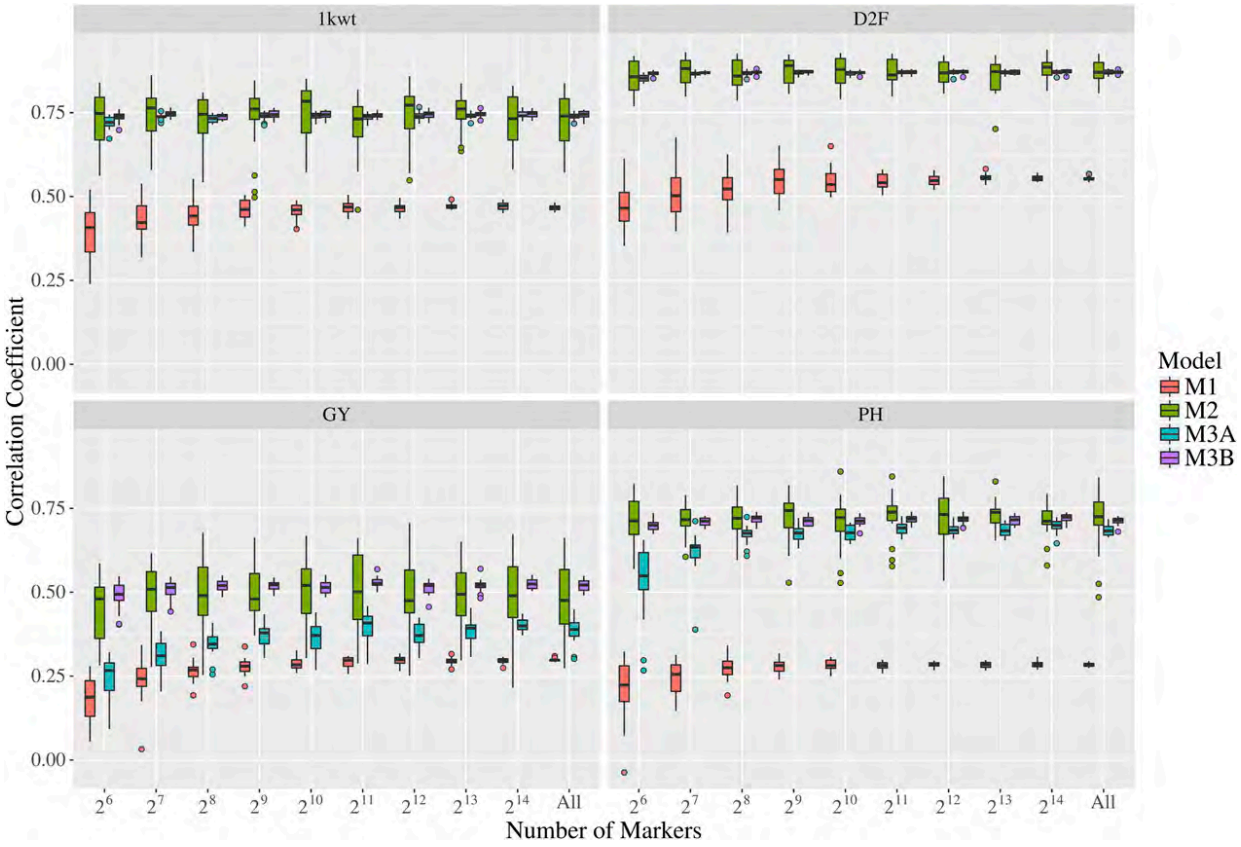


Figure 4 Prediction accuracy for each of four phenotypes scored in this pearl millet population employing the four schemes outlined in Figure 1 using tGBS SNP calls. Scheme 3A (M3A) employed absolute predicted trait values for inbreds and hybrids to train a genomic prediction model, while scheme 3B (M3B) employed predicted trait data for inbreds and hybrids calculated relative to the separate mean trait values for inbred and hybrid lines.

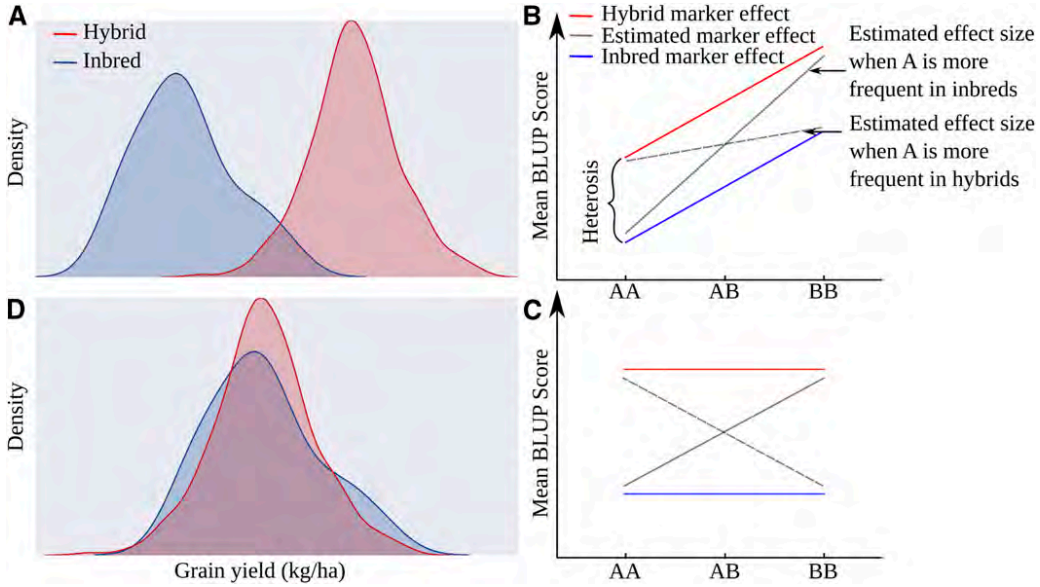


Figure 5 A proposed model for the decrease in genomic prediction accurary for high heterosis traits when inbred individuals are introduced into training populations. A) Distribution of BLUP scores for yield for populations of hybrid and inbred individuals based on a combined BLUP analysis. B) Distribution of scores for a hypothetical marker having an equally large effect size in inbred and hybrid individuals. When allele frequencies differ between these populations, and the ratio of hybrid to inbred individuals may vary between the groups of individuals with genotype AA or with genotype BB. C) Distribution of scores for a hypothetical marker with no effect on trait value. D) Distribution of BLUP scores for yield for populations of hybrid and inbred individuals based on a separate BLUP analysis for hybrid and inbred individuals.

increase prediction accuracy for some traits relative to predicting using data from hybrid lines alone. In addition we found prediction accuracy was equivalent for SNPs generated through either RAD-seq or tGBS. Given the greater number of high MAF (>0.05) SNPs generated per million reads with tGBS, this methodology is likely to be more cost effective in the context of many genomic selection based breeding programs. In addition, the more rapid turn-around time enabled by Ion Proton sequencing (~4 hr) relative to Illumina HiSeq 2000 (~8 days for 2x100 sequencing) increases the feasibility the utilizing genomic prediction to guide real time breeding decisions. However, it should also be noted that there is nothing inherent about either the RAD-seq or tGBS protocol which prevents the adaptation of either protocol to sequencing using either instrument.

Longer growing seasons will generally – in the absence of constraints from temperature, or resources abundance – result in more total fixed carbon Dohleman and Long (2009). Whether this increase in carbon fixation results in an increase in yield depends, among other factors on harvest index, the partitioning of carbon between vegetative and reproductive development. In our data, grain yield displayed only weak correlations with flowering time (Figure 2B). Grain yield and plant height show a strong positive correlation with each other which is not what would be expected based on models of carbon partitioning. One potential explanation is that, in small plot trials with substantial height variation among accessions, tall plots can shade shorter plots if the experimental design is not blocked by height. This shading effect produces an apparent yield penalty for short accessions which does not translate to larger scale yield trials or commercial production. However, there may also be significant room to improve the yield and resource use efficiency of pearl millet through selection for improved harvest index. A second explanation is loci responsible for tolerance of heat and or drought stress are segregating in the population, sensitive genotypes are likely to exhibit both stunted growth and low grain yields, as pearl millet is grown in marginal soils and high levels of abiotic stress (high temperatures and water constraint).

As described above, we found that the naive incorporation of inbred genotype and trait information into training datasets decreased prediction accuacy for high heterosis traits. In Figure 5, we propose a model to explain this finding. When BLUP values for a trait with a high degree of heterosis are calculated in a population containing a mix of inbred and hybrid individuals, most inbreds will be assigned negative BLUP scores and most hybrids postive BLUP scores (Figure 5A). Distributions of allele frequencies will vary between inbred and hybrid populations. As a result, inbred individuals may be relatively more common among the population of individuals with AA or BB genotypes. Alleles more common in inbred individuals relative to hybrid individuals will tend to be assigned a more negative or less positive effect value by a genomic prediction model trained with a mixed hybrid/inbred population than by a genomic prediction model trained on a purely hybrid or purely inbred population (Figure 5B Figure 5C). Calculating BLUPs separately for inbred and hybrid individuals removes this source of bias in the training data by centering the distributions of both inbred and hybrid individuals (Figure 5D).

Consistent with earlier studies in maize, we found that inbred trait values alone were poor predictors of hybrid performance, particularly for yield e Gama and Hallauer (1977); Smith (1986), although the prediction values in scheme 1 (M1), trained only on inbred data were at least statistically significantly greater than zero ($P < 0.05$). Here we found that when using a conventional additive genomic prediction model (RR-BLUP), traits with higher median heterosis (grain yield and plant height) experienced a decrease in prediction accuracy when inbred data were naively incorporated into the training dataset (M3A). Segregating BLUP means for inbreds and hybrids (M3B) statistically significantly moderately enhanced prediction accuracy for grain yield ($P = 5.03e\text{-}9$) and flowering time ($P = 6.00e\text{-}4$), compared to a scheme

which excluded phenotypic data from inbred parents (M2). While statistically significant, the absolute values of the increases in prediction accuracy are moderate: flowering time (M2: 0.87, M3B: 0.87) and grain yield (M2: 0.50, M3B: 0.52). Yield was both the most difficult trait to predict, and the trait where the inclusion of inbred trait data provided the largest increase in prediction accuracy. In this study the number of inbred lines for which genotypic and phenotypic data were available was quite small. Even in scheme 3B, inbred lines made up less than 15% of the training dataset. Given that extensive inbred trait datasets exist for pearl millet, it may be that the incorporation of genotypic and phenotypic data from larger numbers of inbred lines would produce a larger absolute increase in prediction accuracy. As inbred lines must be grown prior to the production of hybrid seed, the collection of trait data from these lines comes at relatively low cost, and may have additional value when integrated into training datasets which also include genotypic and phenotypic data from a sample of hybrid lines.

Even small numbers of selected SNPs can achieve relatively high prediction accuracy in this pearl millet population. The implementation of a hybrid GS/GP guided pearl millet breeding program has the potential to significantly improve the efficiency of breeding efforts (Figure 4). However, it must be noted that in our training set the high representation of haplotypes drawn from 33 common parental lines produces close relationships between sub-sampled training and testing populations, and this could also be a reason to explain why a smaller number set of SNPs can reach plateaus in accuracy for genomic prediction of some studied traits. As a result, our estimates of model prediction accuracy are likely inflated relative to predictions on unrelated populations Isidro *et al.* (2015). To expand the applicability of this genomic prediction model to a wider pearl millet genomic selection assisted breeding program, it will be necessary to incorporate data from a hybrids derived from a broader genetic basis.

ACKNOWLEDGMENTS

This work was supported by an CGIAR-US Universities Linkage Program on Dryland Cereals (404-40-89) to PSS, JCS, SKG and RKS.

LITERATURE CITED

Albrecht, T., H.-J. Auinger, V. Wimmer, J. O. Ogutu, C. Knaak *et al.*, 2014 Genome-based prediction of maize hybrid performance across genetic groups, testers, locations, and years. Theor. Appl. Genet. 127: 1375–1386. https://doi.org/10.1007/s00122-014-2305-z

Dickert, T., and W. Tracy, 2002 Heterosis for flowering time and agronomic traits among early open-pollinated sweet corn cultivars. J. Am. Soc. Hortic. Sci. 127: 793–797.

Dohleman, F. G., and S. P. Long, 2009 More productive than maize in the midwest: how does miscanthus do it? Plant Physiol. 150: 2104–2115. https://doi.org/10.1104/pp.109.139162

e Gama, E. E. G., and A. R. Hallauer, 1977 Relation between inbred and hybrid traits in maize 1. Crop Sci. 17: 703–706. https://doi.org/10.2135/cropsci1977.0011183X001700050007x

Endelman, J. B., 2011 Ridge regression and other kernels for genomic selection with r package rrblup. Plant Genome 4: 250–255. https://doi.org/10.3835/plantgenome2011.08.0024

Gowda, M., Y. Zhao, H. P. Maurer, E. A. Weissmann, T. Würschum *et al.*, 2013 Best linear unbiased prediction of triticale hybrid performance. Euphytica 191: 223–230. https://doi.org/10.1007/s10681-012-0784-z

Gupta, S., K. Rai, P. Singh, V. Ameta, S. K. Gupta *et al.*, 2015 Seed set variability under high temperatures during flowering period in pearl millet (*Pennisetum glaucum* L. (R.) Br.). Field Crops Res. 171: 41–53. https://doi.org/10.1016/j.fcr.2014.11.005

Gupta, S., A. Rathore, O. Yadav, K. Rai, I. Khairwal *et al.*, 2013 Identifying mega-environments and essential test locations for pearl millet cultivar selection in india. Crop Sci. 53: 2444–2453. https://doi.org/10.2135/cropsci2013.01.0053

Hanna, W., 1989 Characteristics and stability of a new cytoplasmic-nuclear male-sterile source in pearl millet. Crop Sci. 29: 1457–1459. https://doi.org/10.2135/cropsci1989.0011183X002900060026x

Holland, J., W. Nyquist, and C. Cervantes-Martínez, 2003 Estimating and interpreting heritability for plant b: an update, pp. 9–112 in *Plant Breeding Reviews*, edited by Janick, J. John Wiley & Sons, Hoboken, NJ.

Isidro, J., J.-L. Jannink, D. Akdemir, J. Poland, N. Heslot *et al.*, 2015 Training set optimization under population structure in genomic selection. Theor. Appl. Genet. 128: 145–158. https://doi.org/10.1007/s00122-014-2418-4

Kadam, D. C., S. M. Potts, M. O. Bohn, A. E. Lipka, and A. J. Lorenz, 2016 Genomic prediction of single crosses in the early stages of a maize hybrid breeding pipeline. G3: Genes, Genomes. Genetics 6: 3443–3453. https://doi.org/10.1534/g3.116.031286

Mace, E. S., K. K. Buhariwalla, H. K. Buhariwalla, and J. H. Crouch, 2003 A high-throughput dna extraction protocol for tropical molecular breeding programs. Plant Mol. Biol. Report. 21: 459–460. https://doi.org/10.1007/BF02772596

Miller, M. R., J. P. Dunham, A. Amores, W. A. Cresko, and E. A. Johnson, 2007 Rapid and cost-effective polymorphism identification and genotyping using restriction site associated dna (rad) markers. Genome Res. 17: 240–248. https://doi.org/10.1101/gr.5681207

Mühleisen, J., H.-P. Piepho, H. P. Maurer, and J. C. Reif, 2015 Yield performance and stability of cms-based triticale hybrids. Theor. Appl. Genet. 128: 291–301. https://doi.org/10.1007/s00122-014-2429-1

Ott, A., S. Liu, J. C. Schnable, C.-T. E. Yeh, K.-S. Wang *et al.*, 2017 tgbs genotyping-by-sequencing enables reliable genotyping of heterozygous loci. Nucleic Acids Res. 45: e178. https://doi.org/10.1093/nar/gkx853

Poland, J., J. Endelman, J. Dawson, J. Rutkoski, S. Wu *et al.*, 2012 Genomic selection in wheat breeding using genotyping-by-sequencing. Plant Genome 5: 103–113. https://doi.org/10.3835/plantgenome2012.06.0006

Ramya, R., L. Ahamed, R. Yadav, P. Katiyar, B. Raj *et al.*, 2017 Towards defining heterotic gene pools in pearl millet [*Pennisetum glaucum* L. (R.) Br.]. Front. Plant Sci. 8: 1934. https://doi.org/10.3389/fpls.2017.01934

Schrag, T. A., H. P. Maurer, A. E. Melchinger, H.-P. Piepho, J. Peleman *et al.*, 2007 Prediction of single-cross hybrid performance in maize using haplotype blocks associated with qtl for grain yield. Theor. Appl. Genet. 114: 1345–1355. https://doi.org/10.1007/s00122-007-0521-5

Smith, O., 1986 Covariance between line per se and testcross performance 1. Crop Sci. 26: 540–543. https://doi.org/10.2135/cropsci1986.0011183X002600030023x

Spindel, J., H. Begum, D. Akdemir, P. Virk, B. Collard *et al.*, 2015 Genomic selection and association mapping in rice (*Oryza sativa*): effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite, tropical rice breeding lines. PLoS Genet. 11: e1004982 (correction: PLoS Genet. 11: e1005350). https://doi.org/10.1371/journal.pgen.1004982

Tabangin, M. E., J. G. Woo, and L. J. Martin, 2009 The effect of minor allele frequency on the likelihood of obtaining false positives. BMC Proc. 3: S41 BioMed Central. https://doi.org/10.1186/1753-6561-3-S7-S41

Technow, F., T. A. Schrag, W. Schipprack, E. Bauer, H. Simianer *et al.*, 2014 Genome properties and prospects of genomic prediction of hybrid performance in a breeding program of maize. Genetics 197: 1343–1355. https://doi.org/10.1534/genetics.114.165860

Varshney, R. K., C. Shi, M. Thudi, C. Mariac, J. Wallace *et al.*, 2017 Pearl millet genome sequence provides a resource to improve agronomic traits in arid environments. Nat. Biotechnol. 35: 969–976. https://doi.org/10.1038/nbt.3943

Windhausen, V. S., G. N. Atlin, J. M. Hickey, J. Crossa, J.-L. Jannink *et al.*, 2012 Effectiveness of genomic prediction of maize hybrid performance in different breeding populations and environments. G3: Genes, Genomes. Genetics 2: 1427–1436. https://doi.org/10.1534/g3.112.003699

Wu, T. D., and S. Nacu, 2010 Fast and snp-tolerant detection of complex variants and splicing in short reads. Bioinformatics 26: 873–881. https://doi.org/10.1093/bioinformatics/btq057

Xing, N., C. Fan, and Y. Zhou, 2014 Parental selection of hybrid breeding based on maternal and paternal inheritance of traits in rapeseed (brassica napus l.). PLoS One 9: e103165. https://doi.org/10.1371/journal.pone.0103165

Yadav, O. P., K. Rai, H. Yadav, B. Rajpurohit, S. Gupta *et al.*, 2016 Assessment of diversity in commercial hybrids of pearl millet in india. Indian J. Plant. Genet. Resour. 29: 130–136. https://doi.org/10.5958/0976-1926.2016.00018.8

Yu, J., X. Li, C. Zhu, C.-T. Yeh, W. Wu *et al.*, 2012 Genic and non-genic contributions to natural variation of quantitative traits in maize. Genome Res. 22: 2436–2444. https://doi.org/10.1101/gr.140277.112

Zhong, S., J. C. Dekkers, R. L. Fernando, and J.-L. Jannink, 2009 Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: a barley case study. Genetics 182: 355–364. https://doi.org/10.1534/genetics.108.098277

Communicating editor: A. Doust

Parallel natural selection in the cold-adapted crop-wild relative *Tripsacum dactyloides* and artificial selection in temperate adapted maize

Lang Yan[1,2,5,+], Sunil Kumar Kenchanmane Raju[1,7,+], Xianjun Lai[1,2,4,+], Yang Zhang[1,8], Xiuru Dai[1], Oscar Rodriguez[3], Samira Mahboub[1,6], Rebecca L. Roston[1,6], and James C. Schnable[1,3*]

[1]Center for Plant Science Innovation, University of Nebraska-Lincoln, NE, 68588, USA

[2]Laboratory of Functional Genome and Application of Potato, Xichang University, Liangshan, 615000, China

[3]Department of Agronomy and Horticulture, University of Nebraska-Lincoln, Lincoln, 68588, USA

[4]Maize Research Institute, Sichuan Agricultural University, Chengdu, 611130, China

[5]College of Life Sciences, Sichuan University, Chengdu, 610065, China

[6]Department of Biochemistry, University of Nebraska-Lincoln, Lincoln, Nebraska, USA

[7]Current Address: Department of Plant Biology, Michigan State University, East Lansing, Michigan, USA

[8]Current Address: St. Jude Children's Research Hospital, Memphis, Tennessee, USA

[*]To whom correspondence should be addressed. Tel: +1(402)472-3192; E-mail: schnable@unl.edu

[+]these authors contributed equally to this work

Running head: parallel selection in maize and tripsacum

ABSTRACT

Artificial selection has produced varieties of domesticated maize which thrive in temperate climates around the world. However, the direct progenitor of maize, teosinte, is indigenous only to a relatively small range of tropical and sub-tropical latitudes and grows poorly or not at all outside of this region. *Tripsacum*, a sister genus to maize and teosinte, is naturally endemic to the majority of areas in the western hemisphere where maize is cultivated. A full-length reference transcriptome for *Tripsacum dactyloides* generated using long-read isoseq data was used to characterize independent adaptation to temperate climates in this clade. Genes related to phospholipid biosynthesis, a critical component of cold acclimation on other cold adapted plant lineages, were enriched among those genes experiencing more rapid rates of protein sequence evolution in *T. dactyloides*. In contrast with previous studies of parallel selection, we find that there is a significant overlap between the genes which were targets of artificial selection during the adaptation of maize to temperate climates and those which were targets of natural selection in temperate adapted *T. dactyloides*. This overlap between the targets of natural and artificial selection suggests genetic changes in crop-wild relatives associated with adaptation to new environments may be useful guides for identifying genetic targets for breeding efforts aimed at adapting crops to a changing climate.

Keywords: parallel evolution, adaptation, cold tolerance, maize, crop wild relatives

Significance Statement

Corn was domesticated in Central America and is very sensitive to cold and freezing temperatures. Eastern gamagrass is a close relative of corn, is native to prairies throughout the United States, east of the rocky mountains, the region we now call the corn belt and can survive the winter. We compared rates of protein sequence evolution across the same genes in seven grass species to identify genes likely to be involved in adapting gamagrass to life in the corn belt. We identified a specific metabolic pathway likely involved in cold and freezing tolerance and also found that many of the same genes were targets of selection when humans started developing new varieties of corn to grow in temperate North America.

Introduction

The common ancestor of maize and *Tripsacum dactyloides* was adapted to a tropical latitude, yet today domesticated maize and wild *T. dactyloides* both grow in large temperate regions of the globe. The adaptation of tropical maize landraces to temperate environments required changes to flowering time regulation and adaption to new abiotic and biotic stresses [1]. As both a leading model for plant genetics and one of the three crops that provides more than one half of all calories consumed by humans around the world, maize (*Zea mays* ssp. *mays*) and its wild relatives have been the subject of widespread genetic and genomic investigations. The closest relatives of maize are the teosintes, which include the direct wild progenitor of the crop (*Z. mays* ssp. *parviglumis*) as well as a number of other teosinte species within the genus *Zea* (Table 1). Outside the genus *Zea*, the closest relatives of maize are the members of the sister genus *Tripsacum* (Figure 1A-H). Together, these two genera form the subtribe *Tripsacinae* within the tribe *Andropogoneae* [2]. Despite their close genetic relationship, some species of the genus *Tripsacum* are adapted to a much wider range of climates than wild members of the genus *Zea* (Figure 1).

The direct progenitor of maize, *Z. m. parviglumis*, is confined to a relatively narrow native range that spans tropical and subtropical areas of Mexico, Guatemala, Nicaragua and Honduras [3]. Other species and subspecies within the genus *Zea* are also largely confined to the same geographic region [4, 3]. In contrast, *T. dactyloides* is widely distributed through temperate regions of both North and South America [5, 6], largely mirroring the distribution of modern agricultural production of maize in the western hemisphere. The common ancestor of *Zea* and *Tripsacum* is predicted to have been adapted to tropical latitudes [5, 7, 8]. Therefore, the study of how natural selection adapted *T. dactyloides* to temperate climates represents an informative parallel to the adaption of maize to temperate climates through artificial selection. *T. dactyloides* is also a potential source of insight into the genetic changes responsible for traits such as disease and insect resistance, drought and frost tolerance, many of which are targets for maize improvement [9, 10, 11].

Until recently, molecular sequence data for *Tripsacum* species was largely been generated to serve as an outgroup for molecular evolution studies in maize (as reviewed [12]). As part of Hapmap2, 8x short read shotgun data generated from *T. dactyloides* [9], additional low pass genomic data has been generated for several other species in the Tripsacum genus [13], and recently, Illumina transcriptome assemblies of two species in the genus Tripsacum – *T. dactyloides* and *T. floridanum* became available [11]. Here we employ PacBio long-read sequencing to generate a set of full length transcript sequences from *T. dactyloides*. Using data from orthologous genes in maize, *T. dactyloides*, *Sorghum bicolor*, *Setaria italica*, and *Oropetium thomaeum*, a set of genes with uniquely high rates of nonsynonymous substitution in *T. dactyloides* were identified. We show that a surprisingly large subset of these genes are also identified as targets of selection during artificial selection for maize lines adapted to temperate climates. A specific metabolic pathway identified through this method – phospholipid metabolism – is linked to cold and freezing tolerance in other species and we demonstrate that the metabolic response of this pathway to cold stress shows functional divergence between maize and *T. dactyloides* .

Results

Sequencing and analysis of full-length *T. dactyloides* transcripts

PacBio Iso-seq of RNA isolated from a single *Tripsacum dactyloides* plant grown from seed collected from the wild in eastern Nebraska (USA) was used to generate 64,326 HQ consensus sequences (See Supplemental Results; Table S1). These sequences were aligned to the maize reference genome, which identified 24,616 isoforms corresponding to 14,401 annotated maize gene models (See Supplemental Results; Table S2). A total of 1,259 high quality consensus sequences from *T. dactyloides* failed to map to maize reference genome, which was generated from the maize inbred B73. This number is roughly consistent with another report of 1,737 *T. dactyloides* transcripts which lack orthologs in the maize genome [11]. Sequences which lacked apparent homologs in the maize reference genome were aligned to NCBI's RefSeq plant database. Two-hundred-and-sixty-three of these sequences aligned to genes from other grass species such as *Sorghum bicolor*, *Setaria italica*, *Oryza sativa* (Table S3). These genes may either represent genes missed when generating the maize reference genome assembly [14], genes present within the maize pan-genome but absent from the specific line used to generate the maize reference genome [15], or genes present in the common ancestor of maize and *T. dactyloides* but lost from the maize lineage sometime after the *Zea/Tripsacum* split. As a partial test of these different models, these sequences were aligned to the published genome sequence of a second

maize inbred (PH207) [16]. Only four of these sequences aligned to the PH207 genome, indicating that the majority of this population of sequences are more likely to represent losses from the maize linage sometime after the *Zea/Tripsacum* split, rather than genes left out of the maize reference genome assembly [14] or genes present within the maize pan-genome but absent from the specific line used to generate the maize reference genome [15].

Patterns of alternative splices observed between multiple *T. dactyloides* transcripts which aligned to the same maize genes followed similar patterns to those identified in studies of alternative splicing in maize based on either short read or long read technology (see supplemental results). However, despite greater sequencing depth and sampling a wider range of tissues, maize Iso-seq dataset did not identify alternative splicing events corresponding to the specific alternative splicing events identified in *T. dactyloides* in 85.7% of cases (2,447 of 2,856 orthologous genes). This result is consistent with the rapid divergence of most AS patterns between even closely related species (see Supplemental Results).

Identification of *T. dactyloides* genes experiencing rapid protein sequence evolution

A total of 6,950 groups of orthologous genes present in seven grass species were identified using the tripsacum-maize orthologous relationships (see supplemental results), plus an existing dataset of syntenic orthologous genes across six grass species with known phylogenetic relationships (Figure 2A) [17]. These groups included consisted of 4,162 one-to-one, 1,436 one-to-two and 1,352 two-to-two orthologous gene sets due to the WGD event shared by maize and *T. dactyloides*. The overall distribution of synonymous substitution (Ks) values for branches leading to individual species scaled with branch length (Figure 2B). However, while *Zea* and *Tripsacum* are sister taxa, the average maize gene showed more synonymous substitutions than the average *T. dactyloides* gene. In 1,775 cases the branch leading to *T. dactyloides* had the highest Ka/Ks ratio of all branches examined and in 1,114 the branch leading to maize had the highest Ka/Ks ratio (Figure 2C, Figure S2). This bias towards more gene groups showing the highest Ka/Ks values in maize or *T. dactyloides* rather than showing the highest values sorghum, foxtail millet, or oropetium, as well as the presence of more extremely high outlier values in these two species (Figure 2C), likely results from the fact that Ka/Ks ratios are based on a smaller absolute counts of substitutions per gene along shorter branches and therefore exhibit greater variance. In the analyses below, the set of genes experiencing accelerated rates of protein sequence evolution in maize were used as a control set for any analysis of patterns observed in fast evolving *T. dactyloides* genes.

Genes with signatures of rapid evolution in *T. dactyloides* tended to be associated with the functional annotations ”stress response” and ”glycerophospholipid metabolic process”, whereas fast-evolving genes in maize were enriched in the functions microtubule cytoskeleton organization, nutrient reservoir activity and ATPase activity. Figure 3A illustrates the distribution of Ka/Ks ratios in *T. dactyloides* and maize for genes where the branch leading to one of these two species exhibited the highest Ka/Ks ratio among the five species examined. Multiple fast-evolving genes involving in cell response to stimulus and stress had extremely high Ka/Ks ratios (> 1) in *T. dactyloides*, consistent with positive selection for increased abiotic stress tolerance in *T. dactyloides* relative to maize and other related grasses. The annotated functions of the maize orthologs of *T. dactyloides* genes experiencing accelerated protein sequence evolution include cold-induced protein, drought-responsive family protein, salt tolerance family protein, etc. (Table S4).

The phospholipid metabolic pathway is a specific focus of accelerated protein sequence evolution in *T. dactyloides*

In the process of identifying *T. dactyloides* genes that might experience accelerated evolution for temperate climate adaptation, we noticed multiple genes annotated as participating in the phospholipid metabolism pathway where the highest Ka/Ks ratio for that gene were observed in the branch leading to *T. dactyloides*. While several genes in this pathway also showed signs of accelerated evolution in maize, the bias towards high rates of protein sequence change in *T. dactyloides* was dramatic (Figure 3A). Using log-transformed Ka/Ks values, genes in the phospholipid biosynthesis pathway exhibited a significantly higher range of Ka/Ks values than the background set of other genes (p-value = 1.87e-04). In contrast, maize genes in the same exhibited significantly lower Ka/Ks values than background maize genes (p-value = 2.38e-04). Comparing the ratio of Ka/Ks values for between same genes in both maize and *T. dactyloides*, genes in the phospholipid biosynthesis pathway showed significantly higher ratios of Ka/Ks than background genes (p-value = 4.16e-05) (Figure S3A).

Phospholipids are a class of lipids that are a major component of cell membranes and include lipids with head groups such as phosphatidate (PA), phosphatidylethanolamine (PE), phosphatidylcholine (PC), phosphatidylglycerol(PG) and phosphatidylserine (PS) which share overlapping biosynthesis pathways and are often inter-convertible (Figure 4a). The set of genes involved in phospholipid metabolism which experienced accelerated evolution in *T. dactyloides* were particularly concentrated in the pathway leading to PC which is the major phospholipid component of non-plastid membranes. PC also tends to control membrane desaturation through acyl-editing [18]. Testing confirmed that *T. dactyloides* seedlings grown from seed collected as part of the same expedition were able to tolerate prolonged 4 °C cold stress while the same temperature stress treatment produced significant levels of cell death in maize seedlings from the reference genotype B73 (Figure S3B-F).

T. dactyloides-specific changes in the response of lipid metabolism to cold stress

RNA-seq and membrane lipid profiles were collected from maize, sorghum, and *T. dactyloides* seedlings under control and cold stressed conditions. The inclusion of sorghum provided a method to ascertain whether maize or *T. dactyloides* represented the ancestral state when metabolic or transcriptional patterns of responses to cold differed between the two species. As previously reported, no GO terms were significantly enriched among differentially regulated orthologs (DROs) of maize and sorghum, three hours after the onset of cold stress. At the same time point, DROs between *T. dactyloides* and maize + sorghum were enriched in genes related to photosynthesis light harvesting, protein-chromophore linkage, and chlorophyll metabolic genes. At 6 hr post-stress, genes related to chloroplast and mitochondrial metabolic processes were differentially regulated in *T. dactyloides* compared to maize and sorghum (Table S5). Specifically, chloroplast and mitochondrial RNA processing, modification and metabolic process genes were up-regulated in *T. dactyloides* at 6 hr post stress, while a suite of abiotic stress responsive genes, annotated as responding to osmotic stress, heat stress or salt stress, as well as genes involved in histone modification and methylation, chromatin organization and ethylene signaling were down-regulated (Table S6). These observations are all consistent with maize and sorghum both experiencing severe impairment of fundamental biological processes within six hours of the onset of cold stress, while *T. dactyloides* seedlings remained relatively healthy under equivalent levels of stress treatment.

The overall pattern changes in the abundances of the major membrane lipids described in (Figure S4A) between control and cold stress conditions were not significantly different across the three species although individual statistically significant changes in lipid abundance in response to cold were observed in each species. However, this assay also allowed the quantification of fatty acid desaturation for individual lipid types. Fatty acids are initially synthesized in a more saturated state. Hence a decrease in desaturation can be an indicator of increases in lipid synthesis. Increase in desaturation can serve as a signaling mechanism and also increases membrane fluidity allowing plants to avoid membrane damage at low temperatures (As Reviewed [19]). Statistically significant changes in lipid desaturation were observed for several lipid classes in individual species (Table S7). However, PC was unique in that the pattern of desaturation change in response to cold was consistent between maize and sorghum, and opposite in *T. dactyloides* (Figure 4b). Fast evolving genes in *T. dactyloides* lipid biosynthesis genes were concentrated in pathways leading to PC (Figure 4a).

Genes evolving rapidly in *T. dactyloides* also experienced selection in temperate adapted maize

Previous studies have identified a large set of maize genes which were targets of artificial selection during the process of adaptation to temperate climates [20]. We hypothesized that the more ancient process of the expansion of *T. dactyloides* into temperate climates may have targeted some of the same genes targeted by artificial selection during the introduction of maize into temperate climates. Tripsacum genes were divided into those where the maize ortholog was identified as likely under selection during the adaption of maize to temperate climates and those where the maize ortholog did not show evidence of being under selection during this process. As Ka/Ks ratios can vary widely across different genes as a result of factors including expression level, gene functional category, and location relative to centromeres [21] all *T. dactyloides* Ka/Ks values were normalized relative to sorghum, a closely related species that is still primarily adapted to tropical latitudes. Genes under selection during the development of temperate maize lines showed significant increases in Ka/Ks values in temperate adapted *T. dactyloides* relative to tropical adapted sorghum (log transformed t-test p-value = 0.027, Wilcoxon signed-rank test (WST) p-value = 0.018) (Figure 3B) [22]. This observation remained significant when using the median Ka/Ks value from orthologs in three different tropically adapted grass species (rice, oropetium and sorghum) (log transformed t-test p-value = 0.038, WST p-value = 0.029).

While the Hufford dataset consisted primarily of temperate elite lines and tropical landraces, it did also include a number of elite tropical lines and temperate landraces. A second dataset consisting of 47 high confidence tropical and subtropical maize lines and 46 high confidence temperate lines from maize HapMap3 [23], provided approximately equivalent results (Table S8). Normalized Ka/Ks values of *T. dactyloides* genes orthologous to the 10% of maize genes with the highest XP-CLR scores were not significant higher than the background (p-value = 0.24 in log transformed t-test, 0.19 in WST). However the normalized Ka/Ks value of top 5% and top 1% genes were significantly higher than background genes (top 5%: p-value = 0.032 in log transformed t-test, 0.018 in WST; top 1%: p-value = 0.028 in log transformed t-test, 0.008 in WST). The increase in significance increased at more stringent cut offs – even as the total number of data points decreases – may indicate that the overlap comes from only a subset of the genes under the strongest selection between tropical and temperate lines in maize and in the temperate adapted *T. dactyloides* relative to tropical-latitude-adapted related species.

Discussion

The potential for data from *Tripsacum* to aid in both basic biological research and applied plant breeding in maize has long been discussed [9, 10]. However, as of December 2017, a total of only 611 published nucleotide sequences existed for the entire *Tripsacum* genus, including 565 for *T. dactyloides*, 12 for *T. andersonii*, and 34 for all other named taxa within the genus. Here we have generated a set of 24,616 full length *T. dactyloides* cDNA sequences covering 22.4%, 31.5% and 60.2% of the annotated, expressed, and syntenically conserved gene space of maize respectively. This larger scale transcriptome resource enables a number of comparative analyses of *Zea* and *Tripsacum* not previously feasible.

Significant evolutionary rate heterogeneity exists among extant grass species. Previously, variation in the rate of divergence between homeologous gene pairs generated during the rho polyploidy [24] in different grass species was employed to detect variation in synonymous substitution rates [25]. However, this approach, which relies on pairwise comparisons between species, provides aggregate estimates for each lineage across the 70-96 million years since the rho WGD [24, 25]. Utilizing known phylogenetic relationships across relatively large numbers of grass species with sequenced genomes or significant genome resources and fitting rates of synonymous and nonsynonymous substitutions for each branch separately [26] demonstrated that even comparing sister genera (*Zea* and *Tripsacum*), maize exhibits significantly more rapid accumulation of synonymous substitutions. One potential explanation is differences in life cycle. Many *Zea* species are annuals [4] while more than 13 *Tripsacum* species are perennials [5] (Table 1) and several analyses have suggested that synonymous substitutions accumulate more rapidly in annual species [27, 28]. Another potential explanation is that the difference in the rate of molecular evolution between maize and *T. dactyloides* may reflect the difference in native range between these genera, as species native to tropical regions have been shown to accumulate nucleotide substitutions up to twice as rapidly as temperate species [29].

Genes are generally considered to show evidence of positive selection if the frequency of nonsynonymous substitutions is significantly higher than that of synonymous substitutions. However, if positive selection is assumed to be episodic rather than constant, elevated Ka/Ks ratios which are less than one can reflect a mixture of positive selection and purifying selection, relaxation of purifying selection, or statistical noise. Episodic positive selection is harder to detect on longer branches where the proportion of evolutionary time a gene spends under positive selection decreases relative to the time spent under purifying selection. Background ratios of Ka/Ks can also vary significantly between different genes reflecting differences in chromosome environment, expression level, function, and the presence or absence of different types of duplicate gene copies [21]. As shown in Figure 2C, the frequency of extreme Ka/Ks ratios decreases as the overall branch length increases. The inclusion of *T. dactyloides* breaks up the long branch between maize and sorghum, permitting the identification of genes experiencing either an interval of positive selection alongside ongoing purifying selection or a relaxation of purifying selection. It must be emphasized that the analysis presented here cannot distinguish between true positives – genes showing elevated rates of protein sequence evolution as a result of positive selection or relaxed selection – and false positives – statistical noise – on a single gene level. Instead the focus must be on the differences observed between the functional classes or pathways of genes which exhibited higher rates of protein sequence evolution in maize and *T. dactyloides*. Here we found that genes involved in phospholipid metabolism and stress response both tended to be experiencing higher rates of protein sequence evolution in *T. dactyloides* than

in maize or in the other grass species tested.

Recent studies of several instances of parallel selection have reported that it often acts on largely unlinked sets of genes at a molecular level [30, 31, 32]. This suggests that there are many different molecular mechanisms which can be employed to achieve the same phenotypic changes, and as a result the same genes will rarely be targeted in independent instances of selection for the same traits. However, the evidence presented above suggests that as the lineage leading to *T. dactyloides* expanded into temperate environments millions of years ago natural selection targeted some of the same genetic loci which would later be targets of artificial selection as the cultivation of maize spread from the center of domestication in Mexico into more temperate regions of North America. This overlap between targets of natural and artificial selection for adaptation to the same environment in sister genera also indicates that genetic changes in crop-wild relatives associated with adaptation to new environments may be useful guides for identifying genetic targets for breeding efforts aimed at adapting crops to a changing climate.

One specific difference between the native environment of wild *Zea* and *Tripsacum* species is that many *Tripsacum* species grow in areas where they are exposed to cold and freezing temperatures. Unlike maize, *T. dactyloides* can survive prolonged cold and freezing temperatures and successfully overwinter. The identification of accelerated protein sequence evolution among genes involved in phospholipid metabolism provides a plausible candidate mechanism for the increased cold and freezing tolerance of *T. dactyloides* relative to maize. While widely grown in temperate regions over the summer, maize remains sensitive to cold. Maize varieties with the ability to be planted significantly earlier, or in the extreme case to overwinter, have the potential to intercept a greater proportion of total annual solar radiation increasing maximum potential yields [33, 11]. This study illustrates how studying the genetic mechanisms responsible for crop-wild relative adaptation to particular climates may guide breeding and genome engineering efforts to adapt crops to a changing climate.

Methods

Plant materials and RNA preparation

T. dactyloides seeds were collected from wild growing plants located in Eastern Nebraska (USA, GPS coordinates: 41.057836, -96.639844). Seeds with brown cupules were selected. For each seed the cupule was removed, followed by a cold treatment at 4°Cfor at least two days, resulting in germination rates between 30% and 50%. A single plant (ID #-1) was selected for transcriptome sequencing. Young leaves were sampled one month after germination. Stem and root were sampled three months after germination. Harvested tissue samples were rinsed with cold distilled water and then immediately frozen in liquid N_2. Total RNA was extracted from each tissue separately by manually grinding each sample in liquid N_2, adding TriPure isolation reagent (Roche Life Science, catalog number #11667157001), followed by separating phase using chloroform, precipitating RNA using isopropanol and washing the RNA pellet using 75% ethanol. The air-dried RNA samples were dissolved in DEPC-treated water. RNA quantity and quality were assessed uisng a NanoDrop 1000 spectrophotometer and electrophoresis on a 1% agarose gel respectively.

Single-molecule sequencing and isoform detection

Equal quantities of total RNA from each sample were pooled prior to library construction and the combined sample was shipped to the Duke Center for Genomic and Computational Biology (GCB), Duke University, USA for sequencing. Three size-fractionated libraries (1-2 kb, 2-3 kb, and 3-6 kb) were constructed and sequenced separately on the PacBio RS II. Each libraries was sequenced using 2 SMRT cells. Raw reads data was analyzed through running the Iso-Seq pipeline included in the SMRT-Analysis software package (SMRT Pipe v2.3.0, https://github.com/PacificBiosciences/cDNA_primer) (for details see Supplemental Methods).

Substitution rate estimation and selection analyses

Codon based alignments were generated using ParaAT2.0 [34] for sets of orthologous genes identified using a dataset of syntenic orthologous genes identified across six grass species with sequenced genomes (maize V3 [35], sorghum v3.1 [36], setaria v2.2 [37], oropetium v1.0 [38], rice v7 [39] and brachypodium v3.1 [40]) [17], plus the maize/tripsacum orthologous relationships defined above. Synonymous nucleotide substitution rates (Ks) were calculated by using the codeml maximum-likelihood method (runmode = -2, CodonFreq = 2) implemented within PAML [26] and the known phylogenetic relationships of the seven species. The divergence time (T) between maize and *T. dactyloides* was estimated following the formula T = Ks/2μ[41] (for more details see Supplemental Methods).

Genome-wide scan for selection between tropical/temperate maize subpopulations

A cross-population composite likelihood approach XP-CLR [42] (updated by Hufford et al.[20] to incorporate missing data), based on the allele frequency differentiation between purely tropical/temperate maize subpopulations was used to identify genes likely to have been targets of selection during the process of adaption from tropical to temperate climates. 47 tropical and subtropical lines (TSS) and 46 temperate lines (NSS and SS) were selected from the Hapmap3 dataset [23], based on having a the probability value of membership into either of these groups greater than than 0.8 (Table S8) [43]. Based on the recombination rates measured using high density genetic maps in maize [44], XP-CLR was run with the following parameters: sliding window size of 0.05 cM; a fixed number of SNPs per window of 100; downweighting of SNPs in high LD (r2 $>$ 0.75); and a set of grid points as the putative selected allele positions were placed with a spacing of 1 kb across the whole genome. Each gene was assigned the maximum XP-CLR values found within the region 5 kb up- and downstream of the gene's annotation transcription start and transcription stop site.

Profiling of lipid and transcriptional responses to cold

Maize, sorghum, and *T. dactyloides* seedlings were grown under 13 hours/11 hours 29 °C /23 °C day/night and 60% relative humidity in a growth chamber at University of Nebraska-Lincoln's Beadle Center facility. At the three leaf stage, one half of the seedlings were moved to a second growth chamber maintained at a constant 6 °C temperature. The initiation of cold stress was timed to coincide with the end of daylight illumination. For RNA, *T. dactyloides* seedlings were harvested from both control and cold stressed conditions at 1, 3, 6, and 24 hours after the initiation of cold stress treatment. For lipid characterization, seedlings were harvested from control and cold stressed plants 24 hours after the initiation of cold stress. Raw

fastq sequences for maize and sorghum cold stress treatment are those published [45]. All data was realigned and reanalyzed using Trimmomatic [46], GSNAP [47] and HTSeq [48]. Differentially expressed genes (DEGs) and differentially regulated orthologs (DROs) were identified from read count data using DESeq2 [49], as described in [45].

Lipid composition determined essentially as described [50] (see Supplemental Methods for details).

Data availability

Raw PacBio sequence data has been deposited in the NCBI SRA under project PRJNA471728. Raw Illumina RNA-seq data has been deposited in the NCBI SRA under project PRJNA471735. A fasta file with the set of non-redundant *T. dactyloides* isoforms employed in this study is available at Zenodo with the identifier http://dx.doi.org/10.5281/zenodo.841005.

Acknowledgements

We thank Dr. Christy Gault (Cornell University) for sharing her protocol for germinating *T. dactyloides* seeds. This work was supported by the National Science Foundation under Grant No.OIA-1557417 to JCS, USDA NIFA award 2016-67013-24613 to RLR and JCS, a Science Foundation of Xichang University awarded to LY and a China Scholarship Council fellowship awarded to XL.

Author contributions

JCS, RLR, LY and XL conceived the project and designed the studies; OR identified and collected the plant material used in this study; LY, YZ, and SM performed the experiments; LY, XL, SKKR, and XD analyzed the data; LY, SKKR, and JCS wrote the paper. All authors reviewed and approved the final manuscript.

References

1. Swarts K, et al. (2017) Genomic estimation of complex traits reveals ancient maize adaptation to temperate north america. *Science* 357(6350):512–515.

2. Group GPW, et al. (2001) Phylogeny and subfamilial classification of the grasses (poaceae). *Annals of the Missouri Botanical Garden* pp. 373–457.

3. Iltis HH, Doebley JF (1980) Taxonomy of *Zea* (gramineae). ii. subspecific categories in the *Zea mays* complex and a generic synopsis. *American Journal of Botany* pp. 994–1004.

4. Doebley J (1990) Molecular evidence for gene flow among *Zea* species. *BioScience* 40(6):443–448.

5. Doebley J (1983) The taxonomy and evolution of tripsacum and teosinte, the closest relatives of maize. *Maize Virus Disease Colloquium and Workshop. The Ohio State University, Ohio Agricultural Research and Development Center, Wooster, Ohio.* pp. 15–28.

6. Forster B, Kole C (2011) Wild crop relatives: Genomic and breeding resources. cereals. *Experimental Agriculture* 47(4):736.

7. Edwards EJ, Smith SA (2010) Phylogenetic analyses reveal the shady history of c4 grasses. *Proceedings of the National Academy of Sciences* 107(6):2532–2537.

8. McKain MR, et al. (2018) Ancestry of the two subgenomes of maize. *bioRxiv*.

9. Chia JM, et al. (2012) Maize hapmap2 identifies extant variation from a genome in flux. *Nature genetics* 44(7):803–807.

10. Wang C, et al. (2013) Genomic resources for gene discovery, functional genome annotation, and evolutionary studies of maize and its close relatives. *Genetics* 195(3):723–737.

11. Gault C, Kremling K, Buckler ES (2018) Tripsacum de novo transcriptome assemblies reveal parallel gene evolution with maize after ancient polyploidy. *bioRxiv*.

12. Blakey C, Costich D, Sokolov V, Islam-Faridi MN (2007) Tripsacum genetics: from observations along a river to molecular genomics. *Maydica* 52(1):81.

13. Zhu Q, Cai Z, Tang Q, Jin W (2016) Repetitive sequence analysis and karyotyping reveal different genome evolution and speciation of diploid and tetraploid tripsacum dactyloides. *The Crop Journal* 4(4):247–255.

14. Lai J, et al. (2010) Genome-wide patterns of genetic variation among elite maize inbred lines. *Nature genetics* 42(11):1027–1030.

15. Hirsch CN, et al. (2014) Insights into the maize pan-genome and pan-transcriptome. *The Plant Cell* 26(1):121–135.

16. Hirsch C, et al. (2016) Draft assembly of elite inbred line ph207 provides insights into genomic and transcriptome diversity in maize. *The Plant Cell Online* pp. tpc–00353.

17. Schnable J, Zang Y, W.C. Ngu D (2016) Pan-grass syntenic gene set (sorghum referenced). *Figshare* p. https://dx.doi.org/10.6084/m9.figshare.3113488.v1.

18. Bates PD, Ohlrogge JB, Pollard M (2007) Incorporation of newly synthesized fatty acids into cytosolic glycerolipids in pea leaves occurs via acyl editing. *Journal of Biological Chemistry* 282(43):31206–31216.

19. Raju SK, Barnes A, Schnable JC, Roston RL (In Review) Low-temperature tolerance in land plants: Are transcript and membrane responses conserved?

20. Hufford MB, et al. (2012) Comparative population genomics of maize domestication and improvement. *Nature genetics* 44(7):808–811.

21. Yang L, Gaut BS (2011) Factors that contribute to variation in evolutionary rate among arabidopsis genes. *Molecular Biology and Evolution* 28(8):2359–2369.

22. He M, Liu P, Lawrence-Dill CJ (2017) A method to assess significant differences in rna expression among specific gene groups. *bioRxiv*.

23. Bukowski R, et al. (2017) Construction of the third generation zea mays haplotype map. *GigaScience*.

24. Paterson A, Bowers J, Chapman B (2004) Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proceedings of the National Academy of Sciences of the United States of America* 101(26):9903–9908.

25. Wang X, et al. (2015) Genome alignment spanning major poaceae lineages reveals heterogeneous evolutionary rates and alters inferred dates for key evolutionary events. *Molecular plant* 8(6):885–898.

26. Yang Z (1997) Paml: a program package for phylogenetic analysis by maximum likelihood. *Bioinformatics* 13(5):555–556.

27. Gaut B, Yang L, Takuno S, Eguiarte LE (2011) The patterns and causes of variation in plant nucleotide substitution rates. *Annual Review of Ecology, Evolution, and Systematics* 42:245–266.

28. Smith SA, Donoghue MJ (2008) Rates of molecular evolution are linked to life history in flowering plants. *science* 322(5898):86–89.

29. Wright S, Keeling J, Gillman L (2006) The road from santa rosalia: a faster tempo of evolution in tropical climates. *Proceedings of the National Academy of Sciences* 103(20):7718–7722.

30. Gaut BS (2015) Evolution is an experiment: Assessing parallelism in crop domestication and experimental evolution: (nei lecture, smbe 2014, puerto rico). *Molecular biology and evolution* 32(7):1661–1671.

31. Takuno S, et al. (2015) Independent molecular basis of convergent highland adaptation in maize. *Genetics* 200(4):1297–1312.

32. Lai X, Yan L, Lu Y, Schnable J (2017) Largely unlinked gene sets targeted by selection for domestication syndrome phenotypes in maize and sorghum. *bioRxiv* p. 184424.

33. Dohleman FG, Long SP (2009) More productive than maize in the midwest: how does miscanthus do it? *Plant physiology* 150(4):2104–2115.

34. Zhang Z, et al. (2012) Paraat: a parallel tool for constructing multiple protein-coding dna alignments. *Biochemical and biophysical research communications* 419(4):779–781.

35. Schnable PS, et al. (2009) The B73 maize genome: complexity, diversity, and dynamics. *science* 326(5956):1112–1115.

36. McCormick RF, et al. (2017) The sorghum bicolor reference genome: improved assembly and annotations, a transcriptome atlas, and signatures of genome organization. *bioRxiv* p. 110593.

37. Bennetzen JL, et al. (2012) Reference genome sequence of the model plant setaria. *Nature biotechnology* 30(6):555–561.

38. VanBuren R, et al. (2015) Single-molecule sequencing of the desiccation-tolerant grass *Oropetium thomaeum*. *Nature* 527(7579):508.

39. Ouyang S, et al. (2006) The tigr rice genome annotation resource: improvements and new features. *Nucleic acids research* 35(suppl_1):D883–D887.

40. Vogel JP, et al. (2010) Genome sequencing and analysis of the model grass brachypodium distachyon. *Nature* 463(7282):763–768.

41. Gaut BS, Morton BR, McCaig BC, Clegg MT (1996) Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene adh parallel rate differences at the plastid gene rbcl. *Proceedings of the National Academy of Sciences* 93(19):10274–10279.

42. Chen H, Patterson N, Reich D (2010) Population differentiation as a test for selective sweeps. *Genome research* 20(3):393–402.

43. Olukolu BA, et al. (2013) A connected set of genes associated with programmed cell death implicated in controlling the hypersensitive response in maize. *Genetics* 193(2):609–20.

44. Ott A, et al. (2017) Tunable genotyping-by-sequencing (tgbs®) enables reliable genotyping of heterozygous loci. *bioRxiv* p. doi.org/10.1101/100461.

45. Zhang Y, et al. (2017) Differentially regulated ortholog analysis demonstrates that early transcriptional responses to cold are more conserved in andropogoneae. *Biorxiv* p. doi: https://doi.org/10.1101/120303.

46. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics* 30(15):2114–2120.

47. Wu TD, Reeder J, Lawrence M, Becker G, Brauer MJ (2016) Gmap and gsnap for genomic sequence alignment: enhancements to speed, accuracy, and functionality. *Statistical Genomics: Methods and Protocols* pp. 283–334.

48. Anders S, Pyl PT, Huber W (2015) Htseq—a python framework to work with high-throughput sequencing data. *Bioinformatics* 31(2):166–169.

49. Love M, Anders S, Huber W (2014) Differential analysis of count data–the deseq2 package. *Genome Biol* 15:550.

50. Wang Z, Benning C (2011) Arabidopsis thaliana polar glycerolipid profiling by thin layer chromatography (tlc) coupled with gas-liquid chromatography (glc). *Journal of visualized experiments: JoVE* (49).

51. De Wet J, Timothy D, Hilu K, Fletcher G (1981) Systematics of south american tripsacum (gramineae). *American Journal of Botany* pp. 269–276.

52. Iltis HH, Doebley JF, Guzmán R, Pazy B (1979) Zea diploperennis (gramineae): a new teosinte from mexico. *Science* 203(4376):186–188.

53. Doebley JF, Iltis HH (1980) Taxonomy of *Zea* (gramineae). i. a subgeneric classification with key to taxa. *American Journal of Botany* pp. 982–993.

54. Eubanks MW (2001) The mysterious origin of maize. *Economic Botany* pp. 492–514.

55. Iltis HH, Benz BF (2000) Zea nicaraguensis (poaceae), a new teosinte from pacific coastal nicaragua. *Novon* pp. 382–390.

56. Wang B, et al. (2016) Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. *Nature communications* 7.

57. Akhunov ED, et al. (2013) Comparative analysis of syntenic genes in grass genomes reveals accelerated rates of gene structure and coding sequence evolution in polyploid wheat. *Plant physiology* 161(1):252–265.

58. Barbazuk WB, Fu Y, McGinnis KM (2008) Genome-wide analyses of alternative splicing in plants: opportunities and challenges. *Genome research* 18(9):1381–1392.

59. Marquez Y, Brown JW, Simpson C, Barta A, Kalyna M (2012) Transcriptome survey reveals increased complexity of the alternative splicing landscape in arabidopsis. *Genome research* 22(6):1184–1195.

60. Yang Z (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Molecular biology and evolution* 15(5):568–573.

61. Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences* 100(16):9440–9445.

62. Wang L, et al. (2013) Cpat: Coding-potential assessment tool using an alignment-free logistic regression model. *Nucleic acids research* 41(6):e74–e74.

63. Tello-Ruiz MK, et al. (2016) Gramene: a resource for comparative analysis of plants genomes and pathways. *Plant Bioinformatics: Methods and Protocols* pp. 141–163.

64. Haas BJ, et al. (2013) De novo transcript sequence reconstruction from rna-seq using the trinity platform for reference generation and analysis. *Nature protocols* 8(8):1494–1512.

65. Davidson RM, et al. (2011) Utility of rna sequencing for analysis of maize reproductive transcriptomes. *The Plant Genome* 4(3):191–203.

66. Foissac S, Sammeth M (2007) Astalavista: dynamic and flexible analysis of alternative splicing events in custom gene datasets. *Nucleic acids research* 35(suppl 2):W297–W299.

67. Mei W, et al. (2017) A comprehensive analysis of alternative splicing in paleopolyploid maize. *Frontiers in Plant Science* 8:694.

Figure legends

Figure 1. Morphological characteristics of inflorescences for representative species in Tripsacum and Zea. *Tripsacum dactyloides* (tripsacum, A-D) and *Zea mays* ssp. *parviglumis* (teosinte, E-H). (A) Early stage inflorescence. (B) Intermediate stage inflorescence with silks exerted from female spikelets. (C) magnified view of female spikelets. (D) mature inflorescence with both silks exerted from female spikelets (base) and anthers exerted from male spikelets (top). (E) Separate male and female inflorescences in teosinte. (F) magnified view of male inflorescence with anthers exerted. (G) magnified view of female inflorescence with silks exerted (H) hard fruitcases surrounding teosinte seeds (absent in domesticated maize). Bottom part present the latitude distribution of wild species in *Tripsacum* and *Zea* (outliers were removed), the order of which is consistent with that in Table 1.

Figure 2. Synonymous substitution rates and ratios of synonymous and nonsynonymous substitution rates across five related grass species.(A) Phylogenetic relationships of the seven species employed in this study. Red, blue and black boxes indicate target species, background species and outgroups respectively. (B) Distribution of synonymous substitution rates (Ks) in orthologous gene groups across each target and background species. Observed synonymous substitution rates in *T. dactyloides* are comparable to or less than those observed in maize. (C) Distribution of Ka/Ks ratios for orthologous genes conserved in maize, *T. dactyloides*, sorghum, foxtail millet (Setaria), and oropetium. (D) Subset of orthologous gene groups displayed in panel C, where the gene copy found in the in *T. dactyloides* transcriptome exhibits a higher Ka/Ks ratio than the same gene in the other four species tested. (E) Subset of orthologous gene groups displayed in panel C, where the gene copy found in the maize genome exhibits a higher Ka/Ks ratio than the same gene in the other four species tested.

Figure 3. Comparative distribution of Ka/Ks ratios between maize and *T. dactyloides*. (A) The relationship between Ka/Ks ratios observed in maize (blue) and those observed in *T. dactyloides* (green) for genes having the higher Ka/Ks ratio in one of these two taxa than in any of the three background taxa (sorghum, setaria, and oropetium). Orange triangles mark genes annotated as involved in glycerophospholipid metabolism, while red triangles mark genes involved in stress response. (B) Plot of the log transformed ratio of ratios between *T. dactyloides* and sorghum Ka/Ks values for two populations of genes. The artificial selection candidate set includes all orthologous gene groups conserved across the seven species employed in this analysis but where the maize gene copy was identified as a target of artificial selection between tropical and temperate maize inbreds. The background gene set includes all other orthologous gene groups conserved across these seven species. The artificial selection candidate gene set exhibits a skew towards higher ratios of Ka/Ks ratios relative to the background gene set.

Figure 4. A partial diagram of glycerophospholipid metabolism including phospholipid synthesis. a) Each name enclosed in a white box corresponds to an intermediate product in glycerophospholipid metabolism. Each name enclosed in a blue box indicates a mature lipid quantified in maize and *T. dactyloides*. If at least one gene encoding an enzyme which catalyzes a specific reaction was found to be evolving faster in *T. dactyloides* than any of the other grass species tested, the arrow indicating that reaction is drawn in red. If at least one gene encoding an enzyme which catalyzes a specific reaction was found to be evolving faster in maize than any of the other grass species tested, the arrow indicating that reaction is drawn in blue. If multiple genes encode different isoforms of the same enzyme and at least one was identified in the maize fast evolving gene set and a

separate gene was identified in the Tripsacum fast evolving gene set, the arrow is drawn in purple. In cases where none of the genes encoding an enzyme catalyzing a particular reaction were identified as fast evolving in either species, the arrow is drawn in gray. b) Estimation of phosphatidyl-choline (PC) desaturation in maize, sorghum, and T. dactyloides before and after one-day of cold treatment. '*' denotes t-test p-value 0.035.

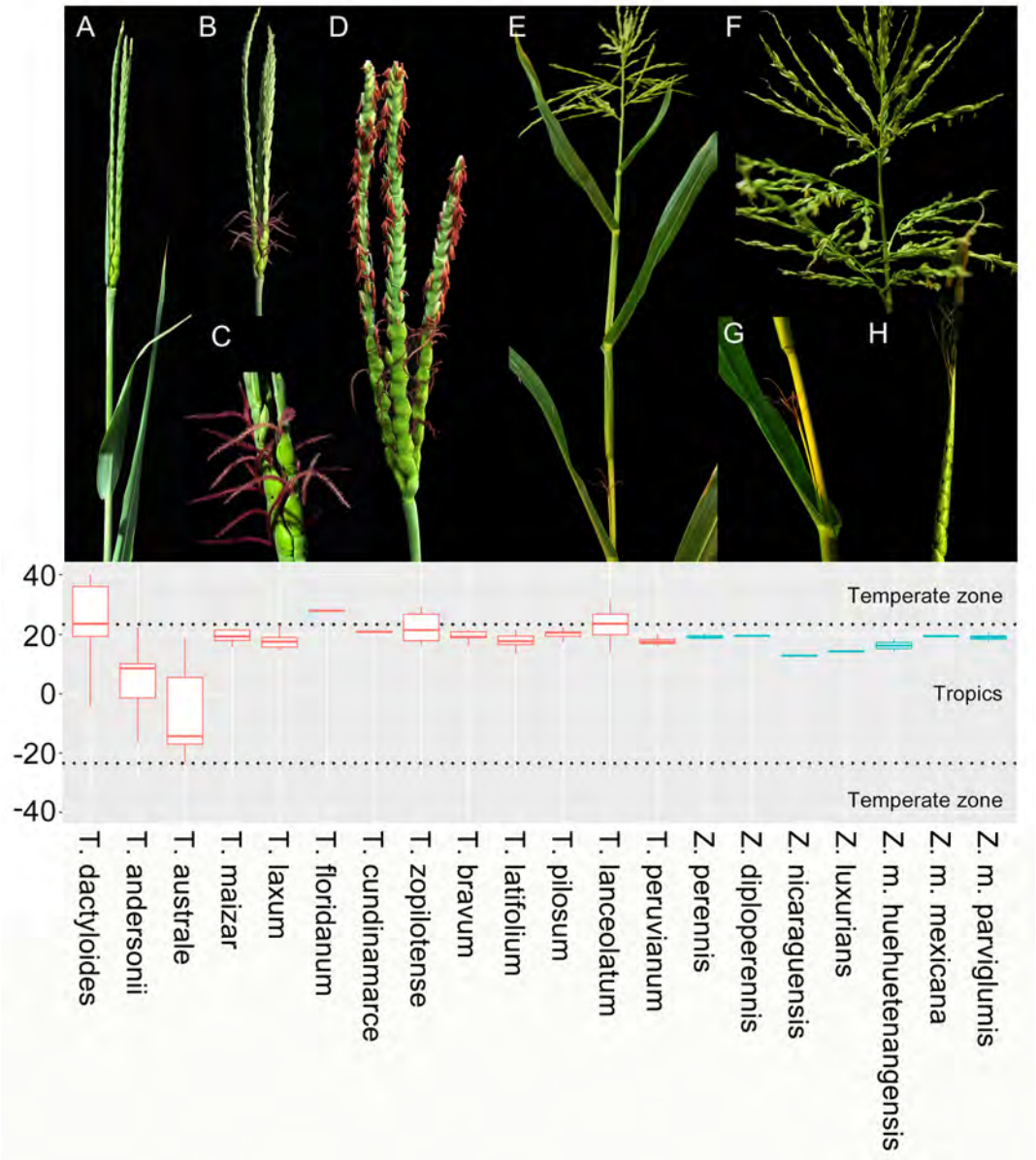


Figure 1. Morphological characteristics of inflorescences for representative species in Tripsacum and Zea. *T. dactyloides* (tripsacum, A-D) and *Zea mays* ssp. *parviglumis* (teosinte, E-H). (A) Early stage inflorescence. (B) Intermediate stage inflorescence with silks exerted from female spikelets. (C) magnified view of female spikelets. (D) mature inflorescence with both silks exerted from female spikelets (base) and anthers exerted from male spikelets (top). (E) Separate male and female inflorescences in teosinte. (F) magnified view of male inflorescence with anthers exerted. (G) magnified view of female inflorescence with silks exerted (H) hard fruitcases surrounding teosinte seeds (absent in domesticated maize). Latitudinal distribution of reported observations in GBIF for wild species within the genus *Tripsacum* (red) and *Zea* (teal). The boundaries of tropical latitudes (ie the tropic of cancer and tropic of capricorn) are marked with dashed black lines. Individual outlier datapoints more than 1.5x the interquartile range are omitted in this display.

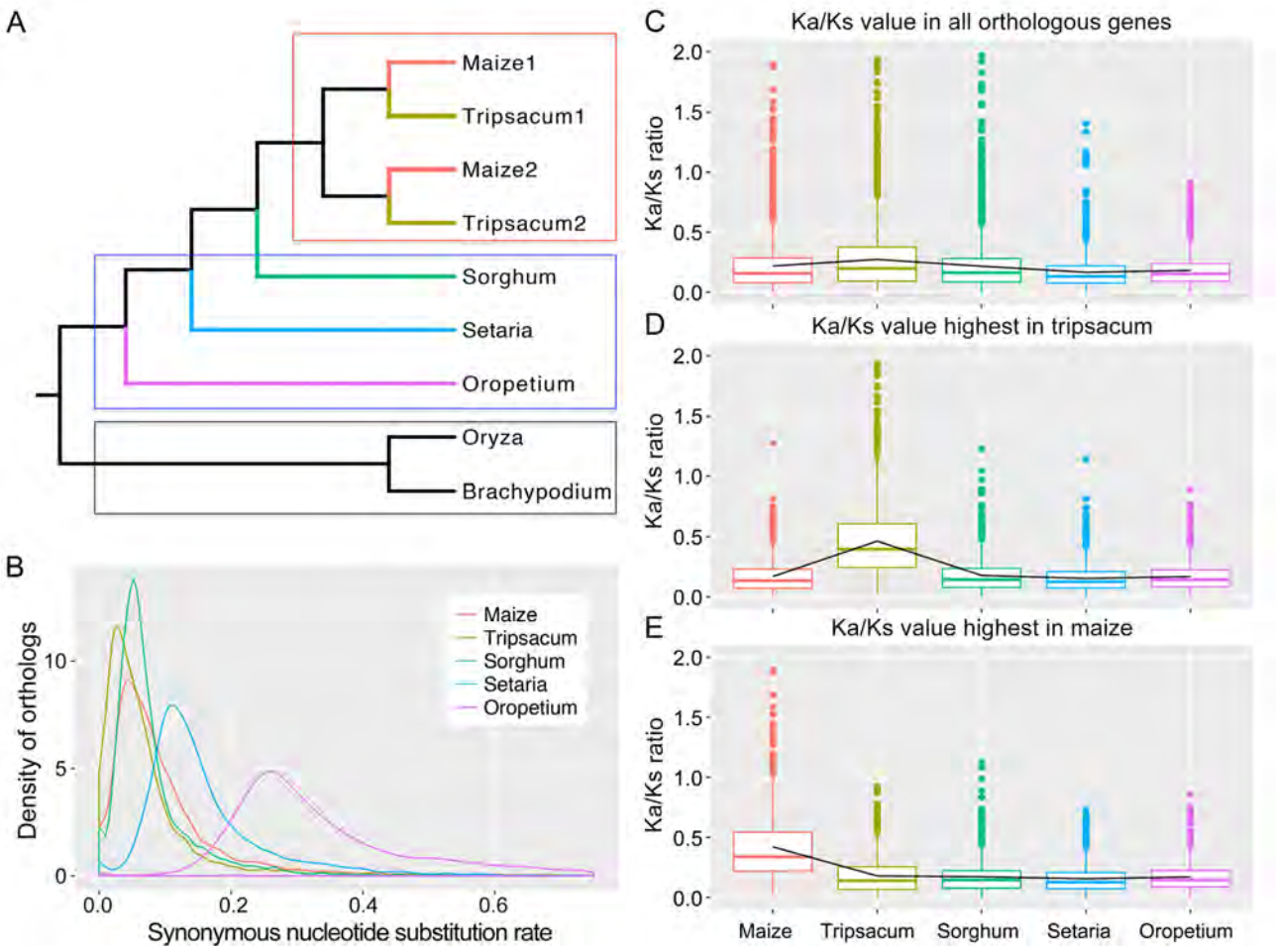


Figure 2. Synonymous substitution rates and ratios of synonymous and nonsynonymous substitution rates across five related grass species.(A) Phylogenetic relationships of the seven species employed in this study. Red, blue and black boxes indicate target species, background species and outgroups respectively. (B) Distribution of synonymous substitution rates (Ks) in orthologous gene groups across each target and background species. Observed synonymous substitution rates in *T. dactyloides* are comparable to or less than those observed in maize. (C) Distribution of Ka/Ks ratios for orthologous genes conserved in maize, *T. dactyloides*, sorghum, foxtail millet (Setaria), and oropetium. (D) Subset of orthologous gene groups displayed in panel C, where the gene copy found in the in *T. dactyloides* transcriptome exhibits a higher Ka/Ks ratio than the same gene in the other four species tested. (E) Subset of orthologous gene groups displayed in panel C, where the gene copy found in the maize genome exhibits a higher Ka/Ks ratio than the same gene in the other four species tested.

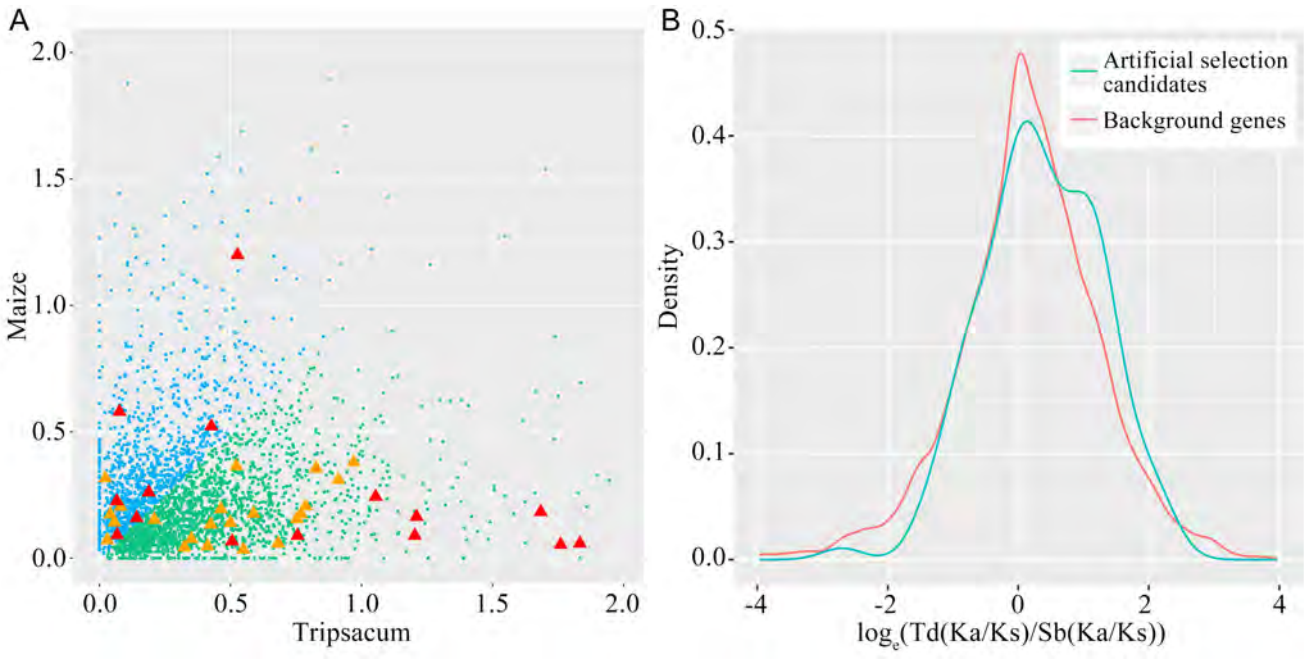


Figure 3. Comparative distribution of Ka/Ks ratios between maize and *T. dactyloides*. (A) The relationship between Ka/Ks ratios observed in maize (blue) and those observed in *T. dactyloides* (green) for genes having the higher Ka/Ks ratio in one of these two taxa than in any of the three background taxa (sorghum, setaria, and oropetium). Orange triangles mark genes annotated as involved in glycerophospholipid metabolism, while red triangles mark genes involved in stress response. (B) Plot of the log transformed ratio of ratios between *T. dactyloides* and sorghum Ka/Ks values for two populations of genes. The artificial selection candidate set includes all orthologous gene groups conserved across the seven species employed in this analysis but where the maize gene copy was identified as a target of artificial selection between tropical and temperate maize inbreds. The background gene set includes all other orthologous gene groups conserved across these seven species. The artificial selection candidate gene set exhibits a skew towards higher ratios of Ka/Ks ratios relative to the background gene set.

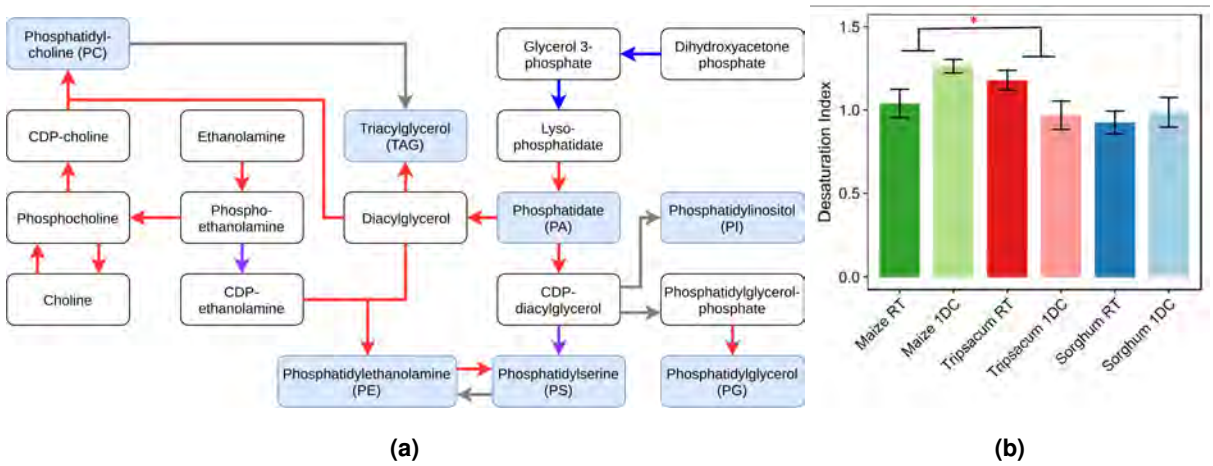


Figure 4. A partial diagram of glycerophospholipid metabolism including phospholipid synthesis. a) Each name enclosed in a white box corresponds to an intermediate product in glycerophospholipid metabolism. Each name enclosed in a blue box indicates a mature lipid quantified in maize and *T. dactyloides*. If at least one gene encoding an enzyme which catalyzes a specific reaction was found to be evolving faster in *T. dactyloides* than any of the other grass species tested, the arrow indicating that reaction is drawn in red. If at least one gene encoding an enzyme which catalyzes a specific reaction was found to be evolving faster in maize than any of the other grass species tested, the arrow indicating that reaction is drawn in blue. If multiple genes encode different isoforms of the same enzyme and at least one was identified in the maize fast evolving gene set and a separate gene was identified in the Tripsacum fast evolving gene set, the arrow is drawn in purple. In cases where none of the genes encoding an enzyme catalyzing a particular reaction were identified as fast evolving in either species, the arrow is drawn in gray. b) Estimation of phosphatidyl-choline (PC) desaturation in maize, sorghum, and *T. dactyloides* before and after one-day of cold treatment. '*' denotes t-test p-value 0.035.

Tables

Table 1. Taxonomic comparison between genus *Tripsacum* and *Zea*.

Table 1. Taxonomic comparison between genus *Tripsacum* and *Zea*

Tripsacum[1]			*Zea*[2]		
Species	Life cycle	Ploidy	Species	Life cycle	Ploidy
T. dactyloides L. (Gamagrass)	perennial	diploid, tetraploid	*Z. perennis*	perennial	tetraploid
T. andersonii (Guatemala grass)	perennial	diploid, tetraploid	*Z. diploperennis*	perennial	diploid
T. australe	perennial	diploid	*Z. nicaraguensis*	annual	diploid
T. maizar	perennial	diploid	*Z. luxurians*	annual	diploid
T. laxum	perennial	diploid	*Z. m. huehuetenangensis*	annual	diploid
T. floridanum	perennial	diploid	*Z. m. mexicana*	annual	diploid
T. cundinamarce	perennial	diploid	*Z. m. parviglumis*	annual	diploid
T. zopilotense	perennial	diploid			
T. bravum	perennial	diploid			
T. latifolium	perennial	tetraploid			
T. pilosum	perennial	tetraploid			
T. lanceolatum	perennial	tetraploid			
T. peruvianum	perennial	tetraploid, pentaploid, hexaploid			

[1]References for *Tripsacum* are [5, 51].

[2]References for *Zea* are [52, 53, 3, 54, 55].

Supplemental Information (SI)

Supplemental Results

A single *Tripsacum dactyloides* plant grown from seed collected from the wild in eastern Nebraska (USA) was used as the donor for all RNA samples. RNA extracted from three tissues (root, leaf, and stem) was used to construct three size fractionated libraries (1-2, 2-3, and 3-6 kb) which were sequenced using a PacBio RS II yielding a total of 532,071 Read Of Inserts (ROIs). The SMRT Pipe v2.3.0 classified more than half (267,186, 50.2%) of the ROIs as full-length and non-chimeric (FLNC) transcripts based on the presence of 5'-, 3'-cDNA primers and polyA tails. Each size-fractionated library had expected average length of FLNC transcripts of 1,364 bp, 2,272 bp, and 3,323 bp, with the average length of total FLNC transcripts of 2,354 bp, ranging from 300 to 29,209 bp Table S1). ICE and Quiver processing of FLNC transcripts produced a total of 64,362 high quality (HQ) consensus transcript sequences with an estimated consensus base call accuracy $\geq$ 99% (Figure S4).

Final consensus tripsacum sequences were mapped to the maize reference genome (RefGen_v3) using GMAP. Consistent with previously reported low overall rate of divergence in gene content and gene sequence between maize and tripsacum [9], 98.04% (63,103 out of 64,362 HQ consensus sequences) could be confidently mapped to the maize reference genome. Pbtranscript-TOFU was used to collapse the consensus sequences into 24,616 unique isoforms, in which differences in the 5' end of the first exon were considered redundant, and otherwise identical isoforms were merged (see Methods). This final set of unique tripsacum isoforms mapped to a total of 13,089 maize genes, including 7,633 maize genes represented by a single tripsacum transcript and 5,456 maize genes represented by two or more transcripts. Among maize genes to which two or more tripsacum isoforms were mapped the average was 3.1 isoforms per gene. Eighty-four maize genes were represented by more than 10 or more tripsacum isoforms, and the single maize gene represented by the most isoforms was GRMZM2G306345, which encodes a pyruvate, phosphate dikinase (PPDK) protein involved in the fixation of carbon dioxide as part of the C4 photosynthetic pathway, with 83 identified tripsacum isoforms (Table S2). A set of 249 confident lncRNAs were identified among the final tripsacum consensus sequences (See Methods) (Figure S6G). With an average length of 1.45 kb (ranging from 0.51 to 3.5 kb), the distribution of lncRNA sequences is notably larger than the average length of the maize lncRNAs identified using pacbio isoseq 0.67 kb (ranging from 0.2–6.6 kb) (Figure S5) [56]. Only 17 of these 249 lncRNAs exhibited high sequence similarity (identity $>$ 80%) with lncRNA sequences identified in maize.

In total, 12,826 out of 13,089 tripsacum transcripts mapped to 14,401 annotated maize gene models (Figure S6A,B). In some cases a single consensus tripsacum sequence spanned two or more maize gene models (Figure S6). Maize genes were sorted based on their expression levels using an existing short read RNA-seq dataset from maize seedling tissue [45]. Of the 13,089 most highly expressed maize genes, 8,191 (62.6%) were aligned with at least one tripsacum transcript indicating that the expression level of a gene in maize is a relatively good predictor of how likely that gene was to be captured in the tripsacum isoseq data. Because genes located in the chromosome arms of maize tend to exhibit higher levels of expression than genes in pericentromeric regions, this sample of tripsacum genes is likely depleted in the types of genes which are over represented in pericentromic regions. Figure S6C and D illustrate the comparative densities of highly expressed maize genes aligned to tripsacum isoseq transcripts across the 10 chromosomes of maize. Tripsacum transcript density was correlated with the density of

maize highly expressed genes (Spearman correlation coefficient r = 0.855, $p < 2.2e-16$). Among the 12,826 tripsacum transcripts mapped to maize gene models, 11,910 were unique one-to-one mappings. Data on conserved maize-sorghum orthologous gene pairs was used to increase the confidence of these mappings (Figure S6E), resulting in a final set of 9,112 putative sorghum-maize-tripsacum orthologous gene groups (available on https://figshare.com/s/6d55867b09e014eb7aed, Figure S6F). These gene pairs were grouped into three categories "one-to-one", single tripsacum/maize orthologs without duplication (5,641 gene pairs); "one-to-two", a single tripsacum gene mapped to one copy of a homeologous maize gene pair (1,964 gene triplet); "two-to-two", unique tripsacum sequences mapped to each copy of a homeologous maize gene pair (1,507 gene quartets, Figure S6H, Figure S7).

A total of 223 transcripts aligned to the maize genome in locations where no maize gene model was present. After additional validation and QC (see Methods), 94 of these cases appeared to be unannotated maize genes which were supported by both an aligned tripsacum transcript and short read RNA-seq data in maize, and another 102 cases appeared to be genomic sequences conserved between maize and tripsacum which were transcribed in tripsacum but lacked expression evidence in maize. More than two thirds of the 94 potentially unannotated maize genes could be confidently aligned to annotated coding sequences in the reference genomes of sorghum (63%) or setaria (66%). In contrast, the sequences present in both the tripsacum and maize genomes but expressed only in tripsacum were much less likely to be present in outgroup species (sorghum (25%) and setaria (30%)), suggesting the transcription of these genomic sequences may be a comparatively recent feature, potentially unique to the tripsacum lineage.

Consistent with observations from other plant systems [57, 58, 59], the most common single AS variants in tripsacum were – in descending order of frequency – intron retention (IntronR), exon skipping (ExonS), alternative donor (AltD), alternative acceptor (AltA), and mutually exclusive exons (MXEs). The remaining isoforms incorporated two or more types of changes and were classified as Complex AS (CompAS) (Figure S10). While a comparable maize dataset generated using the same long read technology contained more total splicing events, likely as a result of approx 4.5x greater sequencing depth [56], the overall proportions of AS events belonging to different categories in maize and tripsacum were more similar to each other than either was to sorghum (Figure S10). Shifting from overall frequency to the conservation of specific splicing events – defined as identical AS codes at the same physical positions when tripsacum and maize full length transcripts are aligned to the maize reference genome – a total of 4,324 (35.3%) tripsacum AS events associated with 1,065 genes were also identified in maize (Figure S10) and more than two third (656, 61.6%) of the conserved AS genes were observed in orthologous gene groups while 409 genes were *Zea-Tripsacum* linage-specific in *Tripsacinae*.

Of the 1,542 gene quartets where a single gene in sorghum is co-orthologous to two maize genes and each maize gene is orthologous to a single tripsacum gene, 212 genes exhibited alternative splicing for both tripsacum genes and 409 genes exhibited alternative splicing for both maize genes. In 52.06% of of gene pairs in tripsacum the pattern of splicing was conserved between homeologous gene pairs as well as in 77.8% of maize homeologous gene pairs. Changes in patterns of splicing which could be more parsimoniously explained by differences in *trans*- regulation of AS were observed substantially substantially for frequently than were changes in changes in patterns of splicing which could be more parsimoniously explained by differences in *cis*- regulation (Figure S11).

Supplemental Methods

Creation of high quality consensus transcript sequences

First, reads of insert consensus sequence (ROIs, previously known as circular consensus sequences, CSCs) were identified using the Reads Of Insert protocol, one of the submodules of the Iso-Seq pipeline. The minimum required number of full passes was set to 0 and predicted consensus accuracy was set to 75%. Raw sequences that passed this step were considered to be the ROIs.

Second, the classify submodule was used to automatically determine which ROIs were full-length. ROIs were considered to be full length if both the 5'- and 3'-cDNA primers as well as a poly(A) tail signal preceding the 3'-primer were detected. The classify submodule also separated chimeric reads from non-chimeric reads through detecting the presence of SMRTbell adapters in the middle of sequences, producing a set of full-length non-chimeric (FLNC) reads.

Third, the isoform-level clustering algorithm ICE (Iterative Clustering for Error Correction), which uses only full length reads, followed by Quiver which polished FLNC consensus sequences from ICE using non-FL reads were used to improve the accuracy of FLNC consensus sequences. These steps resulted in a set of high quality polished consensus with $> 99\%$ post-correction accuracy.

Fourth, HQ polished consensus reads were mapped to the maize reference genome (B73_RefGen_v3) using GMAP and redundant mapped transcripts were merged using using the collapse_isoforms_by_sam.py script from the pbtranscript-ToFU package (https://github.com/PacificBiosciences/cDNA_primer/wiki/tofu-Tutorial-(optional).-Removing-redundant-transcripts/) with the parameter settings of min-coverage = 85% and min-identity =82%. Isoforms differing only at the 5'-sites within the first exon were considered to be redundant and collapsed.

Ka/Ks Analysis Details

Branch-specific Ka/Ks ratios with self-built phylogenetic trees were calculated using codeml program (runmode = 0). The resulting codeml data including Ka, Ks and Ka/Ks for the genes of each branch were obtained and genes with Ka > 0.5, Ks $>$ 2 and Ka/Ks > 2 were discarded. Two models in codeml were used, one required a constant value for ω across all branches (model 0), and the other allowed heterogenous rates on each branch of the phylogeny (model 1). A likelihood ratio test was used to compare likelihood values under model 0 and model 1 to test whether significant variation in Ka/Ks ratios between different branches was present [60]. For each set of orthologous genes the log likelihood values under two models (lnL1 for the alternative and lnL0 for the null model) were obtained. From these values, the LRT was computed using $2 \times$ (lnL1-lnL0). A χ^2 curve with degree of freedom = 1 was used to calculate a p-value for this LRT. Multiple testing was performed to correct these p-values by applying the false discovery rate method (FDR) adjusted in R [61]. A gene was considered to be experiencing accelerated rates of protein evolution in a given lineage if the highest Ka/Ks ratio was detected in the branch leading to that species and the FDR-adjusted p-value for the comparison of the constant ω and heterogeneous models was < 0.05.

Use of statistical tests in comparisons of Ka/Ks ratios between different populations of genes

Like gene expression fold-change data, Ka/Ks ratios and ratios of Ka/Ks ratios exhibit non-normal distributions. Applying a log transformation to either Ka/Ks ratios or ratios of ratios produces roughly normal distributions, as observed for gene

expression fold change data [22]. This method was employed to conduct two comparisons using the independent-samples t-test package within R. In addition, a non-parametric test in R, Wilcoxon signed-ranked tests, were also used test whether Ka/Ks ratios differ significantly between gene groups. In the first test, the distribution of ratio of ratio values between *T. dactyloides* and sorghum Ka/Ks values was compared between genes identified as likely to be under selection in a comparison of tropical and temperate maize lines [20, 32] and genes not identified as likely to be under selection in the same comparison. The second test compared Ka/Ks ratios for genes annotated as involved in lipid metabolism to the population of genes not involves in phospholipid metabolism in *T. dactyloides* and maize, respectively, as well as Ka/Ks ratios between *T. dactyloides* and maize.

Lipid Profiling Details

Briefly, lipid extraction was performed by immediately immersing fresh tissue samples in ice-cold 2:1:0.1 (v/v/v) methanol: chloroform: formic acid and bead beating them for 2 minutes until the tissue was completely disrupted, then phase-partitioning by addition of 0.2M phosphoric acid, 1 M potassium chloride. The organic phase was removed (and stored at -20 degrees Celsius until) derivatization. 2D-thin Layer Chromatography (TLC) was used to separate lipids by their headgroup properties using the following solvent systems, 130:50:10 (v/v/v) Chloroform: Methanol: Ammonia hydroxide in the first dimension and 85:12.5:12.5:4 (v/v/v/v) Chloroform: Methanol: acetic acid: water in the second dimension. Lipids were identified by with respect to standards purchased from Avanti Polar Lipids. Quantitative data was obtained by exposure to iodine vapor and compared with a standard mixture. Each target lipid was scraped from the TLC plate and derivatized to fatty acyl methylesters (FAME Reaction) and the resulting profile of FAME was determined by gas chromatography (GC).

Identification of LncRNAs

The protein coding-potential of individual *T. dactyloides* transcripts was assessed using CPAT (Coding Potential Assessment Tool) [62], which employs a logistic regression model built with four sequence features as predictor variables: open reading frame (ORF) size, ORF coverage, Fickett testcode statistic, and hexamer usage bias. CPAT was trained using 4,900 high-confidence lncRNAs transcripts identified as part of Gramene 52 release (`http://www.gramene.org/release-notes-52/`) [63] and an equal number of known protein-coding transcripts randomly subsampled from the RefGen_v3 annotation 6a to measure the prediction performance of this logic model in maize. The accuracy of the trained CPAT model was assessed by quantifying six parameters using 10-fold cross validation with the maize training dataset: sensitivity (TPR), specificity (1-FPR), accuracy (ACC) and precision (PPV) under the receiver operating characteristic (ROC) curve and precision–recall (PR) curve. Based on there parameters, a probability threshold of 0.425 was identified as providing the best trade off between specificity and sensitivity for identifying lncRNA sequences.

4,095 candidate non-coding RNAs with a length greater than $>$ 200 bp were predicted in CPAT. ORF prediction for non-coding candidate of *T. dactyloides* transcripts was performed using TransDecoder [64] and 2,509 transcripts encoding ORFs longer than 100 amino acids were removed from the set of putative lncRNAs. The remaining lncRNAs were aligned to the NCBI-nr database using BLASTX and sequences showing similarity to existing protein sequencing in the database (e-value $\leqslant$ 1e-10) were removed from the set of putative lncRNAs.

Identification of orthologs

Maize-tripsacum orthologs were defined based on the following criteria. Firstly, the tripsacum sequence must have been uniquely mapped to the target maize gene using GMAP. Secondly, the tripsacum gene must have been uniquely mapped to a single sorghum gene using BLASTN. Thirdly, the maize and sorghum genes must be syntenic orthologs of each other using a previously published dataset [17, 45]. As a result of the maize WGD shared by maize and *T. dactyloides*, in some cases multiple *T. dactyloides* sequences mapped to unique maize genes on different maize subgenomes, but to the single shared co-ortholog of these two maize genes in sorghum.

BLAST analyses for Isoseq data

Isoforms which failed to map to the maize genome aligned to different NCBI refSeq databases using BLASTN with the following parameters: min-coverage = 50%, min-identity = 70%, max_target_seqs = 5 and e-value $\leqslant$ 1e-10. Transcripts were considered to originate from ancestral grass genes not present in the B73 reference genome if none of the top 5 blast hits were to maize sequences, but did include sequences isolated from other grass species.

T. dactyloides sequences which aligned to regions of the maize genome not annotated as genes were also aligned to the full set of annotated maize gene CDS sequences using BLASTN with parameters: min-coverage=80%, min-identity=85%, max_target_seqs=1 and e-value $\leqslant$ 1e-10. Transcripts which did not align to any maize gene models even with these relaxed criteria were then manually proofed for expression evidence in maize using IGV and a set of Illumina short read RNA-seq data from a wide range of maize tissues [65].

Detection of alternative splicing (AS) events

AS analysis was conducted using Astalavista-4.0 (Alternative Splicing Trascriptional Landscape Visualization Too) [66], transforming identified place isoforms into a set of defined AS codes based on their location within the gene and the type of splicing observed. Among the multiple maize genes with multiple aligned *T. dactyloides* isoforms, in 3,566 cases one or more AS events (12,260 in total) was identified between the multiple aligned isoforms, while in the 1,890 remaining cases differences between isoforms were the result of either variation in transcription start sites (VSTs), or additional 3' exons (DSPs). Each AS site is assigned a number according to its relative position in the event and a symbol depending on its type. In addition to the main five single AS types such as Intron retention (IR), Exon skipping (ES), Alternative donor (AD), alternative acceptor (AA) and Mutually exclusive exons (MXEs), other AS events in which multiple types of AS are present between two isoforms were counted as complicated AS. The maize data used to compare AS events between maize and *T. dactyloides* was extracted from the published AS dataset generated by Wang et al [56].

Table S1. Summary of sequence data produced.

Table S2. Number of maize genes with one or more identified tripsacum isoforms.

Table S3. Patial list of conserved grass genes present in tripsacum but not in maize reference genome.

Table S4. Genes experiencing accelerated selections in *T. dactyloides* related to stress response.

Table S5. GO terms associated with Maize-Tripsacum and Sorghum Tripsacum DROs after 3, 6, 12 and 24 hours post cold stress

Table S6. GO terms associated with up-regulated and down-regulated Maize-Tripsacum and Sorghum Tripsacum DROs after 6 hours post cold stress

Table S7. ANOVA table for lipid desaturation levels in maize, sorghum and *T. dactyloides* at room temperature and 24 hours after cold stress.

Table S8. List of purely temperate and tropical maize lines from HapMap3.

Figure S1. Ks distribution of orthologous gene pairs between each two of species pair-wisely (bin size = 0.05). (A) divergence between tripsacum and maize (td-zm), maize and sorghum (zm-sb), sorghum and setaria (sb-si), setaria and oropetium (si-or), their divergence were shown by the peak of each pair. (B) divergence of maize with other species. (C) divergence of tripsacum with other species. (D) divergence of sorghum with other species. (E) divergence of setaria with other species. (F) divergence of oropetium with other species.

Figure S2. Distribution of species-specific Ka/Ks ratios including the homeologous gene quartets in both tripsacum and maize shared the WGD. (A) distribution of Ka/Ks ratios in orthologous genes sets. (B) increased Ka/Ks ratios in maize1. (C) increased Ka/Ks ratios in maize2. (D) increased Ka/Ks ratios in tripsacum1. (E) increased Ka/Ks ratios in tripsacum2.

Figure S3. (A) Ka/Ks distributions in tripsacum between lipid genes and other functional genes. (B-E) Phenotype changes between tripsacum (left) and maize (right) in continuous three days after cold treatment at 4 °C for five days.

Figure S4. Changes in abundance (top) and desaturation (bottom) of specific lipid types in seedlings of maize, T. dactyloides, and sorghum under control conditions (RT) and after exposure to 24 hours of cold stress (1DC).

Figure S5. Workflow of Iso-Seq bioinformaics analysis for tripsacum using the SMRT-Analysis software package (SMRT Pipe v2.3.0).

Figure S6. Tripsacum Iso-seq data mapped onto the maize reference genome (RefGen_v3). (A) 10 chromosomes of maize. (B) Maize gene density in each chromosome. (C) Density of highly expressed maize genes (FPKM $>$ 4.15). (D) Tripsacum transcript density in each chromosome. (E) Density of syntenic gene pairs between maize and sorghum. (F) Density of sorghum-maize-tripsacum gene pairs. (G) Distribution of Long non-coding RNA (lncRNA) in tripsacum. (H) Distributions of homeologous genes pairs following the whole genome duplication which is shared by both tripsacum and maize.

Figure S7. Comparison of length distribution of lncRNAs identified using Pacbio sequencing data between maize and tripsacum.

Figure S8. Example of one tripsacum transcript spanning two maize gene models but correlated with one sorghum gene model through GEvo analysis.

Figure S9. Distribution of maize genes with either 1:1, 2:1 or 2:2 relationships to assembled tripsacum transcripts across the maize genome.

Figure S10. Proportions of alternative splicing (AS) types (Intron retention, IntronR; Exon skipping, ExonS; Alternative donor, AltD; Alternative acceptor, AltA; Mutually exclusive exons, MXEs; Complicated AS, CompAS; More complicated AS, MoreCompAS) found in (A) tripsacum and (B) maize using PacBio long sequences. (C) Proportion of conserved AS types between maize and tripsacum.

Figure S11. Comparison of alternative splicing (AS) distribution in (A) different species using long- and short- reads (long reads data for maize taken from [56] and short reads data of maize and sorghum were from [67]) and (B) subgenomes of maize and tripsacum. Identical shapes indicate genes with a conserved AS event. Solid line boxes mark cases most parsimoniously explained by a change in *trans*-regulation of AS, while dashed line boxes mark cases most parsimoniously explained by a change in *cis*-regulation of AS.

Supplemental Figures

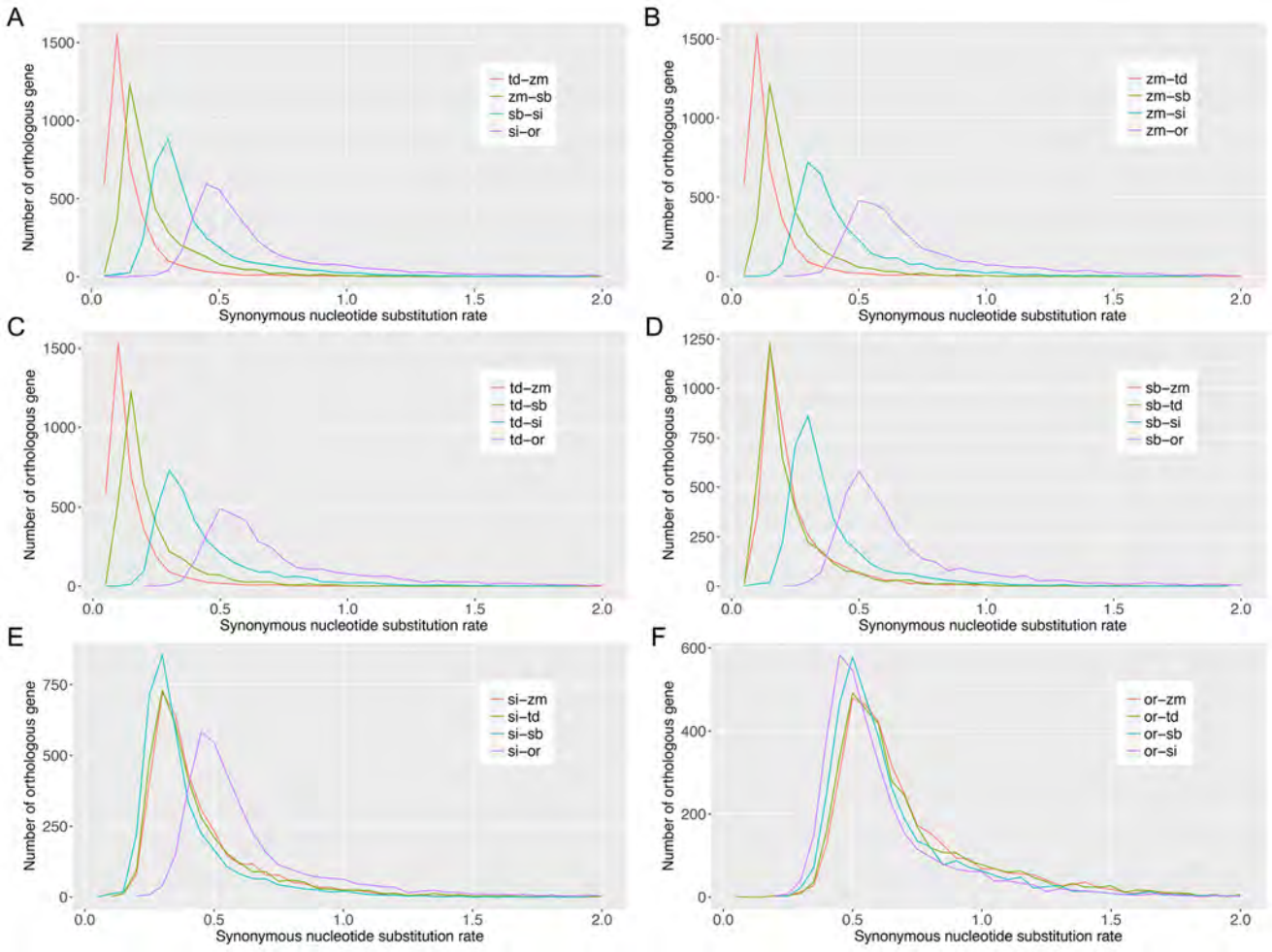


Figure S1. Ks distribution of orthologous gene pairs between each two of species pair-wisely (bin size = 0.05). (A) divergence between tripsacum and maize (td-zm), maize and sorghum (zm-sb), sorghum and setaria (sb-si), setaria and oropetium (si-or), their divergence were shown by the peak of each pair. (B) divergence of maize with other species. (C) divergence of tripsacum with other species. (D) divergence of sorghum with other species. (E) divergence of setaria with other species. (F) divergence of oropetium with other species.

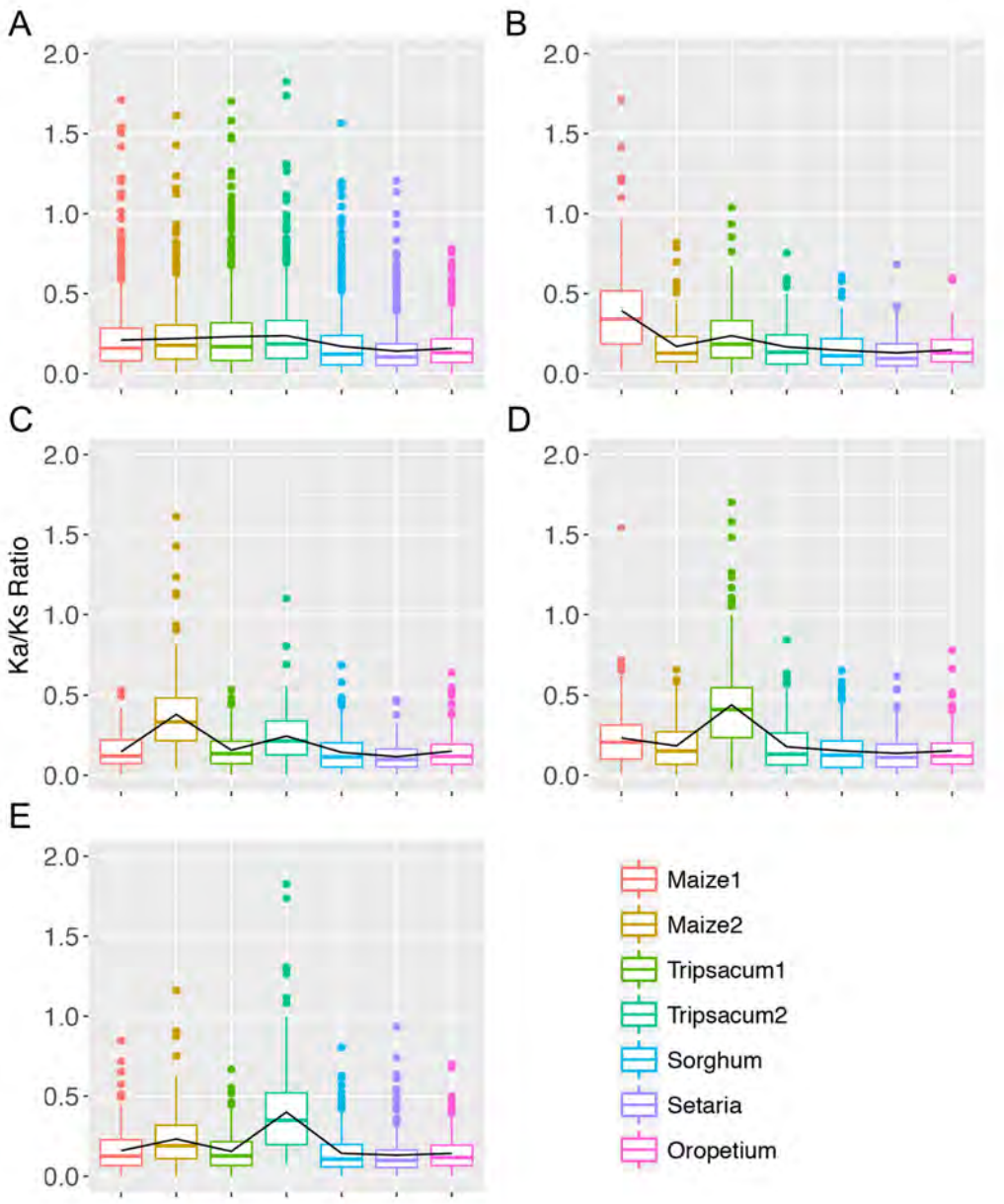


Figure S2. Distribution of species-specific Ka/Ks ratios including the homeologous gene quartets in both tripsacum and maize shared the WGD. (A) distribution of Ka/Ks ratios in orthologous genes sets. (B) increased Ka/Ks ratios in maize1. (C) increased Ka/Ks ratios in maize2. (D) increased Ka/Ks ratios in tripsacum1. (E) increased Ka/Ks ratios in tripsacum2.

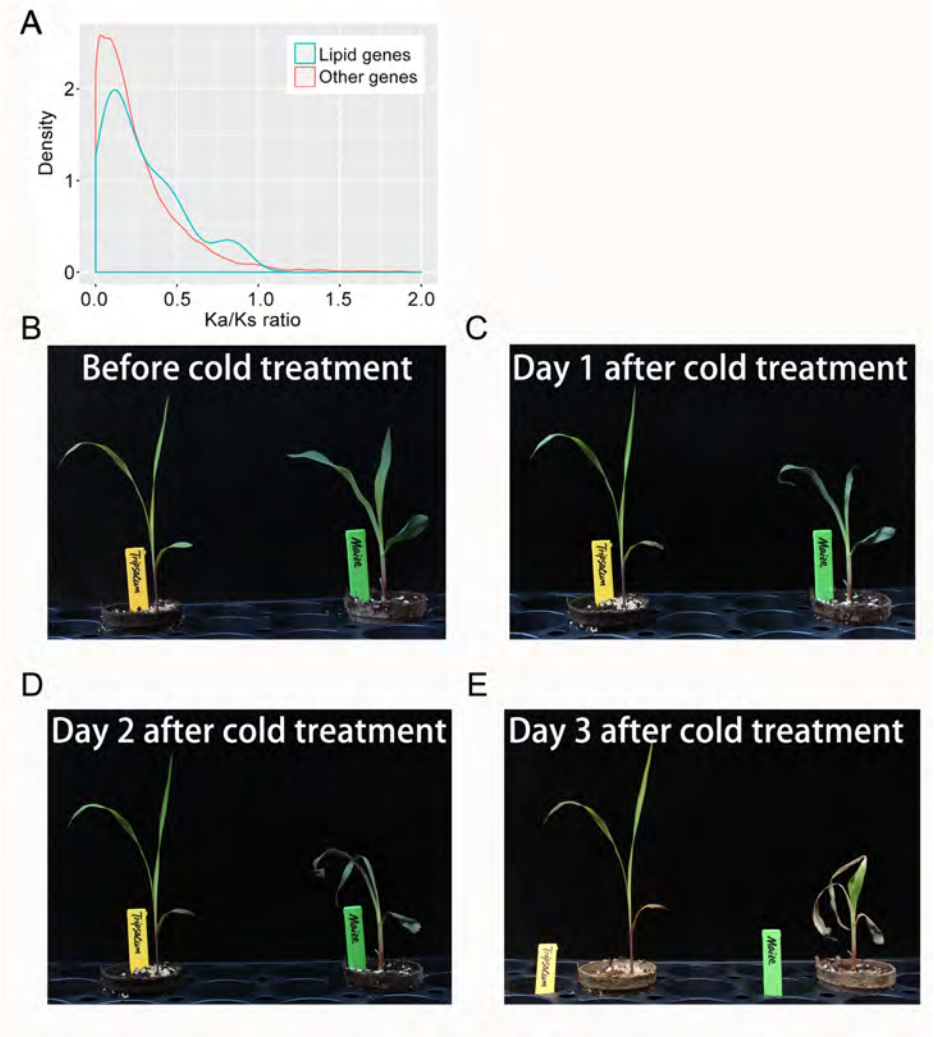


Figure S3. (A) Ka/Ks distributions in tripsacum between lipid genes and other functional genes. (B-E) Phenotype changes between tripsacum (left) and maize (right) in continuous three days after cold treatment at 4 °C for five days.

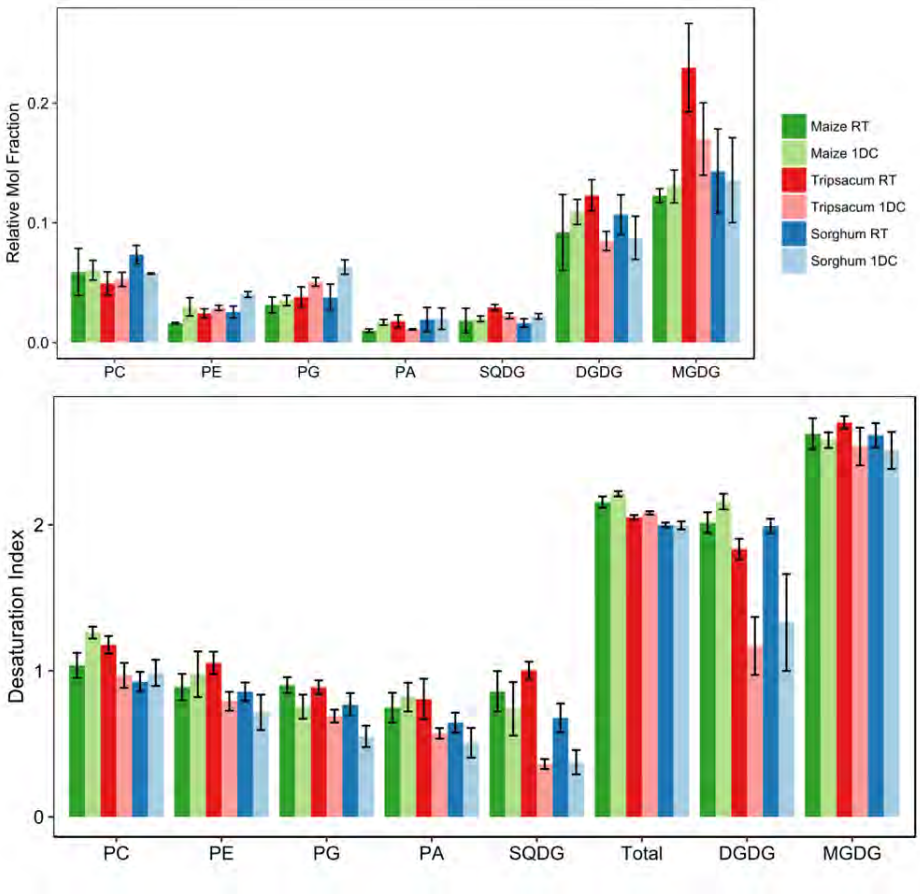


Figure S4. Changes in abundance (top) and desaturation (bottom) of specific lipid types in seedlings of maize, *T. dactyloides*, and sorghum under control conditions (RT) and after exposure to 24 hours of cold stress (1DC).

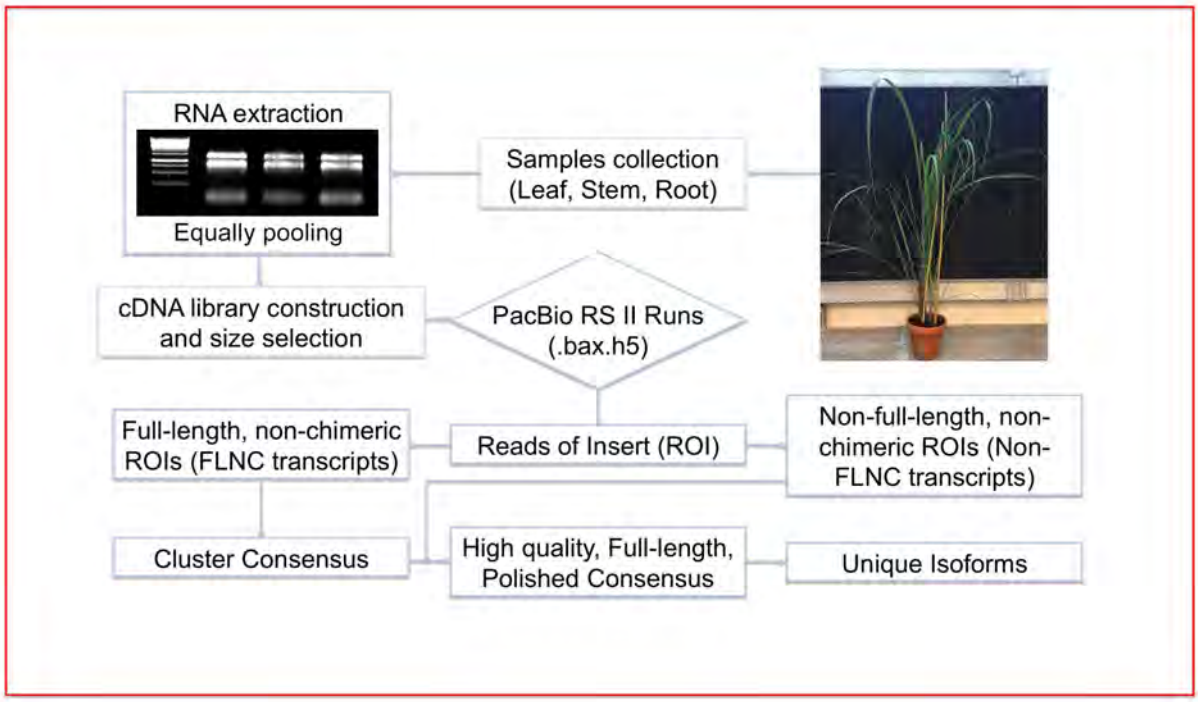


Figure S5. Workflow of Iso-Seq bioinformaics analysis for tripsacum using the SMRT-Analysis software package (SMRT Pipe v2.3.0).

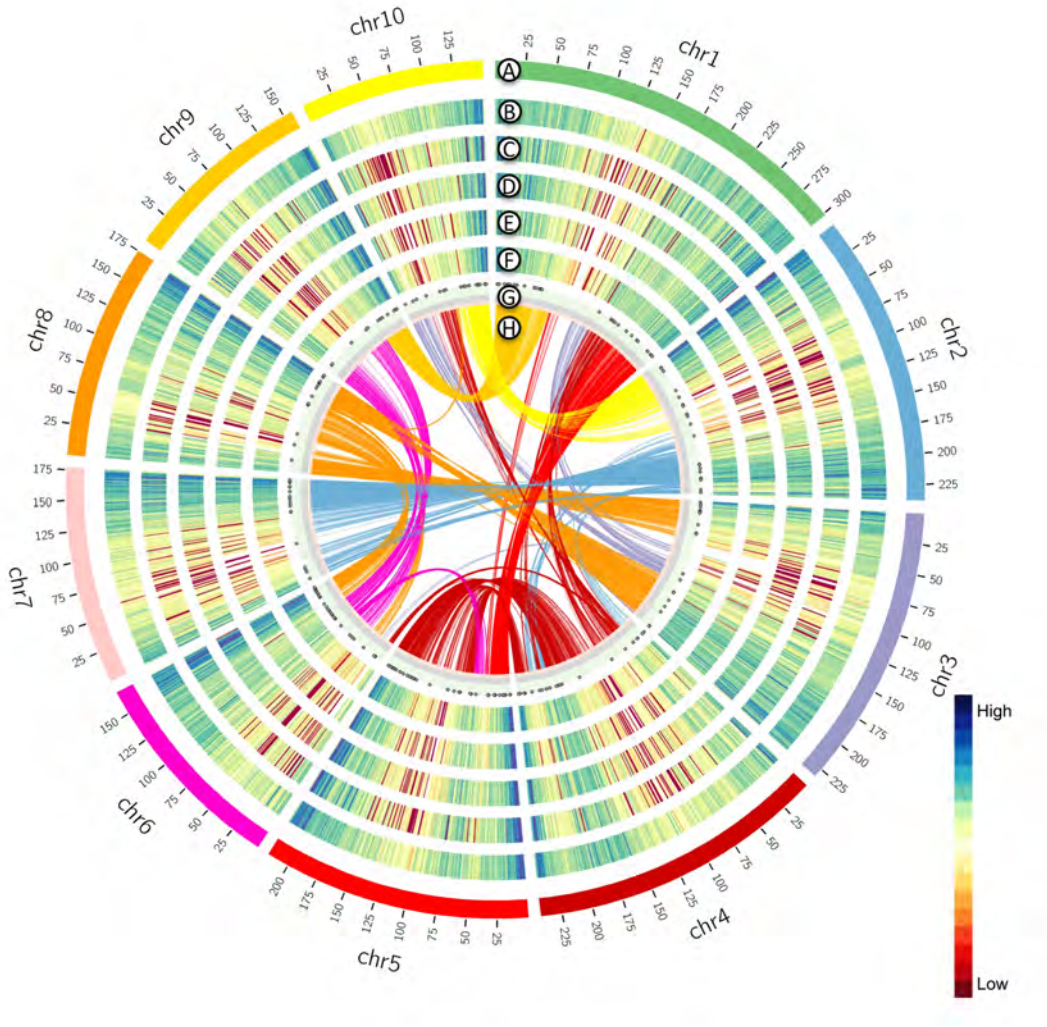


Figure S6. Tripsacum Iso-seq data mapped onto the maize reference genome (RefGen_v3). (A) 10 chromosomes of maize. (B) Maize gene density in each chromosome. (C) Density of highly expressed maize genes (FPKM $>$ 4.15). (D) Tripsacum transcript density in each chromosome. (E) Density of syntenic gene pairs between maize and sorghum. (F) Density of sorghum-maize-tripsacum gene pairs. (G) Distribution of Long non-coding RNA (IncRNA) in tripsacum. (H) Distributions of homeologous genes pairs following the whole genome duplication which is shared by both tripsacum and maize.

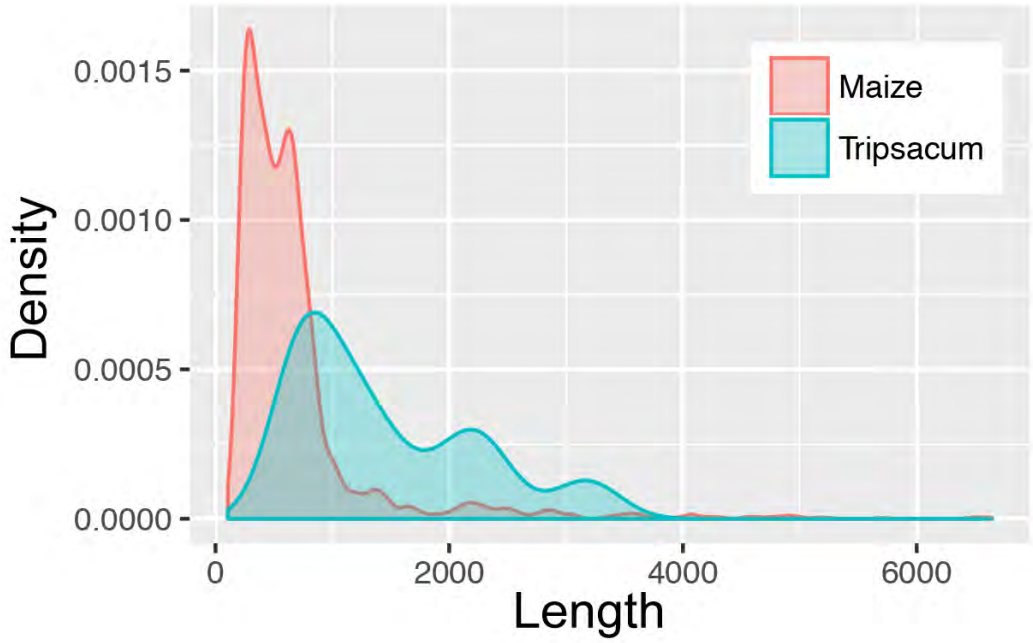


Figure S7. Comparison of length distribution of lncRNAs identified using Pacbio sequencing data between maize and tripsacum.

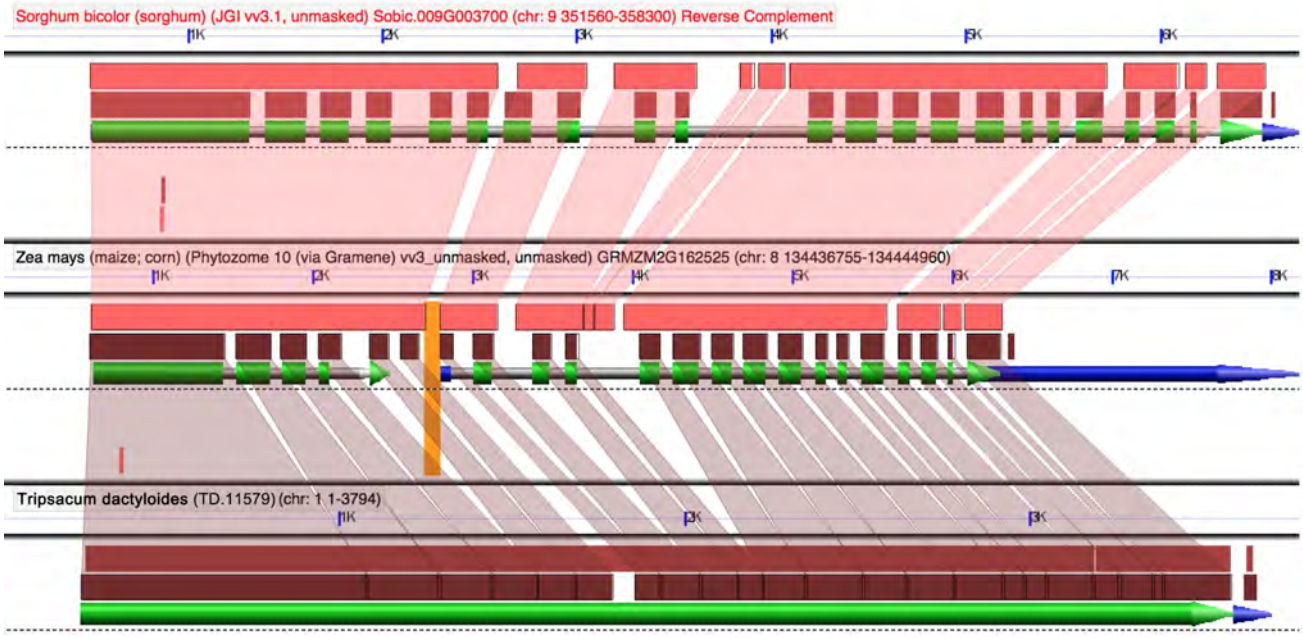


Figure S8. Example of one tripsacum transcript spans two maize gene models but correlated with one sorghum gene model through GEvo analysis.

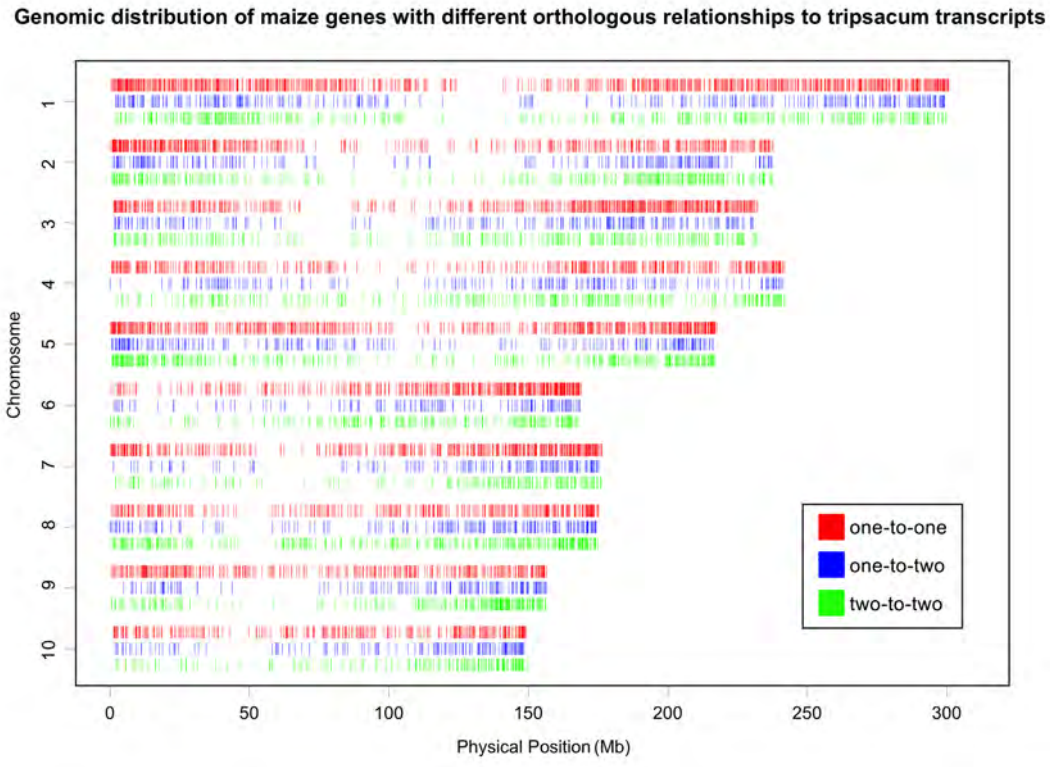


Figure S9. Distribution of maize genes with either 1:1, 2:1 or 2:2 relationships to assembled tripsacum transcripts across the maize genome.

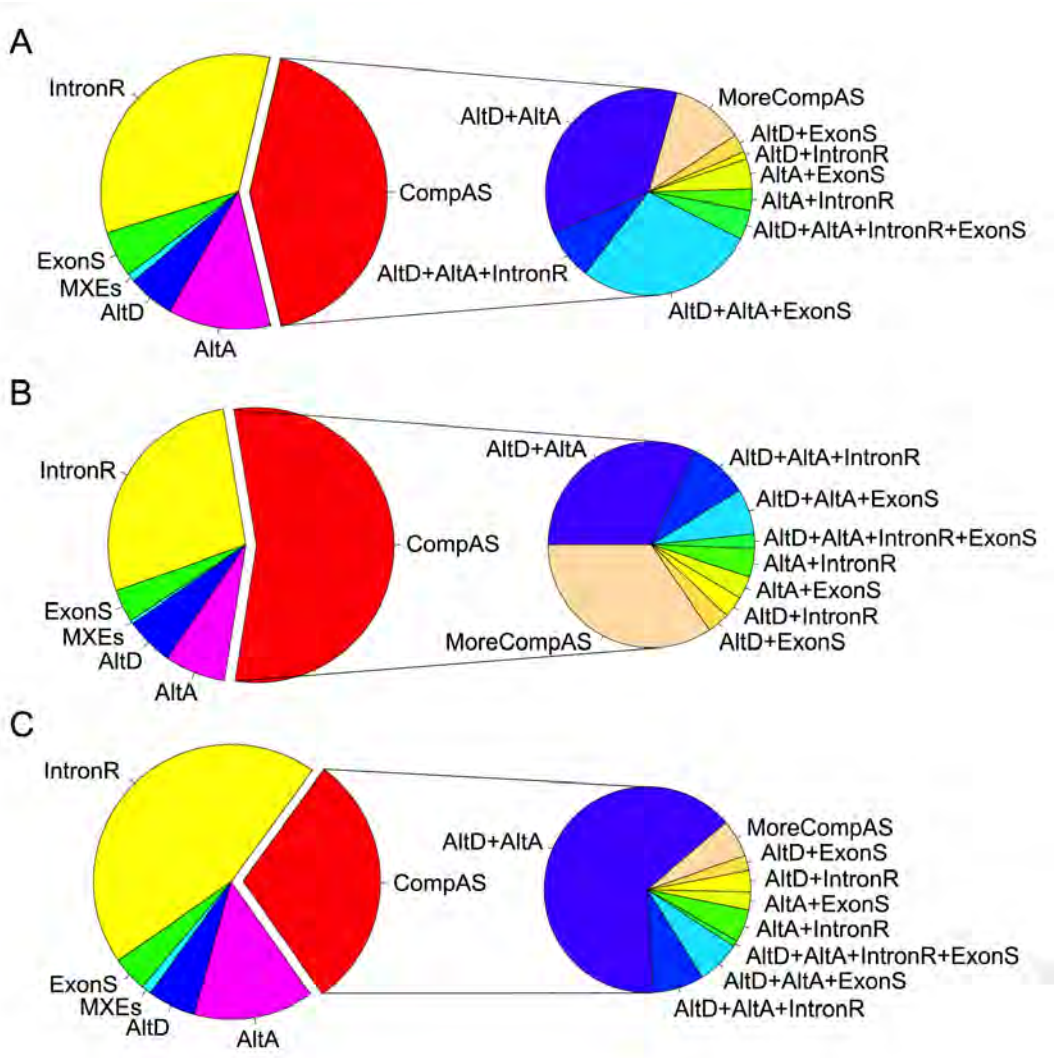


Figure S10. Proportions of alternative splicing (AS) types (Intron retention, IntronR; Exon skipping, ExonS; Alternative donor, AltD; Alternative acceptor, AltA; Mutually exclusive exons, MXEs; Complicated AS, CompAS; More complicated AS, MoreCompAS) found in (A) tripsacum and (B) maize using PacBio long sequences. (C) Proportion of conserved AS types between maize and tripsacum.

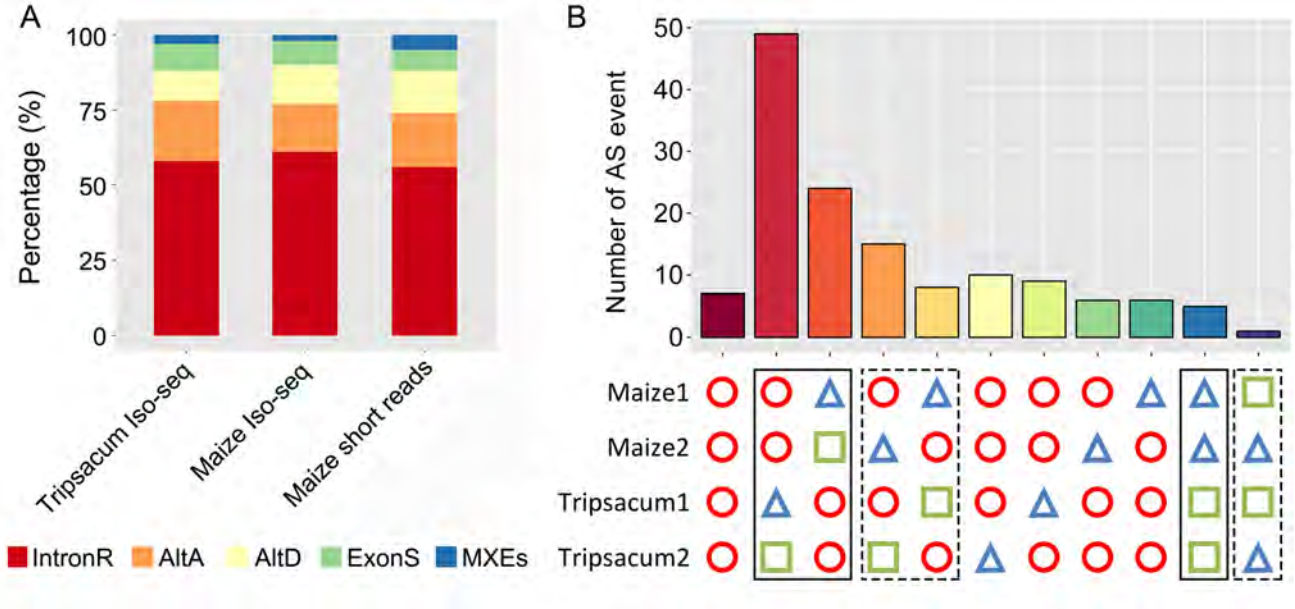


Figure S11. Comparison of alternative splicing (AS) distribution in (A) different species using long- and short- reads (long reads data for maize taken from [56] and short reads data of maize and sorghum were from [67]) and (B) subgenomes of maize and tripsacum. Identical shapes indicate genes with a conserved AS vent. Solid line boxes mark cases most parsimoniously explained by a change in *trans*-regulation of AS, while dashed line boxes mark cases most parsimoniously explained by a change in *cis*-regulation of AS.