the plant journal

SEB SOCIETY FOR EXPERIMENTAL BIOLOGY

The Plant Journal (2023) doi: 10.1111/tpj.16582

A large sequenced mutant library – valuable reverse genetic resource that covers 98% of sorghum genes

Yinping Jiao[1,*], Deepti Nigam[1], Kerrie Barry[2], Chris Daum[2], Yuko Yoshinaga[2], Anna Lipzen[2], Adil Khan[1], Sai-Praneeth Parasa[1], Sharon Wei[3], Zhenyuan Lu[3], Marcela K. Tello-Ruiz[3], Pallavi Dhiman[1], Gloria Burow[4], Chad Hayes[4], Junping Chen[4], Federica Brandizzi[5,6,7], Jenny Mortimer[8,9,10,*], Doreen Ware[3,11,*] and Zhanguo Xin[4,*]

[1]*Department of Plant and Soil Science, Institute of Genomics for Crop Abiotic Stress Tolerance, Texas Tech University, Lubbock, Texas 79409, USA,*

[2]*DOE Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA,*

[3]*Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA,*

[4]*Plant Stress and Germplasm Development Unit, Crop Systems Research Laboratory, USDA-ARS, 3810, 4th Street, Lubbock, Texas 79424, USA,*

[5]*MSU-DOE Plant Research Lab, Michigan State University, East Lansing, Michigan, USA,*

[6]*Great Lakes Bioenergy Research Center, Michigan State University, East Lansing, Michigan, USA,*

[7]*Department of Plant Biology, Michigan State University, East Lansing, Michigan, USA,*

[8]*Joint BioEnergy Institute, Emeryville, California 94608, USA,*

[9]*Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, California 94720, USA,*

[10]*School of Agriculture, Food and Wine, Waite Research Institute, Waite Research Precinct, University of Adelaide, Glen Osmond, South Australia 5064, Australia, and*

[11]*USDA-ARS NAA Robert W. Holley Center for Agriculture and Health, Agricultural Research Service, Ithaca, New York 14853, USA*

Received 16 May 2023; revised 8 September 2023; accepted 29 November 2023.

*For correspondence (e-mail yijiao@ttu.edu; jcmortimer@lbl.gov; ware@cshl.edu; zhanguo.xin@usda.gov).

SUMMARY

Mutant populations are crucial for functional genomics and discovering novel traits for crop breeding. *Sorghum*, a drought and heat-tolerant C4 species, requires a vast, large-scale, annotated, and sequenced mutant resource to enhance crop improvement through functional genomics research. Here, we report a sorghum large-scale sequenced mutant population with 9.5 million ethyl methane sulfonate (EMS)-induced mutations that covered 98% of sorghum's annotated genes using inbred line BTx623. Remarkably, a total of 610 320 mutations within the promoter and enhancer regions of 18 000 and 11 790 genes, respectively, can be leveraged for novel research of *cis*-regulatory elements. A comparison of the distribution of mutations in the large-scale mutant library and sorghum association panel (SAP) provides insights into the influence of selection. EMS-induced mutations appeared to be random across different regions of the genome without significant enrichment in different sections of a gene, including the 5′ UTR, gene body, and 3′-UTR. In contrast, there were low variation density in the coding and UTR regions in the SAP. Based on the K_a/K_s value, the mutant library (~1) experienced little selection, unlike the SAP (0.40), which has been strongly selected through breeding. All mutation data are publicly searchable through SorbMutDB (https://www.depts.ttu.edu/igcast/sorbmutdb.php) and SorghumBase (https://sorghumbase.org/). This current large-scale sequence-indexed sorghum mutant population is a crucial resource that enriched the sorghum gene pool with novel diversity and a highly valuable tool for the Poaceae family, that will advance plant biology research and crop breeding.

Keywords: ethyl methyl sulfone mutants, sorghum, resequencing.

INTRODUCTION

Mutants are critical and highly valuable to plant biology research and crop breeding. Applications and deployments of mutants have increasingly grown in conjunction with genomics. Mutation breeding is a technique used to develop new crop varieties by inducing genetic changes or mutations in crops. For instance, mutations in dwarfing genes have led to the development of high-yielding, dwarf varieties of wheat and rice, which have significantly increased global food production (Hedden, 2003; Peng et al., 1999; Sasaki et al., 2002).

Ethyl methyl sulfone (EMS) has been successfully used in various crops to generate morphological diversity and understand the regulation of these traits (Richardson & Hake, 2022). EMS mutagenesis drove the improvement of many desirable traits, such as yield, fruit quality, disease resistance, and male sterility (Greene et al., 2003; Sato et al., 2006; Siddique et al., 2020). For example, in rice, EMS mutagenesis has proven effective in generating mutant lines with diverse phenotypes related to agronomic traits, such as varied photosynthetic rates of leaves (Feldman et al., 2014), improved abiotic stress tolerance, such as drought resistance based on root length and volume parameters (Mohapatra et al., 2014), and developing heat-tolerant mutants with higher photosystem II efficiency (Poli et al., 2013). Furthermore, EMS-induced mutants are exempt from genetically modified organism (GMO) regulations in many countries, including the United States and Europe, which allows them to be planted and evaluated in fields without undergoing the same regulatory processes as GMO. As a result, EMS mutants can be readily deployed and tested in agricultural settings. These examples demonstrate the potential of EMS mutagenesis to enhance yield and manage stress resistance in various crops.

Sorghum (*Sorghum bicolor* L. Moench) is a highly versatile cereal crop with significant global importance, serving as a subsistence crop for millions of people in sub-Saharan Africa and South Asia (Motlhaodi et al., 2017; Ritter et al., 2008). With a genome size estimated to be ~730 megabyte (Mb) and a diploid species ($2n = 20$), sorghum is an excellent model for studying functional genomics and genetic improvement in the Poaceae family (Paterson et al., 2009). Domesticated and bred for diverse uses, including food, fodder, and bioenergy, sorghum exhibits significant genome-level diversity (Cuevas et al., 2017; Mace et al., 2008; Olatoye et al., 2018). Functional genomics studies in sorghum have been facilitated by the availability of two sequence index mutant populations created through EMS treatment (Addo-Quaye et al., 2018; Xin et al., 2008). The sequencing of approximately 600 mutant lines from one population at 5× coverage revealed mutations covering about 30 285 sorghum genes, including 7979 genes with protein structure truncated mutations (Addo-Quaye et al., 2018; Simons et al., 2022). Our mutant population was created under the genetic background of the reference genome line BTx623 (Xin et al., 2008). A total of 256 mutant lines were first sequenced at an average of 16× coverage. This effort identified over 1.8 million canonical EMS-induced mutations, affecting more than 95% of the genes in the sorghum genome. Notably, most (97.5%) of these mutations were distinct from natural variations, demonstrating the potential for generating diverse genetic variations using this method (Jiao et al., 2016). The second population has been successfully employed to dissect several important sorghum traits, including male sterility (Chen et al., 2019), epicuticular wax (Jiao, Burow, et al., 2018), brown midrib (Tetreault et al., 2021), root (Balasubramanian et al., 2021), grain quality (Khan et al., 2023) and inflorescence development (Dampanaboina et al., 2019; Gladman et al., 2019; Jiao, Lee, et al., 2018; Poursarebani et al., 2020).

Notably, the two mutant populations reported so far focused on mutations in the coding region of genes. However, with rapid development in genomics, the detailed scrutiny of *cis*- and *trans*-acting genetic factors, epigenetic factors, and environmental influences play important roles in controlling complex, heritable traits in plants and humans (Renganaath et al., 2020; Sun et al., 2021). Phenotypic diversity within and between species is often due to variations in gene expression that are heritable (Zheng et al., 2011) and arise from mutations in DNA sequences encoding regulatory elements (e.g., enhancers and promoters) and *trans*-regulatory factors (e.g., signaling molecules, noncoding RNAs, transcription factors [TFs]) (Carroll, 2008; Stern & Orgogozo, 2008). Recent advances in next-generation sequencing (NGS) technology have accelerated research on *cis*-regulatory elements (CREs) (Schmitz et al., 2022; Zhao et al., 2021). TFs typically bind to CREs to regulate the transcription and expression of neighboring genes (Ho & Geisler, 2019; Li et al., 2015). Identifying the function of plant CREs with perturbation (Nigam et al., 2015; Olsen et al., 2014) is crucial in determining the tissue-specific or abiotic stress response expression patterns of the target gene. Mutations in CREs such as promoters, enhancers, silencers, and insulators can significantly impact gene expression by altering the binding of TFs (Galli et al., 2020). For instance, a large insertion in the enhancer of the *teosinte branched1* (tb1) gene has been suggested to underlie the morphological differences between maize (*Zea mays*) and its wild ancestors (*Zea mays* ssp. *parviglumis* and *mexicana*) (Doebley et al., 1995, 1997, 2006). Mutations in CREs for TFs such as *GW8/SPL16*, *GW7*, and *GW6* have the potential to improve grain quality in rice by fine-tuning gene expression (Ding et al., 2021; Zeng et al., 2020). Specifically, many examples have shown that single nucleotide polymorphisms (SNPs) at the promoters or enhancers can lead to significant alterations in plant development (Swinnen et al., 2016). For instance, an SNP in the promoter of the tomato *WRKY33* TF was found to reduce the self-transcription, resulting in decreased cold tolerance (Guo et al., 2022). Similarly, in wheat, an SNP located in the promoter region of the *Vrn-D1* gene has been linked to vernalization response (Zhang et al., 2012). Despite these findings, the prevalence and consequences of mutations in most core promoters and enhancers in major crops, and in sorghum in particular have yet to be explored. Moreover, as noted earlier, only a small fraction of mutations in the coding sequence regions of two sorghum mutant populations were annotated for functional genomics studies.

To create a comprehensive and large-scale sorghum mutant resource, we conducted deep sequencing (with ~38× coverage) on 897 EMS mutants. This study analyzed the impacts of the EMS-induced 9.5 million mutations on gene function and physiological pathways, covering 98% of predicted sorghum genes. For crop improvement applications, this current large-scale mutant population provides multiple alleles for most sorghum genes, allowing for genetic testing and validation of gene function and deployment for causal variants of interest. The mutations identified in the core promoter and enhancer regions will also assist in plant functional genomics research and discoveries for breeding purposes. Using the large number of mutations revealed by our sequencing efforts, we provided new insights into the distribution of EMS-induced mutations and performed comparisons with the distribution of variations in the natural sorghum population.

RESULTS

The sequenced mutant resource covers 98% of the genes with at least one predicted deleterious mutation

In our previous study, we reported a sorghum mutant population established using EMS treatment in the reference genome line BTx623 background via a single seed descent approach (Jiao et al., 2016; Xin et al., 2008). The initial sequencing of 256 mutants, with an average coverage of 16×, facilitated reverse genetics studies using sorghum mutants. However, this effort only identified missense or high-impact mutations (stop codon gains, splicing donor/-acceptor site changes, and start codon losses) in approximately 50% of sorghum genes. Since EMS mutations occur randomly, the relatively small size of the sequencing population also constrained the availability of multiple alleles for most genes.

To create a comprehensive sequence-indexed mutant resource for gene function studies, we performed whole-genome sequencing of 897 mutants. To cover as many mutations as possible, 20 individual M3 plants were randomly selected to represent each M2 mutant line. The average sequencing depth was 37.43×, with a range of 12× to 112× (Figure S1a). Using our previously validated high-accuracy pipeline (Jiao et al., 2016), we identified 9 561 298 mutations, after removing seven outlier mutant lines with an exceptionally high number of variations (more than 150 000 mutations). On average, each mutant line had 12 577 mutations, equivalent to approximately 17 mutations per Mb per line (Figure S1b). In our initial sequencing round, with an average depth of 16×, we detected 7660 mutations per line. It indicated that the increased sequencing depth enlarged the power of mutation detection. Remarkably, approximately 91% of these mutations were unique to individual mutants (Figure S1c), aligning with the random nature of EMS mutations. Among these EMS-induced mutations, only 11.2% were identified in the natural variation data from the recent whole-genome resequencing of the SAP (Boatwright et al., 2022). This finding underscores the potential of our mutant population to introduce new alleles that could be valuable for sorghum improvement efforts.

According to the annotation, approximately 15% of the mutations were found in the genes (Figure 1a). When combined with our previously sequenced 256 mutant lines (Jiao et al., 2016), we discovered that 97.86% (33 388 out of 34 118) of sorghum genes have at least one mutation that could cause an amino acid change, stop gain, start loss, or splicing site mutations (Table 1). On average, there are 12 mutations per gene. The availability of multiple alleles will enable the validation of gene functions through complementation tests. Approximately 46% of the sorghum genes had at least one large effect truncated protein mutations, including splice site acceptor, splice site donor, stop gained, start lost, or stop lost (Figure 1b). Notably, we observed that 14.5% of the amino acid variation can result in a change from nonpolar to polar or acidic to basic, or vice versa (Figure 1c). This finding is interesting since the ratio of polar and non-polar amino acids is a critical feature of the protein (Panja et al., 2015; Yuan et al., 2015). An alteration in this ratio could significantly impact the protein structure, ultimately resulting in a change in phenotype (Jiang et al., 2020).

The coding sequences were further analyzed to identify mutations that could have a significant impact, and they were classified based on different types of mutations, such as stop gain/loss, splicing donor/acceptor, and start loss (Table 1). Out of the 416 503 amino acid change mutations, 64% were predicted to be deleterious, with a SIFT score (Vaser et al., 2016) of less than 0.05 in 33 202 genes (Figure 1d). Additionally, we identified 21 918 cases of stop-gain mutations in 13 098 genes, accounting for 38.4% of the sorghum genes. A total of 6495 splice site donor mutations were identified in 4937 genes (14.5%) and 718 instances of start-lost mutations in 666 genes (2%) (Table S1). These findings suggested that the sorghum mutant population can be

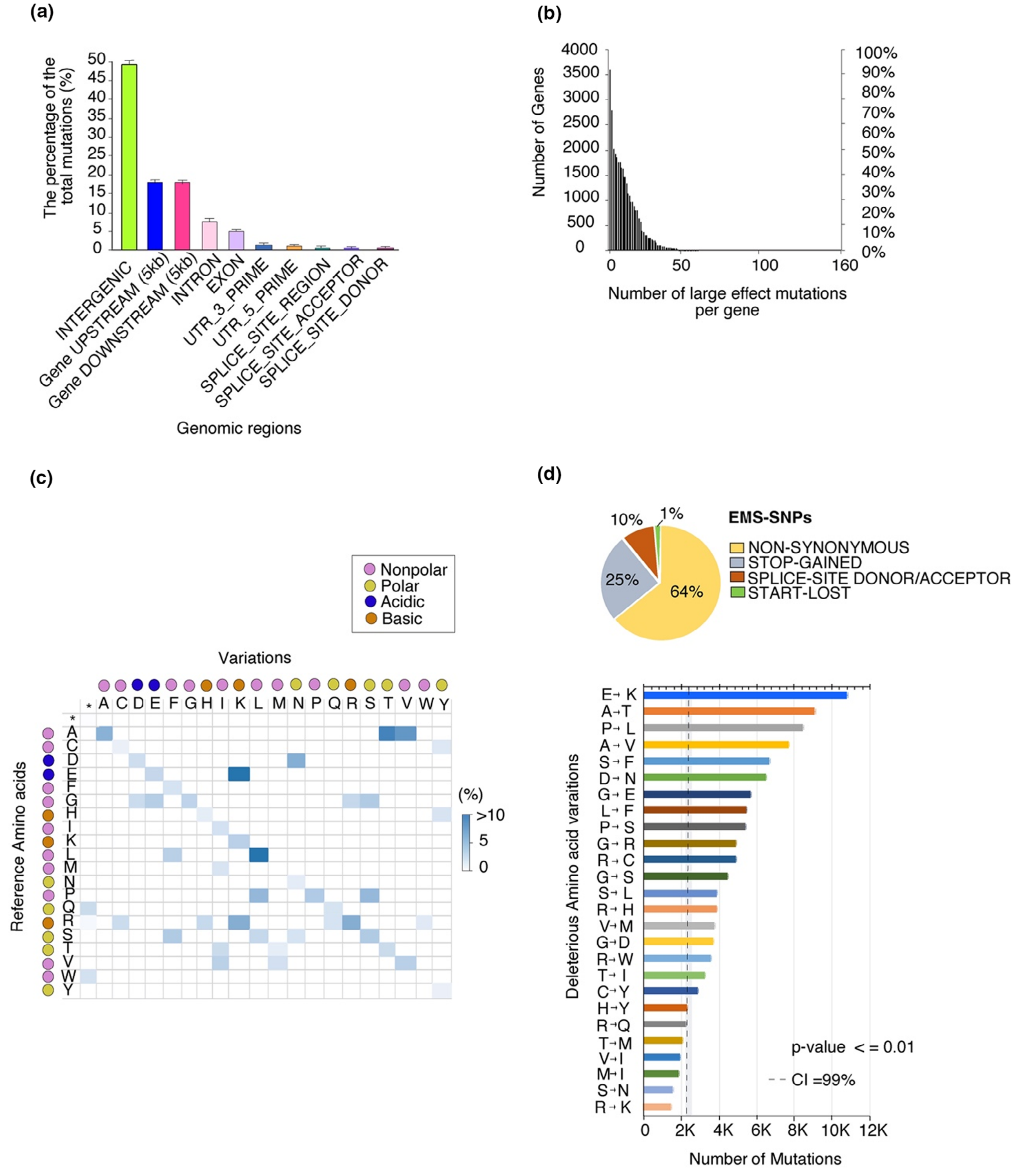

Figure 1. A summary of the mutations.

(a) Distribution of ethyl methane sulfonate (EMS)-induced mutations in different genomic regions of the Sorghum Genome. The abundance is plotted in percent.

(b) Distribution of the number of large effect mutations in genes.

(c) Amino acid changes caused by EMS. The diagonal matrix shows the percent change (%) from one amino acid to another amino acid. The biochemical property of every amino acid is on the top.

(d) A pie chart (upper panel) showing the distribution of high impact mutations in coding regions i.e., non-synonymous, stop-gained, splice-site donor/or acceptor, and start lost. The bar graph shows the frequency of the predicted deleterious mutations (SIFT score < 0.05) deleterious amino acid variation. The dotted line depicts a confidence interval of 99%.

Table 1 Summary of the impacts of the mutations on the genes in the large-scale sequenced sorghum mutant population

Mutation effect	Effect classic	No. of mutations	No. of genes
Moderate	NON_SYNONYMOUS_CODING	416 503	33 202
High	STOP_GAINED	21 918	13 098
High	SPLICE_SITE_DONOR/ ACCEPTOR	6495	4937
High	START_LOST	718	666

used for further gene functional and genetic research, given the availability of multiple alleles.

This comprehensive mutant library provides an excellent resource to bridge the gap between model plants and grass cereals in gene functional studies. For instance, according to our analysis of the KEGG database, our EMS mutations are likely to affect approximately 95% of the genes in 17 identified metabolic pathways (Figure S2; Table 1). This insight can be leveraged to deepen our understanding of metabolic pathways and drive improvements in crop production for the future.

Mutations in the *cis*-regulatory regions

Mutations in *cis*-regulatory sequences can disrupt the interaction between *cis* and *trans* elements, affecting transcriptional initiation and gene expression, and ultimately leading to phenotypic changes (He et al., 2021). As mentioned above, SNP in the promoters and enhancers also have the potential to change the phenotypes (Swinnen et al., 2016). To provide resources to investigate the function of CREs, we first predicted promoters and enhancers for 23 501 and 17 212 genes, respectively (Figure 2a). A total of 375 372 and 234 948 mutations were identified in the promoters and enhancers of 18 000 and 11 790 genes respectively (Figure 2a). The occurrence of TATA-less (−) promoters is significantly higher ($P < 0.0001$) than that of TATA-containing (+) promoters (Figure 2b). Previous studies reported a significant GC-compositional strand bias around the transcription start sites (TSSs) in plant genes (Lis & Walther, 2016; Tatarinova et al., 2003). Our GC-Skew analysis (Gao & Zhang, 2006) of the region encompassing the TSSs revealed a distinct strand bias in GC composition, with the highest values precisely at the TSSs (Figure S3). The GC-Skew diminished rapidly downstream of the TSSs. Furthermore, two prominent GC-Skew peaks emerge in the TATA promoter, specifically at positions −25 bp and +15 bp relative to the TSS, respectively (Figure S3a). Importantly, these peaks exhibit more significant GC-Skew values. In contrast, the AT-Skew does not display a significant bias, except for two minor peaks observed at positions −15 bp. It also suggested that TATA-less promoters exhibited GC-Skew bias, but the peaks were not as pronounced as in the TATA promoter (Figure S3b).

Since the binding of TFs to the promoter and enhancer regions regulates gene expression, we also analyzed mutations in the transcription factor binding sites (TFBSs) based on the TF binding motifs obtained as described in the method (Figure 2c). We found that EMS mutations affected a total of 1238 and 1712 TFBSs in promoter and enhancer regions, respectively. The most prominent perturbed TFs were *bHLH*, *NAC*, *ERF*, *bZIP*, *MYB*, *C2H2*, *WRKY*, and MYB-related factors (Table S2). Our systematic identification of mutations in CREs and TFBSs provides a valuable resource for understanding gene expression regulation in grasses. Although further work is required to establish the causal relationship between mutations in regulatory regions to gene expression and phenotypes, our mutation data provide a much-needed foundational resource for such research.

EMS-induced mutations happened randomly in the sorghum genome

Both natural and induced mutations have been reported to exhibit high and low mutation rates in populations with limited size (Monroe et al., 2022; Yan et al., 2021). For instance, the sequence context and chromatin structure biases of EMS-induced mutations were observed in rice using 17 397 EMS-induced SNPs from 52 mutants (Yan et al., 2021). Our previous studies have shown that GC methylation did not affect the distribution of sorghum EMS mutations (Jiao et al., 2016). With the advantage of our new deep sequencing of 897 mutant lines, we further assessed whether EMS induced mutations differently at specific genic sites within particular motifs. We analyzed the nucleotide frequencies 10 bp upstream and downstream flanking the 9.5 million identified SNPs. Remarkably, no preferred genic motif or bases upstream or downstream of the mutated site were identified (Figure 3a).

Natural mutations have been observed to exhibit non-random biases towards non-essential genes and are influenced by DNA methylation in Arabidopsis (Monroe et al., 2022). However, it is unclear how natural mutations behave in crops that are subject to strong breeding selection. To address this, we compared our large EMS mutant population to the sorghum association panel (SAP). As described earlier, our mutant population represents mutations with less selective pressure, while the SAP is a population with strong selection due to the substantial number of breeding lines present. Furthermore, the sequencing depth of SAP (~38×) is comparable to ours, enabling a direct comparison. We compared the mutation/variation density of the two populations across different genomic regions (coding sequence, intron, 5′ and 3′ UTRs, 5 kb upstream and downstream of genes, and intergenic regions) (Figure 3b). Our mutant population exhibited the highest mutation rate in intergenic regions, consistent with the low likelihood of mutations in the intergenic regions

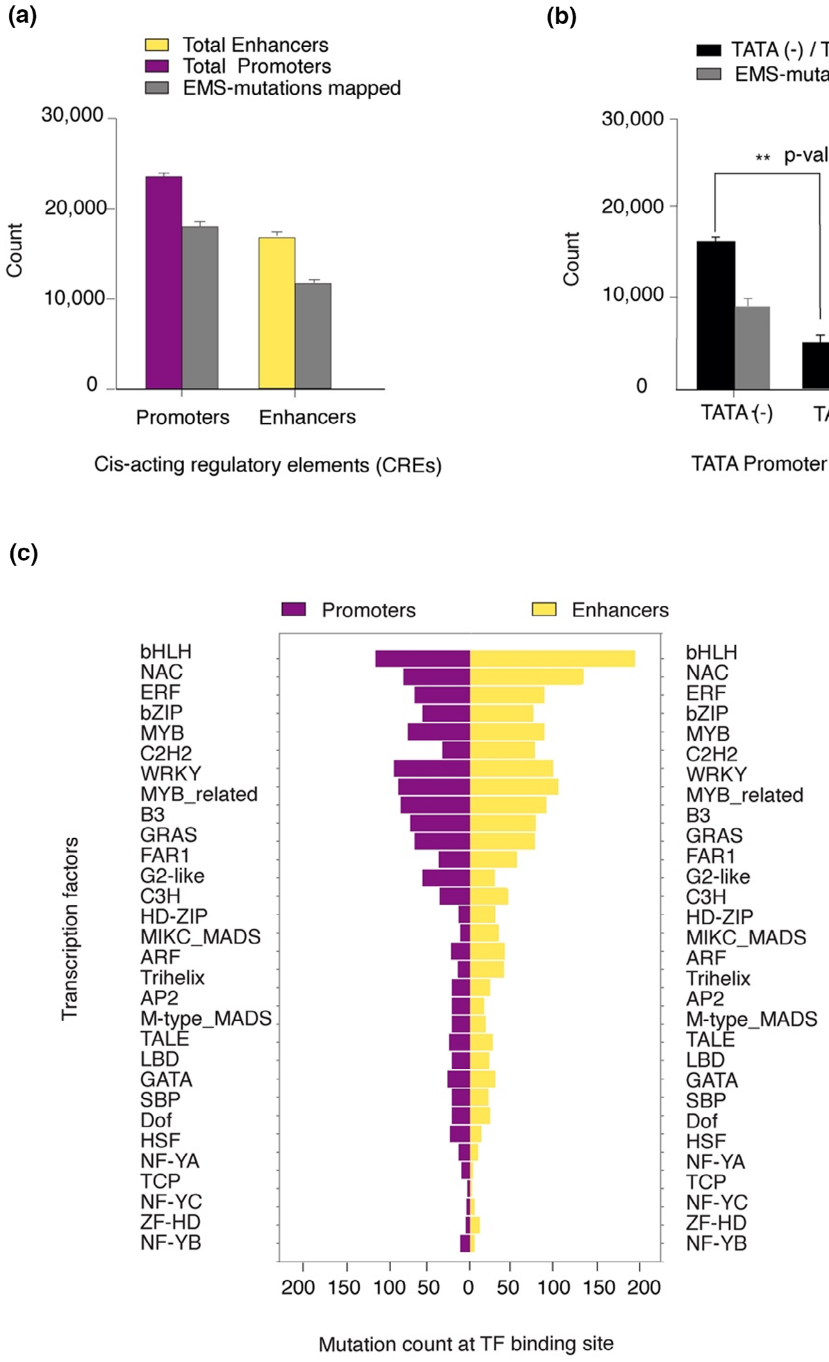

Figure 2. The impact of ethyl methane sulfonate (EMS) mutations on *cis*-regulatory elements.
(a) Distribution of EMS-Mutations on promoters and enhancers: This bar plot showcases the distribution of EMS-induced mutations across the genome's promoters and enhancers. The purple bars represent the total number of identified promoters, while the yellow bars signify the total number of identified enhancers. The gray bar represents the cumulative count of EMS-induced mutations detected within both *cis*-regulatory elements.
(b) Categorization of promoters into TATA-Containing and TATA-Less Categories. Promoters with TATA-box motif in their regulatory regions are shown in black, representing TATA-containing promoters. Those devoid of the TATA-box motif are also visualized in black, symbolizing TATA-less promoters. The gray bars correspond to the quantity of EMS-induced mutations identified in each promoter category. Statistical significance regarding the disparity in mutation rates between TATA-containing and TATA-less promoters was assessed via a chi-square test, with the *P*-value provided (denoted with **).
(c) Impacts of EMS mutations on transcription factor binding sites (TFBSs). TFBSs located within promoters are depicted in purple, while those within enhancers are represented in yellow. The number of EMS-induced mutations affecting TFBSs is specified for each *cis*-regulatory element.

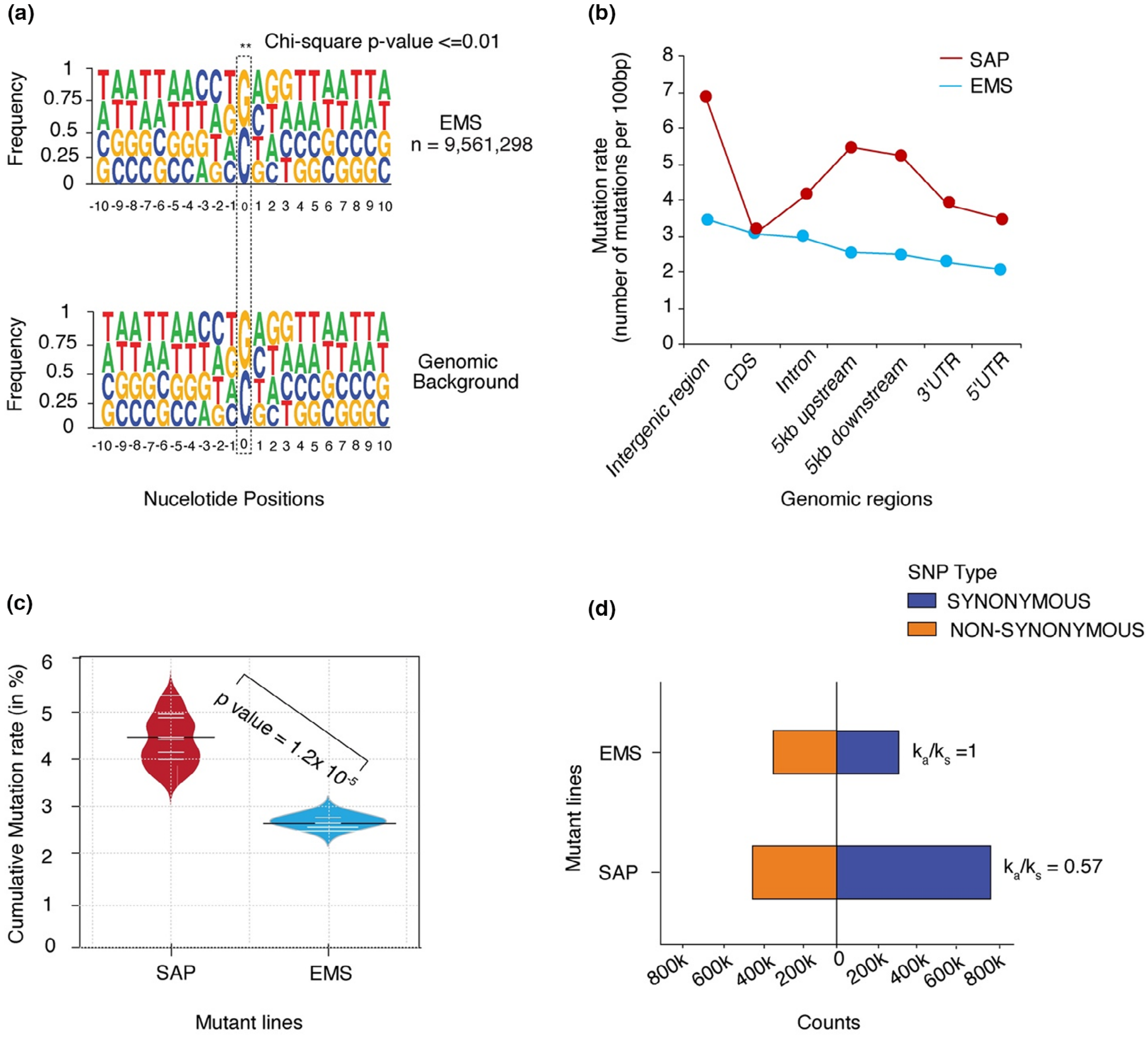

Figure 3. The unbiased distribution of ethyl methane sulfonate (EMS) mutations and the comparison with the natural sorghum population.

(a) The comparison of the frequency of the base pairs within 10-bp flanking region encompassing the EMS mutations and all the GC pairs of the sorghum genome (**$P \leq 0.01$).

(b) The comparison of the mutation/variation rate in the EMS mutant populations and sorghum association panel (SAP) crossing different genomic regions. In the EMS mutant population, only the number of GC base pairs in the genome was used for the calculation of the mutation density.

(c) The deviation of mutation/variation rate crossing different genome regions in the two populations. The chi-square P-value was calculated to measure the difference in their standard deviation.

(d) Comparison of the K_a/K_s between the EMS mutant population and SAP, where K_a is nonsynonymous substitutions and K_s is synonymous substitutions.

causing phenotypic changes. Additionally, if a mutation in the coding sequence region led to lethality or sterility, the homozygous mutant of that gene would not be conserved in the mutant population, which can reduce the frequency of this mutation in the population. The standard deviation of mutation rates in different genomic regions was much smaller in our mutant population (0.0048) compared to SAP (0.013) (Figure 3c), indicating stronger non-uniform selection pressures on variations in different regions of the genome in the natural population than in the EMS mutant population. We also observed that coding sequences had the second-highest mutation rate in the mutant population, but the lowest in SAP (Figure 3b), suggesting that breeding selection in SAP eliminated mutations in coding sequences that could cause unfavorable traits. Finally, we compared the ratio of nonsynonymous to synonymous mutations (K_a/K_s) in the two populations, finding that our EMS mutant population had a K_a/K_s ratio of 1, while SAP had a K_a/K_s ratio of about 0.57 (Figure 3d). The lower proportion of nonsynonymous variations in SAP indicates stronger

© 2023 The Authors.
The Plant Journal published by Society for Experimental Biology and John Wiley & Sons Ltd.,
The Plant Journal, (2023), doi: 10.1111/tpj.16582

selection against nonsynonymous mutations during breeding. In summary, our analysis suggested that EMS-induced mutations occur randomly without bias, and selective pressure causes uniform variation density across different regions of the genome.

EMS mutations changed codon usage bias and amino acid dynamics

The GC content and codon usage bias (CUB) of a genome can vary greatly among different species, due mainly to differences in mutational pressure (Lagerkvist, 1978). To investigate this feature further, we analyzed the impact of nonsynonymous mutations on different types of amino acids (Figure 4; Data S2), As shown in Figure 4(a), the codons with high G/C content are more likely to be mutated by EMS treatment. Additionally, we found that amino acids with the same number of codons had varying mutation rates. For example, alanine (Ala), glycine (Gly), proline (Pro), threonine (Thr), and valine (Val), which are all encoded by four codons, exhibited different mutation rates. Similarly, even though asparagine (Asn), aspartic acid (Asp), lysine (Lys), and phenylalanine (Phe) have two codons each, codons encoding for amino acids such as Asn, the codons of Lys (AAG) and Phe (TTC) were more frequently mutated than Asp (Figure 4a; Data S2). A similar pattern was also observed in arginine (Arg), leucine (Leu), and serine (Ser), which have six codons each. The frequency of codon usage in the reference genome and the mutant population was compared using the Kolmogorov–Smirnov test (Massey, 1951), which showed a statistically significant difference (P-value = 8.832e-7) with a test-statistic D of 0.2375 (Figure 4a, left panel). We also compared the overall frequencies of various amino acids before and after EMS mutagenesis in the population (Figure 4b). As shown in Figure 4(b), Asn, Lys, and Phe showed a significant increase following mutagenesis in the whole population, while Ala, Gly, and Pro were significantly reduced. For example, EMS-induced GC → AT mutations cannot change the two codons of Lys (AAA and AAG) to another amino acid. But other codons, such as AGA, Arg and GAA, Gly, could be mutated into the Lys codons. We have measured the amino acid content in the seeds of 256 mutants (Khan et al., 2023). Increased lysine content was observed in most of the measured mutant lines, which agreed with the changed codon usage frequency in the mutant population (Figure 4c).

All the mutants are publicly available

To provide access to the sorghum EMS mutant resources to the plant biology and sorghum breeding community, all the mutations are searchable through the Sorghumbase (Gladman et al., 2022) (https://www.sorghumbase.org/) and the search page 'SorbMutDB' (https://www.depts.ttu.edu/igcast/SorbMutDB.php). SorghumBase is an emerging genomic resource for the sorghum research community and is under rapid development with a rich suite of comprehensive views and tables for variation data, enabling integration of mutant's genomic context, gene and regulation effect, genotype frequency, and sample genotypes. The SorbMutDB search page allows users to rapidly search their interested genes using both v1 and v3 sorghum reference gene IDs and identify mutations along with the SIFT scores in their target gene(s). The mutation summary for all the accessions can be downloaded from the download data option in the database. The seeds of the mutant lines can be requested as described on the website.

DISCUSSION

Harnessing the sorghum mutant population for functional genomics studies

The sequence-indexed sorghum mutant population, and its remarkable phenotypic diversity, serve as an excellent platform for investigating gene functions within the grass family. Notably, the *Sorghum bicolor* genome has not undergone significant whole-genome duplication events, resulting in a predominance of single-copy genes. This characteristic facilitates the observation of phenotypes without the necessity of generating double or triple mutants in duplicated gene sets. Given the substantial size of this mutant population, which comprises over 6400 independent M2 lines, comprehensive phenotyping across various traits, such as root architecture and responses to abiotic stresses, will require collaborative efforts from the broader plant biology community. So far, we have isolated hundreds of lines with phenotypes related to important traits, such as plant height, the production of epicuticular wax, tiller, inflorescent structure, grain quality, and so on (Khan et al., 2023; Xin et al., 2021). To enhance the utility of this mutant population, we have developed both forward and reverse genetic pipelines, along with accompanying protocols, to facilitate gene function studies. (Jiao, Burow, et al., 2018; Wang et al., 2021).

This mutant population has proven to be a valuable resource for exploring the gene functions related to important traits via the reverse genetics approach. For instance, we have confirmed the involvement of four genes in epicuticular wax biosynthesis by identifying orthologs of Arabidopsis genes and subsequently validating them through co-segregation tests in F_2 populations (Jiao et al., 2016). The mutant in the ortholog of the barley COMPOSITUM 1 gene has been used to study the inflorescence evolution in the grass family (Poursarebani et al., 2020). In our efforts to capture as many mutations as possible, we adopted a strategy of pooling 20 M3 individual plants derived from the same M2 parent for sequencing. The variation calling tool indicated the zygosity of the sequencing data for each mutation. However, it is important to note that it did not

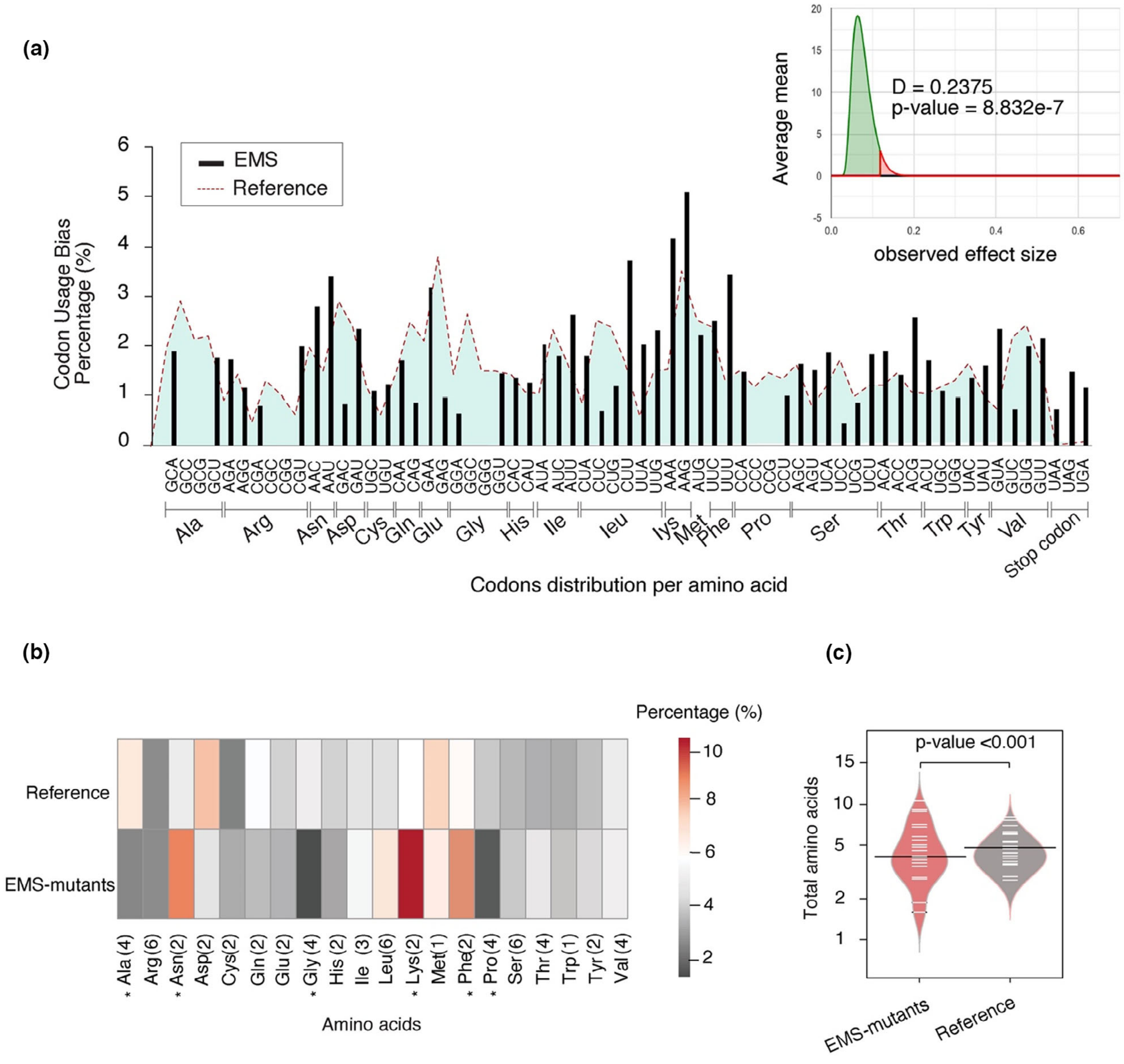

Figure 4. Statistics of the change of codon usage and amino acid in the EMS mutant population.

(a) The change of the codon usage in the EMS mutant population compared to the reference genome. Black bars represent the frequencies of the codons in the mutant population. The red dashed line shows the percentages of each codon in the Sorghum reference genome. The Smirnov test is shown the upper right side of the graph.

(b)The alteration in the frequency of amino acids within the mutant population compared to the reference genome. The star signs refer to the significant differences (P value < 0.001) in the frequency of the amino acid.

(c) The statistical significance of these mutations' impact on amino acid frequency was evaluated through a chi-square test, resulting in a P-value of <0.001.

reflect the zygosity of a mutation in a specific mutant line. To address this, we implemented Competitive allele-specific PCR (Myakishev et al., 2001) as an efficient method to screen the M3 pools, enabling the identification of mutations of interest as the initial step in our reverse genetics studies.

The accuracy of mutation calls from the sequencing data is pivotal for the effective utilization of the mutant population in reverse genetics studies. The bioinformatics pipeline employed in this study was developed and rigorously validated during our initial sequencing phase. Within our mutation detection pipeline, we have implemented several key steps to mitigate the influence of random variations stemming from the sequencing and read alignment processes, as well as background mutations. Background mutation is a major cause of the false positive in the EMS mutation calling. In this context, background mutations refer to the variations in the parental seeds utilized during

EMS treatment when contrasted with the reference genome sequence. We employed two strategies to eliminate the background mutations: First, we conducted sequencing of our parental line, BTx623, achieving a 30X coverage (Jiao et al., 2016). This analysis revealed 359 980 variations, potentially encompassing background mutations and locations prone to alignment errors. To mitigate false positives, any mutations overlapping with these identified variations were excluded. Mutations exhibiting a high allele frequency in the population may be classified as background mutations, given that EMS-induced mutations occur randomly. Through our Sanger sequencing of 1000 mutations randomly chosen, our initial sequencing round achieved a 98.7% accuracy rate (Jiao et al., 2016). With the average sequencing depth of this study having risen from 16× to 38×, we anticipate a corresponding increase or similar in the accuracy of mutation calling within the new dataset. Furthermore, additional published in-depth investigations on calling EMS mutations from short-read sequencing data (Simons et al., 2022) further supported the reliability of our pipeline.

Forward genetics is a potent method for uncovering unique gene functions specific to sorghum or those not extensively explored in other plant species. For instance, the multisided (MSD) mutants represent a distinctive group characterized by the fertility of both sessile spikelet (SS) and pedicellate spikelet (PS). In contrast, in the wild-type sorghum, only SS can set seeds. Through a bulk segregant analysis (BSA) of the *msd1* mutant, we identified a TCP transcriptional factor responsible for regulating PS fertility via the modulation of jasmonic acid levels (Jiao, Lee, et al., 2018). It's worth noting that this TCP gene lacks an ortholog in Arabidopsis and was not previously studied before our investigation in the *msd1* mutant. This highlighted the effectiveness of our sorghum mutant as a valuable platform for uncovering novel gene functions in grass crops. Our highly efficient BSA pipeline is now publicly accessible via SciApps (Wang et al., 2021).

EMS predominantly induced G/C to A/T transition in sorghum

Studies have shown that EMS primarily induces G/C to A/T transitions in sorghum via guanine alkylation, which can cause incorrect pairing during DNA replication (Sega, 1984). However, EMS can also induce non-GC to AT transitions, typically not induced by alkylating agents. One proposed mechanism for EMS-induced non-GC to AT transitions involves replicating damaged DNA templates. It has been suggested that EMS induces DNA damage, which can cause replication forks to stall and subsequently bypass the damaged site through an error-prone mechanism, leading to mutations, including non-GC to AT transitions (Seplyarskiy et al., 2019).

We identified the non-GC to AT mutations potentially induced by EMS by applying all the filtration steps except for the one exclusively retaining GC to AT variations. Our analysis revealed 462 561 non-GC to AT point mutations, representing 4.6% of all mutations detected by our pipeline. Among these mutations, 4.33%, 0.15%, and 0.13% are point mutations, insertions, and deletions, respectively (Table S3). Notably, approximately 42% of the non-typical G/C to A/T mutations were transitions from A/T to G/C that resulted in changes to the coding sequence of up to 7650 genes. Different studies have reported a wide variation in the frequency of non-G/C to A/T mutations, with 1% in *Arabidopsis thaliana* (Greene et al., 2003), 30% in rice (Till et al., 2007), and 20.3% in maize (Lu et al., 2018). The whole-genome sequencing of EMS-treated MicroTom lines revealed 39–76% G/C to A/T transitions (Shirasawa et al., 2016), while the sequencing of 95 tomato mutants through whole-exome sequencing displayed 20.7% G/C to A/T transitions (Yano et al., 2019). The significant variation in the frequency of G/C to A/T transitions among different species is likely a species-specific phenomenon that may be linked to differences in DNA repair efficiency. The percentage of GC to AT mutations discovered in different studies has varied from 99% in Arabidopsis (Greene et al., 2003) to 58% in rice (Yan et al., 2021), depending on EMS treatment protocols, sequencing depth, population size, and data analysis approaches.

Some of the non-GC to AT mutations identified in this large-scale mutant population may have occurred spontaneously as the sorghum plants grew in the field. For instance, exposure to ultraviolet radiation in the natural environment can cause various types of mutations in plants, such as thymine dimers, base substitutions, deletions, and insertions (Frohnmeyer & Staiger, 2003; Nakamura et al., 2021). Since we sequenced the M3 generation, the average mutation rate of non-GC to AT mutations in each line was calculated to be 2.37×10^{-7} mutations per generation. The spontaneous mutation rates of maize and rice have been estimated to be 10^{-8} per site per generation (Li et al., 2014; Yang et al., 2017), which is lower than the non-GC to AT mutation rate in our mutant population. Therefore, we conclude that EMS predominantly induced GC to AT transitions in sorghum, with a rate of over 95%.

Sorghum genes not covered by mutations are predominantly short genes

Another sorghum EMS mutant population has recently become available (Addo-Quaye et al., 2018; Simons et al., 2022). However, this resource was generated from low-depth sequencing (average coverage of 7×) of 589 mutant lines with truncated protein mutations (splice site acceptor, splice site donor, stop gained, start lost, and stop lost) that only covered 7979 sorghum genes. In contrast, our sequencing effort resulted in knock-out mutations for over 15 000 sorghum genes. Merging with this sorghum mutant population revealed that 517 sorghum genes lacked mutations that could at least cause the amino acid change. These genes had a shorter coding sequence than the average length of the entire genome, while no significant difference was observed in the GC content (Figure S4a,b). This suggests that the shorter length of these genes may explain the lower likelihood of mutation caused by EMS treatment. Gene ontology analysis revealed an enrichment of 64 genes associated with mitochondria and chloroplast-related processes, suggesting that mutants in these genes with essential functions may not be viable. For instance, the gene Sobic.002G044500 was annotated as chloroplast 30S ribosomal protein S16, but no mutant is available for its Arabidopsis ortholog ATCG00050 according to The Arabidopsis Information Resource (TAIR, https://www.arabidopsis.org/servlets/TairObject?id=500229539&type=locus). Another gene, Sobic.001G110701, is the ortholog of Arabidopsis gene AT3G01480 cyclophilin 38, which is crucial for the assembly and maintenance of photosystem II. In Arabidopsis, the mutant of this gene fails the dynamic greening process of etiolated leaves (Sirpiö et al., 2008), which explains the absence of a mutant in this gene in sorghum. Taken together, these findings suggest that these 517 sorghum genes that are not covered by any of the EMS mutant populations may also be involved in maintaining essential plant functions, in addition to being small in size.

EXPERIMENTAL PROCEDURES

Generation of the sorghum EMS mutants

The EMS treatment was achieved as previously described (Xin et al., 2008). Briefly, dry *S. bicolor* BTx623 seeds in batches of 100 g (~3300 seeds) were soaked with agitation for 16 h at 3724***g*** on a rotary shaker in 200 ml of tap water containing 0.1–0.3% EMS (v/v). The treated seeds were carefully washed in ~400 ml of tap water for 5 h at ambient temperature, wash water was changed every 30 min. Mutagenized seeds were air-dried and prepared for planting. The air-dried seeds were planted at a density of 120 000 seeds per hectare. Before the anthesis, each panicle was covered with a 400-weight rainproof paper pollination bag (Lawson Bags) to prevent cross-pollination. Each bag was injected with 5 ml Chlorpyrifos (Dow AgroSciences) at 0.5 ml L^{-1} to control maize earworms that might hatch within the bag and destroy the seeds. sorghum panicles were harvested manually and threshed individually, and M2 seeds were planted in one row per head. To ensure high mutagenesis efficiency, only panicles that set 10% or fewer seeds were allowed to propagate. Three panicles from each M2 head row were bagged before the anthesis. Only one fertile panicle was used to produce the M3 seeds. For DNA extraction, duplicate leaf samples were collected from the same fertile plant, and both the leaf samples and the panicle were barcoded. Seeds from the barcoded M2 plants were harvested as M3 families of seeds. For phenotypic evaluation, each M3 family of seeds was planted as one row in the field. Many of the mutant lines exhibited diminished seed production during the M3 generation. Consequently, 10 panicles were bagged for each M3 head row and pooled as M4 seeds, which will be distributed to end users upon request.

DNA sample preparation and sequencing

The DNA extraction method described by Jiao, Lee, et al. (2018) was used to prepare genomic DNA. Due to a lack of M3 seeds from many lines and the extended storage of genomic DNA extracted from the original M2 plants, fresh genomic DNA was obtained from pooled M3 plants to ensure high-quality fresh DNA. To capture most mutations present in the original M2 plants, young leaf tissue samples from 20 random individual plants were pooled for each M3 family. These pooled leaf samples were freeze-dried for 2 days using a Labconco freeze dryer to prepare genomic DNA.

The Hamilton Vantage robotic liquid handling system and the Kapa Biosystems HyperPrep library preparation kit (Roche) were used to prepare plate-based DNA libraries for Illumina sequencing. Genomic sample DNA (100–200 ng) was sheared to 500 bp using a Covaris LE220 focused-ultrasonicator. The resulting DNA fragments were size-selected using double-SPRI with TotalPure NGS beads (Omega Bio-tek, Nocross, GA, USA), and then end-repaired, A-tailed, and ligated with Illumina-compatible unique dual-index sequencing adaptors (IDT, Inc., Coralville, IA, USA). The prepared libraries were quantified using the KAPA Illumina library quantification kit (Roche, Indianapolis, IN, USA) and a LightCycler 480 real-time PCR instrument (Roche). The quantified libraries were then multiplexed and prepared for sequencing on the Illumina NovaSeq 6000 platform using NovaSeq XP v1.5 reagent kits (Illumina, San Diego, CA, USA) with an S4 flow cell, following a 2 × 150 indexed run recipe.

Identification of EMS-Induced SNPs

All high-quality reads were first aligned to the sorghum reference genome v3.1 (McCormick et al., 2018) using BWA (Li & Durbin, 2009). The resulting alignment files were converted to BAM format and sorted using Samtools (Li et al., 2009). Variations were called using Bcftools version 1.9 (Li, 2011) from unique reads with both sequencing and alignment quality >20. The filtration process to identify high-confidence EMS-induced mutations followed the same steps as our previous study. First, any variations with more than one alternative allele were discarded. Although the parental line utilized for the EMS treatment was the reference genome line, it is important to note that there were still some variations due to the spontaneous mutations and the potential low content of the heterozygosity. These variations were defined as background mutations in this study. The parental BTx623 were sequenced to 30X coverage using Illumina sequencing in our previous study (Jiao et al., 2016). The variations between our parental line BTx623 and the reference genome BTx623 were called using the same pipeline described above. In addition to eliminating mutations that overlapped with variations identified in the parental line, we also applied a filter to exclude mutations detected in more than 10 lines. This step is based on the fact that the likelihood of EMS-induced mutations occurring at the same genomic location in multiple independent lines is low. After removing the background variations, high-confidence SNPs were selected based on the following criteria: (i) the SNP is supported by at least three reads or two complementary reads; and (ii) the sequence change is from G/C to A/T. We performed functional annotation of SNPs using SnpEff (Cingolani et al., 2012) based on the sorghum genome annotation v3.1 (McCormick et al., 2018) and used SIFT 4G to predict deleterious mutations (Vaser et al., 2016).

Mutations in the metabolism pathways

The sorghum genes from the BTx623 reference genome annotation v3.1 were classified into various metabolic pathways categories using the KEGG database (Kanehisa et al., 2022). To determine the extent to which our EMS mutations affect each metabolic pathway or category, we further conducted an overlap analysis. The chi-square test was conducted to assess the significance of each metabolic category and filter criteria were set as $P \leq 0.05$. The metabolic network was visualized by Cytoscape version 3.9 (Shannon et al., 2003).

Identification of mutations in the promoter and enhancer regions

In-house Perl and Python scripts were used to extract the promoter regions for each gene from the sorghum reference genome (Singh et al., 2015). We examined several promoter properties, including the presence of a TATA box motif, CpG sites, and islands in the promoter region, which were expected to affect mutational variability. These properties were selected based on their diverse range of effects, as shown in previous studies by Landry et al. (2007) and Hornung et al. (2012). The detailed parameters are below. For TSS prediction in the upstream regions of selected genes, we used the TSSPlant program (Shahmuradov et al., 2017). The program predicted TSSs within the range of sequence extending from −1000 to +250 bp relative to the start codon of the gene. The average length of the selected promoter's sequences was >100, while for enhancers, it was >200 bp. TSSPlant utilizes the expectation–maximization algorithm and neural networks to estimate 17 features for predicting TATA promoters and 15 for predicting TATA-less promoters. TATA promoters were characterized by a conserved TATA box sequence at the −25 position relative to the TSS, while TATA-less promoters typically exhibit other distinct features. To extract the possible TSS prediction for each gene, we used the probability density function (Dai et al., 2006) and selected the start site with the best score (*P*-value cut off 0.01). The GC-content profiles, including GC-skew and AT-skew, of both TATA-containing and TATA-less promoter sequences were analyzed by our in-house Python script. GC-Skew was used as an index for plant promoter prediction (Fujimori et al., 2005). The analysis followed the prescribed parameters, with a halting parameter set at 100 and a minimum segment length of 3000 bp (Gao & Zhang, 2006). Accessible Chromatin Regions were mapped to SNP locations using previously published data (Zhou et al., 2021), and enhancers were identified and verified using iEnhancer-2L (Liu et al., 2016). iEnhancer-2L is a method that classifies both enhancers and their potency based solely on sequence information and has become increasingly popular in genomics analysis.

Identification of the mutations affected TFs and their binding sites

We obtained the sorghum TF genes from the PlantTFDB (Jin et al., 2017). To further evaluate TFBSs, we utilized the Plant Promoter Analysis Navigator (Chow et al., 2019) (PlantPAN3.0, http://plantpan.itps.ncku.edu.tw/). This tool allowed us to identify TFBSs and the corresponding TFs within a promoter or a group of promoter sequences. In this analysis, sorghum was selected as the background database, and the motifs of sorghum TFs from the output of the PlantPAN3.0 analysis were utilized.

Flanking sequences of the induced mutations

The flanking sequences of the induced mutations and all the background GC base pairs in the sorghum genome were extracted by Bedtools (Quinlan & Hall, 2010). The preferential base pair frequency of 10-bp flanking was calculated using k-mer probability logo (kpLogo), a tool for the sensitive discovery, and representation of position-specific sequence motifs from a set of sequences with precisely defined positions (Wu & Bartel, 2017).

Measurements of codon usages and amino acid preferences

The change of the codon and amino acid frequency in the mutant population was predicted by SnpEff. The numbers for each codon and the corresponding original amino acid and substitution amino acid were counted for the calculation of their ratio to the total amino acids. The percentage of each codon and corresponding amino acid in the sorghum reference genome was calculated based on the annotation of genes. To assess the statistical significance of the disparity between two cumulative distributions of mutated codon frequency before and after EMS mutagenesis, we employed the *D*-statistic of the two-sample Kolmogorov–Smirnov test (Massey, 1951). The *D*-statistics were calculated using the R function ks.test, enabling quantification of the dissimilarity (Lopes et al., 2007).

AUTHOR CONTRIBUTIONS

ZX, DW, JM, and YJ conceived the project. ZX, JC, CH, and GB generated the mutant population and collected DNA samples for sequencing. JM, FB, KB, CD, YY, and AL conducted mutant sequencing and the initial data processing. YJ, DN, AL, AK, and PD performed the data analysis. S-PP and DN constructed the mutant search page. SW, ZL, and MKT-R contributed to data release in SorghumBase. YJ and DN prepared the manuscript. All authors edited and approved the final version for publication.

ACKNOWLEDGMENTS

The sequencing of the mutants (proposal: 10.46936/10.25585/60001251) was conducted by the U.S. Department of Energy Joint Genome Institute (https://ror.org/04xm1d337), a DOE Office of Science User Facility, and by the DOE Joint BioEnergy Institute (http://www.jbei.org), both of which are supported by the Office of Science of the U.S. Department of Energy operated under Contract No. DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the U.S. Department of Energy (JCM). The sequencing was also supported by DOE Great Lakes Bioenergy Research Center (DOE BER Office of Science DE-SC0018409). DW, ZL, SW, and MKT-R were supported by the USDA-ARS Current Research Information System (CRIS) grant number: 8062-21000-044-000D. YJ, DN, AK, PD, and S-PP would like to thank the support from the State of Texas' Governor's University Research Initiative (GURI).

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

All raw sequencing reads for the mutants are available from NCBI SRA under the BioProject PRJNA967029 https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA967029. The mutations can be searched and downloaded through SorbMutDB (https://www.depts.ttu.edu/igcast/SorbMutDB.php) and SorghumBase (https://sorghumbase.org).

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Table S1. Summary of the genes covered by the mutant population in the major metabolism pathways.

Table S2. Prediction of the mutations in the TF binding sites.

Table S3. The number of different types of mutations identified in the sequenced sorghum mutant population.

Figure S1. Summary of the mutation identification in 879 sorghum EMS mutant lines. (a) Distribution of the sequencing depth. (b) Distribution of the number of mutations identified in the population. (c) Distribution of the number of lines the mutations appeared in the population. A total of 91% of mutations were only in one independent mutant.

Figure S2. Hub of enriched metabolic pathways in EMS-mutants and mutation coverage. The study investigated the hub of enriched metabolic pathways in the EMS mutants, and their mutation coverage was evaluated. The KEGG database was used to examine the enrichment of various metabolic processes, and 17 clusters were identified based on the chi-square test ($P \leq 0.05$). G/C to A/T mutation coverage was then examined for these clusters. Circular nodes represent distinct categories of metabolic processes, while edges indicate connectivity within each category. The size of the node is proportional to its significance, with larger nodes indicating a more significant metabolic activity in EMS mutants. The *P*-value was categorized as ≤0.05 and <0.01, and 17 categories (C1 to C17) were filtered based on this criterion and illustrated with different colors. Finally, the percentage mutation coverage was evaluated.

Figure S3. GC/AT skew in upstream and downstream regions of the transcription start site (tss) for the sorghum promoters. The GC skew is visualized in black, and the AT skew is illustrated as a dotted red line. The vertical blue dotted line marks the precise location of the TATA motif at −25 bp relative to the TSS.

Figure S4. Characterization of genes without mutant available in any sorghum mutant populations. (a) A density plot showing the length of the Coding sequences (CDS) derived from 517 non-mutant genes in comparison to the whole genomic average level. (b) The cumulative GC content of 517 non-mutant genes compared to the whole genome.

Data S2. (Excel spreadsheet). Codon usage bias in EMS mutants.

Sheet 1. Codon usage in the sorghum reference genome.

Sheet 2. Codon change in EMS population.

Sheet 3. Codon changes in EMS population comparing with the reference sorghum genome.

Sheet 4. Amino acid change and comparison.

REFERENCES

Addo-Quaye, C., Tuinstra, M., Carraro, N., Weil, C. & Dilkes, B.P. (2018) Whole-genome sequence accuracy is improved by replication in a population of mutagenized sorghum. *G3: Genes, Genomes, Genetics*, **8**, 1079–1094.

Balasubramanian, V.K., Dampanaboina, L., Cobos, C.J., Yuan, N., Xin, Z. & Mendu, V. (2021) Induced secretion system mutation alters rhizosphere bacterial composition in *Sorghum bicolor* (L.) Moench. *Planta*, **253**, 1–18.

Boatwright, J.L., Sapkota, S., Jin, H., Schnable, J.C., Brenton, Z., Boyles, R. *et al.* (2022) Sorghum Association Panel whole-genome sequencing establishes cornerstone resource for dissecting genomic diversity. *The Plant Journal*, **111**, 888–904.

Carroll, S.B. (2008) Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell*, **134**, 25–36.

Chen, J., Jiao, Y., Laza, H., Payton, P., Ware, D. & Xin, Z. (2019) Identification of the first nuclear male sterility gene (male-sterile 9) in sorghum. *The Plant Genome*, **12**, 190020.

Chow, C.N., Lee, T.Y., Hung, Y.C., Li, G.Z., Tseng, K.C., Liu, Y.H. *et al.* (2019) PlantPAN3.0: a new and updated resource for reconstructing transcriptional regulatory networks from ChIP-seq experiments in plants. *Nucleic Acids Research*, **47**, D1155–D1163.

Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L. *et al.* (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly*, **6**, 80–92.

Cuevas, H.E., Rosa-Valentin, G., Hayes, C.M., Rooney, W.L. & Hoffmann, L. (2017) Genomic characterization of a core set of the USDA-NPGS Ethiopian sorghum germplasm collection: implications for germplasm conservation, evaluation, and utilization in crop improvement. *BMC Genomics*, **18**, 1–17.

Dai, Y., Zhang, R. & Lin, Y.-X. (2006) The probability distribution of distance TSS-TLS is organism characteristic and can be used for promoter prediction. In: *Advances in applied artificial intelligence: 19th International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2006*. Annecy, France, June 27–30, Proceedings 19. Springer, pp. 927–934.

Dampanaboina, L., Jiao, Y., Chen, J., Gladman, N., Chopra, R., Burow, G. *et al.* (2019) Sorghum MSD3 encodes an omega-3 fatty acid desaturase that increases grain number by reducing jasmonic acid levels. *International Journal of Molecular Sciences*, **20**, 5359.

Ding, Y., Zhu, J., Zhao, D., Liu, Q., Yang, Q. & Zhang, T. (2021) Targeting cis-regulatory elements for rice grain quality improvement. *Frontiers in Plant Science*, **12**, 705834.

Doebley, J., Stec, A. & Gustus, C. (1995) Teosinte branched1 and the origin of maize: evidence for epistasis and the evolution of dominance. *Genetics*, **141**, 333–346.

Doebley, J., Stec, A. & Hubbard, L. (1997) The evolution of apical dominance in maize. *Nature*, **386**, 485–488.

Doebley, J.F., Gaut, B.S. & Smith, B.D. (2006) The molecular genetics of crop domestication. *Cell*, **127**, 1309–1321.

Feldman, A.B., Murchie, E.H., Leung, H., Baraoidan, M., Coe, R., Yu, S.-M. *et al.* (2014) Increasing leaf vein density by mutagenesis: laying the foundations for C4 rice. *PLoS One*, **9**, e94947.

Frohnmeyer, H. & Staiger, D. (2003) Ultraviolet-B radiation-mediated responses in plants. Balancing damage and protection. *Plant Physiology*, **133**, 1420–1428.

Fujimori, S., Washio, T. & Tomita, M. (2005) GC-compositional strand bias around transcription start sites in plants and fungi. *BMC Genomics*, **6**, 26.

Galli, M., Feng, F. & Gallavotti, A. (2020) Mapping regulatory determinants in plants. *Frontiers in Genetics*, **11**, 591194.

Gao, F. & Zhang, C.-T. (2006) GC-Profile: a web-based tool for visualizing and analyzing the variation of GC content in genomic sequences. *Nucleic Acids Research*, **34**, W686–W691.

Gladman, N., Jiao, Y., Lee, Y.K., Zhang, L., Chopra, R., Regulski, M. *et al.* (2019) Fertility of pedicellate spikelets in sorghum is controlled by a jasmonic acid regulatory module. *International Journal of Molecular Sciences*, **20**, 4951.

Gladman, N., Olson, A., Wei, S., Chougule, K., Lu, Z., Tello-Ruiz, M. *et al.* (2022) SorghumBase: a web-based portal for sorghum genetic information and community advancement. *Planta*, **255**, 35.

Greene, E.A., Codomo, C.A., Taylor, N.E., Henikoff, J.G., Till, B.J., Reynolds, S.H. *et al.* (2003) Spectrum of chemically induced mutations from a large-scale reverse-genetic screen in Arabidopsis. *Genetics*, **164**, 731–740.

Guo, M., Yang, F., Liu, C., Zou, J., Qi, Z., Fotopoulos, V. *et al.* (2022) A single-nucleotide polymorphism in WRKY33 promoter is associated with the cold sensitivity in cultivated tomato. *The New Phytologist*, **236**, 989–1005.

He, F., Steige, K., Kovacova, V., Göbel, U., Bouzid, M., Keightley, P. *et al.* (2021) Cis-regulatory evolution spotlights species differences in the adaptive potential of gene expression plasticity. *Nature Communications*, **12**, 3376.

Hedden, P. (2003) The genes of the green revolution. *Trends in Genetics*, **19**, 5–9.

Ho, C.-L. & **Geisler, M.** (2019) Genome-wide computational identification of biologically significant cis-regulatory elements and associated transcription factors from rice. *Plants*, **8**, 441.

Hornung, G., **Bar-Ziv, R.**, **Rosin, D.**, **Tokuriki, N.**, **Tawfik, D.S.**, **Oren, M.** ***et al.*** (2012) Noise–mean relationship in mutated promoters. *Genome Research*, **22**, 2409–2417.

Jiang, L., **Ramamoorthy, R.**, **Ramachandran, S.** & **Kumar, P.P.** (2020) Systems metabolic alteration in a semi-dwarf rice mutant induced by OsCYP96B4 gene mutation. *International Journal of Molecular Sciences*, **21**, 1924.

Jiao, Y., **Burke, J.**, **Chopra, R.**, **Burow, G.**, **Chen, J.**, **Wang, B.** ***et al.*** (2016) A sorghum mutant resource as an efficient platform for gene discovery in grasses. *The Plant Cell*, **28**, 1551–1562.

Jiao, Y., **Burow, G.**, **Gladman, N.**, **Acosta-Martinez, V.**, **Chen, J.**, **Burke, J.** ***et al.*** (2018) Efficient identification of causal mutations through sequencing of bulked F_2 from two allelic bloomless mutants of *Sorghum bicolor*. *Frontiers in Plant Science*, **8**, 2267.

Jiao, Y., **Lee, Y.K.**, **Gladman, N.**, **Chopra, R.**, **Christensen, S.A.**, **Regulski, M.** ***et al.*** (2018) MSD1 regulates pedicellate spikelet fertility in sorghum through the jasmonic acid pathway. *Nature Communications*, **9**, 822.

Jin, J., **Tian, F.**, **Yang, D.C.**, **Meng, Y.Q.**, **Kong, L.**, **Luo, J.** ***et al.*** (2017) PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. *Nucleic Acids Research*, **45**, D1040–D1045.

Kanehisa, M., **Sato, Y.** & **Kawashima, M.** (2022) KEGG mapping tools for uncovering hidden features in biological data. *Protein Science*, **31**, 47–53.

Khan, A., **Khan, N.A.**, **Bean, S.R.**, **Chen, J.**, **Xin, Z.** & **Jiao, Y.** (2023) Variations in total protein and amino acids in the sequenced sorghum mutant library. *Plants*, **12**, 1662.

Lagerkvist, U. (1978) "Two out of three": an alternative method for codon reading. *Proceedings of the National Academy of Sciences of the United States of America*, **75**, 1759–1762.

Landry, C.R., **Lemos, B.**, **Rifkin, S.A.**, **Dickinson, W.** & **Hartl, D.L.** (2007) Genetic properties influencing the evolvability of gene expression. *Science*, **317**, 118–121.

Li, H. (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, **27**, 2987–2993.

Li, H. & **Durbin, R.** (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Li, H., **Handsaker, B.**, **Wysoker, A.**, **Fennell, T.**, **Ruan, J.**, **Homer, N.** ***et al.*** (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Li, J.Y., **Wang, J.** & **Zeigler, R.S.** (2014) The 3,000 rice genomes project: new opportunities and challenges for future rice research. *GigaScience*, **3**, 8.

Li, Y., **Chen, C.-Y.**, **Kaye, A.M.** & **Wasserman, W.W.** (2015) The identification of cis-regulatory elements: a review from a machine learning perspective. *Biosystems*, **138**, 6–17.

Lis, M. & **Walther, D.** (2016) The orientation of transcription factor binding site motifs in gene promoter regions: does it matter? *BMC Genomics*, **17**, 1–21.

Liu, B., **Fang, L.**, **Long, R.**, **Lan, X.** & **Chou, K.-C.** (2016) iEnhancer-2L: a two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition. *Bioinformatics*, **32**, 362–369.

Lopes, R.H., **Reid, I.** & **Hobson, P.R.** (2007) The two-dimensional Kolmogorov-Smirnov test. XI *International Workshop on Advanced Computing and Analysis Techniques in Physics Research*. April 23–27, 2007. Amsterdam, The Netherlands.

Lu, X., **Liu, J.**, **Ren, W.**, **Yang, Q.**, **Chai, Z.**, **Chen, R.** ***et al.*** (2018) Gene-indexed mutations in maize. *Molecular Plant*, **11**, 496–504.

Mace, E.S., **Xia, L.**, **Jordan, D.R.**, **Halloran, K.**, **Parh, D.K.**, **Huttner, E.** ***et al.*** (2008) DArT markers: diversity analyses and mapping in *Sorghum bicolor*. *BMC Genomics*, **9**, 1–11.

Massey, F.J., Jr. (1951) The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association*, **46**, 68–78.

McCormick, R.F., **Truong, S.K.**, **Sreedasyam, A.**, **Jenkins, J.**, **Shu, S.**, **Sims, D.** ***et al.*** (2018) The *Sorghum bicolor* reference genome: improved assembly, gene annotations, a transcriptome atlas, and signatures of genome organization. *The Plant Journal*, **93**, 338–354.

Mohapatra, T., **Robin, S.**, **Sarla, N.**, **Sheshashayee, M.**, **Singh, A.**, **Singh, K.** ***et al.*** (2014) EMS induced mutants of upland rice variety Nagina22: generation and characterization. *Proceedings of the Indian National Science Academy*, **80**, 163–172.

Monroe, J.G., **Srikant, T.**, **Carbonell-Bejerano, P.**, **Becker, C.**, **Lensink, M.**, **Exposito-Alonso, M.** ***et al.*** (2022) Mutation bias reflects natural selection in *Arabidopsis thaliana*. *Nature*, **602**, 101–105.

Motlhaodi, T., **Geleta, M.**, **Chite, S.**, **Fatih, M.**, **Ortiz, R.** & **Bryngelsson, T.** (2017) Genetic diversity in sorghum [*Sorghum bicolor* (L.) Moench] germplasm from Southern Africa as revealed by microsatellite markers and agro-morphological traits. *Genetic Resources and Crop Evolution*, **64**, 599–610.

Myakishev, M.V., **Khripin, Y.**, **Hu, S.** & **Hamer, D.H.** (2001) High-throughput SNP genotyping by allele-specific PCR with universal energy-transfer-labeled primers. *Genome Research*, **11**, 163–169.

Nakamura, M., **Nunoshiba, T.** & **Hiratsu, K.** (2021) Detection and analysis of UV-induced mutations in the chromosomal DNA of Arabidopsis. *Biochemical and Biophysical Research Communications*, **554**, 89–93.

Nigam, D., **Kumar, S.**, **Mishra, D.C.**, **Rai, A.**, **Smita, S.** & **Saha, A.** (2015) Synergistic regulatory networks mediated by microRNAs and transcription factors under drought, heat and salt stresses in *Oryza sativa* spp. *Gene*, **555**, 127–139.

Olatoye, M.O., **Hu, Z.**, **Maina, F.** & **Morris, G.P.** (2018) Genomic signatures of adaptation to a precipitation gradient in Nigerian sorghum. *G3: Genes, Genomes, Genetics*, **8**, 3269–3281.

Olsen, C., **Fleming, K.**, **Prendergast, N.**, **Rubio, R.**, **Emmert-Streib, F.**, **Bontempi, G.** ***et al.*** (2014) Inference and validation of predictive gene networks from biomedical literature and gene expression data. *Genomics*, **103**, 329–336.

Panja, A.S., **Bandopadhyay, B.** & **Maiti, S.** (2015) Protein thermostability is owing to their preferences to non-polar smaller volume amino acids, variations in residual physico-chemical properties and more salt-bridges. *PLoS One*, **10**, e0131495.

Paterson, A.H., **Bowers, J.E.**, **Bruggmann, R.**, **Dubchak, I.**, **Grimwood, J.**, **Gundlach, H.** ***et al.*** (2009) The *Sorghum bicolor* genome and the diversification of grasses. *Nature*, **457**, 551–556.

Peng, J., **Richards, D.E.**, **Hartley, N.M.**, **Murphy, G.P.**, **Devos, K.M.**, **Flintham, J.E.** ***et al.*** (1999) 'Green revolution' genes encode mutant gibberellin response modulators. *Nature*, **400**, 256–261.

Poli, Y., **Basava, R.K.**, **Panigrahy, M.**, **Vinukonda, V.P.**, **Dokula, N.R.**, **Voleti, S.R.** ***et al.*** (2013) Characterization of a Nagina22 rice mutant for heat tolerance and mapping of yield traits. *Rice*, **6**, 1–9.

Poursarebani, N., **Trautewig, C.**, **Melzer, M.**, **Nussbaumer, T.**, **Lundqvist, U.**, **Rutten, T.** ***et al.*** (2020) COMPOSITUM 1 contributes to the architectural simplification of barley inflorescence via meristem identity signals. *Nature Communications*, **11**, 5138.

Quinlan, A.R. & **Hall, I.M.** (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.

Renganaath, K., **Cheung, R.**, **Day, L.**, **Kosuri, S.**, **Kruglyak, L.** & **Albert, F.W.** (2020) Systematic identification of cis-regulatory variants that cause gene expression differences in a yeast cross. *Elife*, **9**, e62669.

Richardson, A.E. & **Hake, S.** (2022) The power of classic maize mutants: driving forward our fundamental understanding of plants. *The Plant Cell*, **34**, 2505–2517.

Ritter, K.B., **Jordan, D.R.**, **Chapman, S.C.**, **Godwin, I.D.**, **Mace, E.S.** & **Lynne McIntyre, C.** (2008) Identification of QTL for sugar-related traits in a sweet × grain sorghum (*Sorghum bicolor* L. Moench) recombinant inbred population. *Molecular Breeding*, **22**, 367–384.

Sasaki, A., **Ashikari, M.**, **Ueguchi-Tanaka, M.**, **Itoh, H.**, **Nishimura, A.**, **Swapan, D.** ***et al.*** (2002) A mutant gibberellin-synthesis gene in rice. *Nature*, **416**, 701–702.

Sato, Y., **Shirasawa, K.**, **Takahashi, Y.**, **Nishimura, M.** & **Nishio, T.** (2006) Mutant selection from progeny of gamma-ray-irradiated rice by DNA heteroduplex cleavage using *Brassica* petiole extract. *Breeding Science*, **56**, 179–183.

Schmitz, R.J., **Grotewold, E.** & **Stam, M.** (2022) Cis-regulatory sequences in plants: their importance, discovery, and future challenges. *The Plant Cell*, **34**, 718–741.

Sega, G.A. (1984) A review of the genetic effects of ethyl methanesulfonate. *Mutation Research/Reviews in Genetic Toxicology*, **134**, 113–142.

Seplyarskiy, V.B., **Akkuratov, E.E.**, **Akkuratova, N.**, **Andrianova, M.A.**, **Nikolaev, S.I.**, **Bazykin, G.A.** ***et al.*** (2019) Error-prone bypass of DNA lesions during lagging-strand replication is a common source of germline and cancer mutations. *Nature Genetics*, **51**, 36–41.

Shahmuradov, I.A., **Umarov, R.K.** & **Solovyev, V.V.** (2017) TSSPlant: a new tool for prediction of plant Pol II promoters. *Nucleic Acids Research*, **45**, e65.

Shannon, P., **Markiel, A.**, **Ozier, O.**, **Baliga, N.S.**, **Wang, J.T.**, **Ramage, D. *et al.*** (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, **13**, 2498–2504.

Shirasawa, K., **Hirakawa, H.**, **Nunome, T.**, **Tabata, S.** & **Isobe, S.** (2016) Genome-wide survey of artificial mutations induced by ethyl methanesulfonate and gamma rays in tomato. *Plant Biotechnology Journal*, **14**, 51–60.

Siddique, M.I., **Back, S.**, **Lee, J.-H.**, **Jo, J.**, **Jang, S.**, **Han, K. *et al.*** (2020) Development and characterization of an ethyl methane sulfonate (EMS) induced mutant population in *Capsicum annuum* L. *Plants*, **9**, 396.

Simons, J.M., **Herbert, T.C.**, **Kauffman, C.**, **Batete, M.Y.**, **Simpson, A.T.**, **Katsuki, Y. *et al.*** (2022) Systematic prediction of EMS-induced mutations in a sorghum mutant population. *Plant Direct*, **6**, e404.

Singh, M., **Bag, S.K.**, **Bhardwaj, A.**, **Ranjan, A.**, **Mantri, S.**, **Nigam, D. *et al.*** (2015) Global nucleosome positioning regulates salicylic acid mediated transcription in *Arabidopsis thaliana*. *BMC Plant Biology*, **15**, 1–21.

Sirpiö, S., **Khrouchtchova, A.**, **Allahverdiyeva, Y.**, **Hansson, M.**, **Fristedt, R.**, **Vener, A.V. *et al.*** (2008) AtCYP38 ensures early biogenesis, correct assembly and sustenance of photosystem II. *The Plant Journal*, **55**, 639–651.

Stern, D.L. & **Orgogozo, V.** (2008) The loci of evolution: how predictable is genetic evolution? *Evolution*, **62**, 2155–2177.

Sun, Z., **Fan, J.** & **Zhao, Y.** (2021) Trans-acting factors and cis elements involved in the human inactive X chromosome organization and compaction. *Genetics Research*, **2021**, 1–7.

Swinnen, G., **Goossens, A.** & **Pauwels, L.** (2016) Lessons from domestication: targeting cis-regulatory elements for crop improvement. *Trends in Plant Science*, **21**, 506–515.

Tatarinova, T., **Brover, V.**, **Troukhan, M.** & **Alexandrov, N.** (2003) Skew in CG content near the transcription start site in *Arabidopsis thaliana*. *Bioinformatics*, **19**, i313–i314.

Tetreault, H.M., **Gries, T.**, **Liu, S.**, **Toy, J.**, **Xin, Z.**, **Vermerris, W. *et al.*** (2021) The sorghum (*Sorghum bicolor*) brown midrib 30 gene encodes a chalcone isomerase required for cell wall lignification. *Frontiers in Plant Science*, **12**, 732307.

Till, B.J., **Cooper, J.**, **Tai, T.H.**, **Colowit, P.**, **Greene, E.A.**, **Henikoff, S. *et al.*** (2007) Discovery of chemically induced mutations in rice by TILLING. *BMC Plant Biology*, **7**, 1–12.

Vaser, R., **Adusumalli, S.**, **Leng, S.N.**, **Sikic, M.** & **Ng, P.C.** (2016) SIFT missense predictions for genomes. *Nature Protocols*, **11**, 1–9.

Wang, L., **Lu, Z.**, **Regulski, M.**, **Jiao, Y.**, **Chen, J.**, **Ware, D. *et al.*** (2021) BSA-seq: an interactive and integrated web-based workflow for identification of causal mutations in bulked F2 populations. *Bioinformatics*, **37**, 382–387.

Wu, X. & **Bartel, D.P.** (2017) k pLogo: positional k-mer analysis reveals hidden specificity in biological sequences. *Nucleic Acids Research*, **45**, W534–W538.

Xin, Z., **Wang, M.**, **Cuevas, H.E.**, **Chen, J.**, **Harrison, M.**, **Pugh, N.A. *et al.*** (2021) Sorghum genetic, genomic, and breeding resources. *Planta*, **254**, 114.

Xin, Z., **Wang, M.L.**, **Barkley, N.A.**, **Burow, G.**, **Franks, C.**, **Pederson, G. *et al.*** (2008) Applying genotyping (TILLING) and phenotyping analyses to elucidate gene function in a chemically induced sorghum mutant population. *BMC Plant Biology*, **8**, 103.

Yan, W., **Deng, X.W.**, **Yang, C.** & **Tang, X.** (2021) The genome-wide EMS mutagenesis bias correlates with sequence context and chromatin structure in rice. *Frontiers in Plant Science*, **12**, 579675.

Yang, N., **Xu, X.W.**, **Wang, R.R.**, **Peng, W.L.**, **Cai, L.**, **Song, J.M. *et al.*** (2017) Contributions of *Zea mays* subspecies *mexicana* haplotypes to modern maize. *Nature Communications*, **8**, 1874.

Yano, R., **Hoshikawa, K.**, **Okabe, Y.**, **Wang, N.**, **Dung, P.T.**, **Imriani, P.S. *et al.*** (2019) Multiplex exome sequencing reveals genome-wide frequency and distribution of mutations in the 'Micro-Tom' Targeting Induced Local Lesions in Genomes (TILLING) mutant library. *Plant Biotechnology*, **36**, 223–231.

Yuan, H., **Owsiany, K.**, **Sheeja, T.**, **Zhou, X.**, **Rodriguez, C.**, **Li, Y. *et al.*** (2015) A single amino acid substitution in an ORANGE protein promotes carotenoid overaccumulation in *Arabidopsis*. *Plant Physiology*, **169**, 421–431.

Zeng, D., **Liu, T.**, **Ma, X.**, **Wang, B.**, **Zheng, Z.**, **Zhang, Y. *et al.*** (2020) Quantitative regulation of Waxy expression by CRISPR/Cas9-based promoter and 5′UTR-intron editing improves grain quality in rice. *Plant Biotechnology Journal*, **18**, 2385–2387.

Zhang, J., **Wang, Y.**, **Wu, S.**, **Yang, J.**, **Liu, H.** & **Zhou, Y.** (2012) A single nucleotide polymorphism at the Vrn-D1 promoter region in common wheat is associated with vernalization response. *Theoretical and Applied Genetics*, **125**, 1697–1704.

Zhao, Y., **Hou, Y.**, **Xu, Y.**, **Luan, Y.**, **Zhou, H.**, **Qi, X. *et al.*** (2021) A compendium and comparative epigenomics analysis of cis-regulatory elements in the pig genome. *Nature Communications*, **12**, 2217.

Zheng, W., **Gianoulis, T.A.**, **Karczewski, K.J.**, **Zhao, H.** & **Snyder, M.** (2011) Regulatory variation within and between species. *Annual Review of Genomics and Human Genetics*, **12**, 327–346.

Zhou, C., **Yuan, Z.**, **Ma, X.**, **Yang, H.**, **Wang, P.**, **Zheng, L. *et al.*** (2021) Accessible chromatin regions and their functional interrelations with gene transcription and epigenetic modifications in sorghum genome. *Plant Communications*, **2**, 100140.