PLOS ONE

RESEARCH ARTICLE

# *De novo* identification and targeted sequencing of SSRs efficiently fingerprints *Sorghum bicolor* sub-population identity

**John P. Baggett[1], Richard L. Tillett[2], Elizabeth A. Cooper[3], Melinda K. Yerka[4]***

**1** Department of Biochemistry and Molecular Biology, University of Nevada, Reno, NV, United States of America, **2** Nevada Center for Bioinformatics, University of Nevada, Reno, NV, United States of America, **3** Department of Bioinformatics and Genomics, University of North Carolina at Charlotte, Charlotte, NC, United States of America, **4** Department of Agriculture, Veterinary, and Rangeland Sciences, University of Nevada, Reno, NV, United States of America

* myerka@unr.edu





**Citation:** Baggett JP, Tillett RL, Cooper EA, Yerka MK (2021) *De novo* identification and targeted sequencing of SSRs efficiently fingerprints *Sorghum bicolor* sub-population identity. PLoS ONE 16(3): e0248213. https://doi.org/10.1371/journal.pone.0248213

**Editor:** David D. Fang, USDA-ARS Southern Regional Research Center, UNITED STATES

**Received:** May 12, 2020

**Accepted:** February 22, 2021

**Published:** March 8, 2021



**Data Availability Statement:** All sequencing data is publically available through the submission to NCBI (https://www.ncbi.nlm.nih.gov/) as BioProject (PRJNA610844), with 53 BioSample accession numbers (SAMN14318439 - SAMN14318491), and 53 Sequence Read Archive (SRA) submission accession numbers (SRR11252447 - SRR11252499).

**Funding:** The authors would like to acknowledge the funding sources used in this project: United States Department of Agriculture National Institute

## Abstract

Recent plant breeding studies of several species have demonstrated the utility of combining molecular assessments of genetic distance into trait-linked SNP genotyping during the development of parent lines to maximize yield gains due to heterosis. SSRs (Short Sequence Repeats) are the molecular marker of choice to determine genetic diversity, but the methods historically used to sequence them have been burdensome. The ability to analyze SSRs in a higher-throughput manner independent of laboratory conditions would increase their utility in molecular ecology, germplasm curation, and plant breeding programs worldwide. This project reports simple bioinformatics methods that can be used to generate genome-wide *de novo* SSRs *in silico* followed by targeted Next Generation Sequencing (NGS) validation of those that provide the most information about sub-population identity of a breeding line, which influences heterotic group selection. While these methods were optimized in sorghum [*Sorghum bicolor* (L.) Moench], they were developed to be applied to any species with a reference genome and high-coverage whole-genome sequencing data of individuals from the sub-populations to be characterized. An analysis of published sorghum genomes selected to represent its five main races (bicolor, caudatum, durra, kafir, and guinea; 75 accessions total) identified 130,120 SSR motifs. Average lengths were 23.8 bp and 95% were between 10 and 92 bp, making them suitable for NGS. Validation through targeted sequencing amplified 188 of 192 assayed SSR loci. Results highlighted the distinctness of accessions from the guinea sub-group margaritiferum from all other sorghum accessions, consistent with previous studies of nuclear and mitochondrial DNA. SSRs that efficiently fingerprinted margaritiferum individuals (*Xgma1 –Xgma6*) are presented. Developing similar fingerprints of other sub-populations (*Xunr1 –Xunr182*) was not possible due to the extensive admixture between them in the data set analyzed. In summary, these methods were able to fingerprint specific sub-populations when rates of admixture between them are low.

of Food and Agriculture (https://nifa.usda.gov/) (Award Number 2019-67014-29174 [MYK]; Award Number 2017-33522-27086 [MYK]) and University of Nevada, Reno (https://www.unr.edu/) new faculty startup funding [MYK]. This publication was also made possible by a grant from the National Institute of General Medical Sciences (https://www.nigms.nih.gov/) (GM103440 [MYK]) from the National Institutes of Health. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

**Competing interests:** The authors have declared that no competing interests exist.

## Introduction

Originating in Africa, sorghum [*Sorghum bicolor* (L.) Moench] ($2n = 2x = 20$) is a drought-tolerant $C_4$ grass species having wide genetic diversity [1]. The fifth-most produced grain crop based on tonnage in the world, sorghum is used for biofuels, livestock feed, and human consumption, particularly in hot, dry regions of the developing world [2]. It is hypothesized that humans have been cultivating sorghum for >8,000 years [3]. Ancient farmers carried sorghum seeds throughout Africa and eventually to India and China, adapting germplasm to each new environment as they went, resulting in the diversity seen today. The five main cultivated races of sorghum (bicolor, caudatum, durra, guinea, and kafir) and working groups within them are typically identified based on panicle architecture and seed morphology [4]. There is an increasing amount of genomic sequence data available in sorghum that can be used to construct SNP-based haplotypes to differentiate between races and working groups (collectively referred to as sub-populations through this manuscript) at the genomic level, but this has yet to be leveraged efficiently in the development of molecular markers that could be used by breeders. For example, a working group within guinea, known as margaritiferum, has been shown to be distinct from the rest of the guinea race by both mitochondrial [5] and genomic [6–8] DNA analyses, and contains a large amount of untapped genetic diversity [9]. This diversity notably includes many rare alleles [8], as well as desirable traits such as high nitrogen use efficiency [10] and aluminum tolerance [11]. Currently, the identification of margaritiferum from other guinea accessions is based on differences in seed size, as set forth by Harlan and de Wet in 1972 [4]. However, seed size in sorghum is a variable phenotype that is strongly influenced by both high [12,13] and low [14] temperatures during reproductive development and the growth environment [15], making phenotypic identification difficult. Recent studies [16,17] have emphasized the need for better molecular markers than SNP haplotypes to more accurately and consistently assign accessions to a race or working group to distinguish among sub-populations of sorghum and assist in developing better heterotic groups.

Short Sequence Repeats (SSRs), also known as microsatellites and short tandem repeats, are the marker of choice for determining genetic diversity due to their high number of alleles per locus and mutation frequency [18–20]. Hypervariability in the quantity of repeated motifs [21] results in higher levels of polymorphism and mutation rates than single nucleotide polymorphisms (SNPs) and provides more information per location sequenced [18]. SSRs can be quickly and inexpensively screened through PCR amplification followed by gel separation. For these reasons, SSRs have been used for many years by plant geneticists, ecologists, and breeders. However, with the advent of next-generation sequencing (NGS) methods in the early 2000s, researchers quickly switched to using single nucleotide polymorphisms (SNPs) from whole-genome sequencing (WGS) and genotyping-by-sequencing (GBS) data sets to replace SSRs because NGS methods are much higher-throughput and, being digital, the data can be stored indefinitely and do not vary across laboratory conditions.

Following the original release of the *Sorghum bicolor* reference genome in 2009 [22], publications have emerged to make WGS and GBS data publicly available for sorghum germplasm [1,23–25]. SNPs from these data sets have been successfully used to construct haplotypes for the analysis of genetic diversity [23,24,26] and, in conjunction with phenotypic data [27–29], to make gene-trait associations for breeding using genomic selection (GS). However, despite these advancements in the throughput of sorghum genetics work, much remains unknown about the molecular basis of heterosis, which derives from population structure within the species. Theoretically, SSR loci would provide better information about population structure than a comparable number of SNP loci [18], but they come with the limitations outlined above. In addition, with all of the published SSRs in sorghum [7,30–37], their primer sites are not

amenable to targeted NGS and they were not developed specifically for their ability to differentiate sub-populations, making *de novo* identification and NGS validation necessary.

Rapid and consistent screening for population structure could be achieved if SSRs that accurately assign sorghum germplasm to races and working groups were optimized for newer and less expensive NGS methods like targeted sequencing. Currently, the volume of publicly available WGS data provides the ideal platform for *in silico* identification of SSR loci to achieve much higher genome coverage than was previously available for the best possible genetic diversity characterizations. Conventional wisdom has held that SSRs may be too long for reliable targeted sequencing due to amplicon length restrictions. In recent years, this has been proven untrue as NGS methodologies for sequencing SSRs have worked in multiple species, such as rice (*Oryza sativa* L.), cucumber (*Cucumis sativus* L.) and golden pompano (*Trachinotus ovatus* Linnaeus, 1758) [38–41]. While these studies showed the utility of NGS for large scale SSR sequencing, their focus was on the creation of new methodologies, technologies, and pipelines rather than routine screening for sub-population assignments. While such information is useful for discovery genetics purposes, a commercially-available panel of SSRs that is accessible to all breeders (including those lacking a molecular lab and NGS capabilities) would be vastly more efficient at standardizing sub-population assignments of sorghum accessions around the world. Custom panels of SSRs identified in this work can be sequenced using a variety of targeted sequencing methods at commercial labs that can also extract DNA and assist with bioinformatics.

Genetic diversity information from Diversity Arrays Technology (DArT) markers has been used in wheat [*Triticum aestivum* L.] to assist in the selection of parent lines having greater genetic distances among them to achieve improved heterosis in grain yield [42]. Jaccard genetic distance coefficients (*d*) among wheat parents in the study ranged from 0 to 0.76, with an overall mean of 0.55. Thus, giving a Jaccard's similarity coefficient of 1 to .24 with an overall mean of .45. In maize [*Zea mays* L.], heterosis observed in grain yield and most yield components was positively correlated with greater genetic distance among parents whose % relatedness ranged from 0.18 to 0.33, as determined by SNP and SilicoDArT markers [43]. Similar studies have been conducted in sorghum over the years using increasingly advanced DNA sequencing technologies. Jordan et al. (2003) [44] used RFLP markers and Mindaye et al. (2016) [45] used SSR markers; both studies reported a positive correlation between genetic distance among sorghum parents and heterosis for grain yield. Conversely, no correlation between genetic distance and heterosis was observed by Amelework et al. (2017) [46], who measured genetic distance with phenotypic and SSR markers; or by Crozier et al. (2020) [16], who used GBS-SNP data. Crozier et al. (2020) reported that genetic similarity among elite grain sorghum parents in the U.S. ranged from 0.63 to 0.79, which is quite high compared to the afore-mentioned wheat and maize studies. The authors noted that the lack of correlation could be due to the decreased genetic diversity information that results from reduced-complexity genotyping or the relatively high degree of relatedness among elite male and female parents in this crop. Either way, improved genotyping methods that efficiently and inexpensively assess population structure could be used in breeding programs to broaden the genetic distance among elite hybrid parents of sorghum to better test for and exploit heterosis in traits related to grain, biomass, or sugar yield [16,17].

The current study conducted a re-analysis of sorghum accessions sequenced from various published works [1,22,25] as a proof of concept that SSRs could be *de novo* identified from WGS data and optimized for targeted sequencing and population structure determination, including the distinction of races and working groups. To the best of the authors' knowledge, this is the first use of version three of the *Sorghum bicolor* genome assembly, which was assembled using long reads (~30 kb) as opposed to previous versions assembled from short reads

(resulting in less ambiguity within repetitive regions where SSRs reside), to identify sub-population-specific SSRs at much wider genome coverage. The methods outlined herein demonstrate that large sets of SSRs can be optimized for targeted sequencing to easily analyze many different genomic locations during single runs. These runs are easily compiled into Microsoft Excel or CSV files that are accessible to those with limited bioinformatics training. This helps breeders to better understand and utilize the genetic diversity within their breeding populations without the need for expensive specialized equipment.

The aim of this project was the *de novo* identification of SSRs for sub-population determination in sorghum. This work provides the first publicly available resource to genetically differentiate sub-populations using targeted sequencing methods that are more economical than genome-wide SNPs and that also provide a more accurate picture of genetic diversity. As a proof of concept, this work identified NGS-validated SSRs that successfully differentiated the guinea working group, margaritiferum, (*Xgma1 –Xgma6*), from the five main races of sorghum. This work also provides an additional 182 NGS-validated SSRs for analyzing genetic diversity (*Xunr1 –Xunr182*). All SSRs are presented with their genomic locations and suitable primers for their extension using either gel-based or NGS techniques. Once additional WGS data is published that evenly and comprehensively represents the five main races, similar fingerprints for each of them will also be possible. Finally, these methods can be deployed in any species with a reference genome to assist molecular ecology, germplasm curation, or conservation programs.

## Materials and methods

### Published sequencing data processing and SSR identification

The raw sequencing data of 75 sorghum accessions [1,22,25] were used for *in silico* analysis, and were obtained from the sequence read archive at NCBI. To distinguish samples by source, accessions from the first published data were named as published [22], while data from subsequent published works had suffix labels added to the accession's names [1,25]. The suffixes added were: _TAMU, _UQ, _BGI, for data generated at Texas A&M University, University of Queensland, and BGI respectively. BWA (version: bwa-0.7.17) [47] was used with default settings to align to the *Sorghum bicolor* reference genome (Sbicolor_454_v3.0.1) [25]. GATK (version: gatk-4.0.5.1) [48] was used to identify variations and call haplotyes with the CreateSequenceDictionary, IndexFeatureFile, HaplotypeCaller, CombineGVCFs, and MergeVcfs tools. Subsequently, htslib (version: 1.8) in Samtools (version: 1.9) [49] was used with the subtool tabix to create an indexed file. Variant sites were then filtered prior to *de novo* SSR identification using PLINK software (version: 1.9) [50] with settings --mind 0.7 --maf 0.05 --geno 0.01 --allow-extra-chr --indep-pairwise 50 10 0.1 --double-id --vcf-inspace-to "_". Filtered variants were processed with HipSTR (version: HipSTR-v0.6.2) [51] for SSR identification using the settings --min-reads 25 --no-rmdup --max-mate-dist 1000 --max-str-len 500 --max-reads 200 --def-stutter-model --require-pair.

### Genetic analysis of populations

For the genetic analysis of sorghum populations, fastSTRUCTURE [52] (version: fastSTRUCTURE-e47212f) was used to map the genetic structure at $K = 2–10$ (--full --seed = 100 --prior = logistic). Calculation of the log-marginal likelihood lower bound (LLBO) was performed in fastSTRUCTURE to determine the optimal value $K$ within the population of study [52]. Plots were generated using the R [53] package ggplot2 (version: 3.2.1) [54] in RStudio [55]. SplitsTree4 (version: 4.15.1) [56] was used to create a split network tree. Colors were overlaid to the

separate clades (putatively corresponding to races) of the tree based on the populations determined in fastSTRUCTURE [52].

## Venn diagram

SSR locations for each of the six populations (bicolor ($n$ = 4), caudatum ($n$ = 29), durra ($n$ = 19, guinea ($n$ = 3), kafir ($n$ = 16), and margaritiferum ($n$ = 4)) were assigned a score of 1 if there was a reference or alternative allele found in the previously published sequencing data [1,22,25] within each population. SSR locations were given a 0 if there was no reference or alternative allele found at that position of the chromosome in the genotype data of the Variant Call Format (VCF) file. Shared and unique sets of SSR candidates were visualized in R [53] using the package venn (version: 1.7) [57] in RStudio [55] with clades again colored to match those in fastSTRUCTURE [52] and SplitsTree4 [56].

## SSR filtering

The DNA sequences for sorghum sub-populations were compiled together and separated into VCF files arranged by chromosome. Through vcftools (version: 0.1.16) [58] VCF files were first filtered for allele quantity (--max-alleles 13 --min-alleles 2 --max-missing 1 --recode), then the vcftools function --singletons was used. Only the 4,179 unique doubletons, locations with a minor allele occurring in only one population homozygous for that allele, were taken from these files for identification of SSRs. Locations were targeted for primer design by LGC Genomics (Teddington, United Kingdom) using proprietary methods. Only locations with high specificity (no off-target sites throughout the genome for both forward and reverse primers) were considered beyond this stage. Reference and alternative alleles were used to calculate the standard deviation of SSR lengths between the populations in R [53]. SSRs with standard deviations above seven were retained for further filtering to focus on locations with maximal diversification within the population. SSRs within 1 Mb of the ends of each chromosome were filtered out to avoid telomeric regions. SSRs with a length less than 13 or greater than 49 were excluded. Finally, all SSRs with a motif length of one were removed. Based upon these criteria, 192 SSRs were validated using NGS sequencing.

## Plant material, DNA extraction, and sequencing

A total of 75 sorghum WGS data sets were published with sufficient sequencing depth for analysis at the time of the study [1,22,25], but a subset of only 53 of these accessions were available through the United States Department of Agriculture U.S National Plant Germplasm System Grin-Global [59] for validation of the new SSRs. Seedlings of these 53 accessions were grown in Pro-Mix® Biofungicide™ growing medium (Premier Tech Horticulture, Rivière-du-Loup, Qc) until they were large enough to provide leaf samples. Leaf discs were harvested from immature tissues using the LGC BioArk Leaf kits according to the manufacturer's instructions. The DNA was extracted by LGC genomics using LGC sbeadex™ chemistry and libraries were prepared using the LGC genomics SeqSNP pipeline. Sequencing was performed by LGC genomics using Illumina® [60] sequencing-by-synthesis technology on an Illumina® NextSeq 550 with paired-end 150-base pair reads. All sequencing data is publicly available through the submission to NCBI (https://www.ncbi.nlm.nih.gov/) as BioProject (PRJNA610844), with 53 BioSample accession numbers (SAMN14318439—SAMN14318491), and 53 Sequence Read Archive (SRA) submission accession numbers (SRR11252447—SRR11252499).

### *In silico* SSR chromosomal mapping

*In silico* SSR chromosome mapping was performed using the R [53] package chromPlot (version: 1.12.0) [61] in RStudio [55]. SSRs were mapped using base settings except for the following: names were plotted at chromosomal location with the "stat" argument set to the SSR labels and the "noHist" argument used to avoid plotting histograms onto the output plots.

### Population genetics statistics

The genetic information statistics for polymorphism information content (PIC) and heterozygosity (H) [62] were calculated in Excel using the following equations:

$$H = 1 - \sum_{i=1}^{l} P_i^2$$

$$PIC = 1 - \sum_{i=1}^{l} P_i^2 - \sum_{i=1}^{l-1} \sum_{j=i+1}^{l} 2P_i^2 P_j^2$$

Wherein $l$ is the number of alleles in the locus and $P_i$ and $P_j$ are the allele frequencies of the $i^{th}$ and $j^{th}$ alleles, respectively.

## Results

In order to investigate genetic diversity in sorghum, a re-analysis was performed of WGS data from 75 published accessions [1,22,25], generating a map of 19,230,634 variants at 18,299,015 sites across the *S. bicolor* (Sbicolor_454_v3.0.1) genome [25]. Populations within the 75 samples were predicted using whole genome polymorphism data through fastSTRUCTURE for multiple *K* values, 2–10 (S1 Fig). By investigating the population size that maximizes the marginal likelihood through the LLBO curve (S2 Fig), the optimum population number $K = 6$ was found. Plotting *K* values two through ten (S1 Fig) demonstrated fastSTRUCTURE's ability to avoid overfitting: $K = 7$–10 would not produce more than six populations with the seventh through tenth populations providing zero contribution to the overall genetic structure (S1 Fig). The use of fastSTRUCTURE with $K = 6$ (Fig 1) showed a distinct population genetic structure. The separation of the accessions into six populations further partitioned margaritiferum into a population distinct from guinea (Fig 1). At $K = 2$ (the first separation among sorghum populations), the genetic structure analysis depicts the margaritiferum accessions as different from the majority of accessions within the other populations (S1 Fig), indicating its genetic differentiation from all of the main races. The margaritiferum accessions remained distinct from $K = 2$–10 (S1 Fig). These findings are consistent with those of other researchers [1,5–8,63] indicating that margaritiferum accessions are distinct from guinea accessions. Independently of, and concurrent with fastSTRUCTURE (Figs 1, S1 and S2), split network analysis with SplitsTree4 also separated the 75 sorghum accessions into six clades using aligned WGS data for each samples [1,22,25] (Fig 2). The spatially-segregated accessions within the split network (Fig 2) again showed a sharp separation of accessions belonging to margaritiferum from guinea accessions, and the margaritiferum clade was the most distinct among all clades.

*In silico* analysis of the *Sorghum bicolor* reference genome predicted 163,943 SSRs, of which 130,120 had mapped sequence data for the 75 accessions analyzed. Of the 130,120 SSRs identified, 18,080, 15,767, 15,857, 13,768, 11,392, 11,496, 11,255, 10,161, 10,895, and 11,449 were located on chromosomes one, two, three, four, five, six, seven, eight, nine, and ten respectively. In the merged VCF files, the majority showed allele presence/absence variation based on sub-

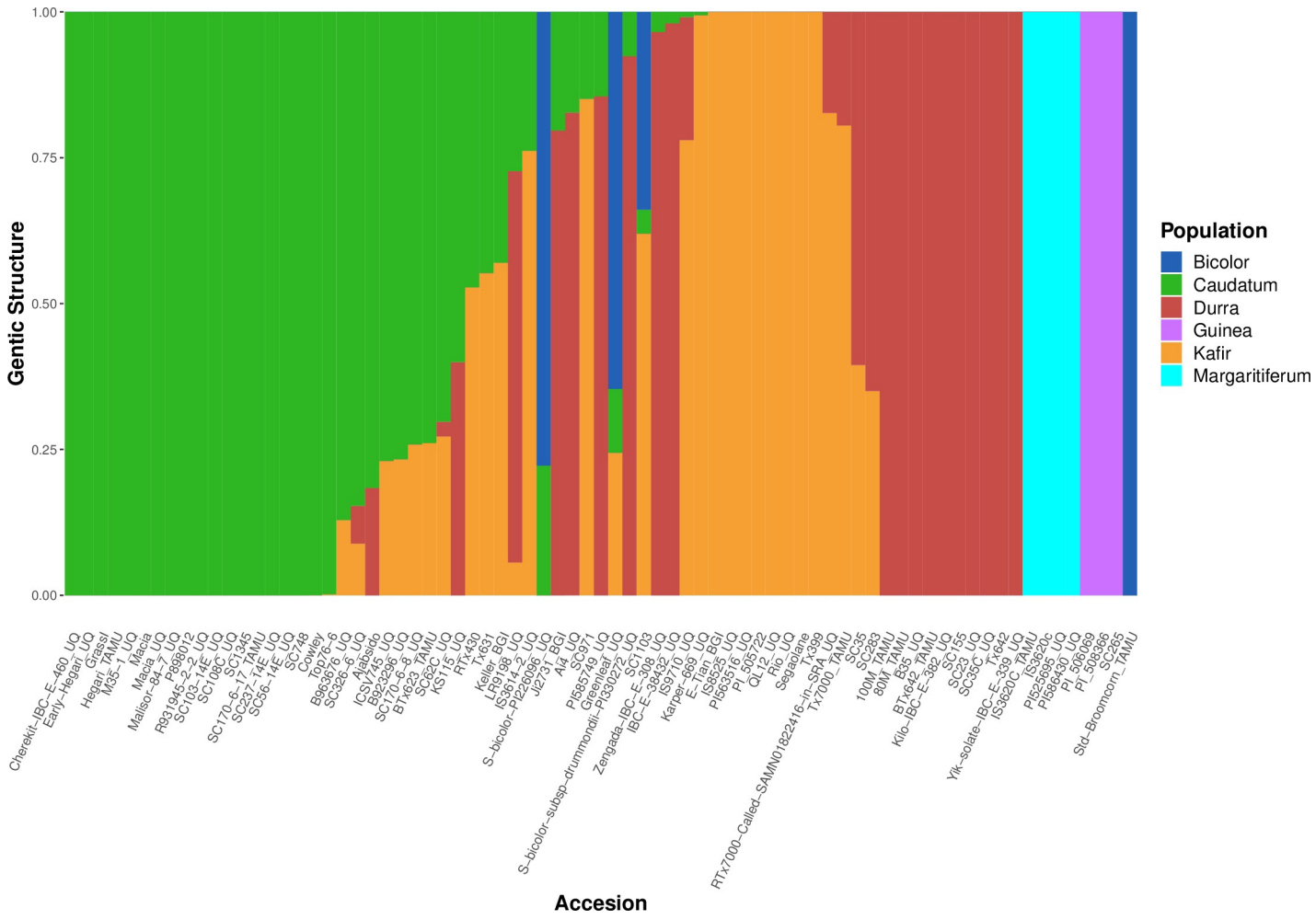


**Fig 1. Six distinct sorghum sub-population distributions predicted by optimal fastSTRUCTURE grouping.** fastSTRUCTURE mapping of the genetic structure of 75 sorghum accessions with optimized $K = 6$. The six sub-populations are depicted by: dark blue, green, red, purple, orange, and light blue representing bicolor, caudatum, durra, guinea, kafir, and margaritiferum respectively. The *x-axis* represents each accession and the *y-axis* represents the proportion of the genetic structure from the sorghum sub-populations.

https://doi.org/10.1371/journal.pone.0248213.g001

population (Fig 3). While many SSR alleles were shared (65,512) between all six sub-populations, each one had at least 300+ SSR alleles that were unique to that sub-population (Fig 3). These unique SSR alleles highlight the divergence that has occurred between the main races of sorghum as well as the guinea working group margaritiferum, and the genetic diversity that is available to plant breeders. The key innovation in these methods was in filtering for doubletons–SSR loci that were unique to one sub-population and every individual in that sub-population was homozygous for that allele. This filtering method maximized the information obtained by the sequencing of each SSR so that more accurate population structure could be determined.

After filtering for doubletons, 4,179 sites passed with 516, 506, 558, 409, 397, 326, 458, 247, 311, and 451 SSRs located on Chromosomes one, two, three, four, five, six, seven, eight, nine, and ten, respectively. Doubleton filtering increased the likelihood of finding SSR alleles that were specific to only one sub-population upon sequencing. Filtering based on allele quantity removed SSRs with too few alleles to be informative or SSRs at sites that were too prone to mutation, which would decrease the heritability of SSR lengths and hence their ability to

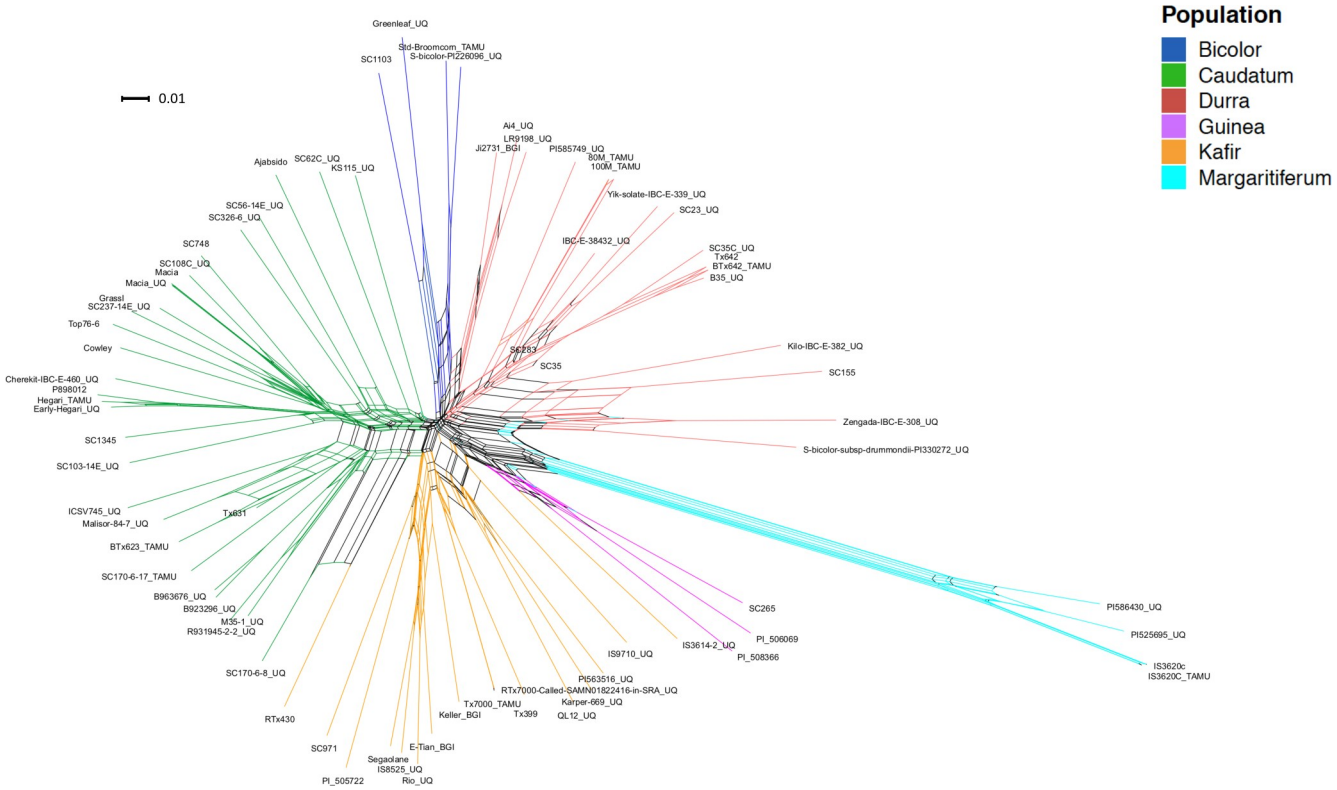


**Fig 2. Split network divides sorghum into six clades.** Split network diagram from SplitsTree4 depicting the genetic separation of 75 sorghum accessions. Splits-identified genetic divergence between accessions. Colors indicate the clade each accession belongs to. Dark blue, green, red, purple, orange, and light blue represent bicolor, caudatum, durra, guinea, kafir, and margaritiferum, respectively. Each node is labeled with the accession it represents. The scale bar is representative of a weight of 0.01 of the corresponding split.

https://doi.org/10.1371/journal.pone.0248213.g002

accurately track the sub-population identity of a breeding line across generations. Filtering out telomeric regions was important to prevent SSRs from being drawn from these gene-rich sites of high recombination and crossing over events [64,65] in the *Sorghum bicolor* genome. Unequal crossing over events along with replication strand slippage are the two main models for SSR length mutation [66]. The goal of the study was not to develop trait-linked SSRs in gene-rich regions of the genome, but to identify SSRs that stably, over many generations, assess genetic distance among sub-populations. This necessitated finding SSRs that mutate less frequently.

The 4,179 doubleton sites were targeted for primer design by LGC Genomics using proprietary methods. Final filtering for sites with high primer specificity, no off-target primer binding, and fragment length identified 192 SSRs that were selected for validation using targeted NGS. SSR length filtering is necessary to identify sites where the entire sequence of the SSR can be ascertained while taking into account the primer length (40 bp) and read lengths (150 bp) used. Selected SSRs were sequenced using Illumina technology [60] in all 53 publicly-available accessions from the USDA Grin-Global [59] out of the original 75 used for *in silico* analysis.

The PI (Plant Introduction) numbers and accession name labels of the 53 accessions used for NGS validation of SSRs are provided in S1 Table. After seedling emergence, genomic DNA from the accessions was used in Illumina® directed paired end DNA NGS to amplify the 192 SSR sites. Sequencing summary statistics are provided in S2 Table. Of the 192 SSR sites 187 were successfully sequenced in the accessions analyzed in this manuscript. One SSR (Xunr43) only amplified in wild sorghum accessions included in sequencing run for a separate analysis,

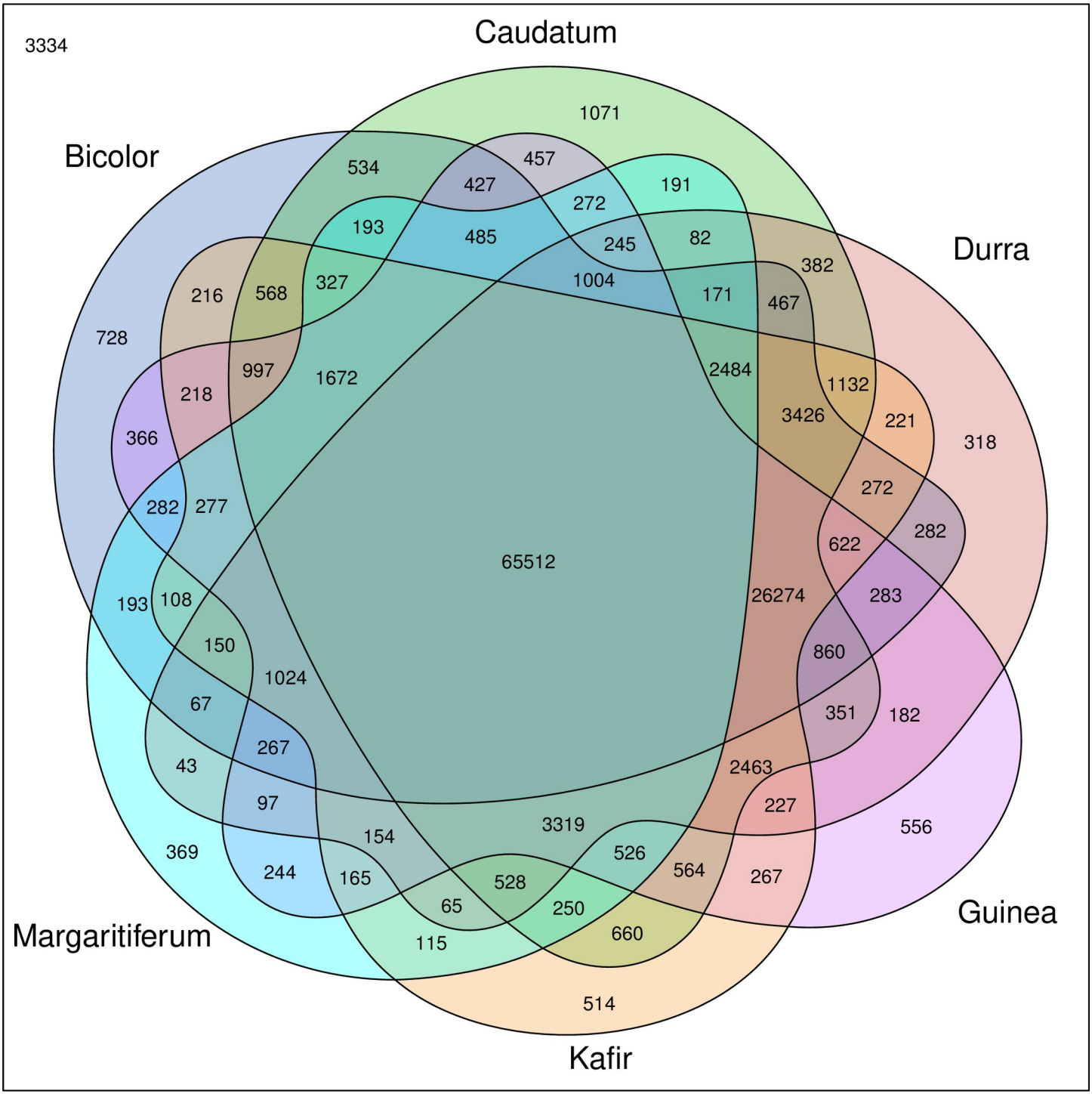


**Fig 3. Venn diagram of *de novo* SSRs shows distinct SSRs for each clade.** SSRs that are specific to only one clade in the merged VCF file are depicted by sections with no overlapping colors. SSRs that are shared by different clades are depicted within overlapping sectors. The SSRs that were identified in the *Sorghum bicolor* reference genome but had no reference or alternative alleles in the accessions evaluated by this study are denoted in the top left corner. The colored Venn regions are labeled with the sub-population they depict. Dark blue, green, red, purple, orange, and light blue represent bicolor ($n = 4$), caudatum ($n = 29$), durra ($n = 19$), guinea ($n = 3$), kafir ($n = 16$), and margaritiferum ($n = 4$), respectively.

https://doi.org/10.1371/journal.pone.0248213.g003

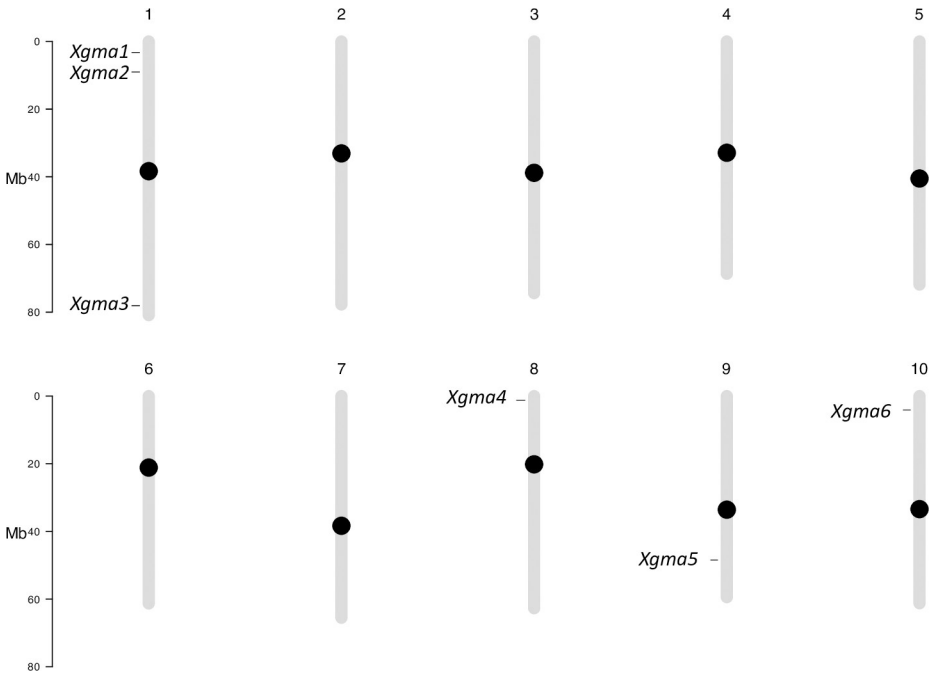


**Fig 4. Chromosomal map of the six SSRs specific to margaritiferum.** The ten chromosomes of *Sorghum bicolor* are represented in sequential order and labeled with the chromosome number above each. The bar on the left of each set of five chromosomes depicts the length in millions of bases (Mb). The black circles indicate centromeric regions. Each SSR is named and marked at its location on the chromosome.

https://doi.org/10.1371/journal.pone.0248213.g004

and six of the sites were unique for margaritiferum (S3 Table). These six SSRs are shown mapped to their chromosomal locations in Fig 4. The SSRs are denoted first with *X* to signify that they are SSRs, then with *gma* to identify their use for determining the guinea working group margaritiferum, and finally numbered one through six in order of chromosomal and genomic location starting from the first base pair on Chromosome one to the final base pair on Chromosome ten based on the current *Sorghum bicolor* reference genome [25]. The six SSRs, *Xgma1* –*Xgma6*, their genomic locations, repeat sequences in the reference genome, repeat sequences specific to margaritiferum, as well as their corresponding forward and reverse primers, are shown in Table 1. The sequences of *Xgma1* –*Xgma6* for the 53 accessions sequenced are presented in S3 Table. The rest of the SSRs identified are denoted as *X* to signify them as SSRs, then with *unr* to identify the location of their development (University of Nevada, Reno), and numbered 1–182 by the same ordination scheme of genomic location (S4 Table) as was used for *Xgma1* –*Xgma6*. Based off the thresholds set by Botstein et al. 1980 [62], 114 of the SSRs presented here are highly informative (PIC $>$ 0.5), 43 are reasonably informative (0.5 $>$ PIC $>$ 0.25), and 30 are slightly informative (PIC $<$ 0.25). It is important to note that the single SSR (*Xunr43*) that amplified only in wild sorghum relatives sequenced for a separate analysis was not used to calculate PIC or H because it was not relevant to differentiating among sub-populations (S5 Table). In agreement with the observation that many SSR alleles are shared among non-margaritiferum accessions (S4 Table), common genetic structure was observed between accessions from the bicolor, caudatum, durra, and kafir sub-populations, whereas no shared genetic structure was found between accessions from the margaritiferum or guinea sub-populations when all variants were analyzed simultaneously (Fig 1). This finding is potentially the result of genetic admixture among bicolor, caudatum, durra, and kafir accessions in U.S. breeding programs, which have largely excluded guinea and margaritiferum. The

**Table 1. NGS-validated primer sites and sequence lengths of margaritiferum-specific SSRs (*Xgma1*—*Xgma6*), based on the data set analyzed.**

| SSR Name | Chromosome | Start Location (bp) | End Location (bp) | Reference Repeat Structure | Repeat Structure in Margaritiferum | Forward Primer (+, 5' to 3')) | Reverse Primer (-, 5' to 3') |
|---|---|---|---|---|---|---|---|
| ***Xgma1*** | 1 | 3,295,349 | 3,295,385 | GCG (12) | GCG (4) | TCGGTCGTGCCGGGAAGGGGGACTGGAGGGTAGGGTCTGG | CCCAACCGCAACAGACCAAACCCAGCCGCAACCGCAGAGC |
| ***Xgma2*** | 1 | 9,033,349 | 9,033,373 | GCC (8) | GCC (10) | GGTCGAGGGCTCCTCTCTCTCTACCGTCCCCGCTCGATCC | GGCAGCGCGACGATAGGCGCGAGTTGGACTGGCGGAGAGC |
| ***Xgma3*** | 1 | 78,120,648 | 78,120,675 | ACT (9) | ACT (6) | ATACATACATATCTGTTAGGCCATGCATGAACATCTAAAC | GAAACCGGAAGAAAGATTGCACTCGATCGTGTAATAGTCG |
| ***Xgma4*** | 8 | 1,256,427 | 1,256,444 | CCACGC (3) | CCACGC (2) | GCCCAGCAGACCAGCCGCGTTGGCTGGCGGGACGGAGCAG | CAGTAGACCGTCGGGACGGCGCCGCTGGGCCTCGGCTTGG |
| ***Xgma5*** | 9 | 48,460,761 | 48,460,787 | GGATG (5) | GGATG (4) | TGTGGGATTTCGCTTCTCGAGGGAACGGAATTACGGGAAG | GCCAAATCAAAGCTAGACTCGACGCTAGTGCCATGTGACG |
| ***Xgma6*** | 10 | 4,107,816 | 4,107,841 | GAG (8) | GAG (5) | GCCGGTGGTGGAGCGGTTGCGGCGGCGGCAGGGCACGGCC | CCACCCACTTGCCCCACTTGCGCATCCGAACGCCGCGATA |

SSR names and genomic locations are listed. Repeat structure with repeat quantity, rounded to whole repeat, in parentheses for both the reference genome sequence and margaritiferum individuals are shown. The forward and reverse primers corresponding to each SSR are provided in 5' to 3' orientation for the + and—strands.

https://doi.org/10.1371/journal.pone.0248213.t001

genetic purity of margaritiferum accessions made it easy to distinguish from other sub-populations. Extensive admixture among accessions of other sub-populations reduced the number of doubletons that were unique to only one sub-population, and therefore the statistical power to generate unique SSR fingerprints for each one.

The numbers of repeats in *Xgma1* –*Xgma6* are highly unique to margaritiferum accessions. For example, the *Xgma2* allele in margaritiferum accessions was ten (GCC) repeats, but no *Xgma2* alleles of this length were observed in non-margaritiferum accessions (S3 Table). *Xgma4* alleles in one margaritiferum accession did not produce sequencing data, but all others were homozygous for two (CCACGC) repeats; whereas no *Xgma4* alleles of this length were observed in non-margaritiferum accessions (S3 Table). The *Xgma6* allele in margaritiferum accessions had five (GAG) repeats and all accessions were homozygous for this length (S3 Table), whereas no *Xgma6* alleles of this length were observed in non-margaritiferum accessions. In combination, the length haplotypes amplified by our novel six SSRs primer sets (*Xgma1* –*Xgma6*) can clearly differentiate the margaritiferum accessions from all other sub-populations that contributed to this analysis.

## Discussion

Other researchers have observed that there are likely more than five primary sub-populations within cultivated sorghum [5,6]. Depending on the number and ancestry of the accessions included in any population structure analysis, the distribution of populations may vary. The current project identified six distinct sub-populations (Figs 1 and 2) via genetic structure analysis and independently by the split network method among the 75 published WGS data sets in sorghum that were used for this analysis. The LLBO curve (S2 Fig), also denoting six as the optimal population size (*K*) for the sampled accessions, validated these findings. *In silico* analysis identified 163,943 SSRs in the *Sorghum bicolor* reference genome. Of those, 130,120 were present in 75 published WGS data sets and selected for further investigation based on read depth and sequence quality. While the six sub-populations identified shared many SSRs (65,512), each sub-population had 318 to 1,071 SSRs that were present in only that sub-population in the analyzed dataset (Fig 3). Nevertheless, one must be cautious interpreting this result, as conclusive fingerprinting of sub-population margaritiferum was likely only possible because it was free of admixture. Larger sample sizes representing the remaining five sub-populations are needed for WGS, particularly accessions with minimal admixture from other sub-populations, in order to develop SSR fingerprints with the greatest ability to assess population structure in sorghum and contribute the most meaningful genetic diversity information to breeding for yield gains through improved heterosis, as has been done in wheat and maize [42,43].

3,334 SSRs were predicted by HipSTR from the *Sorghum bicolor* reference genome that were not found in any of the merged VCF files from the 75 accessions studied at those locations. This is possibly an effect of differing sequencing depths among accessions used in file merging or SSRs that only exist in the individual used for the reference genome. Within the 130,120 SSRs identified in the 75 accessions studied, there were 4,179 unique doubleton locations that were filtered using the methods described to narrow down the NGS validation set to 192. These SSR sites were investigated for sub-population determination among all 53 publicly-available sorghum accessions available from Grin-Global [59] at the time of this project. The unique method doubleton filtering, compared to other SSR-NGS sequencing projects [38–41], was the key to enriching for SSRs that efficiently identified population structure. Doubleton filtering is better at identifying unique alleles specific to an entire sub-population whereas singleton filtering is better suited to identifying rare alleles within sub-populations.

Doubleton filtering to identify sub-populations was achieved by merging accessions within each sub-population prior to analysis of the WGS data, essentially treating the merged file as one individual versus using each accession as its own individual.

Of the 192 genomic locations selected for NGS validation, six unambiguously differentiated margaritiferum accessions from bicolor, caudatum, durra, guinea, and kafir accessions, making them ideal genetic diversity markers for this purpose. These six SSRs, mapped in Fig 4 and described in Table 1, can be used to identify margaritiferum individuals in the analyzed data sets. Additional SSRs may be tested and validated using the data sets provided. The sequences of *Xgma1* –*Xgma6* (S3 Table) show how a simple Excel output file allows SSRs to be easily visually compared among sub-populations, thus providing accessible genetic diversity information to students and researchers with limited bioinformatics training.

DNA markers have long been used to study genetic relatedness in plants and forensics in humans. Along with SSRs other commonly used markers of genetic analysis include restriction fragment length polymorphisms (RFLPs), DArT markers, and SNPs. SSRs have advantages of increased power of discrimination among sub-populations, are more reliable, and have better repeatability than RFLPs [67] and SNPs [18], and are more modular and customizable in format than DArT markers, which come in sets and rely on solid-state sequencing platforms. Nevertheless, SSR sequencing has historically been cumbersome and fell out of favor once GBS-SNP pipelines were developed. Unfortunately, a selection bias is introduced by using the direct identification of genomic regions responsible for desired phenotypes to form the basis of estimating genetic diversity, instead of molecular markers that may or may not be linked to traits under artificial selection [18]. It is worth investigating whether or not cost savings can be introduced in GS pipelines currently relying on GBS-SNPs by diversifying marker types to reduce the total number of loci sequenced. For example, NGS-SSR markers could be used to assess the genetic background of entire genomes or even specific chromosomal regions, as well as the genetic diversity of breeding lines; and SNP markers could be used in parallel to focus on functional phenotypic variation. The methods outlined herein are ideal for developing sets of SSRs specific to the germplasm and needs of individual breeding programs that could be used to test this strategy.

The availability of NGS-SSR pipelines enables the evaluation of genetic diversity at a throughput not previously possible, and they can be deployed in any species with a reference genome. NGS techniques are currently powerful enough that even genetically "identical" twins can be differentiated from one another [68]. The power of NGS coupled with the low cost of targeted sequencing makes methodologies like those described in this paper an exciting frontier in population genetics, heterotic group development in breeding programs, and the genetic identification of patented materials.

## Conclusions

Our results divided sorghum into six distinct sub-populations (bicolor, caudatum, durra, guinea, kafir, and margaritiferum) based both on genetic structure and split network mapping analyses of 75 published WGS data sets and comparisons with the *Sorghum bicolor* reference genome. To help identify margaritiferum from the remaining sub-populations, six novel SSR primer sets (*Xgma1* –*Xgma6*) suited for targeted NGS are presented. 182 additional novel SSR primer sets (*Xunr1* –*Xunr182*) are presented and may be used to develop similar fingerprints for the remaining sub-populations once more WGS information for each one becomes available. These SSRs are reported with their repeat motifs, the quantity of repeats in the reference genome and (where applicable) in margaritiferum accessions; forward and reverse primer sequences, and PIC and H values. This project demonstrates the amenability of SSRs to

targeted sequencing and NGS and lays out a framework for future work standardizing molecular characterizations of sub-populations in sorghum.

## Supporting information

**S1 Fig. Accessions' genetic structure of *K* = 2–10 shows 6 is the maximum population size.** FastSRTUCTURE plotting in ggplot2 of accessions' genetic structure from *K* equals 2 to 10 in order. Color depicting each population (P1-P10) shown in the corresponding legend on the right of each graph.
(PDF)

**S2 Fig. Marginal likelihood identifies 6 as the optimum *K* value.** The log-marginal likelihood lower bound (*y*-axis) calculated in fastSTRUCTURE and plotted against the *K* population size (*x*-axis) shows 6 is the optimum population size to maximize the marginal likelihood. Dashed line drawn at -.709 marginal likelihood for cut off between *K* = 5 & 7 versus *K* = 6.
(PDF)

**S1 Table. Accessions sequenced using NGS.** This table includes the accession names and PI numbers that were grown and used for NGS sequencing.
(XLSX)

**S2 Table. Read count table for 53 sequence accessions on Illumina® NextSeq 550.** The read count table is presented for all of the samples sequenced by NGS.
(XLSX)

**S3 Table. Sequences of accessions based on SSR.** This table includes the sequencing data for each of the 53 accessions at *Xgma1 –Xgma6*. The accessions are sorted based on population. The reference allele and all alternative alleles for the SSRs are included at the bottom of the table.
(XLSX)

**S4 Table. Sequences of all SSRs (*Xgma1 –Xgma6* & *Xunr1 –Xunr182*).** This table includes the sequencing data of the 53 accessions at *Xgma1 –Xgma6* & *Xunr1 –Xunr182*. The chromosomal locations of the SSRs, the primers (forward and reverse) with their associated Tms, and the reference genome sequence are provided.
(XLSX)

**S5 Table. PIC and H values of all SSRs (*Xgma1 –Xgma6* & *Xunr1 –Xunr182*).** This table includes the PIC and H values calculated for *Xgma1 –Xgma6* & *Xunr1 –Xunr182* based on the 53 accessions sequencing data.
(XLSX)

## Acknowledgments

The authors would like to acknowledge Drs. Emma Mace and David Jordan for their helpful discussions of the data generated by this project.

## Author Contributions

**Conceptualization:** John P. Baggett, Melinda K. Yerka.

**Data curation:** John P. Baggett, Richard L. Tillett.

**Formal analysis:** John P. Baggett.

**Funding acquisition:** Melinda K. Yerka.

**Investigation:** John P. Baggett.

**Methodology:** John P. Baggett.

**Resources:** Richard L. Tillett.

**Software:** John P. Baggett.

**Visualization:** John P. Baggett.

**Writing – original draft:** John P. Baggett.

**Writing – review & editing:** Richard L. Tillett, Elizabeth A. Cooper, Melinda K. Yerka.

## References

1. Mace ES, Tai S, Gilding EK, Li Y, Prentis PJ, Bian L, et al. Whole-genome sequencing reveals untapped genetic potential in Africa's indigenous cereal crop sorghum. Nat Commun. 2013; 4. https://doi.org/10.1038/ncomms3320 PMID: 23982223
2. Paterson AH. Genomics of Sorghum. Int J Plant Genomics. 2008; 2008: 1–6. https://doi.org/10.1155/2008/362451 PMID: 18483564
3. Wendorf F, Close AE, Schild R, Wasylikowa K, Housley RA, Harlan JR, et al. Saharan exploitation of plants 8,000 years BP. Nature. 1992; 359: 721–724. https://doi.org/10.1038/359721a0
4. Harlan JR, de Wet JMJ. A Simplified Classification of Cultivated Sorghum. Crop Sci. 1972; 12: 172–176. https://doi.org/10.2135/cropsci1972.0011183X001200020005x
5. Deu M, Hamon P, Chantereau J, Dufour P, D'hont A, Lanaud C. Mitochondrial DNA diversity in wild and cultivated sorghum. Genome. 1995; 38: 635–645. https://doi.org/10.1139/g95-081 PMID: 7672599
6. Oliveira AC de, Richter T, Bennetzen JL. Regional and racial specificities in sorghum germplasm assessed with DNA markers. Genome. 1996; 39: 579–587. https://doi.org/10.1139/g96-073 PMID: 8675002
7. Ramu P, Billot C, Rami J-F, Senthilvel S, Upadhyaya HD, Ananda Reddy L, et al. Assessment of genetic diversity in the sorghum reference set using EST-SSR markers. Theor Appl Genet. 2013; 126: 2051–2064. https://doi.org/10.1007/s00122-013-2117-6 PMID: 23708149
8. Deu M, Rattunde F, Chantereau J. A global view of genetic diversity in cultivated sorghums using a core collection. Genome Ott. 2006; 49: 168–80. https://doi.org/10.1139/g05-092 PMID: 16498467
9. Folkertsma RT, Rattunde HFW, Chandra S, Raju GS, Hash CT. The pattern of genetic diversity of Guinea-race *Sorghum bicolor* (L.) Moench landraces as revealed with SSR markers. Theor Appl Genet. 2005; 111: 399–409. https://doi.org/10.1007/s00122-005-1949-0 PMID: 15965652
10. Massel K, Campbell BC, Mace ES, Tai S, Tao Y, Worland BG, et al. Whole Genome Sequencing Reveals Potential New Targets for Improving Nitrogen Uptake and Utilization in Sorghum bicolor. Front Plant Sci. 2016; 7: 1544. https://doi.org/10.3389/fpls.2016.01544 PMID: 27826302
11. Caniato FF, Guimarães CT, Hamblin M, Billot C, Jean-François R, Hufnagel B, et al. The Relationship between Population Structure and Aluminum Tolerance in Cultivated Sorghum. PLoS One San Franc. 2011; 6: e20830. https://doi.org/10.1371/journal.pone.0020830 PMID: 21695088
12. Prasad PVV, Boote KJ, Allen LH. Adverse high temperature effects on pollen viability, seed-set, seed yield and harvest index of grain-sorghum [*Sorghum bicolor* (L.) Moench] are more severe at elevated carbon dioxide due to higher tissue temperatures. Agric For Meteorol. 2006; 139: 237–251. https://doi.org/10.1016/j.agrformet.2006.07.003
13. Prasad PVV, Pisipati SR, Mutava RN, Tuinstra MR. Sensitivity of Grain Sorghum to High Temperature Stress during Reproductive Development. Crop Sci. 2008; 48: 1911–1917.
14. Maulana F, Tesso TT. Cold Temperature Episode at Seedling and Flowering Stages Reduces Growth and Yield Components in Sorghum. Crop Sci. 2013; 53: 564–574.
15. Griess JK, Mason SC, Jackson DS, Galusha TD, Yaseen M, Pedersen JF. Environment and Hybrid Influences on Food-Grade Sorghum Grain Yield and Hardness. Crop Sci. 2010; 50: 1480–1489.
16. Crozier D, Leo Hoffmann Jr, Klein Patricia E., Klein Robert R., Rooney William L. Predicting heterosis in grain sorghum hybrids using sequence-based genetic similarity estimates. J CROP Improv. 19.

17. Sapkota S, Boyles R, Cooper E, Brenton Z, Myers M, Kresovich S. Impact of sorghum racial structure and diversity on genomic prediction of grain yield components. Crop Sci. 2020; 60: 132–148. https://doi.org/10.1002/csc2.20060.

18. Hamblin MT, Warburton ML, Buckler ES. Empirical Comparison of Simple Sequence Repeats and Single Nucleotide Polymorphisms in Assessment of Maize Diversity and Relatedness. PLOS ONE. 2007; 2: e1367. https://doi.org/10.1371/journal.pone.0001367 PMID: 18159250

19. Mesak F, Tatarenkov A, Earley RL, Avise JC. Hundreds of SNPs vs. dozens of SSRs: which dataset better characterizes natural clonal lineages in a self-fertilizing fish? Front Ecol Evol. 2014; 2. https://doi.org/10.3389/fevo.2014.00066 PMID: 25729749

20. Manechini JRV, da Costa JB, Pereira BT, Carlini-Garcia LA, Xavier MA, Landell MG de A, et al. Unraveling the genetic structure of Brazilian commercial sugarcane cultivars through microsatellite markers. Fang DD, editor. PLOS ONE. 2018; 13: e0195623. https://doi.org/10.1371/journal.pone.0195623 PMID: 29684082

21. Jeffreys AJ, Royle NJ, Wilson V, Wong Z. Spontaneous mutation rates to new length alleles at tandem-repetitive hypervariable loci in human DNA. Nature. 1988; 332: 278–281. https://doi.org/10.1038/332278a0 PMID: 3347271

22. Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, et al. The *Sorghum bicolor* genome and the diversification of grasses. Nature. 2009; 457: 551–556. https://doi.org/10.1038/nature07723 PMID: 19189423

23. Zheng L-Y, Guo X-S, He B, Sun L-J, Peng Y, Dong S-S, et al. Genome-wide patterns of genetic variation in sweet and grain sorghum (*Sorghum bicolor*). Genome Biol. 2011; 12: R114. https://doi.org/10.1186/gb-2011-12-11-r114 PMID: 22104744

24. Evans J, McCormick RF, Morishige D, Olson SN, Weers B, Hilley J, et al. Extensive Variation in the Density and Distribution of DNA Polymorphism in Sorghum Genomes. PLoS ONE. 2013; 8: e79192. https://doi.org/10.1371/journal.pone.0079192 PMID: 24265758

25. McCormick RF, Truong SK, Sreedasyam A, Jenkins J, Shu S, Sims D, et al. The *Sorghum bicolor* reference genome: improved assembly, gene annotations, a transcriptome atlas, and signatures of genome organization. Plant J. 2018; 93: 338–354. https://doi.org/10.1111/tpj.13781 PMID: 29161754

26. Morris GP, Ramu P, Deshpande SP, Hash CT, Shah T, Upadhyaya HD, et al. Population genomic and genome-wide association studies of agroclimatic traits in sorghum. Proc Natl Acad Sci. 2013; 110: 453–458. https://doi.org/10.1073/pnas.1215985110 PMID: 23267105

27. Tao Y, Zhao X, Wang X, Hathorn A, Hunt C, Cruickshank AW, et al. Large-scale GWAS in sorghum reveals common genetic control of grain size among cereals. Plant Biotechnol J. 2020; 18: 1093–1105. https://doi.org/10.1111/pbi.13284 PMID: 31659829

28. Zhang D, Li J, Compton RO, Robertson J, Goff VH, Epps E, et al. Comparative Genetics of Seed Size Traits in Divergent Cereal Lineages Represented by Sorghum (Panicoidae) and Rice (Oryzoidae). G3amp58 GenesGenomesGenetics. 2015; 5: 1117–1128. https://doi.org/10.1534/g3.115.017590 PMID: 25834216

29. Rhodes DH, Hoffmann L, Rooney WL, Herald TJ, Bean S, Boyles R, et al. Genetic architecture of kernel composition in global sorghum germplasm. BMC Genomics. 2017; 18: 15. https://doi.org/10.1186/s12864-016-3403-x PMID: 28056770

30. Brown SM, Hopkins MS, Mitchell SE, Senior ML, Wang TY, Duncan RR, et al. Multiple methods for the identification of polymorphic simple sequence repeats (SSRs) in sorghum [*Sorghum bicolor* (L.) Moench]. Theor Appl Genet. 1996; 93: 190–198. https://doi.org/10.1007/BF00225745 PMID: 24162217

31. Taramino G, Tarchini R, Ferrario S, Lee M, Pe' ME. Characterization and mapping of simple sequence repeats (SSRs) in *Sorghum bicolor*. Theor Appl Genet. 1997; 95: 66–72. https://doi.org/10.1007/s001220050533

32. Kong L, Dong J, Hart GE. Characteristics, linkage-map positions, and allelic differentiation of *Sorghum bicolor* (L.) Moench DNA simple-sequence repeats (SSRs). Theor Appl Genet. 2000; 101: 438–448. https://doi.org/10.1007/s001220051501

33. Schloss SJ, Mitchell SE, White GM, Kukatla R, Bowers JE, Paterson AH, et al. Characterization of RFLP probe sequences for gene discovery and SSR development in *Sorghum bicolor* (L.) Moench. Theor Appl Genet. 2002; 105: 912–920. https://doi.org/10.1007/s00122-002-0991-4 PMID: 12582917

34. Srinivas G, Satish K, Murali Mohan S, Nagaraja Reddy R, Madhusudhana R, Balakrishna D, et al. Development of genic-microsatellite markers for sorghum staygreen QTL using a comparative genomic approach with rice. Theor Appl Genet. 2008; 117: 283–296. https://doi.org/10.1007/s00122-008-0773-8 PMID: 18438637

35. Srinivas G, Satish K, Madhusudhana R, Seetharama N. Exploration and mapping of microsatellite markers from subtracted drought stress ESTs in Sorghum bicolor (L.) Moench. Theor Appl Genet. 2009; 118: 703–717. https://doi.org/10.1007/s00122-008-0931-z PMID: 19034408

36. Li M, Yuyama N, Luo L, Hirata M, Cai H. In silico mapping of 1758 new SSR markers developed from public genomic sequences for sorghum. Mol Breed. 2009; 24: 41–47. https://doi.org/10.1007/s11032-009-9270-2

37. Ramu P, Kassahun B, Senthilvel S, Ashok Kumar C, Jayashree B, Folkertsma RT, et al. Exploiting rice–sorghum synteny for targeted development of EST-SSRs to enrich the sorghum genetic linkage map. Theor Appl Genet. 2009; 119: 1193–1204. https://doi.org/10.1007/s00122-009-1120-4 PMID: 19669123

38. Li L, Fang Z, Zhou J, Chen H, Hu Z, Gao L, et al. An accurate and efficient method for large-scale SSR genotyping and applications. Nucleic Acids Res. 2017; 45: e88. https://doi.org/10.1093/nar/gkx093 PMID: 28184437

39. Li T, Fang Z, Peng H, Zhou J, Liu P, Wang Y, et al. Application of high-throughput amplicon sequencing-based SSR genotyping in genetic background screening. BMC Genomics. 2019; 20: 444. https://doi.org/10.1186/s12864-019-5800-4 PMID: 31159719

40. Yang J, Zhang J, Han R, Zhang F, Mao A, Luo J, et al. Target SSR-Seq: A Novel SSR Genotyping Technology Associate With Perfect SSRs in Genetic Analysis of Cucumber Varieties. Front Plant Sci. 2019; 10: 531. https://doi.org/10.3389/fpls.2019.00531 PMID: 31105728

41. Guo L, Yang Q, Yang J, Zhang N, Liu B, Zhu K, et al. MultiplexSSR: A pipeline for developing multiplex SSR-PCR assays from resequencing data. Ecol Evol. 2020; 10: 3055–3067. https://doi.org/10.1002/ece3.6121 PMID: 32211176

42. Zhang L, Liu D, Guo X, Yang W, Sun J, Wang D, et al. Investigation of genetic diversity and population structure of common wheat cultivars in northern China using DArT markers. BMC Genet. 2011; 12: 42. https://doi.org/10.1186/1471-2156-12-42 PMID: 21569312

43. Tomkowiak A, Bocianowski J, Radzikowska D, Kowalczewski PŁ. Selection of Parental Material to Maximize Heterosis Using SNP and SilicoDarT Markers in Maize. 2019; 15.

44. Jordan D, Tao Y, Godwin I, Henzell R, Cooper M, McIntyre C. Prediction of hybrid performance in grain sorghum using RFLP markers. Theor Appl Genet. 2003; 106: 559–567. https://doi.org/10.1007/s00122-002-1144-5 PMID: 12589557

45. Mindaye TT, Mace ES, Godwin ID, Jordan DR. Heterosis in locally adapted sorghum genotypes and potential of hybrids for increased productivity in contrasting environments in Ethiopia. Crop J. 2016; 4: 479–489. https://doi.org/10.1016/j.cj.2016.06.020

46. Amelework B, Shimelis H, Laing M. Genetic variation in sorghum as revealed by phenotypic and SSR markers: implications for combining ability and heterosis for grain yield. Plant Genet Resour. 2017; 15: 335–347. https://doi.org/10.1017/S1479262115000696

47. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009; 25: 1754–1760. https://doi.org/10.1093/bioinformatics/btp324 PMID: 19451168

48. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010; 20: 1297–1303. https://doi.org/10.1101/gr.107524.110 PMID: 20644199

49. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25: 2078–2079. https://doi.org/10.1093/bioinformatics/btp352 PMID: 19505943

50. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience. 2015; 4: 7. https://doi.org/10.1186/s13742-015-0047-8 PMID: 25722852

51. Willems T, Zielinski D, Yuan J, Gordon A, Gymrek M, Erlich Y. Genome-wide profiling of heritable and de novo STR variations. Nat Methods. 2017; 14: 590–592. https://doi.org/10.1038/nmeth.4267 PMID: 28436466

52. Raj A, Stephens M, Pritchard JK. fastSTRUCTURE: Variational Inference of Population Structure in Large SNP Data Sets. Genetics. 2014; 197: 573–589. https://doi.org/10.1534/genetics.114.164350 PMID: 24700103

53. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2018. Available: http://www.R-project.org.

54. Wickham H. ggplot2: Elegant Graphics for Data Analysis: Book Reviews. Springer-Verl N Y. 2011; 174: 245–246. https://doi.org/10.1111/j.1467-985X.2010.00676_9.x

55. RStudio Team. RStudio: Integrated Development Environment for R Version 1.1.456. RStudio, Inc.; 2016. Available: http://www.rstudio.com/.

56. Huson DH, Bryant D. Application of Phylogenetic Networks in Evolutionary Studies. Mol Biol Evol. 2006; 23: 254–267. https://doi.org/10.1093/molbev/msj030 PMID: 16221896

57. Adrian Dusa. venn: Draw Venn Diagrams. 2018. Available: https://CRAN.R-project.org/package=venn.

58. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011; 27: 2156–2158. https://doi.org/10.1093/bioinformatics/btr330 PMID: 21653522

59. Germplasm Resources Information Network. United States Department of Agriculture, Agricultural Research Service. Beltsville, MD; 2018. Available: http://www.ars-grin.gov/.

60. Bennett S. Solexa Ltd. Pharmacogenomics. 2004; 5: 433–438. https://doi.org/10.1517/14622416.5.4.433 PMID: 15165179

61. Oróstica KY, Verdugo RA. chromPlot: visualization of genomic data in chromosomal context. Bioinformatics. 2016; 32: 2366–2368. https://doi.org/10.1093/bioinformatics/btw137 PMID: 27153580

62. Botstein D, White RL, Skolnick M, Davis RW. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet. 1980; 32: 314–331. PMID: 6247908

63. Sagnard F, Deu M, Dembélé D, Leblois R, Touré L, Diakité M, et al. Genetic diversity, structure, gene flow and evolutionary relationships within the *Sorghum bicolor* wild–weedy–crop complex in a western African region. Theor Appl Genet. 2011; 123: 1231–1246. https://doi.org/10.1007/s00122-011-1662-0 PMID: 21811819

64. Ott A, Trautschold B, Sandhu D. Using Microsatellites to Understand the Physical Distribution of Recombination on Soybean Chromosomes. Ingvarsson PK, editor. PLoS ONE. 2011; 6: e22306. https://doi.org/10.1371/journal.pone.0022306 PMID: 21799819

65. Saintenac C, Falque M, Martin OC, Paux E, Feuillet C, Sourdille P. Detailed Recombination Studies Along Chromosome 3B Provide New Insights on Crossover Distribution in Wheat (Triticum aestivum L.). Genetics. 2009; 181: 393–403. https://doi.org/10.1534/genetics.108.097469 PMID: 19064706

66. Bhargava A, Fuentes FF. Mutational Dynamics of Microsatellites. Mol Biotechnol. 2010; 44: 250–66. https://doi.org/10.1007/s12033-009-9230-4 PMID: 20012711

67. Smith JSC, Chin ECL, Shu H, Smith OS, Wall SJ, Senior ML, et al. An evaluation of the utility of SSR loci as molecular markers in maize (Zea mays L.): comparisons with data from RFLPS and pedigree. Theor Appl Genet. 1997; 95: 163–173. https://doi.org/10.1007/s001220050544

68. Weber-Lehmann J, Schilling E, Gradl G, Richter DC, Wiehler J, Rolf B. Finding the needle in the haystack: Differentiating "identical" twins in paternity testing and forensics by ultra-deep next generation sequencing. Forensic Sci Int Genet. 2014; 9: 42–46. https://doi.org/10.1016/j.fsigen.2013.10.015 PMID: 24528578