G3, 2023, 13(11), jkad196
https://doi.org/10.1093/g3journal/jkad196
Advance Access Publication Date: 31 August 2023
Plant Genetics and Genomics

The *Ga1* locus of the genus *Zea* is associated with novel genome structures derived from multiple, independent nonhomologous recombination events

Amruta R. Bapat,[1,2] Adrienne N. Moran Lauter,[3] Matthew B. Hufford,[4] Nicholas A. Boerman,[2] M. Paul Scott[3,*]

[1]Interdepartmental Genetics and Genomics Program, Iowa State University, Ames, IA 50011, USA
[2]Department of Agronomy, Iowa State University, Ames, IA 50011, USA
[3]Corn Insects and Crop Genetics Research Unit, USDA-ARS, Ames, IA 50011, USA
[4]Department of Ecology, Evolution and Organismal Biology, Iowa State University, Ames, IA 50011, USA

*Corresponding author: Corn Insects and Crop Genetics Research Unit, USDA-ARS, 716 Farmhouse Lane, Ames, IA 50011, USA. Email: paul.scott@usda.gov

The *Ga1* locus controls cross-incompatibility between field corn and popcorn. The *Ga1-S* haplotype contains 2 types of pectin methylesterase (PME) genes, *ZmPme3* and several copies of *ZmGa1P* that are expressed in silk and pollen, respectively. The *ga1* haplotype contains nonfunctional tandem repeat sequences related to *ZmPme3* and *ZmGa1P*. This haplotype can cross-pollinate freely and is widely present in field corn. The primary objective of this study is to characterize the repeat sequences from a diverse collection of maize and teosinte lines and use this information to understand the evolution of the *Ga1* locus. First, we characterized the complexity of the *Ga1* genome region in high-quality maize genome assemblies that led to their categorization into 5 groups based on the number and type of PME-like sequences found at this region. Second, we studied duplication events that led to the *ga1* and *Ga1-S* repeats using maximum likelihood phylogenetic reconstruction. Divergence estimates of the *ga1* haplotype suggest that the duplication events occurred more than 600 KYA whereas those in *Ga1-S* occurred at 3 time points, i.e. >600, ~260, and ~100 KYA. These estimates suggest that the *ga1* and *Ga1-S* tandem duplication events occurred independently. Finally, analysis of *ZmPme3* and *ZmGa1P* homologs in *Zea* and *Tripsacum* genomes suggests that *ga1* and *Ga1-S* repeats originated from an ancestral pair of PME genes that duplicated and diverged through 2 evolutionary branches prior to the domestication of maize.

Keywords: repeated sequences; evolution; pectin methylesterase; pseudogenes; Plant Genetics and Genomics

Introduction

The *Ga1* locus maps to the short arm of maize chromosome 4. The locus contains 2 genes that regulate cross-incompatibility. *ZmPme3* encodes a pectin methylesterase (PME) expressed in silks (Moran Lauter *et al.* 2017; Wang *et al.* 2022; Zhang, Li, *et al.* 2023) that interferes with pollen tube growth, preventing pollination by maize varieties that do not carry a functional version of the second gene of the *Ga1* locus. The second gene is called *ZmGa1P* (Zhang *et al.* 2018) and also encodes a PME. This gene is expressed in pollen, and pollen carrying this gene can overcome the barrier to cross-pollination created by *ZmPme3*. Wang *et al.* (2022) discovered that in addition to the single *ZmGa1P* gene reported initially, 4 additional tandem repeated sequences of *ZmGa1P* constitute the male function and were designated as *ZmGa1P*s-m. More such full-length duplicates of *ZmGa1P* were discovered, and now, a total of 8 functional *ZmGa1P* genes are reported to constitute the male function (Zhang, Li, *et al.* 2023). Similarly, 3 alleles of the *Ga1* locus have been defined based on which of these 2 genes is functional; for example, *Ga1-S* carries functional *ZmPme3* and *ZmGa1P*, while *ga1* carries neither. *Ga1-M* carries a functional *ZmGa1P* but lacks a functional *ZmPme3* (Lu *et al.* 2020). Two other unilateral cross-incompatibility systems called *Ga2* and *Tcb1* are functionally equivalent but not compatible with *Ga1* and map to different genetic loci. The *Ga2* locus has been mapped to a 1.7-Mb region on maize chromosome 5 (Chen, Luo, *et al.* 2022). The *Tcb1* locus is present on chromosome 4, about 44 cM away from the *Ga1* locus (Evans and Kermicle 2001). The female function gene of the *Tcb1* locus, *Tcb1-f*, was described by Lu *et al.* (2019) and encodes a PME protein that differs from ZmPME3 in 9 amino acids. The male function of the *Tcb1* locus, also a PME gene, has been identified recently (Zhang, Li, Zhang, and Chen 2023).

Intriguingly, the genome region around *Ga1* locus has an unusual structure. Maize lines carrying the *ga1* haplotype lack functional copies of either of the 2 *Ga1* genes and have multiple pseudogenes related to each of the 2 active genes of the *Ga1-S* allele. In contrast, the haplotypes containing functional PME genes lack the nonfunctional pseudogenes related to *ZmPme3* but do contain tandem repeats of the *ZmGa1P* gene.

The complexity of the *Ga1* locus together with its role in controlling cross-compatibility makes the evolution of this locus particularly interesting. The objective of this study is to compare the evolutionary history of the *ga1* and *Ga1-S* haplotypes of the *Ga1* locus in the genus *Zea* in order to gain a better understanding of the molecular events that gave rise to this genome region. The results provide insights into key evolutionary events in the development of modern maize.

Received on 17 May 2023; accepted on 16 August 2023
Published by Oxford University Press on behalf of The Genetics Society of America 2023.

Materials and methods

Identification of pseudogenes and gene fragments at the *Ga1* locus in maize genotypes

To identify genomic sequences related to PME genes at the *Ga1* locus, tblastx searches using amino acid sequences of *ZmPme3* and *ZmGa1P* as queries were carried out against Zm-B73 -REFERENCE-NAM-5.0 and Zm-Hp301-REFERENCE-NAM-1.0. Similar tblastx searches were performed in all nested association mapping (NAM) founders and other high-quality maize whole genome assemblies listed in Table 1. All genome assemblies used in the analysis were downloaded from MaizeGDB (Woodhouse *et al.* 2021, https://download.maizegdb.org/).

Self and pairwise alignments and alignment visualization

The genomes included in this study were masked for repeats using RepeatMasker (Tarailo-Graovac and Chen 2009) using the MTEC transposon consensus library (https://github.com/oushujun/MTEC/blob/master/maizeTE02052020). Pangenome single nucleotide polymorphisms flanking the genomic intervals containing the *Ga1* loci were identified using GBrowse from MaizeGDB (Supplementary Tables 1 and 2). Sequences of the genomic regions on chromosome 4 were extracted based on the position information of the markers. Self-alignments of these genomic intervals were constructed using the nucmer alignment script from Mummer version 3.23 (Kurtz *et al.* 2004). The options nucmer --maxmatch and --nosimplify were used to find nonexact alignments to identify repeat sequences within this region of interest. To visualize these alignments, the delta file was used as an input file for the mummerplot script to generate an image (.png) file of the self-alignments (Supplementary Fig. 1a–e). For pairwise dot plots, alignments between repeat masked chromosome 4 of the selected genotypes were made using nucmer --mum option. The alignments were visualized using "mummerplot" with the pangenome marker positions specified for the --xrange and --yrange options (Supplementary Fig. 2a–d).

Examination of grass genomes for *Ga1*-related sequences

A BLAST search using the genomic sequence of *ZmGa1P* from SDG25a (Zhang *et al.* 2018) was conducted against the entire NCBI database using the least stringent parameters and an *e*-value cutoff of 1e − 10. A similar BLAST search was conducted using a transcript sequence of *ZmPme3* and the same parameters as the *ZmGa1P* search. The corresponding predicted protein sequences were also identified. To determine whether the identified significant hits for *ZmGa1P* were more significant to *QRT1* (a PME gene that is not part of the *Ga1* locus but is more closely related to *ZmGa1P* than *ZmPme3*) or *ZmGa1P*, the maize *QRT1* genomic sequence was acquired from MaizeGDB (Zm00001d030643/Zm00001eb028580) and aligned with each respective species' reference genome in which a significant *ZmGa1P* hit was found. A BLAST search was conducted on MaizeGDB using the genomic sequences of *ZmGa1P* from SDG25a and *ZmPme3* from Hp301 as queries against Zx-PI566673 Yan 1.0 assembly (teosinte). All predicted protein sequences are listed in Supplementary Table 6.

Relationship between transposons and pseudogenes and gene fragments

BEDTools option intersect (Quinlan and Hall 2010) was used to identify transposon sequences that are inserted within pseudogenes and gene fragments of interest. Tables 2 and 3 list gene fragments with transposons inserted within or overlapping either 5′ or 3′ terminals of their sequences. Gene fragments with transposons inserted within them were pieced together. Such "joined" sequences were also included in the sequence data set used for phylogenetic tree reconstruction of *ZmPme3*-like sequences in B73 and *ZmGa1P*-like sequences in Hp301.

Maximum likelihood phylogenetic tree reconstruction of duplicated sequences

Maximum likelihood phylogenetic reconstruction was used to create duplication trees for *ZmPme3* sequences in B73 and *ZmGa1P* sequences in Hp301 using RAxML-NG (Kozlov *et al.* 2019). The final data set included 41 *ZmPme3*-like sequences in B73 and 18 *ZmGa1P*-like sequences in Hp301. Multiple sequence alignments were generated using MAFFT. GTR + GAMMA model of rate heterogeneity was selected for the analysis. A default extended majority rule-based bootstrapping test was used to determine a sufficient number of bootstrap replicates (Pattengale *et al.* 2010).

Stop codon analysis

The genomic sequences for each of the *ZmPme3*-like sequences (including the "joined" sequences) were aligned with the coding sequence of *ZmPme3*. The intron was removed during the alignment in MEGA X (Kumar *et al.* 2018). The alignment was translated to the amino acid sequence, and the positions of stop codons resulting from base substitutions were noted.

Determining retrotransposon ages using LTR age of insertion analysis

Retrotransposon annotations for NAM founders were downloaded from https://ftp.maizegdb.org/MaizeGDB/FTP/. Retrotransposons with intact right and left long terminal repeats (LTRs) were selected for this analysis. Sequences of the left and right LTRs of all retrotransposons were extracted using SAMtools. Pairwise alignments between the 2 LTR sequences of each retrotransposon were performed using MUSCLE (Edgar 2004). Pairwise alignments were then used to calculate the divergence distance (d). The substitution rate, $r = 3.3 \times 10^{-8}$ substitutions per site per year, was used for insertion age estimation (Clark *et al.* 2005).

Helitron and TIR age assessment using terminal branch length estimates

The ages of individual helitron and terminal inverted repeat (TIR) transposon insertions were calculated using terminal branch lengths from phylogenetic trees of the corresponding TE families. For each family of helitrons and TIR elements, multiple sequence alignments of all TE sequences in the corresponding genome were made using MAFFT. The directionality of the transposons was maintained using the—adjustdirection option in MAFFT. The alignments were then used for phylogenetic tree reconstruction using RaxML-NG. Terminal branch lengths were used as a measure of divergence distance, and insertion ages were calculated using the same parameters for LTR insertion age estimation.

Results and discussion

Genomic regions encompassing B73 (*ga1*) and Hp301 (*Ga1-S*) loci contain genotype-specific arrays of sequences homologous to PMEs involved in gametophytic cross-incompatibility

It has been reported that inactive alleles (*ga1*) of the *Ga1* locus contain tandem arrays of pseudogenes related to *ZmPme3* and *ZmGa1P*—the 2 PMEs that confer cross-incompatibility in active (*Ga1-S*) alleles of the locus. In this study, we identified several

Table 1. Characteristics of the *Ga1* locus of a diverse set of inbred lines.

Ga1 genotype[a]	Lines	*ZmPme3*-like sequences	*ZmGa1P*-like sequences	Domain length (Mb)	Group designation
Ga1-S/M	Hp301[c], SK[c], **CML333, CML52, NC350, NC358, Tzi8**	2 (1 full length)	22–27 (8 full length)	1.5–1.7	A
ga1	B73,B97,CML69,CML103,CML228, CML247, Il14H[b], Ki3, Ki11, M162W, M37W, Mo18W, Oh7B, Oh43, P39[b],Tx303, Ia453[b], B104, DK105, W22, EP1, F7, Mo17, PE0075, PH207	61–64	25–30	1.1–1.2	B
ga1	MS71	17	8	0.2	C
ga1	Ky21, CML322, A188	126–139	48–59	3.1	D
ga1	CML277	119	35	1.4	E

Bolded *Ga1-M* genotypes can be pollinated by any *Ga1* haplotype (*ga1*, *Ga1-S*, and *Ga1-M*) and can pollinate *Ga1-S* plants (Jones and Goodman 2018).
[a] *ga1* lacks functional copies of *ZmPme3* and *ZmGa1P*; *Ga1-S/M* has intact copies of both.
[b] Sweet corn lines.
[c] Popcorn lines.

ZmPme3-like and *ZmGa1P*-like pseudogenes and gene fragments in the ~1.1-Mb region between 8.56 and 9.6 Mb on chromosome 4 in Zm-B73-REFERENCE-NAM-5.0. In B73, a few of the *ZmPme3*-like sequences are part of a *ZmPme3*-N-*ZmGa1P* repeat ($N = AT_{\sim 250}$) that occurs 16 times in the 1.1-Mb region forming a tandem cassette of pseudogenes and gene fragments. Most of the *ZmGa1P*-like sequences in B73 are truncated to contain 3′ terminal fragments. In contrast, Zm-Hp301-REFERENCE-NAM-1.0 contained several full-length genes as well as partial *ZmGa1P*-like sequences between 8.5 and 9.8 Mb with 1 functional *ZmPme3* sequence and a 350-bp gene fragment. All sequences in these arrays are oriented in the same direction. The distribution of repeat sequences in these 2 genotypes is illustrated in the top 2 sections of Fig. 1. The differences in the genome structure of this region between B73 and Hp301 led us to examine additional lines to gain a better understanding of the variation in genome structure present at this locus.

Variation in genome structure among diverse maize inbred lines

We examined genomic intervals containing the *Ga1* locus in all NAM assemblies (Hufford *et al.* 2021), previously reported high-quality assemblies of European flint lines (Unterseer *et al.* 2017; Haberer *et al.* 2020), and recent assemblies (Yang *et al.* 2019; Lin *et al.* 2021) from MaizeGDB. Supplementary Figure 1a–e shows self-comparisons of the *Ga1* locus of some of the genotypes, selected to illustrate the diversity present among the lines under study. The dot plots reveal distinct genomic patterns of duplications throughout the *Ga1* loci, which appear as signals of the central diagonal. The dot plots illustrate the substantial diversity of size, density, and arrangement of the repeat-containing region.

The NAM founders, European flint lines, and recently added high-quality assemblies together capture a large amount of diversity in maize. This set of inbred lines contained Hp301 and SK, 2 popcorn lines that have an active (*Ga1-S*) genotype, 5 lines with the male function of *Ga1-S*, i.e. *Ga1-M*, and 30 lines with the inactive allele *ga1*. Based on the observed genome structures apparent in the representative dot plots (Supplementary Figs. 1 and 2), the number and type of pseudogenes present, and the length of the repeat region in the genome, these lines were classified into 5 groups designated "A" through "E" as summarized in Table 1. It is interesting to note that Jones and Goodman (2018) classified 2 of the lines we classified in this sequence analysis as *ga1*, P39, and Ki11, as potentially having the *Ga1-M* allele using phenotypic analysis.

Group A contains all the lines with active *Ga1* components, including the alleles *Ga1-S* (found in many popcorn varieties) and *Ga1-M*. In addition to the active genes (*ZmPme3* and *ZmGa1Ps-m*), this group is characterized by the presence of only 1 *ZmPme3* gene fragment and several *ZmGa1P-like* pseudogenes. Group B is the largest and contains *ga1* genotypes, which is the genotype of most cultivated field corn varieties. As described above, this group is characterized by many pseudogenes related to *ZmPme3* and *ZmGa1P*. Three other groups have only 1 or 2 members and contain unusual rearrangements of genome features found in most *ga1* genotypes. Thus, group C has a large deletion and is a truncated version of the group B genotype while group D contains a duplication of the entire *ga1* locus of group B. Group E with only 1 member, i.e. CML277, appears to have an internally expanded version of the group B genome structure with a larger number of *ZmPme3*- and *ZmGa1P*-like sequences. The arrangement of *ZmPme3* and *ZmGa1P* genes and pseudogenes in a representative member of each group is shown in Fig. 1.

Tandem duplications arising from nonhomologous recombination are responsible for the formation of *ga1* and *Ga1-S* sequence clusters

Several types of molecular events can give rise to gene duplications. These include whole genome duplications, transposition mediated by transposons of several types, and tandem duplications arising from nonhomologous recombination events (Panchy *et al.* 2016). Transposition via an RNA intermediate is not likely to be responsible for duplication of *Ga1*-associated sequences because introns are found in all complete and partial-length pseudogene sequences. Regions of microhomology in genomes can be attributed to the presence of transposons and low-complexity repeated sequences. Nonhomologous recombination creates proximal repeats that can be targets for subsequent nonhomologous recombination events, creating several more copies of the sequences arranged in a tandem array. The tandem arrangement of the *Ga1* sequence arrays suggests nonhomologous recombination to be the mechanism for their origin.

To determine the time of these duplication events, we reconstructed a phylogeny of *ZmPme3*-like sequences in B73 using the maximum likelihood phylogeny reconstruction method. A phylogenetic tree for the B73 *ZmGa1P*-like sequences is shown in Supplementary Fig. 3.

Figure 2a shows the topology of the tree for all *ZmPme3*-like sequences from B73. The branch lengths indicate that *ZmPme3*-like sequences are highly diverged relative to each other and are therefore likely to be a result of ancient duplication events. Although the topology of this phylogeny tells us only about the relatedness of the sequences and not the precise order of the duplication events, the tree offers some clues about the events that led to the repeat array. The tree topology and the stop codon

Table 2. Age of transposon insertions within *ga1* pseudogene sequences in B73.

Group	Fragment name	Sequence start	Sequence stop	Transposon insertion(s)	Transposon family	Strand	Start	Stop	Age (MYA)	Method
I	PME3-S2	8772140	8772512	uwum_AC190887_2701	Unknown LTR	–	8772512	8785694	0.611	Terminal branch length
	PME3-S3	8794858	8795785	uwum_AC190887_2701	Unknown LTR	–	8772911	8785317	0.073	Terminal branch length
				uwum_AC213069_12092	Gypsy	+	8774798	8784721	0.661	Terminal branch length
				DTH_ZM00280_consensus	PIF/Harbinger	–	8779539	8779740	0.000	Terminal branch length
				ji_AC213834_12382	Copia	–	8785689	8794798	0.839	LTR-LTR divergence
I	PME3-S11	8924498	8924767	uwum_AC213069_12092	Gypsy	+	8924765	8926893	0.552	Terminal branch length
	PME3-S12	8948456	8949338	huck_AC193313_3542	Gypsy	–	8927033	8935832	2.445	Terminal branch length
				chr4:8935865..8942218	Unknown LTR	+	8935859	8942223	0.041	LTR-LTR divergence
				huck_AC216048_13250	Gypsy	–	8942218	8948455	0.499	Terminal branch length
II	PME3-S15	9028581	9029171	chr4:9029172..9038718	Gypsy	+	9029166	9038723	0.380	LTR-LTR divergence
	PME3-S16	9038719	9039236							
I	PME3-S17	9039599	9040798	gyma_AC189750_2238	Gypsy	–	9040798	9057063	0.739	Terminal branch length
	PME3-S18	9057059	9057151	gyma_AC189750_2238	Gypsy	–	9041975	9047481	0.531	Terminal branch length
				TE_00017050_LTR	Copia	–	9042832	9043093	0.774	Terminal branch length
				TE_00018639_LTR	Unknown LTR	+	9044768	9056211	0.639	Terminal branch length
				TE_00010043_LTR	Unknown LTR	+	9044858	9056766	0.378	Terminal branch length
				TE_00016495_LTR	Unknown LTR	+	9045016	9056520	0.266	Terminal branch length
				TE_00024226_LTR	Unknown LTR	+	9045261	9045407	0.847	Terminal branch length
				naiba_AC195481_139	Unknown LTR	–	9045407	9057060	0.913	Terminal branch length
				gyma_AC198939_6228	Gypsy	+	9047482	9051559	0.894	Terminal branch length
I	PME3-S22	9135959	9136380	uwum_AC213069_12092	Gypsy	+	9136380	9161591	0.181	Terminal branch length
	PME3-S23	9161583	9162469	doke_AC197224_5479	Gypsy	–	9142979	9154412	0.291	LTR-LTR divergence
				DTC_ZM00113_consensus	CACTA	+	9154414	9158206	0.155	Terminal branch length
				DTC_ZM00106_consensus	CACTA	–	9154806	9157882	0.132	Terminal branch length
				TE_00003586_INT	CACTA	+	9155059	9156273	0.016	Terminal branch length
				DTC_ZM00106_consensus	CACTA	+	9157882	9157996	1.239	Terminal branch length
I	PME3-S38	9540015	9540590	prem1_AC196065_4927	Gypsy	–	9540590	9542489	2.765	Terminal branch length
	PME3-S39	9543881	9544091	TE_00008429_LTR	Gypsy	–	9542489	9543344	0.709	Terminal branch length
	PME3-S40	9550960	9551465	TE_00019198_LTR	Unknown LTR	+	9543344	9543812	0.290	Terminal branch length
				uwum_AC190887_2701	Unknown LTR	–	9544091	9550959	0.718	LTR-LTR divergence
I	PME3-S28	9240407	9241363	uwum_AC190887_2701	Unknown LTR	–	9241420	9243879	0.624	Terminal branch length
	PME3-S29	9250047	9250227	uwum_AC177933_415	Gypsy	–	9243760	9244114	0.369	Terminal branch length
				uwum_AC190887_2701	Unknown LTR	–	9244069	9247126	0.350	Terminal branch length
				uwum_AC190887_2701	Unknown LTR	–	9247121	9247523	0.257	Terminal branch length
				uwum_AC177933_415	Gypsy	–	9247406	9247653	0.233	Terminal branch length
				uwum_AC213069_12092	Gypsy	–	9247656	9248846	0.558	Terminal branch length
				uwum_AC177933_415	Gypsy	+	9248846	9250032	1.771	Terminal branch length
I	PME3-S31	9260842	9261257	DTC_ZM00108_consensus	CACTA	–	9261259	9262690	1.175	Terminal branch length
	PME3-S32	9305963	9306800	ZM_Tourist_7	PIF/Harbinger	+	9261556	9261803	1.253	Terminal branch length
				DTC_ZM00063_consensus	CACTA	–	9261803	9262311	0.465	Terminal branch length
				TE_00012109_INT	Gypsy	+	9262694	9262854	0.229	Terminal branch length
				flip_AC208040_9765	Gypsy	+	9266382	9267299	0.538	Terminal branch length
				DTC_ZM00101_consensus	CACTA	+	9267306	9275768	0.023	Terminal branch length
				cinful_zeon_AC207755_9705	Gypsy	+	9268118	9268775	0.124	Terminal branch length
				flip_AC193970_3791	Gypsy	+	9275771	9279524	0.345	Terminal branch length
				DTC_ZM00107_consensus	CACTA	–	9280035	9280628	0.703	Terminal branch length
				cinful_zeon_AC211573_11290	Gypsy	+	9280624	9285480	0.416	Terminal branch length
				prem1_AC196065_4927	Gypsy	–	9285480	9287637	0.936	Terminal branch length
				chr9_P_79032327	Copia	–	9286199	9292005	0.106	Terminal branch length
				TE_00009958_LTR	Unknown LTR	+	9287420	9287551	0.117	Terminal branch length

(continued)

Table 2. (continued)

Group	Fragment name	Sequence start	Sequence stop	Transposon insertion(s)	Transposon family	Strand	Start	Stop	Age (MYA)	Method
				TE_00007970_LTR	Unknown LTR	–	9287551	9288483	0.031	Terminal branch length
				prem1_AC212325_11702	Gypsy	–	9288483	9288902	0.627	Terminal branch length
				TE_00002690_INT	Unknown LTR	+	9288902	9289134	0.644	Terminal branch length
				TE_00016332_INT	Unknown LTR	–	9288927	9289434	0.978	Terminal branch length
				odoj_AC194387_4072	Unknown LTR	+	9289120	9290234	0.436	Terminal branch length
				prem1_AC206253_9147	Gypsy	–	9290234	9291694	0.512	Terminal branch length
				prem1_AC200105_6751	Gypsy	–	9291950	9292172	2.244	Terminal branch length
				TE_00007970_LTR	Unknown LTR	–	9292303	9293235	0.117	Terminal branch length
				TE_00017576_LTR	Unknown LTR	–	9293091	9293683	2.300	Terminal branch length
				cinful_zeon_AC211573_11290	Gypsy	+	9293683	9305730	0.371	Terminal branch length
				chr6_P_87978263	Copia	–	9294693	9304136	0.168	LTR-LTR divergence
				DTC_ZM00108_consensus	CACTA	–	9305729	9305960	1.204	Terminal branch length
I	PME3-S20	9092845	9093811	uwum_AC177933_415	Gypsy	+	9093811	9113071	0.470	Terminal branch length
	PME3-S21	9113072	9113390	uwum_AC177933_415	Gypsy	+	9094300	9104851	0.439	Terminal branch length
				DTC_ZM00107_consensus	CACTA	+	9094784	9095594	0.548	Terminal branch length
				uwum_AC213069_12092	Gypsy	+	9095919	9098282	0.335	Terminal branch length
				TE_00005725_INT	Gypsy	+	9101896	9102033	0.585	Terminal branch length
				DTC_ZM00108_consensus	CACTA	+	9102034	9102624	0.316	Terminal branch length
				DTC_ZM00107_consensus	CACTA	+	9104851	9105505	0.192	Terminal branch length
				TE_00012109_INT	Gypsy	–	9106290	9107919	0.190	Terminal branch length
				DTC_ZM00108_consensus	CACTA	+	9107920	9108514	0.146	Terminal branch length
				uwum_AC177933_415	Gypsy	+	9108514	9111984	0.599	Terminal branch length
I	Ga1P-S4	8864391	8864589	TE_00019959_LTR	Unknown LTR	–	8864588	8864758	0.329	Terminal branch length
	Ga1P-S5	8867265	8867757	uwum_AC177933_415	Gypsy	+	8864758	8864903	0.245	Terminal branch length
	Ga1P-S6	8911569	8911644	uwum_AC190887_2701	Unknown LTR	+	8864810	8867265	0.497	Terminal branch length
				uwum_AC190887_2701	Unknown LTR	–	8867757	8911567	0.647	Terminal branch length
				chr4:8868541..8881257	Gypsy	+	8868535	8881262	0.629	LTR-LTR divergence
				DTC_ZM00061_consensus	CACTA	–	8872192	8875352	0.005	Terminal branch length
				uwum_AC190887_2701	Unknown LTR	–	8881257	8882720	0.799	LTR-LTR divergence
				DTC_ZM00053_consensus	CACTA	+	8882720	8887045	0.358	Terminal branch length
				DTH_ZM00326_consensus	PIF/Harbinger	–	8890240	8890332	0.454	Terminal branch length
				DTC_ZM00091_consensus	CACTA	+	8890690	8894028	0.239	Terminal branch length
				flip_AC208040_9765	Gypsy	–	8894024	8904986	0.388	Terminal branch length
				DTA_ZM00186_consensus	hAT	+	8904985	8905382	0.382	Terminal branch length
				flip_AC193970_3791	Gypsy	–	8905373	8909909	0.358	Terminal branch length
				DTC_ZM00091_consensus	CACTA	+	8909909	8911232	0.314	Terminal branch length
I	Ga1P-S11	9113950	9114288	uwum_AC190887_2701	Unknown LTR	+	9114283	9121174	0.138	LTR-LTR divergence
	Ga1P-S12	9121170	9121592							
I	Ga1P-S13	9175540	9176112	leviathan_AC208826_10024	Gypsy	–	9176117	9179686	0.575	LTR-LTR divergence
	Ga1P-S14	9195230	9195404	TE_00000918_INT	Gypsy	+	9179423	9179718	1.731	Terminal branch length
				uwum_AC177933_415	Gypsy	–	9179718	9183866	0.266	Terminal branch length
				uwum_AC177933_415	Gypsy	–	9183939	9189286	0.116	Terminal branch length
				cinful_zeon_AC203004_7602	Gypsy	–	9189280	9195228	0.418	Terminal branch length
I	Ga1P-S18	9463252	9463528	uwum_AC190887_2701	Unknown LTR	+	9463532	9466113	0.604	Terminal branch length
	Ga1P-S19	9466114	9466578							
I	Ga1P-S20	9553652	9554150	opie_AC217577_13524	Copia	–	9554144	9563239	0.453	LTR-LTR divergence
	Ga1P-S21	9563235	9563498							
I	Ga1P-S26	9801433	9802398	cinful_zeon_AC215255_13029	Gypsy	+	9797524	9801238	0.377	LTR-LTR divergence
	Ga1P-S27	9802409	9802606							

Table 3. Age of transposon insertions within *Ga1-S* pseudogene sequences in Hp301.

Group	Fragment name	Fragment start	Fragment stop	Transposon insertion(s)	Transposon family	Strand	Transposon start	Transposon stop	Age (MYA)	Method
I	Ga1P-S5	9048928	9049189	chr5_P_170822725	Copia	−	9049226	9049676	0.535	Terminal branch length
	Ga1P-SS6	9071697	9072355	grande_AC200214_6803	Gypsy	+	9049676	9057763	0.216	LTR-LTR divergence
	Ga1P-SS7	9092635	9092696	TE_00027235_INT	Unknow LTR	−	9057763	9057912	0.104	Terminal branch length
	Ga1P-S8	9092699	9092896	grande_AC200214_6803	Gypsy	+	9057912	9058482	0.215	Terminal branch length
				grande_AC200214_6803	Gypsy	+	9058464	9063910	0.372	LTR-LTR divergence
				opie_AC217577_13524	Copia	−	9063910	9071697	0.790	Terminal branch length
				ji_AC211489_11215	Gypsy	+	9072567	9081674	0.266	Terminal branch length
				flip_AC203163_7675	Gypsy	+	9081705	9085106	0.778	Terminal branch length
				uwum_AC190887_2701	Gypsy	+	9085101	9091981	0.041	Terminal branch length
				nihep_AC194441_4115	Gypsy	−	9091976	9092632	0.086	Terminal branch length
II	Ga1P-S12	9411371	9411786	uwum_AC177933_415	Gypsy	+	9411781	9420942	0.018	Terminal branch length
	Ga1P-S13	9420938	9421789							
II	Ga1P-S19	9894733	9895166	uwum_AC190887_2701	Gypsy	+	9895161	9902043	0.077	Terminal branch length
	Ga1P-S20	9902036	9902869							

information for all sequences (see Fig. 2a and b) indicate that *ZmPme3*-like sequences can be broadly divided into 2 groups. Sequences in group I on average are farther from the root (i.e. the extent of divergence is greater) than those belonging to group II. Also, group I sequences have a higher number of stop codons (Fig. 2b), some of which are shared by all its members. Group II sequences on the other hand have fewer stop codons as compared to group I, some of which are unique. For example, sequences B73-Pme3-S7 and B73-Pme3-S9 from group II have just 1 unique stop codon each and no other disablements. In addition, all *ZmPme3*-like sequences that are part of the larger repeating motif, *ZmPme3*-N-*ZmGa1P* described above, belong to group I. The topology suggests that group I sequences were generated by proximal duplications first, followed by additional duplications leading to the group II sequences in multiple distinct nonhomologous recombination events.

The relative positions of the sequences in the genome provide some clues about the nature and order of duplication events that gave rise to the repeat sequences. First, sequences that are closely related to each other do not tend to be adjacent to each other in the genome (Fig. 2a and c). This suggests that the duplication events involved duplication of multiple repeats per event. Second, group I and group II sequences are imperfectly interspersed throughout the repeat region (Fig. 2c). This suggests that some duplication events involving members of both groups occurred after the 2 groups were established.

An important question in understanding the duplication history of the array is whether the duplications occurred while the genes were active or after they had been inactivated by mutations. Duplication of active genes may have disrupted reproduction and resulted in strong selection against the duplicated locus, while duplication of inactive genes would be reproductively neutral. The B73 *ZmPme3* phylogeny enriched with stop codon information (Fig. 2b) addresses this question. Sequences in group I have 2 stop codons 774 and 549 that are shared by all except 1 of its members, indicating that duplication events in this group occurred after inactivation of the functional sequences by either or both stop codons. On the other hand, several group II members have unique stop codons suggesting that a second series of multiple nonhomologous recombination events occurred. The presence of stop codons shared by all sequences in group I signifies that nonfunctional sequences were amplified during the nonhomologous recombination events that led to the tandem arrays. This suggests that the role of *Ga1* in reproduction had little impact on the structure of the pseudogene arrays in group I, whereas the reason behind the inactivation of sequences with unique stop codons in group II is unclear. Branch lengths and the nucleotide divergence estimates indicate that both group I and group II duplications occurred >600 KYA. This estimate for the *ga1* pseudogene cluster coincides with the *Tripsacum-Zea* split, which was recently demonstrated to have occurred ~650,000 years ago (Chen, Zhang, *et al.* 2022).

In contrast to the B73 tandem pseudogene array that is dominated by *ZmPme3* pseudogenes, the array in the *Ga1-S* line Hp301 has only 1 full-length *ZmPme3* sequence, only 1 *ZmPme3* fragment, 8 full-length *ZmGa1P* sequences, and 10 *ZmGa1P* pseudogenes. Like the tandem duplications in B73, the *ZmGa1P* sequences present in Hp301 may have also arisen due to proximal duplications from unequal crossover events. The tree for *ZmGa1P* sequences is shown in Fig. 3a. Figure 3b shows stop codon information for the pseudogenes in the Hp301 *ZmGa1P* tree. Unlike the B73 pseudogene array, sequences in the Hp301 array that are most similar to each other tend to be adjacent in the

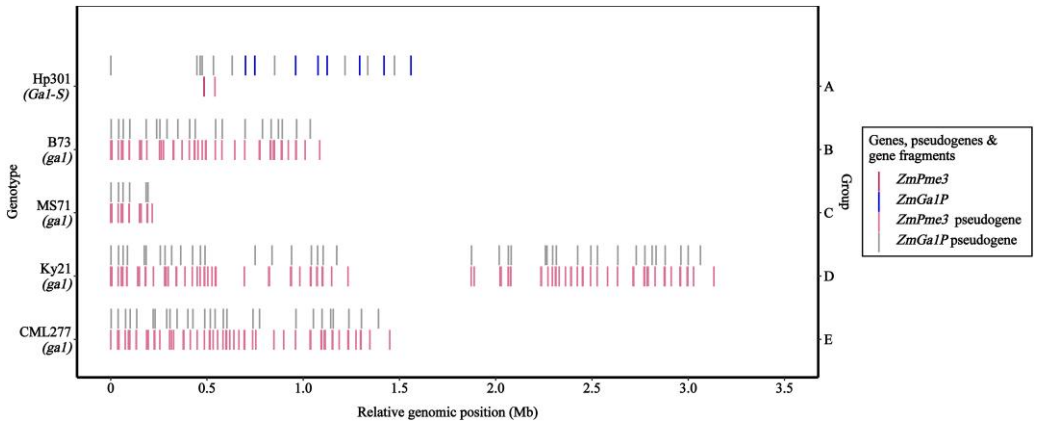

Fig. 1. Position of *Ga1* PME genes, and PME pseudogene and gene fragments in 5 representative inbred lines. The haplotype of the *Ga1* locus is shown in parentheses below the name of each line. Genome positions are adjusted to align with the first base of the cluster in each genotype for ease of comparison. Group letters on the secondary axis are from Table 1.

genome (Fig. 3c). This suggests that duplication events involved 1 gene/pseudogene sequence at a time. The *ZmGa1P* tree topology also shows 2 groups of sequences—sequences in group I are older and duplicated approximately >600 KYA whereas those in group II duplicated at several different time periods during the locus history, i.e. at 260 KYA and between 60 and 100 KYA. The differences in the time and mode of duplications at the *Ga1* region indicate that the B73 (*ga1*) and Hp301 (*Ga1-S*) tandem arrays arose independently.

Transposon insertions leading to splitting of full-length sequences into gene fragments date to different time periods in B73 and Hp301

Transposons have major impacts on genome structure and evolution (Wicker *et al.* 2018). Transposon insertions within full-length repeats of *ga1* and *Ga1-S* regions provide an indirect measure of the age of the sequences they insert into. A duplication event giving rise to a repeat sequence precedes a unique transposon insertion event in the sequence and thereby is older than the insertion event. In the case of sequences with newer and nested insertions, we examined the age of the oldest transposons. We compare the age of transposon insertions between the sequence groupings in both *ga1* (B73) and *Ga1-S* (Hp301) arrays.

Several of the *ZmPme3*-like and *ZmGa1P*-like gene fragments in B73 and *ZmGa1P*-like gene fragments in Hp301 are a result of 1 or more transposon insertions, causing the originally intact sequences to split into 5′ and 3′ terminal gene fragments. When examined further, the 3′ and the 5′ ends of the 5′ and 3′ fragments have direct repeats of 5–7 bp, known as target site duplications, a characteristic feature of LTR retrotransposons and TIR transposons. In case of LTR retrotransposons, LTRs are identical at the time of insertion and diverge with time. The sequence divergence between LTR sequences allows estimation of the age of an insertion event (SanMiguel *et al.* 1998). For an individual TIR or helitron, age of insertion can be estimated using its terminal branch length in the phylogenetic tree of the corresponding transposon family members in the genome. Tables 2 and 3 list TE insertions in *ga1* and *Ga1-S* sequences and their corresponding insertion ages.

In Hp301 group II sequences, the insertion of Gypsy retrotransposon *uwum_AC177933_415* within Ga1P-S12//13 occurred 18,333 years ago. Similarly, the duplication that gave rise to Ga1P-S18//19 was followed by an insertion of another Gypsy retrotransposons *uwum_AC190887_2701*, about 77,121 years ago. Most of the retrotransposon insertions in group I on the other hand are older. The median insertion age of transposons within group I sequence Ga1P-S5//6//7//8 was found to be 454 KYA whereas the 2 insertions in group II sequences occurred in the last 80,000 years. This is expected as group I sequences in Hp301 are older and duplicated ~600 KYA as compared to group II sequences, which originated ~80–100 KYA.

Ages of transposon insertions in sequences belonging to the 2 groups in the B73 *ZmPme3* phylogeny were also examined. The only insertion in the group II sequence has an age of 380 KYA whereas the median age of insertions in group I sequences was found to be 420 KYA. Insertions within group I sequences are older and are more numerous as compared to group II, also indicating that group I sequences originated before group II.

BLAST analysis of the male and female function genes of the *Ga1* locus shows *ga1* and *Ga1-S*-like sequence arrays across all *Zea* genomes

BLAST analysis of *ZmPme3* and *ZmGa1P* gene sequences queried against teosinte genomes from the Pan-And project (https://panandropogoneae.com/) show the presence of sequence arrays like those of *ga1* and *Ga1-S* in various species of the *Zea* genus. Figure 4 depicts the position of these arrays on chromosome 4 of teosinte genomes released in phase I of the Pan-And project. Supplementary Tables 3–5 are a list of *ZmPme3* and *ZmGa1P* BLAST hits in all teosinte genomes.

BLAST results of *ZmPme3* gene sequence queried against teosinte genomes show the presence of 3 *ZmPme3* copies in *Zea mays mexicana* accession TIL18 and a sequence that has 99.92% identity to the *ZmPme3* sequence in *Z. mays parviglumis* accession TIL01. The next closest BLAST hits for *ZmPme3* (99.38–99.61% identities) are present across all other *Zea* genomes except *mexicana* accession TIL25. *ZmGa1P* BLAST hits with sequence identities between 98.9% and 99.8% occur in the same genomic region as *ZmPme3* loci. Together, they form the *Ga1-S*-like haplotype structure in many of the *Zea* genomes studied. Supplementary Figure 4 shows a phylogenetic tree of all *ZmGa1P* BLAST hits in *Zea* and *Tripsacum* genomes. Sequences with ~98% identities to the *ZmPme3* sequence along with *ZmGa1P* BLAST hits with ~96% identities represent the female (*Tcb1-f*) and the male function genes respectively, and together they constitute the *Tcb1* loci in the *Zea* genomes. Figure 4 also depicts the location of the *Tcb1* loci in addition to the *ga1* and *Ga1-S* arrays mentioned above.

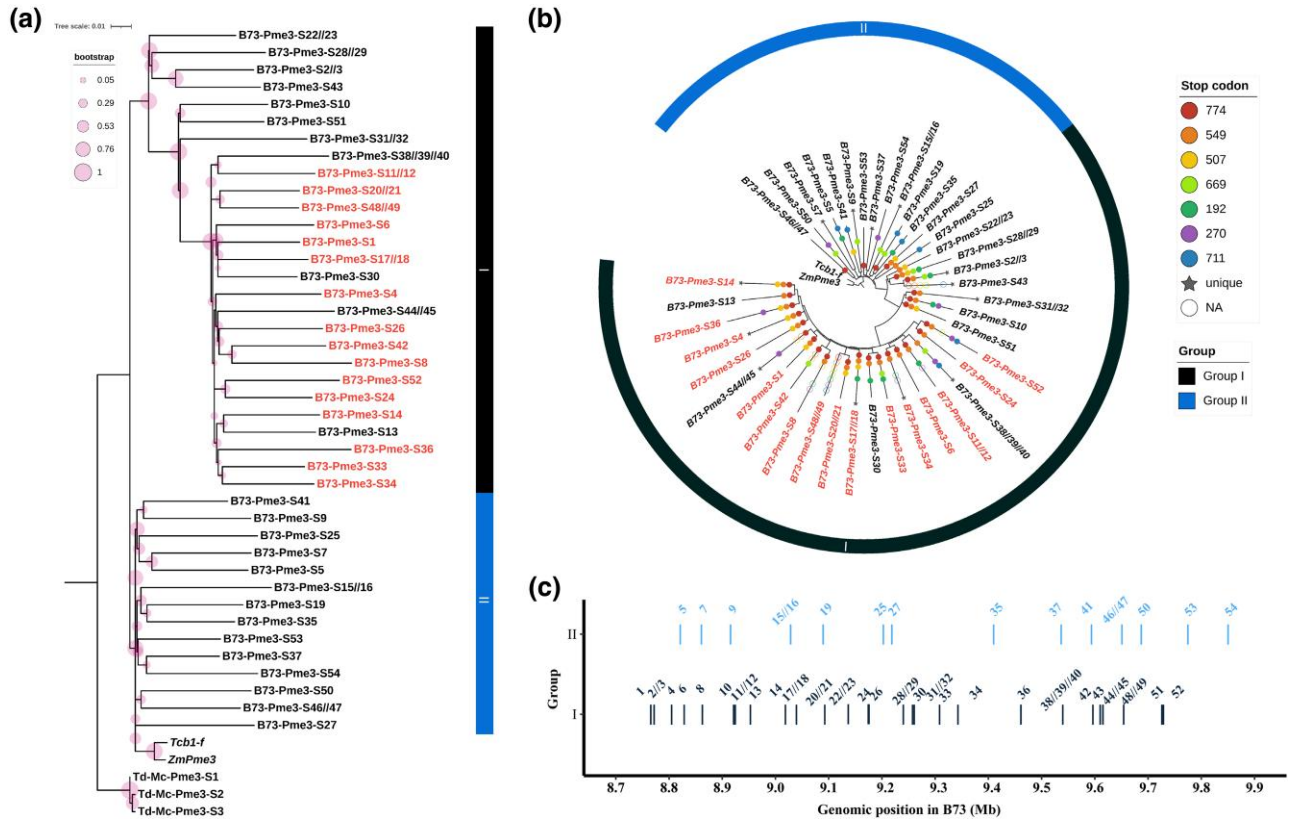

Fig. 2. Analysis of *ZmPme3*-like sequences of the *ga1* haplotype. a) Phylogenetic analysis of *ZmPme3*-like sequences in B73. The tree has been rooted using *Tripsacum ZmPme3* homologs. Sequences are numbered according to their relative positions in the genome. Sequences B73-Pme3-S(1,4,6,8,11//12,14,17//18,20//21,24,26,33,34,36,42,48//49,52) are part of the repeat motif *ZmPme3*-N-*ZmGa1P* and are shown in colored text. The symbol (//) indicates sequences with transposon insertions. *Tcb1-f* is a PME gene similar in sequence to *ZmPme3* but at the *Tcb1* locus. b) Stop codons in sequences of the *ZmPme3* phylogeny. The tree topology shows 2 groups of sequences. In group I, stop codons 774 and 549 are shared by the majority of its members. Group II sequences have more unshared stop codons. c) Genome positions of *ZmPme3* pseudogene sequences in B73 (colored by group).

Figure 5 is a phylogenetic tree of *ZmPme3* BLAST hits. *Tripsacum ZmPme3* BLAST hits form the outgroup of this tree. The pseudogene arrays in *Z. mays parviglumis* accession TIL11, *Zea diploperennis* accession Momo, and *Z. mays* accession B73 form distinct clades in the tree. Full-length sequences on the other hand form 2 other clades—1 with *ZmPme3* and the other with *Tcb1-f* as a member.

TIL11 has a Group B-type (see Table 1) pseudogene array between positions 9.06 and 9.91 Mb in its genome. This array is syntenic to the *ga1* haplotype in B73. Interestingly, among non-*Z. mays* members, only *Z. diploperennis* (Momo) has a pseudogene array like TIL11 and B73. This array is present between 41.71 and 41.86 Mb on chromosome 4 in its genome and has a fewer number of repeats as compared to TIL11 or B73. Figure 6 depicts the relative genomic position of these pseudogene arrays in the genotypes Momo, TIL11 and B73.

Z. mays parviglumis is considered the closest relative of *Z. mays* L., and 2 accessions were sequenced in the Pan-And project. Accession TIL11 contains an arrangement of PME pseudogenes that is syntenic to the *ga1* haplotype of modern field corn varieties. In contrast, accession TIL01 contains a haplotype that is similar to the *Ga1-S* haplotype found in many *Z. mays* L. popcorn varieties (Fig. 4). Thus, *Z. mays parviglumis* contains a locus equivalent to the *Ga1* locus of *Z. mays* L. with haplotypes equivalent to *ga1* and *Ga1-S*.

The presence of 3 tandem *Ga1-S* arrays in the *Z. mays mexicana* accession TIL18 is noteworthy. All 3 copies of the female function gene in these arrays have 100% identities to the *ZmPme3* sequence. The presence of a *ga1*-like array in *Z. diploperennis* accession Momo is also interesting. The widespread occurrence of both alleles in modern maize may be attributed to a weak genome-wide bottleneck during improvement (Hufford *et al.* 2012).

PME genes homologous to those encoded by the *Ga1* locus of maize are widespread and often occur in proximity to each other in several cereal genomes

Predicted PME proteins that share high sequence similarities (>45%) to the male and female determinants of the *Ga1* locus have been reported in several cereal genomes. We used *ZmPme3* and *ZmGa1P* mRNA transcript sequences as queries to conduct BLAST searches in cereal genomes to identify predicted gene and protein sequences that share high sequence similarities with the *Ga1* locus genes. Supplementary Table 6 lists predicted protein homologs identified in cereal genomes. Figure 7 depicts the positions of the *ZmPme3*-like and *ZmGa1P*-like genes in some cereal genomes. Rice (*Oryza sativa japonica*), wild rice (*Oryza brachyantha*), *Sorghum* (*Sorghum bicolor*), and foxtail millet (*Setaria italica*) contain a PME with sequence similarity to the maize *ZmPme3*. Additionally, these species have a *Pme63*-type and occasionally a *QRT1*-type PME with sequence similarity to maize's *ZmGa1P* genes. *QRT1* in *Arabidopsis* is involved in the separation of pollen tetrads after meiosis in the pollen mother cell (Francis *et al.* 2006) and is

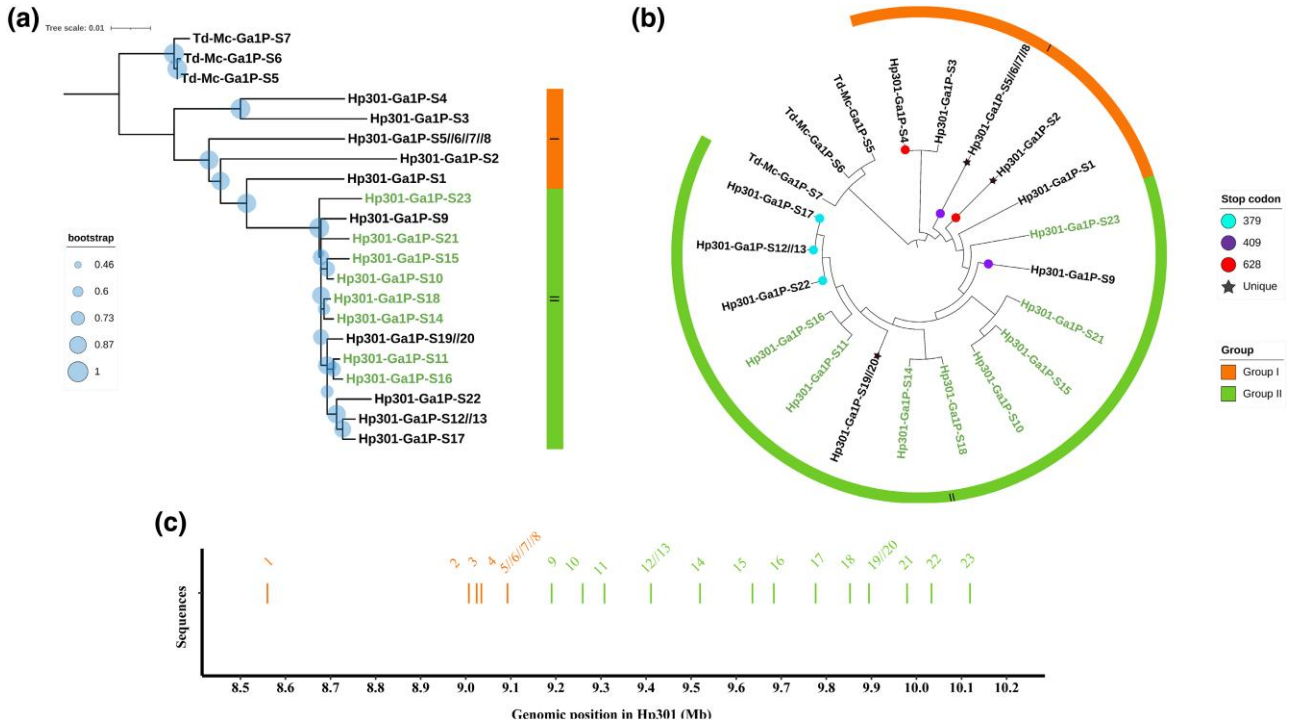

Fig. 3. Analysis of *ZmGa1P*-gene and pseudogene sequences in Hp301. a) Phylogenetic tree for *ZmGa1P* homologs in Hp301. Sequences Hp301-Ga1P-S(10,11,14,15,16,18,21,23) do not contain any disablements and are shown in colored text. The tree has been rooted using *Tripsacum* homologs of *ZmGa1P*. The tree shows 2 groups of sequences that duplicated at 2 different time points. b) Stop codons in the duplicated sequences. c) Genomic positions of *ZmGa1P* repeat sequence array in Hp301 (colored by group).

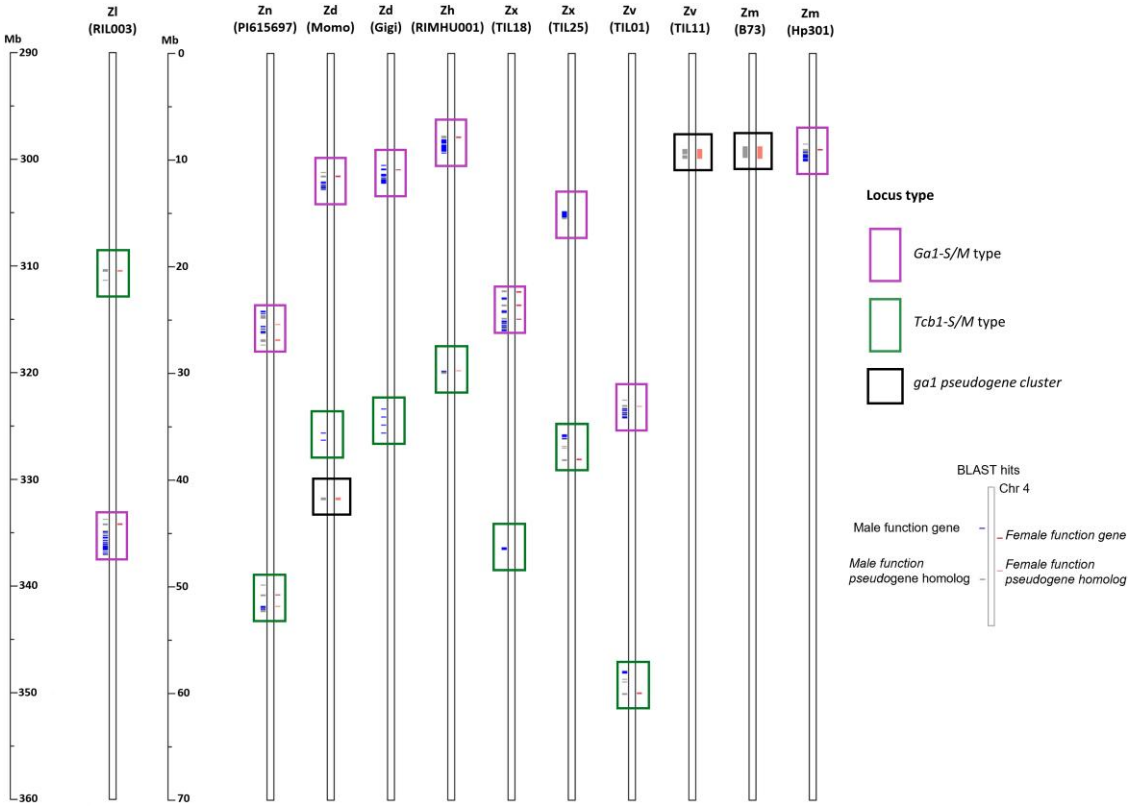

Fig. 4. *ZmPme3* and *ZmGa1P* genes and pseudogene arrays on chromosome 4 in the *Zea* genus. Partial chromosomes (0–70 Mb) are shown for all *Zea* genomes except *Zea luxurians*. For *Zea luxurians*, the genomic region between 290 and 360 Mb has been shown. The figure has been generated using CVit (Cannon and Cannon 2011). Zl, *Z. luxurians*; Zn, *Zea nicaraguensis*; Zd, *Z. diploperennis*; Zh, *Z. mays huehuetenangensis*; Zx, *Z. mays mexicana*; Zv, *Z. mays parviglumis*; Zm, *Z. mays mays*.

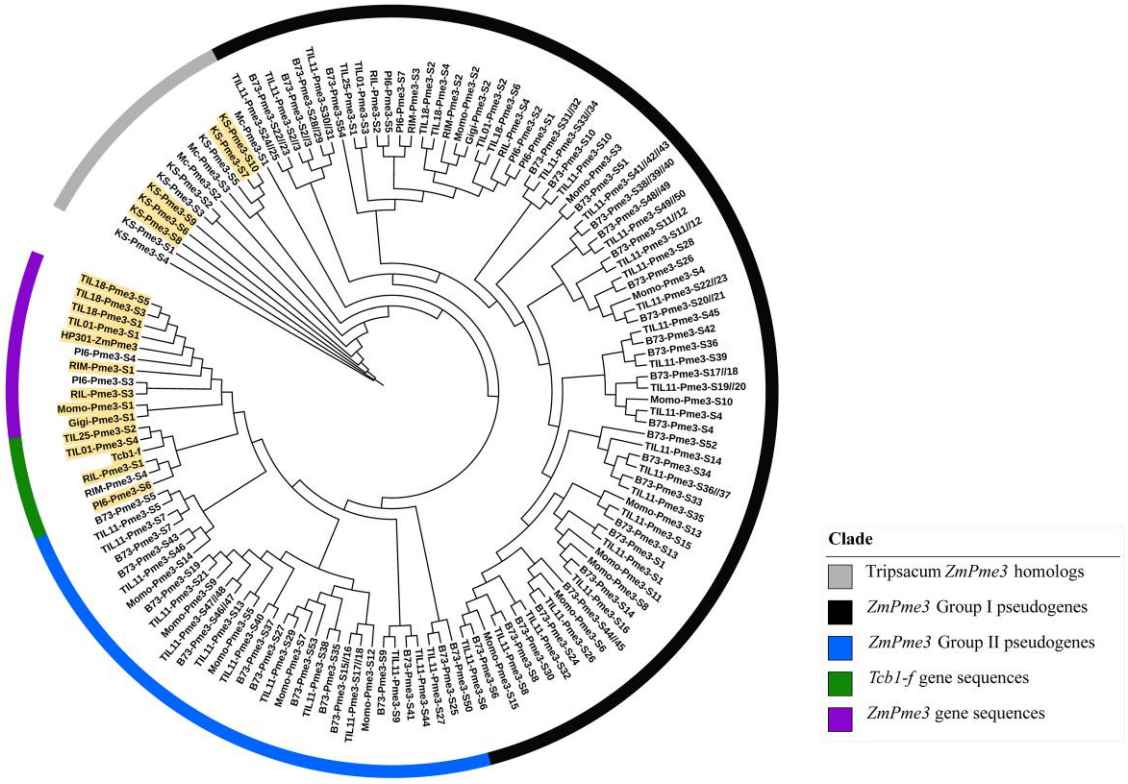

Fig. 5. Phylogenetic tree of *ZmPme3* BLAST hits in all *Zea* genomes with *Tripsacum* homologs as outgroup. Highlighted sequences are full length.

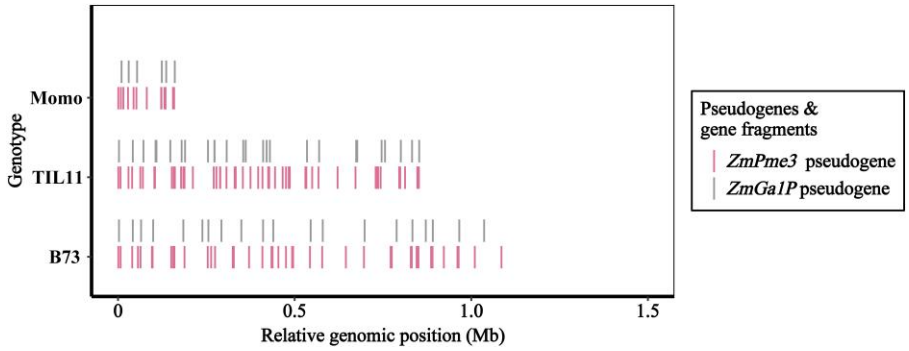

Fig. 6. Relative genomic positions of pseudogenes in B73 (field corn) and TIL11 and Momo accessions (teosintes).

therefore not functionally orthologous to the maize PMEs that control cross-incompatibility. In the case of *Setaria*, the *Pme63*-type *ZmGa1P* homolog is within 0.15 Mb of the *ZmPme3*-like gene (Fig. 7). In *Sorghum* chromosome 4 the distance between the *ZmPme3* homolog and the *QRT1*-type gene is ~11 kb whereas the distance between *ZmPme3* homolog and the *Pme63*-type gene on chromosome 5 is more than 60 Mb. *O. brachyantha* has 2 genes that are like *ZmGa1P* and are situated ~1.8 and 2.3 Mb away, respectively, from the *ZmPme3*-like gene; however, *O. sativa japonica* has 3 genes like *ZmPme3* and 3 genes similar to *ZmGa1P* on chromosome 11. One of the 3 *ZmPme3*-like genes is within ~5.0–6.0 Mb from the genes like *ZmGa1P*. The genes with similarity to the *Ga1* locus in both rice species are located on a portion of chromosome 11 that is syntenic with each other and the short arm of maize chromosome 4 where the *Z. mays Ga1* locus is located (Moore *et al.* 1995; Chen *et al.* 2013; Sun *et al.* 2017). The genes identified in foxtail millet and *Sorghum*, however, are located on chromosomes that are nonsyntenic with maize chromosome 4 (Moore *et al.* 1995; Paterson *et al.* 2009; Zhang *et al.* 2012; Sun *et al.* 2017).

These results indicate that a locus similar to *Ga1* is relatively widespread among grasses, in the sense that homologs of the *Ga1* male and female function genes appear to be present in proximity in numerous cereal genomes examined. Neither *Sorghum* nor *Setaria* has been reported to have unilateral cross-incompatibility. While some *japonica* by *indica* hybrids in rice have been reported to experience UCI, the mechanism is different from that of maize, resulting in aborted embryo and endosperm development and functional pollen tube growth (Matsubara *et al.*, 2003).

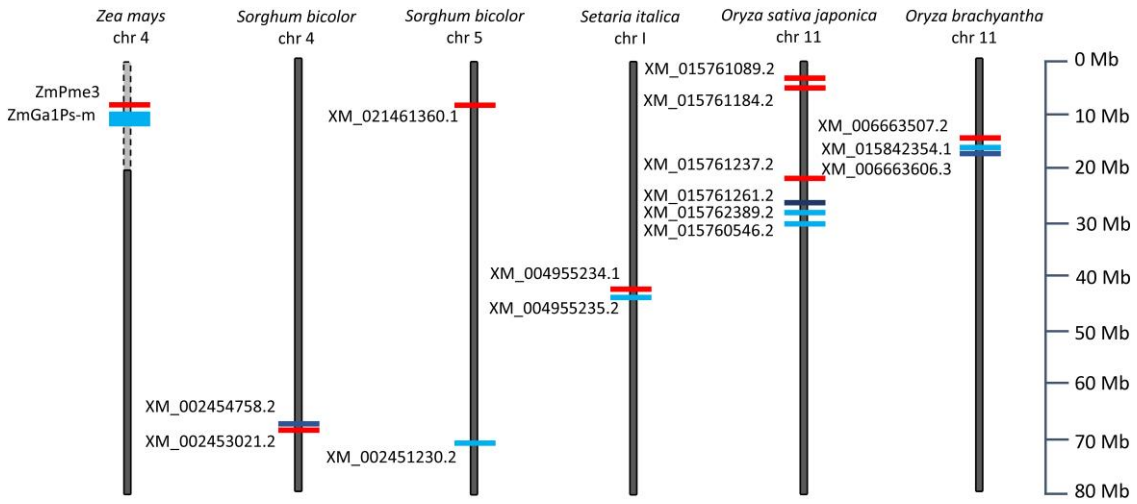

Fig. 7. *ZmPme3* and *ZmGa1P* from BLAST search results in 4 grass species and maize (Supplementary Table 6). Accessions XM_002454758.2, XM_015761261.2 and XM_006663606.3 in sorghum and rice chromosomes have sequence similarity to both *ZmGa1P* and *QRT1* and were more like *QRT1*. The dashed segment of maize chromosome 4 is syntenic with rice chromosome 11. Overall length of chromosomes is not represented; instead, the regions between 0 and 80 Mb are depicted.

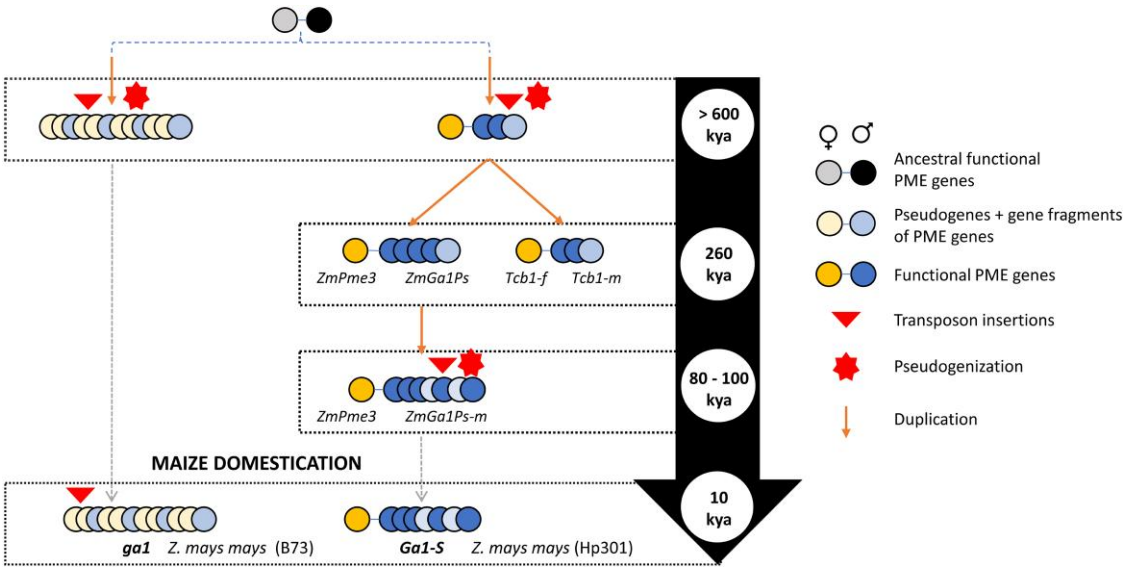

Fig. 8. Model of *Ga1* evolution leading to *ga1* and *Ga1-S* haplotypes.

Conclusion

After examining the duplication histories of the gene and pseudogene sequences, estimating their divergence dates, calculating transposon insertion ages, and analyzing sequences from *Zea* genomes using BLAST, we arrive at a model for the evolution of the *Ga1* locus (see Fig. 8). According to this model, the *ga1* and *Ga1-S* haplotypes evolved from a pair of ancestral PME genes through 2 distinct evolutionary branches. In 1 branch, the gene pair underwent pseudogenization followed by multiple duplication events leading to the sequence arrays present in *Z. diploperennis* and *Z. mays parviglumis* accession TIL11, which later formed the nonfunctional *ga1* haplotype in modern maize. In the second branch of the model, the male function gene underwent a series of duplications at 3 different time periods during its evolution, i.e. >600, ~260, and 80–100 KYA to yield several functional copies as well as copies that underwent pseudogenization. The female function gene is also duplicated at a time corresponding to the second duplication event of the male function gene, and together all these sequences constitute the *Ga1-S* haplotype in modern maize.

The *Ga1* locus controls cross-incompatibility in maize. Two different haplotypes of this locus contain structurally and temporally independent repeat regions. The repeat regions both appear to have evolved through multiple rounds of proximal duplication by nonhomologous recombination. Because 1 of the duplication events in the *ga1* array occurred after inactivation of ancestral functional genes, at least this duplication event may have occurred independent of the function of the *Ga1* locus. This is an important case study that may provide insights into the evolution of repeated regions of genomes.

Data availability

The authors affirm that all data necessary for confirming the conclusions of the article are present within the article, figures,

tables, and supplemental material. Supplemental material is available at figshare: https://doi.org/10.25387/g3.24018756. Raw data files are available at https://github.com/amruta0306/G3-2023-404295.

Acknowledgments

This research was supported in part by the U.S. Department of Agriculture, Agricultural Research Service. USDA is an equal opportunity employer. Mention of trade names or commercial products in this report is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture.

Funding

Work on this manuscript was supported by USDA-Agricultural Research Service base funding, Project 21000-050-066D, and USDA National Institute of Food and Agriculture, Organic Research and Education Initiative grant 2020-51300-3210 to MPS.

Conflicts of interest

The authors declare no conflicts of interest.

Literature cited

Cannon EKS, Cannon SB. Chromosome visualization tool: a whole genome viewer. Int J Plant Genomics. 2011;2011:373875. doi:10.1155/2011/373875.

Chen J, Huang Q, Gao D, Wang J, Lang Y, Liu T, Li B, Bai Z, Luis Goicoechea J, Liang C, *et al.* Whole-genome sequencing of *Oryza brachyantha* reveals mechanisms underlying *Oryza* genome evolution. Nat Commun. 2013;4(1):1595. doi:10.1038/ncomms2596.

Chen L, Luo J, Jin M, Yang N, Liu X, Peng Y, Li W, Phillips A, Cameron B, Bernal JS, *et al.* Genome sequencing reveals evidence of adaptive variation in the genus *Zea*. Nat Genet. 2022;54(11):1736–1745. doi:10.1038/s41588-022-01184-y.

Chen Z, Zhang Z, Zhang H, Li K, Cai D, Zhao L, Liu J, Chen H. A pair of non-Mendelian genes at the Ga2 locus confer unilateral cross-incompatibility in maize. Nat Commun. 2022;13(1):1993. doi:10.1038/s41467-022-29729-z.

Clark RM, Tavaré S, Doebley J. Estimating a nucleotide substitution rate for maize from polymorphism at a major domestication locus. Mol Biol Evol. 2005;22(11):2304–2312. doi:10.1093/molbev/msi228.

Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32(5):1792–1797. doi:10.1093/nar/gkh340.

Evans MMS, Kermicle JL. Teosinte crossing barrier1, a locus governing hybridization of teosinte with maize. Theor Appl Genet. 2001;103(2–3):259–265. doi:10.1007/s001220100549.

Francis KE, Lam SY, Copenhaver GP. Separation of *Arabidopsis* pollen tetrads is regulated by QUARTET1, a pectin methylesterase gene. Plant Physiol. 2006;142(3):1004–1013. doi:10.1104/pp.106.085274.

Haberer G, Kamal N, Bauer E, Gundlach H, Fischer I, Seidel MA, Spannagl M, Marcon C, Ruban A, Urbany C, *et al.* European maize genomes highlight intraspecies variation in repeat and gene content. Nat Genet. 2020;52(9):950–957. doi:10.1038/s41588-020-0671-9.

Hufford MB, Seetharam AS, Woodhouse MR, Chougule KM, Ou S, Liu J, Ricci WA, Guo T, Olson A, Qiu Y, *et al.* De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. Science. 2021;373(6555):655–662. doi:10.1126/science.abg5289.

Hufford MB, Xu X, Van Heerwaarden J, Pyhäjärvi T, Chia JM, Cartwright RA, Elshire RJ, Glaubitz JC, Guill KE, Kaeppler SM, *et al.* Comparative population genomics of maize domestication and improvement. Nat Genet. 2012;44(7):808–811. doi:10.1038/ng.2309.

Jones ZG, Goodman MM. Identification of M-type gametophyte factors in maize genetic resources. Crop Sci. 2018;58(2):719–727. doi:10.2135/cropsci2017.09.0560.

Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. Bioinformatics. 2019;35(21):4453–4455. doi:10.1093/bioinformatics/btz305.

Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: molecular evolutionary genetics analysis across computing platforms. Mol Biol Evol. 2018;35(6):1547–1549. doi:10.1093/molbev/msy096.

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. Versatile and open software for comparing large genomes. Genome Biol. 2004;5(2):R12. http://www.tigr.org/software/mummer. doi:10.1186/gb-2004-5-2-r12.

Lin G, He C, Zheng J, Koo DH, Le H, Zheng H, Tamang TM, Lin J, Liu Y, Zhao M, *et al.* Chromosome-level genome assembly of a regenerable maize inbred line A188. Genome Biol. 2021;22(1):175. doi:10.1186/s13059-021-02396-x.

Lu Y, Hokin SA, Kermicle JL, Hartwig T, Evans MMS. A pistil-expressed pectin methylesterase confers cross-incompatibility between strains of *Zea mays*. Nat Commun. 2019;10(1):2304. doi:10.1038/s41467-019-10259-0.

Lu Y, Moran Lauter AN, Makkena S, Scott MP, Evans MMS. Insights into the molecular control of cross-incompatibility in *Zea mays*. Plant Reprod. 2020;33(3–4):117–128. doi:10.1007/s00497-020-00394-w.

Matsubara K, Khin-Thidar, Sano Y. A gene block causing cross-incompatibility hidden in wild and cultivated rice. Genetics. 2003;165(1):343–352. doi:10.1093/genetics/165.1.343.

Moore G, Devos KM, Wang Z, Gale MD. Cereal genome evolution: grasses, line up and form a circle. Curr Biol. 1995;5(7):737–739. doi:10.1016/s0960-9822(95)00148-5.

Moran Lauter AN, Muszynski MG, Huffman RD, Scott MP. A pectin methylesterase ZmPme3 Is expressed in gametophyte factor1-s (Ga1-s) silks and maps to that locus in maize (Zea mays L.). Front Plant Sci. 2017;8(11):1–11. doi:10.3389/fpls.2017.01926.

Panchy N, Lehti-Shiu M, Shiu SH. Evolution of gene duplication in plants. Plant Physiol. 2016;171(4):2294–2316. doi:10.1104/pp.16.00523.

Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, *et al.* The *Sorghum bicolor* genome and the diversification of grasses. Nature. 2009;457(7229):551–556. doi:10.1038/nature07723.

Pattengale ND, Alipour M, Bininda-Emonds OR, Moret BM, Stamatakis A. How many bootstrap replicates are necessary? J Comput Biol. 2010;17(3):337–354. doi:10.1089/cmb.2009.0179.

Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841–842. doi:10.1093/bioinformatics/btq033.

SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL. The paleontology of intergene retrotransposons of maize. Nat Genet. 1998;20(1):43–45. doi:10.1038/1695.

Sun S, Wang J, Yu J, Meng F, Xia R, Wang L, Wang Z, Ge W, Liu X, Li Y, *et al.* Alignment of common wheat and other grass genomes establishes a comparative genomics research platform. Front Plant Sci. 2017;8(8):1480. doi:10.3389/fpls.2017.01480.

Tarailo-Graovac M, Chen N. Using RepeatMasker to identify repetitive elements in genomic sequences. Curr Protoc Bioinforma. 2009;Chapter 4(3):4.10.1–4.10.14. doi:10.1002/0471250953.bi0410s25.

Unterseer S, Seidel MA, Bauer E, Haberer G, Hochholdinger F, Opitz N, Marcon C, Baruch K, Spannagl M, Mayer KFX, *et al.* 2017. European Flint reference sequences complement the maize pan-genome. bioRxiv.:103747. https://doi.org/10.1101/103747.

Wang Y, Li W, Wang L, Yan J, Lu G, Yang N, Xu J, Wang Y, Gui S, Chen G, *et al.* Three types of genes underlying the gametophyte factor1 locus cause unilateral cross incompatibility in maize. Nat Commun. 2022;13(1):4498. doi:10.1038/s41467-022-32180-9.

Wicker T, Gundlach H, Spannagl M, Uauy C, Borrill P, Ramírez-González RH, De Oliveira R; International Wheat Genome Sequencing Consortium; Mayer KFX, Paux E, *et al.* Impact of transposable elements on genome structure and evolution in bread wheat. Genome Biol. 2018;19(1):103. doi:10.1186/s13059-018-1479-0.

Woodhouse MR, Cannon EK, Portwood JL, Harper LC, Gardiner JM, Schaeffer ML, Andorf CM. 2021. A pan-genomic approach to genome databases using maize as a model system. BMC Plant Biol;21:385. doi:10.1186/s12870-021-03173-5

Yang N, Liu J, Gao Q, Gui S, Chen L, Yang L, Huang J, Deng T, Luo J, He L, *et al.* Genome assembly of a tropical maize inbred line provides insights into structural variation and crop improvement. Nat Genet. 2019;51(6):1052–1059. doi:10.1038/s41588-019-0427-6.

Zhang Z, Li K, Zhang T, Chen H. A pollen expressed PME gene at Tcb1 locus confers maize unilateral cross-incompatibility. Plant Biotechnol J. 2023;21(3):454–456. doi:10.1111/pbi.13962.

Zhang Z, Li K, Zhang H, Wang Q, Zhao L, Liu J, Chen H. A single silk- and multiple pollen-expressed PMEs at the Ga1 locus modulate maize unilateral cross-incompatibility. J Integr Plant Biol. 2023;65(5):1344–1355. doi:10.1111/jipb.13445.

Zhang G, Liu X, Quan Z, Cheng S, Xu X, Pan S, Xie M, Zeng P, Yue Z, Wang W, *et al.* Genome sequence of foxtail millet (*Setaria italica*) provides insights into grass evolution and biofuel potential. Nat Biotechnol. 2012;30(6):549–554. doi:10.1038/nbt.2195.

Zhang Z, Zhang B, Chen Z, Zhang D, Zhang H, Wang H, Zhang Y, Cai D, Liu J, Xiao S, *et al.* A PECTIN METHYLESTERASE gene at the maize Ga1 locus confers male function in unilateral cross-incompatibility. Nat Commun. 2018;9(1):3678. doi:10.1038/s41467-018-06139-8.

Editor: J. Holland