Deleterious Mutation Burden and its Association with Complex Traits in Sorghum (*Sorghum bicolor*)

Ravi Valluru[1*], Elodie E. Gazave[2], Samuel B. Fernandes[3], John N. Ferguson[3], Roberto Lozano[2], Pradeep Hirannaiah[3], Tao Zuo[1,4], Patrick J. Brown[5], Andrew D.B. Leakey[3], Michael A. Gore[2], Edward S. Buckler[1,2,6], Nonoy Bandillo[1]

[1]Institute for Genomic Diversity, 175 Biotechnology Building, Cornell University, Ithaca, New York 14853, USA. [2]Plant Breeding and Genetics Section, School of Integrative Plant Science, Cornell University, Ithaca, New York 14853, USA. [3]Departments of Plant Biology and Crop Sciences, Institute for Genomic Biology, 1402 Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Illinois, USA. [5]Section of Agricultural Plant Biology, Department of Plant Sciences, University of California Davis, Davis, CA 95616, USA. [6]USDA-ARS, R. W. Holley Center, 538 Tower Road, Ithaca, New York 14853, USA. [4]Present address: Monsanto Company, St. Louis, Missouri 63167, USA.

ORCID IDs:
0000-0001-5725-5766 (RV);
0000-0001-8269-535X (SBF);
0000-0003-3603-9997 (JNF);
0000-0003-0760-4977 (RL);
0000-0001-9654-4253 (PH);
0000-0002-6581-1192 (TZ);
0000-0003-1332-711X (PJB);
0000-0001-6251-024X (ABDL);
0000-0001-6896-8024 (MAG);
0000-0002-3100-371X (ESB);
0000-0002-5941-9047 (NB);

SHORT RUNNING TITLE: Deleterious mutations in sorghum

KEYWORDS: deleterious mutations, genetic load, genome-wide predictions, mutation burden, sorghum

*CORRESPONDANCE:
Ravi Valluru
Institute for Genomic Diversity,
175 Biotechnology Building
Cornell University
Ithaca 14853, USA
rv285@cornell.edu; rvalluru285@gmail.com

ABSTRACT

Sorghum (*Sorghum bicolor* L.) is a major food cereal for millions of people worldwide. The sorghum genome, like other species, accumulates deleterious mutations, likely impacting its fitness. The lack of recombination, drift, and the coupling with favorable loci impede the removal of deleterious mutations from the genome by selection. To study how deleterious variants impact phenotypes, we identified putative deleterious mutations among ~5.5M segregating variants of 229 diverse biomass sorghum lines. We provide the whole-genome estimate of the deleterious burden in sorghum, showing that about 33 percent of nonsynonymous substitutions are putatively deleterious. The pattern of mutation burden varies appreciably among racial groups. Across racial groups, the mutation burden correlated negatively with biomass, plant height, specific leaf area (SLA), and tissue starch content (TSC), suggesting deleterious burden decreases trait fitness. Putatively deleterious variants explain roughly half of the genetic variance. However, there is only moderate improvement in total heritable variance explained for biomass (7.6%) and plant height (average of 3.1% across all stages). There is no advantage in total heritable variance for SLA and TSC. The contribution of putatively deleterious variants to phenotypic diversity therefore appears to be dependent on the genetic architecture of traits. Overall, these results suggest that incorporating putatively deleterious variants into genomic models slightly improve prediction accuracy because of extensive linkage. Knowledge of deleterious variants could be leveraged for sorghum breeding through either genome editing and/or conventional breeding that focus on the selection of progeny with fewer deleterious alleles.

Introduction

Plant genomes continually accumulate new mutations due to population demographic history (Brandvain et al. 2013), random drift (Lynch and Gabriel 1990), the mating system (Hartfield and Glémin 2014), domestication (Lu et al. 2006; Ramu et al. 2017), and linked selection due to genetic interactions (Felsenstein 1974). While a sizeable portion of such new mutations are neutral (Shaw et al. 2002; Covert et al. 2013), a small portion of new mutations are likely to be deleterious because they disrupt evolutionarily conserved sites, protein function (Yampolsky et al. 2005; Doniger et al. 2008), or gene expression (Kremling et al. 2018) in a way that results in negative impacts on fitness. The elimination of deleterious mutations from breeding populations has therefore been suggested as a prospective avenue for crop improvement (Morrell et al. 2012; Moyers et al. 2018).

Sorghum (*Sorghum bicolor* (L.), 2*n* = 20) is an important and versatile crop that is grown for food, forage, and fuel. It was domesticated from its wild ancestor about 8,000 years ago in Africa (Wendorf et al. 1992). Five major morphological forms have traditionally been recognized: bicolor, caudatum, durra, guinea, and kafir. While these races are widespread in distinct regions of Africa reflecting the diverse agro-eco-environments (Dillon et al. 2007; Evans et al. 2013), sorghum has maintained minimal genome redundancy due to the absence of any whole genome duplication for over 70 million years (Paterson et al. 2004, 2009). However, inbreeding sorghum is likely to accumulate more weakly deleterious mutations when compared to an outcrossing species, which accumulates strong recessive deleterious mutations that reduce the mean fitness of the species over time (Moyers et al. 2018). Nonetheless, there is accumulating evidence showing that enhanced homozygosity (Kumaravadivel and Rangasamy 1994), relaxed selection (Arunkumar et al. 2015), and low levels of outcrossing (Pamilo et al. 1987; Nakayama et al. 2012) can act to purge deleterious mutations leading to lower mutation burden in selfing populations. Though the relative contributions of these processes to mutation burden has long been debated, both theoretical and experimental evidence suggests that reduced population size effects usually outcompete processes that enhance purging of deleterious mutations caused by selfing (Bustamante et al. 2002; Slotte et al. 2010, 2013; Arunkumar et al. 2015) leading to an influx of deleterious mutations into selfing species.

Modern breeding and domestication results in an increased mutation burden in domesticates when compared to their wild progenitors, and a decreased mutation burden in elite cultivars when compared to landraces (Gaut et al. 2015; Ramu et al. 2017; Yang et al. 2017). The demographic history and inbreeding allow deleterious variants of weaker effect to reach appreciable frequencies owing to random drift, which can contribute significantly to mutation burden and affect fitness-related traits (Kono et al. 2016). An estimated 20 to 30% of nonsynonymous variants are deleterious in rice

(Lu et al. 2006), Arabidopsis (Günther and Schmid 2010), maize (Mezmouk and Ross-Ibarra 2014), and cassava (Ramu et al. 2017). Renaut and Rieseberg (2015) identified an excess of nonsynonymous Single Nucleotide Polymorphisms (SNPs) segregating in domesticated sunflower and globe artichoke relative to natural populations. Similarly, about 20 to 40% of protein-coding SNPs are predicted to have a deleterious allele in maize (Mezmouk and Ross-Ibarra 2014). Indeed, deleterious mutations are predicted to be enriched near regions of strong selection (Chun and Fay, 2011; Gaut et al. 2015; Kono et al. 2016), pointing to a potentially important role for deleterious variants in shaping agronomic phenotypes.

Genomic Selection (GS) can help to accelerate crop breeding when compared to conventional phenotype-based selection approaches. In Genome-Wide Prediction (GWP) models employed in GS, the genetic variance is modeled by accounting for either the biological additive and dominant effects of the markers that can potentially improve the prediction accuracy of phenotypic traits (Vitezica et al. 2013, 2016). Genes associated with complex traits carry an uncertain number of deleterious mutations distributed across the genome, and such a mutation burden contributes significantly to the total phenotypic variation of traits (Yang et al. 2017). Because deleterious mutations can occur in both homozygous and heterozygous states depending on the genetic context, trait-specific and genetic-context based GWP models could be expected to capture the phenotypic effects of deleterious mutations. Therefore, GWP models encompassing deleterious variants are expected to account for the total genetic contribution to and improve the prediction accuracy of complex traits (Yang et al. 2017). However, the improvement of GWP will depend on how strongly correlated deleterious variants are to all other variants.

In this study, we examine the contribution of putatively deleterious variants to phenotypic variation in sorghum. We used a racially, geographically, and phenotypically diverse biomass sorghum population that represents the ancestry of four major sorghum types (Brenton et al. 2016). All accessions were phenotyped for two agronomic traits, dry biomass (DBM) and plant height (PH), and for two physiological traits, specific leaf area (SLA) and tissue starch content (TSC) under field conditions. We performed whole-genome resequencing (WGS) on 229 sorghum lines and identified putative deleterious mutations in the genome. The main objectives of this study were to determine (1) whether empirical patterns of deleterious mutation burden differ among sorghum racial groups; and (2) whether deleterious variants improve prediction accuracy of complex traits, and if so, whether such accuracy differs among phenotypic traits that have different genetic architecture. To address these questions, we first identified the putative deleterious mutations and their biological effect sizes and then, estimated an individual mutation burden and its relationship with phenotypic traits. Taking

advantage of a Bayesian genomic selection framework (Habier et al. 2011), we tested the biological significance of deleterious variants on prediction of DBM, PH, SLA, and TSC.

Materials and methods

Plant material, field experiments and phenotypic data

A biomass sorghum diversity panel assembled for the TERRA-MEPP and TERRA-WEST projects was used in this study. This panel was composed of 869 lines, 339 lines coming from Fernandes et al. (2018), 117 lines coming from Brenton et al. (2016), 273 lines coming from Yu et al. (2016), and 140 additional lines obtained from John Burke (USDA - Lubbock, TX). Although phenotypic data for the entire panel was collected, only a subset of 229 lines for which WGS data were available were used in the study. These 229 lines belong to four major races of sorghum (caudatum, durra, guinea, and kafir) with representatives from the African continent, Asia, and the Americas (Fig. S1).

Field experiments were conducted in Illinois during 2016 in an augmented block design that consisted of 960 four-row plots with a row length of 3 m, 1.5 m alleys and 0.76 m row spacing. All plots were arranged in 40 rows and 24 columns. Target density of plant population was approximately 270,368 plants ha-1 and experiments were planted in late May and harvested in early October. Plant height was measured from the ground to the uppermost leaf whorl at 7 developmental stages starting 4 weeks after planting (WAP) up to 16 WAP with an interval of 2 weeks (7 stages) and averaged across the plot. Biomass data was collected at harvesting using a four-row Kemper head attached to the John Deere 5830 tractor. A plot sampler equipment with near infra-red sensor (model 130S, RCI engineering) was used to measure the wet weight of total biomass (lb) and to quantify biomass moisture (%) and starch (%) contents of plants (Li et al. 2015) in the 2 middle rows of each four-row plot. Biomass yield in dry US tons per acre was calculated as: dry US tons per acre = total plot wet weight (lb) x (1 − plot moisture) / (plot area in acre) x 0.0005. Because some accessions had flowered (38 accessions), flowering data were recorded in 2018 (flowering data was not available for 2016). We conducted an additional set of analyses that had excluded these 38 accessions to assess the potential confounding effect of flowering time on plant height.

To estimate specific leaf area (SLA), the youngest fully expanded leaf from two randomly selected plants of the middle two rows of each plot were excised just above the ligule 60 to 70 days after planting. Damaged leaves were avoided. Excised leaves were then re-cut under water, and the cut surface kept immersed. In the laboratory, three 1.6 cm leaf discs were collected from the middle of each leaf whilst avoiding the mid-rib. Leaf discs were immediately transferred to an oven set at 60°C

for two weeks. The dry mass of leaf discs was determined, and SLA was expressed as the ratio of fresh leaf area to dry leaf mass ($cm^2\ g^{-1}$). Considering 10 days interval among the SLA sampling, we used 'date of sampling' as a term in the model to generate BLUPs.

Statistical analysis of phenotypic data

Phenotypic data analysis was conducted according to experimental design, which consisted of a series of incomplete blocks connected through common checks. The following model was used to get best linear unbiased prediction (BLUPs) for all genotypes included in the field trial:

$$y_{ijk} = \mu + g_i + e_j + b_{k(j)} + ge_{ij} + \varepsilon_{ijk}$$

where μ is the overall mean, g_i is the random effect of the i^{th} genotype, e_j is the random effect of the j^{th} location, $b_{k(j)}$ is the random effect of the k^{th} incomplete block nested within the j^{th} location, ge_{ij} represents the effect of genotype-by-environment interaction, and ε_{ijk} is the residual error for the i^{th} genotype in the k^{th} incomplete block in the j^{th} location.

For specific leaf area (SLA), we fit another model that accounted for the sampling date:

$$y_{ijkl} = \mu + g_i + e_j + b_{k(j)} + d_{l(kj)} + ge_{ij} + \varepsilon_{ijkl}$$

where μ is the overall mean, g_i is the random effect of the i^{th} genotype, e_j is the random effect of the j^{th} location, $b_{k(j)}$ is the random effect of the k^{th} incomplete block nested within the j^{th} location, $d_{l(j)}$ is the random effect of the l^{th} sampling date nested within k^{th} incomplete block and the j^{th} location, ge_{ij} represents the effect of genotype-by-environment interaction, and ε_{ijkl} is the residual error for the i^{th} genotype in the k^{th} incomplete block and l^{th} sampling date in the j^{th} location.

For the purpose of estimating the broad-sense heritability (H^2) of each phenotype, we estimated variance components using the restricted maximum likelihood. All effects were assumed to be random. Broad-sense heritability on an entry-mean basis was calculated as $H^2 = \sigma^2_G / (\sigma^2_G + \sigma^2_{GXE} /$ number of locations $+ \sigma^2_e /$ number of locations x number of replicates), where σ^2_G is the variance among accessions, σ^2_{GXE} is the accession-by-environment variance, and σ^2_e is the error variance. All analyses were conducted in R software (R Development Core Team, 2015).

Genotyping

Genomic DNA (gDNA) was extracted using the CTAB method and quantified using picogreen (Molecular Probes, Eugene Oregon, USA) on a microplate reader of Synergy HT (BioTek, Vermont, USA). After preprocessing steps of the genomic DNA samples, ten libraries were prepared (24 samples in each library) and sequenced on HiSeq 4000 (PE_2x150) using sequencing kit version 1. Fastq files were demultiplexed with the bcl2fastq v2.17.1.14 conversion software of Illumina. We used Sentieon DNAseq (Freed et al. 2017) and a series of custom bash scripts to process the raw reads. Briefly, fastq files were aligned to the Sorghum bicolor reference genome version 3.1 (https://phytozome.jgi.doe.gov). PCR duplicates were removed, base quality was recalibrated based on a 'known SNPs' file, and recalibrated files were processed through the Haplotype Caller (HC). No realignment around indels was performed. The dataset therefore contains 239 samples, corresponding to 229 unique accessions, of which 7 had 1 or 2 replicates.

To create a list of "known SNPs" for the recalibration step, the HC pipeline was run without recalibration on the list of 239 BAM files. The output was filtered removing SNPs that had a number of heterozygote genotypes across all accessions greater than 10% and/or a number of heterozygote genotypes greater than two times the number of minor alleles (hereafter referred to as "homozygosity-based filter" (Chia et al. 2012)). In addition, "SNP clusters", defined as three or more SNPs located within five base pairs (bp) were also filtered out. Clusters of SNPs are often generated by misalignment and were conservatively considered as spurious. The filtered list of SNPs was used as "known SNPs" to recalibrate the BAM files and to generate a final list of SNPs. The vcf file generated by the HC contained biallelic SNPs (n=22,359,733) and were further filtered to only retain SNPs with at least 4X coverage (n=21,865,512), and with a non-missing genotype in at least 40% of the samples (n=14,535,156). After removing SNP clusters and applying homozygosity-based filters, the final dataset contained 5,512,653 SNPs, which were used for further analyses.

Identifying putatively deleterious mutations

The substitution of amino acid effect on protein function was predicted with the SIFT algorithm (Vaser et al. 2016). A nonsynonymous mutation with a SIFT score <0.05 was defined as a putative deleterious mutation. To identify a higher confidence set of deleterious mutations, we used genomic evolutionary rate profiling (GERP>2) (Davydov et al. 2010) estimated from a multi-species whole-genome alignment of six species including *Zea mays, Oryza sativa, Setaria italica, Brachypodium distachyon, Hordeum vulgare*, and *Musa acuminate*. We therefore used both an estimate of sequence conservation (GERP>2) and protein conservation (SIFT<0.05) to identify a more conservative deleterious mutations (hereafter $HGERP_{DEL\text{-}SNPs}$) in constrained portions of the genome. Using these $HGERP_{DEL\text{-}SNPs}$, we estimated the mutation burden, which was defined as the number

of derived deleterious alleles carried by an individual divided by the total number of non-missing alleles (Vitezica et al. 2016) based on putative derived deleterious allele that was defined as a minor allele in the multi-species alignment (Yang et al. 2017). First, we counted the total number of deleterious alleles in a given genotype. Here, each allele was given a score of 0.5. If both are deleterious alleles at a given position, we counted them as 1 (0.5 for each allele). If only one allele is deleterious, then it was counted as 0.5. We sum all these homozygous (1's) and heterozygous (0.5's) deleterious alleles. Second, we counted the total number of alleles used to score deleterious alleles in a given genotype. Finally, the total number of deleterious alleles was divided by the total number of scored alleles, and the resulting ratio was defined as mutation burden.

To account for the effects of linkage, we calculated linkage disequilibrium (LD) between SNPs and identify random variants (nondeleterious) to be used as a control set to compare with deleterious variants. A subset of 100k random SNP markers were selected, and all possible pairwise r^2 values were calculated using Plink 1.9 (Chang et al. 2015). Using the 1% of all the possible pairwise calculations, we calculated the relationship of distance between markers and r^2. To define local LD structure across each chromosome, we also calculated the mean LD score (Bulik-Sullivan et al. 2015) for each marker. LD scores were calculated with a window of 1Mb using the software GCTA (Yang et al. 2011; Bulik-Sullivan et al. 2015). Each LD score was divided by the total number of SNPs within each window (Fig. S2). To identify SNPs in high LD with deleterious variants, we first explored the effect of windows size and r^2 threshold on the number of SNPs selected (Fig. S3). Given the LD pattern observed, we used a window size of 250 kb and an r^2 threshold of 0.9, meaning that if any marker within 250 kb of a deleterious variants has an r^2 of 0.9 or higher, it would be excluded from further analysis. This yielded a list of ~1 million SNPs that were in LD with deleterious SNPs, which were excluded from all SNPs. An equal proportion of 100 sets of random variants with the similar allele frequency range of deleterious variants were selected (Fig. S4).

Estimating effect sizes of deleterious and nondeleterious variants

Despite the different assumptions in genetic architecture made by the different models, and the fact that the QTL effects are not of equal size and have different genetic architectures, the simplest model RR-BLUP often performs just as well in extensive cross-validation and empirical studies. Unless indicated otherwise, effect sizes were estimated using the RR-BLUP model implemented in the R-package rrBLUP version 4.2 (Endelman 2011). We fit a model $y = 1\mu + Zu + e$, where y is a vector of BLUPs of phenotype; 1μ is an intercept vector; Z is an n x p incidence matrix (either deleterious or random variants) containing the allelic states of the p marker loci ($z = \{-1, 0, 1\}$), where −1 represents the minor allele; u is the p x 1 vector of marker effects; and e is a n x 1 vector of residuals.

Under RR-BLUP, u ~ MVN (0, $I\sigma^2_u$) where σ^2_u is the variance of the common distribution of marker effects and was estimated using restricted maximum likelihood.

Partitioning of genetic variance and genome-wide prediction

We compare the variance explained by deleterious variants to that of an equal proportion of randomly sampled variants from the distribution of non-deleterious variants. Following the method of Brenton et al. (2016), we used a two-dimensional sampling approach to create 100 equal-sized datasets of randomly sampled variants matched for minor allele frequency. For each trait, we fit the model separately for each of variant set (either deleterious variant or nondeleterious variant) and estimated phenotypic variance explained.

For each variant set (deleterious variant vs nondeleterious set), we fit a standard GBLUP model including only additive effects by fitting a linear mixed model of the following form: $y = Zg + e$, where y is a vector of BLUPs of phenotype, the vector g is a random effect, the BLUP, which represents the GEBV for each individual, and Z is a design matrix indicating observations of genotype identities, and e is a vector of residuals. The genomic estimated breeding values (GEBV) were obtained by assuming g ~ MVN (0, $K\sigma^2_g$), where σ^2_g is the additive genetic variance, and K is the square genomic relationship matrix based on SNP data, implemented in TASSEL (Bradbury et al. 2007). Predictive abilities for all traits were evaluated using a five-fold cross-validation approach repeated 100 times and were implemented in the R statistical software.

Data availability

Phenotypic and genotypic data are available through the Data Dryad digital repository, doi: ### (pending).

Results

About 33 percent of nonsynonymous substitutions are putatively deleterious

We resequenced the whole genome of 229 diverse biomass sorghum accessions, belonging to four racial groups that were selected to be representative of diverse geographical regions (Fig. S1) (Brown et al. 2011; Thurber et al. 2013). The mean sequencing depth was 5.8X resulting in a data set consisting of ~5.5M single nucleotide polymorphisms (SNPs). Out of 5.5M SNPs, about 6.3% of SNPs are located in coding regions. To determine the distribution of putatively deleterious SNPs in

coding regions of the sorghum genome, we first annotated deleterious SNPs using a SIFT score (SIFT<0.05) that predicts an amino acid substitution effect on protein function (Vaser et al. 2016). Based on SIFT score <0.05, we find that about 33% of the total nonsynonymous substitutions are putatively deleterious (average SIFT score of 0.08), while 67% are predicted as tolerated mutations (average SIFT score of 0.47). We estimated the derived allele frequency (DAF) spectrum based on 'derived allele', which is defined as a minor allele in the multi-species sequence alignment (Yang et al. 2017). Our results reveal that a large proportion of deleterious SNPs have a lower DAF (<0.05; Fig 1a). While DAF shows a negative association with GERP scores (Fig. 1b) (Yang et al. 2017), it has a positively associated pattern with SIFT scores (Fig. S5).

We then combined GERP (>2) and SIFT (<0.05) scores to identify a higher confidence set of deleterious SNPs (hereafter, $HGERP_{DEL\text{-}SNPs}$, Fig. S6). Unless otherwise indicated, all further analyses were performed using $HGERP_{DEL\text{-}SNPs}$. While the majority of $HGERP_{DEL\text{-}SNPs}$ had an average SIFT score of <0.01 (Fig. S6a), they also showed a low overall allele frequency (average MAF=0.07, Fig. S6c) that is consistent with population genetic expectations. All identified $HGERP_{DEL\text{-}SNPs}$ show comparably similar distributions among all chromosomes (P = 0.34; Fig. S6b) and arise from non-centromeric regions of the chromosomes (Fig. S7). Our results corroborate previous studies showing that selection acts on deleterious variants to keep them rare (Mezmouk and Ross-Ibarra 2014) and support a combined use of SIFT and GERP scores (Fig. 1) as effective quantitative measures of an observed variant for its long-term fitness consequences (Yang et al. 2017).

Both deleterious and nondeleterious variants exhibit different effect size distributions

We estimated the additive effect sizes explained by $HGERP_{DEL\text{-}SNPs}$ for all phenotypic traits. An equal number of nondeleterious variants are used as control, which are not in LD but have similar minor allele frequency spectrum of $HGERP_{DEL\text{-}SNPs}$ across the genome (Fig. S4). We compared the full density distribution of the effect sizes of both $HGERP_{DEL\text{-}SNPs}$ and nondeleterious variants to avoid the winner's curse (Zöllner and Pritchard 2007; Jun et al. 2018) and examined whether $HGERP_{DEL\text{-}SNPs}$ effect sizes are overall higher in magnitude compared to nondeleterious variants (Fig 2).

Our results show that the density distribution of the effect sizes of both $HGERP_{DEL\text{-}SNPs}$ and nondeleterious variants follow similar pattern albeit showing some subtle differences in the density peak and distribution. The density distribution of $HGERP_{DEL\text{-}SNPs}$ extends much farther than the distribution of nondeleterious variants both at the highest and lowest range of distribution (Fig. 2), which are similar to results of previous studies (Zöllner and Pritchard 2007; Jun et al. 2018). While such density distributions are consistent across all traits, $HGERP_{DEL\text{-}SNPs}$ show different density

peaks compared to nondeleterious variants. For some traits, $HGERP_{DEL\text{-}SNPs}$ show a reduced density peak while for height at 4WAP, $HGERP_{DEL\text{-}SNPs}$ show higher density peak compared to nondeleterious variants (Fig. 2a-j).

We then compared the empirical cumulative distribution of effect sizes of $HGERP_{DEL\text{-}SNPs}$ and nondeleterious variants. Using the two-sample Kolmogorov-Smirnov (KS) test, we demonstrate that the effect sizes of both $HGERP_{DEL\text{-}SNPs}$ and nondeleterious variants show different density pattern for all phenotypes studied (Fig. S8). This suggests that $HGERP_{DEL\text{-}SNPs}$ have more variable effect sizes compared to nondeleterious variants for all phenotypic traits. Indeed, the observed variance for estimated effects across all traits was two-fold higher for $HGERP_{DEL\text{-}SNPs}$, suggesting that $HGERP_{DEL\text{-}SNPs}$ have substantially larger and more subtle effects overall.

We also compared the means of folded distributions of both $HGERP_{DEL\text{-}SNPs}$ and nondeleterious variants. Across all phenotypes, $HGERP_{DEL\text{-}SNPs}$ have on average 30.14% (ranging 0% to 42.34%) higher effects than that observed for nondeleterious variants (Fig. 3; Fig. S9). The average effect sizes captured by $HGERP_{DEL\text{-}SNPs}$ therefore appears to have a greater effect sizes than the average effect sizes explained by nondeleterious variants, which are consistent with the previous results observed in maize (Yang et al. 2017), human (Marouli et al. 2017; Jun et al. 2018), and mouse (Ji et al. 2016).

Deleterious mutation burden varies among racial groups and negatively correlates with phenotypes

We estimated the mutation burden based on $HGERP_{DEL\text{-}SNPs}$ as the count of derived deleterious alleles carried by an individual divided by the total number of scored (non-missing) alleles (*see* Methods, Fig. 4). This reveals a substantial variation for mutation burden among racial groups ($p = 3.14 \times 10^{-05}$) based on the $HGERP_{DEL\text{-}SNPs}$ (Fig. 4a). We observed that the caudatum group is significantly higher, with an average of 36%, for homozygous mutation burden as compared to other racial groups. Compared to the median burden across all racial groups, the guinea group has a proportionately lower burden (-20%), while the caudatum group has a proportionately higher burden (+49%). On average, an individual typically carries 0.0112 (s.d. 0.006), 0.0124 (s.d. 0.006), 0.0140 (s.d. 0.006), and 0.0178 (s.d. 0.007) mutation burden in the homozygous state in the guinea, durra, kafir and caudatum groups, respectively. Across all racial groups, individual mutation burden ranges from 0.001 to 0.038 based on the $HGERP_{DEL\text{-}SNPs}$, suggesting that all racial groups showed variable mutation burden.

Given that there is a considerable amount of admixture present in sorghum lines, we checked if admixture influenced the mutation burden estimation among racial groups. We plotted the relationship between the homozygous mutation burden and the principal components derived from genome-wide SNP markers (Fig. 4b; Fig. S10). This shows that although there are admixed lines, a tendency towards a higher and lower mutation burden was observed for the caudatum and guinea groups, respectively (Fig. 4a,b). These results indicate that the deleterious mutation burden estimated based on derived deleterious allele is largely due to the genomic architecture of racial groups while it is less biased with admixture.

We further evaluated the underlying relationship of mutation burden with phenotypic traits. Four putative phenotypic fitness traits were selected for this study: dry biomass, plant height (seven developmental stages), SLA, and TSC. We selected these traits because total biomass has been explicitly used as an index of fitness in several species as it can integrate the overall capacity for survival and reproduction (Donovan et al. 2009; Younginger et al. 2017). Plant height is an ecological fitness trait that incorporate processes for coexistence along spectra of light gradients (Falster and Westoby 2003). SLA is generally regarded as a useful summary ecological trait that often strongly correlates with many key plant attributes of ecological interest (Westoby 1998; Meziane and Shipley 1999). Starch production and its utilization on the diurnal basis and its role under diverse growth conditions is regarded as a major integrator in the regulation of plant growth and hence can be considered as a determinant of plant fitness (Sulpice et al. 2009; Thalmann and Santelia 2017).

We observed a substantial phenotypic variation for all traits among racial groups (Fig. S11, biomass: $P < 0.001$; SLA: $P < 0.001$; starch: $P < 0.05$; height: $P = 5.9e^{-5}$ (4WAP), $P = 0.04$ (6WAP), $P = 3.1e^{-6}$ (8WAP), $P = 3.9e^{-6}$ (10WAP), $P = 7.5e^{-5}$ (12WAP), $P = 0.001$ (14WAP), and $P < 0.05$ (16WAP), with highly heritable variation observed for plant height ($H^2 = 0.87$) and biomass ($H^2 = 0.73$), consistent with previous studies (Brenton et al. 2016). We also found strong correlations among traits (Fig. S12).

Using a simple linear regression model between mutation burden and phenotypic traits across all racial groups, we consistently found a negative relationship of mutation burden with all phenotypic traits (Table S1). We also performed a grouped regression combining racial groups that show parallel response and show that the combined slopes further confirmed the negative correlations between mutation burden and phenotypes (Table S1). These results suggest that deleterious variants decrease trait fitness. However, the majority of these correlations are not significant except for plant height (in case of grouped regression only), indicating that the deleterious mutation burden can be strongly linked to the variation in plant height in the biomass sorghum lines studied.

Deleterious variants contribute considerably to phenotypic variation but varies substantially among traits

We tested whether incorporating putatively deleterious variants could inform GS models and improve GWP of phenotype. $HGERP_{DEL\text{-}SNPs}$ identified from WGS were used as priors and integrated into a genomic prediction framework (Fig. 5). We quantified the amount of genetic variance, heritability, and model improvement by deleterious variants and compared with that of random variants. Based on a variance partitioning approach with a two-kernel model (see Methods), the model with putatively $HGERP_{DEL\text{-}SNPs}$ explained roughly half of the genetic variance (biomass: 52%, SLA: 48%, and starch: 46%, plant height: 45%-49% (across all stages)) (Fig. 5). There was a modest improvement in total heritable variance explained for biomass (7.6%, $h^2 = 0.24$ against 0.22 for random variants) and plant height (3.1%, $h^2 = 0.33$ against 0.32 for random variants across seven developmental stages). However, there was no advantage on heritable variance for SLA and TSC (Fig. 6a,b) for $HGERP_{DEL\text{-}SNPs}$ as compared to random variants.

We addressed the potential confounding effects of flowering on plant height. We performed heritability estimates based on non-flowered lines (all flowered lines were excluded) within and across racial groups. We observed only minor nonsignificant differences on heritability and these model results are complimentary to the model results obtained using all genotypes (Fig. S13).

To evaluate the predictive ability, we performed a five-fold cross-validation, repeated 100 times which was implemented in a GBLUP model with either the $HGERP_{DEL\text{-}SNPs}$ or the nondeleterious SNP data sets. Consistent with the results of heritability, we observed an 8.1% and 7.0% improvement on predictive ability for biomass and plant height (at 10-16WAP only), respectively, while there was no improvement either for SLA and TSC or plant height at early stages (at 4-8WAP, Fig. 6c,d). These results suggest that the contribution of putatively $HGERP_{DEL\text{-}SNPs}$ to phenotypic variation varies considerably among traits.

Discussion

Sorghum, a genus that evolved across diverse environments in Africa, exhibits a wide range of phenotypic diversity (Wright 1931; Doggett 1970; Dillon et al. 2007). This raises the question of whether sorghum racial groups carry variable deleterious mutation burden, allowing the mutation consequences to be tested for phenotypic diversity. In this study, we whole-genome resequenced

229 biomass sorghum lines and defined a high confidence set of putative deleterious mutations using SIFT (<0.05) and GERP (>2) scores. All racial groups of sorghum showed variable mutation burden (ranged from 0.001-0.038) that correlated negatively with phenotypic traits. We observed that an average deleterious variant had larger biological effects than an average nondeleterious variant. We further noticed that the prediction ability of the genome-wide prediction models encompassing deleterious variants are largely trait-dependent.

Combining the criteria of SIFT (<0.05) and GERP (>2) scores, we first show that sorghum racial groups accumulate appreciable amounts of deleterious mutations in the genome, estimated to be ~33% of total nonsynonymous substitutions (Fig. 1). Although the number and frequency of such mutations within a population largely depends on effective population size, our results match well with previous studies that estimate 20 to 30% of nonsynonymous variants to be deleterious in several crop species, including model plant species (Lu et al. 2006; Günther and Schmid 2010; Mezmouk and Ross-Ibarra 2014; Ramu et al. 2017). Considering highly frequent (DAF>0.9) mutations, there are 63 nonsynonymous deleterious mutations across racial groups, and distributed across all chromosomes. These mutations could likely be a combination of variants of important domestication targets, recent pseudogenes, and some truly deleterious variants that are the product of drift (Figueiredo et al. 2008; Smith et al. 2018).

We next estimated an individual mutation burden as the count of derived deleterious alleles carried by an individual divided by the total number of scored (non-missing) alleles, which differed considerably among individuals and racial groups (Fig. 4). It is notable but expected given that different racial groups have had varying patterns of population dynamics, selection intensities, and domestication histories that could detectably alter the influx of deleterious mutations (Wendorf et al. 1992; Dillon et al. 2007; Paterson et al. 2009). Contrasting deleterious burden has previously been reported in different populations of crop species (Lu et al. 2006; Renaut and Rieseberg 2015; Ramu et al. 2017), and humans (Lohmueller et al. 2008; Simons et al. 2014; Fu et al. 2014). Comparatively, the caudatum group appears to have a higher mutation burden than the guinea group; the oldest of the specialized sorghum races (Stemler et al. 1975; Harlan et al. 1976). We propose that the higher mutation burden of the caudatum group might be potentially related to the population bottleneck, resulting in a smaller population size that increases the chances of inbreeding, genetic homogeneity, and an increased influx of deleterious mutations (Renaut and Rieseberg 2015; Yang et al. 2017; Moyers et al. 2018). On the other hand, a lower mutation burden in the guinea group might be due partly to its higher outcrossing rates, which can reach up to 20% when compared to other races (Barro-Kondombo et al. 2010; Ranwez et al. 2017). Our results, therefore, suggest that, first,

negative selection is less effective at removing weakly deleterious mutations, yielding variable mutation burden among racial groups. Second, the combined effects of a bottleneck and directional selection during domestication (Hamblin et al. 2006; Lohmueller et al. 2008) can have an important impact on the deleterious mutation burden even in smaller racial groups of sorghum in which founder events can be more frequent (Charlesworth and Wright 2001; Szövényi et al. 2014).

Although informative, our estimation of mutation burden has some important limitations. First, the deleterious mutations identified in the population were based on the degree of sequence conservation that is often poorly constructed. Second, our derivation of deleterious mutations does not include noncoding or structural variants, which can contribute substantially to the total load of deleterious mutations (Huang et al. 2017; Bastarache et al. 2018). Third, our burden estimation assumes equal fitness effects for all mutations, which is unlikely, as mutations can have different fitness effects that can vary with environments (Henn et al. 2016). Fourth, we consider the same sign of the effect when estimating the burden, which would be misestimated, as some deleterious mutations may be locally adaptive, or neutral (Vikram et al. 2015; Bastarache et al. 2018). Nonetheless, despite these caveats, our findings revealed a substantial genomic burden of deleterious mutations in sorghum.

We investigated the phenotypic effects of deleterious mutations (Table S1). We found negative correlations between mutation burden and phenotypic traits, suggesting a considerable cost of deleterious mutations on phenotypic traits (Yang et al. 2017) in a species that has been subjected to recent demographic expansion (Hamblin et al. 2006). Consistently, we find that an average deleterious variant has demonstrably larger biological effect, which could likely have an important impact contributing to heritable phenotypic variation (Fig. 2 & 3). In grasses, it has been previously shown that heritable phenotypic variation can be increased as much as 0.1-1% by new mutations (Sprague et al. 1960; Houle et al. 1996; Bataillon 2000). The fate of such large effect mutations on phenotypes is, however, unclear and it has been actively debated as to whether such mutations are attributable to unconditional deleteriousness or can grant adaptable heritable variation to diverse growing conditions (Glémin and Bataillon 2009). Nonetheless, previous studies reveal novel variations of genes resulted from postdomestication mutations in sorghum and suggest that neodiversity contributed to new adaptations (Figueiredo et al. 2008; Glemin and Bataillon 2009).

Across four traits, we find that putatively deleterious alleles explain roughly half of the genetic variance (46%-49%), but there is only a moderate improvement in total heritable variance explained for biomass (7.6%) and plant height (3.1%). Additionally, there is no advantage for SLA and TSC

(Fig 5 and 6). Such a difference in the contribution of deleterious variants to phenotypic traits was recently observed in maize where dominance contributed substantially to grain yield while phenology traits appeared to be largely additive (Yang et al. 2017). Though the effects of mutations being deleterious or compensatory depends greatly upon the genetic background into which that mutation is incorporated (Moyers et al. 2018), the trivial contributions of mutations to SLA and TSC indicate that such mutations could be either nearly neutral or negatively synergistic. Our results therefore support the proposition that deleterious mutational effects vary with phenotypic traits and appear to be often larger for fitness-related quantitative traits, while they are unclear for traits that are not directly linked to fitness (Park et al. 2011). Fitness-related quantitative traits, which are expected to have a more complex genetic architecture, could potentially carry a higher polygenic mutation burden that could considerably affect phenotypes (Purcell et al. 2014). Also, such expectations are in line with the longstanding understanding that fitness-linked quantitative traits showing directional dominance generally exhibit inbreeding depression (Wright 1931; Kelly 1999; Charlesworth and Charlesworth 1999), which indeed is strongly linked to the degree of deleterious burden in the genome (Mezmouk and Ross-Ibarra 2014).

Finally, although our study did not account for sampling error while estimating an individual deleterious variant effect, which is generally greater for rare variants (Jun et al. 2018), our heritability estimates are consistent with the prediction abilities of phenotypic traits. Our work, therefore, adds to ongoing GWP efforts for exploring the cumulative effects of deleterious mutations on phenotypic diversity (Yang et al. 2017; Moyers et al. 2018). However, since rare deleterious variants are less correlated with each other and their associations greatly suffer from low statistical power (Park et al. 2011; Auer and Lettre 2015), employing either gene- and/or family-based approaches (Auer and Lettre 2015; Ji et al. 2016; Jun et al. 2018), or leveraging the phenotypic patterns, (Bastarache et al. 2018) in which deleterious mutations have detectable phenotypic consequences, would assist in examining how rare deleterious mutations shape an individual phenotype.

Conclusions

We used phenotypic and genomic data from different racial groups of sorghum to show that sorghum accumulates an appreciable number of deleterious mutations in the genome. Mutation burden differs substantially among racial groups that negatively correlate with phenotypes. Genomic selection models encompassing deleterious mutations show variable predictive ability across traits and, given the relatively high level of population structure in sorghum, disentangling deleterious effects at the

single variant level would take a tremendous amount of effort and recombination. Deleterious variants could be prioritized through work with intermediate phenotypes or with more extensive evolutionary analysis among closely related species. Both of these avenues, if combined with high throughput genome editing and conventional breeding approaches involving parental lines with fewer deleterious variants, could be used to systematically start removing deleterious variants from elite sorghum lines.

Acknowledgements

The information, data, or work presented herein was funded in part by the Advanced Research Projects Agency-Energy (ARPA-E), U.S. Department of Energy, under Award Numbers DE-AR0000598 and DE-AR0000661. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof. The support from the U.S. Department of Agriculture, Agricultural Research Service is greatly acknowledged. We thank Robert Bukowski for assistance with SIFT pipeline, Sara Miller for editorial assistance and two reviewers and the Editor for their constructive comments on the earlier version of the manuscript.

References

Arunkumar R., R. W. Ness, S. I. Wright, and S. C. H. Barrett, 2015 The Evolution of Selfing Is Accompanied by Reduced Efficacy of Selection and Purging of Deleterious Mutations. Genetics 199: 817–829. https://doi.org/10.1534/genetics.114.172809

Auer P. L., and G. Lettre, 2015 Rare variant association studies: considerations, challenges and opportunities. Genome Med. 7. https://doi.org/10.1186/s13073-015-0138-2

Barro-Kondombo C., F. Sagnard, J. Chantereau, M. Deu, K. Vom Brocke, et al., 2010 Genetic structure among sorghum landraces as revealed by morphological variation and microsatellite markers in three agroclimatic regions of Burkina Faso. TAG Theor. Appl. Genet. Theor. Angew. Genet. 120: 1511–1523. https://doi.org/10.1007/s00122-010-1272-2

Bastarache L., J. J. Hughey, S. Hebbring, J. Marlo, W. Zhao, et al., 2018 Phenotype risk scores identify patients with unrecognized Mendelian disease patterns. Science 359: 1233–1239. https://doi.org/10.1126/science.aal4043

621 Bataillon T., 2000 Estimation of spontaneous genome-wide mutation rate parameters: whither
622 beneficial mutations? Heredity 84: 497–501. https://doi.org/10.1046/j.1365-
623 2540.2000.00727.x

624 Bradbury P. J., Z. Zhang, D. E. Kroon, T. M. Casstevens, Y. Ramdoss, et al., 2007 TASSEL:
625 software for association mapping of complex traits in diverse samples. Bioinforma. Oxf.
626 Engl. 23: 2633–2635. https://doi.org/10.1093/bioinformatics/btm308

627 Brandvain Y., T. Slotte, K. M. Hazzouri, S. I. Wright, and G. Coop, 2013 Genomic Identification of
628 Founding Haplotypes Reveals the History of the Selfing Species Capsella rubella. PLoS
629 Genet. 9. https://doi.org/10.1371/journal.pgen.1003754

630 Brenton Z. W., E. A. Cooper, M. T. Myers, R. E. Boyles, N. Shakoor, et al., 2016 A Genomic
631 Resource for the Development, Improvement, and Exploitation of Sorghum for Bioenergy.
632 Genetics 204: 21–33. https://doi.org/10.1534/genetics.115.183947

633 Brown P. J., S. Myles, and S. Kresovich, 2011 Genetic Support for Phenotype-based Racial
634 Classification in Sorghum. Crop Sci. 51: 224–230.
635 https://doi.org/10.2135/cropsci2010.03.0179

636 Bulik-Sullivan B. K., P.-R. Loh, H. K. Finucane, S. Ripke, J. Yang, et al., 2015 LD Score regression
637 distinguishes confounding from polygenicity in genome-wide association studies. Nat.
638 Genet. 47: 291–295. https://doi.org/10.1038/ng.3211

639 Bustamante C. D., R. Nielsen, S. A. Sawyer, K. M. Olsen, M. D. Purugganan, et al., 2002 The cost
640 of inbreeding in Arabidopsis. Nature 416: 531–534. https://doi.org/10.1038/416531a

641 Chang C. C., C. C. Chow, L. C. Tellier, S. Vattikuti, S. M. Purcell, et al., 2015 Second-generation
642 PLINK: rising to the challenge of larger and richer datasets. GigaScience 4: 7.
643 https://doi.org/10.1186/s13742-015-0047-8

644 Charlesworth B., and D. Charlesworth, 1999 The genetic basis of inbreeding depression. Genet.
645 Res. 74: 329–340.

646 Charlesworth D., and S. I. Wright, 2001 Breeding systems and genome evolution. Curr. Opin.
647 Genet. Dev. 11: 685–690.

648 Chia J.-M., C. Song, P. J. Bradbury, D. Costich, N. de Leon, et al., 2012 Maize HapMap2 identifies
649 extant variation from a genome in flux. Nat. Genet. 44: 803–807.
650 https://doi.org/10.1038/ng.2313

651 Chun S., and J. C. Fay, 2011 Evidence for hitchhiking of deleterious mutations within the human
652 genome. PLOS Genet. 7(8): e1002240. https://doi.org/10.1371/journal.pgen.1002240
653
654 Covert A. W., R. E. Lenski, C. O. Wilke, and C. Ofria, 2013 Experiments on the role of deleterious
655 mutations as stepping stones in adaptive evolution. Proc. Natl. Acad. Sci. 110: E3171–
656 E3178. https://doi.org/10.1073/pnas.1313424110

Davydov E. V., D. L. Goode, M. Sirota, G. M. Cooper, A. Sidow, et al., 2010 Identifying a High Fraction of the Human Genome to be under Selective Constraint Using GERP++. PLOS Comput. Biol. 6: e1001025. https://doi.org/10.1371/journal.pcbi.1001025

Dillon S. L., F. M. Shapter, R. J. Henry, G. Cordeiro, L. Izquierdo, et al., 2007 Domestication to Crop Improvement: Genetic Resources for Sorghum and Saccharum (Andropogoneae). Ann. Bot. 100: 975–989. https://doi.org/10.1093/aob/mcm192

Doggett H., 1970 Sorghum. [London] : Longmans.

Doniger S. W., H. S. Kim, D. Swain, D. Corcuera, M. Williams, et al., 2008 A Catalog of Neutral and Deleterious Polymorphism in Yeast. PLOS Genet. 4: e1000183. https://doi.org/10.1371/journal.pgen.1000183

Donovan L. A., F. Ludwig, D. M. Rosenthal, L. H. Rieseberg, and S. A. Dudley, 2009 Phenotypic selection on leaf ecophysiological traits in Helianthus. New Phytol. 183: 868–879. https://doi.org/10.1111/j.1469-8137.2009.02916.x

Endelman J. B., 2011 Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. Plant Genome 4: 250–255. https://doi.org/10.3835/plantgenome2011.08.0024

Evans J., R. F. McCormick, D. Morishige, S. N. Olson, B. Weers, et al., 2013 Extensive Variation in the Density and Distribution of DNA Polymorphism in Sorghum Genomes. PLOS ONE 8: e79192. https://doi.org/10.1371/journal.pone.0079192

Falster D. S., and M. Westoby, 2003 Plant height and evolutionary games. Trends Ecol. Evol. 18: 337–343. https://doi.org/10.1016/S0169-5347(03)00061-2

Felsenstein J., 1974 The Evolutionary Advantage of Recombination. Genetics 78: 737–756.

Fernandes S. B., K. O. G. Dias, D. F. Ferreira, and P. J. Brown, 2018 Efficiency of multi-trait, indirect, and trait-assisted genomic selection for improvement of biomass sorghum. Theor. Appl. Genet. 131: 747–755. https://doi.org/10.1007/s00122-017-3033-y

Figueiredo L. F. de A., C. Calatayud, C. Dupuits, C. Billot, J.-F. Rami, et al., 2008 Phylogeographic Evidence of Crop Neodiversity in Sorghum. Genetics 179: 997–1008. https://doi.org/10.1534/genetics.108.087312

Freed D. N., R. Aldana, J. A. Weber, and J. S. Edwards, 2017 The Sentieon Genomics Tools - A fast and accurate solution to variant calling from next-generation sequence data. bioRxiv 115717. https://doi.org/10.1101/115717

Fu W., R. M. Gittelman, M. J. Bamshad, and J. M. Akey, 2014 Characteristics of Neutral and Deleterious Protein-Coding Variation among Individuals and Populations. Am. J. Hum. Genet. 95: 421–436. https://doi.org/10.1016/j.ajhg.2014.09.006

690 Gaut B. S., C. M. Díez, and P. L. Morrell, 2015 Genomics and the Contrasting Dynamics of Annual
691 and Perennial Domestication. Trends Genet. 31: 709–719.
692 https://doi.org/10.1016/j.tig.2015.10.002

693 Glémin S., and T. Bataillon, 2009 A comparative view of the evolution of grasses under
694 domestication. New Phytol. 183: 273–290. https://doi.org/10.1111/j.1469-
695 8137.2009.02884.x

696 Günther T., and K. J. Schmid, 2010 Deleterious amino acid polymorphisms in Arabidopsis thaliana
697 and rice. Theor. Appl. Genet. 121: 157–168. https://doi.org/10.1007/s00122-010-1299-4

698 Habier D., R. L. Fernando, K. Kizilkaya, and D. J. Garrick, 2011 Extension of the bayesian
699 alphabet for genomic selection. BMC Bioinformatics 12: 186. https://doi.org/10.1186/1471-
700 2105-12-186

701 Hamblin M. T., A. M. Casa, H. Sun, S. C. Murray, A. H. Paterson, et al., 2006 Challenges of
702 Detecting Directional Selection After a Bottleneck: Lessons From Sorghum bicolor.
703 Genetics 173: 953–964. https://doi.org/10.1534/genetics.105.054312

704 Harlan J. R., J. M. J. De Wet, A. B. L. Stemler, I. ebrary, I. C. of Anthropological, et al., 1976
705 Origins of African plant domestication. The Hague : Mouton ; Chicago : distributed by
706 Aldine.

707 Hartfield M., and S. Glémin, 2014 Hitchhiking of Deleterious Alleles and the Cost of Adaptation in
708 Partially Selfing Species. Genetics 196: 281–293.
709 https://doi.org/10.1534/genetics.113.158196

710 Henn B. M., L. R. Botigué, S. Peischl, I. Dupanloup, M. Lipatov, et al., 2016 Distance from sub-
711 Saharan Africa predicts mutational load in diverse human genomes. Proc. Natl. Acad. Sci.
712 113: E440–E449. https://doi.org/10.1073/pnas.1510805112

713 Houle D., B. Morikawa, and M. Lynch, 1996 Comparing Mutational Variabilities. Genetics 143:
714 1467–1483.

715 Huang Y.-F., B. Gulko, and A. Siepel, 2017 Fast, scalable prediction of deleterious noncoding
716 variants from functional and population genomic data. Nat. Genet. advance online
717 publication. https://doi.org/10.1038/ng.3810

718 Ji X., R. L. Kember, C. D. Brown, and M. Bućan, 2016 Increased burden of deleterious variants in
719 essential genes in autism spectrum disorder. Proc. Natl. Acad. Sci. 113: 15054–15059.
720 https://doi.org/10.1073/pnas.1613195113

721 Jun G., A. Manning, M. Almeida, M. Zawistowski, A. R. Wood, et al., 2018 Evaluating the
722 contribution of rare variants to type 2 diabetes and related traits using pedigrees. Proc. Natl.
723 Acad. Sci. 115: 379–384. https://doi.org/10.1073/pnas.1705859115

Kelly J. K., 1999 An experimental method for evaluating the contribution of deleterious mutations to quantitative trait variation. Genet. Res. 73: 263–273.

Kono T. J. Y., F. Fu, M. Mohammadi, P. J. Hoffman, C. Liu, et al., 2016 The Role of Deleterious Substitutions in Crop Genomes. Mol. Biol. Evol. 33: 2307–2317. https://doi.org/10.1093/molbev/msw102

Kremling K. A. G., S.-Y. Chen, M.-H. Su, N. K. Lepak, M. C. Romay, et al., 2018 Dysregulation of expression correlates with rare-allele burden and fitness loss in maize. Nature. https://doi.org/10.1038/nature25966

Kumaravadivel N., and S. R. S. Rangasamy, 1994 Plant regeneration from sorghum anther cultures and field evaluation of progeny. Plant Cell Rep. 13: 286–290. https://doi.org/10.1007/BF00233321

Li J., M.-G. C. Danao, S.-F. Chen, S. Li, V. Singh, et al., 2015 Prediction of Starch Content and Ethanol Yields of Sorghum Grain Using near Infrared Spectroscopy. J. Infrared Spectrosc. 23: 85–92.

Lohmueller K. E., A. R. Indap, S. Schmidt, A. R. Boyko, R. D. Hernandez, et al., 2008 Proportionally more deleterious genetic variation in European than in African populations. Nature 451: 994–997. https://doi.org/10.1038/nature06611

Lu J., T. Tang, H. Tang, J. Huang, S. Shi, et al., 2006 The accumulation of deleterious mutations in rice genomes: a hypothesis on the cost of domestication. Trends Genet. TIG 22: 126–131. https://doi.org/10.1016/j.tig.2006.01.004

Lynch M., and W. Gabriel, 1990 Mutation load and the survival of small populations. Evol. Int. J. Org. Evol. 44: 1725–1737. https://doi.org/10.1111/j.1558-5646.1990.tb05244.x

Mace E. S., S. Tai, E. K. Gilding, Y. Li, P. J. Prentis, et al., 2013 Whole-genome sequencing reveals untapped genetic potential in Africa's indigenous cereal crop sorghum. Nat. Commun. 4: 2320. https://doi.org/10.1038/ncomms3320

Marouli E., M. Graff, C. Medina-Gomez, K. S. Lo, A. R. Wood, et al., 2017 Rare and low-frequency coding variants alter human adult height. Nature 542: 186–190. https://doi.org/10.1038/nature21039

Meziane D., and B. Shipley, 1999 Interacting determinants of specific leaf area in 22 herbaceous species: effects of irradiance and nutrient availability. Plant Cell Environ. 22: 447–459. https://doi.org/10.1046/j.1365-3040.1999.00423.x

Mezmouk S., and J. Ross-Ibarra, 2014 The Pattern and Distribution of Deleterious Mutations in Maize. G3 Genes Genomes Genet. 4: 163–171. https://doi.org/10.1534/g3.113.008870

Morrell P. L., E. S. Buckler, and J. Ross-Ibarra, 2012 Crop genomics: advances and applications. Nat. Rev. Genet. 13: 85–96. https://doi.org/10.1038/nrg3097

759 Moyers B. T., P. L. Morrell, and J. K. McKay, 2018 Genetic Costs of Domestication and
760 Improvement. J. Hered. 109: 103–116. https://doi.org/10.1093/jhered/esx069

761 Nakayama S.-I., S. Shi, M. Tateno, M. Shimada, and K. R. Takahasi, 2012 Mutation Accumulation
762 in a Selfing Population: Consequences of Different Mutation Rates between Selfers and
763 Outcrossers. PLOS ONE 7: e33541. https://doi.org/10.1371/journal.pone.0033541

764 Pamilo P., M. Nei, and W.-H. Li, 1987 Accumulation of mutations in sexual and asexual
765 populations. Genet. Res. 49: 135–146. https://doi.org/10.1017/S0016672300026938

766 Park J.-H., M. H. Gail, C. R. Weinberg, R. J. Carroll, C. C. Chung, et al., 2011 Distribution of allele
767 frequencies and effect sizes and their interrelationships for common genetic susceptibility
768 variants. Proc. Natl. Acad. Sci. 108: 18026–18031.
769 https://doi.org/10.1073/pnas.1114759108

770 Paterson A. H., J. E. Bowers, and B. A. Chapman, 2004 Ancient polyploidization predating
771 divergence of the cereals, and its consequences for comparative genomics. Proc. Natl.
772 Acad. Sci. U. S. A. 101: 9903–9908. https://doi.org/10.1073/pnas.0307901101

773 Paterson A. H., J. E. Bowers, R. Bruggmann, I. Dubchak, J. Grimwood, et al., 2009 The Sorghum
774 bicolor genome and the diversification of grasses. Nature 457: 551.
775 https://doi.org/10.1038/nature07723

776 Purcell S. M., J. L. Moran, M. Fromer, D. Ruderfer, N. Solovieff, et al., 2014 A polygenic burden of
777 rare disruptive mutations in schizophrenia. Nature 506: 185–190.
778 https://doi.org/10.1038/nature12975

779 Ramu P., W. Esuma, R. Kawuki, I. Y. Rabbi, C. Egesi, et al., 2017 Cassava haplotype map
780 highlights fixation of deleterious mutations during clonal propagation. Nat. Genet. 49: 959–
781 963. https://doi.org/10.1038/ng.3845

782 Ranwez V., A. Serra, D. Pot, and N. Chantret, 2017 Domestication reduces alternative splicing
783 expression variations in sorghum. PLOS ONE 12: e0183454.
784 https://doi.org/10.1371/journal.pone.0183454

785 Renaut S., and L. H. Rieseberg, 2015 The Accumulation of Deleterious Mutations as a
786 Consequence of Domestication and Improvement in Sunflowers and Other Compositae
787 Crops. Mol. Biol. Evol. 32: 2273–2283. https://doi.org/10.1093/molbev/msv106

788 Shaw F. H., C. J. Geyer, and R. G. Shaw, 2002 A Comprehensive Model of Mutations Affecting
789 Fitness and Inferences for Arabidopsis Thaliana. Evolution 56: 453–463.
790 https://doi.org/10.1111/j.0014-3820.2002.tb01358.x

791 Simons Y. B., M. C. Turchin, J. K. Pritchard, and G. Sella, 2014 The deleterious mutation load is
792 insensitive to recent population history. Nat. Genet. 46: 220–224.
793 https://doi.org/10.1038/ng.2896

Slotte T., J. P. Foxe, K. M. Hazzouri, and S. I. Wright, 2010 Genome-wide evidence for efficient positive and purifying selection in Capsella grandiflora, a plant species with a large effective population size. Mol. Biol. Evol. 27: 1813–1821. https://doi.org/10.1093/molbev/msq062

Slotte T., K. M. Hazzouri, J. A. Ågren, D. Koenig, F. Maumus, et al., 2013 The Capsella rubella genome and the genomic consequences of rapid mating system evolution. Nat. Genet. 45: 831. https://doi.org/10.1038/ng.2669

Smith O., W. V. Nicholson, L. Kistler, E. Mace, A. Clapham, et al., 2018 A domestication history of dynamic adaptation and genomic deterioration in sorghum. bioRxiv 336503. https://doi.org/10.1101/336503

Sprague G. F., W. A. Russell, and L. H. Penny, 1960 Mutations Affecting Quantitative Traits in the Selfed Progeny of Doubled Monoploid Maize Stocks. Genetics 45: 855–866.

Stemler A. B. L., J. R. Harlan, and J. M. J. de Wet, 1975 Evolutionary History of Cultivated Sorghums (Sorghum bicolor [Linn.] Moench) of Ethiopia. Bull. Torrey Bot. Club 102: 325–333. https://doi.org/10.2307/2484758

Sulpice R., E.-T. Pyl, H. Ishihara, S. Trenkamp, M. Steinfath, et al., 2009 Starch as a major integrator in the regulation of plant growth. Proc. Natl. Acad. Sci. 106: 10348–10353. https://doi.org/10.1073/pnas.0903478106

Szövényi P., N. Devos, D. J. Weston, X. Yang, Z. Hock, et al., 2014 Efficient Purging of Deleterious Mutations in Plants with Haploid Selfing. Genome Biol. Evol. 6: 1238–1252. https://doi.org/10.1093/gbe/evu099

Thalmann M., and D. Santelia, 2017 Starch as a determinant of plant fitness under abiotic stress. New Phytol. 214: 943–951. https://doi.org/10.1111/nph.14491

Thurber C. S., J. M.-Y. Ma, R. H. Higgins, and P. J. Brown, 2013 Retrospective genomic analysis of sorghum adaptation to temperate-zone grain production. undefined.

Vaser R., S. Adusumalli, S. N. Leng, M. Sikic, and P. C. Ng, 2016 SIFT missense predictions for genomes. Nat. Protoc. 11: 1. https://doi.org/10.1038/nprot.2015.123

Vikram P., B. P. M. Swamy, S. Dixit, R. Singh, B. P. Singh, et al., 2015 Drought susceptibility of modern rice varieties: an effect of linkage of drought tolerance with undesirable traits. Sci. Rep. 5: 14799. https://doi.org/10.1038/srep14799

Vitezica Z. G., L. Varona, and A. Legarra, 2013 On the Additive and Dominant Variance and Covariance of Individuals Within the Genomic Selection Scope. Genetics 195: 1223–1230. https://doi.org/10.1534/genetics.113.155176

Vitezica Z. G., L. Varona, J.-M. Elsen, I. Misztal, W. Herring, et al., 2016 Genomic BLUP including additive and dominant variation in purebreds and F1 crossbreds, with an application in pigs. Genet. Sel. Evol. 48: 6. https://doi.org/10.1186/s12711-016-0185-1

Wendorf F., A. E. Close, R. Schild, K. Wasylikowa, R. A. Housley, et al., 1992 Saharan exploitation of plants 8,000 years BP. Nature 359: 721. https://doi.org/10.1038/359721a0

Westoby M., 1998 A leaf-height-seed (LHS) plant ecology strategy scheme. Plant Soil 199: 213–227. https://doi.org/10.1023/A:1004327224729

Wright S., 1931 Evolution in Mendelian Populations. Genetics 16: 97–159.

Yampolsky L. Y., F. A. Kondrashov, and A. S. Kondrashov, 2005 Distribution of the strength of selection against amino acid replacements in human proteins. Hum. Mol. Genet. 14: 3191–3201. https://doi.org/10.1093/hmg/ddi350

Yang J., S. H. Lee, M. E. Goddard, and P. M. Visscher, 2011 GCTA: a tool for genome-wide complex trait analysis. Am. J. Hum. Genet. 88: 76–82. https://doi.org/10.1016/j.ajhg.2010.11.011

Yang J., S. Mezmouk, A. Baumgarten, E. S. Buckler, K. E. Guill, et al., 2017 Incomplete dominance of deleterious alleles contributes substantially to trait variation and heterosis in maize. PLOS Genet. 13: e1007019. https://doi.org/10.1371/journal.pgen.1007019

Younginger B. S., D. Sirová, M. B. Cruzan, and D. J. Ballhorn, 2017 Is biomass a reliable estimate of plant fitness?1. Appl. Plant Sci. 5. https://doi.org/10.3732/apps.1600094

Yu X., X. Li, T. Guo, C. Zhu, Y. Wu, et al. 2016. Genomic prediction contributing to a promising global strategy to tubrcharge gene banks. Nat. Plants. 2: 1-7.

Zöllner S., and J. K. Pritchard, 2007 Overcoming the Winner's Curse: Estimating Penetrance Parameters from Case-Control Data. Am. J. Hum. Genet. 80: 605–615.

Legends to Figures

Figure 1 Deleterious mutations in the sorghum genome. (a) Site allele-frequency spectrum of nonsynonymous deleterious mutations and synonymous mutations in the sorghum genome. The Derived Allele Frequency (DAF) distribution of alleles is shown where a minor allele in the multi-species alignment was considered as a derived deleterious allele (Yang et al. 2017). (b) The allele frequency of the derived alleles in bins of different GERP scores. The vertical bars in (b) indicate standard error.

Figure 2 Smoothed estimate of density distribution of regression coefficients associated with high conserved deleterious variants ($HGERP_{DEL\text{-}SNPs}$) and nondeleterious variants for ten phenotypic traits ((a) biomass, (b) specific leaf area (SLA), (c) tissue starch content (TSC), (d-j) plant height 4, 6, 8, 10, 12, 14, and 16 weeks after planting, respectively).

Figure 3 Barplots of means of folded distributions of effect sizes of high conserved deleterious variants ($HGERP_{DEL\text{-}SNPs}$) and nondeleterious variants for ten phenotypic traits ((a) biomass, (b) specific leaf area (SLA), (c) tissue starch content (TSC), (d-j) plant height 4, 6, 8, 10, 12, 14, and 16 weeks after planting, respectively).

Figure 4 Homozygous mutation burden in sorghum. (a) Homozygous mutation burden estimated for different racial groups of sorghum based on high conserved deleterious variants ($HGERP_{DEL\text{-}SNPs}$). The derived allele is defined as a minor allele from multi-species sequence alignments (Yang et al. 2017). The mutation burden was estimated as the count of derived deleterious alleles carried by an individual divided by the total number of scored (non-missing) alleles. The horizontal broken line indicates the mean of homozygous mutation burden across all racial groups. (b) scatter plots of homozygous mutation burden and principal coordinate 1 derived from genome-wide SNP markers. The black circled point indicates the median values for each group.

Figure 5 Heritability estimates for all traits using a two-kernel model. Abbreviations: SLA, specific leaf area; TSC, tissue starch content; PH4 until PH16, plant height at 4, 6, 8, 10, 12, 14, and 16 weeks after planting (WAP), respectively.

Figure 6 Genome-wide prediction models incorporating putatively deleterious variants. (a-b) Heritability estimates for all traits using a single-kernel model. Heritability estimates for nondeleterious variants are derived based on 100 independent sets that are randomly chosen across the genome from variants that are not in LD with deleterious variants. (c-d) Boxplots showing a five-fold cross validation prediction ability estimation for deleterious variants and random variants.

Deleterious Mutation Burden and its Association with Complex Traits in Sorghum

Ravi Valluru, Elodie E. Gazave, Samuel B. Fernandes, John Ferguson, Roberto Lozano, **Pradeep**, Tao Zuo, Patrick J. Brown, Andrew D.B. Leakey, Michael A. Gore, Edward S. Buckler, Nonoy Bandillo

Fig 1

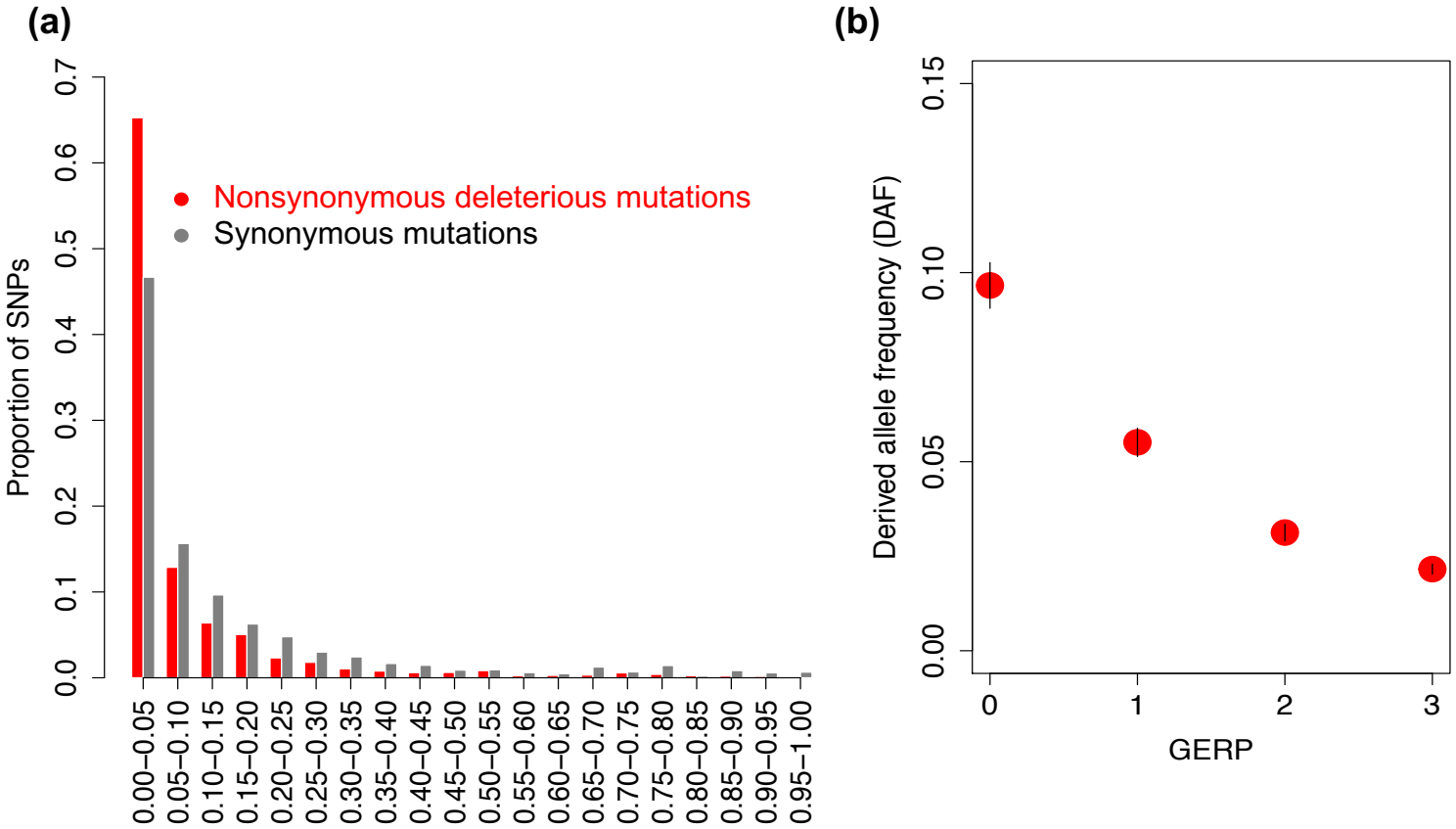

Figure 1 Deleterious mutations in the sorghum genome. (a) Site allele-frequency spectrum of nonsynonymous deleterious mutations and synonymous mutations in the sorghum genome. The Derived Allele Frequency (DAF) distribution of alleles is shown where a minor allele in the multi-species alignment was considered as a derived deleterious allele (Yang et al. 2017). (b) The allele frequency of the derived alleles in bins of different GERP scores. The vertical bars in (b) indicate standard error.

Fig 2

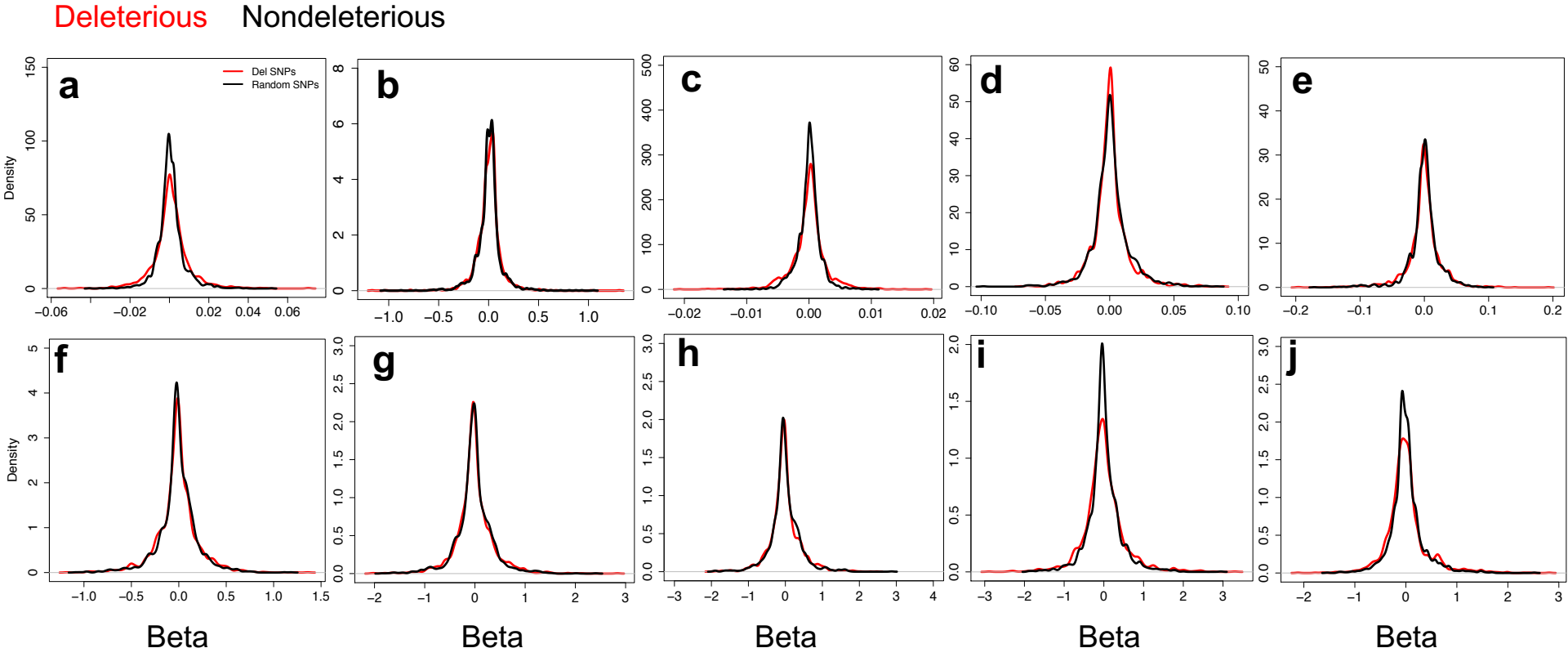

Figure 2 Smoothed estimate of density distribution of regression coefficients associated with high conserved deleterious variants ($HGERP_{DEL\text{-}SNPs}$) and nondeleterious variants for ten phenotypic traits ((a) biomass, (b) specific leaf area (SLA), (c) tissue starch content (TSC), (d-j) plant height 4, 6, 8, 10, 12, 14, and 16 weeks after planting, respectively).

Fig 3

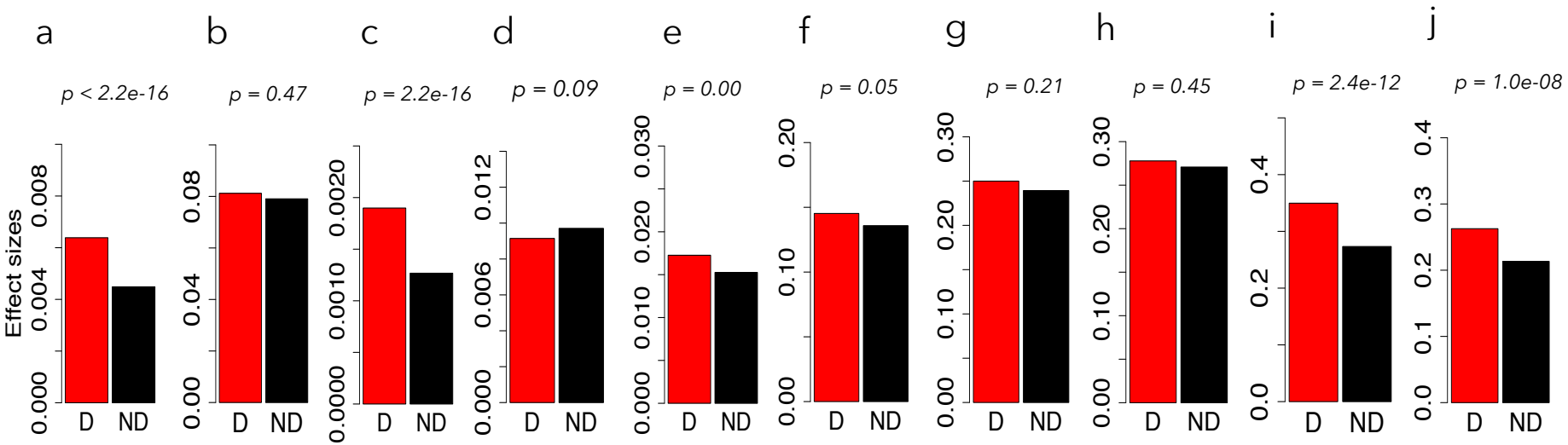

Figure 3 Barplots of means of folded distributions of effect sizes of high conserved deleterious variants ($HGERP_{DEL\text{-}SNPs}$) and nondeleterious variants for ten phenotypic traits ((a) biomass, (b) specific leaf area (SLA), (c) tissue starch content (TSC), (d-j) plant height 4, 6, 8, 10, 12, 14, and 16 weeks after planting, respectively). D: deleterious; ND: nondeleterious.

Fig 4

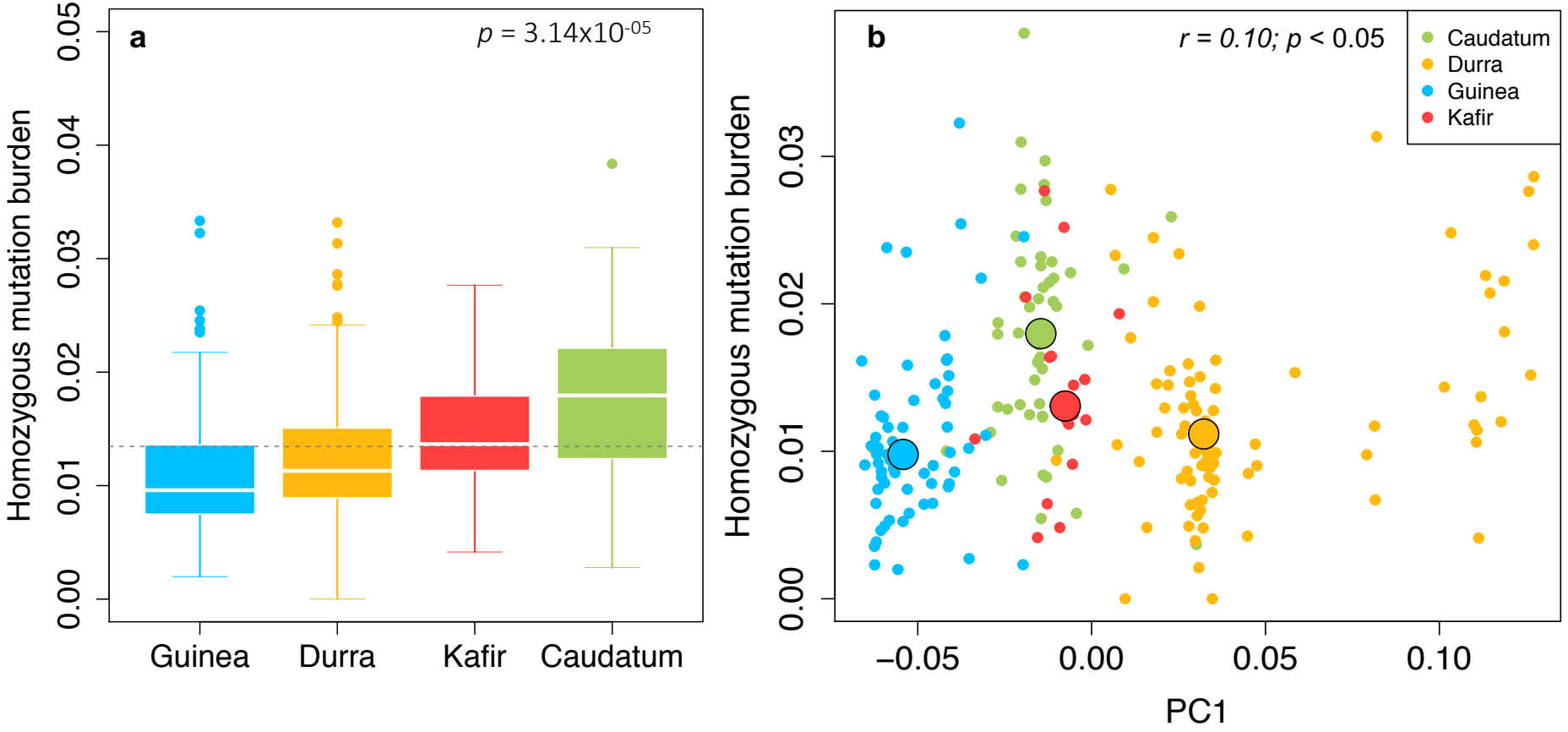

Figure 4 Homozygous mutation burden in sorghum. (a) Homozygous mutation burden estimated for different racial groups of sorghum based on high conserved deleterious variants ($HGERP_{DEL\text{-}SNPs}$). The derived allele is defined as a minor allele from multi-species sequence alignments (Yang et al. 2017). The mutation burden was estimated as the count of derived deleterious alleles carried by an individual divided by the total number of scored (non-missing) alleles. The horizontal broken line indicates the mean of homozygous mutation burden across all racial groups. (b) scatter plots of homozygous mutation burden and principal coordinate 1 derived from genome-wide SNP markers. The black circled point indicates the median values for each group.

Fig 5

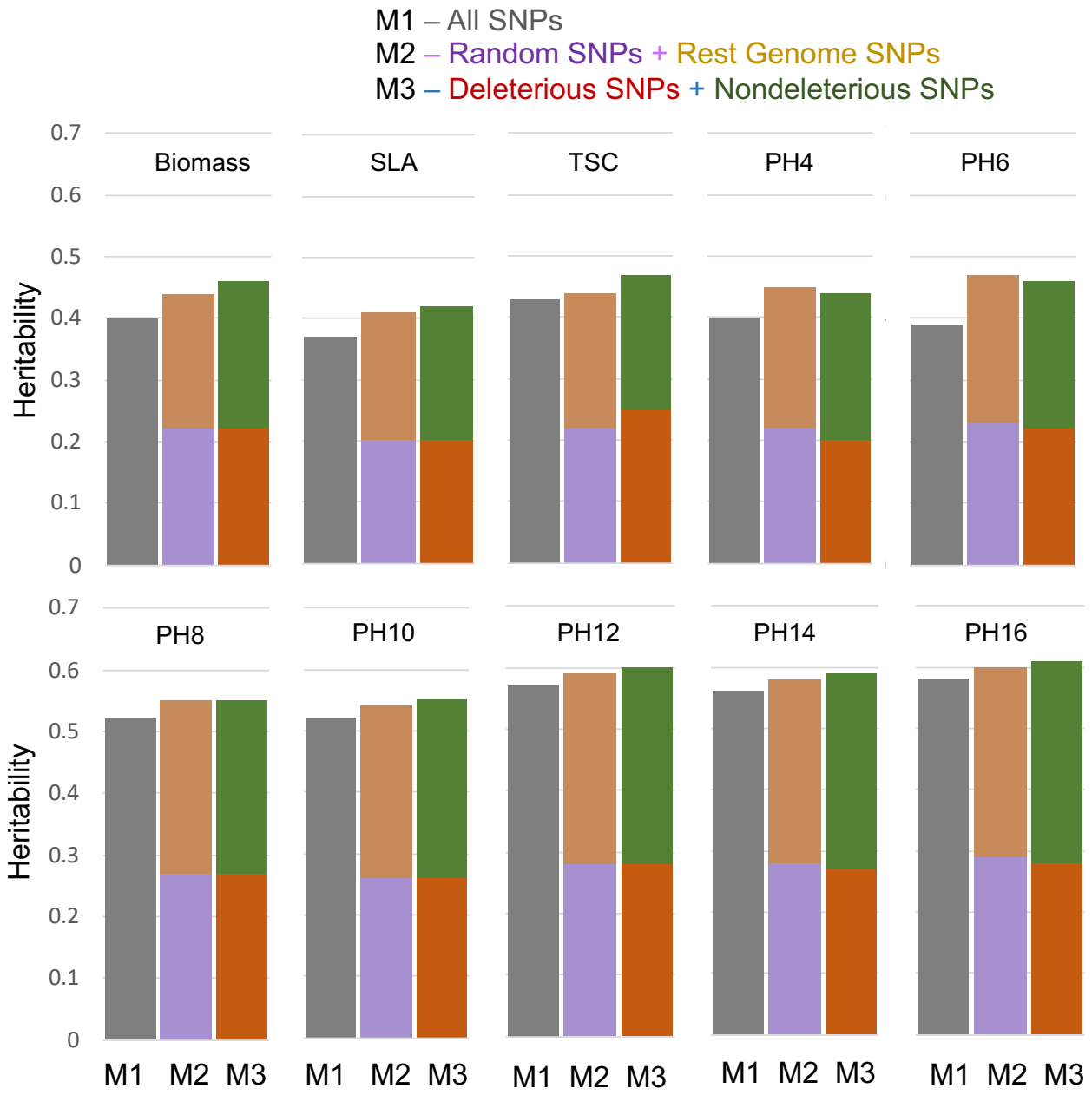

Figure 5 Heritability estimates for all traits using a two-kernel model. Abbreviations: SLA, specific leaf area; TSC, tissue starch content; PH4 until PH16, plant height at 4, 6, 8, 10, 12, 14, and 16 weeks after planting (WAP), respectively.

Fig 6

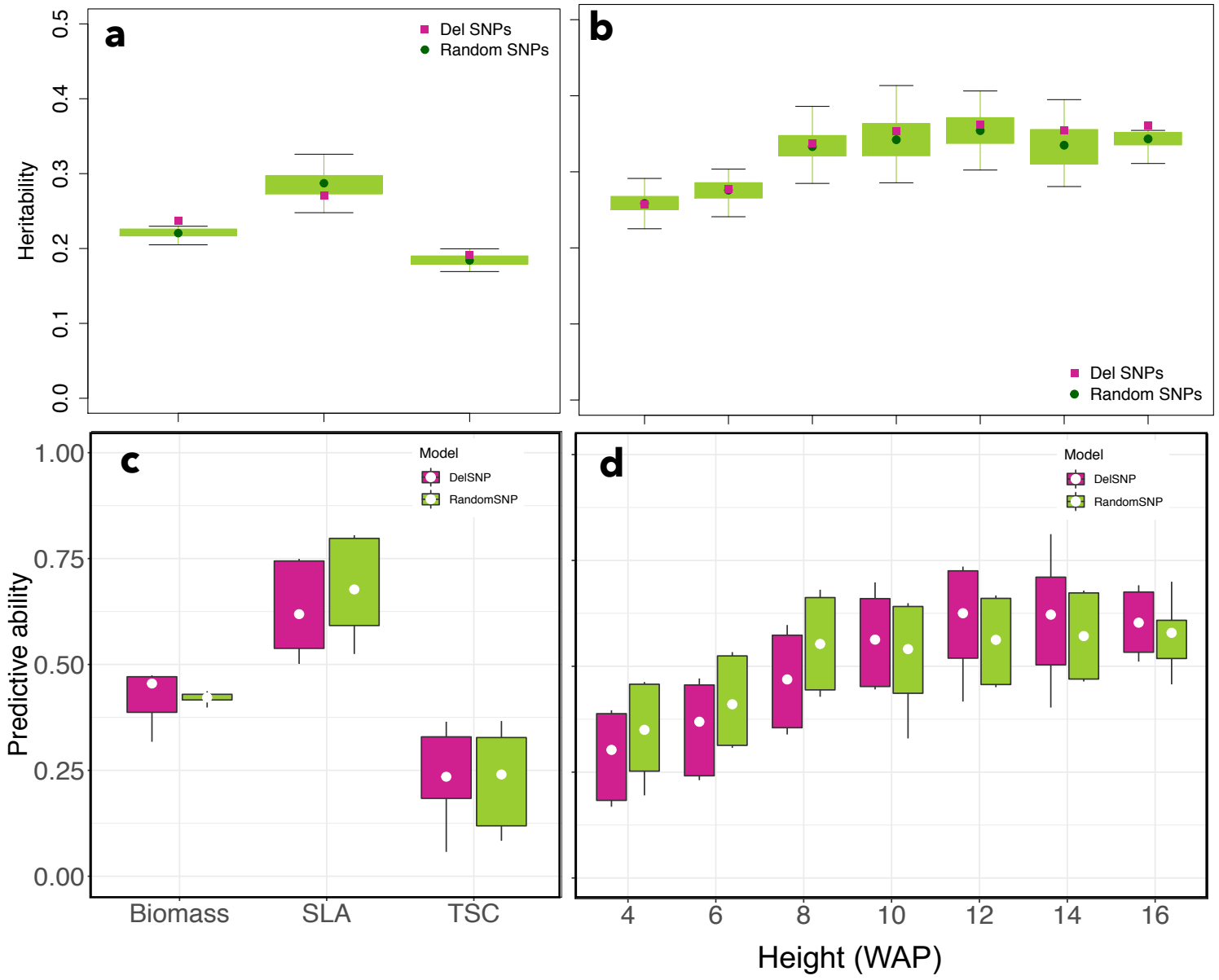

Figure 6 Genome-wide prediction models incorporating putatively deleterious variants. (a-b) Heritability estimates for all traits using a single-kernel model. Heritability estimates for nondeleterious variants are derived based on 100 independent sets that are randomly chosen across the genome from variants that are not in LD with deleterious variants. (c-d) Boxplots showing a five-fold cross validation prediction ability estimation for deleterious variants and random variants.

Supplementary Information

Fig. S1

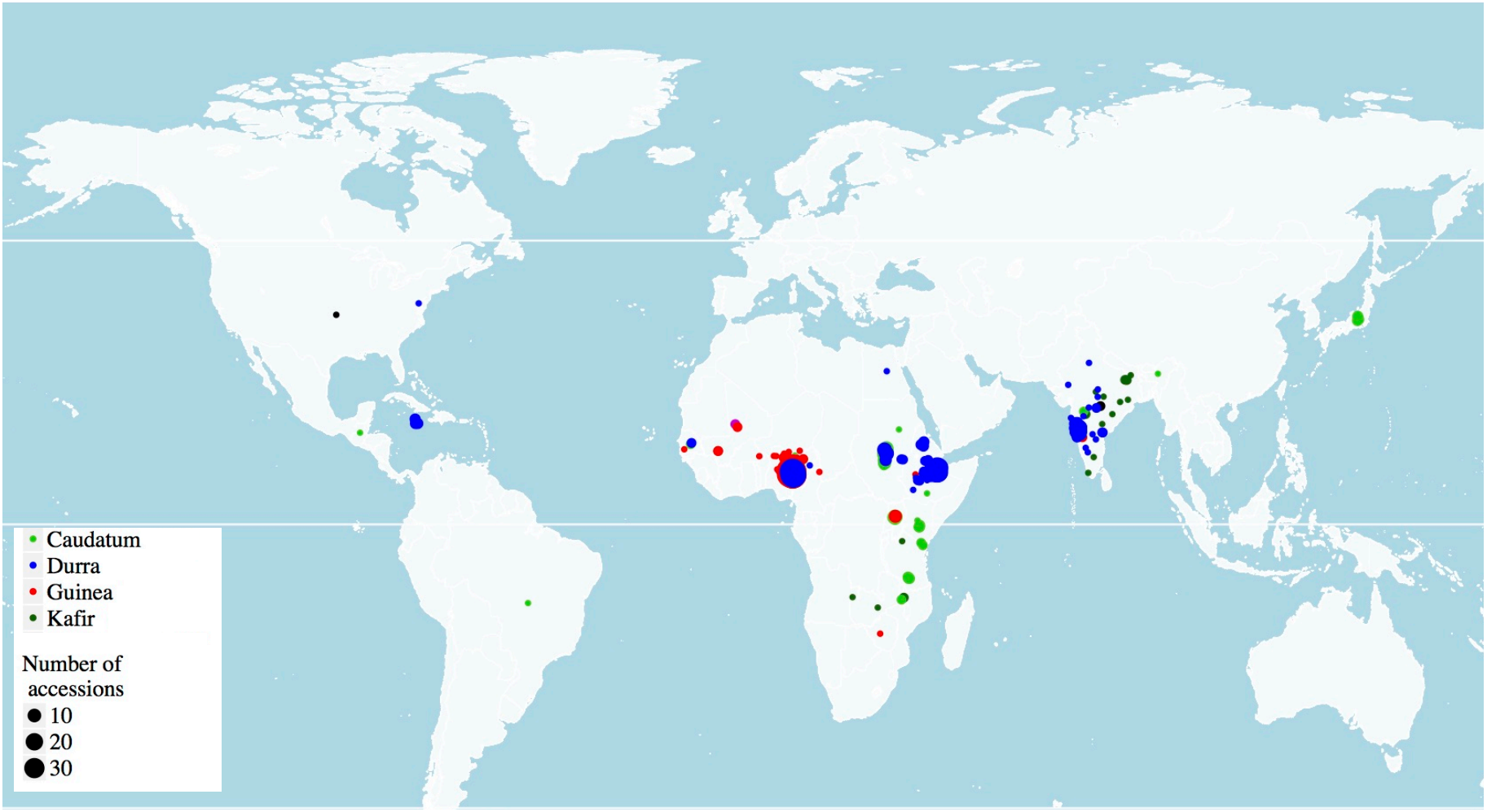

Fig S1 The geographical distribution of diverse sorghum lines (229) used in the study.

Fig. S2

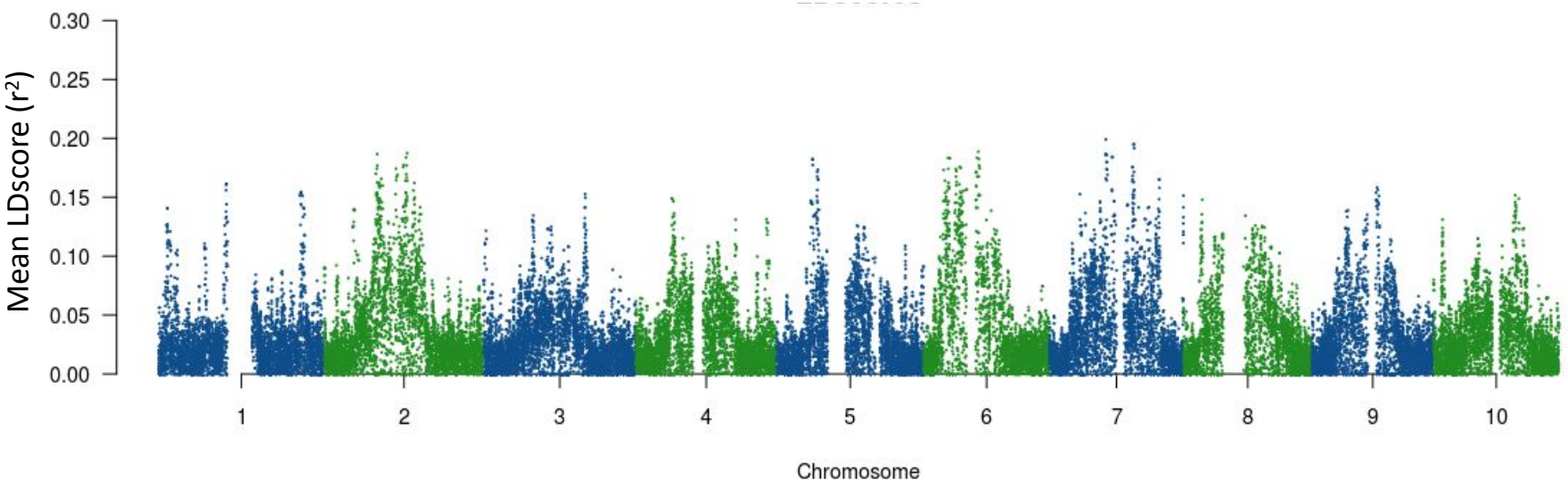

Fig S2 Mean linkage disequilibrium (LD) scores estimated for all chromosomes.

Fig. S3

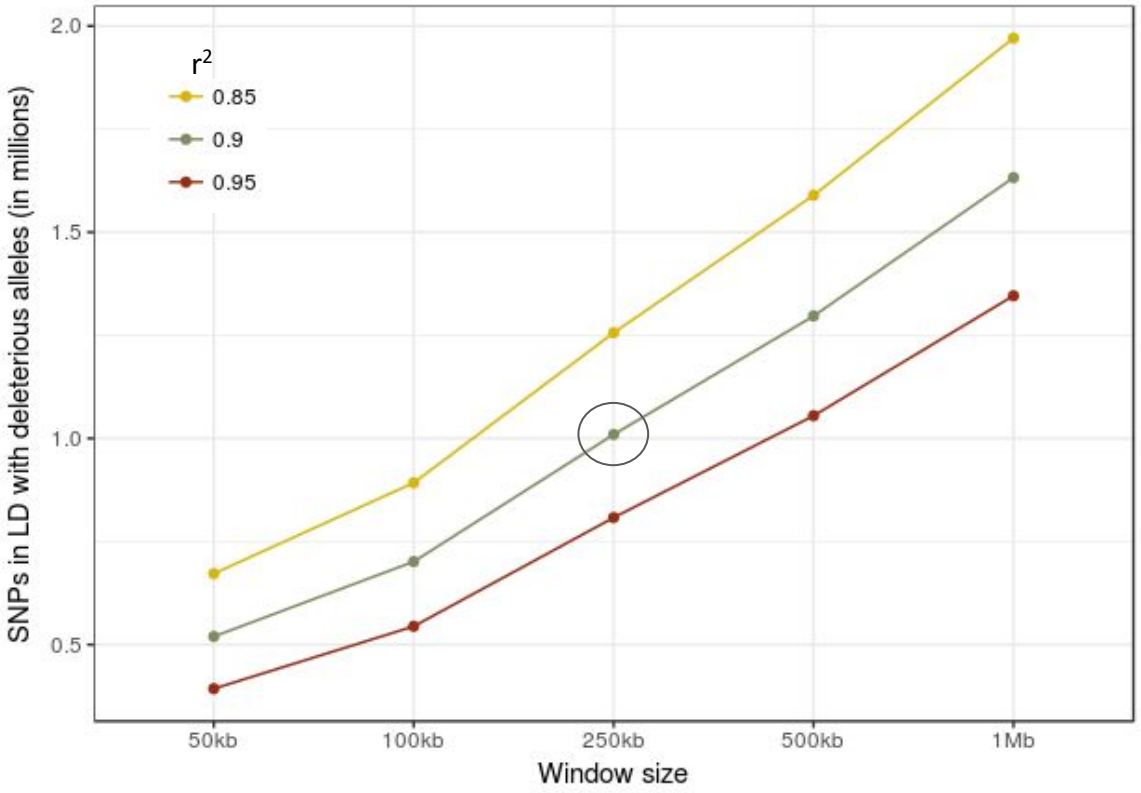

Fig S3 The number of single-nucleotide polymorphisms (SNPs) estimated under different parameters of window size and r^2.

Fig. S4

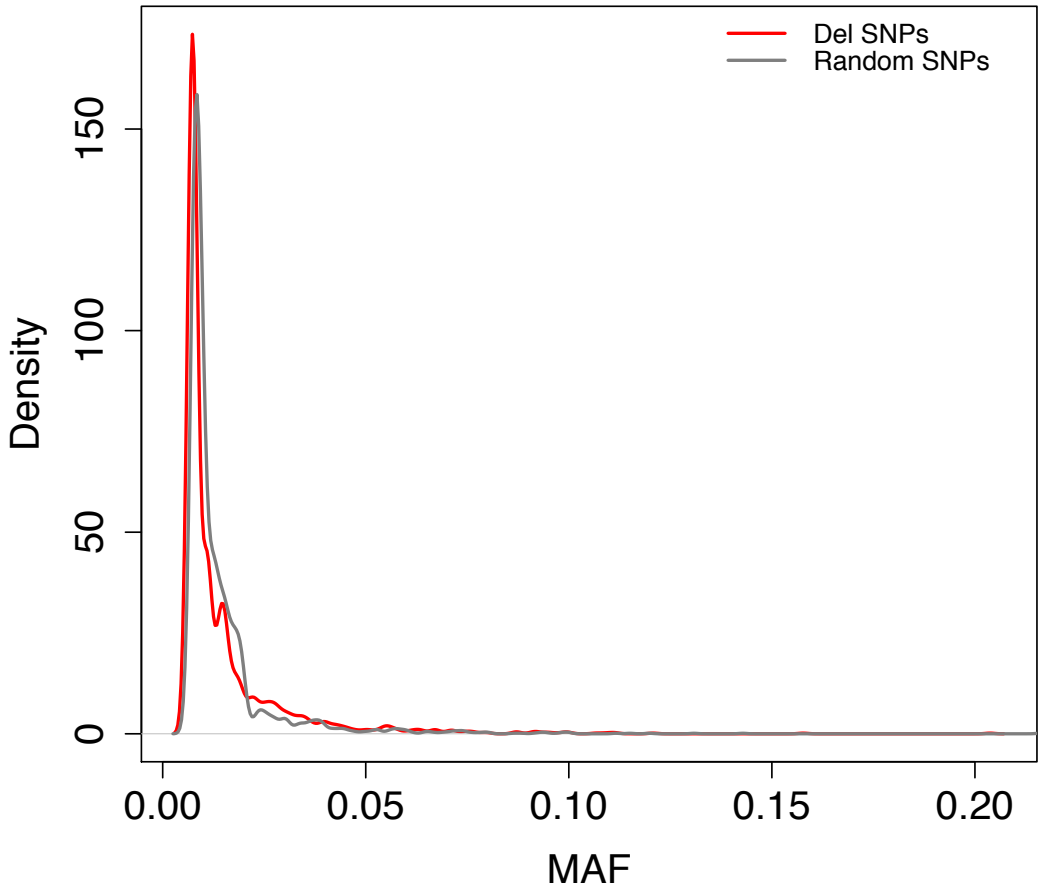

Fig S4 Allele frequency distribution comparison for high conserved deleterious variants ($HGERP_{DEL\text{-}SNPs}$, red) and nondeleterious variants (grey).

Fig. S5

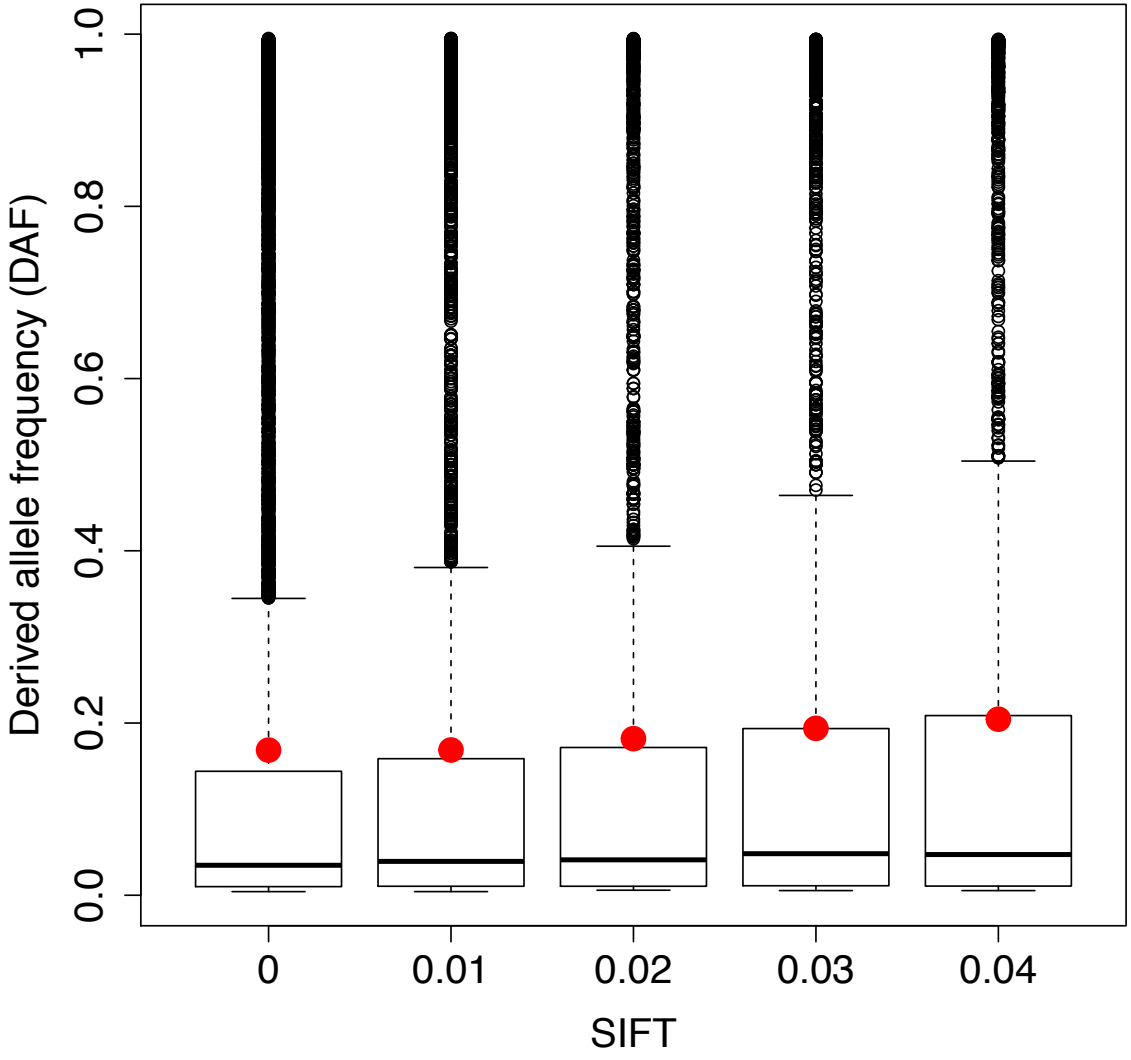

Fig S5 Derived allele frequency (DAF) association with Sorting Intolerant From Tolerant (SIFT) scores, Derived allele was defined as a minor allele from a multi-species sequence alignments. SIFT was estimated using Vaser et al. (2016).

Fig. S6

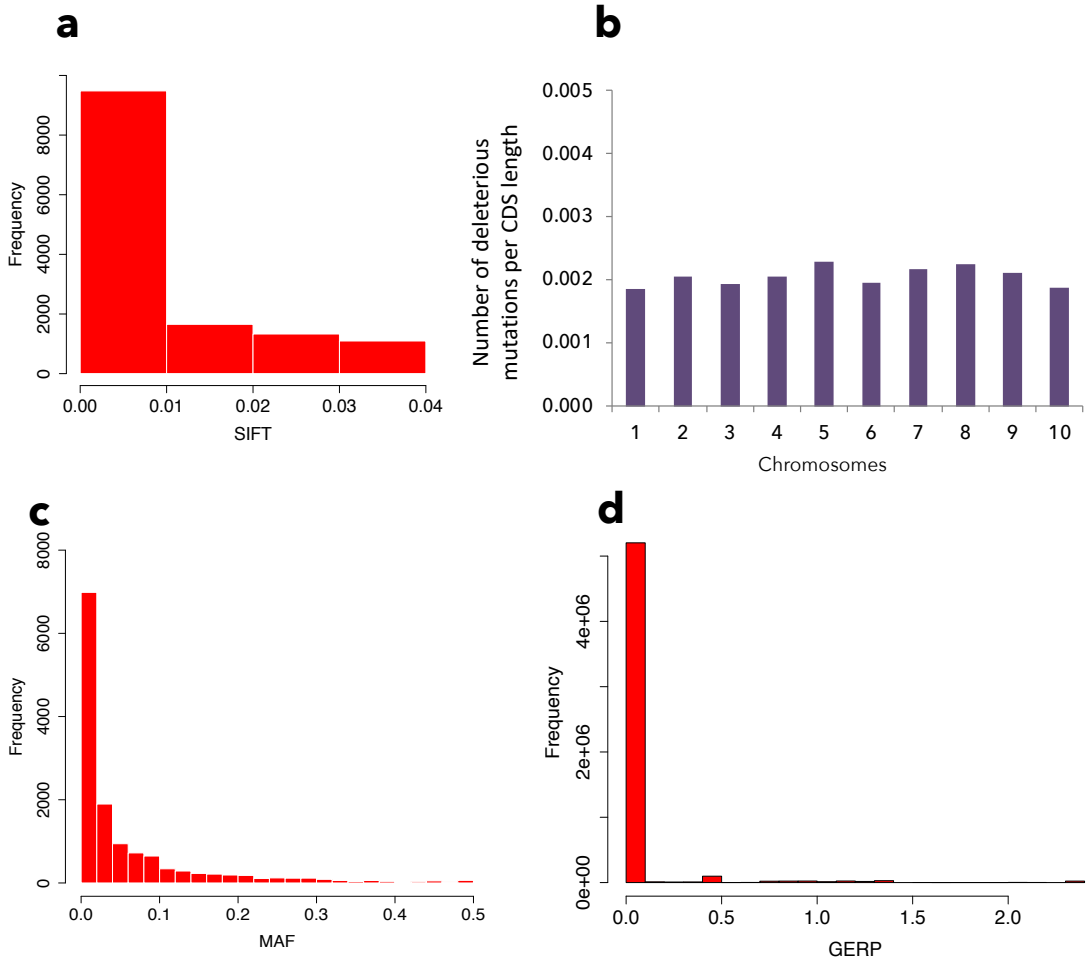

Fig S6 Frequency distributions of highly conserved deleterious mutations (SIFT<0.05 and GERP>2). Sorting Intolerant From Tolerant (SIFT) distributions of (a) all deleterious mutations, (b) number of deleterious mutations estimated per coding regions in all chromosomes, (c) minor allele frequency distribution of deleterious mutations, and (d) GERP distributions of deleterious mutations.

Fig. S7

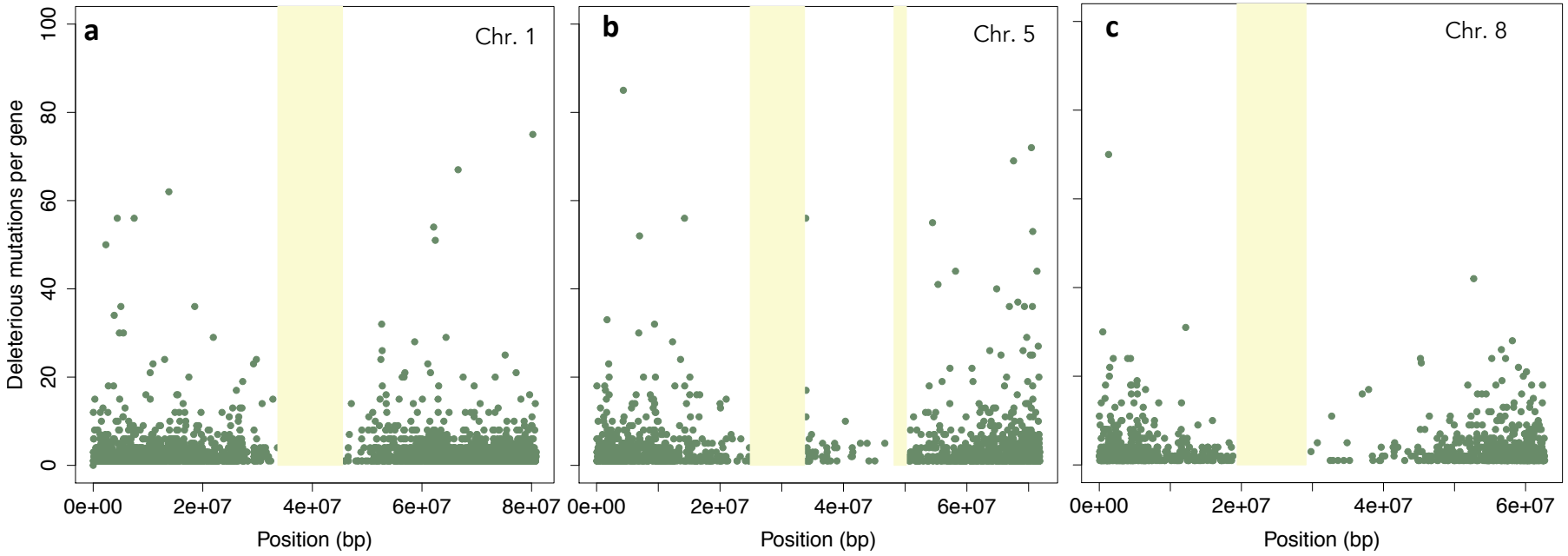

Fig S7 Gene deleterious mutations distribution in chromosomes 1 (a), 5 (b), and 8 (c). The yellow color vertical bar indicates a centromeric regions showing absence of genes or deleterious mutations.

Fig S8

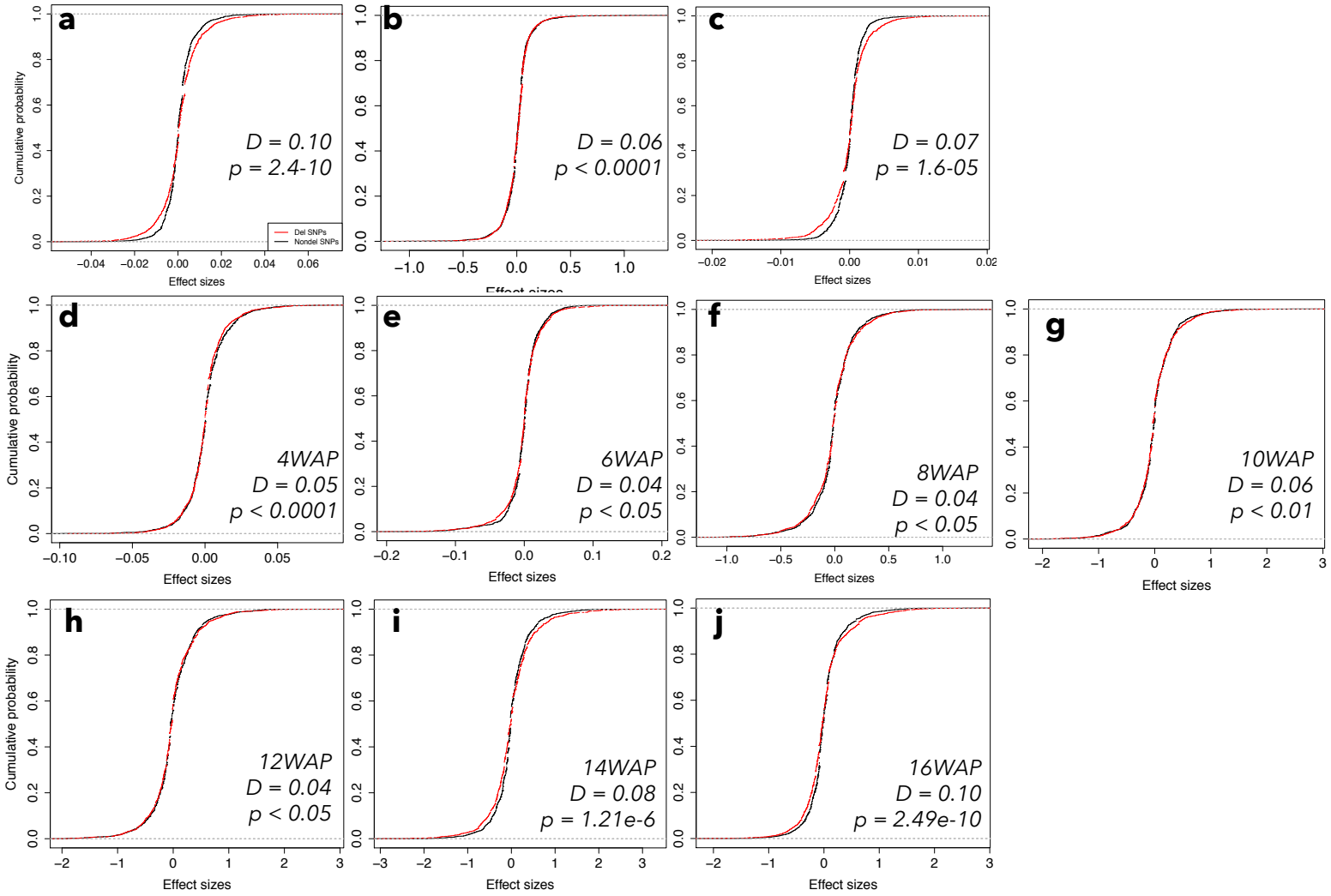

Fig S8 Empirical cumulative distributions (ECD) of the effect sizes for both high conserved deleterious variants ($HGERP_{DEL\text{-}SNPs}$) and nondeleterious variants for four phenotypic traits, biomass (a), specific leaf area (SLA, b), tissue starch content (TSC, c), and plant height 4 (d), 6 (e), 8 (f), 10 (g), 12 (h), 14 (i), and 16 (j) WAP, following a two-sample Kolmogorov-Smirnov (KS) test.

Fig S9

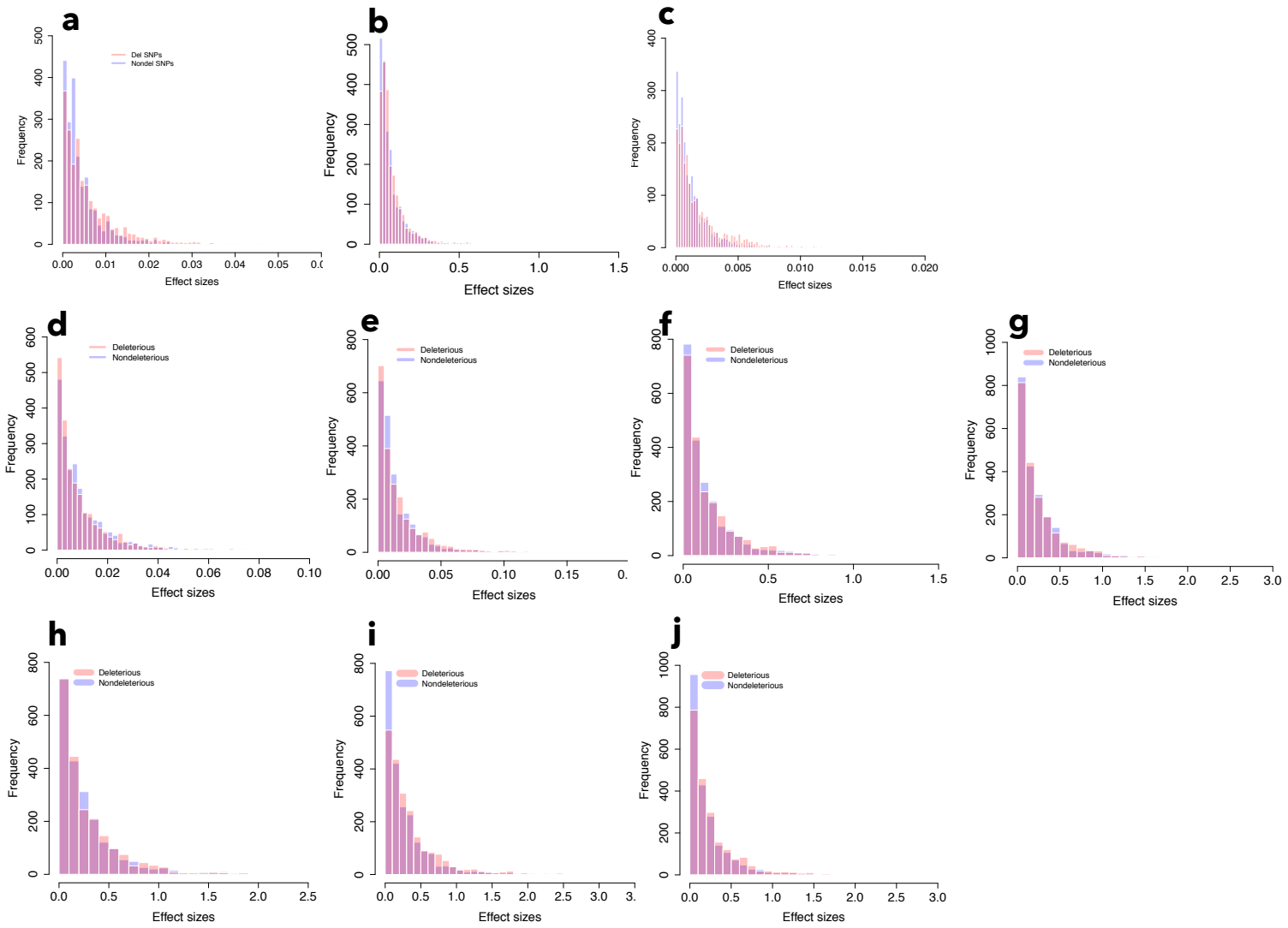

Fig S9 Folded distributions of the effect sizes for both high conserved deleterious variants ($HGERP_{DEL\text{-}SNPs}$) and nondeleterious variants for four phenotypic traits, biomass (a), specific leaf area (SLA, b), tissue starch content (TSC, c), and plant height 4 (d), 6 (e), 8 (f), 10 (g), 12 (h), 14 (i), and 16 (j) WAP.

Fig. S10

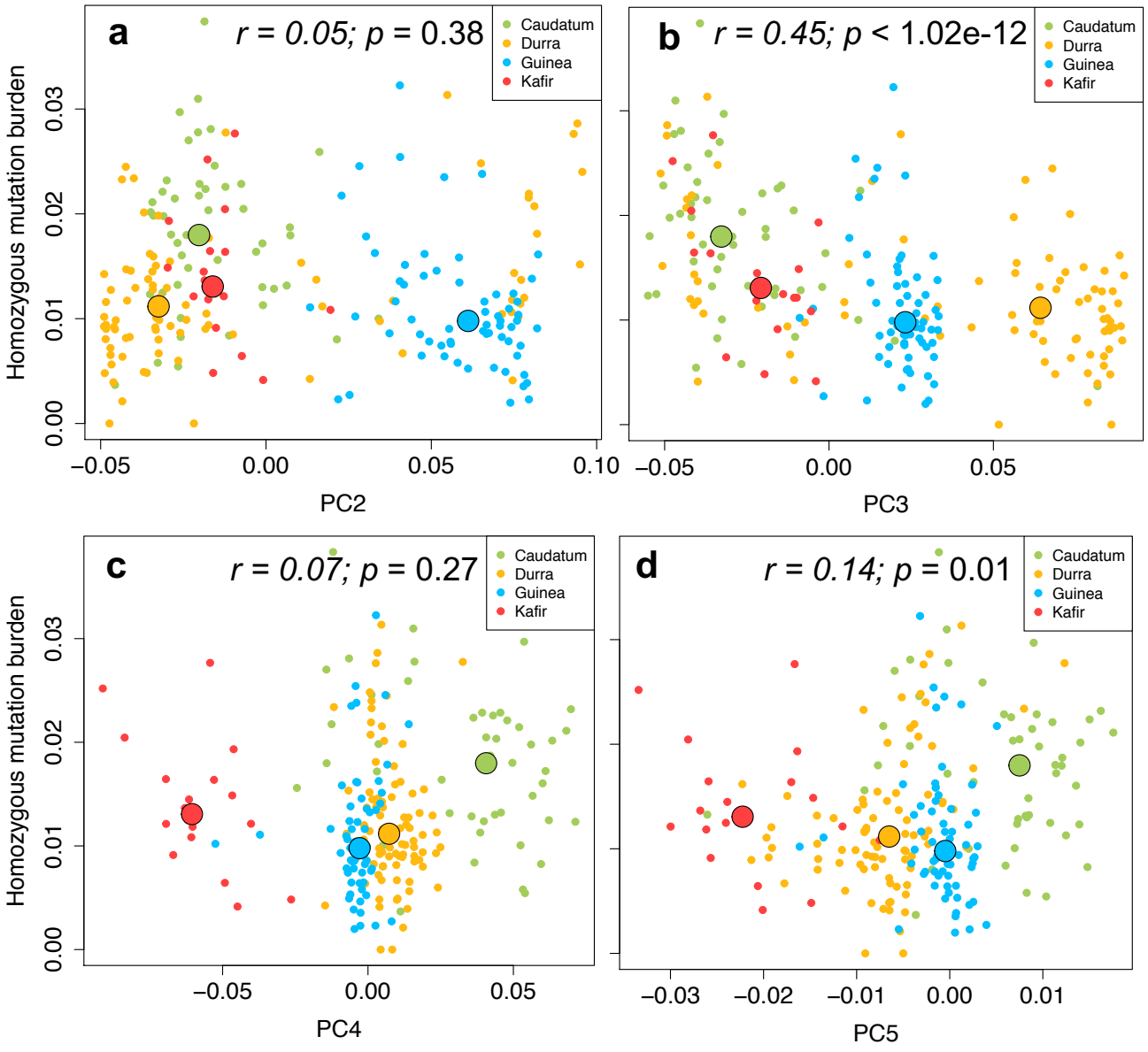

Fig S10 Scatterplots of the relationship between homozygous mutation burden and genetic principal components (PC2-5).

Fig S11

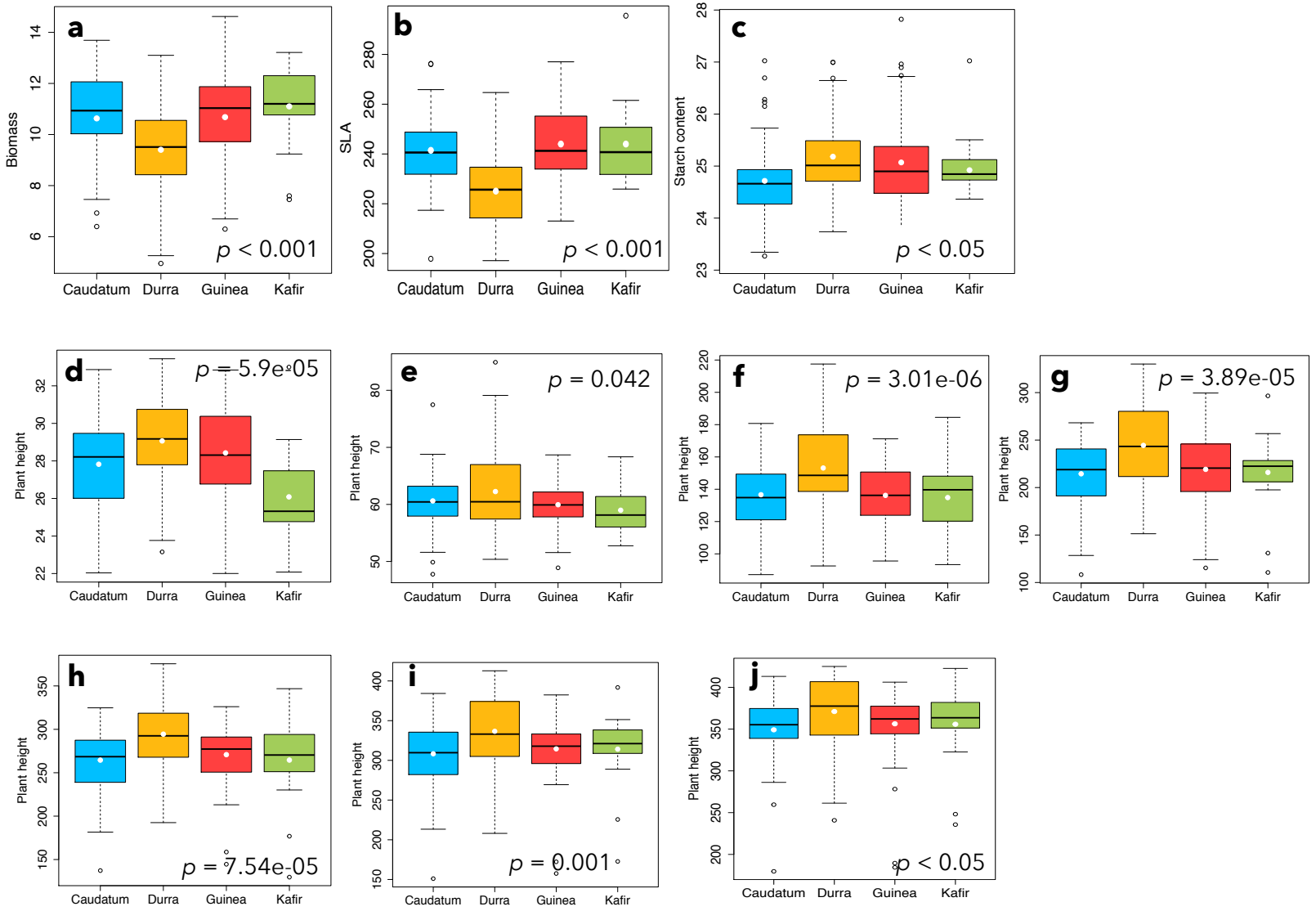

Fig S11 Boxplots of phenotypic data for biomass (a), specific leaf area (SLA, b), tissue starch content (TSC, c), and plant height 4 (d), 6 (e), 8 (f), 10 (g), 12 (h), 14 (i), and 16 (j) WAP under different racial groups of sorghum.

Fig S12

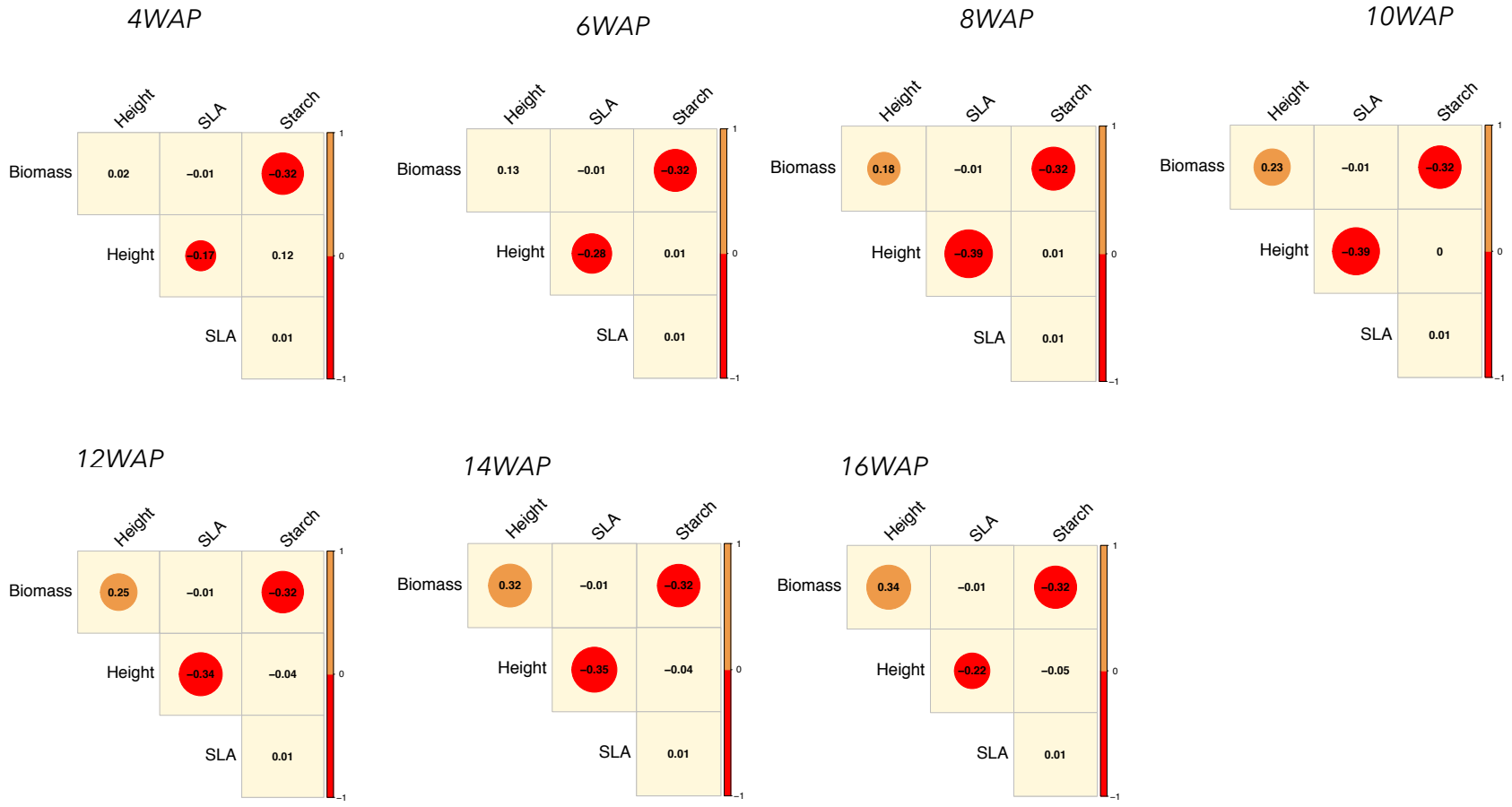

Fig S12 Correlations among traits either across all subpopulations using plant height 4, 6, 8, 10, 12, 14, and 16 WAP. All circles around values indicates significance at $p < 0.05$. Orange circles indicate positive correlation while red circle represents a negative correlation.

Fig S13

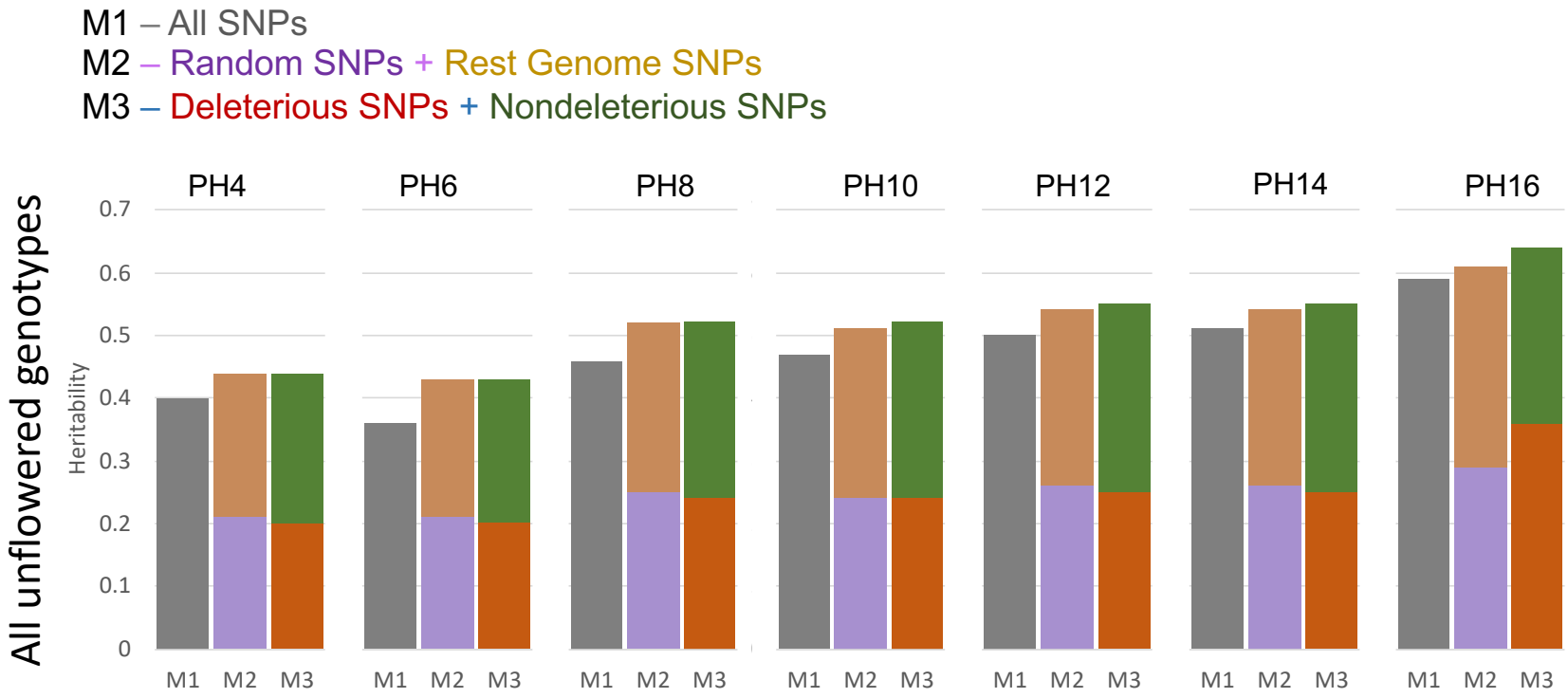

Fig S13 Heritability estimates for all traits using a two-kernel model in non-flowered lines.

Table. S1

Table S1 Slopes estimated between mutation burden and phenotypic traits using simple linear regression and grouped regression.

	Across all races			Grouped regression		
Phenotype	***Estimate***	***r^2***	***P***	***Estimate***	***r^2***	***P***
Biomass	-17.22	0.004	0.367	-48.8	0.04	0.151
PH (4WAP)	-32.64	0.008	0.199	-58.28	0.03	0.058
PH (6WAP)	-6.79	0	0.904	-131.65	0.03	0.195
PH (8WAP)	-15.67	0	0.945	-468.2	0.03	0.207
PH (10WAP)	-213.45	0.001	0.598	-1326.8	0.07	0.056
PH (12 WAP)	-329.48	0.003	0.422	-1301.5	0.06	0.084
PH (14 WAP)	-472.26	0.005	0.283	-1552.5	0.05	0.006**
PH (16 WAP)	-716.27	0.014	0.081	-1592.1	0.07	0.002**
SLA	-237.98	0.009	0.148	-166.09	0	0.520
Starch	-11.6	0.01	0.153	-46.92	0.16	0.102