BMC Genomics

RESEARCH

Open Access

The landscape of regulatory element evolution in a C4 perennial grass

Taslima Haque[1,2*], Govinal Badiger Bhaskara[1,3], Robert J. Schmitz[4] and Thomas E. Juenger[1*]

Abstract

Background Gene regulatory evolution is a well-known source of phenotypic diversity and adaptive evolution. Although *cis*-regulatory elements (CREs) play a vital role in gene expression evolution, the molecular evolution of CREs remains mostly unknown due to the difficulty in identifying and characterizing these functional elements. Comparative genomic analyses of noncoding DNA can be leveraged to identify conserved noncoding sequences (CNS), many of which may harbor functional CREs conserved by purifying selection. However, purely computational inference of CREs from putative CNS can be erroneous due to the complex genomic architecture in plants.

Results One promising experimental approach to identify CREs is by profiling accessible chromatin regions (ACRs) that are often associated with the location of CREs. In this study, we use comparative genomics along with the profiling of ACRs to study the molecular evolution of putative functional noncoding regulatory regions in Panicoid grasses. We identified sets of CNS that varied in relationship to the degree of evolutionary divergence among the studied taxa, including identifying core-Panicoid-CNS. We augmented this analysis by profiling ACRs in *Panicum hallii* ecotypes using ATAC-seq. ACRs had low SNP density at the summit, harbored a high frequency of core-Panicoid-CNS, and were enriched with expression QTL. These data help to annotate the *P. hallii* genome for putative functional elements and suggest that a large proportion of these ACRs are evolving under purifying selection. Turnover in CNS and ACR between ecotypes of *P. hallii* identifies a small set of putatively divergent CREs that may underlie differences in gene regulation between genotypes from inland and coastal habitats.

Conclusions In summary, we profiled ACRs in Panicoid grasses and integrated this data with our putative CNS prediction framework, which provides unique insight into patterns of polymorphism and divergence in CREs in C4 perennial grasses.

Keywords Gene regulatory evolution, Accessible chromatin, ATAC-seq, *Panicum hallii*, Ecotypic differentiation

*Correspondence:
Taslima Haque
taslima@utexas.edu
Thomas E. Juenger
tjuenger@austin.utexas.edu
[1]Department of Integrative Biology, University of Texas at Austin, 2415 Speedway #C0930, Austin, TX 78712, USA
[2]Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, 1105 N University Ave, Ann Arbor, MI 48109, USA
[3]Department of Epigenetics and Molecular Carcinogenesis, University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA
[4]Department of Genetics, University of Georgia, 120 East Green Street, Athens, GA 30602, USA

Introduction

Gene regulatory evolution is the process by which the regulation of gene expression evolves over time. It can result in modifications to the timing, location, and level of gene expression as well as its responsiveness to the environment. It is a major source of genetic variation and is likely a common driver of phenotypic diversification and adaptive evolution [1, 2]. Gene regulatory evolution can occur through changes in *cis*-regulatory elements (CREs) or through *trans*-acting factors that bind to CREs and control gene expression. *Cis*-regulatory evolution is an especially common cause of expression variability within and among species [2, 3]. For example, many expression QTL and allele specific expression studies in plants report a remarkable level of *cis*-genetic variation in expression [4–6]. Despite their clear importance in controlling gene function, we know little about the evolution of plant CREs. This is in part due to the difficulty of identifying CREs. CREs are generally short noncoding DNA elements that contain transcription factor binding sites that are critical for the activation and sustained transcription of genes. Core promoters contain the best known CREs, but a diverse complement of sequence features including transcriptional enhancers, silencers, and insulator elements comprise the *cis*-regulatory modules (CRMs) that ultimately determine patterns of gene expression [7]. The activity states of CRMs are dynamic and correlate with chromatin accessibility, transcription factor binding, and the combinatorial patterns of DNA methylation and histone modification. Outside of a few model plants, there is currently poor information about CREs and little empirical insight into their molecular function or evolution.

One longstanding approach for identifying CREs rests on the idea that their sequence content will be conserved by persistent purifying selection that maintains their important roles in regulating gene expression. Under this hypothesis, putative CREs can be identified by comparing the noncoding DNA sequences of related species to detect regions that have remained largely unchanged over evolutionary time. The growing availability of high-quality plant genome assemblies and annotations afford opportunities for identifying these conserved noncoding sequences (CNS). For example, Song et al. [8] identified core and pan-CNS among six deeply diverged grass genomes (Andropogoneae) that harbored more than 50% of the well-known CREs identified by other empirical methods. However, purely bioinformatic inferences of plant CREs from CNS can be problematic due to the complexity of plant genomes, their high repeat content, challenges caused by polyploidy, or by turnover resulting from CRE divergence. Other tools for identifying CREs center on direct screens for known binding element motifs, often identified from model species, or by using other features that may be indicators of functional CREs. For example, chromatin can be tightly packed or loosely packed, which can affect the ability of transcription factors or regulatory molecules to interact with CREs [9–11]. As such, there is a strong relationship between accessible chromatin states and plant CREs [11–14]. One common avenue for identifying putative CREs is based on the genome-wide characterization of chromatin accessibility. Genome-wide accessible chromatin regions (ACRs) have been identified using high-throughput methods such as Micrococcal Nuclease analysis of chromatin structure (MNase-seq) [15] or Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq) [16, 17]. Studies in plants [10, 14, 18–20] have demonstrated that ACRs tend to harbor more CNS and lower SNP density than the flanking regions or random genomic background. This suggests that the vast majority of ACRs contain functional CREs that are conserved by purifying selection. Therefore, leveraging empirical data such as genome-wide chromatin accessibility profiles or patterns of nucleotide diversity can facilitate the search for putative CREs and ultimately enable studies of their molecular evolution.

The monocot grasses provide a unique opportunity for the study of regulatory sequences. Genomic resources in crops like maize, millet, rice, sugarcane and sorghum have allowed sequence comparisons to identify conserved noncoding DNA and to predict putative CRE motifs. These studies have identified CNS which harbored deeply conserved regulatory elements and also in some cases loss of certain elements in one or more grass lineage [8, 21]. However, one limitation of crop plants is their unique demographic history that includes extreme domestication bottlenecks, strong selective sweeps, and complex breeding histories. Studies of wild grass species may provide a unique opportunity for evaluating the molecular evolution of noncoding regulatory sequences, especially by leveraging the extensive comparative genomic resources of related crops, but affording demographic histories unencumbered by domestication.

We have developed the diploid C4 perennial bunchgrass *Panicum hallii* as a study system for ecology, evolution and plant genomics. Our earlier studies have focused on characterizing the population structure [22, 23], patterns of genetic diversity [6, 24], divergence between ecotypes utilizing two fully assembled and annotated genomes [25], and population-wide transcriptomics [26]. One underlying theme of the research has been to uncover the role of changes in gene expression as a mechanism for tolerating changing environmental conditions, extreme abiotic stresses, and local adaptation. For example, Haque et al. [24] studied *P. hallii* response to salinity in contrasts of coastal and inland adapted genotypes. Lovell et al. [6] characterized the drought-responsive

expression divergence between upland (drought-adapted) and lowland (mesic) ecotypes of *P. hallii* using F1 hybrids and discovered that cis factors contributed to expression divergence in over 47% of genes, 7.2% of which exhibited drought-responsive gene-by-environment interaction. Lovell et al. [25] reported that gene expression networks in a *P. hallii* are dominated by local CREs, with over 50% of transcripts exhibiting cis eQTL underlying expression divergence between ecotypes. Using reciprocal transplant experiments, Bhaskara et al. [26] reported putatively adaptive expression divergence based on outlier tests for several hundred genes with constitutive expression divergence. However molecular evolutionary studies of this expression variation have been limited by a lack of annotation of putative CREs.

In this study, we utilize a comparative genomic approach to identify CNS in several Panicoid grasses. These data were augmented with ATAC-seq analyses to identify ACRs in *P. hallii* with a focus on the identification of shared and unique CREs among inland and coastal adapted ecotypes. These data allow us to evaluate the role of selection in the molecular evolution of putative CREs. Overall, we found ACRs possessed low SNP density and harbored a significantly higher frequency of deeply conserved CNS, suggesting these elements are predominantly under purifying selection. An association of eQTL with ACRs suggested that the presence of functional CRE that might underlie gene expression divergence between locally adapted ecotypes.

Results

CNS in Panicoid grasses

To gain insights into noncoding regulatory sequence evolution in wild grasses, we first aimed to characterize the genomic landscape of CNS with a specific focus on Panicoid grasses. To identify CNS, we used a coding gene based anchored and sensitive alignment pipeline (dCNS) developed by Song et al. [8]. Since earlier reports in plants suggests that most ACRs and CREs are located within the 100 Kb of the nearest gene [10, 14, 27], this pipeline was designed to restrict the search space to 100 Kb up and downstream of intergenic regions of gene models, introns, and UTRs. We implemented this pipeline with the inland and coastal reference genomes of *P. hallii*, the K sub-genome of *Panicum virgatum*, and the *Setaria viridis* genome to identify CNS among Panicoids. The choice of a single sub-genome of *P. virgatum* reduced the complexity due to polyploidy. We chose two recently diverged inland (var. *hallii*) and coastal (var. *filipes*) ecotypes of *P. hallii* [22–25] that represent ~ 1 Mya of divergence. The *Panicum* grasses are C4 perennials native to North America but have different ploidy levels. *P. virgatum* in this study is an allotetraploid whereas *P. hallii* is diploid. The *Panicums* shared a common ancestor ~ 8.3 Mya, whereas the *Panicum* clade and *S. viridis* shared a common ancestor ~ 13.1 Mya [25]. Our pipeline identified CNS that can be categorized by the degree of their conservation across the Panicoid phylogeny. The total CNS length identified using the *P. halli* inland reference genome (HAL2) as the query was negatively correlated with the divergence time for comparison genomes from the reference (Pearson's correlation coefficient, r^2 = -0.84). The greatest shared CNS space (cumulative length of all CNSs) was observed between the two *P. hallli* genomes (140 Mb) whereas the smallest shared space (77 Mb) was with *S. viridis*. We highlight CNS that are shared among all Panicoid genomes (core-*Panicoid*-CNS) and shared among all the Panicoid grasses except *S. viridis* (unique-*Panicum*-CNS). Shared CNS space was 11 Mb (56,512 core-Panicoid-CNS; Fig. 1; Supplementary Table 1) and 8.7 Mb (35,115 unique-*Panicum*-CNS; Supplementary Table 1) among the studied *Panicoid* and *Panicum* grasses, respectively. Together, we identified 91,627 CNS which were conserved among these *Panicoid* or *Panicum* grasses comprising ~ 4% of the *P. hallii* noncoding genome. A much larger sequence space (~ 80 Mb) is conserved between the two *P. hallii* ecotypes which is certainly an overestimate of the conservation of functional elements.

Next, we explored the relative distance of identified CNS to the closest gene model. We observed that approximately 92% and 70% of the core-*Panicoid* CNS resided within the gene model (UTR or Intron) or less than 2 Kb (up or downstream) from a gene (Fig. 1C). Overall, CNS in the core-*Panicoid* class occurred closer to the nearest gene than for unique-*Panicum* CNS (Welch two-sample t-test p-value $<< 0.05$). Much of our *P. hallii* research has focused on divergence between upland and lowland ecotypes. As such, an especially interesting class of predicted CNS are those that are highly conserved across *Panicoid* grasses, but that show turnover since the relatively recent divergence of var *hallii* and *filipes*. In this context, we evaluated a unique set of CNS shared between *S. viridis*, *P. virgatum*, and any one of the *P. hallii* genomes (pan-lost-Coastal-CNS and pan-lost-Inland-CNS). Based on this set, we found 2,565 and 2,539 CNS which exhibited turnover (defined as the loss of homologous blocks of 38 bp; see in method) in either the inland or coastal ecotype respectively (Supplementary Tables 3–4). Approximately 57% and 72% of the pan-lost-Inland and pan-lost-Coastal CNS resided within gene models or were less than 2 Kb from a gene. To understand the relative abundance of CNS in different genomic features, we assigned CNS that resided in either UTR, intron, or proximal to genes (2Kb upstream or downstream from the TSS), or in distal regions to genes (>2Kb upstream or downstream from a gene model) (Supplementary Figure S1). Overall, we observed core-*Panicoid* and

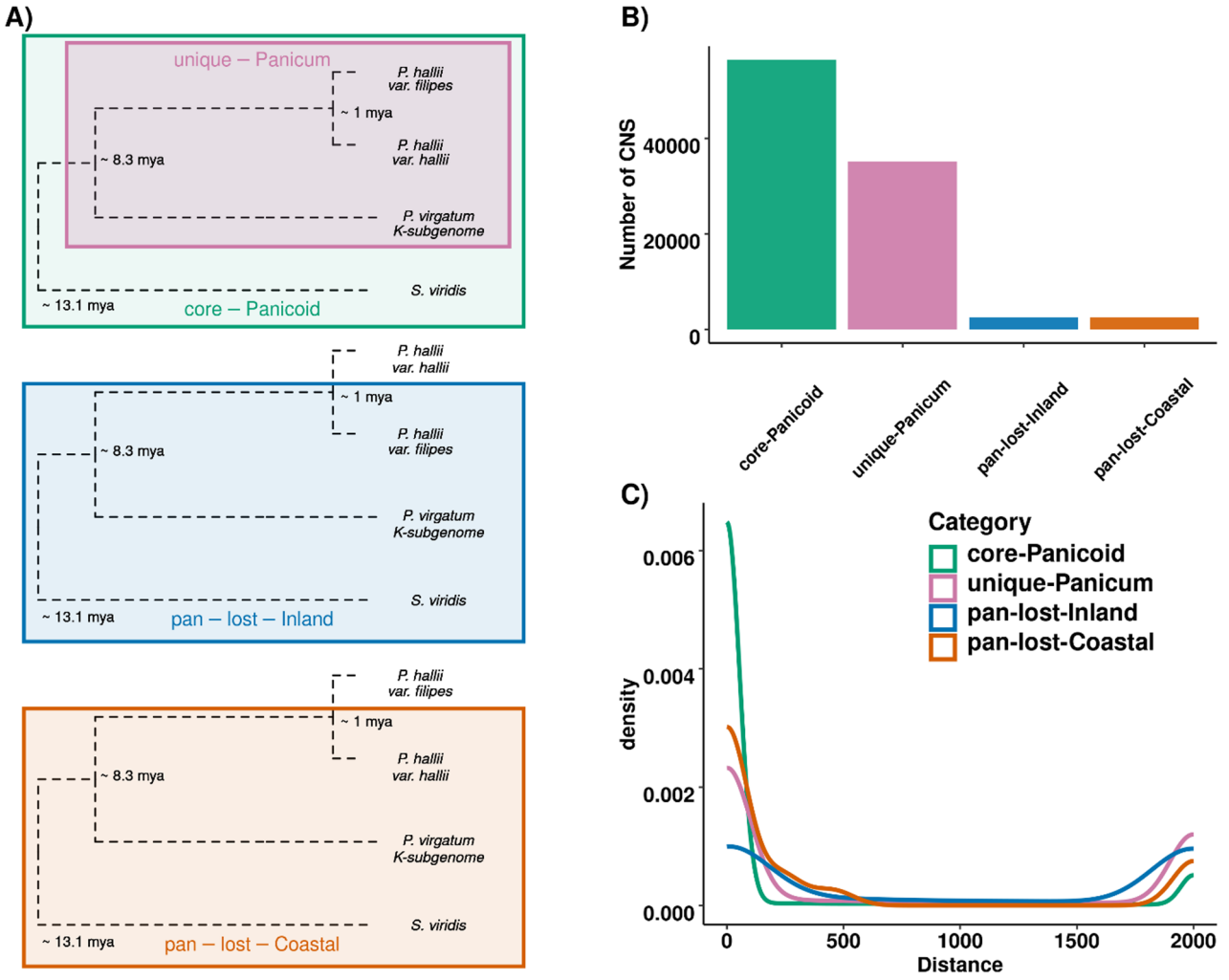

Fig. 1 The genomic landscape of conserved noncoding sequences (CNS) in Panicoid grasses. **A** CNS space shared between Panicoid grass genomes (*P. hallii*, *P. virgatum* (K sub-genome), *Setaria viridis*) was classified as the core-*Panicoid*-CNS. The CNS set shared between the *Panicum* grasses was classified as unique-*Panicum*-CNS. CNS present between *S. viridis*, *P. virgatum* and either of the ecotype of *P. hallii* were classified as either pan-lost-Coastal-CNS or pan-lost-Inland-CNS. **B** Number of CNS identified in different categories and (**C**) Distance (in bp) distribution of CNS to the nearest gene. Note that CNS which had distance more than 2 Kbp from the nearest gene were collapsed into the 2 Kb bin for the purpose of plotting

unique-*Panicum* CNS were highly abundant in intronic regions (~73% and 62% for the core-*Panicoid* and unique-*Panicum* category respectively) whereas turnover CNS had highest abundance in proximal regions (pan-lost-coastal category) or distal regions (pan-lost-inland category). As CNS likely contain Transcription Factor (TF) binding motifs as functional elements, we next asked whether these CNS were enriched with TF-binding motifs. Indeed, we observed a higher frequency of known plant TF-binding motifs from the JASPAR22 database in each category of CNS compared to a random noncoding genomic background (Mann-Whitney U-test p-value < 0.05), similar to results reported in other studies of motif enrichment in plant CNS (Turco, et al. 2013; Song, et al. 2021; Yocca, et al. 2021). We also observed enrichment of TF binding motifs in turnover CNS categories relative to core-*Panicoid* or unique-*Panicum* categories but no difference was observed among two different turnover CNS categories.

Genomic landscape of ACRs

To identify ACRs, we studied leaf tissue of young seedlings from genotypes that are representative of upland and lowland ecotypes of *P. hallii*. We implemented a fluorescence-activated nuclei (FAN) sorting based ATAC-seq protocol [17] developed specifically for plant tissues. We used three biological replicates for each ecotype to identify high-quality conserved ACRs across replicates (mean FRiP score 0.37; range = 0.31–0.41; Supplementary Table S6 for detailed library quality). We identified 18,746 and 28,299 leaf tissue ACRs for the inland and coastal ecotypes, respectively, accounting for approximately 2.5% of

each respective genome. Approximately 44% and 50% of the detected ACRs in the inland and coastal ecotype were found within a gene model (Fig. 2A). The median length of ACRs was larger for the inland ecotype (547 bp) than for the coastal ecotype (502 bp) (Fig. 2B). For both ecotypes, a large fraction of ACRs were observed in genic or the proximal regions of genes (within 2 Kb upstream or downstream of an existing gene model) of genes which is congruent with other plant ACR studies [14, 18, 19, 28] and were classified as genic ACR (gACR) (~ 39%) or proximal ACR (pACR) (~ 23%) respectively (Fig. 2C; Supplementary Table S7-8). However, on average ~ 38% of the ACRs occurred in distal intergenic regions (> 2 Kb from the nearest gene model) and were classified as distal ACR (dACR). ACRs from both the ecotypes were enriched at transcription start sites and exhibited higher GC content compared to proximal up and downstream regions (Fig. 2E-F).

Sequence conservation of ACRs among *P. hallii* population and across Panicoid lineages

The geographic distribution of *P. hallii* spans the mesic coastal habitats of the Gulf of Mexico to the desert southwest of the US and northern Mexico. The major axis of population differentiation corresponds to the taxonomic classification of the coastal and inland ecotypes. However, there is clear secondary population structure occurring within ecotypes that is associated with independent demographic histories and ecoregions [22, 23]. Here, we leverage resequencing data and existing SNP variants [26] to describe molecular evolution in the noncoding regions of the genome. We constrained our population genomics study to the two major genetic clusters of *P. hallii*: 54

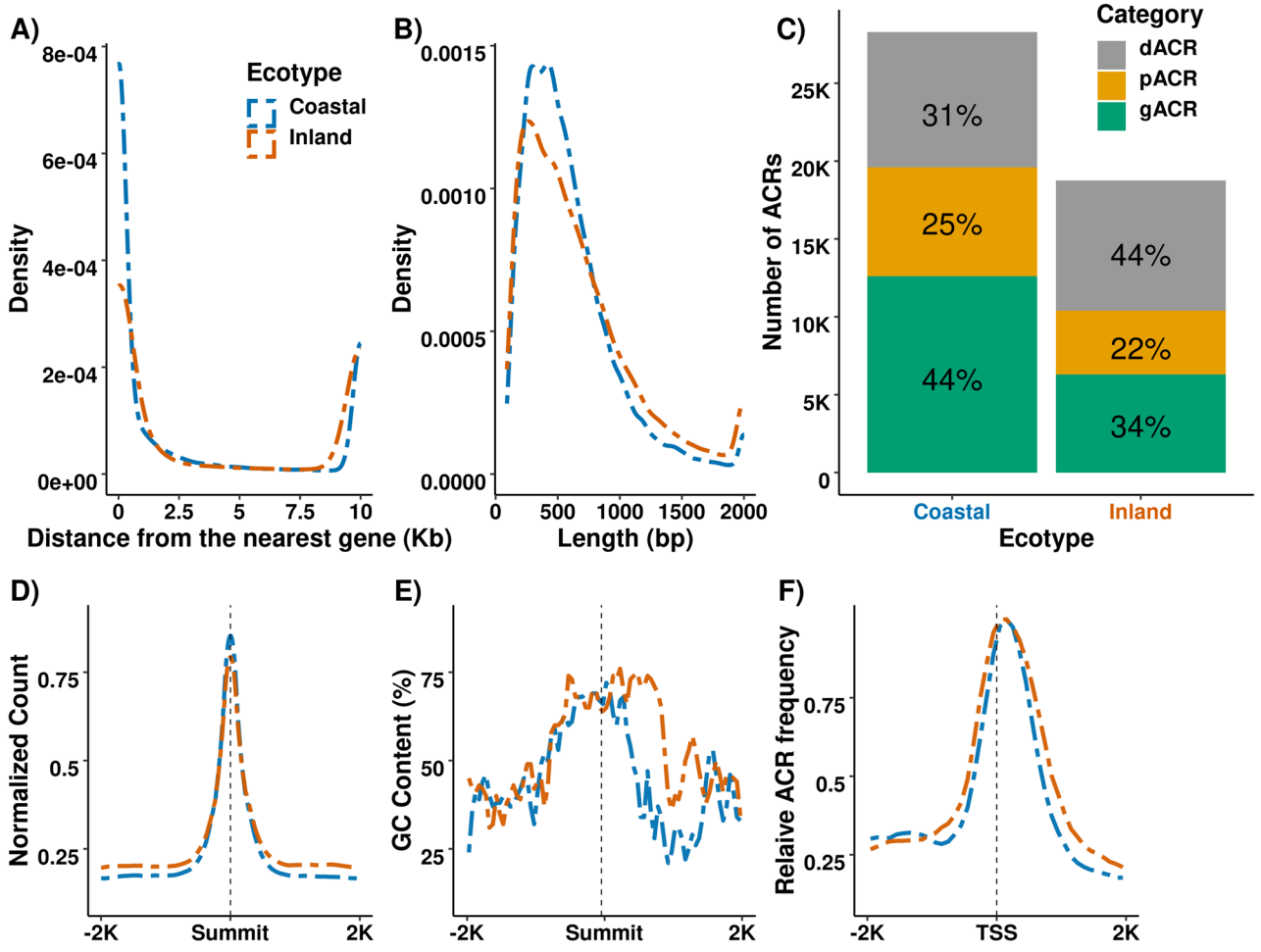

Fig. 2 Genomic landscape of accessible chromatin regions (ACRs) of inland and coastal ecotypes of Panicum hallii. **A** Distribution of the length (bp) to the nearest gene for ACRs. ACRs which resided within a gene model were assigned zero length and ACRs with greater than 10 Kb distance were binned into a single bin. **B** Length distribution of ACRs. ACRs of length greater than or equal to 2 Kb were binned into one single bin (**C**) Number and proportion of ACRs categorized by their distance to the nearest gene. Genic ACR (gACR) reside within a gene model, proximal ACR (pACR) occur within 2 Kb of a gene model and distal ACR (dACR) were more than 2Kb from the nearest gene model. **D-E** Distribution of normalized read count and GC content at ACR summit along with 2 Kb proximal region. **F** Enrichment of ACRs at Transcription Start Site (TSS)

individuals from the inland ecotype and 11 individuals from the coastal ecotype. We profiled the sequence diversity at identified ACRs based on their respective genome references and estimated SNP density (Min-Max normalized number of SNP detected in 200 bp windows) at the ACRs and their flanking genomic regions. We observed lower SNP density at identified ACRs compared to nearby flanking genomic regions for both of the ecotypes (Fig. 3) and for each class of ACR (Supplementary Figure S3). Similar results were reported in both monocot and dicot plant genomes [10, 14, 18, 19] suggesting that the vast majority of CREs evolve under purifying selection.

Earlier plant genome comparisons primarily reported deeply conserved regulatory elements and relatively little turnover of CNS in grass lineages [8, 21]. In this study, we observed a higher frequency of CRE motifs from the JASPAR22 database within ACRs compared to a random genomic background (Mann-Whitney test p-value < 0.05). To further uncover the relationship between CNS with ACR and to evaluate their conservation or evolution, we asked whether ACRs were enriched with CNS relative to a random genomic background. Here, we considered any given ACR which overlapped with at least 25% of a CNS interval (~ 50 bp overlap) to possess that CNS. Globally, ACRs were enriched with unique-*Panicum* and core-Panicoid classes of CNS for both the ecotypes (adjusted p-values for Fisher's exact test < 0.05; Fig. 4; Supplementary Table S9; Supplementary Figure S5) suggesting a conserved relationship. However, we observed no significant enrichment of ACRs with turnover CNS (adjusted p-value > 0.05). This pattern may result from noise in identifying chromatin accessibility across different developmental ages, tissues, cell types or response to different environments, or simply from erroneous inference of CNS. Overall, we observed 8.3% of detected core CNS overlapped with ACRs across the two *P. hallii* ecotypes. This is similar to other studies in Arabidopsis, maize, and rice [14, 18, 29, 30], but contrasts with a depletion in over CNS/ACR overlap observed in *Capsella grandiflora* [19]. As mentioned earlier, different cell types, tissues, developmental ages, or environmental conditions may have variable chromatin accessible profiles which can weaken this relationship. Moreover, CNS detection, which could be influenced by divergence time of a model system, can alter the trend. The absence of ACR enrichment at the turnover CNS can also be attributed to a higher false positive rate due to absence of replicates at the ecotypic level.

Next, to understand this relationship in the context of genomic features, we decomposed ACRs by their position relative to gene models and tested for CNS enrichment. gACRs (Odds-ratio 1.6, 1.1 for core-*Panicoid*-CNS in inland and coastal ecotype respectively; 2.5 and 1.4 for unique-*Panicum*-CNS; Fisher's exact test adjusted p-value < 0.05) and pACR (Odds-ratio 15.5 and 13.6 for core-*Panicoid*-CNS and 8.2 and 6.0 for unique-*Panicum*-CNS; adjusted p-value < 0.05) were enriched with unique-*Panicum*, and core-*Panicoid* CNS classes. However, dACRs were depleted in the inland ecotype for core-*Panicoid* CNS (Odds-ratio 0.4; adjusted p-value < 0.05) and no significant enrichment was found for unique-*Panicum*-CNS. On the contrary, in the coastal ecotype

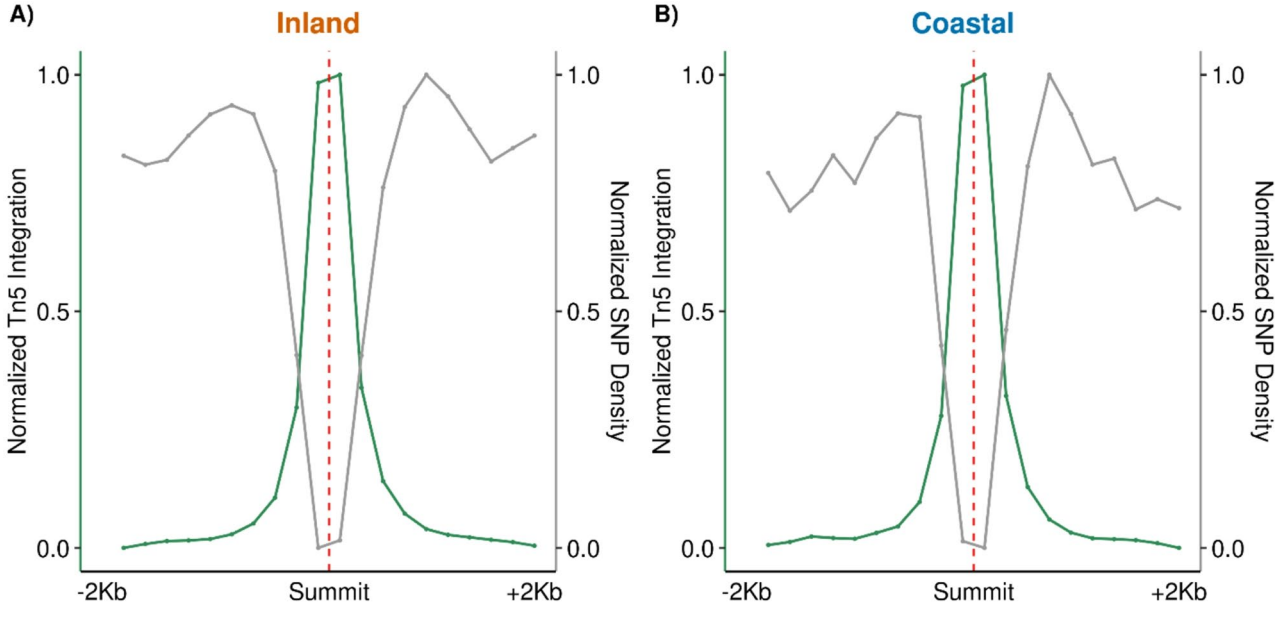

Fig. 3 Sequence diversity at ACR and their proximal regions: **A-B** Metaplot of coastal and inland genetic clusters respectively. Tn5 integration sites and the SNP density from accession resequencing were estimated in 200 bp bins for the upstream and downstream genomic regions from the summit of ACRs. Normalized Tn5 integration sites (green line) and mean SNP density (gray line) of each were plotted with the summit at the center (red dotted line)

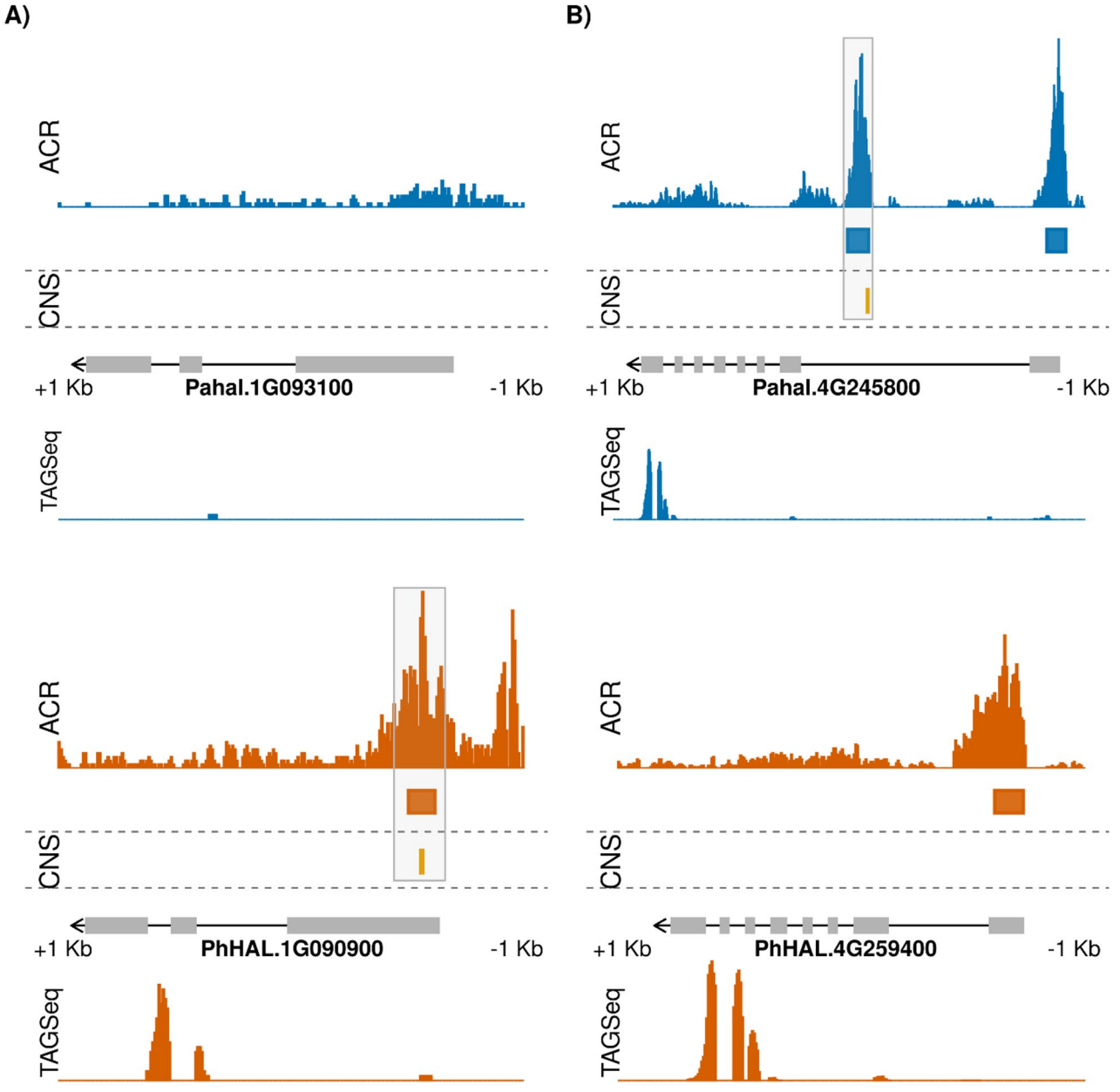

Fig. 4 Example genome browser tracks are shown for turnover CNS for two orthologous gene pairs. **A** Depiction of the loss of both a CNS and ACR for the coastal ecotype (pan-loss-Coastal-CNS) for PhHAL.1G090900 compared to its syntenic orthologue in the inland ecotype (CNS Chr01:5977681-5977718). **B** Depiction of the loss of both a CNS and ACR in the inland ecotype (pan-loss-Inland-CNS) for PhHAL.4G259400 compared to its syntenic orthologue in the coastal ecotype (CNS Chr04:4444902-44449020). Genomic coordinates were plotted relative to each genome reference. Tn5 integration is shown surrounding the orthologous genes (-1 Kb upstream to +1 Kb downstream) with color (blue= Coastal ecotype; red=Inland ecotype). Turnover CNS were plotted as orange rectangles and ACRs were plotted as blue or red rectangles for the coastal and inland ecotype, respectively

dACR were enriched with both core-*Panicoid*-CNS and unique-*Panicum*-CNS (Odds-ratio 1.7 and 2.1; adjusted p-value < 0.05). Notably, dACRs had relatively few CNS compared to other ACR categories which may have led to low power to detect enrichment. Nevertheless, enrichment of ACR with the core-*Panicoid* and unique-*Panicum* classes of CNS for gACR and pACR categories suggests the identification of functional elements proximal to genic regions that evolve primarily under strong purifying selection.

Relationship of chromatin accessibility with expression QTL

ACRs which harbor functional CREs might lead to gene expression divergence between different ecotypes of *P. hallii*. To test this idea, we leveraged results from an earlier eQTL study among the coastal and inland reference

genomes [25]. We asked whether we observe enrichment of eQTL genes with genes that harbor ACRs identified in the present study. Indeed, we detected enrichment of eQTL in ACR (Odds ratio = 1.2, Fisher's exact test adjusted p-value < 0.05; Supplementary Table S10; Supplementary Figure S6). Subsequently, we decomposed eQTL by *cis* and *trans* architecture and observed an excess of *cis*-eQTL (Odds ratio = 1.1, Fisher's exact test p-value < 0.05; Supplementary Table S10) but no enrichment of *trans*-eQTL. Next, we tested for the frequency of ACR occurrence near eQTL (within 2 Kb upstream or downstream) with background genes (genes which were not detected as eQTL and had one-to-one orthologous relationship) and found mean ACR frequency was 1.2 (3,375/2,923) and 1.2 (7,928/6,373) times higher for eQTL in both the inland and coastal ecotype genomes (Welch two-sample t-test p-value < 0.05). Overall, the enrichment of loci involved in gene expression variation in ACRs provides evidence that the identified ACRs harboring CREs that may be functionally divergent. Similar patterns have been reported in other plant species such as maize [14] and *C. grandiflora* [19] suggesting that variation in accessibility or sequence variation in ACRs can contribute to expression variation.

Conservation and turnover of CRE motifs and gene expression evolution

The presence and absence of CNS across the Panicoid phylogeny is indicative of motif evolution, either by deletion and motif loss or sequence divergence to particular motifs. We were especially interested in relatively recent gene expression divergence between ecotypes of *P. hallii*: the groups of pan-*Panicoid* CNS which were present in both *S. viridis* and *P. virgatum*, but were only present in either the coastal ecotype (pan-lost-Inland) or inland ecotype (pan-lost-Coastal) genomes. Turnover in CNS could lead to expression divergence due to loss or gain of CREs. Consistent with this idea, we observed genes with previously identified *cis*-expression divergence (*cis*-eQTL) had marginally higher occurrence of having at least a turnover CNS (loss of homologous block only at either coastal or inland ecotype) at the proximal region (1 Kb up and downstream) compared to a background of expressed genes (Mann-Whitney U-test p-value = 0.057). We detected 113 and 121 ACRs with CNS lost in inland and coastal *P. hallii* respectively. These ACRs harbor putative candidates of noncoding regulatory elements which could lead to expression divergence. Among these, 63 and 133 genes also showed *cis*-expression divergence in an earlier study [25]. As an example, PhHAL.1G090900, a F-box-like protein which lost a CNS in the coastal ecotype at the 5′-UTR region, exhibited chromatin accessibility at that region only in the inland ecotype, and had higher expression in the inland ecotype compared to the coastal ecotype (log2FoldChange = 4.73, q-value < 0.05; Fig. 4A). This CNS contained several putative MYB (MA1767.1; Chr01:5977707–59777717) and ARF (MA1696.1; Chr01:5977695–59777706) regulatory motifs which have been widely known for their regulatory role in water stress. Lovell et al. [25] reported a *cis*-eQTL for this gene that significantly interacted with a drought treatment. It is plausible that the coastal ecotype might have lost the capacity to respond to drought due to the loss of this CNS while adapting to a more mesic habitat. Similarly, a multidrug and toxic compound extrusion (MATE) efflux family protein (PhHAL.4G259400) lost a CNS at the first intron in the inland ecotype, is inaccessible and overexpressed inland ecotype (log2FoldChange = 0.7, q-value < 0.05; Fig. 4B). This CNS in the coastal ecotype had several putative stress or Abscisic acid (ABA) induced motifs including MYBs (MA2024.1; Chr4:4444904–4444916) and RAX3 (MA0576.1; Chr4:4444905–4444917) which could lead to expression divergence between ecotypes. Integrating CNS inference with empirical chromatin accessibility information can facilitate identifying putatively divergent regulatory elements which may underlie gene expression divergence between ecotypes.

Discussion

To understand the molecular evolution of CREs in native plant species, we complemented a comparative genomics analysis with empirical accessible chromatin profiling to identify shared and unique CREs in locally adapted *P. hallii* ecotypes. Broadly, we observed that ACRs harbor a significantly higher frequency of deeply conserved CNS and have lower SNP density. Both of these results suggest that these non-coding regulatory elements are predominantly evolving under purifying selection. This integrative framework opens up possible avenues for studying regulatory element evolution in wild grasses, especially in the context of gene expression divergence leading to adaptation.

We investigated the evolution of CREs across Panicoid grasses. Our analyses highlight the general pattern of conservation of CREs across deep evolutionary time. We identified 56,512 core-*Panicoid*-CNS and 35,115 unique-*Panicum*-CNS. Overall, we estimate that ~ 4% (~ 20 Mb) of the *P.hallii* noncoding genome is conserved among *Panicum* or *Panicoid* species. For comparison, Haudry et al. [31] identified over 90,000 CNS among nine deeply diverged (~ 24 mya) Brassicaceae genomes and inferred that ~ 4% of the noncoding Arabidopsis genome is under purifying selection based on the conservation. Song et al. [8] reported ~ 5% of the maize reference genome is conserved with any one of the five Andropogoneae genomes in their study. In a study of deeper divergence (~ 50 mya), Turco, et al. [21] identified around 15,000 CNS across

the five grass species. Importantly, the number and total space of predicted CNS can vary between study systems due to the variation in divergence time queried as well as the differing filtering and algorithms used to detect and characterize the conserved sequence. In our study, we found a much larger sequence space (~ 80 Mb) is conserved between the two *P. hallii* ecotypes. This is certainly an overestimate of the conservation of functional regulatory elements due to the relatively short divergence time and limited sequence evolution between the upland and lowland ecotypes of this species. One limitation in our study of CNS among *Panicoid* grasses is that the phylogeny was sparse and only a limited number of genomes were compared. Incorporating more species can further fill the gap and provide higher resolution for lineage specific CNS and better power in detected CNS evolution and turnover.

Accessible chromatin regions often harbor CREs and can provide empirical evidence of functional CREs at the regulatory regions. We observed lower SNP density at identified ACRs compared to nearby flanking genomic regions for both of the ecotypes. This result is similar to the earlier studies in both monocot (rice, maize, sorghum) and dicot (Arabidopsis, Capsella) plant genomes [10, 14, 18, 19] showing that ACRs exhibit lower SNP density compared to the nearby non-coding regions. This pattern is commonly interpreted as evidence that the vast majority of the CREs evolve under purifying selection. We identified 56,512 core-*Panicoid*-CNS with 2,573 supported by empirical ACR data in *P. hallii*. Enrichment of high-quality predicted CREs suggest strong purifying selection conserving the function of *cis* element modules. Our study also allows novel contrasts across the relatively shallow divergence between two ecotypes of *P. hallii* and identified over 5,000 CNS which exhibited turnover in either of these ecotypes. Enrichment of turnover CNS near eQTL further corroborated their plausible role in gene expression divergence. Empirical study identified 28,299 and 18,746 ACRs in the leaf tissue of inland and coastal ecotypes which comprises a very small fraction of the total genome harboring CREs. Overall, these leaf ACRs exhibited lower SNP density, enrichment of eQTL [25] and shared homology with CNS identified across *Panicoid* grass. Earlier studies in maize, Arabidopsis, and rice reported similar patterns of CNS enrichment at ACRs [14, 18, 29, 30]. However, Horvath et al. [19] reported a depletion of CNS at ACR in *Capsella grandiflora* and suggested rapid turnover of CREs or a faster mutation rate at noncoding sites in this species. Taken together, these results support a predominance of purifying selection on these noncoding sequences and provides an effective integrative framework to study gene expression divergence and *cis*-regulatory module evolution.

A major goal in plant genomics has been to uncover the role of changes in gene expression as a mechanism for tolerating changing environmental conditions, extreme abiotic stresses, and local adaptation. Coupling CNS, ACR, and gene expression data may provide one avenue for generating and testing hypotheses related to abiotic stress tolerance. For instance, we found that PhHAL.1G090900, a F-box-like protein gene is expressed highly in the inland ecotype compared to the coastal ecotype. This gene lost a CNS in the coastal ecotype at the 5′-UTR region and has the TSS adjacent chromatin region accessible only in the inland ecotype (Fig. 4A). This CNS contains several putative regulatory motifs such as MYB and ARF which are well-known for their regulatory role in water stress. Studies in Arabidopsis reported a F-box protein involved in drought response by modulating ABA-mediated growth inhibition [32]. The inland ecotype of *P. hallii* is widely distributed in dry xeric habitat whereas the coastal ecotype is mostly restricted to more mesic habitat. Earlier studies [6, 25] reported widespread drought-responsive gene expression divergence between these two ecotypes suggesting expression divergence as a predominant mechanism of local adaptation. One of these studies [25]. reported a *cis*-eQTL for this PhHAL.1G090900 gene that demonstrated significant interaction with a drought treatment. The inland allele was downregulated in response to water availability compared to the coastal allele. It is plausible that the coastal ecotype might have lost the CNS over time and the capacity to trigger a plastic response while adapting to more mesic habitats. Likewise, we also report a multidrug and toxic compound extrusion (MATE) efflux family protein (PhHAL.4G259400) as another example of a gene which lost a CNS in the inland ecotype and is mostly inaccessible at that region (Fig. 4B). Several putative stress or Abscisic acid (ABA) induced motifs including MYBs and RAX3 were identified in that pan-lost-Inland CNS and could lead to expression divergence between these two ecotypes. Studies in plants reported this family protein to be involved in various abiotic stresses such as drought and salinity [33]. Lovell et al. [25]. found this gene to be highly expressed in leaf tissue in the inland ecotype relative to the coastal ecotype. In contrast, Haque et al. [24] found the opposite pattern in root tissue suggesting a tissue-specific expression divergence. However, in this study, we specifically profiled chromatin accessibility in the leaf tissue and so are unable to study tissue-specific expression divergence. Incorporating chromatin accessibility data from other tissues can address this expression divergence and the underlying CREs leading to this variation. Broadly, integrating CNS inference with empirical chromatin accessibility information can facilitate identifying putatively divergent regulatory elements which may underlie gene

expression divergence between ecotypes. These genes are strong candidates for functional manipulation to test adaptive hypotheses concerning gene regulatory evolution and abiotic stress tolerance.

As we mentioned earlier, one limitation of our chromatin accessibility study is that it focused on a single tissue type and developmental stage. As such, we are unable to study dynamic aspects of chromatin accessibility which can vary between cell type, tissue, developmental ages, or environmental conditions. Including more data from different tissues, cell types, developmental stages, or environments will provide a clearer picture of the dynamic nature of chromatin accessibility and its role in evolution in wild species. Our study is also limited by contrasting two divergence *P. hallii* genomes. Expanding contrasts to a growing number of pan-genome resources in wild grasses will allow a more careful characterization of functional non-coding regulatory elements leading to expression divergence.

Conclusions

This study is among the first few which evaluate the molecular evolution of noncoding CREs in wild plant species. By implementing an integrative framework inferring putative CNS along with empirical data of ACRs, we reveal that most CREs are under purifying selection. The occurrence of turnover CNS at eQTL intervals may point towards divergent selection on these elements and may underlie divergence between inland and coastal ecotypes.

Materials and methods

Identification of conserved CNSs in Panicoid grasses

To identify CNSs in Panicoid grasses we use the two reference genomes of *P. hallii* along with the K-sub-genome *P. virgatum* (v5.1: https://data.jgi.doe.g.ov/refine-download/phytozome?genome_id=516) and *S. viridis* (v4.1: https://data.jgi.doe.g.ov/refine-download/phytozome?genome_id=726). Only a single sub-genome of *P. virgatum* was chosen for reducing complexity associated with polyploidy. We implemented a coding gene based anchored and sensitive alignment pipeline developed by Song et al. [8]. The detail description of the pipeline can be found at https://github.com/baoxingsong/dCNS. We used the inland *P. hallii* ecotype genome as the base reference. Briefly, this pipeline first maps the protein coding genes for each query genome against the base reference separately using minimap2 [34]. All CDSs and high-frequency *k-mers* were masked from the genomes and only introns, UTRs, and noncoding sequences of 100 kbp upstream and downstream were extracted to perform pairwise alignment. Later, the query sequences were split into shorter fragments with overlapping sliding windows (w = 38 bp and s = 8 bp) and the Smith-Waterman algorithm [35] was employed to align these short fragments against the reference sequence. Alignment were subsequently filtered by alignment score (> = 40), the seed was then extended, and extension was terminated using the X-drop approach [36] and alignments with a dynamic programming score ≥ 54 were defined as CNSs (Supplementary Table S1-4). Later we categorized CNS which were shared among all *Pancoid* genomes (core-*Panicoid*-CNS), shared among the three *Panicum* genomes but not in *S. viridis* (unique-*Panicum*-CNS), and shared between *S. viridis*, *P. virgatum*, and any one of the *P. hallii* genomes (pan-lost-Coastal-CNS and pan-lost-Inland-CNS respectively). To identify the pan-lost-Inland-CNS, we used the same approach as mentioned above with coastal *P. hallii* ecotype genome as the base reference.

Plant material and growth conditions

P. hallii is a perennial C4 bunchgrass that occurs across much of the southwestern USA and northern Mexico [22, 23]. The greatest morphological diversity within *P. hallii* occurs between two well recognized ecotypes that are adapted to southwestern upland (*P. hallii* var. *hallii*) and coastal lowland (*P. hallii* var. *filipes*) habitats. Here, we studied an inbred genotype representative of the upland (*hallii)* ecotype (HAL2) and lowland *(filipes)* ecotype (FIL2). The upland genotype was collected and gifted from the Lady Bird Johnson Wildflower Center (Austin, TX, USA; 30.19°N, 97.87°W) and the lowland genotype was collected from a natural area associated with the Corpus Christi Botanical Garden in South Texas (27.65°N, 97.40°W) and gifted from the USDA E. Kika de la Garza Plant Materials Center in Kingsville TX. *P. hallii* is native to North American and not considered threatened or endangered. These genotypes have been extensively studied and are the representative genotypes that have been used for genome assembly and ecotype characterization [25]. Our experiment was arrayed in a randomized block design with four replicates in three temporal blocks (n = 3 blocks x 4 plants x 2 inbred genotypes). Since library preparation is a very time sensitive technique, blocks were separated in time by splitting the experiment into 3 cohorts in order to minimize the number of ATAC-Seq libraries per cohort. Seeds of inland (var. *hallii;* HAL2 genotype) and coastal (var. *filipe;* FIL2 genotype) ecotypes of *P. hallii* were stratified and germinated in sand for each temporal block separately. Seedlings were transferred to individual 4-inch pots filled with potting media (6:1:1 homogenous mixture of Promix: Turface: Profile) on the fourth day after germination. Plants were grown in a controlled growth chamber at 28/25°C for 12 h/12hr day/night cycle and pots were arrayed in a randomized design with four replicates in each cohort. One replicate from each ecotype from a given cohort was randomly chosen for library preparation. Leaves of a whole seedling

were harvested after two weeks to prepare TAGSeq and corresponding ATAC-Seq library for that given replicate.

TAGSeq library construction

Tissue samples were flash-frozen in liquid nitrogen, stored at -80 °C and later were ground into a fine powder. RNA was extracted using TRIzol Reagent (Thermo Fisher Scientific; Cat # 15596018) following the manufacturer's instructions. TAGSeq libraries were prepared from 1 µg of total RNA following the protocol described by Weng et al. [37]. These strand-specific libraries were sequences at the Genome Sequencing and Analysis Facilities (GSAF) at the University of Texas at Austin on the Illumina NovaSeq Platform as 1 × 100 bp single-end reads for approximately 7 million reads per sample.

ATAC-Seq library preparation

ATAC-Seq libraries were prepared as described by Lu et al. [17]. In brief, approximately 300 mg of freshly harvested leaves were finely chopped with a razor blade in 1 mL of prechilled lysis buffer (Supplementary method) and filtered through Miracloth and a 40-um filter. This crude extract was stained with 4,6-diamidino-2-phenylindole (DAPI) and loaded into a flow cytometer (Sony MA900) to sort nuclei. Sorted nuclei were further washed and treated with 2 uL Tn5 transposase in 40 uL tagmentation buffer (detailed in supplementary method) at 37 °C for 30 min. The tagmentation products were subsequently purified with a NEB Monarch DNA Cleanup Kit and then amplified using Phusion DNA polymerase for 10–12 cycles. Finally, the amplified products were purified using AMPure beads following manufacturer's instructions. Libraries were sequenced at the Genome Sequencing Centre at the HudsonAlpha Institute of Biotechnology using Illumina NovaSeq Platform as 2 × 150 bp reads for approximately 15 million paired-reads per sample.

TAGSeq data processing and gene expression analysis

The sequence quality of all TAGSeq libraries were first evaluated by FastQC v 0.11.5 [38] for various quality metrices to optimize quality filtering parameters of raw sequences. The first 9 bases (including degenerate bases) from the sequence adapter and homopolymers (polyA or polyT) for base count of > = 15mer were removed using Cutadapt v3.7 [39]. Reads with average sequencing quality of > = 20 and a minimum sequence length of 70 bp were retained. Filtered reads were subsequently mapped to corresponding inland and coastal *P. hallii* reference genome (var. *hallii* HAL2 reference genome v2.2: https://data.jgi.doe.g.ov/refine-download/phytozome?genome_id=591/ and var. *filipes* FIL2 reference genome v3.2: https://data.jgi.doe.g.ov/refine-download/phytozome?genome_id=495/) using BWA-mem algorithm [40], filtered by SAMtools v1.10 [41] with mapping quality > = 10, and duplicated alignments were marked by Picard v2.11(http://broadinstitute.g.ithub.io/picard/). Finally, Feature Counts [42] was used to generate an expression count matrix from first strand-specific filtered alignments with fragment length > = 70 bp (--readExtension3 0 –readExtention5 0 -d 50 -s 1) against the existing annotated gene models of the corresponding reference (Supplementary Table S5). Count data was normalized by size factors, transformed using the fitted dispersion-mean relationship (VST method) and later differentially expressed gene were detected by fitting a factorial model for the effect of ecotype in the DESeq2 package [43] in R environment.

ATAC-seq data processing and ACR identification

To identify high-quality ACR, we utilized a conservative approach focused on discovering congruent ACRs identified in at least two biological replicates and their pooled analysis. For each library, raw reads were trimmed using Trimmomatic v0.39 [44] for NexteraPE (2:30:10) based on clipping leading and trailing bases below quality thresholds (LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15). Reads with a minimum length of 36 bp with proper pairs were retained. The reference genomes of the two study ecotypes were masked for non-mappable regions and known repeats. To identify mappable regions, the target genome was simulated by in silico fragmentation into 100 bp fragments which were mapped to reference. Only uniquely mapped regions were considered as mappable. Trimmed reads of each ATAC-Seq library were mapped against the corresponding masked reference genome using bowtie2 v2.4.4 [45] (-X 1000 –very-sensitive –no-mixed –no-discordant), filtered for mapping quality or mapped at organelle genomes and sorted by coordinate using SAMtools v1.13 [41]. Clonal duplicates were removed by Picard v2.25 (http://broadinstitute.g.ithub.io/picard/). Non-duplicated alignments were converted to BED tags using bedtools v2.29.1 [46] and ACR were identified using MACS2 [47] for each library of each ecotype (--keep-dup-all –min-length 90 -q 0.05) using genomic DNA ATAC-Seq libraries as background control (Supplementary Table S6: Quality statistics; Supplementary Figure S3). On average 25,142 and 35,431 ACRs were detected for inland and coastal ecotypes (Supplementary Table S6). We observed ~ 70% of the ACRs overlapped between a pair of biological replicates when each library was analyzed individually. To leverage the shared information across replicates we implemented a conservative approach to identify high-quality ACRs that were detected at least in two replicates. First, we identified ACRs using MACS2 for pooled replicates of each ecotype using the same parameters mentioned earlier. Subsequently, high-quality ACRs were obtained if: (1) ACRs were detected at least in a pair of

biological replicates and (2) ACRs do not overlap with the centromeric repeat regions, known repeat regions, (Supplementary Table S10-11) or have higher local gDNA coverage than the global gDNA mean at the intergenic region to avoid genomic regions with low complexity. Accessibility centered at ACR summits and annotated transcription start sites were calculated using deepTool v3.5.1 [48] for 20 bp bins with a given interval. For each sample, the count matrix was scaled to 0–1 with the lower 2nd and upper 98th quantile value and the mean values of each bin was used to generate metaplots.

Diversity statistics for *P. hallii* population genetics

We used resequencing data from an earlier study by Bhaskara et al. [26] which identified a set of the high-quality SNP variants from a large panel (n = 85) of *P. hallii* natural accession (detailed pipeline of SNP calling can be found here: https://github.com/tahia/SNP_calling_GATK). This study identified three major genetic clusters in *P. hallii*. We limited our population genomic analysis here to only include the two major genetic clusters of *P. hallii* population: Northern inland ecotypes from dry xeric habitats and coastal ecotypes from wet mesic habitats (Supplementary Table 12). We calculated the SNP density from 2 Kb upstream to 2 Kb downstream from the summit of a given ACR by a sliding window approach (window size = 200 bp, step size = 100 bp). Tn5 integration at ACR was also obtained for the same sliding windows by counting BED tags for pooled libraries using bedtools v2.29.1 to construct metaplots (Fig. 3).

Test for enrichment of features

In order to understand the relationship between accessible chromatin regions and gene expression divergence we joined our results with the expression QTL study of Lovell et al. [25]. This allowed us to ask about the overlap or degree to which the observation of an ACR (or CNS) is predictive of the observation of a *cis*- or *trans*-eQTL. These analyses focused only on genes that had one-to-one orthology between the inland and coastal reference genomes. Genes which had one-to-one ortholog pairs, and that were reported as expressed, but that did not exhibit eQTL were considered background genes for eQTL enrichment tests. Genes were subsequently grouped as an ACR carrier if any reported ACR was positioned within 2Kb of the gene model or else grouped as a non-ACR carrier. To test for enrichment, we asked how often eQTL genes were also ACR carriers compared to background genes using Fisher's exact test. P-values were adjusted for multiple tests using a Bonferroni correction method. To test the enrichment of CNS at ACR, we first calculated the frequency of CNS in a given interval (ACR or random genomic background of a given category or globally). Any ACR (or background genomic interval) which overlapped > = 25% of CNS interval was considered to possess that CNS. Similar to our eQTL enrichment test, we tested for the frequency of CNS at a given ACR category compared to the appropriate background genomic regions using Fisher's exact test and p-values were adjustedwith a Bonferroni correction method. We also tested for the enrichment of CNS at ACR globally in a similar fashion.

Random genomic background for ACRs and CNS

To test for enrichment of a given feature (eQTL or CNS) we drew random genomic backgrounds from each ecotype for a given category of ACR. First the genomic intervals of each category of ACRs were obtained using bedtools v2.29.1 [46]. We restricted these genomic intervals to those that do not overlap with any detected ACRs or the centromeric repeat regions (which were blacklisted while calling ACR peak). Later, random genomic regions from each class (window size=median ACR length of a given ecotype) were drawn using bedtools and subsequently n intervals (n= number of ACRs for a given category) were randomly selected as a corresponding genomic background. While testing for motif enrichment at CNS, we drew genomic background of 10,000 random set of 200 bp intervals from each reference ecotype in the same genomic space (no CDS, the centromeric repeat regions or annotated Transposable Elements and restricted to 100 Kb flanking regions of annotated gene models) for CNS except the regions detected as CNS in any category (core- , unique-, or pan-CNS).

Transcription Factor (TF) Motif analyses

To predict putative TF Motifs in ACRs or CNS, we used the MEME FIMO tool [49] to scan the JASPAR22 [50] core plant motifs with default parameters (p-value < $1e^{-4}$) and retained predicted motifs which pass the multiple test correction (q-value < 0.05). To detect the enrichment of TF-motifs at ACRs the frequency of predicted motifs was calculated in a 100 bp window and the mean frequency of putative motifs for different categories of ACRs was compared with random noncoding genomic background.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-026-12785-w.

Supplementary Material 1.

Supplementary Material 2.

Acknowledgements

This work was supported by funds from the National Science Foundation (IOS-1856627) to R.J.S. and (IOS- 1444533) to T.E.J., the Office of Science (BER), U.S. Department of Energy, Grant no DE-SC0021126 to T.E.J, and the Integrative Biology Research Grant for graduate students, UT Austin to T.H.

Authors' contributions
T.H., G.B.B and T.E.J. planned and designed the research. T.H. and G.B.B. performed the experiments and collected data, T.H. analyzed the data, T.H., and T.E.J. wrote the manuscript, with review and contributions from G.B.B and R.J.S.

Funding
This work was supported by funds from the National Science Foundation (IOS-1856627) to R.J.S. and (IOS- 1444533) to T.E.J., the Office of Science (BER), U.S. Department of Energy, Grant no DE-SC0021126 to T.E.J, and the Integrative Biology Research Grant for graduate students, UT Austin to T.H.

Data availability
All raw sequence data of TAGSeq and ATAC-Seq for this study can be found in NCBI SRA archive under bioproject PRJNA1170929 (https://www.ncbi.nlm.nih.gov/sra/?term=PRJNA1170929). Processed data for TAGSeq and ATACSeq libraries is available under GSE316229. All code used to produce this manuscript is available at (https://github.com/tahia/Evo_CREs).

Declarations

Competing interests
The authors declare no competing interests.

Received: 12 October 2025 / Accepted: 19 March 2026
Published online: 30 March 2026

References

1. Stern DL, Orgogozo V. The loci of evolution: how predictable is genetic evolution? Evolution. 2008;62(9):2155–77.
2. Hill MS, Vande Zande P, Wittkopp PJ. Molecular and evolutionary processes generating variation in gene expression. Nat Rev Genet. 2021;22(4):203–15.
3. Necsulea A, Kaessmann H. Evolutionary dynamics of coding and non-coding transcriptomes. Nat Rev Genet. 2014;15(11):734–48.
4. Springer NM, Stupar RM. Allele-Specific Expression Patterns Reveal Biases and Embryo-Specific Parent-of-Origin Effects in Hybrid Maize. Plant Cell. 2007;19(8):2391–402.
5. Shi X, Ng DWK, Zhang C, Comai L, Ye W, Jeffrey Chen Z. *Cis-* and *trans*-regulatory divergence between progenitor species determines gene-expression novelty in Arabidopsis allopolyploids. Nat Commun. 2012;3(1):950.
6. Lovell JT, Schwartz S, Lowry DB, Shakirov EV, Bonnette JE, Weng X, Wang M, Johnson J, Sreedasyam A, Plott C, et al. Drought responsive gene expression regulatory divergence between upland and lowland ecotypes of a perennial C4 grass. Genome Res. 2016;26(4):510–8.
7. Schmitz RJ, Grotewold E, Stam M. *Cis*-regulatory sequences in plants: Their importance, discovery, and future challenges. Plant Cell. 2021;34(2):718–41.
8. Song B, Buckler ES, Wang H, Wu Y, Rees E, Kellogg EA, Gates DJ, Khaipho-Burch M, Bradbury PJ, Ross-Ibarra J, et al. Conserved noncoding sequences provide insights into regulatory sequence and loss of gene expression in maize. Genome Res. 2021;31(7):1245–57.
9. Zhao H, Tu Z, Liu Y, Zong Z, Li J, Liu H, Xiong F, Zhan J, Hu X, Xie W. PlantDeepSEA, a deep learning-based web service to predict the regulatory effects of genomic variants in plants. Nucleic Acids Res. 2021;49(W1):W523–9.
10. Rodgers-Melnick E, Vera DL, Bass HW, Buckler ES. Open chromatin reveals the functional maize genome. Proc Natl Acad Sci. 2016;113(22):E3177–84.
11. Lu Z, Ricci WA, Schmitz RJ, Zhang X. Identification of *cis*-regulatory elements by chromatin structure. Curr Opin Plant Biol. 2018;42:90–4.
12. Klemm SL, Shipony Z, Greenleaf WJ. Chromatin accessibility and the regulatory epigenome. Nat Rev Genet. 2019;20(4):207–20.
13. Oka R, Zicola J, Weber B, Anderson SN, Hodgman C, Gent JI, Wesselink J-J, Springer NM, Hoefsloot HCJ, Turck F, et al. Genome-wide mapping of transcriptional enhancer candidates using DNA and chromatin features in maize. Genome Biol. 2017;18(1):137.
14. Ricci WA, Lu Z, Ji L, Marand AP, Ethridge CL, Murphy NG, Noshay JM, Galli M, Mejía-Guerra MK, Colomé-Tatché M, et al. Widespread long-range *cis*-regulatory elements in the maize genome. Nat Plants. 2019;5(12):1237–49.
15. Zaret K. Micrococcal Nuclease Analysis of Chromatin Structure. Curr Protoc Mol Biol. 1999;45(1):21. 21.21–21.21.17.
16. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods. 2013;10(12):1213–8.
17. Lu Z, Hofmeister BT, Vollmers C, DuBois RM, Schmitz RJ. Combining ATAC-seq with nuclei sorting for discovery of *cis*-regulatory regions in plant genomes. Nucleic Acids Res. 2016;45(6):e41–41.
18. Lu Z, Marand AP, Ricci WA, Ethridge CL, Zhang X, Schmitz RJ. The prevalence, evolution and chromatin signatures of plant regulatory elements. Nat Plants. 2019;5(12):1250–9.
19. Horvath R, Josephs EB, Pesquet E, Stinchcombe JR, Wright SI, Scofield D, Slotte T. Selection on Accessible Chromatin Regions in Capsella grandiflora. Mol Biol Evol. 2021;38(12):5563–75.
20. Marand AP, Chen Z, Gallavotti A, Schmitz RJ. A *cis*-regulatory atlas in maize at single-cell resolution. Cell. 2021;184(11):3041–e30553021.
21. Turco G, Schnable J, Pedersen B, Freeling M. Automated conserved non-coding sequence (CNS) discovery reveals differences in gene content and promoter evolution among grasses. Front Plant Sci. 2013;4.
22. Lowry DB, Purmal CT, Juenger TE. A population genetic transect of Panicum hallii (Poaceae). Am J Bot. 2013;100(3):592–601.
23. Palacio-Mejía JD, Grabowski PP, Ortiz EM, Silva-Arias GA, Haque T, Des Marais DL, Bonnette J, Lowry DB, Juenger TE. Geographic patterns of genomic diversity and structure in the C4 grass Panicum hallii across its natural distribution. AoB Plants. 2021;13(2).
24. Haque T, Bhaskara GB, Yin J, Bonnette J, Juenger TE. Natural variation in growth and leaf ion homeostasis in response to salinity stress in Panicum hallii. Front Plant Sci. 2022;13.
25. Lovell JT, Jenkins J, Lowry DB, Mamidi S, Sreedasyam A, Weng X, Barry K, Bonnette J, Campitelli B, Daum C, et al. The genomic landscape of molecular responses to natural drought stress in Panicum hallii. Nat Commun. 2018;9(1):5213.
26. Bhaskara GB, Haque T, Bonnette JE, Napier JD, Bauer D, Schmutz J, Juenger TE. Evolutionary Analyses of Gene Expression Divergence in Panicum hallii: Exploring Constitutive and Plastic Responses Using Reciprocal Transplants. Mol Biol Evol. 2023;40(10).
27. Tu X, Mejía-Guerra MK, Valdes Franco JA, Tzeng D, Chu P-Y, Shen W, Wei Y, Dai X, Li P, Buckler ES, et al. Reconstructing the maize leaf regulatory network using ChIP-seq data of 104 transcription factors. Nat Commun. 2020;11(1):5089.
28. Maher KA, Bajic M, Kajala K, Reynoso M, Pauluzzi G, West DA, Zumstein K, Woodhouse M, Bubb K, Dorrity MW, et al. Profiling of Accessible Chromatin Regions across Multiple Plant Species and Cell Types Reveals Common Gene Regulatory Principles and New Control Modules. Plant Cell. 2017;30(1):15–36.
29. Zhang W, Wu Y, Schnable JC, Zeng Z, Freeling M, Crawford GE, Jiang J. High-resolution mapping of open chromatin in the rice genome. Genome Res. 2012;22(1):151–62.
30. Yocca AE, Lu Z, Schmitz RJ, Freeling M, Edger PP. Evolution of Conserved Noncoding Sequences in Arabidopsis thaliana. Mol Biol Evol. 2021;38(7):2692–703.
31. Haudry A, Platts AE, Vello E, Hoen DR, Leclercq M, Williamson RJ, Forczek E, Joly-Lopez Z, Steffen JG, Hazzouri KM, et al. An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. Nat Genet. 2013;45(8):891–8.
32. Li Y, Zhang L, Li D, Liu Z, Wang J, Li X, Yang Y. The Arabidopsis F-box E3 ligase RIFP1 plays a negative role in abscisic acid signalling by facilitating ABA receptor RCAR3 degradation. Plant Cell Environ. 2016;39(3):571–82.
33. Khan D, Hui L, Khokhar AA, Hussain MA, Lv W, Zaman QU, Wang H-F. Functional characterization of MATE gene family under abiotic stresses and melatonin-mediated tolerance in dragon fruit (Selenicereus undatus L). Plant Stress. 2024;11:100300.
34. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34(18):3094–100.
35. Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol. 1981;147(1):195–7.
36. Zhang Z, Berman P, Miller W. Alignments Without Low-Scoring Regions. J Comput Biol. 1998;5(2):197–210.
37. Weng X, Lovell JT, Schwartz SL, Cheng C, Haque T, Zhang L, Razzaque S, Juenger TE. Complex interactions between day length and diurnal patterns of gene expression drive photoperiodic responses in a perennial C4 grass. Plant Cell Environ. 2019;42(7):2165–82.
38. FastQC. A Quality Control Tool for High Throughput Sequence Data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

39. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. 2011;17(1):3. https://doi.org/10.14806/ej.17.1.200
40. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009;25(14):1754–60.
41. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Subgroup GPDP. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078–9.
42. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2013;30(7):923–30.
43. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550.
44. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114–20.
45. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357–9.
46. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841–2.
47. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, et al. Model-based Analysis of ChIP-Seq (MACS). Genome Biol. 2008;9(9):R137.
48. Ramírez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, Heyne S, Dündar F, Manke T. deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res. 2016;44(W1):W160–5.
49. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. Bioinformatics. 2011;27(7):1017–8.
50. Castro-Mondragon JA, Riudavets-Puig R, Rauluseviciute I, Berhanu Lemma R, Turchi L, Blanc-Mathieu R, Lucas J, Boddie P, Khan A, et al. Manosalva Pérez N: JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. Nucleic Acids Res. 2021;50(D1):D165–73.

Publisher's Note