Theoretical and Applied Genetics (2021) 134:295–312
https://doi.org/10.1007/s00122-020-03697-8

ORIGINAL ARTICLE

Association mapping and genomic selection for sorghum adaptation to tropical soils of Brazil in a sorghum multiparental random mating population

Karine C. Bernardino[1,2] · Cícero B. de Menezes[2] · Sylvia M. de Sousa[2] · Claudia T. Guimarães[2] · Pedro C. S. Carneiro[1] · Robert E. Schaffert[2] · Leon V. Kochian[3] · Barbara Hufnagel[2,4] · Maria Marta Pastina[2] · Jurandir V. Magalhaes[2]

Received: 17 September 2019 / Accepted: 28 September 2020 / Published online: 14 October 2020

Abstract

Key message **A multiparental random mating population used in sorghum breeding is amenable for the detection of QTLs related to tropical soil adaptation, fine mapping of underlying genes and genomic selection approaches.**

Abstract Tropical soils where low phosphorus (P) and aluminum (Al) toxicity limit sorghum [*Sorghum bicolor* (L.) Moench] production are widespread in the developing world. We report on BRP13R, a multiparental random mating population (MP-RMP), which is commonly used in sorghum recurrent selection targeting tropical soil adaptation. Recombination dissipated much of BRP13R's likely original population structure and average linkage disequilibrium (LD) persisted up to 2.5 Mb, establishing BRP13R as a middle ground between biparental populations and sorghum association panels. Genome-wide association mapping (GWAS) identified conserved QTL from previous studies, such as for root morphology and grain yield under low-P, and indicated the importance of dominance in the genetic architecture of grain yield. By overlapping consensus QTL regions, we mapped two candidate P efficiency genes to a ~ 5 Mb region on chromosomes 6 (*ALMT*) and 9 (*PHO2*). Remarkably, we find that only 200 progeny genotyped with ~ 45,000 markers in BRP13R can lead to GWAS-based positional cloning of naturally rare, subpopulation-specific alleles, such as for *SbMATE*-conditioned Al tolerance. Genomic selection was found to be useful in such MP-RMP, particularly if markers in LD with major genes are fitted as fixed effects into GBLUP models accommodating dominance. Shifts in allele frequencies in progeny contrasting for grain yield indicated that intermediate to minor-effect genes on P efficiency, such as *SbPSTOL1* genes, can be employed in pre-breeding via allele mining in the base population. Therefore, MP-RMPs such as BRP13R emerge as multipurpose resources for efficient gene discovery and deployment for breeding sorghum cultivars adapted to tropical soils.

Communicated by Hai-Chun Jing.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00122-020-03697-8) contains supplementary material, which is available to authorized users.

✉ Maria Marta Pastina
marta.pastina@embrapa.br

✉ Jurandir V. Magalhaes
jurandir.magalhaes@embrapa.br

[1] Universidade Federal de Viçosa, Avenida Peter Henry Rolfs, s/n, Viçosa, MG 36570-900, Brazil

[2] Embrapa Milho e Sorgo, Rodovia MG 424 km 65, Sete Lagoas, MG 35701-970, Brazil

[3] Global Institute for Food Security, University of Saskatchewan, Saskatoon, SK S7N 4J8, Canada

[4] BPMP, CNRS, INRAE, SupAgro, University of Montpellier, 34060 Montpellier, France

Introduction

Crop adaptation to tropical soils relies on tolerance to multiple abiotic stresses rather than to a single stress condition. Hence, populations amenable for the simultaneous detection of favorable alleles at multiple tolerance loci and for selecting transgressive progeny are needed. Here, we explore the potential of using BRP13R, a sorghum multiparental random mating population (MP-RMP) constructed based on the nuclear male sterility gene, ms_3 (Webster 1965), for such an endeavor. Populations such as BRP13R are commonly used in recurrent selection schemes in crop pre-breeding, which may potentially narrow the gap between gene discovery and applications in cultivar development.

Acidic soils ($pH \leq 5$) are prevalent in the tropics and subtropics, occupying more than half of the world arable lands (Von Uexküll and Mutert 1995). In sub-Saharan Africa,

where sorghum is a staple food, 25% of the soils are acidic (FAO 2015; Tully et al. 2015). The highly weathered nature of acidic soils results in enrichment of aluminum (Al) and iron (Fe) oxides in the soil clay fraction (Shaw 2001), which is a central aspect leading to multiple abiotic stresses that significantly reduce crop yields, and hence food security worldwide (reviewed by Magalhaes et al. (2018)).

Under low pH, Al solubilizes into its ionic form, Al^{3+}, which damages the root system and impairs root growth (Kochian 1995) into deeper soils layers. Therefore, Al toxicity reduces grain yield due to restricted uptake of mineral nutrients and water (Foy et al. 1993). Due to Al toxicity in an acidic soil, we showed that sorghum grain yield was reduced by about 23% or one ton ha^{-1}, compared to an adjacent non-Al-toxic field site (Carvalho et al. 2016). Phosphorus (P) diffusion on tropical soils is strongly constrained due to the formation of stable complexes between P and soil Al and Fe oxides (Marschner 1995; Lynch 2011), leading to very low-P availability for crop uptake. Furthermore, P diffusion is severely limited by reductions in soil water content on tropical soils, even when those that are still not nearly enough to cause drought stress (Ruiz et al. 1988). Hence, for non-irrigated crops cultivated on acidic soils, P stress is also a common limiting factor during the crop growth cycle. Therefore, acidic soil impact on crop yields results from a rather complex interplay of different abiotic stresses, which is further worsened by the often ubiquitous occurrence of drought stress. In terms of sorghum adaptation to low-P conditions, we have previously established that grain yield under low-P largely reflects P acquisition efficiency, which is the most important P efficiency component in sorghum (Bernardino et al. 2019). Among root traits, total root surface area (SA) and root diameter (RD) are important for grain yield under low-P availability in the soil (Hufnagel et al. 2014; Bernardino et al. 2019).

Some of the molecular determinants and related physiological mechanisms that contribute to sorghum adaptation to acidic tropical soils have been revealed. The Al-activated citrate transporter, *SbMATE*, which mediates sorghum Al tolerance by promoting Al detoxification via citrate release into the rhizosphere (Magalhaes et al. 2007), has been shown to increase grain yield by over one ton ha^{-1} for both sorghum lines and hybrids harboring superior *SbMATE* alleles, when grown in an Al-toxic soil (Carvalho et al. 2016). In addition, *SbMATE*-specific single nucleotide polymorphism (SNP) markers have been associated with grain yield under low-P availability in West Africa, suggesting a pleiotropic effect of *SbMATE* also enhancing P acquisition (Leiser et al. 2014). We also found that sorghum homologs of the rice (*Oryza sativa*) serine/threonine receptor kinase, *phosphorus-starvation tolerance1* (*OsPSTOL1*) (Gamuyao et al. 2012), were associated with root morphology changes, such as increased root surface area, leading to grain yield increases under low-P availability in the soil (Hufnagel et al. 2014; Bernardino et al. 2019). In addition, either quantitative trait loci (QTLs) or anonymous SNP loci associated with abiotic stress tolerance in sorghum (Mace and Jordan 2011; Leiser et al. 2014; Parra-Londono et al. 2018; Mace et al. 2019), including stay green QTLs that enhance grain yield under drought stress (Harris et al. 2006; Sabadin et al. 2012), are expected to lead to the isolation of novel abiotic stress tolerance genes in sorghum.

In order to efficiently integrate multiple abiotic stress tolerance loci into sorghum breeding, detection strategies and appropriate target populations should be carefully designed. Provided that proper attention is directed to the occurrence of false positives, the population flexibility provided by association mapping approaches (Yu and Buckler 2006) can facilitate tolerance loci detection directly on the breeder's germplasm, within a multi-allelic context, which can facilitate progeny selection. In the case where inferences are made directly in the target population, more readily available applications for crop improvement can be expected (Breseghello and Sorrells 2006). Although they explore a narrower allelic range with in general less resolution, genetic mapping using biparental crosses, such as with recombinant inbred lines (RILs), is an important complementary approach to association mapping, particularly by providing higher detection power for quantitative trait locus (QTL) (Breseghello and Sorrells 2006). A middle ground between biparental crosses and association panels in terms of population structure, genetic diversity, the number of traits that can be investigated, resolution and power are provided by multiparental populations, such as Multiparent Advanced Generation Intercross populations (MAGIC) (Mackay and Powell 2007; Stadlmeier et al. 2018). Eight-parent MAGIC populations have been shown to capture a high proportion of the allelic diversity available in the German wheat breeding gene pool (Stadlmeier et al. 2018) and have been deemed adequate for high-resolution mapping of quantitative trait loci (Mackay et al. 2014). Nested association mapping (NAM) approaches, where diverse founders are crossed to a common parent to produce sets of mapping populations, minimize genetic background effects on QTL detection and increase detection power (Yu et al. 2008). Such approaches have been shown to lead to more consistent detection of phenology QTL compared to association mapping in sorghum (Bouchet et al. 2017), and to enhance detection of putative multiple small-effect alleles influencing flowering time (Mace et al. 2013).

We focus here on exploring the consequences of enhanced recombination via randomly mating multiple parents repeatedly throughout the genesis of BRP13R, focusing on simultaneously detecting loci related to abiotic stress tolerance by GWAS and deploying previously identified tolerance loci into a pre-breeding pipeline. In the context of BRP13R,

we also investigate the potential of genomic prediction as a tool to assist sorghum breeding efforts with the final goal of selecting progeny with broad adaptation to tropical soils with low-P availability and Al toxicity.

Materials and methods

Genetic material

The steps in the development of BRP13R are shown in Fig. S1. Male sterile plants (ms_3ms_3) from the Nebraska Random Mating Population 3 (NRP3R) were crossed with 100 sorghum fertility restorer (R) lines from the world collection selected for grain protein content, giving rise to the Purdue Population 3R (PP3R, Robert Schaffert, personal communication). PPR3 was then subjected to 6 cycles of recombination to generate the Brazilian Random Mating Population 3 (BRP3R); subsequently, male-sterile (ms_3ms_3) plants from BRP3R were crossed with 24 R-lines selected for Al tolerance, P efficiency and other desirable traits (Table S1). This random mating population was designated BRP13R. Fertile F_1 plants were self-pollinated to produce F_2 seeds, which were recombined (first recombination cycle). Seeds of sterile plants were harvested in bulk for the second recombination cycle. After the third recombination cycle, approximately 600 seedlings were phenotyped for Al tolerance in nutrient solution.

Al-tolerant plants were selected and transplanted to pots in the greenhouse. Fertile and sterile plants were then self-pollinated or crossed with a composite pollen sample of the population, respectively. Seeds were harvested in bulk and planted in an isolated field for recombination purposes. Seeds derived from the sterile plants produced BRP13R S_0 progeny. Two hundred and ten fertile S_0 plants, with plant height between 100 and 150 cm (Ms_3ms_3), were self-pollinated producing $S_{0:1}$ progeny. One fertile plant of each $S_{0:1}$ progeny was self-pollinated, originating $S_{0:2}$ progeny.

Phenotyping

Phenotyping in a low-P soil

Two field experiments were conducted at the experimental station of Embrapa Maize and Sorghum, in Sete Lagoas, Minas Gerais, Brazil, during the summer season of 2014. The two experiments were conducted side-by-side in contiguous sub-areas within the same general area and at the same time; thus this division was adopted only for operational reasons, given the field area and the lattice design. Therefore, there is no noticeable between-experiment difference, except for the progeny that composed each experiment (see below). The experimental area is a weathered tropical soil with low-P availability, containing 2.57 ppm P (±0.57 standard deviation, s.d.) (Mehlich 1) in the top soil (0—20 cm) and 1.25 ppm P (±0.30 s.d.) in the subsoil (20—40 cm). Two hundred $S_{0:2}$ progeny were arranged in two experiments each consisting of a 10 (progeny) × 10 (incomplete block) lattice design with two added checks per block (BR007 and SC283) and three replicates. Each plot consisted of two 3-m rows, with 0.45 m between rows and 8 plants m^{-1}. Fertilization was applied as 150 kg ha^{-1} of 20-00-20 (NPK) at sowing and 200 kg ha^{-1} of urea 30 days after, and the experiments were sprinkler-irrigated when necessary.

The traits measured were: grain yield (Gy, ton ha^{-1}), flowering time (FT, days), plant height (PH, cm), plant phosphorus content (Pp, ton ha^{-1}), grain phosphorus content (Pg, ton ha^{-1}), total phosphorus content (Pt, ton ha^{-1}), plant dry matter (PDM, ton ha^{-1}) and grain dry matter (GDM, ton ha^{-1}). Plant and grain tissues, collected by plot, were dried at 65 °C until constant weight, ground and homogenized, and P content was assessed in 20 g subsamples using inductively coupled argon plasma emission spectrometry.

Phenotyping of root system morphology in nutrient solution with low-P availability

Assessment of root system morphology under low-P was undertaken in nutrient solution as described by Sousa et al. (2012) and Hufnagel et al. (2014) in a randomized block design with three replicates. Seeds were sterilized with sodium hypochlorite (5%), washed with distilled water and germinated in paper rolls. After 4 days, uniform seedlings of each progeny were transferred to moistened germination papers placed in paper pouches (24 × 33 × 0.02 cm) (Hund et al. 2009).

Each experimental unit consisted of one pouch with three seedlings per pouch, whose bottom (3 cm) was immersed in containers with 5 L of nutrient solution as described by Magnavaca et al. (1987) at pH 5.6 and 2.5 µM P. The containers were kept in a growth chamber for 13 days with 12 h of photoperiod, 27 °C day and 20 °C night, and continuous aeration.

After 13 days, the root system was photographed with a digital camera Nikon D300S SLR, and the images were analyzed with the RootReader2D (https://www.plantmineralnutrition.net/software/rootreader2d/) software and WinRhizo (https://www.regent.qc.ca/) software. The traits measured were: root length (RL, cm); root diameter (RD, mm); total root surface area (SA, cm^2); surface area of super-fine roots (SA1, cm^2—0 mm < RD ≤ 1 mm); surface area of fine roots (SA2, cm^2—1 mm < RD ≤ 2 mm); surface area of thicker roots (SA3, cm^2—2 mm < RD ≤ 4.5 mm); root volume (RV, cm^3); volume of fine roots (V2, cm^3—1 mm < RD ≤ 2 mm); shoot dry matter (SDM, g); root dry matter (RDM, g); shoot

phosphorus content (Ps, g); and root phosphorus content (Pr, g).

Al tolerance in nutrient solution

Al tolerance was assessed in nutrient solution by measuring Al-inhibition of root growth as described in Caniato et al. (2007). Seed sterilization and germination were as described above but with a 3-day germination period. After germination, uniform seedlings were transferred to containers (49 seedlings per container) in a growth chamber with a photoperiod of 12 h, 27 °C day and 20 °C night temperatures under continuous aeration without stress.

After 24 h, the nutrient solution of half of the trays was replaced by an identical solution without Al (control containers), whereas the remaining trays received nutrient solution with {27} µM Al^{3+} (braces indicate Al^{3+} activity). Aluminum was supplied as $AlK(SO_4)_2$ and the solution pH was adjusted to 4.0 with HCl. The experimental design was an augmented block, in which seven seedlings constituted one experimental plot. Each tray represented one block with seven plots, containing also four seedlings Al-sensitive (ATF13) and three Al-tolerant (ATF14) as controls.

The initial root length (IRL), the final root length after 5 days (FRL_{5d}) and net root growth ($NRG = FRL_{5d} - IRL$) were recorded and relative net root growth (RNRG) was calculated by dividing the NRG Al treatment by the NRG without Al.

Statistical analysis

The model adopted for traits assessed in a low-P soil was:

$$y_{ijkl} = \mu + E_j + R_{k(j)} + B_{l(kj)} + G_i + \varepsilon_{ijkl}.$$

y_{ijkl} is the phenotypic value of progeny i in the block l of the kth replicate, within the experiment j; μ is the overall mean; E_j is the fixed effect of the jth experiment ($j = 1, 2$); $R_{k(j)}$ is the fixed effect of replicate k ($k = 1, \ldots, 3$) in experiment j; $B_{l(kj)}$ is the random effect of block l ($l = 1, \ldots, 10, b_l \sim N(0, \sigma_b^2)$) in the replicate k, within the experiment j; G_i is the genetic effect of progeny i, which can be defined as:

$$G_i = \begin{cases} g_i & i = 1, \ldots, n_g \\ t_i & i = n_g + 1, \ldots, n_g + n_c \end{cases}.$$

g_i is the random effect of progeny i with n_g being the total number of progeny ($g_i \sim N(0, \sigma_g^2)$); t_i is the fixed effect of check i with n_c being the total number of checks.ε_{ijkl} is the experimental error for progeny i in the block l of the kth replicate within the experiment j, assuming $\varepsilon_{ijkl} \sim N(0, \sigma_e^2)$.

The model used for analyzing the hydroponic experiments with low-P conditions was:

$$y_{ij} = \mu + B_j + g_i + \varepsilon_{ij},$$

where y_{ij} is the phenotypic value of the progeny i ($i = 1, \ldots, n_g$) in the block j; μ is the overall mean; B_j is the fixed effect of block j ($j = 1, \ldots, 3$); g_i is the random genetic effect of progeny i($g_i \sim N(0, \sigma_g^2)$; and ε_{ij} is the experimental error for progeny i in the block j ($\varepsilon_{ij} \sim N(0, \sigma_e^2)$).

The model used for analyzing the hydroponic experiments with aluminum stress was:

$$y_{ij} = \mu + B_j + G_i + \varepsilon_{ij},$$

where y_{ij} is the phenotypic value of the progeny i ($i = 1, \ldots, n_g$) in incomplete block j; μ is the overall mean; B_j is the fixed effect of incomplete block j ($j = 1, \ldots, 35$), G_i is the genetic effect of progeny i, which can be defined as:

$$G_i = \begin{cases} g_i & i = 1, \ldots, n_g \\ t_i & i = n_g + 1, \ldots, n_g + n_c \end{cases}.$$

g_i is the random effect of progeny i with n_g being total number of progeny ($g_i \sim N(0, \sigma_g^2)$; t_i is the fixed effect of check i with n_c being the total number of checks; and ε_{ij} is the experimental error for progeny i in the block j, assuming $\varepsilon_{ij} \sim N(0, \sigma_e^2)$.

Fixed and random effects were tested using the Wald statistics (Wald 1943) and the likelihood ratio test (LRT) (Neyman and Pearson 1928), respectively, considering a 5% significance level (α). For all statistical models, the genetic effect of progeny was first taken as random for estimating the genetic variance component (σ_g^2) via restricted maximum likelihood (REML) and the heritability coefficient of each trait. The effect of progeny was then considered as fixed for estimating the adjusted means using best linear unbiased estimators (BLUEs) using the ASReml-R package (Butler et al. 2009). Generalized heritabilities (h^2) were estimated as proposed by Cullis et al. (Cullis et al. 2006):

$$h^2 = 1 - \frac{\bar{v}\text{BLUP}}{2\sigma_g^2},$$

where $\bar{v}BLUP$ is the average variance of the difference between two best linear unbiased predictions (BLUPs). Person's correlation coefficients (Pearson 1895) were estimated based on adjusted means using the package *Hmisc* (Harrell Jr 2015) in R software (R Core Team 2016).

Genotyping

Genomic DNA was isolated from 500 mg of vegetal tissue (eight plants per progeny), as described by Saghai-Maroof et al. (1984). DNA samples were genotyped by sequencing (Elshire et al. 2011). DNA fragments ("reads") obtained during genotyping were aligned against the sorghum reference

genome (version 2.1), using the Burrows Wheeler Aligner (BWA) (Li and Durbin 2009) software and SNP calling was performed with the GBS pipeline (Glaubitz et al. 2014) in TASSEL V (Bradbury et al. 2007). $S_{0:2}$ progeny were genotyped with *SbPSTOL1*- and *SbMATE-specific* markers (Caniato et al. 2014; Hufnagel et al. 2014) using the Allele-Specific PCR genotyping system (KASP, LGC genomics) (Robinson 2006).

Marker imputation

Missing data were imputed with Beagle (Browning and Browning 2007), which has been reported to show higher imputation accuracy for heterozygous populations and reduced computation time compared to other procedures (Nothnagel et al. 2009; Swarts et al. 2014).

At least two reads from different sister chromatids are needed for correctly calling a heterozygous genotype (Swarts et al. 2014). Thus, the probability of miscalling heterozygous genotypes is related to the read depth and can be estimated as $P(\mathrm{AA|Aa}) + P(\mathrm{aa|Aa}) = 0.5^n + 0.5^n$, where AA and aa represent genotypes homozygous for the most and least frequent alleles and n is the sequencing depth (Swarts et al. 2014). Based on that, for heterozygous genotypes with read depth 5, 6 and 7, miscalling percentages are 6.25%, 3.125% and 1.5625%, respectively. As the median read depth prior to imputation in BRP13R was 5, while selecting genotypes with read depth > 5 leads to enhanced imputation accuracy for heterozygotes, it also decreases the total number of markers left for GWAS. We thus set out to identify the imputation conditions that would balance the trade-off between imputation accuracy and the total number of markers left. Imputation accuracy was first calculated by selecting loci with read depths ≥ 5, 6 and 7, window sizes (i.e., physical distance used for haplotype inference) between 10 Kb and 10 Mb and with no filtering for missing data or selecting loci with at most 25%, 50%, 75% missing data.

Accuracy tests were performed with a total of 146,306 biallelic and polymorphic GBS SNPs. Masking was undertaken by randomly replacing twenty percent of data that had genotypic information (homozygous and heterozygous classes) for missing data. Upon imputation, accuracy was calculated by comparing imputed genotypes with the "real," observed genotypes. Finally, loci with MAF < 0.01 were eliminated.

Population structure and relatedness

The genetic relationship or kinship matrix (K) was obtained by the identity-by-state approximation, proposed by Endelman and Jannink (2012), with TASSEL V (Bradbury et al. 2007). Genetic divergence between progeny was calculated in R (R Core Team 2016) based on the Euclidean distance and clustering was undertaken with the unweighted pair group method with arithmetic mean (UPGMA) method (Sokal and Sneath 1963). We also undertook a principal component analysis based on 43,825 SNP markers to investigate the degree of population structure remaining in BRP13R after recombination using the *pcaMethods* package (Stacklies and Redestig 2016) in R.

Linkage disequilibrium

Linkage disequilibrium (LD) was calculated for each sorghum chromosome using squared genotypic correlations between pairs of loci (r^2) (Weir 2008) with the *Hmisc* package in R. To assess the extent of LD per chromosome, we first selected SNPs under significant LD based on a t test ($\alpha = 0.05$) corrected for multiple tests based on the Bonferroni correction for the total number of pairs of SNP loci (0.05/[total number of SNPs × (total number of SNPs-1)/2]). We then plotted the r^2 values of SNPs under significant LD as a function of the physical distance between pairs of SNPs. The extent of LD was determined as the physical region beyond which average r^2 reached constant, basal levels.

Association mapping

Adjusted means (BLUEs) for the different traits were used for GWAS. The association mapping analyses were performed only with markers whose genotypic classes showed frequencies above 0.05, totaling 24,485 markers. We fitted models with no correction for population structure nor relatedness (naïve model), including the kinship matrix (K), population structure (Q, with PC1 scores), or jointly incorporating population structure and relatedness (Q + K). The best model was chosen based on the Akaike information criterion (AIC) (Akaike 1973), the Bayesian information criterion (BIC) (Schwarz et al. 1978) and on Type-I error simulation via inspection of the quantile–quantile (q–q) plots of the *p* values from association analysis plotted against cumulative *p* values. For each locus, the GWAS model includes dominance as the phenotypic deviation between the heterozygous class and the mean of the two homozygous classes. The significance threshold for GWAS was determined with a Bonferroni correction (Bland and Altman 1995), calculated by dividing an alpha level of 0.05 by the number of independent genome blocks based on the estimated LD extent per chromosome.

Genomic selection (GS)

Genomic selection was undertaken for grain yield (Gy, ha^{-1}), plant height (PH, cm), plant dry matter (PDM, ha^{-1}) and aluminum tolerance (RNRG). The genomic best linear unbiased prediction (GBLUP) models examined were:

1. Model 1—GBLUP with the Additive genomic relationship matrix (GBLUP-A) estimated by the method proposed by Van Raden (VanRaden 2008);
2. Model 2—Model 1 incorporating the Dominance genomic relationship matrix (GBLUP-AD) estimated by the method proposed by Vitezica et al. (Vitezica et al. 2013);
3. Model 3—GBLUP-A incorporating Gene-specific SNP markers, for *SbPSTOL1* and *SbMATE* (Caniato et al. 2014; Hufnagel et al. 2014), as Fixed cofactors with additive genetic effects (GF-GBLUP-A);
4. Model 4—Model 3 incorporating the dominance genomic relationship matrix, and gene-specific SNP markers, for *SbPSTOL1* and *SbMATE*, as fixed cofactors with both additive and dominance genetic effects (GF-GBLUP-AD);
5. Model 5—GBLUP-A with SNP markers associated with different traits by GWAS as fixed cofactors with additive genetic effects (GWAS-GBLUP-A);
6. Model 6—Model 5 incorporating the dominance genomic relationship matrix and SNP markers associated with different traits by GWAS as fixed cofactors with both additive and dominance genetic effects (GWAS-GBLUP-AD).

Additive and dominance genomic relationship matrices were calculated using the R package *AGHmatrix* (Amadeu et al. 2016). For models incorporating GWAS or gene-specific SNPs as fixed cofactors (models 3–6), SNPs within the same LD block were removed from the estimation process of the genomic additive and dominance relationship matrices. LD blocks were defined based on the estimated LD extent per chromosome. The general model fitted was:

$$y = \mu 1 + X_1 a_f + X_2 d_f + Z_1 a_r + Z_2 d_r + e,$$

where $y(p \times 1)$ is a vector of adjusted means via BLUE, obtained by correcting the phenotypic progeny means for nuisance variables from the experimental design, for p progeny; μ is the overall mean; a_f and d_f are the vectors of additive and dominance fixed effects, respectively, for g *SbPSTOL1* and *SbMATE* genes, $a_f(g \times 1)$, or s GWAS SNPs, $a_f(s \times 1)$; $a_r(p \times 1)$ is the vector of random additive genetic effects of p progeny, with $a_r \sim N(0, A\sigma_a^2)$; $d_r(p \times 1)$ is the vector of random dominance genetic effects of p progeny, with $d_r \sim N(0, D\sigma_d^2)$; and $e(p \times 1)$ is the vector of residuals, with $e \sim N(0, I\sigma_e^2)$. $X_1(p \times g$ or $p \times s)$, $X_2(p \times g$ or $p \times s)$, $Z_1(p \times p)$ and $Z_2(p \times p)$ are incidence matrices for their respective effects, 1 is a vector of ones $(p \times 1)$, A and D are $p \times p$ additive and dominance genomic relationship matrixes, respectively, and I is a $p \times p$ identity matrix.

To avoid bias that could artificially inflate accuracy estimates, the SNP markers included as fixed effect cofactors in the genomic selection models v–vi were selected based on one hundred rounds of GWAS, using different populations constructed from randomly sampled 160 BRP13R progeny in each round. SNP loci with association signals exceeding the Bonferroni threshold in more than half of the GWAS rounds were selected as fixed cofactors. For genomic selection, BRP13R was randomly split into training/validation sets, and four distinct population sizes were compared: 100/100, 120/80, 140/60 and 160/40. The different GS models were fitted considering 100 replicates for each size of the training/validation sets. Then, the optimum size of training/validation sets was selected based on the maximization of the predictive accuracy, which was calculated as the correlation between the adjusted means obtained by BLUE and the predicted means via GBLUP models. For that, the adjusted means via BLUE of 40, 60, 80 or 100 progeny (dependent of the validation population size) taken at random were masked and compared to the predicted means using the GBLUP models. All analyses were performed with ASReml-R (Butler et al. 2009). Only markers whose genotypic classes showed frequencies above 5% were used.

Results

Trait correlations and heritability estimates

We focused on the following performance traits on a low-P tropical soil field site: grain yield, plant dry matter and plant height assessed under low-P availability in the soil. In addition, we assessed root morphology traits related to P acquisition and Al tolerance in nutrient solution and Al tolerance was assessed based on relative net root growth (RNRG) in hydroponics. Heritability estimates for root diameter and total root surface area were 0.37 and 0.51, respectively, 0.61 for grain yield and 0.57 for plant dry matter (PDM) (Table S2). Additionally, Al tolerance assessed in controlled conditions was highly heritable (0.73). Grain yield was highly correlated both with grain P content (Pg, $r = 0.76$) and total P (Pt, $r = 0.63$), as well as with plant dry matter (PDM, $r = 0.56$) (Table S3).

Genotyping-by-sequencing (GBS) and marker imputation

Single nucleotide polymorphism (SNP) markers distributed genome-wide were genotyped via GBS (Elshire et al. 2011). Before imputation, the median number of reads per genotype (read depth) was 5. About half of the reads had a depth between 1 and 5, and 43% of the reads had depths exceeding 6 (Fig. S2). There was a tendency of higher read depth toward the end of the sorghum chromosomes compared to

the centromeres, and average minor allele frequency (MAF) was in general low.

For genotypes covered by 5 reads, the probability of GBS to mistakenly call as homozygous a heterozygous genotype is 0.0625 (Swarts et al. 2014). Due to the partially heterozygous nature of half-sib progeny in BRP13R, we only kept homozygous genotypes with read depth ≥ 6 to minimize miscalling of heterozygotes (expected miscalling frequency = 0.03), while keeping adequate marker coverage in the genome. We then replaced 20% of known genotypes by missing data (i.e., masking) and imputed missing data with Beagle (Browning and Browning 2007). Accuracy for all genotypic classes (global accuracy) was very high, about 97%, irrespective of the imputation window size (Table S4), and was the highest for genotypes homozygous for major alleles at SNP loci. For imputation, we selected a window size of 500 Kb, which maximized imputation accuracy for heterozygous genotypes. Inspection of our masking procedure indicated that imputation errors for heterozygous genotypes occasionally caused them to be imputed as genotypes homozygous for the major allele. Therefore, the main effect of the incorrect imputation of heterozygotes, which was likely due to their lower frequency and sparse distribution in the genome, was to reduce the frequency of this class. After imputation, BRP13R was found to consist of 20% heterozygous genotypes with and average MAF of 0.14. Imputation more than doubled the number of markers, totaling 43,825 markers, and improved genome coverage, especially in centromeric regions (Fig. 1a and b).

Linkage disequilibrium

Linkage disequilibrium (LD) was measured using squared genotypic correlations between pairs of loci (r^2) (Weir 2008). The number of SNPs under significant LD was plotted as a function of physical distance between pairs of loci, and LD extent per chromosome was determined as the physical distance after which average r^2 values reached constant, basal levels (Fig. S3). Based on this method, LD was found to decay in a remarkable homogeneous way across chromosomes, with LD extending to 2.5 Mb on average (± 0.5 Mb, Fig. 1c). LD persisted the longest on chromosome 6 (3.5 Mb, Fig. S3), probably due to selection during breeding (Bouchet et al. 2017) acting on the linked plant height and maturity loci, *Dw2* and *Ma1* (Sabadin et al. 2012), respectively. The shortest LD extent of 2 Mb was found for chromosomes 1, 2 and 3.

Population structure and relatedness

Population structure may absorb phenotypic variance and reduce the detection power in association mapping (Kang et al. 2008). We used an identity-by-state (IBS)-based method (Endelman and Jannink 2012) implemented in TASSEL V (Bradbury et al. 2007) to assess genetic relatedness between half-sib progeny in BRP13R (Fig. 2a). The progeny kinship coefficients were tightly clustered around a mean value of 0.5 (0.5 ± 0.02). Consequently, the kinship heatmap was rather homogeneous, with the absence of strongly differentiated groups in the population. Based on the kinship heatmap and on UPGMA clustering, five clusters were detected, but average kinship for these groups was in general only slightly above the population mean (between 0.53 and 0.58). Group 1 showed an average kinship (0.75) that was higher than the population average, but this group had only three progeny. Next, we conducted a principal component analysis with 43,825 SNP markers and plotted progeny scores for the first two principal components (Fig. 2b). The groups detected by UPGMA were in general separated by the two principal components, with groups 3 and 4 tending to overlap. Group 2 was the most well-defined group, whereas progeny within the other groups were rather disperse. The 24 restorer lines used in the formation of BRP13R comprised different morphological races and geographical origins, which largely govern population structure in sorghum (Caniato et al. 2011; Bouchet et al. 2012), as well as breeding materials (Table S1). In conjunction, the population structure and relatedness results indicate that the probable substantial population structure in the initial base population was likely dissipated by recombination, resulting in little structure left in BRP13R.

Association mapping

Model selection

We initially conducted a series of model selection steps to define the most adequate model for GWAS. Inspection of the Bayesian Information Criterion (BIC) (Schwarz et al. 1978) determined that the naïve model (without correction for population structure nor relatedness) performed poorly in terms of goodness-of-fit to grain yield on low-P soil data (BIC = 10,229, Fig. S4). The model including the kinship matrix (K) produced a slight decrease in model performance compared to the naïve model, which is likely due to the highly homogeneous relatedness among BRP13R progeny (Fig. 2a). The best performing model (Fig. S4) included the first PC of our principal component analysis (PC1, Fig. 2b). The PC + K model showed reduced performance compared to the PC model, indicating that including progeny scores for PC1 alone efficiently captured the remaining population structure in BRP13R. Next, for each tested model, the probability distribution under the null hypothesis was inspected based on the quantile–quantile (q–q) plots of the *p* values from association analysis plotted against cumulative *p* values (Fig. S4). Consistent with the model fitness results, we

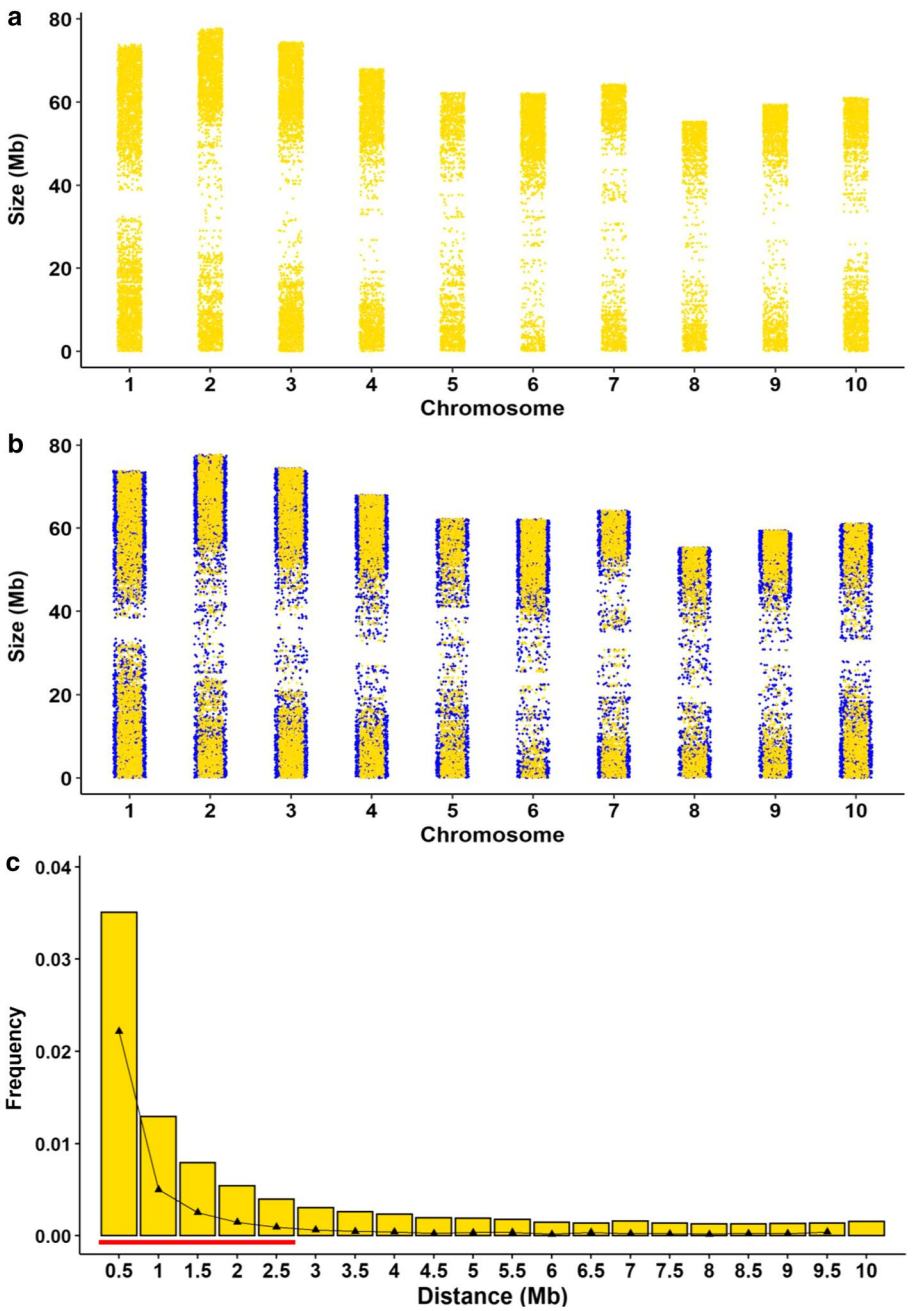

Fig. 1 Chromosome distribution of SNP loci before and after imputation and linkage disequilibrium decay. **a** Unimputed data. A maximum of 20% missing data per site and read depth ≥ 6 were allowed and the dataset contained 20,506 SNPs. **b** Imputed dataset with Beagle (Browning and Browning 2007). A maximum of 50% missing data per site was allowed and the window size was 500 Kb. The imputed dataset contained 43,825 SNPs. Blue points in panel (**b**) depict imputed markers. Both panels show the distribution of biallelic, polymorphic loci, without insertions and deletions, with a read depth ≥ 6 and MAF ≥ 0.01. Mb: megabase pairs. **c** Genomic linkage disequilibrium decay in BRP13R. The histogram shows the mean, genomic frequency of loci under significant LD, which was measured as squared genotypic correlations (r^2) between pairs of SNP loci (Weir 2008). A thin line was used to connect the difference between the proportion of loci in significant LD of the current and previous chromosome physical interval (distance in Mb). The Fisher exact test was used to assess significance followed by a Bonferroni ($\alpha = 0.05$) multiple test correction. The thick red line above the x-axis indicates the LD extent, defined as the physical distance where the average r^2 values reached constant, basal levels

observed substantial inflation of type-I error in the naïve model compared to the PC, K and PC + K model. Both the K and PC + K models showed below-diagonal *p* values, indicating reduction in detection power caused by the K matrix. Therefore, we selected the PC model for GWAS.

Genome-wide association mapping

The significance threshold for GWAS was based on the Bonferroni correction for multiple tests (Bland and Altman 1995). The number of independent tests was defined based on the extent of linkage disequilibrium estimated for each sorghum chromosome (Fig. S3) and the resulting $-\log(p)$ threshold was 3.74 ($alpha = 0.05$). As an additional false positive control, GWAS was performed only with markers whose genotypic classes showed frequencies above 5%. The GWAS profiles for the selected traits are shown in Fig. 3 and the additive and dominance effects for the respective associated SNPs, in addition to effects for SNPs associated with different auxiliary traits, are shown in Table S5. We found in total 78 significant SNP loci (Fig. 3), within which 18 SNPs associated with grain yield on low-P soil were distributed across all sorghum chromosomes, except for chromosome 7. A local, pairwise LD analysis for SNPs located within

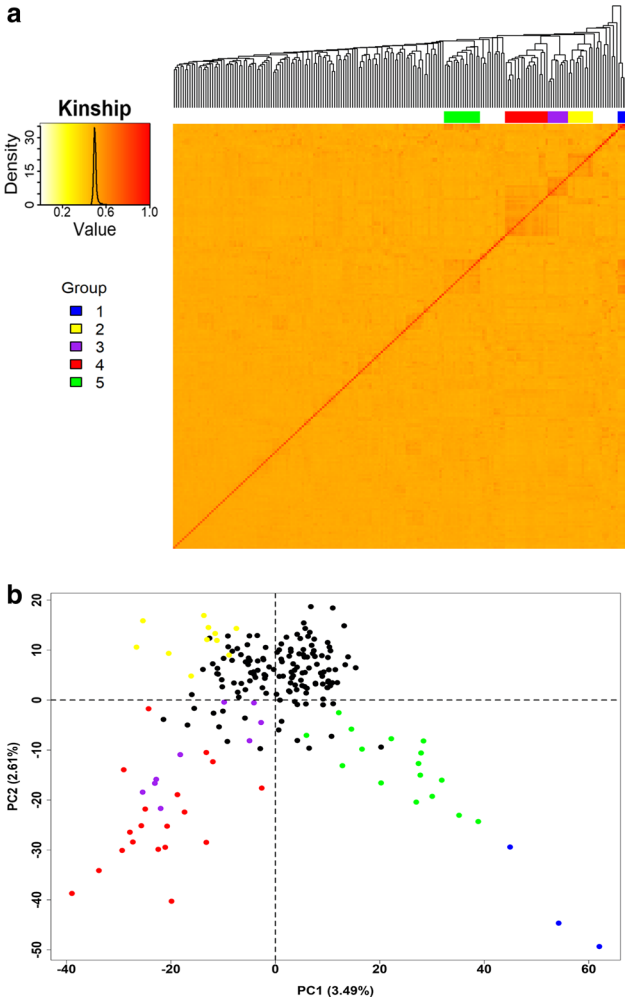

Fig. 2 Genetic relationship and population structure in 200 BRP13R progeny estimated with 43,825 SNP markers. **a** The kinship matrix was calculated with TASSEL (Bradbury et al. 2007) using an identity by state (IBS, (Endelman and Jannink 2012)) method and displayed as a heatmap and the frequency distribution of genetic relationship values are depicted (left). The unweighted pair group method with arithmetic mean (UPGMA) clustering of BRP13R progeny based on Euclidian distances is shown above the kinship heatmap. A colored scale was used to depict five differentiated groups. **b** Graphical display of progeny scores obtained by principal component analyses (PCA). Progeny belonging to the five groups identified in **a** were depicted by the same colors. The percentages of variance explained by the two PCs are shown in the axis titles

the same general physical region indicated that 16 of the 18 grain yield QTLs are in linkage equilibrium, hence possibly constituting independent loci. The grain yield effect for the associated SNPs varied from 110 to 430 kg ha^{-1} and explained 6.78 to 12.41% of the phenotypic variance. About 65% of the SNPs whose effect could be partitioned between dominance and additivity (i.e., all three genotypic classes were present) had predominantly dominant effects (Table S5). This strongly contrasts with Al tolerance, which was controlled largely by SNPs acting additively. The SNP with the highest association signal and effect on Al tolerance is located at position 71.1 Mb on chromosome 3 and co-localizes with the major Al tolerance gene, *SbMATE* (Magalhaes et al. 2007). *SbMATE* has been previously shown to control Al tolerance in hydroponics and in the field in a semi-dominant and additive way, respectively (Magalhaes et al. 2004; Carvalho et al. 2016). The effects for SNPs associated with total root surface area were evenly partitioned into additive and dominance, whereas those associated with root diameter (RD) acted mostly in an additive manner.

Association mapping with gene-specific markers

We genotyped BRP13R with a set of *SbMATE*-specific markers and SNPs within the Alt_{SB} locus where *SbMATE* is located (Magalhaes et al. 2004, 2007), which were strongly associated with Al tolerance (Caniato et al. 2014). Genotyping was also performed with markers tagging *SbPSTOL1* genes, which are sorghum homologs of rice *phosphorus-starvation tolerance1* that have been previously associated with root morphology and/or grain yield under low-P (Hufnagel et al. 2014). For gene-specific marker loci, the genotypic means for grain yield (Fig. S5a) and the association results (Fig. S5b) support a functional role of *SbMATE* and *SbPSTOL1* genes in BRP13R. *SbMATE* SNPs (in red in Fig. 3), in addition to other GBS SNPs near *SbMATE* (*Sb03g043890*) at position 71.1 Mb, were highly associated with Al tolerance (RNRG, Fig. S5b and Table S5, respectively). The association probabilities for *SbPSTOL1* SNPs with grain yield under low-P availability in the soil in BRP13R were in the same general range detected previously in a diverse sorghum association panel (Hufnagel et al. 2014). Furthermore, inspection of adjusted phenotypic means indicated that the *SbPSTOL1* alleles increasing grain yield reported in Hufnagel et al. (2014) were consistently associated with grain yield advantage in BRP13R (Fig. S5a). For example, within the *SbPSTOL1* gene *Sb03g006765*, the SNP loci (favorable allele in parenthesis), 1912 (A), 1998 (C), 2042 (G), 2067 (G), 2073 (C) and 2141 (T) were in complete LD in Hufnagel et al. (2014). In BRP13R, grain yield means for all the favorable alleles was consistently higher than that of the respective alternate alleles (Fig. S5a). In addition, although the association signal for this SNP was below significance, the A allele at the 1.541 SNP within *Sb03g031680* increased grain yield in Hufnagel et al. (2014) and the grain yield mean for this allele was again higher than the alternative allele in BRP13R (Fig. S5a), pointing toward functionality of *Sb03g031680* in BRP13R.

Next, we compared the positions of the QTLs detected in BRP13R to those previously detected in a large RIL population (Bernardino et al. 2019) and observed many instances of likely QTL conservation in the two populations (Table S6 and Fig. S6). QTLs detected in BRP13R coincide with those

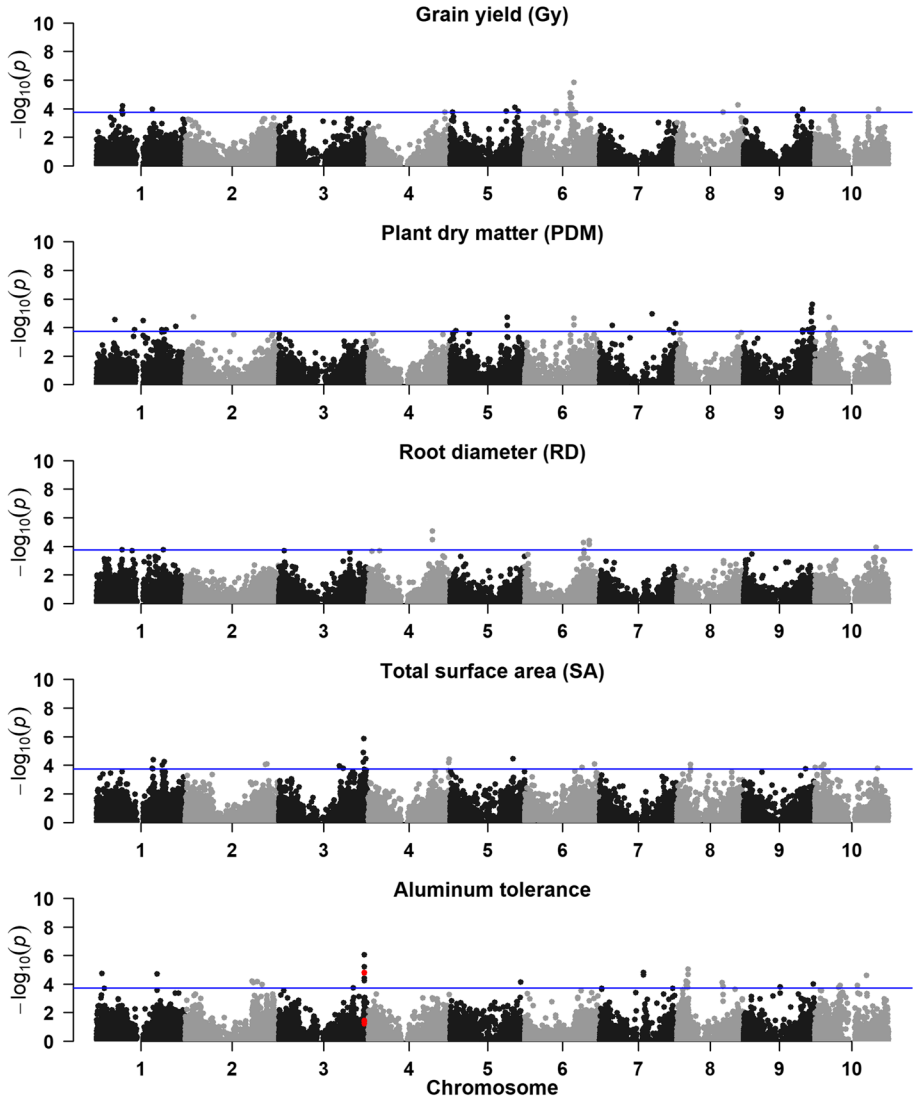

Fig. 3 GWAS profiles for grain yield (Gy, ton ha^{-1}), plant dry matter (PDM, ton ha^{-1}), root morphology traits and Al tolerance. The root morphology traits, root diameter (RD, in mm) and total surface area (SA, in cm^2), were assessed after 13 days in nutrient solution with low-P. Al tolerance was measured by relative net root growth after 5 days of ± Al exposure in nutrient solution with an Al^{3+} activity of {27} µM at pH 4.0. Colored in red are SNPs within the Alt_{SB} locus where *SbMATE* is located and within *SbMATE* itself (Caniato et al. 2014). The negative log of *p* values ($-\log_{10}(p)$) were obtained with a GWAS model including principal component 1 (PC1, Fig. 2b). The horizontal line in blue depicts the significance threshold based on the Bonferroni correction for multiple, independent tests (*alpha* = 0.05), which was defined based on the extent of LD for each sorghum chromosome

in the RILs, but BRP13R QTLs were more comprehensive genome-wide, covering regions were QTLs were not found in the RILs, either by single- or by multi-trait mapping (Fig. S6). Co-localized QTLs for grain yield within a 15 Mb window were found on chromosomes 1 (50–65), 4 (55–70 Mb), 6 (35–45 Mb), 8 (55–60 Mb), 9 (45–60 Mb) and 10 (5–20 Mb). We also found co-localized QTL for grain yield and P content under low-P availability, such as a QTL for grain P content (Pg) on chromosome 1 (5–25 Mb), and grain yield/P content QTL on chromosomes 1 (55–65 Mb), 3 (0–5 Mb), 4 (0–10 Mb), 6 (0–5 Mb), 7 (0–10 Mb), 8 (55–65 Mb), 9 (55–60 Mb) and 10 (5–20 Mb). Over half of the grain yield QTL detected in BRP13R co-localize with root morphology QTL, mainly with QTL for root surface area and occasionally with root diameter QTL, which supports the importance of root morphology in P acquisition on low-P soils (Bernardino et al. 2019). We detected more QTLs for root surface area in BRP13R in comparison with the RIL analyses, but clear instances of conserved QTL were also observed, for example, at 60–70 Mb on chromosome 2 and at 5–15 Mb and 65–75 Mb on chromosome 3, among other cases. Our design did not allow for the study of GxE. Leiser et al. (2012) looked at GxE for sorghum grain yield in multi-environment trials and found GxE to be small. The GxL (Genotype x Location) and GxY (Genotype x Year) variance components were 13% and 23%, respectively, of the genotype variance component. GxE in that study was mainly affected by the amount of annual rainfall, which was not a problem in our irrigated trials. Because of that, the −P environments were considered as one population in the Leiser et al. (2012) study, and there was also tight correspondence between genotypic performance under −P and +P conditions. Hence, in a condition such as that, as our phenotypic traits were assessed with reasonable precisions based on our heritability estimates, we do not expect dramatic impacts of additional trials, particularly for QTL

detection. Furthermore, finding significant conservation of grain yield QTL in two population contexts that are very different in terms of allele diversity, population structure and linkage disequilibrium provides support for the QTL detected in BRP13R.

Genetic makeup of selected progeny

Next, we explored the genetic constitution of BRP13R progeny selected for grain yield using a 10% selection pressure for high (designated henceforth as top 10% for simplicity) and low (bottom 10%) grain yield (Fig. 4). This analysis was undertaken with markers associated with grain yield via GWAS (Fig. 4a) as well as with our gene-specific markers for *SbSPTOL1* genes and *SbMATE* (Fig. 4b). The frequency of heterozygotes for SNP loci associated with grain yield was much higher in the top 10% of the BRP13R progeny compared to low-yielding progeny, whereas fewer loci homozygous for the favorable alleles were present in top 10% progeny (Fig. 4a). In addition, the top 10% group had much fewer progeny that were homozygous for the inferior allele. In conjunction, these results are consistent with the predominance of dominance effects for SNPs associated with grain yield and suggested a relevant role of overdominance on grain yield (Table S5). In fact, while the results in Table S5 indicated that dominance deviations were very common for loci associated with grain yield (7 in 12 SNPs), 4 of those loci apparently act strictly in an overdominant fashion (Table S7). For all those 4 loci, the *p values* contrasting the two homozygous classes were not significant (*p values* ranging from 0.54 to 0.99), suggesting the absence of additive effects and strict overdominance.

In contrast, for gene-specific markers, both loci that are homozygous for the favorable allele as well as heterozygotes were more frequent in top 10% progeny compared to low-yielding progeny (Fig. 4b). The frequency of loci in

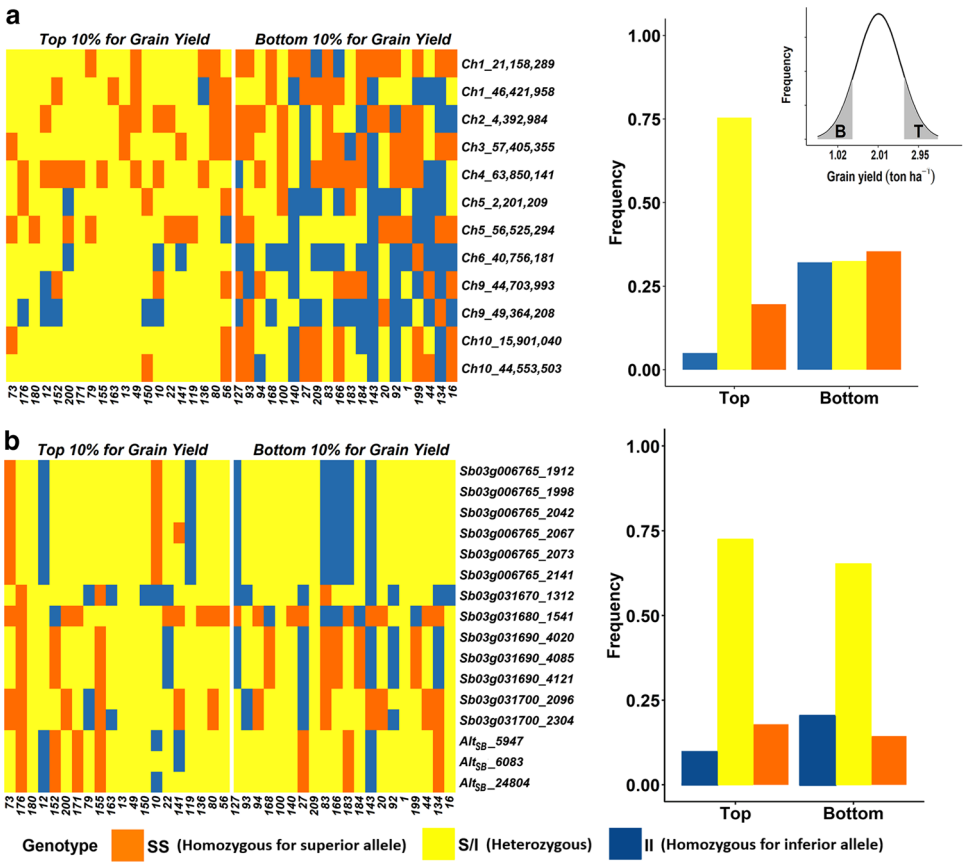

Fig. 4 Genotypic makeup of BRP13R progeny selected based on superior and inferior grain yield assessed on a low-P soil. A 10% selection threshold was imposed to select 20 progeny each with the highest (top 10%) and lowest (bottom 10%) grain yields on a low-P soil. The adjusted grain yield means for BRP13R and for the progeny in the top (T) and bottom (B) groups are shown in the histogram cartoon (top right). For each BRP13R progeny, homozygous genotypes for superior (SS) and inferior (II) alleles and heterozygotes (S/I) are depicted for: **a** SNP loci significantly associated with grain yield (Fig. 3) and (**b**) SNPs previously associated with Al tolerance (Caniato et al. 2014) and grain yield under low-P (Hufnagel et al. 2014) within the Alt_{SB} locus or *SbPSTOL1* genes (*Sb03g006765*, *Sb03g031670*, *Sb03g031680*, *Sb03g031690* and *Sb03g031700*), respectively. Genotypic frequencies for the "top" and "bottom" progeny are shown in the respective bar charts

homozygosity for the inferior allele was, in turn, higher in the low-yielding progeny. We also looked at allele frequencies per gene and calculated frequency shifts between high- and low-yielding BRP13R progeny as Δ_f (Fig. S7). Based on this analysis, all *SbPSTOL1* genes showed increased frequency of the favorable allele in high-yielding progeny. This frequency divergence was the highest in *Sb03g006765* and *Sb03g31680* and neglectable for the Al tolerance gene, *SbMATE*. An analysis of molecular variance confirmed that the high and low-yielding groups differed for allele frequencies ($p < 0.10$) for both GWAS- and gene-specific loci within the *SbPSTOL1* genes, *Sb03g006765*, *Sb03g031680*, which are associated with the highest Δ_f for grain yield (Fig. S7).

Genomic selection

Due to the intrinsically quantitative nature of sorghum adaptive traits to abiotic stresses on tropical soils, we explored the adequacy of BRP13R for genomic selection (GS), targeting grain yield under low-P availability in the soil. For that, we used models that accommodate dominance effects to take advantage of the residual heterozygosity in the multiparental population. We also studied whether the inclusion of loci associated with grain yield by GWAS as fixed effects (GWAS-SNPs) could increase prediction accuracies via GBLUP. To avoid artificially inflating accuracy estimates, the population used for GWAS to identify SNP cofactors was different than that used for genomic selection. First, SNP loci most frequently associated with grain yield after multiple rounds of GWAS, conducted in different BRP13R subsets of randomly sampled 160 progeny, were selected as cofactors. Then, for genomic selection, BRP13R was randomly split into training and prediction sets, which consisted of 100/120 and 100/80 progeny, respectively. Genomic selection was also undertaken in multiple rounds, varying the constitution of the training and the prediction sets across rounds. Accuracy was calculated as the correlation between the adjusted and the predicted means via BLUE and GBLUP, respectively, for grain yield.

In the absence of dominance effects and fixed cofactors (GBLUP-A), prediction accuracy varied from 0.22 for grain yield to 0.35 for Al tolerance (RNRG) and the traits with the highest heritability (plant height and Al tolerance, $h^2 = \sim 0.75$) also showed the highest accuracies (Fig. 5 and Table S8). Inclusion of dominance effects (GBLUP-AD) increased prediction accuracies for grain yield and plant dry matter, but only slightly. There was no advantage in using gene-specific markers for *SbPSTOL1* and *SbMATE* as cofactors (GF-GBLUP-A and -D) except for Al tolerance, where accuracy was increased in GF-GBLUP-A. In general, when used as fixed cofactors, SNPs associated with the different traits via GWAS (GWAS-GBLUP) increased prediction accuracies, except for plant dry matter. The highest prediction accuracies of 0.28, 0.53 and 0.45 for grain yield, plant height and Al tolerance, respectively, resulted from the GBLUP model which included dominance effects and GWAS-SNPs as cofactors (GWAS-GBLUP-AD). The strongest impact of including GWAS-SNPs as cofactors was observed for plant height and appears to be closely related to the presence of underlying loci with dominance effects. Although there was no advantage in including dominance effects in the absence of GWAS-SNPs (GBLUP-A vs GBLUP-AD), dominance increased prediction accuracies from 0.40 to 0.53 in the presence of GWAS-derived cofactors (GWAS-GBLUP-A vs. AD). Strikingly, accuracies increased by 90% after inclusion of GWAS-SNP cofactors in the presence of dominance effects (GBLUP-AD vs. GWAS-GBLUP-AD).

Discussion

Gene discovery and sorghum breeding with a multiparental random matting population

Different from populations such as some recombinant inbred lines and diverse association panels, the multiparental, partially selfed random mating population (MP-RMP), BRP13R, is intrinsically a breeding resource. BRP13R has been designed to dynamically incorporate into a pre-breeding pipeline new sources of alleles for desirable agronomic traits and to allow for the identification of transgressive progeny accumulating favorable alleles at multiple loci, particularly those related to sorghum adaptation to tropical soils, where abiotic stresses are common.

One significant advantage of BRP13R emerges from its power to positionally clone abiotic stress tolerance genes whose favorable alleles are present in rather low frequencies and are specific to certain subgroups, which is an enormous challenge for GWAS approaches (Brachi et al. 2011). The Al tolerance gene, *SbMATE*, has been shown to increase grain yield by over one ton ha^{-1} on an Al-toxic acid soil (Carvalho et al. 2016) and is the major determinant of Al tolerance in sorghum. Favorable alleles of *SbMATE* are rather rare and mostly specific mostly to guinea sorghums from their primary and secondary domestication centers, in West and South East Africa, respectively (Caniato et al. 2011). A GWAS approach targeting Al tolerance was performed in a diverse and highly structured association panel (Melo et al. 2019), with a model jointly including population structure and relatedness (Yu et al. 2006). Accordingly, many SNP loci distributed within most of the sorghum chromosomes showed association signals either similar to or even higher than the GBS-SNP that showed the highest association signal in the *SbMATE* region (Melo et al. 2019). Therefore, without previous knowledge, GWAS in the association panel

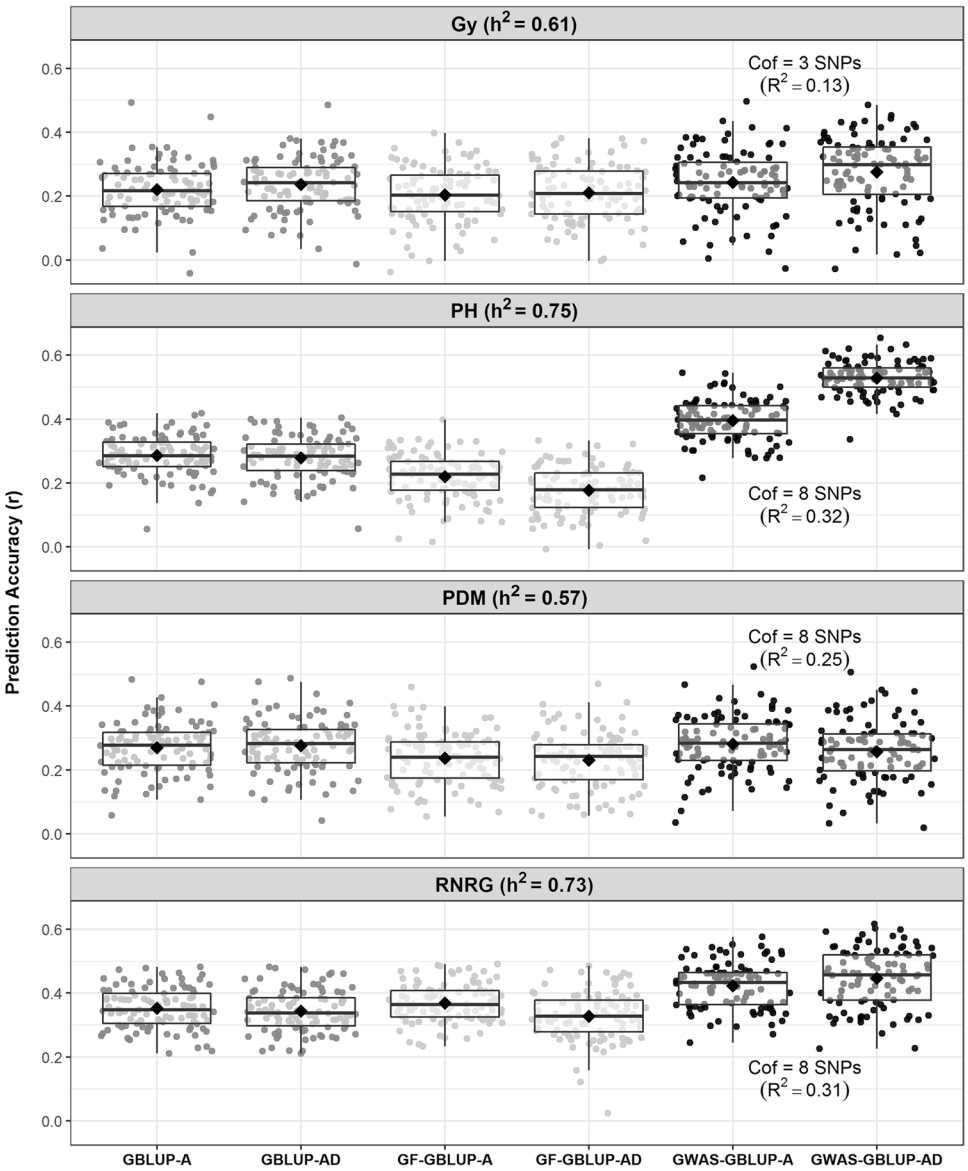

Fig. 5 Prediction accuracy (*r*) for genomic selection for grain yield (Gy), plant height (PH), plant dry matter (PDM) and Al tolerance (RNRG). Heritability (h^2) coefficients are shown for each trait. GBLUP-A is a GBLUP model with an additive genomic relationship matrix. The inclusion of a dominance genomic relationship matrix to GBLUP-A gives rise to GBLUP-AD. Gene-specific markers for *SbPSTOL1* and *SbMATE* (GF) and SNPs associated with the different traits by GWAS (GWAS-SNPs) were included as fixed cofactors both in the presence of an additive genomic relationship matrix (GF-GBLUP-A and GWAS-GBLUP-A) or including a dominance genomic relationship matrix (GF-GBLUP-AD and GWAS-GBLUP-AD) associated with the progeny random effect. Prediction accuracies were calculated as the correlation between the grain yield adjusted means via BLUE and the predicted means using GBLUP models. Cof. corresponds to the number of cofactors, that is, the number of GWAS-SNP markers fitted as fixed effects, whereas R^2 is the coefficient of determination of the full GWAS model including all the selected fixed effect cofactors for each trait

used by Melo et al. (Melo et al. 2019) would have been rather inefficient to directly positionally clone *SbMATE*, as many other candidate regions would have to be considered for gene discovery and further validation.

Roughly 265,000 SNPs have been considered adequate for GWAS approaches in sorghum, even in a highly diverse association panel, which was likely based on an average LD extent estimated to be under 10 Kb (Morris et al. 2013). Nevertheless, this assertion should be viewed with extreme caution, as LD coefficients typically show extremely high variance (Hedrick 1987) and LD fluctuations in the genome are also common due to heterogeneous recombination (Flint-Garcia et al. 2003). For example, in sorghum, Alencar Figueiredo et al. (2008) reported on remarkably variable within-gene LD, encompassing whole genes (>4 Kb) for *Opaque2*, intensive intragenic recombination happening within only 244 bp in *Waxy*, and very weak LD along *Brittle2*. Lack of salient associations of GBS-SNPs near *SbMATE* with Al tolerance is influenced by the need for population structure cofactors in the association model, the low frequency of Al tolerance and low LD in the *SbMATE* region, which was found to persist to up to around 500 bp, resulting in intragenic recombination and great haplotype diversity for the Al tolerance gene (Caniato et al. 2014; Hufnagel et al. 2018). Because rare alleles are not efficiently sampled by the skim sequencing of GBS using moderate population sizes, *SbMATE*-specific markers identified via a targeted associating mapping approach (Caniato et al. 2014), rather than any GBS SNP marker, showed by far the highest association signals for Al tolerance in the diverse association panel used by Melo et al. (2019). In contrast, using BRP13R, an extremely strong and prominent probability

peak for association between GBS markers and Al tolerance was observed kilobases away *SbMATE* and, in fact, two GBS-SNPs overlapping with *SbMATE* SNPs (in red in Fig. 3) showed even stronger association signals compared to the *SbMATE*-specific markers.

Because r^2 reflects statistical power to detect LD (Balding 2006) and is inversely proportional to the sample size required for detection (Zondervan and Cardon 2004; Wang et al. 2005), we infer that a much larger population size would have been needed to positionally clone *SbMATE* only with GBS markers via GWAS in the diverse association panel used by Melo et al. (2019). Alternatively, a GWAS approach with only 200 individuals, genotyped with ~ 44,000 SNPs in BRP13R, would have been enough to directly positionally clone *SbMATE* as previously discussed, which contrasts with the ~ 235,000 markers and 254 accessions used for GWAS by Melo et al. (2019). Hence, BRPR13R appears to offset some of the hurdles of other related multiparental designs such as MAGIC and NAM populations, some of which require long development time and large population sizes (Mackay et al. 2014; Bouchet et al. 2017) that can significantly constrain phenotyping, particularly in field experiments. The reduced population size in BRP13R leading to these results is remarkable, as in the original positional cloning of *SbMATE*, a 354-member RIL population and over 2,085 F_2 individuals had to be screened (Magalhaes et al. 2007). This extraordinary advantage of BRP13R over other populations is likely associated with its multi-allelic nature, intermediate levels of LD compared to highly diverse association panels and biparental populations, highly reduced population structure (that may also reduce detection power (Kang et al. 2008)) via random mating and selection for the target trait, which were all but natural consequences of the breeders' effort to identify transgressive segregants for hybrid development.

Genomic selection with BRP13R

Our predictive ability for grain yield in the absence of dominance effects or fixed cofactors, in the range of 0.22, was slightly lower than what was reported in previous publications on this in sorghum (Velazco et al. 2019). However, the inclusion of dominance effects and GWAS-derived fixed cofactors raised grain yield accuracies to ~ 0.3. This indicates that the multiparental, random mating nature and the residual heterozygosity in BRP13R do not preclude its effective use for genomic selection approaches, particularly if markers in LD with major genes with dominant effects are identified by GWAS and included in the GS model as fixed effects. The most dramatic increase in accuracy was achieved with a GBLUP model including dominance effects and GWAS cofactors for plant height, yielding maximum accuracy of 0.53, which is a value similar to that in previous reports (Velazco et al. 2019). In general, modeling SNPs in LD with major genes as having fixed instead of random effects has been shown to improve accuracies for traits such as plant height and flowering time in rice (Spindel et al. 2016), rust resistance in wheat (Rutkoski et al. 2014) and carotenoid levels in maize (Owens et al. 2014).

The fraction of the genetic variance jointly explained by the GWAS-derived cofactors in the GBLUP models applied to BRP13R varied from 0.13 for grain yield to 0.32 for plant height, which showed the highest heritability among all traits ($h^2 = 0.75$). Thus, our results in sorghum, particularly based on the substantial improvement in plant height prediction accuracy with GWAS-derived cofactors, agree with simulations by Bernardo et al. (Bernardo 2014) in maize. Accordingly, these authors concluded that adding fixed cofactors to the GS models is helpful, particularly for oligogenic traits and when each major gene explains more than 10% of the genetic variance.

Validation and deployment of gene-specific markers with BRP13R

Many important traits in plant breeding are quantitative in nature and are controlled by several genes, each with modest effects on the phenotype. Due to the large number of hypotheses to be tested and the consequent need to correct for multiple tests, detecting minor-effect loci is a substantial limitation of GWAS approaches, and population sizes in the range of thousands may be needed, even for alleles with a frequency of 0.15 (Hirschhorn and Daly 2005). From this perspective, we set out to explore if and how SNPs discovered by targeted approaches and explaining a much smaller portion of the genetic variance could be integrated into a pre-breeding pipeline including genomic selection under random mating.

Sorghum homologs of rice *Phosphorus-starvation tolerance1* (*OsPSTOL1)* (Gamuyao et al. 2012) were found to be associated with root morphology traits that have been shown to enhance root P acquisition and grain yield under low-P availability in the soil (Hufnagel et al. 2014), and co-localized with the respective QTLs in a sorghum RIL population (Bernardino et al. 2019). Six SNPs in total LD within the *SbPSTOL1* gene, *Sb03g006765*, in addition to one SNP within *Sb03g031680*, which individually explained ~ 3–4% of the genetic variance, were associated with increases in grain yield of about 154–200 kg ha^{-1} in a diverse sorghum association panel cultivated in a low-P soil, which likely results from increases in root surface area leading to enhanced P uptake (Hufnagel et al. 2014). We fit a linear model for grain yield under low-P including as factors the three genotypic classes for each *SbPSTOL1*-specific locus (Fig. S5) and found that the association probabilities for those SNPs in BRP13R are in the same ballpark as we

had found previously (Hufnagel et al. 2014). Although these association probabilities were not nearly as close as our GWAS threshold in BRP13R, the same six alleles within *Sb03g006765* and the single allele in *Sb03g031680* that increased grain yield in the Hufnagel et al. (2014) study also resulted in higher grain yield average compared to the alternative alleles (Fig. S5), indicating that the *SbPSTOL1* effect on grain yield in BRP13R is sound. In further support of SbPSTOL1 functionality enhancing grain yield under low-P in BRP13R, we found that the allele frequency for all *SbPSTOL1* SNPs shifted upward, favoring a higher frequency of favorable alleles (positive Δ_f) (Fig. S7) in the top 10% lines selected for grain yield compared to low-yielding BRP13R progeny. Furthermore, this shift was most dramatic for *Sb03g006765* (0.13) and *Sb03g031680* (0.10), which are exactly the two *SbPSTOL1* genes that had been previously associated with grain yield under low-P availability in the soil via targeted association mapping (Hufnagel et al. 2014). In addition, on average, we observed enrichment of genotypes homozygous for favorable alleles over progeny homozygous for the unfavorable alleles in high-yielding progeny. This pattern, which was primarily driven by *Sb03g006765*, was reversed in low-yielding progeny, where progeny homozygous for the unfavorable alleles predominate. The predominantly additive gene action that emerges from these results is again consistent with our statistical analysis (Fig. S5), which shows significance only for additive effects for gene-specific loci.

Collectively, our findings confirm that genes with more subtle effect on the target traits, consequently explaining smaller fractions of the phenotypic variance, such as *SbPSTOL1* genes, are in principle not useful for boosting accuracies in genomic selection approaches as previously predicted (Bernardo 2014). Notwithstanding, our genotype and allele frequency analyses indicate that such genes identified via targeted association mapping and other approaches, should enter into GS pipelines via allele mining and characterization of the founder germplasm and, if needed, marker-assisted introgression in the base population.

The allele shift for SNPs tagging the Al tolerance gene, *SbMATE*, was negligible between high- and low-yielding progeny and our regression analysis did not detect significant associations with grain yield under low-P for *SbMATE* loci. Conversely, our *SbMATE* markers were highly associated with Al tolerance (Fig. S5), as expected. These results can be explained by the fact that Al saturation in the soil surface was below toxicity levels for sorghum, as we wanted our low-P site to isolate the effect of low-P availability from Al toxicity on BRP13R performance. A previous study with *SbMATE*-specific markers suggested that citrate release mediated by the root plasma membrane SbMATE protein may also benefit P uptake and grain yield under low-P availability in West Africa (Leiser et al. 2014). However, P stress on tropical soils occurs via P fixation on the surfaces of Al and iron oxides in the soil clays, impairing the diffusive flux of P from the soil toward the root surface (reviewed by Magalhaes et al. (2018)). Since BRP13R was assessed for grain yield in a clay low-P soil under irrigation in Brazil, it is possible that a stronger P stress occurred in the West Africa experiments, which might have potentiated the effect of citrate release mediated by *SbMATE* on P uptake. In addition, the *SbMATE* effect on grain yield in the Leiser et al. (2014) study may result from a combined effect of *SbMATE* enhancing both P acquisition and Al tolerance (Leiser et al. 2014; Magalhaes et al. 2018). While we cannot generalize our results due to the complex nature of tropical soils, in by far the most widespread situation where Al toxicity and low-P availability co-exist on acidic soils, progeny combining favorable alleles of both *SbMATE* and *SbSPTOL1* genes are expected to be more adapted and hence show enhanced yield stability.

BRP13R as a multipurpose, integrative resource for genomics and plant breeding

We have found reassuring evidence for QTL conservation between BRP13R and a large RIL population (Bernardino et al. 2019), which were both phenotyped for grain yield under low-P availability and root morphology traits (Table S6). For example, we have previously reported on the presence of a sorghum homolog of the wheat Al tolerance gene, *ALMT*, within a grain yield QTL on chromosome 6, and of a *PHOSPHATE2* (*PHO2*) homolog co-localized with a grain yield QTL on chromosome 9 (Bernardino et al. 2019). The aluminum-activated malate transporter, ALMT1, has been recently shown to influence root growth in low-P conditions in Arabidopsis (Mora-Macías et al. 2017), while *PHO2* has been implicated in maize P efficiency (Du et al. 2018). In BRP13R, the QTL with the strongest association signal for grain yield was located on chromosome 6 at 37–45 Mb. This region overlaps at 40–45 Mb with the grain yield QTL detected both by single- and multi-trait mapping in the RIL population, and the overlapping region includes the sorghum homolog of *ALMT* at position ~ 44 Mb. At position ~ 57 Mb, *PHO2* is near a grain yield QTL in the end region of chromosome 9 and is within overlapping QTL for P content, grain, root and plant dry matter, in addition to plant height in BRP13R.

Some instances of QTLs found exclusively in BRP13R may highlight another substantial advantage of such population steaming from its residual heterozygosity. Loci acting strictly overdominantly on grain yield are not expected to be detected in a homozygous population, such as in RILs. Indeed, based on the positions of the 4 SNP loci inferred to be acting strictly overdominantly in BRP13R (Table S7), conserved QTL in our RIL population

(Bernardino et al. 2019) were in general not found, except for S1_613208087 on chromosome 10 (Table S6). Despite its largely self-pollinating nature, this supports the importance of heterosis in sorghum as previously reported (Quinby and Karper 1946), which can be caused by overdominant loci controlling grain yield (Ben-Israel et al. 2012). Therefore, in allowing for the detection of such loci, the residual heterozygosity present in BRP13R may nicely complement QTL mapping approaches in homozygous populations, thereby unlocking overdominance as a major contributor to genetic gains in sorghum grain yield.

The cloning of genes important for crop breeding, particularly the challenging ones with rather minor effects, will benefit from integrative approaches that explore complementarities between different types of populations, such as recombinant inbred lines, diverse association panels and multiparental populations. We can anticipate that such integrative resources may balance advantages and drawbacks of each type of population taken alone, arising from historical aspects influencing genetic structure, demography and diversity, which ultimately translate into variable levels of linkage disequilibrium (Nordborg and Tavaré 2002). Our study with BRP13R indicates that this type of random mating population—where many founders and derived progeny were intensively recombined—emerges as a multipurpose resource useful both for genomics and breeding applications. Such highly recombined multiparental populations increase the chances of cloning important genes by GWAS, serving as a vehicle for bridging gene discovery and cultivar development via deployment of gene-specific markers into pre-breeding efforts. Finally, boosted by genomic selection, this approach benefits cultivar development via selection of progeny transgressively accumulating favorable alleles at many loci that are important for a broader adaptation to acidic soils, such as those conferring Al tolerance and P efficiency.

Electronic supplementary material

Relevant data for genome-wide association mapping and genomic selection are available in the Electronic Supplementary Material.

Acknowledgements We acknowledge grants from the CGIAR Generation Challenge Program, the Embrapa Macroprogram, the Fundação de Amparo a Pesquisa do Estado de Minas Gerais (FAPEMIG) and the National Council for Scientific and Technological Development (CNPq). The funding body had no role in the design of the study and collection, analysis and interpretation of data and in writing the manuscript. We also thank Gislene Braga Cristeli and all the staff and trainees of Embrapa Maize and Sorghum that indirectly collaborated in the execution of this work.

Author Contribution statement JVM conceived, supervised the study and contributed to manuscript writing and revision, KCB performed experiments, analyzed the data and contributed to manuscript writing, CBM, SMS and RES contributed to sorghum phenotyping, MMP designed the statistical framework and contributed to data analysis and interpretation, BH contributed to integrative analysis of QTL conservation between RILs and BRP13R, LVK, CTG and PCSC revised the manuscript, and all authors read and approved the final manuscript.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval On behalf of all co-authors, the corresponding author states that the work described is original, previously unpublished research. All the authors listed have approved the manuscript.

References

Akaike H (1973) Information theory and an extension of maximum likelihood principle. In: Petrov BN, Czaki F (eds) Proceedings 2nd International Symposium on Information Theory. Akademia Kiado, Budapest, pp 267–281

Amadeu RR, Cellon C, Olmstead JW et al (2016) AGHmatrix: R package to construct relationship matrices for autotetraploid and diploid species: a blueberry example. Plant Genome 9:1–10

Balding DJ (2006) A tutorial on statistical methods for population association studies. Nat Rev Genet 7:781

Ben-Israel I, Kilian B, Nida H, Fridman E (2012) Heterotic trait locus (HTL) mapping identifies intra-locus interactions that underlie reproductive hybrid vigor in Sorghum bicolor. PLoS ONE 7:e38993

Bernardino KC, Pastina MM, Menezes CB et al (2019) The genetic architecture of phosphorus efficiency in sorghum involves pleiotropic QTL for root morphology and grain yield under low phosphorus availability in the soil. BMC Plant Biol 19:87

Bernardo R (2014) Genomewide selection when major genes are known. Crop Sci 54:68–75

Bland JM, Altman DG (1995) Multiple significance tests: the Bonferroni method. BMJ 310:170

Bouchet S, Olatoye MO, Marla SR et al (2017) Increased power to dissect adaptive traits in global sorghum diversity using a nested association mapping population. Genetics 206:573–585

Bouchet S, Pot D, Deu M et al (2012) Genetic structure, linkage disequilibrium and signature of selection in sorghum: lessons from physically anchored DArT markers. PLoS ONE 7:e33470

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

Bradbury PJ, Zhang Z, Kroon DE et al (2007) TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23:2633–2635. https://doi.org/10.1093/bioinformatics/btm308

Breseghello F, Sorrells ME (2006) Association mapping of kernel size and milling quality in wheat (*Triticum aestivum* L.) cultivars. Genetics 172:1165–1177

Browning SR, Browning BL (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet 81:1084–1097

Butler DG, Cullis BR, Gilmour AR, Gogel BJ (2009) ASReml-R reference manual. State Queensland, Dep Prim Ind Fish Brisbane
Caniato FF, Guimarães CT, Hamblin M et al (2011) The relationship between population structure and aluminum tolerance in cultivated sorghum. PLoS ONE 6:e20830
Caniato FF, Guimaraes CT, Schaffert RE et al (2007) Genetic diversity for aluminum tolerance in sorghum. Theor Appl Genet 114:863–876
Caniato FF, Hamblin MT, Guimaraes CT et al (2014) Association mapping provides insights into the origin and the fine structure of the sorghum aluminum tolerance locus. AltSB PLoS One 9:e87438
Carvalho G, Schaffert RE, Malosetti M et al (2016) Back to acid soil fields: The citrate transporter SbMATE is a major asset for sustainable grain yield for sorghum cultivated on acid soils. G3 Genes Genomes Genet 6:475–484. https://doi.org/10.1534/g3.115.025791
Cullis BR, Smith AB, Coombes NE (2006) On the design of early generation variety trials with correlated data. J Agric Biol Environ Stat 11:381
De Alencar Figueiredo LF, Calatayud C, Dupuits C et al (2008) Phylogeographic evidence of crop neodiversity in sorghum. Genetics 179:997–1008
de Sousa SM, Clark RT, Mendes FF et al (2012) A role for root morphology and related candidate genes in P acquisition efficiency in maize. Funct Plant Biol 39:925–935
Du Q, Wang K, Zou C et al (2018) The PILNCR1-miR399 regulatory module is important for low phosphate tolerance in maize. Plant Physiol 177:1743–1753
Elshire RJ, Glaubitz JC, Sun Q et al (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS ONE 6:e19379
Endelman JB, Jannink J-L (2012) Shrinkage estimation of the realized relationship matrix. G3 Genes Genomes Genet 2:1405–1413
FAO 2015 Status of the world's soil resources (SWSR)–main report Food Agric Organ United Nations Intergov Tech panel soils Rome, Italy 650
Flint-Garcia SA, Thornsberry JM, Buckler ES IV (2003) Structure of linkage disequilibrium in plants. Annu Rev Plant Biol 54:357–374
Foy CD, Duncan RR, Waskom RM, Miller DR (1993) Tolerance of sorghum genotypes to an acid, aluminum toxic Tatum subsoil. J Plant Nutr 16:97–127. https://doi.org/10.1080/01904169309364517
Gamuyao R, Chin JH, Pariasca-Tanaka J et al (2012) The protein kinase Pstol1 from traditional rice confers tolerance of phosphorus deficiency. Nature 488:535
Glaubitz JC, Casstevens TM, Lu F et al (2014) TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. PLoS ONE 9:e90346. https://doi.org/10.1371/journal.pone.0090346
Harrell Jr FE (2015) With contributions from Charles Dupont and many others. Hmisc: Harrell Miscellaneous. R package version 3.17-4
Harris K, Subudhi PK, Borrell A et al (2006) Sorghum stay-green QTL individually reduce post-flowering drought-induced leaf senescence. J Exp Bot 58:327–338
Hedrick PW (1987) Gametic disequilibrium measures: proceed with caution. Genetics 117:331–341
Hirschhorn JN, Daly MJ (2005) Genome-wide association studies for common diseases and complex traits. Nat Rev Genet 6:95
Hufnagel B, de Sousa SM, Assis L et al (2014) Duplicate and conquer: multiple homologs of phosphorus-starvation tolerance1 enhance phosphorus acquisition and sorghum performance on low-phosphorus soils. Plant Physiol 166:659–677
Hufnagel B, Guimaraes CT, Craft EJ et al (2018) Exploiting sorghum genetic diversity for enhanced aluminum tolerance: Allele mining based on the Alt SB locus. Sci Rep 8:10094
Hund A, Trachsel S, Stamp P (2009) Growth of axile and lateral roots of maize: I development of a phenotying platform. Plant Soil 325:335–349. https://doi.org/10.1007/s11104-009-9984-2
Kang HM, Zaitlen NA, Wade CM et al (2008) Efficient control of population structure in model organism association mapping. Genetics 178:1709–1723
Kochian LV (1995) Cellular mechanisms of aluminum toxicity and resistance in plants. Annu Rev Plant Biol 46:237–260. https://doi.org/10.1146/annurev.pp.46.060195.001321
Leiser WL, Rattunde HFW, Piepho H-P et al (2012) Selection strategy for sorghum targeting phosphorus-limited environments in West Africa: analysis of multi-environment experiments. Crop Sci 52:2517–2527
Leiser WL, Rattunde HFW, Weltzien E et al (2014) Two in one sweep: aluminum tolerance and grain yield in P-limited soils are associated to the same genomic region in West African sorghum. BMC Plant Biol 14:206. https://doi.org/10.1186/s12870-014-0206-6
Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25:1754–1760
Lynch JP (2011) Root phenes for enhanced soil exploration and phosphorus acquisition: tools for future crops. Plant Physiol 156:1041–1049. https://doi.org/10.1104/pp.111.175414
Mace E, Innes D, Hunt C et al (2019) The Sorghum QTL Atlas: a powerful tool for trait dissection, comparative genomics and crop improvement. Theor Appl Genet 132:751–766
Mace ES, Hunt CH, Jordan DR (2013) Supermodels: sorghum and maize provide mutual insight into the genetics of flowering time. Theor Appl Genet 126:1377–1395
Mace ES, Jordan DR (2011) Integrating sorghum whole genome sequence information with a compendium of sorghum QTL studies reveals uneven distribution of QTL and of gene-rich regions with significant implications for crop improvement. Theor Appl Genet 123:169
Mackay I, Powell W (2007) Methods for linkage disequilibrium mapping in crops. Trends Plant Sci 12:57–63
Mackay IJ, Bansept-Basler P, Barber T et al (2014) An eight-parent multiparent advanced generation inter-cross population for winter-sown wheat: creation, properties, and validation. G3 Genes Genomes Genet 4:1603–1610
Magalhaes J, Piñeros MA, Maciel L, Kochian L (2018) Emerging pleiotropic mechanisms underlying aluminum resistance and phosphorus acquisition on acidic soils. Front Plant Sci 9:1420
Magalhaes JV, Garvin DF, Wang Y et al (2004) Comparative mapping of a major aluminum tolerance gene in sorghum and other species in the Poaceae. Genetics 167:1905–1914
Magalhaes JV, Liu J, Guimarães CT et al (2007) A gene in the multidrug and toxic compound extrusion (MATE) family confers aluminum tolerance in sorghum. Nat Genet 39:1156–1161. https://doi.org/10.1038/ng2074
Magnavaca R, Gardner CO, Clark RB (1987) Inheritance of aluminum tolerance in maize in Genetic aspects of plant mineral nutrition. Springer, Berlin, pp 201–212
Marschner H (1995) Adaptation of plants to adverse chemical soil conditions. In: Mineral nutrition of higher plants. Academic Press, London, pp 596–657
Melo JO, Martins LGC, Barros BA et al (2019) Repeat variants for the SbMATE transporter protect sorghum roots from aluminum toxicity by transcriptional interplay in cis and trans. Proc Natl Acad Sci 116:313–318. https://doi.org/10.1073/PNAS.1808400115
Mora-Macías J, Ojeda-Rivera JO, Gutiérrez-Alanís D et al (2017) Malate-dependent Fe accumulation is a critical checkpoint in the root developmental response to low phosphate. Proc Natl Acad Sci 114:E3563–E3572
Morris GP, Ramu P, Deshpande SP et al (2013) Population genomic and genome-wide association studies of agroclimatic traits in sorghum. Proc Natl Acad Sci 110:453–458

Neyman J, Pearson ES (1928) On the use and interpretation of certain test criteria for purposes of statistical inference: part II. Biometrika 20A:263–294. https://doi.org/10.2307/2331945
Nordborg M, Tavaré S (2002) Linkage disequilibrium: what history has to tell us. Trends Genet 18:83–90
Nothnagel M, Ellinghaus D, Schreiber S et al (2009) A comprehensive evaluation of SNP genotype imputation. Hum Genet 125:163–171
Owens BF, Lipka AE, Magallanes-Lundback M et al (2014) A foundation for provitamin a biofortification of maize: genome-wide association and genomic prediction models of carotenoid levels. Genetics 198:1699–1716
Parra-Londono S, Kavka M, Samans B et al (2018) Sorghum root-system classification in contrasting P environments reveals three main rooting types and root-architecture-related marker–trait associations. Ann Bot 121:267–280
Pearson K (1895) Mathematical contributions to the theory of evolution. III. regression, heredity, and panmixia. Philos Trans R Soc London Ser A, Contain Pap a Math or Phys Character 187:253–318. https://doi.org/10.1098/rsta.1896.0007
Quinby JR, Karper RE (1946) Heterosis in sorghum resulting from the heterozygous condition of a single gene that affects duration of growth. Am J Bot 33:716–721
R Core Team (2016) R: A language and environment for statistical computing
Robinson PS (2006) EP1726664B1—Detection system for PCR assay
Ruiz HA, Fernandes B, Novais RF, Alvaes V (1988) VH Efeito da umidade do solo sobre o volume e o conteúdo de fósforo no exsudato xilemático de soja. R Bras Ci Solo 12:39–42
Rutkoski JE, Poland JA, Singh RP et al (2014) Genomic selection for quantitative adult plant stem rust resistance in wheat. Plant Genome 7:1–10
Sabadin PK, Malosetti M, Boer MP et al (2012) Studying the genetic basis of drought tolerance in sorghum by managed stress trials and adjustments for phenological and plant height differences. Theor Appl Genet 124:1389–1402
Saghai-Maroof MA, Soliman KM, Jorgensen RA, Allard RWL (1984) Ribosomal DNA spacer-length polymorphisms in barley: Mendelian inheritance, chromosomal location, and population dynamics. Proc Natl Acad Sci 81:8014–8018. https://doi.org/10.1073/pnas.81.24.8014
Schwarz G et al (1978) Estimating the dimension of a model. Ann Stat 6:461–464
Shaw JN (2001) Iron and aluminum oxide characterization for highly-weathered Alabama ultisols. Commun Soil Sci Plant Anal 32:49–64. https://doi.org/10.1081/CSS-100102992
Sokal RR, Sneath PHA (1963) Principles of numerical taxonomy. Freeman, San Francisco
Spindel JE, Begum H, Akdemir D et al (2016) Genome-wide prediction models that incorporate de novo GWAS are a powerful new tool for tropical rice improvement. Heredity (Edinb) 116:395
Stacklies W, Redestig H (2016) The pcaMethods package. Diambil dari< https//www. bioconductor. org/packages/3.3/bioc/manuals/pcaMethods/man/pcaMethods. pdf
Stadlmeier M, Hartl L, Mohler V (2018) Usefulness of a multiparent advanced generation intercross population with a greatly reduced mating design for genetic studies in winter wheat. Front Plant Sci 9:1825
Swarts K, Li H, Romero Navarro JA et al (2014) Novel methods to optimize genotypic imputation for low-coverage, next-generation sequence data in crop plants. Plant Genome 7:1–12
Tully K, Sullivan C, Weil R, Sanchez P (2015) The state of soil degradation in Sub-Saharan Africa: baselines, trajectories, and solutions. Sustainability 7:6523–6552
VanRaden PM (2008) Efficient methods to compute genomic predictions. J Dairy Sci 91:4414–4423
Velazco JG, Malosetti M, Hunt CH et al (2019) Combining pedigree and genomic information to improve prediction quality: an example in sorghum. Theor Appl Genet 132:1–13
Vitezica ZG, Varona L, Legarra A (2013) On the additive and dominant variance and covariance of individuals within the genomic selection scope. Genetics 195:1223–1230
Von Uexküll HR, Mutert E (1995) Global extent, development and economic impact of acid soils. Plant Soil 171:1–15
Wald A (1943) Tests of statistical hypotheses concerning several parameters when the number of observations is large. Trans Am Math Soc 54:426–482
Wang WYS, Barratt BJ, Clayton DG, Todd JA (2005) Genome-wide association studies: theoretical and practical concerns. Nat Rev Genet 6:109
Webster OJ (1965) Genetic studies in Sorghum vulgare (Pers.) 1. Crop Sci 5:207–210
Weir BS (2008) Linkage disequilibrium and association mapping. Annu Rev Genomics Hum Genet 9:129–142
Yu J, Buckler ES (2006) Genetic association mapping and genome organization of maize. Curr Opin Biotechnol 17:155–160
Yu J, Holland JB, McMullen MD, Buckler ES (2008) Genetic design and statistical power of nested association mapping in maize. Genetics 178:539–551
Yu J, Pressoir G, Briggs WH et al (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38:203
Zondervan KT, Cardon LR (2004) The complex interplay among factors that influence allelic association. Nat Rev Genet 5:89