







Article

Low-Density Reference Fingerprinting SNP Dataset of CIMMYT Maize Lines for Quality Control and Genetic Diversity Analyses

Jingtao Qu ^{1,2}, Alberto A. Chassaigne-Ricciulli ², Fengling Fu ¹, Haoqiang Yu ¹ , Kate Dreher ², Sudha K. Nair ³, Manje Gowda ⁴ , Yoseph Beyene ⁴, Dan Makumbi ⁴ , Thanda Dhliwayo ² , Felix San Vicente ² , Michael Olsen ⁴, Boddupalli M. Prasanna ⁴ , Wanchen Li ^{1,*} and Xuecai Zhang ^{2,*}

¹ Maize Research Institute, Sichuan Agricultural University, Chengdu 611130, China

² International Maize and Wheat Improvement Center (CIMMYT), El Batan, Texcoco 56237, Mexico

³ Asia Regional Maize Program, International Maize and Wheat Improvement Center (CIMMYT), ICRISAT Campus, Patancheru, Hyderabad 502324, Telangana, India

⁴ International Maize and Wheat Improvement Center (CIMMYT), Village Market, P.O. Box 1041, Nairobi 00621, Kenya

* Correspondence: aumdmys@sicau.edu.cn (W.L.); xc.zhang@cgiar.org (X.Z.)

Abstract: CIMMYT maize lines (CMLs), which represent the tropical maize germplasm, are freely available worldwide. All currently released 615 CMLs and fourteen temperate maize inbred lines were genotyped with 180 kompetitive allele-specific PCR single nucleotide polymorphisms to develop a reference fingerprinting SNP dataset that can be used to perform quality control (QC) and genetic diversity analyses. The QC analysis identified 25 CMLs with purity, identity, or mislabeling issues. Further field observation, purification, and re-genotyping of these CMLs are required. The reference fingerprinting SNP dataset was developed for all of the currently released CMLs with 152 high-quality SNPs. The results of principal component analysis and average genetic distances between subgroups showed a clear genetic divergence between temperate and tropical maize, whereas the three tropical subgroups partially overlapped with one another. More than 99% of the pairs of CMLs had genetic distances greater than 0.30, showing their high genetic diversity, and most CMLs are distantly related. The heterotic patterns, estimated with the molecular markers, are consistent with those estimated using pedigree information in two major maize breeding programs at CIMMYT. These research findings are helpful for ensuring the regeneration and distribution of the true CMLs, via QC analysis, and for facilitating the effective utilization of the CMLs, globally.

Keywords: tropical maize; CIMMYT maize line (CML); reference SNP dataset; fingerprinting; quality control; heterotic group



Citation: Qu, J.; Chassaigne-Ricciulli, A.A.; Fu, F.; Yu, H.; Dreher, K.; Nair, S.K.; Gowda, M.; Beyene, Y.; Makumbi, D.; Dhliwayo, T.; et al. Low-Density Reference Fingerprinting SNP Dataset of CIMMYT Maize Lines for Quality Control and Genetic Diversity Analyses. *Plants* **2022**, *11*, 3092. <https://doi.org/10.3390/plants11223092>

Academic Editors: Jeong-Dong Lee, Ivan Kreft and Hyun Jo

Received: 20 October 2022

Accepted: 10 November 2022

Published: 14 November 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Maize is one of the most important crops and is widely cultivated all over the world. Maize production of the tropically adapted germplasm occupied 30% of global maize production [1], and it is the preferred staple food for 900 million people living on less than 2 US\$ a day, primarily in sub-Saharan Africa, Latin America, and Asia. The International Maize and Wheat Improvement Center (CIMMYT), a non-profit international agricultural research and training organization, seeks to develop improved maize inbred lines and hybrids, with superior performance and multiple stress tolerance, in various agro-ecological zones across Africa, Latin America, and Asia in order to improve maize productivity for resource-constrained smallholder farmers [2]. Between 1984 and 2022, 615 CIMMYT maize lines (CMLs), adapted to the tropical/subtropical maize production environments, had been developed and released as international public goods and were accessible to all. These CMLs represent the broad genetic diversity of improved tropical maize germplasm in the

world, and they are freely available to both public and private sector breeders worldwide under the standard material transfer agreement, via the CIMMYT Germplasm Bank (<https://www.cimmyt.org/resources/seed-request/#maize> (accessed on 13 May 2021)). The development of a reference fingerprinting molecular marker dataset of all currently released CMLs is helpful in order to perform molecular marker-based quality assurance (QA) and quality control (QC) analysis on the increased seed stocks and to assure their genetic purity and identity. In addition, it is also useful for the breeders to better understand how to utilize the currently released CMLs in their breeding programs, based on the genetic diversity analysis and using the reference fingerprinting molecular marker dataset.

The CMLs have been widely requested and used by both public and private institutions all over the world [3]; they are public goods and have significantly contributed to improving abiotic and biotic stress tolerance and grain quality in both temperate and tropical maize [2,4]. The large number of seed requests coming from all over the world each year bring great challenges to the preservation, maintenance, and distribution of the germplasm of CMLs. The QC of seeds in maize breeding and seed production aims to ensure that the currently used inbred/parental lines are genetically identical to the original source, the seed has required genetic purity, the hybrids are indeed derived from the specified parents, and that the seed lots are meeting all the seed quality standards, etc. [5,6].

Compared to the traditional QC approaches of morphological comparison and isozymes electrophoresis, the molecular marker-based QC—i.e., DNA fingerprinting performed by genotyping at defined molecular markers that cover all 10 chromosomes of maize—is rapid, accurate, cost-effective, and independent to the environment. Over the past several decades, different molecular marker genotyping technologies, including restriction fragment length polymorphisms, simple sequence repeats, and single nucleotide polymorphisms (SNPs), have been applied in DNA fingerprinting and molecular marker-based QC analysis in maize [7–9]. Recently, SNPs have become the most promising genotyping technology for DNA fingerprinting, molecular marker-based QC analysis, and database construction, due to their high coverage of the genome, codominance, known chromosomal locations, and the potential for high-throughput and automated analysis. The KBiosciences' Competitive Allele-Specific PCR system (KASPar; <http://www.kbioscience.co.uk> (accessed on 1 May 2022)) is a single-plex SNP genotyping platform that has been widely used for molecular marker-based QC analysis, linkage mapping, fine-mapping, and marker-assisted selection in maize [10–12]. The KASPar is an ideal SNP genotyping platform for molecular marker-based QC analysis because it can be used to quickly genotype tens of thousands of samples, with tens to hundreds of markers, and it does not require large computing resources and time-consuming bioinformatics analyses, thus allowing breeders to receive the genotypic data and analysis results in a shorter time than other genotyping platforms. In addition, the cost per sample of KASPar genotyping platforms is more cost-effective than chip-based SNP technologies, such as Infinium, GoldenGate, and Affymetrix Axiom [13–20]. A framework of QC genotyping, with a recommended 50 or 100 KASPar SNPs, has been established in several practical maize breeding programs at CIMMYT [5,6]. A QC genotyping approach was also developed to genotype CML1-561 with the DArT-Seq technology [21], and a total of 88,600 SNP markers were generated on each CML, in which 80 and 10 SNPs were selected as the “broad QC” and “rapid QC” sets, respectively. However, the reference fingerprints of all of the released CMLs to date are not available publicly.

The results of numerous germplasm characterization studies have shown that chip-based SNP genotyping platforms are promising and effective for performing genetic diversity analysis in collections of maize germplasms with different genetic backgrounds, although the cost per sample of chip-based SNP genotyping platforms is still high [10]. Three generations of the maize haplotype map (HapMap) had been reported by re-sequencing 27 to 1218 maize inbred lines, representing the global maize genetic diversity [22–24]. Genotyping-by-Sequencing (GBS) is one of the most widely used simplified reduced-representation sequencing approaches, it is high-throughput and cost-effective; the geno-

typing cost per sample was decreased by reducing genome complexity with restriction enzymes, and multiplex sequencing the samples by using the barcode system [25–28]. A molecular characterization study on all of the CMLs before CML539 was carried out using GBS SNPs to better understand how to utilize these CMLs [2]. However, new CMLs are continuously being released and the genotypic dataset has to be updated to include new lines. A cost-effective genotyping platform with a short turnaround time, such as KASPar, is required to genotype all of the CML candidates before release in order to generate the reference fingerprinting SNP dataset for all CMLs, as a whole. This can then be used to examine the genetic purity and identity of the CML candidates, and will avoid the new CMLs having close genetic relationships with existing CMLs.

In the present study, all of the currently released 615 CMLs (available till 2022) were genotyped with 180 KASPar QC SNPs, routinely used at CIMMYT, and two to five biological replications per CML were genotyped. The main objectives of the present study were to: (1) develop the reference fingerprinting SNP dataset of all currently released 615 CMLs; (2) implement QC analysis on two to five biological replications of each CML in order to identify suspicious CMLs in the CIMMYT gene bank with issues of genetic purity, identity, seed contamination or mislabeling during genotyping; (3) conduct genetic diversity analysis on all of the currently released 615 CMLs and 15 temperate maize inbred lines.

2. Results

2.1. Summary of KASP Marker Characteristics

All 180 KASP SNP markers were evenly anchored on the reference genome of B73_RefGen_V4, based on their flanking sequence information. In total, 179 SNP markers were anchored on the ten maize chromosomes, and one SNP marker was anchored on the B73_RefGen_V4_Contig104 (Figure 1). The number of SNPs on each chromosome ranged from 12, on chromosome 4, to 22 on chromosome 6, with a mean of 17.9 markers per chromosome (Table 1). The physical distance of two adjacent SNPs on each chromosome varied from 114 bp on chromosome 7, between SNP PZE-101093951 and SNP PZE0186065237, to 58.54 Mb on chromosome 4, between SNP PZE-101093951 and SNP PZE0186065237, with an average value of 12.30 Mb (Table S1). Among the 180 KASP SNPs, seven SNPs were breeder-ready markers for provitamin A (PVA); three SNPs were breeder-ready markers for resistance to maize streak virus (MSV); 10 SNPs were breeder-ready markers for resistance to maize lethal necrosis (MLN); 10 SNPs were breeder-ready markers for quality protein maize (QPM); and 2 SNPs were breeder-ready markers for resistance to tar spot complex (TSC).

Table 1. Summary statistics for the 180 KASP markers in the first round of genotyping.

Chromosome	No. of Markers	Minor Allele Frequency	Heterozygosity Rate (%)	Missing Rate (%)
1	21	0.34	1.16	2.27
2	18	0.35	1.18	3.52
3	19	0.32	1.49	3.77
4	12	0.32	1.07	2.44
5	19	0.34	0.86	2.76
6	22	0.31	2.07	3.07
7	15	0.28	0.68	4.77
8	20	0.37	1.09	3.62
9	13	0.39	0.79	2.41
10	20	0.19	0.73	2.52
Ctg104	1	0.29	0.51	2.83
Total	180	-	-	-
Avg.	17.90	0.32	1.14	3.11

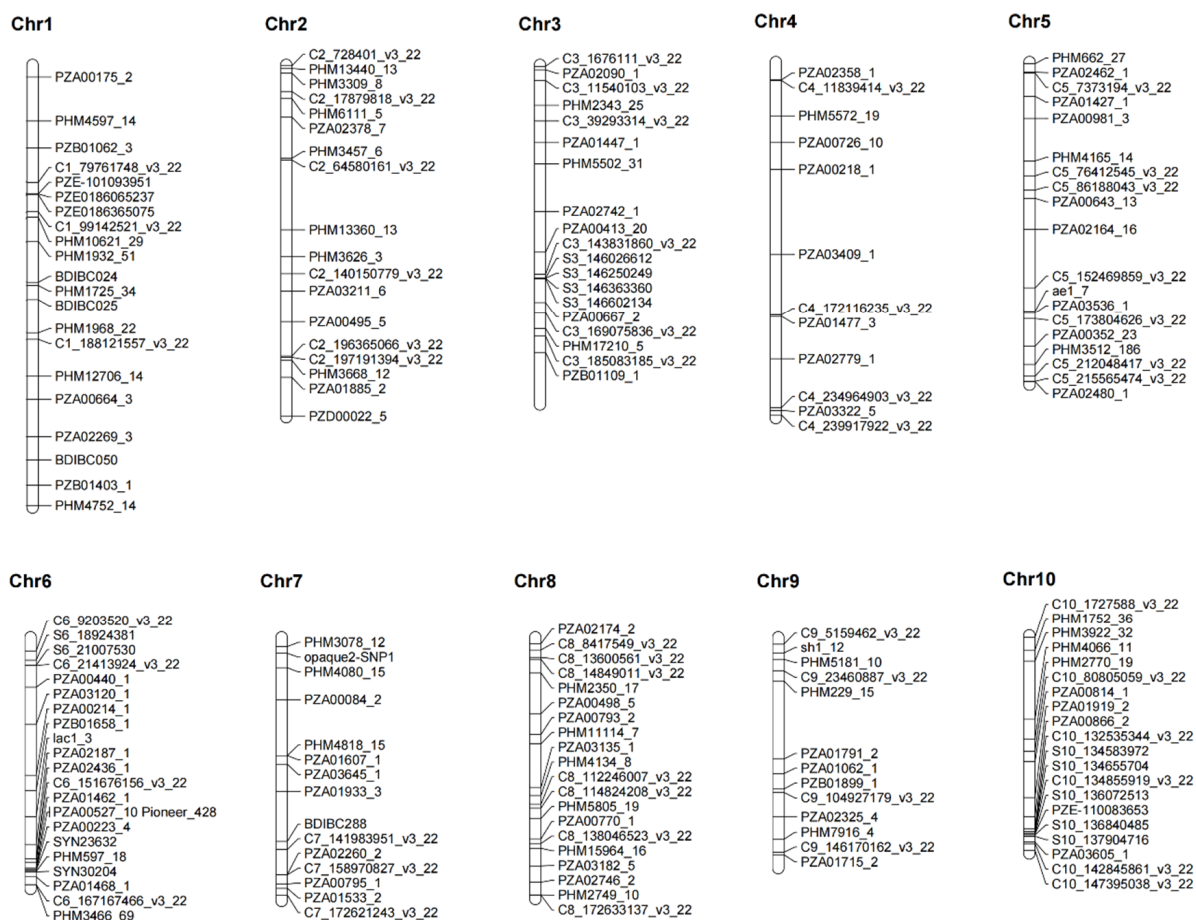


Figure 1. The distribution of 180 SNPs on the ten maize chromosomes.

In total, 360 alleles were detected for all the 180 KASP SNPs, with two alleles per SNP, as was expected. The SNP mutation type analysis indicated that 59.9% of the SNP mutations belong to the A/G, T/C type; 20.9% of the SNP mutations belong to the A/C, G/T type; and 19.2% of the SNP mutations belong to the A/T, C/G type. SNP is a type of mutation that occurs in a single nucleotide in the genome. In the present study, the number of KASP markers with the allelic types of A/G and T/C was higher than those with other allelic types. The summary information for all 180 KASP SNPs in the first round of genotyping dataset are shown in Figure 2a. The average heterozygosity rate across all the SNPs was 1.14%, with a range from 0 to 19.73% per SNP. In total, 172 of the 180 SNPs had a heterozygosity rate lower than 5% (Table S1). The remaining eight SNPs with heterozygosity rates greater than 5%—i.e., C6_9203520_v3_22, PZA02746_2, C3_169075836_v3_22, C2_64580161_v3_22, PZA01919_2, PHM2350_17, C6_151676156_v3_22, and PZA00413_20—were excluded from the further QC analysis and development of the reference fingerprinting SNP dataset. The average missing rate across all the SNPs was 5.55%, with a range from 0.42% to 37.75% per SNP. In total, 164 of the 180 SNPs had their missing rates lower than 10%. The remaining 16 SNPs with missing rates > 10%—i.e., C3_11540103_v3_22, C7_141983951_v3_22, PZA03211_6, C7_158970827_v3_22, PZE-110083653, PZA02746_2, C2_64580161_v3_22, C6_151676156_v3_22, PZA00223_4, C10_132535344_v3_22, C6_167167466_v3_22, C8_114824208_v3_22, C4_239917922_v3_22, PHM13440_13, C5_152469859_v3_22, and C5_86188043_v3_22—were excluded from further QC analysis and development of the reference fingerprinting SNP dataset. The average MAF across all the SNPs was 0.32, with a range from 0 to 0.50. In total, 170 of the 180 SNPs had MAF values greater than 0.05. The remaining 10 SNPs with MAF values lower than 0.05—i.e., PZE-101093951, PZE-110083653, C7_141983951_v3_22, C7_158970827_v3_22, C10_132535344_v3_22, S6_18924381, S10_134655704, S10_134583972, S10_136840485, and S10_137904716—were excluded from

further QC analysis and development of the reference fingerprinting SNP dataset. Finally, 152 SNPs, which had heterozygosity rates lower than 5%, missing rates lower than 10%, and MAF values greater than 0.05, were retained and used in the further QC analysis and development of the reference fingerprinting SNP dataset.

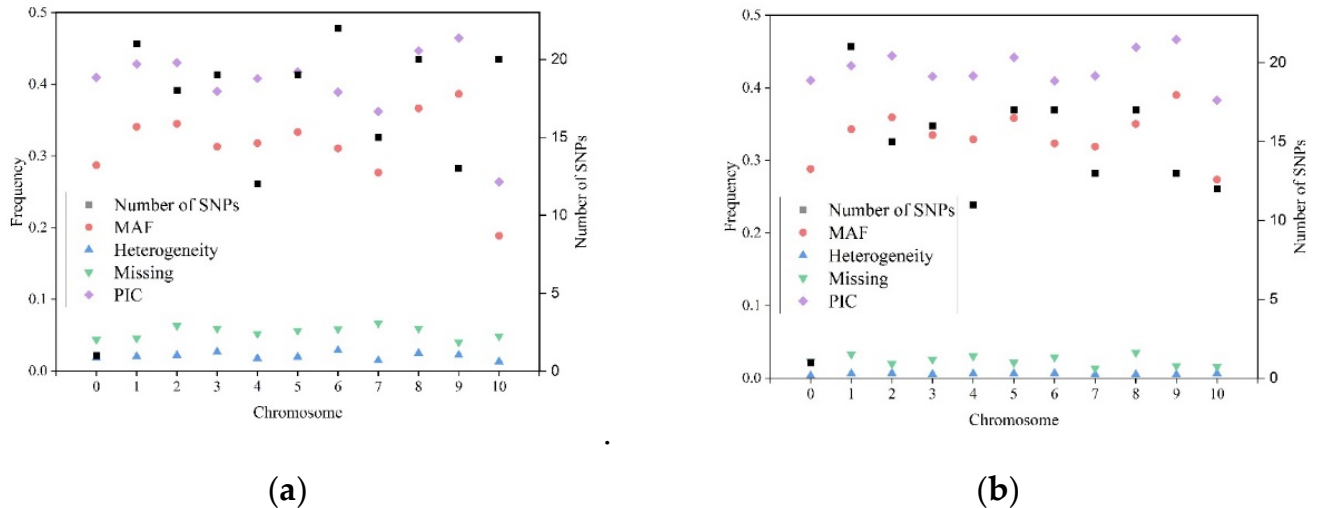


Figure 2. Summary of the heterogeneity, minor allele frequency (MAF), missing rate, and polymorphic information content (PIC) of (a) 180 SNPs across 1200 samples and (b) 152 high-quality SNPs across 603 CMLs. Chromosome assignments are indicated; where position anchored on non-chromosome, the chromosome is designated as “0”; The heterogeneity, MAF, percentage of Missing value, PIC was shown in left y -axis, the number of SNPs for each chromosome was shown in right y -axis.

The genes associated with the KASP markers were used to perform GO analysis. In total, 164 of the 180 KASP SNP markers were associated with 208 genes, including 151 KASP SNP markers located in the genic region, 15 KASP SNP markers located within the two kb upstream intergenic region of a gene, and 34 KASP SNP markers located within the two kb upstream or downstream intergenic region of a gene. Among the 208 genes, 181 of them were annotated with potential function, such as Zm00001d018971, a regulatory protein of opaque-2 and Zm00001d034385, aldehyde oxidase 3, associated with drought stress tolerance (Table S2). In total, 204 of the 208 genes were annotated into 134 GO terms, divided into three categories: 66 terms involved in biological processes, 43 GO terms associated with cellular components, and 25 GO terms correlated with molecular functions (Figure 3; Table S3). In addition, four GO terms were associated with the development and regulation of quality protein maize, including GO:0044267 (cellular protein metabolic process), GO:0006464 (cellular protein modification process), GO:0019538 (protein metabolic process), and GO:0005515 (protein binding).

2.2. Quality Control Analysis Results

The genetic purity analysis revealed that the average heterozygosity rate across two biological replications of all the 603 CMLs was 2.29%. In total, 556 of the 1206 samples, i.e., 46.10%, had heterozygosity rates lower than 1%, and the seeds of these samples achieved the required purity level. In addition, 1058 of the 1206 samples, i.e., 87.73%, had heterozygosity rates lower than 5%, indicating that the seeds of these samples achieved the basic purity level. The 151 CMLs with a heterozygosity rate of at least one biological replication greater than 3% were used for the second round of genotyping, and are listed in Table S4.

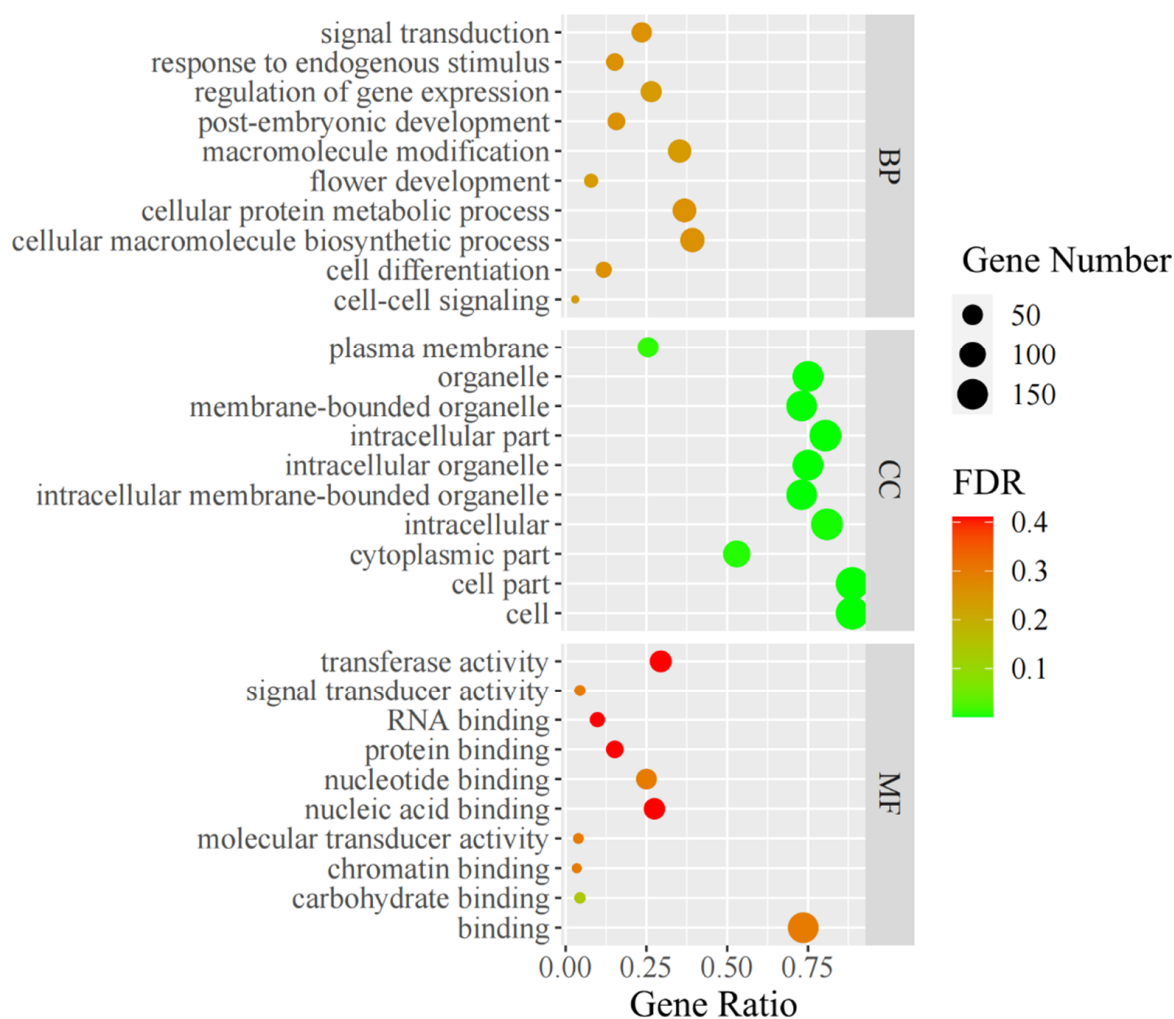


Figure 3. GO enrichment of genes associated with KASP markers. The top 10 of each GO class were shown. BP indicated the class of biological processes, CC indicated the class of cellular components, and MF indicated the class of molecular functions.

The genetic identity analysis revealed that the average similarity rate between the two biological replications across all the 603 CMLs was 98.54%, indicating that most of the CMLs were essentially identical for the two biological replications genotyped in the first round of genotyping. The similarity rates between the two biological duplications of 485 CMLs were greater than 99%, indicating that these CMLs were essentially identical. The similarity rates between the two biological duplications of 93 CMLs were greater than 95%, indicating that these CMLs were identical, with slight genetic divergences caused by genetic drift. The rest of the 25 CMLs had similarity rates lower than 95% between the two independent biological replications, indicating either a genetic identity issue of the seeds or mislabeling during genotyping. Among these 25 CMLs, eleven CMLs had relatively high similarity rates, of greater than 85%, and relatively high heterozygosity rates, of greater than 5% in at least one biological replication; this result indicated that the genetic identity issue of these CMLs, including CML51B, CML95B, CML196A, CML208A, CML437A, CML473B, CML474A, CML487B, CML519B, CML521B, and CML523A, were probably caused by the issue of impurity seeds. The rest of the fourteen CMLs, including CML76A, CML80A, CML82A, CML83A, CML96B, CML100A, CML101A, CML107A, CML111A, CML112A, CML113A, CML116B, CML240A, and CML418B, had similarity rates lower than 85% between the two independent biological replications, and six of these fourteen CMLs, including CML76A, CML96B, CML100A, CML107A, CML111A, and

CML240A, had relative high heterozygosity rates of greater than 5%. The genetic identity issue of these six CMLs was most likely caused by the issue of impurity seeds, and the genetic identity issue of the other eight CMLs was most likely caused by mislabeling during genotyping. In these 25 CMLs, the genetic distance between the two biological duplications of a few CMLs was greater than the between different CMLs, and they were selected in the second round of genotyping.

Among the 540 samples of 180 CMLs genotyped in the second round, the heterozygosity rates of 482 samples, corresponding to at least one biological replication of 179 CMLs, were lower than 5%, including 323 samples with heterozygosity rates lower than 1%, and 111 samples with heterozygosity rates between 1% to 3%. The biological replications with heterozygosity rates lower than 3% from both rounds of genotyping were used to build the reference SNP dataset for the fingerprinting and QA/QC analyses. In addition, 13 of the 25 CMLs with similarity rates lower than 95% between the two biological replications in the first round of genotyping were identified with mislabeling issues in one biological replication in the first round of genotyping (Figure 4), based on the neighbor-joining tree results across all five biological replications. The reference fingerprinting SNP dataset for these thirteen CMLs, including CML76A, CML80A, CML82A, CML83A, CML96B, CML100A, CML101A, CML107A, CML111A, CML112A, CML113A, CML116B, CML240A, was built with the reliable genotypic data of the four biological replications from both rounds of genotyping.

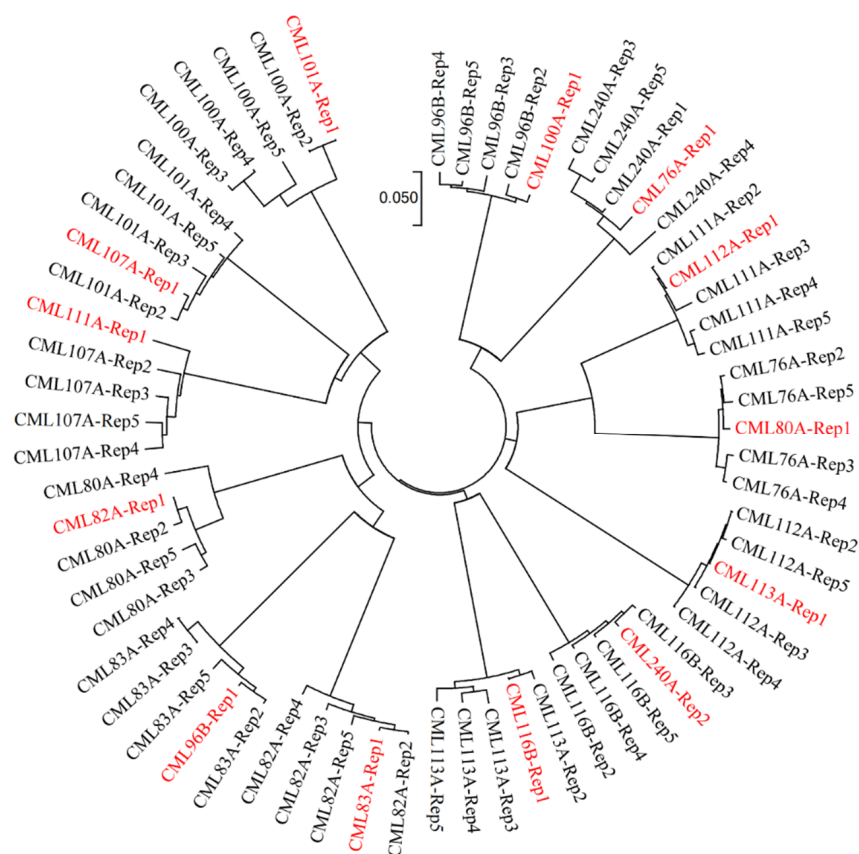


Figure 4. Neighbor-joining tree for the thirteen mislabeled CMLs in the first round of genotyping, estimated with the fingerprinting SNP dataset of five biological duplications from both rounds of genotyping. Rep1 and Rep2 represent the two biological duplications from the first round of genotyping, and Rep3, Rep4, and Rep5 represent the three biological duplications from the second round of genotyping. Mislabeled biological duplications were indicated in red color.

2.3. Low-Density Reference SNP Dataset for Fingerprinting and QA/QC Analysis

The low-density reference fingerprinting SNP dataset was developed for all of the currently released 615 CMLs through the integration of the reliable genotypic data of all the biological replications for each CML. For CML1A to CML603A, 152 SNPs were used to develop the reference SNP dataset; the SNPs in the first round of genotyping with a heterozygosity rate greater than 5%, a missing rate greater than 10%, and a MAF lower than 0.05, were excluded from the reference SNP dataset. For CML604A to CML615A, only one biological duplication for each CML was genotyped with 85 of the 152 SNPs to develop the reference SNP dataset. In total, the reference SNP dataset included 584 CMLs with a heterozygosity rate lower than 5% and an average similarity rate greater than 95% between the biological replications for each CML; 22 CMLs with a heterozygosity rate lower than 5%, and an average similarity rate lower than 95% between the biological replications for each CML; and CML179A with a heterozygosity rate of 10.08% and an average similarity rate lower of 97.94% between the biological replications (Table S5). Genotyping data of CML543B, CML569B, CML585A, and CML605B were missing in the current reference SNP dataset. Among the 22 CMLs with heterozygosity rates lower than 5% and average similarity rates lower than 95% between the biological replications, the similarity rates between the biological duplications of eight CMLs were lower than the similarity rate with the other CML.

The summary information of the 152 SNPs in the reference fingerprinting genotypic dataset in all the 584 reliable CMLs is shown in Figure 2b. The number of SNPs per chromosome ranged from 11 on chromosome 4 to 21 on chromosome 1, the average missing rate across all the SNPs was 2.50%, the average heterozygosity rate across all the SNPs was 0.60%, and the average MAF across all the SNPs was 0.34. The 152 SNPs in the reference fingerprinting genotypic dataset had lower missing and heterozygosity rates, and higher MAF than those in the original 180 SNPs dataset. In the reference fingerprinting SNP dataset, the average missing rate across all the reliable 584 CMLs was 2.50%, and the average heterozygosity rate across all the reliable 584 CMLs was 0.60% (Table S5).

The genetic distance result for the 25 pairs of CML with pedigree-based close relationships is shown in Table 2, where the genetic distance between each pair of CMLs varied from 0.02 to 0.44, with a mean of 0.14. This result indicates that the high-quality reference SNP dataset developed in the present study is appropriate for fingerprinting, QA/QC, and genetic diversity analyses, and it can accurately reflect the relationships of the genetically close related CMLs.

2.4. Genetic Diversity Analysis Conducted with the Reference SNP Dataset

The result of the principal component analysis in all of the currently released 615 CMLs, along with the fourteen temperate maize inbred lines, is shown in Figure 5. The first three principal components explained 7.74% of the genetic variations, in total, and the genetic variation explained by the first principal component, the second principal component, and the third component was 3.12%, 2.58%, and 2.24%, respectively. A clear genetic divergence was observed between temperate and tropical inbred lines, while a clear clustering pattern was not observed between the BSSS and NSSS heterotic groups in the temperate maize, due to the small sample size of the temperate subgroup affecting the accuracy of cluster pattern analysis within this subgroup. The Highland Tropical subgroup, on the whole, was separated from the Lowland Tropical and Subtropical/Mid-altitude subgroups, whereas the Lowland Tropical subgroup and the Subtropical/Mid-altitude subgroup overlaid partially with each other.

Table 2. The genetic distance between the original parent and the converted line.

Original Parent	Converted Line	Relationship	Genetic Distance
CML78A	CML512A	CML78A IR ^a	0.05
CML202B	CML513B	CML202B IR	0.44
CML204B	CML514B	CML204A IR	0.04
CML247A	CML515A	CML247A IR	0.08
CML254A	CML516A	CML254A IR	0.08
CML312A	CML517A	CML312A IR	0.18
CML373B	CML518B	CML373B IR	0.05
CML384B	CML519B	CML384B IR	0.05
CML390A	CML520A	CML390A IR	0.15
CML395B	CML521B	CML395B IR	0.03
CML444B	CML522B	CML444B IR	0.09
CML445A	CML523A	CML445A IR	0.02
CML264A	CML503A	CML264A Q ^b	0.04
CML242A	CML524A	CML242A Q	0.04
CML244A	CML525A	CML244A Q	0.07
CML246B	CML526B	CML246B Q	0.10
CML349A	CML527B	CML349A Q	0.41
CML352A	CML528A	CML352A Q	0.07
CML354A	CML529A	CML354A Q	0.08
CML312A	CML537A	MAS(CML206A/CML312A)-23-2-1-1-B	0.31
CML312A	CML539A	MAS(MSR/CML312A)-117-2-2-1-B	0.20
CML444B	CML566B	(LAPOSTASEQ-C7-F96-1-2-1-1-B*3/CML444B//CML444B)-DH16-B	0.11
CML539A	CML567A	(LAPOSTASEQ-C7-F71-1-2-1-2-B*3/CML539A//CML539A)-DH3-B	0.42
CML539A	CML568A	(LAPOSTASEQ-C7-F71-1-2-1-2-B*3/CML539A//CML539A)-DH20-B	0.19
CML444B	CML570B	(LAPOSTASEQ-C7-F71-1-2-1-2-B*3/CML444B//CML444B)-DH49-B	0.14

^a IR: imidazolinone resistant. ^b Q: quality protein maize. * indicated harvest with bulk method.

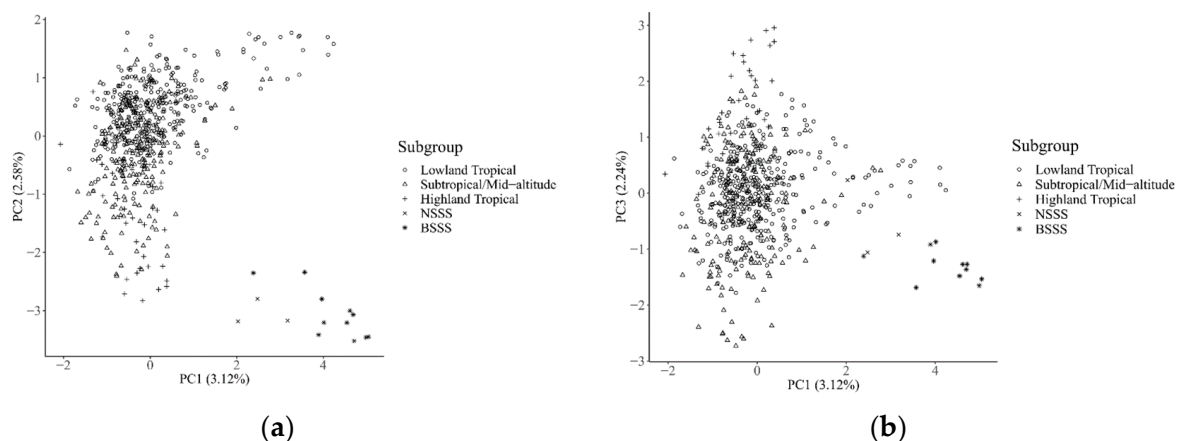


Figure 5. The scatter plot of the first three principal components for 584 reliable CIMMYT inbred lines according to adaptation: Lowland Tropical, Subtropical/Mid-Latitude, and Highland Tropical, and Temperate subgroup of 14 inbred lines belonging to BSSS and NSSS heterotic groups. (a) indicated PC1 versus PC2, and (b) indicated PC1 versus PC3.

The pairwise genetic distance between all the 584 reliable CMLs varied between 0 and 0.63, with an average of 0.43 (Table S6). The results showed that 99.43% of the pairs of CMLs had genetic distances greater than 0.30, including 92.49% that had genetic distances between 0.30 to 0.50, and 6.94% with genetic distances greater than 0.50. Less than 1% of the pairs of CMLs had genetic distances lower than 0.30, including 0.49% ranging between 0.10 and 0.30; 0.08% of them ranging between 0 and 0.10; and ten pairwise of CMLs with genetic distances equal to 0, i.e., CML117A-CML133A, CML138B-CML367B, CML171B-CML172B, CML204B-CML208A, CML214A-CML216A, CML283A-CML284A, CML377B-CML383B, CML387A-CML504B, CML472B-CML474A, and CML563A-CML579A. Of the 86 pairwise of CMLs with genetic distances lower than 0.05, 34 pairwise of CMLs were siblings, and 14 pairwise of CMLs were selected from the same population. This information indicates that most of the lines in the entire panel are either not related or only distantly related to each other and reflects the high genetic diversity of the CML germplasm source. However, it also indicates that the genetic distances between the CML candidates and the existing CMLs should be greater than 0.05 to avoid releasing genetically close-related new CMLs.

The average genetic distances between the different subgroups were calculated. The results indicated that the genetic distance between the tropical and temperate maize was greater than those among the three tropical subgroups. The genetic distance between the Temperate subgroup and each of the three tropical subgroups was 0.51. The genetic distance among each pair of all three tropical subgroups was 0.43 or 0.44.

2.5. Heterotic Pattern in Two Major Maize Breeding Programs

The analysis of the heterotic groups for each CIMMYT maize breeding program can facilitate the exchange and utilization of global germplasm resources. The 615 CMLs that are currently available have been divided into heterotic group A (HGA) and heterotic group B (HGB), according to the information of pedigrees and combining abilities. Within the Lowland tropical-Mexico (LLT) and East Africa (EA) CIMMYT maize breeding programs, a collection of key founder lines, including the most frequently used breeding lines and testers, were selected to investigate their genetic relatedness with each other (Figures 6 and 7). For LLT, 18 key founder CMLs, including 11 CMLs belonging to HGA and 7 CMLs belonging to HGB, were used to analyze the heterotic pattern in this breeding program. The average pairwise genetic distance within HGA, within HGB, and between HGA and HGB, were 0.28, 0.40, and 0.43, respectively. Within HGA, the CMLs are primarily derived from the Tuxpeño population, all the CMLs were closely related to each other, with a maximum genetic distance of 0.34, with the exception of two early-developed lines of CML247A and CML254A (Figure 6). Within HGB, fewer CMLs were selected, and most CMLs were distantly related to each other. In the LLT breeding program, the average pairwise genetic distance between the heterotic groups was greater than those within the heterotic groups, and the CMLs from the same heterotic group were more closely related to each other, particularly in HGA; this indicates that the heterotic pattern is being developed in this breeding program through breeding selection and recycling of a core set of key founder line.

For EA, 44 key founder CMLs, including 28 CMLs belonging to HGA and 16 CMLs belonging to HGB, were used to analyze the trends of heterotic patterns in this breeding program. The pedigree-based heterotic pattern was consistent with that estimated using the reference fingerprinting SNP dataset, with the exception of three CMLs mixed between the two heterotic groups, i.e., CML519B, CML544B, CML567A. This indicates that the CMLs from the same heterotic group were more closely related, and heterotic patterns are being developed in EA. In addition, a few key founder lines developed by the LLT breeding program, i.e., CML495A, CML494B, and CML576B, are also being used in EA to broaden the genetic diversity. The heterotic group information of these introduced lines is consistent with that of the EA lines, showing the feasibility of harmonization of heterotic patterns and utilization of founder lines across breeding programs, via the global germplasm exchange.

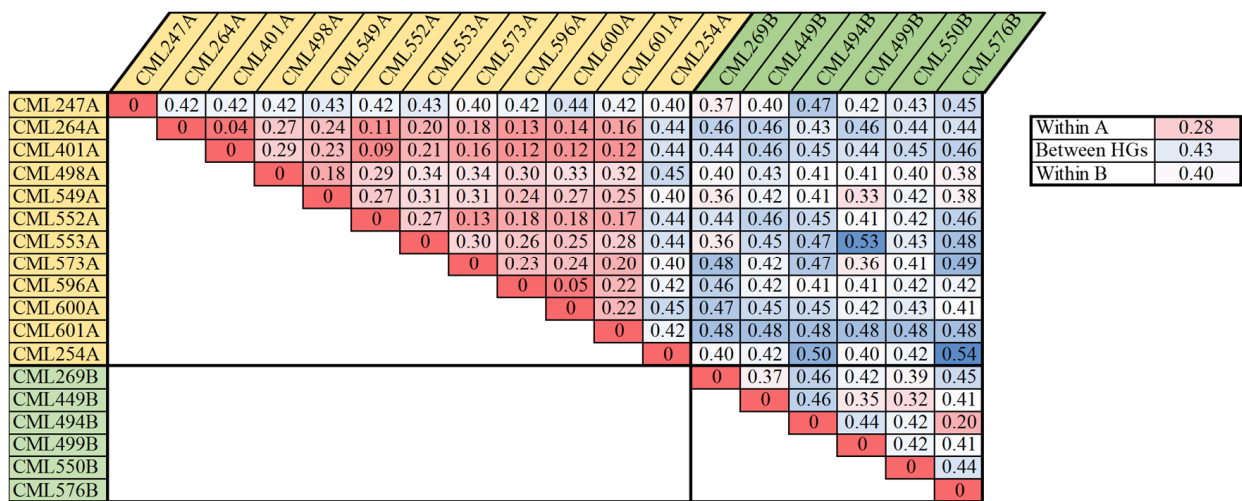


Figure 6. Heat map of the pairwise genetic distances between the 18 key founder CMLs of the Lowland tropical-Mexico (LLT) breeding program, 11 CMLs belonging to Heterotic Group “A”, and 7 CMLs belonging to Heterotic Group “B”.

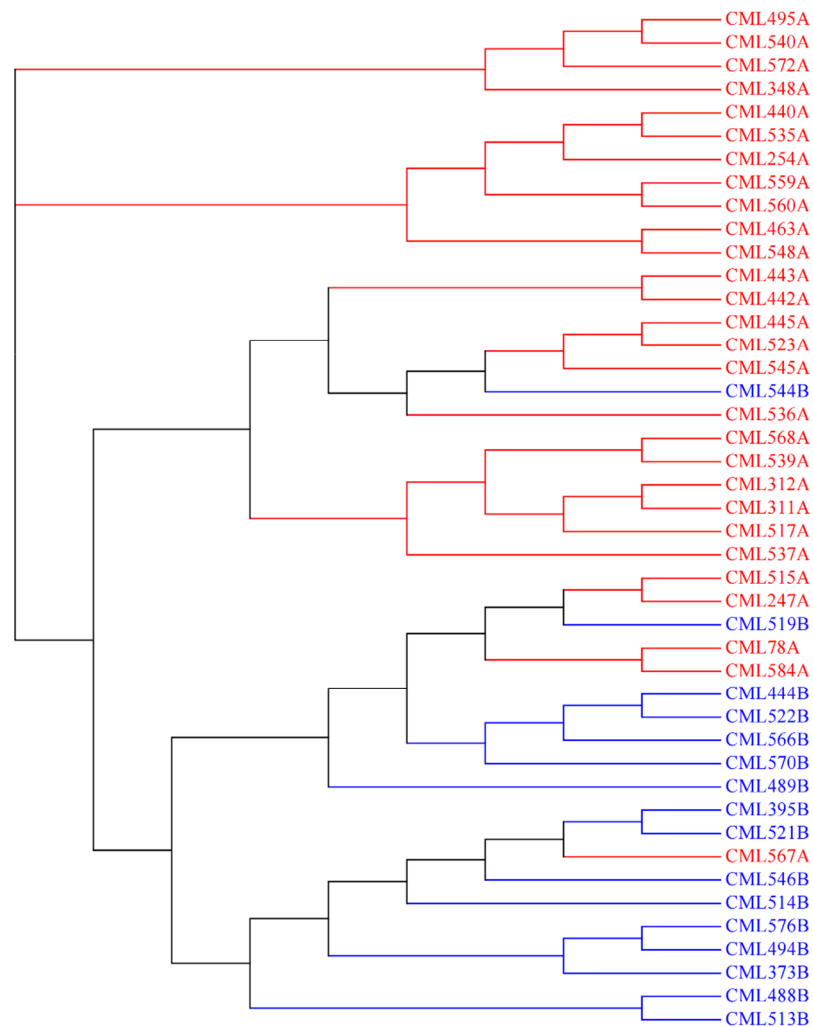


Figure 7. Neighbor-Joining tree based on IBS' genetic distance for the 44 key founder CMLs of the East Africa (EA) maize breeding program. CMLs from heterotic groups A and B are represented by red and blue, respectively.

3. Discussion

The collection of currently released CMLs, which represent the diversity of improved tropical maize germplasm, has been distributed to various public and private sectors worldwide to improve a significant number of value-added traits in temperate and tropical maize. The reference fingerprinting SNP dataset for all the currently released CMLs developed in this study is helpful for the CIMMYT germplasm bank; in order to regenerate, maintain, and distribute the CML germplasm worldwide, they can genotype the regeneration seeds with the reference fingerprinting SNP dataset to assess their genetic purity and identity. Seed requesters can also use the reference fingerprinting SNP dataset to evaluate whether the obtained CML seeds are what they expect. Moreover, the reference fingerprinting SNP dataset developed by the present study is also useful for the breeders from both CIMMYT and partner institutes to better understand how to utilize the currently released CMLs in their breeding programs to select parental lines, replace testers, assign heterotic groups and create a core set of breeding germplasm [2].

Previous studies only genotyped part of the currently released CMLs, using different genotyping platforms to perform QC and genetic diversity analyses [2,6,21,29]. To the best of our knowledge, the reference fingerprinting molecular marker dataset had been not developed for all the currently released CMLs with any genotyping platform. In addition, single replication was genotyped for each CML, and a few CMLs were selected as technical replications in the previous studies. In the present study, two rounds of genotyping were applied to generate the reference fingerprinting SNP dataset for all the currently released 615 CMLs, and each CML was genotyped with at least two biological duplications to ensure the reliability of this reference fingerprinting SNP dataset. Using the genotypic data of the five biological duplications from both rounds of genotyping, mislabeled samples corresponding to thirteen CMLs in the first round of genotyping were also identified, and the reference fingerprinting SNP dataset for these thirteen CMLs was improved by removing the genotypic data of the mislabeled duplications. Furthermore, 28 SNPs with heterozygosity rates greater than 5%, missing rates greater than 10%, and MAF values lower than 0.05, were excluded from the reference fingerprinting SNP dataset. For the rest of the 152 SNPs, the average missing was 2.50%, the average heterozygosity rate was 0.60%, and the average MAF was 0.34. These results indicate that these 152 SNPs are high quality, and they are appropriate for further genetic diversity analysis, as well as for future QC studies.

Due to a large number of germplasm requests all over the world, regenerating, maintaining, and distributing the CML germplasms has become a significant challenge. During seed regeneration, there is a possibility of contamination with seeds or pollen [3]. The QC genotyping is important for the regeneration, maintenance, and distribution of CML seeds. The genetic purity analysis of the present study showed that 12.27% of the samples in the first round of genotyping did not achieve the basic purity level. Most of these samples correspond to the early CMLs, developed with the recurrent method of population improvement, and the original seeds of these CMLs submitted to the CIMMYT maize germplasm bank were not purified completely. For preserving the genetic diversity of the CMLs, the bulk pollen method was applied for regeneration seeds, and it also maintained the heterozygosity in the regeneration seeds. The early CMLs, with a possibly high level of heterozygosity rate identified in the present study, need to be purified by the CIMMYT germplasm bank and restock the regeneration seeds. For the CMLs with a possibly high level of heterozygosity rate released in the past two decades, and still being frequently used by the breeding program, the CIMMYT maize breeding team take responsibility for purifying them and re-submitting the purified increased seeds to the CIMMYT germplasm bank. Then, the reference fingerprinting SNP dataset for these CMLs needs to be updated by sampling the purified new seed sources.

Based on the first round of genotypic data, 25 CMLs were identified with the genetic identity issue, and eleven of these CMLs have both identity and purity issues; therefore, these CMLs need to first be purified, and then the reference fingerprinting SNP dataset

of these CMLs need to be updated by sampling the purified new seed sources. Among these 25 CMLs, the genetic distance of seven pairs of CMLs, estimated with the reference fingerprinting SNP dataset, was zero, and most of these pairs of CMLs had similar genetic backgrounds and closer pedigree relationships. These results indicated that the CMLs with close pedigree relationships could not be distinguished by the low-density reference fingerprinting dataset of 152 SNPs and that, before release, all the CMLs candidates have to be genotyped with the same reference fingerprinting dataset as that of the currently released CMLs. In addition, the genetic relationships between the CML candidates and all currently released CMLs need to be estimated to avoid releasing any new CMLs that have close genetic relationships with the existing CMLs, and the seed purity of the CML candidates should be tested before being submitted to the CIMMYT maize germplasm bank for release. Moreover, the reference fingerprinting SNP dataset for all the CMLs can be updated as soon as the new CMLs are announced for release.

The results of the genetic diversity analysis performed in the present study showed that the average genetic distance between all paired CMLs was 0.43, and 99.43% of pairs of CMLs had genetic distances greater than 0.30, indicating that high genetic diversity existed in the collection of the currently released CMLs, and most CMLs are distantly related. These results are similar to the previous studies [2]. The result of the principal component analysis and the average genetic distances between the different subgroups revealed that a clear genetic divergence was observed between temperate and tropical maize; the Highland Tropical subgroup can be distinguished from the Lowland Tropical and Subtropical/Mid-altitude subgroups. The Lowland Tropical subgroup and the Subtropical/Mid-altitude subgroup partially overlapped with each other; this reflects the current germplasm exchange flow from the Lowland Tropical breeding programs to the Subtropical/Mid-altitude breeding programs. The currently released CMLs are clustered according to their environmental adaptation groups and pedigree relationships.

The heterotic patterns in maize are discovered based on whether phylogenetic and geographical isolation exists, as well as being enhanced by breeders through long-term selection [29]. The heterotic groups within a collection of maize germplasm can be estimated based on pedigree information and by combining ability tests, genetic distance, or kinship coefficient analysis. Using the molecular markers provides an alternative approach for revealing the heterotic patterns in a collection of maize germplasm. A previous study showed that only half of CMLs have heterotic information [2], which was provided by the individual breeding program based on the information of pedigree and combining ability. In the present study, the heterotic group information for all of the currently released CMLs was assigned to specific heterotic groups of CIMMYT: A and B. This will facilitate the germplasm exchange and utilization among the different maize breeding programs of CIMMYT. The pairwise genetic distances between all of the currently released CMLs were also estimated based on the reference fingerprinting SNP dataset, which is complementary to the updated heterotic group information to help breeders to better understand the utilization of the currently released CMLs in their breeding programs. In two major maize breeding programs at CIMMYT, the pedigree-based heterotic pattern of the important CMLs, widely used in breeding and tests, were mostly consistent with those estimated using the reference fingerprinting SNP dataset. There were only a few exceptions of inbred lines, mixed between the two heterotic groups, indicating that the CMLs from the same heterotic group were more closely related, and heterotic patterns are being developed through breeding selection and recycling of a core set of key founder lines. The correct parental combinations to maximize heterosis can be identified to develop high yield potential hybrids.

4. Materials and Methods

4.1. Plant Materials

In the present study, 615 CMLs and 14 temperate maize lines were used for genotyping in order to perform QC and genetic diversity analysis. Seeds of 615 CMLs were obtained

from the CIMMYT Maize Germplasm Bank. The modified CTAB (cetyltrimethylammonium bromide) method was applied for DNA extraction from the leaf tissues [30]. All of the CMLs were planted at the Tlaltizapan experimental station of CIMMYT, with 25 seeds per line per plot. The leaf tissues of each CML were sampled by mixing 10 individual plants, and two independent biological replications were collected for each inbred line. For the second round of genotyping, a subset of 180 CMLs was selected; the heterozygosity rate of at least one biological duplication of these CMLs was greater than 3%, and the genetic distance between the biological replications of the same CML was greater than that between different CMLs. In the second round of genotyping, these 180 CMLs were planted in the greenhouse for tissue collection, a leaf sample from a single plant was collected for each independent biological duplication, and three independent biological duplications were collected for each CML. The latest batch of CMLs (CML604 to CML615), released in May 2021, was genotyped with a subset of 90 SNPs, and each CML was sampled by mixing three individual plants.

Based on the environmental adaptation, all of the 615 CMLs developed by the different breeding programs were classified into three categories: 336 CMLs adapted to the lowland tropical environment, 230 CMLs adapted to the subtropical/mid-altitude environment, and 37 CMLs adapted to the highland tropical environment (Table 3). The maize breeding programs in Latin America released 464 CMLs, including 285 lowland tropical CMLs, 144 subtropical/mid-altitude CMLs, and 35 highland tropical CMLs. The maize breeding programs in Southern and Eastern Africa released 110 CMLs, including 22 lowland tropical CMLs, 86 subtropical/mid-altitude CMLs, and 2 highland tropical CMLs. The maize breeding program in Asia released 29 lowlands tropical CMLs (Table 3). To increase the utilization of the CMLs in maize breeding programs globally, all the CMLs in the present study had been assigned to specific heterotic groups of CIMMYT: A and B, according to the information of pedigrees and combining abilities (personal communication, Felix San Vicente, 2021). The heterotic group assignment is represented for each CML after the CML number, for example, CML312A or CML444B.

Table 3. Environmental adaptation, geographic subset, and phenotypic characterization information of 615 CMLs and 14 temperate maize inbred lines.

Environmental Adaptation/ Geographic Subset	No. of Lines	Grain Color			Grain Texture			QPM
		Y	W	D	SD	F	SF	
Mexico Subtropical	146	34	112	51	38	36	21	22
Africa Lowland	22	8	14	6	0	16	0	0
Latin Am Lowland	14	4	10	1	3	6	4	4
Highland Tropical	36	6	30	2	18	3	13	6
Asia Lowland	33	31	2	3	9	17	4	0
Mexico Lowland	249	96	153	69	51	98	31	38
Africa Mid-altitude	93	0	92	9	32	23	35	0
South America	22	17	5	1	2	13	6	0
Temperate	14							

In total, 14 temperate maize inbred lines requested from the US Department of Agriculture (USDA)-Agricultural Research Service (ARS) North Central Regional Plant Introduction Station (NCRPIS) in Ames, Iowa, were used in the present study as controls to assess the genetic divergence between temperate and tropical maize. These inbred lines were primarily the Germplasm Enhancement Maize (GEM) accessions and the maize inbred lines, with expired plant variety protection. The fourteen temperate maize inbred lines, including 11 inbred lines from the Stiff Stalk Synthetic (BSSS) heterotic group, and 3 inbred lines from the non-Stiff Stalk Synthetic (NSSS) heterotic group, were referred to as the Temperate subgroup in the present study.

4.2. KASP SNP Markers Used for Genotyping

A set of over 1200 KASPar SNP genotyping assays was developed from the mapping data against the B73 reference genome [31] and a SNP mining study from EST sequences [32], which is now publicly available to the global maize genetics community, and it has been routinely applied at the Global Maize Program of CIMMYT for QA/QC, fingerprinting analysis, marker-assisted selection (MAS), and genetic diversity analysis. In total, 180 SNPs prioritized by CIMMYT for QC analysis were selected for genotyping all the currently released 615 CMLs and the fourteen temperate maize inbred lines in the present study (Table S1). The 180 SNPs were selected because of their high polymorphic information content (PIC), high minor allele frequency (MAF), and uniform coverage of the maize genome. Among these 180 SNPs, subsets of 50 or 100 SNPs were used in previous studies as the core set for routine QA/QC analysis [6], and dozens of KASPar SNPs converted from the DArT-Seq platform were developed for “broad” or “rapid” QA/QC analysis of CMLs [21]. In addition, some breeder-ready markers for implementing MAS were also converted into the KASPar SNPs and included in this marker set. These breeder-ready markers are associated with several traits for tropical maize improvement (<https://excellenceinbreeding.org/toolbox/tools/kasp-low-density-genotyping-platform> (accessed on 1 May 2022)) [33], including resistance to maize streak virus [12], resistance to maize lethal necrosis [34], kernel provitamin A content [35], quality protein maize [36], and resistance to tar spot complex [11,37] etc.

4.3. Characteristics of the SNPs Used for Genotyping

The summary statistics of each SNP in the genotypic dataset across 1206 samples of CMLs in the first round of genotyping, including PIC, MAF, missing rate, and heterozygosity rate, was analyzed using an in-house Perl script. To evaluate the uniformity of distribution, 180 SNPs were anchored to maize reference genome B73_RefGen_v4 using BLAST software and an in-house Perl script [38,39]. The genes harboring the SNP or in a flanking region of two Kb were extracted and annotated using gene functional information, downloaded from the MaizeGBD database [40]. The physical position of the SNPs on the reference genome, and the candidate genes nearby these QC SNPs and their functions, are shown in Table S1.

The gene ontology (GO) functional analysis was performed with the genes harboring the KASP markers using the singular enrichment analysis (SEA) with the online software AgriGO (<http://bioinfo.cau.edu.cn/agriGO/index.php> (accessed on 1 May 2022)) [41]. In the GO enrichment analysis, the reference gene set of “Maize AGPv4 (Maize-GAMER)” was chosen as the background and the “plant GO slim” was used as the gene ontology type. The *p*-value was adjusted using the hypergeometric test with a Benjamini-Hochberg false discovery rate (FDR). The visualization of GO enrichment was performed using the ggplot2 package in R.

4.4. Quality Control Analysis

The QC analysis was performed on 603 CMLs. First, the missing rate and the heterozygosity rate, representing the genetic purity, were calculated on each biological replication for CMLs using the in-house Perl scripts (www.perl.org (accessed on 1 May 2022)). The genetic purity of each biological replication for all the CMLs was classified into three levels by their heterozygosity rates: purity level with a heterozygosity rate of less than 1%; basic purity level with a heterozygosity rate ranging from 1% to 5%; and heterozygosity with a heterozygosity rate greater than 5%. The CMLs with a heterozygosity rate of at least one biological replication greater than 3% were included in the second round of genotyping, with the leaf tissue collected from the individual plant and three biological replications per CML. For these CMLs, the reference fingerprinting SNP dataset was obtained by integrating the fingerprinting genotypic data of each CML from the five independent biological replications.

Second, the genetic identity analyses were conducted between the two independent biological replications and among all the CMLs by evaluating the similarities between the two replications for the same CML or between the different CMLs. The similarity rate between the two replications or the two genotypes was calculated as the total number of identical alleles divided by the total number of non-missing alleles. The maximum identity value is 1, and the minimum identity value is 0. For the CML with the genetic distance between the two biological replications greater than that between different CMLs, the genotypic data from all the five independent biological replications was used to remove the mislabeled genotyping samples, and develop the reference fingerprinting SNP dataset.

Third, the parent-offspring tests were performed on 25 pairs of CMLs with close relationships to test the accuracy of the reference fingerprinting SNP dataset. Some CMLs are converted from the existing CMLs through marker-assisted selection or backcrossing to improve some specific target traits, such as QPM, imidazolinone resistance for controlling Striga, MSV resistance, and drought tolerance. The genetic distances between the original CMLs and the corresponding converted CMLs were calculated with TASSEL V5.0 software [42].

4.5. High-Quality and Low-Density Reference SNP Dataset for Fingerprinting

The reference SNP dataset for all the currently released 615 CMLs was obtained by integrating the fingerprinting genotypic data of each CML from the two independent biological replications in the first round of genotyping. The criteria for integration were as follow: (1) the SNP site containing the consistent alleles in both biological replications, was merged directly; (2) the SNP site containing homozygous alleles in one biological replication and heterozygous alleles in the other biological replication kept the homozygous alleles; (3) the rest of SNP sites were set as missing. The missing rate and heterozygosity rate of each CML across the reference dataset were calculated using the in-house Perl scripts. The SNPs with heterozygosity rates greater than 5%, the missing rate greater than 10%, and the MAF (minor allele frequency) values lower than 0.05 were excluded to develop the reference fingerprinting SNP dataset and to perform further genetic diversity analyses.

4.6. Genetic Diversity Analysis

Principal component analysis (PCA) was performed with all of the reference fingerprinting SNPs, on all tropical and temperate maize inbred lines genotyped in the present study, using the TASSEL V5.0 software. The first three principal components were plotted using the Scatter3dplot package in R to show the genetic relationships among the subgroups adapted to different environments.

To reveal the genetic relatedness among all CMLs and the temperate maize inbred lines, the identity by state (IBS) genetic distances were estimated with all the reference fingerprinting SNPs in the TASSEL V5.0 software. The average pairwise genetic distances were also estimated within each subgroup, adapted to different environments, i.e., Lowland Tropical subgroup, Subtropical/Mid-altitude subgroup, Highland Tropical subgroup, and Temperate subgroup.

Based on the pedigree information and the breeding knowledge, a core set of CMLs used by the Lowland Tropical maize breeding program in Mexico and the Subtropical maize breeding programs in East Africa, including the most frequently used CMLs and testers, was selected to investigate the heterotic patterns developed by Lowland tropical-Mexico (LLT) and East Africa (EA) CIMMYT maize breeding program. The phylogenetic tree was constructed based on IBS's genetic distance matrix for core CMLs in EA using the neighbor-joining algorithm in the ape package in R. The visualization of the phylogenetic tree was performed by MEGA7 software [43].

5. Conclusions

In the present study, a set of 180 KASP SNPs was used to genotype all the currently released 615 CMLs to perform QC analysis, build the reference fingerprinting SNP

dataset, and conduct the genetic diversity analysis. The QC analysis identified 25 CMLs having purity, identity, or mislabeling issues, further field observation, purification, and re-genotyping work on these CMLs are required. The reference fingerprinting SNP dataset was developed for all the currently released CMLs with 152 high-quality SNPs from 180 KASP SNP markers. Results of principal component analysis and average genetic distances between subgroups showed a clear genetic divergence between temperate and tropical maize, whereas the three tropical subgroups overlaid partially with each other. More than 99% of pairs of CMLs had genetic distances greater than 0.30, showing their high genetic diversity, and most CMLs distantly related. Heterotic patterns estimated with the molecular markers are consistent with those estimated with pedigree information in two major maize breeding programs at CIMMYT. These research findings are helpful to ensure the regeneration and distribution of the true CMLs via QC analysis, and facilitate the effective utilization of the CMLs globally.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/plants11223092/s1>, Table S1: The summary information of 180 KASP SNP markers; Table S2: Annotation information of genes associated with 180 KASP SNP markers; Table S3: Gene Ontology annotation information of genes associated with 180 KASP SNP markers; Table S4: Similarity and heterozygosity information of 600 CMLs with 2 replications; Table S5: The reference fingerprinting SNP dataset of 615 CMLs; Table S6: The pairwise genetic distance for 600 CMLs and 14 ex-pvp lines calculated using 152 SNPs.

Author Contributions: X.Z., W.L. and B.M.P. initiated and designed the overall study; X.Z., J.Q., A.A.C.-R., T.D. and F.S.V. performed and coordinated the field experiments; X.Z., K.D., S.K.N. and M.G. contributed to the genotypic data generation, J.Q., S.K.N., M.G. and X.Z. analyzed the data; J.Q., X.Z., F.F., H.Y., Y.B., D.M., T.D., F.S.V., M.O. and B.M.P. interpreted the results, J.Q., W.L. and X.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the Bill & Melinda Gates Foundation (B&MGF) and U.S. Agency for International Development (USAID) and Foundation for Agricultural Research (FFAR) [INV-003439]. Under the grant conditions of the Foundation, a Creative Commons Attribution 4.0 Generic License has already been assigned to the Author Accepted Manuscript version that might arise from this submission. The authors also thank the Sichuan Science and Technology Program (2022YFH0067) and the China Scholarship Council for their support.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset presented in this study can be found online at <https://hdl.handle.net/11529/10548818> (accessed on 2 November 2022).

Acknowledgments: The authors gratefully acknowledge the financial support from the CGIAR Research Program (CRP) on MAIZE, the Accelerating Genetic Gains in Maize and Wheat for Improved Livelihoods (AGG) project, the AGG supplemental project, One CGIAR Initiative on Accelerated Breeding. We also wish to thank Denise E. Costich for sharing the seeds of CMLs from the CIMMYT germplasm bank.

Conflicts of Interest: The authors declare that they have no conflict of interest.

References

1. Campos, H.; Caligari, P.D. *Genetic Improvement of Tropical Crops*; Springer International Publishing: Cham, Switzerland, 2017. [[CrossRef](#)]
2. Wu, Y.; Vicente, F.S.; Huang, K.; Dhliwayo, T.; Costich, D.E.; Semagn, K.; Sudha, N.; Olsen, M.; Prasanna, B.M.; Zhang, X.; et al. Molecular characterization of CIMMYT maize inbred lines with genotyping-by-sequencing SNPs. *Theor. Appl. Genet.* **2016**, *129*, 753–765. [[CrossRef](#)] [[PubMed](#)]
3. Warburton, M.L.; Setimela, P.; Franco, J.; Cordova, H.; Pixley, K.; Bänziger, M.; Dreisigacker, S.; Bedoya, C.; MacRobert, J. Toward a Cost-Effective Fingerprinting Methodology to Distinguish Maize Open-Pollinated Varieties. *Crop Sci.* **2010**, *50*, 467–477. [[CrossRef](#)]

4. Lu, Y.; Shah, T.; Hao, Z.; Taba, S.; Zhang, S.; Gao, S.; Liu, J.; Cao, M.; Wang, J.; Prakash, A.B.; et al. Comparative SNP and Haplotype Analysis Reveals a Higher Genetic Diversity and Rapider LD Decay in Tropical than Temperate Germplasm in Maize. *PLoS ONE* **2011**, *6*, e24861. [[CrossRef](#)] [[PubMed](#)]
5. Gowda, M.; Woruku, M.; Nair, S.K.; Palocios-Rojas, N.; Prasanna, B.M. *Quality Assurance/Quality Control (QA/QC) in Maize Breeding and Seed Production: Theory and Practice*; CIMMYT: Veracruz, Mexico, 2017; ISBN 9789966197191.
6. Semagn, K.; Beyene, Y.; Makumbi, D.; Mugo, S.; Prasanna, B.M.; Magorokosho, C.; Atlin, G. Quality control genotyping for assessment of genetic identity and purity in diverse tropical maize inbred lines. *Theor. Appl. Genet.* **2012**, *125*, 1487–1501. [[CrossRef](#)]
7. Coe Edward, H.; Gardiner, J.M. RFLP Maps of Maize. In *DNA-Based Markers in Plants*; Phillips Ronald, L., Vasil, I.K., Eds.; Springer: Dordrecht, The Netherlands, 1994; pp. 240–245, ISBN 978-94-011-1104-1.
8. Sharopova, N.; McMullen, M.D.; Schultz, L.; Schroeder, S.; Sanchez-Villeda, H.; Gardiner, J.; Bergstrom, D.; Houchins, K.; Melia-Hancock, S.; Musket, T.; et al. Development and mapping of SSR markers for maize. *Plant Mol. Biol.* **2002**, *48*, 463–481. [[CrossRef](#)]
9. Ertiro, B.T.; Ogugo, V.; Worku, M.; Das, B.; Olsen, M.; Labuschagne, M.; Semagn, K. Comparison of Kompetitive Allele Specific PCR (KASP) and genotyping by sequencing (GBS) for quality control analysis in maize. *BMC Genom.* **2015**, *16*, 908. [[CrossRef](#)]
10. Semagn, K.; Babu, R.; Hearne, S.; Olsen, M. Single nucleotide polymorphism genotyping using Kompetitive Allele Specific PCR (KASP): Overview of the technology and its application in crop improvement. *Mol. Breed.* **2014**, *33*, 1–14. [[CrossRef](#)]
11. Ren, J.; Wu, P.; Huestis, G.M.; Zhang, A.; Qu, J.; Liu, Y.; Zheng, H.; Alakonya, A.E.; Dhliwayo, T.; Olsen, M.; et al. Identification and fine mapping of a major QTL (qRtsc8-1) conferring resistance to maize tar spot complex and validation of production markers in breeding lines. *Theor. Appl. Genet.* **2022**, *135*, 1551–1563. [[CrossRef](#)]
12. Nair, S.K.; Babu, R.; Magorokosho, C.; Mahuku, G.; Semagn, K.; Beyene, Y.; Das, B.; Makumbi, D.; Kumar, P.L.; Olsen, M.; et al. Fine mapping of Msv1, a major QTL for resistance to Maize Streak Virus leads to development of production markers for breeding pipelines. *Theor. Appl. Genet.* **2015**, *128*, 1839–1854. [[CrossRef](#)]
13. Ganal, M.W.; Durstewitz, G.; Polley, A.; Bérard, A.; Buckler, E.S.; Charcosset, A.; Clarke, J.D.; Graner, E.-M.; Hansen, M.; Joets, J.; et al. A Large Maize (*Zea mays* L.) SNP Genotyping Array: Development and Germplasm Genotyping, and Genetic Mapping to Compare with the B73 Reference Genome. *PLoS ONE* **2011**, *6*, e28334. [[CrossRef](#)]
14. Rousselle, Y.; Jones, E.; Charcosset, A.; Moreau, P.; Robbins, K.; Stich, B.; Knaak, C.; Flament, P.; Karaman, Z.; Martinant, J.-P.; et al. Study on Essential Derivation in Maize: III. Selection and Evaluation of a Panel of Single Nucleotide Polymorphism Loci for Use in European and North American Germplasm. *Crop Sci.* **2015**, *55*, 1170–1180. [[CrossRef](#)]
15. Unterseer, S.; Bauer, E.; Haberer, G.; Seidel, M.; Knaak, C.; Ouzunova, M.; Meitinger, T.; Strom, T.M.; Fries, R.; Pausch, H.; et al. A powerful tool for genome analysis in maize: Development and evaluation of the high density 600 k SNP genotyping array. *BMC Genom.* **2014**, *15*, 823. [[CrossRef](#)] [[PubMed](#)]
16. Xu, C.; Ren, Y.; Jian, Y.; Guo, Z.; Zhang, Y.; Xie, C.; Fu, J.; Wang, H.; Wang, G.; Xu, Y.; et al. Development of a maize 55 K SNP array with improved genome coverage for molecular breeding. *Mol. Breed.* **2017**, *37*, 20. [[CrossRef](#)] [[PubMed](#)]
17. Wu, X.; Li, Y.; Shi, Y.; Song, Y.; Wang, T.; Huang, Y.; Li, Y. Fine genetic characterization of elite maize germplasm using high-throughput SNP genotyping. *Theor. Appl. Genet.* **2014**, *127*, 621–631. [[CrossRef](#)]
18. Ruanjaichon, V.; Khammona, K.; Thunnom, B.; Suriharn, K.; Kerd Sri, C.; Aesomnuk, W.; Yongsuwan, A.; Chaomueang, N.; Thammapichai, P.; Arikrit, S.; et al. Identification of Gene Associated with Sweetness in Corn (*Zea mays* L.) by Genome-Wide Association Study (GWAS) and Development of a Functional SNP Marker for Predicting Sweet Corn. *Plants* **2021**, *10*, 1239. [[CrossRef](#)]
19. Li, S.; Zhang, C.; Yang, D.; Lu, M.; Qian, Y.; Jin, F.; Liu, X.; Wang, Y.; Liu, W.; Li, X. Detection of QTNs for kernel moisture concentration and kernel dehydration rate before physiological maturity in maize using multi-locus GWAS. *Sci. Rep.* **2021**, *11*, 1764. [[CrossRef](#)]
20. Tian, H.-L.; Wang, F.-G.; Zhao, J.-R.; Yi, H.-M.; Wang, L.; Wang, R.; Yang, Y.; Song, W. Development of maizeSNP3072, a high-throughput compatible SNP array, for DNA fingerprinting identification of Chinese maize varieties. *Mol. Breed.* **2015**, *35*, 136. [[CrossRef](#)]
21. Chen, J.; Zavala, C.; Ortega, N.; Petroli, C.; Franco, J.; Burgueño, J.; Costich, D.E.; Hearne, S.J. The Development of Quality Control Genotyping Approaches: A Case Study Using Elite Maize Lines. *PLoS ONE* **2016**, *11*, e0157236. [[CrossRef](#)]
22. Gore, M.A.; Chia, J.-M.; Elshire, R.J.; Sun, Q.; Ersoz, E.S.; Hurwitz, B.L.; Peiffer, J.A.; McMullen, M.D.; Grills, G.S.; Ross-Ibarra, J.; et al. A First-Generation Haplotype Map of Maize. *Science* **2009**, *326*, 1115–1117. [[CrossRef](#)]
23. Chia, J.-M.; Song, C.; Bradbury, P.J.; Costich, D.; De Leon, N.; Doebley, J.; Elshire, R.; Gaut, B.; Geller, L.; Glaubitz, J.C.; et al. Maize HapMap2 identifies extant variation from a genome in flux. *Nat. Genet.* **2012**, *44*, 803–807. [[CrossRef](#)]
24. Bukowski, R.; Guo, X.; Lu, Y.; Zou, C.; Bicheng, Y.; Rong, Z.; Wang, B.; Xu, D.; Yang, B.; Xie, C.; et al. Construction of the third-generation *Zea mays* haplotype map. *GigaScience* **2018**, *7*, gix134. [[CrossRef](#)] [[PubMed](#)]
25. Elshire, R.J.; Glaubitz, J.C.; Sun, Q.; Poland, J.A.; Kawamoto, K.; Buckler, E.S.; Mitchell, S.E. A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. *PLoS ONE* **2011**, *6*, e19379. [[CrossRef](#)] [[PubMed](#)]
26. Zhang, X.; Pérez-Rodríguez, P.; Semagn, K.; Beyene, Y.; Babu, R.; López-Cruz, M.A.; Vicente, F.S.; Olsen, M.; Buckler, E.; Jannink, J.-L.; et al. Genomic prediction in biparental tropical maize populations in water-stressed and well-watered environments using low-density and GBS SNPs. *Heredity* **2015**, *114*, 291–299. [[CrossRef](#)] [[PubMed](#)]

27. Tello, J.; Roux, C.; Chouiki, H.; Laucou, V.; Sarah, G.; Weber, A.; Santoni, S.; Flutre, T.; Pons, T.; This, P.; et al. A novel high-density grapevine (*Vitis vinifera* L.) integrated linkage map using GBS in a half-diallel population. *Theor. Appl. Genet.* **2019**, *132*, 2237–2252. [[CrossRef](#)]
28. Wang, N.; Yuan, Y.; Wang, H.; Yu, D.; Liu, Y.; Zhang, A.; Gowda, M.; Nair, S.K.; Hao, Z.; Lu, Y.; et al. Applications of genotyping-by-sequencing (GBS) in maize genetics and breeding. *Sci. Rep.* **2020**, *10*, 16308. [[CrossRef](#)]
29. Guo, R.; Chen, J.; Petroli, C.D.; Pacheco, A.; Zhang, X.; Vicente, F.S.; Hearne, S.J.; Dhliwayo, T. The genetic structure of CIMMYT and U.S. inbreds and its implications for tropical maize breeding. *Crop Sci.* **2021**, *61*, 1666–1681. [[CrossRef](#)]
30. Saghai-Marouf, M.A.; Soliman, K.M.; Jorgensen, R.A.; Allard, R.W. Ribosomal DNA spacer-length polymorphisms in barley: Mendelian inheritance, chromosomal location, and population dynamics. *Proc. Natl. Acad. Sci. USA* **1984**, *81*, 8014–8018. [[CrossRef](#)]
31. Jones, E.; Chu, W.-C.; Ayele, M.; Ho, J.; Bruggeman, E.; Yourstone, K.; Rafalski, A.; Smith, O.S.; McMullen, M.D.; Bezawada, C.; et al. Development of single nucleotide polymorphism (SNP) markers for use in commercial maize (*Zea mays* L.) germplasm. *Mol. Breed.* **2009**, *24*, 165–176. [[CrossRef](#)]
32. Batley, J.; Barker, G.; O’Sullivan, H.; Edwards, K.J.; Edwards, D. Mining for Single Nucleotide Polymorphisms and Insertions/Deletions in Maize Expressed Sequence Tag Data. *Plant Physiol.* **2003**, *132*, 84–91. [[CrossRef](#)]
33. Prasanna, B.M.; Cairns, J.E.; Zaidi, P.H.; Beyene, Y.; Makumbi, D.; Gowda, M.; Magorokosho, C.; Zaman-Allah, M.; Olsen, M.; Das, A.; et al. Beat the stress: Breeding for climate resilience in maize for the tropical rainfed environments. *Theor. Appl. Genet.* **2021**, *134*, 1729–1752. [[CrossRef](#)]
34. Gowda, M.; Das, B.; Makumbi, D.; Babu, R.; Semagn, K.; Mahuku, G.; Olsen, M.S.; Bright, J.M.; Beyene, Y.; Prasanna, B.M. Genome-wide association and genomic prediction of resistance to maize lethal necrosis disease in tropical maize germplasm. *Theor. Appl. Genet.* **2015**, *128*, 1957–1968. [[CrossRef](#)] [[PubMed](#)]
35. Babu, R.; Rojas, N.P.; Gao, S.; Yan, J.; Pixley, K. Validation of the effects of molecular marker polymorphisms in LcyE and CrtRB1 on provitamin A concentrations for 26 tropical maize populations. *Theor. Appl. Genet.* **2013**, *126*, 389–399. [[CrossRef](#)]
36. Babu, R.; Prasanna, B.M. Molecular Breeding for Quality Protein Maize (QPM). In *Genomics of Plant Genetic Resources*; Springer: Dordrecht, The Netherlands, 2014; Volume 2. [[CrossRef](#)]
37. Cao, S.; Loladze, A.; Yuan, Y.; Wu, Y.; Zhang, A.; Chen, J.; Huestis, G.; Cao, J.; Chaikam, V.; Olsen, M.; et al. Genome-Wide Analysis of Tar Spot Complex Resistance in Maize Using Genotyping-by-Sequencing SNPs and Whole-Genome Prediction. *Plant Genome* **2017**, *10*, 2016. [[CrossRef](#)] [[PubMed](#)]
38. Jiao, Y.; Peluso, P.; Shi, J.; Liang, T.; Stitzer, M.C.; Wang, B.; Campbell, M.S.; Stein, J.C.; Wei, X.; Chin, C.-S.; et al. Improved maize reference genome with single-molecule technologies. *Nature* **2017**, *546*, 524–527. [[CrossRef](#)]
39. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403–410. [[CrossRef](#)]
40. Portwood, J.L., II; Woodhouse, M.R.; Cannon, E.K.; Gardiner, J.M.; Harper, L.C.; Schaeffer, M.L.; Walsh, J.R.; Sen, T.Z.; Cho, K.T.; Schott, D.A.; et al. MaizeGDB 2018: The maize multi-genome genetics and genomics database. *Nucleic Acids Res.* **2019**, *47*, D1146–D1154. [[CrossRef](#)] [[PubMed](#)]
41. Du, Z.; Zhou, X.; Ling, Y.; Zhang, Z.; Su, Z. agriGO: A GO analysis toolkit for the agricultural community. *Nucleic Acids Res.* **2010**, *38*, W64–W70. [[CrossRef](#)]
42. Bradbury, P.J.; Zhang, Z.; Kroon, D.E.; Casstevens, T.M.; Ramdoss, Y.; Buckler, E.S. TASSEL: Software for association mapping of complex traits in diverse samples. *Bioinformatics* **2007**, *23*, 2633–2635. [[CrossRef](#)]
43. Kumar, S.; Stecher, G.; Tamura, K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol. Biol. Evol.* **2016**, *33*, 1870–1874. [[CrossRef](#)]