A High-Resolution Method for the Systematic Detection of EMS-Induced Mutations in a Sequenced Population

Jared M. Simons¹, Tim C. Herbert¹, Coleby Kauffman¹, Marc Y. Batete¹, Andrew T. Simpson¹, Yuka Katsuki¹, Dong Le¹, Danielle Amundson¹, Elizabeth M. Buescher², Clifford Weil³, Mitch Tuinstra³ and Charles Addo-Quaye^{1*}

¹Division of Natural Sciences and Mathematics, Lewis-Clark State College, 500 8th Avenue, Lewiston, ID 83501, USA

²Land O'Lakes, INC., 4001 Lexington Avenue N, Arden Hills, MN 55126, USA

³Department of Agronomy, Purdue University, 915 W. State Street, West Lafayette, IN 47907, USA

* To whom correspondence should be addressed. Tel: +1 2087922341; Fax: +1 2087922878; Email: caaddoquaye@lcsc.edu

Present Address: Charles Addo-Quaye, Division of Business and Computer Science, Lewis-Clark State College, 500 8th Avenue, Lewiston, ID 83501, USA

Present Address: Andrew T. Simpson, Department of Biological Sciences, University of Idaho, 875 Perimeter Dr, Moscow, ID 83844, USA

29 Abstract

30 The precise detection of causal DNA mutations is very crucial for forward genetic studies. Several
 31 sources of errors contribute to false-positive detections by current variant-calling algorithms, and
 32 these impact associating phenotypes with genotypes. To improve the accuracy of mutation detection
 33 we propose and implemented a high-resolution binning method for the accurate detection of likely
 34 EMS-induced mutations in a sequenced mutant population. The approach also incorporates a novel
 35 clustering algorithm for detecting likely false negatives with high accuracy. *Sorghum bicolor* is a very
 36 valuable crop species with tremendous potential for uncovering novel gene functions associated with
 37 highly desirable agronomical traits. We demonstrate the precision of the proposed method in the
 38 detection of likely EMS-induced mutations in the publicly available low-cost sequencing of the M₃
 39 generation from 600 sorghum BTx623 mutants. The method detected 3,274,606 single nucleotide
 40 polymorphisms (SNPs) of which 96% (3,141,908) were G/C to A/T DNA substitutions, as expected by
 41 EMS-mutagenesis action. We demonstrated the general applicability of the method, and showed a
 42 high concordance, 94% (3,074,759) SNPs overlap between SAMtools-based and GATK-based
 43 variant-calling algorithms. We also implemented a novel clustering algorithm which uncovered
 44 evidence for an additional 223,048 likely false-negative shared EMS-induced mutations. The final
 45 3,497,654 SNPs represents an 87% increase in SNPs detected in the previous analysis of the
 46 sorghum mutant population. Annotation of the final SNPs revealed 10,263 high impact and 136,639
 47 moderate impact SNPs, including 7,217 stop-gained mutations, and an average of 12 stop-gained
 48 mutations per mutant. We have implemented a public search database for this new genetic resource
 49 of 30,285 distinct sorghum genes containing medium or high impact EMS-induced mutations.
 50 Seedstock for a select 486 of the 600 described mutants are publicly available in the Germplasm
 51 Resources Information Network (GRIN) database.

52

53 Keywords

54 Sorghum, Variant-Calling, EMS Mutagenesis, Single Nucleotide Polymorphism, SAMtools, GATK

55

56

57

58

59 Introduction

60 Plant forward and reverse genetics studies play a vital role in harnessing the untapped potential of
 61 plants in the search for new sources of fuel energy, improving crop productivity, discovering desirable
 62 agronomic traits, and enhancing ecosystem services (1-11). Despite the tremendous progress in plant
 63 gene function studies, a significant number of genes found in plant genomes still lack experimentally
 64 determined function. The rapid advances in cost-effective, high throughput sequencing techniques
 65 such as next-generation sequencing (NGS) technologies (12,13), and modern mutation breeding
 66 techniques, such as TILLING (4), had been pivotal in recent large-scale mutant studies that have
 67 provided expansive resources for crop forward genetics (14-26). A diverse collection of physical
 68 mutagens (including fast neutrons, carbon ion beam, X-rays, gamma rays and ultraviolet light) and
 69 chemical mutagens (including ethyl methanesulfonate [EMS], (N-ethyl-N-nitrosourea [ENU], and N-
 70 methyl-N-nitrosourea [MNU]) are available for inducing mutations in living organisms (27-33). A
 71 disadvantage of traditional mutagens is the modification of random genomic positions, which results in
 72 the uncertainty of genomic consequences. Although precise genome-editing techniques are on the
 73 ascendency, their general utility is still limited (34-37), and hence traditional mutagenesis techniques
 74 still remain useful. Various variant-calling algorithms are currently available for identifying induced
 75 mutations from next-generation sequencing datasets (38-42). The detection accuracy of these *in silico*
 76 tools are impacted by the inherently higher sequencing error rate of NGS technologies and other
 77 sources of error such as reference genome mis-assembly, imprecise sequence alignments and
 78 sample and library preparation errors (43). To expose false positives, present in the initial variant-
 79 calling, various practices such as the removal of shared variants, low phred-scale variant-quality
 80 score variants, extreme strand-biased variants and excessively high-coverage depth variants had
 81 been proposed (43-48). Although these filtering approaches are very potent, a possible unintended
 82 consequence is the elimination of bona fide false-negative variants. In the era of rapid advances in
 83 high-throughput phenotyping (phenomics) for both in-field and greenhouse experiments (3, 49-55),
 84 precisely detecting true variants, while filtering false-positive variants and recovering bona fide false-
 85 negative variants would greatly improve the precision of mutation detection, and greatly enhance
 86 phenotype to genotype association (52).

87

88 *Sorghum bicolor* is a grass species which is very closely related to maize and has a relatively
 89 compact genome size of around 730 million bases (56-59). Sorghum is globally cultivated for food
 90 (grain sorghum), feed (forage sorghum), beverage and fuel (cellulosic sorghum), and is characterized
 91 by the ability to grow in nutrient-poor soils (60-67). In this article we describe and implement a high-
 92 resolution approach to detect likely EMS-induced mutations in low-coverage NGS sequencing of
 93 mutant populations. This generally applicable method leverages the well-described action of EMS
 94 mutagenesis, which results in predominantly G/C to A/T DNA transitions (68,69). We also
 95 implemented a clustering algorithm for uncovering bona fide false-negative shared variants in
 96 sequenced mutant populations. Applying the method, we describe a highly improved indexed mutant
 97 collection and genetic resources for the publicly available M₃ generation from 600 EMS-mutagenized
 98 sorghum individuals.

99

100 **Results**

101 **NGS Mapping and Initial Variant-Calling**

102 We implemented a high-resolution variant-calling and filtering pipeline to detect likely EMS-induced
 103 mutations in a sequenced mutant population (Figure 1). A total of 30,027,070,132 NGS paired-end
 104 reads were retrieved from the NCBI short reads archive (SRA) with an average of 50,045,117 reads
 105 per sample. Of these 98.3% (29,521,006,672) mapped to the current sorghum reference genome
 106 (version 3.1) and 28,132,018,400 were properly paired mapped reads (Table 1). Initial variant-calling
 107 using SAMtools detected 85,940,118 variants including 82,645,413 single nucleotide polymorphisms
 108 (SNPs). The percentage of SNPs that were G/C to A/T substitutions was 36.4% (30,096,471 SNPs;
 109 Table 2). Similarly using GATK variant-calling algorithm with single individual genotyping and GATK
 110 with joint genotyping, 89,338,788 and 106,488,567 variants were initially detected respectively, and
 111 these included 84,193,419 and 95,488,475 SNPs, respectively. The corresponding percentages of
 112 GATK SNPs that were G/C to A/T substitutions were 37.3% (single individual genotyping) and 37.9%
 113 (joint genotyping) respectively (Table 2). From here we simply refer to GATK variant-calling for single
 114 mutant individual genotyping as GATK, and GATK analysis with joint genotyping for all 600 mutant
 115 individuals as GATK-Joint.

116

117

SNPs Filtering

For independently mutagenized individuals it is highly unlikely for an EMS-induced mutation to be detected in two or more mutants. The subtraction of shared (likely false-positive) variants retained 6,373,258, 3,886,456, 3,750,600 SNPs for SAMtools, GATK and GATK-Joint, respectively. Interestingly for both GATK-based approaches 88 and 90% (3,412,446 and 3,356,788) of the retained unique (non-shared) SNPs were G/C to A/T substitutions (Table 2), while the percentage for SAMtools was only 65.4% (4,166,924 SNPs). The ranges of phred-scaled SNP quality-scores for the mutations detected were 3.0 - 228.0 (for SAMtools), 10.0 - 214,334.8 (for GATK) and 10.0 - 65,576,475.2 (for GATK-Joint). Visual inspection of the mutation spectrums for all variant-quality score bin categories (Figure 2, Supplemental Figures S1 and S2) shows that in all three approaches, the percentage of SNPs which were G/C to A/T transitions increased with increasing SNP quality-score values. As shown in Figure 2, for SAMtools, the G/C to A/T percentage showed a striking increase from 40.91% (at quality-score=11) to 63.73% (at quality-score=12). Similarly, for GATK and GATK-Joint the G/C to A/T percentages jumped to 64.04% (at quality-score=26) and 66.36% (at quality-score=28) respectively (Supplemental Figures S1 and S2). Based on these observations, the following SNP-quality score thresholds were selected for subsequent quality-score based filtering: 12 (for SAMtools), 26 (for GATK) and 28 (for GATK-Joint). Applying the selected SNP quality-score thresholds, 96% (3,141,908 out of 3,274,606) of SAMtools SNPs, 92.4% (3,248,054 out of 3,514,943) of GATK SNPs and 93.5% (3,211,794 out of 3,435,790) of GATK-Joint SNPs retained were G/C to A/T transitions (Figure 3; Table 2) and these are most likely EMS-induced mutations. Comparison of the detected likely EMS-induced SNPs showed a consensus (intersection) of 3,074,759 SNPs detected by all three approaches (Figure 4). The high concordance of SNPs between the SAMtools- and GATK-based algorithms suggests the filtering method is generally applicable. A total of 181,432 SNPs were SAMtools-specific, while 77,957 and 14,981 SNPs were respectively GATK- and GATK-Joint-specific. A total of 344,930 SNPs were detected in both GATK-based approaches but not by SAMtools (Figure 4). Examination of the GATK-based SNPs that were not detected by SAMtools, 55% (232,234 out of 422,887) of GATK SNPs and 58% (207,491 out of 359,911 SNPs) of GATK-Joint SNPs were G/C to A/T substitutions (Supplemental Table S1, Supplemental Files S1 and S2). Hence it is presumable that the vast majority of these GATK-specific SNPs may not be true EMS-induced SNPs. In our previous analysis of the 600 mutants, where the SAMtools SNP calling was based on

version 2.1 of the sorghum reference genome, we detected 1,753,403 SNPs of which 96.7% (1,695,973 SNPs) were G/C to A/T substitutions (16). Hence the 3,274,606 SNPs detected using the currently proposed method represents an 87% (1,521,203 SNPs) increase in the number of SNPs predicted. It is possible the increase in SNPs detected could be due to differences in reference genome versions used in the separate studies. As a control, we repeated the analysis using the current approach with the older reference genome version 2.1. The initial number of SAMtools SNPs predicted in the genome version 2.1 was 60,407,348 and these included 6,313,576 unique SNPs (Table 2). Similarly, at a SNP-quality score threshold of 12, 96.7% (3,182,253 out of 3,289,616 SNPs) were G/C to A/T substitutions (Figure 3; Table 2). Hence, the increase in the number of SNPs predicted is the result of a superior method and not due to an improvement in genome assembly. Finally, we investigated the mutation spectrum for SNPs associated with extreme strand-biased reads mapping at the corresponding SNP position in the genome. Of the 307,434 strand-biased SNPs where nearly all the mapped reads (95%) at the SNP position were derived entirely from only the forward or only the reverse strand, over 94% (290,110 SNPs) were G/C to A/T substitutions. (Figure 5; Supplemental Table S2; Supplemental File S3). These supposedly extreme strand-biased SNPs are most likely EMS-induced mutations lacking sufficient read depth, because of the low-coverage (6X) sequencing of the mutant genomes.

Discovery of Likely False-Negative Shared Variants

For multiple independently mutagenized individuals in a population, the pragmatic and very effective technique of subtracting shared variants could result in the unintended consequence of the removal of false-negative shared variants. The false-negative shared variants could originate from pollen contamination, sample mislabeling, and other sources of sample or library preparation errors. Using the initial unfiltered SAMtools-predicted SNPs, we categorized all SNPs based on the number of mutant individuals they were detected in, and then computed the mutation spectrum for each category. A comparison of the G/C to A/T transition percentages for the various SNPs categorizes indicated a non-trivial signal for likely EMS-induced SNPs shared by two or three mutants (Supplemental Figure S3). We next analyzed the initially subtracted shared SNPs from the earlier SNPs filtering step, for evidence of true EMS-induced mutations. Using a clustering algorithm we implemented, we clearly distinguished between localized clusters of mutant individuals with shared

mutations (likely false-negatives) and very large clusters of mutant individuals with global or near-global shared variants (likely false-positives). Overall, we identified 930,352 non-empty clusters with an average 78 mutant individuals per cluster, and an average of 1.27 shared mutations per cluster (Supplemental File S4). Ranking the clusters by the number of shared mutations within a cluster, the top 43 clusters (Figure 6) contained an average of 2,831 shared mutations, and the highest ranked cluster (*clust-39453_2*) contained 6,837 mutations shared by two mutant individuals. Most of the top-ranking clusters (41 out of 43) were of member size two (i.e., contained mutations shared by only two mutant individuals). Of the remaining two clusters, one cluster (*clust-930352_600*) contained 4,487 mutations found (or globally shared) in all 600 mutants, while the final of the top 43 clusters contained 1,238 mutations shared by a specific group of three mutants (*clust-125781_3*). Excluding the mutations from the global cluster *clust-930352_600*, the mutation spectrum for each set of recovered shared mutations for the 77 impacted mutants showed G/C to A/T percentages ranging from 96.7%-99.7% (with an average of 98.7%), and that suggests these are likely false-negative (shared) EMS-induced mutations (Figure 5). On the contrary, only 17.61% (473,979 of 2,692,200) of the SNPs originating from the 4,487 genomic positions detected in all 600 mutants were G/C to A/T substitutions. In total, we recovered a total of 235,736 likely false-negative shared SNPs originating from 117,249 distinct genomic positions, and these include 223,048 SNPs (~95%) having a SNP quality-score of at least 12 (Supplemental File S5). This shows that, of the initially subtracted shared SNPs, 35% (232,022 of 661,156) of shared SNPs found in pairs of mutants, and only 1% (3,714 of 291,966) shared SNPs initially predicted in triplet of mutants, are likely bona fide EMS-induced mutations. Including the recovered, likely false-negative shared SNPs with variant quality-score of at least 12, we retained a total of 3,497,654 likely EMS-induced SNPs for further downstream analysis. The final retained SNPs consisted of 2,157,409 homozygous and 1,340,245 heterozygous mutations. Overall, the final SAMtools-predicted SNPs represent an average of 5,829 SNPs per mutant, and about one mutation per 125kb in the sorghum BTx623 genome.

SNP Impact Annotation

Using the snpEff software, we annotated 10,263 distinct high impact SNP positions affecting 7,979 distinct sorghum genes. This includes 7,217 stop-gained (nonsense) mutations and represent averages of 17 high impact SNPs and 12 stop-gained mutations per mutant (Table 3). A total of

136,639 distinct missense SNPs were predicted to have a moderate impact on gene function in 29,835 distinct sorghum genes, and this represents an average of 228 missense mutations per mutant. A total of 83,166 SNPs were synonymous (silent) substitutions. Overall, 30,285 distinct sorghum genes were impacted by moderate-to-high impact SNPs. The remaining SNP effect annotations include 293,332 and 2,884,020 SNPs derived from introns and intergenic regions respectively (Supplemental File S6). Using the SIFT program and the UNIREF90 protein database, analysis of the non-distinct 291,043 missense SNPs (overlapping various isoforms of gene transcripts), showed 106,520 were likely deleterious, 90,892 were likely tolerated, while the remaining 93,631 were indeterminate (Supplemental File S7). The detailed annotation of all the high and medium impact SNPs, including the description of the impacted genes has been included in the analysis (Supplemental File S8).

Discussion

Although several mutant resources are currently available for sorghum genetics (15,16,78), gene function discovery in sorghum and other crop species would greatly benefit from improvements in computational techniques for the precise detection of DNA mutations⁶¹. Several techniques had been suggested for filtering false positives, and these include subtraction of shared variants, applying prescribed variant-quality score thresholds, and the removal of variant calls associated with either extreme strand-bias reads or excessive read coverage depth (43-48, 79). The inappropriate application of these best practices could result in the exclusion of bona fide DNA variants as false negatives or the retention of false-positive detections. As an example, in our previous studies, using the variant quality-score threshold of 20 in the initial filtering of SNPs showed marginal improvement (16,80,81), while in another study a quality-score threshold of 100 was effective in the detection bona fide fast neutrons-induced mutations in rice (20). In the specific case of EMS mutagenesis, the likelihood of detecting the same mutation in two or more independently mutagenized individuals in a population is very low, since EMS action impacts random genomic positions. Also, in comparison to other mutagens such as fast neutrons, EMS mutagenesis results in predominantly G/C to A/T DNA transitions (68,69) and hence is limited to a lower diversity in categories of genome target sites (16,19,20). However, the well-known EMS mutation spectrum can be leveraged in determining the suitability or effectiveness of any of the above-mentioned false-positive variant filtering techniques for

specific studies. Guided by these well-described characteristics, and the expected mutation spectrum of EMS we proposed and implemented a high-resolution method (Figure 1) for detecting induced mutations in the low-coverage NGS sequencing data for the M₃ generation from 600 EMS-mutagenized sorghum individuals (Table 1). Using the proposed method, we empirically determined the phred-scale variant quality-score threshold precisely when the retained SNPs in the mutant population were predominantly G/C to A/T substitutions (Figure 2). Using the proposed method we predicted 3,274,606 likely EMS-induced SNPs in the 600 mutant individuals (Table 2 and Figure 3) of which 96% (3,141,908) were G/C to A/T substitutions as expected by EMS action. This result represents an 87% increase (1,521,203 additional SNPs) over the previous analysis which detected 1,753,403 likely EMS-induced mutations (of which 96.7% were G/C to A/T substitutions) in the same mutant population (16). The previous study was based on analysis using the sorghum reference genome version 2.1, while the current analysis used genome version 3.1. To control for genome version being the source for the increased number of mutations predicted, we re-analyzed the entire 600 genomes by using the reference genome version 2.1. The re-analysis for reference genome version 2.1, similarly predicted 3,289,616 likely EMS-induced SNPs including 3,182,253 (96.7%) G/C to A/T substitutions (Table 2 and Figure 3). Coincidentally, the previous study using version 2.1 of the reference genome also predicted 96.7% (1,695,973 out of 1,753,403) were G/C to A/T substitutions (16). Hence the increase in number of SNPs detected is due to the proposed (improved) method and not the improved reference genome assembly version. This high-resolution analysis also suggests that using a phred-scale variant quality-score threshold of 20 for SNPs filtering in the previous study may have been too conservative and probably caused widespread false-negatives. The new findings from the proposed method were primarily based on the SAMtools variant-calling algorithm. To determine the general applicability of the proposed method, we implemented the methodology by using GATK variant-calling algorithm and re-analyzed the NGS sequencing datasets for the whole mutant population. Using GATK with single individual genotyping, 92.4% (3,248,054) of the 3,514,943 SNPs were G/C to A/T mutations (Table 2; Figure 3 and Supplemental Figure S1). Similarly, using GATK with joint genotyping 93.5% (3,211,794) of the 3,435,790 likely EMS-induced SNPs were G/C to A/T mutations (Table 2; Figure 3 and Supplemental Figure S2). A previous benchmark study using high-confident variants showed ~92% concordance among three variant callers using Illumina exome sequencing datasets (82). Our results from both GATK- and SAMtools-based algorithms also showed

a high-concordance, with 3,074,759 likely EMS-induced SNPs consistently detected in all three implementations (Figure 4). Our analysis similarly revealed a high concordance with 94% of SAMtools SNPs overlapping GATK SNPs at a whole-genome mutant population scale. Despite the high similarity in results, there was a striking marked difference in the initial filtering between SAMtools and GATK. The initial subtraction of shared variants had an immediate desirable impact on the GATK-based approaches, with 88% and 90% (3,412,446 and 3,356,788) of retained unique SNPs being G/C to A/T substitutions (Table 2). In comparison, only 65.4% (4,166,924) of the retained, non-shared SAMtools SNPs were G/C to A/T transitions (Table 2). Also, estimating the phred-scale variant quality-score threshold for SAMtools was relatively straightforward, and resulted in a slightly higher percentage of G/C to A/T substitutions (96% versus 94%). We also observed that the vast majority of GATK-specific SNPs which were not predicted by SAMtools, were less likely to be EMS-induced mutations (Supplemental Table S1). These differences in characteristics reflect the underlying assumptions in the two variant-calling algorithms (39,83). Although subtraction of shared variants is a very potent technique, it could inadvertently result in bona fide false-negative mutations being removed. Pollen contamination is quite often a source of shared variants in a sequenced population. Previous attempts to recover false-negative shared variants involves retaining shared G/C to A/T substitutions with high variant-quality score that were detected in up to two mutants (16). We examined the EMS mutation spectrums for various categories of initially predicted SNPs. We observed that among the various SNP categories the G/C to A/T substitution percentage was highest for shared SNPs detected in pairs of mutants; even higher than the percentage for unique SNPs (68 % vs 65%; Supplemental Figure S3). We implemented a clustering algorithm for recovering false-negative shared variants in the mutant population. The algorithm uncovered 42 clusters containing 235,736 likely false-negative SNPs, of which 98.7% were G/C to A/T substitutions. Using the clustering algorithm, we were able to determine that at least 232,022 of 661,156 initially subtracted (shared) SNPs originating from pairs of mutants, and only 3,714 of 291,966 initially subtracted (shared) SNPs detected in triplets of mutant individuals, are most likely bona fide EMS-induced mutations (Figures 5 and 6). Previous studies had shown SNPs associated with extreme strand-biased reads as a potential source of false-positives in Illumina NGS sequencing data (84). Next, we investigated the evidence of extreme strand-bias impact in the filtered SNPs dataset. Of the 307,434 final SNPs having least 95% of the reads derived from a single strand, 94% (290,110 SNPs) were

G/C to A/T substitutions, and most likely EMS-induced mutations (Figure 5). This suggests that for low-coverage whole-genome NGS sequencing datasets, the lack of sufficient read depth might result in artificial extreme strand-biased reads mapping at some SNP positions, this might result in these SNPs being discarded as false positives. The snpEff annotation of the final 3,497,654 SNPs predicted 10,263 distinct high-impact and 136,639 moderate impact SNPs in the sequenced mutant population (Table 3). This represents an average of 17 high impact and 228 moderate impact mutations per mutant line. Previous studies revealed that known variants that alter phenotypes usually corroborate their earlier prediction as deleterious mutations, and as such computational predicted deleterious mutations may likely alter phenotypes⁸⁵. Using the UNIREF90 reference proteome dataset, our analysis predicted 106,520 missense SNPs to be likely deleterious, 90,892 as likely tolerated, while the impact of the remaining 93,631 missense SNPs could not be reliably determined. In summary, we have created an improved large-scale mutant resource for sorghum genetics and created a light-weight web portal for querying the mutant database and seamless link to requesting seed stock for the collections in the Germplasm Resources Information Network (GRIN) database.

Materials and Methods

NGS Sequence Retrieval and Reads Mapping

The Illumina next-generation sequencing data for the M₃ generation from 600 EMS-mutagenized sorghum mutants were retrieved from the NCBI Sequence Read Archive (SRA Accession Number SRP065118) by using the *fastq-dump* program (version 2.9.0) with non-default parameters: "--split-3" (70,71). We retrieved the reference genome sequences and annotations for the sorghum BTx623 genome assembly versions 2.1 and 3.1, from the Phytozome web portal (72). Prior to NGS reads mapping, the reference genome sequences were indexed using *index* command in the BWA short reads aligner (version 0.7.15) (73). The NGS reads for each sequenced sample were mapped to the sorghum reference genome using the BWA *mem* command with non-default parameter: "-t 64".

SAMtools-Based Variant-Calling

The SAMtools variant-calling was performed using version 1.8 of the software package (38). Prior to variant-calling, the sorghum reference genome sequences were indexed using the SAMtools *faidx* command with default parameters. The BWA-mapped NGS reads were sorted using the SAMtools *sort* command with non-default parameters: "-@ 64 -m 2G -O bam -T". Duplicate NGS reads were identified by using GATK *MarkDuplicates* command (version 4.0.5.0) with default parameters (41). The resulting BAM-formatted files were indexed using the *index* command with non-default parameters: "-b -@ 64". The genotype likelihoods were computed using the BCFtools *mpileup* command (version 1.8) with non-default parameter: "-Ou". Final variant detection was performed using the BCFtools *call* command with non-default parameters: "-vmO z". Based on the SAMtools zygosity call encoding in the output variant call files SNPs were characterized as either homozygous or heterozygous.

GATK-Based Per-Sample Single Individual Variant-Calling

For the GATK-based variant-calling the sorghum reference genome sequences were indexed using the GATK *CreateSequenceDictionary* command (version 4.0.5.0) with default parameters (39). Using the above BWA-sorted and indexed NGS reads, the initial GATK single sample variant-calling for single individuals was performed by invoking the GATK *HaplotypeCaller* with non-default parameters: '--java-options "-Xmx4g" HaplotypeCaller'.

GATK-based Joint Genotyping

For joint genotyping with GATK, the initial steps were similar to the above described GATK approach for per-sample individual variant calling, with the exception of invoking the GATK *HaplotypeCaller* with non-default parameters: '--java-options "-Xmx4g" HaplotypeCaller -ERC GVCF'. Next, we created two additional input files: the first was the map file which contained the sequenced sample ID names and their corresponding individual genotyping call filenames. The second file contained the list of chromosome and super scaffold names extracted from the reference genome sequence assembly file. Specifying these two input files, we executed the GATK *GenomicsDBImport* command with non-default parameters: '--java-options "-Xmx4g" --batch-size 100 --reader-threads 50' to merge the individual sample variant call files, which were then subsequently imported into a *GenomicsDB*

workspace. Joint genotype calling was performed using the GATK *GenotypeGVCFs* command with non-default parameters: "-G StandardAnnotation -new-qual".

Variants Filtering

For both SAMtools-based and GATK-based approaches, all the initially predicted SNPs for the 600 mutants were combined and sorted by their genomic positions. Using this information all variants detected at the same genomic position in multiple mutant individuals were subtracted (44). We calculated the mutation spectrum for the retained unique SNPs to check the effect of the likely false-positive subtraction. We next empirically determined the phred-scaled variant-quality score threshold for further SNPs filtering, by using a high-resolution binning method as follows. All the retained SNPs from the shared variants subtraction were sorted into bin categories corresponding to their respective variant-quality scores, and as such SNPs with the same quality-score were assigned to the same bin category. With emphasis on the percentage of mutations that were G/C to A/T substitutions, we plotted the mutation spectrum for each quality-score bin category. Visual inspection by starting from the lowest to the highest phred-scaled quality-score bin category, we determined the threshold variant quality-score, at which the percentage of SNPs that were G/C to A/T transitions approached the expected results of EMS action. The subset of SNPs with variant quality-score less than the threshold score were filtered. We separately estimated the SNPs quality-score thresholds for the SAMtools- and GATK-based variant-calling results.

Detection of Likely False-Negative Variants

We implemented a clustering algorithm for detecting and recovering likely false-negative shared variants that were subtracted in the initial SNPs filtering step. We hypothesize that if a set of mutants are genetically related, then a sufficiently large number of (shared) mutations would be detected exclusively in the specific set of mutants and absent in the remaining unrelated mutants. All the initial SAMtools SNP calls in each variant call file were tagged (prefixed) with their respective sample ID names. Next, the sample ID-tagged SNPs from all 600 mutants were combined into a single file. The combined file was sorted by using the genome positions of the SNPs, and the sorted file was used in creating a summary table including the following three column entries: *genome position*, *number of samples*, and *list of sample IDs*. For example, the table entry "Chr04:23080886 3

SRR2759335|SRR2759504|SRR2759646" means the SNP at genome position 23,080,886 on chromosome 4, was detected in three mutants with sample ID names *SRR2759335*, *SRR2759504*, and *SRR2759646*. Each grouping of sample IDs on each line in the summary table was treated as a candidate cluster of related mutants. For each candidate cluster, we extend the search by examining the rest of the file to count how many times the same n-tuple of mutants occur exclusively together (i.e., determine the remaining set of mutations that are shared exclusively by the specific set of mutants). For each candidate cluster we recorded the number of shared mutations and the percentage of mutations that were G/C to A/T substitutions. Next, the detected clusters were sorted and ranked based on their number of shared mutations. The top-ranked clusters with markedly high number of shared mutation counts were retained and their corresponding percentages of G/C to A/T substitution was used as a benchmark evidence of likely EMS-induced origins. The identified, likely false negatives were then added to the unique (non-shared) filtered SNPs. For each affected mutant individual, the merging of the likely false-negatives and the likely true-positives was accomplished by appending the shared variants file to their corresponding filtered unique variants file and re-sorting the resulting variant call file by using the BEDTools *sort* command (version v2.26.0) with the non-default parameter "-faidx chromNames" (74).

SNP Impact Annotation

We used the snpEff variant annotation package version 4.3t (75) to predict the impact of the detected SNPs on gene function. We created a custom snpEff database for the sorghum reference genome version 3.1 as follows. We edited the snpEff configuration file to include an entry for sorghum genome version 3.1. Next, we added the sorghum genome reference sequences, genome annotation .gff files and the protein sequences to the required subfolders in the snpEff package, and we run the snpEff *build* command with non-default option "-gff3 -v Sbicolor_454_v3.0.1". Using the variant call files as input, the variant impact annotation was performed by using the snpEff *eff* command.

SIFT Annotation of Missense SNP Effect

For the final set of retained SNPs we selected all missense substitutions, including the genome position and the DNA base substitution information from their corresponding variant call files. Next, we retrieved the subset of sorghum protein sequences which contained the predicted missense

mutations. We next downloaded the UNIREF90 set of protein sequences (109,653,977 sequences in release version 2020_02; 22-Apr-2020) from the UniProt web portal (76). Using the impacted sorghum protein sequences and the UNIREF90 FASTA sequences as query and database options respectively, we used the SIFT4G algorithm (77) with non-default parameter "-t 64", to predict the likelihood of the missense substitutions having a tolerated or deleterious effect in the genome. Based on the SIFT analysis results, each missense variant was assigned one of the following designations: *TOLERATED*, *DELETERIOUS* or *UNKNOWN*. For all SNPs with predicted moderate or high impact on gene function, we appended the SIFT predictions, and the gene function description from the sorghum reference genome version 3.1.

List of abbreviations

SNPs: Single Nucleotide Polymorphisms
GRIN: Germplasm Resources Information Network
NGS: Next-Generation Sequencing
EMS: Ethyl Methanesulfonate
ENU: N-ethyl-N-nitrosourea
MNU: N-methyl-N-nitrosourea
SRA: Short Reads Archive

Declarations

Ethics approval and consent to participate

Not Applicable

Consent for publication

Not Applicable

Availability of data and materials

All data generated or analyzed during this study are included in this published article (and its supplementary information files).

The NGS datasets for the sequenced sorghum mutants are available in the NCBI SRA repository (Study Accession Number: SRP065118; BioProject ID: PRJNA297450).

Supplemental data files are available via the DRYAD repository

(<https://doi.org/10.5061/dryad.hmgqnk9hj>).

The described sorghum mutations can be accessed at the following link:

(<http://heliamphora.lcsc.edu/sorghumgenetics2.0/>)

Competing interests

Not Applicable

Funding

This work was supported by an Institutional Development Award (IDeA) from the National Institute of General Medical Sciences of the National Institutes of Health [P20GM103408 to C.A.-Q]; and the Idaho Higher Education Research Council (HERC).

Authors' contributions

C.A.-Q, C.W, M.T and E.M.B designed the experiments. J.M.S, M.Y.B, A.T.S, D.A and C.A.-Q, performed the variant-calling and annotation. T.C.H implemented the webserver and Y.K implemented the databases. C.K, D.L and C.A.-Q performed the clustering analysis. C.A.-Q drafted the manuscript. C.A.-Q, E.M.B, C.W, and M.T edited the manuscript.

Acknowledgements

The authors would like to thank the following Lewis-Clark State College administrators for their research support: Jane Finan, Matthew Johnston, Heather Henson-Ramsey, Jerry Hindberg, Lori Stinson, and Martin Gibbs.

References

1. Doebley, J. F., Gaut, B. S. & Smith, B. D. The molecular genetics of crop domestication. *Cell* **127**, 1309–21 (2006).
2. Fernie, A. R. & Yan, J. De Novo Domestication: An Alternative Route toward New Crops for the Future. *Mol. Plant* **12**, 615–631 (2019).
3. Olsen, K. M. & Wendel, J. F. A bountiful harvest: genomic insights into crop domestication phenotypes. *Annu. Rev. Plant Biol.* **64**, 47–70 (2013).
4. McCallum, C. M., Comai, L., Greene, E. a & Henikoff, S. Targeting induced local lesions IN genomes (TILLING) for plant functional genomics. *Plant Physiol.* **123**, 439–442 (2000).

- 476 5. Meyer, R. S. & Purugganan, M. D. Evolution of crop species: genetics of domestication and
477 diversification. *Nat. Rev. Genet.* **14**, 840–52 (2013).
- 478 6. Mickelbart, M. V, Hasegawa, P. M. & Bailey-Serres, J. Genetic mechanisms of abiotic stress
479 tolerance that translate to crop yield stability. *Nat. Rev. Genet.* **16**, 237–251 (2015).
- 480 7. Li, X. *et al.* A fast neutron deletion mutagenesis-based reverse genetics system for plants.
481 *Plant J.* **27**, 235–42 (2001).
- 482 8. Schneeberger, K. & Weigel, D. Fast-forward genetics enabled by new sequencing
483 technologies. *Trends Plant Sci.* **16**, 282–8 (2011).
- 484 9. Schneeberger, K. Using next-generation sequencing to isolate mutant genes from forward
485 genetic screens. *Nat. Rev. Genet.* **15**, 662–676 (2014).
- 486 10. Hartwig, B., James, G. V., Konrad, K., Schneeberger, K. & Turck, F. Fast isogenic mapping-
487 by-sequencing of ethyl methanesulfonate-induced mutant bulks. *Plant Physiol.* **160**, 591–600
488 (2012).
- 489 11. Grierson, C. S. *et al.* One hundred important questions facing plant science research. *New*
490 *Phytol.* **192**, 6–12 (2011).
- 491 12. Metzker, M. L. Sequencing technologies - the next generation. *Nat. Rev. Genet.* **11**, 31–46
492 (2010).
- 493 13. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: Ten years of next-
494 generation sequencing technologies. *Nat. Rev. Genet.* **17**, 333–351 (2016).
- 495 14. Chia, J.-M. *et al.* Maize HapMap2 identifies extant variation from a genome in flux. *Nat. Genet.*
496 **44**, 803–7 (2012).
- 497 15. Jiao, Y. *et al.* A Sorghum Mutant Resource as an Efficient Platform for Gene Discovery in
498 Grasses. *Plant Cell* **28**, tpc.00373.2016 (2016).
- 499 16. Addo-Quaye, C., Tuinstra, M., Carraro, N., Weil, C. & Dilkes, B. P. Whole Genome Sequence
500 Accuracy Is Improved by Replication in a Population of Mutagenized Sorghum. *G3*
501 *Genes|Genomes|Genetics* **8**, 1079–1094 (2018).
- 502 17. Szurman-Zubrzycka, M. E. *et al.* HorTILLUS—a rich and renewable source of induced
503 mutations for forward/reverse genetics and pre-breeding programs in barley (*Hordeum vulgare*
504 L.). *Front. Plant Sci.* **9**, 1–16 (2018).

- 505 18. Cooper, J. L. *et al.* TILLING to detect induced mutations in soybean. *BMC Plant Biol.* **8**, 9
506 (2008).
- 507 19. Henry, I. M. *et al.* Efficient Genome-Wide Detection and Cataloging of EMS-Induced Mutations
508 Using Exome Capture and Next-Generation Sequencing. *Plant Cell* **26**, 1382–1397 (2014).
- 509 20. Li, G. *et al.* The sequences of 1504 mutants in the model rice variety Kitaake facilitate rapid
510 functional genomic studies. *Plant Cell* **29**, 1218–1231 (2017).
- 511 21. Bolon, Y.-T. *et al.* Phenotypic and genomic analyses of a fast neutron mutant population
512 resource in soybean. *Plant Physiol.* **156**, 240–253 (2011).
- 513 22. Krasileva, K. V. *et al.* Uncovering hidden variation in polyploid wheat. *Proc. Natl. Acad. Sci.*
514 **114**, E913–E921 (2017).
- 515 23. Lethin, J. *et al.* Development and characterization of an EMS-mutagenized wheat population
516 and identification of salt-tolerant wheat lines. *BMC Plant Biol.* **20**, 1–15 (2020).
- 517 24. Till, B. J. *et al.* Discovery of induced point mutations in maize genes by TILLING. *BMC Plant*
518 *Biol.* **4**, 12 (2004).
- 519 25. Talamè, V. *et al.* TILLMore, a resource for the discovery of chemically induced mutants in
520 barley. *Plant Biotechnol. J.* **6**, 477–485 (2008).
- 521 26. Caldwell, D. G. *et al.* A structured mutant population for forward and reverse genetics in Barley
522 (*Hordeum vulgare* L.). *Plant J.* **40**, 143–150 (2004).
- 523 27. Waugh, R., Leader, D. J., McCallum, N. & Caldwell, D. Harvesting the potential of induced
524 biological diversity. *Trends Plant Sci.* **11**, 71–9 (2006).
- 525 28. Westergaard, M. Chemical mutagenesis in relation to the concept of the gene. *Experientia* **13**,
526 224–234 (1957).
- 527 29. Loveless, a. Increased rate of plaque-type and host-range mutation following treatment of
528 bacteriophage in vitro with ethyl methane sulphonate. *Nature* **181**, 1212–3 (1958).
- 529 30. Giles, N. The effect of fast neutrons on the chromosomes of *Tradescantia*. *Proc. Natl. Acad.*
530 *Sci. U. S. A.* **26**, 567 (1940).
- 531 31. Mikaelsen, K., Kiss, I. & Osone, K. Some effects of fast neutrons and gamma radiation on rice.
532 *Neutron Irradiat. Seeds II, IAEA, Vienna* **10**, 49–54 (1968).
- 533 32. Nagai, M. & Locher, G. L. Production of mutations by neutrons. *Nature* **140**, 111–112 (1937).

- 534 33. Sax, K. An analysis of X-ray induced chromosomal aberrations in *Tradescantia*. *Genetics* **25**,
535 41 (1940).
- 536 34. Van de Wiel, C. C. M., Schaart, J. G., a.P. Lotz, L. & Smulders, M. J. M. New traits in crops
537 produced by genome editing techniques based on deletions. *Plant Biotechnol. Rep.* **11**, 1–8
538 (2017).
- 539 35. Anzalone, A. V., Koblan, L. W. & Liu, D. R. Genome editing with CRISPR–Cas nucleases,
540 base editors, transposases and prime editors. *Nat. Biotechnol.* **38**, 824–844 (2020).
- 541 36. Lemmon, Z. H. *et al.* Rapid improvement of domestication traits in an orphan crop by genome
542 editing. *Nat. Plants* **4**, 766–770 (2018).
- 543 37. Biswas, S., Zhang, D. & Shi, J. CRISPR/Cas systems: opportunities and challenges for crop
544 breeding. *Plant Cell Rep.* **40**, 979–998 (2021).
- 545 38. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–9
546 (2009).
- 547 39. DePristo, M. a *et al.* A framework for variation discovery and genotyping using next-generation
548 DNA sequencing data. *Nat. Genet.* **43**, 491–8 (2011).
- 549 40. Koboldt, D. C. *et al.* VarScan 2 : Somatic mutation and copy number alteration discovery in
550 cancer by exome sequencing. 1–9 (2012). doi:10.1101/gr.129684.111.22
- 551 41. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-
552 generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
- 553 42. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. *arXiv*
554 *Prepr. arXiv1207.3907* (2012).
- 555 43. Robasky, K., Lewis, N. E. & Church, G. M. The role of replicates for error mitigation in next-
556 generation sequencing. *Nat. Rev. Genet.* **15**, 56–62 (2014).
- 557 44. Uchida, N., Sakamoto, T., Kurata, T. & Tasaka, M. Identification of EMS-induced causal
558 mutations in a non-reference *Arabidopsis thaliana* accession by whole genome sequencing.
559 *Plant Cell Physiol.* **52**, 716–22 (2011).
- 560 45. Flibotte, S. *et al.* Whole-genome profiling of mutagenesis in *Caenorhabditis elegans*. *Genetics*
561 **185**, 431–441 (2010).

- 562 46. Nordström, K. J. V *et al.* Mutation identification by direct comparison of whole-genome
563 sequencing data from mutant and wild-type individuals using k-mers. *Nat. Biotechnol.* **31**, 325–
564 30 (2013).
- 565 47. Sarin, S. *et al.* Analysis of multiple ethyl methanesulfonate-mutagenized *Caenorhabditis*
566 *elegans* strains by whole-genome sequencing. *Genetics* **185**, 417–30 (2010).
- 567 48. Der Auwera, G. A. *et al.* From FastQ data to high-confidence variant calls: the genome
568 analysis toolkit best practices pipeline. *Curr. Protoc. Bioinforma.* **43**, 10–11 (2013).
- 569 49. Houle, D., Govindaraju, D. R. & Omholt, S. Phenomics: the next challenge. *Nat. Rev. Genet.*
570 **11**, 855–66 (2010).
- 571 50. White, J. W. *et al.* Field-based phenomics for plant genetics research. *F. Crop. Res.* **133**, 101–
572 112 (2012).
- 573 51. Furbank, R. T. & Tester, M. Phenomics--technologies to relieve the phenotyping bottleneck.
574 *Trends Plant Sci.* **16**, 635–44 (2011).
- 575 52. Araus, J. L., Kefauver, S. C., Zaman-Allah, M., Olsen, M. S. & Cairns, J. E. Translating high-
576 throughput phenotyping into genetic gain. *Trends Plant Sci.* **23**, 451–466 (2018).
- 577 53. Bolger, A. M. *et al.* Computational aspects underlying genome to phenome analysis in plants.
578 *Plant J.* **97**, 182–198 (2019).
- 579 54. Tardieu, F., Cabrera-Bosquet, L., Pridmore, T. & Bennett, M. Plant Phenomics, From Sensors
580 to Knowledge. *Curr. Biol.* **27**, R770–R783 (2017).
- 581 55. Weischenfeldt, J., Symmons, O., Spitz, F. & Korbel, J. O. Phenotypic impact of genomic
582 structural variation: insights from and for human disease. *Nat. Rev. Genet.* **14**, 125–38 (2013).
- 583 56. Paterson, A. H. *et al.* The Sorghum bicolor genome and the diversification of grasses. *Nature*
584 **457**, 551–6 (2009).
- 585 57. Price, H. J. *et al.* Genome evolution in the genus Sorghum (Poaceae). *Ann. Bot.* **95**, 219–27
586 (2005).
- 587 58. Paterson, A. H. Genomics of sorghum. *Int. J. Plant Genomics* **2008**, (2008).
- 588 59. Swigoňová, Z. *et al.* Close split of sorghum and maize genome progenitors. *Genome Res.* **14**,
589 1916–1923 (2004).
- 590 60. Zheng, L.-Y. *et al.* Genome-wide patterns of genetic variation in sweet and grain sorghum
591 (*Sorghum bicolor*). *Genome Biol.* **12**, R114 (2011).

- 592 61. Boyles, R. E., Brenton, Z. W. & Kresovich, S. Genetic and genomic resources of sorghum to
593 connect genotype with phenotype in contrasting environments. *Plant J.* **97**, 19–39 (2019).
- 594 62. Scully, E. D. *et al.* Characterization of novel Brown midrib 6 mutations affecting lignin
595 biosynthesis in sorghum. *J. Integr. Plant Biol.* (2015). doi:10.1111/jipb.12375
- 596 63. Pedersen, J. F., Bean, S. R., Graybosch, R. a., Park, S. H. & Tilley, M. Characterization of
597 waxy grain sorghum lines in relation to granule-bound starch synthase. *Euphytica* **144**, 151–
598 156 (2005).
- 599 64. Wu, Y., Yuan, L., Guo, X., Holding, D. R. & Messing, J. Mutation in the seed storage protein
600 kafirin creates a high-value food trait in sorghum. *Nat. Commun.* **4**, 2217 (2013).
- 601 65. Xin, Z., Wang, M. L., Burow, G. & Burke, J. An Induced Sorghum Mutant Population Suitable
602 for Bioenergy Research. *BioEnergy Res.* **2**, 10–16 (2009).
- 603 66. Dillon, S. L. *et al.* Domestication to crop improvement: Genetic resources for Sorghum and
604 Saccharum (Andropogoneae). *Ann. Bot.* **100**, 975–989 (2007).
- 605 67. Griebel, S. *et al.* Mutations in sorghum SBEIIb and SSIIa affect alkali spreading value, starch
606 composition, thermal properties and flour viscosity. *Theor. Appl. Genet.* **132**, 3357–3374
607 (2019).
- 608 68. Loveless, A. Possible relevance of O-6 alkylation of deoxyguanosine to the mutagenicity and
609 carcinogenicity of nitrosamines and nitrosamides. *Nature* **223**, 206–207 (1969).
- 610 69. Greene, E. a *et al.* Spectrum of chemically induced mutations from a large-scale reverse-
611 genetic screen in Arabidopsis. *Genetics* **164**, 731–40 (2003).
- 612 70. Kodama, Y., Shumway, M. & Leinonen, R. The Sequence Read Archive: explosive growth of
613 sequencing data. *Nucleic Acids Res.* **40**, D54–D56 (2012).
- 614 71. Leinonen, R., Sugawara, H., Shumway, M. & Collaboration, I. N. S. D. The sequence read
615 archive. *Nucleic Acids Res.* **39**, D19–D21 (2010).
- 616 72. Goodstein, D. M. *et al.* Phytozome: a comparative platform for green plant genomics. *Nucleic
617 Acids Res.* **40**, D1178–86 (2012).
- 618 73. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform.
619 *Bioinformatics* **25**, 1754–1760 (2009).
- 620 74. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic
621 features. *Bioinformatics* **26**, 841–842 (2010).

- 622 75. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide
623 polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2;
624 iso-3. *Fly (Austin)*. **6**, 80–92 (2012).
- 625 76. Consortium, T. U. Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res.*
626 **42**, D191–8 (2014).
- 627 77. Vaser, R., Adusumalli, S., Leng, S. N., Sikic, M. & Ng, P. C. SIFT missense predictions for
628 genomes. *Nat. Protoc.* **11**, 1–9 (2016).
- 629 78. Xin, Z. *et al.* Applying genotyping (TILLING) and phenotyping analyses to elucidate gene
630 function in a chemically induced sorghum mutant population. *BMC Plant Biol.* **8**, 103 (2008).
- 631 79. James, G. V. *et al.* User guide for mapping-by-sequencing in Arabidopsis. *Genome Biol.* **14**,
632 R61 (2013).
- 633 80. Addo-Quaye, C. *et al.* Forward Genetics by Sequencing EMS Variation-Induced Inbred Lines.
634 *G3 Genes|Genomes|Genetics* **7**, 413–425 (2017).
- 635 81. Thapa, R. *et al.* New Alleles of FAD3A Lower the Linolenic Acid Content of Soybean Seeds.
636 *Crop Science*, **58**, 713–718 (2018).
- 637 82. Hwang, S., Kim, E., Lee, I. & Marcotte, E. M. Systematic comparison of variant calling
638 pipelines using gold standard personal exome variants. *Sci. Rep.* **5**, 1–8 (2015).
- 639 83. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and
640 population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–93
641 (2011).
- 642 84. Guo, Y. *et al.* The effect of strand bias in Illumina short-read sequencing data. *BMC Genomics*
643 **13**, 1–11 (2012).
- 644 85. Kono, T. J. Y. *et al.* The Role of Deleterious Substitutions in Crop Genomes. *Mol. Biol. Evol.*
645 **33**, 2307–2317 (2016).

646

647

648

649

650

651

652 Figures and Figure legends

653
654 **Figure 1.** The proposed high-resolution computational pipeline for the detection of likely EMS-induced
655 variants in the mutant population.

656
657
658
659
660
661

Figure 2. The binning approach to computing the variant quality-score filtering threshold for the initial SAMtools-predicted mutations in the 600 sorghum BTx623 mutants. The initially predicted SNPs are assigned to bin categories corresponding to their variant-quality score. The EMS mutation spectrum for all SNPs in each bin category is calculated. At a variant-quality threshold of 12, the percentages of non-G/C to A/T DNA substitutions decrease drastically, while G/C to A/T mutations become the dominant mutation type.

Figure 3. The mutation spectrums for the likely EMS-induced SNPs detected in the 600 sorghum BTx623 mutants. Using the SAMtools variant-caller and sorghum reference genome versions 2.1 (shown in above figure as Sb_v2.1) and 3.1, 96.7% and 96% of detected SNPs were G/C to A/T mutations, respectively. Using sorghum reference genome version 3.1 and GATK with single individual genotyping and joint-genotyping 92.4% and 93.5% of SNPs were G/C to A/T mutations, respectively.

Figure 4. The high concordance of high-quality SNPs detected by the different variant-calling approaches, in the 600 sorghum mutants. A total of 3,274,606, 3,514,943 and 3,435,790 filtered SNPs were predicted by SAMtools, GATK (using single individual genotyping), and GATK (using joint genotyping; GATK-Joint) respectively, and 3,074,759 of these SNPs were found by all three approaches.

Figure 5. The mutation spectrums for the SAMtools-predicted extreme strand-biased SNPs and the likely false-negative shared SNPs detected by the clustering algorithm. Almost 99% (232,662) of the 235,736 likely false-shared SNPs and 94% (290,110 SNPs) of the 307,434 extreme strand-biased SNPs were G/C to A/T substitutions, respectively.

Figure 6. The detection of groups of mutants with shared variants using a clustering algorithm. A total of 930,352 clusters were identified, and these contained an average 78 mutant individuals per a cluster, and an average of 1.3 shared mutation positions per cluster. The top 43 clusters contained an average of 2,831 shared mutation positions, with 41 of the clusters defined by only pairs of mutant individuals. The number of mutant individuals in a cluster is appended to the end of the cluster ID name. Cluster clust-930352_600 (shown in red box) contained 4,487 mutation positions shared by all 600 mutants. Cluster clust-125781_3 (shown in red box) also contained 1,238 mutations shared by three mutants.

718 Tables

Table 1. Illumina NGS paired-end reads mapping statistics for the 600 EMS-mutagenized sorghum BTx623 individuals.

	Total	Mean
All NGS Reads	30,027,070,132	50,045,117
Mapped Reads	29,521,006,672	49,201,678
Properly Paired Reads	28,132,018,400	46,886,697
Number of Mutants	600	-

Table 2. Summary variant-filtering statistics for the 600 EMS-mutagenized sorghum BTx623 individuals.

	SAMtools (Sb_v2.1) ^a		SAMtools		GATK-Single ^b		GATK-Joint ^c	
	Total	GCAT(%) ^d	Total	GCAT(%)	Total	GCAT(%)	Total	GCAT(%)
Initial Variants	65,384,440	33.7	85,940,118	36.4	89,338,788	37.3	106,488,567	37.9
Initial SNPs	60,407,348	33.7	82,645,413	36.4	84,193,419	37.3	95,488,475	37.9
Unique SNPs^e	6,313,576	65.9	6,373,258	65.4	3,886,456	87.8	3,750,600	89.5
Filtered SNPs^f	3,289,616	96.7	3,274,606	95.9	3,514,942	92.4	3,435,789	93.5

^aSAMtools SNPs-calling using Sorghum BTx623 reference genome version 2.1.

^bGATK SNPs-calling with single individual genotyping.

^cGATK SNPs-calling with joint genotyping.

^dPercentage of SNPs which are G/C to A/T DNA substitutions.

^eSubset of SNPs detected only once in the entire mutant population.

^fSNPs retained after SNP quality-score filtering.

Table 3. Summary predicted impact of likely EMS-induced SNPs on gene function for the 600 sorghum BTx623 mutants.

SNP Impact Type	Total	Mean
High Impact	10,263	17
Stop Gained	7,217	12
Splice Site Donor	1,370	2
Splice Site Acceptor	1,432	2
Stop Lost	19	0
Start Lost	226	0
Moderate Impact	136,639	228
Missense	136,639	228
Low Impact	98,623	164
Silent	83,166	139
Start Gained	7,491	12
Intron variant	293,332	489
Intergenic Region	2,884,020	4,807
Modifier	3,417,516	5,696

Supplemental Information

Supplemental Figure S1. The binning approach to computing the variant quality-score filtering threshold for the initial GATK-predicted single individual genotyping mutations in the 600 sorghum BTx623 mutants. The initially predicted SNPs are assigned to bin categories corresponding to their variant-quality score. The EMS mutation spectrum for all SNPs in each bin category is calculated. At a variant-quality threshold of 26, the percentages of non-G/C to A/T DNA substitutions decrease drastically, while G/C to A/T mutations become the dominant mutation type.

Supplemental Figure S2. The binning approach to computing the variant quality-score filtering threshold for the initial GATK-predicted joint genotyping mutations in the 600 sorghum BTx623 mutants. The initially predicted SNPs are assigned to bin categories corresponding to their variant-quality score. The EMS mutation spectrum for all SNPs in each bin category is calculated. At a variant-quality threshold of 28, the percentages of non-G/C to A/T DNA substitutions decrease drastically, while G/C to A/T mutations become the dominant mutation type.

Supplemental Figure S3. Categories of initially predicted SNPs and their corresponding mutation spectrums. The mutation categories are based on the number of times a mutation is detected in the mutant population. For example, category 1 and category 2 refer to mutations detected once and twice, respectively in the mutant population. The percentage of G/C to A/T DNA substitutions for category 2 is unexpectedly higher than category 1.

Supplemental Table S1. The mutation spectrums for the final filtered GATK-specific SNPs that were not detected by SAMtools.

1 Supplemental Information

2 Supplemental Figures

3

4 **Supplemental Figure S1.** The binning approach to computing the variant quality-score filtering

5 threshold for the initial GATK-predicted single individual genotyping mutations in the 600 sorghum

6 BTx623 mutants. The initially predicted SNPs are assigned to bin categories corresponding to their

7 variant-quality score. The EMS mutation spectrum for all SNPs in each bin category is calculated.

8 At a variant-quality threshold of 26, the percentages of non-G/C to A/T DNA substitutions decrease

9 drastically, while G/C to A/T mutations become the dominant mutation type.

10

11 **Supplemental Figure S2.** The binning approach to computing the variant quality-score filtering
12 threshold for the initial GATK-predicted joint genotyping mutations in the 600 sorghum BTx623
13 mutants. The initially predicted SNPs are assigned to bin categories corresponding to their variant-
14 quality score. The EMS mutation spectrum for all SNPs in each bin category is calculated. At a
15 variant-quality threshold of 28, the percentages of non-G/C to A/T DNA substitutions decrease
16 drastically, while G/C to A/T mutations become the dominant mutation type.

17

18

19 **Supplemental Figure S3.** Categories of initially predicted SNPs and their corresponding mutation
 20 spectrums. The mutation categories are based on the number of times a mutation is detected in the
 21 mutant population. For example, category 1 and category 2 refer to mutations detected once and
 22 twice, respectively in the mutant population. The percentage of G/C to A/T DNA substitutions for
 23 category 2 is unexpectedly higher than category 1.

24

25 Supplemental Tables

Supplemental Table S1. The mutation spectrums for the final filtered GATK-specific SNPs that were not detected by SAMtools.

	GATK-Single ^a	GATK-Joint ^b
G/C > A/T	232,234	207,491
G/C > T/A	32,465	26,505
G/C > C/G	22,990	18,618
A/T > G/C	74,779	58,456
A/T > C/G	21,043	16,461
A/T > T/A	39,376	32,380

^aGATK SNPs-calling with single individual genotyping.

^bGATK SNPs-calling with joint genotyping.

26