Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Genome-wide and haplotype mapping reveals genetic drivers of biomass productivity per unit water and nitrogen in sorghum

Pramod Sargar
International Crops Research Institute for the Semi-Arid Tropics (ICRISAT)
Niranjan Thakur
International Crops Research Institute for the Semi-Arid Tropics (ICRISAT)
Shivaji Mehtre
Vasantrao Naik Marathwada Krishi Vidyapeeth
Pradeep Ruperao
International Crops Research Institute for the Semi-Arid Tropics (ICRISAT)
Christian Hermans
Université libre de Bruxelles
Hirakant Kalpande
Vasantrao Naik Marathwada Krishi Vidyapeeth
Godawari Pawar
Vasantrao Naik Marathwada Krishi Vidyapeeth
Anil Kumar Vemula
International Crops Research Institute for the Semi-Arid Tropics (ICRISAT)
Ephrem Habyarimana
Ephrem.Habyarimana@icrisat.org

International Crops Research Institute for the Semi-Arid Tropics (ICRISAT)

Research Article

Keywords: candidate gene expression, genome-wide association study, haplotype-based selection, nitrogen use efficiency, partial factor productivity of applied nitrogen, Sorghum bicolor × S. halepense, water use efficiency, wild relative introgression

Posted Date: November 2nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11280428/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Improving crop productivity per unit of water and nitrogen is central to climate resilient sorghum improvement, yet the genetic architecture of biomass-based resource productivity remains poorly resolved across contrasting production environments. This study integrated multi-environment phenotyping with 1.75 million SNPs derived from whole genome resequencing of 172 cultivated *Sorghum bicolor* and *S. bicolor* × *S. halepense* derivative lines evaluated across temperate Mediterranean and semi-arid tropical environments.

Results

Biomass-based water productivity and partial factor productivity of applied nitrogen displayed extensive genetic variation and high broad-sense heritability of 0.75 to 0.81, whereas grain-based indices were more strongly influenced by genotype × environment interaction. Multi-model genome wide association analyses identified 27 significant SNPs defining 22 genomic regions for biomass-based productivity. A recurrent chromosome 9 signal at S09_55037491 was detected across three analytical models and biomass-derived indices and colocalized with Sobic.009G200500, a nitrate-responsive HAT3.1 homeobox transcription factor. Haplotype analysis independently prioritized two additional loci: a chromosome 9 haplotype associated with a 2.7-fold difference in biomass-based water productivity and containing an AHP-like cytokinin-signaling gene, and a chromosome 5 haplotype associated with a 1.86-fold difference and encompassing DAD1, a phospholipase implicated in jasmonate biosynthesis. Candidate gene prioritization also implicated aquaporin-mediated transport, trehalose-6-phosphate metabolism, oxidative signaling and, provisionally, C4 carbon assimilation. Most favorable common haplotypes occurred within cultivated germplasm rather than the introgressed lines.

Conclusions

Together, these findings reveal genetically tractable loci and biological pathways underlying sorghum biomass productivity per unit of supplied water and nitrogen. Following independent validation, the identified SNPs and haplotypes could support diagnostic marker development and haplotype informed selection, while genomic selection remains more appropriate for environmentally sensitive grain-based productivity traits.

Background

Agriculture must produce more food with less water and less fertilizer. Irrigated agriculture accounts for about 70% of global freshwater withdrawals, and roughly four billion people face severe water scarcity for at least one month a year [1, 2]. Nitrogen (N) fertilizer poses a parallel problem: less than half of the N applied to cropland is recovered in the harvested product, and the residue is lost as nitrate to ground and surface water and as nitrous oxide, an ozone-depleting greenhouse gas with about 270 times the warming potential of CO_2 [3–5]. Nitrogen flows are now among the most severely transgressed Earth environmental boundaries [6]. Closing yield gaps by adding more water and N is therefore no longer an option, and further gains must come from the crop itself.

Resource use efficiency (RUE) describes how effectively a crop converts an acquired resource into biomass or grain [7]. Its two agronomically decisive components require separate definitions. Water use efficiency (WUE) is the biomass or grain produced per unit of water [8]. At field scale this is an agronomic water productivity, which differs from the transpiration based or leaf level intrinsic WUE measured by gas exchange or carbon isotope discrimination. Nitrogen use efficiency (NUE) describes how well a crop takes up N, assimilates it and converts it into harvestable product, and it can be decomposed into several components [9]. In this study, NUE is represented by the partial factor productivity of applied N (PFPAN), the yield obtained per unit of N applied [10, 11]. Because sorghum is grown for both grain and biomass, WUE and PFPAN were each calculated on a dry biomass (DBM) basis and on a grain yield (Gyld) basis, giving four indices that cover vegetative and reproductive resource conversion. Biomass based indices matter for forage and dual purpose systems, and grain-based indices for food grain production [12, 13].

Sorghum is well suited to the study of RUE. It is the world's fifth most important cereal and a staple for more than 500 million people across semi-arid Africa and South Asia [14]. It combines C_4 photosynthesis, a deep and plastic root system and a well documented capacity to yield on marginal, low input soils [13, 15, 16]. Gains in RUE in sorghum would therefore benefit directly the production systems that need them most, rather than adding small increments in environments that are already optimized.

The main obstacle is allelic rather than agronomic. Domestication and a century of selection under high inputs have narrowed the cultivated sorghum gene pool, and alleles that improve growth per unit of water and N have rarely been targeted by selection. *Sorghum halepense* (L.) Pers. (Johnsongrass), the closest perennial wild relative of cultivated sorghum, retains traits absent from cultivated germplasm, including rhizomatous perenniality, deep roots and vigorous regrowth [17–19]. Long treated as a weed, it is better regarded as a genetic reservoir. Whole genome resequencing (WGRS) of the *S. bicolor* × *S. halepense* (Sb×Sh) panel used here showed that the introgressed sub-population carries more SNPs and structural variants than cultivated sorghum, and that its private genes are enriched in Gene Ontology categories relevant to RUE, including N compound transport, response to abiotic stimulus, hormone signaling and root development [20]. Wild relative introgression has already contributed useful alleles to rice, maize and wheat [21–23], but it has rarely been exploited for RUE in sorghum. The first question addressed here is therefore what wild relative introgressions contribute to RUE.

The second question is whether water use and N use share a genetic basis. There are mechanistic reasons to expect overlap. Aquaporins transport water and, in several characterized members, also ammonia and urea, so a single channel protein can influence the uptake of both resources [24, 25]; the nodulin-

26-like intrinsic proteins (NIPs) additionally transport a range of small solutes [26]. Trehalose 6-phosphate is a signaling metabolite that reports carbon status and, through it, matches growth to N availability [27, 28]. Cytokinin phosphorelay links nitrate availability to root architecture and shoot growth [29, 30]. If such nodes carry natural variation, single genomic regions could influence both WUE and PFPAN, and such regions would be particularly valuable because water and N supply commonly vary together in the field.

Converting either answer into a breeding tool needs strong associations. Linkage disequilibrium (LD) decays rapidly in sorghum (1–20 kb; [31–33]), so WGRS density genotyping can narrow an association to a few genes. Dissecting haplotypes within LD defined windows captures allele combinations that single markers miss and provides more robust tags for marker assisted selection (MAS) [34–36]. Tissue resolved transcriptome atlases add an independent filter, a candidate whose expression corresponds to the organ in which the trait is expressed, such as roots for water uptake or seed for grain-based indices, is more likely to be causal than one whose expression does not [37–39]. The third question is therefore which biological pathways emerge as promising breeding targets for resource-efficient sorghum.

A recent review of sorghum root system architecture and RUE identified persistent gaps that limit progress on these questions. RUE and root traits are rarely combined in a single genetic framework, *S. halepense* introgressed germplasm is underused, and water and N productivity indices have rarely been dissected jointly on both a biomass and a grain basis across environments [16]. Here, these gaps are addressed by phenotyping a diverse sorghum panel comprising S. bicolor and Sb×Sh genotypes for RUE in temperate Mediterranean and semi-arid tropical environments, associating the indices with WGRS derived SNPs, and following each robust signal through to candidate genes, tissue specific expression and haplotype structure. The outcome is a ranked set of biological hypotheses on how sorghum converts water and N into biomass and grain.

Methods

Plant material

The germplasm from the parents fell into two categories. The annual parents were lines of *S. bicolor* comprising four types, which included standard diploids (2n = 20), induced tetraploids (2n = 40), cytoplasmic genetic male sterile lines and genetic male sterile lines. The perennial parents were *S. halepense* plants together with their tetraploid derivatives. Interspecific recombinant inbred lines (RILs) were obtained from crosses between annual and perennial (A/P), from A/P backcrosses, from crosses between perennial and perennial and from A/P × perennial crosses, and then the resulting lines were subjected to repeated selfing and phenotypic selection until the F_4–F_7 generations were reached. As selection was carried out over the generations, the undesirable traits associated with the wild parents such as seed shattering, small kernels, excessive tillering and abnormal inflorescence shape were eliminated, while the introgressed genomic background was retained [20, 40].

The panel of genotypes included 172 lines, consisting of 153 inbred lines of the cultivated *S. bicolor* and tropical landraces from Africa and Asia, as well as 19 Sb×Sh RILs which had been derived from A/P crosses (there were 12 of these), from A/P crosses that had been backcrossed to the annual parent (six of these), and one cross that involved an A/P and a perennial plant. The genotypes, together with their pedigrees and cross categories, are given in Additional file 1: Table S1. The whole genome resequencing (WGRS) for this panel was previously reported by Habyarimana et al. [20] and the variant set from that study was used in this case.

Field environments, treatments and environmental combinations

Trials were conducted in three environments, each denoted by a single acronym throughout:

- **IT21**: CREA Research Centre, Anzola dell'Emilia, Bologna, Italy (44°32' N, 11°12' E, 24 m a.s.l.), 2021. The area has a temperate Mediterranean climate. The soil is alluvial silt loam.
- **IND22**: ICRISAT, Patancheru, India (17°32' N, 78°16' E, 549 m a.s.l.), 2022–23 season. Semi-arid tropical climate. Vertisol soil.
- **IND23**: The ICRISAT site in Patancheru, during the 2023–24 season; at the same location as IND22 in the preceding season.

IT21 stands for the temperate, radiation rich, spring sown conditions found in southern Europe, where sorghum is primarily grown as fodder and for bioenergy production [41]. IND22 and IND23 refer to the semi-arid tropical conditions that occur after the rainy season and are characterized by decreasing moisture in peninsular India, where sorghum is used as a food grain and as a dual-purpose crop [19]. By comparing these systems, this study tested whether the RUE alleles are consistent across the range of environments in which sorghum is grown or only within a single type of production system. Although IND22 and IND23 were conducted at the same location, the two successive seasons were treated as separate environments.

Each environment used an augmented randomized complete block design [42], with check cultivars replicated in every block and unreplicated test genotypes. The checks were used to estimate block effects and to obtain best linear unbiased estimates (BLUEs) for the test entries. Trial size differed among environments. IT21 had 172 test genotypes, 6 checks and 5 blocks; IND22 had 170 test genotypes, 13 checks and 5 blocks; and IND23 had 156 test genotypes, 4 checks and 12 blocks. Two input treatments were applied in each environment: $+H_2O$+N (full irrigation plus the recommended N dose) and $-H_2O$–N (reduced irrigation and no N fertilizer). Standard sorghum agronomy was applied at each site, and no major pest or disease pressure was recorded. The primary traits were recorded in both treatments (Additional file 1: Tables S3, S5), whereas the RUE indices analyzed here refer to the $+H_2O$+N treatment, because PFPAN is undefined when no N is applied.

The total seasonal water input (W, in mm) consisted of the rainfall that occurred during the season plus the supplemental irrigation given to each treatment. W is thus a quantity of applied water which is the same for all the genotypes within a given treatment and not a measurement specific to the transpiration of any genotype. At IT21 the volume of irrigation was measured using an automated in line flow meter and then converted into water depth per plot. At IND22

and IND23 the irrigation depth was estimated from the volume applied and the area of the plot in accordance with the FAO methodology [43]. For the fully rainfed plots W was set equal to the cumulative seasonal rainfall. In the +H_2O+N treatment W amounted to 262 mm at IT21, 1,268 mm at IND22 and 1,750 mm at IND23, and nitrogen (N) was applied at a rate of 140 kg ha^{-1} at IT21 and 160 kg ha^{-1} at IND22 and 180 kg ha^{-1} IND23.

Grain yield and the grain-based indices were only recorded in India, which is why two environmental combinations were established (Additional file 1: Table S2). Combination 1 (IT21 + IND22 + IND23) is the one that supports the biomass-based traits, these being those with the suffix "_1" (for example, WUE_DBM_1). Combination 2 (IND22 + IND23) supports both the biomass and the grain based traits, the latter not having any suffix (for example, WUE_DBM). This practice is applied consistently throughout the manuscript, the tables and the figures.

Phenotyping and calculation of RUE indices

The main measurements taken for each plot were the number of days to flowering (DF), plant height (PlHt, in cm), dry matter content (DM, %), aboveground dry biomass (DBM, in kg ha^{-1}) and grain yield (Gyld, in kg ha^{-1}). WUE and PFPAN were each calculated based on both DBM and Gyld according to Moll et al. [10] and Cassman et al. [11], which resulted in four RUE indices: WUE_DBM, WUE_Gyld, PFPAN_DBM and PFPAN_Gyld (Table 1). Since grain yield was only recorded in India, the grain-based indices are available only for Combination 2, and six trait × combination variables were analyzed: WUE_DBM_1 and PFPAN_DBM_1 for Combination 1, and WUE_DBM, WUE_Gyld, PFPAN_DBM and PFPAN_Gyld for Combination 2.

Table 1. Abbreviations, descriptions, units and formulae of the agronomic and resource-use efficiency (RUE) traits analyzed in this study.

Abbreviation	Trait (unit)	Definition or formula
DF	Days to flowering (d)	Number of days from sowing to 50% of plants at mid-flowering
PlHt	Plant height (cm)	Average height from the soil surface to the tip of the panicle
DM	Dry matter content (%)	(Dry mass/fresh mass) × 100 (oven-drying at 60°C to constant weight)
DBM	Aboveground dry biomass (kg ha^{-1})	Total dry biomass produced per hectare
Gyld	Grain yield (kg ha^{-1})	Total grain weight harvested per hectare
WUE_DBM	Water use efficiency, dry biomass basis (kg ha^{-1} mm^{-1})	DBM/W
WUE_Gyld	Water use efficiency, grain yield basis (kg ha^{-1} mm^{-1})	Gyld/W
PFPAN_DBM	Partial factor productivity of applied N, dry biomass basis (kg kg^{-1} N)	DBM/N applied
PFPAN_Gyld	Partial factor productivity of applied N, grain yield basis (kg kg^{-1} N)	Gyld/N applied

W: the total seasonal water input (i.e. the rainfall during the season plus the irrigation, in mm); N: nitrogen fertilizer applied (in kg ha–1). Traits that are estimated from Combination 1 of the three environments are followed by the suffix "_1" (for example, WUE_DBM_1).

The interpretation of these indices depends on two factors. The first of these is that the WUE calculated here refers to the amount of biomass or grain obtained per unit of water supplied, which means it is an agronomic water productivity figure and not transpiration efficiency or leaf-level intrinsic WUE. The second point is that all the genotypes in a given environment were given the same amounts of W and N; consequently, for any particular environment the WUE and the PFPAN are the same yield divided by different constants; the differences between WUE and PFPAN in the multi-environment BLUEs result from the different way in which the environments are weighted according to their water and nitrogen inputs.

Statistical analysis

For each environment, analysis of variance (ANOVA) was carried out using the augmented RCBD model $y_{ij} = \mu + b_j + g_i + \varepsilon_{ij}$, with the block term entered before the genotype (type I sums of squares), in R v4.3 [44] with the aid of the lme4 package [45]. The variance components were estimated by the method of moments and the per-environment ANOVA tables are provided in Additional file 1: Table S3. The combined model for all environments was

$$y_{ijk} = \mu + E_j + B_{k(j)} + g_i + (gE)_{ij} + \varepsilon_{ijk}$$

The values of σ^2_G, $\sigma^2_{G\times E}$, σ^2_{Env}, $\sigma^2_{B(E)}$ and σ^2_ε were obtained from the expected mean squares [46]. The results of the combined ANOVA are provided in Additional file 1: Table S4. Broad-sense heritability as well as the genotypic and phenotypic coefficients of variation were estimated as

$$GCV\ (\%) = 100 \times \sqrt{\sigma^2_G}/\bar{x};\ PCV\ (\%) = 100 \times \sqrt{\sigma^2_P}/\bar{x}$$

$$H^2 = \sigma^2_G/(\sigma^2_G + \sigma^2_{G\times E}/e + \sigma^2_\varepsilon/(e\cdot\bar{r}))$$

where $\bar{x}$ is the trait mean, e the number of environments and $\bar{r}$ the harmonic-mean replication, and σ^2_P is the phenotypic variance of genotype means, per-environment heritability was estimated as $H^2 = \sigma^2_G/(\sigma^2_G + \sigma^2_\varepsilon/\bar{r})$. H^2 was classified as high (≥ 0.60), moderate (0.30–0.60) or low (< 0.30) following Robinson et al. [47], and interpreted jointly with GCV following Burton [48]. Per-environment genetic parameters are given in Additional file 1: Table S5.

The BLUEs were obtained in two stages. First, the per-environment BLUEs were derived from the augmented RCBD mixed model in which genotype was treated as fixed and block as random. Then, the multi environment BLUEs were obtained from the per environment adjusted means by taking genotype as fixed and both environment and block within environment as well as the G × E interaction as random [49], separately for Combinations 1 and 2 (Additional file 1: Table S6). These BLUEs were used as the response variables in all the subsequent analyses.

The relationships among the various traits were examined using principal component analysis (PCA) of the genotypic BLUE values, with the analysis carried out separately for the six traits in Combination 1, the nine traits in Combination 2 and the combined set of 15 trait × combination variables; missing BLUEs were imputed by regularized iterative PCA (missMDA; [50]). Biplots of PC1 and PC2 were produced using matplotlib [51], grouping the genotypes by sub-population (Sb and Sb×Sh).

Whole-genome resequencing and variant calling

The genotypic data were sourced from Habyarimana et al. [20]. For this, high molecular weight DNA was extracted from 10 day old, etiolated seedlings using the CTAB method and then sequenced on the DNBseq PE150 platform at a depth of 20×. The reads were quality filtered using SOAPnuke [52], aligned to the *S. bicolor* BTx623 v3.1.1 reference sequence with BWA-MEM [53], and duplicates were identified using Picard v1.118. Variants were detected with GATK HaplotypeCaller and the genotypes were called jointly using GenotypeGVCFs. After filtering with VCFtools v0.1.16 [54], only biallelic SNPs with a minor allele frequency (MAF) of at least 0.05, a site quality Q of at least 40 and a missingness rate of 20% or less were retained. The remaining missing genotypes were imputed using Beagle v5.1 [55]. The final matrix consisted of 1,752,624 high-confidence SNPs.

Population structure, kinship, and linkage disequilibrium

Genetic differentiation between the Sb and Sb×Sh groups was assessed using TASSEL HapMap genotypes [56], with individuals assigned to populations based on pedigree. SNPs were filtered for minor allele frequency ≥0.05, ≤20% missing data, and at least eight genotyped individuals per population. Genome-wide and chromosome level differentiation (FST; [57]) was estimated using the Weir and Cockerham [58] method, with Hudson's estimator used for comparison. Overall FST was calculated from average values, with 5 Mb block jackknife confidence intervals to account for linkage among nearby SNPs [59]. Population structure was assessed using a model based admixture approach [60, 61] after SNP filtering, physical thinning, and LD pruning ($r^2 > 0.2$). The number of ancestral groups (K) was selected using cross entropy. A neighbor joining tree, genomic relationship matrix [62], and hierarchical clustering were also generated. The first three principal components were included as covariates in all association analyses, and the VanRaden kinship matrix was fitted as a random effect in MLMM, whereas BLINK and FarmCPU control relatedness through iteratively selected pseudo-QTNs [63, 64]. LD decay was estimated from SNP pairs within 500 kb on each chromosome using PopLDdecay [65]. The estimated half-decay distance was 13.78 kb, which was used as the ±13.78 kb window for candidate-gene annotation and haplotype analysis.

Genome-wide association analysis and model selection

The association analyses were carried out using version 3 of the R package GAPIT [66]. Three multi locus methods were applied in parallel, BLINK [67], FarmCPU [68] and MLMM [69]. The genome wide significance level was set at a Bonferroni-corrected p value of $p < 0.05/m = 2.85 \times 10^{-8}$ ($-\log_{10}p > 7.54$), with m equal to 1,752,624. Manhattan and quantile quantile (Q–Q) plots were generated using the R packages CMplot [70] and qqman [71].

The fit of the model was evaluated using the genomic inflation factor λ_{GC}, which is the ratio of the median observed to the median expected χ^2 statistic [72]. Any combinations of trait and model for which λGC was greater than or equal to 1.35 (indicating residual inflation) or less than or equal to 0.90 (indicating over-correction) were excluded [73], and only the associations from the models that remained were included (Table 3). Associations involving SNPs that were detected by two or more of the retained models were classified as high-confidence associations.

The amount of phenotypic variance accounted for by a single SNP (PVE) is given in the text, but it is likely to be biased upwards in a panel of this size, a phenomenon recognized as the Beavis effect [74–76]. Because PVE values represent upper limits, conclusions were based on the consistency of associations across models, traits, and environments rather than on the exact PVE estimates.

Candidate gene identification, annotation and expression profiling

All annotated gene models within ± 13,778 bp of each significant SNP were extracted from the BTx623 v3.1.1 annotation in Phytozome v13 [77] and compared with Ensembl Plants (https://plants.ensembl.org); overlapping windows from multi model hits were merged. Gene Ontology (GO) annotations were retrieved from Phytozome, UniProt [78] and PANTHER [79]. Candidates were prioritized on three criteria: (i) an annotated function relevant to N metabolism, water transport, stomatal regulation, hormone signaling or carbon assimilation; (ii) tissue appropriate expression in the JGI Plant Gene Atlas and SorghumBase [80, 81], which cover 33 tissues and developmental stages, including nitrate and urea treatment series [38]; and (iii) orthology to functionally characterized genes in *Arabidopsis*, rice or maize. All genes in the annotation windows are listed in Additional file 1: Table S8, and the priority candidates in Table 5. Tissue specific $\log_2$ FPKM profiles were visualized with TBtools [82].

Haplotype based analysis of associated loci

For each significant SNP, WGRS markers within ±13.78 kb were extracted from the imputed matrix and coded as 0/1/2, with the major allele coded as 0. Haplotypes were constructed for each individual, and haplotypes with <5% frequency or fewer than five carriers were grouped as residual and excluded. The remaining haplotypes were labelled H1, H2, etc., according to their frequency.

Haplotype phenotype associations were tested using the model Y = HapID + PC1 + PC2 + PC3 with genotypic BLUEs. HapID effects were tested using type II ANOVA, followed by Tukey HSD comparisons with emmeans and multcomp [83]. P values across all locus × trait tests were adjusted using both the Benjamini–Hochberg (BH) method [84] and Bonferroni correction, and both adjusted values were reported. Because rare haplotype filtering can change the population composition, the analyzed and excluded individuals were recorded for each locus. For the recurrent chromosome 9 locus, a separate analysis was performed for its three related traits, including a median joining haplotype network based on the 27.6 kb window (pegas; [85]). Boxplots and forest plots of haplotype effect sizes were also generated [86].

Results

Two distinct gene pools and gene level mapping resolution

Following quality filtering, 1,752,624 high confidence SNPs were kept and the panel divided into two clearly separated gene pools. The unrooted phylogeny showed that the Sb×Sh RILs formed a distinct and well supported clade which branched off from the *S. bicolor* cultivars and landraces (Fig. 1B), and PC1 to PC3 separated the RILs as a tight cluster from the more widely scattered cultivated accessions (Fig. 1A). The ancestry estimation revealed a clear split. Cross entropy was lowest at K = 5, with K = 4 showing a nearly identical value, while all 18 introgression lines formed one cluster (mean ancestry = 0.999), and the 148 cultivated entries were distributed among the remaining four, with 93 showing a maximum ancestry coefficient > 0.70 (Fig. 1C). Marker density varied across chromosomes, with > 4,000 SNPs/Mb in distal regions of chromosomes 1, 8, 9 and 10 and < 1,150 SNPs/Mb in centromeric regions of chromosomes 2, 3 and 7 (Fig. 2A); genome-wide LD half-decay was 13.78 kb, which was used as the ± 13.78 kb window for candidate-gene and haplotype analyses (Fig. 2B).

Contrasting genetic architecture of biomass and grain based efficiency traits

The BLUE values were calculated for the main agronomic traits (DF, PlHt, DM, DBM and Gyld) and for the RUE indices (WUE_DBM, WUE_Gyld, PFPAN_DBM and PFPAN_Gyld), along with the corresponding values for Combination 1 (WUE_DBM_1 and PFPAN_DBM_1), under both environmental combinations (Additional file 1: Table S6). For most of the morphological traits the distributions were almost symmetrical (skewness ranging from − 0.39 to 0.77, except DF and DBM in Combination 2, with 1.05 and 2.02, respectively) while the RUE indices were right-skewed, the biomass-based ones being strongly so (1.11 to 2.30) and the grain-based ones to a moderate extent (0.48 to 0.56), and showing wide genetic ranges (Additional file 2: Figs. S1, S2).

For nearly all the traits, under both combinations, the effects of genotype, environment, the block within the environment and the interaction between genotype and environment were highly significant ($p < 0.001$) (Additional file 1: Table S4). In Combination 1, the environment F values exceeded the genotype F values 38- to 224-fold (for instance, F_{Env} = 3,610.5 compared to F_G = 54.3 for DF_1), which shows that IT21 and the Indian environments are indeed contrasting and therefore supports the use of the multi-environment BLUE strategy. In Combination 2 the contrast was much smaller, and for PFPAN_DBM the genotype F value exceeded that of the environment (14.76 vs 11.64). The genotype by environment interaction was also significant ($p < 0.001$) for all six of the RUE traits by combination variables (Additional file 1: Table S4).

Table 2
Mean, range, variance components, coefficients of variation and broad-sense heritability of agronomic and RUE traits in the + H_2O + N treatment, from the combined multi-environment analyses.

Trait	Mean	Range	CV (%)	σ^2_G	$\sigma^2_{G\times E}$	σ^2_E	GCV (%)	PCV (%)	H^2
Combination 1 (IT21 + IND22 + IND23)									
DF_1	76.63	51.00–164.0	3.17	87.88	40.08	5.91	12.23	13.25	0.85
PlHt_1	242.2	90.00–480.0	3.98	1,621.59	709.9	93.04	16.63	17.94	0.86
DM_1	40.22	21.93–65.62	8.27	22.20	6.00	11.05	11.71	13.10	0.80
DBM_1	11,279.06	981.3–67,949.48	25.29	38,070,000	22,790,000	8,139,000	54.70	61.60	0.79
WUE_DBM_1	27.29	1.28–257.4	36.76	453.4	353.8	100.6	78.04	90.04	0.75
PFPAN_DBM_1	73.99	7.01–485.4	26.81	1,897.70	1,176.56	393.4	58.88	66.43	0.79
Combination 2 (IND22 + IND23)									
DF	71.08	51.00–164.0	2.17	78.86	36.69	2.38	12.49	13.95	0.80
PlHt	226.8	122.0–427.8	3.72	881.8	478.5	71.28	13.09	14.99	0.76
DM	43.74	23.00–65.62	8.05	17.44	12.79	12.39	9.55	12.47	0.59
Gyld	2,782.38	833.3–10,000.00	7.64	406,900	746,000	45,152.62	22.93	32.18	0.51
DBM	7,305.94	2,241.03–24,290.34	14.87	7,823,000	2,981,000	1,181,000	38.28	43.02	0.79
WUE_DBM	4.97	1.28–19.16	17.32	3.63	1.05	0.74	38.37	42.77	0.81
WUE_Gyld	1.89	0.48–5.71	7.93	0.18	0.33	0.02	22.45	31.48	0.51
PFPAN_DBM	42.59	12.11–151.8	16.12	267.3	86.20	47.14	38.39	42.84	0.80
PFPAN_Gyld	16.21	4.50–54.05	7.75	13.58	24.43	1.58	22.74	31.80	0.51

CV, coefficient of variation; σ^2_G, genotypic variance; $\sigma^2_{G\times E}$, genotype × environment variance; σ^2_E, residual variance; GCV and PCV, genotypic and phenotypic coefficients of variation; H^2, broad-sense heritability = $\sigma^2_G/(\sigma^2_G + \sigma^2_{G\times E}/e + \sigma^2_E/(e\cdot\bar{r}))$, where e is the number of environments and $\bar{r}$ the harmonic-mean replication. Heritability classes follow Robinson et al. [47]: *high ($H^2 \geq 0.60$), moderate (0.30–0.60) and low ($H^2 < 0.30$).*

The heritability patterns (Table 2) constitute the main phenotypic finding and divide the RUE indices into two groups. The indices based on biomass were highly heritable, ranging from $H^2 = 0.75$ (WUE_DBM_1) to 0.81 (WUE_DBM), and showed high variability (GCV = 38–78%), the PCV/GCV ratios being between 1.11 and 1.15. It follows that the phenotype is mostly governed by genetic factors and the range of genotypes is wide enough to allow selection to produce a substantial shift. The agronomic traits DF, PlHt and DM showed a similar pattern ($H^2 = 0.80$–0.86 in Combination 1 and 0.59–0.80 in Combination 2). The grain-based indices were different. For Gyld, WUE_Gyld and PFPAN_Gyld, H^2 was 0.51, because their G × E variance exceeded the genotypic variance (Table 2); within each Indian environment, however, their heritability was high ($H^2 = 0.96$–0.97; Additional file 1: Table S5).

What the RUE indices measure

The multivariate trait structure (Fig. 3) provides the framework for interpreting the association results. In Combination 1, PC1 and PC2 accounted for 49.4% and 28.3% of the variance respectively (Fig. 3A). PC1 was associated with biomass accumulation since DBM_1, WUE_DBM_1 and PlHt_1 all had positive loadings (0.73 to 0.97) and their vectors were very closely aligned. PC2 mainly showed phenological characteristics, with DM_1 having a positive loading (0.87) and DF_1 (− 0.84) as well as PFPAN_DBM_1 (− 0.33) showing a negative loading to a lesser degree. In Combination 2, PC1 and PC2 explained 50.7% and 23.8% of the variance (Fig. 3B). PC1 reflected overall productivity, with DBM, Gyld and the four RUE indices (0.78 to 0.81). PC2 separated out a group of biomass related traits (DBM, WUE_DBM, PFPAN_DBM and PlHt; loadings 0.48 to 0.58) from a group of grain related traits (Gyld, WUE_Gyld and PFPAN_Gyld; −0.52 to − 0.53). The fact that the vectors were closely aligned within each group shows that there are high correlations among the traits derived from the same yield component. DM was almost in the opposite direction to the grain related traits, which shows that there is an inverse relationship between dry matter content and grain allocation.

In the pooled analysis of the 15 trait × combination variables, PC1 and PC2 accounted for 44.9% and 24.4% of the variance (Fig. 3C). As before, the biomass and grain related groups were separated, the biomass based traits had positive loadings on PC2 (0.19 to 0.53) and the grain based traits negative ones (– 0.47 to − 0.49). This partial separation shows that the production of vegetative biomass and the efficiency of reproductive resource use are not completely linked, and therefore it is reasonable to treat them as separate targets in the association analysis. The traits measured under both combinations were found in similar areas (DM and DM_1, DF and DF_1, PlHt and PlHt_1 and, to a lesser extent, DBM and DBM_1), which indicates that the trait patterns are consistent across the different environments. The Sb×Sh introgression lines were in a limited area of the ordination, situated towards positive PC2, that is, towards higher dry matter content and plant height and moved away from flowering time and the grain related traits (Figs. 1, 3C).

Model calibration and mappable traits

To reduce the number of false positives, the statistical fit of each association model was assessed for every trait using λ_{GC} (Fig. 4), and only those trait × model combinations that were well calibrated were kept (Table 3). The biomass-based WUE traits were well calibrated regardless of the three algorithms used (λ_{GC} = 0.92–1.08). For the biomass based PFPAN traits, only MLMM was acceptable (with λGC values of 1.148 and 1.132), since BLINK exhibited residual inflation (with λGC values of 1.36–1.37) and FarmCPU showed over correction (with λGC values of 0.77–0.81).

Table 3
Genomic inflation factors (λ_{GC}) and trait-specific model selection across the three multi-locus association algorithms (BLINK, FarmCPU and MLMM) for the six RUE trait × combination variables.

Trait	BLINK (λ_{GC})	FarmCPU (λ_{GC})	MLMM (λ_{GC})	Models retained
WUE_DBM_1	1.078	1.062	0.983	BLINK, FarmCPU, MLMM
PFPAN_DBM_1	1.369	0.810	1.148	MLMM
WUE_DBM	0.927	0.922	1.073	BLINK, FarmCPU, MLMM
PFPAN_DBM	1.357	0.772	1.132	MLMM
WUE_Gyld	—	—	—	No significant MTA
PFPAN_Gyld	—	—	—	No significant MTA

λ_{GC}, genomic inflation factor [72]; *MTA, marker–trait association. Models with $\lambda_{GC} \geq 1.35$ (residual inflation) or $\lambda_{GC} \leq 0.90$ (over-correction) were excluded from downstream analysis.*

Following the filtering process, four of the six RUE trait × combination variables WUE_DBM_1, WUE_DBM, PFPAN_DBM_1 and PFPAN_DBM showed significant associations. None of the SNPs exceeded the Bonferroni threshold in any of the models for WUE_Gyld or PFPAN_Gyld. This is consistent with the lower across environment heritability of the two grain-based indices (Table 2) and with a polygenic, environment dependent architecture whose individual effects fall below the detection power of a panel of this size.

Twenty-two genomic regions, and one that recurs

The retention of significant SNP–trait associations after λGC-based model filtering resulted in 29 such associations at 27 different SNPs among the four mappable RUE variables. These associations are found in 31 rows specific to each model in Table 4 and lie within 22 non-overlapping genomic regions when the ± 13.78 kb merge rule is applied (Table 4 and Fig. 4). The associations were spread across seven of the ten chromosomes, chromosome 9 being the one with the highest number of signals (7 SNPs), then chromosomes 2 and 5 with 5 SNPs each. The full list of significant SNPs prior to λ_{GC} filtering, together with links to the Ensembl Plants browser, is provided in Additional file 1: Table S7.

With respect to WUE_DBM_1, ten SNPs were found in eight regions on chromosomes 2, 4, 6, 9 and 10. The most significant association was observed at S10_8605850 (FarmCPU, $p = 1.88 \times 10^{-14}$). The largest single SNP effects were noted at S04_372161 (PVE = 74.9%, BLINK), S09_57598861 (41.2%, MLMM) and S02_20342977 (40.7%, MLMM), together with additional signals at S06_14836893 (23.2%), S09_8412463 (10.8%), S09_55177729 and S02_70155056.

For WUE_DBM, 17 SNPs in 14 regions were identified on chromosomes 1, 2, 4, 5, 6, 9 and 10. The notable characteristic was a cluster of three SNPs on chromosome 9 between 54.99 and 56.55 Mb (S09_54994173, S09_55037491 and S09_56548597). This cluster contained the most significant association for the trait, S09_55037491 ($p = 1.13 \times 10^{-15}$, PVE = 54.5%, BLINK), and the one with the largest effect, S09_56548597 (PVE = 73.8%, MLMM). Moderate effect signals were observed at S04_51635502 (24.3%, BLINK), S06_2580708 (15.9%), S01_56842887 (6.9%), S10_48413013 (6.2%), which is near an aquaporin gene, and S02_2597978 (3.4%).

For the two PFPAN traits, MLMM identified one very strong signal at S09_55037491 (PFPAN_DBM: $p = 1.13 \times 10^{-14}$, PVE = 73.8%; PFPAN_DBM_1: $p = 1.33 \times 10^{-13}$, PVE = 73.9%). The minor allele at this SNP is rare (MAF = 0.056 among the 162 genotypes analyzed, i.e. approximately nine lines; Additional file 1: Table S7). The same SNP was also found independently by BLINK and FarmCPU in the case of WUE_DBM. Thus, three traits and three algorithms point to the same variant, the most reproducible signal in the present study; detection in both environmental combinations, however, is not independent replication, because Combination 1 includes both environments of Combination 2. This degree of convergence is to be expected due to the common biomass numerator among these traits and therefore does not constitute independent evidence of pleiotropy; the single SNP PVE remains an upper limit in a panel of this size.

Table 4
Marker trait associations (MTAs) for the RUE traits identified by multi-locus association mapping after filtering on the genomic inflation factor.

Trait	Model	SNP	Chr.	Position (bp)	P-value	Allele effect	PVE (%)
PFPAN_DBM	MLMM	S09_55037491	9	55,037,491	× 10 − 14	−23.717	73.80
PFPAN_DBM_1	MLMM	S09_55037491	9	55,037,491	× 10 − 13	−22.559	73.94
WUE_DBM	FarmCPU	S01_3770917	1	3,770,917	× 10 − 9	−0.692	< 0.01
WUE_DBM	FarmCPU	S01_56842887	1	56,842,887	× 10 − 9	−0.709	6.90
WUE_DBM	FarmCPU	S02_2596347	2	2,596,347	× 10 − 11	−0.854	< 0.01
WUE_DBM	FarmCPU	S02_2597978	2	2,597,978	× 10 − 10	−0.258	3.41
WUE_DBM	BLINK	S04_51635502	4	51,635,502	× 10 − 11	−1.251	24.31
WUE_DBM	FarmCPU	S04_51635502	4	51,635,502	× 10 − 10	−0.876	2.16
WUE_DBM	BLINK	S05_2886668	5	2,886,668	× 10 − 10	−0.689	3.48
WUE_DBM	FarmCPU	S05_2933660	5	2,933,660	× 10 − 9	−0.633	4.93
WUE_DBM	FarmCPU	S05_6703509	5	6,703,509	× 10 − 12	−1.083	< 0.01
WUE_DBM	FarmCPU	S05_6708487	5	6,708,487	× 10 − 10	2.406	< 0.01
WUE_DBM	FarmCPU	S05_7734857	5	7,734,857	× 10 − 9	−0.517	< 0.01
WUE_DBM	FarmCPU	S06_2557367	6	2,557,367	× 10 − 8	−0.525	< 0.01
WUE_DBM	FarmCPU	S06_2580708	6	2,580,708	× 10 − 10	−0.042	15.92
WUE_DBM	FarmCPU	S06_39951727	6	39,951,727	× 10 − 8	−0.453	< 0.01
WUE_DBM	FarmCPU	S09_54994173	9	54,994,173	× 10 − 15	0.285	< 0.01
WUE_DBM	BLINK	S09_55037491	9	55,037,491	× 10 − 15	−2.408	54.47
WUE_DBM	FarmCPU	S09_55037491	9	55,037,491	× 10 − 12	−1.412	0.01
WUE_DBM	MLMM	S09_56548597	9	56,548,597	× 10 − 14	−2.871	73.81
WUE_DBM	BLINK	S10_48413013	10	48,413,013	× 10 − 8	0.815	6.22
WUE_DBM_1	MLMM	S02_20342977	2	20,342,977	× 10 − 8	−12.707	40.66
WUE_DBM_1	FarmCPU	S02_70155056	2	70,155,056	× 10 − 10	−6.909	< 0.01
WUE_DBM_1	FarmCPU	S02_70156467	2	70,156,467	× 10 − 10	4.897	< 0.01
WUE_DBM_1	BLINK	S04_372161	4	372,161	× 10 − 9	−21.999	74.86
WUE_DBM_1	FarmCPU	S06_14836893	6	14,836,893	× 10 − 8	−5.197	23.19
WUE_DBM_1	FarmCPU	S09_8412463	9	8,412,463	× 10 − 8	4.844	10.81
WUE_DBM_1	FarmCPU	S09_55177729	9	55,177,729	× 10 − 8	−5.744	4.43
WUE_DBM_1	FarmCPU	S09_55196670	9	55,196,670	× 10 − 10	1.059	< 0.01
WUE_DBM_1	MLMM	S09_57598861	9	57,598,861	× 10 − 10	−16.159	41.24
WUE_DBM_1	FarmCPU	S10_8605850	10	8,605,850	× 10 − 14	−5.633	1.98

The amount of phenotypic variance accounted for by the focal SNP (an upper bound estimate for a panel of this size) is given; only SNPs that remained after the λ_{GC} *filtering process are included (*Table 3*). In the case of SNPs detected by more than one model, each model-specific entry retains the original allele-effect estimates; high-confidence associations are those SNPs detected by two or more retained models (S04_51635502 and S09_55037491).*

Candidate genes: a nitrate responsive regulator, an aquaporin and a putative C4 enzyme

The LD defined windows of the significant loci contain 52 protein coding genes (Additional file 1: Table S8), eight of which were selected as priority candidates based on considerations of function, orthology, and tissue appropriate expression (Table 5).

The SNP on chromosome 9 at position S09_55037491 is located within the gene Sobic.009G200500, which codes for HOMEOBOX PROTEIN HAT3.1, a class III HD-Zip transcription factor that has been annotated for its ability to bind DNA, chromatin and metal ions as well as for its role in the negative epigenetic regulation of gene expression. According to the SorghumBase Gene Atlas, this gene is part of the shoot nitrate treatment specific co-expression subnetwork, linking the most prominent genetic signal in the study to nitrate-responsive transcriptional control. In the neighboring genomic window at S09_54994173, 43

kb away, the gene Sobic.009G200200 encodes α,α-trehalose phosphate synthase (TPS), the enzyme responsible for producing the carbon-nitrogen signaling metabolite trehalose 6-phosphate.

The signal S10_48413013 on chromosome 10 is situated 9.3 kb away from Sobic.010G164100, a gene that codes for the aquaporin NIP1-4 (nodulin-26-like intrinsic protein), a transmembrane channel which is noted in the Gene Atlas as being specific to the root. Since some aquaporins are also capable of transporting ammonia or urea together with water [24, 25], this protein family lies at a potential point of convergence between water and nitrogen uptake.

The remaining WUE_DBM windows contain Sobic.002G028100 (alpha-amylase inhibitor 5/cationic peroxidase SPC4; hydrogen peroxide catabolism) and Sobic.005G032400 (Set1/Ash2 histone methyltransferase complex subunit ASH2; histone H3K4 methylation), and the window including S05_2933660 contains Sobic.005G032700 (phospholipase A1, DAD1; supply of jasmonate precursors). The S02_20342977 location on chromosome 2 of the WUE_DBM_1 locus is 13.7 kb away from Sobic.002G135750, which is annotated as having L-malate dehydrogenase ($NADP^+$) activity and being involved in the malate metabolic process, a reaction central to the NADP-ME-type C_4 cycle (provisional assignment; see Section 4.4). There are two additional WUE_DBM_1 candidates on chromosome 9: Sobic.009G202900, a histidine-containing phosphotransfer (AHP-like) protein belonging to the cytokinin phosphorelay, and Sobic.009G238000, SUPPRESSOR OF PHYA-105 1 (SPA1), which is a regulator of red/far-red light phototransduction and photomorphogenesis.

Table 5
Priority candidate genes selected from linkage disequilibrium-defined annotation windows based on functional relevance to nitrogen metabolism, water trans hormone signaling and carbon assimilation.

Trait	Model	SNP	Candidate gene	Position (bp)	Strand	Protein/annotation	Biological process (GO)	Gene Atl expressi
WUE_DBM	FarmCPU	S02_2596347	Sobic.002G028100	2,590,650–2,593,117	+	Bifunctional inhibitor/plant lipid transfer protein/seed storage helical domain-containing protein; cationic peroxidase SPC4; alpha-amylase inhibitor 5	Hydrogen peroxide catabolic process	—
WUE_DBM_1	MLMM	S02_20342977	Sobic.002G135750	20,327,359–20,329,281	–	CCHC-type domain-containing protein; zinc finger, CCHC-type superfamily; malate dehydrogenase [NADP] 1, chloroplastic	Malate metabolic process; positive regulation of cytoplasmic translation	—
WUE_DBM	BLINK	S05_2886668	Sobic.005G032400	2,892,613–2,896,122	+	Set1/Ash2 histone methyltransferase complex subunit ASH2	Histone H3-K4 methylation	Down-regulate in seed a grain maturity (imbibed vs. dry)
WUE_DBM	FarmCPU	S09_54994173	Sobic.009G200200	55,004,317–55,013,213	+	α,α-Trehalose-phosphate synthase (UDP-forming)	Trehalose biosynthetic process	—
PFPAN_DBM, PFPAN_DBM_1 (WUE_DBM)	MLMM (BLINK, FarmCPU)	S09_55037491	Sobic.009G200500	55,030,822–55,038,173	+	HOMEOBOX PROTEIN HAT3.1	Negative regulation of gene expression, epigenetic	Compon of the shoot nitrate-treatmer specific expressi subnetw
WUE_DBM_1	FarmCPU	S09_55177729	Sobic.009G202900	55,183,741–55,190,886	–	Histidine-containing phosphotransfer protein	Phosphorelay signal transduction system; cytokinin-activated signaling pathway	—
WUE_DBM_1	MLMM	S09_57598861	Sobic.009G238000	57,602,900–57,611,460	–	Protein SUPPRESSOR OF PHYA-105 1 (SPA1)	Red, far-red light phototransduction; photomorphogenesis	—
WUE_DBM	BLINK	S10_48413013	Sobic.010G164100	48,401,814–48,403,745	+	AQUAPORIN NIP1-4	Transmembrane transport	Root-specific expressi

The priority candidates were selected based on (i) the annotated function relating to RUE-related processes, (ii) appropriate tissue expression as shown in the SorghumBase Gene Atlas and (iii) orthology with genes that have been functionally characterized in Arabidopsis, rice or maize. The gene identifiers correspond to

the Sorghum bicolor BTx623 v3.1.1 annotation, and the molecular-function and cellular-component assignments, the full list of Gene Ontology accessions, and all the genes located within the annotation windows are provided in Additional file 1: Table S8.

Expression profiles of the candidate genes

The log2 FPKM profiles across 33 tissues and developmental stages (Additional file 2: Figs. S3–S5), together with hierarchical clustering of the priority candidates (Fig. 5), were used to evaluate these assignments. WUE_DBM_1 candidates were most abundant in roots, with higher log2 FPKM values, across all developmental stages and tested conditions, including control, nitrate, urea and fertilizer treatments, compatible with a potential role in water uptake (Additional file 2: Fig. S3); this root enrichment was not evident for the three WUE_DBM_1 priority candidates shown in Fig. 5 (Sobic.002G135750, Sobic.009G202900 and Sobic.009G238000). The WUE_DBM candidates were more widely distributed, showing increased expression in the seeds at dry maturity and in the shoots when subjected to nitrate or urea treatment (Additional file 2: Fig. S4). The genes within the chromosome 9 PFPAN region showed a limited pattern of expression that was confined to the internodes (both mature and at floral initiation), the growing internodes and the seeds, a pattern compatible with stem storage and remobilization (Additional file 2: Fig. S5), although HAT3.1 itself was expressed in all tissues (Fig. 5). The WUE_DBM candidates on chromosome 1 (Sobic.001G291500 and Sobic.001G291600) were preferentially expressed in the leaves and shoots at the stage of floral initiation and at grain maturity, and they also responded to nitrate and urea treatment.

Hierarchical clustering (Fig. 5) divided the priority set into two groups. One group comprised broadly and moderately expressed genes: TPS (Sobic.009G200200), ASH2 (Sobic.005G032400), SPA1 (Sobic.009G238000) and HAT3.1 (Sobic.009G200500). The other group comprised genes with low expression across the atlas: the AHP-like protein (Sobic.009G202900), the NIP1-4 aquaporin (Sobic.010G164100), Sobic.002G028100 and Sobic.002G135750, of which only NIP1-4 showed tissue-preferential (root) expression. Both clusters contained WUE_DBM and WUE_DBM_1 candidates, so the expression clusters did not correspond to trait classes.

Haplotype dissection distinguishes candidate haplotype tags from single SNP signals

Regional haplotype analysis was attempted at all 27 significant SNPs. After applying regional quality control and filtering out the rare haplotypes, 15 of these yielded at least two common haplotypes, resulting in 17 locus × trait tests each with two to four common haplotypes (Additional file 1: Table S9); the other 12 did not yield two or more haplotypes meeting the frequency (≥ 5%) and carrier (≥ 5) thresholds and could not be tested. Two of the tests were found to be significant following both BH and Bonferroni adjustments, one was nominally significant (unadjusted $p < 0.05$) and 14 were not significant. The number of genotypes entering each test varied from 32 to 113 out of 166 (the median being 64) and tended, although not significantly, to decrease as the number of window SNPs used to define the haplotypes (ranging from 25 to 109) increased (Spearman $\rho = -0.41$ across the 14 loci, $p = 0.14$).

The most prominent haplotype effect was observed at position S09_55177729 for WUE_DBM_1 (Fig. 6A). Twenty-eight regional SNPs identified two common haplotypes, H1 (carried by 94 individuals in the entire panel; frequency 0.566) and H2 (carried by 12 individuals; frequency 0.072), together with 60 genotypes possessing rare combinations. Altogether, 104 genotypes had both a common haplotype and a BLUE for the trait and were included in the model (H1, 93; H2, 11). The association was extremely strong ($p = 1.55 \times 10^{-11}$; BH and Bonferroni-adjusted $p = 2.63 \times 10^{-10}$). Carriers of H2 had 2.7 times higher WUE_DBM_1 than carriers of H1 (PC adjusted emmean ± SE = 51.44 ± 3.97 versus 18.93 ± 1.18 kg DBM ha^{-1} mm^{-1}; Tukey groups b versus a), and, taking ± 1 SE of both means, the fold change ranges from about 2.4 to 3.1. The window includes Sobic.009G202900, the AHP-like cytokinin signaling component (Table 6), together with four other gene models (Additional file 1: Table S8).

Table 6
Haplotype characterization, effect sizes and Tukey HSD groupings at the two loci with significant haplotype–trait associations after Benjamini–Hochberg and Bonferroni correction.

SNP	Trait	Candidate gene	Hap. ID	n (panel/ analyzed)	Frequency	emmean ± SE	Tukey group/ direction
S09_55177729	WUE_DBM_1	Sobic.009G202900 (AHP-like)	H1	94/93	0.566	± 1.18	a, unfavorable
S09_55177729	WUE_DBM_1	Sobic.009G202900 (AHP-like)	H2	12/11	0.072	± 3.97	b, favorable
S09_55177729	WUE_DBM_1	Sobic.009G202900 (AHP-like)	Rare	60/—	0.361	—	Excluded (< 5%)
S05_2933660	WUE_DBM	Sobic.005G032700 (DAD1)	H1	42/41	0.253	± 0.27	a, intermediate
S05_2933660	WUE_DBM	Sobic.005G032700 (DAD1)	H2	31/29	0.187	± 0.40	a, unfavorable
S05_2933660	WUE_DBM	Sobic.005G032700 (DAD1)	H3	9/8	0.054	± 0.47	b, favorable
S05_2933660	WUE_DBM	Sobic.005G032700 (DAD1)	Rare	84/—	0.506	—	Excluded (< 5%)

The emmean is the principal component adjusted estimated marginal mean from the linear model Y = HapID + PC1 + PC2 + PC3. Haplotypes within a locus that share a Tukey letter do not show a significant difference at the 0.05 level of α. The 'Direction' classification designates each haplotype as favorable (i.e. having the highest emmean), unfavorable (i.e. having the lowest emmean) or intermediate. Rare haplotypes (those with a frequency of less than 5% or with fewer than five carriers) were combined and then excluded from the association test.

Another association that stayed significant after both BH and Bonferroni adjustments was observed at S05_2933660 (chromosome 5, 2.93 Mb) for WUE_DBM (Fig. 6B). Forty-six regional SNPs defined three common haplotypes, namely H1 (42 carriers in the entire panel; 0.253), H2 (31; 0.187) and H3 (9; 0.054), and 78 genotypes were included in the model (41 for H1, 29 for H2 and 8 for H3). The effect of the haplotype was significant ($p = 2.42 \times 10^{-3}$; BH-adjusted $p = 2.06 \times 10^{-2}$; Bonferroni-adjusted $p = 4.11 \times 10^{-2}$). According to the Tukey HSD test, H3 (5.55 ± 0.47 kg DBM ha^{-1} mm^{-1}, group b) was separated from H1 (4.30 ± 0.27, group a) and H2 (2.98 ± 0.40, group a), showing a 1.86-fold difference between the favorable and the unfavorable haplotypes, within a region that includes Sobic.005G032700 (DAD1). Sobic.005G032400 (ASH2) is located 37.5 kb away from S05_2933660 and is therefore outside the ± 13.78 kb interval and so is not a candidate at this locus; it is the candidate at the adjacent SNP S05_2886668.

The locus S09_55037491, which was identified by three different algorithms when considering three traits and had a single SNP PVE ranging from 54.5 to 73.9 per cent, showed no significant association with haplotypes for any of the traits (with raw p values of 0.52, 0.55 and 0.69; Additional file 2: Fig. S6). The 27.6 kb region included 53 SNPs that had passed the quality control and these resolved three common haplotypes (H1, n = 58; H2, 18; H3, 16), and the PC-adjusted marginal means followed the anticipated order (H3 > H2 > H1 for the PFPAN traits) although this did not reach statistical significance. Thus, the single SNP effect estimate and the haplotype block test are in disagreement. A forest plot of the haplotype effect sizes across the tests and a median joining haplotype network for this region are given in Additional file 2: Figs. S7, S8, respectively.

Across all 17 tests, only one of the 18 genotyped introgression lines carried a common haplotype at S09_55177729, corresponding to the unfavorable H1 haplotype. In contrast, the favorable haplotypes S09_55177729 H2 and S05_2933660 H3 were found exclusively in cultivated lines. The introgression lines mainly carried rare haplotype combinations that were grouped into a residual category and excluded from the association tests (Additional file 1: Table S9).

Discussion

Biomass drives agronomic RUE, which is what these indices measure

The stronger genetic signal for biomass based indices provides a more tractable basis for genomic dissection and selection than the corresponding grain based indices. Biomass-based RUE is highly heritable ($H^2 = 0.75–0.81$), having a GCV of 38–78% and PCV/GCV ratios close to one, a combination which Robinson et al. [47] and Burton [48] found to be indicative of a large genetic advance when selecting directly; these figures are similar to those recorded for biomass and yield characteristics in sorghum [87] and wheat [88]. The grain-based indices showed an H^2 of 0.51, in agreement with the findings in sorghum [13] and wheat [34], and with the well known difficulty of detecting large effect loci for grain yield [89]. The angle between the biomass and grain based trait vectors in the ordination (Fig. 3) indicates that the conversion of resources between the vegetative and reproductive phases is only partly decoupled, with harvest index (not measured here) as the likely link [90]. Hence, the two bases are not interchangeable targets. The lack of significant associations between WUE_Gyld and PFPAN_Gyld is in line with the lower heritability and greater G × E variance of these indices in this kind of panel.

The fact that WUE and PFPAN are collinear is a general feature of agronomic efficiency indices and the relationship between such indices and their yield numerator is well known in the field of cereal agronomy [11, 91]. It does, nevertheless, place limitations on how the results can be interpreted. When S09_55037491 is recovered for WUE_DBM, PFPAN_DBM and PFPAN_DBM_1, one locus is found three times for a single underlying quantity. These data provide no evidence that it simultaneously governs water-use and nitrogen-use physiology, and it would therefore be an artefact of the index construction to describe it as a major locus across different efficiencies. It is better interpreted as a candidate locus for biomass-based resource capture. This is a legitimate and worthwhile breeding objective: when water and nitrogen are in short supply, the plant which achieves greater biomass per unit of input is the more productive one, regardless of the mechanism involved.

Wild-relative introgression and the allelic space for RUE

The initial question was about the contribution of introgressions from *S. halepense*. The panel structure sets this precondition. The phylogeny, PCA, ancestry and kinship all show that the Sb×Sh RILs belong to a gene pool that is separate from that of cultivated sorghum (Figs. 1, 2C), something which is in agreement with the higher levels of SNP and structural diversity and the enrichment of private genes relevant to RUE that have previously been reported for this material [20]. *S. halepense* is a rhizomatous perennial possessing deep roots and strong regrowth [17, 18].

The association results, however, do not show a broad functional contribution of the introgressed material. None of the prioritized candidates can yet be attributed to *S. halepense*-derived alleles, and the favorable haplotype at S09_55177729 (H2; 2.7-fold difference in WUE_DBM_1 at a frequency of 7%) was carried only by cultivated lines (Table 6; Additional file 1: Table S9). This low-frequency, large-effect variant therefore represents standing variation within cultivated sorghum that panel-wide phenotypic selection is unlikely to capture, rather than a wild-relative introgression of the kind that has broadened the genetic base of rice [21, 92], maize [23] and wheat [22].

In each model, population structure (the first three principal components) was included as a covariate. Although this helps to control for confounding it also means that any signal which is perfectly confounded with the Sb/Sb×Sh split is eliminated, so the analysis is conservative when it comes to effects specific to introgression. The design also makes it impossible to separate out the contribution of an introgressed segment from that of the *S. bicolor* background in which it is situated. The allele-frequency comparison (Fig. 7) refines this picture. At a number of the associated SNPs the minor allele was missing from the introgression lines and thus some of the signal is based on cultivated variation, while at only a small number of SNPs the minor allele was found in higher amounts in the introgressed material. Among these are S04_372161, which has the highest single SNP PVE in the study and whose minor allele is essentially unique to the introgression lines, and S09_55177729, which marks the strongest haplotype effect. Therefore, wild-relative germplasm appears to contribute differentiated alleles at only a small number of associated loci rather than a general expansion of the allelic space for these traits. There are two cautions to note. The SNPs that are enriched in the introgression material are also among the most differentiated markers in the set of associated markers, so an

association at such a marker could be due either to a causal introgressed allele or to residual confounding with sub-population membership. Furthermore, the assertion that a specific *S. halepense* allele is responsible for a specific gain still requires a near isogenic comparison.

The LD decay distance of 13.78 kb (Fig. 2B) agrees with that found in global sorghum diversity panels [31, 32, 93] and is similar to the value observed in maize [94, 95], but it is one to two orders of magnitude shorter than in wheat [96] or barley [97]. Since each significant SNP identifies a small interval that includes only a few genes, the mapping resolution is improved and this is the main reason why a panel of this size was able to yield biologically interpretable candidates.

A candidate nitrate responsive transcriptional node on chromosome 9

The signal that was most reproducible in the study, S09_55037491, is located in Sobic.009G200500, a gene that encodes HAT3.1, a member of the class III HD-Zip transcription factors. HD-Zip III proteins act as master regulators of meristem maintenance, the organization of shoots, vascular patterning and organ polarity [98, 99] and they combine hormonal and environmental inputs into decisions regarding growth [100, 101]. According to the SorghumBase Gene Atlas, Sobic.009G200500 is included in the shoot nitrate treatment specific co-expression subnetwork [38]. A transcription factor which coordinates shoot growth and is part of a nitrate responsive module could plausibly function as a regulator of biomass accumulation per unit of nitrogen supplied, since this is precisely the quantity that the relevant indices measure.

The signal need not reflect a single gene. At about 18 kb from HAT3.1, within the region covered by the adjacent WUE_DBM SNP S09_54994173, the gene Sobic.009G200200 codes for trehalose-6-phosphate synthase. Trehalose 6-phosphate is a signaling metabolite which indicates the status of sucrose and allows growth to proceed accordingly [27], and altering its levels has been shown to increase grain yield in maize and wheat [102, 103]. A putative transcriptional regulator of nitrate-responsive shoot growth located about 18 kb from an enzyme that produces a carbon-status signal forms a plausible gene cluster that links carbon and nitrogen signaling to biomass allocation, an arrangement that has been reported for several cereal yield-stability QTLs [89, 104]. To find out whether one or both two genes is responsible and to determine their individual contributions, finer mapping and functional tests are required; the co-location remains a hypothesis and not a conclusion.

Where water and nitrogen converge: shared molecular machinery

The second question, whether genomic regions affect both water use and N use, cannot be answered from the coincidence of WUE and PFPAN signals. It can, however, be addressed at the level of gene function.

The best example is Sobic.010G164100 (NIP1-4). Aquaporins mediate the movement of water across membranes and influence both root hydraulic conductance and stomatal behavior [105, 106]; some aquaporins also transport ammonia and urea [24, 25], and the NIP proteins are capable of transporting a variety of small solutes [26]. According to the Gene Atlas, Sobic.010G164100 is expressed in the root (Fig. 5; Additional file 2: Fig. S4), so it is located at the main site of water and nutrient uptake and thus could be involved in coordinating the acquisition of both of these resources. It has not yet been determined whether NIP1-4 itself transports nitrogen compounds; if it does, then this would be an example of water–N convergence in the sense meant here, that is, not two efficiencies relating to one interval, but a single gene product transporting two substrates.

There is also a metabolic convergence point. Trehalose 6-phosphate (through Sobic.009G200200) acts as the standard indicator of carbon status, and since carbon status determines whether assimilated nitrogen can be employed for growth, the signaling involving trehalose 6-phosphate is situated at the interface between carbon and nitrogen [28]. A third convergence point is hormonal. The gene Sobic.009G202900 codes for an AHP-like phosphotransfer protein which is part of the cytokinin multistep phosphorelay and is involved in the regulation of root architecture, leaf senescence and nitrate-dependent shoot growth [29, 107], cytokinin signaling thus linking nitrogen availability to growth [30]. This gene also lies in the window that showed the strongest haplotype effect in the study, which contains four other gene models (Additional file 1: Table S8); on the combined evidence it is the best-supported candidate, although the haplotype effect cannot be attributed to it specifically.

Two further candidate mechanisms relate to water use. Sobic.002G135750 is noted to have L-malate dehydrogenase ($NADP^+$) activity, referring to the reaction in which the mesophyll chloroplasts of NADP-ME-type C_4 plants, such as sorghum, reduce oxaloacetate to malate and transfer the reducing equivalents to the bundle sheath [108, 109]. Genotypic variation in this process could affect the amount of carbon assimilated per unit of water lost, since this ratio is what defines intrinsic WUE. If confirmed, the association with WUE_DBM_1 (S02_20342977; PVE = 40.7%) would connect a core C4 enzyme to biomass-based WUE in sorghum. However, the lead SNP is rare (MAF = 0.046; Additional file 1: Table S7), and agronomic WUE as measured here does not isolate intrinsic WUE. This gene model also includes a CCHC-type zinc-finger annotation and is assigned to the cytoplasm as a cellular component, so its functional assignment must be regarded as provisional. Moreover, Sobic.002G028100, annotated both as a lipid-transfer/α-amylase-inhibitor-domain protein and as cationic peroxidase SPC4, may belong to the reactive oxygen species (ROS)-processing system involved in the regulation of stomatal aperture [110, 111], and Sobic.005G032700 (DAD1 phospholipase A1) provides linolenic acid for jasmonate biosynthesis [112]. Both of these place oxidative and lipid signaling among the WUE candidates.

Expression data provide only partial support for these assignments. The priority genes separated into a broadly expressed cluster and a low-expression cluster (Fig. 5), a pattern common in plant gene networks [113], but the clusters did not correspond to trait categories, and root-preferential expression was clear only for NIP1-4. Tissue–function agreement is a widely used criterion for prioritizing GWAS candidates in cereals [37, 39], but here it supports prioritization rather than causality.

Haplotypes distinguish candidate tags from single-SNP signals

Haplotype dissection was carried out to determine whether single SNP signals remain at the LD block level and to produce tags for MAS [34–36]. It achieved both objectives, although the two results pointed in opposite directions.

The haplotype test at S09_55177729 reinforced the locus since a moderate single SNP association (PVE = 4.4%, FarmCPU) was replaced by a haplotype association with a BH adjusted p value of 2.63×10^{-10} and a 2.7-fold difference, in a window that includes the AHP-like cytokinin component. The favorable H2 haplotype is rare (7%; 12 carriers, 11 analyzed), which reduces the precision of the effect estimate and makes it prone to upward bias; its low frequency is consistent with, but does not demonstrate, an absence of selection on this combination. At S05_2933660, the 1.86-fold difference lies in a window containing DAD1 and three other gene models (Additional file 1: Table S8), which makes lipid signaling a candidate additional layer of WUE. Both are candidate targets that need to be confirmed in independent populations; haplotype-derived candidate genes similar to these have been reported in wheat [34] and rice [114].

The result from the haplotype test at S09_55037491 was the opposite one and this is the more informative result. Even though the single SNP PVE was between 54.5 and 73.9%, no association was found for any of the three traits when using a haplotype. There are three possible explanations, which are not mutually exclusive. The first is that the PVE is overstated in a panel consisting of 172 genotypes because of the Beavis effect [74, 75] and so the real effect is likely to be considerably smaller than the estimated value. The second is that the focal SNP may already include the causal variation, in which case combining the dense group of surrounding markers into a categorical haplotype factor tends to reduce rather than enhance the signal. The third is that, since there were only 16 to 18 carriers of minor haplotypes, the haplotype ANOVA had less power than the single SNP mixed model test. A fourth is that the single-SNP signal rests on a rare allele (MAF = 0.056; approximately nine lines) in strongly right-skewed traits, so a few extreme genotypes may drive it; its robustness to phenotype transformation should be tested.

The breeding effects therefore vary depending on the locus; at S09_55037491 the SNP itself would be the more suitable diagnostic marker, while at S09_55177729 and S05_2933660 it is the haplotypes that do so; in each of the three cases deployment must be based on independent validation. More broadly speaking, haplotype analysis has functioned in this study as a confirmatory and prioritization tool showing how a signal should be deployed, rather than as an independent means of discovery [115].

Pathways and priorities for breeding resource-efficient sorghum

The third question was about which biological pathways should be pursued further. The evidence presented here identifies four such pathways, ranked in order of the strength of the support. The cytokinin phosphorelay (Sobic.009G202900, AHP-like) has the most comprehensive body of evidence: it is supported by a GWAS association, lies in the window with the largest haplotype effect (2.7-fold) and is expressed, albeit at low levels, in roots and other tissues across developmental stages and N treatments (Fig. 5). Aquaporin mediated transport (Sobic.010G164100, NIP1-4) offers the most plausible mechanistic route to water–N convergence because it is root preferential and belongs to a channel family which includes transporters of nitrogen compounds. Nitrate responsive transcriptional control (Sobic.009G200500, HAT3.1, with Sobic.009G200200/TPS nearby) has the most consistently reproduced association and the clearest connection with N dependent shoot growth, although it has the weakest haplotype support; it is a hypothesis concerning a regulatory node rather than being based on a validated allele. C_4 core assimilation (Sobic.002G135750) would provide a direct mechanistic link to carbon gain per unit of water, but it rests on a single association at a low-frequency SNP (MAF = 0.046) and on annotation that requires curation.

In terms of translation, this creates a hierarchy of possible target genes, each one depending on validation. S09_55037491 is a potential single-SNP MAS target for the capture of biomass-based resources and could be incorporated into elite genetic backgrounds once its effect and its single SNP character have been confirmed. The mid effect loci (PVE = 15–41%) on chromosomes 2, 4, 6 and 9, many of which have candidates that can be interpreted from a mechanistic point of view, are suitable for use in marker-assisted recurrent selection [116]. The two significant haplotypes could be turned into KASP or AmpSeq diagnostic markers for use in marker-assisted backcrossing [117] following independent confirmation. For the grain-based indices WUE_Gyld and PFPAN_Gyld, for which no significant associations were found, genomic selection is the appropriate method, since these indices appear to have a polygenic architecture and genomic prediction has already been applied to other traits in this panel [40, 118, 119].

One agronomic implication deserves emphasis, with its caveat. Because the biomass-based indices are collinear with biomass, selection at a major biomass locus would improve resource capture rather than resource use physiology. Where water and N are limited, that is the outcome that matters, and it avoids the tradeoffs that can arise when efficiency components are selected in isolation [120]. This is particularly relevant to the dual purpose sorghum grown across semi-arid Africa and Asia. Stacking the favorable allele at S09_55037491 with the S09_55177729 and S05_2933660 haplotypes is an promising strategy to evaluate, but its magnitude, additivity and timeline must be established empirically rather than assumed, as pyramiding experience in rice shows [121].

Limitations and perspectives

Some limitations should be considered. The panel size limits statistical power for the polygenic grain based indices and biases the estimated effects of detected loci upwards (the Beavis effect), so the reported single SNP PVE values are upper bounds instead of direct variance estimates. The WUE indices measure productivity per unit of water supplied, that is, agronomic water productivity, not transpiration efficiency or leaf level intrinsic WUE, and PFPAN does not separate N uptake from N utilization. Independent multi-environment trials, ideally with a factorial N × water design, together with larger panels, are needed to confirm the candidate loci and haplotypes reported here. Candidate gene prioritization relied on annotation, orthology and tissue expression, so functional validation of the top priorities, the AHP-like phosphotransfer protein (Sobic.009G202900), HAT3.1 (Sobic.009G200500) and the NIP1-4 aquaporin (Sobic.010G164100), together with curation of the putative NADP-MDH gene model (Sobic.002G135750), through CRISPR–Cas9 knockout [122–124] or allele specific introgression is the critical next step to demonstrate causality. Because field RUE reflects canopy level processes that genotype means do not completely capture, combining these loci with high throughput phenotyping of canopy spectral and thermal properties and of root systems [125] would enable multi-omics RUE prediction models. Finally, although the Sb×Sh sub-population increases the allelic spectrum, controlled experiments are needed to separate the contribution of *S. halepense* introgressions from the *S. bicolor* background and to quantify any agronomic penalty of carrying the favorable alleles in elite backgrounds.

Conclusions

This study integrates multi-environment phenotyping, whole genome association mapping, regional haplotype analysis and public tissue expression evidence to dissect the genetic basis of biomass productivity under contrasting water and nitrogen inputs in sorghum. Biomass-based productivity indices displayed stronger genetic control than grain-based indices, supporting their use as selection targets while recognizing that they largely reflect yield scaled by treatment level inputs rather than intrinsic physiological efficiencies. A recurrent chromosome 9 signal and two significant regional haplotypes provide priority targets for independent validation. Candidate gene analyses additionally identified plausible regulatory and functional pathways involving cytokinin signaling, nitrate responsive growth control, aquaporin-mediated transport, trehalose-6-phosphate metabolism and, provisionally, C4 carbon assimilation. However, expression profiles support candidate prioritization rather than establish causality, and because water and nitrogen inputs were not varied factorially, their individual effects cannot be separated. Most detected variation was associated with cultivated germplasm, with a smaller subset of differentiated alleles enriched in the introgressed material. Overall, the findings provide testable genomic hypotheses and candidate markers for improving sorghum biomass productivity under low-input conditions. Validation in larger populations and factorial water-by-nitrogen experiments will be essential before marker deployment or causal inference.

Abbreviations

A/P
annual × perennial cross
AHP
histidine-containing phosphotransfer protein
ANOVA
analysis of variance
ASH2
Set1/Ash2 histone methyltransferase complex subunit
BH
Benjamini–Hochberg
BLINK
Bayesian-information and linkage-disequilibrium iteratively nested keyway
BLUE
best linear unbiased estimate
CTAB
cetyltrimethylammonium bromide
CV
coefficient of variation
DAD1
DEFECTIVE IN ANTHER DEHISCENCE1
DBM
aboveground dry biomass
DF
days to flowering
DM
dry matter content
FarmCPU
fixed and random model circulating probability unification
FPKM
fragments per kilobase of transcript per million mapped reads
FST
fixation index
GAPIT
Genome Association and Prediction Integrated Tool
GATK
Genome Analysis Toolkit
GCV
genotypic coefficient of variation
GO
Gene Ontology
G × E
genotype × environment interaction
GWAS
genome-wide association study
Gyld

grain yield
H2
broad-sense heritability
HAT3.1
HOMEOBOX PROTEIN HAT3.1
HD-Zip
homeodomain-leucine zipper
IND22
ICRISAT Patancheru, 2022–23 season
IND23
ICRISAT Patancheru, 2023–24 season
IT21
CREA Anzola dell'Emilia, 2021
KASP
Kompetitive allele-specific PCR
LD
linkage disequilibrium
MAF
minor allele frequency
MAS
marker-assisted selection
MLMM
multi-locus mixed model
MTA
marker–trait association
N
nitrogen
NADP-ME
NADP-malic enzyme
NIP
nodulin-26-like intrinsic protein
NUE
nitrogen use efficiency
PC
principal component
PCA
principal component analysis
PCV
phenotypic coefficient of variation
PFPAN
partial factor productivity of applied nitrogen
PlHt
plant height
PVE
phenotypic variance explained
Q–Q
quantile–quantile
QTL
quantitative trait locus
RCBD
randomized complete block design
RIL
recombinant inbred line
ROS
reactive oxygen species
RUE
resource use efficiency
Sb
Sorghum bicolor
Sb×Sh

Sorghum bicolor × S. halepense
SNP
single nucleotide polymorphism
SPA1
SUPPRESSOR OF PHYA-105 1
TPS
trehalose-6-phosphate synthase
W
total seasonal water input
WGRS
whole-genome resequencing
WUE
water use efficiency
λGC
genomic inflation factor.

Declarations

Ethics approval and consent to participate

Not applicable. This study did not involve human participants, human data or human tissue, and no endangered or protected plant species were used. All sorghum germplasm was obtained from the institutional breeding collections of ICRISAT, CREA and Vasantrao Naik Marathwada Krishi Vidyapeeth and was handled in accordance with institutional and national guidelines.

Consent for publication

Not applicable.

Availability of data and materials

All phenotypic data supporting the conclusions of this article are included within the article and its additional files (Additional file 1: Table S6). The whole-genome resequencing variant set analysed in this study was generated and reported previously by Habyarimana et al. [20] and is available from the corresponding author on reasonable request. Gene models and annotations were retrieved from Phytozome v13 (https://phytozome-next.jgi.doe.gov), Ensembl Plants (https://plants.ensembl.org) and SorghumBase (https://www.sorghumbase.org).

Competing interests

The authors declare that they have no competing interests.

Funding

This work received support from the SusCrop ERA-NET project C4Future, which funded the field experiments conducted in Italy and the first year in India. PS was supported by a PhD fellowship for research and training, awarded by MahaJyoti (Government of Maharashtra, India). The funders had no role in the design of the study, in the collection, analysis or interpretation of data, or in writing the manuscript.

Authors' contributions

PS: conceptualization, methodology, investigation, data curation, formal analysis, writing-original draft. NT: visualization, investigation, writing-review and editing. SM: writing-review and editing. HK: writing-review and editing. CH: writing-review and editing, funding acquisition, project administration, resources. GP: writing-review and editing. PR: formal analysis. AKV: formal analysis. EH: conceptualization, methodology, investigation, formal analysis, supervision, writing-review and editing. All authors read and approved the final manuscript.

Acknowledgements

The authors thank the field and laboratory teams at ICRISAT (Patancheru, India), CREA (Anzola dell'Emilia, Italy) and Vasantrao Naik Marathwada Krishi Vidyapeeth (Parbhani, India) for managing the trials and for assistance with phenotyping.

Clinical trial number: Not applicable

References

1. Mekonnen MM, Hoekstra AY. Four billion people facing severe water scarcity. Sci Adv. 2016;2:e1500323. https://doi.org/10.1126/sciadv.1500323.
2. FAO. The state of the world's land and water resources for food and agriculture: systems at breaking point. Rome: Food and Agriculture Organization of the United Nations; 2021. https://doi.org/10.4060/cb7654en

3. Anas M, Liao F, Verma KK, Sarwar MA, Mahmood A, Chen ZL, et al. Fate of nitrogen in agriculture and environment: agronomic, eco-physiological and molecular approaches to improve nitrogen use efficiency. Biol Res. 2020;53:47. https://doi.org/10.1186/s40659-020-00312-4.
4. Tian H, Xu R, Canadell JG, Thompson RL, Winiwarter W, Suntharalingam P, et al. A comprehensive quantification of global nitrous oxide sources and sinks. Nature. 2020;586:248–56. https://doi.org/10.1038/s41586-020-2780-0.
5. Gu B, Zhang X, Lam SK, Yu Y, van Grinsven HJM, Zhang S, et al. Cost-effective mitigation of nitrogen pollution from global croplands. Nature. 2023;613:77–84. https://doi.org/10.1038/s41586-022-05481-8.
6. Rockström J, Gupta J, Qin D, Lade SJ, Abrams JF, Andersen LS, et al. Safe and just Earth system boundaries. Nature. 2023;619:102–11. https://doi.org/10.1038/s41586-023-06083-8.
7. Singh H, Kumar N, Kumar A. Enhancing resource use efficiency in crops through plant functional traits. Plant functional traits for improving productivity. Singapore: Springer Nature; 2024. pp. 97–117.
8. Yan M, Lian H, Zhang C, Chen Y, Cai H, Zhang S. The role of root size and root efficiency in grain production, and water- and nitrogen-use efficiency in wheat. J Sci Food Agric. 2023;103:7083–94.
9. Congreves KA, Otchere O, Ferland D, Farzadfar S, Williams S, Arcand MM. Nitrogen use efficiency definitions of today and tomorrow. Front Plant Sci. 2021;12:637108. https://doi.org/10.3389/fpls.2021.637108.
10. Moll RH, Kamprath EJ, Jackson WA. Analysis and interpretation of factors which contribute to efficiency of nitrogen utilization. Agron J. 1982;74:562–4. https://doi.org/10.2134/agronj1982.00021962007400030037x.
11. Cassman KG, Dobermann A, Walters DT. Agroecosystems, nitrogen-use efficiency, and nitrogen management. Ambio. 2002;31:132–40. https://doi.org/10.1579/0044-7447-31.2.132.
12. Habyarimana E, Bonardi P, Laureti D, Di Bari V, Cosentino S, Lorenzoni C. Multilocational evaluation of biomass sorghum hybrids under two stand densities and variable water supply in Italy. Ind Crops Prod. 2004;20:3–9.
13. Boyles RE, Brenton ZW, Kresovich S. Genetic and genomic resources of sorghum to connect genotype with phenotype in contrasting environments. Plant J. 2019;97:19–39. https://doi.org/10.1111/tpj.14113.
14. Thakur NR, Gorthy S, Vemula A, Odeny DA, Ruperao P, Sargar PR, et al. Genome-wide association study and expression of candidate genes for Fe and Zn concentration in sorghum grains. Sci Rep. 2024;14:12729. https://doi.org/10.1038/s41598-024-63308-0.
15. Mullet J, Morishige D, McCormick R, Truong S, Hilley J, McKinley B, et al. Energy sorghum—a genetic model for the design of C4 grass bioenergy crops. J Exp Bot. 2014;65:3479–89. https://doi.org/10.1093/jxb/eru229.
16. Sargar P, Jwala P, Thakur N, Mehtre S, Kalpande H, Pawar G, et al. Decoding sorghum root system architecture for resource use efficiency and climate resilience under multifactorial stress conditions. Physiol Plant. 2025;177:e70672. https://doi.org/10.1111/ppl.70672.
17. Paterson AH, Kong W, Johnston RM, Nabukalu P, Wu G, Poehlman WL, et al. The evolution of an invasive plant, Sorghum halepense L. ('Johnsongrass'). Front Genet. 2020;11:317. https://doi.org/10.3389/fgene.2020.00317.
18. Ananda GKS, Myrans H, Norton SL, Gleadow R, Furtado A, Henry RJ. Wild sorghum as a promising resource for crop improvement. Front Plant Sci. 2020;11:1108. https://doi.org/10.3389/fpls.2020.01108.
19. Thakur NR, Ingle KP, Sargar PR, Baraskar SS, Kasanaboina K, Awio B, et al. Sustainable utilization of wild germplasm resources. In: Al-Khayri JM, Jain SM, Penna S, editors. Sustainable utilization and conservation of plant genetic diversity. Sustainable development and biodiversity. Volume 35. Singapore: Springer; 2024.
20. Habyarimana E, Gorthy S, Baloch FS, Ercisli S, Chung G. Whole-genome resequencing of Sorghum bicolor and S. bicolor × S. halepense lines provides new insights for improving plant agroecological characteristics. Sci Rep. 2022;12:5556. https://doi.org/10.1038/s41598-022-09433-0.
21. Brar DS, Khush GS. Alien introgression in rice. Plant Mol Biol. 1997;35:35–47. https://doi.org/10.1023/A:1005825519998.
22. Mujeeb-Kazi A, Gul A, Farooq M, Rizwan S, Ahmad I. Rebirth of synthetic hexaploids with global implications for wheat improvement. Aust J Agric Res. 2008;59:391–8. https://doi.org/10.1071/AR07226.
23. Lennon JR, Krakowsky M, Goodman M, Flint-Garcia S, Balint-Kurti PJ. Identification of alleles conferring resistance to gray leaf spot in maize derived from its wild progenitor species teosinte. Crop Sci. 2017;57:1973–83. https://doi.org/10.2135/cropsci2016.12.0980.
24. Liu LH, Ludewig U, Gassert B, Frommer WB, von Wirén N. Urea transport by nitrogen-regulated tonoplast intrinsic proteins in Arabidopsis. Plant Physiol. 2003;133:1220–8. https://doi.org/10.1104/pp.103.027409.
25. Loqué D, Ludewig U, Yuan L. von Wirén N. Tonoplast intrinsic proteins AtTIP2;1 and AtTIP2;3 facilitate NH3 transport into the vacuole. Plant Physiol. 2005;137:671–80. https://doi.org/10.1104/pp.104.051268.
26. Pommerrenig B, Diehn TA, Bienert GP. Metalloido-porins: essentiality of Nodulin 26-like intrinsic proteins in metalloid transport. Plant Sci. 2015;238:212–27. https://doi.org/10.1016/j.plantsci.2015.06.002.
27. Paul MJ, Primavesi LF, Jhurreea D, Zhang Y. Trehalose metabolism and signaling. Annu Rev Plant Biol. 2008;59:417–41. https://doi.org/10.1146/annurev.arplant.59.032607.092945.
28. Figueroa CM, Lunn JE. A tale of two sugars: trehalose 6-phosphate and sucrose. Plant Physiol. 2016;172:7–27. https://doi.org/10.1104/pp.16.00417.
29. To JPC, Kieber JJ. Cytokinin signaling: two-components and more. Trends Plant Sci. 2008;13:85–92. https://doi.org/10.1016/j.tplants.2007.11.005.
30. Kant S, Bi YM, Rothstein SJ. Understanding plant response to nitrogen limitation for the improvement of crop nitrogen use efficiency. J Exp Bot. 2011;62:1499–509. https://doi.org/10.1093/jxb/erq297.
31. Mace ES, Tai S, Gilding EK, Li Y, Prentis PJ, Bian L, et al. Whole-genome sequencing reveals untapped genetic potential in Africa's indigenous cereal crop sorghum. Nat Commun. 2013;4:2320. https://doi.org/10.1038/ncomms3320.

32. Morris GP, Ramu P, Deshpande SP, Hash CT, Shah T, Upadhyaya HD, et al. Population genomic and genome-wide association studies of agroclimatic traits in sorghum. Proc Natl Acad Sci USA. 2013;110:453–8. https://doi.org/10.1073/pnas.1215985110.
33. Gorthy S, Sargar P, Hugar C, Li J, Habyarimana E. Summary of the potential and exploitation of omics and biotechnological breakthroughs in sorghum. In: Habyarimana E, Nadeem MA, Baloch FS, Zencirci N, editors. Omics and biotechnological approaches for product profile-driven sorghum improvement. Singapore: Springer; 2024. https://doi.org/10.1007/978-981-97-4347-6_17.
34. Sehgal D, Mondal S, Guzman C, Garcia Barrios G, Franco C, Singh R, et al. Validation of candidate gene-based markers and identification of novel loci for thousand-grain weight in spring bread wheat. Front Plant Sci. 2019;10:1189. https://doi.org/10.3389/fpls.2019.01189.
35. Voss-Fels KP, Cooper M, Hayes BJ. Accelerating crop genetic gains with genomic selection. Theor Appl Genet. 2019;132:669–86. https://doi.org/10.1007/s00122-018-3270-8.
36. Bhat JA, Yu D, Bohra A, Ganie SA, Varshney RK. Features and applications of haplotypes in crop breeding. Commun Biol. 2021;4:1266. https://doi.org/10.1038/s42003-021-02782-y.
37. Bandillo NB, Anderson JE, Kantar MB, Stupar RM, Specht JE, Graef GL, et al. Dissecting the genetic basis of local adaptation in soybean. Sci Rep. 2017;7:17195. https://doi.org/10.1038/s41598-017-17342-w.
38. McCormick RF, Truong SK, Sreedasyam A, Jenkins J, Shu S, Sims D, et al. The Sorghum bicolor reference genome: improved assembly, gene annotations, a transcriptome atlas, and signatures of genome organization. Plant J. 2018;93:338–54. https://doi.org/10.1111/tpj.13781.
39. Mascher M, Schreiber M, Scholz U, Graner A, Reif JC, Stein N. Genebank genomics bridges the gap between the conservation of crop diversity and plant breeding. Nat Genet. 2019;51:1076–81. https://doi.org/10.1038/s41588-019-0443-6.
40. Habyarimana E, Lopez-Cruz M, Baloch FS. Genomic selection for optimum index with dry biomass yield, dry mass fraction of fresh material, and plant height in biomass sorghum. Genes. 2020;11:61. https://doi.org/10.3390/genes11010061.
41. Awio B, Sargar P, Aghuru A, Chavan S, Kasanaboina K, Reddy VK, et al. Breeding advancements of sorghum in Europe. In: Habyarimana E, Nadeem MA, Baloch FS, Zencirci N, editors. Omics and biotechnological approaches for product profile-driven sorghum improvement. Singapore: Springer Nature; 2024. pp. 287–311. https://doi.org/10.1007/978-981-97-4347-6_12.
42. Federer WT. Augmented designs. Hawaii Planters' Rec. 1956;55:191–208.
43. Brouwer C, Prins K, Heibloem M. Irrigation water management: irrigation scheduling. Training manual no. 4. Rome: Food and Agriculture Organization of the United Nations; 1989.
44. R Core Team. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. 2023. https://www.R-project.org/. Accessed 5 Oct 2026.
45. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. J Stat Softw. 2015;67:1–48. https://doi.org/10.18637/jss.v067.i01.
46. Piepho HP, Möhring J. Computing heritability and selection response from unbalanced plant breeding trials. Genetics. 2007;177:1881–8. https://doi.org/10.1534/genetics.107.074229.
47. Robinson HF, Comstock RE, Harvey PH. Estimates of heritability and the degree of dominance in corn. Agron J. 1949;41:353–9. https://doi.org/10.2134/agronj1949.00021962004100080005x.
48. Burton GW. Quantitative inheritance in grasses. In: Proceedings of the 6th International Grassland Congress, vol. 1. State College, PA: Pennsylvania State College; 1952. pp. 277–283.
49. Malosetti M, Ribaut JM, van Eeuwijk FA. Front Physiol. 2013;4:44. https://doi.org/10.3389/fphys.2013.00044. The statistical analysis of multi-environment data: modeling genotype-by-environment interaction and its genetic basis.
50. Josse J, Husson F. missMDA: a package for handling missing values in multivariate data analysis. J Stat Softw. 2016;70:1–31. https://doi.org/10.18637/jss.v070.i01.
51. Hunter JD. Matplotlib: a 2D graphics environment. Comput Sci Eng. 2007;9:90–5. https://doi.org/10.1109/MCSE.2007.55.
52. Chen Y, Chen Y, Shi C, Huang Z, Zhang Y, Li S, et al. SOAPnuke: a MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. GigaScience. 2018;7:gix120. https://doi.org/10.1093/gigascience/gix120.
53. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009;25:1754–60. https://doi.org/10.1093/bioinformatics/btp324.
54. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27:2156–8. https://doi.org/10.1093/bioinformatics/btr330.
55. Browning BL, Zhou Y, Browning SR. A one-penny imputed genome from next-generation reference panels. Am J Hum Genet. 2018;103:338–48. https://doi.org/10.1016/j.ajhg.2018.07.015.
56. Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics. 2007;23:2633–5. https://doi.org/10.1093/bioinformatics/btm308.
57. Wright S. The genetical structure of populations. Ann Eugen. 1951;15:323–54. https://doi.org/10.1111/j.1469-1809.1949.tb02451.x.
58. Weir BS, Cockerham CC. Estimating F-statistics for the analysis of population structure. Evolution. 1984;38:1358–70. https://doi.org/10.1111/j.1558-5646.1984.tb05657.x.
59. Bhatia G, Patterson N, Sankararaman S, Price AL. Estimating and interpreting FST: the impact of rare variants. Genome Res. 2013;23:1514–21. https://doi.org/10.1101/gr.154831.113.

60. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000;155:945–59. https://doi.org/10.1093/genetics/155.2.945.
61. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009;19:1655–64. https://doi.org/10.1101/gr.094052.109.
62. VanRaden PM. Efficient methods to compute genomic predictions. J Dairy Sci. 2008;91:4414–23. https://doi.org/10.3168/jds.2007-0980.
63. Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet. 2006;38:203–8. https://doi.org/10.1038/ng1702.
64. Sul JH, Martin LS, Eskin E. Population structure in genetic studies: confounding factors and mixed models. PLoS Genet. 2018;14:e1007309. https://doi.org/10.1371/journal.pgen.1007309.
65. Zhang C, Dong SS, Xu JY, He WM, Yang TL. PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. Bioinformatics. 2019;35:1786–8. https://doi.org/10.1093/bioinformatics/bty875.
66. Wang J, Zhang Z. GAPIT version 3: boosting power and accuracy for genomic association and prediction. Genomics Proteom Bioinf. 2021;19:629–40. https://doi.org/10.1016/j.gpb.2021.08.005.
67. Huang M, Liu X, Zhou Y, Summers RM, Zhang Z. BLINK: a package for the next level of genome-wide association studies with both individuals and markers in the millions. GigaScience. 2019;8:giy154. https://doi.org/10.1093/gigascience/giy154.
68. Liu X, Huang M, Fan B, Buckler ES, Zhang Z. Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. PLoS Genet. 2016;12:e1005767. https://doi.org/10.1371/journal.pgen.1005767.
69. Segura V, Vilhjálmsson BJ, Platt A, Korte A, Seren Ü, Long Q, et al. An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. Nat Genet. 2012;44:825–30. https://doi.org/10.1038/ng.2314.
70. Yin L, Zhang H, Tang Z, Xu J, Yin D, Zhang Z, et al. rMVP: a memory-efficient, visualization-enhanced, and parallel-accelerated tool for genome-wide association study. Genomics Proteom Bioinf. 2021;19:619–28.
71. Turner SD. qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. J Open Source Softw. 2018;3:731. https://doi.org/10.21105/joss.00731.
72. Devlin B, Roeder K. Genomic control for association studies. Biometrics. 1999;55:997–1004. https://doi.org/10.1111/j.0006-341X.1999.00997.x.
73. Zeng P, Zhao Y, Qian C, Zhang L, Zhang R, Gou J, et al. Statistical analysis for genome-wide association study. J Biomed Res. 2015;29:285–97. https://doi.org/10.7555/JBR.29.20140007.
74. Beavis WD. The power and deceit of QTL experiments: lessons from comparative QTL studies. In: Proceedings of the 49th Annual Corn and Sorghum Industry Research Conference. Washington, DC: American Seed Trade Association; 1994. pp. 250–266.
75. Xu S. Theoretical basis of the Beavis effect. Genetics. 2003;165:2259–68. https://doi.org/10.1093/genetics/165.4.2259.
76. King EG, Long AD. The Beavis effect in next-generation mapping panels in Drosophila melanogaster. G3 (Bethesda). 2017;7:1643–52. https://doi.org/10.1534/g3.117.041426
77. Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, et al. Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res. 2012;40:D1178–86. https://doi.org/10.1093/nar/gkr944.
78. The UniProt Consortium. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 2017;45:D158–69. https://doi.org/10.1093/nar/gkw1099.
79. Mi H, Muruganujan A, Ebert D, Huang X, Thomas PD. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. Nucleic Acids Res. 2019;47:D419–26. https://doi.org/10.1093/nar/gky1038.
80. Gladman N, Olson A, Wei S, Chougule K, Lu Z, Tello-Ruiz M, et al. SorghumBase: a web-based portal for sorghum genetic information and community advancement. Planta. 2022;255:35. https://doi.org/10.1007/s00425-022-03821-6.
81. Sreedasyam A, Plott C, Hossain MS, Lovell JT, Grimwood J, Jenkins JW, et al. JGI Plant Gene Atlas: an updateable transcriptome resource to improve functional gene descriptions across the plant kingdom. Nucleic Acids Res. 2023;51:8383–401. https://doi.org/10.1093/nar/gkad616.
82. Chen C, Chen H, Zhang Y, Thomas HR, Frank MH, He Y, et al. TBtools: an integrative toolkit developed for interactive analyses of big biological data. Mol Plant. 2020;13:1194–202. https://doi.org/10.1016/j.molp.2020.06.009.
83. Lenth RV. emmeans: estimated marginal means, aka least-squares means. R package version 1.8.5. 2023. https://CRAN.R-project.org/package=emmeans. Accessed 5 Oct 2026.
84. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B Stat Methodol. 1995;57:289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.
85. Paradis E. pegas: an R package for population genetics with an integrated–modular approach. Bioinformatics. 2010;26:419–20. https://doi.org/10.1093/bioinformatics/btp696.
86. Shin JH, Blay S, McNeney B, Graham J. LDheatmap: an R function for graphical display of pairwise linkage disequilibria between single nucleotide polymorphisms. J Stat Softw. 2006;16. https://doi.org/10.18637/jss.v016.c03. :Code Snippet 3.
87. Spindel JE, Dahlberg J, Colgan M, Hollingsworth J, Sievert J, Staggenborg SH, et al. Association mapping by aerial drone reveals 213 genetic associations for Sorghum bicolor biomass traits under drought. BMC Genomics. 2018;19:679. https://doi.org/10.1186/s12864-018-5055-5.
88. Sukumaran S, Reynolds MP, Sansaloni C. Genome-wide association analyses identify QTL hotspots for yield and component traits in durum wheat grown under yield potential, drought, and heat stress environments. Front Plant Sci. 2018;9:81. https://doi.org/10.3389/fpls.2018.00081.

89. Bouchet S, Olatoye MO, Marla SR, Perumal R, Tesso T, Yu J, et al. Increased power to dissect adaptive traits in global sorghum diversity using a nested association mapping population. Genetics. 2017;206:573–85. https://doi.org/10.1534/genetics.116.198499.
90. Fernandes SB, Dias KOG, Ferreira DF, Brown PJ. Efficiency of multi-trait, indirect, and trait-assisted genomic selection for improvement of biomass sorghum. Theor Appl Genet. 2018;131:747–55. https://doi.org/10.1007/s00122-017-3033-y.
91. Sadras VO, Angus JF. Benchmarking water-use efficiency of rainfed wheat in dry environments. Aust J Agric Res. 2006;57:847–56. https://doi.org/10.1071/AR05359.
92. Sanchez PL, Wing RA, Brar DS. The wild relative of rice: genomes and genomics. In: Zhang Q, Wing RA, editors. Genetics and genomics of rice. New York: Springer; 2014. pp. 9–25. https://doi.org/10.1007/978-1-4614-7903-1_2.
93. Kasule F, Alladassi BME, Aru CJ, Adikini S, Biruma M, Ugen MA, et al. Genetic diversity, population structure, and a genome-wide association study of sorghum lines assembled for breeding in Uganda. Front Plant Sci. 2024;15:1458179. https://doi.org/10.3389/fpls.2024.1458179.
94. Yan J, Shah T, Warburton ML, Buckler ES, McMullen MD, Crouch J. Genetic characterization and linkage disequilibrium estimation of a global maize collection using SNP markers. PLoS ONE. 2009;4:e8451. https://doi.org/10.1371/journal.pone.0008451.
95. Romay MC, Millard MJ, Glaubitz JC, Peiffer JA, Swarts KL, Casstevens TM, et al. Comprehensive genotyping of the USA national maize inbred seed bank. Genome Biol. 2013;14:R55. https://doi.org/10.1186/gb-2013-14-6-r55.
96. Maccaferri M, Canè MA, Sanguineti MC, Salvi S, Colalongo MC, Massi A, et al. A consensus framework map of durum wheat (Triticum durum Desf.) suitable for linkage disequilibrium analysis and genome-wide association mapping. BMC Genomics. 2014;15:873. https://doi.org/10.1186/1471-2164-15-873.
97. Bayer MM, Rapazote-Flores P, Ganal M, Hedley PE, Macaulay M, Plieske J, et al. Development and evaluation of a barley 50k iSelect SNP array. Front Plant Sci. 2017;8:1792. https://doi.org/10.3389/fpls.2017.01792.
98. McConnell JR, Emery J, Eshed Y, Bao N, Bowman J, Barton MK. Role of PHABULOSA and PHAVOLUTA in determining radial patterning in shoots. Nature. 2001;411:709–13. https://doi.org/10.1038/35079635.
99. Ariel FD, Manavella PA, Dezar CA, Chan RL. The true story of the HD-Zip family. Trends Plant Sci. 2007;12:419–26. https://doi.org/10.1016/j.tplants.2007.08.003.
100. Prigge MJ, Otsuga D, Alonso JM, Ecker JR, Drews GN, Clark SE. Class III homeodomain-leucine zipper gene family members have overlapping, antagonistic, and distinct roles in Arabidopsis development. Plant Cell. 2005;17:61–76. https://doi.org/10.1105/tpc.104.026161.
101. Carlsbecker A, Lee JY, Roberts CJ, Dettmer J, Lehesranta S, Zhou J, et al. Cell signalling by microRNA165/6 directs gene dose dependent root cell fate. Nature. 2010;465:316–21. https://doi.org/10.1038/nature08977.
102. Nuccio ML, Wu J, Mowers R, Zhou HP, Meghji M, Primavesi LF, et al. Expression of trehalose-6-phosphate phosphatase in maize ears improves yield in well-watered and drought conditions. Nat Biotechnol. 2015;33:862–9. https://doi.org/10.1038/nbt.3277.
103. Griffiths CA, Sagar R, Geng Y, Primavesi LF, Patel MK, Passarelli MK, et al. Chemical intervention in plant sugar signalling increases yield and resilience. Nature. 2016;540:574–8. https://doi.org/10.1038/nature20591.
104. Borrell AK, Mullet JE, George-Jaeggli B, van Oosterom EJ, Hammer GL, Klein PE, et al. Drought adaptation of stay-green sorghum is associated with canopy development, leaf anatomy, root growth, and water uptake. J Exp Bot. 2014;65:6251–63. https://doi.org/10.1093/jxb/eru232.
105. Chaumont F, Tyerman SD. Aquaporins: highly regulated channels controlling plant water relations. Plant Physiol. 2014;164:1600–18. https://doi.org/10.1104/pp.113.233791.
106. Maurel C, Boursiac Y, Luu DT, Santoni V, Shahzad Z, Verdoucq L. Aquaporins in plants. Physiol Rev. 2015;95:1321–58. https://doi.org/10.1152/physrev.00008.2015.
107. Immanen J, Nieminen K, Duchens Silva H, Rojas Rodríguez F, Meisel LA, Silva H, et al. Characterization of cytokinin signaling and homeostasis gene families in two hardwood tree species: Populus trichocarpa and Prunus persica. BMC Genomics. 2013;14:885. https://doi.org/10.1186/1471-2164-14-885.
108. Scheibe R. Malate valves to balance cellular energy supply. Physiol Plant. 2004;120:21–6. https://doi.org/10.1111/j.0031-9317.2004.0222.x.
109. Edwards GE, Voznesenskaya EV. C4 photosynthesis: Kranz forms and single-cell C4 in terrestrial plants. In: Raghavendra AS, Sage RF, editors. C4 photosynthesis and related CO2 concentrating mechanisms. Dordrecht: Springer; 2011. pp. 29–61. https://doi.org/10.1007/978-90-481-9407-0_4.
110. Passardi F, Cosio C, Penel C, Dunand C. Peroxidases have more functions than a Swiss army knife. Plant Cell Rep. 2005;24:255–65. https://doi.org/10.1007/s00299-005-0972-6.
111. Mittler R, Vanderauwera S, Suzuki N, Miller G, Tognetti VB, Vandepoele K, et al. ROS signaling: the new wave? Trends Plant Sci. 2011;16:300–9. https://doi.org/10.1016/j.tplants.2011.03.007.
112. Ishiguro S, Kawai-Oda A, Ueda J, Nishida I, Okada K. The DEFECTIVE IN ANTHER DEHISCENCE1 gene encodes a novel phospholipase A1 catalyzing the initial step of jasmonic acid biosynthesis, which synchronizes pollen maturation, anther dehiscence, and flower opening in Arabidopsis. Plant Cell. 2001;13:2191–209. https://doi.org/10.1105/tpc.010192.
113. Aoki K, Ogata Y, Shibata D. Approaches for extracting practical information from gene co-expression networks in plant biology. Plant Cell Physiol. 2007;48:381–90. https://doi.org/10.1093/pcp/pcm013.
114. Abbai R, Singh VK, Nachimuthu VV, Sinha P, Selvaraj R, Vipparla AK, et al. Haplotype analysis of key genes governing grain yield and quality traits across 3K RG panel reveals scope for the development of tailor-made rice with enhanced genetic gains. Plant Biotechnol J. 2019;17:1612–22. https://doi.org/10.1111/pbi.13087.

115. Abed A, Belzile F. Comparing single-SNP, multi-SNP, and haplotype-based approaches in association studies for major traits in barley. Plant Genome. 2019;12:190036. https://doi.org/10.3835/plantgenome2019.05.0036.
116. Bernardo R, Charcosset A. Usefulness of gene information in marker-assisted recurrent selection: a simulation appraisal. Crop Sci. 2006;46:614–21. https://doi.org/10.2135/cropsci2005.05-0088.
117. Steele KA, Quinton-Tulloch MJ, Amgai RB, Dhakal R, Khatiwada SP, Vyas D, et al. Accelerating public sector rice breeding with high-density KASP markers derived from whole genome sequencing of indica rice. Mol Breed. 2018;38:38. https://doi.org/10.1007/s11032-018-0777-2.
118. Habyarimana E, Lopez-Cruz M. Genomic selection for antioxidant production in a panel of Sorghum bicolor and S. bicolor × S. halepense lines. Genes. 2019;10:841. https://doi.org/10.3390/genes10110841.
119. Crossa J, Pérez-Rodríguez P, Cuevas J, Montesinos-López O, Jarquín D, de los Campos G, et al. Genomic selection in plant breeding: methods, models, and perspectives. Trends Plant Sci. 2017;22:961–75. https://doi.org/10.1016/j.tplants.2017.08.011.
120. Sinclair TR, Rufty TW, Lewis RS. Increasing photosynthesis: unlikely solution for world food problem. Trends Plant Sci. 2019;24:1032–9. https://doi.org/10.1016/j.tplants.2019.07.008.
121. Khanna A, Sharma V, Ellur RK, Shikari AB, Gopala Krishnan S, Singh UD, et al. Development and evaluation of near-isogenic lines for major blast resistance gene(s) in Basmati rice. Theor Appl Genet. 2015;128:1243–59. https://doi.org/10.1007/s00122-015-2502-4.
122. Char SN, Wei J, Mu Q, Li X, Zhang ZJ, Yu J, et al. An Agrobacterium-delivered CRISPR/Cas9 system for targeted mutagenesis in sorghum. Plant Biotechnol J. 2020;18:319–21. https://doi.org/10.1111/pbi.13229.
123. Massel K, Lam Y, Wong ACS, Hickey LT, Borrell AK, Godwin ID. Hotter, drier, CRISPR: the latest edit on climate change. Theor Appl Genet. 2021;134:1691–709. https://doi.org/10.1007/s00122-020-03764-0.
124. Sargar P, Baraskar S, Altaf MT, Liaqat W, Jamil A, Maqbool A. Molecular mechanisms and current progress of CRISPR/Cas in plant development. In: Basit A, Mohamed HI, Tanveer Shah S, Dawood MFA, editors. CRISPR-Cas9 as a revolutionary gene editing tool: genome engineering, advances and applications in agriculture. Boca Raton, FL: CRC; 2026. pp. 19–33. https://doi.org/10.1201/9781003518204-2.
125. Araus JL, Cairns JE. Field high-throughput phenotyping: the new crop breeding frontier. Trends Plant Sci. 2014;19:52–61. https://doi.org/10.1016/j.tplants.2013.09.008.

Figures

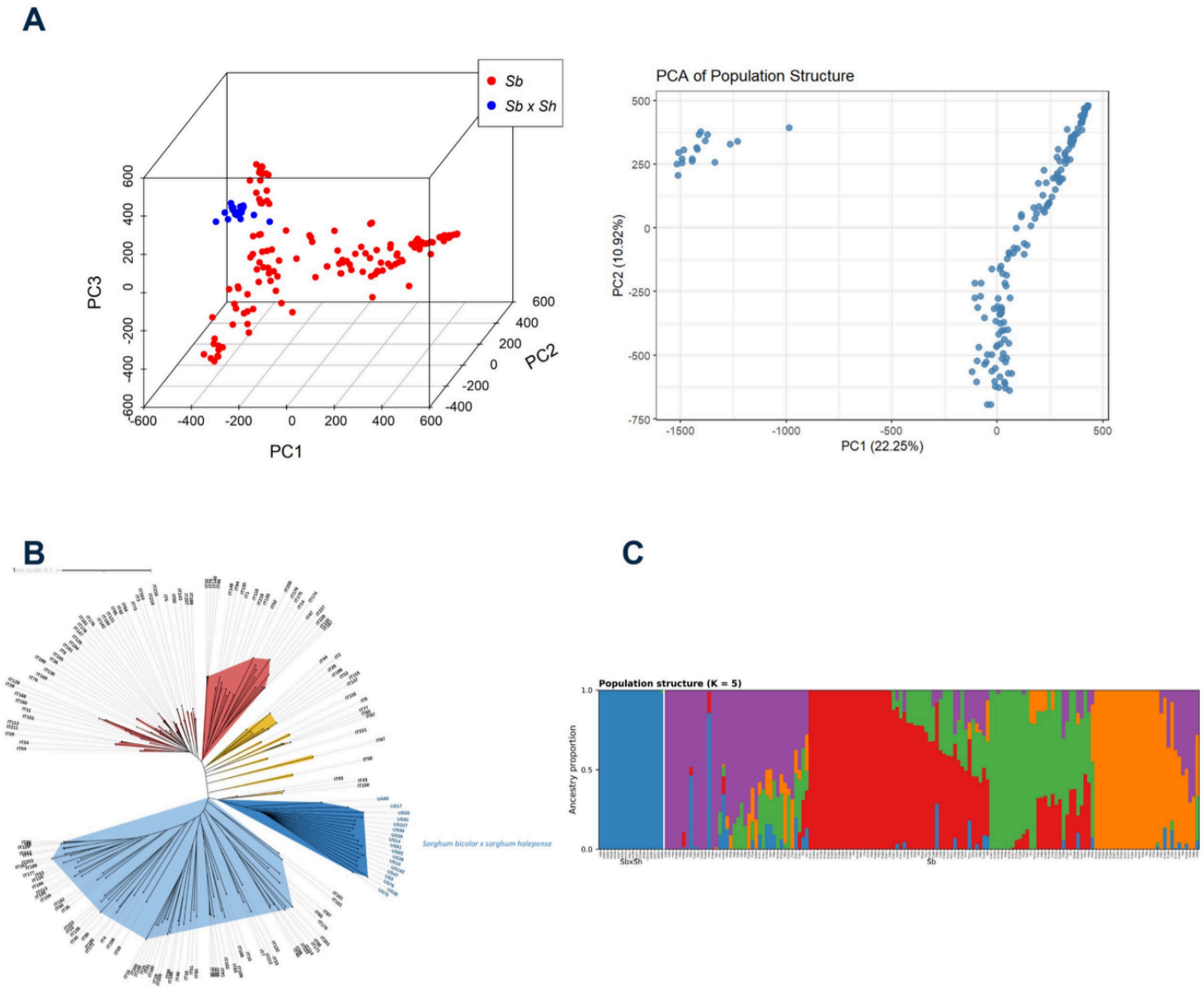


Figure 1

Population structure of the sorghum diversity panel. (A) Three and two dimensional principal component analysis separating *S. bicolor*(Sb) from *S. bicolor* × *S. halepense* (Sb×Sh) lines on genome-wide SNPs. (B) Neighbor-joining tree of the panel, with the Sb×Sh RILs highlighted in dark blue. (C) Ancestry coefficients at K = 5, the cross-entropy optimum; all 18 genotyped introgression lines fall in a single cluster.

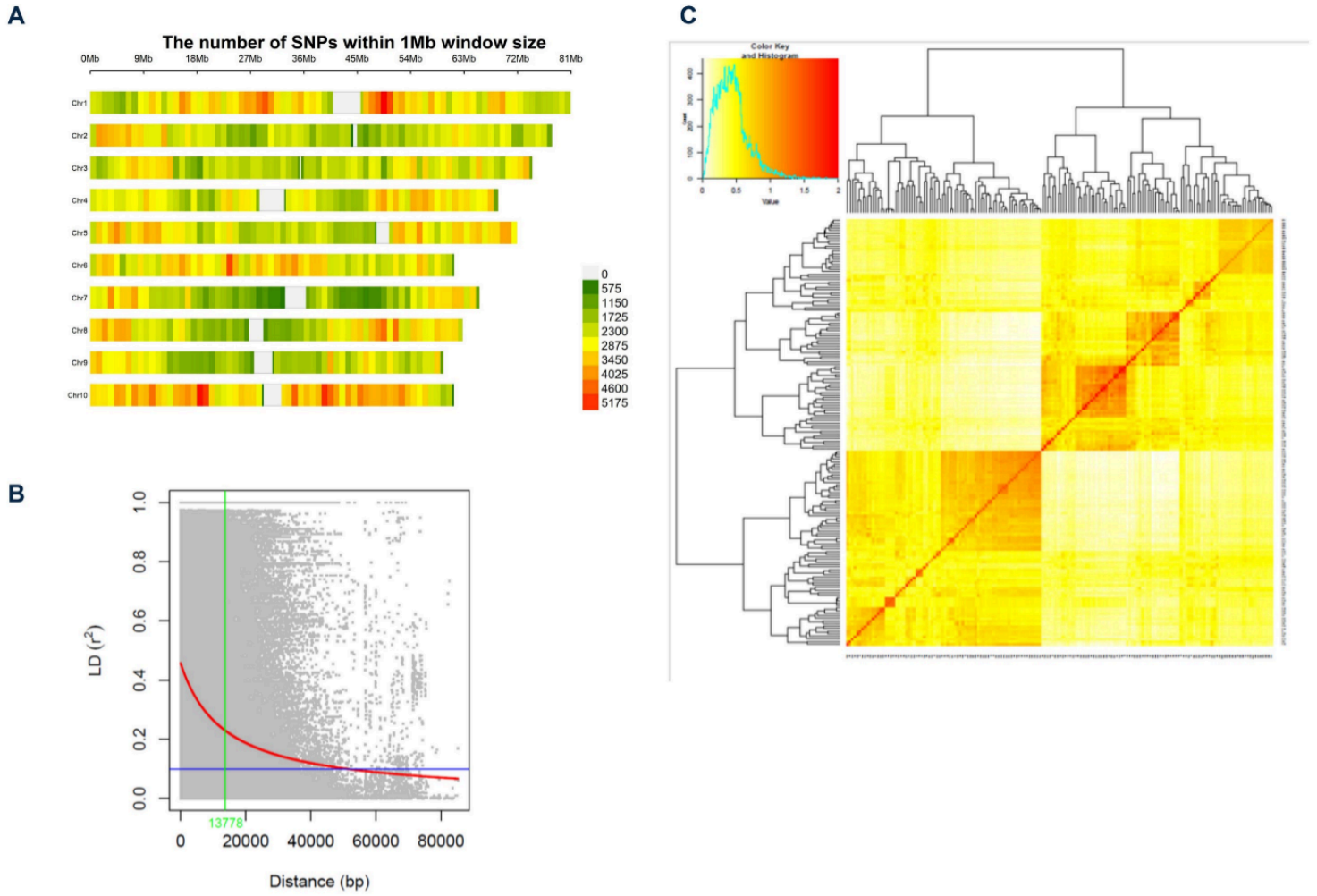


Figure 2

Genomic characteristics and relatedness of the sorghum panel. (A) Genome wide SNP density in 1-Mb windows (MAF ≥ 0.05; n = 1,752,624 SNPs). (B) Decay of linkage disequilibrium (r^2) with physical distance; the half-decay distance is 13.78 kb. (C) Pairwise kinship matrix with hierarchical clustering.

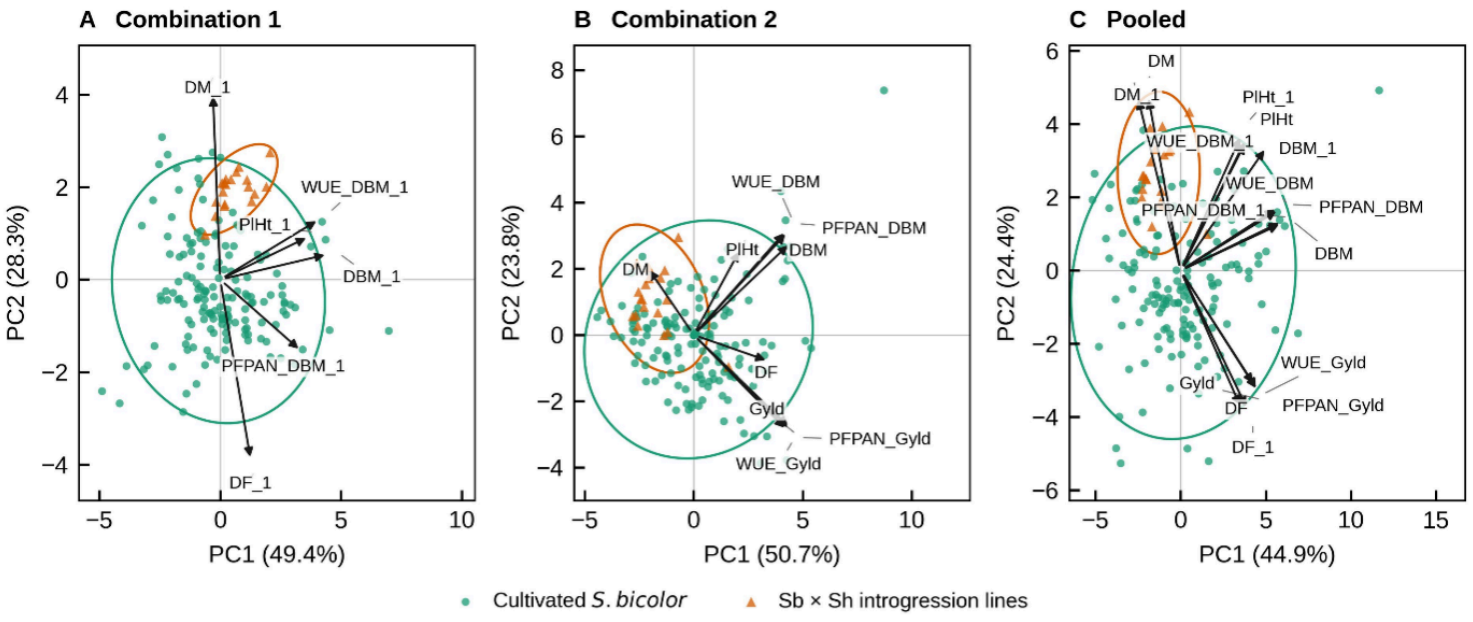


Figure 3

Principal component analysis (PCA) biplots of agronomic and resource-use efficiency traits. (A) Combination 1 (IT21 + IND22 + IND23; six traits). (B) Combination 2 (IND22 + IND23; nine traits). (C) Pooled analysis of the 15 trait × combination variables. Points are genotypes (n = 172) colored by sub-population (green circles, cultivated *S. bicolor*; orange triangles, Sb×Sh introgression lines), with a 95% normal-probability ellipse for each sub-population; arrows are trait loading vectors on the first two components. Axis labels give the percentage of total variance explained by each component. Missing BLUEs were imputed by regularized iterative PCA before analysis.

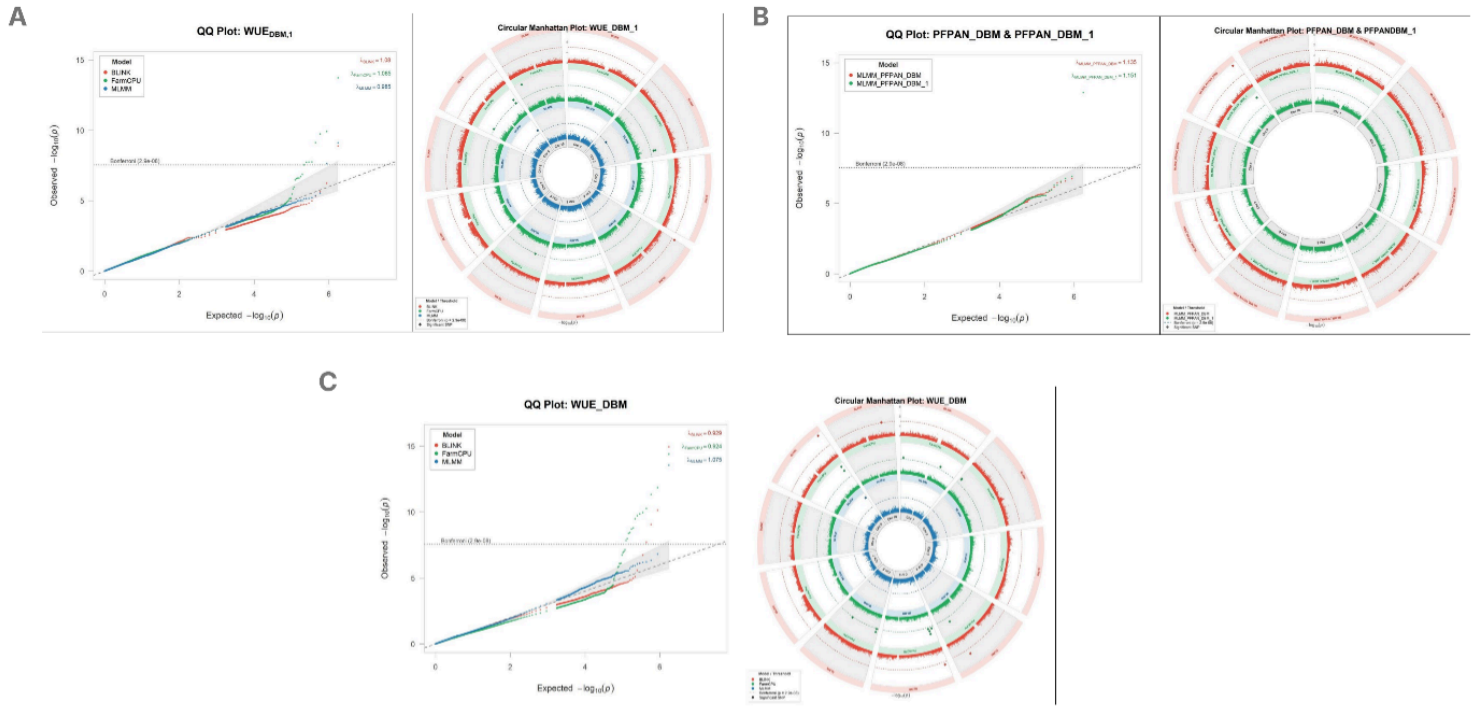


Figure 4

Quantile–quantile and circular Manhattan plots from the genome-wide association analysis of the four resource-use efficiency traits with significant associations (WUE_DBM_1, PFPAN_DBM_1, WUE_DBM and PFPAN_DBM). (A) WUE_DBM_1 (BLINK, FarmCPU and MLMM); (B) PFPAN_DBM and PFPAN_DBM_1 (MLMM only, the single well-calibrated model for these traits; Table 3); (C) WUE_DBM (BLINK, FarmCPU and MLMM). Genomic inflation factors (λ) are shown in each Q–Q plot, Bonferroni thresholds ($p = 2.85 \times 10^{-8}$) are drawn as dotted (Q–Q) or dashed (Manhattan) lines, and diamonds mark significant SNPs.

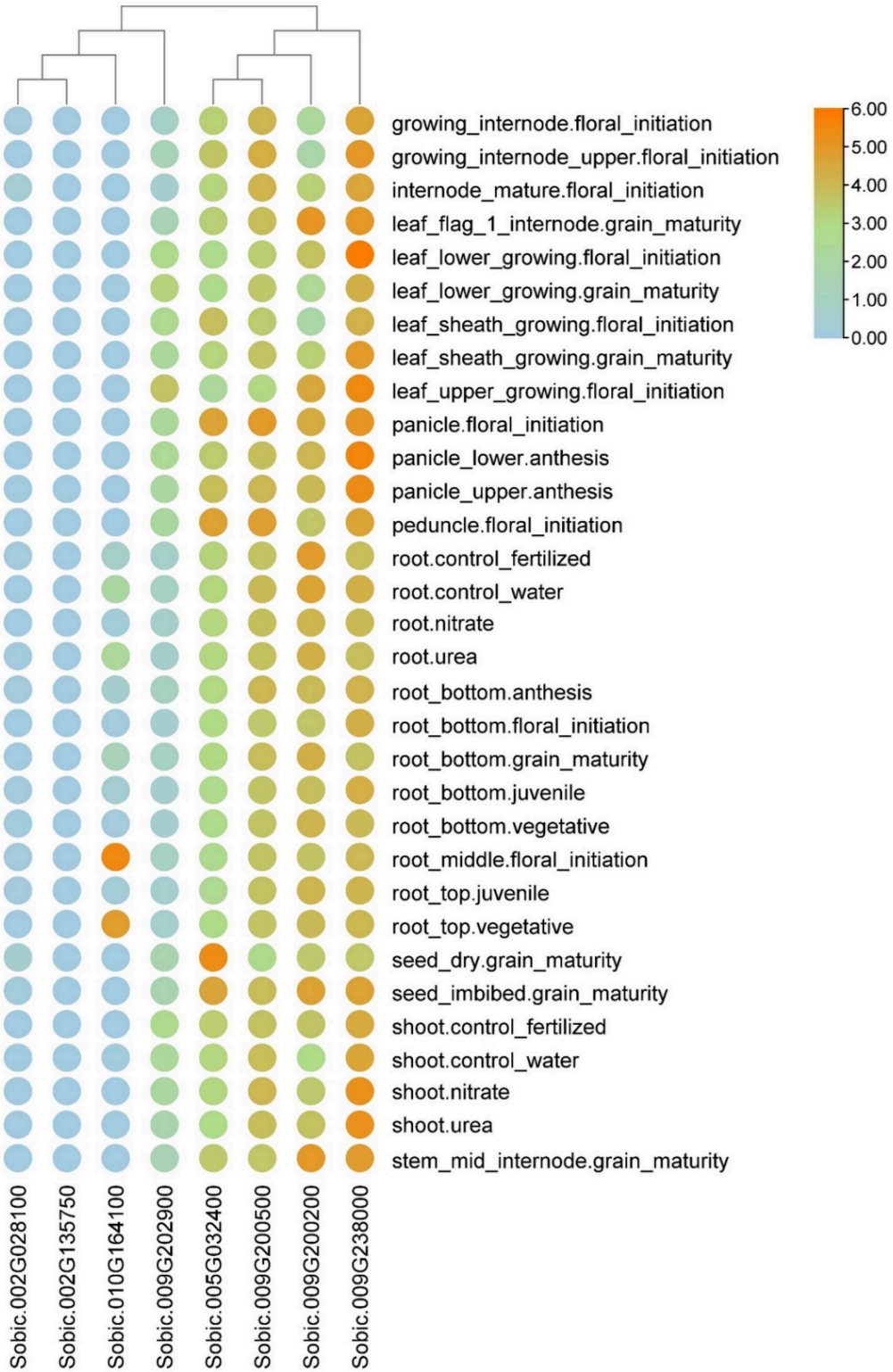


Figure 5

Hierarchically clustered expression heatmap of the eight prioritized candidate genes across 33 sorghum tissues and developmental stages. Color scale: $\log_2$ FPKM.

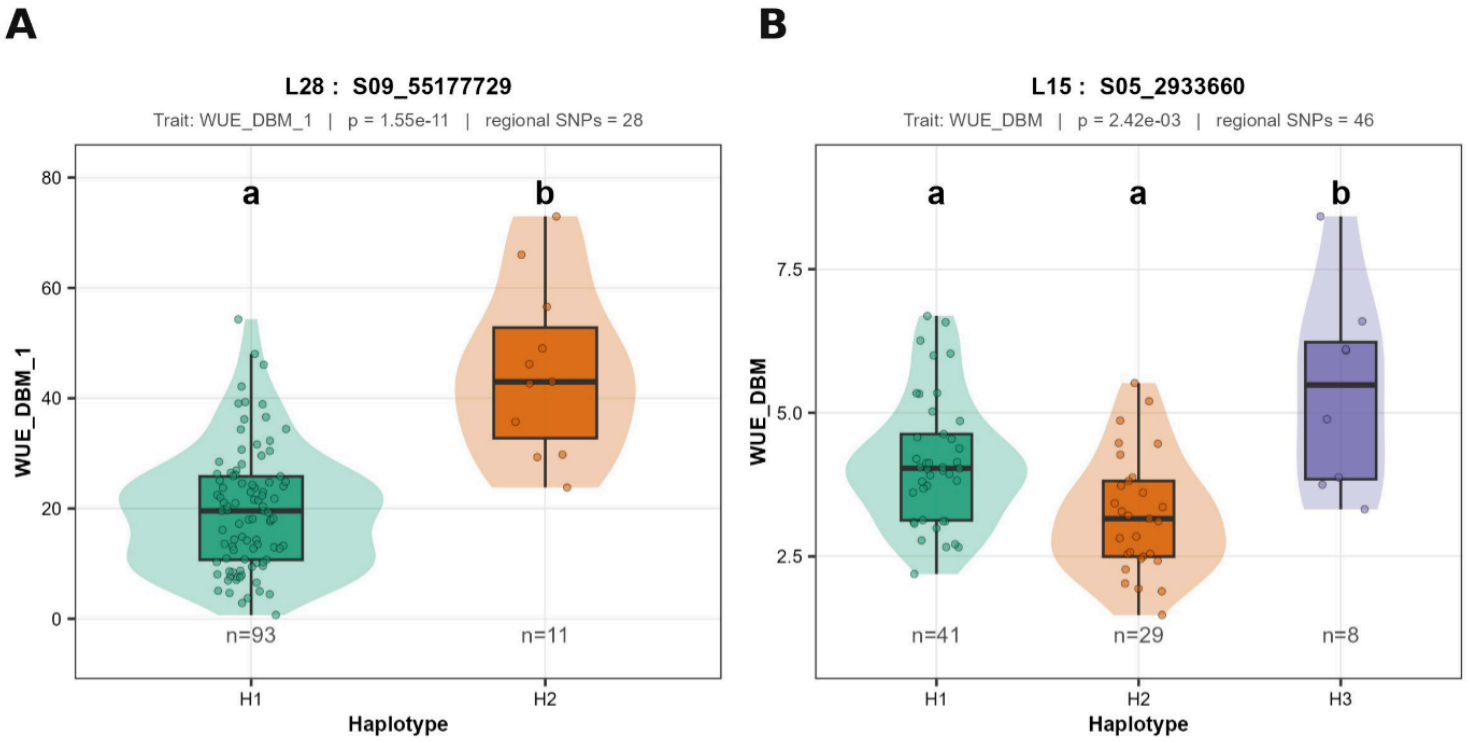


Figure 6

Haplotype–trait associations at the two loci that remained significant after both Benjamini–Hochberg and Bonferroni correction across the 17 locus × trait tests. (A) S09_55177729 (chromosome 9, 55.18 Mb) for WUE_DBM_1, showing a strong contrast between the favorable haplotype H2 and the unfavorable H1. (B) S05_2933660 (chromosome 5, 2.93 Mb) for WUE_DBM, showing the favorable haplotype H3 separated from H1 and H2. Each panel reports the Tukey HSD compact letter display, the haplotype-level p-value, the number of regional QC-passing SNPs in the ± 13,778 bp window and the number of genotypes analyzed per haplotype (n). Boxes show the median and interquartile range; violins show the full density distribution. L28 and L15 are locus identifiers used in Additional file 1: Table S9; WUE units, kg DBM ha^{-1} mm^{-1}.

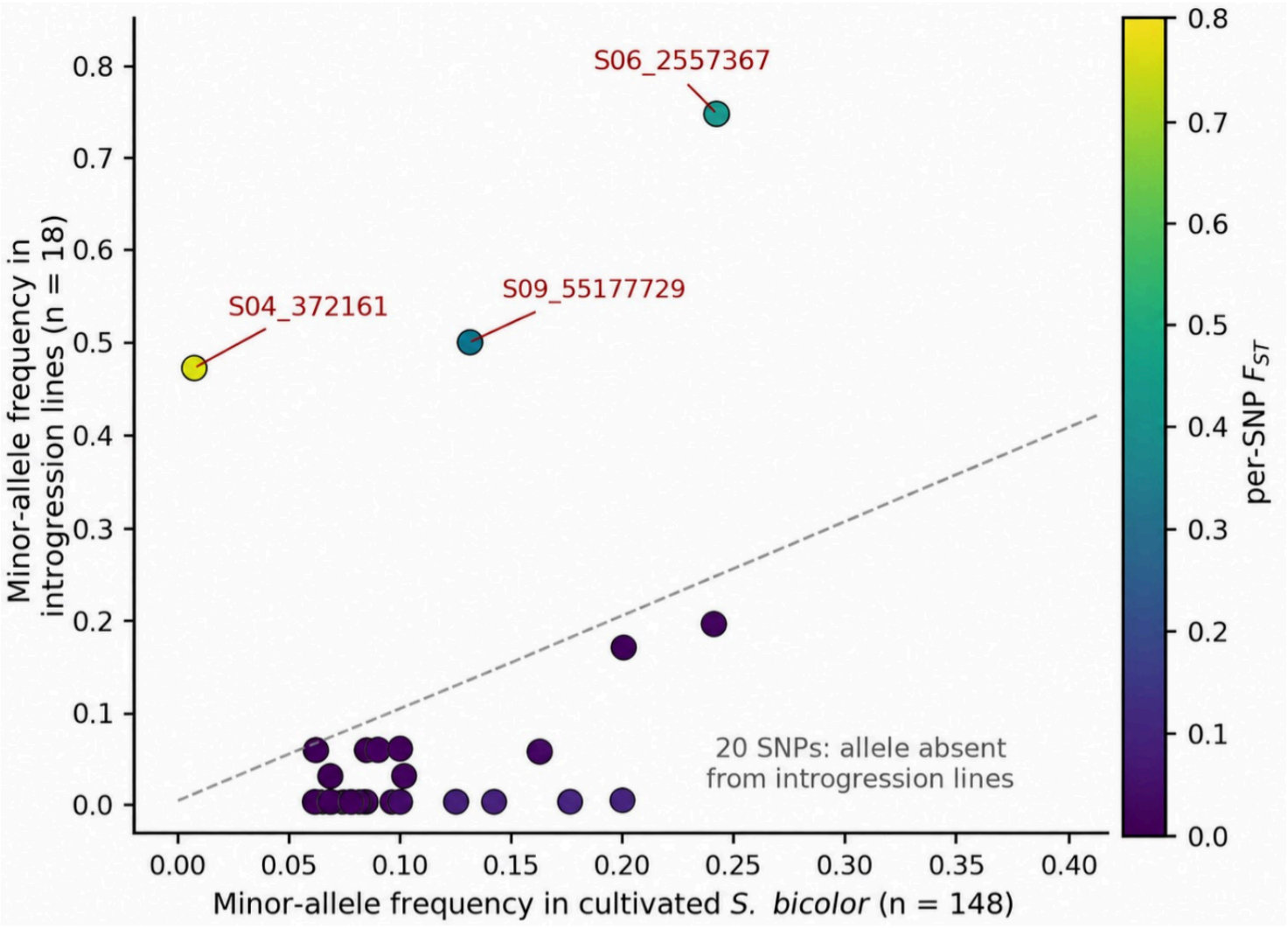


Figure 7

Sub-population distribution of the alleles associated with WUE and PFPAN. Minor-allele frequency in the 148 cultivated entries is plotted against that in the 18 introgression lines, colored by per-SNP FST; the dashed line indicates parity. SNPs on the horizontal axis carry an allele absent from the introgressed sub-population, and SNPs above the diagonal carry an allele enriched there.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1TablesS1S9.xlsx
- Additionalfile2FiguresS1S8.pdf