



Molecular analysis of *opaque2* gene governing accumulation of lysine and tryptophan in maize endosperm

Gulab Chand · Vignesh Muthusamy · Rajkumar U. Zunjare ·
Subhra J. Mishra · Gaurav Sharma · Brijesh K. Mehta · Sohini Singh ·
Tanu Allen · Firoz Hossain

Received: 30 December 2023 / Accepted: 12 September 2024
© The Author(s), under exclusive licence to Springer Nature B.V. 2024

Abstract The recessive *opaque2* (*o2*) gene enhances lysine and tryptophan by ~twofold in maize kernels. Despite >6 decades of research, no functional marker for *o2* has been reported yet. Here, we sequenced 2848 bp of entire *o2* gene in 10 diverse inbreds, and identified 7 exonic SNPs and 3 exonic InDels that distinguished the normal and quality protein maize (QPM) inbreds. Of these, three potential SNPs (SNP601_T/C, SNP2359_C/T and SNP2360_G/T) and three InDels (InDel_1, InDel_8 and InDel_9) were responsible for modification of amino acids. Clustering analysis grouped 24 QPM and 24 normal inbreds into two clusters, and led to the identification of 24 haplotypes of *o2*. Insertion of alanine was found to regulate the binding of *o2* protein with promoter of zein genes. QPM inbreds

possessed significantly higher lysine (0.271–0.406%) and tryptophan (0.069–0.101%) compared to normal (lysine: 0.124–0.213%; tryptophan: 0.029–0.047%). A functional marker (MGU-InDel10-*o2*) was developed, and it clearly differentiated the QPM and normal inbreds. MGU-InDel10-*o2* successfully identified desirable plants in BC₁F₁, BC₂F₁ and BC₂F₂ populations. BC₂F₃ progenies possessed higher lysine (0.379%) and tryptophan (0.084%) over original inbreds (lysine: 0.189% and tryptophan: 0.039%). Among haplotypes, Hap-3 showed the highest average lysine (0.380%) and tryptophan (0.094%), while Hap-14 and Hap-16 had the lowest lysine (0.124%) and tryptophan (0.029%), respectively. This is the first report on understanding allelic variation, haplotypes and functional marker development for *o2* gene in maize.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s10681-024-03414-2>.

G. Chand · V. Muthusamy · R. U. Zunjare · S. J. Mishra ·
G. Sharma · F. Hossain (✉)
ICAR-Indian Agricultural Research Institute (IARI),
New Delhi 110012, India
e-mail: fh_gpb@yahoo.com

G. Chand · S. Singh · T. Allen
Amity Institute of Biotechnology, AMITY University,
Noida 201303, India

B. K. Mehta
ICAR- Indian Grassland and Fodder Research Institute,
Jhansi 284003, India

Keywords Maize · Allelic diversity · *opaque2* ·
Lysine · Tryptophan

Introduction

Maize has emerged as the most important cereal crops for food and nutrition globally (Zhang et al. 2023). It provides at least 30% of the food calories to more than 4.5 billion people in 94 developing countries (Shiferaw et al. 2011; Prasanna et al. 2020). Traditional maize endosperm protein contains low level of lysine and tryptophan, which is less than one-half

of the recommended dose required for proper growth and development in humans (Gupta et al. 2015). The deficiency of lysine and tryptophan affects protein biosynthesis thereby aggravates the symptoms of protein energy malnutrition (PEM) (Hossain et al. 2023). They also cause loss of appetite, depression; delayed growth, and anxiety in children (Nuss and Tanumihardjo 2010). PEM is considered to be one of the most lethal forms of malnutrition, as it recorded the highest number of deaths among children worldwide (Nyakurwa et al. 2017). It is commonly observed in the least developing countries where maize is consumed as a staple food (Shiferaw et al. 2011; Bouis et al. 2019). Thus, efforts directed towards biofortification of maize for protein quality assume great impact for nutritional security (Neeraja et al. 2022). The traditional/ normal inbreds possess dominant *Opaque2* (*O2*) gene (Hossain et al. 2018). While, successful utilization of naturally available recessive *o2* gene coupled with endosperm modifier loci led to the development of nutritionally-rich maize cultivars popularly known as quality protein maize (QPM) (Vasal et al. 1980). Wide array of QPM cultivars adaptable to various agro ecologies have been released, and its beneficial effects on growth and development of humans as well as poultry and piggery have been well documented worldwide (Gunaratne et al. 2010; Hossain et al. 2019a; Maqbool et al. 2021).

The recessive *o2* allele nearly doubles the lysine and tryptophan in maize kernels, and thereby significantly improves the protein quality of maize over the genotypes possessing dominant *O2* allele (Mertz et al. 1964). The *o2* gene has been cloned through transposon tagging (Schmidt et al. 1987; Motto et al. 1988). It encodes a polypeptide containing a basic leucine zipper (bZIP) transcription factor, thereby regulates 19-kD and 22-kD α -zeins (Hunter et al. 2002). In *o2* based germplasm, lysine-poor α -zein is decreased with compensation by lysine-rich non-zeins (Gupta et al. 2013). Furthermore, in *o2* kernels, the expression and accumulation levels of lysine keto glutarate (LKR) are lower than those of the normal, resulting in reduced LKR activity and a higher free lysine content (Kemper et al. 1999).

Though, phenotypic selection of opaqueness in the kernels on 'light box' has been the main stay of QPM breeding, introgression of *o2* into elite germplasm using molecular markers provides great opportunity to accelerate the breeding cycle (Prasanna et al.

2020; Hossain et al. 2018). Presence of insertions or deletions (InDels), single nucleotide polymorphisms (SNPs), and length variation in motifs within simple sequence repeats (SSRs) serve as a valuable source for the development of effective marker (Borevitz et al. 2003). Three SSR markers (*phi057*, *umc1066* and *phi112*) present inside the *o2* gene as reported by Pioneer Corporation and the University of Missouri, have been abundantly used for the selection of *o2* gene in the breeding programme (Gupta et al. 2015; Prasanna et al. 2020). Several researchers have used these SSRs to introgress recessive *o2* allele into elite inbreds through marker-assisted selection (MAS) (Babu et al. 2005; Gupta et al. 2013; Jompuk et al. 2011; Hossain et al. 2018; Chand et al. 2022a). Though gene-based marker is preferred over the linked marker, which often causes selection of false positives due to crossing over between the gene and marker, it does not always differentiate the dominant (*O2*) and recessive (*o2*) alleles in several genetic background (Zhao et al. 2012; Chen et al. 2014; Hossain et al. 2018). Functional marker developed based on polymorphism causing the change in gene function always causes precise selection of the target gene, besides unambiguously identify the QPM genotypes. So far, no functional marker for the *o2* gene has been reported yet in maize. Further, information on allelic diversity and haplotype analysis on *o2* gene especially among the subtropically adapted maize inbreds is limited. The present investigation was therefore undertaken to (i) study the allelic variation of *o2* gene, (ii) analyze haplotype diversity of *o2* gene, and (iii) develop and validate functional marker specific to *o2* gene among diverse set of maize inbreds.

Materials and methods

Genetic materials

A set of five diverse inbreds of normal (Zm_O2-Normal1, Zm_O2-Normal2, Zm_O2-Normal3, Zm_O2-Normal4, and Zm_O2-Normal5) and QPM (Zm_o2-QPM1, Zm_o2-QPM2, Zm_o2-QPM3, Zm_o2-QPM4 and Zm_o2-QPM5) were selected for analysis of allelic variation at nucleotide and amino acids level (Table S1). The traditional/ normal maize inbreds possessed dominant *O2* gene in homozygous condition, while the QPM inbreds had the recessive

o2 gene in homozygous conditions. Further, a diverse panel of subtropically adapted 48 inbreds including 24 normal and 24 QPM was used for assessing haplotype analysis of the *o2* gene (Table S2). These QPM inbreds have been developed by various centres of India and CIMMYT, Mexico through introgression of recessive *o2* allele. QPM lines were initially selected based on presence of kernel opacity in the seeds when kept on the light-box (Hossain et al. 2008). While the normal lines did not show any opacity. BC₁F₁, BC₂F₁ and BC₂F₂ populations generated by crossing three normal inbreds (HKI1344, HKI1378 and HKI1348-6-2) with a *o2* donor (PMI-102) were used for genotyping using marker specific to *o2* gene. The segregating generations (BC₁F₁, BC₂F₁ and BC₂F₂) were raised at two locations viz., ICAR-Indian Agricultural Research Institute (IARI), New Delhi (28 °0' N, 77 °10' E, 216 m MSL) and IIMR-Winter Nursery Centre (WNC), Hyderabad (17 °21' N, 78 °29' E, 536 m MSL). The F₁ crosses were generated during the rainy season (2016) at IARI, New Delhi. The true F₁s were backcrossed with their recurrent parents during winter season (2016–17) at IIMR-WNC, Hyderabad to generate BC₁F₁ populations. The BC₁F₁ progenies were further raised at IARI, New Delhi during rainy season (2017). The BC₂F₁ seeds were planted at WNC, Hyderabad during winter season (2017–18), while BC₂F₂ populations were grown at Delhi during rainy season (2018) (Chand et al. 2022a). The purpose of selecting diverse inbreds to check the effectiveness of newly developed marker specific to *o2* gene in differentiating the 24 normal and 24 QPM inbreds. While, the aim to use three backcross populations using three recipient parents and one donor parent, was to check the utility of newly developed markers in molecular breeding.

DNA isolation, PCR protocol and primer designing for sequencing *o2* gene

DNA for sequencing and gene-based diversity analysis was isolated from matured seeds using sodium dodecyl sulphate (SDS) extraction procedure described by Dellaporta et al. (1983). The *o2* genomic DNA sequence of GenBank accession no. KF831425.1 (recessive allele of *o2*, hereafter *o2-Ref*) was used to design the 10 overlapping primer pairs using Primer3 online software, and were custom

synthesized from M/s G-Bioscience (Table S3). These primers covered entire 2848 bp of *o2* gene with an amplicon size ranging from 200 to 850 bp. The 50 µl PCR reaction mixture was prepared using 200 ng DNA, 25 µl PCR master mix (M/s G-Biosciences) and 1.87 µM each of forward and reverse primers. The PCR amplicons were checked using 2% agarose gel, and after confirmation of expected products, the remaining PCR product was processed for sequencing at M/s. Barcode Bio-Sciences (BBS) Pvt. Ltd.

Sequence analysis and phylogenetic tree construction of *o2*

The DNA and amino acid sequence were analyzed on BioEdit software using ClustalW alignment and MEGA-X tool using MUSCLE alignment to identify SNPs or InDels variation among inbreds (Kumar et al. 2018). After identifying variations, the alignment sequence file was then used to compute the number of SNPs, InDels, polymorphic sites and nucleotide diversity using DnaSP 6 software version 6.12.03 (Rozas et al. 2017). Putative InDels and SNPs that clearly distinguished the normal (*O2O2*) and QPM (*o2o2*) inbreds were analyzed manually. Online available bioinformatics tools viz; SOFTBERRY, Gene Structure Display Server (GSDS-2.0) were used for exon–intron organization analysis and domain was searched through scan Prosite (de Castro et al. 2006; Hu et al. 2015). The 10 diverse inbred sequences (five normal and five QPM) of *o2* along with *o2-Ref* (KF831425.1: QPM) and B73 (normal) were used for tree construction. A total 12 sequences were used for phylogenetic clustering of both nucleotide and amino acid sequences using MUSCLE algorithm in MEGA-X software.

Development of molecular marker specific to *o2* gene

A set of 12 pair of primers with amplicon size ranging from 143 to 783 bp were designed to develop InDel markers for molecular diversity analysis (Table S4). Further potential allele specific SNP-based PCR markers were also designed to differentiate the recessive (QPM) and dominant (normal) allele of *o2* gene (Table S5). To enhance the specificity of the SNP based primers, a nucleotide mismatch was inserted to replace the third nucleotide from the 3' end of the

forwards primers (Yang et al. 2012). The 20 μ l PCR reaction containing 100 ng DNA, 10 μ l M/s G-Biosciences master mixture and 0.7 μ M each of forward and reverse primer were used for genotyping. All PCR reactions were accomplished on a T-100 thermal cycler (M/s Bio-Red) with the touch down 60 (TD-60) protocol (Duo et al. 2021). Amplicons were visualized in 2–4% agarose gel using a gel documentation system (M/s Alpha Innotech, USA).

InDel-based diversity analysis in *o2* gene among diverse inbreds

The marker profile of each InDel marker among the diverse panel of 48 genotypes was obtained using 3–4% agarose gel electrophoresis. The obtained marker data based on gel profile was analyzed in DARwin v6.0 to estimate genetic dissimilarity based on Jaccard's coefficient and constructed dendrogram using hierarchical-clustering method (Perrier et al. 2003). The PowerMarker v3.25 was deployed to estimate parameters viz. total number of alleles, major allele frequency, gene diversity, heterozygosity, polymorphism information content (PIC) and total no. of haplotypes (Liu and Muse 2005). The genotypic score of each accession was further exploited through iTOL (Interactive tree of life) online software to construct dendrogram. Model-based population structure analysis was performed using STRUCTURE software version 2.3.4 (Pritchard et al. 2000); the software was run in 10 replicates by setting K (the number of populations) from 1 to 5, with a length of burn-in period of 1,00,000 and number Markov Chain Monte Carlo (MCMC) replications at 1,00,000 (Evanno et al. 2005). An online tool, Structure Harvester (<http://taylor0.biology.ucla.edu>) was used to calculate the most probable genetic diverse population of the inbreds. The analysis of molecular variance (AMOVA) was undertaken using the program GenAlEx 6.5 (Peakall and Smouse 2006).

Estimation of lysine and tryptophan from endosperm

A panel of 48 diverse inbreds was grown in randomized complete block design (RCBD) with three replications at two locations viz., IARI, Delhi and ICAR- Indian Grassland and Fodder Research Institute (IGFRI), Jhansi during rainy season (2022). Each of the plants were selfed to avoid xenia effects

to avoid contamination by foreign pollen upon cross pollination. Selfed seeds were ground into fine powder for estimation of lysine and tryptophan using UHPLC (Dionex Ultimate 3000 System, Thermo Scientific, Massachusetts, USA) (Sarika et al. 2018). The flour of the grains was acid hydrolyzed using 800 μ l of 6N HCl, 100 μ l of 0.1N HCl, 100 μ l of nor-leucine and 10 μ l of phenol for 16 h at 110 °C. The two mobile phases, –A and –B consisted of buffer and organic phase in the ratio of 9:1 (v/v) and 1:9 (v/v), respectively were used for lysine. The buffer phase for lysine contained tetra-methyl ammonium chloride and sodium acetate trihydrate (pH 3.5), while organic phase had acetonitrile and methanol (49:1, v/v). In case of tryptophan, alkaline hydrolysis (2 ml of 4 M NaOH and 200 μ l of 0.1% ascorbic acid for 16 h at 110 °C) was performed. The mobile phase for tryptophan consisted of water and acetonitrile in the ratio of 95:5. The samples were injected separately in UHPLC through Acclaim™ 120 C₁₈ column (5 μ m, 120 Å, 4.6×150 mm) with a flow rate of 1.0 and 0.7 ml/min, and detected using RS3000 photodiode array detector at 265 and 280 nm, respectively. The concentration of lysine and tryptophan was estimated by standard regression curve derived using dilutions of external standards (AAS 18-5ML, Sigma Aldrich).

Homology modelling and molecular docking for *o2* protein

The *o2* protein homology models for QPM and normal were built using automated protein structure homology-modelling servers viz; SWISS-MODEL (Waterhouse et al. 2018) and I-TASSER (Yang and Zhang 2015). The best models for normal and QPM were selected for further analysis based on various parameters like global model quality estimation (GMQE) score and identity score > 50%. Stereo chemical property of top predicted model was analyzed through structure assessment tool in SWISS-MODEL online server via Ramachandran Plot. The modelled protein structures of normal and QPM inbreds were visualized in PyMOL (DeLano 2009) and UCSF-Chimera1.15 (Goddard et al. 2005).

The 22-kD α -zein (GRMZM2G160739), 19-kD α -zein (AF546188.1_FG003) and 27-kD γ -zein (GRMZM2G138727) were selected for the molecular docking with normal and QPM *o2* protein structures. The three-dimensional (3D) conformers of the *o2*

proteins were subjected to I-TASSER servers to get homology models in PDB format for docking studies. Promoter sequences of zeins were downloaded from the MaizeGDB and were converted to PDB format using Discovery Studio Visualizer (Biovia 2020). DNA and Protein interactions were performed using HDock server based on hybrid docking algorithm of template-based modeling and ab initio free docking (Yan et al. 2017). The PDB files of the DNA and protein were submitted to HDock server and best results based on docking score and confidence score were selected. Finally, the docked complex structures were analyzed in PyMOL (DeLano 2009) and UCSF-Chimera1.15 (Goddard et al. 2005).

Statistical analysis

The haplotypes analysis among diverse inbreds was graphically depicted using Graphical Genotype (GGT2) software version 2.0 (Van-Berloo 2008). The amplicon of markers used in generated haplotypes scored as 'A' for the favourable (recessive) allele, 'B' and 'C' for the unfavourable (dominant) allele1 and allele2. Graphical representation based on the mean of lysine and tryptophan was established by MS-Office Excel (2019). The analysis of variance (ANOVA) for nutritional quality was analyzed using CropStat7.2 software available in the public domain.

Result

DNA sequence variations in *o2* gene

The entire nucleotide sequence of *o2* gene among QPM (Zm_o2-QPM1 to Zm_o2-QPM5) and normal (Zm_O2-normal1 to Zm_O2-normal5) inbred along with recessive *o2-Ref* sequence [GenBank accession no. KF831425.1, deposited by Chen et al. 2014] and B73 (normal) retrieved from maizeGDB were aligned. Sequence based structure organization analysis of *o2* among 12 diverse inbreds revealed six exonic regions in *o2* gene. A set of 248 SNPs were discovered in the maize *o2* gene (Table S6). Of these, 22 SNPs (7 SNPs in exonic region and 15 SNPs in intronic region) clearly differentiated dominant allele (*O2*) from recessive allele (*o2*) (Table 1). The DNA sequence analysis also revealed the presence of 61 InDels in *o2* with a mean nucleotides length of 7.31

and diversity (ki) of 25.50 (Table S7). A total of 9 InDels (3 InDels in exonic regions and 6 in intronic regions) also differentiated the dominant allele (*O2*) from recessive allele (*o2*). These included 3, 6 and 105 bp insertions, and 2, 3, 5, 6 and 26 bp deletions (Table 1).

Peptide sequence variation in *o2* gene

Multiple peptide sequences alignment of *o2* proteins revealed distinct variability among normal and QPM amino acids sequences. Of these, five modifications (due to SNP) revealed amino acids substitutions and three modifications (due to InDels) showed insertions in amino acids sequence (Fig. S1). Amino acids substitutions due to SNP polymorphisms included changes viz., (i) SNP601_T/C (exon-1) causing valine (V) to alanine (A) conversion at 89th position, (ii) dinucleotide change (CG/TT) due to SNP2359_C/T and SNP2360_G/T (exon-5) conferring alanine (A) to serine (S) conversion at 295th position, (iii) SNP2663_C/T (exon-6) with proline (P) to valine (V) conversion at 361st position, (iv) SNP 2792_G/C (exon-6) converting glycine (G) to alanine (A) at 404th position and (v) SNP2865_G/C (exon-6) modifying glutamine (Q) to histidine (H) at 428th position of amino acid in the polypeptide chain. Among these, (i) alanine (A) to serine (S) and (ii) glutamine (Q) to histidine (H) belonged to modification in properties of amino acids such as alanine was hydrophobic, while serine and glutamine were hydrophilic polar uncharged with histidine being hydrophilic basic in nature. Further, three InDel modifications viz., (i) InDel-1 having 3 bp InDel at 656–658 bp position (valine at position 108), (ii) InDel-8 having 3 bp InDel at 2594–2596 bp position (alanine at position 339) and (iii) InDel-9 having 6 bp InDel at 2657–2662 bp position (proline-valine at position 356–357) were produced due to InDels present in exon-1 and exon-6.

In addition, the impact of amino acid substitutions was studied through modelling of 3-D structure among QPM and normal protein using BZIP domain containing protein from maize (AlphaFoldDB: A0A1D6H7P7) as a template. Both QPM (liao2345/o2-2; GenBank accession no. KF831425.1) and normal (Zm_O2-Normal5) protein structures were superimposed through TM alignment and obtained superimposed protein structure with TM-score of 0.41835

Table 1 Identified SNPs/InDels differentiating normal and QPM maize inbreds

| S. No. | Position (bp) | | Nucleotide modification | | Amino acids modification | Length (bp) | Types of mutation | Region |
|--------|---------------|------|-------------------------|-----------|-----------------------------|-------------|-------------------|---------|
| | Start | End | Normal | QPM | | | | |
| 1 | 601 | 601 | T | C | Valine to Alanine | 1 | SNP601 | Exon1 |
| 2 | 656 | 658 | Deletion | Insertion | Insertion of Valine | 3 | InDel-1 | Exon1 |
| 3 | 1093 | 1197 | Deletion | Insertion | – | 105 | InDel-2 | Intron1 |
| 4 | 1229 | 1229 | T | C | – | 1 | SNP1229 | Intron1 |
| 5 | 1233 | 1234 | Insertion | Deletion | – | 2 | InDel-3 | Intron1 |
| 6 | 1285 | 1285 | A | G | – | 1 | SNP1285 | Intron1 |
| 7 | 1286 | 1286 | G | T | – | 1 | SNP1286 | Intron1 |
| 8 | 1306 | 1306 | C | A | – | 1 | SNP1306 | Intron1 |
| 9 | 1376 | 1381 | Insertion | Deletion | – | 6 | InDel-4 | Intron1 |
| 10 | 1382 | 1382 | A | G | – | 1 | SNP1382 | Intron1 |
| 11 | 1384 | 1384 | G | T | – | 1 | SNP1384 | Intron1 |
| 12 | 1385 | 1385 | A | C | – | 1 | SNP1385 | Intron1 |
| 13 | 1388 | 1388 | C | G | – | 1 | SNP1388 | Intron1 |
| 14 | 1389 | 1389 | A | G | – | 1 | SNP1389 | Intron1 |
| 15 | 1434 | 1434 | A | G | – | 1 | SNP1434 | Intron1 |
| 16 | 1495 | 1520 | Insertion | Deletion | – | 26 | InDel-5 | Intron1 |
| 17 | 1676 | 1680 | Insertion | Deletion | – | 5 | InDel-6 | Intron2 |
| 18 | 1687 | 1689 | Insertion | Deletion | – | 3 | InDel-7 | Intron2 |
| 19 | 1704 | 1704 | C | T | – | 1 | SNP1704 | Intron2 |
| 20 | 1728 | 1728 | A | G | – | 1 | SNP1728 | Intron2 |
| 21 | 1760 | 1760 | T | A | – | 1 | SNP1760 | Intron2 |
| 22 | 1761 | 1761 | C | T | – | 1 | SNP1761 | Intron2 |
| 23 | 2359 | 2359 | C | T | Alanine to Serine | 1 | SNP2359 | Exon5 |
| 24 | 2360 | 2360 | G | T | – | 1 | SNP2360 | Exon5 |
| 25 | 2497 | 2497 | G | A | – | 1 | SNP2497 | Intron5 |
| 26 | 2594 | 2596 | Deletion | Insertion | Insertion of Alanine | 3 | InDel-8 | Exon6 |
| 27 | 2656 | 2656 | C | G | No change in amino acid | 1 | SNP2656 | Exon6 |
| 28 | 2657 | 2662 | Deletion | Insertion | Insertion of Proline-Valine | 6 | InDel-9 | Exon6 |
| 29 | 2663 | 2663 | C | T | Proline to Valine | 1 | SNP2663 | Exon6 |
| 30 | 2792 | 2792 | G | C | Glycine to alanine | 1 | SNP2792 | Exon6 |
| 31 | 2865 | 2865 | G | C | Glutamine to Histidine | 1 | SNP2865 | Exon6 |

Normal inbreds (*O2O2*): UMI1200, PMI-PV3, CML342, CML454 and LM13; QPM inbreds (*o2o2*): CML150, HKI193-1, HKI161, CML165 and PMI-PV6

and RMSD value of 6.5 (Fig. S2). The 3-D structure analysis also revealed structural deformation due to insertion of alanine (at 108 amino acid position) and proline-valine (at 356–357 amino acid position) but no structural change was present due to substitutions of amino acids. Molecular docking to determine the specificity of *o2* encoded bZIP protein with promoter sequences of 22-kD α -zein (GRMZM2G160739), 19-kD α -zein (AF546188.1_FG003), 27-kD γ -zein

(GRMZM2G138727) revealed that DNA bound to recessive *o2* protein had less binding affinity for 22-kD α -zein (-242.26) and 19-kD α -zein (-230.54) as compared to dominant *O2* protein interaction with 22-kD α -zein (-296.47) and 19-kD α -zein (-266.79) (Table S8). While recessive *o2* protein interaction with 27-kD γ -zein showed high binding affinity (-225.89) over the dominant *O2* protein (-192.11). Further docking score is best explaining through

confidence score. The 3D visualization of protein-DNA complex interactions is depicted in Fig. S3.

Genetic relationships among inbreds

Based on phylogenetic analysis of DNA sequences of the 10 maize inbreds along with KF831425.1 (*o2-Ref*) and B73 (normal) sequences, inbred were divided into two clusters, –A and –B. Cluster-A had five QPM inbreds (*Zm_o2-QPM1*, *Zm_o2-QPM2*, *Zm_o2-QPM3*, *Zm_o2-QPM4* and *Zm_o2-QPM5*), as well as *o2-Ref* (KF831425.1), while Cluster-B had five normal inbreds (*Zm_O2-Normal1*, *Zm_O2-Normal2*, *Zm_O2-Normal3*, *Zm_O2-Normal4*, and *Zm_O2-Normal5*) along with B73 (normal) (Fig. 1A). Similarly amino acids sequences based clustering also divided into the two clusters, –C and –D. Cluster C had all QPM inbreds while cluster D contained all the normal inbreds (Fig. 1B).

InDels based diversity, clustering and haplotype analysis

Among the InDels identified from *o2* sequence alignment analysis, selected InDels (> 2 bp covering full-length gene) were exploited to develop the

gene-based InDel markers (Table S4). Molecular diversity using InDels (ranged from 3–105 bp) generated a total of 27 alleles with an allelic mean of 2.25 (range: 2–3) in a set of 48 diverse inbreds. PIC varied from 0.17 to 0.50 with a mean of 0.35. Among the 12 InDels based-markers, one marker (MGU-InDel10-*o2*) had a PIC of ≥ 0.5 (Table 2). Major allele frequency varied from 0.50 to 0.90 with mean of 0.78. The genetic diversity ranged from 0.19 (MGU-InDel12-*o2*) to 0.58 (MGU-InDel10-*o2*) with mean of 0.43. The dissimilarity matrix varied from 0 to 0.96.

Cluster analysis grouped the 48 diverse inbreds into two major clusters, namely –E and –F (Fig. 2). All the 24 QPM inbreds were grouped into cluster-E, while all 24 normal inbreds were grouped into cluster-F. Clustering patterns also revealed that genotypes originated from similar/ related pedigree grouped together. Haplotype analysis based on InDel markers revealed the presence of 24 haplotypes of *o2* existed among the 48 inbreds with haplotype diversity 0.894. The QPM inbreds formed four haplotypes (Hap-1, Hap-2, Hap-3 and Hap-4), while other haplotypes (Hap-5 to Hap-24) were generated by the normal inbreds (Fig. S4).

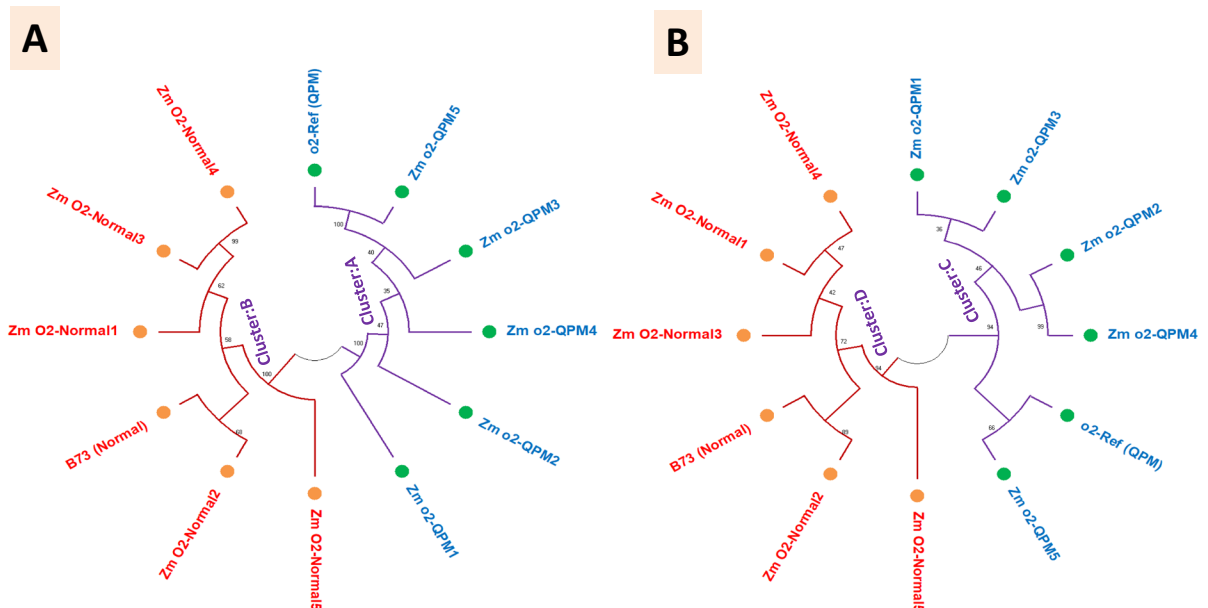
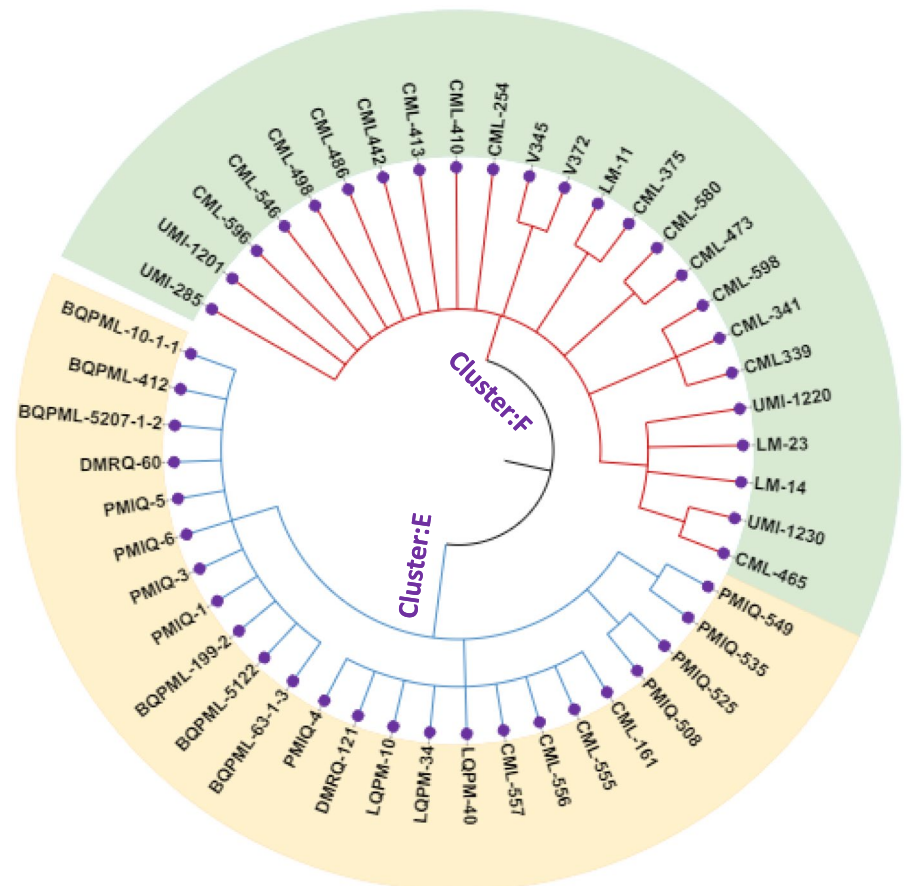


Fig. 1 Sequenced based phylogenetic relationship among maize inbreds; **A** Nucleotide based phylogenetic tree **B** Protein-based phylogenetic tree

Table 2 Molecular diversity parameters for gene-based InDel markers

| S. No. | Marker | Major allele frequency | No. of alleles | Gene diversity | Hetero-zygosity | PIC |
|--------|------------------------|------------------------|----------------|----------------|-----------------|------|
| 1 | MGU-InDel1- <i>o2</i> | 0.67 | 3.00 | 0.50 | 0.00 | 0.44 |
| 2 | MGU-InDel2- <i>o2</i> | 0.85 | 3.00 | 0.26 | 0.00 | 0.24 |
| 3 | MGU-InDel3- <i>o2</i> | 0.52 | 2.00 | 0.50 | 0.00 | 0.37 |
| 4 | MGU-InDel4- <i>o2</i> | 0.73 | 2.00 | 0.39 | 0.00 | 0.32 |
| 5 | MGU-InDel5- <i>o2</i> | 0.50 | 2.00 | 0.50 | 0.00 | 0.38 |
| 6 | MGU-InDel6- <i>o2</i> | 0.69 | 2.00 | 0.43 | 0.00 | 0.34 |
| 7 | MGU-InDel7- <i>o2</i> | 0.50 | 2.00 | 0.50 | 0.00 | 0.38 |
| 8 | MGU-InDel8- <i>o2</i> | 0.50 | 2.00 | 0.50 | 0.00 | 0.38 |
| 9 | MGU-InDel9- <i>o2</i> | 0.50 | 2.00 | 0.50 | 0.00 | 0.38 |
| 10 | MGU-InDel10- <i>o2</i> | 0.50 | 3.00 | 0.58 | 0.00 | 0.50 |
| 11 | MGU-InDel11- <i>o2</i> | 0.77 | 2.00 | 0.35 | 0.00 | 0.29 |
| 12 | MGU-InDel12- <i>o2</i> | 0.90 | 2.00 | 0.19 | 0.00 | 0.17 |
| 13 | Mean | 0.64 | 2.25 | 0.43 | 0.00 | 0.35 |

Fig. 2 Clustering pattern of the 48 diverse QPM and normal inbreds using *InDel* markers



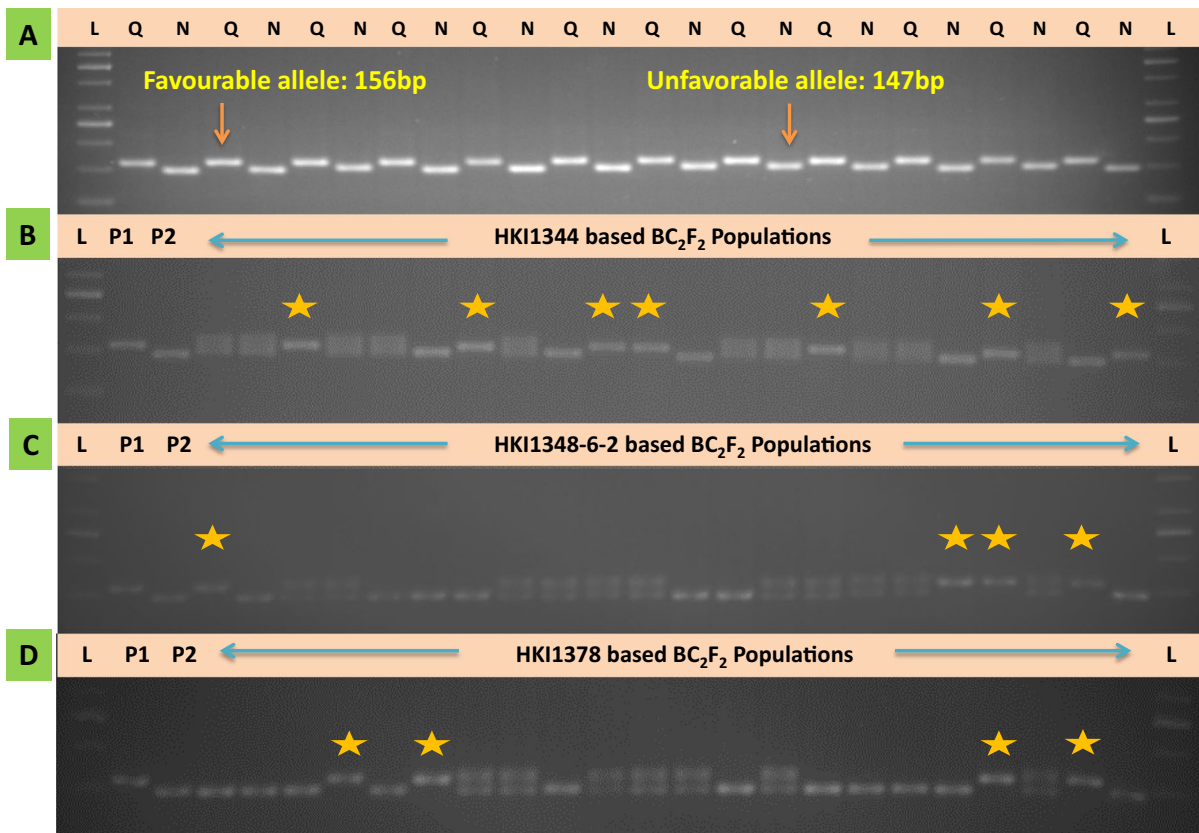


Fig. 3 Validation of Functional marker (*MGU-InDel10-o2*) in inbred panel and utilization in BC_2F_2 . **A:** Diverse inbred panel; **B:** HKI1344 based BC_2F_2 Population; **C:** HKI1348-6-2 based

BC_2F_2 Population; **D:** HKI1378 based BC_2F_2 Population; Q: QPM; N: Normal; Star indicate recessive *o2* segregants

Population structure and PCoA analysis

Structure harvester identified the two gene-based populations ($K=2$) in panel of 48 maize inbreds (Fig. S5). Population-I was composed of QPM inbreds, while population-II had only normal inbreds. The mean genetic differentiation value (F_{st}) of population-I, and population-II were 0.585 and 0.870, respectively, and the mean α -value was 0.028. The pairwise allele-frequency divergence between population-I and -II was 0.5623. Analysis of molecular variance (AMOVA) also illustrated significant divergence between the populations (28%), and within populations (72%) (Table S9). PCoA analysis revealed that normal inbreds were found to be uniformly distributed among the two quadrangles (-II and -IV) in a scatter plot, while QPM inbreds were clustered in the other two quadrangles (-I and -III). The first three PCoA axes together explained 81.80% cumulative

variance, with first, second and third axes explained 67.38%, 9.11%, and 5.31% of genetic variation, respectively (Fig. S5).

Development of functional marker specific to *o2*

Three potential SNPs (SNP601_T/C, SNP2359_C/T and SNP2360_G/T) responsible for substitution of amino acids and three exonic InDels (InDel_1 at position 656–658 bp, InDel_8 at position 2594–2596 bp and InDel_9 at position 2657–2662 bp) were exploited to develop two SNP based marker and two Indels markers. The InDel based markers viz., *MGU-InDel3-o2* and *MGU-InDel10-o2* were designed from exon-1 and exon-6. Of these two, a functional marker (*MGU-InDel10-o2*) covering InDel-8 and InDel-9 belonged to exon-6 was developed (Fig. S6). The marker (*MGU-InDel10-o2*) amplified 147 bp and 153 bp amplicons among the normal inbreds, while

156 bp amplicon was generated among QPM inbreds (Fig. 3). The MGU-InDel10-*o2* marker clearly distinguished the normal (dominant allele) and QPM (recessive allele) inbreds. While InDel based marker MGU-InDel3-*o2* unable to distinguish the normal and QPM inbreds. The SNP causing change from G (normal) to C (QPM) followed by substitution of glutamine (Q) to histidine (H), and dinucleotide CG (normal) to TT (QPM) coding for substitution of alanine (A) to serine (S), were used as the basis for the development of allelic specific SNP marker (MGU-SNP_G/C-*o2* and MGU-SNP_CG/TT-*o2*). The markers viz., MGU-SNP_G/C-*o2* and MGU-SNP_CG/TT-*o2* produced an amplicon of 93 bp and 87 bp, respectively. However, SNP-based PCR markers could not clearly distinguish the recessive allele from the dominant allele of *o2* (Fig. S7). Besides, three earlier reported *o2* gene-based SSRs viz., *phi112*, *phi057* and *umc1066* were also used for analysis among the 48 diverse inbreds. These SSRs also did not show clear-cut polymorphism between normal and QPM inbreds in the panel (Fig. S8).

Variation in lysine and tryptophan among diverse inbreds

Pooled ANOVA revealed significant variation for lysine and tryptophan among inbreds was present due to genotypes and locations×genotypes. Locations also showed significant variation for tryptophan (Table S10). However, contribution [(sum of square/ total sum of square)×100%] of location and location×genotype was very low (<1%) for lysine and tryptophan. Across locations, 24 QPM inbreds possessed higher lysine (0.350%) and tryptophan (0.085%) over the normal inbreds (lysine: 0.159%,

tryptophan: 0.037%). QPM inbred, BQPML-5122 had the highest lysine (0.406%), while PMIQ-549 possessed the highest tryptophan (0.101%). Among normal inbreds, CML-498 and CML-580 had the lowest lysine (0.124%) and tryptophan (0.029%) (Fig. S9; Table S11). Mean of haplotype-3 (Hap-3) based inbreds showed the highest average lysine (0.380%) and tryptophan (0.094%), while Hap-14 and Hap-16 had the lowest lysine (0.124%) and tryptophan (0.029%), respectively (Fig. S10).

Validation of novel functional marker in backcross populations

We analyzed the segregation of newly developed functional marker, MGU-InDel10-*o2* in three backcrossed-derived populations (BC₁F₁, BC₁F₂, and BC₂F₂) in three genetic backgrounds. In BC₁F₁ population, MGU-InDel10-*o2* identified of 46, 53, and 49 heterozygous plants in HKI344×PMI-102, HKI378×PMI-102, and HKI348-6-2×PMI-102 populations, respectively. Segregation of *o2* gene in all the three crosses followed Mendelian inheritance ratio of 1:1 (Table 3). In BC₂F₁ population the MGU-InDel10-*o2* identified of 51, 50, and 46 heterozygous plants in HKI344×PMI-102, HKI378×PMI-102, and HKI348-6-2×PMI-102 populations, respectively. Segregation of *o2* gene in all the three crosses followed Mendelian inheritance ratio of 1:1 (Table 3). In BC₂F₂ generation, MGU-InDel10-*o2* identified of 21, 14, and 23 homozygous recessive (*o2o2*), 25, 22, and 19 homozygous dominant (*O2O2*) and 48, 58, and 52 heterozygous plants in HKI344×PMI-102, HKI378×PMI-102 and HKI348-6-2×PMI-102 populations, respectively. Mendelian segregation pattern of 1:2:1 was followed in HKI344×PMI-102 and

Table 3 Segregation pattern of functional marker (MGU-InDel10-*o2*) specific to *o2* gene in different backcrosses and self-generations

| S. No. | Cross | Generations | NP | <i>O2O2</i> | <i>O2o2</i> | <i>o2o2</i> | χ^2 | P-value |
|--------|---------------------|--------------------------------|----|-------------|-------------|-------------|----------|--------------------|
| 1 | HKI1344×PMI-102 | BC ₁ F ₁ | 94 | 48 | 46 | – | 0.042 | 0.84 ^{NS} |
| | | BC ₂ F ₁ | 94 | 43 | 51 | – | 0.681 | 0.41 ^{NS} |
| | | BC ₂ F ₂ | 94 | 25 | 48 | 21 | 0.383 | 0.83 ^{NS} |
| 2 | HKI1378×PMI-102 | BC ₁ F ₁ | 94 | 41 | 53 | – | 1.532 | 0.22 ^{NS} |
| | | BC ₂ F ₁ | 94 | 44 | 50 | – | 0.383 | 0.54 ^{NS} |
| | | BC ₂ F ₂ | 94 | 22 | 58 | 14 | 6.511 | 0.05* |
| 3 | HKI1344-6-2×PMI-102 | BC ₁ F ₁ | 94 | 45 | 49 | – | 0.170 | 0.68 ^{NS} |
| | | BC ₂ F ₁ | 94 | 48 | 46 | – | 0.042 | 0.84 ^{NS} |
| | | BC ₂ F ₂ | 94 | 19 | 52 | 23 | 1.404 | 0.49 ^{NS} |

NP Number of plants, *O2* dominant allele, *o2* recessive allele, NS Non-significant

*Significant at P value of 0.5%

HKI348-6-2×PMI-102, while HKI378×PMI-102 showed segregation distortion in relation to *o2* gene (Table 3; Fig. 3). Selected three BC₂F₃ progenies having favourable allele of *o2* identified by MGU-InDel10-*o2* from three genetic background along with their original inbreds were analysed for lysine and tryptophan in the endosperm. Analysis revealed significantly higher accumulation of lysine (0.379%) and tryptophan (0.084%) in MAS derived QPM (*o2o2*) inbreds as compared to original (*O2O2*) inbreds (lysine: 0.189% and tryptophan: 0.039%) (Table S12).

Discussion

The importance of balanced-nutrition on human health has emerged as the most priority area after COVID-19 pandemic (WFP, WHO and UNICEF 2022). Of these protein with proper balance of essential amino acids is the key to the overall growth and development in humans (Hossain et al. 2023). Thus, efforts directed towards biofortification of maize with higher lysine and tryptophan assumes great significance (Gupta et al. 2015; Yadava et al. 2018). Molecular breeding provides great opportunity to develop *o2*-based QPM hybrids nearly half of the time required in conventional breeding (Hossain et al. 2018). Though three gene-based SSRs have been routinely used in the MAS programme, their usage is quite often limited due to lack of functional polymorphisms between the recipient and donor lines. Here, we characterized a set of normal and QPM inbreds for *o2* gene, and identified different haplotypes of *o2* followed by development and validation of functional marker for utilization in molecular breeding.

Nucleotides and amino acids variation analysis in *o2*

The successful development of suitable plant cultivars depends on utilization of diversity in genetic materials (Sood et al. 2014; Mayer et al. 2020). The sequence analysis of *o2* revealed that the differentiating SNPs and InDels were found more frequently in non-coding regions than coding regions of *o2* gene. Mutation in non-coding region possesses little influence on the gene's fitness, thus they are quickly fixed in the population (Chhabra et al. 2022). Several authors also performed the nucleotide-based *o2* analysis and found high degree of nucleotides variability

between QPM and normal inbreds (Schmidt et al. 1990; Henry et al. 2005; Gavazzi et al. 2007).

Amino acids substitutions sites viz., (i) alanine to serine, (ii) proline to valine, (iii) glutamine to histidine, (iv) valine to alanine, and (v) glycine to alanine, and insertion of valine and alanine were identified by comparing the normal and QPM amino acid sequence. Similar amino acids substitutions sites were also identified by Chen et al. (2014) in QPM inbreds. Heterocyclic amino acid such as proline (insertion of proline and substitution of valine in place of proline) influence the activity of the *o2* protein as the position of the proline residue affects the direction and degree of turns in the 3-D structure of the *o2* protein (Yang et al. 2004). The single amino acid substitution mutation in peptide sequence from a nonpolar hydrophobic (alanine) to a polar hydrophilic (serine) at 295 bp might cause loss in its spatial conformation, thereby leading to its decreased function of transcriptional activation of *o2* protein (Chen et al. 2014). Domain analysis showed that the C-terminal of basic leucine-zipper (bZIP-C) showed highest E-value (9.1e−18) among the identified domains present in the alignment sequence from position 316 to 435. In this domain region, we identified three substitutions (proline to valine, glycine to alanine and glutamine to histidine) of amino acids. While, Aukerman et al. (1991) showed an arginine to lysine substitution in the *o2* (bZIP) domain abolished specific DNA binding in QPM. In this study, SNPs based markers belonging to potential amino acids substitutions (alanine to serine and glutamine to histidine) did not show clear cut polymorphism between normal and QPM inbreds.

The formation of 3D models provides the better understanding of the structural differences of the proteins (Tufchi and Singh 2018). In the recessive *o2* protein structure, we found insertion of amino acids (alanine and proline-valine) in domain region which affect the 3D confirmation of *o2* protein. The *o2* motif recognized the core binding sites viz; TCACAT GTGT, TCATGCATGT and TCCACGTAGAT sites from the 22-kD α -zein gene promoter (Schmidt et al. 1992; Muth et al. 1996); ACGT and ACAT sites from the 19-kD α -zein gene promoter (Li et al. 2015) and the *O2*-like-box (TTTACGTAGAT) from the 27-kD γ -zein gene promoter (Wu and Messing 2012; Li et al. 2015) in maize. The molecular docking study of *o2* protein with promoter (DNA sequence) of zein protein revealed that most of the interactions were

having docking scores of more than the threshold cut-off of 200 kcal/mol with confidence score more than 0.7 except 27-kD γ -zein protein. The highest docking score for the interaction of maize protein with DNA sequence of promoter region in zein protein suggests that it requires less energy to bind protein with DNA (Yan et al. 2017). In present study, recessive *o2* protein interaction with promoter sequence of α -zein protein required more binding energy, while recessive *o2* protein interaction with promoter sequence of γ -zein protein required less binding energy as compared to dominant *O2* protein. Chromatin Immunoprecipitation Sequencing (ChIP-seq) study also revealed that *O2* directly is bound to the promoters of known targets (22-kD α -zein, 19-kD α -zein and 27-kD γ -zein gene) (Vicente-Carbajosa et al. 1997; Wu and Messing 2012; Li et al. 2015; Li and Song 2020). Transactivation of *O2* on the promoters of zein genes was earlier confirmed by dual-luciferase transactivation assay (Li et al. 2015). This information may play a key role in unveiling the structure behaviour and regulatory framework of *o2* protein for improving the nutritional value of the maize. Phylogenetic analysis revealed that normal and QPM inbreds were diverse. Yang et al. (2004) also analyzed 14 maize accessions of QPM along with normal maize using three *o2* gene-based markers (*umc1066*, *phi057* and *phi112*), and observed close relationship between recessive *o2*-based accessions.

Molecular diversity, population structure and haplotype analysis using InDel markers

Genetic diversity analyses based on gene-based markers identified an average of 2.25 alleles/loci, which was higher than Hossain et al. (2019b) (1.81 alleles/loci) and Chhabra et al. (2021) (2.00 alleles/loci) and comparable to Katral et al. (2022) (2.27 alleles/loci). PIC, major allele frequency and gene diversity were comparable to earlier study by Chand et al. (2022b) in QPM maize, Hossain et al. (2019b) in waxy maize and Chhabra et al. (2022) in sweet corn. Zero heterozygosity was observed in the study which indicated that all loci were homozygous and the alleles were fixed. Population structure analysis identified the two gene-based populations (QPM and normal) with the mean alpha value was 0.028. Das et al. (2019) also identified two populations for *vte4* gene based inbreds. A total of 24 haplotypes of *o2*

was identified among a set of 48 inbreds in the present study. Chhabra et al. (2021, 2022) identified 44 haplotypes for *sugary1* (*su1*) gene and 47 haplotypes for *shrunk2* (*sh2*) gene among 48 genotypes based on candidate gene-based markers. Katral et al. (2022) also reported 41 haplotypes from 48 accessions using InDel based markers specific to *fatb* gene.

Superiority of lysine and tryptophan in *o2* based inbreds

The recessive *o2* based inbreds (QPM) possessed an average of 2.19-fold and 2.26-fold more lysine and tryptophan over normal inbreds. The most promising genotypes had 2.55-fold and 2.70-fold lysine and tryptophan enhancement over the normal inbreds. All the QPM genotypes with higher lysine (0.350%) and tryptophan (0.085%) were grouped in Cluster-E, while most of the genotypes with low lysine (0.159%) and tryptophan (0.037%) were grouped in Cluster-F. The four haplotypes (Hap-1, Hap-2, Hap-3, and Hap-4) covering all QPM inbreds possessed more lysine and tryptophan, while the other haplotypes (Hap-5 to Hap-24) normal inbreds had lower lysine and tryptophan. Further, within QPM inbreds, Hap-3 showed the highest average lysine (0.380%) and tryptophan (0.094%), while Hap-4 had the lowest average lysine (0.292%) and tryptophan (0.076%) within QPM inbreds. In case of normal inbreds, Hap-13 and Hap-22 accumulated the highest lysine (0.213%) and tryptophan (0.047%) respectively, while Hap-14 and Hap-16 haplotypes accumulated and lowest lysine (0.124%) and tryptophan (0.029%), respectively. These InDel-based markers facilitated easy differentiation using agarose gel electrophoresis. PCR-based marker development for the SNPs are challenging as primer binding with a difference of only one base at the 3' end may not be precise leading to faulty assay. Further, use of next generation-based SNP genotyping of large number of lines are quite costly (Semagn et al. 2014; Talukder et al. 2023). The haplotype analysis using InDels based marker assay is simple and cheap, and can be used by any breeder having simple lab facility (Chhabra et al. 2021, 2022). The *o2* gene codes a bZIP protein that acts as a transcriptional factor for expression of α -zeins (Ueda et al. 1992). The recessive *o2* protein causes reduction in synthesis of zein protein by 50–70% primarily due to its less affinity of binding to the promoter regions (Kodrzycki

et al. 1989). The enhancement of nutritional quality in QPM inbreds is mainly due to reduction of lysine deficient zein proteins followed by enhanced synthesis of lysine-rich non-zein proteins (Habben et al. 1993). Recessive *o2* also significantly reduces transcription of LKR, the enzyme that degrades lysine in maize endosperm, thereby enhancing the concentration of lysine (Kemper et al. 1999). Further, *o2* is involved in regulation of various metabolic pathways and causes enhanced synthesis of various lysine-rich proteins and enzymes (Jia et al. 2013).

Functional marker and utilization in molecular breeding

Functional markers associated with *o2* can effectively help in breeding for higher accumulation of lysine and tryptophan. Here, we developed a functional marker (MGU-InDel10-*o2*) covering a 3 bp insertion (TGC termed as InDel-8) and 6 bp insertion (TGCCGG termed as InDel-9) in exon-6 region of *o2* gene. Marker-assisted backcross breeding has emerged as the most preferred approach for introgression of recessive *o2* allele as heterozygotes can be easily identified (Gupta et al. 2013). Additionally, lengthy progeny testing after each backcross is avoided (Hossain et al. 2018). Microsatellite markers present within *o2* genes available in public domain (Yang et al. 2004), however, may not necessarily be polymorphic between normal and QPM inbreds. Further, modification of repeat numbers due to replication slippage often leads to generation of new allele in the segregating populations. The functional marker developed here efficiently identified the foreground positive plants in all the segregating generations (BC₁F₁, BC₂F₁ and BC₂F₂) due to its codominance nature. The analysis also revealed that in most of populations, the marker segregated according to the predicted Mendelian ratio of 1:1 in the backcross generations and 1:2:1 in the selfed generation. Further, the functional marker unambiguously separated the recessive *o2* (QPM) inbreds from the normal inbreds.

The germplasm base of QPM inbreds with recessive *o2* gene is quite limited compared to the normal maize inbreds with dominant *O2* allele (Gupta et al. 2013). Thus, development of new diverse *o2*-based lines would strengthen the QPM breeding programme. The functional marker (MGU-InDel10-*o2*) can also be used to develop new QPM inbreds for

their utilization in the breeding programme. Large number of elite normal inbreds (*O2O2*) with high per se performance and desirable combining ability can be crossed with QPM donor lines (*o2o2*), and F₂ populations can be genotyped using the functional marker. The segregants homozygous for recessive *o2* gene can be selected and advanced to fix the inbreds for their utilization in the breeding programme. Similar effort has been undertaken in developing new inbreds with *crtRBI* gene governing provitamin-A accumulation in maize (Duo et al. 2021).

The sequencing of the entire *o2* gene (2,848 bp) costs ~US\$ 120 per sample. Thus, sequencing the entire *o2* gene among large set of inbreds involves considerable cost. On the contrary, the PCR-based functional marker developed here was effective in identifying genotypes with recessive *o2* allele with 100% accuracy. Further, the marker-assay involves only US\$1–1.5 per sample, thereby suggesting its utility in cost-effectiveness (Gain et al. 2023). In addition, functional marker developed here is also breeder-friendly as it requires basic lab facility with PCR machines, gel electrophoresis and gel documentation system. On the other hand, though next-generation sequence (NGS) based marker system is high throughput, it requires highly skilled personnel and considerable cost (Talukder et al. 2023). While, NGS-based such as Kompetitive Allele Specific PCR (KASP) involves high cost (US\$246 per sample) (Semagn et al. 2014). Further, accessibility of such NGS-based facilities in remote places further limits its utilization. The PCR-based functional marker system developed in the present study is quite affordable by the breeding programme especially in underdeveloped and developing countries where resources are limited.

Conclusion

The analysis revealed presence of wide diversity in *o2* allele in diverse maize inbreds. Specific SNPs and InDels present within the *o2* gene differentiated the normal and QPM inbreds. The most promising *o2* haplotype with higher accumulation of lysine and tryptophan can be the best source for the QPM breeding. The study provided the novel insight in to the DNA–protein interactions. The functional breeder-friendly marker developed here could identify the

desirable homozygotes from the heterozygotes in backcross populations. It also precisely identified the normal and QPM inbreds of diverse pedigree. The marker would help in accelerating the development of QPM hybrids worldwide. This is the first report on understanding natural allelic variation, haplotype analysis and functional marker development for *o2* gene regulating the enhancement of lysine and tryptophan in maize kernels.

Acknowledgements Financial support from the ICAR-Indian Agricultural Research Institute (IARI), New Delhi and ICAR-Consortia Research Platform on ‘Molecular breeding for improvement of tolerance to biotic and abiotic stresses, yield and quality traits in crops - Maize component’ (IARI Project Code No.: 12-143C) is duly acknowledged. We sincerely acknowledge IIMR-WNC, Hyderabad for providing the off-season nursery.

Author contributions Conduct of experiment: GC, Marker development: VM, Generation of segregating populations: RUZ, Estimation of lysine and tryptophan: SJM, Molecular docking analysis: GS, Conduct of inbred trial: BKM and VM, Writing of manuscript: GC and FH, Editing of manuscript: SS and TA, Designing of experiment: FH and TA.

Data availability No datasets were generated or analysed during the current study.

Declarations

Competing interests The authors declare no competing interests.

References

- Aukerman MJ, Schmidt RJ, Burr B, Burr FA (1991) An arginine to lysine substitution in the bZIP domain of an opaque-2 mutant in maize abolishes specific DNA binding. *Genes Dev* 5(2):310–320
- Babu R, Nair SK, Kumar A, Venkatesh S, Sekhar JC, Singh NN, Srinivasan G, Gupta HS (2005) Two-generation marker-aided backcrossing for rapid conversion of normal maize lines to quality protein maize (QPM). *Theor Appl Genet* 111:888–897
- Biovia DS (2020) Discovery studio visualizer. Dassault Systemes, San Diego
- Borevitz JO, Nordborg M (2003) The impact of genomics on the study of natural variation in Arabidopsis. *Plant Physiol* 132(2):718–725
- Bouis HE, Saltzman A, Birol E (2019) Improving nutrition through biofortification. *Agriculture for improved nutrition: Seizing the momentum*. CAB International, Wallingford, pp 47–57
- Chand G, Muthusamy V, Allen T, Zunjare RU, Mishra SJ, Singh B, Mehta BK, Talukder ZA, Ismail MR, Sarika K, Kamboj MC (2022a) Composition of lysine and tryptophan among biofortified-maize possessing novel combination of opaque2 and opaque16 genes. *J Food Compos Anal* 107:104376
- Chand G, Zunjare RU, Allen T, Muthusamy V, Mishra SJ, Kasana RK, Bhatt V, Ismail MR, Sarika K, Guleria SK, Hossain F (2022b) Genetic analysis of maize genotypes possessing novel combination of opaque2 and opaque16 genes affecting higher accumulation of lysine and tryptophan in kernels. *J Cereal Sci* 107:103534
- Chen Y, Zhou Z, Zhao G, Li X, Song L, Yan N, Weng J, Hao Z, Zhang D, Li M, Zhang S (2014) Transposable element rbg induces the differential expression of opaque-2 mutant gene in two maize *o2* NILs derived from the same inbred line. *PLoS ONE* 9(1):e85159
- Chhabra R, Muthusamy V, Baveja A, Katral A, Mehta B, Zunjare RU, Hossain F (2022) Allelic variation in shrunken2 gene affecting kernel sweetness in exotic-and indigenous-maize inbreds. *PLoS ONE* 17(9):e0274732
- Chhabra R, Muthusamy V, Gain N, Katral A, Prakash NR, Zunjare RU, Hossain F (2021) Allelic variation in sugary1 gene affecting kernel sweetness among diverse-mutant and-wild-type maize inbreds. *Mol Genet Genom* 296:1085–1102
- Das AK, Jaiswal SK, Muthusamy V, Zunjare RU, Chauhan HS, Chand G, Saha S, Hossain F (2019) Molecular diversity and genetic variability of kernel tocopherols among maize inbreds possessing favourable haplotypes of γ -tocopherol methyl transferase (*ZmVTE4*). *J Plant Biochem Biotechnol* 28:253–262
- De Castro E, Sigrist CJ, Gattiker A, Bulliar V, Langendijk-Genevaux PS, Gasteiger E, Bairoch A, Hulo N (2006) ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. *Nucleic Acids Res* 34:W362–W365
- DeLano WL (2009) The PyMOL molecular graphics system. Available online at: <http://www.pymol.org>
- Dellaporta SL, Wood J, Hicks JB (1983) A plant DNA mini-preparation: version II. *Plant Mol Biol Rep* 1:19–21
- Duo H, Hossain F, Muthusamy V, Zunjare RU, Goswami R, Chand G, Mishra SJ, Chhabra R, Gowda MM, Pal S, Baveja A (2021) Development of sub-tropically adapted diverse provitamin-A rich maize inbreds through marker-assisted pedigree selection, their characterization and utilization in hybrid breeding. *PLoS ONE* 16(2):e0245497
- Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol Ecol* 14(8):2611–2620
- Gain N, Chhabra R, Chandra S, Zunjare RU, Dutta S, Chand G, Sarika K, Devi EL, Kumar A, Madhavan J, Muthusamy V (2023) Variation in anthocyanin pigmentation by R1-navajo gene, development and validation of breeder-friendly markers specific to C1-inhibitor locus for in-vivo haploid production in maize. *Mol Biol Rep* 50(3):2221–2229
- Gavazzi F, Lazzari B, Ciceri P, Gianazza E, Viotti A (2007) Wild-type opaque2 and defective opaque2 polypeptides form complexes in maize endosperm cells and bind the opaque2-zein target site. *Plant Physiol* 145(3):933–945
- Goddard TD, Huang CC, Ferrin TE (2005) Software extensions to UCSF chimera for interactive visualization of large molecular assemblies. *Structure* 13(3):473–482

- Gunaratna NS, De Groote H, Nestel P, Pixley KV, McCabe GP (2010) A meta-analysis of community-based studies on quality protein maize. *Food Policy* 35(3):202–210
- Gupta HS, Hossain F, Muthusamy V (2015) Biofortification of maize: an Indian perspective. *Indian J Genet Plant Breed* 75(01):1–22
- Gupta HS, Raman B, Agrawal PK, Mahajan V, Hossain F, Thirunavukkarasu N (2013) Accelerated development of quality protein maize hybrid through marker-assisted introgression of opaque-2 allele. *Plant Breed* 132(1):77–82
- Habben JE, Kirleis AW, Larkins BA (1993) The origin of lysine-containing proteins in opaque-2 maize endosperm. *Plant Mol Biol* 23:825–838
- Henry AM, Manicacci D, Falque M, Damerval C (2005) Molecular evolution of the Opaque-2 gene in *Zea mays* L. *J Mol Evol* 61:551–558
- Hossain F, Muthusamy V, Pandey N, Vishwakarma AK, Baveja A, Zunjare RU, Thirunavukkarasu N, Saha S, Manjaiah KM, Prasanna BM, Gupta HS (2018) Marker-assisted introgression of opaque2 allele for rapid conversion of elite hybrids into quality protein maize. *J Genet* 97:287–298
- Hossain F, Muthusamy V, Zunjare RU, Gupta HS (2019) Biofortification of maize for protein quality and provitamin-a content. Nutritional quality improvement in plants. Springer International Publishing, Cham, pp 115–136
- Hossain F, Chhabra R, Devi EL, Zunjare RU, Jaiswal SK, Muthusamy V (2019) Molecular analysis of mutant granule-bound starch synthase-I (waxy1) gene in diverse waxy maize inbreds. *3 Biotech* 9:1–10
- Hossain F, Prasanna BM, Kumar R, Singh BB (2008) Genetic analysis of kernel modification in quality protein maize (QPM) genotypes. *Indian J Genet Plant Breed* 68(1):1–9
- Hossain F, Zunjare RU, Muthusamy V, Kumar A, Madhavan J, Ikkurti G, Katral A, Talukder ZA, Chhabra R, Chand G, Bhatt V (2023) Genetic improvement of specialty corn for nutritional quality traits. *Maize Improvement: current advances in yield, quality, and stress tolerance under changing climatic scenarios*. Springer International Publishing, Cham, pp 235–257
- Hu B, Jin J, Guo AY, Zhang H, Luo J, Gao G (2015) GSDB 2.0: an upgraded gene feature visualization server. *Bioinformatics* 31(8):1296–1297
- Hunter BG, Beatty MK, Singletary GW, Hamaker BR, Dilkes BP, Larkins BA, Jung R (2002) Maize opaque endosperm mutations create extensive changes in patterns of gene expression. *Plant Cell* 14(10):2591–2612
- Jia M, Wu H, Clay KL, Jung R, Larkins BA, Gibbon BC (2013) Identification and characterization of lysine-rich proteins and starch biosynthesis genes in the opaque2mutant by transcriptional and proteomic analysis. *BMC Plant Biol* 13(1):1–14
- Jompuk C, Cheuchart P, Jompuk P, Apisitwanich S (2011) Improved tryptophan content in maize with opaque-2 gene using marker assisted selection (MAS) in backcross and selfing generations. *Agric Nat Resour* 45(4):666–674
- Katral A, Muthusamy V, Zunjare RU, Chhabra R, Maman S, Yadava DK, Hossain F (2022) Allelic variation in Zmfatb gene defines variability for fatty acids composition among diverse maize genotypes. *Front Nutr* 9:1–9
- Kemper EL, Neto GC, Papes F, Moraes KCM, Leite A, Arruda P (1999) The role of opaque2 in the control of lysine-degrading activities in developing maize endosperm. *Plant Cell* 11(10):1981–1993
- Kodrzycki R, Boston RS, Larkins BA (1989) The opaque-2 mutation of maize differentially reduces zein gene transcription. *Plant Cell* 1:105–114
- Kumar S, Stecher G, Li M, Knyaz C, Tamura K (2018) MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol Biol Evol* 35(6):1547
- Li C, Song R (2020) The regulation of zein biosynthesis in maize endosperm. *Theor Appl Genet* 133(5):1443–1453
- Li C, Qiao Z, Qi W, Wang Q, Yuan Y, Yang X, Tang Y, Mei B, Lv Y, Zhao H, Xiao H (2015) Genome-wide characterization of cis-acting DNA targets reveals the transcriptional regulatory framework of opaque2 in maize. *Plant Cell* 27(3):532–545
- Liu K, Muse SV (2005) PowerMarker: an integrated analysis environment for genetic marker analysis. *Bioinformatics* 21(9):2128–2129
- Maqbool MA, Beshir IA, Khokhar ES (2021) Quality protein maize (QPM): importance, genetics, timeline of different events, breeding strategies and varietal adoption. *Plant Breed* 140(3):375–399
- Mayer M, Hölker AC, González-Segovia E, Bauer E, Prestler T, Ouzunova M, Melchinger AE, Schön CC (2020) Discovery of beneficial haplotypes for complex traits in maize landraces. *Nat Commun* 11(1):4954
- Mertz ET, Bates LS, Nelson OE (1964) Mutant gene that changes protein composition and increases lysine content of maize endosperm. *Sci* 145(3629):279–280
- Motto M, Maddaloni M, Ponziani G, Brembilla M, Marotta R, Di Fonzo N, Soave C, Thompson R, Salamini F (1988) Molecular cloning of the o2-m5 allele of *Zea mays* using transposon marking. *Mol Gen Genet* MGG 212:488–494
- Muth JR, Müller M, Lohmer S, Salamini F, Thompson RD (1996) The role of multiple binding sites in the activation of zein gene expression by Opaque-2. *Mol Gen Genet* MGG 252:723–732
- Neeraja CN, Hossain F, Hariprasanna K, Sewa R, Satyawathi CT, Longvah T, Raghu P, Voleti SR, Sundaram RM (2022) Towards nutrition security of India with biofortified cereal varieties. *Curr Sci* 123(271):271–277
- Nuss ET, Tanumihardjo SA (2010) Maize: a paramount staple crop in the context of global nutrition. *Compr Rev Food Sci Food Saf* 9(4):417–436
- Nyakurwa CS, Gasura E, Mabasa S (2017) Potential for quality protein maize for reducing proteinenergy undernutrition in maize dependent Sub-Saharan African countries: a review. *Afr Crop Sci J* 25(4):521–537
- Peakall R, Smouse PE (2006) Genalex 6: genetic analysis in excel. *Population genetic software for teaching and research*. *Mol Ecol Notes* 6:288–295
- Perrier X, Flori A, Bonnot F (2003) Data analysis methods. In: Hamon P, Seguin M, Perrier X, Glaszmann JC (eds) *Genetic diversity of cultivated tropical plants*. Science Publishers, Enfield, pp 43–76
- Prasanna BM, Palacios-Rojas N, Hossain F, Muthusamy V, Menkir A, Dhliwayo T, Ndhlela T, San Vicente F, Nair SK, Vivek BS, Zhang X (2020) Molecular breeding for

- nutritionally enriched maize: status and prospects. *Front Genet* 10:1392
- Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genet* 155(2):945–959
- Rozas J, Ferrer-Mata A, Sánchez-DelBarrio JC, Guirao-Rico S, Librado P, Ramos-Onsins SE, Sánchez-Gracia A (2017) DnaSP 6: DNA sequence polymorphism analysis of large data sets. *Mol Biol Evol* 34(12):3299–3302
- Sarika K, Hossain F, Muthusamy V, Zunjare RU, Baveja A, Goswami R, Bhat JS, Saha S, Gupta HS (2018) Marker-assisted pyramiding of opaque2 and novel opaque16 genes for further enrichment of lysine and tryptophan in subtropical maize. *Plant Sci* 272:142–152
- Schmidt RJ, Burr FA, Burr B (1987) Transposon tagging and molecular analysis of the maize regulatory locus opaque-2. *Sci* 238:960–963
- Schmidt RJ, Burr FA, Aukerman MJ, Burr B (1990) Maize regulatory gene opaque-2 encodes a protein with a “leucine-zipper” motif that binds to zein DNA. *Proc Natl Acad Sci USA* 87(1):46–50
- Schmidt RJ, Ketudat M, Aukerman MJ, Hoschek G (1992) Opaque-2 is a transcriptional activator that recognizes a specific target site in 22-kD zein genes. *Plant Cell* 4(6):689–700
- Semagn K, Babu R, Hearne S, Olsen M (2014) Single nucleotide polymorphism genotyping using kompetitive allele specific PCR (KASP): overview of the technology and its application in crop improvement. *Mol Breed* 33:1–14
- Shiferaw B, Prasanna BM, Hellin J, Bänziger M (2011) Crops that feed the world 6. Past successes and future challenges to the role played by maize in global food security. *Food Sec* 3:307–327
- Sood S, Flint-Garcia S, Willcox MC, Holland JB (2014) Mining natural variation for maize improvement: selection on phenotypes and genes. *Genomics of plant genetic resources: managing, sequencing and mining genetic resources*. Springer, Netherlands, Dordrecht, pp 615–649
- Talukder ZA, Chhabra R, Muthusamy V, Zunjare RU, Hossain F (2023) Development of novel gene-based markers for waxy1 gene and their validation for exploitation in molecular breeding for enhancement of amylopectin in maize. *J Appl Genet* 64:1–10
- Tufchi M, Singh NK (2018) In silico prediction of 3D structure of Opaque-2 protein in maize. *Res Crops* 19(3):526–536
- Ueda T, Wawerczak W, Ward K, Sher N, Ketudat M, Schmidt RJ, Messing J (1992) Mutations of the 22-and 27-kD zein promoters affect transactivation by the Opaque-2 protein. *Plant Cell* 4(6):701–709
- Van Berloo R (2008) GGT 2.0: versatile software for visualization and analysis of genetic data. *J Hered* 99(2):232–236
- Vasal SK, Villegas E, Bjarnason M, Gelaw B, Goertz P (1980) Genetic modifiers and breeding strategies in developing hard endosperm opaque-2 materials. *Improv Qual Traits Maize Grain Silage* 37:73
- Vicente-Carbajosa J, Moose SP, Parsons RL, Schmidt RJ (1997) A maize zinc-finger protein binds the prolamin box in zein gene promoters and interacts with the basic leucine zipper transcriptional activator Opaque2. *Proc Natl Acad Sci USA* 94(14):7685–7690
- Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, Heer FT, de Beer TAP, Rempfer C, Bordoli L, Lepore R (2018) SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res* 46(W1):W296–W303
- WFP, WHO, UNICEF (2022) The state of food security and nutrition in the world 2022.
- Wu Y, Messing J (2012) Rapid divergence of prolamin gene promoters of maize after gene amplification and dispersal. *Genet* 192(2):507–519
- Yadava DK, Hossain F, Mohapatra T (2018) Nutritional security through crop biofortification in India: status & future prospects. *Indian J Med Res* 148(5):621
- Yan Y, Zhang D, Zhou P, Li B, Huang SY (2017) HDOCK: a web server for protein–protein and protein–DNA/RNA docking based on a hybrid strategy. *Nucleic Acids Res* 45(W1):W365–W373
- Yang J, Zhang Y (2015) I-TASSER server: new development for protein structure and function predictions. *Nucleic Acids Res* 43(W1):W174–W181
- Yang W, Zheng Y, Ni S, Wu J (2004) Recessive allelic variations of three microsatellite sites within the O₂ gene in maize. *Plant Mol Biol Rep* 22:361–374
- Yang Z, Bai Z, Li X, Wang P, Wu Q, Yang L, Li L, Li X (2012) SNP identification and allelic-specific PCR markers development for TaGW2, a gene linked to wheat kernel weight. *Theor Appl Genet* 125:1057–1068. <https://doi.org/10.1007/s00122-012-1895-6>
- Zhang Z, Lu C (2023) Assessing changes in potential yields and yield gaps of summer maize in the North China Plain. *Food Energy Secur* 13(1):e489
- Zhao G, Li M, Zhang D, Li X, Wu Z, Ci X, Xie C, Bai L, Lu Z, Chen L, Hao Z (2012) Kernel lysine content does not increase in some maize opaque2 mutants. *Planta* 235:205–215

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.