nature portfolio

Peer Review File

Constraint of accessible chromatins maps regulatory loci involved in maize speciation and domestication

Corresponding Author: Dr Junpeng Shi

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Attachments originally included by the reviewers as part of their assessment can be found at the end of this file.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
see attached

Reviewer #2

(Remarks to the Author)
The study explores the functional constraint on maize accessible chromatin regions. This study used publicly available ATAC-seq data for six tissues and added six new tissues to create a “pan atlas” of ACRs. These ACRs were used to study evolutionary constraint, trait associated SNPs, possible domestication of ACRs as well as their evolution in subgenomes of maize. Overall, this is an interesting study that contributes some novelty to the field and will be of interest to others. My only major concern is that some of the results have been published previously such as conservation of maize ACRs with sorghum and other grasses, exploration of subgenome dominance, trait-associated SNPs, comparison to teosinte etc. (Lu et al, Nature Plants 2019/Hufford et al Science, 2021/Yan H et al bioRxiv, 2024/Oka et al, Genome Biology, 2017/Marand et al bioRxiv, 2024/Melnyk-Rodgers et al, PNAS, 2016). Surprisingly, only the Hufford/Melnyk-Rodgers studies were cited, but not for their results on ACRs (exception being Melnyk-Rodgers). I think the current study has plenty to offer, but it should do a better job integrating its findings with that of the other labs in the field. Additional comments are listed below.

1. I wouldn’t refer to the ACRs as a pan-atlas. 12 tissues of data are great, but it is far from being a pan-atlas as there are so many additional tissues, cell types, environments and genotypes yet to study.
2. I was unable to find any sequencing summary stats for the ATAC-seq samples. What is the FRiP scores for each sample? How many reads were sequenced? How many aligned? etc.
3. The manuscript is understandable, but it would benefit from an English scientific editor.
4. It’s already been established that there are long-range distal ACRs in maize and some of these are now being redefined as “ultra-distal”. However, the support for these is using H3K4me3 HiChIP data. H3K4me3 localizes to promoters of expressed genes and is not generally found at distal ACRs. When it is found at distal ACRs it almost always indicates an unannotated or misannotated gene. I would include this limitation when defining ultra-distal regions as its well known the maize genome is challenging to annotate.
5. Numerous times it is mentioned that maize is a “giant” “large” genome. I don’t see a need for this, as its not exceptional especially when compared to wheat, barley, etc.
6. Given that 50% of the data for the “pan atlas” is from other labs I don’t think the authors should reference it using possessive language “our” etc.
7. Many of the citations are listed as “invalid citations”

Reviewer #3

(Remarks to the Author)
This manuscript presents a comprehensive study on the accessible chromatins in maize, providing insights into its speciation, domestication, and regulatory sequences. The authors utilize assay for transposase-accessible chromatin using sequencing (ATAC-seq) to profile over 80,000 maize accessible chromatin regions (ACRs), offering a novel perspective on

the regulatory elements in maize genome. This study demonstrates that ACRs evolve faster than coding genes, with a significant portion being unique to the maize lineage, impacting speciation. It also highlights the role of TEs in driving intraspecific innovation of ACRs, showing their impact on gene expression and regulatory evolution in maize. Additionally, it identifies ACRs associated with maize domestication, pointing to potential regulatory loci that have been under selective pressure during the transition from wild teosinte to modern maize. Overall, the study provides a valuable resource for the research community, including a pan-tissue map of ACRs and evolutionary trajectories of regulatory elements in maize and related grasses. This manuscript is concise and well prepared. However, I do have some comments and concerns as follows:

1, Some concepts are very confusing and misleading. For example, the "pan-tissue", which is very confusing, makes me think that the study is about to investigate the tissue variation across different inbred lines/varieties. But the authors focused on the ACR variation across different tissues/stages of a reference inbred line -B73 in maize. Accordingly, the authors adopted some concept from "pan-genomics" and wrongly used in the description of expression patterns/ACR variation across different tissues/stages across the whole life cycle. In terms of expression patterns, we routinely use tissue-specific or constitutively expression for the variation depiction rather than the concepts/definition (Core/variable/tissue-specific in Fig 2g) from pan-genomics.

2, This study has many statistic issues. Actually, the whole study lacks decent controls for all the enrichment analyses, which is the major bioinformatics tool used in this study. For example, the authors claim that "Gypsy accounted for over half of TEs in dACRs, while Helitrons enhanced substantially in pACRs and gACRs" in line151-152. Gypsy is widespread across the whole genome, so it is expected that it accounted for high proportion of the TEs. What I would strongly recommend is to randomly generate an decent control like randomly fractionization of the genome into similar number of genomics regions with the same fragment length distribution to ACRs, then compare the proportion of the different TEs between randomly fractionized genomic dataset and the observed ACRs. And so on for the other enrichment analyses across the whole study.

3, Sub-genome issue. Maize is an ancient allotetraploid species, of while the genome can be classified into two sub-genomes. This is promoted and curated by Dr. James Schnable in 2010 (Schnable, PNAS, 2010). Dr. Schnable kept curating of the sub-genome information, which I believe, is much accurate than the current version of assembly by the authors themselves.

4, This manuscript over-sells their conclusion especially for the maize speciation and domestication. The authors only compared the ACRs of a reference inbred line-B73 with the ancient wild relatives like Sorghum. It is impossible to detect the ACRs involved in the maize speciation and domestication, which need a population study. Although the authors claimed that this is a pan-tissue map of ACRs, it is not a Pan-ACR study. So, Please weaken the conclusion claim in terms of maize speciation and domestication.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
see attached

Reviewer #2

(Remarks to the Author)
I appreciate the response my comments. Overall, this will make a nice contribution to the field. I will state, that I agree with the two reviewers that the claims about domestication and speciation are far too strong. Identifying loci involved in domestication and speciation is challenging and isn't simply done by looking at differences between teosinte and sorghum. In my opinion, this should be removed from the title and the rest of the paper it should be more carefully worded.

Reviewer #3

(Remarks to the Author)
The authors did a very good job to modify the manuscript, which have addressed all my concerns. I would love to read such research paper with significant achievements published in Nature Communications.

Reviewer #1 (Remarks to the Author):

Liu et al. present an analysis of accessible chromatin and comparative genomics between maize, teosinte, and related taxa. The goal of this work is laudable, and the authors are asking a lot of great questions. The ATAC data seem good as well. In the end, however, I am struck by two large concerns with the paper. First, the authors ignore a large swath of relevant literature of comparative genomics in plants and other work looking at regulatory regions. They need to much more to put their current work in context – what is new, how this builds off of or differs from previous results, etc. The second main concern I have is with methods. The authors are not necessarily doing anything wrong alignment approaches etc. all are widely used methods. But the grass genomes are complex, and having played with some of these methods myself I know that they can be biased and problematic. I think the authors need to do more to convince readers that their results are real and not an artifact of whole-genome or short-read alignment limitations.

Response: We really appreciate the thoughtful comments and constructive suggestions from the reviewer. We have revised our manuscript by adding the relevant paper citations suggested by you and other two reviewers in appropriate section. We also offered additional data or references to further substantiate the feasibility of our methods. Please see our detailed point-by-point response below.

- **Major**

1. How much the conclusions about TEs depend on mapping? For example, L148 only 25% of ACRs overlap TEs, but most short reads won't map to TEs (see https://www.biorxiv.org/content/10.1101/2024.07.10.602892v1)

Response: We thank the reviewer for this thoughtful comment. Mapping short reads onto TEs will introduce a certain extent of bias. However, ATAC-seq remains the most commonly used method for assessing chromatin accessibility not only in maize[1,2] (https://doi.org/10.1016/j.cell.2021.04.014; https://doi.org/10.1038/s41477-019-0547-0), but also in other complex species e.g., wheat[3] (https://doi.org/10.1186/s13059-020-02093-1), cotton[4] (https://doi.org/10.1073/pnas.2209743119), human[5]

(https://doi.org/10.1038/s41467-024-51921-6) and mouse[6] (https://doi.org/10.1186/s13059-020-02164-3) with substantially high TEs.

The abovementioned biorxiv paper applied the recently developed Fiber-seq to map accessible chromatin in maize, which based on PacBio sequencing and theoretically had higher accuracy in identifying ACRs over ATAC-seq within TEs. However, we analyzed the ACRs released in this Fiber-seq paper and found a comparable level of TE content within the Fiber-seq ACRs (17.0%) with those called by ATAC-seq (18.3%) in our leaf tissue data. Therefore, we still kept the conclusion of "the enrichment of low-copy sequences in maize" in our manuscript (Line 154). Since the maize Fiber-seq paper had great significance, we re-organized a new sentence in Discussion section by adding it as a reference (Line 481).

2. The authors cite several mammal papers but throughout seem to suggest there has been little work on conservation in noncoding regions in plants. There are several such papers, including work by e.g. Baoxing Song who specifically looks at constrained noncoding sequences in maize https://genome.cshlp.org/content/31/7/1245, or even databases of noncoding regulatory sequence across more species: https://academic.oup.com/nar/article/52/D1/D1569/7332064. The authors need to do a lot more to explain the uniqueness of their work and put it in the context of a considerable body of previous work.

Response: Thank you for this suggestion. We admit that there were several published papers discussing the evolutionary constraint of non-coding sequences in plants[7-9] (https://genome.cshlp.org/content/31/7/1245; https://academic.oup.com/nar/article/52/D1/D1569/7332064; https://doi.org/10.1093/plcell/koae064;). As exemplified by the reviewer, Song et al. have studied the constraint of maize non-coding sequences across Andropogoneae, while they only identified the functional CNSs according to the degree of sequence constraint since they did not incorporate enough maize ATAC-seq or other similar data. Similarly, the major goal of cross-species CNS paper published in NAR was to study the constraint of gene promoters, by comparison the distal regulatory elements were less investigated. Generally, most published plant studies have addressed the evolurionary constraint of

CNSs, while they often lacked enough data to further annotate and classify CREs among these CNSs. We re-organized some sentences (Line 102-106) in Introduction to summarize these related works.

3. Comparative analysis of ACRs in other genomes requires quality alignments. My experience with Cactus in grass genomes is that much of the genome doesn't align. The number of constrained ACRs in the phylogeny (Fig S9) doesn't seem to correlate with phylogenetic distance outside of Zea, as one might expect (e.g. 4,000 in Pennisetum vs only 1,000 in Hyperrhenia but 7,000 in the equally distant Sorghum). I think it important for the authors to characterize the quality of their alignments and discuss some how they deal with this. For example, it looks to me that the authors are treating "no alignment" as "ACR not present"? It might be better to only assay ACR presence/absence if flanking sequences do align. The alignment strategy may also explain the large number of maize-only ACR, especially those that are in TEs (and thus don't align well with other genomes).

Response: We really appreciate your valuable comments. Since the genome quality of non-maize species might have a major impact on Cactus results, we reanalyzed the genome sequences we previously selected, finding approximately one-third of them were poorly assembled (contig N50 < 100 kb). Therefore, we re-screened the genomes of Poaceae species with high sequence quality (contig N50 > 900 kb, all were diploids, Supplementary Table 4 & Response_Table_R1), and updated our Catcus alignment of grass species. The major conclusion that “the number of highly constrained ACRs was negatively correlated with their pairwise divergence” remained unchanged, and the correlation between the number of constrained ACRs and pairwise phylogenetic distance also improved. Please also see our updated Supplementary Fig. 9 for details.

Response_Table_R1: The genome quality, Cactus alignment results and number of constrained ACRs

Species	Divergence time (mya)	Contig N50 (kb)	Genome size (bp)	Ungapped alignment (bp)	Number of constrained ACRs
Coix aquatica	11.19	2,240	1,615,469,127	108,873,576	6793
Coix lacryma-jobi	11.19	3,190	1,731,458,735	115,057,653	6308

Erianthus rufipilus	11.19	76,900	858,700,241	109,104,217	7473
Sorghum bicolor	11.19	3,000	709,344,700	100,978,885	7117
Echinochloa haploclada	22.8	930	443,511,613	82,852,388	3385
Panicum hallii	22.8	1,100	535,889,049	84,475,945	3546
Pennisetum glaucum	22.8	22,300	1,962,243,091	90,124,065	2973
Setaria italica	22.8	20,596	405,868,399	81,822,467	3592
Setaria viridis	22.8	11,200	395,869,604	81,602,436	3610
Paspalum notatum	27.8	49,000	540,949,507	82,470,740	3412
Paspalum vaginatum	27.8	1,500	517,974,767	80,681,790	3238
Eleusine indica	48	42,100	522,528,889	65,225,404	1992
Aegilops bicornis	49.1	3,800	5,902,718,208	65,544,428	942
Aegilops tauschii	49.1	1,900	4,121,004,028	61,631,809	1068
Alopecurus myosuroides	49.1	1,800	3,571,731,728	63,230,326	1000
Brachypodium distachyon	49.1	22,000	271,298,618	50,882,981	1313
Dactylis glomerata	49.1	930	1,781,320,693	57,227,315	1183
Dasypyrum villosum	49.1	4,270	4,053,807,853	64,473,977	1094
Hordeum marinum	49.1	6,830	3,815,965,547	61,615,746	1113
Hordeum spontaneum	49.1	3,525	5,110,990,204	69,603,204	822
Hordeum vulgare	49.1	69,600	4,225,713,981	60,814,777	1065
Lolium perenne	49.1	11,740	2,276,788,877	57,623,847	1114
Olyra latifolia	49.1	13,700	639,189,835	59,248,352	1610
Oryza barthii	49.1	14,700	347,715,970	52,753,311	1386
Oryza glaberrima	49.1	15,200	347,453,675	52,870,053	1414
Oryza rufipogon	49.1	26,000	462,580,640	57,075,624	1307
Oryza sativa	49.1	31,111	374,422,835	53,776,963	1421
Poa infirma	49.1	84,500	1,151,728,964	54,103,590	1157
Poa supina	49.1	1,600	648,010,102	51,430,476	1148
Raddia guianensis	49.1	3,500	607,833,752	56,096,176	1379
Thinopyrum elongatum	49.1	2,150	4,634,136,514	62,132,290	1081
Triticum monococcum	49.1	55,904	5,154,690,404	64,377,223	1122
Zizania latifolia	49.1	4,480	547,397,560	69,355,698	776
Pharus latifolius	66.2	4,600	1,002,922,384	53,864,250	1053
Arabidopsis thaliana	159.6	11,200	119,668,634	18,085,739	209
Glycine max	159.6	51,200	978,941,695	28,534,899	79
Gossypium hirsutum	159.6	98,600	2,306,070,423	47,332,417	50
Solanum lycopersicum	159.6	41,700	827,962,994	20,184,018	179

Regarding the feasibility of Cactus, it has been commonly used to align the genome between two different crop species, e.g., maize and sorghum (by Baoxing Song, https://pubmed.ncbi.nlm.nih.gov/38796435/), or among Solanaceae species (By Sanwen Huang, https://www.cell.com/cell/fulltext/S0092-8674(23)00405-1). Therefore, we believe that this method is applicable. Notably, the conserved sequences between maize and sorghum we identified here (101 Mb, Supplementary Table 4) were even higher than

the CNS spaces (67.1 Mb) identified by Baoxing Song between maize and sorghum (the genic spaces were excluded in his analysis). We estimate comparable Cactus results if the genic regions were incorporated in Baoxing's analysis.

We also tried to assess the presence or absence of ACRs only if the flanking regions (1-kb down and upstream) of ACRs can be aligned (>10% alignment) as suggested by the reviewer. However, this filtering resulted in only 4.8% to 19% of ACRs being available for further analysis (see Response_Table_R2). Meanwhile, as another independent validation of maize-specific ACRs, we utilized Minimap2 to realign cactus-identified maize-specific ACRs onto non-maize grass genomes and found that approximately 96% of maize-specific ACRs still could not find orthologous sequences on other grass genomes (alignment coverage <=10%). Collectively, these analysis indicated good reliability of cactus-identified maize-specific ACRs.

Response_Table_R2**.** The number of ACRs only if their flanking regions (1-kb down and upstream) can be aligned (>10% alignment).

Species	**Divergence time (mya)**	**The number of ACR with their flanking regions aligned**
Coix aquatica	11.19	14246
Coix lacryma-jobi	11.19	13015
Erianthus rufipilus	11.19	15685
Sorghum bicolor	11.19	15637
Echinochloa haploclada	22.8	9778
Panicum hallii	22.8	10622
Pennisetum glaucum	22.8	8423
Setaria italica	22.8	10487
Setaria viridis	22.8	10553
Paspalum notatum	27.8	10036
Paspalum vaginatum	27.8	9631
Eleusine indica	48	7359
Aegilops bicornis	49.1	3249
Aegilops tauschii	49.1	3761
Alopecurus myosuroides	49.1	3434
Brachypodium distachyon	49.1	4877

Dactylis glomerata	49.1	4101
Dasypyrum villosum	49.1	3804
Hordeum marinum	49.1	3681
Hordeum spontaneum	49.1	2806
Hordeum vulgare	49.1	3627
Lolium perenne	49.1	3856
Olyra latifolia	49.1	5446
Oryza barthii	49.1	5346
Oryza glaberrima	49.1	5312
Oryza rufipogon	49.1	4941
Oryza sativa	49.1	5352
Poa infirma	49.1	4094
Poa supina	49.1	4031
Raddia guianensis	49.1	4924
Thinopyrum elongatum	49.1	3566
Triticum monococcum	49.1	3774
Zizania latifolia	49.1	2106
Pharus latifolius	66.2	3892

4. The sampling of genomes for domestication comparison I think is an issue for making claims about domestication, and at least needs to be discussed. Many loci changed in frequency dramatically due to the bottleneck associated with domestication, and with only one genome from each teosinte, it's hard to rule out that a locus that is fixed in maize isn't at, for example, 80% frequency in parviglumis (and was just missed with 20% probability in the one genome sampled). Such a locus would not widely be considered a key domestication gene, however!

Response: Thanks. To address this question, we further incorporated all available high-quality teosinte genomes (3 *Zea mays* ssp. *parviglumis* and 2 *Zea mays* ssp. *mexicana* genome) in MaizeGDB into our intraspecific analysis. We updated our analysis which resulted in 856 domestication-related ACRs (conserved in over 90% of analyzed maize genomes but absent in at least 3 teosinte genomes), finding the major conclusion regarding the genes that regulated by these domestication-related ACRs remained unchanged. Please see Line 314-353 in updated manuscript for details.

5. The grammar is frequently awkward, occasionally complicating or preventing clear interpretation. This needs careful revision.

Response: Thanks. We have used a professional English editing service to revise our manuscript thoroughly. We attached the editorial certificates below.

Charlesworth Author Services

EDITORIAL CERTIFICATE

This document certifies that the manuscript below was edited for correct English language usage, grammar, punctuation and spelling by qualified native English speaking editors at Charlesworth Author Services.

Paper Title:

Inter- and intraspecific constraint of accessible chromatins maps regulatory loci involved in maize speciation and domestication

Author:

Yuting Liu

Date certificate issued:

December 16, 2024

cwauthors.com

- **Minor**

1. L407 -- I think this is probably right (most TFBS are shared between m1 and m2) but perhaps worth the caveat that the authors have only looked at a small fraction. From my reading of the methods the authors can only identify ~300 TF families of the thousands that are in maize.

Response: Thanks. We have collected motif information of maize TFs from PlantTFDB and JASPER (two commonly used PlantTF databases), complemented by motif discovery

of an additional set of 12 TFs by DAP-seq, resulting in 306 TFs with high-quality motif information. To our knowledge, these motif collections encompassed nearly all publicly available maize TF motifs to date. We do agree that they represented a small fraction of all maize TFs (~2,300), while they belonged to 36 TF families which have accounted for more than half (65%, 36/56) of total maize TF families. In addition, the core motifs of TFs within a same TF family were usually highly similar[10] (https://doi.org/10.1016/j.cell.2018.01.029). Therefore, we think that our conclusion of "most TFBS are shared between m1 and m2" is reasonable (Line 396-399).

2. {{[[TODO]]}} Perhaps a lot to ask to integrate, but RAMPAGE data from the parviglumis TIL11 was made available by maize ENCODE http://www.maizecode.org/data/til11.html Even if the authors don't use the data, citing and discussing the work of the project is likely relevant https://www.biorxiv.org/content/10.1101/2024.02.22.581585v1

Response: We have added this preprint as a citation in relevant sections. Please see Line 316 and Line 500.

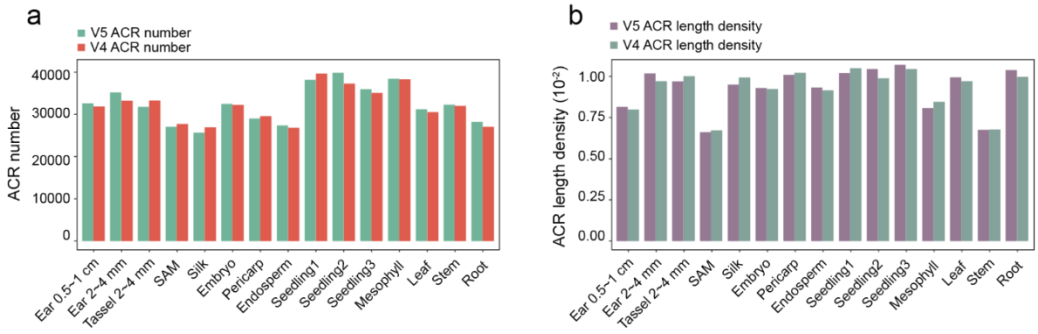
3. I'm curious why the authors are using v4 of B73 instead of the updated v5.

Response: We integrated many published data in our work which were based on B73_V4 genome coordinates (e.g., the selective intervals identified by Wenkang Chen[11] (https://www.science.org/doi/full/10.1126/science.abg7985), as well as the GWAS SNPs identified by Chunhui Li[12] (https://doi.org/10.1038/s41477-022-01190-2). To avoid potential coordinate conversion errors between V4 and V5, we choose to map our ATAC-seq data to B73_V4 genome. In fact, we found minimal differences when ATAC-seq reads were mapped to V4 and V5 genomes. Please see Response_Table_R3, Response_Figure_R1 and Response_Figure_R2 for details.

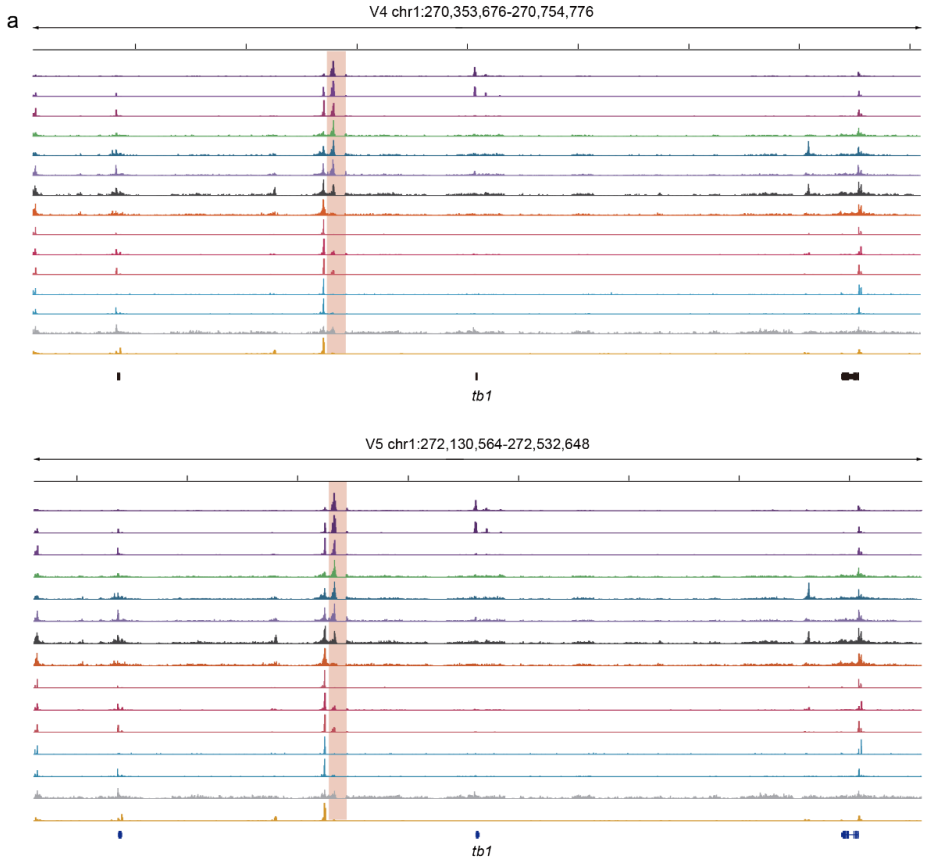
Response_Table_R3. Mapping statistics of ATAC-seq data using B73v4 and v5 genomes.

Sample	Tissue name	Raw reads	Alignment rate (V4, %)	Alignment rate (V5, %)

B7312DAPPrep1	Embryo rep1	76,824,777	99.25	98.67
B7312DAPPRrep1	Endosperm rep1	78,040,450	99.15	98.51
B7312DAPZPrep1	Pericarp rep1	70,116,627	99.12	98.96
B73SAMrep1	SAM rep1	69,353,055	98.31	97.43
B73Silkrep1	Silk rep1	66,182,655	99.1	99.3
B73Stemrep1	Stem rep1	65,149,485	99.35	99.28
SRR10873334	Tassel 2-4mm rep1	30,981,941	98.64	98.07
SRR10873336	Ear 2-4mm rep1	35,278,782	98.48	97.81
SRR11955290	Root 6days rep1	88,756,880	55.36	55.31
SRR12321695	Seedling1 rep1	8,212,025	95.06	95.6
SRR8742453	Leaf 7days rep1	95,964,379	92	91.71
SRR7889829	Ear 0.5-1cm rep1	76,811,222	85.71	85.05
SRR14327895	Seedling2 rep1	128,767,853	89.42	89.73
SRR14327897	Seedling3 rep1	244,130,154	95.12	95.04
SRR5748809	Mesophyll rep1	27,933,187	97.34	97.64



Response_Figure_R1. The number and length of ACRs identified in each sample between B73v4 (a) and B73v5 (b).



Response_Figure_R2. The chromatin accessibility of *tb1* and its ~200 kb flanking region using V4 and V5 reference genomes.

4. none of the MOA-seq work by Hank Bass or Thomas Hartwig are cited, but this literature seems relevant

Response: Thanks. We added this work as a reference. Please see Line 108.

5. L487 why does a bottleneck result in a loss-of-function? the logic here is unclear to me

Response: Thank you for pointing out this. Our logic here was to describe the loss of functional DNA sequences due to domestication bottleneck. We revised this sentence to make clarification. Please see Line 498.

6. I somehow missed where the assemblies of the other species come from and what quality they are.

Response: Please see our revised Supplementary Table 4 for details.

7. Comparison of trait variation with ACRs should probably reference Rogers-Melnick 2016 https://pubmed.ncbi.nlm.nih.gov/27185945/ and Englehorn 2024 https://www.biorxiv.org/content/10.1101/2023.08.08.551183v2

Response: Done. Please see Line 108 in Intro section.

8. L591 says 38 species and 4 outgroups

Response: A typo. Please see revised Line 600.

9. Why highlight that figure 2b is a triangle like in mammals (L212)? This seems to suggest something about comparable results, but is there any other shape this figure could possibly take?

Response: Thanks. This figure can only be a triangle since 0 <= (N1 + N2) =< 34. We delete this sentence in our revised manuscript (Line 217).

10. L147 the maize genome is below average size for angiosperms, it is not "bloated"

Response: Thanks. We have made this revision (Line 153).

11. gACR, pACR, dACR acronyms should be defined on L137

Response: Done. Please see revised Line 140-144.

12. I would encourage authors to use "Ty3" instead of "Gypsy" see here: https://osf.io/preprints/osf/fma57

Response: Thank you for your valuable suggestions, we have update it in the revised manuscript. Please see Line 156 and Line 227.

13. L155 an alternative explanation to purifying selection is TE insertion preferentially into other TEs or closed chromatin. With a single genome this can't be disentangled.

Response: We deleted this sentence to make more clarification. Please see Line 161.

14. L177 is Michaelis-Menton standardly used this way? It seems like doing a rarefaction subsampling analysis might be more appropriate? But OK if this is standard for ATAC and similar analyses.

Response: Michaelis-Menton analysis was standardly used to estimate the saturation of ATAC-seq data, for example, in Bob's maize single-cell ATAC-seq paper[2] (https://doi.org/10.1016/j.cell.2021.04.014).

15. L189 Genic regions of maize have lower diversity than non-genie. See e.g. https://www.nature.com/articles/nplants201684

Response: Thanks. In our original manuscript, we cited the early year (2012) Maize HapMap2 paper (https://www.nature.com/articles/ng.2309), which reported higher genetic diversity in maize genic regions over non-genic regions (Supplementary Table 2 in that paper). However, this result might come from the higher alignment ratio of short reads (<=100 bp) in genic over non-genic regions. With the improvement of read length, quality and mapping algorithm, it is now acknowledged that genic regions should have lower genetic diversity than non-genic regions as suggested by the NP paper (https://www.nature.com/articles/nplants201684) as exemplified by the reviewer. We modified this sentence to avoid ambiguity. Please see Line 192-193.

Reviewer #2 (Remarks to the Author):

The study explores the functional constraint on maize accessible chromatin regions. This study used publicly available ATAC-seq data for six tissues and added six new tissues to create a "pan atlas" of ACRs. These ACRs were used to study evolutionary constraint, trait associated SNPs, possible domestication of ACRs as well as their evolution in subgenomes of maize. Overall, this is an interesting study that contributes some novelty to the field and will be of interest to others. My only major concern is that some of the results have been published previously such as conservation of maize ACRs with

sorghum and other grasses, exploration of subgenome dominance, trait-associated SNPs, comparison to teosinte etc. (Lu et al, Nature Plants 2019/Hufford et al Science, 2021/Yan H et al bioRxiv, 2024/Oka et al, Genome Biology, 2017/Marand et al bioRxiv, 2024/Melnyk-Rodgers et al, PNAS, 2016). Surprisingly, only the Hufford/Melnyk-Rodgers studies were cited, but not for their results on ACRs (exception being Melnyk-Rodgers). I think the current study has plenty to offer, but it should do a better job integrating its findings with that of the other labs in the field. Additional comments are listed below.

Response： Thank you for your support and valuable suggestion. We have incorporated these papers in relavent sections (Lu et al, Nature Plants 2019. Oka et al, Genome Biology, 2017. Melnyk-Rodgers et al, PNAS, 2016. Line 108. & Yan H et al bioRxiv, 2024. Marand et al bioRxiv, 2024. Line 108 and Line 476. & Hufford et al Science, 2021. Line 476.).

1. I wouldn't refer to the ACRs as a pan-atlas. 12 tissues of data are great, but it is far from being a pan-atlas as there are so many additional tissues, cell types, environments and genotypes yet to study.

Response: Thanks. We have revised "pan-tissue" to "multi-tissue" throughout our manuscript.

2. I was unable to find any sequencing summary stats for the ATAC-seq samples. What is the FRiP scores for each sample? How many reads were sequenced? How many aligned? etc.

Response: Thanks. Please see our revised **Supplementary Table 1** for details.

3. The manuscript is understandable, but it would benefit from an English scientific editor.

Response: Thanks. We have used a professional English editing service to revise our manuscript thoroughly. Please see the editorial certificates attached in reviewer1' section.

4. It's already been established that there are long-range distal ACRs in maize and some of these are now being redefined as "ultra-distal". However, the support for these is using H3K4me3 HiChIP data. H3K4me3 localizes to promoters of expressed genes and is not generally found at distal ACRs. When it is found at distal ACRs it almost always indicates an unannotated or misannotated gene. I would include this limitation when defining ultra-distal regions as its well known the maize genome is challenging to annotate.

Response: We agreed that H3K4me3 was a canonical histone mark of promoter of expressed genes. However, H3K4me3 can also be associated with the promoter of expressed lncRNAs in distal intergenic regions (e.g., https://www.nature.com/articles/s41467-018-07500-7). Additionally, Bob's lab also used this H3K4me3 HiChIP data to build the chromatin interactions between dACRs and promoters (Lu, et al. Nat Plants, 2019). Anyway, we revised the relevant sentences and weakened the statement of "indicating" to "suggesting". Please see Line 151.

5. Numerous times it is mentioned that maize is a "giant" "large" genome. I don't see a need for this, as its not exceptional especially when compared to wheat, barley, etc.

Response: Thank you for pointing out this. We made corrections throughout the manuscript.

6. Given that 50% of the data for the "pan atlas" is from other labs I don't think the authors should reference it using possessive language "our" etc.

Response: Thank you for pointing out this. We revised these sentences (Line 182, Line 473).

7. Many of the citations are listed as "invalid citations"

Response: Done.

Reviewer #3 (Remarks to the Author):

This manuscript presents a comprehensive study on the accessible chromatins in maize, providing insights into its speciation, domestication, and regulatory sequences. The authors utilize assay for transposase-accessible chromatin using sequencing (ATAC-seq) to profile over 80,000 maize accessible chromatin regions (ACRs), offering a novel perspective on the regulatory elements in maize genome. This study demonstrates that ACRs evolve faster than coding genes, with a significant portion being unique to the maize lineage, impacting speciation. It also highlights the role of TEs in driving intraspecific innovation of ACRs, showing their impact on gene expression and regulatory evolution in maize. Additionally, it identifies ACRs associated with maize domestication, pointing to potential regulatory loci that have been under selective pressure during the transition from wild teosinte to modern maize. Overall, the study provides a valuable resource for the research community, including a pan-tissue map of ACRs and evolutionary trajectories of regulatory elements in maize and related grasses. This manuscript is concise and well prepared. However, I do have some comments and concerns as follows:

Response: We really appreciate your support and insightful suggestions. Please find our point-by-point response below.

1. Some concepts are very confusing and misleading. For example, the "pan-tissue", which is very confusing, makes me think that the study is about to investigate the tissue variation across different inbred lines/varieties. But the authors focused on the ACR variation across different tissues/stages of a reference inbred line -B73 in maize. Accordingly, the authors adopted some concept from "pan-genomics" and wrongly used in the description of expression patterns/ACR variation across different tissues/stages across the whole life cycle. In terms of expression patterns, we routinely use tissue-specific or constitutively expression for the variation depiction rather than the concepts/definition (Core/variable/tissue-specific in Fig 2g) from pan-genomics.

Response: Thanks. As suggested by you and reviewer2, we revised the term of “pan-tissue” to “multi-tissue” throughout our manuscript. We also renamed the ACRs into

constitutive, variable and tissue-specific ACRs according to their existence in different tissues. Please see Line 185-193.

2. This study has many statistic issues. Actually, the whole study lacks decent controls for all the enrichment analyses, which is the major bioinformatics tool used in this study. For example, the authors claim that "Gypsy accounted for over half of TEs in dACRs, while Helitrons enhanced substantially in pACRs and gACRs" in line151-152. Gypsy is widespread across the whole genome, so it is expected that it accounted for high proportion of the TEs. What I would strongly recommend is to randomly generate an decent control like randomly fractionization of the genome into similar number of genomics regions with the same fragment length distribution to ACRs, then compare the proportion of the different TEs between randomly fractionized genomic dataset and the observed ACRs. And so on for the other enrichment analyses across the whole study.

Response: Thank you for your valuable suggestion. We added controls for all the enrichment analyses in the revised manuscript. Please see revised Fig. 1e, Fig.2c, 2d, Fig.3b and 3c, and Supplementary Fig. 5a.

3. Sub-genome issue. Maize is an ancient allotetraploid species, of while the genome can be classified into two sub-genomes. This is promoted and curated by Dr. James Schnable in 2010 (Schnable, PNAS, 2010). Dr. Schnable kept curating of the sub-genome information, which I believe, is much accurate than the current version of assembly by the authors themselves.

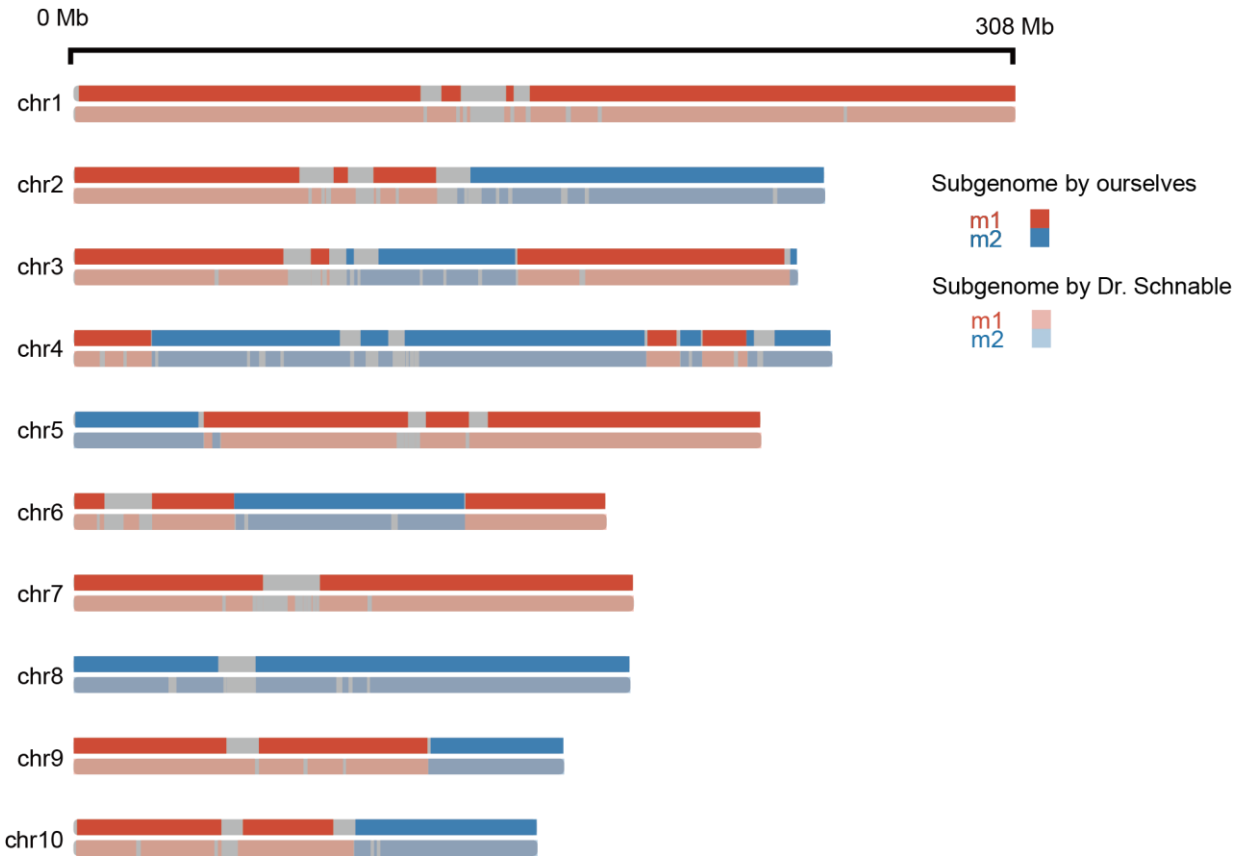
Response: Thanks. The subgenome information in Schnable, PNAS, 2010 was based on B73_V2 genome. We communicated with Dr. Schnable, and he provided us a link for the subgenome information in B73_V4 version that curated by his lab (https://figshare.com/articles/dataset/Grass_Syntenic_Gene_List_sorghum_v3_maize_v3_4_with_teff_and_oropetium_v2/7926674/1?file=14750297). We then compared our subgenome information with Dr. Schnable's, finding minimal differences (1%~2%) that were largely due to ambiguous assignment near breakpoints (see Response_Figure_R3). In our work, we identified the subgenome of B73_V4 by CoGE, an easy-to-use online tool that developed by Mike Freeling and James Schnable, and we have published some

works using this tool in subgenome identification (e.g. the maize Mo17 Nat Genet paper in 2018 and the broomcorn millet PlantCom paper in 2023). Please see the Response_Table_R4 and Response_Figure_R3 below for details.

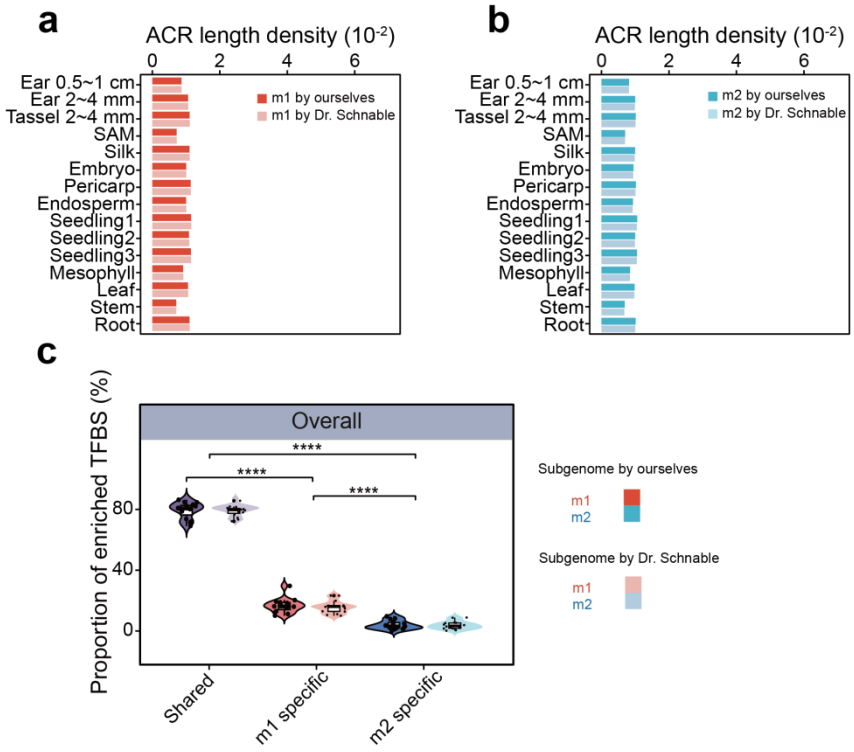
We also compared with ACR statistics using the subgenome information we identified and those provided by Dr. James, finding minimal differences. Please also see Response_Figure_R4 for details.

Response_Table_R4. A comparison of subgenome information.

	m1	**m2**
Our's results	1,180,984,534	725,578,502
Dr. Schnable	1,200,668,308	710,348,070
Difference in length	~1.6%	~2.1%



Response_Figure_R3. Detailed distribution of different subgenome versions across maize chromosomes.



Response_Figure_R4. A comparison of ACRs statistics between the subgenomes we identified and those provided by Dr. Schnable. (a) ACR length density across 15 samples of different subgenome1 versions. (b) ACR length density across 15 samples of different subgenome2 versions. (c) The proportion of enriched TFBSs that shared between two subgenomes and specifically existed in m1 or m2 subgenome.

4. This manuscript over-sells their conclusion especially for the maize speciation and domestication. The authors only compared the ACRs of a reference inbred line-B73 with the ancient wild relatives like Sorghum. It is impossible to detect the ACRs involved in the maize speciation and domestication, which need a population study. Although the authors claimed that this is a pan-tissue map of ACRs, it is not a Pan-ACR study. So, Please weaken the conclusion claim in terms of maize speciation and domestication.

Response: Thanks. The primary goal of our work was to profile the evolutionary trajectory of maize ACRs. When designing this work at early time, we had two different

options. One was to generate ATAC-seq for multiple tissues for a single genotype which we finally did in this study, the other was to generate ATAC-seq data for multiple genotypes in a few (one or two) tissues. No matter what strategy was used, the method we applied to trace the interspecific and intraspecific constraint of ACRs was applicable[13] (https://doi.org/10.1126/science.abn7930). On the other hand, the ~80000 ACRs we identified here should have accounted for a considerable proportion of maize ACRs, offering statistically enough data to characterize ACRs that involved in maize specification and domestication, although we could not exclude the existence of ACRs that specifically existed in other non-B73 genotypes. Therefore, we expressed our anticipation of the forthcoming ATAC-seq data of diverse maize genotypes[14,15] (https://www.science.org/doi/10.1126/science.abg5289; https://www.biorxiv.org/content/10.1101/2024.10.14.618191v1.full) in the Discussion section (see Line 476).

References

1. Ricci, W.A. *et al.* Widespread long-range cis-regulatory elements in the maize genome (vol 5, pg 1237, 2019). *Nature Plants* **6**, 328-328 (2020).
2. Marand, A.P., Chen, Z., Gallavotti, A. & Schmitz, R.J. A cis-regulatory atlas in maize at single-cell resolution. *Cell* **184**(2021).
3. Jordan, K.W., He, F., Soto, M.F.D., Akhunova, A. & Akhunov, E. Differential chromatin accessibility landscape reveals structural and functional features of the allopolyploid wheat chromosomes. *Genome Biology* **21**(2020).
4. Han, J. *et al.* Genome-wide chromatin accessibility analysis unveils open chromatin convergent evolution during polyploidization in cotton. *Proceedings of the National Academy of Sciences* **119**, e2209743119 (2022).
5. Du, A.Y., Chobirko, J.D., Zhuo, X., Feschotte, C. & Wang, T. Regulatory transposable elements in the encyclopedia of DNA elements. *Nature Communications* **15**, 7594 (2024).
6. Miao, B. *et al.* Tissue-specific usage of transposable element-derived promoters in mouse development. *Genome Biology* **21**, 255 (2020).

7. Song, B. *et al.* Conserved noncoding sequences provide insights into regulatory sequence and loss of gene expression in maize. *Genome Res* **31**, 1245-1257 (2021).
8. Chow, C.-N. *et al.* PlantPAN 4.0: updated database for identifying conserved non-coding sequences and exploring dynamic transcriptional regulation in plant promoters. *Nucleic Acids Research* **52**, D1569-D1578 (2024).
9. Xin, H. *et al.* Identification and functional characterization of conserved cis-regulatory elements responsible for early fruit development in cucurbit crops. *Plant Cell* **36**, 2272-2288 (2024).
10. Lambert, S.A. *et al.* The Human Transcription Factors. *Cell* **172**, 650-665 (2018).
11. Chen, W. *et al.* Convergent selection of a WD40 protein that enhances grain yield in maize and rice. *Science* **375**, eabg7985 (2022).
12. Li, C. *et al.* Genomic insights into historical improvement of heterotic groups during modern hybrid maize breeding. *Nature Plants* **8**, 750-763 (2022).
13. Andrews, G. *et al.* Mammalian evolution of human cis-regulatory elements and transcription factor binding sites. *Science* **380**, eabn7930 (2023).
14. Hufford, M.B. *et al.* De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. *Science* **373**, 655-+ (2021).
15. Zhu, Y. *et al.* Pan-cistrome analysis of the leaf accessible chromatin regions of 214 maize inbred lines. *bioRxiv*, 2024.10.14.618191 (2024).

Reviewer #1 (Remarks to the Author):

Major:

1. The authors have done a nice job addressing most of my concerns about TEs (looking at fiber-seq) and alignment (using flanking sequences, filtering for genome quality). I think still worth noting in the discussion the potential caveat that conserved ACRs are more strongly correlated with the size of the alignment in bp (r=0.89) than the divergence time (r=-0.64). Of course, the amount that can be aligned is negatively correlated with time (r=-0.82) so it's hard to disentangle these. But I could imagine an orthologous gene with conserved regulation between two species – same upstream TF binding site with same TFs binding – but where there is no "conserved ACR" because the intergenic DNA between the gene and the TF binding site has changed so much that alignment is not possible. This seems plausible given how few of the flanking regions of ACRs Can such a scenario be ruled out? If not, worth adding a sentence or so in the discussion that this is difficult to disentangle.

Response: Thank you for your support and thoughtful suggestions. We do agree with your assumption that a conserved TF binding site in the upstream of an orthologous gene pair might be miss-identified due to drastic changes of intergenic sequences between TF binding site and gene. It requires more sophisticated method to trace the constraint of binding motifs or CREs, which are usually short (tens of base-pairs) that composed a small proportion of ACRs. We have added a new sentence (Line 496-499) in Discussion.

2. The writing is improved, but still awkward and at sometimes hard to follow. I find that LLMs actually do a wonderful job of clearing this up. Literally pasting paragraphs from the paper into CoPilot, for example, cleared up minor grammatical issues that remained and clarified wording. I recommend this to folks all the time, and see no difference between this and e.g. Microsoft Word's grammar check or editing suggestions.

For example, CoPilot returns:

The intricate regulation of gene expression is crucial for executing complex biological processes in higher organisms. Maize, a classical genetic model in gene regulation, has been extensively studied since Dr. Barbara McClintock's groundbreaking discovery of transposable elements (TEs), also known as "jumping genes," which cause the mosaic pigmentation seen on maize kernels. Recent assembly and analysis of maize genomes have revealed a highly sophisticated and hierarchical
gene regulation system, characterized by the prevalence of distal cis-regulatory elements (CREs), particularly enhancers. This complexity is likely due to the large maize genome (~2.3 Gb), with over 85% of its DNA composed of TEs. Genetic mapping in maize has shown an enrichment of quantitative trait nucleotides (QTNs) within non-genic regulatory regions, including well-known loci such as tb1, vgt1, KRN4, ZmCCT9, ZmCCT10, and UPA1. On a higher-dimensional scale, the widespread distal CREs in maize can coordinate with transcription factors (TFs) and other trans-acting elements while interacting with promoters or other gene-proximal regions to regulate gene expression through chromatin interactions. These findings highlight the crucial role of regulatory sequences in controlling maize development and trait variations.

Or this one:

The proliferation of the maize genome has led to a significant number of distal accessible chromatin regions (dACRs) being located in "gene deserts" (regions 100 kb away from the nearest gene). Interestingly, these ultra-

distal ACRs showed high conservation across the tissues we profiled and were independently verified by maize single-cell ATAC-seq data and ChIP-seq data of various active histone modifications. They were also found to be involved in chromatin interactions, with some linked to genes in 3D genomic data, suggesting that maize ACRs play a cis-regulatory role both locally and distally.

Response: Thank you for your suggestion. We used CoPilot to revise our abstract, introduction and some sentences in discussion. However, we minimized changes of Results to maintain result accuracy as much as possible.

Minor:

1. L61 why not call these QTL, since for most of the examples the causal locus (if known) is not a single nucletotide

Response: Thank you for your suggestion. We made revisions in Line 55.

2. L192 I'm not sure I follow the authors' interpretation of selection here. Lower SNP density in constitutive does perhaps imply conservation. But those same regions have higher SNP density around, and I don't see an explanation for this.

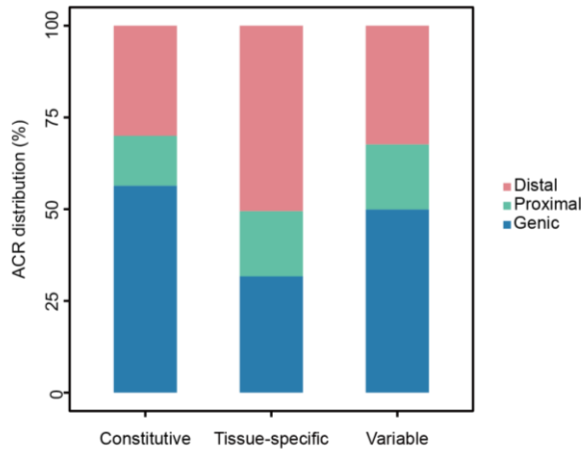
Response: Thanks. We calculated the genome-wide density of HapMap3 SNPs in genic, proximal, and distal regions respectively, finding highest SNP density in gene proximal region and lowest in distal region (Response_Table_R1). We also filtered out high-quality HapMap3 SNPs (MAF <= 0.01) and got same results (Response_Table_R1). Meanwhile, the constitutive and variable ACRs were more enriched in genic and proximal regions than tissue-specific ACRs (Response_Figure_R1), which might explain the significantly higher SNP density in flanking regions of constitutive and variable ACRs than tissue-specific ACRs.

In the original version of our manuscript, we explained these data as "the flanking regions of core (constitutive) and variable ACRs shown significantly higher SNP

density than private (tissue-specific) ACRs, possibly due to their correlation with previously reported higher genetic diversity of genic and proximal regions in maize". You questioned about the genetic diversity of non-genic regions which should be higher than genic regions, and there were papers reporting higher genetic diversity of non-genic regions in maize (https://www.nature.com/articles/nplants201684), which was contrary to the HapMap3 data. Therefore, different SNP datasets characterized differently of genetic diversity in genic and non-genic regions. Anyway, the conclusion regarding the core regions of ACRs having signature of purifying selection was well-supported, and we didn't discuss more about the higher SNP density in flanking regions.

Response_Table_R1. The SNP density in genic, proximal and distal region.

	Genic	Proximal	Distal
Total length	161,698,345	142,680,844	1,817,941,258
SNP	7,972,977	9,913,714	58,091,683
SNP density	0.049	0.069	0.032
SNP after filtering	3,872,622	5,973,917	35,032,433
SNP density after filtering	0.024	0.042	0.019



Response_Figure_R1. The distribution of genomic features of constitutive, tissue-specific and variable ACR.

3. Please cite https://www.biorxiv.org/content/10.1101/2025.01.22.633974v1 for the teosinte genomes from maizegdb instead of ref. 63.

Response: Done. Please see ref. 62.

4. L307. Is the lack of distal interspecfici ACRs just due to the fact you can't align these regions? That is, are there fewer distal ACRs given the length of alignable distal genome? I might expect the reverse – the only distal stuff you can map is conserved, and those regions are enriched for ACR. More succinctly, what is the null expectation?

Response: Thanks. To address this question, we selected 3 species (*Sorghum bicolor*, *Coix aquatica* and *Hordeum vulgare*) for additional interspecific analysis. We found that the conserved/constrained distal regions could account for 10% to 18% of the total conserved sequences when compared with these 3 species (see Response_Table_R2). By contrast, the length of constrained distal ACRs only comprised 2% to 5% of the length of all constrained ACRs. These data suggested that the interspecifically constrained ACRs were enriched in genic and proximal regions.

Response_Table_R2**.** The Length and proportion of conserved distal region and distal ACR.

	Coix aquatica	*Sorghum bicolor*	*Hordeum vulgare*
Divergence time (mya)	11.19	11.19	49.1
Conserved distal region (bp)	17,887,415	10,756,243	11,073,556
Total conserved region (bp)	108,873,576	100,978,885	60,814,777
Proportion of conserved distal region	16.43%	10.65%	18.21%
Constrained distal ACRs (bp)	220,371	229,968	10,840
Total constrained ACRs (bp)	4,254,236	4,445,392	465,754

Proportion of constrained distal ACRs	5.18%	5.17%	2.32%

Reviewer #2 (Remarks to the Author):

I appreciate the response my comments. Overall, this will make a nice contribution to the field. I will state, that I agree with the two reviewers that the claims about domestication and speciation are far too strong. Identifying loci involved in domestication and speciation is challenging and isn't simply done by looking at differences between teosinte and sorghum. In my opinion, this should be removed from the title and the rest of the paper it should be more carefully worded.

Response: We sincerely appreciate your support and kind suggestions. During first round of review, the revierwer1 concerned about the representativeness of a single teosinte genome in identifying regulatory loci involved in domestication, so we further incorporated 4 additional teosinte genomes (5 teosinte genomes in our revised manuscript) to strengthen the quality of these analysis. Regarding to speciation analysis, we compared 34 Poaceae species instead of a single sorghum genome to identify loci involved in speciation. Also, it should be noted that these loci were just "involved" in maize speciation and domestication, implying "association" instead of "causality". So we just slightly modified the title to "Constraint of accessible chromatins maps regulatory loci involved in maize speciation and domestication" to meet the editorial requirement of less than 15 words.

Reviewer #3 (Remarks to the Author):

The authors did a very good job to modify the manuscript, which have addressed all my concerns. I would love to read such research paper with significant achievements published in Nature Communications.

Response: We sincerely appreciate your support and valuable suggestions.

Round 1 Reviewer 1 attachment:
Liu et al. present an analysis of accessible chromatin and comparative genomics between maize, teosinte, and related taxa. The goal of this work is laudable, and the authors are asking a lot of great questions. The ATAC data seem good as well. In the end, however, I am struck by two large concerns with the paper. First, the authors ignore a large swath of relevant literature of comparative genomics in plants and other work looking at regulatory regions. They need to to much more to put their current work in context – what is new, how this builds off of or differs from previous results, etc. The second main concern I have is with methods. The authors are not necessarily doing anything wrong – alignment approaches etc. all are widely used methods. But the grass genomes are complex, and having played with some of these methods myself I know that they can be biased and problematic. I think the authors need to do more to convince readers that their results are real and not an artifact of whole-genome or short-read alignment limitations.

- **Major**
    - How much the conclusions about TEs depend on mapping? For example, L148 only 25% of ACRs overlap TEs, but most short reads won't map to TEs (see https://www.biorxiv.org/content/10.1101/2024.07.10.602892v1)
    - the authors cite several mammal papers but throughout seem to suggest there has been little work on conservation in noncoding regions in plants. There are several such papers, including work by e.g. Baoxing Song who specifically looks at constrained noncoding sequences in maize https://genome.cshlp.org/content/31/7/1245, or even databases of noncoding regulatory sequence across more species: https://academic.oup.com/nar/article/52/D1/D1569/7332064. The authors need to do a lot more to explain the uniqueness of their work and put it in the context of a considerable body of previous work.
    - Comparative analysis of ACRs in other genomes requires quality alignments. My experience with Cactus in grass genomes is that much of the genome doesn't align. The number of constrained ACRs in the phylogeny (Fig S9) doesn't seem to correlate with phylogenetic distance outside of Zea, as one might expect (e.g. 4,000 in Pennisetum vs only 1,000 in Hyperrhenia but 7,000 in the equally distant Sorghum). I think it important for the authors to characterize the quality of their alignments and discuss some how they deal with this. For example, it looks to me that the authors are treating "no alignment" as "ACR not present"? It might be better to only assay ACR presence/absence if flanking sequences do align. The alignment strategy may also explain the large number of maize-only ACR, especially those that are in TEs (and thus don't align well with other genomes).
    - The sampling of genomes for domestication comparison I think is an issue for making claims about domestication, and at least needs to be discussed. Many loci changed in frequency dramatically due to the bottleneck associated with domestication, and with only one genome from each teosinte, it's hard to rule out that a locus that is fixed in maize isn't at, for example, 80% frequency in parviglumis (and was just missed with 20% probability in the one genome

sampled). Such a locus would not widely be considered a key domestication gene, however!
    - The grammar is frequently awkward, occasionally complicating or preventing clear interpretation. This needs careful revision.
- **Minor**
    - L407 -- I think this is probably right (most TFBS are shared between m1 and m2) but perhaps worth the caveat that the authors have only looked at a small fraction. From my reading of the methods the authors can only identify ~300 TF families of the thousands that are in maize.
    - {{[[TODO]]}} Perhaps a lot to ask to integrate, but RAMPAGE data from the parviglumis TIL11 was made available by maize ENCODE http://www.maizecode.org/data/til11.html Even if the authors don't use the data, citing and discussing the work of the project is likely relevant https://www.biorxiv.org/content/10.1101/2024.02.22.581585v1
    - I'm curious why the authors are using v4 of B73 instead of the updated v5.
    - none of the MOA-seq work by Hank Bass or Thomas Hartwig are cited, but this literature seems relevant
    - L487 why does a bottleneck result in a loss-of-function? the logic here is unclear to me
    - I somehow missed where the assemblies of the other species come from and what quality they are.
    - Comparison of trait variation with ACRs should probably reference Rogers-Melnick 2016 https://pubmed.ncbi.nlm.nih.gov/27185945/ and Englehorn 2024 https://www.biorxiv.org/content/10.1101/2023.08.08.551183v2
    - L591 says 38 species and 4 outgroups
    - Why highlight that figure 2b is a triangle like in mammals (L212)? This seems to suggest something about comparable results, but is there any other shape this figure could possibly take?
    -
    - L147 the maize genome is below average size for angiosperms, it is not "bloated"
    - gACR, pACR, dACR acronyms should be defined on L137
    - I would encourage authors to use "Ty3" instead of "Gypsy" see here: https://osf.io/preprints/osf/fma57
    - L155 an alternative explanation to purifying selection is TE insertion preferentially into other TEs or closed chromatin. With a single genome this can't be disentangled.
    - L177 is Michaelis-Menton standardly used this way? It seems like doing a rarefaction subsampling analysis might be more appropriate? But OK if this is standard for ATAC and similar analyses.
    - L189 Genic regions of maize have lower diversity than non-genie. See e.g. https://www.nature.com/articles/nplants201684

Round 2 Reviewer 1 attachment:
Major:

The authors have done a nice job addressing most of my concerns about TEs (looking at fiber-seq) and alignment (using flanking sequences, filtering for genome quality). I think still worth noting in the discussion the potential caveat that conserved ACRs are more strongly correlated with the size of the alignment in bp (r=0.89) than the divergence time (r=-0.64). Of course the amount that can be aligned is negatively correlated with time (r=-0.82) so it's hard to disentangle these. But I could imagine an orthologous gene with conserved regulation between two species – same upstreadm TF binding site with same TFs binding – but where there is no "conserved ACR" because the intergenic DNA between the gene and the TF binding site has changed so much that alignment is not possible. This seems plausible given how few of the flanking regions of ACRs Can such a scenario be ruled out? If not, worth adding a sentence or so in the discussion that this is difficult to disentangle.

The writing is improved, but still awkward and at sometimes hard to follow. I find that LLMs actually do a wonderful job of clearing this up. Literally pasting paragraphs from the paper into CoPilot, for example, cleared up minor grammatical issues that remained and clarified wording. I recommend this to folks all the time, and see no difference between this and e.g. Microsoft Word's grammar check or editing suggestions.

For example, CoPilot returns:

> The intricate regulation of gene expression is crucial for executing complex biological processes in higher organisms. Maize, a classical genetic model in gene regulation, has been extensively studied since Dr. Barbara McClintock's groundbreaking discovery of transposable elements (TEs), also known as "jumping genes," which cause the mosaic pigmentation seen on maize kernels. Recent assembly and analysis of maize genomes have revealed a highly sophisticated and hierarchical gene regulation system, characterized by the prevalence of distal cis-regulatory elements (CREs), particularly enhancers. This complexity is likely due to the large maize genome (~2.3 Gb), with over 85% of its DNA composed of TEs. Genetic mapping in maize has shown an enrichment of quantitative trait nucleotides (QTNs) within non-genic regulatory regions, including well-known loci such as tb1, vgt1, KRN4, ZmCCT9, ZmCCT10, and UPA1. On a higher-dimensional scale, the widespread distal CREs in maize can coordinate with transcription factors (TFs) and other trans-acting elements while interacting with promoters or other gene-proximal regions to regulate gene expression through chromatin interactions. These findings highlight the crucial role of regulatory

> sequences in controlling maize development and trait variations.

Or this one:

> The proliferation of the maize genome has led to a significant number of distal accessible chromatin regions (dACRs) being located in "gene deserts" (regions 100 kb away from the nearest gene). Interestingly, these ultra-distal ACRs showed high conservation across the tissues we profiled and were independently verified by maize single-cell ATAC-seq data and ChIP-seq data of various active histone modifications. They were also found to be involved in chromatin interactions, with some linked to genes in 3D genomic data, suggesting that maize ACRs play a cis-regulatory role both locally and distally.

Minor:

L61 why not call these QTL, since for most of the examples the causal locus (if known) is not a single nucletotide

L192 I'm not sure I follow the authors' interpretation of selection here. Lower SNP density in constitutive does perhaps imply conservation. But those same regions have higher SNP density around, and I don't see an explanation for this.

Please cite https://www.biorxiv.org/content/10.1101/2025.01.22.633974v1 for the teosinte genomes from maizegdb instead of ref. 63.

L307. Is the lack of distal interspecfici ACRs just due to the fact you can't align these regions? That is, are there fewer distal ACRs given the length of alignable distal genome? I might expect the reverse – the only distal stuff you can map is conserved, and those regions are enriched for ACR. More succinctly, what is the null expectation?