SCIENCE CHINA PRESS

SCIENCE CHINA Life Sciences
life.scichina.com link.springer.com
https://doi.org/10.1007/s11427-024-2725-1

LETTER TO THE EDITOR

Genomic analysis of modern maize inbred lines reveals diversity and selective breeding effects

Kaijian Fan[1,2,3†], Mohsin Ali[1,2†], Kunhui He[1,2†], Yingwei Feng[1,2], Tingxi Yu[1,2], Hao Zhang[1,2], Tai An[1,3], Weiwei Zeng[1,3], Junjie Fu[1], Ying Zhou[1,2], Yanfang Heng[1,3], Fei Gu[4,5], Jiankang Wang[1,2], Changling Huang[1,2,3], Liang Li[1,3*] & Huihui Li[1,2*]

[1]State Key Laboratory of Crop Gene Resources and Breeding, Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, CIMMYT-China Office, Beijing 100081, China;
[2]Nanfan Research Institute, Chinese Academy of Agricultural Sciences, Sanya 572024, China;
[3]National Engineering Research Center of Crop Molecular Breeding, Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing 100081, China;
[4]DAMO Academy, Alibaba Group, Hangzhou 310023, China;
[5]Hupan Lab, Hangzhou 310023, China

†Contributed equally to this work
*Corresponding authors (Huihui Li, email: lihuihui@caas.cn; Liang Li, email: liliang05@caas.cn)

Received 17 June 2024; Accepted 10 September 2024; Published online 15 November 2024

Dear Editor,

Maize (*Zea mays* L.) is one of the most widely cultivated crops worldwide (Ranum et al., 2014), accounting for about 37.2% of the global cereal crop production. Maize production yield has increased over sevenfold since single-cross maize hybrids were deployed in the beginning of the last century (Andorf et al., 2019). To date, intensive breeding efforts have improved the density tolerance, yield potential, disease resistance, stress tolerance, and adaptability of maize hybrids (Duvick, 2005; Luo et al., 2023; Wang et al., 2011). In the present study, we have conducted genome wide selection analysis and machine learning feasibility test in a large maize germplasm panel consisting of 2,430 modern inbred lines from diverse geographic regions with varying desirable traits. Deep neural network genomic prediction (DNNGP) (Wang et al., 2023) models were built to predict plant architecture and yield-related traits using rare and favorable loci across different germplasms.

A diversity panel of 2,933 maize inbred lines consisted of 2,430 modern inbred lines, which were derived from commercial hybrids and 503 diverse inbred lines, were widely used in breeding (Table S1 in Supporting Information). These inbred lines were utilized to explore rare and favorable allelic variations across different germplasms with an ambition to understand the missing genetic variations in the current maize breeding population and further to design breakthrough varieties in maize breeding. By resequencing of the entire panel, calling and filtering single-nucleotide polymorphism (SNP), we selected and used 437,081 high quality SNP markers in the following analyses (Table S2 in Supporting Information). The SNP distribution plot revealed that chromosome (Chr) 1 (16.43%) and Chr 6 (7.27%) had the highest and the least number of SNPs, respectively (Figure 1A; Table S2 in Supporting Information). These SNPs were distributed almost in the entire genome, covered a physical distance of 2,105.41 Mb ranging from 150.94 Mb of Chr 10 to 306.96 Mb of Chr 1. The average distance between neighbored pair-wise SNPs across each chromosome varied from 4.27 kb in Chr 1 to 5.47 kb in Chr 6.

Genetic variations in the maize diversity panel were assessed by annotating SNPs with genetic location (e.g., intron or exon) and potential effects on gene function (e.g., missense). A total of 2,648,748 putative genomic variations were identified between the 2,933 lines and the B73 genome assembly (Table S3 in Supporting Information). Of these variations, missense and frameshift mutations accounted for 5.14% and 0.08%, respectively. The rest of the variations either located in introns/intergenic regions (85.53%) or resulted in synonymous mutations (9.26%). Downstream region of genes held the most of SNPs, followed by introns, upstream, exons, intergenic, 3′ UTR, and transcript variant regions, covering 23.57%, 18.85%, 16.64%, 13.18%, 9.89%, 6.80%, and 6.57% of the total SNPs (Table S4 in Supporting Information). Relatively, a few SNPs were distributed in the 5′UTR, splice regions, acceptor splice sites, and splice donor sites. The synonymous/non-synonymous SNPs ratio is 136,064/213,000 (i.e., 0.64), indicating adaptive selections in the panel.

Inferring the population genetic structure of the panel through unsupervised ADMIXTURE analysis revealed a low-diversity clustering, which contradicts previous reports (Yang et al., 2011; Yang et al., 2010; Zhang et al., 2018). Subsequently, semi-supervised ADMIXTURE analyses produced 11 clusters denoted as K1 to K11 (Figure S1 and Table S1 in Supporting Information), with the highest number of inbred lines clustered in K2 (830) and the fewest in K9 (45). The K1 subgroup, referred to as Lvda Red Cob, contained 180 (6.14%) inbred lines including the elite ones, E28 and Lv28. The K2 (28.30%) and K3 (9.96%) were mainly derived from Monsanto and Pioneer hybrids, which included few traditional inbred lines. The K4 subgroup, referred to as X (including Jing724), was the second largest group consisting of 676 lines mainly derived from XY335 hybrid. The K5 and K6 subgroups, referred to as Lancaster (145) and BSSS (218), contained the male parents of Mo17 and Zi330, and female parents of B73 and B37. The K7 subgroup, referred to as P78599 (98), consisted of P138 and Qi319 disease-resistant inbred lines that were mainly derived from P78599 hybrid. The lines in K8 subgroup, referred to as

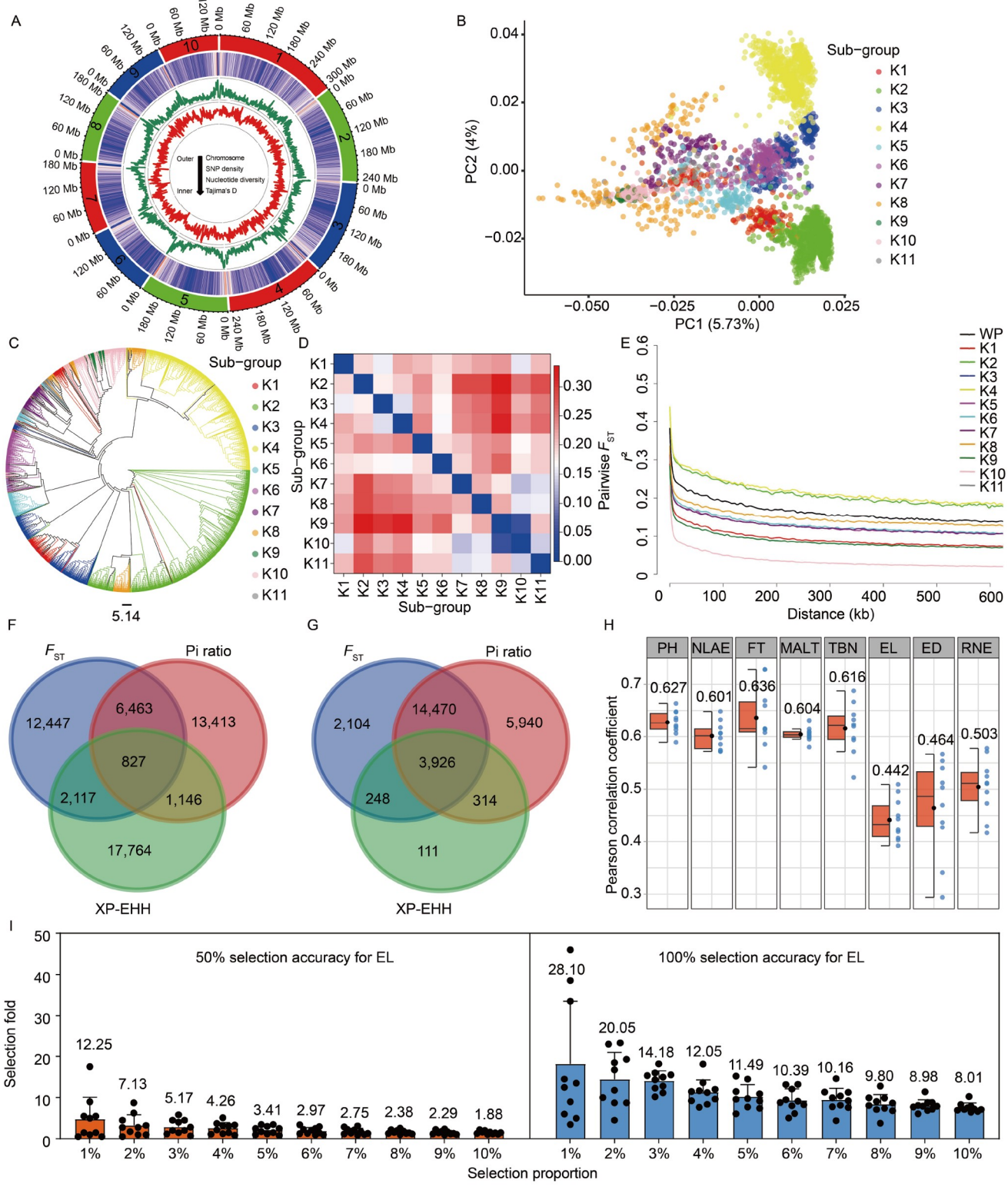

Figure 1. Genomic features, population structure, and genomic prediction across a maize panel containing 2,933 lines. A, The distribution of various genomic features across 2,933 maize panels. B, PCA plot of the first two components (PC1 and PC2). C, Neighbor-joining tree of all 2,933 inbred lines was generated based on whole-genome SNPs. D, Heatmap of pairwise fixation index (F_{ST}) values among the 11 subgroups. E, Distance of linkage disequilibrium estimated from the whole panel and the 11 subgroups measured using the squared correlation coefficient (r^2) against distance. WP represents the whole panel. F, Venn diagram showing the number of overlapping or unique loci with a large difference in term of allelic frequencies across germplasm among F_{ST}, cross population extended haplotype homozygosity (XP-EHH) and nucleotide diversity (Pi) ratio values. G, Venn diagram showing the number of overlapping and unique candidate genes identified using F_{ST}, XP-EHH and Pi ratio values. H, The box plot showing the accuracies (Pearson correlation coefficient) of DNNGP models for eight plant architecture and yield-related traits. PH, plant height; NLAE, number of leaves above the ear; FT, flowering time; MALT, main axis length of the tassel; TBN, tassel branch number; EL, ear length; ED, ear diameter; RNE, number of rows per ear. I, The required individual fold (selection fold) of EL under 50% and 100% selection accuracy.

https://doi.org/10.1007/s11427-024-2725-1

TSPT (184), were mainly derived from the Huangzaosi inbred line. The K9 and K10 subgroups, referred to as Tropical I (45) and Tropical II (211), contained tropical and sub-tropical inbred lines. The lines in the K11 cluster, referred to as Reid (54), were primarily developed from American commercial hybrids.

Principal component analysis (PCA) based on semi-supervised population structure showed that the top 10 PCs accounted for about 25% of the genomic variation, in which the first three PCs explained 5.73%, 4.00%, and 2.45% (Figure S2 in Supporting Information). Scatter plots of the top three PCs revealed a weak differentiation between subgroups, with high overlaps among K1, K5, K6, K7, K10 and K11. In contrast, K2 and K3 were clearly differentiated from the others (Figure 1B; Figure S3 in Supporting Information). A distance-based phylogenetic analysis revealed that clustering patterns were consistent with the PCA results (Figure 1C).

The findings of this study suggested new heterotic groups through introducing novel germplasm into the panel. To evaluate their genetic relationships, pairwise kinship coefficients and genetic distances were calculated (Tables S5–S6 and Figure S4 in Supporting Information). The average relative kinship between any two lines was 0.0065, ranging from 0 to 3.52 across the panel (Table S5 in Supporting Information). Sixty-five percent of the relative kinship values were 0, indicating a weak genetic similarity between the lines (Figure S4A in Supporting Information). The average genetic distance among inbred lines was 0.03, varying from 0 to 0.37 across the panel (Table S6 in Supporting Information). We found that 87% of genetic distance comparisons among inbred lines fell between 0.10 and 0.30, suggesting a relatively large genetic distance among them. In alignment with the PCA results, the genome-wide pairwise genetic fixation index (F_{ST}) between the 11 subgroups was high, except among K1, K5, K6, K7, K10, and K11 (Figure 1D). The analysis of molecular variation (AMOVA) revealed that 77.13% of the variance was explained within subgroups, compared to 20.16% between subgroups, indicating a relatively rich genetic diversity within subgroups (Table S7 in Supporting Information).

Genome-wide linkage disequilibrium (LD) was examined for the 11 subgroups, revealing a rapid decrease in LD with an increasing physical distance (Figure 1E). K9 and K10 had the slowest LD decay, indicating that they may have a higher genetic diversity. To further investigate the genetic diversity among the 2,933 inbred lines and 11 subgroups, we conducted analyses of heterozygosity (Ho), expected heterozygosity (He), polymorphic information content (PIC), and nucleotide diversity (Pi) using 437,081 SNPs (Table S8 and Figures S5–S7 in Supporting Information). The average of Ho was 0.01, ranging from 0 to 0.15, which was lower than that of He (0.23, ranging from 0.09 to 0.50), suggesting a potential directional selection. The average PIC of subgroups K9 (0.21) and K10 (0.24) was higher than that of the whole panel (0.20, ranging from 0.09 to 0.38), which indicates a higher genetic diversity. This result is consistent with the results of Pi and LD analysis. Furthermore, the Tajima's D values were positive (Tajima's D>0) for the whole panel and all subgroups, indicating a balancing selection. However, the lower average values for K2 (0.96) and K3 (1.06) implied a relatively lower selection pressure (Table S8 in Supporting Information).

Results of this study revealed that a large proportion of genetic variation was utilized better in modern maize breeding lines. The F_{ST}, cross population extended haplotype homozygosity (XP-EHH), and Pi ratio values (Park et al., 2012) were computed in the diversified panel of maize inbred lines (Figure S8 in Supporting Information). Implementing the top 5% as a cutoff, the three methods identified 21,854 loci with a large allelic frequency difference across the germplasm (Figure 1F; Tables S9–S11 in Supporting Information). To facilitate demonstration, these loci were named as enriched loci hereafter. Among them, 827 were detected by all of the three methods. A total of 6,463, 2,117, and 1,146 were detected by F_{ST} and Pi ratio, F_{ST} and XP-EHH, and Pi ratio and XP-EHH, respectively. A total of 54,177 non-redundant loci, accounting for 12.40% of all SNPs, were identified from the three analyses (Figure 1F).

The confidence intervals (CIs) for enriched loci encompassed 27,113 genes, representing 69.5% of all maize genes (Tables S9–S11 in Supporting Information). Of the encompassed genes, 3,926 were located within the CIs of 827 enriched loci (Figure 1G). In addition, 14,470 genes were identified by both F_{ST} and Pi ratio. Fewer overlapping genes were found between F_{ST} and XP-EHH (248) or Pi ratio and XP-EHH (314). Moreover, 605 genes with known functions (72.7% of total functional genes) were found in the CIs of enriched loci. GO analysis suggested the enrichment of these genes for various biological processes, including flowering time, root development, stress resistance, yield, and plant architecture building.

To verify whether machine learning models could facilitate genomic selection in plant breeding programs, we constructed genomic selection models for eight plant architecture and yield-related traits using DNNGP algorithm with the 54,177 non-redundant enriched loci for 2,430 modern maize inbred lines. The Pearson correlation coefficient of the eight traits ranged from 0.442 to 0.636. Flowering time (FT) showed the highest predictive accuracy, while ear length (EL) had the lowest accuracy (Figure 1H; Table S12 in Supporting Information). Plant architecture traits showed a significantly higher accuracy than yield-related traits.

We conducted simulations for yield-related traits to determine the required size of the breeding populations at different selection proportions. The results reveal that the required individual fold significantly decreases with the selection proportion increases. Taking EL for example, an average of 12.25- and 28.10-time of the individual number is required to be planted in a 1% selection proportion to achieve 50% and 100% selection accuracies, respectively. By contrast, only 1.88 and 8.01 times individuals were required to achieve 50% and 100% selection accuracies under a 10% selection proportion. This indicates that a much larger breeding population is required to achieve 100% selection accuracy compared to that is required to achieve 50% selection accuracy, regardless of selection proportion (Figure 1I; Figure S9 and Table S13 in Supporting Information). Hence, selection proportion is one of the key factors in genomic selection-based breeding strategy to ensure improved genetic gain of yield-related traits.

This study constructed a large and diversified maize germplasm panel, representing the advanced commercial germplasm. Two new subgroups were identified from this panel, which could potentially complement the widely used heterotic groups in maize breeding pro-

grams. A large proportion of the rare and favorable loci were deciphered and validated using deep neural network prediction models, which could be pyramided in genomic design breeding to develop new maize varieties.

Compliance and ethics

The authors declare that they have no conflict of interest.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (32361143514, 32272190), Nanfan Special Project, Chinese Academy of Agricultural Sciences (YBXM2407), Alibaba Foundation, and Innovation Program of Chinese Academy of Agricultural Sciences (CAAS-CSIAF-202303).

Supporting information

The supporting information is available online at https://doi.org/10.1007/s11427-024-2725-1. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

Andorf, C., Beavis, W.D., Hufford, M., Smith, S., Suza, W.P., Wang, K., Woodhouse, M., Yu, J., and Lübberstedt, T. (2019). Technological advances in maize breeding: past, present and future. Theor Appl Genet 132, 817–849.

Duvick, D.N. (2005). The contribution of breeding to yield advances in maize (*Zea mays* L.). Adv Agron 86, 83–145.

Luo, N., Meng, Q., Feng, P., Qu, Z., Yu, Y., Liu, D.L., Müller, C., and Wang, P. (2023). China can be self-sufficient in maize production by 2030 with optimal crop management. Nat Commun 14, 2637.

Park, D.J., Lukens, A.K., Neafsey, D.E., Schaffner, S.F., Chang, H.H., Valim, C., Ribacke, U., Van Tyne, D., Galinsky, K., Galligan, M., et al. (2012). Sequence-based association and selection scans identify drug resistance loci in the *Plasmodium falciparum* malaria parasite. Proc Natl Acad Sci USA 109, 13052–13057.

Ranum, P., Peña-Rosas, J.P., and Garcia-Casal, M.N. (2014). Global maize production, utilization, and consumption. Ann N Y Acad Sci 1312, 105–112.

Wang, K., Abid, M.A., Rasheed, A., Crossa, J., Hearne, S., and Li, H. (2023). DNNGP, a deep neural network-based method for genomic prediction using multi-omics data in plants. Mol Plant 16, 279–293.

Wang, T., Ma, X., Li, Y., Bai, D., Liu, C., Liu, Z., Tan, X., Shi, Y., Song, Y., Carlone, M., et al. (2011). Changes in yield and yield components of single-cross maize hybrids released in China between 1964 and 2001. Crop Sci 51, 512–525.

Yang, X., Gao, S., Xu, S., Zhang, Z., Prasanna, B.M., Li, L., Li, J., and Yan, J. (2011). Characterization of a global germplasm collection and its potential utilization for analysis of complex quantitative traits in maize. Mol Breeding 28, 511–526.

Yang, X., Yan, J., Shah, T., Warburton, M.L., Li, Q., Li, L., Gao, Y., Chai, Y., Fu, Z., Zhou, Y., et al. (2010). Genetic analysis and characterization of a new maize association mapping panel for quantitative trait loci dissection. Theor Appl Genet 121, 417–431.

Zhang, R., Xu, G., Li, J., Yan, J., Li, H., and Yang, X. (2018). Patterns of genomic variation in Chinese maize inbred lines and implications for genetic improvement. Theor Appl Genet 131, 1207–1221.