

Crain, J., Bajgain, P., Anderson, J., Zhang, X., DeHaan, L., and Poland, J. Enhancing Crop Domestication Through Genomic Selection, a Case Study of Intermediate Wheatgrass. *Frontiers in Plant Science*, doi: 10.3389/fpls.2020.00319.

Data Set:

R scripts (as RMarkdown) for phenotypic analysis and genomic selection for Enhancing Crop Domestication Through Genomic Selection, a Case Study of Intermediate Wheatgrass. All RCode is documented in RMarkdown. The entire workflow progress from 0_0 to 8_0 sequentially, with the RMarkdown files documenting analysis and files.

All fastq files to recreate SNP calling can be found in the NCBI sequence read archive (SRA) (<https://www.ncbi.nlm.nih.gov/bioproject/>) BioProject accession numbers PRJNA563706, PRJNA609095, and PRJNA608473.

Authors:

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

Prabin Bajgain
Department of Agronomy and Plant Genetics, University of Minnesota
St. Paul, MN 55108

James Anderson
Department of Agronomy and Plant Genetics, University of Minnesota
St. Paul, MN 55108

Xiaofei Zhang
The Alliance of Bioversity International and International Center for Tropical
Agriculture, Cali, Colombia

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

Jesse Poland
Wheat Genetics Resource Center
Department of Plant Pathology, Kansas State University
4024 Throckmorton Plant Sciences Center, Manhattan KS, USA 66506

Files and Description:

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

File or Directory Name	Description of Contents
README.pdf	README of file contents for Enhancing Crop Domestication Through Genomic Selection, a Case Study of Intermediate Wheatgrass.
0_0_GBS_Pipeline_Phenotypic_Documentation.Rmd	Documents GBS pipeline and SNP calling.
1_0_GBS_Filtering_and_Imputation.Rmd	Filters GBS data, code for imputation.
2_0_Population_Structure.Rmd	Evaluates population structure using PCA and Beagle.
3_0_Phenotypic_Data.Rmd	Set up ASREML phenotypic BLUP models. Code to set up files and then ran on ASREML standalone
4_0_Phenotypic_Compile_Models.Rmd	Compile ASREML results in R.
5_0_Genomic_Selection_Models.Rmd	Setup genomic selection models. Code developed in R and then ran on high performance cluster.
6_0_Genomic_Selection_Compile_Results.Rmd	Curate GS results in R.
7_0_Results.Rmd	Develop results and figures for publication.
8_0_Tables_Supplemental.Rmd	Code for extracting data for tables and supplemental figures.
0_0_GBS_Pipeline_Phenotypic_Documentation.html	
1_0_GBS_Filtering_and_Imputation.html	
2_0_Population_Structure.html	
3_0_Phenotypic_Data.html	
4_0_Phenotypic_Compile_Models.html	
5_0_Genomic_Selection_Models.html	
6_0_Genomic_Selection_Compile_Results.html	
7_0_Results	
8_0_Tables_Supplemental.Rmd.html	
File_List.txt	Knitted Rmarkdown of each particular file. Provides package and software versions used in analysis.
beocat/	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 Catalina Version 10.15.2.
bulk/	Contain original genotyping-by-sequencing results, imputed files, and STRUCTURE data.
data/	Extracted results from ASREML-R models. Directory that holds all original data, processed data, R objects, tables and figures.

output/

Directory holding genomic selection output
from high performance cluster.

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_Wild_Plants_FPS_SE. Some files, particularly figures, that can be recreated easily from the code have been removed to reduce file size.