Conserved non-coding sequences provide insights into regulatory sequence and loss of gene expression in maize

Baoxing Song[1*], Edward S. Buckler[1,2,3], Hai Wang[1, 4], Yaoyao Wu[1,5], Evan Rees[2], Elizabeth A. Kellogg[6], Daniel J Gates[7], Merritt Khaipho-Burch[2], Peter J. Bradbury[3], Jeffrey Ross-Ibarra[7,8], Matthew B. Hufford[9], M. Cinta Romay[1*]

1 Institute for Genomic Diversity, Cornell University, Ithaca, NY, 14853, USA

2 Section of Plant Breeding and Genetics, Cornell University, Ithaca, NY, 14853, USA

3 Agricultural Research Service, United States Department of Agriculture, Ithaca, NY, 14853, USA

4 National Maize Improvement Center, Key Laboratory of Crop Heterosis and Utilization, Joint Laboratory for International Cooperation in Crop Molecular Breeding, China Agricultural University, Beijing, 100193, China

5 Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, 518124, China

6 Donald Danforth Plant Science Center, St. Louis, MO, 63132, USA

7 Department of Evolution and Ecology, University of California Davis, Davis, CA 95616, USA

8 Center for Population Biology and Genome Center, University of California Davis, Davis, CA 95616, USA

9 Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, 50011, USA

*Corresponding authors

Abstract

Thousands of species will be sequenced in the next few years; however, understanding how their genomes work without an unlimited budget requires both molecular and novel evolutionary approaches. We developed a sensitive sequence alignment pipeline to identify conserved non-coding sequences (CNSs) in the Andropogoneae tribe (multiple crop species descended from a common ancestor ~18 million years ago). The Andropogoneae share similar physiology while being tremendously genomically diverse, harboring a broad range of ploidy levels, structural variation, and transposons. These contribute to the potential of Andropogoneae as a powerful system for studying CNSs and are factors we leverage to understand the function of maize CNSs. We found that 86% of CNSs were comprised of annotated features, including introns, UTRs, putative *cis*-regulatory elements, chromatin loop anchors, non-coding RNA genes, and several transposable element superfamilies. CNSs were enriched in active regions of DNA replication in the early S phase of the mitotic cell cycle and showed different DNA methylation ratios compared to the genome-wide background. More than half of putative *cis*-regulatory sequences (identified via other methods) overlapped with CNSs detected in this study. Variants in CNSs were associated with gene expression levels, and CNS absence contributed to loss of gene expression. Furthermore, the evolution of CNSs was associated with the functional diversification of duplicated genes in the context of maize subgenomes. Our results provide a quantitative understanding of the molecular processes governing the evolution of CNSs in maize.

Introduction

The genomes of a million eukaryote species will likely be sequenced within the next decade (Lewin et al. 2018), but understanding how these genomes work without ENCODE-scale projects and data (The ENCODE Project Consortium 2012) for each species will require that we also use evolutionary approaches to identify key functional regions. In general, non-coding sequences occupy a larger portion of the genome than coding regions. Most genome-wide association hits have been reported to be located in the non-coding regions in, for example, maize and humans, and are enriched in putative gene expression regulatory sequences (Wallace et al. 2014; Nishizaki and Boyle 2017; F. Zhang and Lupski 2015; Giral, Landmesser, and Kratzer 2018). Comparison of non-coding sequences across species can identify regions under purifying selection to reveal functional constraint (Finucane et al. 2015; Xiang et al. 2019; Haudry et al. 2013; Polychronopoulos et al. 2017; Guo and Moose 2003; Vandepoele, Casneuf, and Van de Peer 2006). However, detection of conserved non-coding sequences (CNSs) in plants is an ongoing challenge (Van de Velde et al. 2016), receiving extensive recent attention in a broad range of species (J. Xie et al. 2018; Inada et al. 2003; Algama et al. 2017; Polychronopoulos et al. 2017; Freeling and Subramaniam 2009). A genome-wide comparison of features of putative functional elements (Lu et al. 2019; Ricci et al. 2019; Rodgers-Melnick et al. 2016; Oka et al. 2017; W. Zhang et al. 2012; Z. Li et al. 2019; M. Wang et al. 2017; Tu et al. 2020) with CNSs could provide new insight into understudied non-coding fractions of the genome.

Genomes of the grass tribe Andropogoneae provide a valuable and powerful system for the study of conserved sequences. Species of the Andropogoneae tribe have diverged in a relatively short time frame, sharing a common ancestor ~16-20 million years ago (Vicentini et al. 2008).

Andropogoneae species include maize, sorghum, sugarcane, and silvergrass, some of the most productive grain, sugar, and biofuel crops worldwide (Brosse et al. 2012; Manners 2011). Andropogoneae species share the NADP-ME C_4 photosynthesis system (Sage and Zhu 2011; Black, Chen, and Brown 1969) and similar development patterns, while their genomes are highly diverse with frequent polyploidization (Estep et al. 2014) and extremely active transposable elements (TEs) (Ramachandran et al. 2020). Nevertheless, despite rapid sequence turnover elsewhere in Andropogoneae genomes, functional sequences are expected to be under purifying selection, making the tribe an ideal system in which to identify and understand the role of CNS.

Results

An atlas of CNSs in the Andropogoneae tribe

Inspired by recent studies of regulatory architecture (Ricci et al. 2019; Tu et al. 2020; Parvathaneni et al. 2020; Oka et al. 2017; Lu et al. 2019), we used coding genes as anchors and developed a sensitive sequence alignment pipeline to identify CNSs in Andropogoneae genomes which have undergone genome-wide duplications, numerous rearrangements, and gene loss (J. C. Schnable, Springer, and Freeling 2011) (Fig. 1). Andropogoneae genomes (see below) were aligned to the maize B73 v4 assembly (Jiao et al. 2017), which was used as a reference. First, we lifted over maize protein coding genes to each query genome (Fig. 1, step 1) by mapping coding sequences (CDSs) using minimap2 (H. Li 2018). Gene copy number often varied between the maize genome and the query genome, so all minimap2 mapping hits with a similarity larger than 60% were used as anchors, where similarity is the number of identical base-pairs to the CDS length in maize. All CDSs and high-frequency *k*-mers were then removed from the genomes (Fig. 1, step 2). Next, introns, sequences of 100 kbp upstream of the translation start codon and

sequences of 100 kbp downstream of the translation stop codon were extracted to perform pairwise alignment. The selection of a 100 kbp range was based on previous observations that almost all open chromatin regions and transcription factor binding sites (TFBSs) are located within 100 kbp of the nearest gene (Rodgers-Melnick et al. 2016; Tu et al. 2020; Ricci et al. 2019). Pairwise alignment was conducted following the widely used seed-and-extend process (Altschul et al. 1990; H. Li and Homer 2010). Briefly, the query sequences were cut into fragments with overlapping sliding windows using a window size of 38 bp and a step size of 8 bp. The Smith-Waterman algorithm (T. F. Smith and Waterman 1981) was employed to align the fragment in each window against the reference sequence, and any alignments with an alignment score ≥ 40 were used as seeds (Fig. 1, step 3). Every seed was then extended, and extension was terminated using the X-drop approach (Z. Zhang, Berman, and Miller 1998) (Fig. 1, step 4). Alignments with a dynamic programming score ≥ 54 were defined as CNSs. This score threshold corresponds to a *p*-value < 0.1 assuming that a pair of sequences with a length of 100 kbp were aligned (Karlin and Altschul 1990) (Fig. 1, step 5). Finally, the removed CDS and *k*-mer sequences were put back into aligned CNSs (Fig. 1, step 6) and the SAM-format alignments were generated (H. Li et al. 2009) (Fig. 1, step 7).

Three publicly available genomes, sorghum (*Sorghum bicolor*) (McCormick et al. 2018), maiden silvergrass (*Miscanthus sinensis*) (J. Zhang et al. 2018), and wild sugarcane (*Saccharum spontaneum*) (Mitros et al. 2020) were used as queries (see Supplemental Table S1 for more information regarding all the genomes used in this study). In addition, the genomes of two heterozygous Andropogoneae species, *Hyparrhenia diplandra* and *Chrysopogon serrulatus*, were assembled to supplement the above mentioned three genomes (Fig. 2A). Benchmarking

Universal Single-Copy Orthologs (BUSCO) completeness scores (Simão et al. 2015) of 0.94 demonstrated that these two newly assembled genomes had high completeness and low sequencing error rates. The median distance between genes and contig edges were 35 kbp and 77 kbp in the genomes of *H. diplandra* and *C. serrulatus*, respectively, indicating the contiguity of these two assemblies was suitable for the detection of CNSs within 10 kbp of >85% of coding genes (Supplemental Fig S1).

Total CNS length identified using each query genome was positively correlated with the query genome size (Supplemental Fig S2). The sorghum genome is the only monoploid assembly in our dataset that has not experienced a genome-wide duplication event after its divergence from the common ancestor of Andropogoneae (J. Schnable and Lyons 2015). Aligning the genome sequence of sorghum against that of maize generated the smallest CNS space (67.07 Mbp, by counting matched base-pairs in the maize genome). The largest CNS space was observed by aligning genome sequences between wild sugarcane and maize (86.97 Mbp). The sizes of CNSs ranged from 27 bp to 15 kbp (Supplemental Fig S3). The total length of CNSs present in maize and at least one other species was 106.52 Mbp, accounting for ~5% of the maize genome. Hereafter, those CNSs were designated as pan-And-CNSs (pan-Andropogoneae CNSs). A total of 42.27 Mbp CNSs were present in all species and were termed as core-And-CNSs (core-Andropogoneae CNSs). However, more species are needed to better describe pan- and core-And-CNSs (Fig. 2B).

Pan-And-CNSs were enriched with putative *cis*-regulatory elements (Supplemental Fig S4) and overlap with 52-78% of sequences with putative *cis*-regulatory features (TFBSs, open chromatin

regions, acetylation of histone 3 lysine 9 (H3K9ac) ChIP-seq peaks, micrococcal nuclease hypersensitive regions (MNase HS) or DNase I hypersensitive sites (DHSs)) identified in different plant tissues (Supplemental Table S2). We also confirmed that a few known regions were correctly classified as CNS: the third intron of *Knotted1* (Greene, Walko, and Hake 1994; Lai et al. 2017) (Supplemental Fig S5), *Vgt1* (Salvi et al. 2007) (Supplemental Fig S6), and *tb1* (Clark et al. 2006; Studer et al. 2011) (Supplemental Fig S7) suggesting the identification of functional non-coding sequences using our approach.

A large proportion of CNSs overlap with putative regulatory elements

A large proportion (86.8%) of genes have pan-And-CNS detected within 2 kbp upstream. When compared with genes without pan-And-CNS detected within 2 kbp upstream, genes with CNS have higher expression levels and less tissue expression specificity (Kadota et al. 2006; Yanai et al. 2005) across 23 maize B73 tissues (Walley et al. 2016) (Supplemental Fig S8). 51% of the pan-And-CNSs overlap with the untranslated region (UTR) or intron of coding genes (Supplemental Fig S9) (herein genic CNS, otherwise intergenic CNS); this proportion is comparable to that of *Arabidopsis thaliana* (Haudry et al. 2013). Since introns and UTRs have a wide range of conserved functional roles (e.g., promote gene expression, guide splicing, produce non-coding RNA) (Chorev and Carmel 2012; Akua, Berezin, and Shaul 2010; Rigau et al. 2019; Greene, Walko, and Hake 1994; Ritchie and Flicek 2014), we did not classify genic CNSs further by potential function. The overlap between intergenic pan-And-CNS and low DNA methylation loci (Xu et al. 2020) was higher than the genome-wide random expectation (Supplemental Fig S10). 56% of intergenic pan-And-CNS records overlap with open chromatin regions, TFBSs, or H3K9ac ChIP-seq peaks (Fig. 3A), a 6.5- to 19.5-fold enrichment relative to

random expectation (Fig. 3B). Compared to a previously released list of conserved elements in the maize intergenic region (Tian et al. 2020), we generated a larger dataset and our intergenic CNSs have greater overlap with open chromatin regions, TFBSs, and H3K9ac ChIP-seq peaks (Supplemental Fig S11).

To further uncover the function of intergenic pan-And-CNSs, we analyzed colocalization of intergenic pan-And-CNSs with chromatin loop anchors, TEs, and non-coding RNA (ncRNA) genes. We started with chromatin loop anchors because they were inferred to be conserved and their dynamics are associated with transcription activity (P. Dong et al. 2020; Q. Dong et al. 2018; Liu et al. 2017; Szabo, Bantignies, and Cavalli 2019; M. Wang et al. 2018; Harmston et al. 2017; Polychronopoulos et al. 2017; Delaneau et al. 2019; P. Dong et al. 2017). Chromatin loop anchors identified by Hi-C and HiChIP (Ricci et al. 2019; Peng et al. 2019) overlapped with 47% of intergenic pan-And-CNS records, with an enrichment of 6.6-fold compared to the genome-wide background (Fig. 3B). In addition to open chromatin regions, TFBSs, and H3K9ac ChIP-seq peaks, chromatin loop anchors overlapped with an additional 12% intergenic pan-And-CNS records (Supplemental Fig S12). Regulatory elements derived from TEs have been described by several independent studies (Dupeyron et al. 2019; A. M. Smith et al. 2008; X. Xie, Kamal, and Lander 2006), and cases of TEs acting as regulatory sequence have been reported (Makarevitch et al. 2015; Noshay et al. 2020; Studer et al. 2011; Hainan Zhao et al. 2018; Dupeyron et al. 2019; A. M. Smith et al. 2008; X. Xie, Kamal, and Lander 2006). Enrichment of CNSs overlapping TEs was not observed (Fig. 3B), but when looking at particular TE superfamilies (Stitzer et al. 2019), 1.08- to 1.78-fold enrichments of intergenic pan-And-CNSs were observed in 4 of the 13 TE superfamilies (RIL, RST, DHH, and DTM) (Supplemental Fig S13). However, these 4 TE

superfamilies only accounted for another 2.09% of intergenic pan-And-CNSs. Finally, enrichment of CNSs in ncRNA genes (Han et al. 2019) was also observed (Fig. 3B), which accounted for only 1.04% of the total intergenic pan-And-CNSs.

CNSs have diverse functions

Due to the tissue-specific activity of some *cis*-regulators, a full dataset of putative *cis*-regulatory sequences has not been generated until the present study. It is essential to know if CNSs that do not overlap with features of *cis*-regulatory elements might have alternative functions. We therefore classified intergenic CNSs into the following groups: (1) CNSs overlapping with open chromatin, TFBS, or H3K9ac ChIP-seq peaks (*cis* CNS); (2) CNSs overlapping with chromatin loop anchors not part of the group *cis* (non-*cis* loop CNS); and (3) remaining CNS not included in (1) and (2) (rest CNS). We then investigated the DNA methylation ratio, guanine-cytosine (GC) content and DNA replication activity in the early S phase of the mitotic cell cycle in each CNS group. Different DNA methylation ratios, GC content, and DNA replication activity in the early S phase were obtained for these CNS groups, indicating these CNSs may have a diversity of functions.

In terms of DNA methylation, the *cis* pan-And-CNSs showed a low DNA methylation ratio, which supports the previous observation that putative *cis*-regulatory sequences correspond to a low DNA methylation ratio in plants (Ricci et al. 2019; Rodgers-Melnick et al. 2016; Oka et al. 2017; W. Zhang et al. 2012; Suzuki and Bird 2008; Zilberman et al. 2007; X. Zhang et al. 2006). The non-*cis* loop pan-And-CNSs exhibited a medium DNA methylation ratio (Fig. 4A) and could be divided into two distinct subgroups according to DNA methylation ratios (Supplemental

Fig S14). This might be due to different functions of CNSs between these two subgroups; alternatively, some may actually be *cis* CNSs with tissue specificity but were incorrectly classified into the non-*cis* loop pan-And-CNS group. DNA methylation is associated with GC content (Mugal et al. 2015) and GC content is associated with chromatin accessibility (Hammelman et al. 2020; Parker, Margulies, and Tullius 2008; Schwartz et al. 2019). Compared to the genome-wide background, the *cis* pan-And-CNSs showed a higher GC content, while non-*cis* loop pan-And-CNSs exhibited a lower GC content. This is again suggestive of diverse functions across *cis* pan-And-CNSs and non-*cis* loop pan-And-CNSs. The GC content of the rest of the pan-And-CNSs was similar to that of the genome-wide background (Fig. 4B). Active regions of DNA replication in the early S phase (Wear et al. 2017) were associated with both *cis* pan-And-CNSs and non-*cis* loop pan-And-CNSs. The activity of the rest pan-And-CNSs was higher than that of the genome-wide background but lower than that of *cis* pan-And-CNSs and non-*cis* loop pan-And-CNSs (Fig. 4C, Supplemental Fig S15).

Overall, 14% of pan-And-CNS records did not fall in any putative features (i.e., open chromatin regions, TFBSs, introns, UTRs, non-coding RNA genes, chromatin loop anchors, H3K9ac ChIP-seq peaks, and TEs) (Fig. 4D, Supplemental Fig S16). These CNSs were shorter and had lower alignment scores (Supplemental Fig S17) than the 86% that could be assigned to a feature and may be enriched for false positives. These unannotated pan-And-CNSs might also be tissue-specific regulators. A more comprehensive investigation of functional non-coding features in different tissues may provide a better understanding of CNS functions.

Variants in CNSs impact gene expression

We further investigated the CNS function by testing if genotypic variants within the identified CNS affected maize gene expression. Using the expression quantitative trait loci (eQTLs) identified by Kremling et al. (Kremling et al. 2018), the intergenic pan-And-CNSs were enriched 4.02-fold among the "lead" eQTLs compared to the genome-wide background. We selected the lead eQTL as the one with the strongest association (lowest *p*-value) with the expression of its target gene. The pan-And-CNS regions harbored a larger proportion of maize HapMap3 variants (Bukowski et al. 2018) with low minor allele frequency (MAF) when compared to intergenic regions (Fig. 5A). This result suggests that variants in the CNS regions are under stronger purifying selection compared to those in the intergenic regions, likely because variants in CNSs could negatively impact functional elements.

We identified CNS presence/absence variants (PAVs) using the maize HapMap3 second-generation sequencing reads (Bukowski et al. 2018) and CNSs identified using the sorghum genome as the query. Most CNS PAVs were rare (MAF < 0.1), especially genic CNSs and *cis* CNSs (Fig. 5B). Previous studies suggested that non-coding rare variants contribute to the dysregulation of nearby downstream genes and are negatively associated with organism fitness (Flint-Garcia et al. 2005; Kremling et al. 2018). By analyzing the CNS absence within the 2 kbp upstream region of a gene and its expression, we observed loss of CNSs was associated with loss of gene expression (Fig. 5C). However, further upstream, the association between loss of CNSs and loss of gene expression was much weaker (Supplemental Fig S18); beyond 2 kbp, the regulatory activity of CNSs on downstream genes likely diminishes. To further test the impact of CNS PAVs on gene expression, we used gene expression profiles of seven tissues from a modern inbred population (Kremling et al. 2018) along with the CNS PAVs with MAF ≥ 0.1 to perform

an expression genome-wide association study (Supplemental Fig S19). The result showed that more than half of significant CNS PAVs were located within 2.5 Mbp of associated genes (Fig. 5D, Supplemental Fig S20).

Finally, we investigated the evolution of CNSs in maize during domestication and subsequent local adaptation. We identified regions likely selected during domestication using genome-wide selective sweeps using RaiSD (Alachiotis and Pavlidis 2018) in a panel of 31 maize landraces (L. Wang et al. 2017), and regions important for local adaptation in maize using PCAdapt (Luu, Bazin, and Blum 2017). The overlap between these regions and pan-And-CNS was substantially less than would be expected by chance, suggesting recent selection in maize has mainly favored variants that do not modify these constrained functional regions (Supplemental Fig S21).

CNS variation is associated with functional diversity between maize subgenomes

To further study the effect of CNS variation on gene expression, we took advantage of the genome-wide duplication event that occurred in maize after divergence from sorghum (Swigonová et al. 2004), to investigate CNS variation between the two maize subgenomes. Syntenic homologous genes between genomes of maize and sorghum were identified using the quota-alignment implementation (Tang et al. 2011) with parameters that keep every two maize genes corresponding to one sorghum gene (--quota, see methods). First, for each orthogroup (including two maize genes and one sorghum gene), the total CNS length within the 2 kbp upstream region of a gene was recorded if this CNS site was present within the 2 kbp region of the sorghum homologous gene and at least one maize homologous gene copy (Supplemental Fig S22). Second, the proportion of shared CNS sites was calculated as the number of CNS sites

present in both maize gene copies to that of the total CNS length. Third, the expression similarity between two maize gene copies was calculated using Pearson's correlation of their expression levels across 23 tissues of maize B73 (Walley et al. 2016). Then, a correlation analysis of the proportion of shared CNS sites and expression similarity was conducted. The proportion of CNS sites shared between the two paralogous genes was positively correlated with the expression similarity between them (r^2=0.10, *p*-value< 2.2×10^{-16}, Pearson's correlation, Fig. 6A, Supplemental Fig S23). Moreover, maize paralogs with negatively correlated expression patterns shared a significantly smaller proportion of CNS sites than positively correlated paralogs (Fig. 6B) (*p*-value<2×10^{-16}, Wilcoxon rank-sum test). Here we defined the gene copy with a longer CNS (within 2 kbp upstream region) as the major copy and the gene with a shorter CNS as the minor copy. In the context of maize subgenomes, for each pair of maize syntenic paralogous genes with negative expression correlation, in addition to spatiotemporal expression patterns, we also observed a higher nonsynonymous to synonymous mutations ratio for the minor CNS approximation gene (Supplemental Fig S24), which may indicate neofunctionalization or pseudogenization. The size difference between the major and minor CNSs was smaller in positively correlated paralog pairs than in negatively correlated paralog pairs (Fig. 6C) (*p*-value=5×10^{-10}, Wilcoxon rank-sum test).

In addition, for those negatively correlated paralog pairs, we examined the correlation between the expression of maize genes and their sorghum homologs in the shoot tissues. The normalized shoot RNA-seq data of B73 (maize) (Kremling et al. 2018) and sorghum (NCBI BioProject PRJNA503076) were retrieved from (Washburn et al. 2019). We observed a higher correlation between the expression levels of maize major copies and the homologous sorghum genes than

between maize minor copies and the homologous sorghum genes in the shoot (Fig. 6D). Overall, these results suggest that CNS variation is associated with expression and functional diversity between duplicated genes.

Discussion

Despite the challenges of working on genomes with vast numbers of transposons and frequent duplications, a novel sensitive alignment approach shows that non-coding regions under purifying selection can be identified by comparing genomes of related species. In our CNS identification pipeline, the aim of using coding genes as anchors is to narrow down the sequence alignment scope and reduce false positives. There are some scenarios where our pipeline will fail to identify the target gene. The nearest gene of a regulatory element is not necessarily its target, and the distance between a regulatory element and its target genes might be longer than 100 kbp. In addition, the proximity of a functional non-coding sequence and its target gene may be impacted by large insertions or chromosomal rearrangements. Previous studies suggested that the total number of predicted TFBSs is much smaller than that of transcription factor (TF) recognition sites. TFs often work cooperatively and the sequence contexts of TF recognition sites or combinatorial recognition of *cis*-elements are key for TF binding specificity (Tu et al. 2020; Gerstein et al. 2012; O'Malley et al. 2016; Dror et al. 2015; Levo et al. 2015; Avsec et al. 2021). CNS records identified in this study are much longer than a single TF recognition site, suggesting the possibility of recognition of ordered combinations of non-overlapping TFBS (Viturawong et al. 2013; Shen et al. 2020). CNSs identified using our approach were highly enriched in TFBSs, open chromatin regions, H3K9ac ChIP-seq peaks, chromatin loop anchors, and eQTLs.

Using the genome sequences of six Andropogoneae species, we were able to identify a set of core-And-CNSs and pan-And-CNSs. Core-And-CNSs and pan-And-CNSs showed similar enrichment in potential regulatory sequences, suggesting that not all functional non-coding elements are conserved across all the species. The presence/absence of those elements may be related to species-specific traits (Khalturin et al. 2008; Won et al. 2019). As expected based on the role of core-And-CNS in more essential functions (Cvijović, Good, and Desai 2018) and the likelihood of mutations in these regions carrying a higher deleterious burden (Kistler et al. 2018), we observed a larger number of variants with low MAF in core-And-CNS regions than in pan-And-CNS regions in a maize population (Fig. 5A), indicating stronger purifying selection.

The overlap between CNSs and features of putative *cis*-regulatory sequences has been reported previously (Lai et al. 2017; Haudry et al. 2013; Warnefors et al. 2016; Viturawong et al. 2013; Ricci et al. 2019; Oka et al. 2017; Lu et al. 2019), although not all *cis*-regulatory sequences are evolutionarily conserved (Ross, Fong, and Cavener 1994; Wittkopp, Vaccaro, and Carroll 2002; McLean et al. 2011). We observed enrichment of lead eQTLs in CNS regions and association between CNS absence and gene expression (Fig. 5C-D). In terms of the regulation of CNSs on target genes, we were able to show that upstream 2 kbp CNS absence was strongly correlated with loss of expression, both in the context of natural variation segregating within a maize population and between duplicated genes in the maize subgenomes. Our results indicate that a longer conserved region proximal to a maize gene correlates with higher expression and similarity of its orthologous gene in sorghum (Fig. 6D), suggesting pseudogenization or neofunctionalization of the gene copy with a shorter CNS. In summary, our study shows that a

meta-analysis using CNSs, and gene expression levels combined with open chromatin regions, TFBSs, chromatin loop anchors, and low DNA methylation loci can provide a more comprehensive view of the molecular mechanisms underlying the regulation of gene expression.

Analysis of the identified CNSs extended our knowledge of CNS function. Around half of the identified intergenic CNSs overlapped with different groups of putative *cis*-regulatory features (e.g., open chromatin regions, TFBSs, and H3K9ac ChIP-seq peaks). Chromatin looping is important for gene regulation (Kadauke and Blobel 2009) and subtle changes in chromatin loop anchors are associated with differential gene regulation and expression (Diehl, Ouyang, and Boyle 2020; Greenwald et al. 2019). GC content and DNA methylation ratio might influence DNA stiffness/flexibility and has been reported to be correlated with DNA supercoiling and recombination (Jabbari et al. 2019; Rodgers-Melnick et al. 2015; Naughton et al. 2013). In this study, we also observed distinguishable GC content and DNA methylation ratios between *cis* CNS with non-*cis* loop CNSs, suggesting diverse functions between these two types of CNSs. ncRNAs can interact with DNAs, RNAs, and proteins, and have been implicated in the regulation of gene transcription and translation, as well as in response to stresses and stimuli (Statello et al. 2021; Han et al. 2019; Chen and Aravin 2015; Yoon, Abdelmohsen, and Gorospe 2013; Yao, Wang, and Chen 2019). The enrichment of CNSs in ncRNA genes indicates sequence conservation of transcribed but untranslated DNA sequences. The enrichment of CNS in active regions of DNA replication in the early S phase suggests that regions of DNA replication initialization might be evolutionarily conserved. CNS knockout lines should be generated to gain a comprehensive understanding of their function (Gasperini, Tome, and Shendure 2020). Taking advantage of the observation of GC content and DNA methylation

pattern across different groups of CNS, a model to classify CNSs into different functional groups might be useful to provide a guide for the verification of CNS functions via molecular approaches.

Methods:

Plant materials collection:

Hyparrhenia diplandra was collected in Kenya by Rémy Pasquet (Pasquet 1126) and *Chrysopogon serrulatus* was obtained from the USDA Germplasm Repository Information Network (PI 219580; seed originally from Pakistan). Both plants were grown in the greenhouse at the Donald Danforth Plant Science Center. Vouchers of flowering specimens were deposited at the herbarium of the Missouri Botanical Garden; full specimen data are available through www.tropicos.org.

DNA preparation and sequencing:

Total DNA was extracted from young leaf tissues. Long-read sequencing was conducted on a Nanopore MinION platform at the Institute of Biotechnology, Cornell University. DNA with a size of 20-80 kbp was selected following the Blue Pippin protocol, and the selected DNA were cleaned using AMPure XP beads. DNA repair and end-prep were performed with NEB enzyme kits. After adapter ligation, MinKNOW software was used for quality control of the MinION sequencing library. Sequencing was performed following the manufacturer's instructions.

A total of 1.0 μg of DNA per sample was used as input material to generate second-generation sequencing reads. Sequencing libraries were conducted using the NEBNext® DNA LibraryPrep

Kit following the manufacturer's recommendations, and barcodes were added to each sample. Genomic DNA was randomly fragmented to a size of 350 bp. Then DNA fragments were end polished, A-tailed, and ligated with the NEBNext adapter for Illumina sequencing, and further PCR enriched by P5 and indexed P7 oligos. PCR products were purified (AMPure XP system) and the resulting libraries were analyzed for size distribution by an Agilent 2100 Bioanalyzer and quantified using real-time PCR. The qualified libraries were fed into Illumina NovaSeq sequencers after pooling according to their effective concentration and expected data volume.

Genome assembly:

We used the NanoPlot (De Coster et al. 2018) and porechop package (Wick n.d.) to respectively check and filter the MinION raw reads. The MinION clean reads were then assembled using Flye v1.4.2 (Kolmogorov et al. 2019) with the genome size estimated via flow cytometry (Supplemental Fig S1F) as reference. The MinION clean reads were mapped to the assembly using minimap2 (H. Li 2018) with a setting of “-x map-ont”, and racon v1.3.1 (Vaser et al. 2017) was used to polish the assembly with default parameters. Assembly polishing using MinION reads was repeated three times. The MEM module of BWA v 0.7.17 (H. Li and Durbin 2009) was used to map Illumina reads to the MinION polished assembly with parameters “-k11 -r10”. The “markdup” command implemented in SAMtools v1.09 (H. Li et al. 2009) was used to remove duplicated Illumina reads. Pilon v1.23 (Walker et al. 2014) with parameter “--diploid --fix bases" was used for error correction (Supplemental Fig S1A). Assembly correction using Illumina reads was repeated three times.

Genome assembly evaluation:

To evaluate the contiguities of our new assemblies, we identified 5,592 Benchmarking Andropogoneae Single-Copy Orthologs (BASCO) genes shared by four Andropogoneae genomes, maize (*Zea mays*), sorghum (*Sorghum bicolor*), maiden silvergrass (*Miscanthus sinensis*), and wild sugarcane (*Saccharum spontaneum*), as well as an outgroup species, foxtail millet (*Setaria italica*). The URLs to access these genomes are listed in Supplemental Table S1. The following procedures were conducted to identify the BASCO genes:

1) Syntenic genes: CDSs of the B73 genome (v4.34) were first aligned against the sequences of sorghum, wild sugarcane, maiden silvergrass, and foxtail millet using BLASTN (Altschul et al. 1990) with parameters "-outfmt 6 -strand plus -task blastn -evalue 5 -word_size 7 -max_target_seqs 1000". Next, the quota-alignment pipeline (Tang et al. 2011) was used to detect syntenic genes using parameters "--tandem_Nmax=5 --cscore=0.2 --no_strip_names --filter_repeats" for blast_to_raw.py. The parameters of quota_align.py were set as "--merge --Dm=20 --min_size=3". In addition, for specific species the "--quota" parameter was set as "2:2" for maiden silvergrass, "1:2" for foxtail millet, "1:2" for sorghum, and "4:2" for wild sugarcane according to their different genome assembly ploidy or genome-wide duplication levels. Overall, 25,155 maize genes were found in at least one syntenic region.

2) Orthogroups: OrthoFinder (Emms and Kelly 2019) with parameter "-S blast -M msa" was used to find orthogroups. Only those orthogroups with 1-2 maize genes, 1 sorghum gene, 1-2 maiden silvergrass genes, 1-4 wild sugarcane genes, and 1 foxtail millet gene were kept. The CDS of each transcript in the selected orthogroups was mapped to the corresponding genome sequence using minimap2 with parameters "-ax splice -a -uf -C1 --cs". Any transcripts with a higher than expected number of hits were removed.

3) Intersect syntenic genes and orthogroups: Orthogroups with at least one syntenic maize gene were kept, otherwise dropped.

4) Double check gene copy numbers: Similar to step 2, CDSs of the genes that passed the previous filter were mapped to the five genomes using minimap2 and filtered by the number of hits within each genome. Orthogroups that passed this last check were defined as BASCO. CDSs of BASCO genes were mapped to the *H. diplandra* and *C. serrulatus* assemblies using minimap2 with parameters "-ax splice -a -uf -C 1 -k 12 -P -t 12 --cs". To evaluate the contiguity of the assemblies, we defined the minimum extent of flanking for mapped genes (Supplemental Fig S1C) as:

- n1 = start position of mapped CDS
- n2 = contig length – end position of mapped CDSs
- minimum extent of flanking regions = minimum (n1, n2)

BUSCO v3.1.0 with parameters "-m geno -sp maize -f -r -l" and the liliopsida_odb10 database were used to evaluate the completeness of the assemblies.

Parameter optimization for identification of CNS:

The pipeline started by finding anchor points using minimap2 (H. Li 2018) with parameters "-x splice -a -uf -C 1 -k 12 -P --cs". Keeping in mind the high diversity of the Andropogoneae non-coding regions, for the following alignment steps we used a match score "2", mis-match "-3", gap opening penalty "-4", and gap extension penalty "-2". For step 3 (Fig. 1), we kept alignments with a minimum Smith-Waterman score of 40 as seeds, thus the minimum seed length was 20 bp (minimum seed score/match score). A sliding window size of 38 bp was selected to ensure there

was only one seed in each window. To reduce the computational time and minimize the number of missing seeds, the sliding step size was set as 8.

To identify high-frequency *k*-mers, we counted the frequency of 20-mers (minimum seed length) using KAT v2.4.2 (Mapleson et al. 2017). For each genome, the secondary derivative of the *k*-mer frequency density distribution was calculated and the point with minimum distance to zero was identified. The *k*-mer frequency at that point was used as a threshold to define and remove high-frequency *k*-mers.

To find the alignment score that corresponds to a *p*-value < 0.1, we randomly extracted 10,000 fragments with a length of 1,000 bp from the unmasked reference genome and query genomes separately. A total of 10,000 maximum Smith-Waterman scores were calculated by aligning those fragments, and results were fit into a non-linear least square regression. The final k (0.006662) and λ (0.382291) values were determined by using maize as the reference against sequences extracted from the other five species randomly. Based on those k and λ values, alignments with a Smith-Waterman score of 54 or higher were kept.

Overlap and enrichment analysis:

The output SAM files were reformatted into BAM files using SAMtools, and the "depth" command of SAMtools was used to check how many unique base-pairs were classified as conserved. We counted how many unique CNS base-pairs overlapped with open chromatin regions (Ricci et al. 2019; Tu et al. 2020). Then, enrichment values were calculated as:

$$\frac{\frac{genome\ size}{total\ base{-}pairs\ of\ open\ chromatin}}{\frac{CNS\ base{-}pairs}{open\ chromatin\ CNS\ base{-}pairs}}.$$

The overlap and enrichment values for TFBSs (Tu et al. 2020), H3K9ac ChIP-seq peaks (Oka et al. 2017), chromatin loop anchors (Ricci et al. 2019; Peng et al. 2019), TEs (Stitzer et al. 2019), ncRNA (Han et al. 2019; Jiao et al. 2017) genes, genome-wide DNA methylation (Ricci et al. 2019), DNA replication profiles (Wear et al. 2017), and lead eQTLs (Kremling et al. 2018) were calculated in the same way.

The above mentioned features were obtained from the original publications without further processing. For the datasets using the B73 v3 genome assembly coordinates, CrossMap v0.2.8 (Hao Zhao et al. 2014) and the chain file released from Ensembl (Howe et al. 2020) were used to uplift to the maize B73 v4 genome assembly coordinates.

CNS and other features were considered overlapping if they shared ≥ 1 bp.

Comparison of gene expression for genes with and without CNS in 2 kbp upstream:

A total of 28,950 gene IDs from the expression matrix from 23 maize tissues (Walley et al. 2016) were uniquely lifted from the maize B73 v3 to the v4 genome annotation using the table released by MaizeGDB (gene_model_xref_v4.txt, https://www.maizegdb.org/search/gene/download_gene_xrefs.php?relative=v4) (Lawrence et al. 2004). Among those genes, 25,127 have upstream CNSs detected within a 2 kbp range. The tissue specificity of gene expression was measured using τ (Yanai et al. 2005) and entropy (Kadota et al. 2006).

CNS PAVs in a maize population:

For each accession in the maize Goodman panel (Flint-Garcia et al. 2005), paired-end Illumina reads (NCBI BioProject PRJNA389800) (Bukowski et al. 2018) were mapped to the maize B73 v4 reference genome using BWA-MEM v0.7.13 (H. Li and Durbin 2009) with default parameters. To save computational time, we used GATK v3.8 (McKenna et al. 2010) to classify each base-pair as "callable" or "noncallable". For the "noncallable" base-pairs, coverage was checked using the "samtools mpileup" command. Some CNS regions were shorter than the reads and only a single end of the paired-end reads fell into the conserved region. Therefore, we also mapped the reads to the CNS fragments using BWA-MEM with parameters "-a -c 200000 -S -P" and calculated the coverage of the CNS fragments using "samtools mpileup". A base-pair of a CNS fragment was classified as "present" if it was "callable", if it had coverage from reads mapping to the genome-wide, or if it had coverage from reads mapping to the CNS fragment; otherwise, it was classified as "absent". Maize accessions with low sequencing coverage (≤ 14 Gbp of reads) were excluded from the analysis, as there was not enough information to accurately conduct calling (Supplemental Fig S25).

Association analysis using CNS PAVs as independent variables:

To define CNS PAVs variables, the number of present base pair/CNS length was calculated for each CNS in each maize accession. If the ratio was ≥ 0.8, it was encoded as 1; if the ratio was ≤ 0.2, it was encoded as 0; otherwise unknown. To associate CNS PAVs with gene expression, we used CNS PAVs with a minor allele count ≥ 15 and a known allele count ≥ 35. The list of maize accessions analyzed is available as Supplemental Table S3.

Twenty-five PEER factors and three principal components for population structure generated by Kremling et al. (Kremling et al. 2018) were used as co-variants and association analysis was

conducted using a fixed linear model. Association *p*-values were calculated using an *F*-test, with a significance threshold 1×10^{-6}. The association test was performed using a custom Python script (https://github.com/baoxingsong/dCNS/blob/master/scripts/CnsBasedGwasFixedModelV2.py).

We further investigated the significant associations between CNS PAVs and gene expression level. To further check the possible functional/loss-of-function states of alleles, especially those with a presence ratio <0.8 and >0.2:

1) For each significantly association, we grouped those accessions with CNS PAV encoded as 1 as group ‘presence‘, and grouped those accessions with CNS PAV encoded as 0 as group ‘absence‘. For the significant associated gene, we calculated the median expression level of group ‘presence‘ accessions and group ‘absence‘ accessions separately.
2) For each of those accessions with the CNS presence ratio <0.8 and >0.2; if its expression level is closer to the group ‘presence’ median value, we classified it into group ‘presence’. If its expression is closer with group ‘absence’ median value, we classified it into group ‘absence’.
3) We then compared the CNS reads mapping coverage ratio of group ‘absence’ accessions with group ‘presence’ accessions.

Detection of selective sweeps:

We mapped the raw reads from resequenced landraces (L. Wang et al. 2017) to the maize B73 v4 reference genome using BWA-MEM (H. Li and Durbin 2009) with default parameters, and conducted SNP calling using the GATK v3.8 (McKenna et al. 2010) germline SNP calling best practices. Specifically, first HaplotypeCaller was used to call variants per sample and create GVCF files. Following this, we used GenomicsDBImport to consolidate GVCF files and joint-called genotypes from these with GenotypeGVFs. SelectVariants was used to output SNPs, as the sweep detection software currently cannot handle indels. We conducted genome-wide scans for selective sweep patterns using RaiSD v 2.5 (Alachiotis and Pavlidis 2018) with default

parameters, correcting for the number of base-pairs with usable sequence data using mop (https://github.com/RILAB/mop) with default parameters

In order to screen for adaptive loci as a complement to our domestication loci screen, we used the program PCAdapt (Luu, Bazin, and Blum 2017). We used the program on its default settings, but removed inversion Inv4m and known inversions on Chromosome 1 and 3, as well as a region of low recombination at the end of Chromosome 10 that is likely the abnormal 10 region (Mroczek et al. 2006). We removed these regions as low recombination regions interfere with the way that PCAdapt defines the background relatedness for comparisons.

CNS similarity between maize two subgenomes:

Syntenic genes between maize and sorghum, introduced above, were used here. To calculate the proportion of CNS sites shared in the 2 kbp upstream of the duplicated genes, we calculated the total length of the sorghum CNSs and the number of matched base-pairs shared between each maize gene copy and sorghum gene. Then, we checked the proportion of those shared base-pairs that overlapped (Supplemental Fig S22).

Data access:

All the CNS files in SAM format and BEDformat have been submitted to figshare (https://figshare.com/) under URL https://figshare.com/articles/dataset/CNS/14129006 and as Supplemental Data.

The CNS identification program source code has been submitted to GitHub (https://github.com/) under URL https://github.com/baoxingsong/dCNS and as Supplemental Code.

All the sequence reads and *de novo* assembled genomes generated in this study have been submitted to the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/) under accession number PRJNA543119.

Competing interests:

The authors declare that they have no competing interests.

Acknowledgments:

We thank Maria Katherine Mejia-Guerra for suggestions and comments on the open chromatin and TFBS analysis and William F. Thompson for insights on the DNA replication active regions. We thank Rémy Pasquet for providing plant material of *Hyparrhenia diplandra* and the USDA Germplasm Repository Information Network for material of *Chrysopogon serrulatus*. We thank Emre Cimen for help with statistical analysis, Guillaume Ramstein for discussion of eQTL and GWAS analysis, and Qi Sun and Cheng Zhou for advice about the design of sequence alignment algorithms. We also thank Robert Bukowski, Qi Sun, and Arcadio Valdes Franco for sequencing read mapping and variant calling of the maize data. Gen Xu provided suggestions on identification of low-methylated loci. Sara Miller assisted with proofreading the manuscript. We would like to thank the three anonymous reviewers for insightful suggestions and comments. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation #ACI-1548562. This project is supported by the USDA-ARS and National Science Foundation #1822330 to E.S.B., E.A.K. J.R.I. and M.B.H. and M.C.R. B.S., M.C.R., and E.S.B. designed the experiments and wrote the manuscript. B.S. and E.S.B. implemented the dCNS software. B.S. and Y.W. identified CNS. B.S. and E.R.

performed genome assembly. B.S. and H.W. conducted maize subgenome analysis. P.B., B.S and M.B. conducted association analysis. E.A.K. and M.C.R. generated new sequencing data. J.R.I. and D.J.G. conducted the domestication and local adaptation scan. J.R.I., E.A.K., and M.B.H. contributed ideas and provided technical help. All authors revised and reviewed the manuscript.

Figure legends:

Fig. 1. Procedures to identify CNSs in Andropogoneae. The maize B73 v4 genome was used as reference (red lines), while the other five genomes were individually used as a query (green lines). First, full-length CDS of each maize protein coding gene was mapped to the query genome (CDSs belonging to the same gene are linked with '>' in the cartoon) (1); then we deleted CDSs (orange lines) and high frequency *k*-mers (blue lines) (2). Next, upstream, intron, and downstream sequences were pairwise aligned using a dynamic programming algorithm (3-4). Candidate fragments below a *p*-value threshold (0.1) were defined as CNSs (5-7).

Fig. 2. Pan-Andropogoneae CNSs. (A) Phylogenetic relationships of Andropogoneae species used in this study. Andropogoneae species are in the green shaded portion of the phylogeny. **(B)** Simulation of the total length of pan-And-CNSs and core-And-CNSs by iterative random sampling of taxa. Red and blue lines indicate the pan- and core-And-CNS curves fit using points from all combinations.

Fig. 3. CNSs are primarily putative regulatory sequences. (A) Proportions of intergenic pan-And-CNSs overlapping with features of putative *cis*-regulatory sequence. **(B)** Enrichment of intergenic pan-And-CNSs in open chromatin regions, TFBSs (transcription factor binding sites), H3K9ac (acetylation of histone 3 lysine 9) ChIP-seq peaks, chromatin loop anchors, TEs (transposable elements), and non-coding RNA (ncRNA) genes.

Fig. 4. Patterns of DNA methylation and GC content in CNSs suggest diverse functions. (A) Different DNA methylation ratios among pan-And-CNS groups. Red dots correspond to CG DNA methylation, green dots are CHG DNA methylation, and blue dots represent CHH DNA methylation (where 'H' indicates A, C, or T). "other genome regions" on the horizontal axis represents DNA methylation sites located in the intergenic regions that were not defined as CNSs, and "protein coding genes" denotes DNA methylation sites located within CDSs, introns, or UTR regions of coding genes. **(B)** Different groups of pan-And-CNSs (indicated in orange, brown, and green) have distinct GC content when compared with CDSs (blue) or the genome-wide (red). **(C)** Overlap of CDS regions, *cis*, non-*cis* loop, and rest pan-And-CNSs with active regions of DNA replication in the early S phase of the mitotic cell cycle. Sequences that did not overlap with coding genes or CNSs were used as background (intergenic). **(D)** The proportion of pan-And-CNSs overlapping with annotated features. Each CNS can overlap with multiple features. Unknown CNSs are those CNSs that do not overlap with any used features.

Fig. 5. Variants in CNS regions impact gene expression. (A) MAF distribution of HapMap3 variants in CNS regions, genome-wide CDS regions and genome-wide intergenic regions. **(B)** MAF distribution of CNS PAVs in genic, *cis*, non-*cis* loop, and rest CNS groups. **(C)**

Comparison of the proportion of maintained CNSs in the 2 kbp upstream regions of the top 1,500 expressed genes in root tissues in each maize accession. Dotted lines indicate the 99% one-tailed confidence interval calculated by shuffling the gene expression ranks and CNS maintained proportions 1,000 times. Red dots are beyond the 99% one-tailed intervals. Similar patterns were observed across different tissues (Supplemental Fig S18). **(D)** Histogram of the distance between CNS PAVs and associated genes for root expression data when a PAV and its associated genes are on the same chromosome. The vertical dotted line indicates a distance of 2.5 Mbp.

Fig. 6. CNS variation was associated with expression diversity between paralogous genes in maize. (A) Correlation of CNS similarity and expression similarity of paralogous gene pairs. Red dots indicate negatively correlated genes; blue dots indicate positively correlated genes across tissues. **(B)** The shared proportion of CNS sites for negatively (red) and positively (blue) correlated paralogous gene pairs. **(C)** The diversity of CNS maintained by the maize major copy and minor copy for negatively (blue) and positively (red) correlated gene pairs. **(D)** Correlation of expression levels of the maize major copy genes with their sorghum homologous genes (red) and minor copy genes with their sorghum homologous genes (green) in shoots for genes with negatively correlated expression patterns across maize tissues in Fig. A.

References:

Akua, Tsofit, Irina Berezin, and Orit Shaul. 2010. “The Leader Intron of AtMHX Can Elicit, in the Absence of Splicing, Low-Level Intron-Mediated Enhancement That Depends on the Internal Intron Sequence.” *BMC Plant Biology* 10 (May): 93.

Alachiotis, Nikolaos, and Pavlos Pavlidis. 2018. “RAiSD Detects Positive Selection Based on Multiple Signatures of a Selective Sweep and SNP Vectors.” *Communications Biology* 1 (June): 79.

Algama, Manjula, Edward Tasker, Caitlin Williams, Adam C. Parslow, Robert J. Bryson-

Richardson, and Jonathan M. Keith. 2017. “Genome-Wide Identification of Conserved Intronic Non-Coding Sequences Using a Bayesian Segmentation Approach.” *BMC Genomics* 18 (1): 259.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. “Basic Local Alignment Search Tool.” *Journal of Molecular Biology* 215 (3): 403–10.

Avsec, Žiga, Melanie Weilert, Avanti Shrikumar, Sabrina Krueger, Amr Alexandari, Khyati Dalal, Robin Fropf, et al. 2021. “Base-Resolution Models of Transcription-Factor Binding Reveal Soft Motif Syntax.” *Nature Genetics* 53 (3): 354–66.

Black, C. C., T. M. Chen, and R. H. Brown. 1969. “Biochemical Basis for Plant Competition.” *Weed Science* 17 (3): 338–44.

Brosse, Nicolas, Anthony Dufour, Xianzhi Meng, Qining Sun, and Arthur Ragauskas. 2012. “Miscanthus : A Fast-Growing Crop for Biofuels and Chemicals Production.” *Biofuels, Bioproducts & Biorefining* 6 (5): 580–98.

Bukowski, Robert, Xiaosen Guo, Yanli Lu, Cheng Zou, Bing He, Zhengqin Rong, Bo Wang, et al. 2018. “Construction of the Third-Generation Zea Mays Haplotype Map.” *GigaScience* 7 (4): 1–12.

Chen, Yung-Chia Ariel, and Alexei A. Aravin. 2015. “Non-Coding RNAs in Transcriptional Regulation: The Review for Current Molecular Biology Reports.” *Current Molecular Biology Reports* 1 (1): 10–18.

Chorev, Michal, and Liran Carmel. 2012. “The Function of Introns.” *Frontiers in Genetics* 3 (April): 55.

Clark, Richard M., Tina Nussbaum Wagler, Pablo Quijada, and John Doebley. 2006. “A Distant Upstream Enhancer at the Maize Domestication Gene tb1 Has Pleiotropic Effects on Plant and Inflorescent Architecture.” *Nature Genetics* 38 (5): 594–97.

Cvijović, Ivana, Benjamin H. Good, and Michael M. Desai. 2018. “The Effect of Strong Purifying Selection on Genetic Diversity.” *Genetics* 209 (4): 1235–78.

De Coster, Wouter, Svenn D’Hert, Darrin T. Schultz, Marc Cruts, and Christine Van Broeckhoven. 2018. “NanoPack: Visualizing and Processing Long-Read Sequencing Data.” *Bioinformatics* 34 (15): 2666–69.

Delaneau, O., M. Zazhytska, C. Borel, G. Giannuzzi, G. Rey, C. Howald, S. Kumar, et al. 2019. “Chromatin Three-Dimensional Interactions Mediate Genetic Effects on Gene Expression.” *Science* 364 (6439). https://doi.org/10.1126/science.aat8266.

Diehl, Adam G., Ningxin Ouyang, and Alan P. Boyle. 2020. “Transposable Elements Contribute to Cell and Species-Specific Chromatin Looping and Gene Regulation in Mammalian Genomes.” *Nature Communications* 11 (1): 1796.

Dong, Pengfei, Xiaoyu Tu, Po-Yu Chu, Peitao Lü, Ning Zhu, Donald Grierson, Baijuan Du, Pinghua Li, and Silin Zhong. 2017. “3D Chromatin Architecture of Large Plant Genomes Determined by Local A/B Compartments.” *Molecular Plant* 10 (12): 1497–1509.

Dong, Pengfei, Xiaoyu Tu, Haoxuan Li, Jianhua Zhang, Donald Grierson, Pinghua Li, and Silin Zhong. 2020. “Tissue-Specific Hi-C Analyses of Rice, Foxtail Millet and Maize Suggest Non-Canonical Function of Plant Chromatin Domains.” *Journal of Integrative Plant Biology* 62 (2): 201–17.

Dong, Qianli, Ning Li, Xiaochong Li, Zan Yuan, Dejian Xie, Xiaofei Wang, Jianing Li, et al. 2018. “Genome-Wide Hi-C Analysis Reveals Extensive Hierarchical Chromatin Interactions in Rice.” *The Plant Journal: For Cell and Molecular Biology* 94 (6): 1141–56.

Dror, Iris, Tamar Golan, Carmit Levy, Remo Rohs, and Yael Mandel-Gutfreund. 2015. “A

Widespread Role of the Motif Environment in Transcription Factor Binding across Diverse Protein Families." *Genome Research* 25 (9): 1268–80.
Dupeyron, Mathilde, Kumar S. Singh, Chris Bass, and Alexander Hayward. 2019. "Evolution of Mutator Transposable Elements across Eukaryotic Diversity." *Mobile DNA* 10 (March): 12.
Emms, David M., and Steven Kelly. 2019. "OrthoFinder: Phylogenetic Orthology Inference for Comparative Genomics." *Genome Biology* 20 (1): 238.
ENCODE Project Consortium. 2012. "An Integrated Encyclopedia of DNA Elements in the Human Genome." *Nature* 489 (7414): 57–74.
Estep, Matt C., Michael R. McKain, Dilys Vela Diaz, Jinshun Zhong, John G. Hodge, Trevor R. Hodkinson, Daniel J. Layton, Simon T. Malcomber, Rémy Pasquet, and Elizabeth A. Kellogg. 2014. "Allopolyploidy, Diversification, and the Miocene Grassland Expansion." *Proceedings of the National Academy of Sciences of the United States of America* 111 (42): 15149–54.
Finucane, Hilary K., Brendan Bulik-Sullivan, Alexander Gusev, Gosia Trynka, Yakir Reshef, Po-Ru Loh, Verneri Anttila, et al. 2015. "Partitioning Heritability by Functional Annotation Using Genome-Wide Association Summary Statistics." *Nature Genetics* 47 (11): 1228–35.
Flint-Garcia, Sherry A., Anne-Céline Thuillet, Jianming Yu, Gael Pressoir, Susan M. Romero, Sharon E. Mitchell, John Doebley, Stephen Kresovich, Major M. Goodman, and Edward S. Buckler. 2005. "Maize Association Population: A High-Resolution Platform for Quantitative Trait Locus Dissection: High-Resolution Maize Association Population." *The Plant Journal: For Cell and Molecular Biology* 44 (6): 1054–64.
Freeling, Michael, and Shabarinath Subramaniam. 2009. "Conserved Noncoding Sequences (CNSs) in Higher Plants." *Current Opinion in Plant Biology* 12 (2): 126–32.
Gasperini, Molly, Jacob M. Tome, and Jay Shendure. 2020. "Towards a Comprehensive Catalogue of Validated and Target-Linked Human Enhancers." *Nature Reviews. Genetics* 21 (5): 292–310.
Gerstein, Mark B., Anshul Kundaje, Manoj Hariharan, Stephen G. Landt, Koon-Kiu Yan, Chao Cheng, Xinmeng Jasmine Mu, et al. 2012. "Architecture of the Human Regulatory Network Derived from ENCODE Data." *Nature* 489 (7414): 91–100.
Giral, Hector, Ulf Landmesser, and Adelheid Kratzer. 2018. "Into the Wild: GWAS Exploration of Non-Coding RNAs." *Frontiers in Cardiovascular Medicine* 5 (December): 181.
Greene, B., R. Walko, and S. Hake. 1994. "Mutator Insertions in an Intron of the Maize knotted1 Gene Result in Dominant Suppressible Mutations." *Genetics* 138 (4): 1275–85.
Greenwald, William W., He Li, Paola Benaglio, David Jakubosky, Hiroko Matsui, Anthony Schmitt, Siddarth Selvaraj, et al. 2019. "Subtle Changes in Chromatin Loop Contact Propensity Are Associated with Differential Gene Regulation and Expression." *Nature Communications* 10 (1): 1054.
Guo, Hena, and Stephen P. Moose. 2003. "Conserved Noncoding Sequences among Cultivated Cereal Genomes Identify Candidate Regulatory Sequence Elements and Patterns of Promoter Evolution." *The Plant Cell* 15 (5): 1143–58.
Hammelman, Jennifer, Konstantin Krismer, Budhaditya Banerjee, David K. Gifford, and Richard I. Sherwood. 2020. "Identification of Determinants of Differential Chromatin Accessibility through a Massively Parallel Genome-Integrated Reporter Assay." *Genome Research* 30 (10): 1468–80.
Han, Linqian, Zhenna Mu, Zi Luo, Qingchun Pan, and Lin Li. 2019. "New lncRNA Annotation Reveals Extensive Functional Divergence of the Transcriptome in Maize." *Journal of*

Integrative Plant Biology 61 (4): 394–405.
Harmston, Nathan, Elizabeth Ing-Simmons, Ge Tan, Malcolm Perry, Matthias Merkenschlager, and Boris Lenhard. 2017. “Topologically Associating Domains Are Ancient Features That Coincide with Metazoan Clusters of Extreme Noncoding Conservation.” *Nature Communications* 8 (1): 441.
Haudry, Annabelle, Adrian E. Platts, Emilio Vello, Douglas R. Hoen, Mickael Leclercq, Robert J. Williamson, Ewa Forczek, et al. 2013. “An Atlas of over 90,000 Conserved Noncoding Sequences Provides Insight into Crucifer Regulatory Regions.” *Nature Genetics* 45 (8): 891–98.
Howe, Kevin L., Bruno Contreras-Moreira, Nishadi De Silva, Gareth Maslen, Wasiu Akanni, James Allen, Jorge Alvarez-Jarreta, et al. 2020. “Ensembl Genomes 2020-Enabling Non-Vertebrate Genomic Research.” *Nucleic Acids Research* 48 (D1): D689–95.
Inada, Dan Choffnes, Ali Bashir, Chunghau Lee, Brian C. Thomas, Cynthia Ko, Stephen A. Goff, and Michael Freeling. 2003. “Conserved Noncoding Sequences in the Grasses.” *Genome Research* 13 (9): 2030–41.
Jabbari, Kamel, Johannes Wirtz, Martina Rauscher, and Thomas Wiehe. 2019. “A Common Genomic Code for Chromatin Architecture and Recombination Landscape.” *PloS One* 14 (3): e0213278.
Jiao, Yinping, Paul Peluso, Jinghua Shi, Tiffany Liang, Michelle C. Stitzer, Bo Wang, Michael S. Campbell, et al. 2017. “Improved Maize Reference Genome with Single-Molecule Technologies.” *Nature* 546 (7659): 524–27.
Kadauke, Stephan, and Gerd A. Blobel. 2009. “Chromatin Loops in Gene Regulation.” *Biochimica et Biophysica Acta* 1789 (1): 17–25.
Kadota, Koji, Jiazhen Ye, Yuji Nakai, Tohru Terada, and Kentaro Shimizu. 2006. “ROKU: A Novel Method for Identification of Tissue-Specific Genes.” *BMC Bioinformatics* 7 (June): 294.
Karlin, S., and S. F. Altschul. 1990. “Methods for Assessing the Statistical Significance of Molecular Sequence Features by Using General Scoring Schemes.” *Proceedings of the National Academy of Sciences of the United States of America* 87 (6): 2264–68.
Khalturin, Konstantin, Friederike Anton-Erxleben, Sylvia Sassmann, Jörg Wittlieb, Georg Hemmrich, and Thomas C. G. Bosch. 2008. “A Novel Gene Family Controls Species-Specific Morphological Traits in Hydra.” *PLoS Biology* 6 (11): e278.
Kistler, Logan, S. Yoshi Maezumi, Jonas Gregorio de Souza, Natalia A. S. Przelomska, Flaviane Malaquias Costa, Oliver Smith, Hope Loiselle, et al. 2018. “Multiproxy Evidence Highlights a Complex Evolutionary Legacy of Maize in South America.” *Science* 362 (6420): 1309–13.
Kolmogorov, Mikhail, Jeffrey Yuan, Yu Lin, and Pavel A. Pevzner. 2019. “Assembly of Long, Error-Prone Reads Using Repeat Graphs.” *Nature Biotechnology* 37 (5): 540–46.
Kremling, Karl A. G., Shu-Yun Chen, Mei-Hsiu Su, Nicholas K. Lepak, M. Cinta Romay, Kelly L. Swarts, Fei Lu, Anne Lorant, Peter J. Bradbury, and Edward S. Buckler. 2018. “Dysregulation of Expression Correlates with Rare-Allele Burden and Fitness Loss in Maize.” *Nature* 555 (7697): 520–23.
Lai, Xianjun, Sairam Behera, Zhikai Liang, Yanli Lu, Jitender S. Deogun, and James C. Schnable. 2017. “STAG-CNS: An Order-Aware Conserved Noncoding Sequences Discovery Tool for Arbitrary Numbers of Species.” *Molecular Plant* 10 (7): 990–99.
Lawrence, Carolyn J., Qunfeng Dong, Mary L. Polacco, Trent E. Seigfried, and Volker Brendel.

2004. “MaizeGDB, the Community Database for Maize Genetics and Genomics.” *Nucleic Acids Research* 32 (Database issue): D393–97.

Levo, Michal, Einat Zalckvar, Eilon Sharon, Ana Carolina Dantas Machado, Yael Kalma, Maya Lotam-Pompan, Adina Weinberger, Zohar Yakhini, Remo Rohs, and Eran Segal. 2015. “Unraveling Determinants of Transcription Factor Binding Outside the Core Binding Site.” *Genome Research* 25 (7): 1018–29.

Lewin, Harris A., Gene E. Robinson, W. John Kress, William J. Baker, Jonathan Coddington, Keith A. Crandall, Richard Durbin, et al. 2018. “Earth BioGenome Project: Sequencing Life for the Future of Life.” *Proceedings of the National Academy of Sciences of the United States of America* 115 (17): 4325–33.

Li, Heng. 2018. “Minimap2: Pairwise Alignment for Nucleotide Sequences.” *Bioinformatics* 34 (18): 3094–3100.

Li, Heng, and Richard Durbin. 2009. “Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform.” *Bioinformatics* 25 (14): 1754–60.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. “The Sequence Alignment/Map Format and SAMtools.” *Bioinformatics* 25 (16): 2078–79.

Li, Heng, and Nils Homer. 2010. “A Survey of Sequence Alignment Algorithms for next-Generation Sequencing.” *Briefings in Bioinformatics* 11 (5): 473–83.

Liu, Chang, Ying-Juan Cheng, Jia-Wei Wang, and Detlef Weigel. 2017. “Prominent Topologically Associated Domains Differentiate Global Chromatin Packing in Rice from Arabidopsis.” *Nature Plants* 3 (9): 742–48.

Li, Zijuan, Meiyue Wang, Kande Lin, Yilin Xie, Jingyu Guo, Luhuan Ye, Yili Zhuang, et al. 2019. “The Bread Wheat Epigenomic Map Reveals Distinct Chromatin Architectural and Evolutionary Features of Functional Genetic Elements.” *Genome Biology* 20 (1): 139.

Luu, Keurcien, Eric Bazin, and Michael G. B. Blum. 2017. “Pcadapt: An R Package to Perform Genome Scans for Selection Based on Principal Component Analysis.” *Molecular Ecology Resources* 17 (1): 67–77.

Lu, Zefu, Alexandre P. Marand, William A. Ricci, Christina L. Ethridge, Xiaoyu Zhang, and Robert J. Schmitz. 2019. “The Prevalence, Evolution and Chromatin Signatures of Plant Regulatory Elements.” *Nature Plants* 5 (12): 1250–59.

Makarevitch, Irina, Amanda J. Waters, Patrick T. West, Michelle Stitzer, Candice N. Hirsch, Jeffrey Ross-Ibarra, and Nathan M. Springer. 2015. “Transposable Elements Contribute to Activation of Maize Genes in Response to Abiotic Stress.” *PLoS Genetics* 11 (1): e1004915.

Manners, John M. 2011. “Chapter 3 - Functional Genomics of Sugarcane.” In *Advances in Botanical Research*, edited by Jean-Claude Kader and Michel Delseny, 60:89–168. Academic Press.

Mapleson, Daniel, Gonzalo Garcia Accinelli, George Kettleborough, Jonathan Wright, and Bernardo J. Clavijo. 2017. “KAT: A K-Mer Analysis Toolkit to Quality Control NGS Datasets and Genome Assemblies.” *Bioinformatics* 33 (4): 574–76.

McCormick, Ryan F., Sandra K. Truong, Avinash Sreedasyam, Jerry Jenkins, Shengqiang Shu, David Sims, Megan Kennedy, et al. 2018. “The Sorghum Bicolor Reference Genome: Improved Assembly, Gene Annotations, a Transcriptome Atlas, and Signatures of Genome Organization.” *The Plant Journal: For Cell and Molecular Biology* 93 (2): 338–54.

McKenna, Aaron, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, et al. 2010. "The Genome Analysis Toolkit: A MapReduce Framework for Analyzing next-Generation DNA Sequencing Data." *Genome Research* 20 (9): 1297–1303.

McLean, Cory Y., Philip L. Reno, Alex A. Pollen, Abraham I. Bassan, Terence D. Capellini, Catherine Guenther, Vahan B. Indjeian, et al. 2011. "Human-Specific Loss of Regulatory DNA and the Evolution of Human-Specific Traits." *Nature* 471 (7337): 216–19.

Mitros, Therese, Adam M. Session, Brandon T. James, Guohong Albert Wu, Mohammad B. Belaffif, Lindsay V. Clark, Shengqiang Shu, et al. 2020. "Genome Biology of the Paleotetraploid Perennial Biomass Crop Miscanthus." *Nature Communications* 11 (1): 5442.

Mroczek, Rebecca J., Juliana R. Melo, Amy C. Luce, Evelyn N. Hiatt, and R. Kelly Dawe. 2006. "The Maize Ab10 Meiotic Drive System Maps to Supernumerary Sequences in a Large Complex Haplotype." *Genetics* 174 (1): 145–54.

Mugal, Carina F., Peter F. Arndt, Lena Holm, and Hans Ellegren. 2015. "Evolutionary Consequences of DNA Methylation on the GC Content in Vertebrate Genomes." *G3* 5 (3): 441–47.

Naughton, Catherine, Nicolaos Avlonitis, Samuel Corless, James G. Prendergast, Ioulia K. Mati, Paul P. Eijk, Scott L. Cockroft, Mark Bradley, Bauke Ylstra, and Nick Gilbert. 2013. "Transcription Forms and Remodels Supercoiling Domains Unfolding Large-Scale Chromatin Structures." *Nature Structural & Molecular Biology* 20 (3): 387–95.

Nishizaki, Sierra S., and Alan P. Boyle. 2017. "Mining the Unknown: Assigning Function to Noncoding Single Nucleotide Polymorphisms." *Trends in Genetics: TIG* 33 (1): 34–45.

Noshay, Jaclyn M., Alexandre P. Marand, Sarah N. Anderson, Peng Zhou, Maria Katherine Mejia Guerra, Zefu Lu, Christine O'Connor, et al. 2020. "Cis-Regulatory Elements within TEs Can Influence Expression of Nearby Maize Genes." https://doi.org/10.1101/2020.05.20.107169.

Oka, Rurika, Johan Zicola, Blaise Weber, Sarah N. Anderson, Charlie Hodgman, Jonathan I. Gent, Jan-Jaap Wesselink, et al. 2017. "Genome-Wide Mapping of Transcriptional Enhancer Candidates Using DNA and Chromatin Features in Maize." *Genome Biology* 18 (1): 137.

O'Malley, Ronan C., Shao-Shan Carol Huang, Liang Song, Mathew G. Lewsey, Anna Bartlett, Joseph R. Nery, Mary Galli, Andrea Gallavotti, and Joseph R. Ecker. 2016. "Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape." *Cell* 165 (5): 1280–92.

Parker, Stephen C. J., Elliott H. Margulies, and Thomas D. Tullius. 2008. "The Relationship between Fine Scale DNA Structure, GC Content, and Functional Elements in 1% of the Human Genome." *Genome Informatics. International Conference on Genome Informatics* 20: 199–211.

Parvathaneni, Rajiv K., Edoardo Bertolini, Md Shamimuzzaman, Daniel L. Vera, Pei-Yau Lung, Brian R. Rice, Jinfeng Zhang, et al. 2020. "The Regulatory Landscape of Early Maize Inflorescence Development." *Genome Biology* 21 (1): 165.

Peng, Yong, Dan Xiong, Lun Zhao, Weizhi Ouyang, Shuangqi Wang, Jun Sun, Qing Zhang, et al. 2019. "Chromatin Interaction Maps Reveal Genetic Regulation for Quantitative Traits in Maize." *Nature Communications* 10 (1): 2632.

Polychronopoulos, Dimitris, James W. D. King, Alexander J. Nash, Ge Tan, and Boris Lenhard. 2017. "Conserved Non-Coding Elements: Developmental Gene Regulation Meets Genome

Organization." *Nucleic Acids Research* 45 (22): 12611–24.
Ramachandran, Dhanushya, Michael R. McKain, Elizabeth A. Kellogg, and Jennifer S. Hawkins. 2020. "Evolutionary Dynamics of Transposable Elements Following a Shared Polyploidization Event in the Tribe Andropogoneae." *G3* 10 (12): 4387–98.
Ricci, William A., Zefu Lu, Lexiang Ji, Alexandre P. Marand, Christina L. Ethridge, Nathalie G. Murphy, Jaclyn M. Noshay, et al. 2019. "Widespread Long-Range Cis-Regulatory Elements in the Maize Genome." *Nature Plants* 5 (12): 1237–49.
Rigau, Maria, David Juan, Alfonso Valencia, and Daniel Rico. 2019. "Intronic CNVs and Gene Expression Variation in Human Populations." *PLoS Genetics* 15 (1): e1007902.
Ritchie, Graham Rs, and Paul Flicek. 2014. "Computational Approaches to Interpreting Genomic Sequence Variation." *Genome Medicine* 6 (10): 87.
Rodgers-Melnick, Eli, Peter J. Bradbury, Robert J. Elshire, Jeffrey C. Glaubitz, Charlotte B. Acharya, Sharon E. Mitchell, Chunhui Li, Yongxiang Li, and Edward S. Buckler. 2015. "Recombination in Diverse Maize Is Stable, Predictable, and Associated with Genetic Load." *Proceedings of the National Academy of Sciences of the United States of America* 112 (12): 3823–28.
Rodgers-Melnick, Eli, Daniel L. Vera, Hank W. Bass, and Edward S. Buckler. 2016. "Open Chromatin Reveals the Functional Maize Genome." *Proceedings of the National Academy of Sciences of the United States of America* 113 (22): E3177–84.
Ross, J. L., P. P. Fong, and D. R. Cavener. 1994. "Correlated Evolution of the Cis-Acting Regulatory Elements and Developmental Expression of the Drosophila Gld Gene in Seven Species from the Subgroup Melanogaster." *Developmental Genetics* 15 (1): 38–50.
Sage, Rowan F., and Xin-Guang Zhu. 2011. "Exploiting the Engine of C(4) Photosynthesis." *Journal of Experimental Botany* 62 (9): 2989–3000.
Salvi, Silvio, Giorgio Sponza, Michele Morgante, Dwight Tomes, Xiaomu Niu, Kevin A. Fengler, Robert Meeley, et al. 2007. "Conserved Noncoding Genomic Sequences Associated with a Flowering-Time Quantitative Trait Locus in Maize." *Proceedings of the National Academy of Sciences of the United States of America* 104 (27): 11376–81.
Schnable, James C., Nathan M. Springer, and Michael Freeling. 2011. "Differentiation of the Maize Subgenomes by Genome Dominance and Both Ancient and Ongoing Gene Loss." *Proceedings of the National Academy of Sciences of the United States of America* 108 (10): 4069–74.
Schnable, James, and Eric Lyons. 2015. "Plant Paleopolyploidy." https://doi.org/10.6084/m9.figshare.1538627.v1.
Schwartz, Uwe, Attila Németh, Sarah Diermeier, Josef H. Exler, Stefan Hansch, Rodrigo Maldonado, Leonhard Heizinger, Rainer Merkl, and Gernot Längst. 2019. "Characterizing the Nuclease Accessibility of DNA in Human Cells to Map Higher Order Structures of Chromatin." *Nucleic Acids Research* 47 (3): 1239–54.
Shen, Zeyang, Jenhan Tao, Gregory J. Fonseca, and Christopher K. Glass. 2020. "Natural Genetic Variation Affecting Transcription Factor Spacing at Regulatory Regions Is Generally Well Tolerated." *Cold Spring Harbor Laboratory*. https://doi.org/10.1101/2020.04.02.021535.
Simão, Felipe A., Robert M. Waterhouse, Panagiotis Ioannidis, Evgenia V. Kriventseva, and Evgeny M. Zdobnov. 2015. "BUSCO: Assessing Genome Assembly and Annotation Completeness with Single-Copy Orthologs." *Bioinformatics* 31 (19): 3210–12.
Smith, Aileen M., Maria-Jose Sanchez, George A. Follows, Sarah Kinston, Ian J. Donaldson,

Anthony R. Green, and Berthold Göttgens. 2008. “A Novel Mode of Enhancer Evolution: The Tal1 Stem Cell Enhancer Recruited a MIR Element to Specifically Boost Its Activity.” *Genome Research* 18 (9): 1422–32.

Smith, T. F., and M. S. Waterman. 1981. “Identification of Common Molecular Subsequences.” *Journal of Molecular Biology* 147 (1): 195–97.

Statello, Luisa, Chun-Jie Guo, Ling-Ling Chen, and Maite Huarte. 2021. “Gene Regulation by Long Non-Coding RNAs and Its Biological Functions.” *Nature Reviews. Molecular Cell Biology* 22 (2): 96–118.

Stitzer, Michelle C., Sarah N. Anderson, Nathan M. Springer, and Jeffrey Ross-Ibarra. 2019. “The Genomic Ecosystem of Transposable Elements in Maize.” *bioRxiv*. https://doi.org/10.1101/559922.

Studer, Anthony, Qiong Zhao, Jeffrey Ross-Ibarra, and John Doebley. 2011. “Identification of a Functional Transposon Insertion in the Maize Domestication Gene tb1.” *Nature Genetics* 43 (11): 1160–63.

Suzuki, Miho M., and Adrian Bird. 2008. “DNA Methylation Landscapes: Provocative Insights from Epigenomics.” *Nature Reviews. Genetics* 9 (6): 465–76.

Swigonová, Zuzana, Jinsheng Lai, Jianxin Ma, Wusirika Ramakrishna, Victor Llaca, Jeffrey L. Bennetzen, and Joachim Messing. 2004. “Close Split of Sorghum and Maize Genome Progenitors.” *Genome Research* 14 (10A): 1916–23.

Szabo, Quentin, Frédéric Bantignies, and Giacomo Cavalli. 2019. “Principles of Genome Folding into Topologically Associating Domains.” *Science Advances* 5 (4): eaaw1668.

Tang, Haibao, Eric Lyons, Brent Pedersen, James C. Schnable, Andrew H. Paterson, and Michael Freeling. 2011. “Screening Synteny Blocks in Pairwise Genome Comparisons through Integer Programming.” *BMC Bioinformatics* 12 (April): 102.

The ENCODE Project Consortium. 2012. “An Integrated Encyclopedia of DNA Elements in the Human Genome.” *Nature* 489 (7414): 57–74.

Tian, Feng, De-Chang Yang, Yu-Qi Meng, Jinpu Jin, and Ge Gao. 2020. “PlantRegMap: Charting Functional Regulatory Maps in Plants.” *Nucleic Acids Research* 48 (D1): D1104–13.

Tu, Xiaoyu, María Katherine Mejía-Guerra, Jose A. Valdes Franco, David Tzeng, Po-Yu Chu, Wei Shen, Yingying Wei, et al. 2020. “Reconstructing the Maize Leaf Regulatory Network Using ChIP-Seq Data of 104 Transcription Factors.” *Nature Communications* 11 (1): 5089.

Vandepoele, Klaas, Tineke Casneuf, and Yves Van de Peer. 2006. “Identification of Novel Regulatory Modules in Dicotyledonous Plants Using Expression Data and Comparative Genomics.” *Genome Biology* 7 (11): R103.

Van de Velde, Jan, Michiel Van Bel, Dries Vaneechoutte, and Klaas Vandepoele. 2016. “A Collection of Conserved Noncoding Sequences to Study Gene Regulation in Flowering Plants.” *Plant Physiology* 171 (4): 2586–98.

Vaser, Robert, Ivan Sović, Niranjan Nagarajan, and Mile Šikić. 2017. “Fast and Accurate de Novo Genome Assembly from Long Uncorrected Reads.” *Genome Research* 27 (5): 737–46.

Vicentini, Alberto, Janet C. Barber, Sandra S. Aliscioni, Liliana M. Giussani, and Elizabeth A. Kellogg. 2008. “The Age of the Grasses and Clusters of Origins of C 4 Photosynthesis.” *Global Change Biology* 14 (12): 2963–77.

Viturawong, Tar, Felix Meissner, Falk Butter, and Matthias Mann. 2013. “A DNA-Centric Protein Interaction Map of Ultraconserved Elements Reveals Contribution of Transcription

Factor Binding Hubs to Conservation." *Cell Reports* 5 (2): 531–45.
Walker, Bruce J., Thomas Abeel, Terrance Shea, Margaret Priest, Amr Abouelliel, Sharadha Sakthikumar, Christina A. Cuomo, et al. 2014. "Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement." *PloS One* 9 (11): e112963.
Wallace, Jason G., Peter J. Bradbury, Nengyi Zhang, Yves Gibon, Mark Stitt, and Edward S. Buckler. 2014. "Association Mapping across Numerous Traits Reveals Patterns of Functional Variation in Maize." *PLoS Genetics*. https://doi.org/10.1371/journal.pgen.1004845.
Walley, Justin W., Ryan C. Sartor, Zhouxin Shen, Robert J. Schmitz, Kevin J. Wu, Mark A. Urich, Joseph R. Nery, et al. 2016. "Integration of Omic Networks in a Developmental Atlas of Maize." *Science* 353 (6301): 814–18.
Wang, Li, Timothy M. Beissinger, Anne Lorant, Claudia Ross-Ibarra, Jeffrey Ross-Ibarra, and Matthew B. Hufford. 2017. "The Interplay of Demography and Selection during Maize Domestication and Expansion." *Genome Biology* 18 (1): 215.
Wang, Maojun, Lili Tu, Min Lin, Zhongxu Lin, Pengcheng Wang, Qingyong Yang, Zhengxiu Ye, et al. 2017. "Asymmetric Subgenome Selection and Cis-Regulatory Divergence during Cotton Domestication." *Nature Genetics* 49 (4): 579–87.
Wang, Maojun, Pengcheng Wang, Min Lin, Zhengxiu Ye, Guoliang Li, Lili Tu, Chao Shen, Jianying Li, Qingyong Yang, and Xianlong Zhang. 2018. "Evolutionary Dynamics of 3D Genome Architecture Following Polyploidization in Cotton." *Nature Plants* 4 (2): 90–97.
Warnefors, Maria, Britta Hartmann, Stefan Thomsen, and Claudio R. Alonso. 2016. "Combinatorial Gene Regulatory Functions Underlie Ultraconserved Elements in Drosophila." *Molecular Biology and Evolution* 33 (9): 2294–2306.
Washburn, Jacob D., Maria Katherine Mejia-Guerra, Guillaume Ramstein, Karl A. Kremling, Ravi Valluru, Edward S. Buckler, and Hai Wang. 2019. "Evolutionarily Informed Deep Learning Methods for Predicting Relative Transcript Abundance from DNA Sequence." *Proceedings of the National Academy of Sciences of the United States of America* 116 (12): 5542–49.
Wear, Emily E., Jawon Song, Gregory J. Zynda, Chantal LeBlanc, Tae-Jin Lee, Leigh Mickelson-Young, Lorenzo Concia, et al. 2017. "Genomic Analysis of the DNA Replication Timing Program during Mitotic S Phase in Maize (Zea Mays) Root Tips." *The Plant Cell* 29 (9): 2126–49.
Wick, Ryan. n.d. *Porechop*. Github. Accessed December 24, 2019. https://github.com/rrwick/Porechop.
Wittkopp, Patricia J., Kathy Vaccaro, and Sean B. Carroll. 2002. "Evolution of Yellow Gene Regulation and Pigmentation in Drosophila." *Current Biology: CB* 12 (18): 1547–56.
Won, Hyejung, Jerry Huang, Carli K. Opland, Chris L. Hartl, and Daniel H. Geschwind. 2019. "Human Evolved Regulatory Elements Modulate Genes Involved in Cortical Expansion and Neurodevelopmental Disease Susceptibility." *Nature Communications* 10 (1): 2396.
Xiang, Ruidong, Irene van den Berg, Iona M. MacLeod, Benjamin J. Hayes, Claire P. Prowse-Wilkins, Min Wang, Sunduimijid Bolormaa, et al. 2019. "Quantifying the Contribution of Sequence Variants with Regulatory and Evolutionary Significance to 34 Bovine Complex Traits." *Proceedings of the National Academy of Sciences of the United States of America* 116 (39): 19398–408.
Xie, Jianbo, Kecheng Qian, Jingna Si, Liang Xiao, Dong Ci, and Deqiang Zhang. 2018.

"Conserved Noncoding Sequences Conserve Biological Networks and Influence Genome Evolution." *Heredity* 120 (5): 437–51.

Xie, Xiaohui, Michael Kamal, and Eric S. Lander. 2006. "A Family of Conserved Noncoding Elements Derived from an Ancient Transposable Element." *Proceedings of the National Academy of Sciences of the United States of America* 103 (31): 11659–64.

Xu, Gen, Jing Lyu, Qing Li, Han Liu, Dafang Wang, Mei Zhang, Nathan M. Springer, Jeffrey Ross-Ibarra, and Jinliang Yang. 2020. "Evolutionary and Functional Genomics of DNA Methylation in Maize Domestication and Improvement." *Nature Communications* 11 (1): 5539.

Yanai, Itai, Hila Benjamin, Michael Shmoish, Vered Chalifa-Caspi, Maxim Shklar, Ron Ophir, Arren Bar-Even, et al. 2005. "Genome-Wide Midrange Transcription Profiles Reveal Expression Level Relationships in Human Tissue Specification." *Bioinformatics* 21 (5): 650–59.

Yao, Run-Wen, Yang Wang, and Ling-Ling Chen. 2019. "Cellular Functions of Long Noncoding RNAs." *Nature Cell Biology* 21 (5): 542–51.

Yoon, Je-Hyun, Kotb Abdelmohsen, and Myriam Gorospe. 2013. "Posttranscriptional Gene Regulation by Long Noncoding RNA." *Journal of Molecular Biology* 425 (19): 3723–30.

Zhang, Feng, and James R. Lupski. 2015. "Non-Coding Genetic Variants in Human Disease." *Human Molecular Genetics* 24 (R1): R102–10.

Zhang, Jisen, Xingtan Zhang, Haibao Tang, Qing Zhang, Xiuting Hua, Xiaokai Ma, Fan Zhu, et al. 2018. "Allele-Defined Genome of the Autopolyploid Sugarcane Saccharum Spontaneum L." *Nature Genetics* 50 (11): 1565–73.

Zhang, Wenli, Yufeng Wu, James C. Schnable, Zixian Zeng, Michael Freeling, Gregory E. Crawford, and Jiming Jiang. 2012. "High-Resolution Mapping of Open Chromatin in the Rice Genome." *Genome Research* 22 (1): 151–62.

Zhang, Xiaoyu, Junshi Yazaki, Ambika Sundaresan, Shawn Cokus, Simon W-L Chan, Huaming Chen, Ian R. Henderson, et al. 2006. "Genome-Wide High-Resolution Mapping and Functional Analysis of DNA Methylation in Arabidopsis." *Cell* 126 (6): 1189–1201.

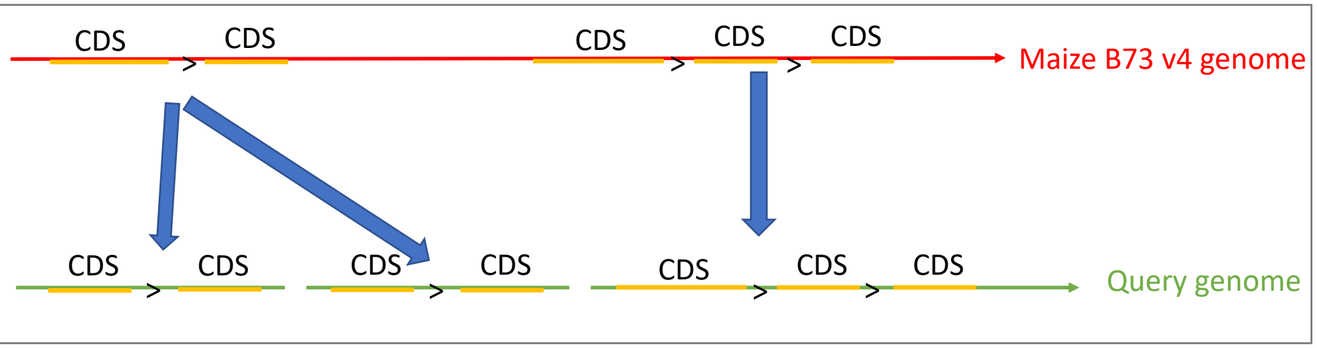
Zhang, Z., P. Berman, and W. Miller. 1998. "Alignments without Low-Scoring Regions." *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology* 5 (2): 197–210.

Zhao, Hainan, Wenli Zhang, Lifen Chen, Lei Wang, Alexandre P. Marand, Yufeng Wu, and Jiming Jiang. 2018. "Proliferation of Regulatory DNA Elements Derived from Transposable Elements in the Maize Genome." *Plant Physiology* 176 (4): 2789–2803.

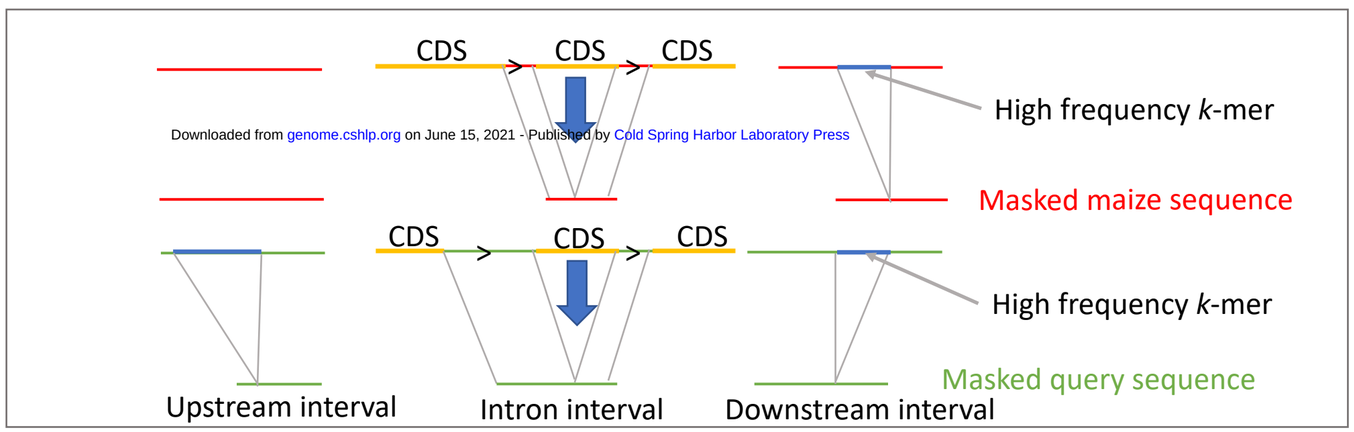
Zhao, Hao, Zhifu Sun, Jing Wang, Haojie Huang, Jean-Pierre Kocher, and Liguo Wang. 2014. "CrossMap: A Versatile Tool for Coordinate Conversion between Genome Assemblies." *Bioinformatics* 30 (7): 1006–7.

Zilberman, Daniel, Mary Gehring, Robert K. Tran, Tracy Ballinger, and Steven Henikoff. 2007. "Genome-Wide Analysis of Arabidopsis Thaliana DNA Methylation Uncovers an Interdependence between Methylation and Transcription." *Nature Genetics* 39 (1): 61–69.

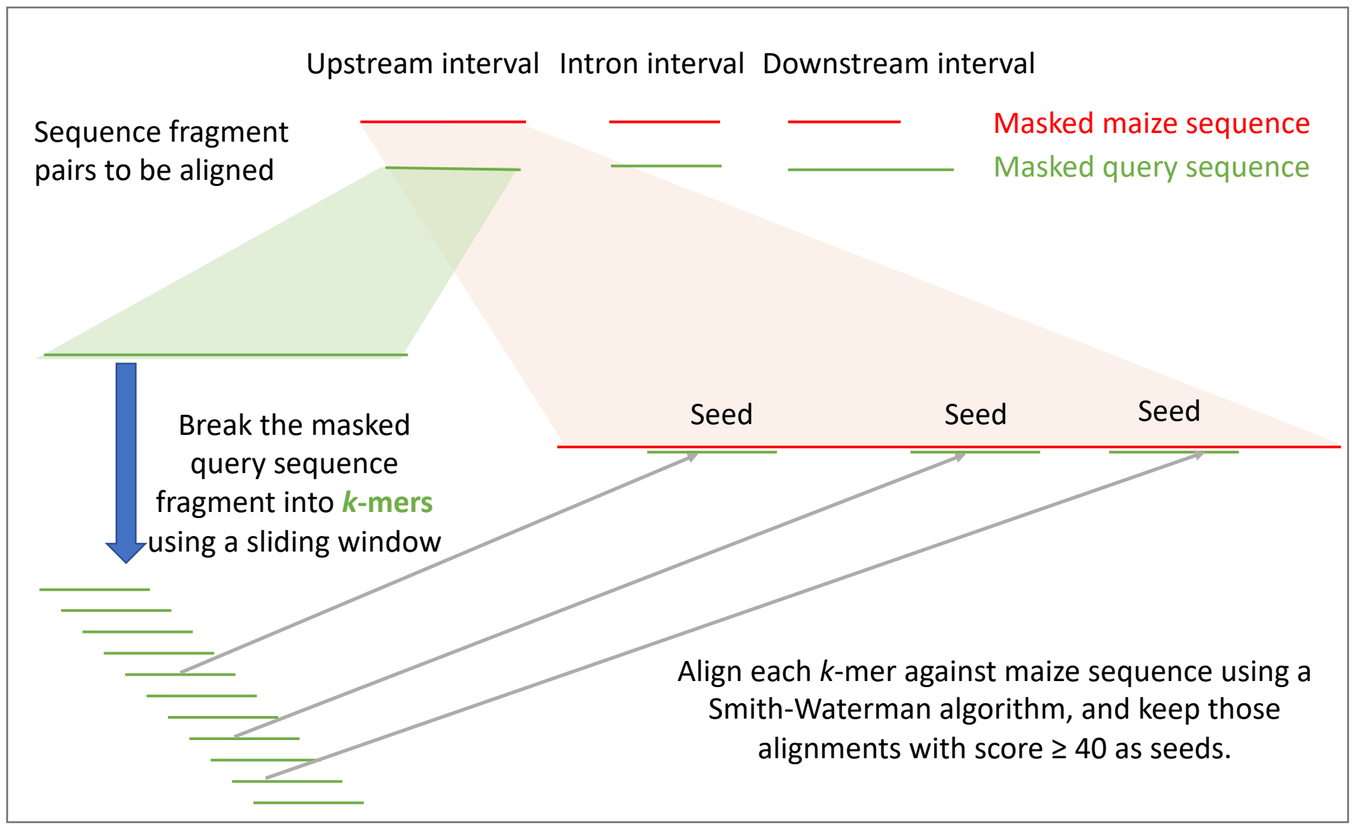
1) Lift over of the maize B73 v4 gene structure annotation to the query genome using minimap2.



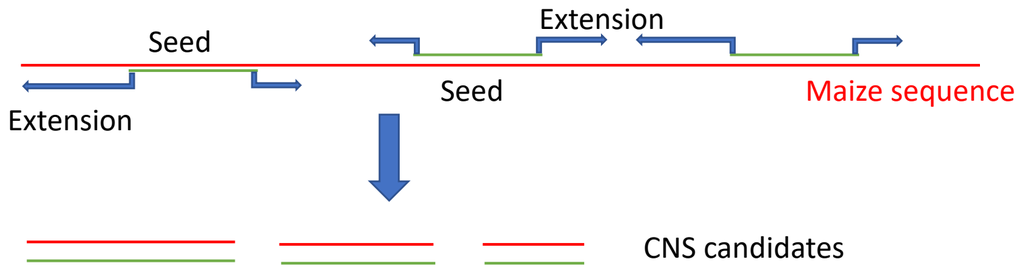
2) Delete CDS and high frequency *k*-mers.



3) Generate sequence alignment seed for each pair of interval fragment.



4) Alignment extension for each seed with a Smith-Waterman algorithm.

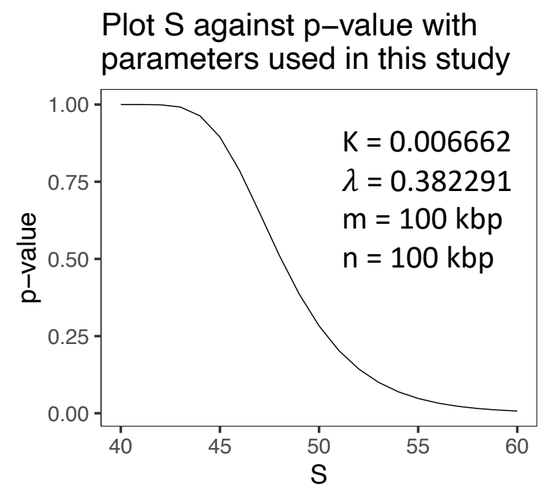


5) Calculate a p-value for each extend sequence alignment, and keep those alignments with p-value smaller than 0.1.

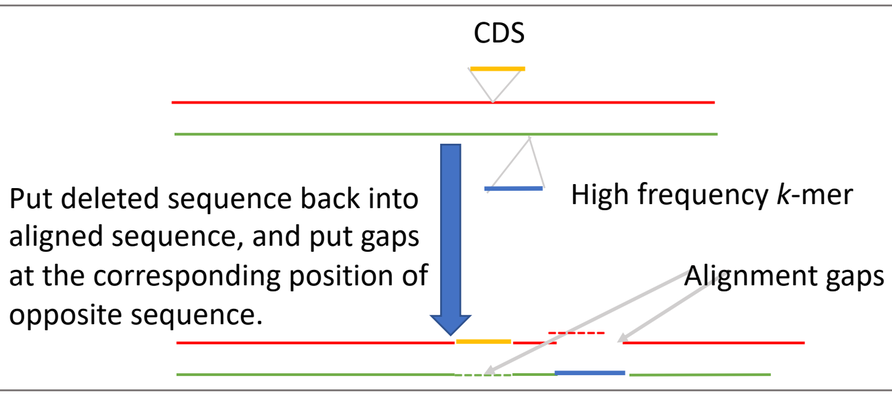
$$E = Kmne^{-\lambda S}$$

$$\text{p-value} = e^{-E}$$

K and λ were calculated using random sequence fragments. m is the length of maize B73 v4 sequence, and n is the length of sorghum sequence. e is the natural base. S is the Smith-Waterman sequence alignment score.



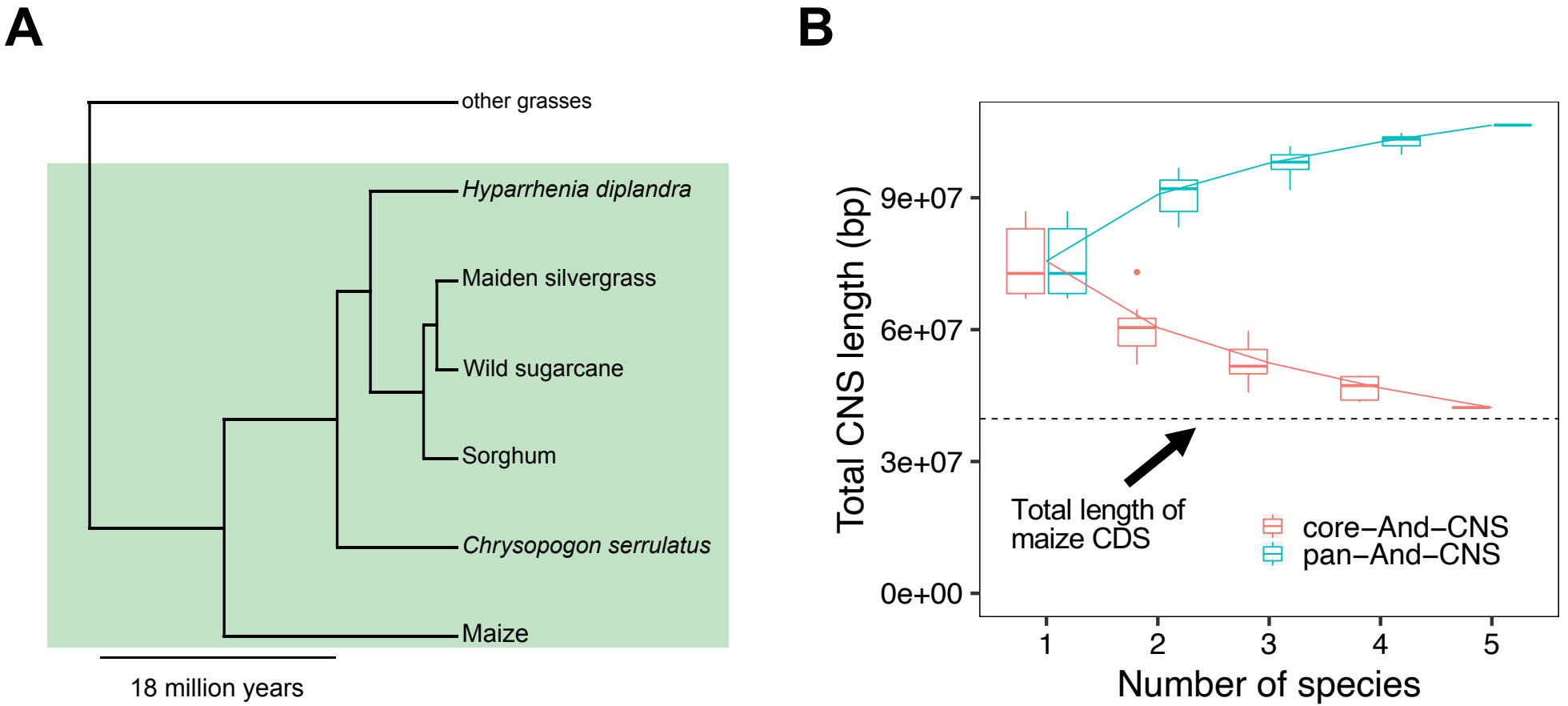
6) Put deleted *k*-mers and CDSs back into significant sequence alignments.

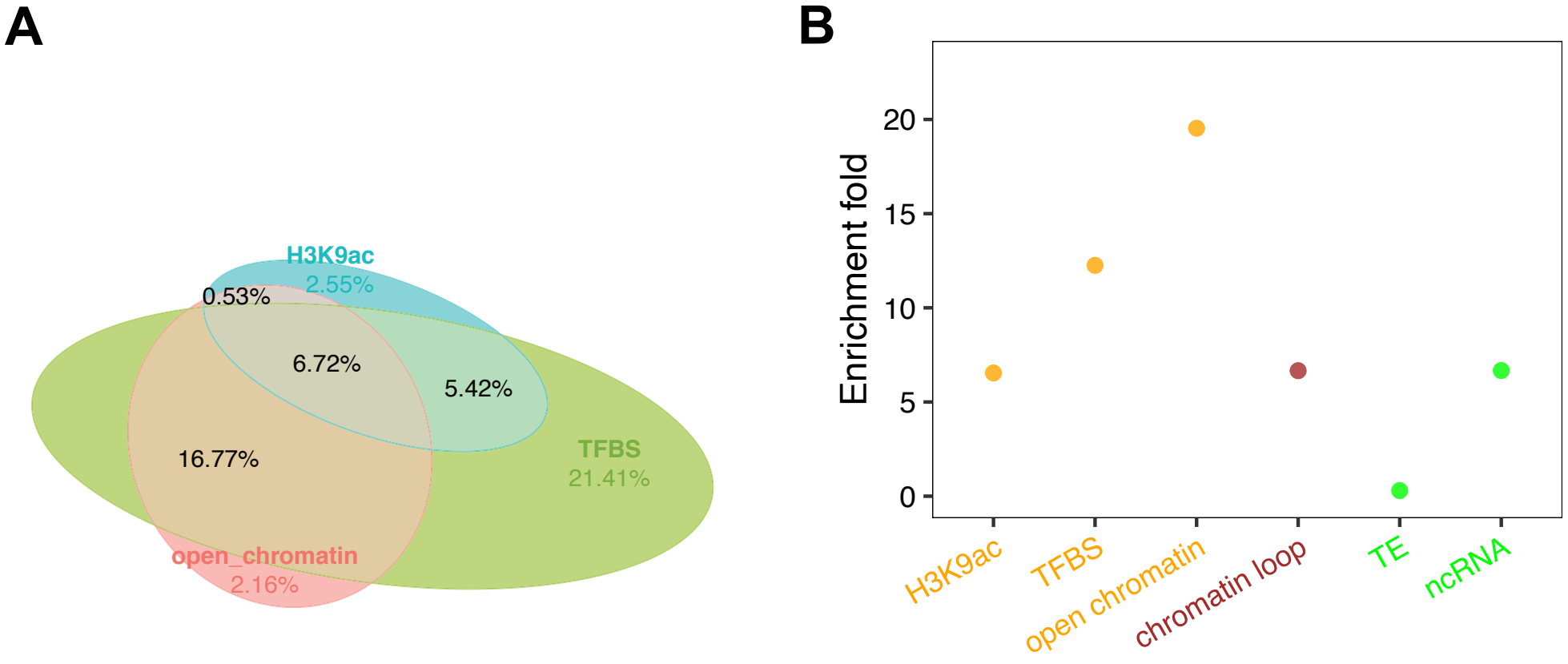


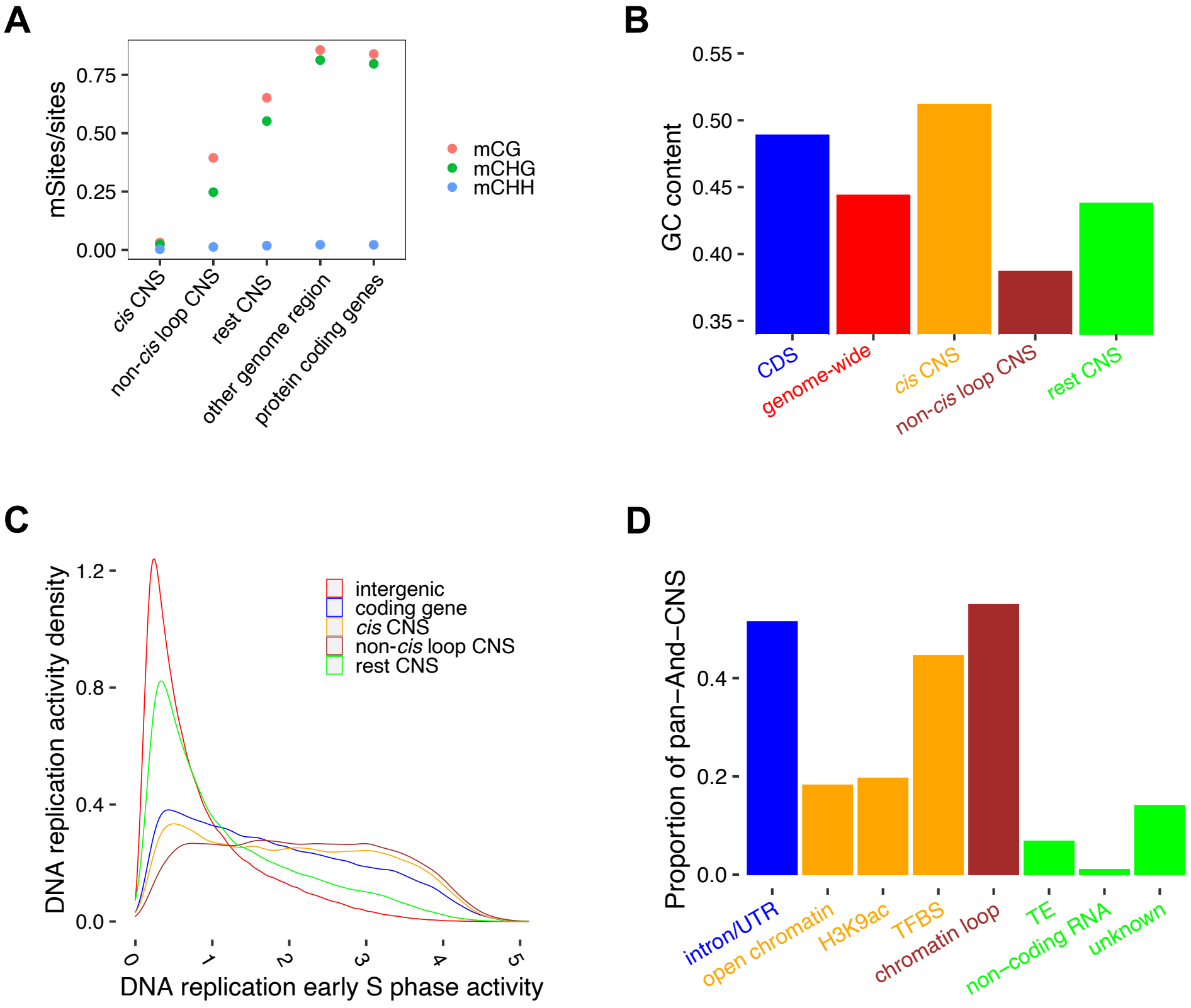
7) Output sequence alignment in SAM format.

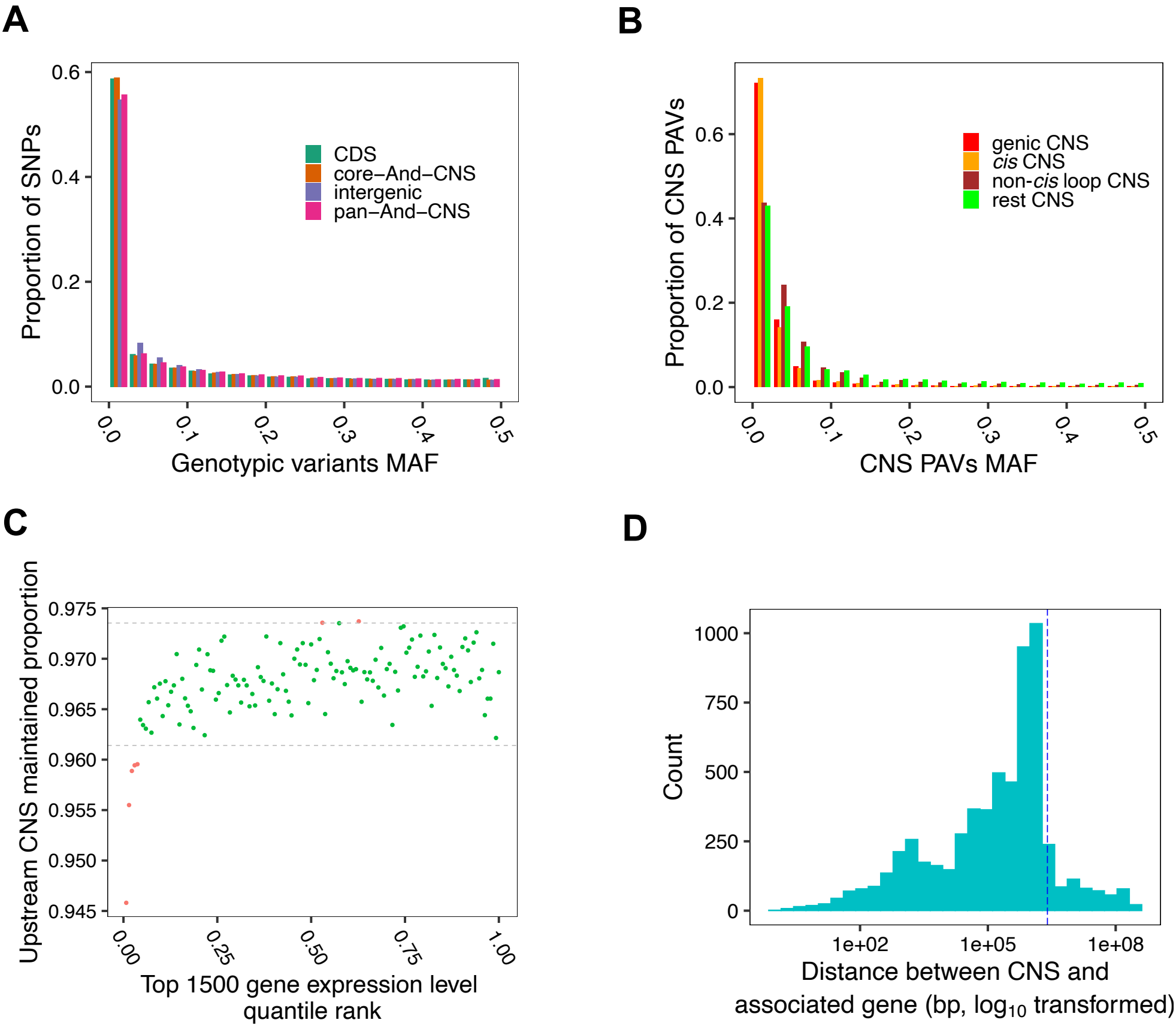
Visualize the result using IGV etc.

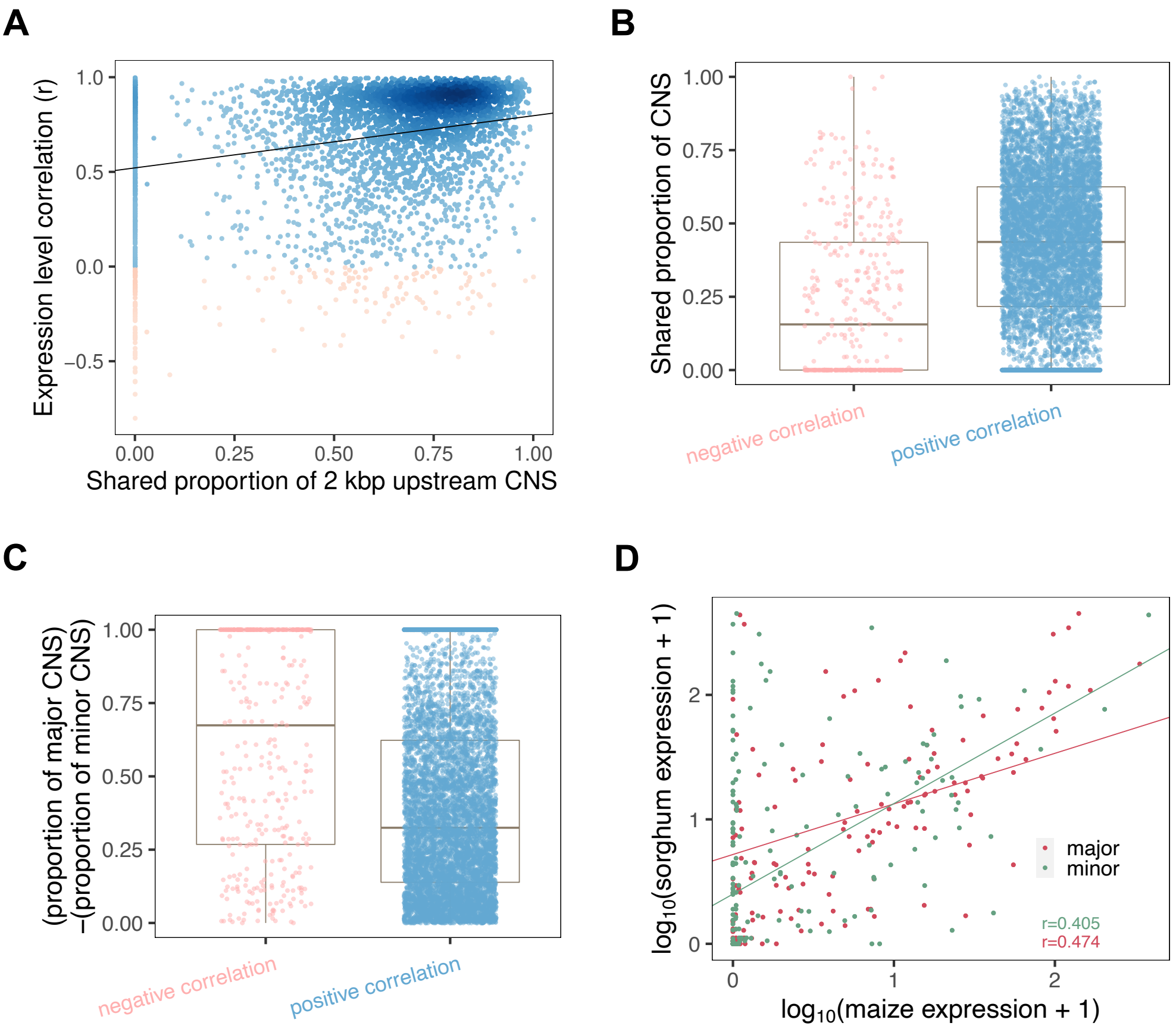
Integrative Genomics Viewer


A
other grasses
Hyparrhenia diplandra
Maiden silvergrass
Wild sugarcane
Sorghum
Chrysopogon serrulatus
Maize
18 million years
B
Total CNS length (bp)
9e+07
6e+07
3e+07
0e+00
Total length of
maize CDS
core–And–CNS
pan–And–CNS
1
2
3
4
5
Number of species


A
H3K9ac
2.55%
0.53%
6.72%
5.42%
16.77%
TFBS
21.41%
open_chromatin
2.16%
B
Enrichment fold
20
15
10
5
0
H3K9ac
TFBS
open chromatin
chromatin loop
TE
ncRNA


A
mSites/sites
0.75
0.50
0.25
0.00
mCG
mCHG
mCHH
cis CNS
non-cis loop CNS
rest CNS
other genome region
protein coding genes
B
GC content
0.55
0.50
0.45
0.40
0.35
CDS
genome-wide
cis CNS
non-cis loop CNS
rest CNS
C
DNA replication activity density
1.2
0.8
0.4
0.0
intergenic
coding gene
cis CNS
non-cis loop CNS
rest CNS
0
1
2
3
4
5
DNA replication early S phase activity
D
Proportion of pan–And–CNS
0.4
0.2
0.0
intron/UTR
open chromatin
H3K9ac
TFBS
chromatin loop
TE
non–coding RNA
unknown


A
Proportion of SNPs
0.6
0.4
0.2
0.0
CDS
core–And–CNS
intergenic
pan–And–CNS
0.0
0.1
0.2
0.3
0.4
0.5
Genotypic variants MAF
B
Proportion of CNS PAVs
genic CNS
cis CNS
non-cis loop CNS
rest CNS
CNS PAVs MAF
C
Upstream CNS maintained proportion
0.975
0.970
0.965
0.960
0.955
0.950
0.945
0.00
0.25
0.50
0.75
1.00
Top 1500 gene expression level quantile rank
D
Count
1000
750
500
250
0
1e+02
1e+05
1e+08
Distance between CNS and associated gene (bp, log10 transformed)


A
Expression level correlation (r)
1.0
0.5
0.0
−0.5
0.00
0.25
0.50
0.75
1.00
Shared proportion of 2 kbp upstream CNS
B
Shared proportion of CNS
1.00
0.75
0.50
0.25
0.00
negative correlation
positive correlation
C
(proportion of major CNS)
−(proportion of minor CNS)
1.00
0.75
0.50
0.25
0.00
negative correlation
positive correlation
D
log10(sorghum expression + 1)
2
1
0
0
1
2
log10(maize expression + 1)
major
minor
r=0.405
r=0.474

Conserved noncoding sequences provide insights into regulatory sequence and loss of gene expression in maize

Baoxing Song, Edward S. Buckler, Hai Wang, et al.

Genome Res. published online May 27, 2021
Access the most recent version at doi:10.1101/gr.266528.120