"Population genomic and genome-wide association studies of agroclimatic traits in sorghum"

Geoffrey P. Morris (a,1,2), Punna Ramu (b,1), Santosh P. Deshpande (b), C. Thomas Hash (c), Trushar Shah(b), Hari D. Upadhyaya (b), Oscar Riera-Lizarazu (b), Patrick J. Brown (d), Charlotte B. Acharya (e), Sharon E. Mitchell (e), James Harriman (e), Jeffrey C. Glaubitz (e), Edward S. Buckler (e,f,g), and Stephen Kresovich (a)

(a) Department of Biological Sciences, University of South Carolina, Columbia, SC 29208;
(b)International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Hyderabad 502 324, Andhra Pradesh, India;
(c)ICRISAT-Sadoré, BP 12404 Niamey, Niger;
(d)Department of Crop Sciences, University of Illinois, Urbana, IL 61801;
(e)Institute for Genomic Diversity and
(f)Department of Plant Breeding and Genetics, Cornell University, Ithaca, NY 14853; and
(g)Agricultural Research Service, Department of Agriculture, Ithaca, NY 14853

Proceedings of the National Academy of Science of the United States of America (PNAS)

http://dx.doi.org/10.1073/pnas.1215985110

PNAS December 24, 2012 201215985

Target journal: PNAS

Classification: BIOLOGICAL SCIENCES: Agricultural Sciences

Population genomic and genome-wide association studies of agroclimatic traits in sorghum

Short title: Population genomics and GWAS in sorghum

Geoffrey P. Morris[1*†], Punna Ramu[2†], Santosh P. Deshpande[2], C. Thomas Hash[3], Trushar Shah[2], Hari D. Upadhyaya[2], Oscar Riera-Lizarazu[2], Charlotte B. Acharya[4], Sharon E. Mitchell[4], James Harriman[4], Jeffrey C. Glaubitz[4], Edward S. Buckler[4,5,6], Patrick J. Brown[7], and Stephen Kresovich[1]

[1] Department of Biological Sciences, University of South Carolina, Columbia, SC 29201, USA.

[2] International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Patancheru PO, Hyderabad 502 324, Andhra Pradesh, India.

[3] ICRISAT - Sadoré, BP 12404 Niamey, Niger.

[4] Institute for Genomic Diversity, Cornell University, Ithaca, NY 14850, USA.

[5] Department of Plant Breeding and Genetics, Cornell University, Ithaca, NY 14850, USA.

[6] United States Department of Agriculture, Agricultural Research Service, Ithaca, NY 14850, USA.

[7] Department of Crop Sciences, University of Illinois, Urbana, IL 61801, USA.

[†] G.P.M. and P.R. contributed equally to this work.

* To whom correspondence may be addressed: morrisgp@mailbox.sc.edu

Abstract

Accelerating crop improvement in sorghum, a staple food for people in semi-arid regions across the developing world, is key to ensuring global food security in the context of climate change. To facilitate gene discovery and molecular breeding in sorghum, we have characterized ~265,000 single nucleotide polymorphisms (SNPs) in 1,000 worldwide accessions that have adapted to diverse agroclimatic conditions. Using this genome-wide SNP map, we have characterized population structure with respect to geographic origin and morphological type, and identified patterns of ancient crop diffusion to diverse agroclimatic regions across Africa and Asia. To better understand the genomic patterns of diversification in sorghum, we quantified variation in nucleotide diversity, linkage disequilibrium, and recombination rates across the genome. Analyzing nucleotide diversity in landraces, we find evidence of selective sweeps around starch metabolism genes, while in landrace-derived introgression lines we find introgressions around known height and maturity loci. To identify additional loci underlying variation in major agroclimatic traits, we performed genome-wide association studies (GWAS) on plant height components and inflorescence architecture. GWAS maps several classical loci for plant height, as well as a priori and novel candidate genes for inflorescence architecture. Finally, we trace the independent spread of multiple haplotypes carrying alleles for short stature or long inflorescence branches. This genome-wide map of SNP variation in sorghum provides a basis for crop improvement through marker-assisted breeding and genomic selection.

Introduction

Agricultural production and food security in the developing world face numerous threats, particularly in semi-arid regions, which are acutely vulnerable to climate change {Lobell et al. 2008}. Sorghum [*Sorghum bicolor* (L.) Moench.] is an important crop species for farmers in semi-arid and arid regions because it can sustain high yields where precipitation is low or erratic. Thus, sorghum has become the major cereal crop in semi-arid regions and a dietary staple for over 500 million people, predominantly in sub-saharan Africa and south Asia {NRC 1996}. Worldwide, sorghum is grown for food (grain and syrup), animal feed, fiber, and fuel in both subsistence and commercial agriculture systems. As rising temperatures and reduced precipitation due to climate change make some areas unsuitable for maize and rice production, the importance of drought-tolerant crops like sorghum is likely to increase {Lobell et al. 2008}. Current breeding priorities in sorghum seek to mitigate climate-dependent stressors, both abiotic [e.g. drought {Vadez et al. 2011} and acid soils {Zheng 2010}], and biotic [e.g. insect pests {Sharma et al. 2005} and fungal diseases {Sharma et al. 2010}]. In order to meet the projected doubling of global food demand over the next few decades in the context of global change the pace of crop improvement must be accelerated {Foley et al. 2011}.

Sorghum has a wide range of adaptation, and traditional varieties from across Africa and Asia provide a rich source of morphological and physiological traits for crop improvement {Deu et al. 2006; Casa et al. 2008; Upadhyaya et al. 2009}. The primary domestication of sorghum occurred near present-day Sudan around 10,000 years ago, and diffusion occurred to diverse climates across Africa, India, the Middle East, and east Asia between 8,000 and 1,500 years ago {Harlan and Stemler 1976; Kimber 2000; Dillon et al. 2007}. Because of this ancient origin and diffusion, adaptation to local climates and cultural practices is reflected in morphological and physiological variation among and within the five major types (races) of domesticated sorghum. For instance, in parts of West Africa where rainy periods are long and erratic, open panicle guinea types are preferred in order to reduce grain mold and insect damage. Conversely, in parts of South and East Africa where rainy seasons are relatively short and predictable, dense panicle kafir and durra types are preferred in order to increase grain yield per plant {NRC, 1996}. Further natural and human selection has occurred in the United States over the past ~150 years as temperate and tropical germplasm from Africa and Asia has been adapted for use in combine-harvested commercial agriculture {Vinall et al. 1936; Quinby 1975; Klein et al. 2008}.

Genomic analysis of diverse populations is increasingly being used to uncover the genetic basis of complex traits, including agroclimatic traits of crop species. Genome-wide SNP scans of population genetic parameters have been used in domesticated species to identify loci under selection {Jiao et al. 2012} and dissect quantitative traits {Vaysse et al. 2011; Kijas et al. 2012}. In addition, genome-wide association studies (GWAS) have been used to elucidate the genetic basis of agronomic traits in rice {Huang et al. 2010; Huang et al. 2012; Jiao et al. 2012} and maize {Tian et al. 2011; Kump et al. 2011}. Nucleotide diversity scans {Casa et al. 2006; Bouchet et al. 2012} and association studies {Casa et al. 2008; Figueiredo et al. 2010; Upadhyaya et al. 2012} have been carried out in sorghum, but the resolution and sensitivity of these studies has been limited by the small number of markers {Bouchet et al. 2012}. Thus, compared to maize and rice, less is known about the genetic basis of agronomic traits in sorghum. Among the four classical dwarfing loci that have been studied in sorghum for over 70 years {Quinby and Karper 1945}, only one has been cloned (*Dw*3/SbPGP1) {Multani et al. 2003}. Recently, it has become feasible to genotype thousands of markers rapidly and at low cost through the application of barcode multiplexing and high-throughput sequencing {Baird et al. 2008; Elshire et al. 2011}. To better understand the diversity of sorghum, facilitate the genetic dissection of agroclimatic traits, and accelerate marker-assisted breeding, we characterized 1,000 sorghum accessions at 265,487 SNPs using genotyping-by-sequencing. Here, we describe a genome-wide map of SNP variation, trace patterns of crop diffusion to diverse agroclimatic regions, and use GWAS to identify genes underlying natural variation in agroclimatic traits.

Results and discussion

A genome-wide map of SNP variation

To represent the genetic, geographic, and morphological diversity of sorghum, we used 1,000 accessions from the world germplasm collections, combining three previously-defined sorghum diversity panels {Deu et al. 2006; Casa et al. 2008; Upadhyaya et al. 2009}. The majority of these accessions consist of source-identified landraces or traditional cultivars from across Africa and Asia (Fig. 1a). Of these, ~200 are landrace-derived sorghum conversion lines, in which alleles for short stature and early maturity were introgressed into tropical landraces to facilitate the use of tropical germplasm in temperate breeding programs {Stephens et al. 1967}. The remainder consists of wild/weedy relatives or elite lines and breeding materials, many of which have unknown geographic origin and/or mixed ancestry. For each accession, we constructed *ApeKI* reduced representation libraries using the GBS method {Elshire et al. 2011} and generated a total ~21 Gbp of sequence on the Illumina Genome Analyzer*IIx*/HiSeq. In total, 6.13 million unique 64 bp tags were identified across all sorghum accessions. Of these tags, 85% aligned to the reference sorghum genome and containing a total of 384,561 putative SNPs. After filtering for local LD and coverage (10% of taxa), 265,487 SNPs were retained for use in present study, at an average density of one SNP per 2.7 kbp. Of 27,412 annotated genes in the reference sorghum genome, 72% were tagged by a SNP within the gene and 99% were tagged by a SNP within 10 kb. Importantly, this genome-wide map of SNP variation is of sufficient resolution for GWAS in sorghum, given >100,000 SNPs is estimated to be required {Bouchet et al. 2012}. Additionally, due to simultaneous SNP discovery and genotyping, this sequencing-based SNP map will have little ascertainment bias and greater power for mapping studies {Myles et al. 2009}.

Linkage disequilibrium and recombination rates

Characterizing patterns of linkage disequilibrium (LD) is critical for the design of association studies {Kim et al. 2007; Mather et al. 2007}, interpretation of association peaks {Huang et al. 2010}, and the transfer of alleles in marker-assisted selection {Casa et al. 2006}. To characterize the mapping resolution for genome scans and GWAS, we quantified the average extent of LD decay and localized patterns of LD for each chromosome (Fig. S2). On average, LD decays to 50% of its initial value by 1 kb and to background levels (r^2 < 0.1) within 150 kb. These estimates are higher than previous published LD decay values in sorghum of 15-20 kb {Hamblin et al. 2005} and 50-100 kb {Bouchet et al. 2012}. This may be attributed to low genome coverage of markers and fewer genotypes in previous studies. Since sorghum is a predominantly selfing species, but readily outcrosses, we expect a moderate LD decay as compared to obligate selfers or obligate out-crossers. Accordingly, the extent of LD is greater in sorghum (150 kb) as compared to maize (2 kb) {Yan et al. 2009}, which is an outcrosser. LD decay in sorghum is almost equivalent to LD decay in rice {Mather et al. 2007}, another self-pollinated crop, but much greater than in *Arabidopsis* (10 kb) {Kim et al. 2007}. Sliding window (1 Mb) estimates of pairwise LD show that telomeric regions have lower LD than centromeric regions (Fig. S2b). This suggests more recombination occurs in telomeric regions. This is supported by estimation of historical recombination rates (Fig. 2c). The average recombination rate across the genome is 1.4 ρ/kb, with considerable variation across chromosome (Fig. 2c). The average recombination rate in sorghum (1.4/kb) is intermediate relative to recent estimates in plants, such as *Arabidopsis* (0.8/kb) {Kim et al. 2007} and *Medicago* (2.6/kb) {Paape et al. 2012}. Based on these results we expect mapping resolution to range widely across the genome, from single-gene resolution in some telomeric regions to Mb-level resolution near the centromeres.

Population structure and geographic differentiation

Next, we used the genome-wide map of SNP variation to characterize genetic relatedness among the 1,000 sorghum accessions representing worldwide diversity. The resulting neighbor joining trees (Fig. 1b) and Bayesian clustering analysis (Fig. S1) show population structuring along both morphological type and geographic origin, as has previously been observed {Brown et al. 2011; Bouchet et al. 2012}. Of the five morphological types, the kafir sorghums that predominate in southern Africa show the strongest pattern of population subdivision relative to other races (Fig. 1b; Fig. S1). Durra type sorghums, which are found in warm semi-arid or warm desert climates of the Horn of Africa, Sahel, Arabian peninsula, and west central India, form a distinct cluster and further cluster according to geographic origin. As a whole, bicolor types are not notably clustered. Bicolor types from China (known as kaoling) do form a distinct subgroup, which shows genetic similarity to the durra types, particularly those from Yemen. Caudatum types, which are primarily found in tropical savanna climates of central Africa, are very diverse and show only modest clustering according to geographic distribution. Finally, guinea types, which are widely distributed in tropical savanna climates, show five distinct subgroups, four of which cluster according their geographic origin (far-west Africa, west-central Africa, southern Africa, and India). A fifth guinea subgroup, which includes guinea margaritiferum types, forms a separate cluster along with wild genotypes from western Africa (Fig. 1b) and may represent an independent domestication {Deu et al. 2006}. In the neighboring joining analysis but not the Bayesian clustering, Indian guinea types cluster with durra types, likely due to admixture with sympatric Indian durra populations (Fig. S1).

The structure of sorghum populations provides insight into historical processes of crop diffusion within and across agroclimatic zones of Africa and Asia. Diffusion across agroclimatic zones is expected to be rare relative to diffusion within agroclimatic zones {Diamond 2002}. Indeed, the patterns of relatedness among sorghum populations suggest that agroclimatic constraints have been at least as important as geographic isolation in shaping the diffusion process. Among the four phylogenetically-supported sorghum types (kafir, durra, guinea, and caudatum), there is least population structure among caudatum types, which range primarily in the ancestral region of domestication or adjacent areas with similar climate (Fig. 1a). The one geographically-structured subpopulation of caudatum is the latitudinal-diffused subpopulation from highland areas in east Africa. While durra types diffused widely across the Africa and Asia, they are restricted to semi-arid and desert climates (Fig. 1). This includes kaoling sorghums, which are likely derived from durra populations of the Middle East, but are found in cold semi-arid regions of northern China (Fig. 1; Fig. S1). Consistent with overland diffusion of durra sorghums along semi-arid/desert climates in central Asia {Kimber 2000}, the single Afghani landrace in the current data set (*Nai-Shaker*, collected in 1954 at Mazar-i-Sharif) is intermediate between the Yemeni and Chinese populations (Fig. S1). Similarly, while guinea types have diffused over long distances, from western Africa to southeastern Africa and eastern India, they remain restricted to tropical savanna climates. Interestingly, Bayesian clustering analysis suggest that the temperate/subtropical-adapted kafir type is derived from (or at least shares ancestry with) guinea types of east African populations (k=2 through k=6; Fig. S1). In this case, the kafir type may represent major phenotypic divergence and genetic bottlenecking resulting from a shift to a contrasting agroclimatic zone.

Genomic patterns of nucleotide variation

To investigate the genomic signatures of domestication and diversification in sorghum we quantified genome-wide nucleotide variation across sorghum landraces. Overall, average nucleotide diversity (π) was 0.00037/kb, $\theta = 0.00017$/kb and *Tajima's D* value of 3.6. A scan of expected heterozygosity, H_e, values across the genome revealed many Mb-scale regions of low heterozygosity including a ~40 Mb region of reduced nucleotide variation on around the centromere of chromosome 7 (Fig. 2a). Of six starch-related genes previously studied {Whitt et al. 2002; de Alencar Figueiredo et al. 2008, de Alencar Figueiredo et al. 2010}, two are found at regions with notably low heterozygosity (Fig. 2a). The starch biosynthesis enzyme *brittle endosperm 2* (*bt2*) gene, which has previously been shown to be a likely domestication locus in maize {Whitt et al. 2002} and sorghum {de Alencar Figueiredo et al. 2008, de Alencar Figueiredo et al. 2010}. The large size and extensive LD (Fig. S2) of the low heterozygosity region on chromosome 7 may be due to low recombination rates in this peri-centromeric region (Fig. 2c), or additional loci under selection. Another a priori domestication candidate from starch metabolic pathways, transcription factor *opaque*2 {de Alencar Figueiredo et al. 2008}, is found at the base of a low heterozygosity region on chromosome 2. The large footprints of selection we observe here (up to several Mb) are consistent with the predominance of inbreeding in sorghum. Selective sweeps in out-crossing maize left smaller footprints (< 100 kb) {Wright et al. 2005;

Tian et al. 2009} than in self-pollinating rice (250 kb to 1 Mb) {Olsen et al. 2006; Sweeney et al. 2007}.

In sorghum conversion lines that carry introgressions of early maturity and short stature alleles, we also observed major reduction in heterozygosity in several genomic regions (Fig. 2b). These regions colocalize to previously-mapped height [*Dw2* {Lin et al. 1995}, *Dw3* {Multani et al. 2003}, and *Dw1*/SbHT9.1 {Brown et al. 2008}] and [*Ma1*/SbPRR37 {Murphy et al. 2011}] maturity loci that are recessive in the introgression donor BTx406. In contrast, another classical maturity locus, Ma3/phyB {Childs et al. 1997}, which is wild-type in BTx406 and therefore was not under selection during the conversion process, shows no such reduction in heterozygosity. On chromosome 6, the low heterozygosity region extends from about 6.6 Mb to the *Ma1*/*Dw2* locus at 42 Mb suggesting that the fourth classical dwarfing loci, *dw4*, may be localized here (SI Appendix, Note). As was seen in the landraces, we find that large LD blocks result when selection occurs in low recombination regions around the centromere (Fig. 2b-c).

Genome-wide association studies

The genome-wide map of SNP variation we have generated permits the dissection of complex traits in sorghum using GWAS. To elucidate the genetic basis of plant height in sorghum, we determined associations between SNPs and plant height components using data from ~300 lines in the Sorghum Association Panel (SAP; Fig. 3; Fig. S2) {Brown et al. 2008}. Plant height is important component for many agroclimatic traits such as competitive growth with weeds, resistance to lodging, and, in the case of temperate-adapted grain sorghums, the efficiency of combine harvest. Since this panel incorporates a large fraction of sorghum conversion lines with introgressions of dwarfing (*Dw*) alleles, we know that much of the variation for height in this panel has a common genetic basis. We identified SNPs associated with total plant height, and two height components; pre-flag height, which quantifies elongation in the lower portion of the stem, and flag-to-apex length, which quantifies elongation in the upper portion of the stem. The *Dw*3 (SbPGP1) gene, the only dwarfism gene that has been cloned in sorghum {Multani et al. 2003}, provides a positive control for GWAS. As it is known that the reduced height of *dw*3 mutants is due to reduced elongation of lower internodes, we considered pre-flag leaf height as a measure lower internode elongation {Brown et al. 2008}. The third most significant association peak for pre-flag height is found at the *dw3* locus (within 12 kb for GLM and 22 kb for CMLM). We also refined the mapping location of *dw1* and *dw2*, and determined the likely location of *dw4* (SI Appendix, Note).

Since the SAP includes a large fraction of converted lines, with large introgressions around height and maturity loci, the previous analysis does not reflect a typical GWAS case. To validate the broader applicability of GWAS in sorghum, we also sought to dissect a trait that was not a target of selection in the sorghum conversion program. Inflorescence architecture is a major agroclimatic trait that, in part, defines the major morphological types in sorghum. Moreover, since the genetic basis of inflorescence architecture is well-studied in maize, rice, and Arabidopsis, there are many a priori candidate genes that can be considered to evaluate the mapping approach. Indeed, several of the significant association peaks for inflorescence branch length were located in or near a priori candidate genes for inflorescence architecture, which are homologous to known maize, rice, or Arabidopsis floral regulators (Table 1; Fig. 3). For instance, two peaks are in, and one is near (47 kb), C2H2 zinc finger transcription factors homologous to the classical maize floral development gene INDETERMINATE 1 (ID1) (Fig. 3; Table 1) {Colasanti et al. 1998}. Another association peak was found in a sorghum ortholog of Arabidopsis UNUSUAL FLORAL ORGAN (UFO) and rice ABERRANT PANICLE ORGANIZATION 1 (APO1) {Ikeda et al. 2005}. In rice, *apo1* mutants exhibit small panicles and fewer branches {Ikeda et al. 2005}. The top association peak for branch length (whether or not population structure is controlled) is a SNP found in ID1 homolog Sb02g019110. The minor allele at this SNP is restricted to the three broomcorn varieties in the panel, which display the most extreme branch length phenotypes, with inflorescence branches over 0.5 m in length or >8 standard deviations above the species-wide mean.

In sorghum, strong population structure among the morphological types presents a challenge for mapping the genetic basis of the inflorescence architecture and other population-associated traits. While statistical controls for population structure have proven effective here, better control of population structure will be achieved using regional mapping {Brachi et al. 2011; Horton et al. 2012} or nested-association mapping (NAM) lines {Tian et al. 2011}, which are under development in sorghum {Jordan et al. 2011}. Due to the use of introgression lines, the SAP captures some aspects of the NAM approach. The introgressions increase mapping power for height and maturity loci by increasing the frequency of rare alleles {Brown et al. 2008}, and reduce the confounding effects of maturity differences in diverse germplasm. However, since the introgression originate from same donor line (BTx406), the large blocks of linked non-causative variation reduces the resolution of the association analysis (Fig. 2 A). Also, the low diversity around height and maturity loci on chromosomes 6, 7, and 9 may prevent the mapping of other QTLs which co-localize to these regions, especially on chromosome 6 most of the chromosome has been introgressed in SC lines (Fig. 2 B). In some cases, therefore, mapping populations without converted lines will be more

Geographic distribution of haplotypes

To gain further insight into the origin and spread of haplotypes linked to agroclimatic traits, we characterized the geographic distribution of QTL SNPs in 330 source-identified landraces that are independent of the lines used for GWAS. The three major height QTL identified by genome scan and GWAS, *dwarf*2, *dwarf*3, and *SbHt*9.1, have distinct allelic distributions across Africa and Asia (Figure 3, e-g). One of the alleles at the SNP (Fig. 3b) closest to the putative *SbHt*9.1 causative gene (GA2-oxidase) {Wang et al. 2011} is found at high frequency (>90%) in East African durra, Indian durra, and Chinese accessions and low frequency (<10%) in all other accessions. Likewise, one allele at the *Dw*2 QTL peak is found common in northeast Africa and Asia and rare elsewhere. This is consistent with a common genetic basis for semi-dwarfism in east African and Asian sorghums, conferred by at least two causative polymorphisms, and originating from ancestral East African durra populations. Did the mutations that underlie classical dwarfing alleles arise *de novo* in the early U.S. grain sorghums, as is suggested in classical breeding literature {Vinall et al. 1936; Quinby 1975}, or were they recruited from standing variation present in African or Asian landraces? The haplotypes associated with dwarfism at *dw1*/SbHt9.1, *dw2*, and *dw3* are indeed widely distributed in among African and Asian landraces (Fig. 3), but these could represent ancestral haplotypes on which new dwarfing mutations occurred. The classical literature, however, confirms that dwarfing alleles were already present in African landraces at the time that dwarf alleles were being adopted in U.S. grain sorghums. For instance, dwarf durra varieties collected near Khartoum, Sudan c. 1920 carry the *dw4* allele (*Gahan dura*) or both *dw1* and *dw4* (*hegari*) {USDA 1928; Quinby and Karper 1954}. Taken together, the evidence suggests that the dwarfing alleles were likely selected from standing variation in durra (*dw1*, *dw2, dw4*) and kafir (*dw3*) landraces.

The geographic distribution of alleles at inflorescence branch length QTL also reveals evidence of the independent spread of multiple alleles controlling branch length (Fig. 3 B - E). In general, the minor allele associated with longer branches is found at high frequency in west African guinea population, and in a number of cases it also found in other geographically distant populations (Fig. 3 B - D). Interestingly, none of the alleles at top branch length association peaks were restricted to durra accessions (Fig. 4), suggesting that we were able to identify QTL for the long-branch phenotype in guinea types, but not QTL for the short-branch phenotype in durra types. This may be due to the fact that there is stronger population structuring of durra populations, as compared to guinea, which can confound mapping of traits {Brachi et al. 2011}. Given the statistical correction for population structure, we did not map QTL underlying the branch length differences among major morphological types, rather we mapped QTL for branch length segregating within the morphological types, which are globally rare but locally common {Jiao et al. 2012} (Table 1; Fig 4 B - D).

Conclusion

A better understanding of genetic diversity in sorghum will support *in situ* conservation efforts, enhance the use of germplasm collections, and guide ongoing collection efforts {Ramanatha Rao and Hodgkin 2002}. The genome-wide map of SNP variation will accelerate molecular breeding, by expanding the diversity of germplasm accessible to crop improvement programs and increasing the resolution of GWAS and marker-assisted selection. A genome-wide SNP map will also allow genomic selection, a form of marker-assisted selection based on whole-genome prediction that can accelerate gains during breeding {Morrell et al. 2012}. By facilitating crop improvement in locally-adapted and locally-improved cultivars, genomic analysis of diverse crop germplasm can play an important role in supporting sustainable agriculture in Africa, Asia, and semi-arid regions worldwide. We will need to combine a better understanding of the genetic basis of agroclimatic traits with genomic-accelerated breeding in order to meet increasing global food requirements in context of climate change.

Materials and Methods

Plant materials

Diverse sorghum germplasm lines from world-wide collections which defined to serve as community resources for allele mining and association studies were used, incorporates three previously-defined diversity panels: (1) sorghum association panel (SAP) {Casa et al. 2008}, (2) sorghum mini core collection (MCC) {Upadhyaya et al. 2009} and (3) the sorghum reference set (RS) (http://www.icrisat.org/what-we-do/crops/sorghum/Sorghum_Reference.htm). We were able to obtain appropriate plant material for 1,000 accessions (SI Table 1). The SAP was obtained from GRIN (http://www.ars-grin.gov). Country of origin, latitude, and longitude for source-identified accessions were obtained from the SINGER crop germplasm database (http://singer.cgiar.org).

Genotyping-by-sequencing

DNA from MCC and RS lines (5-6 plants per accession) was isolated using the CTAB protocol {Mace et al. 2003) from 12-days old seedlings. SAP lines DNA was isolated using DNeasy Plant Mini Kit (QIAGEN). Genotyping was carried out using multiplexed (96- or 384-plex) genotyping-by-sequencing {Elshire et al. 2011} with ApeKI restriction enzyme (recognition site: G|CWCG) on an Illumina HiSeq/Genome Analyzer *IIx*. Sequences were mapped to the BTx623 sorghum reference genome {Paterson et al. 2009} using BWA tool {Li and Durbin 2009} and SNPs were called using the TASSEL 3.0 GBS pipeline (http://www.tassel.svn.sourceforge.net/svnroot/tassel). Tags, unique sequence of 64 bp length that included a leading 4 bp C[T/A]GC signature from the cut site, were identified and tags with at least 10X coverage were retained. Missing data were imputed with NPUTE {Roberts et al. 2007}.

Population genetic analysis

Hierarchical population structure was estimated using the ADMIXTURE program {Alexander et al. 2009}, a model-based estimation of ancestry in unrelated individuals using maximum-likelihood method. The neighboring joining trees was built and heterozygosity calculated using the *ape* package in R {Paradis 2010}. Pairwise LD was calculated (r^2) separately for each chromosome as a full matrix and was plotted as distance between adjacent SNPs vs r^2 using TASSEL 3.0 {Bradbury et al. 2007}. To avoid confounding effects of shared introgressions from conversion, the SAP lines were not included in the LD analysis. LD decay was calculated where r^2 drops down less than a threshold level (r^2<0.1). Mean r^2 values were used to calculate LD in 100 kb sliding window-based approach. Recombination rates were inferred using Bayesian reversible-jump MCMC under the cross-over model as implemented in *rhomap* program of LDhat {Auton and McVean 2007} with 1 million iterations and 1 million burn-ins and default parameters. Rates were estimated separately for each of subgroup identified by ADMIXTURE at K=10, excluding introgression lines from the sorghum conversion program.

Genome-wide association studies (GWAS)

Previously published phenotypes for plant height components and inflorescence branch length for the SAP were used for GWAS {Brown et al. 2008}. GWAS was carried out in Genomic Association and Prediction Integrated Tool (GAPIT) {Lipka et al. 2012} using a (i) general linear model (GLM) or (ii) compressed mixed linear model model (CMLM) with population parameters previously determined (P3D) {Zhang et al. 2010} to control for population structure. Bonferroni correction factor was used to identify significant associations for the traits under study.

Figures and tables

Figure 1 [Sorghum diversity]

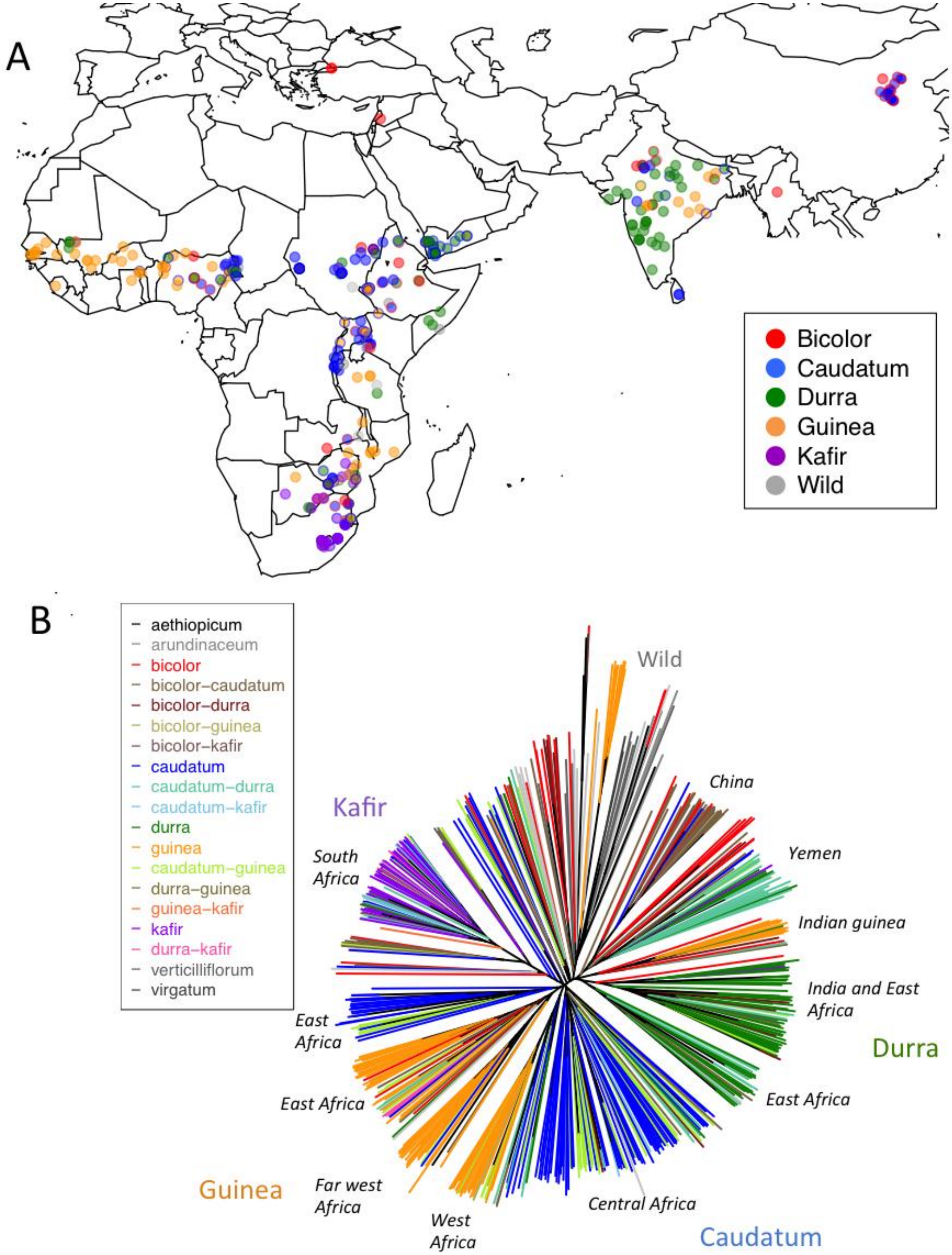

A
Bicolor
Caudatum
Durra
Guinea
Kafir
Wild
B
aethiopicum
arundinaceum
bicolor
bicolor–caudatum
bicolor–durra
bicolor–guinea
bicolor–kafir
caudatum
caudatum–durra
caudatum–kafir
durra
guinea
caudatum–guinea
durra–guinea
guinea–kafir
kafir
durra–kafir
verticilliflorum
virgatum
Wild
China
Yemen
Indian guinea
India and East Africa
Durra
East Africa
Central Africa
Caudatum
West Africa
Far west Africa
Guinea
East Africa
East Africa
South Africa
Kafir

Figure 1: **Germplasm origin and genetic relationships among worldwide sorghum accessions.** (a) Origin for 330 of 1,000 worldwide accessions for which source location is known, color-coded by morphological type. Intermediate morphological types are indicated by two-color symbols. (b) Genetic relatedness among 556 accessions of known morphological type, assessed by neighboring joining method. Worldwide sorghum populations show structuring by morphological type and geographic origin.

Figure 2 [Genome scan]

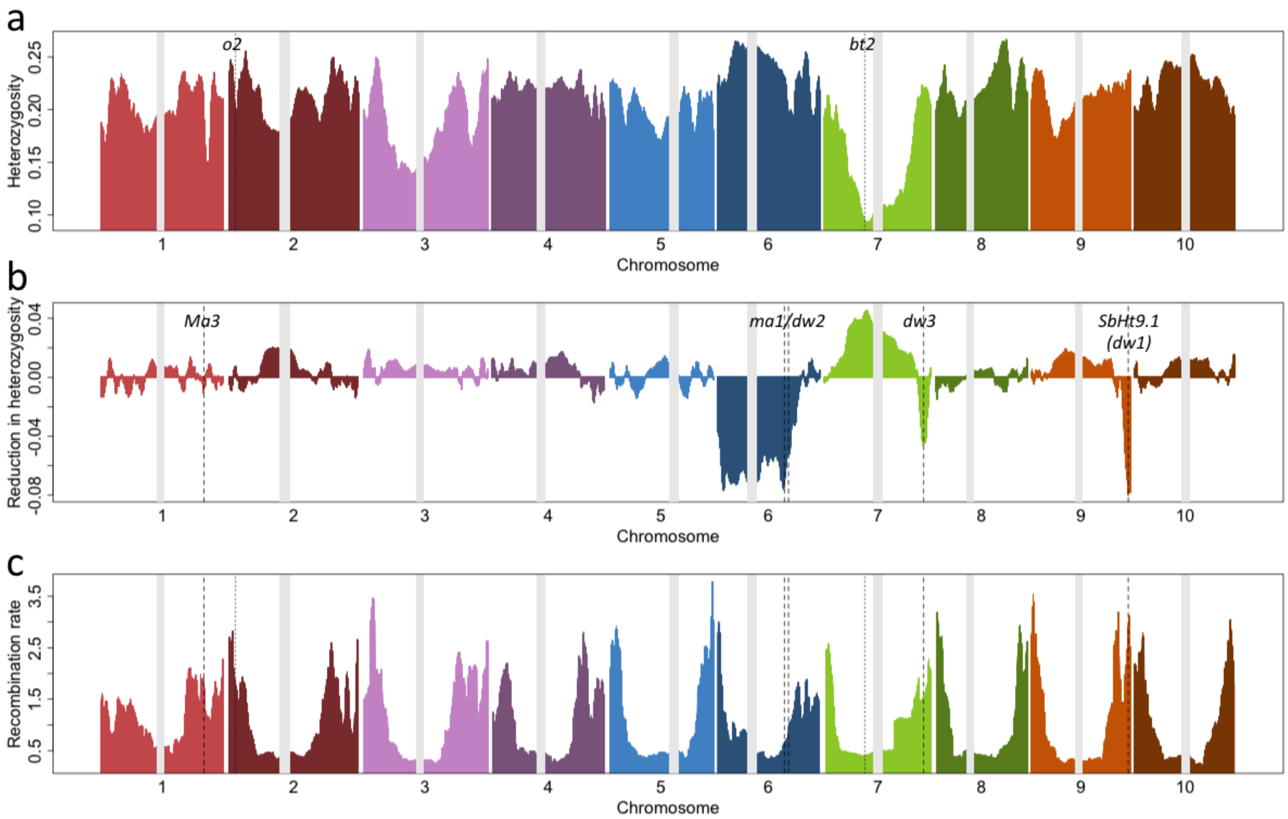


Figure 2: Genome-wide patterns of SNP variation. (a) Genome-wide variation of expected heterozygosity for sorghum landraces, smoothed with a 2000 SNP moving average. The location of the centromeres are noted by the gray bars. Two orthologs of starch-related domestication loci from maize (opaque2 and brittle endosperm2) co-localize with regions of reduced diversity. Panel (b) indicates the reduction in heterozygosity due to introgressions of short stature and early maturity alleles, with known dwarfing (*dw*) and maturity (*ma*) loci noted. (c) Genome-wide variation in inferred recombination rates averaged over ten subpopulations. Wider regions of reduced heterozygosity occur in regions near centromeres with low recombination rates .

Figure 3 [Height GWAS]

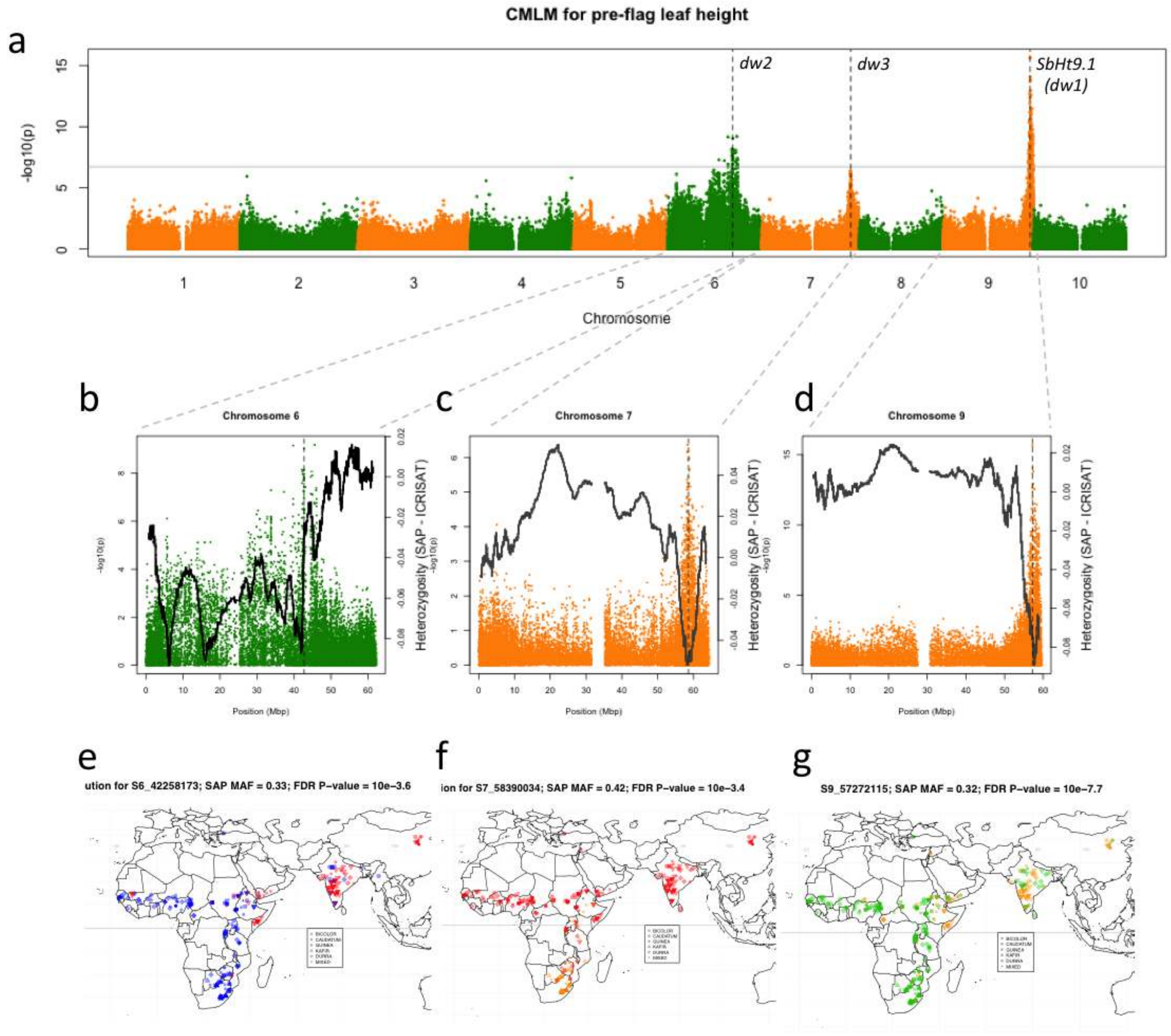


Figure 3: GWAS of pre-flag leaf height using landraces and introgression lines. (a) Manhattan plot for compressed mixed linear model with known dwarfing loci indicated (b-d) GWAS peaks for height colocalize with reductions in heterozygosity in the sorghum association panel due to introgression of short stature and early maturity alleles. (e-g) Geographic distribution of alleles at QTL, color-coded by allele (A=green, C=blue, T=red, G=orange)

Figure 4 [Branch length GWAS]

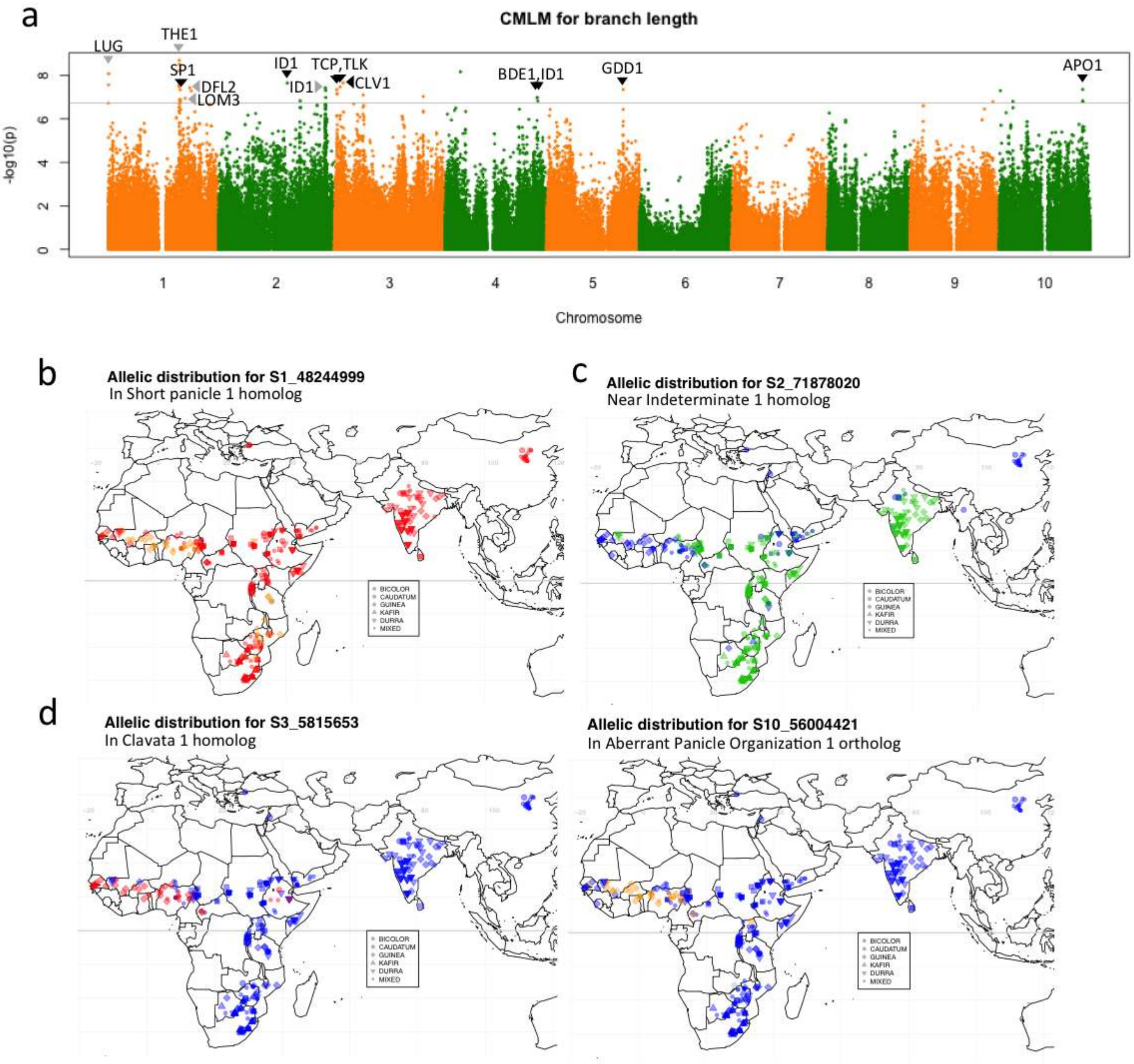


Figure: 4: GWAS on inflorescence branch length and geographic distribution QTL alleles. (a) Compressed Mixed Linear Model using first three principal components of population structure as co-variates. Candidate genes at peaks are indicated (see Table 1), with association peaks in the given gene denoted by black triangles and outside the given gene by grey triangles. (b-d) Worldwide distribution of alleles at four branch length QTL demonstrate the spread of haplotypes associated with variation in branch length, color-coded by allele (A=green, C=blue, T=red, G=orange).

Table 1 [Branch Length GWAS]

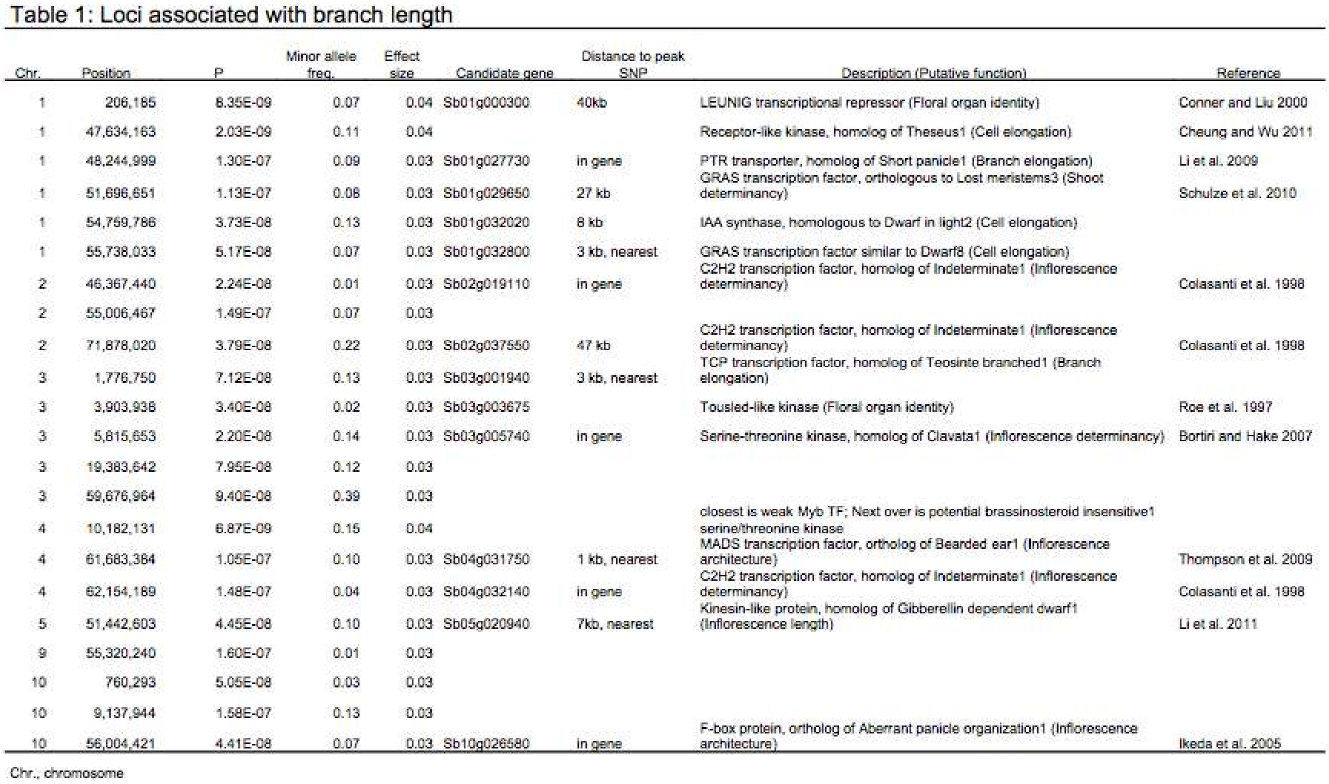


Table 1: Loci associated with branch length

Chr.	Position	P	Minor allele freq.	Effect size	Candidate gene	Distance to peak SNP	Description (Putative function)	Reference
1	206,185	8.35E-09	0.07	0.04	Sb01g000300	40kb	LEUNIG transcriptional repressor (Floral organ identity)	Conner and Liu 2000
1	47,634,163	2.03E-09	0.11	0.04			Receptor-like kinase, homolog of Theseus1 (Cell elongation)	Cheung and Wu 2011
1	48,244,999	1.30E-07	0.09	0.03	Sb01g027730	in gene	PTR transporter, homolog of Short panicle1 (Branch elongation)	Li et al. 2009
1	51,696,651	1.13E-07	0.08	0.03	Sb01g029650	27 kb	GRAS transcription factor, orthologous to Lost meristems3 (Shoot determinancy)	Schulze et al. 2010
1	54,759,786	3.73E-08	0.13	0.03	Sb01g032020	8 kb	IAA synthase, homologous to Dwarf in light2 (Cell elongation)	
1	55,738,033	5.17E-08	0.07	0.03	Sb01g032800	3 kb, nearest	GRAS transcription factor similar to Dwarf8 (Cell elongation)	
2	46,367,440	2.24E-08	0.01	0.03	Sb02g019110	in gene	C2H2 transcription factor, homolog of Indeterminate1 (Inflorescence determinancy)	Colasanti et al. 1998
2	55,006,467	1.49E-07	0.07	0.03				
2	71,878,020	3.79E-08	0.22	0.03	Sb02g037550	47 kb	C2H2 transcription factor, homolog of Indeterminate1 (Inflorescence determinancy)	Colasanti et al. 1998
3	1,776,750	7.12E-08	0.13	0.03	Sb03g001940	3 kb, nearest	TCP transcription factor, homolog of Teosinte branched1 (Branch elongation)	
3	3,903,938	3.40E-08	0.02	0.03	Sb03g003675		Tousled-like kinase (Floral organ identity)	Roe et al. 1997
3	5,815,653	2.20E-08	0.14	0.03	Sb03g005740	in gene	Serine-threonine kinase, homolog of Clavata1 (Inflorescence determinancy)	Bortiri and Hake 2007
3	19,383,642	7.95E-08	0.12	0.03				
3	59,676,964	9.40E-08	0.39	0.03				
4	10,182,131	6.87E-09	0.15	0.04			closest is weak Myb TF; Next over is potential brassinosteroid insensitive1 serine/threonine kinase	
4	61,683,384	1.05E-07	0.10	0.03	Sb04g031750	1 kb, nearest	MADS transcription factor, ortholog of Bearded ear1 (Inflorescence architecture)	Thompson et al. 2009
4	62,154,189	1.48E-07	0.04	0.03	Sb04g032140	in gene	C2H2 transcription factor, homolog of Indeterminate1 (Inflorescence determinancy)	Colasanti et al. 1998
5	51,442,603	4.45E-08	0.10	0.03	Sb05g020940	7kb, nearest	Kinesin-like protein, homolog of Gibberellin dependent dwarf1 (Inflorescence length)	Li et al. 2011
9	55,320,240	1.60E-07	0.01	0.03				
10	760,293	5.05E-08	0.03	0.03				
10	9,137,944	1.58E-07	0.13	0.03				
10	56,004,421	4.41E-08	0.07	0.03	Sb10g026580	in gene	F-box protein, ortholog of Aberrant panicle organization1 (Inflorescence architecture)	Ikeda et al. 2005

Chr., chromosome

Supplementary Note

GWAS on plant height

As *dw*3 mutants are also known to show increased upper stem elongation, we mapped *dw*3 based on associations with flag leaf to apex distance (flag-to-apex) as well. In this case, the top association peak is found near the *dw*3 locus (XX kb; p < 10XX). The strongest association peak for plant height traits (Supp. table XX) is a narrow peak on Chr. 9 between XX and 57.2 Mb, which co-localizes with previously described plant height locus dw1/SbHt9.1 {Brown et al. 2008; Wang et al. 2011}. The top association for plant height is 29 kb from a GA2-oxidase, a catabolic enzyme in the gibberellin pathway, which has been proposed as the gene underlying the plant height QTL SbHt9.1/*Dw1* {Brown et al. 2008; Wang et al. 2011}. Over-expression of GA2ox in rice leads to semi-dwarf phenotypes {Huang et al. 2010}. The second most significant peak maps to Chr. 6 between 39.7 Mb and 42.6 Mb near the classical dwarfing locus *Dw2* {Quinby and Karper 1945; Lin et al. 1995}. *Dw*2 has previously been mapped adjacent to *Ma*1 on chr. 6 {Klein et al. 2008}, to a region of ~100kb around 42.2 Mb, but the gene underlying this QTL has not been cloned. The association peak for total plant height and pre-flag leaf height maps to a histone deacetylase (Sb06g015420), which is homologous to well-studied global transcriptional regulators in human, yeast (*RDP*3), and plants (hda) (ref). In maize and *Arabidopsis*, down-regulation of closely-related histone deacetylases (*hda*101 and *AtHD*1, respectively) results in reduced plant height and a variety of changes in inflorescence architecture {Tian and Chen 2001; Rossi et al. 2007}. In rice, over-expression of *OsHDAC*1 increases plant height {Jang et al. 2003}, while the knock-down of many genes in the *OsHDAC* gene family lead to semi-dwarf phenotypes {Hu et al. 2009}. Therefore, we propose that *dw2* phenotype is a result of loss of function in a sorghum histone deacetylase. The fourth classical dwarfing loci in sorghum, dw4, has not been genetically mapped, but is known to be unlinked to the other dwarfing loci {Quinby and Karper 1954}. Based on the location of the next most significant peak in the height GWAS and heterozygosity scan, the likely physical position of the *Dw4* locus is at ~6.6 Mbp on chromosome 6 (Fig 3b).

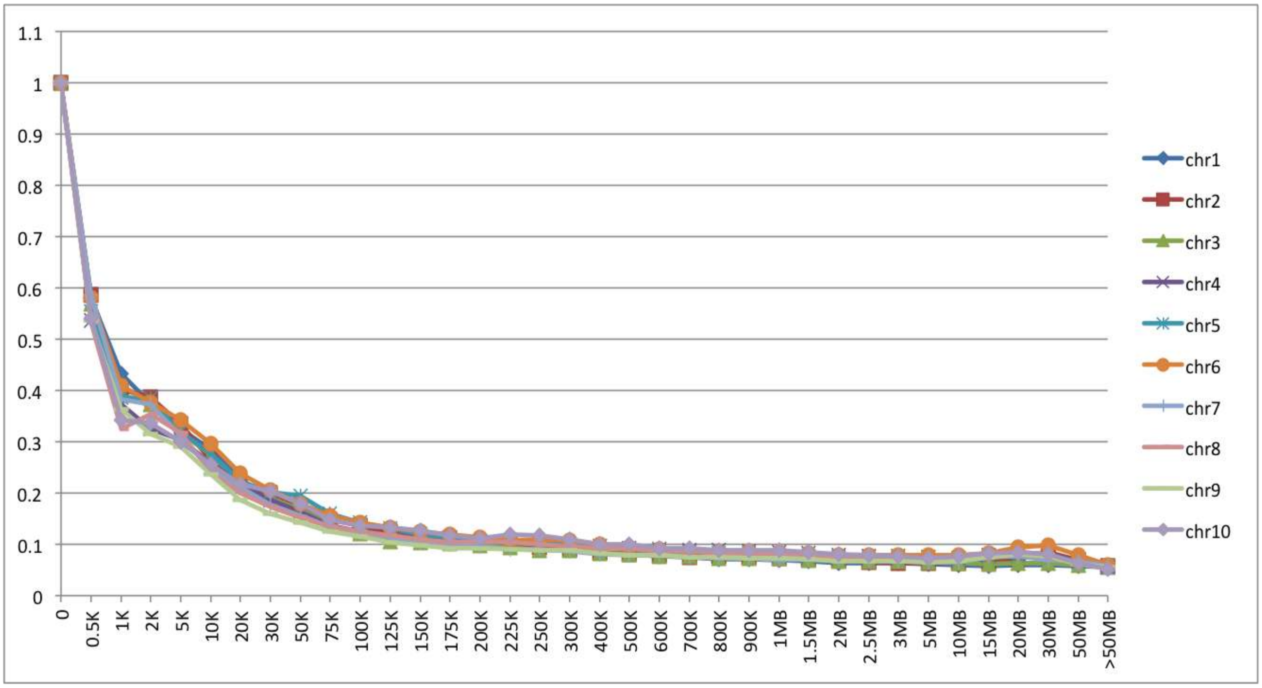
Supplemental table 1: World-wide accessions of sorghum used in this study

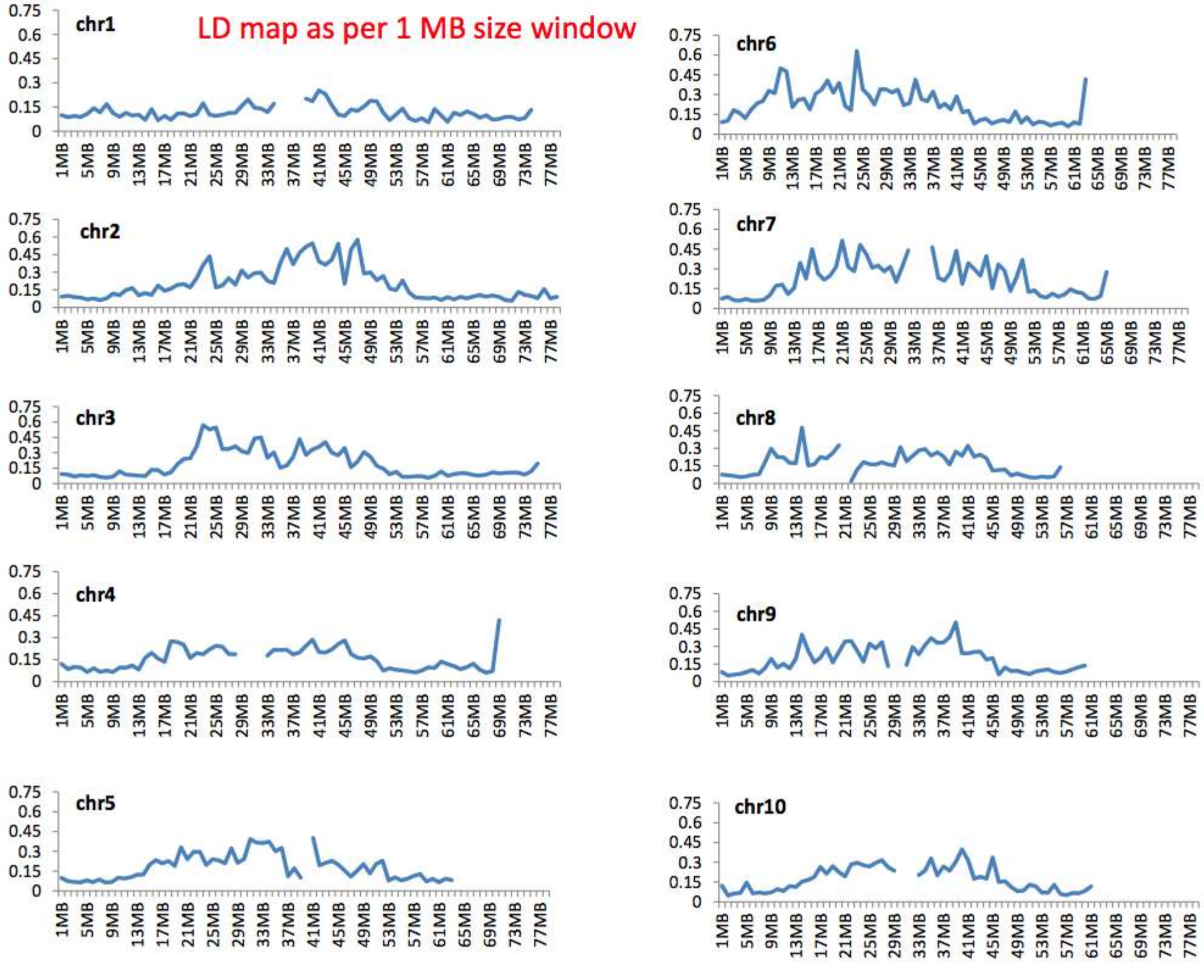
[Still getting the supplement ready...]

Supplemental figure 1: Bayesian hierarchical clustering of sorghum accessions based on 265,000 SNPs. Posterior probability of membership (Q) in each population for K=2 to K=19. Color-coding of Q-value bar plots is arbitrary, while color-coding for rug plots indicates morphological type (see Fig. 1 for key). For clarity, only African and Asian source-identified accessions are displayed. The lowest cross-validation error was observed at K=16.

[See high-resolution version attached separately]

Supplemental figure 2: Genome-wide patterns of linkage disequilibrium (a) LD decay maps for all chromosomes. (b) Average level of LD in 1-Mb windows along each chromosome.





Supplemental table 2: Summary of GWAS results with candidate genes under QTL peaks

[Still getting the supplement ready...]

Supplemental figure 3: Genome-wide association studies for 11 architecture and traits

Manhattan plots and QQ plots for compressed mixed linear model GWAS

1. Plant height (2 locations)
2. Pre-flag leaf height
3. Pre-flag to flag distance
4. Branch length

[Still getting the supplement ready...]

Acknowledgments

We thank National Science Foundation (NSF) for providing funds (ID: IOS-0965342) to carry out this research under Basic Research to Enable Agricultural Development (BREAD) project. SK and GPM were supported by USDA []...

Authors contribution

E.S.B., S.E.M., C.T.H., and S.K. designed the project; H.D.U. contributed plant materials; P.R., D.S.P., O.R.-L., C.B.A. and S.E.M. were responsible for genotyping; J.H., J.C.G., and E.S.B. developed the SNP-calling pipeline; P.J.B contributed phenotype data; P.R., T.S., and G.P.M. analyzed data; G.P.M. wrote the paper with significant contributions from P.R.

The authors declare no conflict of interest.

References

Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. Genome Research **19**:1655-1664

Armstead I, Huang L, Ravagnani A, Robson P, Ougham H (2009) Bioinformatics in the orphan crops. *Briefings in Bioinformatics*, **10**, 645–653.

Atwell S, Huang YS, Vilhjalmsson BJ, *et al.* (2010) Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. *Nature*, **465**, 627–631.

Auton A, McVean G (2007) Recombination rate estimation in the presence of hotspots. Genome Res **17**: 1219 - 1227.

Baird NA, Etter PD, Atwood TS, *et al.* (2008) Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. *PLoS ONE*, **3**, e3376.

Blackman BK, Scascitelli M, Kane NC, *et al.* (2011) Sunflower domestication alleles support single domestication center in eastern North America. *Proceedings of the National Academy of Sciences*, **108**, 14360–14365.

Bouchet S, Pot D, Deu M, *et al.* (2012) Genetic Structure, Linkage Disequilibrium and Signature of Selection in Sorghum: Lessons from Physically Anchored DArT Markers. *PLoS ONE*, **7**, e33470.

Brachi B, Morris GP, Borevitz JO (2011) Genome-wide association studies in plants: the missing heritability is in the field. *Genome Biology*, **12**, 232.

Bradbury, P.J. et al. TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23, 2633-2635 (2007).

Brown P, Klein P, Bortiri E, *et al.* (2006) Inheritance of inflorescence architecture in sorghum. *TAG Theoretical and Applied Genetics*, **113**, 931–942.

Brown PJ, Rooney WL, Franks C, Kresovich S (2008) Efficient Mapping of Plant Height Quantitative Trait Loci in a Sorghum Association Population With Introgressed Dwarfing Genes. *Genetics*, **180**, 629 –637.

Brown PJ, Myles S, Kresovich S (2011) Genetic Support for Phenotype-based Racial Classification in Sorghum. *Crop Science*, **51**, 224.

Buckler ES, Thornsberry JM (2002) Plant molecular diversity and applications to genomics. Current Opinion in Plant Biology 5: 107-111

Casa AM, Pressoir G, Brown PJ, *et al.* (2008) Community Resources and Strategies for Association Mapping in Sorghum. *Crop Science*, **48**, 30.

Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH (2011) Stacks: Building and Genotyping Loci De Novo From Short-Read Sequences. *G3: Genes, Genomes, Genetics*, **1**, 171 –182.

Cheung AY, Wu H-M (2011) THESEUS 1, FERONIA and relatives: a family of cell wall-sensing receptor kinases? *Current Opinion in Plant Biology*, **14**, 632–641.

Childs LH, Witucka-Wall H, Gunther T, Sulpice R, Korff MV, Stitt M, Walther D, Schmid KJ, Altmann T (2010) Single feature polymorphism (SFP)-based selective sweep identification and association mapping of growth-related metabolic traits in Arabidopsis thaliana. BMC Genomics 11: 188

Colasanti J, Yuan Z, Sundaresan V (1998) The indeterminate Gene Encodes a Zinc Finger Protein and Regulates a Leaf-Generated Signal Required for the Transition to Flowering in Maize. *Cell*, **93**, 593–603.

Conner J, Liu Z (2000) LEUNIG, a putative transcriptional corepressor that regulates AGAMOUS expression during flower development. *Proceedings of the National Academy of Sciences*, **97**, 12902–12907.

Crawford DC, Bhangale T, Li N, Hellenthal G, Rieder MJ, Nickerson DA, Stephens M (2004) Evidence for substantial fine-scale variation in recombination rates across the human genome. Nat Genetics 36: 700-706

de Alencar Figueiredo LF, Calatayud C, Dupuits C, *et al.* (2008) Phylogeographic Evidence of Crop Neodiversity in Sorghum. *Genetics*, **179**, 997–1008.

de Alencar Figueiredo L, Sine B, Chantereau J, *et al.* (2010) Variability of grain quality in sorghum: association with polymorphism in Sh2, Bt2, SssI, Ae1, Wx and O2. *TAG Theoretical and Applied Genetics*, **121**, 1171–1185.

Deu M, Gonzalez-de-Leon D, Glaszmann J-C, Degremont I, Chantereau J, Lanaud C, Hamon P (1994) RFLP diversity in cultivated sorghum in relation to racial differentiation. Theor Appl Genet **88**:838–844

Deu M, Rattunde F, Chantereau J (2006) A global view of genetic diversity in cultivated sorghums using a core collection. *Genome*, **49**, 168–180.

Diamond J (2002) Evolution, consequences and future of plant and animal domestication. *Nature*, **418**, 700–707.

Dillon SL, Shapter FM, Henry RJ, Cordeiro G, Izquierdo L, Lee LS (2007) Domestication to Crop Improvement: Genetic Resources for Sorghum and Saccharum (Andropogoneae). Annals of Botany **100**: 975-989

Elshire RJ, Glaubitz JC, Sun Q, *et al.* (2011) A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. *PLoS ONE*, **6**, e19379.

Fode B, Siemsen T, Thurow C, Weigel R, Gatz C (2008) The Arabidopsis GRAS Protein SCL14 Interacts with Class II TGA Transcription Factors and Is Essential for the Activation of Stress-Inducible Promoters. *The Plant Cell Online*, **20**, 3122–3135.

Foley JA, Ramankutty N, Brauman KA, *et al.* (2011) Solutions for a cultivated planet. *Nature*, **478**, 337–342.

Gao S, Wang Y, Li G. Sorghum breeding and production in China. (2010) *Cereals in China*. He ZH, Bonjean APA (Eds.) CIMMYT, Mexico, D.F.

Gepts P (2006) Plant Genetic Resources Conservation and Utilization. *Crop Science*, **46**, 2278–2292.

Goff SA, Ricke D, Lan TH, Presting G, Wang R et al. (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). Science 296:92-100

Hamblin MT, Salas Fernandez MG, Casa AM, *et al.* (2005) Equilibrium Processes Cannot Explain High Levels of Short- and Medium-Range Linkage Disequilibrium in the Domesticated Grass Sorghum bicolor. *Genetics*, **171**, 1247–1256..

Harlan JR, Stemler A (1976) The Races of Sorghum in Africa. In: *Origins of African Plant Domestication* (eds Harlan JR, Wet JMJD, Stemler ABL), pp. 465–478. DE GRUYTER MOUTON, Berlin, New York.

Hu Y, Qin F, Huang L, *et al.* (2009) Rice histone deacetylase genes display specific expression patterns and developmental functions. *Biochemical and Biophysical Research Communications*, **388**, 266–271.

Huang X, Wei X, Sang T, *et al.* (2010) Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat Genet*, **42**, 961–967.

Huang J, Tang D, Shen Y, *et al.* (2010) Activation of gibberellin 2-oxidase 6 decreases active gibberellin levels and creates a dominant semi-dwarf phenotype in rice (Oryza sativa L.). *Journal of Genetics and Genomics*, **37**, 23–36.

Huang X, Zhao Y, Wei X, *et al.* (2012) Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. *Nature Genetics*, **44**, 32–39.

Ikeda K, Nagasawa N, Nagato Y (2005) ABERRANT PANICLE ORGANIZATION 1 temporally regulates meristem identity in rice. *Developmental Biology*, **282**, 349–360.

Jang I-C, Pahk Y-M, Song SI, *et al.* (2003) Structure and expression of the rice class-I type histone deacetylase genes OsHDAC1–3: OsHDAC1 overexpression in transgenic plants leads to increased growth rate and altered architecture. *The Plant Journal*, **33**, 531–541.

Jiao Y, Zhao H, Ren L, Song W, Zeng B, Guo J, Wang B, Liu Z, Chen J, Li W, Zhang M, Xie S, Lai J (2012) Genome-wide genetic changes during modern breeding of maize. Nature Genetics 44: 812-815.

Jones FC, Chan YF, Schmutz J, *et al.* (2012) A Genome-wide SNP Genotyping Array Reveals Patterns of Global and Repeated Species-Pair Divergence in Sticklebacks. *Current Biology*, **22**, 83–90.

Kijas JW, Lenstra JA, Hayes B, *et al.* (2012) Genome-Wide Analysis of the World's Sheep Breeds Reveals High Levels of Historic Mixture and Strong Recent Selection. *PLoS Biol*, **10**, e1001258.

Lawit SJ, Wych HM, Xu D, Kundu S, Tomes DT (2010) Maize DELLA proteins dwarf *plant8* and dwarf *plant9* as modulators of plant development. Plant and Cell Physiol 51: 1854–1868.

Li H. and Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754-60.

Li S, Qian Q, Fu Z, *et al.* (2009) Short panicle1 encodes a putative PTR family transporter and determines rice panicle size. *The Plant Journal*, **58**, 592–605.

Lin YR, Schertz KF, Paterson AH (1995) Comparative analysis of QTLs affecting plant height and maturity across the Poaceae, in reference to an interspecific sorghum population. *Genetics*, **141**, 391–411.

Lin Z, Li X, Shannon LM, *et al.* (2012) Parallel domestication of the Shattering1 genes in cereals. *Nature Genetics*, **44**, 720–724.

Lipka AE, Tian F, Wang Q, Peiffer J, Li M, Bradbury PJ, Gore M, Buckler ES, Zhang Z (2012) GAPIT: Genome Association and Prediction Integrated Tool. *Bioinformatics* doi: 10.1093/bioinformatics/bts444.

Lobell DB, Burke MB, Tebaldi C, *et al.* (2008) Prioritizing Climate Change Adaptation Needs for Food Security in 2030. *Science*, **319**, 607–610.

Lukens L, Doebley J (2001) Molecular evolution of the teosinte branched gene among maize and related grasses. Mol Biol Evol 18: 627-638.

Mace ES, Buhariwalla HK, Crouch JH (2003) A high throughput DNA extraction protocol for molecular breeding programs. Plant Mol Biol Rep 21:459a–459h.

Mather KA, Caicedo AL, Polato NR, Olsen KN, McCouch S, Purugganan MD (2007) The Extent of Linkage Disequilibrium in Rice (Oryza sativa L.). Genetics **177**: 2223–2232.

Moritsuka E, Hisataka Y, Tamura M, Uchiyama K, Watanabe A, Tsumura Y, Tachida H (2012) Extended linkage disequilibrium in non-coding regions in a conifer, *Cryptomeria japonica*. Genetics **190**: 1145-1148

Morrell PL, Buckler ES, Ross-Ibarra J (2012) Crop genomics: advances and applications. *Nature Review Genetics* **13**:85-96

Multani DS, Briggs SP, Chamberlin MA, *et al.* (2003) Loss of an MDR Transporter in Compact Stalks of Maize Br2 and Sorghum Dw3 Mutants. *Science*, **302**, 81–84.

Murphy RL, Klein RR, Morishige DT, *et al.* (2011) Coincident light and clock regulation of pseudoresponse regulator protein 37 (PRR37) controls photoperiodic flowering in sorghum. *Proceedings of the National Academy of Sciences*, **108**, 16469–16474.

Myles S, Peiffer J, Brown PJ, *et al.* (2009) Association Mapping: Critical Considerations Shift from Genotyping to Experimental Design. *The Plant Cell*, **21**, 2194–2202.

Myles S, Boyko AR, Owens CL, *et al.* (2011) Genetic structure and domestication history of the grape. *Proceedings of the National Academy of Sciences*, **108**, 3530 –3535.

National Research Council (1996) *Lost Crops of Africa: Volume I: Grains*. National Academy Press, Washington, D.C.

Olsen KM, Caicedo AL, Polato N, McClung A, McCouch S, Purugganan MD (2006) Selection under domestication: evidence for a sweep in the rice waxy genomic region. Genetics 173: 975–983

Paape T, Zhou P, Branca A, Briskine R, Young N, Tiffin P (2012) Fine-scale population recombination rates, hotspots, and correlates of recombination in the medicago truncatula genome. Genome Biol Evol 4(5):726–737

Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, Schmutz J, Spannagl M, Tang H, Wang X, Wicker T, Bharti AK, Chapman J, Feltus FA, Gowik U, Grigoriev IV, Lyons E, Maher CA, Martis M, Narechania A, Otillar RP, Penning BW, Salamov AA, Wang Y, Zhang L, Carpita NC, Freeling M, Gingle AR, Hash CT, Keller B, Klein P, Kresovich S, McCann MC, Ming R, Peterson DG, Mehboob-ur-Rahman, Ware D, Westhoff P, Mayer KFX, Messing J, Rokhsar DS (2009) The Sorghum bicolor genome and the diversification of grasses. Nature 457:551-556.

Paradis E (2010) pegas: an R package for population genetics with an integrated–modular approach. *Bioinformatics*, **26**, 419 –420.

Peng J, Richards DE, Hartley NM, *et al.* (1999) “Green revolution” genes encode mutant gibberellin response modulators. *Nature*, **400**, 256–261.

Ramanatha Rao V, Hodgkin T (2002) Genetic diversity and conservation and utilization of plant genetic resources. *Plant Cell, Tissue and Organ Culture*, **68**, 1–19.

Ramu P, Kassahun B, Senthilvel S, Kumar CA, Jayashree B, Folkertsma RT, Reddy LA, Kuruvinashetti MS, Haussmann BIG, Hash CT (2009) Exploiting rice-sorghum synteny for targeted development of EST-SSRs to enrich the sorghum genetic linkage map. Theor Appl Genet 119:1193–1204

Roberts A, McMillan L, Wang W, *et al.* (2007) Inferring Missing Genotypes in Large SNP Panels Using Fast Nearest-Neighbor Searches Over Sliding Windows. *Bioinformatics*, **23**, i401–i407.

Roe JL, Nemhauser JL, Zambryski PC (1997) TOUSLED participates in apical tissue formation during gynoecium development in Arabidopsis. *The Plant Cell Online*, **9**, 335–353.

Schulze S, Schäfer BN, Parizotto EA, Voinnet O, Theres K (2010) LOST MERISTEMS genes regulate cell differentiation of central zone descendants in Arabidopsis shoot meristems. *The Plant Journal*, **64**, 668–678.

Stephens JC, Miller FR, Rosenow DT (1967) Conversion of Alien Sorghums to Early Combine Genotypes. *Crop Science*, **7**, 396.

Sweeney MT, Thomson MJ, Cho YG, Park YJ, Williamson SH, Bustamante CD, McCouch SR (2007) Global dissemination of a single mutation conferring white pericarp in rice. PLoS Genet 3:e133

The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408: 796-815

Thompson BE, Bartling L, Whipple C, *et al.* (2009) bearded-ear Encodes a MADS Box Transcription Factor Critical for Maize Floral Development. *The Plant Cell Online*, **21**, 2578–2590.

Thornsberry JM, Goodman MM, Doebley J, *et al.* (2001) Dwarf8 polymorphisms associate with variation in flowering time. *Nature Genetics*, **28**, 286–289.

Tian L, Chen ZJ (2001) Blocking histone deacetylation in Arabidopsis induces pleiotropic effects on plant gene regulation and development. *Proceedings of the National Academy of Sciences*, **98**, 200–205.

Tian C, Wan P, Sun S, Li J, Chen M (2004) Genome-Wide Analysis of the GRAS Gene Family in Rice and Arabidopsis. *Plant Molecular Biology*, **54**, 519–532.

Tian F, Stevens NM, Buckler ES (2009) Tracking footprints of maize domestication and evidence for a massive selective sweep on chromosome 10. Proc. Natl. Acad. Sci 106: 9979–9986.

Tian F, Bradbury PJ, Brown PJ, *et al.* (2011) Genome-wide association study of leaf architecture in the maize nested association mapping population. *Nat Genet*, **43**, 159–162.

Upadhyaya HD, Pundir RPS, Dwivedi SL, Gowda CLL, Reddy VG, Singh S (2009) Developing a mini-core collection of sorghum (Sorghum bicolor (L.) Moench) for diversified utilization of germplasm. Crop Sci. 49:1769-1780.

Upadhyaya HD, Wang Y-H, Sharma S, Singh S (2012) Association mapping of height and maturity across five environments using the sorghum mini core collection. Genome **55**: 471–479

Vadez V, Krishnamurthy L, Hash CT, Upadhyaya HD, Borrell AK (2011) Yield, transpiration efficiency, and water-use variations and their interrelationships in the sorghum reference collection. Crop & Pasture Science 62: 645-655.

Varshney RK, Chen W, Li Y, *et al.* (2012) Draft genome sequence of pigeonpea (Cajanus cajan), an orphan legume crop of resource-poor farmers. *Nature Biotechnology*, **30**, 83–89.

Vigouroux Y, McMullen M, Hittinger CT, *et al.* (2002) Identifying Genes of Agronomic Importance in Maize by Screening Microsatellites for Evidence of Selection During Domestication. *Proceedings of the National Academy of Sciences*, **99**, 9650–9655.

Wang Y-H, Bible P, Loganantharaj R, Upadhyaya H (2011) Identification of SSR markers associated with height using pool-based genome-wide association mapping in sorghum. *Molecular Breeding*, 1–12.

Wendorf F, Close AE, Schild R, *et al.* (1992) Saharan exploitation of plants 8,000 years BP. *Nature*, **359**, 721–724.

Whitt SR, Wilson LM, Tenaillon MI, Gaut BS, Buckler ES (2002) Genetic diversity and selection in the maize starch pathway. *Proceedings of the National Academy of Sciences*, **99**, 12959–12962.

Wijnker E, de Jong H (2008) Managing meiotic recombination in plant breeding. Trends in Plant Science **13**: 640-646

Wright SI, Bi IV, Schroeder SG, Yamasaki M, Doebley JF, McMullen MD, Gaut BS (2005) The Effects of Artificial Selection on the Maize Genome. Science 308, 1310–1314

Zhang Z, Ersoz E, Lai C-Q, *et al.* (2010) Mixed linear model approach adapted for genome-wide association studies. *Nature Genetics*, **42**, 355–360.

Zheng L-Y, Guo X-S, He B, *et al.* (2011) Genome-wide patterns of genetic variation in sweet and grain sorghum (Sorghum bicolor). *Genome Biology*, **12**, R114.

Zheng SJ (2010) Crop production on acidic soils: overcoming aluminum toxicity and phosphorus deficiency. Annals of Botany 106: 183-184.