# The Genetic Makeup of a Global Barnyard Millet Germplasm Collection

Jason G. Wallace,* Hari D. Upadhyaya, M. Vetriventhan, Edward S. Buckler, C. Tom Hash, and Punna Ramu

## Abstract

Barnyard millet (*Echinochloa* spp.) is an important crop for many smallholder farmers in southern and eastern Asia. It is valued for its drought tolerance, rapid maturation, and superior nutritional qualities. Despite these characteristics there are almost no genetic or genomic resources for this crop in either cultivated species [*E. colona* (L.) Link and *E. crus-galli* (L.) P. Beauv.]. Recently, a core collection of 89 barnyard millet accessions was developed at the genebank at the International Crops Research Institute for the Semi-Arid Tropics (ICRISAT). To enhance the use of this germplasm and genomic research in barnyard millet improvement, we report the genetic characterization of this core collection using whole-genome genotyping-by-sequencing. We identified several thousand single-nucleotide polymorphisms segregating in the core collection, and we use them to show patterns of population structure and phylogenetic relationships among the accessions. We determine that there are probably four population clusters within the *E. colona* accessions and three such clusters within *E. crus-galli*. These clusters match phylogenetic relationships but by and large do not correspond to classification into individual races or clusters based on morphology. Geospatial data available for a subset of samples indicates that the clusters probably originate from geographic divisions. In all, these data will be useful to breeders working to improve this crop for smallholder farmers. This work also serves as a case study of how modern genomics can rapidly characterize crops, including ones with little to no prior genetic data.

THE genus *Echinochloa* includes 20 species that are distributed widely in the warmer parts of the world. Barnyard millet is the common name for several *Echinochloa* species, all of them native to southern or eastern Asia. Several of these are aggressive weeds, while two are cultivated as cereals: *E. crus-galli* (L.) P. Beauv. (Japanese barnyard millet) is a temperate grass with awned spikelets, is native to Eurasia, and was domesticated in Japan some 4000 yr ago; and *E. colona* (L.) Link (Indian barnyard millet) occurs widely in tropical and subtropical areas with awnless spikelets and was domesticated in India (de Wet, 1983). Both cultivated species have two subspecies each—*colona* and *frumentacea* in *E. colona* and *crus-galli* and *utilis* in *E. crus-galli*—and each subspecies is further divided into zero to four different races (Upadhyaya et al., 2014). Weedy relatives of barnyard millets are known to infest farmers' fields in Japan, India, the United States, and other locations (Wanous, 1990). More distant relatives include several other cultivated plants, including switchgrass (*Panicum virgatum* L.), foxtail millet (*Setaria italica* subsp. *italica*), and pearl millet [*Pennisetum glaucum* (L.) R. Br.].

Barnyard millet is mainly grown in India, China, Japan, and Korea for human consumption as well as fodder (Upadhyaya et al., 2014). The crop is valued for its drought tolerance (Dwivedi et al., 2012), short growth period (sometimes in as little as 6 wk; Wanous 1990), and superior

Published in The Plant Genome 8
doi: 10.3835/plantgenome2014.10.0067

5585 Guilford Rd., Madison, WI 53711 USA
An open-access publication



J. Wallace, E. Buckler, and P. Ramu, Institute for Genomic Diversity, Cornell University, Ithaca, NY, 14853. E. Buckler, USDA–ARS, Ithaca, NY. H. Upadhyaya and M. Vetriventhan, International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Patancheru 502 324, Telangan, India. C. Hash, International Crops Research Institute for the Semi-Arid Tropics (ICRISAT) Sahelian Center, (ISC), Niamey, Niger. Received 20 Oct. 2014. Accepted 9 Dec. 2014. *Corresponding author (jason.wallace@cornell.edu).

Abbreviations: GBS, genotyping-by-sequencing; MDS, multidimensional scaling; PCA, principal components analysis; SNP, single-nucleotide polymorphism; VCF, variant call format.

nutrition value (Saleh et al., 2013). These characteristics make it an important supplemental crop for small-scale farmers because they can plant and harvest it between major crop growing seasons. It can also be used as a substitute crop in emergencies when the major crop fails.

Barnyard millet is also highly nutritious, consisting of 55% carbohydrate, 11% protein, 3.9% fat, and 13.6% crude fiber, with significant amounts of both calcium and iron (Saleh et al., 2013). Its fiber and iron contents are higher than those of rice, wheat, and other millets, and its low glycemic index makes it an ideal food for management of diabetes mellitus (Ugare, 2008; Saleh et al., 2013; Sharma et al., 2013).These characteristics also make it a good candidate for manufactured food products such as baby foods, snacks, and dietary foods (Ugare, 2008; Surekha et al., 2013; Anju and Sarita, 2010; Vijayakumar et al., 2010).

As with most minor crops, there has been very little attention and few resources devoted to the study of barnyard millet. Nonetheless, it is increasingly recognized that these minor crops are an important component of food security for smallholder farmers (Nelson et al., 2004; Naylor et al., 2004; Godfray et al., 2010; Varshney et al., 2010). The application of modern genomic methods to these crops holds the potential to increase food security and independence among many of the world's poorest (Naylor et al., 2004; Nelson et al., 2004).

An important part of these resources are well-defined core collections of source germplasm. To this end, ICRISAT recently developed a core collection of barnyard millet (Upadhyaya et al., 2014) containing 89 accessions of the two cultivated species of barnyard millet: *E. colona* and *E. crus-galli*. These accessions represent 12% of ICRISAT's entire barnyard millet collection and capture a large amount of its phenotypic diversity, but there is no public data on their genetic composition. To address this gap, we genotyped 95 barnyard millet accessions, including the entire core collection, using genotyping-by-sequencing (GBS) (Elshire et al., 2011). This resulted in a genomewide set of >21,000 single-nucleotide polymorphisms (SNPs) segregating across the entire collection and several thousand SNPs segregating within each species.

One challenge of barnyard millet genetics is that all members are polyploid. Both *E. colona* and *E. crus-galli* are usually reported as hexaploids, with $2n = 6x = 54$ (Prasada Rao et al., 1993; Upadhyaya et al., 2008). However, other numbers have been reported (Wanous, 1990 and references therein), possibly indicating heterogeneity in the species. Total genome size has been estimated by flow cytometry to be roughly 1.4 gigabases (Bennett et al., 1998, 2000).

Despite these challenges, we successfully used our SNP dataset to perform a comprehensive analysis of the genetics, population structure, and phylogenies of the complete barnyard millet core collection. These analyses were done both on the collection as a whole and also for *E. crus-galli* and *E. colona* separately. The specific results should be useful to almost any researcher working on barnyard millet.

## Materials and Methods

### Plant Materials

All plant materials were taken from the barnyard millet collection available at ICRISAT in Patancheru, India. These consisted of the 89 accessions of the barnyard millet core collection (Upadhyaya et al., 2014), along with six additional accessions chosen from the ICRISAT gene bank (additional samples plus one blank were used to fill out a 96-well plate). Thus there were a total of 95 accessions: 65 from *E. colona* and 30 from *E. crus-galli*. Passport data for all samples is included in Supplemental File S1 and was gathered from both Genesys (https://www.genesys-pgr.org/) and ICRISAT. Seedlings were grown in the greenhouse and leaf tissue harvested 10 d after emergence.

### DNA Extraction and Genotyping-by-Sequencing

DNA was extracted using modified cetyltrimethylammonium bromide (CTAB) methodology (Mace et al., 2003). Lyophilized DNA was then sent to the Institute for Genomic Diversity (Cornell University, Ithaca, New York, USA) for genotyping with GBS. Library preparation and sequencing followed the protocol described in Elshire et al. (2011), with ApeKI restriction enzyme for genomic digestion. The barcoded samples were then pooled in 96-plex and sequenced in three lanes of an Illumina HiSeq 2500 (Illumina, Inc.).

### Single-Nucleotide Polymorphism Calling

Single-nucleotide polymorphisms were identified using the TASSEL-GBS pipeline (Glaubitz et al., 2014) in TASSEL v4.3.11, with the TASSEL-UNEAK variant pipeline (Lu et al., 2013) in TASSEL v5.0.9 used to align sequencing tags for SNP calling. Complete scripts and key files used to call SNPs are available in Supplemental File S1. Raw FASTQ data is available from the Sequence Read Archive (SRA, http://www.ncbi.nlm.nih.gov/sra), accession SRX734221.

Quality filtering was performed primarily using built-in functions in VCFtools (Danecek et al., 2011), with the exceptions of filtering by coverage and heterozygosity. To filter by coverage, we first used the "--missing-indv" option in VCFtools to generate a report of missingness across samples, which was then trimmed in R (R Core Team, 2014) to generate a list of individuals that failed the cutoff. These individuals were then removed with the "--remove" option in VCFtools. Filtering by heterozygosity was similar, except that we used TASSEL (Bradbury et al., 2007) to generate the report ("–genoSummary site") and keep only the sites that passed the R filter ("–includeSiteNamesInFile").

All bioinformatics and subsequent analyses were performed on an 8-core Intel i7 desktop workstation with 32 GB of RAM running Linux Mint 16. All bioinformatic scripts used in this study are available in Supplemental File S1. The VCF-formatted files of all SNP sets (including a master file of less stringently filtered ones) are available in Supplemental File S2.

## Population Structure Analysis

Population structure was determined with fastSTRUCTURE (Raj et al., 2014) [github commit f94d4e53ca, https://github.com/rajanil/fastStructure]. After choosing a clustering level, individual samples were assigned to clusters if they had at least 60% membership in that cluster.

## Phylogenetics

Phylogeny for each dataset was determined with PHYLIP 3.695 (Felsenstein 1989) by maximum parsimony using 100 bootstrap iterations of all genotypes. The resulting phylogenetic trees were merged with SplitsTree4 (Huson and Bryant, 2006) into a consensus network using the mean edge weight with a threshold of 0.2.

## Multidimensional Scaling

To perform multidimensional scaling, SNP datasets were converted to distance matrices in TASSEL (Bradbury et al., 2007) before applying singular value decomposition in Python using "linalg.svd()" in the numpy package (van der Walt et al., 2011).

# Results

## Rapid Genotyping of the Barnyard Millet Core Collection

Genotyping-by-sequencing was performed on the 89-member ICRISAT core collection (Upadhyaya et al., 2014) plus six additional samples, for a total of 95 unique accessions (see Methods for details). Libraries were prepped using ApeKI restriction enzyme, both because it cuts frequently and because it has a history of performing well for GBS in many different grass species (Sharon Miller, personal communication, 2013). Read depth was relatively constant across samples (Fig. 1a), with a median depth of 6.84 million reads from each accession.

After assigning reads, SNPs were called using the TASSEL-GBS pipeline (Glaubitz et al., 2014). The TASSEL-UNEAK filter (Lu et al., 2013) was used to align reads in the absence of a reference genome. Raw SNP calls were filtered to include only sites with 80% coverage across samples and minor allele frequencies ≥0.05, and only samples with ≥25% coverage across the remaining sites.

Limited funding is the defining feature of orphan crops, so we investigated the effect of reducing the number of sequencing runs on the number of final SNPs. The results of rerunning our SNP-calling pipeline using only one or two flowcells at a time is shown in Fig. 1b. While there is a significant increase in SNPs when going from one to two flowcells, including all three flowcells has only a minor effect on the total SNPs recovered. Most of the SNPs that are recovered at higher depth are probably not truly new; instead, the increased sequencing depth pushes their coverage high enough that they pass filtering instead of being removed. Regardless of the exact reason, these data suggest that running the barnyard millet samples in 96-plex on two flowcells—functionally the equivalent of 48-plex in one flowcell—is probably sufficient, at least

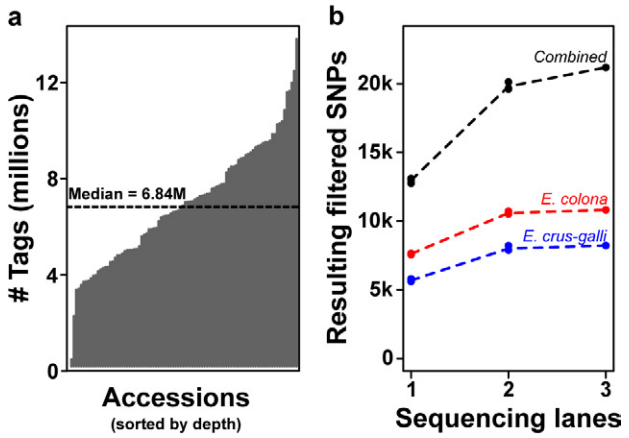


Figure 1. Genotyping statistics. (a) Read depth per individual. The number of good 64-base pair tags (sequencing reads) for each individual is shown in order of increasing depth. One sample is not shown because it had <5000 reads. (b) The number of single-nucleotide polymorphisms (SNPs) recovered at the end of our pipeline as a function of sequencing depth. Even with just three flowcells, the number of SNPs recovered has begun to plateau, especially in the individual species' datasets. Further depth would be unlikely to recover more SNPs. (See Supplemental Fig. S1 for a similar graph at different filtering stringencies).

**Table 1. Single-nucleotide polymorphism (SNP) statistics.**

| | No. samples† | Total read depth | Median read depth | No. filtered SNPs | No. discriminating SNPs |
|---|---|---|---|---|---|
| All samples | 92 | 643 M | 6.84 M | 21186 | 2579‡ |
| *E. colona* | 65 | 465 M | 6.92 M | 10816 | 1299 |
| *E. crus-galli* | 22 | 150 M | 7.00 M | 8217 | 1444 |

† The number of samples remaining after applying initial filters.

‡ The discriminating SNPs dataset across all samples was made by combining the discriminating SNPs from the two individual species instead of filtering the complete (nondiscriminating) SNP set.

for the filtering level given here. Plots for both higher and lower missing data amounts are shown in Supplemental Fig. S1; these largely follow the same pattern, and they may be useful when considering applications that have higher or lower tolerance for missing data. Sequencing these samples to greater depth will probably keep giving diminishing returns. This also implies that our dataset recovers a majority of the SNPs in these samples that are accessible via GBS with ApeKI. (See the Discussion for suggestions on finding additional SNPs.)

The final SNP counts for our dataset are shown in Table 1. After filtering, the complete dataset contains 21,186 sites across 92 accessions. We also split these accessions into the two species based on their population structure assignments (see below) and applied the same SNP filters. This resulted in 10,816 SNPs across 65 accessions for *E. colona* and 8217 SNPs across 22 accessions for *E. crus-galli*. The scripts and support files for genotyping and analysis are in Supplemental File S1, and all SNP datasets are available in Supplemental File S2.

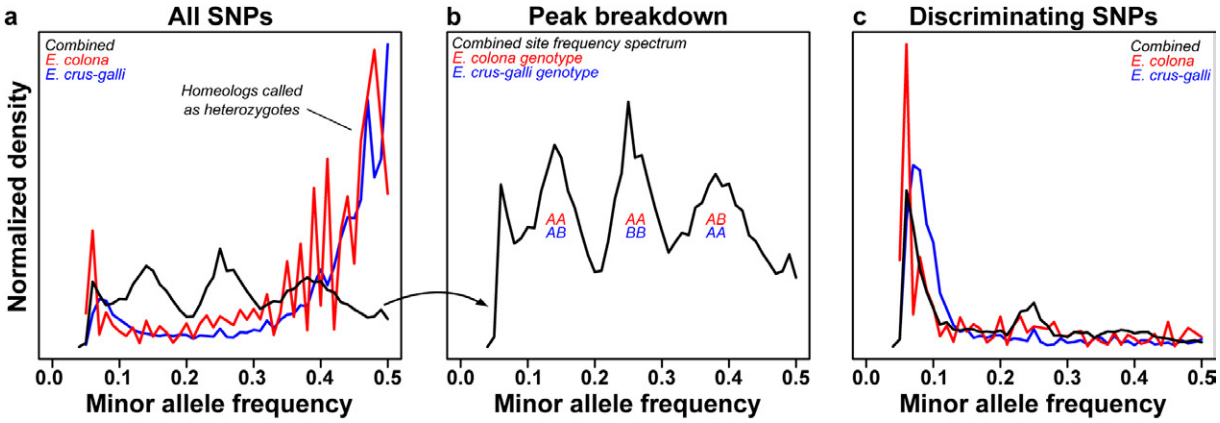


Figure 2. Site frequency spectra of the filtered single-nucleotide polymorphisms (SNPs). (a) The site frequency spectrum of the full SNP dataset for all samples (black), *Echinochloa colona* (red), and *E. crus-galli* (blue). The excess of sites near 0.5 in the two individual species is due to homeologous alleles lining up against each other and being called as heterozygous. The jaggedness of the spectra is due to the small number of samples. (b) Zoom-in of the combined site frequency spectrum from (a). The three largest peaks correspond to SNPs that are segregating differently in the different species; theoretical genotypes are marked below each peak and colored according to the lines in (a). (c) Site frequency spectrum for the discriminating SNP dataset. All the high-heterozygosity SNPs have been removed, leading to a distribution that is much closer to expectation. Note the slight peak in the combined dataset at ~0.25, indicating SNPs that are almost or completely fixed in both species, but for different alleles.

## Dealing with Homeologs

One of the challenges of working with polyploid species is that homeologous sequences often align together, resulting in many heterozygous calls that are actually due to separate (and potentially non-recombining) homeologs. This is evident in the site frequency spectrum of our filtered SNPs, where the spectra for the individual species shows a large excess of allele frequencies near 0.5 (Fig. 2a). The spectrum across the entire collection does not show this pattern, but is instead dominated by SNPs that are differentially polymorphic in the two species (Fig. 2b); that is, at least one species shows no variation among its homeologs, while the other species either has variation among them (right and left peaks) or is fixed for a different allele (central peak).

We considered assigning copy number values to genotypes based on their read depth, but the values show too much spread for us to be confident in the results (Supplemental Fig. S2). Since we could not confidently call copy number, we took advantage of the fact that barnyard millet is mostly self-fertilizing (Potvin, 1991; Dwivedi et al., 2012), meaning that very few sites will be genuinely heterozygous. This allows us to filter out the sites that appear highly heterozygous to enrich for single-copy regions of the genome and for sites that have become fixed across homeologs. We call the resulting SNP sets discriminating SNPs because they should have more power to discriminate among the different accessions.

The number of discriminating SNPs that result from removing sites with >20% heterozygosity from each dataset is shown in Table 1. Note that applying this filter across the entire collection would bias the data toward SNPs that are monomorphic in *E. colona* due to its larger representation. To avoid this, we first made discriminating SNP sets for each species separately, then filtered the combined dataset to contain only the sites in either species' individual sets. This will tend to make the two species look more similar to each other than the raw data, but since the discriminating SNPs still strongly separate the species (compare Fig. 3 and Supplemental Fig. S3), the issue is not severe.

## Population Structure and Phylogenetic Analysis

After finding quality SNPs by the methods above, we analyzed the population structure and phylogenetic relationships among the samples (Fig. 3). Those interested in seeing the effect of removing the homeologous SNPs can compare these results with those in Supplemental Fig. S3, where the same analyses were run with the entire (non-discriminating) SNP set.

For population structure, we used the program fastSTRUCTURE (Raj et al., 2014), an updated version of the program STRUCTURE (Pritchard et al., 2000) designed to handle large SNP datasets rapidly. While fastSTRUCTURE includes a script to identify a range for the optimum number of clusters ($K$), we found that it always selected a $K$ value of exactly 2. This occurred even when visual inspection of the results showed apparently better splits of the data at different $K$ values (see Supplemental Fig. S4). Because of this, we decided to choose the optimal level of $K$ based on how cleanly it separated different populations within the data. The choice of clustering level in Fig. 3 is thus somewhat arbitrary but nonetheless shows good correlation to the phylogenetic analyses. For comparison, population divisions at all levels of $K$ from 2 to 10 are shown in Supplemental Fig. S4.

For the phylogenetic analysis, we used PHYLIP (Felsenstein 1989) to calculate maximum parsimony trees over 100 bootstrap iterations of each dataset. The bootstrap trees were merged with SplitsTree4 (Huson and Bryant, 2006) to form a consensus network, which

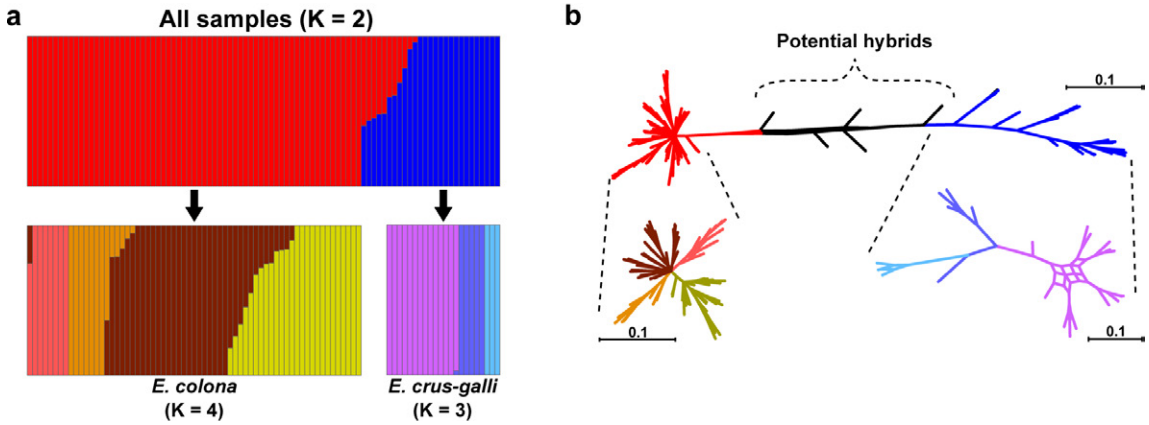


Figure 3. Population structure and phylogeny. (a) Population structure analysis with fastSTRUCTURE (Raj et al., 2014) strongly separates the two species of barnyard millet, along with four primary clusters in *Echinochloa colona* and three primary clusters in *E. crus-galli*. See Supplemental Fig. S4 for the results of clustering at different levels. (b) Phylogenetic analysis closely corresponds with the structure analysis, with inferred clusters generally matching major branch points in the phylogeny. Webbing among branches indicates ambiguity where at least 20% of trees shown an alternate arrangement, and scale bars show phylogenetic distance as calculated by PHYLIP (Felsenstein, 1989). Phylogenetic webs with full sample names are in Supplemental Fig. S5. For comparison, these same analyses were also performed on the full (nondiscriminating) single-nucleotide polymorphism dataset (Supplemental Fig. S3).

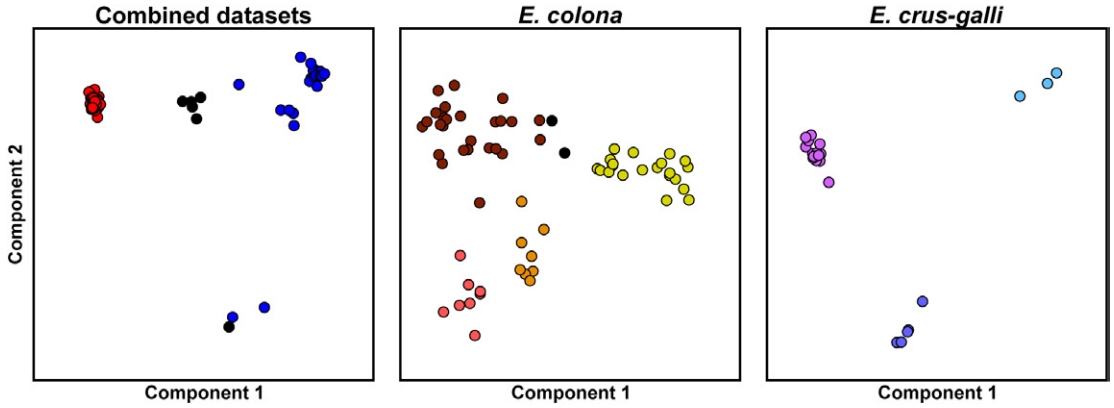


Figure 4. Multidimensional scaling of the population datasets. Multidimensional scaling was performed on each of the single-nucleotide polymorphism datasets. Points are plotted along the first two dimensions (*x* and *y* axes, respectively) and have been colored to match their cluster assignment from Fig. 3. The clusters in each set are largely consistent with the population structure from Fig. 3.

is similar to a consensus tree but with ambiguous splits shown as webbing among the branches. The resulting phylogenetic webs are shown in Fig. 3b and are colored to match the corresponding fastSTRUCTURE results.

As expected, both structure and phylogenetic analysis clearly separates the two species of barnyard millet from each other. Five samples show almost exactly 50% membership in each species cluster; these are probably hybrids between accessions. Such hybrids are known to be sterile (Prasada Rao et al., 1993; Upadhyaya et al., 2008), so these samples may be the result of seed contamination and were excluded from further analysis.

Based on the population structure results, we split the accessions into groups of *E. colona* and *E. crus-galli* and performed structure and phylogenetic analysis on each species separately. We identify four major clusters in *E. colona* and three in *E. crus-galli* that match the corresponding phylogenetic results well.

## Multidimensional Scaling to Confirm Population Structure

As further confirmation of the population structure in each species, we also performed multidimensional scaling (MDS) on each dataset (Fig. 4). Multidimensional scaling is a dimensionality-reduction technique that tries to reduce high-dimensional data to a smaller number of significant dimensions. It is extremely similar to principal components analysis (PCA), with the main difference being that for MDS the raw SNP scores are first converted into a matrix of distances between all the samples. This conversion is necessary because PCA does not function on datasets where some elements are missing, and the stochastic nature of GBS ensures that essentially every dataset will have at least some missing data, and frequently quite a bit.

Plotting the first two MDS dimensions of each dataset (Fig. 4) shows a clear separation by subpopulation, confirming the cluster divisions mentioned above. Interestingly, three *E. crus-galli* samples cluster far from

both *E. colona* and the remaining *E. crus-galli* samples (Fig. 4, left). These three samples correspond to the pale blue group from Fig. 3, a monophyletic group with long branch lengths. These data imply that this group has a relatively high proportion of unique alleles, an important consideration when selecting germplasm for breeding.

## Origins of Population Structure

We compared the population structure assignments with each accession's passport data (included in Supplemental File S1) to see how well our analyses match up with expected classifications. The division into *E. colona* and *E. crus-galli* largely matches the existing classification (Supplemental Fig. S6a); the small number of exceptions may be misclassified or may simply be due to mislabeling (at seed storage, DNA preparation, or some other step). These classifications also closely mirror the country of origin, since *E. colona* is primarily an Indian crop while *E. crus-galli* is largely from Japan (Supplemental Fig. S6b). Additionally, each species has four races represented among our samples: Intermedia, Laxa, Robusta, and Stolonifera for *E. colona*; and Crus-galli, Intermedia, Macrocarpa, and Utilis for *E. crus-galli*. With the exception of *E. crus-galli* race Crus-galli, though, none of them appear to cluster phylogenetically (Supplemental Fig. S7a). This matches previous observations that the races of *E. colona* do not correspond to geographic, ecological, or ethnological divisions, but are instead based on morphology (Prasada Rao et al., 1993). In a similar vein, the morphological clusters used to create the core collection (Upadhyaya et al., 2014) are also only weakly correlated with phylogeny (Supplemental Fig. S7b). Given the very low relationship between these races and the population genetics, it may be useful to create a new classification scheme within each species, either in parallel to or as a replacement for the existing race designations.

In contrast, we see a distinct correlation between the phylogeny and collection locations (Fig. 5). While only 18 accessions—all *E. colona*—have geospatial coordinates recording where they were collected, putting these on a map of India clearly shows geographic segregation. We do not have geospatial data for *E. crus-galli*, so the origin of its population substructure is unknown. Given that it is native to Japan, however, it would not be surprising to find that its population structure is also geographic in origin and probably stems from the different Japanese islands.

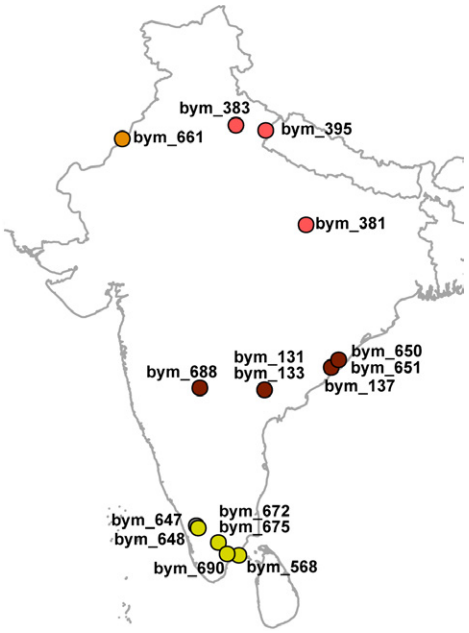


Figure 5. Geographic origin of population clusters. Plotting the collection location for the *Echinochloa colona* samples with known latitude and longitude coordinates reveals that the clusters probably originate from geographic separations. One accession is not shown because it was collected in Africa; it clusters with the southernmost (yellow) group. No *E. crus-galli* accessions have geospatial coordinates, so any correlation between their population structure and geography remains speculative.

# Discussion

We have generated a dataset of several thousand SNPs for barnyard millet, a neglected, polyploid crop important to smallholder farmers in southern and eastern Asia. As expected, these SNPs clearly separate the two primary species of barnyard millet and reveal different levels of population structure and phylogenic relationships within each species.

These data provide a jumping-off point for future breeding work in barnyard millet. Probably the most important point is simply establishing working parameters for GBS in these species, since this opens the door to many other analyses that rely on extensive genotyping (genomewide association, marker-assisted backcrossing, genomic selection, further diversity analysis, etc.). Our population structure and phylogenetic analyses can help guide breeders when selecting germplasm, especially since the existing racial designations have a poor correlation to the underlying genetics. The existence of several potential hybrids (Fig. 3) also is worth investigating. Such hybrids are supposed to be sterile (Prasada Rao et al., 1993, Upadhyaya et al., 2008), so while seed contamination is still the most likely explanation in this case, the possibility that these are actual viable hybrids (and thus could be used to bridge germplasm) is worth investigating.

While our dataset represents a very useful collection of SNPs, there are obviously many more SNPs in these samples that we did not identify. Since further sequencing depth would probably not yield many more SNPs (Fig. 1b), anyone seeking more would need to use a different methodology. The most straightforward way would be to use a different restriction enzyme when preparing the genotyping libraries. A complementary approach would be to do whole-genome shotgun sequencing on one line to assemble several thousand short contigs. This pseudo-reference genome could then be used to align the sequence reads, and in our experience even a very rough reference can dramatically increase the number of SNPs recovered. As yet another option, one could expand the

number of samples to include more diverse genotypes. Adding more samples would be especially useful because it could boost the number of accessions up to the point where one could perform a meaningful genomewide association for traits. (The current core collection is too small to do a useful association analysis, especially if done within a single species.)

An important point in our analysis is the relative ease and low cost of generating these markers. The total costs for DNA extraction, library preparation, and sequencing came to less than US$10,000. (All SNP calling was done in-house, but had it also been outsourced the price would still have been roughly within this range.) This is a fraction of what similar data would have cost 5 yr ago, and it brings it into the range where it is now feasible to perform genomewide association, genomic selection, and other techniques that previously would have been too expensive for a neglected crop like barnyard millet. We expect that these methods will soon be applied across many other orphan crops, and that this will lead to faster and better breeding practices to enhance food security, especially among smallholder famers.

## Supplemental Material

Supplemental Figures S1-S7, Supplemental File S1 (bioinformatics scripts and support files), and Supplemental File S2 (genotype data) are available online.

### Acknowledgments

This work was supported by NSF grants DBI-0820619 and IOS-1238014, ICRISAT, and the USDA–ARS. This work has been undertaken as part of the CGIAR Research Program on Dryland Cereals. Punna Ramu, Hari D. Upadhyaya, and M. Vetriventhan selected materials and prepared the samples for analysis. Jason G. Wallace performed bioinformatic analyses and had primary responsibility for writing the paper. C. Tom Hash and Edward S. Buckler provided oversight and direction. All authors had responsibility for editing the manuscript for publication.

### References

Anju, T., and S. Sarita. 2010. Suitability of foxtail millet (*Setaria italica*) and barnyard millet (*Echinochloa frumentacea*) for development of low glycemic index biscuits. Malaysian J. Nutr. 16:361–368.

Bennett, M.D., P. Bhandol, and I.J. Leitch. 2000. Nuclear DNA amounts in angiosperms and their modern uses—807 new estimates. Ann. Bot. (Lond.) 86:859–909. doi:10.1006/anbo.2000.1253

Bennett, M.D., I.J. Leitch, and L. Hanson. 1998. DNA amounts in two samples of angiosperm weeds. Ann. Bot. (Lond.) 82:121–134. doi:10.1006/anbo.1998.0785

Bradbury, P.J., Z. Zhang, D.E. Kroon, T.M. Casstevens, Y. Ramdoss, and E.S. Buckler. 2007. TASSEL: Software for association mapping of complex traits in diverse samples. Bioinformatics 23:2633–2635. doi:10.1093/bioinformatics/btm308

Danecek, P., A. Auton, and G. Abecasis. 2011. The variant call format and VCFtools. Bioinformatics 27:2156–2158. doi:10.1093/bioinformatics/btr330

de Wet, J.M.J., E.K. Prasada Rao, M.H. Mengesha, and D.E. Brink. 1983. Domestication of sawa millet (*Echinochloa colona*). Econ. Bot. 37:283–291. doi:10.1007/BF02858883

Dwivedi, S., H. Upadhyaya, S. Senthilvel, C. Hash, K. Fukunaga, X. Diao, D. Santra, D. Baltensperger, and M. Prasad. 2012. Millets: Genetic and genomic resources. Plant Breed. Rev. 35:247–375.

Elshire, R.J., J.C. Glaubitz, Q. Sun, J. Poland, K. Kawamoto, E.S. Buckler, and S.E. Mitchell. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS ONE 6(5):E19379. doi:10.1371/journal.pone.0019379

Felsenstein, J. 1989. PHYLIP—Phylogeny inference package (Version 3.2). Cladistics 5:164–166.

Glaubitz, J., T. Casstevens, and F. Lu. 2014. TASSEL-GBS: A high capacity genotyping by sequencing analysis pipeline. PLoS ONE 9(2):E90346. doi:10.1371/journal.pone.0090346

Godfray, H., J. Beddington, and I. Crute. 2010. Food security: The challenge of feeding 9 billion people. Science 327:812–818. doi:10.1126/science.1185383

Huson, D., and D. Bryant. 2006. Application of phylogenetic networks in evolutionary studies. Mol. Biol. Evol. 23:254–267. doi:10.1093/molbev/msj030

Lu, F., A. Lipka, J. Glaubitz, and R. Elshire. 2013. Switchgrass genomic diversity, ploidy, and evolution: Novel insights from a network-based SNP discovery protocol. PLoS Genet. 9(1):E1003215. doi:10.1371/journal.pgen.1003215

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. Rep. 21:459–460. doi:10.1007/BF02772596

Naylor, R.L., W.P. Falcon, R.M. Goodman, M.M. Jahn, T. Sengooba, H. Tefera, and R.J. Nelson. 2004. Biotechnology in the developing world: A case for increased investments in orphan crops. Food Policy 29:15–44. doi:10.1016/j.foodpol.2004.01.002

Nelson, R.J., R.L. Naylor, and M.M. Jahn. 2004. The role of genomics research in improvement of "orphan" crops. Crop Sci. 44:1901–1904. doi:10.2135/cropsci2004.1901

Potvin, C. 1991. Temperature-induced variation in reproductive success: Field and control experiments with the C4 grass *Echinochloa crus-galli*. Can. J. Bot. 69:1577–1582. doi:10.1139/b91-201

Prasada Rao, K.E. J.M.J de Wet, V. Gopal Reddy, and M.H. Mengesha. 1993. Diversity in the small millets collection at ICRISAT. In: K.W. Riley, S.C. Gupta, A. Seetharam, and J.N. Mushonga, editors, Advances in small millets. Oxford and IBM Publ., New Delhi, India. p. 331–346.

Pritchard, J., M. Stephens, and P. Donnelly. 2000. Inference of population structure using multilocus genotype data. Genetics 155:945–959.

R Core Team. 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Raj, A., M. Stephens, and J.K. Pritchard. 2014. fastSTRUCTURE: Variational inference of population structure in large SNP datasets. Genetics 197:573–589. doi:10.1534/genetics.114.164350

Saleh, A.S.M., Q. Zhang, J. Chen, and Q. Shen. 2013. Millet grains: nutritional quality, processing, and potential health benefits. Compr. Rev. Food Sci. Food Saf. 12:281–295. doi:10.1111/1541-4337.12012

Sharma, A., S. Sood, and R.K. Khulbe. 2013. Millets—Food for the future. Biotech Today 3:52. doi:10.5958/j.2322-0996.3.1.010

Surekha, N., R.S. Naik, S. Mythri, and R. Devi. 2013. Barnyard millet (*Echinochloa frumentacea* Link) Cookies: Development, value addition, consumer acceptability, nutritional, and shelf life evaluation. IOSR J. Environ. Sci. Toxicol. Food Technol. 7:1–10.

Ugare, R. 2008. Health benefits, storage quality and value addition of barnyard millet (*Echinochloa frumentacaea* Link). Master's diss., University of Agricultural Sciences, Dharwad, India.

Upadhyaya, H.D., C.L.L. Gowda, V.G. Reddy, and S. Singh. 2008. Diversity of small millets germplasm in genebank at ICRISAT. In: J. Smartt and N. Haq, editors, 5th International Symposium on New Crops and Uses: Their role in a rapidly changing world, 3-4 Sept. 2007, University of Southampton, Southampton, UK. p. 173-185.

Upadhyaya, H., S. Dwivedi, and S. Singh. 2014. Forming core collections in barnyard, kodo, and little millets using morpho-agronomic descriptors. Crop Sci. 54:1–10. doi:10.2135/cropsci2012.12.0710

van der Walt, S., S.C. Colbert, and G. Varoquaux. 2011. The NumPy array: A structure for efficient numerical computation. Comput. Sci. Eng. 13:22–30. doi:10.1109/MCSE.2011.37

Varshney, R.K., J.C. Glaszmann, H. Leung, and J.M. Ribaut. 2010. More genomic resources for less-studied crops. Trends Biotechnol. 28:452–460. doi:10.1016/j.tibtech.2010.06.007

Vijayakumar, T.P., J.B. Mohankumar, and T. Srinivasan. 2010. Quality evaluation of noodles from millet flour blend incorporated composite flour. J. Sci. Ind. Res. (India) 69:48–54.

Wanous, M.K. 1990. Origin, taxonomy and ploidy of the millets and minor cereals. Plant Var. Seeds 3:99–112.