Predicted protein 3D structure provides essential insights into the genetic architecture underlying phenotypic diversity in maize

Shuai Wang, Merritt Khaipho-Burch, Lynn C. Johnson, et al.

Genome Res. published online October 31, 2025
Access the most recent version at doi:10.1101/gr.280514.125

Predicted protein 3D structure provides essential insights into the genetic architecture underlying phenotypic diversity in maize

Shuai Wang[1,2], Merritt Khaipho-Burch[3], Lynn C. Johnson[4], Zachary R. Miller[4], Peter J. Bradbury[5], Doug Speed[6], William J. Allen[7], M. Cinta Romay[4], Jiquan Xue[1], Edward S. Buckler[3,4,5], Guillaume P. Ramstein[6*], Baoxing Song[1,2*]

1 Key Laboratory of Maize Biology and Genetic Breeding in Arid Areas of the Northwest Region, College of Agronomy, Northwest A&F University, Yangling, Shaanxi 712100, China

2 Peking University Institute of Advanced Agricultural Sciences, Shandong Laboratory of Advanced Agriculture Sciences in Weifang, Weifang, Shandong 261325, China

3 Section of Plant Breeding and Genetics, Cornell University, Ithaca, New York 14853, USA

4 Institute for Genomic Diversity, Cornell University, Ithaca, New York 14853, USA

5 Agricultural Research Service, United States Department of Agriculture, Ithaca, New York 14853, USA

6 Center for Quantitative Genetics and Genomics, Aarhus University, Aarhus 8000, Denmark

7 Texas Advanced Computing Center, University of Texas at Austin, Austin, Texas 78758, USA

Abstract:

Variation in protein 3D structures reflects genetic variation and contributes to phenotypic diversity, yet its underlying genetic mechanisms remain unclear. To investigate the relationship between protein 3D structure and phenotype, we predicate the 3D structures of 795,649 proteins from 26 maize (*Zea mays* L.) inbred lines using AlphaFold2. Population genetics analysis of these protein 3D structures reveal that buried residues held greater genomic evolutionary rate profiling (GERP) scores than exposed residues, indicating that buried residues are under stronger purifying selection. The design of the maize nested association mapping population makes it possible to

utilize haplotype information and protein 3D structural variation to reveal the molecular mechanisms linking genetic diversity and phenotypic variation for a population with ~5,000 individuals. Associating protein 3D structure variation with phenotypes (structure-based proteome-wide association study, PWAS) identifies 15.7% more (96 vs. 83) significant proteins compared to associating protein sequence with phenotypes (sequence-based PWAS) using 32 agronomic traits. Moreover, structure-based PWAS identifies 24 additional significant proteins unique to predicted structures, while sequence-based PWAS identifies 12 additional significant proteins. Structure-based proteome-wide predictions (PWP) improves genomic prediction accuracy by an average of 3.8% compared to sequence-based PWP. In general, predicted protein 3D structures represent a powerful approach for understanding the natural diversity of protein haplotypes.

Introduction:

Tremendous progress has been made over the past decade in linking genomic variation to phenotypic variance (Xiao et al. 2017; Tam et al. 2019; Klein et al. 2010; Atwell et al. 2010; Togninalli et al. 2020; Wu et al. 2023; Wang et al. 2023; Ramstein and Buckler 2022). However, understanding the molecular mechanisms underlying sequence phenotype relationships remains a major challenge. Genome-wide predictions can now reach high accuracy within narrow germplasm pools but are less accurate for many complex traits in individuals with greater genetic distance (Windhausen et al. 2012; Desta and Ortiz 2014). This lack of transferability has been observed in both human and agricultural contexts. Moreover, different populations may exhibit distinct architectures of the causal variants underlying complex traits (Werner et al. 2005; Song, Mott, and Gan 2018; Keys et al. 2020), with rare alleles and different allele frequencies. In addition, even the most accurate whole-genome prediction models tend to overlook molecular mechanistic functions and fail to model aspects such as disruptions in protein activity.

Phenotypic variation broadly results from the modification of molecular processes within cells (Doerge 2002) where proteins are among the most important functional molecules. The biological function of a given protein is dictated by the arrangement of the atoms and functional groups in its three-dimensional (3D) structure. Therefore, variation in protein 3D structure is an essential source of information to explain how coding sequence variants impact gene activity (Hicks et al. 2019). Due to the extremely high cost and throughput bottleneck of protein 3D structure investigation technology, omics-wide comparisons of 3D structure have rarely been conducted for alleles at the population scale within eukaryotic species or across related species. AlphaFold2 (Jumper et al. 2021) enables researchers to perform proteome-wide 3D structure predictions without having to generate crystals for all individual proteins encoded by a genome.

Proteins carry diverse selection signatures resulting from structural and functional constraints (MacGowan et al. 2024). One prime determinant of structural constraints is the relative solvent accessible surface area (RSA) of residues. Buried residues accept lower evolutionary substitutions and are more structurally constrained than exposed residues (Chothia and Lesk 1986). Structural constraints are reflected in the allele frequency spectrum, which is commonly used as an indication of natural selection (Vishnoi et al. 2011). Evolutionary measures of sequence conservation provide important insights into structural constraints (Sun et al. 2023; Kistler et al. 2018; Davydov et al. 2010). Therefore, understanding how natural selection shaped protein 3D architecture is crucial for identifying functionally important regions that contribute to genetic adaptation to the environment.

Genome-Wide Association Studies (GWASs) have identified many candidate causal loci in protein-coding genes and regulatory regions. These genomic variations impact traits largely by modulating the dosage or functional activity of proteins. Association studies based on gene expression have also been conducted at population scale. These analyses are often based on

mRNA abundance (G. Yang et al. 2024), which demonstrates a moderate correlation with protein abundance (Mergner et al. 2020). Protein 3D structure, as determinant of protein function, also influences various important biological activities. An improved understanding of protein 3D structural variation might provide critical insights into the molecular mechanisms underlying complex traits (Gerasimavicius, Teichmann, and Marsh 2025). To capture such effects, association analysis and genomic prediction integrating 3D structural variation into quantitative genetics might offer a concrete, functionally interpretable framework for analyzing phenotypic variation.

Considering the high computational cost and lack of annotated protein sequences, applying AlphaFold2 to large-scale natural populations is challenging. Artificially designed populations with a limited number of founder lines makes it possible to use AlphaFold2 to explore the genetic mechanisms underlying phenotypic diversity. In this context, the maize (*Zea mays* L.) nested association mapping (NAM) population (McMullen et al. 2009; Gage et al. 2020) provides a valuable resource to link genotypic variation with phenotypic variation using computationally folded protein structures with affordable computing resources. The genomes of 26 NAM founder lines have been *de novo* assembled using long sequencing reads, and their protein sequences have been annotated (Hufford et al. 2021). The NAM population has been characterized by over 100 different phenotypes in different environments over tens of millions of plants, ranging from agronomic characteristics to omics profiles (Gage et al. 2020; Khaipho-Burch et al. 2023). The use of this unique genetic resource for powerful AI-based structure prediction offers a valuable opportunity to investigate the genetic mechanisms underlying phenotypic diversity.

In this study, we predicted protein 3D structures for 26 maize founder inbred lines and projected these structures onto the NAM inbred population with ~5,000 individuals. Based on these predictions, we investigated signatures of natural selection acting on protein structure. To

examine the impact of structural variation on phenotypes, we utilized the increased association power of protein 3D structural variation by conducting a structure-based proteome-wide association study (PWAS). Additionally, we explored the contribution of protein 3D structural variation to the accuracy of phenotypic prediction using proteome-wide prediction (PWP). Our findings suggest that protein structures have significant potential for enhancing quantitative and population genetics analyses.

Results:

Protein structure predictions for 26 diverse maize inbred lines

To investigate the structural diversity, we used AlphaFold2 to predict the structure of proteins for the maize inbred line B73 and the 25 other NAM founder lines. Whether AlphaFold2 can predict the effects of single amino-acid polymorphisms (SAPs) is controversial (Stein and Mchaourab 2024), so we first compared the protein sequence from the same core pan-gene group to characterize the sequence diversity of orthologous proteins among the 26 maize NAM founder lines (Hufford et al. 2021). We identified sequence variants in approximately 70% of pairwise protein comparisons, while only ~10% differed by a single point mutation (Fig. 1A), indicating the high level of sequence diversity within this panel.

We predicted the 3D structures of 68,262 proteins from B73 (v5.0) (including canonical proteins from 37,724 genes) using AlphaFold2 (Jumper et al. 2021), as B73 has been widely used as the reference maize line for over a decade (Schnable et al. 2009). For the other 25 founder lines, we merged their canonical protein sequences into a nonredundant collective protein set (excluding protein sequences identical to their respective B73 reference). Overall, we predicted the structures of 313,639 unique protein sequences, corresponding to 795,649 proteins for all 26 NAM founder lines (Supplemental Fig. S1A, B). The folded protein structures represent 94.1% of all proteins from B73 and approximately 73% of proteins from the NAM founder lines (Fig. 1B and

Supplemental Table S1). To assess the prediction confidence for each protein and each residue, we used the predicted local distance difference test score (pLDDT) (Jumper et al. 2021). At a per-protein level, 59.2% and 49.1% of all proteins showed high confidence (median pLDDT ≥ 70) in B73 and the NAM founder lines, respectively (Fig. 1B and Supplemental Table S1). A more stringent cut-off (median pLDDT ≥ 90) predicted 33.6% (B73) and 29.6% (NAM founder lines) of folded proteins with very high confidence (Fig. 1B). At the per-residue level, 61.0% (B73; Supplemental Fig. S2A) and 64.4% (NAM founder lines; Fig. 1C) of residues were predicted with high confidence (pLDDT ≥ 70, Supplemental Table S2). The pLDDT scores for each residue followed a clear bimodal distribution (Fig. 1C and Supplemental Fig. S2A), similar to that of humans (Alderson et al. 2023).

RSA is a measure of the exposure of an amino acid residue to the protein's solvent (Ramsey et al. 2011; Savojardo et al. 2020), with values ranging from 0 for a totally buried residue to 1 for a totally exposed residue. RSA followed a bimodal distribution, with low pLDDT scores mainly distributed in the high RSA region and high pLDDT scores mainly distributed in the low RSA region for both B73 (Supplemental Fig. S2B) and the NAM founder lines (Fig. 1D). Furthermore, RSA was significantly negatively correlated with pLDDT (Supplemental Fig. S3). Residues with high confidence scores are potentially functional domains and are tightly folded (Akdel et al. 2022). Structural domains are fundamental units critical to biological function, typically ranging in length from 100 to 500 residues (Wheelan, Marchler-Bauer, and Bryant 2000). Our results reveal that high confidence levels are predominantly enriched in domain-like length regions in B73 (Supplemental Fig. S2C) and the NAM founder lines (Fig. 1E). The folded maize proteins structures with high sequence diversity represent valuable resources to investigate the evolution of protein structures and conduct quantitative genetics analysis.

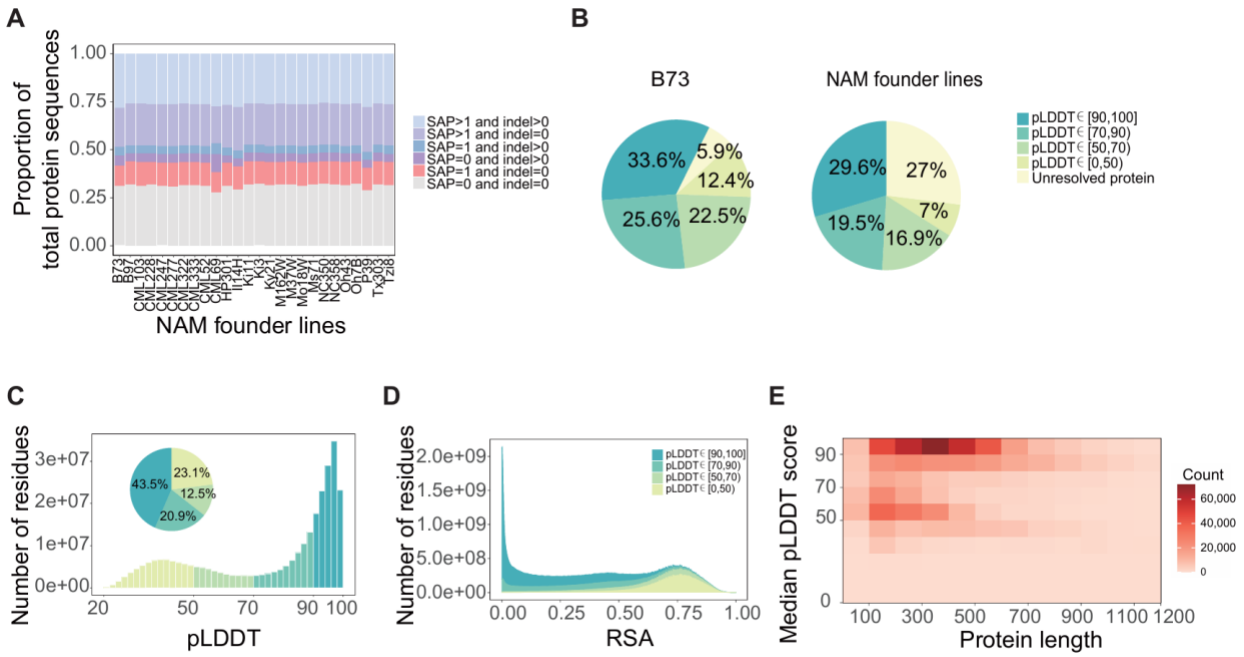

Fig. 1. Protein 3D structure predictions for 26 maize NAM founder lines using AlphaFold2. (A) Protein sequence diversity in maize NAM founder lines. Each NAM founder line was compared with other NAM founder lines in pairs. SAP, single amino acid polymorphism; indel, insertion/deletion. (B) Percentage of predicted protein 3D structures at different confidence levels in B73 (left, total count: 72,539) and NAM founder lines (right, total count: 1,089,242); colors represent median confidence scores. Unresolved protein structures were not predicted; see Methods. (C) Distribution and percentage of confidence scores for all NAM founder lines residues. The histogram shows the density distribution of confidence scores, and the pie chart shows the percentage of confidence scores. pLDDT, predicted local distance difference test. (D) Distribution of RSA for residues in all NAM founder lines. RSA, relative solvent accessible surface area. (E) Heatmap representing median confidence scores of protein structures as a function of protein length in all NAM founder lines.

Buried residues are under stronger purifying selection than exposed residues

Natural selection acting on protein 3D structures plays key roles in environmental adaptation and fitness with structural constraints representing one of the main determinants of the rates of evolution (Wolf, Wolf, and Koonin 2008). Therefore, we aimed to explore the molecular mechanisms underlying natural selection via predicted 3D structures. We identified 495,357 nonsynonymous and 490,815 synonymous single nucleotide polymorphisms (SNPs) (Bukowski et al. 2018) from the maize 282 association panel (Flint-Garcia et al. 2005). High confidence and buried regions (with high pLDDT and low RSA) held higher proportions of synonymous SNPs than nonsynonymous SNPs, indicating that these regions are likely subject to stronger purifying selection pressure (p-value < 2.2×10^{-16}, Fisher's Exact Test, Fig. 2A, 2B). To further investigate how natural selection acted on protein structure, we used *Zea mays ssp. mexicana* and *Zea mays ssp. parviglumis* as outgroups to infer ancestral states and observed that nonsynonymous SNPs in the lowest derived allele frequency bin held the lowest mean RSA value compared to the other bins (p-value < 0.05, one-way analysis of variance and Fisher's least significant difference test, Fig. 2C), suggesting that purifying selection acts on buried residues.

Genomic sites with higher genomic evolutionary rate profiling (GERP) scores are interpreted to be under stronger purifying selection (J. Yang et al. 2017; Davydov et al. 2010). Genomic sites at buried residues (RSA < 0.5) had greater GERP scores than those at exposed residues (RSA ≥ 0.5), suggesting that buried residues are under stronger purifying selection (p-value < 2.2×10^{-16}, t-test, Supplemental Fig. S4 and Fig. 2D). RSA is negatively correlated with pLDDT (Supplemental Fig. 2), and pLDDT decreases substantially when the median alignment depth is less than ~30 sequences in the multiple sequence alignment (MSA) from the AlphaFold2 model (Jumper et al. 2021). A high depth of MSA is likely to reflect sequence conservation. To

avoid this source of confounding (Chakravarty and Porter 2022), we analyzed the strength of purifying selection conditionally on pLDDT bins. In each bin, residues with low solvent accessibility held higher GERP scores than residues with high solvent accessibility (Fig. 2E), suggesting that they are subjected to stronger purifying selection than those with high solvent accessibility, even after accounting for the confounding effect of pLDDT. Moreover, $\pi N/\pi S$ (the ratio of nonsynonymous to synonymous nucleotide site diversity) was lower in buried regions than exposed regions, even when considering the confounding effect of pLDDT scores, confirming purifying selection in buried regions (Supplemental Fig. S5). Taken together, these observations indicate that predicted protein 3D structure by AlphaFold2 at population scale could reveal associations between structural features of protein and evolutionary constraint, providing key functional insights into natural selection.

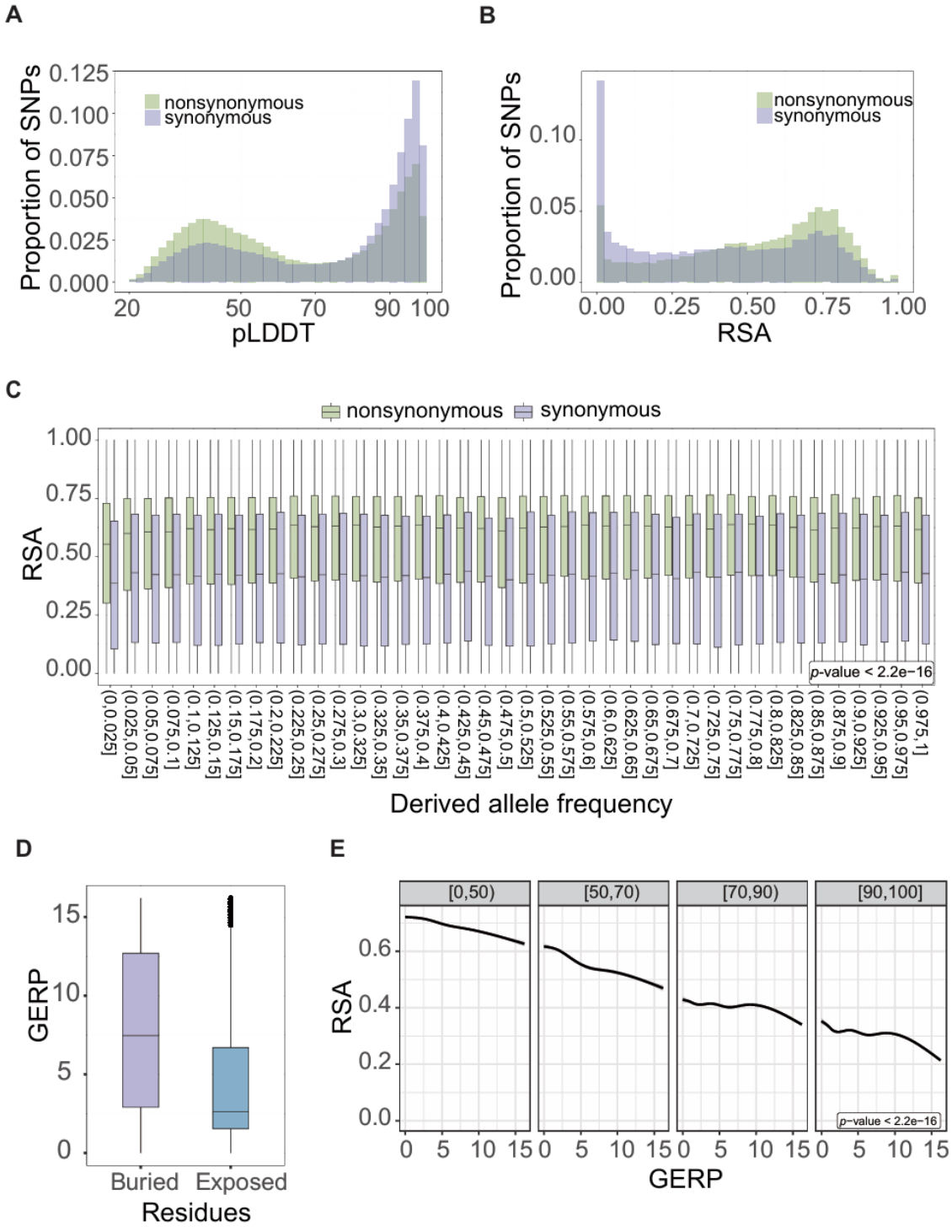

Fig. 2 Protein structure predictions provide functional insights into natural selection in maize. (A) Density distribution of pLDDT scores between synonymous and nonsynonymous SNPs. (B) Density distribution of RSA between synonymous and nonsynonymous SNPs. (C) Distribution of

RSA for different derived allele frequencies. (D) Relationship between GERP scores and buried (RSA < 0.5) or exposed residues. Median value is shown as a solid middle line within the boxplot, which spans from the 25th to 75th percentiles. For definition of buried and exposed residues, please see Supplemental Fig S4. (E) Correlation between RSA and GERP scores divided by pLDDT score bins (as indicated by the gray shading above the plots).

Predicted structural variants provide novel insights into genetic effects on phenotypes

To conduct association analysis based on protein variation, we measured the structural similarity as well as sequence similarity for the canonical isoforms of each core pan-gene group. We used the principal components of the structure- and sequence-based similarity matrices as independent variables to perform association analysis on 32 traits. To account for population structure and relatedness, we included a genome wide relationship matrix (based on genomic SNPs) and a genome-wide IBD matrix (based on protein haplotypes) in a linear mixed model (Supplemental Fig. S6). Our PWAS focused on PCs (Principal Components) with eigenvalues higher than a predefined threshold (1×10^{-5}), which included most PCs with non-zero variance (Supplemental Fig. S7 A). Lowering this predefined threshold resulted in near-identical results from structure-based tests (Supplemental Fig. S7 B, C, D, E). It is noteworthy that this lower threshold also resulted in spurious associations with extremely high significance values (e.g., pangene 9228, with $-\log10(p) > 200$; Supplemental Table S3). On average, the associations between predicted structure and phenotype were stronger than those between sequence and phenotype (Fig. 3A and Supplemental Fig. S8), suggesting higher statistical power through structure-based associations. We identified 108 significant proteins through both structure-based and sequence-based association analyses (96 and 84 by structure- and sequence-based tests, respectively). Of these, 72 proteins were shared between the two methods, while structure-based association identified 24 additional significant protein-trait associations unique to predicted structures (Fig.

3B; Supplemental Fig. S9 and Supplemental Table S4), suggesting that structure-based association analysis is a valuable and informative complement to sequence-based association analysis. The variability of principal components was almost identical with different significant levels for all significant proteins (Supplemental Fig. S10). To test whether the relative advantage of structure-based PWAS was not due to the use of specific sequence-based MSA software, we conducted the analysis using two other MSA tools: MUSCLE and T-Coffee (Notredame, Higgins, and Heringa 2000; Edgar 2004). The protein sequence similarity matrices generated from different MSA approaches were highly correlated with each other but different from the structural similarity matrices (Supplemental Fig. S11). The observed improvement in statistical significance from predicted protein structures was consistent across MSA software (Supplemental Fig. S12; Supplemental Fig. S13). Furthermore, we investigated whether the significant proteins identified via structure-based PWAS could be identified via standard SNP-based GWAS. The results suggested structured-based PWAS identified additional loci that were not identified by SNP-based GWAS (Fig. 3C, D).

To assess our ability to identify candidate causal genes via structure-based PWAS, we inspected associated protein structures encoded by known genes with phenotypes (Supplemental Fig. S14). The predicted protein structure of the transcription factor LIGULELESS1 (encoded by *Zm00001eb067740*) was significantly associated with upper leaf angle. *LIGULELESS1* regulates ligule and leaf angle development and the mutant phenotype of this gene has been genetically studied (Li et al. 2017; Mantilla-Perez and Salas Fernandez 2017). The predicted protein structure of TUBTF6 (encoded by *Zm00001eb188370*) was significantly associated with upper leaf angle. This gene was reported to be enriched in a leaf gene regulatory network (Bertolini et al. 2025). DWARF and IRREGULAR LEAF1 (DWIL1; encoded by *Zm00001eb287100*) was associated with node number above the ear and was previously reported to influence plant height, internode length, and potentially node number (Jiang et al. 2012). GNARLEY (encoded by

Zm00001eb117820) was associated with node number above the ear. Its mutation alters the overall plant architecture (Foster et al. 1999), and its ortholog in rice*, Oryza sativa homeobox 15 (OSH15)*, affects node development (Sato et al. 1999). Due to linkage disequilibrium (LD), the extra power gained from structured PWAS might not point to the causal genes. The predicted structure of GOLDEN2-LIKE 37 (GLK37, encoded by *Zm00001eb073790*) showed a stronger association with ear row number than its sequence, and *KRN2* (Zm00001eb073740) is in high LD with *GLK37*. *KRN2* was previously reported to regulate ear row number (Wenkang Chen et al. 2022). Genetic research is needed to confirm and further understand the underlying molecular mechanisms of these significant protein-traits associations.

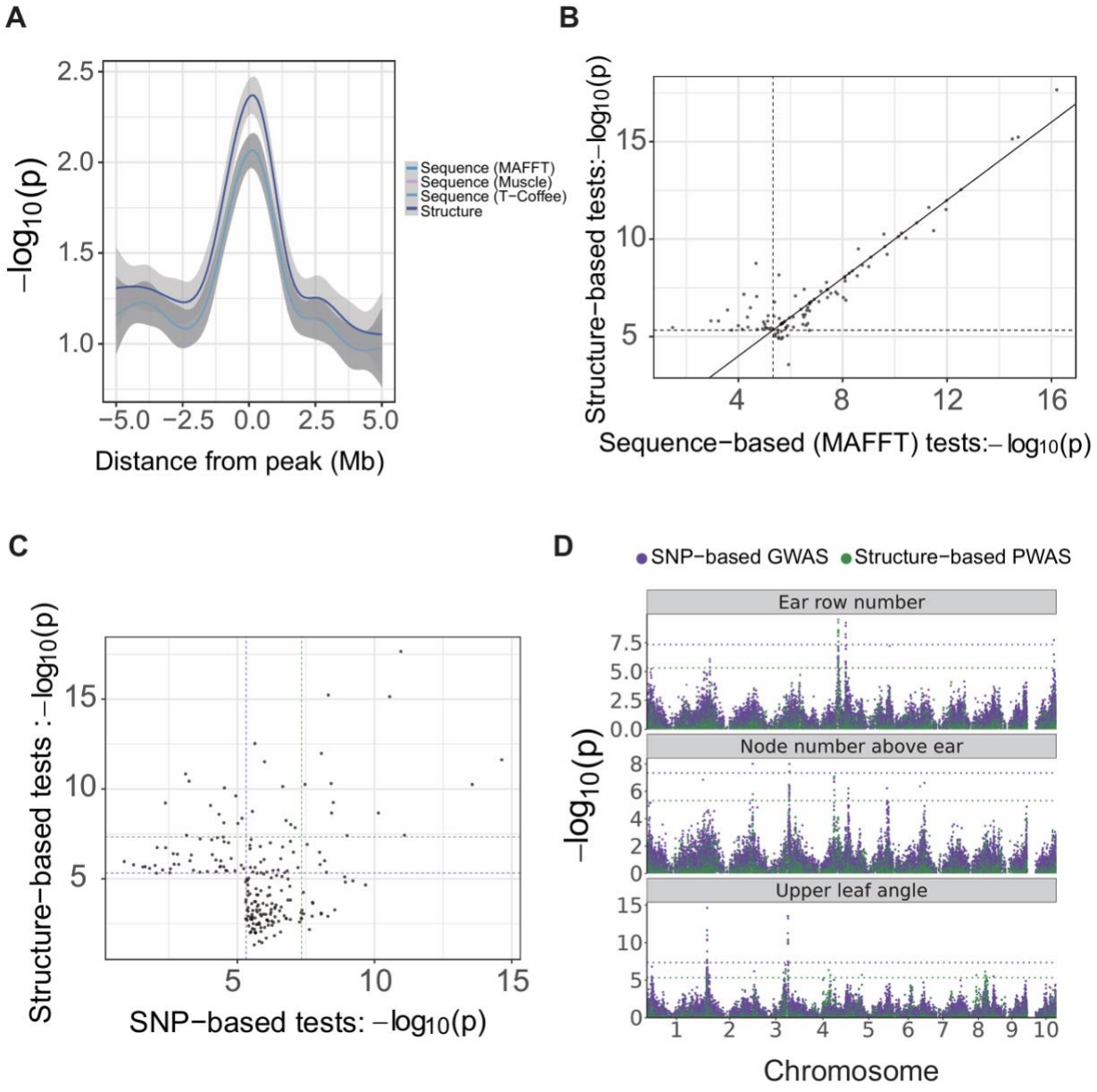

Fig. 3 Comparison between association analyses.

(A) Decay of significance over physical distance from significance peaks in PWAS. The peak is at the starting point of a significant gene. (B) Significant association signals from structure- and sequence-based PWAS. The dashed lines correspond to a 5% significance threshold with Bonferroni correction. (C) Significant loci from structure-based PWAS and SNP-based GWAS. The dashed lines correspond to a 5% significance threshold with Bonferroni correction (5.32 in

green for structure-based PWAS, 7.34 in purple for SNP-based GWAS). In SNP-based GWAS, significant genes are selected based on the most significant SNP in coding regions. (D) Manhattan plot of structure-based PWAS and SNP-based GWAS. Traits shown are ear row number, node number above ear and upper leaf angle. The red dashed lines are identical to those shown in Fig. 3C.

Protein 3D structure improves genomic prediction accuracy

We assessed the additional contributions of protein structure variation to the genomic variance of the 32 traits in the NAM population, by performing variance partitioning using a genomic prediction model. Since the genotypic variants used for association are not fully independent of each other, we estimated their effect sizes (i.e., the amount of phenotypic variance explained) using a mixed model and included the whole genome SNP-based genomic value as an independent variable. We constructed three proteome-wide similarity matrices using structure, sequence, and IBD from all pan-gene groups. We first estimated the variance explained by the sequence similarity matrix, then added the protein structural similarity matrix as an additional source of variance (Fig. 4A). The inclusion of the structural similarity matrix provided a better fit compared to using the sequence similarity matrix alone (likelihood ratio test, p-value <0.05, Fig. 4A). Structure-based PWP improved prediction accuracy by 3.8% on average compared to sequence-based PWP (p-value =1.1×10^{-7}, paired t-test, Fig. 4B), where node number below ear, fructose content, and glucose content yielded more than 10% improvement in prediction accuracy. Structure-based PWP incorporating sequence information improved prediction accuracy by 4.1% compared to sequence-based PWP on average (p-value = 2.7×10^{-7}, paired t-test, Fig. 4C), where fumarate, node number below ear, protein, fructose content and glucose content yielded more than a 10% improvement in prediction accuracy. The increased prediction accuracy was independent of MSA methods used (Supplemental Fig. S15 and Supplemental Fig. S16). Structure-based similarities, alone or combined with IBD-based similarities, resulted

in increased prediction accuracy compared to IBD-based similarity alone (1.9% and 2.5%, respectively; Supplemental Fig. S17), while node number below ear led to the highest improvement in prediction accuracy (12.0% and 11.4%). These results indicate that protein structures explain genotypic variation not captured by protein sequence or IBD information, which translates into higher genomic prediction accuracy.

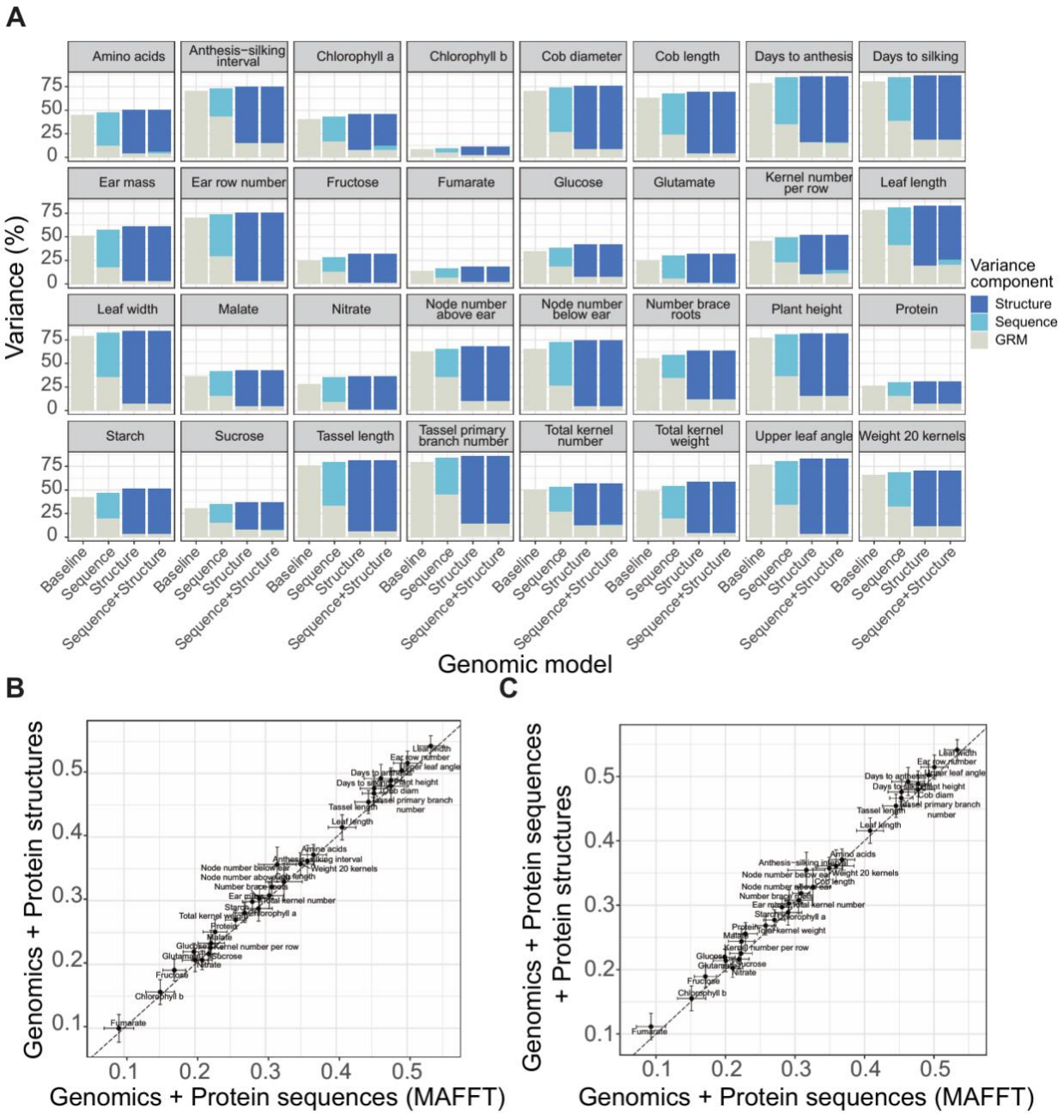

Fig. 4 Protein 3D structure improving prediction accuracy compared to protein sequence. (A) Variance partitioning in different genomic models for 32 traits. Sequences were aligned using MAFFT. Likelihood values are listed in Supplemental Table S5. (B) Comparison of prediction accuracy between sequence-based PWP and structure-based PWP. (C) Comparison of prediction accuracy between sequence-based PWP and sequence-based plus structure-based PWP. Dots represent the mean values of 25 leave-one-family-out repeats, and error bars indicate their standard errors.

Discussion:

Although functional features have been incorporated into association testing or genomic prediction, leading to significant improvements (Weiwei Chen et al. 2024; Ramstein and Buckler 2022), the use of protein 3D structure for PWAS and PWP remains unexplored. Information on protein 3D structure provided molecular insights into the consequences of natural selection. The structure variants at population scale enabled us to perform structure-based PWAS and PWP to identify genes associated with important agronomic traits. Our results represent a pioneering exploration of how predicted protein 3D structural diversity can enhance our understanding of phenotypic variation.

Prediction of protein structure requires high-quality genome sequence and annotation

Prediction of protein structure heavily depends on high-quality protein sequences, as the 3D conformation of a protein is determined by the arrangement of its amino acids. Sequencing errors or gaps in the sequence can lead to inaccurate or misleading predicted structures. Therefore, ensuring the completeness and accuracy of the genomic sequence data is crucial for reliable structure prediction. Highly accurate gene annotation (the process of finding and designating the locations of individual genes) is also required to improve structure predictions (Salzberg 2019). With the advent of Oxford Nanopore Technology (ONT) and PacBio HiFi

sequencing technologies, more and more genomes are being assembled at the Telomere-to-Telomere (T2T) level (Xie et al. 2024). As a result, the acquisition of high-quality protein sequences has become increasingly feasible, enabling large-scale protein structure predictions and opening new avenues for the exploration of functional mechanisms.

Genetic effects are captured by high confidence protein 3D structure

Proteome-wide protein 3D structure prediction accurately captures structural differences between genes in the reference genome of human and other model organisms (Tunyasuvunakool et al. 2021; Akdel et al. 2022). As protein sequence alleles are conserved across mammalian species, AlphaFold2 might not have the power to predict the impact of SAPs (Stein and Mchaourab 2024). However, AlphaFold2 may effectively predict allelic differences in protein structures in plant species, which are significantly more diverse than mammal species (Song, Buckler, and Stitzer 2023; Buckler, Gaut, and McMullen 2006). In this study, we extend proteome-wide protein 3D structure prediction to a population scale by taking advantage of the design of the maize NAM population. This allowed us to use predicted structure to analyze phenotypic variation in quantitative genetics studies, thereby enabling the decomposition of genotypic variability which would otherwise not be possible.

Predicted protein structural similarity is more informative than protein sequence similarity

Structural similarity between proteins is a strong predictor of functional similarity (Erdin, Lisewski, and Lichtarge 2011). Vastly different amino acid sequences can lead to similar structures, while even two closely related sequences may fold into distinctly different structures (Krissinel 2007). Our structure-based PWAS results suggest that 3D structures predicted from sequences are more useful than sequences alone for discovering candidate genes. Our method

filtered some PCs with low variance, which introduced extra false positives in PWAS. We examined several candidate genes that contribute to phenotypic variation. Although we identified promising candidates, sequence differences in proteins do not necessarily affect their predicted 3D structure. For omic-wide associations, we applied the canonical Bonferroni multiple tests correction to adjust significance thresholds for SNPs and pan-gene groups separately. Owing to the substantially larger number of SNPs compared to pan-gene groups, the resulting significance thresholds (expressed as -log10(*p*-value)) were higher for SNPs. Nevertheless, even applying the same threshold to SNP-based GWAS and PWAS, the observation that structure-based PWAS provides complementary information to SNP-based GWAS still holds (Fig. 3C). In our analysis, we only kept pan-genes that were present in all individuals, while additional genes are dispensable or private to some founder lines. Our analysis focused on sequence variability and did not consider potential effects of gene absence. Moreover, the dosage of functional molecules in cells, which is also essential for trait regulation (Wingo et al. 2021), could not be captured by predicting protein structure. Including dosage regulation and protein complexes in the model might further improve our understanding of the genetic mechanisms underlying trait diversity. Predicted 3D structure may only describe one component of gene activity, independent from gene regulation. Therefore, our proposed structure-based PWAS approach should ultimately be combined with other approaches (e.g., proteome-wide dosage association studies) in order to obtain a detailed biological description of genotypic variability.

Predicted protein structural variants capture genotypic variability

There have been mixed results about the usefulness of AlphaFold2 for predicting the effect of single-residue variants on protein structure (Buel and Walters 2022; McBride et al. 2023). Our PWAS and PWP results clearly indicate that AlphaFold2 can capture genotypic variability for agronomic, morphological, and compositional traits in maize. Importantly, predicted protein

structural variants in our study consisted of haplotype variants, not necessarily single-residue variants. Based on our results, we hypothesize that AlphaFold2 can accurately predict allelic variants for protein structure, provided that there are large enough differences among protein alleles. This hypothesis calls for further research to quantify the structural differences between protein alleles, and to test the accuracy of their prediction.

In our analyses, precision was prioritized over power, as our PWAS approach was essential to account for biases introduced by LD, by including both SNP-based relationship and haplotype-based IBD matrices (Supplemental Fig. S6). A previous study predicted mRNA abundance for each haplotype in the NAM population, similar to our prediction of protein structures (Giri et al. 2021). Importantly, in this study permutations of predicted gene activity across the 26 haplotypes in the NAM population were predictive of phenotypes, as long as haplotypes were still correctly assigned to individual NAM lines. This result highlights the potential source of confounding due to haplotype variation alone, and the necessity to account for haplotype IBD in association studies of gene activity imputed by haplotype.

Although our PWP results suggest that predicted 3D structure provides useful functional information about genetic effects, the gains in prediction accuracy realized in our study are certainly too low to justify routine applications of AlphaFold2 in genomic selection programs. However, the PHG used in this study may still be used for imputing haplotypes in any breeding population, and our structure-based similarity matrices may then be used without running AlphaFold2 again. Further research is needed to determine whether haplotype imputation by this PHG database (or future databases containing more genome assemblies) is accurate enough for effective incorporation of predicted protein structure information into genomic prediction models.

Methods:

Protein sequence diversity investigation

To investigate sequence diversity, we extracted the pan-genes from previously published pan-gene annotations (Hufford et al. 2021) with at least one predicted protein structure for each of the 26 founder lines (see below), resulting in 17,633 pan-genes. MSA was performed using MAFFT to investigate the sequence diversity (Katoh and Standley 2013) for each pan-gene group. Protein sequences from the 26 NAM founder lines were iteratively used as the reference. The query sequences were compared with the reference sequence for each pan-gene of each iteration to count the number of SAPs and indels.

Protein 3D structure prediction

Protein structure prediction was performed using AlphaFold v2.0.0 with the preset reduced_dbs parameter as described in the AlphaFold database publication (Tunyasuvunakool et al. 2021). The genome annotations and pan-gene results were generated by Hufford et al. (Hufford et al. 2021). All B73 v5.0 protein sequences and canonical protein sequences from the other NAM founder lines were selected as input. Sequences with residue codes '*' (premature stop codon) or 'X' (unknown residue) were also excluded. The 3D structures of all B73 protein sequences were predicted, and canonical protein sequences from the other NAM founder lines were extracted for 3D structure prediction. All the kept protein sequences were merged to check for duplicated protein sequences. By default, AlphaFold v2.0.0 outputs five structural models for each input sequence; the model with the highest pLDDT was selected for subsequent analyses. Due to high computational cost, we only folded proteins with less than 1200 amino acids for the 25 NAM founder lines, except for B73. Due to high GPU memory requirements, some proteins failed to fold. Protein structures were defined as resolved if they were successfully predicted by

AlphaFold2 and were otherwise considered to be unresolved. Protein folding was conducted using high-performance computing (HPC) computational nodes with 4 NVIDIA Quadro RTX 5000 GPU cards.

Calculation of RSA

The accessible surface area (ASA) for each amino acid residue in a protein structure was calculated using DSSP v2.3.0 (Touw et al. 2015). The maximum possible solvent accessible surface area (MaxASA) was obtained from a previous publication (Tien et al. 2013). The RSA of each amino acid residue was calculated using the formula RSA = ASA/MaxASA.

SNP annotation

The variant calling file for 282 association panels was downloaded from maize HapMap 3.2.1 (Bukowski et al. 2018) and uplifted to B73 V5.0 reference coordinates via CrossMap (Zhao et al. 2014). Allele frequencies were counted using VCFtools (Danecek et al. 2011). Based on genome annotation information in general feature format (GFF3), SNPs in protein-coding regions were extracted and classified as either synonymous SNPs or nonsynonymous SNPs. The SNP dataset was annotated by GERP score from a previous publication (Kistler et al. 2018) uplifted to B73 V5.0 coordinates.

Calculation of nucleotide diversity (π)

The bi-allelic SNP file for each line in the 282 association panel was converted to a genomic sequence file using B73 V5.0 reference coordinates as the reference. Protein-coding sequences were extracted using GffRead (Pertea and Pertea 2020), and the same transcripts from all lines were merged into multiple sequence alignment files. Each codon was classified based on pLDDT bin and RSA bins. RSA value was divided into 10 bins ranging from 0 to 1 at intervals of 0.1. πN and πS (π of nonsynonymous and synonymous mutations, respectively) were

457 calculated by bppsuite (Guéguen et al. 2013), with multiple sequence alignment files used as
458 input. The average πN and πS values for all the codons in each bin were used to perform
459 statistics analysis and generate the plot (Supplemental Fig. S5).

460 **Estimation of derived allele frequency**

461 To infer the derived allele for each SNP in the maize population, the genomes of *Zea mays ssp.*
462 *mexicana* (TIL18) and *Zea mays ssp. parviglumis* (TIL11) were used as outgroups. The
463 sequences of two genomes are available at MaizeGDB (https://download.maizegdb.org/). The
464 genomes of the two outgroups were aligned against the B73 v5.0 reference genome (Hufford et
465 al. 2021) using AnchorWave (Song et al. 2022). The genome alignments were converted into a
466 GVCF format file using the MAFToGVCF plugin of TASSEL (Peter J. Bradbury et al. 2007). For
467 each SNP in the maize population, if the reference allele matched either the TIL18 or TIL11
468 allele, the reference allele was categorized as an ancestral allele. Otherwise, if the alternative
469 allele was identical to the TIL18 or TIL11 allele, the reference allele was defined as a derived
470 allele. For a SNP, whose reference position was not aligned to either of those two outgroups
471 and was therefore likely located in newly derived sequence fragments, the SNP was excluded
472 from the analysis.

473 **Variant calling for the NAM population and kinship matrix construction**
474 SNP variant calling of the NAM founder lines was conducted using PHG v1 (P. J. Bradbury et al.
475 2022). A PHG database was constructed from the genomes of 26 NAM founder lines (Hufford et
476 al. 2021), using the B73 v5.0 assembly as the reference. GBS reads from 4,736 accessions
477 were mapped to the pangenome, and SNP variants were imputed in the NAM RIL population.
478 The PLINK (Purcell et al. 2007) program was used to retain only biallelic variants with the
479 parameter '--max-alleles 2', resulting in 42,329,376 SNPs for 4,736 inbred lines. The SNP-

based kinship matrix **G** was computed using the Balding-Nichols (BN) methods in the EMMAX software (Kang et al. 2010).

Protein similarity matrix construction

The NAM population was used for PWAS and PWP. To predict protein structures at population scale, all metrics for NAM founder lines were projected onto the NAM population with 25 families using the PHG (Supplemental Fig. S18, Bradbury et al. 2022), assuming no recombination crossover in the middle of coding genes. To measure the structural similarity within each core pan-gene group *j*, structure alignments were performed using US-align, and the structural similarity matrices were computed as TM scores with values in [0,1], using structure alignment results as input (C. Zhang et al. 2022). To construct sequence similarity matrices, protein sequences were aligned via one of three MSA approaches, i.e., MAFFT (Katoh and Standley 2013), MUSCLE (Edgar 2004), or T-Coffee (Notredame, Higgins, and Heringa 2000). The similarity scores for each pair of sequences were calculated as numerical values in [0,1] as well. Each pair of amino acids was treated as similar to each other if they had a positive score from the BLOSUM62 matrix; otherwise, they were considered to be dissimilar to each other. For each pair of protein sequences, the similarity score was calculated as the number of similar amino acids divided by the average length of these two protein sequences. The IBD matrix for each pan-gene group was constructed using an identity matrix to NAM founder haplotypes.

For each type of similarity (structure-, sequence-, or IBD-based) and pan-gene group *j*, $\mathbf{K}_j$ was the 26×26 relationship matrix among the 26 founder haplotypes computed from similarity matrix $\mathbf{S}_j$ (https://github.com/shuaiwang2/Protein3D_QG/blob/main/PWAS/PWAS-

data_processing.R). First, each relationship matrix was obtained by centering the corresponding similarity matrix: $\mathbf{K}_j = \mathbf{H}\mathbf{S}_j\mathbf{H}'$, where $\mathbf{H}$ is the centering matrix $\mathbf{H} = \mathbf{I}_{26} - \mathbf{J}_{26}/26$. The pan-gene groups were then filtered, according to the following criteria: (*i*) no missing haplotypes among the 26 NAM founders; (*ii*) successful protein structure prediction for all observed haplotypes; (*iii*) trace of similarity matrix (sum of diagonal elements after centering) greater than 1/25. This filtering step resulted in *m*=11,927 retained pan-gene groups across similarity metrics (US-align, MAFFT, MUSCLE, T-Coffee, IBD). The distribution of pan-gene is almost even on the chromosomes, except for a missing portion of Chromosome 10 (Supplemental Fig. S19).

The input to PWAS consisted of one or more principal components per retained pan-gene group, obtained by eigen decomposition of $\mathbf{K}_j$ for each pan-gene group *j*. To retain eigenvalues that contribute substantially to the overall variance of the matrix, for each pan-gene group *j*, principal components were included in PWAS if their variance explained in $\mathbf{K}_j$ was greater than 1×10^{-5}. If there were no principal components kept, the pan-gene group would not be used for the following analyses. The average number of structured-based principal components included (*k*) was 7, with a range from 1 to 25. The average number of sequence-based principal components included was 6, with a range from 1 to 25. For each pan-gene group *j* and relationship matrix $\mathbf{K}_j$, the matrix of principal components used in PWAS was $\mathbf{X}_j = \mathbf{Z}_j\mathbf{U}_j\mathbf{\Lambda}_j^{1/2}$, where $\mathbf{Z}_j$ is the $n \times 26$ design matrix assigning each individual to one of the 26 founder haplotypes at the pan-gene group *j*, $\mathbf{U}_j$ is the 26 x *k* matrix of retained eigenvectors of $\mathbf{K}_j$ and $\mathbf{\Lambda}_j$ is the *k* x *k* diagonal matrix of the corresponding eigenvalues.

The input to PWP consisted of the sum of sequence-based, IBD-based, or structure-based similarity matrices across all pan-gene groups. To ensure mathematical validity, similarity matrices were adjusted to positive definite matrices using the nearPD function from the Matrix package in R (Higham 2002; R Core Team 2025). To normalize proteome-wide relationship matrices, the mean of their diagonal elements was set to 1, as follows:

$$\mathbf{P}_{sum} = \sum_{j=1}^{m} \mathbf{Z}_j \mathbf{K}_j \mathbf{Z}_j'$$

$$\mathbf{P} = \frac{\mathbf{P}_{sum}}{trace\left(\mathbf{P}_{sum}\right)/n}$$

where trace is the sum of diagonal elements of a matrix.

PWAS

PWAS was performed using the following linear mixed model for each pan-gene group *j*:

$$\mathbf{y} = \mathbf{Q}\alpha + \mathbf{X}_j\beta_j + \mathbf{g} + \mathbf{h} + \mathbf{e}$$

where **y** is an *n* × 1 vector of phenotypes; $\mathbf{Q}$ is an n × 25 design matrix assigning each individual to its NAM family, and α is a 25 × 1 vector of fixed NAM family means; $\mathbf{X}_j$ is an n × *k* matrix of principal components for pan-gene group *j*, and β_j is the *k* x 1 vector of their fixed effects; $\mathbf{g}$ are random genomic values, as captured by the SNP-based kinship matrix; $\mathbf{h}$ are random proteome-wide effects, as captured by pan-gene IBD-based similarities; and $\mathbf{e}$ are random residual effects, *n* is the number of individuals, and *k* is the number of principal components. The model accounts for the population structure and NAM family relationship through α, $\mathbf{g}$ and $\mathbf{h}$. PWAS was conducted using a two-stage approach (Z. Zhang et al. 2010). First, variance parameters were estimated in the null model, excluding $\mathbf{X}_j\beta_j$, using qgg (Rohde, Fourie Sørensen, and Sørensen 2020). Then, the effects of principal components $\mathbf{X}_j\beta_j$ were included for each pan-

gene group *j*, and the significance of β_j estimates was assessed by a joint Wald test (H_0: $\beta_j = 0$). A pan-gene group was deemed significant if the *p*-value from its joint Wald test was less than 4.7×10^{-6}, according to a Bonferroni multiple-testing correction. The candidate proteins of significant pan-gene groups were reported for each trait (Supplemental Table S4).

A total of 11,927 pan-genes were investigated in PWAS separately. A total of 32 traits were analyzed in PWAS. The phenotypes used were selected from previous publications (Buckler et al. 2009; Peiffer et al. 2013; Hung et al. 2012; Wallace et al. 2014; Poland et al. 2011; Brown et al. 2011; Bian and Holland 2017; Kump et al. 2011; Cook et al. 2012; Olukolu et al. 2014; Benson et al. 2015; Diepenbrock et al. 2017; Tian et al. 2011; Foerster et al. 2015), of which the most highly correlated traits were excluded (Supplemental Table S6).

SNP-based GWAS

GWAS was implemented by GEMMA (Zhou and Stephens 2012) using default parameters. A total of 19,458,794 SNPs were used for SNP-based GWAS. SNPs with minor allele frequency more than 0.05 and missing rates less than 20% were retained using PLINK (Purcell et al. 2007). We calculated the GWAS significant threshold using Bonferroni multiple-testing correction based on the number of effective SNPs. The effective SNP count was derived with an R^2 threshold of 0.99, a window size of 100,000, and a step size of 50,000, resulting into 1,090,989 SNPs. The GWAS used the same phenotypic traits as those used in the PWAS.

Variance partitioning

To assess the significance of structure-based PWP, variance partitioning was performed using the MMEst function implemented by the MM4LMM package (Laporte, Charcosset, and Mary-Huard 2022). Proteome-wide relationship matrices were used as input for variance partitioning.

The 'baseline' prediction model only included a SNP-based kinship matrix **G**. The 'sequence' model included a **G** matrix and a proteome-wide relationship matrix **P** based on protein sequences. The 'IBD' model included a **G** matrix and a **P** matrix based on haplotype IBD. The 'structure' model included a **G** matrix and a **P** matrix based on predicted structures. The 'sequence+structure' model included a **G** matrix and two **P** matrices based on sequences and predicted structures. The 'IBD+structure' model included a **G** matrix and two **P** matrices based on haplotype IBD and predicted structures

To test the statistical significance of additional random effects (e.g., structure-based genomic values), a likelihood ratio test was conducted to compare different genomic models using the following formula:$\lambda = -2 \times (logL_0 - logL_1)$

where L_0 and L_1 are the likelihoods of the null and alternative genomic models, respectively. The resulting test statistic (λ) was compared to a chi-squared distribution with corresponding degrees of freedom to obtain a *p*-value (Supplemental Table S5).

PWP

To estimate the improvement of prediction accuracy using structure information, proteome-wide prediction was carried out using the greml function implemented in the qgg package (Rohde, Fourie Sørensen, and Sørensen 2020). The model applied was as follows:

$$\mathbf{y} = \mu + \mathbf{g} + \mathbf{p} + \mathbf{e}$$

where $\mathbf{y}$ is the vector of phenotypes; μ is the intercept (grand mean) as a fixed effect; $\mathbf{g}$ are random genomic values, captured by genome-wide SNP-based relationships. $\mathbf{p}$ are random

proteome-wide effects, captured by one or more proteome-wide relationship matrices; $\mathbf{e}$ is a vector of residuals.

In models with only one proteome-wide relationship matrix, $\mathbf{p}$ is a vector of "proteomic" values, such that $\mathbf{p} \sim \mathcal{N}(\mathbf{0}, \mathbf{P}\sigma_p^2)$, where $\mathbf{P}$ is based on either structure, sequence, or IBD. In models with two proteome-wide relationship matrices, $\mathbf{p} = \mathbf{p}_1 + \mathbf{p}_2$, where $\mathbf{p}_1 \sim \mathcal{N}(\mathbf{0}, \mathbf{P}_1\sigma_{p_1}^2)$, $\mathbf{p}_2 \sim \mathcal{N}(\mathbf{0}, \mathbf{P}_2\sigma_{p_2}^2)$, with $\mathbf{P}_1$ and $\mathbf{P}_2$ being two different proteome-wide relationship matrices. Considering the population properties of the NAM population, PWP models were validated by leave-one-family-out cross-validation. Each NAM family was left out as a validation dataset to compute the prediction accuracy separately, while the other 24 NAM families were used for training. Prediction accuracy was computed as $Cor(\mathbf{y}, \hat{\mathbf{y}})$, where $\hat{\mathbf{y}}$ is the vector of predicted genotypic values: $\hat{\mathbf{y}} = \hat{\mathbf{g}} + \hat{\mathbf{p}}$. This process was repeated for 25 iterations.

Statistical analysis

All statistical analyses used for this paper were performed in R (R Core Team 2025) and are indicated in the main text and figure legends.

Data access

All predicted protein 3D structures are available in Figshare (https://doi.org/10.25452/figshare.plus.29176679).Genotype data for the NAM association panel and protein IDs projected to NAM population are available in Figshare (https://doi.org/10.6084/m9.figshare.28349087). Custom scripts and codes used in this study are available as Supplemental Code and in GitHub (https://github.com/shuaiwang2/Protein3D_QG).

Competing interest statement

The authors declare no competing interests.

Acknowledgments

This project was supported by the United States Department of Agricultural Research Service, NSF No. 1822330, the Shandong Provincial Natural Science Fund for Excellent Young Scientists Fund Program (Overseas) (2023HWYQ-109), the USDA NIFA AFRI predoctoral fellowship (MK: 2022-67011-36458) and project SYS202206 supported by Shandong Provincial Natural Science Foundation. We thank the members of the E.S.B. laboratory (Cornell University, U.S.) for helpful discussions. We thank Zhiwu Zhang (Washington State University, U.S.) and Jiabo Wang (Southwest Minzu University, China) for helpful discussions of quantitative genetics analysis.

Author contributions

E.B., B.S., G.P.R. and M.R. conceived the study. S.W., B.S. and G.P.R. wrote the manuscript. B.S., L.J., Z.M. and P.J. performed variant calling using *de novo* genome assemblies. B.S. and W. A. performed protein structures prediction. M.K. curated the phenotypes. G.P.R., D.S. and S.W. conducted the PWAS and genome prediction analysis. All authors revised and reviewed the manuscript.

Reference

Akdel, Mehmet, Douglas E. V. Pires, Eduard Porta Pardo, Jürgen Jänes, Arthur O. Zalevsky, Bálint Mészáros, Patrick Bryant, et al. 2022. “A Structural Biology Community Assessment of AlphaFold2 Applications.” *Nature Structural & Molecular Biology* 29 (11): 1056–67.

Alderson, T. Reid, Iva Pritišanac, Đesika Kolarić, Alan M. Moses, and Julie D. Forman-Kay. 2023. “Systematic Identification of Conditionally Folded Intrinsically Disordered Regions by AlphaFold2.” *Proceedings of the National Academy of Sciences of the United States of America* 120 (44): e2304302120.

Atwell, Susanna, Yu S. Huang, Bjarni J. Vilhjálmsson, Glenda Willems, Matthew Horton, Yan Li,

652 Dazhe Meng, et al. 2010. “Genome-Wide Association Study of 107 Phenotypes in
653 Arabidopsis Thaliana Inbred Lines.” *Nature* 465 (7298): 627–31.
654 Benson, Jacqueline M., Jesse A. Poland, Brent M. Benson, Erik L. Stromberg, and Rebecca J.
655 Nelson. 2015. “Resistance to Gray Leaf Spot of Maize: Genetic Architecture and
656 Mechanisms Elucidated through Nested Association Mapping and near-Isogenic Line
657 Analysis.” *PLoS Genetics* 11 (3): e1005045.
658 Bertolini, Edoardo, Brian R. Rice, Max Braud, Jiani Yang, Sarah Hake, Josh Strable, Alexander
659 E. Lipka, and Andrea L. Eveland. 2025. “Regulatory Variation Controlling Architectural
660 Pleiotropy in Maize.” *Nature Communications* 16 (1): 2140.
661 Bian, Y., and J. B. Holland. 2017. “Enhancing Genomic Prediction with Genome-Wide
662 Association Studies in Multiparental Maize Populations.” *Heredity* 118 (6): 585–93.
663 Bradbury, Peter J., Zhiwu Zhang, Dallas E. Kroon, Terry M. Casstevens, Yogesh Ramdoss, and
664 Edward S. Buckler. 2007. “TASSEL: Software for Association Mapping of Complex Traits in
665 Diverse Samples.” *Bioinformatics* 23 (19): 2633–35.
666 Bradbury, P. J., T. Casstevens, S. E. Jensen, L. C. Johnson, Z. R. Miller, B. Monier, M. C.
667 Romay, B. Song, and E. S. Buckler. 2022. “The Practical Haplotype Graph, a Platform for
668 Storing and Using Pangenomes for Imputation.” *Bioinformatics* 38 (15): 3698–3702.
669 Brown, Patrick J., Narasimham Upadyayula, Gregory S. Mahone, Feng Tian, Peter J. Bradbury,
670 Sean Myles, James B. Holland, et al. 2011. “Distinct Genetic Architectures for Male and
671 Female Inflorescence Traits of Maize.” *PLoS Genetics* 7 (11): e1002383.
672 Buckler, Edward S., Brandon S. Gaut, and Michael D. McMullen. 2006. “Molecular and
673 Functional Diversity of Maize.” *Current Opinion in Plant Biology* 9 (2): 172–76.
674 Buckler, Edward S., James B. Holland, Peter J. Bradbury, Charlotte B. Acharya, Patrick J.
675 Brown, Chris Browne, Elhan Ersoz, et al. 2009. “The Genetic Architecture of Maize
676 Flowering Time.” *Science* 325 (5941): 714–18.
677 Buel, Gwen R., and Kylie J. Walters. 2022. “Can AlphaFold2 Predict the Impact of Missense
678 Mutations on Structure?” *Nature Structural & Molecular Biology* 29 (1): 1–2.
679 Bukowski, Robert, Xiaosen Guo, Yanli Lu, Cheng Zou, Bing He, Zhengqin Rong, Bo Wang, et
680 al. 2018. “Construction of the Third-Generation Zea Mays Haplotype Map.” *GigaScience* 7
681 (4): 1–12.
682 Chakravarty, Devlina, and Lauren L. Porter. 2022. “AlphaFold2 Fails to Predict Protein Fold
683 Switching.” *Protein Science: A Publication of the Protein Society* 31 (6): e4353.
684 Chen, Weiwei, Xuhui Li, Xiangbo Zhang, Zaid Chachar, Chuanli Lu, Yongwen Qi, Hailong
685 Chang, and Qinnan Wang. 2024. “Genome-Wide Association Study of Trace Elements in
686 Maize Kernels.” *BMC Plant Biology* 24 (1): 724.
687 Chen, Wenkang, Lu Chen, Xuan Zhang, Ning Yang, Jianghua Guo, Min Wang, Shenghui Ji, et
688 al. 2022. “Convergent Selection of a WD40 Protein That Enhances Grain Yield in Maize
689 and Rice.” *Science* 375 (6587): eabg7985.
690 Chothia, C., and A. M. Lesk. 1986. “The Relation between the Divergence of Sequence and
691 Structure in Proteins.” *The EMBO Journal* 5 (4): 823–26.
692 Cook, Jason P., Michael D. McMullen, James B. Holland, Feng Tian, Peter Bradbury, Jeffrey
693 Ross-Ibarra, Edward S. Buckler, and Sherry A. Flint-Garcia. 2012. “Genetic Architecture of
694 Maize Kernel Composition in the Nested Association Mapping and Inbred Association
695 Panels.” *Plant Physiology* 158 (2): 824–34.
696 Danecek, Petr, Adam Auton, Goncalo Abecasis, Cornelis A. Albers, Eric Banks, Mark A.
697 DePristo, Robert E. Handsaker, et al. 2011. “The Variant Call Format and VCFtools.”
698 *Bioinformatics* 27 (15): 2156–58.
699 Davydov, Eugene V., David L. Goode, Marina Sirota, Gregory M. Cooper, Arend Sidow, and
700 Serafim Batzoglou. 2010. “Identifying a High Fraction of the Human Genome to Be under
701 Selective Constraint Using GERP++.” *PLoS Computational Biology* 6 (12): e1001025.
702 Desta, Zeratsion Abera, and Rodomiro Ortiz. 2014. “Genomic Selection: Genome-Wide

703 Prediction in Plant Improvement." *Trends in Plant Science* 19 (9): 592–601.
704 Diepenbrock, Christine H., Catherine B. Kandianis, Alexander E. Lipka, Maria Magallanes-
705 Lundback, Brieanne Vaillancourt, Elsa Góngora-Castillo, Jason G. Wallace, et al. 2017.
706 "Novel Loci Underlie Natural Variation in Vitamin E Levels in Maize Grain." *The Plant Cell*
707 29 (10): 2374–92.
708 Doerge, Rebecca W. 2002. "Mapping and Analysis of Quantitative Trait Loci in Experimental
709 Populations." *Nature Reviews. Genetics* 3 (1): 43–52.
710 Edgar, Robert C. 2004. "MUSCLE: Multiple Sequence Alignment with High Accuracy and High
711 Throughput." *Nucleic Acids Research* 32 (5): 1792–97.
712 Erdin, Serkan, Andreas Martin Lisewski, and Olivier Lichtarge. 2011. "Protein Function
713 Prediction: Towards Integration of Similarity Metrics." *Current Opinion in Structural Biology*
714 21 (2): 180–88.
715 Flint-Garcia, Sherry A., Anne-Céline Thuillet, Jianming Yu, Gael Pressoir, Susan M. Romero,
716 Sharon E. Mitchell, John Doebley, Stephen Kresovich, Major M. Goodman, and Edward S.
717 Buckler. 2005. "Maize Association Population: A High-Resolution Platform for Quantitative
718 Trait Locus Dissection: High-Resolution Maize Association Population." *The Plant Journal*
719 44 (6): 1054–64.
720 Foerster, Jillian M., Timothy Beissinger, Natalia de Leon, and Shawn Kaeppler. 2015. "Large
721 Effect QTL Explain Natural Phenotypic Variation for the Developmental Timing of
722 Vegetative Phase Change in Maize (Zea Mays L.)." *Theoretical and Applied Genetics* 128
723 (3): 529–38.
724 Foster, T., J. Yamaguchi, B. C. Wong, B. Veit, and S. Hake. 1999. "Gnarley1 Is a Dominant
725 Mutation in the knox4 Homeobox Gene Affecting Cell Shape and Identity." *The Plant Cell*
726 11 (7): 1239–52.
727 Gage, Joseph L., Brandon Monier, Anju Giri, and Edward S. Buckler. 2020. "Ten Years of the
728 Maize Nested Association Mapping Population: Impact, Limitations, and Future Directions."
729 *The Plant Cell* 32 (7): 2083–93.
730 Gerasimavicius, Lukas, Sarah A. Teichmann, and Joseph A. Marsh. 2025. "Leveraging Protein
731 Structural Information to Improve Variant Effect Prediction." *Current Opinion in Structural*
732 *Biology* 92 (103023): 103023.
733 Giri, Anju, Merritt Khaipho-Burch, Edward S. Buckler, and Guillaume P. Ramstein. 2021.
734 "Haplotype Associated RNA Expression (HARE) Improves Prediction of Complex Traits in
735 Maize." *PLoS Genetics* 17 (10): e1009568.
736 Guéguen, Laurent, Sylvain Gaillard, Bastien Boussau, Manolo Gouy, Mathieu Groussin, Nicolas
737 C. Rochette, Thomas Bigot, et al. 2013. "Bio++: Efficient Extensible Libraries and Tools for
738 Computational Molecular Evolution." *Molecular Biology and Evolution* 30 (8): 1745–50.
739 Hicks, Michael, Istvan Bartha, Julia di Iulio, J. Craig Venter, and Amalio Telenti. 2019.
740 "Functional Characterization of 3D Protein Structures Informed by Human Genetic
741 Diversity." *Proceedings of the National Academy of Sciences of the United States of*
742 *America* 116 (18): 8960–65.
743 Higham, N. J. 2002. "Computing the Nearest Correlation Matrix--a Problem from Finance." *IMA*
744 *Journal of Numerical Analysis* 22 (3): 329–43.
745 Hufford, Matthew B., Arun S. Seetharam, Margaret R. Woodhouse, Kapeel M. Chougule,
746 Shujun Ou, Jianing Liu, William A. Ricci, et al. 2021. "De Novo Assembly, Annotation, and
747 Comparative Analysis of 26 Diverse Maize Genomes." *Science* 373 (6555): 655–62.
748 Hung, Hsiao-Yi, Laura M. Shannon, Feng Tian, Peter J. Bradbury, Charles Chen, Sherry A.
749 Flint-Garcia, Michael D. McMullen, et al. 2012. "ZmCCT and the Genetic Basis of Day-
750 Length Adaptation Underlying the Postdomestication Spread of Maize." *Proceedings of the*
751 *National Academy of Sciences of the United States of America* 109 (28): E1913–21.
752 Jiang, Fukun, Mei Guo, Fang Yang, Keith Duncan, David Jackson, Antoni Rafalski, Shoucai
753 Wang, and Bailin Li. 2012. "Mutations in an AP2 Transcription Factor-like Gene Affect

754 Internode Length and Leaf Shape in Maize." *PloS One* 7 (5): e37040.
755 Jumper, John, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf
756 Ronneberger, Kathryn Tunyasuvunakool, et al. 2021. "Highly Accurate Protein Structure
757 Prediction with AlphaFold." *Nature* 596 (7873): 583–89.
758 Kang, Hyun Min, Jae Hoon Sul, Susan K. Service, Noah A. Zaitlen, Sit-Yee Kong, Nelson B.
759 Freimer, Chiara Sabatti, and Eleazar Eskin. 2010. "Variance Component Model to Account
760 for Sample Structure in Genome-Wide Association Studies." *Nature Genetics* 42 (4): 348–
761 54.
762 Katoh, Kazutaka, and Daron M. Standley. 2013. "MAFFT Multiple Sequence Alignment
763 Software Version 7: Improvements in Performance and Usability." *Molecular Biology and*
764 *Evolution* 30 (4): 772–80.
765 Keys, Kevin L., Angel C. Y. Mak, Marquitta J. White, Walter L. Eckalbar, Andrew W. Dahl, Joel
766 Mefford, Anna V. Mikhaylova, et al. 2020. "On the Cross-Population Generalizability of
767 Gene Expression Prediction Models." *PLoS Genetics* 16 (8): e1008927.
768 Khaipho-Burch, Merritt, Taylor Ferebee, Anju Giri, Guillaume Ramstein, Brandon Monier, Emily
769 Yi, M. Cinta Romay, and Edward S. Buckler. 2023. "Elucidating the Patterns of Pleiotropy
770 and Its Biological Relevance in Maize." *PLoS Genetics* 19 (3): e1010664.
771 Kistler, Logan, S. Yoshi Maezumi, Jonas Gregorio de Souza, Natalia A. S. Przelomska,
772 Flaviane Malaquias Costa, Oliver Smith, Hope Loiselle, et al. 2018. "Multiproxy Evidence
773 Highlights a Complex Evolutionary Legacy of Maize in South America." *Science* 362 (6420):
774 1309–13.
775 Klein, Robert J., Xing Xu, Semanti Mukherjee, Jason Willis, and James Hayes. 2010.
776 "Successes of Genome-Wide Association Studies." *Cell* 142 (3): 350–51.
777 Krissinel, Evgeny. 2007. "On the Relationship between Sequence and Structure Similarities in
778 Proteomics." *Bioinformatics* 23 (6): 717–23.
779 Kump, Kristen L., Peter J. Bradbury, Randall J. Wisser, Edward S. Buckler, Araby R. Belcher,
780 Marco A. Oropeza-Rosas, John C. Zwonitzer, et al. 2011. "Genome-Wide Association
781 Study of Quantitative Resistance to Southern Leaf Blight in the Maize Nested Association
782 Mapping Population." *Nature Genetics* 43 (2): 163–68.
783 Laporte, Fabien, Alain Charcosset, and Tristan Mary-Huard. 2022. "Efficient ReML Inference in
784 Variance Component Mixed Models Using a Min-Max Algorithm." *PLoS Computational*
785 *Biology* 18 (1): e1009659.
786 Li, Chuxi, Changlin Liu, Xiantao Qi, Yongchun Wu, Xiaohong Fei, Long Mao, Beijiu Cheng,
787 Xinhai Li, and Chuanxiao Xie. 2017. "RNA-Guided Cas9 as an in Vivo Desired-Target
788 Mutator in Maize." *Plant Biotechnology Journal* 15 (12): 1566–76.
789 MacGowan, Stuart A., Fábio Madeira, Thiago Britto-Borges, and Geoffrey J. Barton. 2024. "A
790 Unified Analysis of Evolutionary and Population Constraint in Protein Domains Highlights
791 Structural Features and Pathogenic Sites." *Communications Biology* 7 (1): 447.
792 Mantilla-Perez, Maria Betsabe, and Maria G. Salas Fernandez. 2017. "Differential Manipulation
793 of Leaf Angle throughout the Canopy: Current Status and Prospects." *Journal of*
794 *Experimental Botany* 68 (21-22): 5699–5717.
795 McBride, John M., Konstantin Polev, Amirbek Abdirasulov, Vladimir Reinharz, Bartosz A.
796 Grzybowski, and Tsvi Tlusty. 2023. "AlphaFold2 Can Predict Single-Mutation Effects."
797 *Physical Review Letters* 131 (21): 218401.
798 McMullen, Michael D., Stephen Kresovich, Hector Sanchez Villeda, Peter Bradbury, Huihui Li,
799 Qi Sun, Sherry Flint-Garcia, et al. 2009. "Genetic Properties of the Maize Nested
800 Association Mapping Population." *Science* 325 (5941): 737–40.
801 Mergner, Julia, Martin Frejno, Markus List, Michael Papacek, Xia Chen, Ajeet Chaudhary,
802 Patroklos Samaras, et al. 2020. "Mass-Spectrometry-Based Draft of the Arabidopsis
803 Proteome." *Nature* 579 (7799): 409–14.
804 Notredame, C., D. G. Higgins, and J. Heringa. 2000. "T-Coffee: A Novel Method for Fast and

805 Accurate Multiple Sequence Alignment." *Journal of Molecular Biology* 302 (1): 205–17.
806 Olukolu, Bode A., Guan-Feng Wang, Vijay Vontimitta, Bala P. Venkata, Sandeep Marla, Jiabing
807 Ji, Emma Gachomo, et al. 2014. "A Genome-Wide Association Study of the Maize
808 Hypersensitive Defense Response Identifies Genes That Cluster in Related Pathways."
809 *PLoS Genetics* 10 (8): e1004562.
810 Peiffer, Jason A., Aymé Spor, Omry Koren, Zhao Jin, Susannah Green Tringe, Jeffery L. Dangl,
811 Edward S. Buckler, and Ruth E. Ley. 2013. "Diversity and Heritability of the Maize
812 Rhizosphere Microbiome under Field Conditions." *Proceedings of the National Academy of*
813 *Sciences of the United States of America* 110 (16): 6548–53.
814 Pertea, Geo, and Mihaela Pertea. 2020. "GFF Utilities: GffRead and GffCompare."
815 *F1000Research* 9 (304): 304.
816 Poland, Jesse A., Peter J. Bradbury, Edward S. Buckler, and Rebecca J. Nelson. 2011.
817 "Genome-Wide Nested Association Mapping of Quantitative Resistance to Northern Leaf
818 Blight in Maize." *Proceedings of the National Academy of Sciences of the United States of*
819 *America* 108 (17): 6893–98.
820 Purcell, Shaun, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A. R. Ferreira, David
821 Bender, Julian Maller, et al. 2007. "PLINK: A Tool Set for Whole-Genome Association and
822 Population-Based Linkage Analyses." *American Journal of Human Genetics* 81 (3): 559–
823 75.
824 Ramsey, Duncan C., Michael P. Scherrer, Tong Zhou, and Claus O. Wilke. 2011. "The
825 Relationship between Relative Solvent Accessibility and Evolutionary Rate in Protein
826 Evolution." *Genetics* 188 (2): 479–88.
827 Ramstein, Guillaume P., and Edward S. Buckler. 2022. "Prediction of Evolutionary Constraint by
828 Genomic Annotations Improves Functional Prioritization of Genomic Variants in Maize."
829 *Genome Biology* 23 (1): 183.
830 R Core Team. 2025. "R: A Language and Environment for Statistical Computing." Vienna,
831 Austria: R Foundation for Statistical Computing. https://www.R-project.org.
832 Rohde, Palle Duun, Izel Fourie Sørensen, and Peter Sørensen. 2020. "Qgg: An R Package for
833 Large-Scale Quantitative Genetic Analyses." *Bioinformatics* 36 (8): 2614–15.
834 Salzberg, Steven L. 2019. "Next-Generation Genome Annotation: We Still Struggle to Get It
835 Right." *Genome Biology* 20 (1): 92.
836 Sato, Y., N. Sentoku, Y. Miura, H. Hirochika, H. Kitano, and M. Matsuoka. 1999. "Loss-of-
837 Function Mutations in the Rice Homeobox Gene OSH15 Affect the Architecture of
838 Internodes Resulting in Dwarf Plants." *The EMBO Journal* 18 (4): 992–1002.
839 Savojardo, Castrense, Matteo Manfredi, Pier Luigi Martelli, and Rita Casadio. 2020. "Solvent
840 Accessibility of Residues Undergoing Pathogenic Variations in Humans: From Protein
841 Structures to Protein Sequences." *Frontiers in Molecular Biosciences* 7:626363.
842 Schnable, Patrick S., Doreen Ware, Robert S. Fulton, Joshua C. Stein, Fusheng Wei, Shiran
843 Pasternak, Chengzhi Liang, et al. 2009. "The B73 Maize Genome: Complexity, Diversity,
844 and Dynamics." *Science* 326 (5956): 1112–15.
845 Song, Baoxing, Edward S. Buckler, and Michelle C. Stitzer. 2023. "New Whole-Genome
846 Alignment Tools Are Needed for Tapping into Plant Diversity." *Trends in Plant Science* 29
847 (3): 355–69.
848 Song, Baoxing, Santiago Marco-Sola, Miquel Moreto, Lynn Johnson, Edward S. Buckler, and
849 Michelle C. Stitzer. 2022. "AnchorWave: Sensitive Alignment of Genomes with High
850 Sequence Diversity, Extensive Structural Polymorphism, and Whole-Genome Duplication."
851 *Proceedings of the National Academy of Sciences of the United States of America* 119 (1):
852 e2113075119.
853 Song, Baoxing, Richard Mott, and Xiangchao Gan. 2018. "Recovery of Novel Association Loci in
854 Arabidopsis Thaliana and Drosophila Melanogaster through Leveraging INDELs
855 Association and Integrated Burden Test." *PLoS Genetics* 14 (10): e1007699.

Stein, Richard A., and Hassane S. Mchaourab. 2024. “Rosetta Energy Analysis of AlphaFold2 Models: Point Mutations and Conformational Ensembles.” *bioRxivorg*. https://www.biorxiv.org/content/10.1101/2023.09.05.556364v2.

Sun, Shichao, Baobao Wang, Changyu Li, Gen Xu, Jinliang Yang, Matthew B. Hufford, Jeffrey Ross-Ibarra, Haiyang Wang, and Li Wang. 2023. “Unraveling Prevalence and Effects of Deleterious Mutations in Maize Elite Lines across Decades of Modern Breeding.” *Molecular Biology and Evolution* 40 (8): msad170.

Tam, Vivian, Nikunj Patel, Michelle Turcotte, Yohan Bossé, Guillaume Paré, and David Meyre. 2019. “Benefits and Limitations of Genome-Wide Association Studies.” *Nature Reviews. Genetics* 20 (8): 467–84.

Tian, Feng, Peter J. Bradbury, Patrick J. Brown, Hsiaoyi Hung, Qi Sun, Sherry Flint-Garcia, Torbert R. Rocheford, Michael D. McMullen, James B. Holland, and Edward S. Buckler. 2011. “Genome-Wide Association Study of Leaf Architecture in the Maize Nested Association Mapping Population.” *Nature Genetics* 43 (2): 159–62.

Tien, Matthew Z., Austin G. Meyer, Dariya K. Sydykova, Stephanie J. Spielman, and Claus O. Wilke. 2013. “Maximum Allowed Solvent Accessibilites of Residues in Proteins.” *PloS One* 8 (11): e80635.

Togninalli, Matteo, Ümit Seren, Jan A. Freudenthal, J. Grey Monroe, Dazhe Meng, Magnus Nordborg, Detlef Weigel, Karsten Borgwardt, Arthur Korte, and Dominik G. Grimm. 2020. “AraPheno and the AraGWAS Catalog 2020: A Major Database Update Including RNA-Seq and Knockout Mutation Data for Arabidopsis Thaliana.” *Nucleic Acids Research* 48 (D1): D1063–68.

Touw, Wouter G., Coos Baakman, Jon Black, Tim A. H. te Beek, E. Krieger, Robbie P. Joosten, and Gert Vriend. 2015. “A Series of PDB-Related Databanks for Everyday Needs.” *Nucleic Acids Research* 43 (Database issue): D364–68.

Tunyasuvunakool, Kathryn, Jonas Adler, Zachary Wu, Tim Green, Michal Zielinski, Augustin Žídek, Alex Bridgland, et al. 2021. “Highly Accurate Protein Structure Prediction for the Human Proteome.” *Nature* 596 (7873): 590–96.

Vishnoi, Anchal, Praveen Sethupathy, Daniel Simola, Joshua B. Plotkin, and Sridhar Hannenhalli. 2011. “Genome-Wide Survey of Natural Selection on Functional, Structural, and Network Properties of Polymorphic Sites in Saccharomyces Paradoxus.” *Molecular Biology and Evolution* 28 (9): 2615–27.

Wallace, Jason G., Peter J. Bradbury, Nengyi Zhang, Yves Gibon, Mark Stitt, and Edward S. Buckler. 2014. “Association Mapping across Numerous Traits Reveals Patterns of Functional Variation in Maize.” *PLoS Genetics* 10 (12): e1004845.

Wang, Jian, Wu Yang, Shaohong Zhang, Haifei Hu, Yuxuan Yuan, Jingfang Dong, Luo Chen, et al. 2023. “A Pangenome Analysis Pipeline Provides Insights into Functional Gene Identification in Rice.” *Genome Biology* 24 (1): 19.

Werner, Jonathan D., Justin O. Borevitz, N. Henriette Uhlenhaut, Joseph R. Ecker, Joanne Chory, and Detlef Weigel. 2005. “FRIGIDA-Independent Variation in Flowering Time of Natural Arabidopsis Thaliana Accessions.” *Genetics* 170 (3): 1197–1207.

Wheelan, S. J., A. Marchler-Bauer, and S. H. Bryant. 2000. “Domain Size Distributions Can Predict Domain Boundaries.” *Bioinformatics* 16 (7): 613–18.

Windhausen, Vanessa S., Gary N. Atlin, John M. Hickey, Jose Crossa, Jean-Luc Jannink, Mark E. Sorrells, Babu Raman, et al. 2012. “Effectiveness of Genomic Prediction of Maize Hybrid Performance in Different Breeding Populations and Environments.” *G3* 2 (11): 1427–36.

Wingo, Aliza P., Yue Liu, Ekaterina S. Gerasimov, Jake Gockley, Benjamin A. Logsdon, Duc M. Duong, Eric B. Dammer, et al. 2021. “Integrating Human Brain Proteomes with Genome-Wide Association Data Implicates New Proteins in Alzheimer’s Disease Pathogenesis.” *Nature Genetics* 53 (2): 143–46.

Wolf, Maxim Y., Yuri I. Wolf, and Eugene V. Koonin. 2008. “Comparable Contributions of

907 Structural-Functional Constraints and Expression Level to the Rate of Protein Sequence
908 Evolution." *Biology Direct* 3 (1): 40.
909 Wu, Yaoyao, Dawei Li, Yong Hu, Hongbo Li, Guillaume P. Ramstein, Shaoqun Zhou, Xinyan
910 Zhang, et al. 2023. "Phylogenomic Discovery of Deleterious Mutations Facilitates Hybrid
911 Potato Breeding." *Cell* 186 (11): 2313–28.e15.
912 Xiao, Yingjie, Haijun Liu, Liuji Wu, Marilyn Warburton, and Jianbing Yan. 2017. "Genome-Wide
913 Association Studies in Maize: Praise and Stargaze." *Molecular Plant* 10 (3): 359–74.
914 Xie, Lingjuan, Xiaojiao Gong, Kun Yang, Yujie Huang, Shiyu Zhang, Leti Shen, Yanqing Sun, et
915 al. 2024. "Technology-Enabled Great Leap in Deciphering Plant Genomes." *Nature Plants*
916 10 (4): 551–66.
917 Yang, Guang, Yan Pan, Wenqiu Pan, Qingting Song, Ruoyu Zhang, Wei Tong, Licao Cui, et al.
918 2024. "Combined GWAS and eGWAS Reveals the Genetic Basis Underlying Drought
919 Tolerance in Emmer Wheat (*Triticum Turgidum* L.)." *The New Phytologist* 242 (5): 2115–31.
920 Yang, Jinliang, Sofiane Mezmouk, Andy Baumgarten, Edward S. Buckler, Katherine E. Guill,
921 Michael D. McMullen, Rita H. Mumm, and Jeffrey Ross-Ibarra. 2017. "Incomplete
922 Dominance of Deleterious Alleles Contributes Substantially to Trait Variation and Heterosis
923 in Maize." *PLoS Genetics* 13 (9): e1007019.
924 Zhang, Chengxin, Morgan Shine, Anna Marie Pyle, and Yang Zhang. 2022. "US-Align:
925 Universal Structure Alignments of Proteins, Nucleic Acids, and Macromolecular
926 Complexes." *Nature Methods* 19 (9): 1109–15.
927 Zhang, Zhiwu, Elhan Ersoz, Chao-Qiang Lai, Rory J. Todhunter, Hemant K. Tiwari, Michael A.
928 Gore, Peter J. Bradbury, et al. 2010. "Mixed Linear Model Approach Adapted for Genome-
929 Wide Association Studies." *Nature Genetics* 42 (4): 355–60.
930 Zhao, Hao, Zhifu Sun, Jing Wang, Haojie Huang, Jean-Pierre Kocher, and Liguo Wang. 2014.
931 "CrossMap: A Versatile Tool for Coordinate Conversion between Genome Assemblies."
932 *Bioinformatics* 30 (7): 1006–7.
933 Zhou, Xiang, and Matthew Stephens. 2012. "Genome-Wide Efficient Mixed-Model Analysis for
934 Association Studies." *Nature Genetics* 44 (7): 821–24.

935