



VERSION 2

APR 18, 2023

OPEN  ACCESS

DOI:
dx.doi.org/10.17504/protocols.io.4r3l27jzjg1y/v2

Protocol Citation: Blair L Waldron, B Shaun Bushman, Alexander J Hernandez, Kevin B Jensen, Thomas A Jones, Steven R Larson, Thomas A Monaco, Michael D Peel, Matthew D Robbins, Joseph G. Robins, Richard R-C Wang 2023. Forage and Range Research Laboratory Standard Operating Procedures (SOP) and Protocols. **protocols.io** <https://dx.doi.org/10.17504/protocols.io.4r3l27jzjg1y/v2> Version created by [Joseph G. Robins](#)

License: This is an open access protocol distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

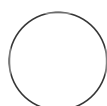
Protocol status: Working
We use this protocol and it's working

Created: Apr 14, 2023

Forage and Range Research Laboratory Standard Operating Procedures (SOP) and Protocols V.2

Blair L Waldron¹, B Shaun Bushman¹, Alexander J Hernandez¹, Kevin B Jensen¹, Thomas A Jones¹, Steven R Larson¹, Thomas A Monaco¹, Michael D Peel¹, Matthew D Robbins¹, Joseph G. Robins¹, Richard R-C Wang¹

¹United States Department of Agriculture, Agricultural Research Service, Forage and Range Research Unit



Joseph G. Robins

ABSTRACT

The United States Department of Agriculture, Agricultural Research Service, Forage and Range Research Laboratory conducts basic and applied research aimed at improving plant materials and management alternatives for sustainable stewardship of rangelands, pasture, and turf in the western U.S. This research is a multi-disciplinary effort combining expertise from plant breeding and quantitative genetics, molecular biology and genomics, plant physiology, and ecology. This research also relies on standard protocols and procedures to ensure experimental rigor, repeatability, and confidence. This document contains the standard operating procedures and protocols included in the 215 Waldron 2080-21000-018-000D project plan.

Breeding Nursery and Plant Trait Measurements

1 Field Plots/Seeded

Seeded plots are 1-m x 2-m on irrigated field sites and 1.5-m x 40-m on rangeland sites. Randomize plot assignments using a randomized complete block design with three to four complete blocks, unless otherwise specified. Seed plots using either a single-row Wintersteiger cone-seeder or a 5-row Hege precision cone-seeder with a row spacing of 22 cm, at a depth of 0.6 cm and seeding rate of 118 pure live seeds per linear meter of row.

2 Field Plots/Spaced-Plant

Randomize plot assignments using a randomized complete block design with three to four complete blocks, unless otherwise specified. Spaced-plant nurseries are comprised of 5 to 10 plants plot⁻¹ and established by hand or mechanically transplanting greenhouse-started seedlings with a spacing of 0.5 m between plants in a plot and 1.0 m between rows of plots. *Spaced-plant-in-sward nurseries* are a modification to mimic swards by either transplanting into an existing grass or legume sward, or the spaced-plants are over-seeded with such (Van Dijk and Winkelhorst, 1978).

3 Flowering Date

Record the date (Julian days) on which fifty percent of inflorescences have at least one fully formed flower.

4 Foliar Cover

To estimate bare ground, weed, and grass canopy cover, capture plot images (smartphone or digital camera) using close-range RGB imagery taken from approximately 1-m above ground level. Place at least three reference rectangular PVC frames (2-ft²) on the ground before taking the camera shots. Record plot numbers and picture IDs in the field. Use Samplepoint software to convert digital imagery to analyzable data (Booth et al., 2006).

5 Forage Nutritive Value

Using a Thomas Wiley Laboratory Model 4 mill (Arthur H Thomas Co, Swedesboro, NJ, USA), grind dried forage samples (see Forage Mass protocol) to pass through a 1-mm screen. Place ground sample into near-infrared spectroscopy sample cups. Using a Foss XDS near-infrared reflectance spectroscopy instrument (Foss, Eden Prairie, MN, USA), scan ground forage samples. Using the standardized NIRS consortium calibration equations, estimate forage sample values of crude protein (CP; N x 6.25), neutral detergent fiber (aNDF), acid detergent fiber (ADF), acid detergent lignin (ADL), in vitro true digestibility (IVTD), fatty acids (FA), NDF digestibility (NDFD), water soluble carbohydrates (WSC), ash, total digestible nutrients (TDN) (computed from NDF and NDFD, metabolizable energy (ME) (computed from TDN [Saha et al., 2010]), and Net energy

for gain (NEg) (National Research Council, 2000) are determined with standardized NIRS calibrations from the NIRS Forage and Feed Testing Consortium.

6 Leaf Area Index (LAI)

Use a Ceptometer, an instrument consisting of a data logger and a 90-cm long probe with photosensors that measures photosynthetically active radiation (PAR), to obtain plot LAI values (Parsons et al., 2011). At least five measurements are on each plot - a single measurement above the canopy and four measurements below the canopy at ground level. Measurements are perpendicular to the direction of the plot.

7 Salinity Tolerance Screening

Grow seedlings in a greenhouse under optimal conditions in silica sand-filled cups and watered with a complete nutrient solution until the three-leaf stage. Salt treatments consist of irrigation every three days by adding a complete nutrient solution with elevated salt (NaCl) concentrations that are increased every week by an electrical conductivity (EC) of 3 dS m^{-1} until an EC level of 24 dS m^{-1} is reached (Peel et al., 2004). Plants are scored as dead (no green present) or alive (some green available) until 95% of plants within a plot exhibit no green growth. Putative salt-tolerant plants are removed following the completion of the study and revived using fresh water. Salt imbalance in nutrient solutions is a frequent deficiency in screening studies and cultivar assessments (Shannon, 1984). To avoid an imbalance, NaCl and CaCl_2 are used in proportions to maintain a sodium adsorption ratio (SAR) of 3.5. The direct EC is measured with an Orin Model 120 conductivity meter (ThermoElectron Inc., Beverly, MA). When salinity tolerance is evaluated under field conditions, the soil EC is monitored throughout the study to identify both spatial and temporal variability in soil salinity.

8 Seed Yield and Mass

Measure under spaced-plant conditions by allowing pollination and full seed maturity for each genotype. Seed are hand harvested, threshed, cleaned, and weighed. Seed mass is determined based upon the weight of 1000 seeds or measured separately by MarviTech seed analyzer.

9 Stand Establishment and Persistence

Seedling or plant frequency is determined using the grid system described by Vogel and Masters (2001) in which the number of 12.5 cm^2 squares in a 1 m^2 grid that contain rooted plant(s) of the assigned treatment are counted. This number is converted to a percentage by dividing it by the total number of squares possible. Data are collected from each plot during the establishment year, and up to 10 years thereafter, to estimate stand establishment and stand persistence. In field-sized plots, stand establishment and subsequent plant persistence of target species are determined using the frequency grid method described above by using 10 grids along a 30-m transect in each plot and multiplying the frequency of occurrence by a fixed constant of 0.51 to obtain a conservative estimate of plants m^{-1} .

10 Turf Digital Image Analysis for Turf Quality

Turf digital image analysis uses a mobile light/camera box to collect images from each plot on a weekly basis from late April to early September (Karchner and Richardson, 2005). Digital images

are analyzed using ImageJ software (imagj.net) and SigmaScan (macro TurfAnalysis.bas) to determine ground cover, density, green color, and turf quality.

Geospatial Analysis and Modeling of High-Throughput (HTP...

11 Modeling Matrix

Regardless of classification or regression problems, a matrix composed of m columns (dependent variable and geospatial predictors) and n rows (individual sample observations) is prepared. Predictors will be the original spectral reflectance bands, derived vegetation indices VIs, digital surface model DSM and topographic derivatives. The values at each column-row ($m \times n$) intersection can be individual pixel values or zonal aggregations (i.e., median, mean).

12 Extraction of Matrix Values

In very rare cases, unique pixel values (multispectral, VIs) are associated with plant traits measured at the field. More often, HTP sample units are of planar nature (polygons). Obtain the boundaries of each sample polygon by collecting the corners using GPS rover units or by automatic delineation using high-resolution orthophotos, whichever is more efficient. For each polygon and for each geospatial predictor, extract all pixel values that intersect the sample polygon. Use zonal statistics algorithms such as `exactextr` (Baston, 2022) that can deal with pixels that are completely contained in a polygon or that can estimate the fractional coverage of a pixel within a polygon. If case raster predictors (VIs, topographic) have different spatial resolutions (pixel size), then use the largest grain size for the rest of the analysis, and the rest of the predictor rasters will be resampled to match the largest pixel size.

13 Modeling Schemes

Use non-parametric approaches, including support vector machines and random forests, to optimize predictions accuracies, as opposed to the model's interpretability (Cutler et al., 2007; Sheykhoumousa et al., 2020; Kok et al., 2021). Randomly subdivide the modeling matrix into training and validation subsets. Fit model(s) with the training subset, and their structure(s) will be simplified to prevent overfitting. Then, independently validate the model(s) using the validation subset.

14 Preparation of Spatially Explicit Response Variables

Once an acceptable error has been achieved with the proposed model structure, input of apply the model to the raster predictors included in the chosen model structure (Freeman & Frescino, 2018). Polygon-level predictions are obtained for the entire universe of plots for each experiment. This is possible because the UAS imagery have comprehensive spatial coverage of the plots and experiment.

Plant Breeding and Genetic/Genomic Protocols

15 Recurrent Selection

Forage breeding uses recurrent selection methodologies to capture the heritable variation in traits of interest. Recurrent selection is a cyclical plant breeding process used to increase the frequency of desirable alleles within a population for the trait of interest (Vogel and Pedersen, 1993). A *cycle*

of recurrent selection (abbreviated as C_n) for perennial forages consists of four years: one year of establishment, two years of phenotypic and family evaluation, and one year to intermate the selected genotypes in an isolated polycross nursery. Selection intensity (SI) is the percentage of families or individuals that are selected and polycrossed to complete the cycle. Methodologies include variations on current phenotypic selection, such as restricted recurrent phenotypic selection (*RRPS*), half-sib (HS), half-sib progeny test (*HSPT*), and conventional among-and-within-family selection (*AWFS*), also called between and within half-sib family selection (Casler, 2008; Casler and Brummer, 2008).

16 Genotype-by-sequencing (GBS)

Extract genomic DNA from samples using DirectZol DNA extraction kits (Zymo Research, Irvine, CA) or similar DNA extraction kits. Assess DNA quantity with spectrophotometry or fluorometry, and normalize samples to equal amounts across each experiment. Prepare sequencing libraries using a PstI-MspI two-enzyme protocol with custom barcodes (Poland et al., 2012). Sequence libraries as single-end 75- or 100-cycle flow cells on a NextSeq Illumina instrument at a licensed core facility. Determine single nucleotide polymorphism (SNP) genotypes with the reference genome-based v2 pipeline of TASSEL software (Bradbury et al., 2007). Genotype assignment for autopolyploid plants requires PolyRAD (Clark et al., 2019) to assign proper dosages as posterior probabilities. PolyRAD empirically derives overdispersion and expected heterozygosity values to remove multiallelic and multiple loci-mapping SNPs. Filter all biallelic SNP loci to keep those that align to only one place in the genome, contain a minimum of five (diploid or allopolyploid) or 10 (autopolyploid) read counts per individual to call a homozygous genotype, contain at least two sequence reads with different alleles per individual to call a heterozygous genotype, have less than 30% missing data per locus, and have a minor-allele frequency greater than 5%.

17 Genome Assembly and Scaffolding

Create sequencing libraries from an individual plant from high molecular weight genomic DNA and sequenced at third party core facilities with PacBio SMRT cells to produce HiFi reads with at least 20X coverage for each haplotype. Obtain estimates of genome size and heterozygosity by k-mer analysis using Jellyfish (Marçais and Kingsford 2011) and GenomeScope2 (Ranallo-Benavidez et al. 2020). For pseudohaploid assemblies, assemble HiFi reads using hifiasm (Cheng et al. 2021) or HiCanu (Nurk et al. 2020) and purge using hifiasm or purge_dups (Guan et al. 2020). For phased assemblies, assemble HiFi reads using hifiasm with proximity ligation data from commercial kits (Arima Genomics, Carlsbad, CA) or service providers (Phase Genomics, Seattle WA). Scaffold assembled contigs using proximity ligation data by a service provider or using juicer (Durand et al. 2016) and 3D-DNA (Dudchenko 2017). Evaluate assembly completeness and contiguity using assembly metrics from QUAST (Mikheenko et al. 2018) software and survey single copy orthologous genes using BUSCO (Manni et al. 2021) software, respectively. Remove scaffolds that are contaminants as identified by Blobtools2 (Challis et al. 2020) and Kraken (Wood et al. 2019), chloroplast or mitochondria as identified by BLAST (Camacho et al. 2009) to the NCBI RefSeq database, highly repetitive (< 10 Kb non-repetitive sequence) as identified by RepeatModeler2 (Flynn et al. 2020) and RepeatMasker (<http://www.repeatmasker.org>), or shorter than 10 Kb to obtain the final assembly.

18 **Genome Resequencing**

Construct Illumina 2X150 paired-end sequencing libraries from genomic DNA of each sample and sequence them at third-party core facilities to at least 10X coverage per sample. Check sequences for quality using FastQC

(<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>), trim them using Trimmomatic (Bolger et al., 2014), and align them to reference genomes using minimap2 (Li 2018). Plot coverage across the genome using the WGScoveragePlotter of jvarkit (Lindenbaum 2015). Identify presence-absence variants using a SGSelneloss-based method (Tay Fernandez, et al. 2022). Identify and annotate large-scale structural variants (chromosomal rearrangements, insertions-deletions, duplications) using coverage plots and visual inspection of mappings in a genome browser. Identify single nucleotide polymorphisms (SNPs) from the mapping files of all samples by removing duplicates using MarkDuplicates tool of the Picard toolkit (<https://broadinstitute.github.io/picard/>), then calling SNPs using the mpileup and call functions of bcftools (Danecek et al. 2021). Filter SNPs for minor allele frequency > 0.1, 70% max missing data, and minimum quality > 30 using vcftools (Danecek et al. 2011).

19 **Genome-wide Association Analysis (GWAS)**

Associate SNP genotypes with phenotypes using GWASpoly (Endelman & Rosyara, 2022) or rrBLUP (Endelman, 2011), modeling either disomic or polysomic (autoploid) inheritance. Use the 'Leave One Chromosome Out' method of determining population structure and kinship with the same markers used for GWAS (Yang et al., 2014). Association with phenotypes uses a false discovery rate (FDR) threshold of 0.05 and includes genetic structure and kinship covariates calculated in the respective R packages (R Core Team, 2019). Filter the significant loci using backward elimination to remove redundant loci.

20 **Genomic Prediction**

Most of the statistical analyses can be completed in the R programming language (R Core Team, 2019) using methods published for intermediate wheatgrass (Crain et al., 2021a; Crain et al., 2021b). Use a mixed model to analyze experimental data and develop BLUPs for each genet using ASReml 4.2 (Gilmour et al., 2021). Then use the BLUPs as empirical observations for genomic prediction. The model fits the inverse genomic relationship matrix, female parent, and units (factor with level for each experimental unit such as plant or half-sib family) with an autoregressive order 1 (AR1 x AR1) model of residual variance to correct for spatial effects. Create the genomic relationship matrix from a mixture of empirical and imputed SNP genotype calls using the A.mat function of the rrBLUP package (Endelman, 2011) to get the additive relationship matrix. Invert the additive relationship matrix from its Choleski decomposition using the chol2inv function, of the R base package. If necessary, use replicated genotypes to construct the additive relationship matrix so that the size and order of the matrix match plots or clones that are replicated in the field. Use the unit term to fit the 'nugget' variance when a correlation structure, such as AR1 x AR1, is applied to the residual. Use the Akaike Information Criterion and Bayesian Information Criterion to compare and choose the best model (Isik et al., 2017). Although BLUPs are predicted for missing observations, use only those BLUPs based on experimental observation for the development and validation of genomic estimated breeding values (GEBVs) described below. To make predictions for plants that have not been phenotyped, estimate GEBVs by line effects (G-BLUP model) using the kin.blup function of the rrBLUP package (Endelman,

2011) with an additive relationship matrix to determine the genetic covariance (G). Here the additive relationship matrix, computed using the A.mat function (Endelman, 2011), includes plants from the training population(s) that have been phenotyped and potentially large numbers of plants that have not been phenotyped. Determine the predictive ability of the G-BLUP models for each trait by a five-fold cross-validation procedure where a unique subset comprising 20% of the BLUPs, based on real experimental observations, is deleted five times so that each BLUP gets deleted once. Use the five resulting BLUP datasets to develop GEBVs for the missing observations (deleted BLUPs). Use the correlation of GEBVs for missing observations (deleted BLUPs) versus BLUPs based on experimental observations to determine the predictive ability of the G-BLUP model for each trait.

21 High-density Genotype Imputation

Breeding and genetic studies comprised of a large number of samples require inexpensive and highly informative genotyping methods. High-density haplotype reference panels of phased genome sequences allow efficient imputations for humans and even some livestock species, but these resources are not yet available for most plants (Browning and Browning 2013, Browning and Browning 2016, Browning, Zhou et al. 2018, Das, Abecasis et al. 2018, Davies, Kucka et al. 2021). However, relatively new methods for high-density genotype imputation based on extremely low sequence coverage (down to 0.15x coverage), without the need for additional reference panels, are also being developed (Davies, Flint et al. 2016, Ros-Freixedes, Whalen et al. 2020, Whalen, Gorjanc et al. 2020, Whalen and Hickey 2020, Browning, Tian et al. 2021). The Sequencing To Imputation Through Constructing Haplotypes (STITCH) algorithm (Davies, Flint et al. 2016) shows promise for any species with a high-quality reference assembly for read mapping, especially for populations with recent strong bottlenecks that limit the number of founder haplotypes. A pipeline involving standard methods of read mapping, variant calling, variant filtering, and STITCH is being developed and tested for IWG with promising results. The costs are competitive with GBS, but the STITCH imputations produce more than two million high-confidence SNPs per chromosome in *Kernza* germplasm. The density of SNPs from this approach vastly exceeds that of GBS and provides a powerful approach for breeding and genetic research.

22 Whole Transcriptome Gene Expression Analysis (RNA-seq)

Extract total RNA from individual plants of appropriate tissue types and treatments depending on research objectives. Construct and sequence Illumina 2x150 paired-end sequencing libraries at third-party core facilities. At least 20 million sequence reads are obtained from each sample.

Check sequences for quality using FastQC

(<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>), trim them using Trimmomatic (Bolger et al., 2014), and align them to reference genomes using HISAT2 (Kim et al., 2019). Count reads with features using HTSeq (Putri et al. 2022) and identify differentially expressed genes (DEG) from counts using predefined contrasts with DESeq2 (Love et al., 2014). Only consider as DEGs transcripts with at least a two-fold difference in expression values between control and treated samples, and with adjusted P-values (false discovery rate; FDR) less than 0.05. Cluster the resulting DEGs based on expression levels in treatment contrasts and visualize them using heat maps, a Venn diagram, expression profile charts, and volcano plots. Classify gene ontologies and functional enrichments of DEGs using the Trinotate pipeline (Bryant et al., 2017) and Omics Box Blast2GO (BioBam, Valencia, Spain). Use the WGCNA R package (Langfelder & Horvath, 2008)

to empirically determine co-expression networks among DEGs, based on their expression profiles and to visualize the results.

Statistical Analyses

- 23** Analyses of all studies rely on generalized linear mixed models. In most cases, experience shows that errors tend to be normally distributed, allowing us to use normal theory and the reductionist models available in general linear mixed models. In other cases, where errors are not normally distributed, use generalized approaches to model proper link functions and distributions of both raw data and errors. Model experiments that are repeated in time or space, including sampling dates or times, harvests, seasons, and/or years, using various covariance structures to appropriately model correlations among residuals. Subject larger field trials to spatial analyses to model spatial autocorrelations using various models, to account for the inevitable within-block variability that cannot be accounted for by experimental design.

Unmanned Aerial Systems (UAS) Data Collection and Post-p...

- 24 Global Positioning System (GPS) Baseline**
At each farm/site for which UAS imagery is required, establish a base station by logging a minimum of four hours of continuous data streaming. Upload these data to the website <https://geodesy.noaa.gov/OPUS/> to obtain a high-accuracy (± 0.02 m) GPS solution for each farm/site. Place a metal survey marker at the location of each base station for easy identification in future flights.
- 25 Flight Mission Planning**
Once a spatial domain for the experiment has been mapped, prepare flight missions using QGroundControl (<http://qgroundcontrol.com/>) software. Depending on the individual requirements of the experiment and plant trait(s) to be evaluated, define a ground sampling distance (GSD) to be used in conjunction with the multispectral camera specifications to adjust the elevation and speed of the UAS to guarantee a minimum of 75% overlap between camera captures (sidelap and frontlap). If imagery is being collected in non-agricultural settings, then plan the mission so that is oriented perpendicular to the general slope of the area. The area to be flown will be at least 10% larger than the area of interest so that the edges of the fields are completely covered. Plan all missions using terrain-following protocols.
- 26 Flight Execution**
Conduct flights only on sunny days to minimize the effects of changing illuminations and only within two hours of local solar noon. Prior to launching the flight mission, place a minimum of six highly visible ground control points (GCPs) throughout the area to be flown. Record accurate coordinates of these GCPs using a rover GPS that is receiving corrections from the GPS base station via LORA (long-range) radio communications. In addition, capture images of calibrated reflectance panels (for visible and Near Infrared bands) immediately before and after every flight. Point sensors onboard the UAS straight down (nadir or close to nadir). If a particular experiment requires calibrated thermal readings, then set up three infrared radiometers to

capture continuous readings over three (3-m x 3-m type 822 fabric) airborne sensor calibration targets during the flight duration. These are diffuse hemispherical reflectivity tarps calibrated at 6.5%, 36% and 64% reflectance, representing dark (hot), medium, and white (cold) objects. Use temperature readings from these three objects to calibrate the thermal estimations from the UAS Forward Looking InfraRed (FLIR) sensors. Once the flight is finished, copy imagery files to a field laptop computer for later transfer to FRR storage and computational servers.

27 Pre-photogrammetry Imagery Post-processing

Individual photos digital numbers DN captured using the Micasense sensors (i.e., Altum, RedEdge) will be converted to reflectance values using the Micasense imagery processing Python scripts (<https://github.com/micasense/imageprocessing>).

28 Photogrammetry Processing

Stitch together photos depicting reflectance values using Pix4D or the OpenDroneMap drone mapping software (<https://www.opendronemap.org/webodm/>). Georectify imagery using the GCPs coordinates obtained prior to the execution of flight missions. Generate multispectral reflectance orthophotos, digital surface models (DSM), as well as several vegetation indices (VIs), such as NDVI, NDRE, VARI, and RVI, for each flight mission. GeoTiff is the file format for the outputs. Experiment-Farm-Date (YYYY-MM-DD) is the file naming convention