Crain, J., Larson, S., Sthapit, S., Jensen, K., Poland, J., Dorn, K., Thomas, A., and DeHaan. Genomic insights into the NPGS intermediate wheatgrass germplasm collection.

Data Set:
R scripts (as RMarkdown) for phenotypic analysis and genome wide analysis of intermediate wheatgrass (*Thinopyrum intermedium*) National Plant Germplasm System (NPGS) collections. All RCode is documented in RMarkdown. The entire workflow progress from 0_0 to 7_0 sequentially, with the RMarkdown files documenting analysis and files.

Original sequence fastq files are a part of the NCBI sequence read archive (SRA) (https://www.ncbi.nlm.nih.gov/bioproject/) BioProject PRJNA866171. Sequence files are only needed if single nucleotide polymorphisms (SNPs) are called using the TASSEL pipeline.

Authors:

Jared Crain
Department of Plant Pathology, Kansas State University
4024 Throckmorton Plant Sciences Center, Manhattan, KS 66506, USA
jcrain@ksu.edu

Steve Larson
USDA-ARS, Forage and Range Research, Utah State University
Logan, UT 84322, USA

Sajal Sthapit
The Land Institute
2440 E. Water Well Rd, Salina, KS 67401, USA

Kevin Jensen
USDA-ARS, Forage and Range Research, Utah State University
Logan, UT 84322, USA

Jesse Poland
Center for Desert Agriculture, King Abdullah University of Science and Technology,
Thuwal Saudi Arabia

Kevin Dorn
USDA-ARS, Soil Management and Sugarbeet Research
Fort Collins, CO 80526, USA

Aaron Thomas
Department of Animal, Dairy and Veterinary Sciences, Utah State University
Logan, UT 84322, USA

Lee DeHaan
The Land Institute
2440 E. Water Well Rd, Salina, KS 67401, USA

Files and Description:

Files and folders located in the main directory and their contents:

File or Directory Name	Description of Contents
README.pdf	README file for workflow for manuscript evaluating IWG NPGS accessions.
IWG_PI.Rproj	R Project for manuscript.
0_0_Verify_GWAS_GBS_Files.Rmd	Documents original genotype SNP calling.
0_1_Compile_Germplasm.Rmd	Compiles IWG NPGS germplasm.
1_0_IWG_PI_Population_Structure.Rmd	Population STRUCTURE analysis for IWG NPGS collection.
1_0_1_IWG_PI_Population_Structure_Deprecated.Rmd	STRUCTURE analysis with more markers, results in conclusions but takes longer to run.
1_0_2_IWG_PI_Population_Structure_Deprecated.Rmd	STRUCTURE analysis with a subset of markers but not LD pruned, results in same conclusions.
1_1_IWG_PI_Population_Structure_Evaluation.Rmd	Population genetics—Mantel test, PCA, linear discriminant analysis.
2_0_IWG_PI_Amova.Rmd	Analysis of Molecular Variance.
3_0_IWG_PI_Phenotypic.Rmd	Statistical models for phenotypic data.
3_1_IWG_PI_Phenotypic_ANOVA.Rmd	Analysis of Variance evaluation of phenotypes between two STRUCTURE groups.
4_0_GWAS.Rmd	Genome-wide association study (GWAS) for phenotypic traits in 331 IWG NPGS accessions.
5_0_XP_GWAS_Pools.Rmd	Code to set up phenotypic pools for extreme-phenotype (XP)-GWAS and analyzing results.
6_0_XP_GWAS_Validation.Rmd	Validation of XP-GWAS with data from Crain et al. (2022) “Genetic architecture and QTL selection response for Kernza perennial grain domestication traits”.
7_0_Manuscript_Tables_Figures.Rmd	Document of tables, figures, and facts used throughout the manuscript.

0_0_Verify_GWAS_GBS_Files.html 0_1_Compile_Germplasm.html 1_0_IWG_PI_Population_Structure.html 1_0_1_IWG_PI_Population_Structure_Deprecated.html 1_0_2_IWG_PI_Population_Structure_Deprecated.html 1_1_IWG_PI_Population_Structure_Evaluation.html 2_0_IWG_PI_Amova. html 3_0_IWG_PI_Phenotypic. html 3_1_IWG_PI_Phenotypic_ANOVA. html 4_0_GWAS. html 5_0_XP_GWAS_Pools. html 6_0_XP_GWAS_Validation.html 7_0_Manuscript_Tables_Figures.html	Knitted Rmarkdown of each particular file. Provides package and software versions used in analysis.
File_List.txt	Contains a list of files contained in the directory. For each file relative path and MD5 checksum is provided to verify against data corruption. MD5 calculated with md5 on Macintosh MacOS Monterey Version 12.6.
beocat/	Contain original genotyping-by-sequencing (GBS) results, imputed files, and STRUCTURE results.
data/	Directory that holds all original data, processed data, final files, tables and figures.
scripts/	All scripts that were used to process files during pipeline processes or on HPC.

Note: All workflow can be traced using the Rmarkdown files in a sequential manner from 0_{sub_number}_Descriptor.Rmd to {max}_{sub_number}_File.Rmd. To run code, the working directory must be set to the main directory IWG_PI, which should set automatically if the IWG_PI.Rproj is opened in RStudio. Some files, particularly figures, that can be recreated easily from the code have been removed to reduce file size. **To run shell scripts:** shell scripts have been included, but file paths and directories must be updated for appropriate systems.