bioRxiv preprint doi: https://doi.org/10.64898/2026.05.27.728270; this version posted May 29, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Graph-based pangenome provides insights into the adaptive evolution of *Cucurbita pepo*

Xuebo Zhao[1,9], Kailiang Bo[2,9], Honghe Sun[1,3], Jie Zhang[4], Arthur A. Schaffer[5], Adi Faigenboim[5], Harry S. Paris[6], Elad Oren[6], Amit Gur[6], Rebecca Grumet[7], Changlin Wang[2,*], Yong Xu[4,*], Zhangjun Fei[1,*]

[1]Boyce Thompson Institute, Cornell University, Ithaca, NY, USA
[2]Institute of Vegetables and Flowers, Chinese Academy of Agricultural Sciences, Beijing, China
[3]Plant Biology Section, School of Integrative Plant Science, Cornell University, Ithaca, NY, USA
[4]National Engineering Research Center for Vegetables, Beijing Academy of Agriculture and Forestry Science, Beijing, China
[5]Plant Sciences Institute, Agricultural Research Organization, The Volcani Center, Rishon LeZiyyon, Israel
[6]Plant Science Institute, Agricultural Research Organization, Newe Ya'ar Research Center, Ramat Yishay, Israel
[7]Department of Horticulture, Michigan State University, East Lansing, MI, USA
[9]These authors contributed equally.
*Email: wangchanglin@caas.cn; xuyong@nercv.org; zf25@cornell.edu

Abstract

Understanding how crops respond to environmental variation is crucial for biodiversity conservation and food security. *Cucurbita pepo* (pumpkin, squash, gourd) is one of the first domesticated crop species and exhibits remarkable phenotypic and ecological diversity, making it a powerful system for investigating the genomic basis of adaptation. Here, we constructed a graph-based *C. pepo* pangenome using nine chromosome-level assemblies and identified 229,431 high-confidence structural variants (SVs) that were genotyped across 206 wild and cultivated accessions. Our results demonstrate that *C. pepo* underwent parallel domestication, with two deeply diverged gene pools independently giving rise to the *pepo* and *ovifera* cultivated lineages, followed by expansion into diverse environments that produced strong signatures of differentiation largely mediated by young adaptive alleles. Single-nucleotide polymorphisms (SNPs) and SVs contribute complementary dimensions of environmental responsiveness. Biogeographical modeling predicts continued range contraction of wild relatives and elevated genetic offset in Eastern North American populations under projected future climates. Populations with higher genetic load harbor fewer adaptive variants and exhibit greater predicted maladaptation, indicating that deleterious mutations constrain adaptive potential. These findings highlight how genomic variants, adaptive diversity, and genetic load together shape environmental adaptation and inform conservation and crop improvement.

Introduction

Environmental variation shapes species distributions, influences biodiversity, and affects global food security[1]. As environmental conditions deviate from historical norms, both natural populations and agricultural species face increasing challenges in maintaining productivity, persistence, and local adaptation. These pressures are particularly pronounced for crops, where rising global demand coincides with diverse environmental constraints, underscoring the urgency of developing well-adapted varieties[2]. Understanding how plants respond to varying environments

bioRxiv preprint doi: https://doi.org/10.64898/2026.05.27.728270; this version posted May 29, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

and how evolutionary potential is distributed across cultivated crops and their wild relatives, is therefore critical for harnessing genetic resources to sustain agricultural productivity[3].

Population genomic studies have provided valuable insights into environment-associated genetic variation[4,5], yet most genomic analyses focus primarily on single-nucleotide polymorphisms (SNPs), which represent fine-scale variation but fail to capture large-effect structural variants (SVs) that can strongly influence gene function and phenotype[6]. SVs–including insertions, deletions, inversions, and translocations–constitute a major component of genomic diversity and play critical roles in phenotypic plasticity and environmental adaptation[7]. Recent advances in sequencing technologies and computational approaches have enabled large-scale SV identification, providing new opportunities to investigate their functional significance[8]. However, the contribution of SVs to local adaptation and their potential role in predicting the responses of species to future climates remain poorly understood. Moreover, adaptive potential is shaped not only by adaptive variants but also by genome-wide genetic load, raising the unresolved question of how deleterious mutations interact with adaptive diversity to constrain or facilitate environmental responses. Integrating SNPs and SVs within predictive genomic frameworks therefore provides a valuable opportunity to link genomic architecture, adaptive variation, and projected future environmental conditions[9].

Cucurbita pepo L. (*C. pepo*) is one of the first species to have been domesticated, with archaeological evidence suggestive of domestication dating back 10,000 years[10]. This species encompasses a wide range of cultivated, edible forms, including pumpkins and summer and winter squash, and is of considerable economic and nutritional importance[11,12]. Now cultivated globally across tropical and temperate regions, *C. pepo* exhibits remarkable morphological and genetic diversity[13]. Its long domestication history[14] and broad ecological distribution spanning diverse environments[15] make this crop species a powerful system for dissecting the genomic basis of environmental adaptation.

Here, we combined high-quality genome assemblies and population resequencing to construct a graph-based pangenome for *C. pepo*, enabling comprehensive identification and genotyping of SVs across diverse wild and cultivated accessions. By integrating genomic variants, including both SVs and SNPs, with environmental gradients and future climate projections, we investigate the genetic basis of local adaptation, assess population vulnerability under projected future climates, and evaluate how genetic load may constrain adaptive potential. Together, this work establishes a predictive framework that links genomic variation and adaptive potential, providing insights into the evolutionary dynamics of *C. pepo* and guiding germplasm conservation and crop improvement.

Results

Chromosome-scale genome assembly and gene-based pangenome construction

To enable accurate identification of SVs, we generated high-quality reference genomes for nine *C. pepo* accessions using PacBio HiFi long-read sequencing. This panel included the two major subspecies, *C. pepo* subsp. *ovifera* and subsp. *pepo*, comprising four accessions of *C. pepo* subsp. *ovifera* var. *ovifera* and five accessions of *C. pepo* subsp. *pepo* (**Fig. 1a and Supplementary Table 1**). In total, 97.03 Gb of HiFi data were produced, with an average of 28.6 Gb per accession (30.69× coverage; **Supplementary Table 1**). Using these data, we successfully assembled chromosome-scale genomes for all nine accessions. Final assembly sizes ranged from 332.1 Mb to 378.55 Mb (average 359.81 Mb), with contig N50 values between 3.51 Mb and 10.73 Mb (average 6.85 Mb) (**Supplementary Table 1**). On average, 93.85% of contigs were anchored to the 20 chromosomes, with unanchored sequences primarily composed of ribosomal DNA and other repetitive elements.

bioRxiv preprint doi: https://doi.org/10.64898/2026.05.27.728270; this version posted May 29, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

BUSCO[16] analysis revealed an average completeness of 98.06% of these assemblies, while k-mer based evaluation using Merqury[17] showed an average completeness of 97.48% and quality values exceeding 60.74, corresponding to approximately one base error per million bases. The long terminal repeat (LTR) assembly index (LAI) ranged from 9.56 to 12.78 (average 10.98), meeting the criteria for reference-level assemblies[18]. Collectively, these metrics confirmed the high accuracy and completeness of the assemblies. Between 30,412 and 32,484 protein-coding genes were annotated per genome (average 31,501), with BUSCO completeness values ranging from 95.20% to 97.90% (average 96.94%) (**Supplementary Table 1**). Repetitive sequences accounted for an average of 53.81% of the assemblies, ranging from 51.35% to 55.58% (**Supplementary Table 2**). Comparative genomic analysis revealed extensive macrosynteny among the nine assemblies, while large-scale chromosomal rearrangements including inversions and translocations were detected on chromosomes 6, 8, 11, and 13 (**Fig. 1a**).

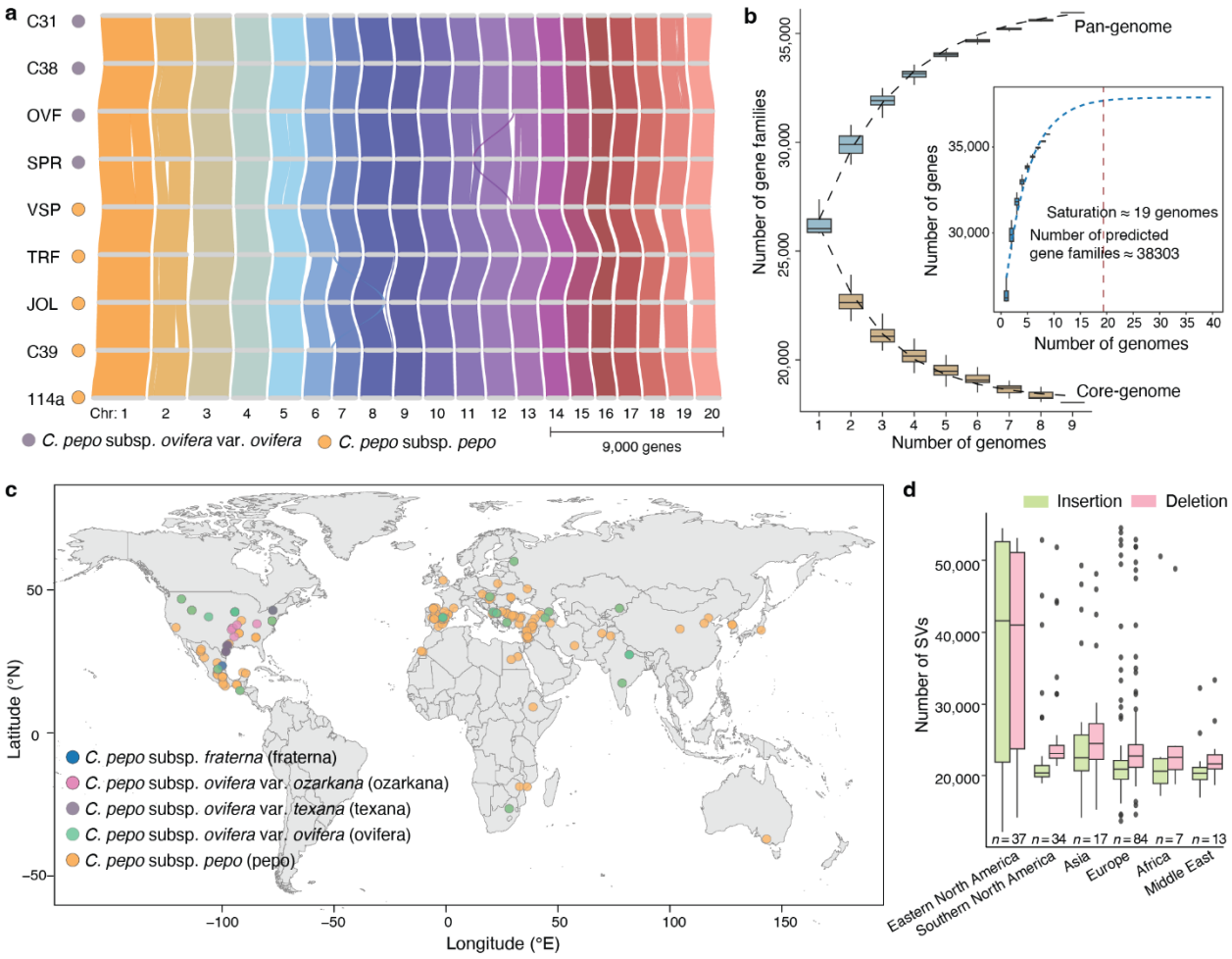

Figure 1 Pangenome of *C. pepo*. a, Chromosome-scale synteny map across the nine *C. pepo* assemblies, generated using GENESPACE (https://github.com/jtlovell/GENESPACE). **b**, Simulations of pangenome expansion and core-genome contraction. For each given number of accessions, 100 random combinations of accessions were sampled. The inset illustrates the predicted pangenome size and estimated saturation. **c**, Geographic distribution of 206 resequenced *C. pepo* accessions. Each colored dot represents an accession. The map was generated using ggplot2 (https://ggplot2.tidyverse.org/). **d,** Number of SVs identified in different *C. pepo* populations. For each boxplot, the lower and upper bounds indicate the first and third

bioRxiv preprint doi: https://doi.org/10.64898/2026.05.27.728270; this version posted May 29, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

quartiles, respectively, the center line indicates the median, and the whiskers extend to 1.5× the interquartile range.

Using protein-coding genes from these nine assemblies, we constructed a gene-based pangenome to characterize gene presence–absence variation. Clustering of all protein-coding genes identified 35,956 distinct gene families, comprising 18,041 (50.2%) core, 2,746 (7.6%) softcore, 12,091 (33.6%) shell, and 3,078 (8.6%) private gene families. The accumulation of new gene families increased rapidly with the addition of the first few genomes and gradually approached a plateau. A nonlinear regression model fitted to the accumulation curve estimated that the pangenome would approach saturation at approximately 19 genomes, corresponding to around 38,303 gene families, indicating that our current sampling captured most of the diversity in *C. pepo* (~93.87%) (**Fig. 1b**).

Population-scale variants revealed by graph-based pangenome

Leveraging the advantages of graph-based pangenomes in capturing complex genomic variation[19], we constructed a graph-based pangenome for *C. pepo* using nine chromosome-scale assemblies. This framework enabled the identification of 229,431 non-redundant SVs (≥20 bp), including 131,373 insertions and 98,058 deletions, collectively spanning a total of 148.83 Mb. Most SVs were small, with 91.52% of insertions and 92.96% of deletions shorter than 1 kb, whereas only approximately 1% exceeded 10 kb in length (**Supplementary Tables 3 and 4**). A substantial proportion of SVs (36.33%) originated from transposable elements (TEs), with DNA transposons containing terminal inverted repeats (DNA-TIRs) representing the most abundant class. Notably, 18.70% of SVs overlapped genic regions, collectively affecting 69.01% of all annotated protein-coding genes (**Supplementary Tables 5**).

To explore SV diversity at the population level, we performed deep genome resequencing of a core collection of 206 *C. pepo* accessions[20] at an average depth of 53.4× (**Supplementary Table 6**). Misidentification and taxonomic inconsistency are well-recognized challenges in plant germplasm collections and are particularly pronounced in *C. pepo* due to its bewildering phenotypic variation. Among the 206 accessions, one was initially annotated as *C. pepo* subsp. *fraterna*, seven as *C. pepo* subsp. *ovifera* var. *ozarkana*, six as *C. pepo* subsp. *ovifera* var. *texana*, and two as *C. pepo* subsp. *pepo*, whereas the remaining 190 were broadly classified as *C. pepo* without further taxonomic resolution (**Supplementary Table 6**). To refine these assignments, we integrated phylogenetic relationships inferred from genome-wide SNPs[20] (**Supplementary Fig. 2**) with manual inspection of fruit morphology. This framework enabled the confident classification of all 190 previously unresolved accessions into specific subspecies and/or cultivar groups. In addition, two accessions were identified as misclassified and reassigned to their correct taxonomic groups (**Supplementary Table 6**). After taxonomic correction, this global collection comprised 12 wild accessions representing three wild subspecies (1 *C. pepo* subsp. *fraterna*, 5 *C. pepo* subsp. *ovifera* var. *texana*, and 6 *C. pepo* subsp. *ovifera* var. *ozarkana*) and 194 cultivated accessions (33 *C. pepo* subsp. *ovifera* var. *ovifera* and 161 *C. pepo* subsp. *pepo*), including both landraces and modern cultivars. These accessions spanned five continents and a broad geographic range, from ~37°S to 60°N latitude and from ~120°W to 142°E longitude, encompassing tropical, subtropical, temperate, and continental climatic regimes (**Fig. 1c and Supplementary Table 6**).

Using the high-depth resequencing data, we genotyped SVs in the graph-based pangenome across all 206 accessions with PanGenie[21]. We observed marked variation in both the number and length of SVs among populations and individual accessions (**Fig. 1d, Supplementary Fig. 1 and**

bioRxiv preprint doi: https://doi.org/10.64898/2026.05.27.728270; this version posted May 29, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Supplementary Table 7). In addition, we identified 5,040,318 high-quality SNPs from the same resequencing dataset (**Supplementary Table 8**). Together, these SNPs and SVs constitute an integrated population-scale genomic resource for investigating the evolutionary history and population differentiation in *C. pepo*.

Parallel domestication of two subspecies of *C. pepo*

The domestication of *C. pepo* is among the earliest known plant domestication events, with archaeological evidence indicating human use of *C. pepo*, as food, as early as 10,000 years before present[12]. Over millennia of cultivation and selection, *C. pepo* has developed remarkable morphological diversity and broad ecological adaptability[15]. Archaeological and molecular evidence indicates that, prehistorically, *C. pepo* was widely distributed in North America, from southern Mexico and perhaps Guatemala to northwards and eastwards through the eastern U.S.A. perhaps as far as the eastern states bordering Canada. The population of Southern North America, subsp. *pepo*, and the population of Eastern North America, subsp. *ovifera*, underwent at least two independent domestication events, the former giving rise to pumpkins and the latter to squash[10,13,14]. However, the evolutionary relationships among wild relatives and cultivated lineages, as well as the extent to which these domestication events were truly independent, have remained unclear.

Phylogenetic analyses based on genome-wide variants have placed a third, undomesticated taxon, subsp. *fraterna*, as allied with subsp. *ovifera* (including var. *texana*, var. *ozarkana*, and var. *ovifera*) or basal to both of the other subspecies, with subsp. *ovifera* and subsp. *pepo* forming two major derived clades (**Supplementary Figs. 2** and **3**). Within the *ovifera* lineage, the cultivated form (*C. pepo* subsp. *ovifera* var. *ovifera*) clustered most closely with the wild taxon var. *texana*, supporting a potential direct domestication origin from a *texana*-like ancestor. In contrast, subsp. *pepo*, the globally distributed and economically dominant cultivated lineage, lacks an extant wild progenitor, obscuring its origin in simple phylogenetic reconstructions. Population structure analyses further supported a clear separation between the two major subspecies (**Supplementary Fig. 4**). At $K = 4$, subsp. *ovifera* and subsp. *pepo* formed distinct genetic clusters, while the wild relatives within the *ovifera* lineage (var. *texana* and var. *ozarkana*) showed finer-scale differentiation from the cultivated *ovifera* group. Notably, Southern North American accessions of subsp. *pepo* formed a partially distinct cluster, suggesting geographic isolation and subsequent diversification following a southward expansion from a more northern origin (**Fig. 2a** and **Supplementary Fig. 4**). Similar patterns were observed in principal component analysis (PCA) and genome-wide genetic differentiation (F_{ST}). PC1 (38.35%) clearly separated the two subspecies, whereas PC2 (3.03%) captured within-subspecies variation, including differentiation among varieties and geographic groups (**Fig. 2b**). Consistently, F_{ST} values were low within subspecies but markedly elevated between them (**Fig. 2c**), indicating strong divergence between the two domesticated gene pools.

Although these patterns revealed substantial divergence between subsp. *pepo* and subsp. *ovifera*, they did not by themselves establish whether the two cultivated lineages originated independently. To address this question, we evaluated alternative evolutionary scenarios using admixture graph models implemented in qpGraph[22]. Models assuming a single domestication origin–either deriving both cultivated lineages from a *texana*-like ancestor or treating subsp. *pepo* as derived from cultivated *ovifera*–were strongly rejected based on large absolute Z-scores (**Supplementary Figs. 5a,b**). In contrast, models in which subsp. *pepo* originates from an unsampled ancestral population and evolves in parallel with the *ovifera* lineage showed a substantially improved fit, supporting independent domestication (**Supplementary Fig. 5c**).

bioRxiv preprint doi: https://doi.org/10.64898/2026.05.27.728270; this version posted May 29, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Across the 92 models tested, the best-fitting topology (|Z| = 2.057) consistently supported a parallel domestication scenario (**Fig. 2d and Supplementary Figs. 5d-f**). In this model, subsp. *pepo* forms an early-diverging lineage relative to the *ovifera* clade, whereas cultivated *ovifera* is derived primarily from a *texana*-like ancestor with a substantial contribution from an additional ancestral lineage. Together, these results reject a single-origin model and indicate that the two cultivated subspecies arose independently from distinct ancestral gene pools.

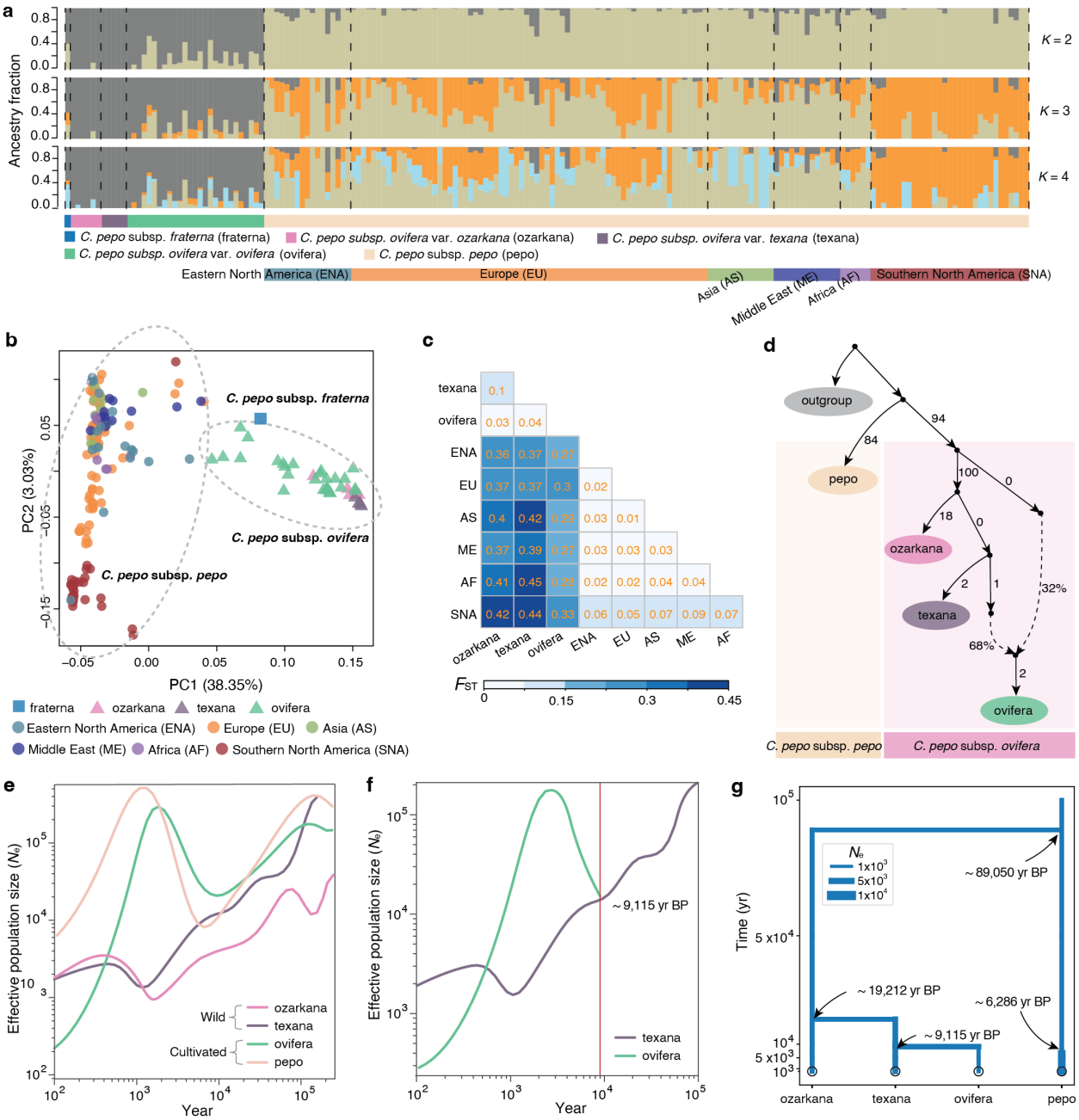

Figure 2 Genetic structure, divergence, and demographic history in *C. pepo*. a, Population structure inferred at $K = 2$–4. Each bar represents an accession and is colored according to ancestry proportion. **b,** Principal component analysis (PCA) of *C. pepo* accessions. **c,** Pairwise fixation index (F_{ST}) among populations. **d**, Inferred admixture graph showing relationships among *C. pepo* groups. Numbers on

bioRxiv preprint doi: https://doi.org/10.64898/2026.05.27.728270; this version posted May 29, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

branches indicate drift parameters, and dashed arrows indicate admixture events. **e**, Changes in effective population size (N_e) over time in *C. pepo* groups. **f,** Estimated divergence time between cultivated *C. pepo* subsp. *ovifera* var. *ovifera* and its closest wild relative *C. pepo* subsp. *ovifera* var. *texana*. **g**, Schematic representation of divergence times among major *C. pepo* groups.

Demographic inference further supported this scenario. Both cultivated lineages (*pepo* and *ovifera*) underwent pronounced bottlenecks between ~5,000 and 10,000 years ago, followed by population expansion associated with human cultivation and dispersal (**Fig. 2e**). In contrast, wild populations (*ozarkana* and *texana*) showed a gradual decline in effective population size over time (**Fig. 2e**). Integrating SMC++ [23] and momi2[24] analyses, we reconstructed a temporal framework for the diversification of *C. pepo* (**Fig. 2f,g**). The two major subspecies diverged ~89,050 years before present, followed by the split between var. *texana* and var. *ozarkana* at ~19,212 years before present. The *ovifera* lineage was domesticated ~9,115 years ago from a *texana*-like ancestor (**Fig. 2f**), coinciding with the onset of agriculture in North America[25]. We further estimated that subsp. *pepo* was domesticated around 6,286 years ago, based on the timing of its domestication bottleneck, likely from an unsampled or extinct wild progenitor, followed by global expansion. This estimate is more recent than archaeological evidence[10] (~8,000–10,000 years) and likely reflects a later phase of population contraction rather than the initial onset of domestication. Together, these results support a model of parallel domestication in *C. pepo*, in which two deeply diverged ancestral gene pools independently gave rise to distinct cultivated lineages.

Differential selective constraints on SNPs and SVs in *C. pepo*

To characterize the evolutionary forces shaping genomic variation in *C. pepo*, we investigated the contributions of both SNPs and SVs to adaptive variation across the wild populations and across each of the two cultivated subspecies, *ovifera* and *pepo*. We estimated the distribution of fitness effects (DFE) from population allele frequency spectra using synonymous SNPs as a neutral reference[26]. We observed that in all three populations, SVs experienced stronger purifying selection than SNPs, with a larger proportion of SVs showing strongly deleterious effects ($N_eS < -100$) (**Fig. 3a,b**). This pattern is consistent with the larger functional impacts of SVs, which frequently disrupt coding or regulatory regions and are therefore rapidly purged by selection, as also reported in cucumber[27].

Despite pervasive purifying selection, a small subset of SVs persisted at moderate to high frequencies, suggesting that these variants may confer substantial fitness advantages that allow them to escape elimination. To further quantify the adaptive variation, we estimated the proportion of fixed adaptive mutations (α) and found that SNPs contribute more to adaptive variation than SVs, consistent with the stronger purifying selection acting on SVs (**Fig. 3c**). Therefore, incorporating SVs into population genomic frameworks provides a more comprehensive understanding of the evolutionary forces that have shaped the genomic architecture and adaptive potential of *C. pepo*.

Climate-driven genomic differentiation in *C. pepo*

Accessions in the *C. pepo* core collection span a broad geographic range and encompass diverse climatic backgrounds. Moreover, most cultivated accessions represent landraces and historical or older varieties, with many collected before the mid-1980s (**Supplementary Table 6**). As a result, these accessions have likely been maintained locally for extended periods, enabling them to retain genetic signatures associated with local environmental conditions. Together, this core collection provides a valuable framework for investigating the role of climate in shaping genomic variation

bioRxiv preprint doi: https://doi.org/10.64898/2026.05.27.728270; this version posted May 29, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

across *C. pepo* populations. Accordingly, as in previous studies[28–32], we integrated both wild and cultivated accessions to identify genomic variants associated with environmental gradients across the species' range.

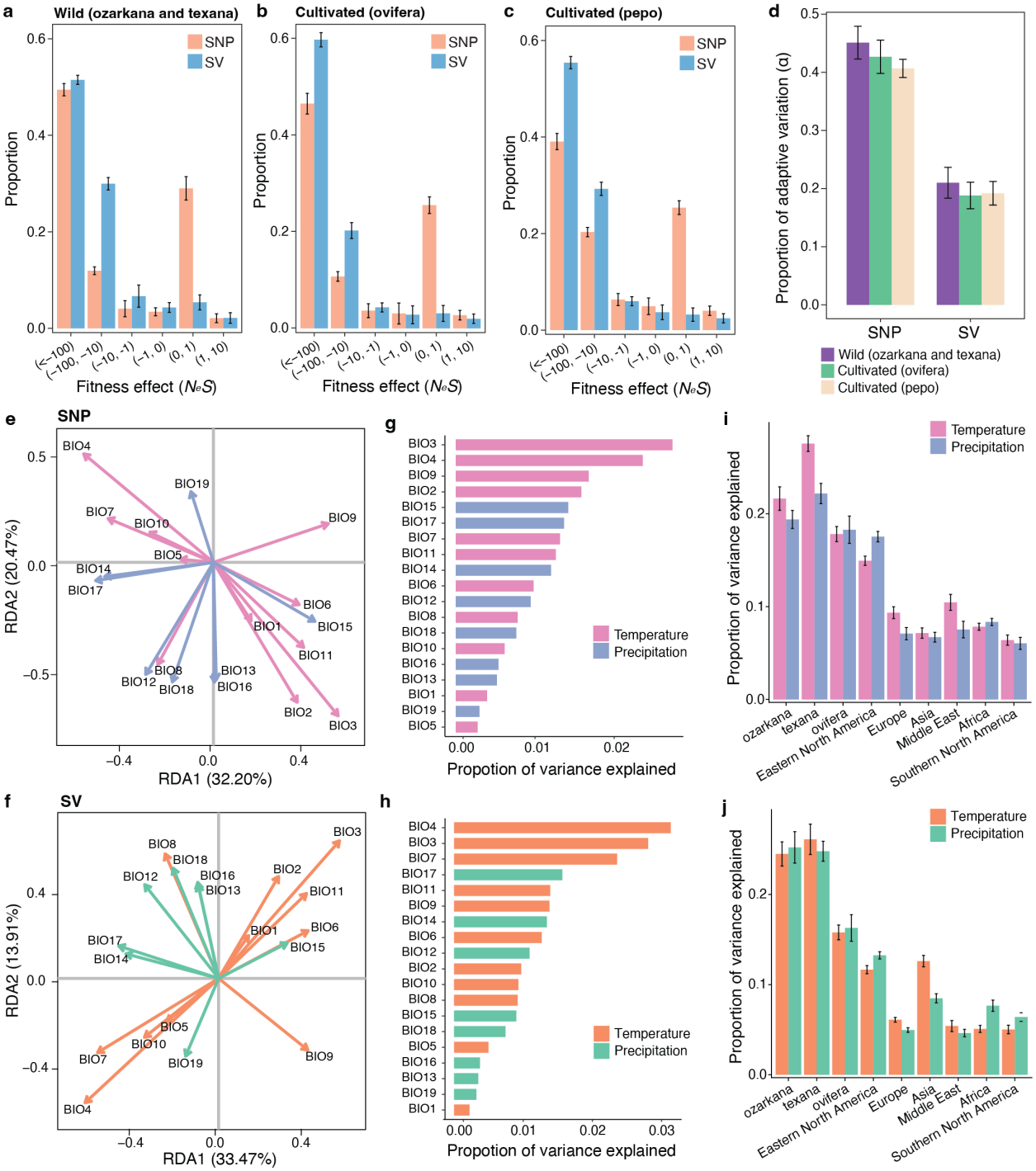

Figure 3 Climate-driven genomic differentiation in *C. pepo*. a–c, Distribution of fitness effects (N_eS) for SNPs and SVs in wild (**a**), cultivated *ovifera* (**b**) and cultivated *pepo* (**c**) populations. **d**, Proportion of adaptive variation (α) contributed by SNPs and SVs in wild and two cultivated groups. For **a-d**, values are derived from 20 bootstrap replicates; bars represent mean ± SD. **e-f,** Contribution of bioclimatic variables

bioRxiv preprint doi: https://doi.org/10.64898/2026.05.27.728270; this version posted May 29, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

(BIO1–BIO19) to the first two principal components derived from redundancy analysis (RDA) based on SNPs (**e**) and SVs (**f**). **g-h,** Proportion of genomic variance explained by individual bioclimatic variables for SNPs **(g)** and SVs **(h)**. **i-j,** Climate-associated genomic variance partitioned across subspecies and geographic regions for SNPs (**i**) and SVs (**j**).

We performed redundancy analysis (RDA)[33] using 19 bioclimatic variables, including 11 temperature-related (BIO1–BIO11) and 8 precipitation-related variables (BIO12–BIO19) (**Supplementary Table 9**). Both SNP- and SV-based RDAs revealed strong associations between genomic variation and multiple climatic predictors, with the first two RDA axes explaining substantial proportions of total genomic variance (**Fig. 3e,f**). The full RDA models were highly significant for both SNP and SV datasets ($P < 0.001$). Because environmental variables are often spatially structured, we further assessed the robustness of these climate–genome associations by accounting for spatial autocorrelation using distance-based Moran's eigenvector maps (PCNM)[34]. After incorporating spatial structure, climatic predictors remained highly significant ($P < 0.001$), demonstrating that climate significantly contributes to genomic differentiation in *C. pepo*.

Among climatic predictors, temperature-related variables explained more genomic variation than precipitation-related variables. In particular, BIO3 (isothermality) and BIO4 (temperature seasonality) showed the greatest explanatory power, highlighting that temperature gradients contribute more strongly to genomic differentiation in *C. pepo* (**Fig. 3g,h**). Notably, climatic factors explained a larger proportion of genetic variation in wild populations than in cultivated populations (**Fig. 3i,j**). This pattern likely reflects the longer evolutionary exposure of wild relatives to diverse environmental conditions, as well as relaxed environmental selection under cultivated conditions.

Genomic loci associated with climatic gradients

Given the strong contribution of climatic gradients to genome-wide differentiation, we next sought to identify specific genomic loci associated with climatic variation using genome–environment association (GEA) analyses. Because the 19 bioclimatic variables related to temperature and precipitation are highly correlated (**Supplementary Fig. 6**), we reduced dimensionality by performing separate principal component analyses (PCAs) for temperature-related and precipitation-related variables. The first two principal components explained 74.17% and 84.98% of the total variance for temperature and precipitation, respectively, and were retained as composite climatic predictors for downstream analyses (**Supplementary Fig. 7**).

Using these major climatic principal components, we identified climate-associated loci with Bayenv2 [35]. Both SNPs and SVs exhibited abundant associations with environmental variables **(Fig. 4a,b** and **Supplementary Fig. 8a,b**), with effect sizes exceeding neutral expectations based on permutation analyses, indicating robust environmental signals beyond neutral demographic structure. We identified 12,946 climate-associated SNPs and 7,382 SVs, which were broadly distributed across the genome rather than clustered in a few genomic regions, suggesting widespread environmentally driven differentiation. This dispersed pattern supports a largely polygenic architecture of climatic adaptation in *C. pepo*, likely involving many loci of small to intermediate effect rather than a limited set of major-effect variants.

To further assess the robustness of climate–allele associations, we applied Gradient Forest models[36], which quantify genetic turnover by measuring how rapidly allele frequencies change along environmental gradients. Both adaptive SNPs and SVs exhibited pronounced nonlinear allele turnover across climatic gradients, with significantly higher turnover rates than randomly sampled loci (**Fig. 4c–f** and **Supplementary Fig. 8c–f**). Temperature-related variables produced

bioRxiv preprint doi: https://doi.org/10.64898/2026.05.27.728270; this version posted May 29, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

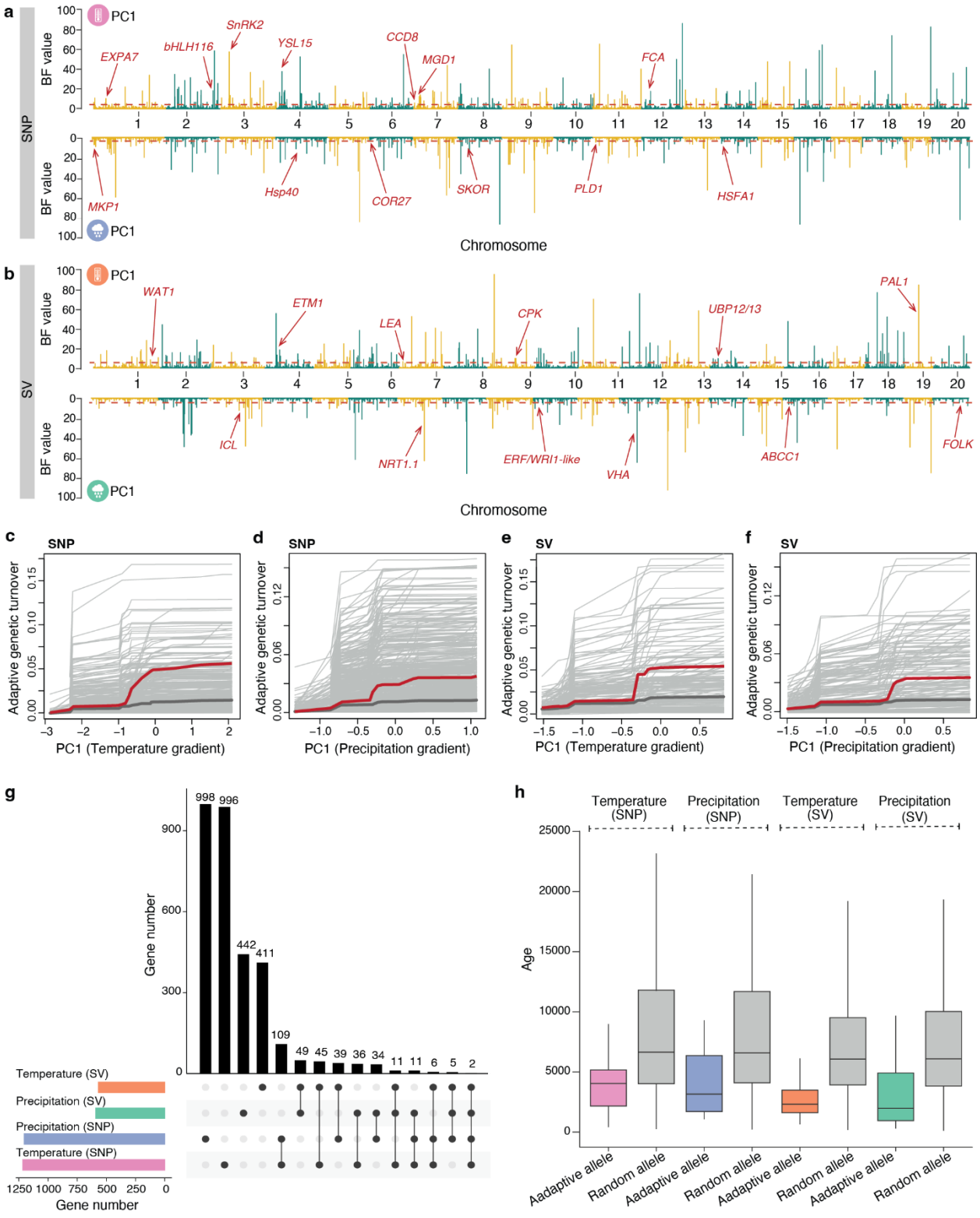

Figure 4 Climate-associated adaptive genetic variation inferred from SNPs and SVs. a,b, Genome-wide distribution of climate-associated SNPs (**a**) and SVs (**b**) along PC1 of temperature (top panels) and precipitation (bottom panels) gradients. The dashed red line indicates the significance threshold, and

bioRxiv preprint doi: https://doi.org/10.64898/2026.05.27.728270; this version posted May 29, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

representative candidate genes near the top signals are shown. Details of the representative candidate genes are provided in **Supplementary Table 12**. **c–f,** Turnover of adaptive SNPs (**c**,**d**) and SVs (**e**,**f**) along PC1 of temperature and precipitation gradients. Thin light-grey lines represent turnover trajectories of individual loci, thick dark-grey lines indicate the mean cumulative turnover of random loci, and colored thick lines denote the mean turnover of climate-associated adaptive loci. **g,** UpSet plot showing overlap among candidate genes associated with climate adaptation. **h,** Estimated variant ages of temperature- and precipitation-associated adaptive alleles and random alleles. For each boxplot, the lower and upper bounds indicate the first and third quartiles, respectively, the center line indicates the median, and the whiskers extend to 1.5× the interquartile range.

stronger genetic turnover than precipitation-related variables, consistent with temperature explaining a larger fraction of genomic variation (**Fig. 3g,h**). The consistent turnover patterns observed for SNPs and SVs further support the robustness of the identified climate-associated loci and reinforce the conclusion that climatic gradients have played an important role in shaping adaptive genomic variation in *C. pepo*.

In total, we identified 3,194 genes associated with climatic gradients, spanning functional categories related to genome maintenance, transcriptional regulation, cell wall biogenesis, and metabolic and transport processes (**Supplementary Tables 10-12** and **Supplementary Fig. 9**). For example, we detected an ethylene-responsive transcription factor (*ERF/WRI1-like*) involved in abiotic stress responses, as well as a gene encoding cell wall–modifying protein homologous to the rice expansin *OsEXPA7* that was previously implicated in salinity tolerance[37] (**Fig. 4a,b**). Notably, most climate-associated genes were specific to particular variant-environment combinations, whereas only a small subset was shared between SNP- and SV-based signals or between temperature- and precipitation-associated loci (**Fig. 4g**). SNPs and SVs therefore capture largely complementary sets of climate-responsive genes, highlighting their distinct contributions to climate adaptation.

We then analyzed the ages of climate-associated variants and found that adaptive variants are generally younger than genome-wide background variants for both SNPs and SVs (**Fig. 4h**), suggesting that many climate-associated alleles arose relatively recently. SVs tended to be younger than SNPs, consistent with their larger functional impacts and the stronger purifying selection that preferentially removes older deleterious SVs from populations, as reported in cucumber[27]. Together, these patterns suggest that recent environmental pressures have strongly shaped the adaptive landscape of *C. pepo*, particularly through the retention of recently arisen, functionally impactful variants.

Predicting climate vulnerability in *C. pepo*

To forecast how *C. pepo* populations may respond to future climates, we integrated ecological niche modeling with genetic offset analyses informed by climate-associated SNPs and SVs. Species distribution models (SDMs)[38] projected substantial rearrangements of suitable habitat under both mid-century (2040–2060) and late-century (2080–2100) climate scenarios (**Fig. 5a,b**). The wild relatives *C. pepo* subsp. *ovifera* var. *ozarkana* and *var. texana* are projected to experience pronounced range contractions accompanied by northward shifts, particularly across the southern United States, indicating heightened vulnerability to warming and increasing aridity (**Fig. 5a** and **Supplementary Fig. 10**). In contrast, the two cultivated subspecific taxa (var. *ovifera* and subsp. *pepo*) displayed more complex responses, with projected expansions into higher latitudes in both

bioRxiv preprint doi: https://doi.org/10.64898/2026.05.27.728270; this version posted May 29, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

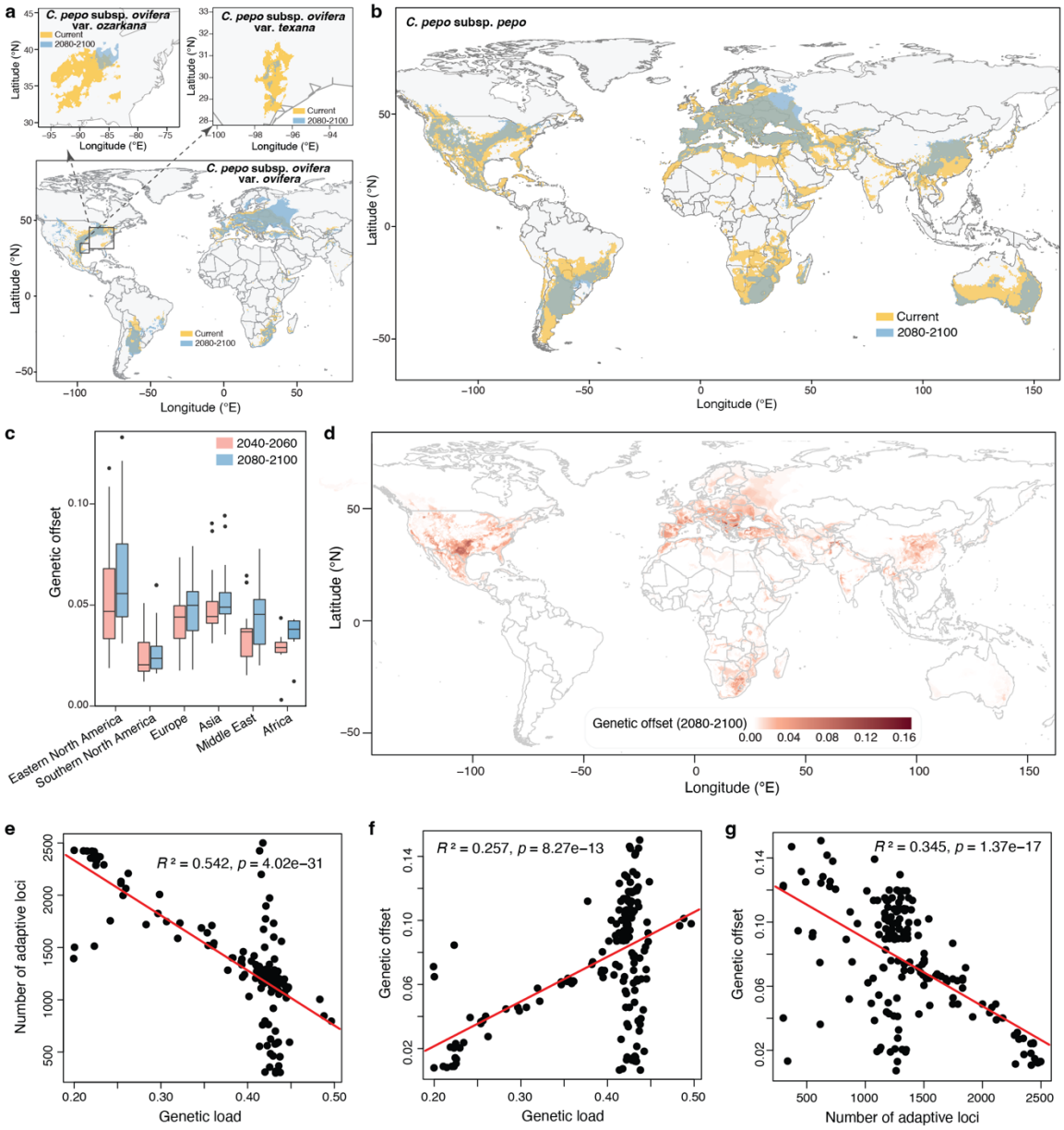

Figure 5 | Projected range shifts and genomic vulnerability for *C. pepo* under future climate scenarios. **a**, Predicted current and late-century (2080–2100) climatically suitable areas for wild *C. pepo* subsp. *ovifera* var. *ozarkana* and var. *texana* (top panels) and cultivated *C. pepo* subsp. *ovifera* var. *ovifera* at the global scale (bottom panel), inferred from species distribution models. Boxes in the bottom panel indicate Eastern North America regions shown in the top panels. **b**, Predicted current and late-century (2080–2100) climatically suitable areas for *C. pepo* subsp. *pepo* at the global scale. **c**, Regional comparison of genetic offset across continents under mid-century (2040–2060) and late-century climate scenarios. For each boxplot, the lower and upper bounds indicate the first and third quartiles, respectively, the center line indicates the median, and the whiskers extend to 1.5× the interquartile range. **d**, Spatial distribution of genetic offset projected for late century for the entire *C. pepo* population. **e–g**, Pearson correlations between genetic load, number of adaptive loci, and genetic offset. Maps in **a, b** and **d** were created using the R package rworldmap (https://cran.r-project.org/web/packages/rworldmap/index.html).

bioRxiv preprint doi: https://doi.org/10.64898/2026.05.27.728270; this version posted May 29, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

hemispheres and contractions in warmer regions such as Central America and northern Africa (**Fig. 5a,b** and **Supplementary Figs. 10,11**). Together, these projections indicate that future climatic regimes may substantially reshape the ecological and geographical landscape of *C. pepo*, underscoring the need for conservation strategies that protect its full diversity, particularly wild relatives that harbor unique reservoirs of adaptive genetic variation.

To incorporate evolutionary potential into assessments of future maladaptation risk, we estimated genetic offset, which quantifies the magnitude of allele-frequency shifts required for populations to remain adapted under projected future climates. Because SNPs and SVs capture complementary components of climatic adaptation (**Fig. 4g**), both variant classes were jointly integrated into the genomic prediction framework. Genetic offset increased from mid-century to late-century climate scenarios, indicating a progressively rising risk of maladaptation over time (**Fig. 5c**). Notably, populations from Eastern North America exhibited the highest genetic offset values, suggesting that they would require substantial allele-frequency changes to maintain local adaptation under future climatic conditions (**Fig. 5d** and **Supplementary Fig. 12**).

To investigate how intrinsic genomic constraints shape vulnerability to future climates, we examined relationships among genetic load, the number of adaptive loci, and genetic offset (**Fig. 5e–g**). Genetic load–defined as the cumulative burden of deleterious alleles within a population–was negatively correlated with the number of adaptive loci ($R^2 = 0.542$, $P = 4.02 \times 10^{-31}$; **Fig. 5e**), consistent with the possibility that populations with higher mutational burdens tend to harbor fewer adaptive variants available for selection. Genetic load was also positively associated with genetic offset ($R^2 = 0.247$, $P = 8.27 \times 10^{-13}$; **Fig. 5f**), suggesting that populations with elevated load would require larger genomic shifts to adapt to future climates. Conversely, the number of adaptive loci was negatively correlated with genetic offset ($R^2 = 0.345$, $P = 1.37 \times 10^{-17}$; **Fig. 5g**), highlighting a potential buffering role of adaptive variation in mitigating climate-induced maladaptation. Together, these results suggest that genetic load represents an important constraint on adaptive potential under future climates and provides a quantitative basis for identifying *C. pepo* populations that may face greater challenges under changing environmental conditions.

Discussion

Environmental variation poses a major challenge to both biodiversity conservation and agricultural sustainability[1]. Understanding how genetic variation underlies responses of species to environmental heterogeneity is therefore key to predicting evolutionary persistence and informing breeding strategies[2]. By integrating genome-wide variation with climate-associated differentiation and predictive modeling, our study provides a comprehensive understanding of how *C. pepo* has evolved, adapted, and may respond to changing environmental conditions.

Our integrative analyses provide strong evidence for parallel domestication in *C. pepo*, in which two deeply diverged gene pools independently gave rise to the cultivated *pepo* and *ovifera* lineages. The two subspecies diverged approximately 89,000 years ago, followed by the split of *texana* and *ozarkana* around 19,000 years ago. Domestication of *ovifera* occurred ~9,100 years ago from a texana-like ancestor, whereas *pepo* was domesticated later (~6,300 years ago), likely from an unsampled or extinct progenitor. Despite their independent origins, both lineages experienced similar demographic trajectories, including domestication bottlenecks followed by expansion associated with human cultivation and dispersal. As these lineages expanded across diverse environments, climatic gradients emerge as an important force shaping genomic differentiation. Genome–environment association analyses identified numerous loci correlated with climatic principal components, broadly distributed across the genome and indicative of highly

bioRxiv preprint doi: https://doi.org/10.64898/2026.05.27.728270; this version posted May 29, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

polygenic climatic adaptation. Gradient Forest modeling further reveals nonlinear allele turnover along temperature and precipitation gradients, demonstrating that adaptive loci respond more strongly to changing environmental conditions than randomized variant sets. Importantly, SNPs and SVs capture complementary dimensions of this environmental responsiveness. Temporal analyses provided additional insight into the dynamics of adaptation: adaptive variants are consistently younger than genome-wide background variants, suggesting that much of the adaptive genomic architecture in *C. pepo* has arisen or been favored relatively recently, either during post-domestication expansion or in response to Holocene environmental variability.

Projections under future climate scenarios indicate increasing genetic offset, particularly in Eastern North American populations, along with continued range contraction of wild relatives, highlighting potential challenges for maintaining local adaptation under rapid environmental change. In addition to environmental drivers, intrinsic genomic properties also play an important role in shaping adaptive capacity. We identify genetic load as a key constraint on adaptation: populations with higher mutational burdens harbor fewer adaptive loci and exhibit greater predicted maladaptation. This pattern emphasizes the buffering role of standing adaptive variation and reveals an evolutionary link among genetic diversity, mutational load, and adaptive potential. Such constraints are especially relevant for understanding why some populations are disproportionately vulnerable to future climates, even under similar environmental pressures.

Beyond providing insights into the evolutionary history and climatic adaptation of *C. pepo*, our study highlights the value of integrating multiple layers of genomic variation into predictive frameworks of environmental response. Most population genomic studies of crop adaptation have focused primarily on SNPs[39,40], yet our results show that SVs capture additional dimensions of environmental responsiveness that would otherwise remain undetected. By jointly analyzing SNPs and SVs within a graph-based pangenome framework, we provide a more complete view of the genomic architecture underlying environmental adaptation. Such integrative approaches will be increasingly important for understanding how complex genomes respond to environmental changes and for identifying genetic variants that contribute to responses to future environmental conditions. More broadly, our findings emphasize that adaptive potential emerges from the interplay between beneficial and deleterious variation. While standing adaptive variation provides the raw material for evolutionary responses to changing environments, the accumulation of deleterious mutations can simultaneously constrain this process. The strong association observed here among genetic load, adaptive variation, and predicted maladaptation highlights the importance of considering both evolutionary opportunities and genomic constraints when evaluating adaptive potential in crop populations.

Overall, our findings reveal how the interplay among selection, genomic variation, and genetic constraints shapes the adaptive trajectory of *C. pepo* across temporal and environmental scales. Conserving genetically diverse and locally adapted populations will be critical for sustaining its evolutionary potential. This integrative genomic framework also offers a broadly applicable approach for assessing adaptive capacity and guiding conservation and breeding strategies in other crops.

Methods

Genome sequencing, assembly and annotation

High-molecular-weight genomic DNA was extracted from young fresh leaves using the cetyltrimethylammonium bromide (CTAB) method[41]. SMRTbell libraries were constructed following the standard protocol provided by PacBio and sequenced on a PacBio Sequel II platform

bioRxiv preprint doi: https://doi.org/10.64898/2026.05.27.728270; this version posted May 29, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

using the circular consensus sequencing mode to generate HiFi reads. For all accessions, HiFi reads were processed using HiFiAdapterFilt[42] (v2.0.1) to remove adapter sequences. The cleaned HiFi reads were assembled into contigs using hifiasm[43] (v0.14.2-r315). Shorter contigs that were represented by a longer contig with at least 99% identity and coverage were identified and removed as haplotypic duplications according to BLAST+ [44] (v2.12.0+) alignments. Potential contaminant sequences from microorganisms and organelle genomes were also identified and removed by comparing contig sequences against the NCBI nt database using BLAST+ [44] (v2.12.0+). For the accession C39, cleaned contigs were anchored to chromosomes by integrating three published genetic maps[45–47], a Hi-C dataset from DNA Zoo using the 3D-DNA pipeline[48] (v180922), as well as the genome assembly of *C. pepo* accession MU-CU-16 [49] for reference-guided anchoring using RagTag[50] (v2.0.1). For the remaining eight accessions (C31, C38, JOL, OVF, SPR, TRF, 114a, and VSP), chromosome-level assemblies were generated using RagTag[50] (v2.0.1) with the C39 genome as reference.

Repeat sequences across the nine *C. pepo* genome assemblies were identified using RepeatModeler[51] (v2.0.4) and RepeatMasker[52] (v4.1.1). Protein-coding genes were initially predicted from the repeat-masked genomes using the MAKER pipeline[53] (v3.01.03), which integrates *ab initio* predictions, transcript evidence, and homologous protein evidence. Specifically, AUGUSTUS[54] (v3.4.0) and SNAP[55] (v2006-07-28) were used for *ab initio* gene prediction, and RNA-seq data from various tissues downloaded from NCBI (accessions PRJNA239659, PRJNA752874, PRJNA339848, PRJNA386743, PRJNA437911, and PRJNA439198) were used for transcript evidence. For protein homology prediction, protein sequences from Arabidopsis (TAIR10), melon (*Cucumis melo*)[56], cucumber (*Cucumis sativus*)[57], watermelon (*Citrullus lanatus*)[58], wax gourd (*Benincasa hispida*)[59], snake gourd (*Trichosanthes anguina*)[60], and chayote (*Sechium edule*)[61], *Cucurbita pepo*[49], *Cucurbita maxima*[62], *Cucurbita moschata*[62], and the UniProt (Swiss-Prot plant division) database were aligned to the genomes using Spaln[63] (v3.0.1). Predicted gene models were subsequently transferred across assemblies using Liftoff[64] (v1.5.1). Furthermore, Helixer (v0.3.2), a gene prediction tool for eukaryotes that combines deep learning and a hidden Markov model[65], was also applied to all nine assemblies. Finally, results from all approaches were integrated using EVidenceModeler[66] (v1.1.1) to generate the final set of predicted protein-coding genes for each genome assembly.

Gene-based and graph-based pangenome construction

Protein sequences from the nine *C. pepo* genomes were clustered into gene families (pan-gene clusters) using OrthoFinder[67] (v2.5.5). Based on their presence across the nine accessions, pan-gene clusters were classified into four groups: core (present in all nine accessions), soft core (present in eight accessions), shell (present in two to seven accessions), and private (present in only one accession).

The graph-based pangenome was constructed using the Minigraph-Cactus package[68] with the nine reference-level genome assemblies. Pangenome construction began with the C39 genome as the initial graph, to which the remaining eight assemblies were sequentially mapped. The resulting genome graph was indexed and exported to VCF format using the vg toolkit[69] and filtered using vcfbub (https://github.com/pangenome/vcfbub) with parameters '-l 0 -r 10000'.

Short-read sequencing, SNP calling, and SV genotyping

Genomic DNA was isolated from young fresh leaf tissues using the Omega Mag-Bind Plant DNA DS Kit (M1130, Omega Bio-Tek, Norcross, GA) following the manufacturer's instructions.

bioRxiv preprint doi: https://doi.org/10.64898/2026.05.27.728270; this version posted May 29, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Shotgun libraries were constructed from the extracted DNA and sequenced on the BGISEQ-500 platform to generate 150-bp paired-end reads. Raw sequencing reads were processed to remove adaptors and low-quality sequences using Trimmomatic[70] (v0.39). For SNP calling, cleaned reads were aligned to the C39 genome using BWA-MEM[71] (v0.7.17-r1188) with default parameters, and duplicate read pairs were marked. SNPs were called using the Sentieon package (https://www.sentieon.com/) and filtered using GATK[72] (v4.2.5.0) with parameters 'QD < 2.0 || FS > 60.0 || MQ < 40.0 || MQRankSum < -12.5 || ReadPosRankSum < -8.0'. Furthermore, only biallelic SNPs with minor alleles present in at least two accessions were retained. SVs in the graph-based pangenome were genotyped across the resequenced *C. pepo* accessions with cleaned resequencing reads using PanGenie[21] (v3.0.2) and default parameters. SNPs and SVs were functionally annotated using SnpEff[73] (v5.1).

Population genomic analyses

Maximum-likelihood phylogenetic trees were constructed separately with the complete set of SVs and 109,659 SNPs located at fourfold degenerate sites (4DTv) using RAxML[74] (v8.2.13) under the GTRGAMMA model with 100 bootstrap replicates. Three watermelon accessions[75] (GenBank accession numbers: SRR8751845, SRR8751846, and SRR8751851) were used as the outgroup. Phylogenetic trees were visualized using iTOL[76] (v7). Population structure was inferred using ADMIXTURE[77] (v1.3.0) with SNPs having a minor allele frequency (MAF) ≥ 0.05 and linkage disequilibrium (LD) ≤ 0.2. Principal component analysis (PCA) was performed using PLINK[78] (v1.9). Populations fixation index (F_{ST}) was calculated separately with SNPs and SVs using VCFtools[79] (v0.1.15) in 100-kb sliding windows with a 10-kb step size.

Reconstruction of demographic history

Effective population sizes were inferred using SMC++ [23] (v1.15.4.dev18+gca077da). The genome of each accession was treated as a single haplotype, with occasional heterozygous sites resolved by randomly selecting one allele. For each population, haplotypes from different accessions were combined to create pseudo-diploid genotypes. A total of 15 pseudo-diploid genotypes were randomly selected for analysis. A mutation rate of 6.5×10^{-9} per site per generation[80] and a generation time of 1 year were applied. The SMC++ analysis was repeated 20 times for each population, with pseudo-diploid genotypes resampled in each iteration. Admixture graph analyses were performed using AdmixTools[22]. The qpGraph function was used to evaluate model fit, and a heuristic search implemented in qpBrute[81] was used to explore graph space. A total of 92 candidate models were tested, and those with $|Z| < 3$ were retained. Among these, the best-fitting model was identified using the admixturegraph package[82] in R based on marginal likelihood and Bayes factors. Finally, demographic models were further explored using momi2[24] (v2.1.19). Folded site frequency spectra (SFS) were computed and divided into 100 blocks for jackknife and bootstrap resampling. Divergence times inferred from SMC++, such as the split between *ovifera* and *texana*, were used as priors to constrain momi2 model fitting.

The ages of SNPs and SVs were estimated using GEVA[83] (v1beta), applying the same mutation rate. Variant ages were computed using the joint clock model, which integrates both mutation and recombination clock models to improve estimation accuracy.

Genetic load and distribution of fitness effects calculation

Ancestral and derived alleles for each variant were determined using watermelon (*Citrullus lanatus*)[58] and *Cucurbita argyrosperma*[84]–the former distantly related and the latter closely related

bioRxiv preprint doi: https://doi.org/10.64898/2026.05.27.728270; this version posted May 29, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

to *Cucurbita pepo*–as outgroup species. SNP and SV alleles shared among *C. pepo*, *C. lanatus*, and *C. argyrosperma* were considered ancestral. Genetic load for each accession was calculated as the ratio of derived alleles to derived synonymous alleles. The distribution of fitness effects (DFE) and the proportion of adaptive variants (α) for nonsynonymous SNPs (nSNPs) and SVs were estimated using polyDFE[85] (v2.15) based on the unfolded site frequency spectrum (SFS). The 95% confidence intervals of the α estimates were derived from discretized DFEs inferred through 20 bootstrap replicates.

Climatic variables collection and redundancy analysis

Climatic variables were obtained from the WorldClim database (https://www.worldclim.org/), which provides high-resolution global climate surfaces derived from long-term weather station records. Nineteen standard bioclimatic variables (BIO1–BIO19), including 11 temperature-related (BIO1–BIO11) and 8 precipitation-related variables (BIO12–BIO19), were downloaded at a spatial resolution of 30 arc-seconds (~1 km²), representing average climatic conditions for 1970–2000. Future climate projections were obtained from downscaled CMIP5 climate data available under the RCP8.5 greenhouse gas emission scenario, representing a high-emission trajectory[86]. Bioclimatic variables for two future periods–mid-century (2040-2060) and late-century (2080-2100)–were downloaded from the WorldClim database to assess potential environmental variation across the geographic range of *C. pepo*.

Geographic coordinates for each accession were used to extract corresponding climatic values from raster layers using the extract function in the R package raster (v3.3.13) (https://cran.r-project.org/web/packages/raster). To evaluate the relationship between genomic variation and climatic gradients, we performed redundancy analysis (RDA)[33] using the R package vegan (https://cran.r-project.org/web/packages/vegan/index.html). RDA was conducted separately for SNPs and SVs, and only variants without missing data and with minor allele frequency (MAF) > 0.05 were retained as response variables to minimize the influence of rare alleles. The 19 bioclimatic variables were used as explanatory variables in the RDA models. To quantify the contribution of climatic predictors to genomic variation, we further conducted variance partitioning within the RDA framework to estimate the proportion of genetic variation explained by the full set of climatic variables. Statistical significance of the RDA models was assessed using permutation tests implemented in the vegan package.

Genome–environment association analysis

Associations between environmental variables and genome-wide variants were identified using Bayenv2[35] with the 19 bioclimatic variables. To account for population structure, a covariance matrix describing population allele frequency covariance was estimated using 100,000 MCMC iterations. For each SNP or SV, Bayes factors (BFs) were computed across 10,000 iterations to evaluate associations between allele frequencies and environmental variables. To ensure robustness, the analysis was repeated five times, and the median BF across runs was used as the final statistic for each variant. Empirical null BF distributions were generated by permuting environmental variables across populations. Variants with BF values exceeding the top 0.1% of the permutation-based BF distribution for each environmental variable were retained as climate-associated loci.

Species distribution modelling

bioRxiv preprint doi: https://doi.org/10.64898/2026.05.27.728270; this version posted May 29, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Species distribution models (SDMs)[38] were constructed using the R package dismo (https://cran.r-project.org/web/packages/dismo/index.html) to estimate the potential habitat range of *C. pepo* under current and future climate conditions. *C. pepo* occurrence records, defined as geographic coordinates of accession collection sites, were obtained from the US National Plant Germplasm System (NPGS) database (https://npgsweb.ars-grin.gov/gringlobal/) and used to build models. The 19 bioclimatic variables were used as environmental predictors. For model calibration, 70% of the occurrence records were randomly selected for training and the remaining 30% for model evaluation. Each model was run 100 times, and performance was evaluated using the true skill statistic (TSS)[87]. The final models were then used to project habitat suitability under current and future climate scenarios for mid-century (2040–2060) and late-century (2080–2100).

Estimating genetic offset under future climates

Genetic offset under future climate scenarios was estimated using Gradient Forest modeling[36]. Climate-associated SNPs and SVs identified from genome–environment association analyses were used as response variables, and bioclimatic variables were used as environmental predictors. Gradient Forest modeling was conducted using the R package gradientForest (https://gradientforest.r-forge.r-project.org/). The fitted models were then used to predict genomic turnover under both current and future climatic conditions. Future climate layers representing mid-century (2040–2060) and late-century (2080–2100) conditions were projected onto the Gradient Forest models to obtain predicted genomic compositions. Genetic offset was calculated as the Euclidean distance between predicted genomic compositions under current and future climates. Spatial patterns of genetic offset were visualized using the R package rasterVis (https://cran.r-project.org/web/packages/rasterVis/).

Data availability

Raw genome resequencing reads have been deposited in the NCBI BioProject database under the accession number PRJNA1422823. Raw HiFi reads and genome assemblies have been deposited in the NCBI Bioproject database under the accession number PRJNA1209910. Genome assemblies and annotations, SNPs, small indels and SVs in VCF format are available at CuGenDBv2 (http://cucurbitgenomics.org/v2/ftp/pan-genome/squash/).

Author contributions

Z.F., Y.X. and C.W. conceived, designed and supervised the study. X.Z. and H.S. contributed to genome assembly and annotation, pangenome construction, and SV genotyping. X.Z. performed population genetic analyses and genome–environment association analyses. K.B., J.Z., A.A.S., A.F., H.S.P., E.O., A.G., R.G., C.W. and Y.X. contributed to sample collection, DNA extraction, and genome sequencing. X.Z. wrote the manuscript. Z.F. revised the manuscript.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgements

The authors thank Michael Mazourek for providing seeds of the *C. pepo* core collection and for critical reading of this manuscript. This research was supported by grants from USDA National Institute of Food and Agriculture Specialty Crop Research Initiative (2015-51181-24285 and 2020-51181-32139).

bioRxiv preprint doi: https://doi.org/10.64898/2026.05.27.728270; this version posted May 29, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

References

1. Clark, D. *et al.* Accelerating extinction risk from climate change. *Science* **348**, 571–573 (2015).
2. Hickey, L. T. *et al.* Breeding crops to feed 10 billion. *Nat. Biotechnol.* **37**, 744–754 (2019).
3. Bailey-Serres, J., Parker, J. E., Ainsworth, E. A., Oldroyd, G. E. D. & Schroeder, J. I. Genetic strategies for improving crop yields. *Nature* **575**, 109–118 (2019).
4. Aguirre-Liguori, J. A., Ramírez-Barahona, S. & Gaut, B. S. The evolutionary genomics of species' responses to climate change. *Nat. Ecol. Evol.* **5**, 1350–1360 (2021).
5. Bernatchez, L., Ferchaud, A. L., Berger, C. S., Venney, C. J. & Xuereb, A. Genomics for monitoring and understanding species responses to global climate change. *Nat. Rev. Genet.* **25**, 165–183 (2024).
6. Alkan, C., Coe, B. P. & Eichler, E. E. Genome structural variation discovery and genotyping. *Nat. Rev. Genet.* **12**, 363–376 (2011).
7. Ho, S. S., Urban, A. E. & Mills, R. E. Structural variation in the sequencing era. *Nat. Rev. Genet.* **21**, 171–189 (2020).
8. Li, H. & Durbin, R. Genome assembly in the telomere-to-telomere era. *Nat. Rev. Genet.* **25**, 658–670 (2024).
9. Schreiber, M., Jayakodi, M., Stein, N. & Mascher, M. Plant pangenomes for crop improvement, biodiversity and evolution. *Nat. Rev. Genet.* **25**, 563–577 (2024).
10. Smith, B. D. The initial domestication of *Cucurbita pepo* in the Americas 10,000 years ago. *Science* **276**, 932–934 (1997).
11. Castellanos-Morales, G. *et al.* Tracing back the origin of pumpkins (*Cucurbita pepo ssp. pepo* L.) in Mexico. *Proc. R. Soc. B Biol. Sci.* **286**, 20191440 (2019).
12. Chomicki, G., Schaefer, H. & Renner, S. S. Origin and domestication of Cucurbitaceae crops: insights from phylogenies, genomics and archaeology. *New Phytol.* **226**, 1240–1255 (2020).
13. Paris, H. S., Lebeda, A., Křistkova, E., Andres, T. C. & Nee, M. H. Parallel evolution under domestication and phenotypic differentiation of the cultivated subspecies of *Cucurbita pepo* (Cucurbitaceae). *Econ. Bot.* **66**, 71–90 (2012).
14. Kates, H. R., Soltis, P. S. & Soltis, D. E. Evolutionary and domestication history of *Cucurbita* (pumpkin and squash) species inferred from 44 nuclear loci. *Mol. Phylogenet. Evol.* **111**, 98–109 (2017).
15. Dhatt, A. S. *et al.* Comprehensive review of pumpkin (*Cucurbita* spp.): Domestication, global distribution, genetic characterization, breeding strategies, and genomic insights. *Veg. Sci.* **51**, 196–210 (2024).
16. Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A. & Zdobnov, E. M. BUSCO update: Novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol. Biol. Evol.* **38**, 4647–4654 (2021).
17. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: Reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 245 (2020).
18. Ou, S., Chen, J. & Jiang, N. Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Res.* **46**, e126 (2018).
19. Liao, W. W. *et al.* A draft human pangenome reference. *Nature* **617**, 312–324 (2023).
20. Hernandez, C. O. *et al.* Characterization of the USDA *Cucurbita pepo*, *C. moschata*, and *C. maxima* germplasm collections. *Front. Plant Sci.* **14**, 1130814 (2023).
21. Ebler, J. *et al.* Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes. *Nat. Genet.* **54**, 518–525 (2022).
22. Patterson, N. *et al.* Ancient admixture in human history. *Genetics* **192**, 1065–1093 (2012).
23. Terhorst, J., Kamm, J. A. & Song, Y. S. Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nat. Genet.* **49**, 303–309 (2017).
24. Kamm, J., Terhorst, J., Durbin, R. & Song, Y. S. Efficiently inferring the demographic history of many populations with allele count data. *J. Am. Stat. Assoc.* **115**, 1472–1487 (2020).
25. Purugganan, M. D. & Fuller, D. Q. The nature of selection during plant domestication. *Nature* **457**, 843–848 (2009).
26. Casillas, S. & Barbadilla, A. Molecular population genetics. *Genetics* **205**, 1003–1035 (2017).

bioRxiv preprint doi: https://doi.org/10.64898/2026.05.27.728270; this version posted May 29, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

27. Zhao, X. *et al.* Graph-based pangenome reveals structural variation dynamics during cucumber breeding. *Nat. Genet.* **58**, 643–654 (2026).
28. Gutaker, R. M. *et al.* Genomic history and ecology of the geographic spread of rice. *Nat. Plants* **6**, 492–502 (2020).
29. Haupt, M. & Schmid, K. Using landscape genomics to infer genomic regions involved in environmental adaptation of soybean genebank accessions. *BMC Plant Biol.* **25**, 1175 (2025).
30. Lasky, J. R. *et al.* Genome-environment associations in sorghum landraces predict adaptive traits. *Sci. Adv.* **1**, e1400218 (2015).
31. Russell, J. *et al.* Exome sequencing of geographically diverse barley landraces and wild relatives gives insights into environmental adaptation. *Nat. Genet.* **48**, 1024–1030 (2016).
32. Li, F. *et al.* Environmental data provide marginal benefit for predicting climate adaptation. *PLOS Genet.* **21**, e1011714 (2025).
33. Van Den Wollenberg, A. L. Redundancy analysis an alternative for canonical correlation analysis. *Psychometrika* **42**, 207–219 (1977).
34. Dray, S., Legendre, P. & Peres-Neto, P. R. Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbour matrices (PCNM). *Ecol. Model.* **196**, 483–493 (2006).
35. Günther, T. & Coop, G. Robust identification of local adaptation from allele frequencies. *Genetics* **195**, 205–220 (2013).
36. Ellis, N., Smith, S. J. & Roland Pitcher, C. Gradient forests: Calculating importance gradients on physical predictors. *Ecology* **93**, 156–168 (2012).
37. Jadamba, C., Kang, K., Paek, N.-C., Lee, S. I. & Yoo, S.-C. Overexpression of rice *Expansin7* (*Osexpa7*) confers enhanced tolerance to salt stress in rice. *Int. J. Mol. Sci.* **21**, 454 (2020).
38. Elith, J. & Leathwick, J. R. Species distribution models: Ecological explanation and prediction across space and time. *Annu. Rev. Ecol. Evol. Syst.* **40**, 677–697 (2009).
39. Ferrer Obiol, J. *et al.* Evolutionarily distinct lineages of a migratory bird of prey show divergent responses to climate change. *Nat. Commun.* **16**, 3503 (2025).
40. Jiang, J. *et al.* Incorporating genetic load contributes to predicting *Arabidopsis thaliana*'s response to climate change. *Nat. Commun.* **16**, 2752 (2025).
41. Doyle, J. J. & Doyle, J. L. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem. Bull.* **19**, 11–15 (1987).
42. Sim, S. B., Corpuz, R. L., Simmonds, T. J. & Geib, S. M. HiFiAdapterFilt, a memory efficient read processing pipeline, prevents occurrence of adapter sequence in PacBio HiFi reads and their negative impacts on genome assembly. *BMC Genomics* **23**, 157 (2022).
43. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).
44. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
45. Esteras, C. *et al.* High-throughput SNP genotyping in *Cucurbita pepo* for map construction and quantitative trait loci mapping. *BMC Genomics* **13**, 80 (2012).
46. Montero-Pau, J. *et al.* An SNP-based saturated genetic map and QTL analysis of fruit-related traits in Zucchini using Genotyping-by-sequencing. *BMC Genomics* **18**, 94 (2017).
47. Michael, V. N., Fu, Y., Shrestha, S. & Meru, G. A novel QTL for resistance to *Phytophthora crown* rot in squash. *Plants* **10**, 2115 (2021).
48. Dudchenko, O. *et al.* De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
49. Montero-Pau, J. *et al.* De novo assembly of the zucchini genome reveals a whole-genome duplication associated with the origin of the *Cucurbita* genus. *Plant Biotechnol. J.* **16**, 1161–1171 (2018).
50. Alonge, M. *et al.* Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. *Genome Biol.* **23**, 258 (2022).
51. Flynn, J. M. *et al.* RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 9451–9457 (2020).

bioRxiv preprint doi: https://doi.org/10.64898/2026.05.27.728270; this version posted May 29, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

52. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinforma.* **Chapter 4,** 4.10.1-4.10.14 (2009).
53. Campbell, M. S., Holt, C., Moore, B. & Yandell, M. Genome annotation and curation using MAKER and MAKER-P. *Curr. Protoc. Bioinforma.* **12**, 11–39 (2014).
54. Keller, O., Kollmar, M., Stanke, M. & Waack, S. A novel hybrid gene prediction method employing protein multiple sequence alignments. *Bioinformatics* **27**, 757–763 (2011).
55. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004).
56. Castanera, R., Ruggieri, V., Pujol, M., Garcia-Mas, J. & Casacuberta, J. M. An improved melon reference genome with single-molecule sequencing uncovers a recent burst of transposable elements with potential impact on genes. *Front. Plant Sci.* **10**, 1815 (2020).
57. Li, Q. *et al.* A chromosome-scale genome assembly of cucumber (*Cucumis sativus* L.). *GigaScience* **8**, giz072 (2019).
58. Renner, S. S. *et al.* A chromosome-level genome of a Kordofan melon illuminates the origin of domesticated watermelons. *Proc. Natl. Acad. Sci. U. S. A.* **118**, e2101486118 (2021).
59. Xie, D. *et al.* The wax gourd genomes offer insights into the genetic diversity and ancestral cucurbit karyotype. *Nat. Commun.* **10**, 5158 (2019).
60. Ma, L. *et al.* The genome and transcriptome analysis of snake gourd provide insights into its evolution and fruit development and ripening. *Hortic. Res.* **7**, 199 (2020).
61. Fu, A. *et al.* Combined genomic, transcriptomic, and metabolomic analyses provide insights into chayote (*Sechium edule*) evolution and fruit development. *Hortic. Res.* **8**, 35 (2021).
62. Sun, H. *et al.* Karyotype stability and unbiased fractionation in the paleo-allotetraploid *Cucurbita* genomes. *Mol. Plant* **10**, 1293–1306 (2017).
63. Iwata, H. & Gotoh, O. Benchmarking spliced alignment programs including Spaln2, an extended version of Spaln that incorporates additional species-specific features. *Nucleic Acids Res.* **40**, e161 (2012).
64. Shumate, A. & Salzberg, S. L. Liftoff: Accurate mapping of gene annotations. *Bioinformatics* **37**, 1639–1643 (2021).
65. Holst, F. *et al.* Helixer: ab initio prediction of primary eukaryotic gene models combining deep learning and a hidden Markov model. *Nat. Methods* **23**, 732–739 (2025).
66. Haas, B. J. *et al.* Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, R7 (2008).
67. Emms, D. M. & Kelly, S. OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
68. Hickey, G. *et al.* Pangenome graph construction from genome alignments with Minigraph-Cactus. *Nat. Biotechnol.* **42**, 663–673 (2024).
69. Garrison, E. *et al.* Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat. Biotechnol.* **36**, 875–881 (2018).
70. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
71. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* https://doi.org/10.48550/arXiv.1303.3997 (2013).
72. McKenna, A. *et al.* The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
73. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly (Austin)* **6**, 80–92 (2012).
74. Stamatakis, A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
75. Guo, S. *et al.* Resequencing of 414 cultivated and wild watermelon accessions identifies selection for fruit quality traits. *Nat. Genet.* **51**, 1616–1623 (2019).
76. Letunic, I. & Bork, P. Interactive tree of life (iTOL) v3: An online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.* **44**, 242–245 (2016).

bioRxiv preprint doi: https://doi.org/10.64898/2026.05.27.728270; this version posted May 29, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

77. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
78. Purcell, S. *et al.* PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
79. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
80. Ossowski, S. *et al.* The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science* **327**, 92–94 (2010).
81. Máire Ní, L. *et al.* The evolutionary history of dogs in the Americas. *Science* **85**, 81–85 (2018).
82. Leppälä, K., Nielsen, S. V. & Mailund, T. Admixturegraph: An R package for admixture graph manipulation and fitting. *Bioinformatics* **33**, 1738–1740 (2017).
83. Albers, P. K. & McVean, G. Dating genomic variants and shared ancestry in population-scale sequencing data. *PLoS Biol.* **18**, e3000586 (2020).
84. Barrera-Redondo, J. *et al.* The domestication of *Cucurbita argyrosperma* as revealed by the genome of its wild relative. *Hortic. Res.* **8**, 109 (2021).
85. Tataru, P. & Bataillon, T. PolyDFEv2.0: Testing for invariance of the distribution of fitness effects within and across species. *Bioinformatics* **35**, 2868–2869 (2019).
86. van Vuuren, D. P. *et al.* The representative concentration pathways: An overview. *Clim. Change* **109**, 5–31 (2011).
87. Allouche, O., Tsoar, A. & Kadmon, R. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). *J. Appl. Ecol.* **43**, 1223–1232 (2006).