nature genetics

Article

https://doi.org/10.1038/s41588-023-01302-4

Pangenomic analysis identifies structural variation associated with heat tolerance in pearl millet

Received: 8 December 2021

Accepted: 18 January 2023

Published online: 2 March 2023

Check for updates

Haidong Yan[1,2,3,17], Min Sun[1,17], Zhongren Zhang[4,17], Yarong Jin[1,17], Ailing Zhang[1], Chuang Lin[1], Bingchao Wu[1], Min He[5], Bin Xu[6], Jing Wang[7], Peng Qin[8], John Pablo Mendieta[3], Gang Nie[1], Jianping Wang[9], Chris S. Jones[10], Guangyan Feng[1], Rakesh K. Srivastava[11], Xinquan Zhang[1], Aureliano Bombarely[12], Dan Luo[1], Long Jin[13], Yuanying Peng[14], Xiaoshan Wang[1], Yang Ji[15], Shilin Tian[4,16] ✉ & Linkai Huang[1,5] ✉

Pearl millet is an important cereal crop worldwide and shows superior heat tolerance. Here, we developed a graph-based pan-genome by assembling ten chromosomal genomes with one existing assembly adapted to different climates worldwide and captured 424,085 genomic structural variations (SVs). Comparative genomics and transcriptomics analyses revealed the expansion of the RWP-RK transcription factor family and the involvement of endoplasmic reticulum (ER)-related genes in heat tolerance. The overexpression of one *RWP-RK* gene led to enhanced plant heat tolerance and transactivated ER-related genes quickly, supporting the important roles of RWP-RK transcription factors and ER system in heat tolerance. Furthermore, we found that some SVs affected the gene expression associated with heat tolerance and SVs surrounding ER-related genes shaped adaptation to heat tolerance during domestication in the population. Our study provides a comprehensive genomic resource revealing insights into heat tolerance and laying a foundation for generating more robust crops under the changing climate.

Global warming has severely affected crop productivity, which seriously threatens world food security[1]. The change in temperature from the historical average in 1900 is expected to exceed 2 °C by the end of the twenty-first century[2]. With every 1 °C increase in the global average temperature, wheat (*Triticum aestivum*) production is estimated to decrease by 6%, rice (*Oryza sativa*) production is estimated to decrease by 3.2% and corn (*Zea mays*) production is estimated to decrease by 7.4%[3]. Therefore, an understanding of heat tolerance in plants is

[1]College of Grassland Science and Technology, Sichuan Agricultural University, Chengdu, China. [2]School of Plant and Environmental Sciences, Virginia Tech, Blacksburg, VA, USA. [3]Department of Genetics, University of Georgia, Athens, GA, USA. [4]Novogene Bioinformatics Institute, Beijing, China. [5]State Key Laboratory of Crop Gene Exploration and Utilization in Southwest China, Sichuan Agricultural University, Chengdu, China. [6]College of Grassland Science, Nanjing Agricultural University, Nanjing, China. [7]Key Laboratory of Bio-Source and Environmental Conservation, School of Life Science, Sichuan University, Chengdu, China. [8]Rice Research Institute, Sichuan Agricultural University, Chengdu, China. [9]Agronomy Department, University of Florida, Gainesville, FL, USA. [10]Feed and Forage Development, International Livestock Research Institute, Nairobi, Kenya. [11]International Crops Research Institute for the Semi-Arid Tropics, Hyderabad, India. [12]Instituto de Biologia Molecular y Celular de Plantas, UPV-CSIC, Valencia, Spain. [13]College of Animal Science and Technology, Sichuan Agricultural University, Chengdu, China. [14]Triticeae Research Institute, Sichuan Agricultural University, Chengdu, China. [15]Sichuan Animal Science Academy, Chengdu, China. [16]Department of Ecology, Hubei Key Laboratory of Cell Homeostasis, College of Life Sciences, Wuhan University, Wuhan, China. [17]These authors contributed equally: Haidong Yan, Min Sun, Zhongren Zhang, Yarong Jin. ✉e-mail: tianshilin@novogene.com; huanglinkai@sicau.edu.cn

urgently required to develop crops that can withstand rising global temperatures and could thus be used to maximize agricultural production to help satisfy the food demands of an increasing population.

Pearl millet (*Pennisetum glaucum* (L.) R. Br., syn. *Cenchrus americanus* (L.) Morrone) ($2n = 2x = 14$) is a C_4 cereal crop that is important in safeguarding the security of food and forage in the arid and semiarid tropics due to its superior tolerance to high temperatures[4–8]. It is also a staple food of more than 90 million farmers living in poverty and is grown on more than 31.2 million hectares[9]. Pearl millet is an ideal model for understanding how plants use heat-related genes and mechanisms to thrive at warmer temperatures. However, few studies have investigated the molecular mechanisms underlying the regulation of heat stress responses (HSRs) in pearl millet relative to other major crops[10,11] and the underlying mechanisms are not well understood.

Recent studies revealed that many genes involved in environmental stress responses are strongly affected by structural variations (SVs)[12–14]; however, the causal relationship of SVs with HSRs is poorly understood. SVs have roles in gene expression alterations linked to important plant phenotypes[15]. However, the detection of SVs is challenging when relying on short-read sequencing data[16,17]. This challenge has promoted the development of new approaches for SV detection using graph-based pan-genomes that are based on multiple high-quality assemblies[17–19]. Therefore, building graph-based pan-genomic resources has the potential to advance the characterization and understanding of the biological impact of SVs on phenotypic variations and accelerate the breeding of pearl millet.

In this study, we generated de novo genome assemblies of ten pearl millet accessions and constructed a graph-based pan-genome assembly to identify genomic SVs. We leveraged SVs, transcriptomics and in vivo validation to reveal the relationship between SVs and gene expression under heat stress conditions. With this approach, we identified SVs that contributed to heat adaptation during crop domestication. By integrating multi-omics analyses, we suggested a possible mechanism in which the resistance of pearl millet to heat stress depends mainly on the endoplasmic reticulum (ER) and validated an RWP-RK (https://www.ebi.ac.uk/interpro/entry/pfam/PF02042/) transcription factor as a positive coregulator of heat tolerance along with the ER pathway. Our findings advance the conceptual understanding of heat tolerance in pearl millet, promise to expedite genomics-assisted breeding for heat tolerance in this important crop and will benefit comparative and functional genomics studies of other crops.

Results

Genome assembly and pan-genomic analysis of representative pearl millet accessions

We selected ten representative accessions from eight major geographical regions based on the phylogenetic relationships of a 394-line core collection of pearl millet[7,20] (Fig. 1a,b, Supplementary Figs. 1 and 2 and Supplementary Table 1). We assembled their chromosome-level genomes by integrating PacBio high-fidelity (HiFi) long-read sequences, Bionano optical mapping data, high-throughput chromosome conformation capture (Hi-C) data and Illumina paired-end sequences (Fig. 1a,b, Extended Data Fig. 1, Supplementary Table 2 and Supplementary Note 1). These genomes ranged in size from 1.89 Gb to 2.00 Gb, with scaffold N50 values ranging from 193.80 Mb to 286.98 Mb, corresponding to 95.85–99.47% of the genome sizes estimated by *k*-mer analysis (1.97–2.01 Gb), which is consistent with the genome sizes predicted by flow cytometry (Extended Data Fig. 2). The contig N50 values were substantially increased from 155 to 3,959-fold over those of the previously published pearl millet reference genome[7] (Table 1 and Supplementary Table 2).

To measure the quality of these ten newly assembled genomes, we realigned high-quality paired-end reads against the assemblies and observed alignment rates ranging from 95.62% to 99.57%, covering 94.92–99.90% of the genomes (Supplementary Table 2). Additionally, more than 91.60% of the embryophyte Benchmarking Universal Single-Copy Orthologs (BUSCOs) were present in each genome (Supplementary Table 2). The long terminal repeat (LTR) assembly index (LAI) scores all exceeded 24 and thus met the criterion standard[21] (Table 1). Further evaluation using Merqury showed a quality value (QV) over 40 for our ten assemblies, which exceeded the Vertebrate Genomes Project standard of QV40[22] (Supplementary Table 2). These results demonstrate the accuracy, completeness and contiguity of the ten pearl millet genome assemblies. In addition, we predicted an average of 36,847 gene models for each assembly, among which more than 99.30% showed matches with the known functional database (Supplementary Table 2). Transposable elements (TEs) constituted 71.58% of each genome, ranging from 70.44% to 72.62% (Supplementary Tables 2–4 and Supplementary Note 1).

We constructed a pan-genome using 11 pearl millet assemblies, including the previously released genome[7]. Among the total gene family sets, 14,608 core gene families were obtained across all accessions, accounting for more than half (46.60–52.08%) of the total sets; dispensable families (39.75–49.94%), in which genes were present in 2–10 accessions, constituted the second-largest proportion. The smallest proportion consisted of private gene sets, which were only detected in one genome and accounted for 0.73–8.73% of the total sets (Fig. 1c).

To further evaluate the representativeness of the pan-genome, we compared the distribution of SNPs between the 11 accessions and the aforementioned 394 core lines. They displayed a similar pattern across the genome and showed strong significant correlations in SNP density, nucleotide diversity (π) and synonymous (d_S) and nonsynonymous (d_N) substitution rates (SNP density, *rho* = 0.95; π, *rho* = 0.89; d_S, *rho* = 0.98; d_N, *rho* = 0.98) (Extended Data Fig. 3a,b). The number of added gene families declined quickly, with only 301 (0.64% of all gene families; 301 out of 47,344) additional gene families being identified when the eleventh accession was included (Extended Data Fig. 3c,d). Moreover, the accessions used to generate the pan-genome showed a similar Shannon's diversity index (H) and π to the 394 accessions (H: 8.07782 versus 8.03436; π: 0.0001327 versus 0.0001209). In general, these results suggest that the pan-genome accessions are genetically diverse and representative of the diversities of the pearl millet population. We further observed that core genes were more functionally conserved and enriched in general biological processes than the dispensable and private genes, as with previous findings in other plants[17,23,24] (Fig. 1d, Extended Data Fig. 4 and Supplementary Note 2). In total, we built a high-quality pan-genome resource that will contribute to pearl millet improvement.

Graph-based genome and SV identification

A total of 744,364 SVs were identified by realigning the assemblies against the PI537069 reference genome as this accession comes from the geographical origin (Northwest Africa) of pearl millet[25] and has a relatively high assembly quality (Table 1 and Supplementary Table 2). These SVs included 622,584 presence and absence variations (PAVs) consisting of 306,679 presence and 315,905 absence cases, 2,177 inversions (INVs), 91,852 copy number variations (CNVs) and 27,751 translocations (TRANS) (Fig. 2a and Supplementary Table 5). Approximately 37.94% of PAVs were less than 2 kb in length, INVs (68.11%) were concentrated within 100 kb, CNVs (62.53%) were enriched in the size range of less than 4 kb and most TRANS (91.10%) were less than 20 kb in length (Extended Data Fig. 5a).

To build the graph-based genome, the SVs from all the pearl millet accessions were merged to yield 424,085 non-redundant SVs. PAVs accounted for 74.70% of private SVs present in only one accession but constituted a relatively high proportion (87.51%) of the non-private SVs. Similar trends were observed for CNVs and TRANS (Fig. 2b). We observed that the SVs were enriched in repeat regions (Fig. 2c). Across these genomes, 37–44% of SVs overlapped with genic and flanking regions (5 kb) (Fig. 2d), suggesting potential roles of SVs in gene regulation. In addition, the SVs and graph-based genome were validated by

https://doi.org/10.1038/s41588-023-01302-4

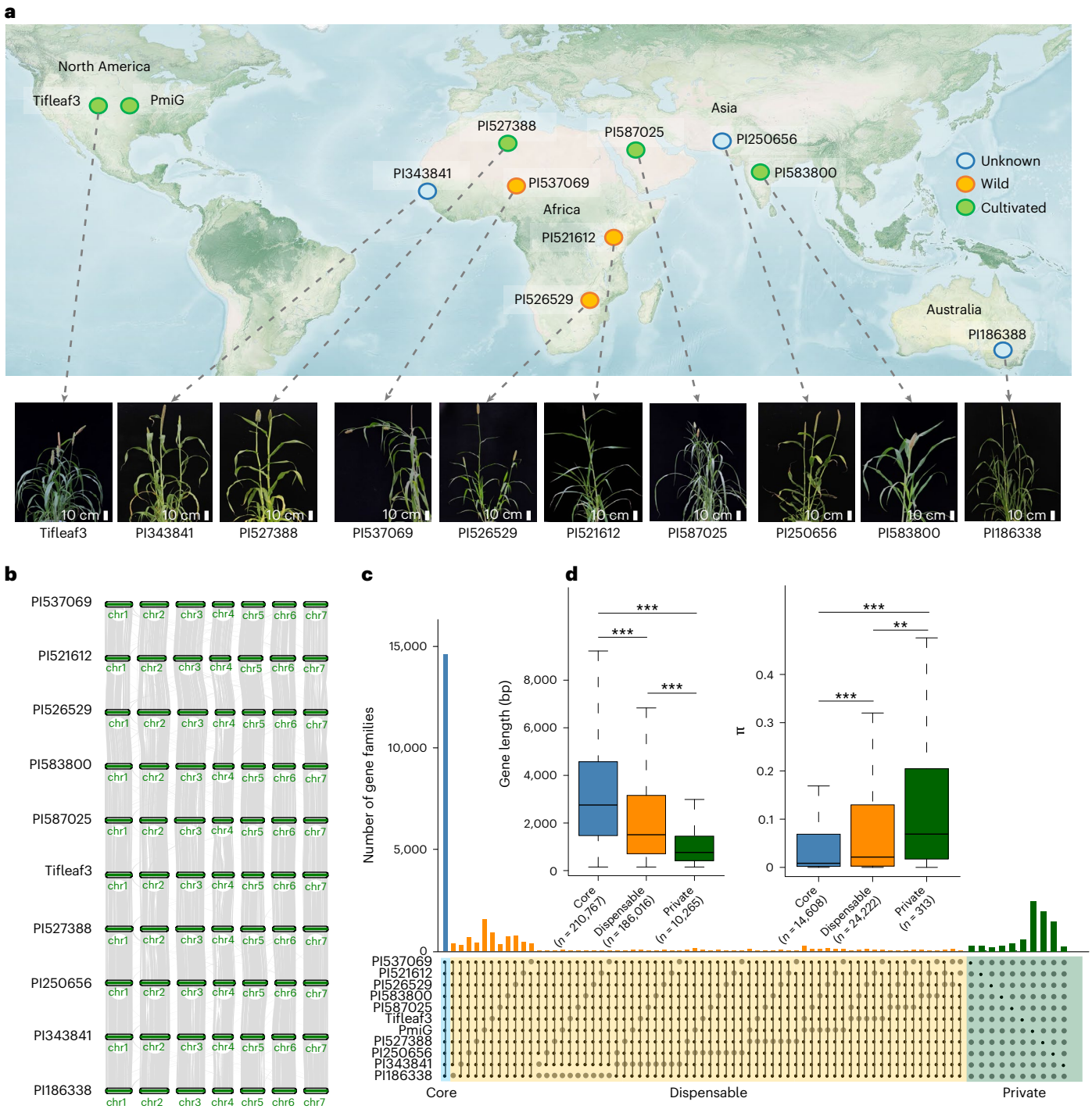


Fig. 1 | Ten high-quality assembled genomes and pan-genome construction in pearl millet. a, The pearl millet accessions are derived from geographically representative regions (PmiG: Tift 23D2B1-P1-P5). The geographical map was adapted from the one provided by the NASA Earth Observatory (https://visibleearth.nasa.gov/images/147190/explorer-base-map/147191w/). **b**, Synteny plot across the ten genomes. **c**, Core gene clusters and pan-genome of pearl millet. The histogram illustrates the core gene clusters (present in all genomes), dispensable gene clusters (present in 2–10 genomes) and private gene clusters (present in one genome). **d**, Composition of gene and nucleotide diversity (*pi*) in core, dispensable and private genes. The center line represents the median; the box limits represent the upper and lower quartiles; the whiskers represent 1.5 times the interquartile range (IQR). Significant differences were tested by two-tailed *t*-test (**$P < 0.01$, ***$P < 0.0005$).

evaluating the performances of different SV calling tools, by conducting PCR and checking read coverage over the possible variant paths (Extended Data Fig. 5b, Supplementary Tables 6–8 and Supplementary Note 3). Overall, this graph-based pan-genome is an essential genomic resource supporting the study of SVs and will provide a prominent reference for the discovery of SVs in pearl millet populations.

Expansion of the RWP-RK transcription factor family contributes to heat tolerance

Pearl millet was shown to be very tolerant to high-temperature conditions based on our phenotypic and physiological data (Fig. 3a). In particular, the leaves of pearl millet seedlings only showed wilting after 21 d of heat treatment (40 °C in light, 35 °C in darkness) (Extended Data Fig. 6a).

https://doi.org/10.1038/s41588-023-01302-4

Table 1 | Summary of genome assembly and annotation

Accession no.	Contig N50 (Mb)	Scaffold N50 (Mb)	Contig length (Mb)	Scaffold length (Mb)	Chromosome anchoring rate (%)	Repeat ratio (%)	Gene no.	LAI
PI537069	61.62	266.84	1,908.34	1,913.80	96.68	71.58	35,486	27.90
PI521612	5.40	278.46	1,891.01	1,891.08	95.98	70.44	37,906	26.15
PI587025	5.15	257.50	1,911.08	1,911.21	94.39	71.58	38,076	27.38
PI583800	3.10	261.45	1,937.87	1,937.98	97.52	72.21	35,826	27.53
Tifleaf3	25.57	279.17	1,950.21	1,950.23	95.00	71.30	37,280	26.22
PI526529	79.18	286.98	1,974.39	1,974.39	98.48	71.88	36,451	26.53
PI186338	3.80	284.64	1,999.44	1,999.53	95.30	72.63	36,343	26.47
PI343841	5.10	263.66	1,962.06	1,962.24	94.23	72.17	36,312	26.76
PI527388	3.10	193.80	1,937.79	1,938.01	94.51	71.02	37,866	27.79
PI250656	4.20	276.63	1,895.51	1,895.77	95.11	70.72	36,923	24.74
Tift 23D2B1-P1-P5/PmiG[7]	0.02	0.88	1,556.18	1,793.24	NA	77.20	38,579	2.09

NA, not applicable.

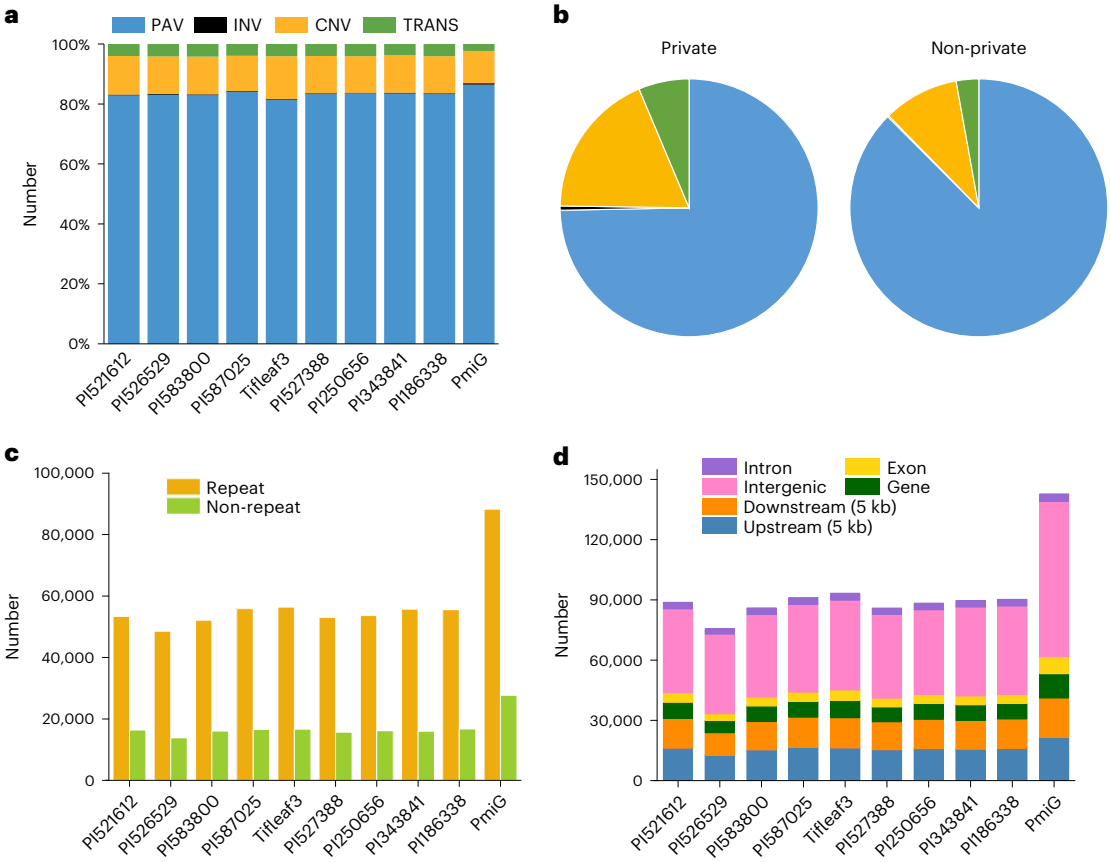


Fig. 2 | Identification of SVs. a, Composition of SVs in each genome. **b**, Comparisons of the proportions of private (present in only one accession) and non-private SVs (present in at least two accessions). **c**, Numbers of SVs in the repeat and non-repeat regions of each genome. **d**, Numbers of SVs overlapping with different genomic features in each genome.

The relative water content, relative electrical conductivity (REC) and malondialdehyde (MDA) content did not change significantly ($P > 0.05$) until 21 d of heat treatment (Extended Data Fig. 6b,c), while in maize leaves, the relative water content decreased and the MDA content increased significantly under 4 h of heat stress (40 °C)[26]. The slower responses might indicate better heat tolerance in pearl millet than in maize.

To dissect the molecular mechanism underlying heat tolerance in pearl millet, we first conducted comparative genomic analyses, which revealed that expanded, positively selected and species-specific gene families, as well as genes located near recently expanded LTR TEs (LTRs) were enriched in stress-related pathways in pearl millet (Extended Data Figs. 2g and 7a,b and Supplementary Note 4). Notably, one transcription factor family (RWP-RK) was identified as expanding

https://doi.org/10.1038/s41588-023-01302-4

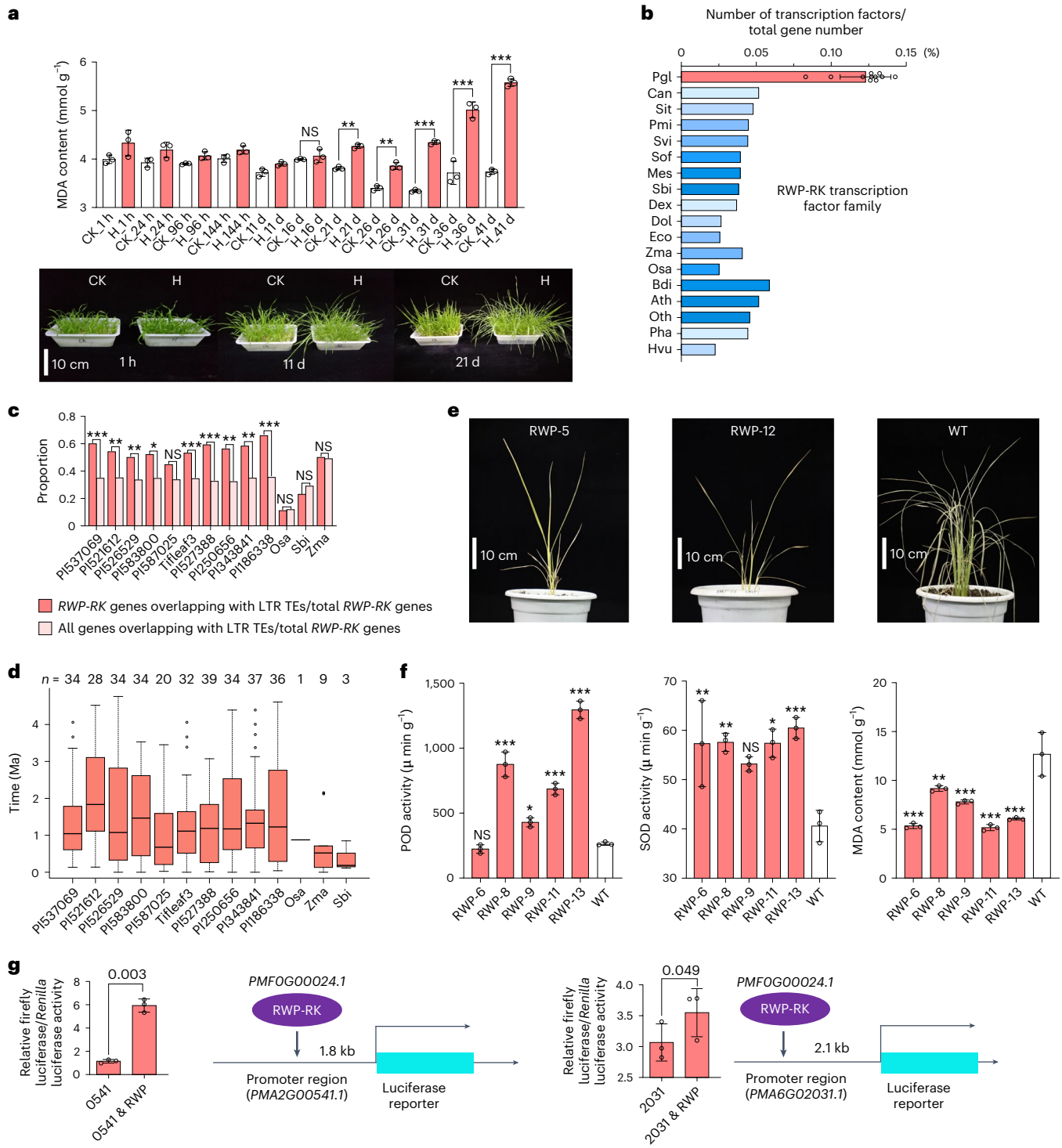


Fig. 3 | Expansion of the RWP-RK transcription factor family contributes to heat tolerance. a, Comparison of MDA levels between the control (CK) and heat treatment (H) groups of pearl millet (Tifleaf3). The error bars indicate the mean ± s.d.; $n = 3$ biological replicates. Significant differences were tested by two-tailed t-test (**$P < 0.01$, ***$P < 0.0005$; NS, not significant). **b**, Proportion of RWP-RK transcription factor family members among all the genes in the 11 pearl millet genomes and ten other genomes. Ath, *Arabidopsis thaliana*; Bdi, *Brachypodium distachyon*; Can, *Capsicum annuum*; Dex, *Digitaria exilis*; Dol, *Dichanthelium oligosanthes*; Eco, *Eleusine coracana*; Hvu, *Hordeum vulgare*; Mes, *Manihot esculenta*; Osa, *O. sativa*; Oth, *Oropetium thomaeum*; Pgl, *P. glaucum* (pearl millet); Pha, *Panicum hallii*; Pmi, *Panicum miliaceum*; Sbi, *Sorghum bicolor*; Sit, *Setaria italica*; Sof, *Saccharum officinarum*; Svi, *Setaria viridis*; Zma, *Z. mays*. **c**, Comparisons of *RWP-RK* gene numbers overlapping with intact LTR TEs among pearl millet, rice, sorghum and maize. Significant differences were tested by one-tailed binomial test (*$P < 0.05$, **$P < 0.01$, ***$P < 0.0005$). **d**, Estimated insertion times of LTR TEs encompassing the *RWP-RK* genes shown in **c**. The center line represents the median; the box limits represent the upper and lower quartiles; the whiskers represent 1.5 times the IQR; the dots represent the outliers. **e**, Phenotypes of two transgenic lines and a control (WT) after 72 h of heat treatment. **f**, POD and SOD activity after 12 h of heat treatment and MDA content after 72 h of heat treatment in transgenic lines and WT plants. The error bars represent the mean ± s.d.; $n = 3$ biological replicates. Significant differences were tested using a two-tailed t-test (*$P < 0.05$, **$P < 0.01$, ***$P < 0.0005$). **g**, Dual luciferase assays (firefly luciferase to *Renilla* luciferase ratio) were applied to verify that *PMF0G00024.1* (RWP) could transactivate *PMA2G00541.1* (541) and *PMA6G02031.1* (2031). The error bars represent the mean ± s.d.; $n = 3$ biological replicates. Significant differences were tested using a two-tailed t-test and are shown as P values.

in the genomes of the 11 pearl millet accessions (Fig. 3b, Supplementary Fig. 3 and Supplementary Table 9). This family responded to biotic or abiotic stresses[27–30], supporting the potential roles of its members in heat tolerance. We investigated LTRs located near the *RWP-RK* genes and found that early LTR expansion might be associated with RWP-RK transcription factor family expansion and probably caused increases in specific *RWP-RK* genes in pearl millet (Fig. 3c,d, Extended Data Fig. 7c, Supplementary Fig. 3 and Supplementary Note 5).

To further characterize the roles of *RWP-RK* genes in response to heat stress, we sequenced leaf and root transcriptomes after high-temperature treatment (Supplementary Table 3). A total of ten differentially expressed *RWP-RK* genes were predicted, including two specific and eight nonspecific transcription factors (Extended Data Fig. 7d, Supplementary Table 10 and Supplementary Note 5). When overexpressing an *RWP-RK* (*PMF0G00024.1*) in rice, we found that the leaves of the transgenic lines (*RWP-RKox*) were less withered than the leaves of wild-type (WT) plants under high temperature (Fig. 3e and Extended Data Fig. 7e). The *RWP-RKox* plants showed significantly higher peroxidase (POD) and superoxide dismutase (SOD) activities and lower MDA contents after exposure to heat stress conditions than the WT plants (Fig. 3f), which provides a potential avenue for the future molecular breeding of heat-tolerant crops. We also characterized this RWP-RK transcription factor in a coregulated network and used a dual luciferase assays to verify that this transcription factor could transactivate two stress-related genes, *PMA2G00541.1* and *PMA6G02031.1* (Fig. 3g, Supplementary Table 11 and Supplementary Note 5). Taken together, these results indicate that the expansion of the RWP-RK transcription factor family has potentially contributed to heat tolerance in pearl millet.

RWP-RK coregulates a fast heat response with ER-related genes

To further dissect the molecular mechanism underlying heat tolerance in pearl millet, we sequenced the leaf and root transcriptomes of Tifleaf3 under high-temperature treatments at eight time points (dataset A) and selected six accessions to perform leaf transcriptome sequencing under stress for 1 and 24 h (dataset B; Supplementary Table 3). Based on gene functional enrichment analyses, the two transcriptome datasets revealed differentially expressed genes (DEGs) that were enriched mainly in ER-related pathways involved in the repair and elimination of misfolded proteins (Fig. 4a, Extended Data Fig. 8a,b, Supplementary Table 12 and Supplementary Note 6.1). We analyzed the RNA sequencing (RNA-seq) data from maize[31] and rice[32] and identified greater proportions of upregulated ER-related and heat shock factor (HSF) (https://www.ebi.ac.uk/interpro/entry/pfam/PF00447/) genes in pearl millet than in these two crops under heat treatment (1 h and 24 h; Fig. 4b).

In addition, the aforementioned ten *RWP-RK* genes exhibited significant correlations (Pearson's *rho* ≥ 0.6, $P < 0.05$) with most ER-related genes (60.2%; 325 out of 540) and *HSF* genes (50%; 16 out of 32) in response to heat stress (Supplementary Table 12), suggesting that *RWP-RK* genes might coregulate the heat tolerance of pearl millet with some ER-related genes and *HSF* genes. We further predicted potential RWP-RK binding sites upstream of these genes and found that higher proportions of ER-related genes had binding sites in pearl millet than in maize and rice (Extended Data Fig. 8c). The transient coexpression of the aforementioned *RWP-RK* (*PMF0G00024.1*) and two ER-related genes, encoding an immunoglobulin protein (BiP) (https://www.kegg.jp/entry/K09490; *PMA2G00107.1*) and the oligosaccharyltransferase complex (OST) (https://www.kegg.jp/entry/K12669; *PMA4G03758.1*), further confirming that *RWP-RK* functions at least partially by transactivating ER-related genes (Fig. 4c). Collectively, these results indicate that pearl millet may quickly respond to heat stress at the gene transcription level via the coregulation of *RWP-RK* genes with *HSF* genes and ER-related genes to eliminate proteins with temperature-induced misfolding (Fig. 4d).

Several focal SVs are associated with heat-related gene expression

Previous reports revealed that SVs could affect the transcription of nearby genes[16,17,33]; our data showed that nearly half of SVs were near genes (Fig. 2d). Therefore, we investigated the influence of SVs on the expression of nearby genes that responded to heat stress. The results showed that SVs were enriched in nearby genes showing changes in gene expression in all accessions and that genes located near SVs are probably more responsive to heat stress (Fig. 5a,b, Extended Data Fig. 9a–d and Supplementary Note 7). We further validated two SVs that could cause transcriptional changes in nearby genes via a transient gene expression experiment in tobacco (*Nicotiana tabacum*) leaves and used PCR to confirm these two SVs (Fig. 5c–e, Extended Data Fig. 9e–h and Supplementary Note 8).

To identify potential SVs related to transcriptional changes of particular heat-related genes, we distinguished four HR (Tifleaf3, PI583800, PI526529 and PI587025) and two HS (PI521612 and PI537069) accessions based on the distinct phenotypes and physiological indicators of these accessions when grown under heat treatment (Fig. 5f, Extended Data Fig. 9i,j and Supplementary Note 6.2). Considering that different breeds in the same group may use different genes to respond to heat stress, we focused on 2,354 SVs present in only three or all four HR accessions and nearby 2,769 genes. We designed an analysis pipeline to screen out 44 candidate SVs potentially related to the expression changes of 34 heat-related genes (Extended Data Fig. 9k, Supplementary Table 13 and Supplementary Note 7). Almost all these genes (33 out of 34) were responsive to heat stress based on our RNA-seq data and 11 genes (32.35%) were included in ER-related gene pathways (Supplementary Table 14), suggesting potential contributions of the neighboring SVs to the HSR. Notably, we found four fixed SVs between the HR and HS groups in the vicinity of *PMA1G04478.1* and *PMA7G02533.1* encoding two HSP70 proteins (https://www.kegg.jp/entry/K03283) and *PMA5G02838.1* encoding one heat shock chaperonin-binding protein, which were associated with differences in gene expression in the HR group than those in the HS group (Fig. 5g and Extended Data Fig. 9l). Interestingly, *PMA1G04478.1* and *PMA5G02838.1* in the ER-related pathway were also identified and the main response of pearl millet to heat stress was found in this pathway (Fig. 5g and Supplementary Tables 13 and 14). In general, the transcription levels of these three genes, which have essential roles in the HSR, were probably affected by their nearby SVs, further demonstrating that these SVs might have important roles in the heat tolerance of pearl millet.

Contributions of SVs to heat adaptation and domestication

To characterize the SVs underlying heat tolerance during adaptation in a pearl millet population (SRP063925)[7], we genotyped SVs by mapping all of the resequences to our graph-based pan-genome and identified a total of 124,532 SVs. We focused on the SVs with population frequency differences (fdSVs) between accessions from tropical and temperate zones by applying a sliding window methodology[34] (Supplementary Note 7). In total, 1,471 genes were annotated against 269 selection sweep regions harboring 4,411 fdSVs (Fig. 6a). Interestingly, we found that 27 of these genes were significantly ($P = 0.038$; chi-squared test) and functionally annotated as belonging to ER-related pathways (that is, the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway ko04141) (Supplementary Tables 15 and 16). From the 591 genes whose expression was previously shown to be associated with SVs (Supplementary Table 13), we identified 25 genes near 27 fdSVs that were present only in the HR group; their expression levels were significantly correlated with the presence of fdSVs (Supplementary Table 17). Notably, one of the fdSVs was positioned close to (360 bp) and upstream of *PMA2G02653.1*, a gene encoding a protein in the zinc finger family that has a role in the ER system[35–37]. This gene was enriched in Gene Ontology (GO) terms associated with the response to temperature

https://doi.org/10.1038/s41588-023-01302-4

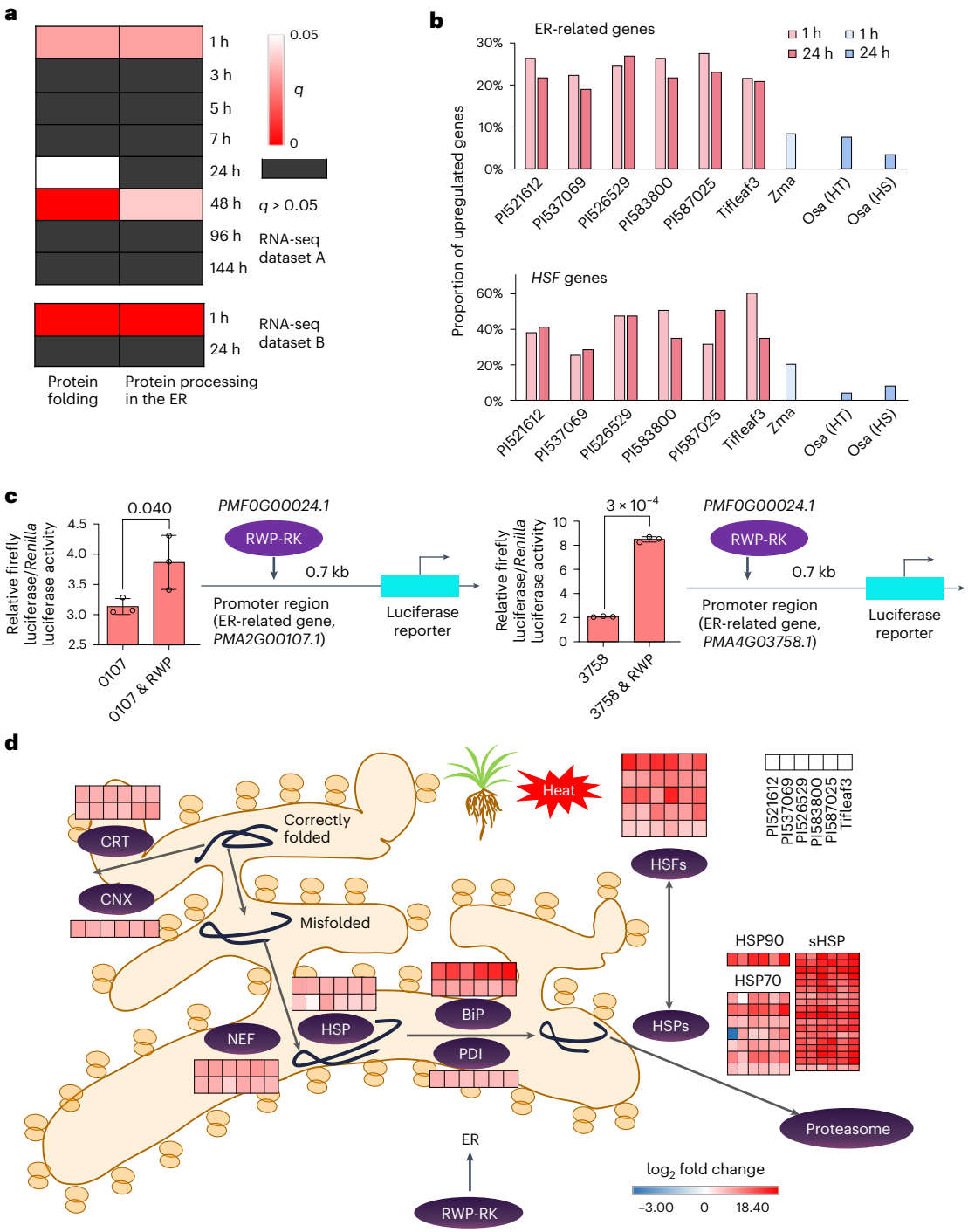


Fig. 4 | Transcriptome analyses reveal that pearl millet responds to heat stress via ER-related pathways. a, Functional enrichment of DEGs coexisting in Tifleaf3 under high temperature stress (40 °C and 35 °C) at eight time points (1–144 h; dataset A in Supplementary Table 3) and in six accessions (PI521612, PI537069, PI526529, PI583800, PI587025 and Tifleaf3) under high temperature stress (45 °C under light and 40 °C in darkness) for 1 h and 24 h (dataset B in Supplementary Table 3). **b**, Comparison of the proportions of upregulated ER-related and *HSF* genes after 1 h and 24 h of heat treatment in pearl millet, maize and rice. Heat-resistant (HR) and heat-susceptible (HS) rice samples, respectively. **c**, Dual luciferase assays were applied to verify that *PMF0G00024.1* (RWP) could transactivate the *PMA2G00107.1* (107) and *PMA4G03758.1* (3758) genes. The error bars represent the mean ± s.d.; $n = 3$ biological replicates. Significant differences were tested using a two-tailed *t*-test and are shown as *P* values. **d**, Proposed activation network of pearl millet in response to growth under heat stress. After 1 h of high-temperature stress in six pearl millet accessions, many misfolded proteins activated the expression of degradation-related genes in the ER, such as genes encoding recognition proteins, including calnexin (CNX) and calreticulin (CRT), and degradation-related proteins, including heat shock proteins (HSPs), thereby correcting or degrading misfolded proteins to maintain protein homeostasis in cells. In addition, the *HSF* and *RWP-RK* genes potentially participate in this process to coregulate *HSP* genes.

stress (GO: 0050826) and was also responsive to heat stress (Extended Data Fig. 10a). We further identified this fdSV as present in accessions that were preferentially located in higher-latitude regions (Fig. 6a and Supplementary Note 7). In general, these results revealed the contributions of SVs possibly associated with the ER system to heat stress adaptation.

https://doi.org/10.1038/s41588-023-01302-4

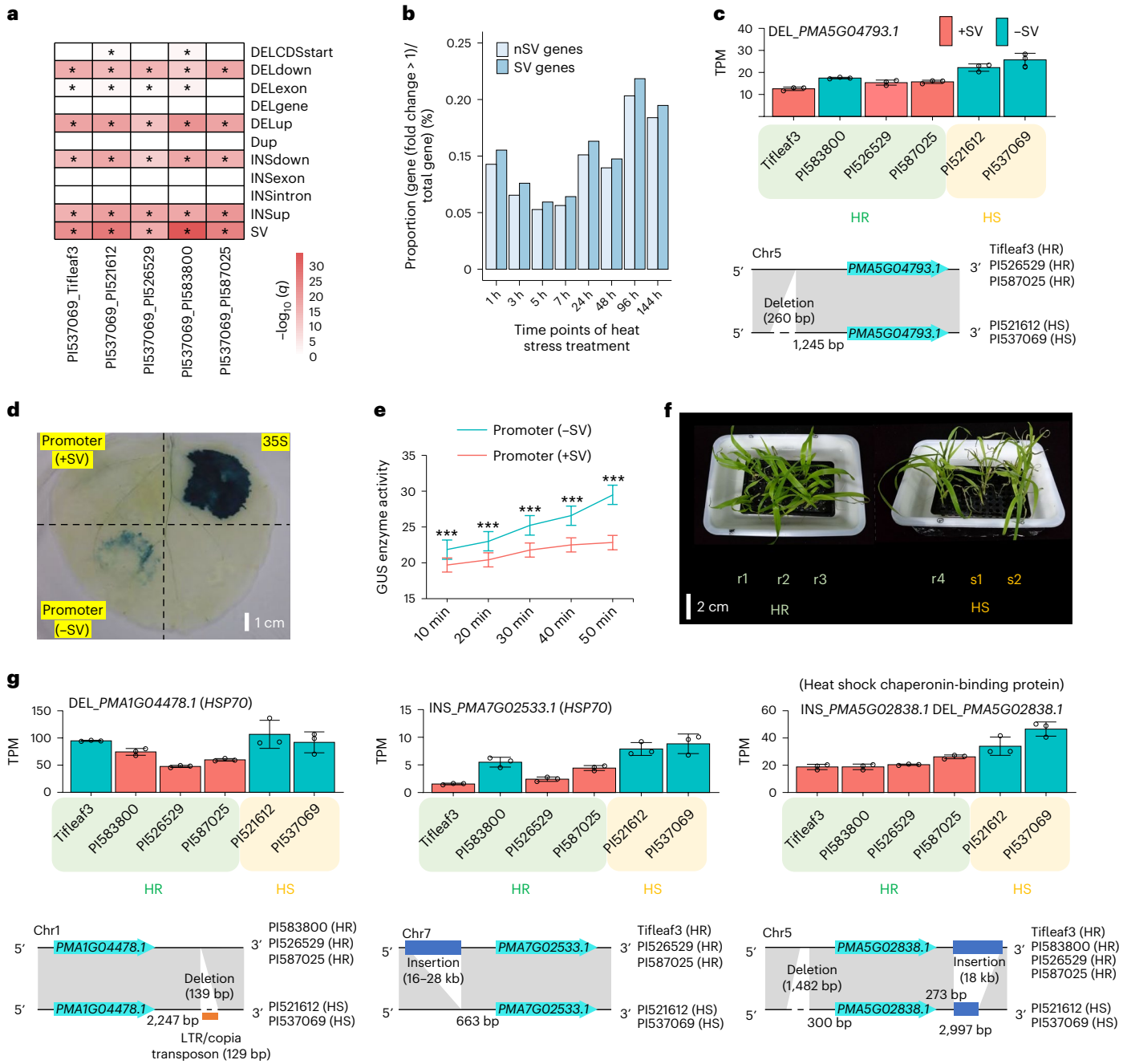


Fig. 5 | Impact of SVs on genes and their contributions to heat tolerance in pearl millet. a, Enrichment of SVs near genes with altered expression in each accession relative to PI537069. The asterisk indicates significance ($q < 0.05$). DEL, deletion; Dup, duplication; INS, insertion. **b**, Proportions of DEGs among total genes overlapping with SVs (SV genes) and those not overlapping with SVs (nSV genes) in leaf tissue under heat treatment. **c**, Deletion near *PMA5G04793.1* with expression changes in different accessions. TPM, transcripts per million. +SV and −SV: accessions with and without SVs, respectively. The error bars indicate the mean ± s.d.; $n = 3$ biological replicates. **d**,**e**, Transformation of the *PMA5G04793.1* promoter in tobacco leaves. **d**, Glucuronidase (*GUS*) reporter gene expression observed by histochemical staining. **e**, Quantitative detection of the GUS enzyme in leaves inoculated with different recombinant vectors at different time points using a microplate plate-based GUS fluorescence assay. The error bars indicate the mean ± s.d.; $n = 3$ biological replicates. Significant differences were tested using a two-tailed *t*-test (***$P < 0.0005$). **f**, Phenotypic comparison of six accessions under heat treatment for 96 h. r1–r4: Tifleaf3, PI583800, PI526529 and PI587025 (HR group). s1 and s2: PI521612 and PI537069 (HS group). The HS plants were more wilted than the HR plants. **g**, Three examples showing the presence of fixed SVs in the HR groups (≥3 accessions) near three heat-related protein genes. Descriptive gene models are presented below the bar charts. The error bars indicate the mean ± s.d.; $n = 3$ biological replicates.

To characterize the domestication of pearl millet with a shift toward higher heat tolerance, we used the above pearl millet population (SRP063925) to identify 113 selection sweep regions harboring 3,952 fdSVs overlapping with 1,285 genes between the landrace and improved cultivars relative to the wild accessions (Extended Data Fig. 10c and Supplementary Table 1). Functional enrichment analyses showed that these genes were associated mainly with stress-related GO terms, including temperature, abiotic stimulus and isoprenoid biosynthetic process (Extended Data Fig. 10c). We also found that 79.3% of those genes (1,019 out of 1,285) exhibited transcriptional

https://doi.org/10.1038/s41588-023-01302-4

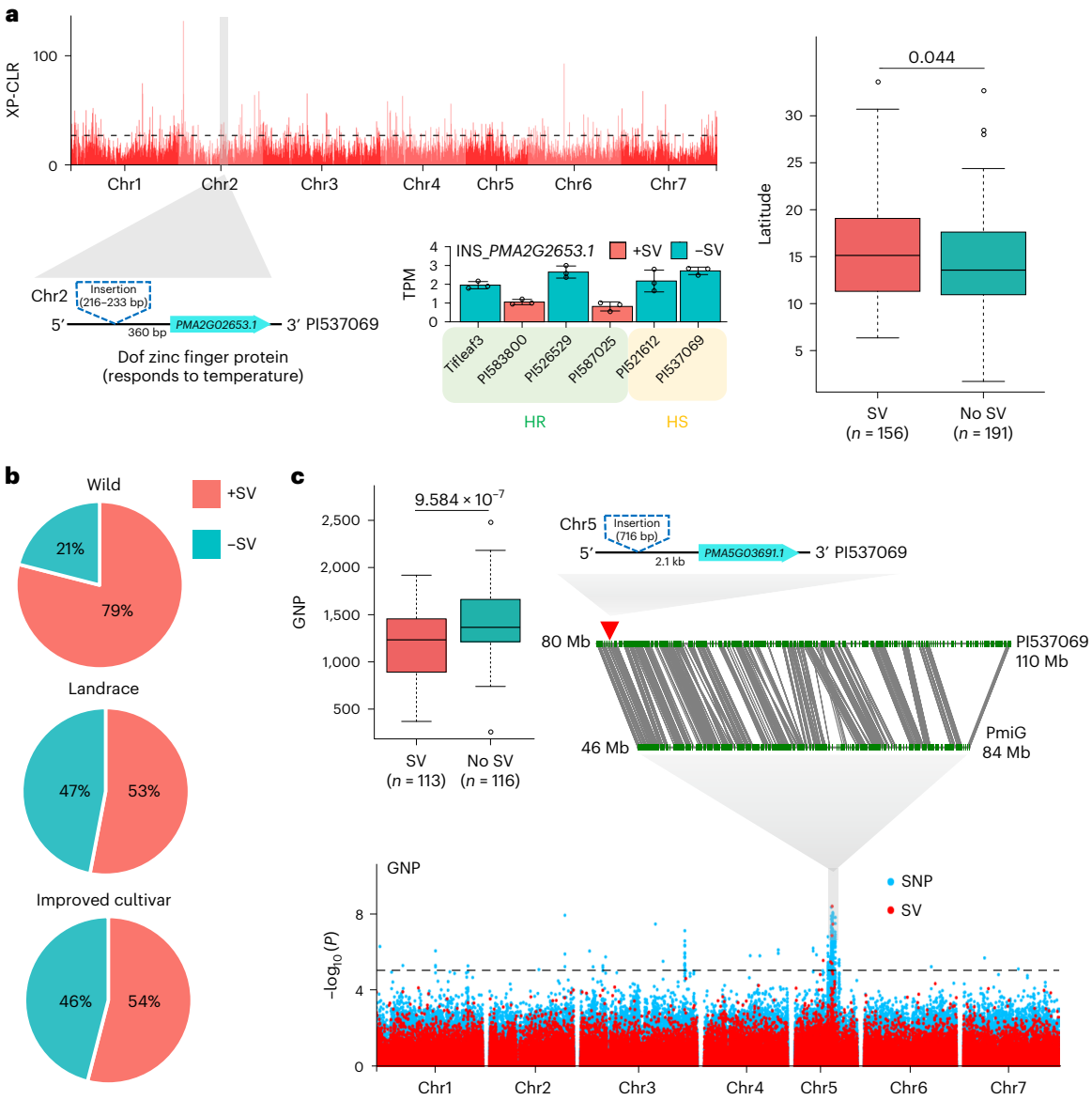


Fig. 6 | SVs contribute to heat tolerance adaptation and domestication. **a**, Upper left, results of an SV-based selection sweep analysis between two groups located in tropical and temperate zones. The black dashed line represents a cutoff window in which the top 1% data points were selected as sweep regions. Bottom left, the expression of *PMA2G02653.1* is influenced by a nearby SV. The error bars indicate the mean ± s.d.; $n = 3$ biological replicates. Right, comparison of the latitudinal distribution of the two SV-related haplotypes in pearl millet accessions. The center line indicates the median; the box limits indicate the upper and lower quartiles; the whiskers indicate 1.5 times the IQR; the dots represent the outliers. Significant differences were tested using a two-tailed *t*-test and are shown as *P* values. XP-CLR, cross-population composite likelihood ratio. **b**, Frequencies of SVs in wild and landrace accessions and improved cultivars. **c**, PAV–GWAS of SVs associated with the GNP trait based on the graph-based pan-genome. The center line indicates the median; the box limits indicate the upper and lower quartiles; the whiskers indicate 1.5 times the IQR; the dots represent the outliers. Significant differences were tested using a two-tailed *t*-test and are shown as *P* values. PmiG: Tift 23D2B1-P1-P5. The black dashed line represents the significance threshold based on a $-\log_{10}(P) > 5$.

changes (Supplementary Table 18), indicating that fdSVs potentially influence domestication genes under heat stress. In addition, 17 of these genes near 16 fdSVs were present only in the HR group and the fdSVs were significantly correlated with their gene expression levels; among these genes, *PMA2G02653.1* was also related to temperature adaptation (Fig. 6a).

Additionally, we found that a 716-bp insertion (SV) was present in a higher proportion of the wild accessions than the landrace ones and improved cultivars (Fig. 6b). This insertion was positioned 2.1 kb upstream of *PMA5G03691.1*, which encodes a coiled-coil 90B-like protein that is probably responsible for pollen germination and is associated with the grain number per panicle (GNP) trait. The presence of this insertion was possibly correlated with heat-induced gene expression (Extended Data Fig. 10d). We then conducted a genome-wide association study (GWAS) examining the associations of the 124,532 PAVs and 1,455,924 SNPs with GNP in a population reported by Varshney et al.[7] (Supplementary Table 19). An association peak on chromosome 5 showed an overlap between PAVs and SNPs. This quantitative trait locus corresponds to grain number[7]. In our study, we found *PMA5G03691.1* and an insertion in the close vicinity of this quantitative trait locus (Fig. 6c, Supplementary Table 19 and Supplementary Note 7). We next observed this insertion in 113 accessions with lower GNP values than 116 accessions without the SV (Fig. 6c). These results suggested that this insertion was probably under positive selection

https://doi.org/10.1038/s41588-023-01302-4

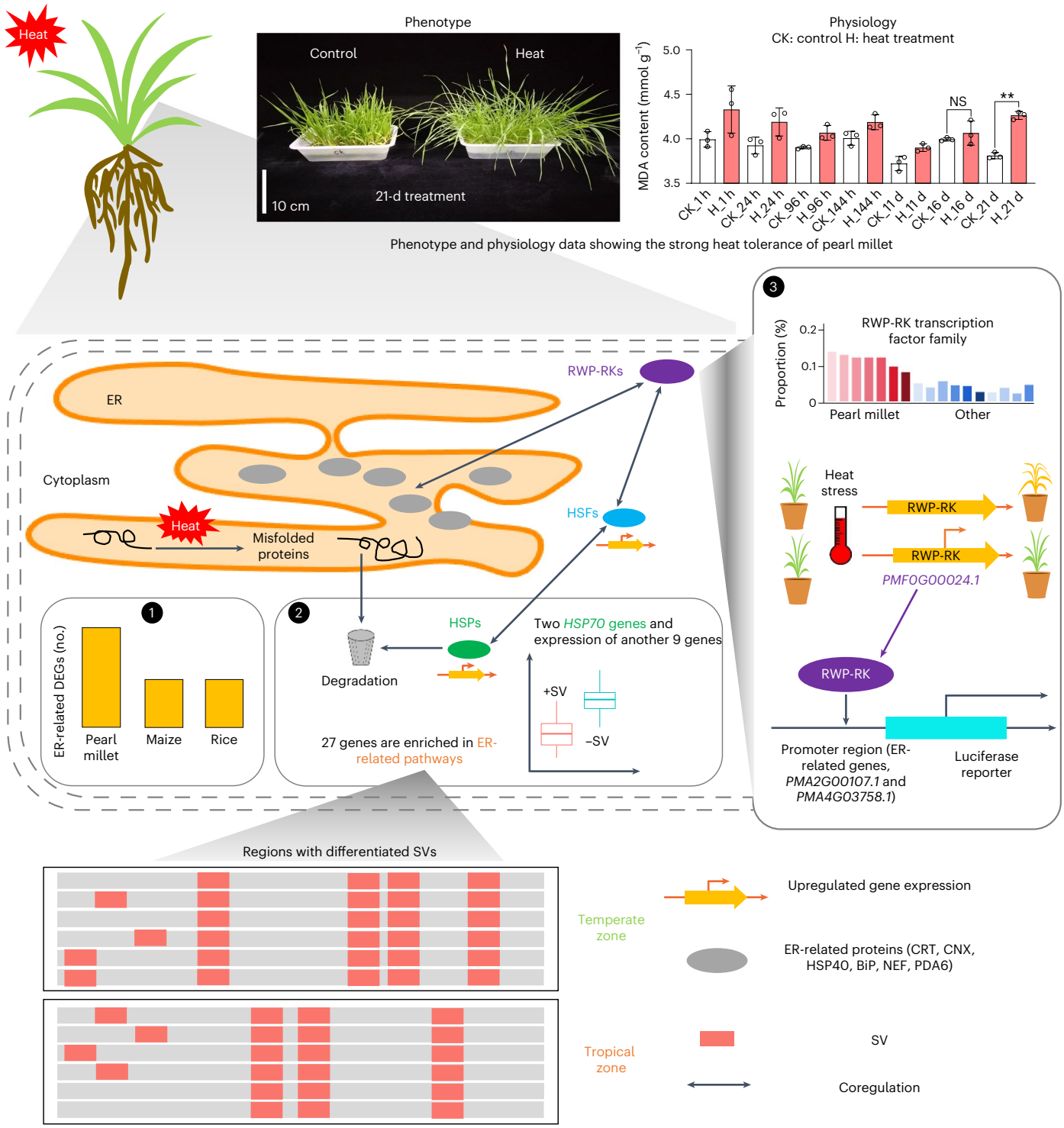


Fig. 7 | A proposed mechanism by which heat tolerance is integrally related to the transport system of the ER. After heat stress (H) for 21 d, only a small proportion of leaves exhibited wilting relative to the control (CK) group; leaves showed physiological changes (MDA) for up to 21 d, suggesting that pearl millet exhibits strong heat resistance. Significant differences were tested using a two-tailed *t*-test (**$P < 0.01$). We leveraged multi-omics analyses to reveal a possible mechanism of heat tolerance in the ER transport system: (1) compared with maize and rice, pearl millet showed a higher proportion of ER-related genes that were differentially expressed, indicating a quicker response to heat stress in this system; (2) this heat stress led to the production of misfolded proteins that could be recognized and degraded via the cooperation of ER-related proteins such as CRT, CNX, BiP, NEF, PDA6 and HSP. *HSF* genes might be involved in the heat response because their expression is upregulated and they can coregulate *HSP* genes[45,46]. SVs surrounding 11 ER-related genes probably contributed to this response. For instance, one SV was associated with the expression of an ER-related gene, *HSP70*, which plays a role in the degradation of misfolded proteins. Additionally, 27 genes enriched in the ER system were located in regions with differentiated SV distributions between two populations in temperate and tropical zones; (3) furthermore, an *RWP-RK* gene (*PMF0G00024.1*) from an expanded transcription factor family was confirmed as a positive regulator involved in heat resistance and was coregulated by ER-related and *HSF* genes. We finally used dual luciferase assays to confirm that this *PMF0G00024.1* gene could transactivate genes encoding ER-related BiP (*PMA2G00107.1*) and OST (*PMA4G03758.1*).

during domestication and influenced the responsiveness of nearby genes to heat, possibly contributing to seed production in pearl millet grown at higher temperatures. Furthermore, we identified a total of 142 PAVs that were each associated with one or more traits (20 traits in total), which might provide insights into the contributions of these SVs to pearl millet molecular breeding (Supplementary Table 19 and Supplementary Note 7). Collectively, these results demonstrate the utility of pearl millet graph-based pan-genome analysis for the identification of both heat tolerance adaptation and its relationship to domestication.

Resistance to heat in pearl millet depends on the ER system

We performed integrated multi-omics analyses supplemented with *cis*-genetic functional verification to propose a possible mechanism by which the superior heat tolerance of pearl millet is related to the expansion and altered expression of genes involved in the ER system (Fig. 7). In particular, the ER system showed a quicker response to high temperature in pearl millet than in maize and rice. Abundant evidence has shown that SVs participate in the heat tolerance response by affecting gene regulation; for example, SVs between HR and HS materials led to differential expression levels of 11 ER-related genes. Several other distinctly differentiated SVs in ER-related genes were also associated with the heat stress adaptation of pearl millet populations at different temperatures. Moreover, by means of functional analysis, we confirmed that one gene (*PMF0G00024.1*) from an expanded RWP-RK transcription factor family acted as a positive regulator of heat resistance; this transcription factor also transactivated one ER-related gene. These observations indicate that SVs and *RWP-RK* genes may coregulate the quick response to heat stress with ER-related genes in pearl millet.

Discussion

Pearl millet is an ideal model for investigating the mechanisms underlying plant heat resistance[9]. We identified distinctly differentiated SVs in ER-related genes that were associated with the heat stress adaptation of pearl millet populations at different temperatures (Fig. 6a); however, we did not find genes in SNP-based selection sweep regions that showed significant enrichment in ER-related pathways. These findings indicate that SV-based population analyses can capture genetic variations complementary to SNPs, providing additional information about the diversity losses caused by population bottlenecks during plant adaptation[38]. In addition, the expansion of RWP-RK transcription factors was likely related to LTR and these factors coregulated heat tolerance with ER-related and heat stress-related genes (Figs. 3 and 4). RWP-RK transcription factors have an important role in the nitrogen starvation response and gametophyte development in plants[39,40]. However, no heat tolerance-related functions of these transcription factors have been reported. Our findings expand the possible functions of RWP-RK transcription factors and illustrate a possible diversification in which this family of transcription factors is responsible for multiple stress condition responses in plants. This finding supports a previous hypothesis that pearl millet probably includes abundant heat tolerance-related genetic resources[6].

The graph-based pan-genome resource offers several potential tools to improve the breeding process in pearl millet. We developed a comprehensive SV map of pearl millet to identify signals associated with phenotypes (that is, GNP) (Fig. 6c), which enables us to investigate potential mechanisms influencing nearby genes that are challenging to detect based only on SNP genotyping. This pan-genome also provides a new window for identifying evolutionary processes, such as the formation of adaptative SVs, to elucidate demographic and selection processes in pearl millet. The dispensable genome within the pan-genome resource offers a pathway for identifying genes associated with traits such as abiotic stress resistance or production, which would benefit the selection of suitable materials for use as breeding targets in pearl millet. In our pan-genome, PmiG showed a higher ratio of private gene families relative to the other assemblies (Fig. 1c), possibly caused by the relatively fragmented sequences generated by previous short-read sequencing or assembly techniques[7,41–43]. A similar result was reported in a soybean pan-genome study[17]. The relatively lower contig N50 value intuitively suggested that the PmiG genome sequence is more fragmented (Table 1), which would lead to a lower average length of genes and coding sequences and a higher proportion of short genes (<1 kb) (Extended Data Fig. 4c,d). Thus, fragmentation of assembled sequences would result in incomplete prediction of genes, potentially contributing to the private gene set in the PmiG. Nonetheless, the PmiG, as the first published pearl millet genome, has been widely used as a reference genome in the pearl millet community[20,25,44]. Including it in our pan-genome research will help to refer to the basis of previous research and provide a smooth transition to the era of high-quality pearl millet genome research.

In conclusion, our study uses a pan-genome approach coupled with multi-omics to comprehensively investigate plant response mechanisms to heat stress. This work provides an excellent reference for future studies on stress tolerance, especially in non-model plants. Our study also offers an approach for breeding crop varieties with enhanced tolerance to various stresses that can cope with the diverse challenges imposed by the changing climate.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-023-01302-4.

References

1. Lesk, C., Rowhani, P. & Ramankutty, N. Influence of extreme weather disasters on global crop production. *Nature* **529**, 84–87 (2016).
2. National Research Council *Advancing the Science of Climate Change* (National Academies Press, 2010).
3. Zhao, C. et al. Temperature increase reduces global yields of major crops in four independent estimates. *Proc. Natl Acad. Sci. USA* **114**, 9326–9331 (2017).
4. Pucher, A. et al. Agro-morphological characterization of West and Central African pearl millet accessions. *Crop Sci.* **55**, 737–748 (2015).
5. Jukanti, A., Gowda, C. L., Rai, K. N., Manga, V. K. & Bhatt, R. K. Crops that feed the world 11. Pearl millet (*Pennisetum glaucum* L.): an important source of food security, nutrition and health in the arid and semi-arid tropics. *Food Secur.* **8**, 307–329 (2016).
6. Satyavathi, C. T., Ambawat, S., Khandelwal, V. & Srivastava, R. K. Pearl millet: a climate-resilient nutricereal for mitigating hidden hunger and provide nutritional security. *Front. Plant Sci.* **12**, 659938 (2021).
7. Varshney, R. K. et al. Pearl millet genome sequence provides a resource to improve agronomic traits in arid environments. *Nat. Biotechnol.* **35**, 969–976 (2017).
8. James, D. et al. Development and characterization of a high temperature stress responsive subtractive cDNA library in pearl millet *Pennisetum glaucum* (L.) R. Br. *Indian J. Exp. Biol.* **53**, 543–550 (2015).
9. Mohammed, R., Gangashetty, P. I., Karimoune, L. & Ba, N. M. Genetic variation and diversity of pearl millet [*Pennisetum glaucum* (L.)] genotypes assessed for millet head miner, *Heliocheilus albipunctella* resistance, in West Africa. *Euphytica* **216**, 158 (2020).
10. Huang, D. et al. Transcriptional changes in pearl millet leaves under heat stress. *Genes* **12**, 1716 (2021).

11. Sun, M. et al. Transcriptome analysis of heat stress and drought stress in pearl millet based on Pacbio full-length transcriptome sequencing. *BMC Plant Biol.* **20**, 323 (2020).
12. Fuentes, R. R. et al. Structural variants in 3000 rice genomes. *Genome Res.* **29**, 870–880 (2019).
13. Catacchio, C. et al. Transcriptomic and genomic structural variation analyses on grape cultivars reveal new insights into the genotype-dependent responses to water stress. *Sci. Rep.* **9**, 2809 (2019).
14. Cardone, M. F. et al. Inter-varietal structural variation in grapevine genomes. *Plant J.* **88**, 648–661 (2016).
15. Yuan, Y., Bayer, P. E., Batley, J. & Edwards, D. Current status of structural variation studies in plants. *Plant Biotechnol. J.* **19**, 2153–2163 (2021).
16. Qin, P. et al. Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations. *Cell* **184**, 3542–3558 (2021).
17. Liu, Y. et al. Pan-genome of wild and cultivated soybeans. *Cell* **182**, 162–176 (2020).
18. Zhou, Y. et al. A platinum standard pan-genome resource that represents the population structure of Asian rice. *Sci. Data* **7**, 113 (2020).
19. Della Coletta, R., Qiu, Y., Ou, S., Hufford, M. B. & Hirsch, C. N. How the pan-genome is changing crop genomics and improvement. *Genome Biol.* **22**, 3 (2021).
20. Serba, D. D. et al. Genetic diversity, population structure, and linkage disequilibrium of pearl millet. *Plant Genome* **12**, 1–12 (2019).
21. Ou, S., Chen, J. & Jiang, N. Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Res.* **46**, e126 (2018).
22. 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).
23. Gordon, S. P. et al. Extensive gene content variation in the *Brachypodium distachyon* pan-genome correlates with population structure. *Nat. Commun.* **8**, 2184 (2017).
24. Zhao, Q. et al. Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. *Nat. Genet.* **50**, 278–284 (2018).
25. Burgarella, C. et al. A western Sahara centre of domestication inferred from pearl millet genomes. *Nat. Ecol. Evol.* **2**, 1377–1380 (2018).
26. Cetinkaya, H., Tasci, E. & Seckin Dinler, B. Regulation of glutathione S-transferase enzyme activity with salt pre-treatment under heat stress in maize leaves. *Res. Plant Biol.* **4**, 45–56 (2014).
27. Yeh, S.-H., Lin, C.-S., Wu, F.-H. & Wang, A.-Y. Analysis of the expression of *BohLOL1*, which encodes an LSD1-like zinc finger protein in *Bambusa oldhamii*. *Planta* **234**, 1179–1189 (2011).
28. Zhang, X. et al. *ScMED7*, a sugarcane mediator subunit gene, acts as a regulator of plant immunity and is responsive to diverse stress and hormone treatments. *Mol. Genet. Genomics* **292**, 1363–1375 (2017).
29. Huang, B. et al. Molecular characterization and functional analysis of tumor necrosis factor receptor-associated factor 2 in the Pacific oyster. *Fish Shellfish Immunol.* **48**, 12–19 (2016).
30. Jagadhesan, B. et al. Genome wide analysis of NLP transcription factors reveals their role in nitrogen stress tolerance of rice. *Sci. Rep.* **10**, 9368 (2020).
31. He, J. et al. Genome-wide transcript and small RNA profiling reveals transcriptomic responses to heat stress. *Plant Physiol.* **181**, 609–629 (2019).
32. Fang, C., Dou, L., Liu, Y., Yu, J. & Tu, J. Heat stress-responsive transcriptome analysis in heat susceptible and tolerant rice by high-throughput sequencing. *Ecol. Genet. Genom.* **6**, 33–40 (2018).
33. Alonge, M. et al. Major impacts of widespread structural variation on gene expression and crop improvement in tomato. *Cell* **182**, 145–161 (2020).
34. Chen, H., Patterson, N. & Reich, D. Population differentiation as a test for selective sweeps. *Genome Res.* **20**, 393–402 (2010).
35. Ge, W. et al. Main regulatory pathways, key genes and micro RNAs involved in flower formation and development of moso bamboo (*Phyllostachys edulis*). *Plant Biotechnol. J.* **15**, 82–96 (2017).
36. Ueda, H. et al. Endoplasmic reticulum (ER) membrane proteins (LUNAPARKs) are required for proper configuration of the cortical ER network in plant cells. *Plant Cell Physiol.* **59**, 1931–1941 (2018).
37. Min, M. K. et al. Overexpression of *Arabidopsis* AGD7 causes relocation of Golgi-localized proteins to the endoplasmic reticulum and inhibits protein trafficking in plant cells. *Plant Physiol.* **143**, 1601–1614 (2007).
38. Siol, M., Wright, S. I. & Barrett, S. C. The population genomics of plant adaptation. *New Phytol.* **188**, 313–332 (2010).
39. Chardin, C., Girin, T., Roudier, F., Meyer, C. & Krapp, A. The plant RWP-RK transcription factors: key regulators of nitrogen responses and of gametophyte development. *J. Exp. Bot.* **65**, 5577–5587 (2014).
40. Sakuraba, Y., Zhuo, M. & Yanagisawa, S. RWP-RK domain-containing transcription factors in the Viridiplantae: their biology and phylogenetic relationships. *J. Exp. Bot.* **73**, 4323–4337 (2022).
41. Logsdon, G. A., Vollger, M. R. & Eichler, E. E. Long-read human genome sequencing and its applications. *Nat. Rev. Genet.* **21**, 597–614 (2020).
42. Hu, T., Chitnis, N., Monos, D. & Dinh, A. Next-generation sequencing technologies: an overview. *Hum. Immunol.* **82**, 801–811 (2021).
43. Mallick, S., Gnerre, S., Muller, P. & Reich, D. The difficulty of avoiding false positives in genome scans for natural selection. *Genome Res.* **19**, 922–933 (2009).
44. Chanwala, J. et al. Genome-wide identification and expression analysis of WRKY transcription factors in pearl millet (*Pennisetum glaucum*) under dehydration and salinity stress. *BMC Genomics* **21**, 231 (2020).
45. Khan, Z. & Shahwar, D. In *Sustainable Agriculture in the Era of Climate Change* 211–234 (Springer, 2020).
46. Liu, J.-X. & Howell, S. H. Endoplasmic reticulum protein quality control and its relationship to environmental stress responses in plants. *Plant Cell* **22**, 2930–2942 (2010).

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

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2023

Methods

Sampling and sequencing

Ten pearl millet accessions (PI537069, PI521612, PI526529, PI587025, PI583800, PI343841, PI186338, PI250656, PI527388 and Tifleaf3) were obtained as representative plants from different geographical regions. All ten accessions were planted in a greenhouse at a density of three plants per pot (filled with nutrient soil), including nine plants of each accession, and grown at a temperature of 26 °C during the light period (14 h of light) and 22 °C during the dark period (10 h of darkness). Thirteen-week-old leaves were collected and immediately frozen in liquid nitrogen for the extraction of genomic DNA using a DNAsecure Plant Kit (TIANGEN). Library construction and Illumina, Hi-C, PacBio and Bionano sequencing were performed at Novogene (Supplementary Note 1).

Genome survey

The genome size of pearl millet was estimated using *k*-mer frequency analysis based on the Lander–Waterman algorithm[47]. We divided the total length of sequence reads by the sequencing depth represented by the peak value of the frequency curve. The following formula was used to estimate genome size: $(N \times (L - k + 1) - B)/D = G$, where N is the total number of sequence reads, L is the average length of the sequence reads, k is the *k*-mer length (17 bp), B is the total number of low-frequency *k*-mers (frequency ≤ 1 in this analysis), G is the genome size and D is the overall estimated depth based on the *k*-mer distribution[48]. Additionally, flow cytometry was used to confirm the estimated genome size according to a reported method[49] with a BD FACSCalibur flow cytometer and the fluorochrome propidium iodide.

Initial assembly

The PacBio HiFi reads were used to assemble the initial contigs in the Hifiasm (v.0.13-r308)[50] package with default parameters. The Pruge_haplotig (v.1.1.0)[51] tool was used to process genomic heterozygous regions to remove redundancy in the genomes using default parameters with several exceptions: -a 50.

Scaffolding with Bionano optical maps

The filtered raw DNA molecules in BNX format were aligned, clustered and assembled into a Bionano optical map using the Bionano Genomics assembly pipeline. Then, a BNX file recorded the basic labeling and DNA length information was converted with the AutoDetect in Bionano Solve package (v3.5.1) (https://bionanogenomics.com/support/software-downloads/). The initial assemblies were aligned to the Bionano data and then analyzed with RefAligner in Bionano Solve package (v3.5.1). The alignments were visualized with a snapshot in IrysView in Bionano Solve package (v3.5.1). Finally, genome maps were combined with the initial assembly to produce hybrid scaffold genome maps using the Bionano Solve package (v.3.5.1) with the parameters -B 1 -N 1.

Pseudochromosome construction

Linkage information for the scaffold and initial assembly was obtained by aligning high-quality Hi-C data to the preceding assemblies using the Burrows–Wheeler Aligner (BWA) software (v.0.7.8)[52]. Chromosome-scale scaffolds were anchored based on linkage information, restriction enzyme sites and the string graph formulation using the ALLHIC (v.0.9.8)[53] package with the following parameters: -K 7 -minREs 50--maxlinkdensity 3--NonInformativeRatio 0. Placement and orientation errors showing obvious discrete chromatin interaction patterns were adjusted manually. For those accessions without Hi-C data, we used collinearity with the PI537069 assembly for clustering and orientation to generate chromosome-level assemblies.

Genome assessment

To evaluate the assembly quality of the genomes, BUSCO (v.4.1.2; http://busco.ezlab.org/)[54] and the CEGMA (v.2.5) (http://korflab.ucdavis.edu/dataseda/cegma/)[55] were used to check the completeness of the genome assembly or annotation. The draft assemblies were further evaluated by mapping the high-quality Illumina paired-end reads to the genome assembly using the BWA–MEM (v.0.7.8)[52] algorithm. The quality of the genome assemblies was further evaluated using LTR TE completeness based on the LAI tool wrapped in LTR_retriever (v.2.8)[21] and using Merqury (v.1.3)[22] with the default parameters.

Annotation of repetitive sequences

Transposons were annotated by combining two strategies, that is, homolog and de novo predictions. For the homology-based approach, the Repbase TE library[56] and the TE protein database (http://www.repeatmasker.org/cgi-bin/RepeatProteinMaskRequest) were used to mask TEs with the RepeatMasker (v.4.0.5)[57] and RepeatProteinMask (v.4.0.5)[57] tools. Under the de novo-based method, LTR_FINDER (v.1.0.7) (https://github.com/xzhub/LTR_Finder)[58], PILER (v.1.0) (https://www.drive5.com/piler/)[59], RepeatScout (v.1.0.5) (https://github.com/mmcco/RepeatScout)[60] and RepeatModeler (v.1.0.8) (http://www.repeatmasker.org/RepeatModeler.html)[61] were used to build a de novo repeat library. This new library was used to mask TEs with the RepeatMasker tool[57]. We estimated the insertion times of the intact LTR retrotransposons. Sequences from the 5′ and 3′ LTRs were aligned with MUSCLE[62] (v.3.8.31). Nucleotide variations (λ) in the 5′ and 3′ ends of intact LTR retrotransposons were calculated and DNA substitution rates (K) were calculated using $K = -0.75\ln(1 - 4\lambda/3)$. The insertion time of these LTR retrotransposons was estimated based on $T = K/2r$, where r is 1.3×10^{-8} per site and per year[63].

Annotation of gene structure

Gene annotation was conducted by combining de novo-, homolog- and transcriptome-based predictions. For the homolog-based approach, we downloaded homologous proteins from the *A. thaliana*, *Z. mays*, *S. bicolor*, *O. sativa*, *S. italica* and pearl millet genomes (Phytozome 13, https://phytozome.jgi.doe.gov/pz/portal.html; NCBI, https://www.ncbi.nlm.nih.gov/) and aligned them to the pearl millet genome with Tblastn (v.2.2.26)[64] using an expected value of 1×10^{-5}. Solar (v.0.9.6)[65] was used to combine the BLAST hits (Homo-set), which were used to predict the exact gene structures of the corresponding genomic regions with GeneWise (v.2.4.1)[66] (https://www.ebi.ac.uk/Tools/psa/genewise). For the transcriptome-based approach, RNA-seq data from Illumina were mapped to the assembled genome with TopHat (v.2.0.13)[67], followed by Cufflinks (v.2.1.1)[68]. In addition, Trinity (v.2.1.1)[69] was used to assemble the RNA-seq data and its output was used to create pseudo-expressed sequence tags, which were then mapped to the assembly. Gene models were predicted by using the Program to Assemble Spliced Alignments (PASA) genome annotation tool[70]. This gene set was denoted as the PASA-T-set and was used to train ab initio gene prediction programs. For the de novo-based approach, five ab initio gene prediction programs, including AUGUSTUS (v.3.2.3) (http://augustus.gobics.de/)[71], GENSCAN (v.1.0) (http://genes.mit.edu/GENSCAN.html)[72], GlimmerHMM (v.3.0.1) (http://ccb.jhu.edu/software/glimmerhmm/)[73], geneid (v.1.4) (http://genome.crg.es/software/geneid/)[74] and SNAP (v.2013.11.29) (http://korflab.ucdavis.edu/software.html)[75] were used to predict coding regions from the repeat-masked genome. Finally, EVidenceModeler (v.1.1.1)[76] was used to combine all gene model evidence obtained from these three strategies.

Functional annotation of protein-coding genes

Two protein sequence databases, Swiss-Prot (http://web.expasy.org/docs/swiss-prot_guideline.html) and the NR Protein Sequence database (ftp://ftp.ncbi.nih.gov/blast/db/) were used to annotate protein-coding genes. Protein domains were predicted using InterProScan (v.4.8) and HMMER (v.3.1) (http://www.hmmer.org/) based on the InterPro (v.32.0) (http://www.ebi.ac.uk/interpro/) and Pfam (v.27.0) (https://pfam-legacy.xfam.org/) databases, respectively[77–80].

These two databases provide a portal for obtaining GO terms (http://geneontology.org/ http://www.geneontology.org/page/go-database)[81]. The pathways of the genes were identified via BLAST searches against the KEGG database (v.53) (http://www.kegg.jp/kegg/kegg1.html)[82] with an expected value cutoff of 1×10^{-5}.

Pan-genome construction

We constructed a pan-genome using the 11 pearl millet assemblies. The core and dispensable gene sets among the 11 pearl millet genomes were estimated based on gene family clustering using OrthoFinder (v.2.3.1)[83]. All protein sequences were subjected to homologous searches using BLASTP with an expected value of 1×10^{-5}. Protein sequences were clustered into paralogous and orthologous sequences using OrthoFinder with an inflation parameter of 1.5.

SV identification

To build a genetic variance atlas for the 11 pearl millet genomes, we aligned the other ten genomes to the PI537069 reference genome using MUMmer (v.4.0.0)[84]. The alignment of the genomes was performed using NUCmer[84] (--c 1000--maxgap=500) and the alignment block filter was implemented using a delta filter in one-to-one alignment mode (−1). Blocks longer than 1,000 bp were used for further analysis. We used the SV function of the MUMmer (SVMU) pipeline to automate PAV discovery by parsing the results of NUCmer. From the SVMU results, SV-based insertions or deletions (with the tag INS or DEL) were treated as PAVs and CNVs were treated as CNVs. Inversion events (referring to SVs more than 1 kb in length) were identified by SVMU. SyRI (v.1.6.3) (https://github.com/schneebergerlab/syri)[85] was used to identify translocation regions. We also used PI537069 as a reference to construct a graph-based genome with the vg tool (v.1.25.0) (https://github.com/vgteam/vg)[86]. To genotype the population SVs, the Illumina short reads (SRP063925) of each accession were mapped to the graph-based genome using the vg tool with default parameters.

Transcription factor family identification and analysis

To identify and compare transcription factor families in pearl millet and other species, we collected the protein sequences of *A. thaliana* (TAIR10)[87], *Z. mays* (B73_RefGen_v4)[88], *B. distachyon* (v.3.1) (https://phytozome-next.jgi.doe.gov/info/Bdistachyon_v3_1)[89], *O. thomaeum* (v.1.0)[90], *P. hallii* (PHallii_v3.1)[91], *D. oligosanthes* (ASM163321v2)[92], *O. sativa* (IRGSP-1.0)[93], *S. bicolor* (Sorghum_bicolor_NCBIv3)[94], *H. vulgare* (Hvulgare_462_r1)[95], *S. italica* (Setaria_italica_v2.0)[96], *S officinarum* (v.1.0)[97], *M. esculenta* (v.1.0)[98], *C. annuum* (v.1.6)[99], *P. miliaceum* (v.2.0)[100], *E. coracana* (v.2.0)[101], *D. exilis* (DiExil)[102] and *S. viridis* (v.2.0)[103]. The iTAK tool (v.1.7a)[104] was used for transcription factor prediction with default parameters. To avoid bias caused by differences in the number of genes among the different plants[105], we calculated the proportion of transcription factor as N_{TF}/N_{total}, where N_{TF} is the number of transcription factors and N_{total} is the total number of genes in the corresponding plant. Moreover, we predicted the binding sites of RWP-RK transcription factors with the FIMO tool (v.5.3.2) (https://meme-suite.org/meme/meme_5.3.2/doc/fimo.html)[106].

Contributions of SVs to nearby gene expression

To investigate whether the SVs could broadly influence nearby gene expression, we used RNA-seq dataset B for the six accessions subjected to 1 h of control conditions (Supplementary Table 3). The SVs were divided into 11 categories: deletion of coding DNA sequence start (DELCDSstart); deletion overlapping the 5-kb downstream region (DELdown); deletion of exons (DELexons); deletion of the whole gene (DELgene); deletion overlapping the 5-kb upstream region (DELup); duplication (Dup); insertion in the 5-kb downstream region (INSdown); insertion in exons (INSexons); insertion in introns (INSintrons); insertion in the 5-kb upstream region (INSup); and the presence of SVs (PresenceSVs).

PAV–GWAS

To explore the usefulness of the graph-based genome and identify SV-driven alterations of genes controlling important agronomic traits, we conducted a PAV–GWAS analysis. After PAV filtration (removal of PAVs with a minor allele frequency < 0.05 or missing rate > 0.1), a total of 124,532 PAVs were used to perform PAV–GWAS in 242 accessions. Association analysis was conducted using the GEMMA (v.0.94.1) software package[107]. For the mixed linear model analysis, we used the equation $y = X\alpha + S\beta + K\mu + e$, where y represents the phenotype, X represents the genotype, S is the structure matrix and K is the relative kinship matrix. $X\alpha$ and $S\beta$ represent fixed effects and $K\mu$ and e represent random effects. The top three principal components were used to build the S matrix for population structure correction. The matrix of simple matching coefficients was used to build the K matrix.

Determination of physiological indicators

Seeds (2.00 g) of Tifleaf3 were cultured in a plastic box (10 × 15 × 6 cm) under growth conditions of 14 h light at 26 °C and 10 h darkness at 22 °C. The 13-day-old seedlings (V3 stage: third leaf visible at the vegetative stage) were divided into three groups: a high-temperature treatment group (45 °C under light for 14 h and 40 °C in darkness for 10 h), a heat treatment group (40 °C under light for 14 h and 35 °C in darkness for 10 h) and a control group (26 °C under light for 14 h and 22 °C in darkness for 10 h). After 1, 24, 96 and 144 h, and 11, 16, 21, 26, 31, 36 and 41 d of heat treatment or control conditions, leaves were subjected to the measurement of relative water content, relative conductivity and MDA content. In addition, the materials (PI537069, PI521612, PI526529, PI587025, PI583800 and Tifleaf3) used for pan-genome sequencing were cultured under the same conditions described above and divided into a high-temperature treatment group and a control group. After treatment for 1, 24, 60 and 96 h, leaves were collected for the determination of relative water content, electrical conductivity and MDA content. Transgenic rice and WT rice were cultured at 26 °C under light for 14 h and 22 °C in darkness for 10 h each day for 45 d and then divided into two groups: a high-temperature treatment group (45 °C under light for 14 h and 45 °C in darkness for 10 h) and a control group (26 °C under light for 14 h and 22 °C in darkness for 10 h). MDA content and POD and SOD enzyme activities were quantified in the plants after 12 h and 72 h of heat treatment.

Measurement of POD, MDA and REC

Leaves (0.1 g) were ground and 1.5 ml of PBS solution (150 mM) was added. The mixtures were centrifuged at 12,879.36*g* for 20 min at 4 °C. The supernatant was then collected. For the determination of MDA activity, 0.5 ml of enzyme extract was added to 1 ml of reaction solution (20% trichloroacetic acid and 0.5% thiobarbituric acid) and the mixture was incubated in a 95 °C water bath for 30 min. Thereafter, the mixture was placed in an ice bath at room temperature (25 °C) and centrifuged at 12,879.36*g* for 10 min. The absorbance was recorded at 532 nm and 600 nm using a spectrophotometer (Sorvall ST 16). For the determination of POD activity, a 1.5 ml reaction system was used. First, 925 µl sodium acetate (100 mM) was added, after which 0.5 ml guaiacol (0.25%) and 25 µl enzyme extract were added. After mixing, 50 µl of hydrogen peroxide (0.75%) was added to the mixture. The absorbance was recorded at 470 nm every 10 s. SOD enzymatic activity was determined as described by Dhindsa et al.[108]. Starting with 50 µl of crude enzyme solution, 1.1 ml of 50 mM phosphate buffer, 100 µl of 0.06 mM riboflavin, 100 µl of 195 mM L-methionine, 50 µl of 0.003 mM EDTA and 100 µl of 1.125 mM nitroblue tetrazolium were added. In addition, two tubes without enzyme extract were included as controls. The reaction was performed under 3000 lx light for 30 min and the reaction was terminated in the dark. Absorbance was recorded at 560 nm. For the measurement of REC, 0.1 g samples of fresh leaves were collected with six biological replicates. The leaves were wrapped using gauze and placed in a 50-ml Eppendorf tube and 20 ml of pure water was added

to completely cover the leaves. The tube was placed in an incubator at room temperature (25 °C). After 25 h, the S1 EC was measured and the sample was kept in a boiling water bath for 30 min. The S2 EC was measured when the water had cooled to room temperature (25 °C). The REC was calculated using the following equation: REC = S1/S2 × 100%.

Transcriptomic analyses of pearl millet under high temperature

Seeds (2.00 g) of six accessions of pearl millet were cultivated in a 10 × 15 × 6 cm plastic basin filled with quartz sand and placed in a growth chamber (26 °C under light for 14 h and 22 °C in darkness for 10 h). The culture conditions were as described by Sun et al.[109]. The V3 stage seedlings were equally divided into two groups: a high-temperature treatment group and a control (CK) group. The conditions of the high-temperature treatment group were 14 h under light at 45 °C and 10 h in darkness at 40 °C, while the CK group was cultured under unchanged conditions (26 °C and 22 °C). After 1 and 24 h of treatment, leaves were collected and stored at −80 °C. In addition, the seeds (2.00 g) of Tifleaf3 were grown under similar conditions and seedlings were divided into treatment and control groups as described above. The culture conditions of the heat-treated group were 14 h under light at 40 °C and 10 h in darkness at 35 °C; the control group was kept under unchanged conditions (26 °C and 22 °C). After treatment for 1, 3, 5, 7, 24, 48, 96 and 144 h, the roots and leaves of the seedlings were collected and stored at −80 °C. A total of 168 samples were collected and three biological replicates were set for each treatment and control. Each replicate consisted of the mixed tissues of 16 seedlings. To obtain the materials used for the annotation of gene structure, the ten accessions were planted in a greenhouse, with nine plants of each accession (26 °C under light for 14 h and 22 °C in darkness for 10 h). We collected leaves (three biological replicates), stems (one sample) and roots (one sample) 5 weeks after the planting of each accession to build 30 RNA-seq libraries. A Total RNA Kit (QIAGEN) was used to extract RNA from these samples to build a complementary DNA library (NEBNext Ultra Directional RNA Library Prep Kit for Illumina) in preparation for RNA-seq. After sequencing, the raw data were filtered with FastQC (v.0.11.9) (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/)[110]. Transcripts were quantified with the Kallisto (v.0.46.2)[111] software using PI537069 as a reference. Finally, DEGs ($|\log_2(\text{group 1/group 2})| \geq 1, P_{adj} < 0.05$) were identified with DESeq2 (v.1.26.0)[112]. GO and KEGG enrichment analyses were performed using the OmicShare tools (http://omicshare.com/tools) ($P < 0.05$). Moreover, for the processing of published maize and rice transcriptomic data, we downloaded raw reads from maize in the V3 stage under 38 °C (14 h under light and 10 h in darkness) stress and normal conditions (25 °C; 14 h under light and 10 h in darkness)[31] and raw data from rice in the V3 stage grown under either 45 °C (13 h under light and 11 h in darkness) stress or normal culture condition (25 °C; 13 h under light and 11 h in darkness)[32]. The same methods and parameters were applied to the RNA-seq analysis of published maize and rice data.

Transgenic plant validation

The *PMF0G00024.1* gene sequence was synthesized via synthetic gene sequence generation and was introduced to the pBWA (V)HS-CCDB vector under the control of the 35S promoter. Three hundred rice seeds without mildew spots that showed normal buds were sterilized with 75% alcohol for 1 min, soaked in sodium hypochlorite for 20 min, washed with sterile water three times and then placed into a culture medium to culture calluses. The culture was conducted under light at 26 °C for 20 d. In addition, a single *Agrobacterium* colony was cultured in medium in a shake flask to obtain an *Agrobacterium* resuspension with an OD_{600} of 0.2. The calluses were added to the *Agrobacterium* suspension step. After 10–15 min of infection, calluses were picked, placed in a cocultivation medium and incubated at 20 °C for 48–72 h. Subsequently, cultured calluses were transferred to a selection medium containing hygromycin and cultured for 20–30 d (26 °C in darkness) for the first selection. After the first selection, 180 calluses were transferred to a new culture medium and cultured for 7–10 d (26 °C in darkness) for the second selection step. Ninety callus tissues were obtained and differentiation and rooting were induced. Finally, a total of 20 seedlings were obtained. The resistant calluses were differentiated into seedlings and PCR detection was performed using the primers listed in Supplementary Table 20. The PCR-positive seedlings were transplanted into the soil (26 °C under light for 14 h and 22 °C in darkness for 10 h). When they reached the four-leaf stage, quantitative PCR with reverse transcription was performed with the primers *RWP1* and *RWP2*, with three technical repeats for each sample (Supplementary Table 20).

Dual luciferase assays to assess the interaction between *RWP-RK* and ER-related genes

The open reading frames of *RWP-RK* (*PMF0G00024.1*) were inserted into the pGreenII62-SK vector to generate effector plasmids. The promoter sequence of *PMA2G00107.1* was synthesized by Hzykang and then cloned into the pGreenII 0800-LUC vector to generate reporter plasmids. Effector and reporter plasmids were expressed in tobacco leaves, mediated by *Agrobacterium* injection. Tobacco leaves in the injection area were collected and fluorescence activity was measured using a luciferase assay kit (cat. no. DL101, Vazyme Biotech). The primers used in this section are shown in Supplementary Table 20.

Tobacco leaf transformation assays to assess the impact of SVs on gene expression

The promoter sequences were cloned into the T vector using the 5 min TA/Blunt-Zero Cloning Kit (cat. no.C601, Vazyme Biotech). We used PCR (enzyme mix, cat. no. P520, Vazyme Biotech) to add the vector sequence at the end of the promoter fragment and obtained the PBI121-GUS linearized vector (Supplementary Table 20). Circularization was performed according to the instructions of the Clone Kit (cat no. MC40101, Monad). The recombinant vectors were injected into *Nicotiana benthamiana* leaf cells using an *Agrobacterium*-mediated transfection system (GV3101). GV3101-pBI121-35s-GUS, GV3101-pBI121-Promoter-GUS and GV3101-pBI121-Promoter_SV-GUS were cultured to an OD_{600} of 0.6 before injection. Two hundred microliters of liquid from each treatment was infiltrated into the tobacco leaves. Gloves were changed after the infiltration of each construct to prevent contamination. Tobacco was pretreated at a high temperature for 24 h (40 °C for 8 h and 35 °C for 16 h) and then cultured under the same conditions for 2 d after injection. The blank group was cultured at 25 °C (8 h under light and 16 h in darkness) and sampled by injection. The histochemical staining and quantitative analysis of GUS in three independent biological replicates were performed as described by Jefferson et al.[113].

PCR validation of SVs

Genomic DNA was extracted from fresh leaves using a DP360 kit (TIANGEN) and PCR was performed using 2× Phanta Flash Master Mix (cat. no. P520, Vazyme Biotech). Five SVs were analyzed by PCR genotyping (condition: followed by 35 cycles of denaturation at 98 °C for 10 s, annealing at 60 °C for 5 s and extension at 72 °C for 5 s kb^{-1}) using the primers indicated in Supplementary Table 20.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The raw sequencing data and transcriptome data of PI186338, PI250656, PI343841, PI521612, PI526529, PI527388, PI537069, PI583800, PI587025 and Tifleaf3 have been deposited in the NCBI Sequence Read Archive under BioProject accession no. PRJNA749489, PRJNA689619 and PRJNA756390. The assemblies of ten pearl millet have been deposited in NCBI GenBank under the accession no. JAMZRY000000000

(PI343841), JAMOAQ000000000 (PI250656), JAMKQL000000000 (PI186338), JAMKQK000000000 (PI527388), JAJHQD000000000 (PI587025), JAIFIR000000000 (PI537069), JAINUP000000000 (Tifleaf3), JAINUO000000000 (PI583800), JAINUN000000000 (PI526529) and JAINUM000000000 (PI521612). These assemblies are also available at http://117.78.45.2:91/download. The raw genome assembly data are available under accession no. PRJNA749489. The transcriptomic data are available under accession nos. PRJNA749489, PRJNA689619 and PRJNA756390. The public RNA-seq data used were downloaded from the NCBI and the BioProject accession no. is PRJNA520822. The public resequencing data used were downloaded from the NCBI and the accession no. is SRP063925. Source data are provided with this paper.

Code availability

All the analysis tools used in this study have been published before as described in the Methods and Reporting Summary.

References

47. Liu, B. et al. Estimation of genomic characteristics by analyzing *k*-mer frequency in de novo genome projects, v2. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1308.2012 (2013).
48. Zhang, Q. et al. The genome of *Prunus mume*. *Nat. Commun.* **3**, 1318 (2012).
49. Dolezel, J., Greilhuber, J. & Suda, J. Estimation of nuclear DNA content in plants using flow cytometry. *Nat. Protoc.* **2**, 2233–2244 (2007).
50. 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).
51. Roach, M. J., Schmidt, S. A. & Borneman, A. R. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics* **19**, 460 (2018).
52. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
53. Li, D. et al. Population genomics identifies patterns of genetic diversity and selection in chicken. *BMC Genomics* **20**, 263 (2019).
54. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
55. Parra, G., Bradnam, K. & Korf, I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**, 1061–1067 (2007).
56. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* **6**, 11 (2015).
57. Nishimura, D. RepeatMasker. *Biotech. Softw. Internet Rep.* **1**, 36–39 (2000).
58. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, W265–W268 (2007).
59. Edgar, R. C. & Myers, E. W. PILER: identification and classification of genomic repeats. *Bioinformatics* **21**, i152–i158 (2005).
60. Price, A. L., Jones, N. C. & Pevzner, P. A. De novo identification of repeat families in large genomes. *Bioinformatics* **21**, i351–i358 (2005).
61. Hubley, R. & Smit, A. RepeatModeler; http://www.repeatmasker.org/RepeatModeler/
62. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
63. Ma, J. & Bennetzen, J. L. Rapid recent growth and divergence of rice nuclear genomes. *Proc. Natl Acad. Sci. USA* **101**, 12404–12410 (2004).
64. Altschul, S. F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
65. Yu, X.-J., Zheng, H.-K., Wang, J., Wang, W. & Su, B. Detecting lineage-specific adaptive evolution of brain-expressed genes in human using rhesus macaque as outgroup. *Genomics* **88**, 745–751 (2006).
66. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res.* **14**, 988–995 (2004).
67. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009).
68. Trapnell, C. et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7**, 562–578 (2012).
69. Grabherr, M. G. et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
70. Haas, B. J. et al. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666 (2003).
71. Stanke, M. & Morgenstern, B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res.* **33**, W465–W467 (2005).
72. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**, 78–94 (1997).
73. Majoros, W. H., Pertea, M. & Salzberg, S. L. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* **20**, 2878–2879 (2004).
74. Guigó, R. Assembling genes from predicted exons in linear time with dynamic programming. *J. Comput. Biol.* **5**, 681–702 (1998).
75. Johnson, A. D. et al. SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics* **24**, 2938–2939 (2008).
76. Haas, B. J. et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, R7 (2008).
77. Finn, R. D. et al. InterPro in 2017—beyond protein family and domain annotations. *Nucleic Acids Res.* **45**, D190–D199 (2017).
78. Finn, R. D. et al. Pfam: the protein families database. *Nucleic Acids Res.* **42**, D222–D230 (2014).
79. Zdobnov, E. M. & Apweiler, R. InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17**, 847–848 (2001).
80. Finn, R. D. et al. HMMER web server: 2015 update. *Nucleic Acids Res.* **43**, W30–W38 (2015).
81. Harris, M. A. et al. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* **32**, D258–D261 (2004).
82. Kanehisa, M. The KEGG database. *Novartis Found. Symp.* **247**, 91–103 (2002).
83. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
84. Marçais, G. et al. MUMmer4: a fast and versatile genome alignment system. *PLoS Comput. Biol.* **14**, e1005944 (2018).
85. Goel, M., Sun, H., Jiao, W.-B. & Schneeberger, K. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol.* **20**, 277 (2019).
86. Garrison, E. et al. Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat. Biotechnol.* **36**, 875–879 (2018).
87. Lamesch, P. et al. The *Arabidopsis* Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.* **40**, D1202–D1210 (2012).

88. Jiao, Y. et al. Improved maize reference genome with single-molecule technologies. *Nature* **546**, 524–527 (2017).
89. Vogel, J. P. et al. Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature* **463**, 763–768 (2010).
90. VanBuren, R. et al. Single-molecule sequencing of the desiccation-tolerant grass *Oropetium thomaeum*. *Nature* **527**, 508–511 (2015).
91. Lovell, J. T. et al. The genomic landscape of molecular responses to natural drought stress in *Panicum hallii*. *Nat. Commun.* **9**, 5213 (2018).
92. Studer, A. J. et al. The draft genome of the C_3 panicoid grass species *Dichanthelium oligosanthes*. *Genome Biol.* **17**, 223 (2016).
93. Kawahara, Y. et al. Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. *Rice* **6**, 4 (2013).
94. Cooper, E. A. et al. A new reference genome for *Sorghum bicolor* reveals high levels of sequence similarity between sweet and grain genotypes: implications for the genetics of sugar metabolism. *BMC Genomics* **20**, 420 (2019).
95. Beier, S. et al. Construction of a map-based reference genome sequence for barley, *Hordeum vulgare* L. *Sci. Data* **4**, 170044 (2017).
96. Bennetzen, J. L. et al. Reference genome sequence of the model plant *Setaria*. *Nat. Biotechnol.* **30**, 555–561 (2012).
97. Zhang, J. et al. Allele-defined genome of the autopolyploid sugarcane *Saccharum spontaneum* L. *Nat. Genet.* **50**, 1565–1573 (2018).
98. Bredeson, J. V. et al. Sequencing wild and cultivated cassava and related species reveals extensive interspecific hybridization and genetic diversity. *Nat. Biotechnol.* **34**, 562–570 (2016).
99. Kim, S. et al. Genome sequence of the hot pepper provides insights into the evolution of pungency in *Capsicum* species. *Nat. Genet.* **46**, 270–278 (2014).
100. Zou, C. et al. The genome of broomcorn millet. *Nat. Commun.* **10**, 436 (2019).
101. Hatakeyama, M. et al. Multiple hybrid de novo genome assembly of finger millet, an orphan allotetraploid crop. *DNA Res.* **1**, 39–47 (2018).
102. Wang, X. et al. Genome sequence and genetic diversity analysis of an under-domesticated orphan crop, white fonio (*Digitaria exilis*). *Gigascience* **10**, giab013 (2021).
103. Mamidi, S. et al. A genome resource for green millet *Setaria viridis* enables discovery of agronomically valuable loci. *Nat. Biotechnol.* **38**, 1203–1210 (2020).
104. Zheng, Y. et al. iTAK: a program for genome-wide prediction and classification of plant transcription factors, transcriptional regulators, and protein kinases. *Mol. Plant* **9**, 1667–1670 (2016).
105. Zhang, H., Zhao, Y. & Zhu, J.-K. Thriving under stress: how plants balance growth and the stress response. *Dev. Cell* **55**, 529–543 (2020).
106. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018 (2011).
107. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* **44**, 821–824 (2012).
108. Dhindsa, R. S., Plumb-Dhindsa, P. & Thorpe, T. A. Leaf senescence: correlated with increased levels of membrane permeability and lipid peroxidation, and decreased levels of superoxide dismutase and catalase. *J. Exp. Bot.* **32**, 93–101 (1981).
109. Sun, M. et al. Transcriptome sequencing revealed the molecular mechanism of response of pearl millet root to heat stress. *J. Agron. Crop Sci.* **207**, 768–773 (2021).
110. Bittencourt, S. A. *FastQC: a Quality Control Tool for High Throughput Sequence Data* (Babraham Institute, 2010); https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
111. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
112. Wang, L., Feng, Z., Wang, X., Wang, X. & Zhang, X. DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics* **26**, 136–138 (2010).
113. Jefferson, R. A., Kavanagh, T. A. & Bevan, M. W. GUS fusions: beta-glucuronidase as a sensitive and versatile gene fusion marker in higher plants. *EMBO J.* **6**, 3901–3907 (1987).

Acknowledgements

This work was supported by the earmarked fund for CARS (CARS-34 to L.H.), the Modern Agricultural Industry System Sichuan Forage Innovation Team (no. SCCXTD-20201-16 to L.H.), the Sichuan Province Research Grant (no. 2021YFYZ0013 to L.H.) and the National Natural Science Foundation of China (nos. 31771866 and 32071867 to L.H.). We thank X. Chen (State Key Laboratory of Crop Gene Exploration and Utilization in Southwest China, Sichuan Agricultural University), R. J. Schmitz (Department of Genetics, University of Georgia) and Q. Tang (College of Animal Science and Technology, Sichuan Agricultural University) for providing valuable suggestions on early versions of the manuscript. We thank G. Vellidis (The Georgia Coastal Plain Experiment Station) for providing some material resources.

Author contributions

L.H., H.Y. and S.T. designed and managed the project. M.S., C.L., A.Z., Y. Jin and B.W. participated in material collection and processing. Z.Z., M.S. and H.Y. performed the bioinformatics analyses. H.Y., M.S. and S.T. wrote the manuscript. Y. Jin, M.S. and C.L. contributed to the validation work. S.T., L.H., M.H., B.X., Jing Wang., Jianping Wang, P.Q., J.P.M., G.N., C.S.J., G.F., R.K.S., X.Z., A.B., Z.Z., B.W., A.Z., D.L., L.J., Y.P., X.W. and Y. Ji revised the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41588-023-01302-4.

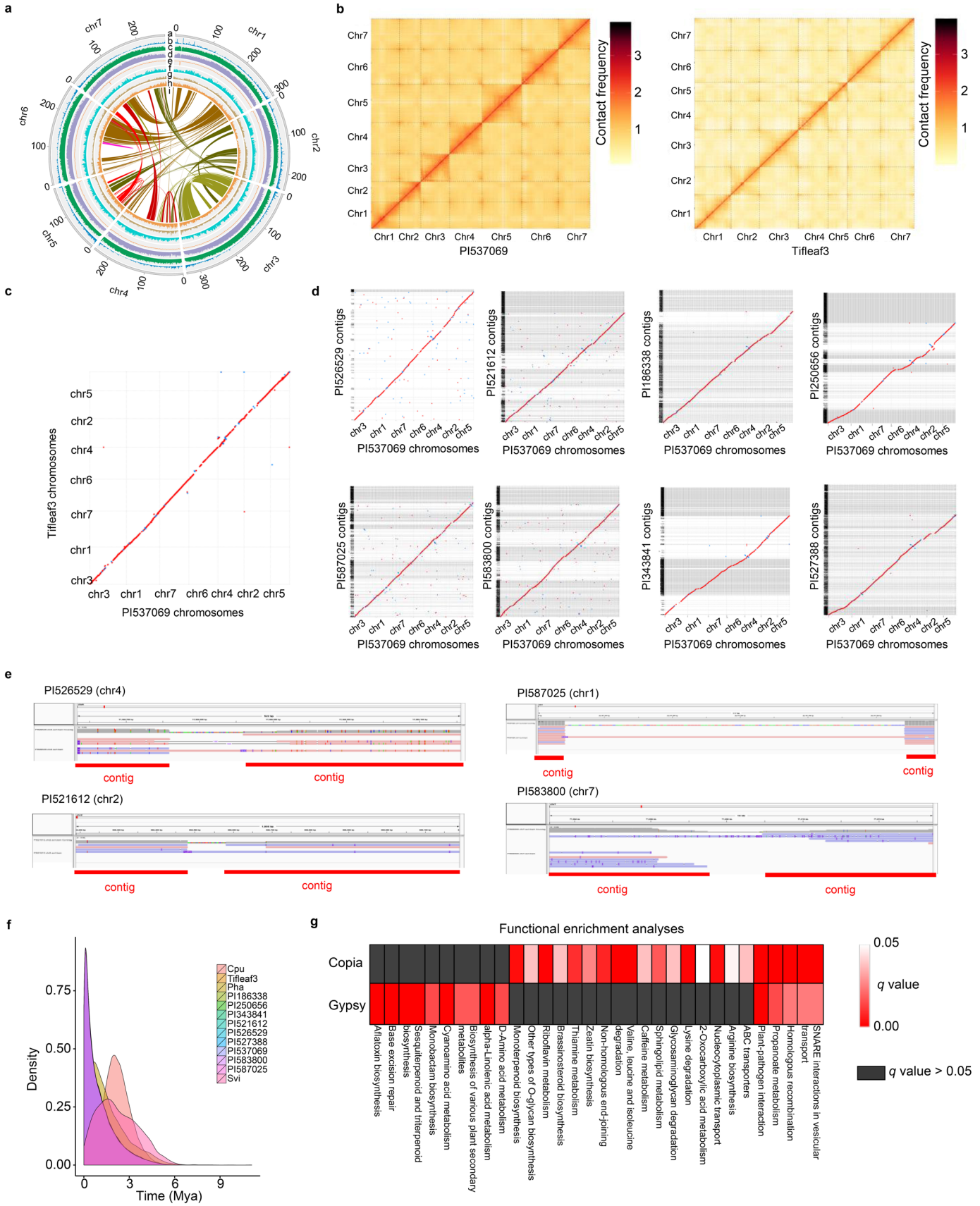
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-023-01302-4.

Correspondence and requests for materials should be addressed to Shilin Tian or Linkai Huang.

Peer review information *Nature Genetics* thanks the anonymous reviewers for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

https://doi.org/10.1038/s41588-023-01302-4

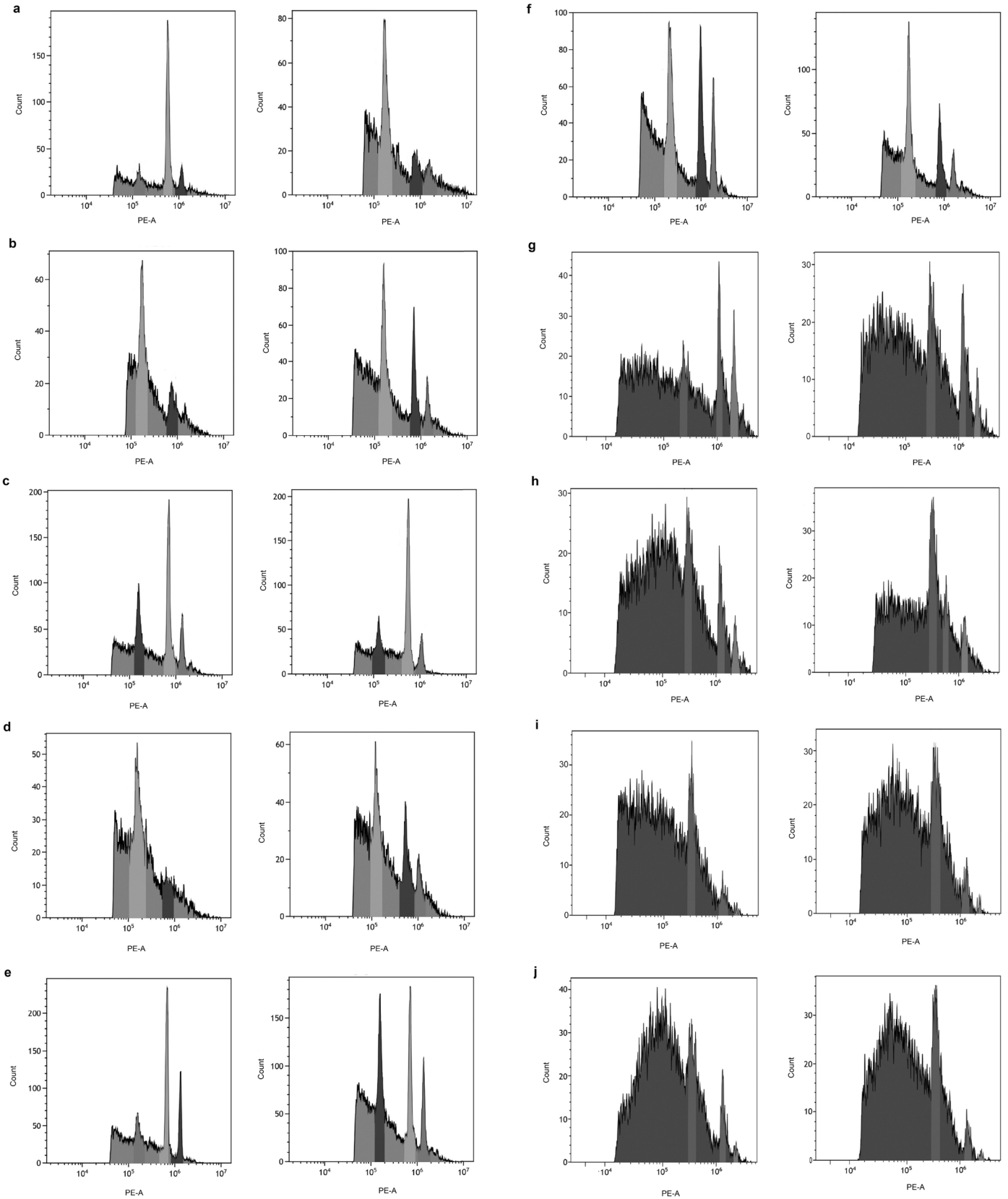


Extended Data Fig. 1 | See next page for caption.

https://doi.org/10.1038/s41588-023-01302-4

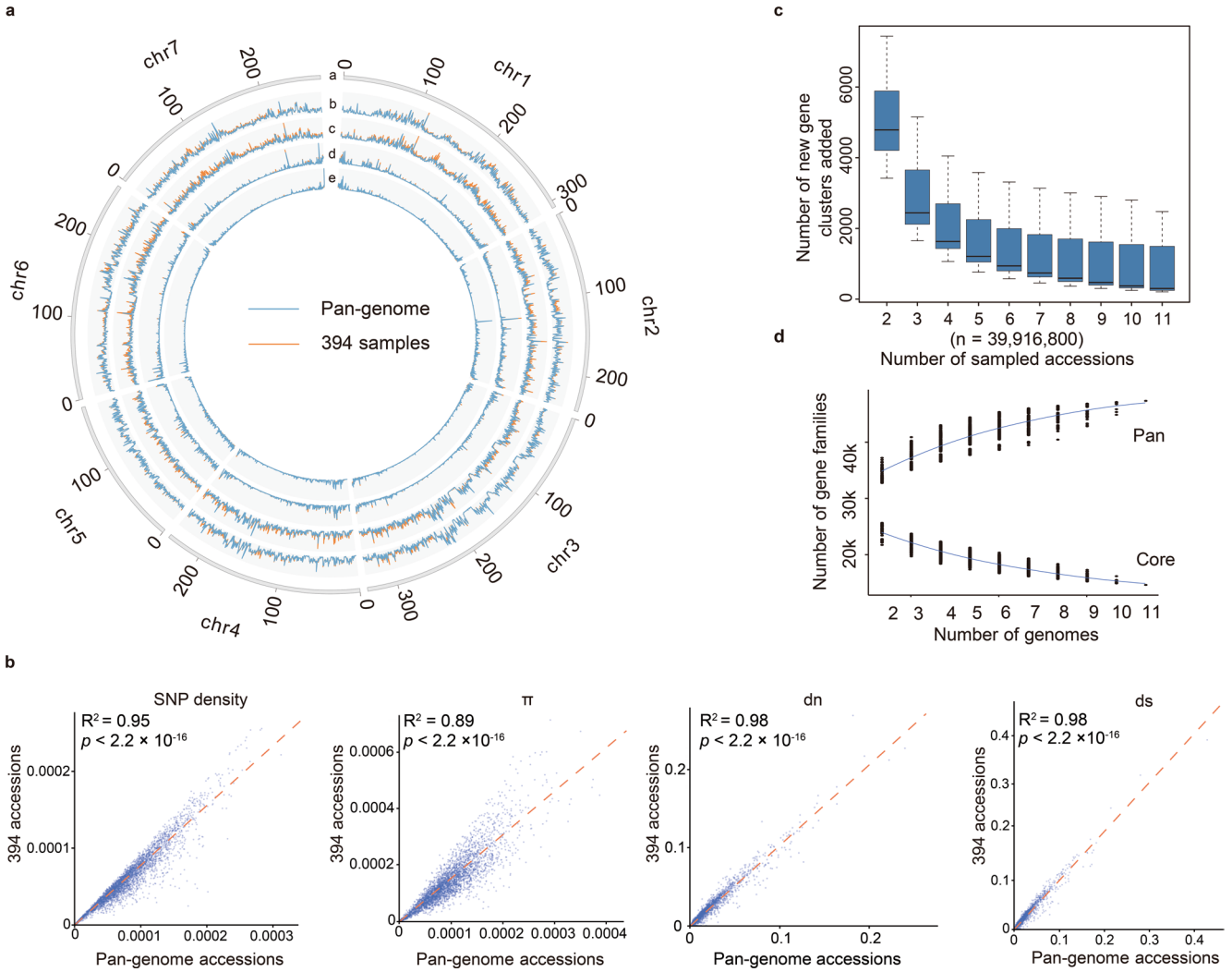
Extended Data Fig. 1 | High quality assembled genomes in pearl millet. **a**, Genome landscape of pearl millet (PI537069). Track 'a': the seven chromosomes at the Mb scale; track 'b': chromosomal distribution of gene models in which the gene density ranged from 201 to 503,365 bp/Mb; track 'c': chromosomal distribution of TEs for which the density was 58,334 to 945,223 bp/Mb; track 'd': TE/LTR distribution, ranging from 122,500 to 920,183 bp/Mb; track 'e': GC content along the assembled genome, which ranged between 20% and 60%/Mb; track 'f' and 'g': numbers of SNPs (5 to 27,243/Mb) and indels (1 to 5,952/Mb); track 'h': number of SVs (6 to 422/Mb); track 'i': collinear blocks of the pearl millet (PI537069) genome. **b**, Hi-C contact matrices of PI537069 and Tifleaf3 assemblies. **c**, Collinearity between the PI537069 and Tifleaf3 genomes. The contigs of Tifleaf3 were connected with Hi-C data only. This plot showed good synteny of the pseudochromosomes of Tifleaf3 and another assembly based on Hi-C data from PI537069. **d**, Synteny of eight contig-level assemblies with the PI537069 chromosome-level assembly. **e**, Assessment of the contig connections based on HiFi reads. Here the four examples show a long HiFi read across two contigs. **f**, Insertion times of LTRs in pearl millet and three related species, *Cenchrus purpureus* (Cpu), *Panicum hallii* (Pha) and *Setaria viridis* (Svi). **g**, Functional enrichment analyses of genes encompassed by intact Copia and Gypsy TEs.

https://doi.org/10.1038/s41588-023-01302-4



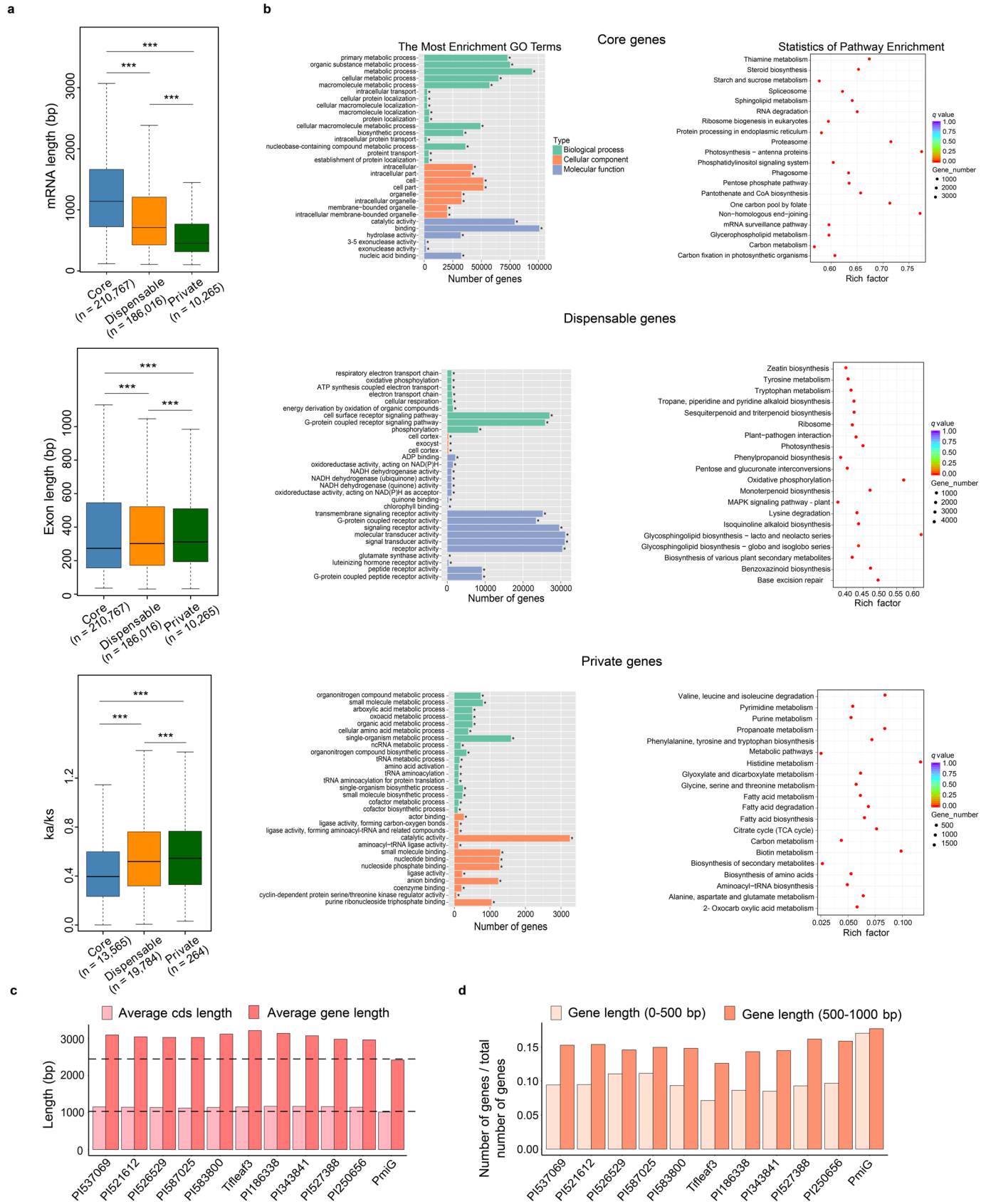
Extended Data Fig. 2 | Estimation of genome size using flow cytometry. Estimation of genome size using flow cytometry. **a-j**, Histogram of relative fluorescence intensities from flow cytometric analysis of Propidium iodide (PI)-stained nuclei of Tifleaf-3, PI521612, PI526529, PI537069, PI583800, PI587025, PI186338, PI250656, PI343841, PI527388, which were isolated, stained and analyzed simultaneously (two biological repeats). The *Oryza sativa* cv. Nipponbare genome (2n = 2x = ~420 Mb) served as an internal reference standard. The average ratio of peak value between pearl millet to *O. sativa* was equal 4.5, hence the estimated average genome size of pearl millet was ~1890 Mb. Source data are provided as a Source Data file.

https://doi.org/10.1038/s41588-023-01302-4



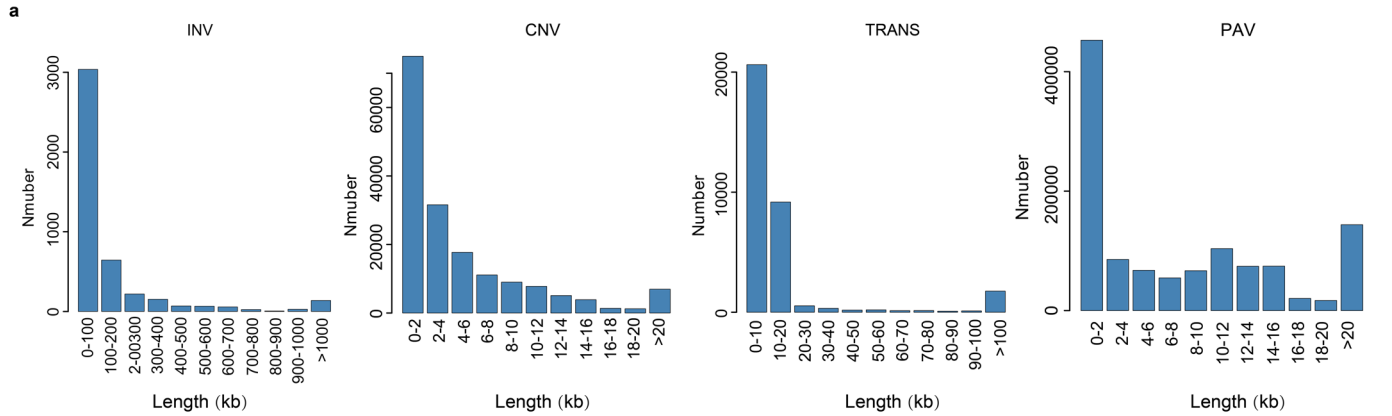
Extended Data Fig. 3 | Evaluation of the representativeness of the pearl millet pan-genome. a, Distribution of genetic variations from 11 genomes and 394 re-sequenced pearl millet accessions. Track 'a': the seven chromosomes at the Mb scale; track 'b': SNP density; track 'c', 'd', and 'e' represent nucleotide diversity (π), nonsynonymous (d_N), and synonymous (d_S) substitution rates, respectively. **b**, Correlation of SNP density, π, d_N, and ds from the 11 de novo assembled genomes and the 394 re-sequenced accessions. Significant differences were tested by two-tailed Pearson correlation test and shown by *p* value. **c**, Number of newly added gene families when adding more accessions. The median number of new gene families is 301 when the last accession comes in. Center line, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range. **d**, The variation in gene families in the core and pan-genomes along with additional pearl millet genomes.

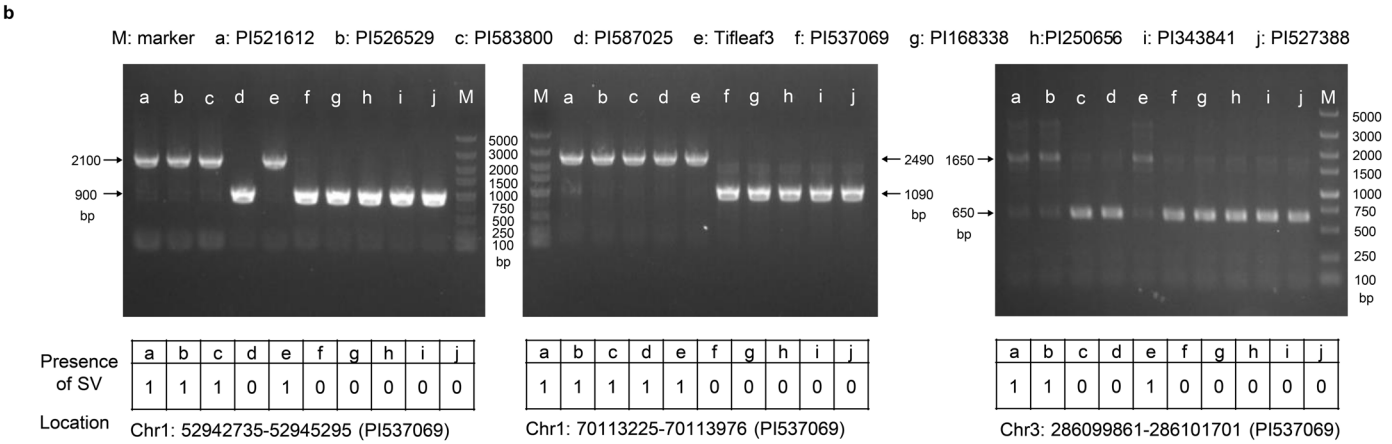
https://doi.org/10.1038/s41588-023-01302-4



Extended Data Fig. 4 | Characterization of core, dispensable, and unique gene sets. a, mRNA length, exon length, and ka/ks values of the three gene categories. Center line, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range; and dots represent outliers. Significant differences were tested by two-tailed *t*-test (***$P < 0.0005$). **b**, Gene Ontology (GO) (left) and Kyoto Encyclopedia of Genes and Genomes (KEGG) (right) enrichment results. Significant differences were tested by hypergeometric test (*FDR-adjusted $P < 0.05$). **c**, Comparison of average CDS and gene length across all accessions. The dotted lines indicate average CDS and gene lengths in the PmiG genome. **d**, Comparison of proportion of genes in two length ranges across all accessions.

https://doi.org/10.1038/s41588-023-01302-4





	a	b	c	d	e	f	g	h	i	j
Presence of SV	1	1	1	0	1	0	0	0	0	0
Location	Chr1: 52942735-52945295 (PI537069)									

a	b	c	d	e	f	g	h	i	j
1	1	1	1	1	0	0	0	0	0

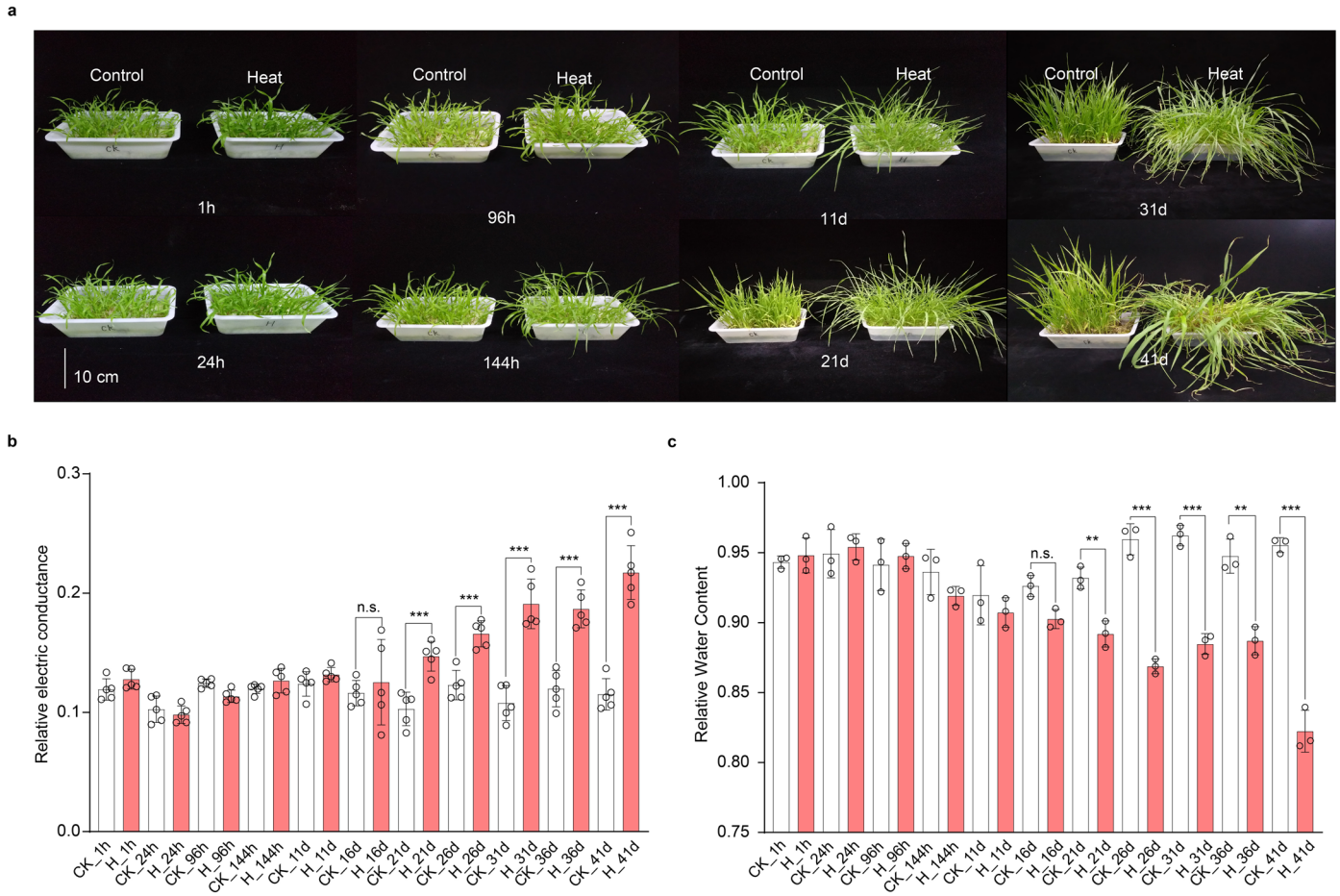
Chr1: 70113225-70113976 (PI537069)

a	b	c	d	e	f	g	h	i	j
1	1	0	0	1	0	0	0	0	0

Chr3: 286099861-286101701 (PI537069)

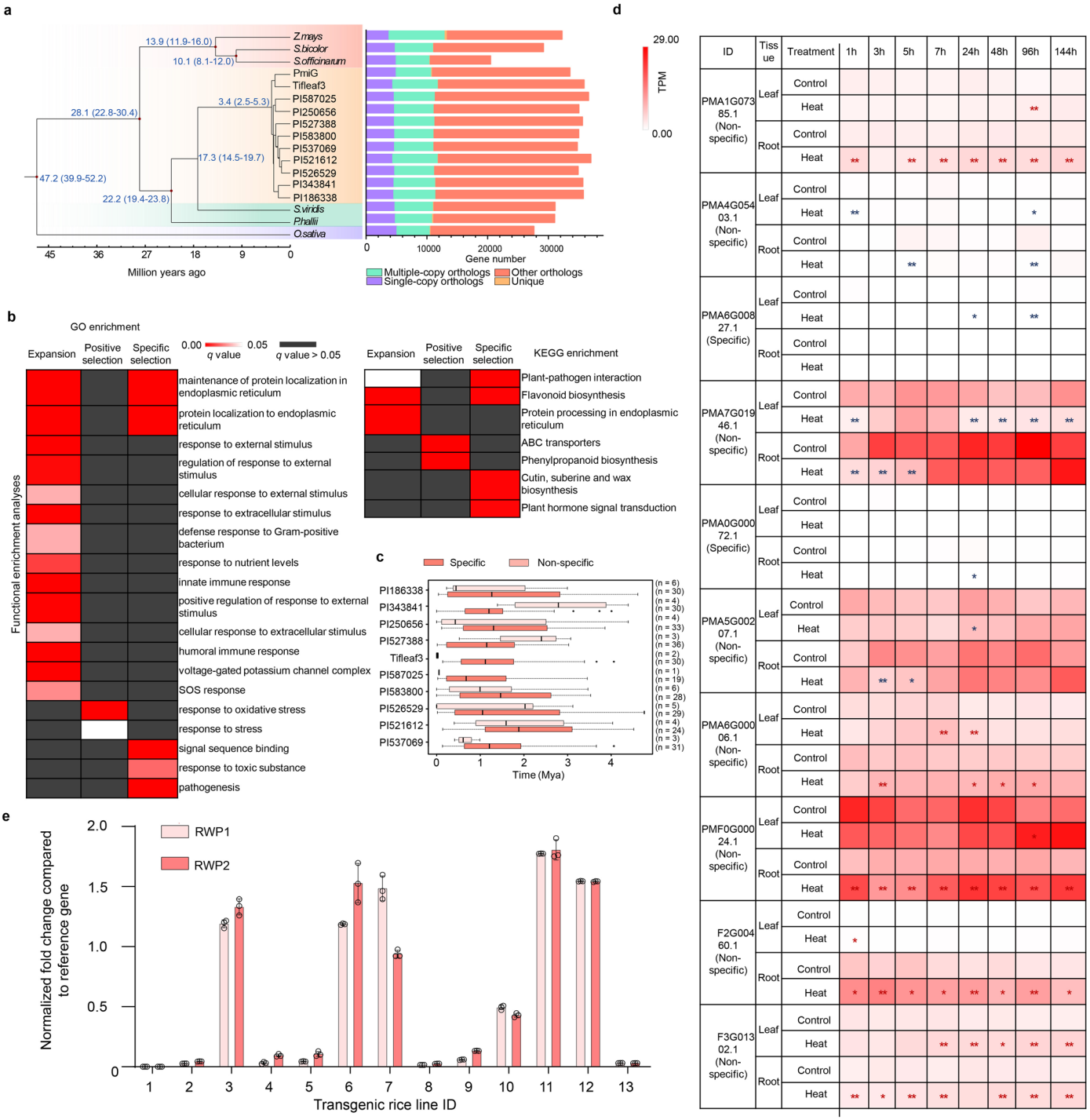
Extended Data Fig. 5 | Characterization of structural variations (SVs). a, Length distributions of four major SV types. PAV, TRANS, CNV, INV: presence and absence variation, translocation, copy number variation, and inversion, respectively. **b**, PCR validation of three SVs in 10 pearl millet accessions. Each experiment is performed once. The images were cropped from the images in the Source data file. Source data are provided as a Source Data file.

https://doi.org/10.1038/s41588-023-01302-4



Extended Data Fig. 6 | Phenotypes and physiological indicators under heat stress conditions. a, Phenotypic changes in pearl millet (Tifleaf3) under heat stress (40°C/35°C) at eight time points. Scale bar indicates 10 cm. **b-c**, Relative conductance (b) and relative water content (c) of the Tifleaf3 accession at 11 time points under control and heat treatments. Error bars, mean ± s.d.; n = 5 biological replicates for b, and n = 3 biological replicates for c. Significant differences were tested by two-tailed *t*-test (*$P < 0.05$, **$P < 0.01$, ***$P < 0.0005$; n.s., not significant).

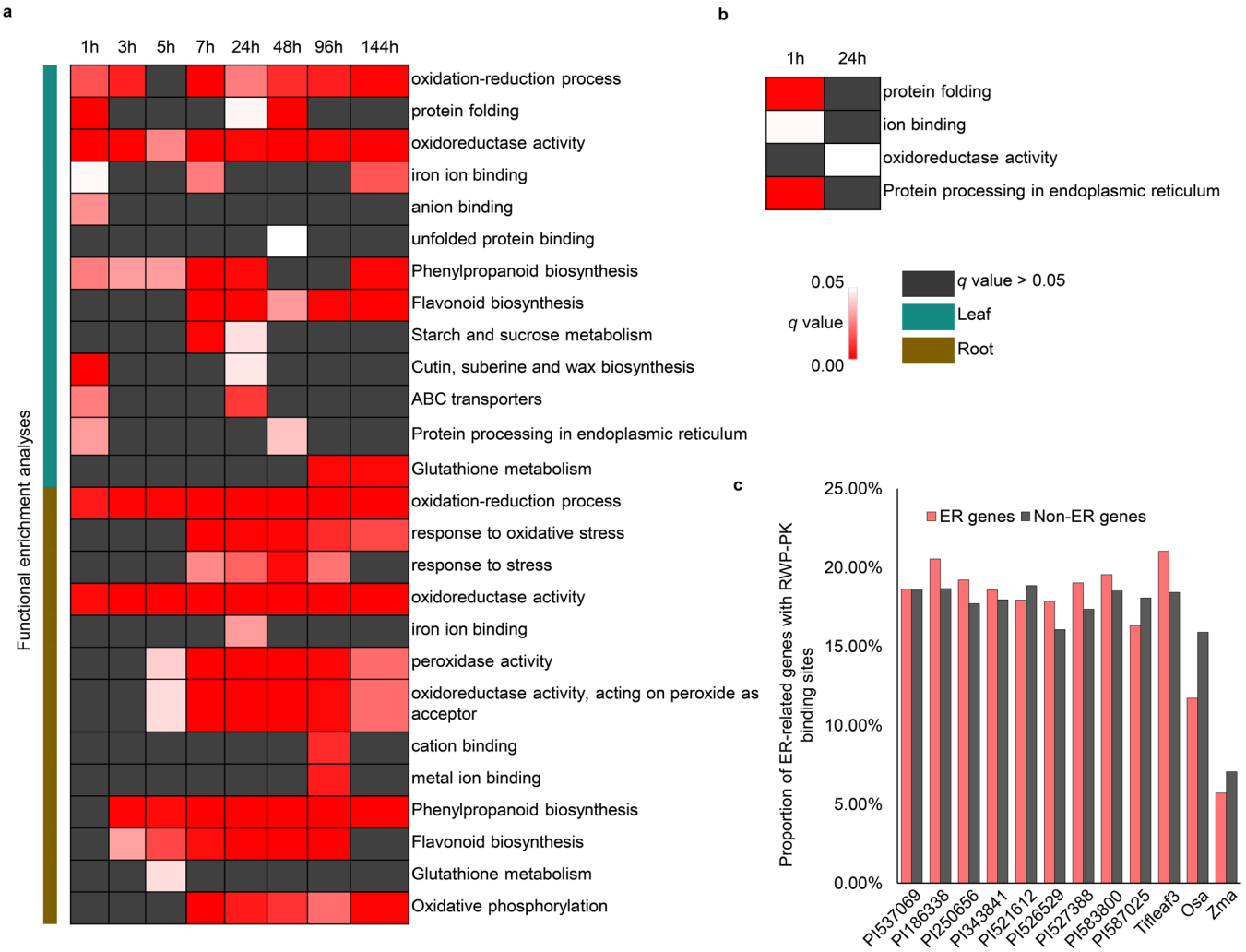
https://doi.org/10.1038/s41588-023-01302-4



ID	Tissue	Treatment	1h	3h	5h	7h	24h	48h	96h	144h
PMA1G07385.1 (Non-specific)	Leaf	Control								
		Heat							**	
	Root	Control								
		Heat	**		**	**	**	**	**	**
PMA4G05403.1 (Non-specific)	Leaf	Control								
		Heat	**						*	
	Root	Control								
		Heat			**				**	
PMA6G00827.1 (Specific)	Leaf	Control								
		Heat					*		**	
	Root	Control								
		Heat								
PMA7G01946.1 (Non-specific)	Leaf	Control								
		Heat	**				**	**	**	**
	Root	Control								
		Heat	**	**	**					
PMA0G00072.1 (Specific)	Leaf	Control								
		Heat								
	Root	Control								
		Heat					*			
PMA5G00207.1 (Non-specific)	Leaf	Control								
		Heat					*			
	Root	Control								
		Heat		**	*					
PMA6G00006.1 (Non-specific)	Leaf	Control								
		Heat				**	**			
	Root	Control								
		Heat		**			*	*	*	
PMF0G00024.1 (Non-specific)	Leaf	Control								
		Heat							*	
	Root	Control								
		Heat	**	**	**	**	**	**	**	**
F2G00460.1 (Non-specific)	Leaf	Control								
		Heat	*							
	Root	Control								
		Heat	*	**	*	*	**	*	**	*
F3G01302.1 (Non-specific)	Leaf	Control								
		Heat				**	**	*	**	**
	Root	Control								
		Heat	**	*	**	**		**	**	**

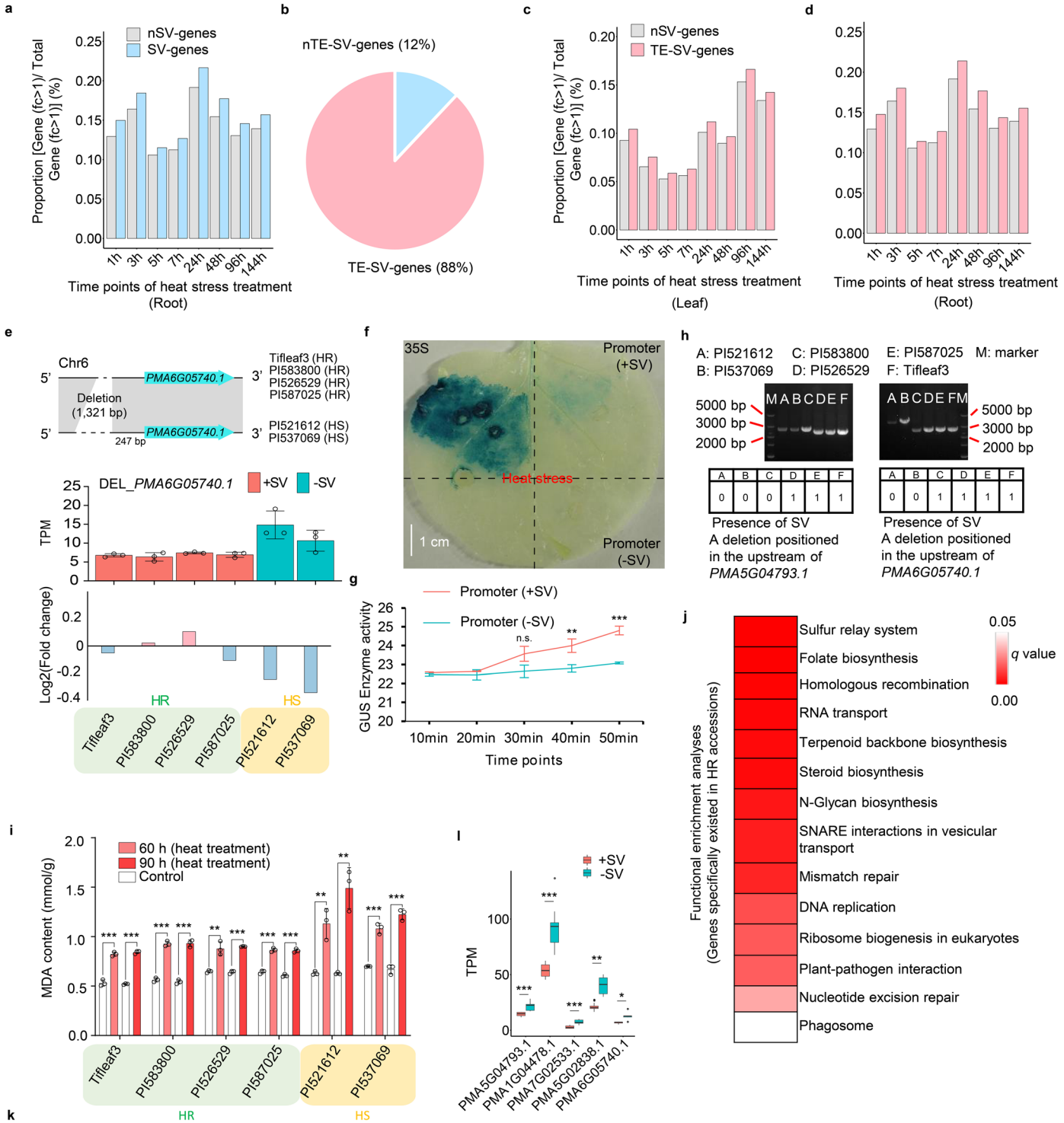
Extended Data Fig. 7 | Expansion and expression of RWP-RK transcription factor (TF) family members. a, Estimation of the divergence times of seven pearl millet accessions and six closely related species. PmiG represents Tift 23D2B1-P1-P5. The right panel displays the distribution of single-copy, multiple-copy, unique and other gene orthologs. **b**, Functional enrichment analyses showed the enrichment of some genes in stress-related processes or pathways among expanded family, specific, and positively selected gene sets. **c**, Estimated insertion times of LTR TEs encompassing the specific RWP-RKs of pearl millet belonging to clade A and the nonspecific RWP-RKs belonging to clade B. Center line, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range; and dots represent outliers. **d**, Expression of seven RWP-RK genes under control (CK) and heat treatments (H). Significant differences were performed by DESeq2 to determine fold change (FC) of gene expressions between the CK and H groups (two-tailed Wald test). The FC was determined by * ($0 < |\log2FC| < 1$, FDR-adjusted $P < 0.05$) or ** ($|\log2FC| >= 1$, FDR-adjusted $P < 0.05$). The blue and red */** represent gene down- and upregulation, respectively. **e**, Expression of two RWP fragments (RWP1 and RWP2) in 13 transgenic rice lines. A value above 0 indicates the upregulation of this gene. Error bars, mean ± s.d.; n = 3 biological replicates.

https://doi.org/10.1038/s41588-023-01302-4



Extended Data Fig. 8 | Gene functional enrichment analyses in transcriptome analyses. a, Functional enrichment analyses of differentially expressed genes (DEGs) identified in all six pearl millet materials after heat stress. **b**, Functional enrichment analyses of DEGs under continuous heat treatment at eight time points. **c**, Proportions of ER-related and non-ER-related genes in which RWP-RK binding sites are found in the promoter region (3 kb).

https://doi.org/10.1038/s41588-023-01302-4

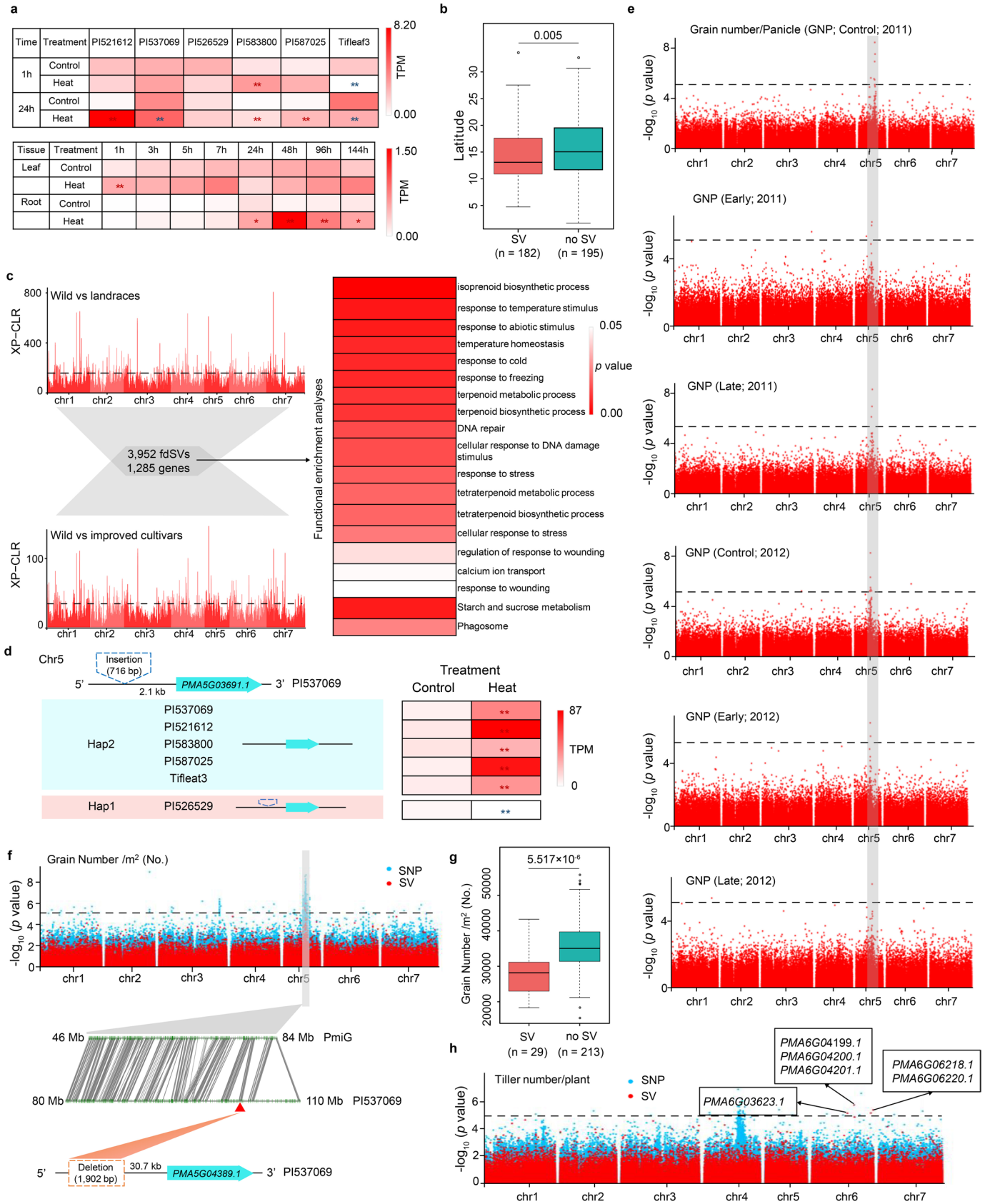


Extended Data Fig. 9 | See next page for caption.

https://doi.org/10.1038/s41588-023-01302-4

Extended Data Fig. 9 | Potential contributions of structural variations (SVs) to heat tolerance in pearl millet. a, Proportion of genes with expression fold changes over 1 compared to total gene expression. nSV-genes and SV-genes: genes not located near SVs (over 5 kb) and genes located near SVs (within 5 kb), respectively. **b**, Composition of genes close to SVs overlapping with transposons (TE-SV-genes) or not overlapping with transposons (nTE-SV-genes). **c-d**, These two panels are similar to panel **a** but compares TE-SV-genes to nSV-genes. **e**, One example shows the presence of a fixed SV in the HR group near *PMA6G05740.1*. The lower panel shows the gene expression changes in *PMA6G05740.1* under 24 h heat treatment in the HR and HS groups. Error bars, mean ± s.d.; n = 3 biological replicates. TPM: transcripts per million. '+SV' and '-SV': accessions with and without SVs, respectively. **f-g**, Transformation of the *PMA6G05740.1* promoter in tobacco leaves. f: GUS phenotype observed by histochemical staining. Scale bar indicates 1 cm. g: Quantitative detection of GUS enzyme levels with a fluorescence microplate in leaves. Error bars, mean ± s.d.; n = 3 biological replicates. Significant differences were tested by two-tailed *t*-test (*$P < 0.05$, ***$P < 0.0005$; n.s., not significant). **h**, PCR validation of two SVs positioned in upstream regions of *PMA5G04793.1* and *PMA6G05740.1*. Each experiment is performed once. The images were cropped from the images in the Source data file. **i**, Malondialdehyde (MDA) levels of four heat-resistant (HR) and two heat-susceptible (HS) accessions. Error bars, mean ± s.d.; n = 3 biological replicates. Significant differences were tested by two-tailed *t*-test (**$P < 0.01$, ***$P < 0.0005$). **j**, Functional enrichment analyses of genes specifically found in the HR group. **k**, Pipeline of identifying focal candidate SVs potentially related to expression changes of nearby heat-related genes. **l**, Comparisons of the expression (TPM) of five genes between accessions with (+SV) and without SVs (-SV). These five genes are from panel e in this figure and from Fig. 5c and g. *PMA5G04793.1*, *PMA1G04478.1*, and *PMA7G02533.1*: n = 9 biological replicates for '+SV' and '-SV', respectively; *PMA5G02838.1* and *PMA6G05740.1*: n = 12 biological replicates for '+SV' and n = 6 for '-SV'. Center line, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range; and dots represent outliers. Significant differences were tested by two-tailed *t*-test (*$P < 0.05$, **$P < 0.01$, ***$P < 0.0005$). Source data are provided as a Source Data file.

https://doi.org/10.1038/s41588-023-01302-4



Extended Data Fig. 10 | See next page for caption.

https://doi.org/10.1038/s41588-023-01302-4

Extended Data Fig. 10 | Contributions of structural variations (SVs) to heat tolerance adaptation and domestication. a, Expression patterns of *PMA2G02653.1*. Significant differences were performed by DESeq2 to determine fold change (FC) of gene expressions between control and heat-treatment groups (two-tailed Wald test). The FC was determined by * ($0 < |log2FC| < 1$, FDR-adjusted $P < 0.05$) or ** ($|log2FC| >= 1$, FDR-adjusted $P < 0.05$). The blue and red */** represent gene down- and upregulation, respectively. **b**, Comparison of the latitudinal distributions of the two SV-related haplotypes of pearl millet accessions. Center line, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range; and dots represent outliers. Significant differences were tested by two-tailed *t*-test and shown by *p* value. **c**, SV-based selection sweep analyses. A total of 3,952 SVs with population frequency differences (fdSVs) were identified in both comparisons and harbored 1,285 genes. The black dotted line represents a cut-off window in which the top 1% of data points were selected as the sweep region. The right panel represents the functional enrichment analyses of the 1,285 genes. **d**, Expression levels of one candidate gene (*PMA5G03691.1*)-related haplotype. Significant differences were performed by DESeq2 to determine fold change (FC) of gene expressions between control and heat-treatment groups (two-tailed Wald test). The FC was determined by * ($0 < |log2FC| < 1$, FDR-adjusted $P < 0.05$) or ** ($|log2FC| >= 1$, FDR-adjusted $P < 0.05$). The blue and red ** represent gene down- and upregulation, respectively. **e**, PAV-GWAS of SVs associated with grain number/panicle (GNP). Control: condition without stress. Early: early drought stress inhibiting irrigation from one week before flowering until maturity. Late: late drought stress conducted during early grain-filling by inhibiting irrigation from 50% flowering time until maturity. The regions marked by grey box indicate the GNP-related QTL could be captured by different conditions. **f**, PAV-GWAS of SVs associated with the grain number/m^2 (GNM2) trait. **g**, Comparison of GNM2 between accessions with SVs and without SVs. Center line, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range; and dots represent outliers. Significant differences were tested by two-tailed *t*-test and shown by *p* value. **h**, PAV-GWAS of SVs associated with the tiller number/plant (Till) trait. PmiG: Tift 23D2B1-P1-P5. The dotted-line represents the significance threshold of -log10 (*p* value) > 5.

nature portfolio

Corresponding author(s): Linkai Huang, Shilin Tian

Last updated by author(s): Jan 11, 2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used in data collection.
Data analysis	Flow cytometry analysis: Kaluza (v2.1.3). Initial assembly: Hifiasm package (v0.13-r308), Pruge_haplotig (v1.1.0), Bionano Solve (v3.5.1). Pseudochromosome construction: BWA (v0.7.8), ALLHIC package (v0.9.8). Genome assessment: BUSCO (v4.1.2), CEGMA (v2.5), BWA (v0.7.8), Merqury (v1.3), LTR_retriever (v2.8). Annotation of repetitive sequences: RepeatMasker (v4.0.5), LTR_FINDER (v1.0.7), Piler (v3.3.0), RepeatScout (v1.0.5), RepeatModeler (v1.0.8), MUSCLE (v3.8.31). Annotation of gene structure: TblastN (v2.2.26), Solar (v0.9.6), GeneWise (v2.4.1), TopHat (v2.0.13), Cufflinks (v2.1.1), Trinity (v2.1.1), PASA, Augustus (v3.2.3), GENSCAN (v1.0), GlimmerHMM (v3.0.1), EVidenceModeler (v1.1.1), SNAP (v2013.11.29), geneid (v1.4). Functional annotation of protein-coding genes: InterProScan (v4.8), HMMER (v3.1), InterPro (v32.0), Pfam (v27.0). Comparative genomic analysis across species: BLASTP (v2.2.26), Orthofinder (v2.3.1), MUSCLE (v3.8.31), RAxML (v8.0.19), MCMCTree program (v4.5). Pan-genome construction: Orthofinder (v2.3.1). SV identification: MUMmer (v4.0.0), SyRI (v1.6.3), vg (v1.25.0). Validation of structural variations: SyRI (v1.6.3), Assmeblytics, smartie-sv, vg (v1.25.0), HISAT2 (v2.2.1). TF family identification and analysis: Fimo (v5.3.2), iTAK tool (v1.7 a). PAV-GWAS: GEMMA (v0.94.1), LightGBM. RNA-seq: FastQC (v0.11.9), Kallisto (v0.46.2), DESeq2 (1.26.0).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The raw sequencing data and transcriptome data of PI186338, PI250656, PI343841, PI521612, PI526529, PI527388, PI537069, PI583800, PI587025, and Tifleaf3 have been deposited in the NCBI Sequence Read Archive under BioProject accession numbers PRJNA749489, PRJNA689619, and PRJNA756390. The assemblies of ten pearl millet have been deposited in NCBI GenBank under the accession numbers JAMZRY000000000 (PI343841), JAMOAQ000000000 (PI250656), JAMKQL000000000 (PI186338), JAMKQK000000000 (PI527388), JAJHQD000000000 (PI587025), JAIFIR000000000 (PI537069), JAINUP000000000 (Tifleaf3), JAINUO000000000 (PI583800), JAINUN000000000 (PI526529), and JAINUM000000000 (PI521612). These assemblies are also available at a website (http://117.78.45.2:91/download). The raw genome assembly data are available under accession number PRJNA749489. The transcriptomic data are available under accession numbers PRJNA749489, PRJNA689619, and PRJNA756390. The public RNA-seq data used was downloaded from NCBI and the bioproject accession numbers is PRJNA520822. The public re-sequence data used was downloaded from NCBI and the accession number is SRP063925. Source data are provided with this paper.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	not applicable
Population characteristics	not applicable
Recruitment	not applicable
Ethics oversight	not applicable

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Bionano: 1 sample; Hi-C: 2 samples; Pacbio HiFi: 10 samples; Illumina: 228 samples. no sample size calculation was performed. We built the pan-genome based on 11 representative accessions where 10 samples are de-novo assembled in our study and one sample downloaded from a published study. We used Bionano and Hi-C sequencing for PI537069 accession, aiming to obtain a high-quality assemble that could be used as the reference genome for the SV discoveries in the downstream analysis. For the 228 samples of Illumina sequencing, we did bulk RNA-seq analyses including leaf and root tissues and eight time points underlying heat stressful conditions (Supplementary Table 1: Overview of RNA-seq).
Data exclusions	For PAV-GWAS we excluded samples without phenotype data. For temperature adaptation analyses, we excluded samples without latitude data.
Replication	Three biological and three technical replicates for Dual-luciferase assays. Two biological and one technical replicates for Tobacco leaf transformation assays. One biological and technical replicate for PCR validation. Three biological replicates for physiological analysis. Two replicates for flow cytometry.
Randomization	Plants were randomly allocated in the greenhouse. tobacco leaves were randomly collected from individuals with same growth stages. For evaluation of contig connections, we randomly picked several to present in the Extended Fig. 1e. To further validate the SVs, we performed a PCR genotyping to validate three SVs randomly picked from the SV pool. For RNA-seq, the leaves or roots of 16 seedlings with consistent growth were randomly selected and stored in cryogenic vials. For Physiological indicators, the leaves of plants with consistent growth were randomly selected and stored in cryogenic vials.
Blinding	The experiments were conducted blindly. All genotypes were only labeled by numbers when planting, so the investigators did not know the

Blinding | exact accession names

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☐	☒ Flow cytometry
☒	☐ MRI-based neuroimaging

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☐ All plots are contour plots with outliers or pseudocolor plots.

☐ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	About 20mg leaves, add 1ml MGB, add 500µl lysis buffer, 25µl 50µg/ml PI and 25µl 50µg/ml RNase, mix and shade before use.
Instrument	Beckman CytoFLEX.
Software	CytExpert (version:2.3.0.84).
Cell population abundance	CytoFLEX flow cytometer automatically collects cells and counts the number.
Gating strategy	Use FSC-A/SSC-A to select cells, use PE-A/PE-H to exclude cell debris, and select the location of the positive result of PI staining.

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.