Phenotypic variation in maize can be largely explained by genetic variation at transcription factor binding sites

Julia Engelhorn[1,2], Samantha J. Snodgrass[4], Amelie Kok[1,2,3], Arun S. Seetharam[4], Michael Schneider[5], Tatjana Kiwit[1,2], Ayush Singh[6], Michael Banf[7], Merritt Khaipho-Burch[8], Daniel E. Runcie[9], Victor Sánchez Camargo[10], J. Vladimir Torres-Rodriguez[11], Guangchao Sun[11], Maike Stam[10], Fabio Fiorani[12], James C. Schnable[11], Hank W. Bass[6], Matthew B. Hufford[4], Benjamin Stich[3,5,13], Wolf B. Frommer[1,2,3], Jeffrey Ross-Ibarra[14], Thomas Hartwig[1,2,3*]

[1] Institute for Molecular Physiology, Cluster of Excellence on Plant Sciences, Heinrich-Heine University, Düsseldorf, Germany

[2] Independent Research Groups, Max Planck Institute for Plant Breeding Research, Cologne, Germany

[3] Cluster of Excellence on Plant Sciences, Düsseldorf, Germany

[4] Dept. of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, USA

[5] Institute for Quantitative Genetics and Genomics of Plants, Cluster of Excellence on Plant Sciences, Heinrich-Heine University, Düsseldorf, Germany

[6] Dept. of Biological Science, Florida State University, Tallahassee, USA

[7] EducatedGuess.ai, Neunkirchen, Germany

[8] School of Integrative Plant Sciences, Plant Breeding and Genetics Section, Cornell University, Ithaca, USA

[9] Dept. of Plant Sciences, University of California, Davis, USA

[10] Swammerdam Institute for Life Sciences, University of Amsterdam, Amsterdam, Netherlands

[11] Dept. of Agronomy and Horticulture, University of Nebraska-Lincoln, Lincoln, USA

[12] Institute of Bio- and Geosciences 2, Plant Sciences, Research Institute Jülich, Germany

[13] Institute for Breeding Research on Agricultural Crops, Federal Research Centre for Cultivated Plants, Groß Lüsewitz, Germany

[14] Dept. of Evolution and Ecology, Center for Population Biology and Genome Center, University of California, Davis, USA

Abstract

Comprehensive maps of functional variation at transcription factor (TF) binding sites (*cis*-elements) are crucial for elucidating how genotype shapes phenotype. Here we report the construction of a pan-cistrome of the maize leaf under well-watered and drought conditions. We quantified haplotype-specific TF footprints across a pan-genome of 25 maize hybrids and mapped nearly two-hundred thousand genetic variants (termed binding-QTLs) linked to *cis*-element occupancy. The functional significance of binding-QTLs is supported by three lines of evidence: i) they coincide with known causative loci that regulate traits, including novel alleles of *Upright Plant Architecture2*, *Trehalase1*, and the MITE transposon near *ZmNAC111* under drought; ii) their footprint bias is mirrored between inbred parents and by ChIP-seq; iii) partitioning genetic variation across genomic regions demonstrates that binding-QTLs capture the majority of heritable trait variation across ~70% of 143 phenotypes. Our study provides a promising approach to make previously hidden *cis*-variation more accessible for genetic studies and multi-target engineering of complex traits.

Background

Over the past two decades, genome-wide association studies (GWAS) transformed our understanding of the inheritance of many complex traits in important agricultural crops such as maize. Several studies estimated that non-coding variation accounts for about 50% of the additive genetic variance underlying phenotypic diversity in plants (Chia *et al.* 2012; Rodgers-Melnick *et al.* 2016; Lorant *et al.* 2020; Song *et al.* 2021). Identification of functional non-coding variants is advancing along with the development of new genomics technologies (Marand *et al.* 2023). However, it remains challenging to discern functional variants that impact *cis*-elements efficiently and at cistrome scale. Knowing which loci to target has become one of the biggest obstacles for trait improvement via targeted genome editing (Wolter and Puchta 2018; Sharon *et al.* 2018; Marand *et al.* 2023). Scalable methods to construct comprehensive cis-element maps are critical for elucidating complex transcriptional networks that underlie growth, behavior, and disease. The potential of *cis*-element maps has been demonstrated by the ENCODE projects which exist for many eukaryotes, including humans. However, genome-wide, high-resolution maps of functional variants are currently lacking in plants (Lane *et al.* 2014). Despite many successes, GWAS generally suffer from insufficient resolution which limits identification of individual causal single nucleotide polymorphisms (SNPs), insertions or deletions (INDELs), and GWAS cannot provide independent molecular information on the potential function of variants, which then requires laborious follow-up analyses of numerous individual loci (Sharon *et al.* 2018).

An alternative approach to identify functional polymorphisms would be to annotate non-coding variants within a GWAS region based on their association with transcription factor (TF) binding. This approach has considerable potential, as TF activity plays an important role in the regulation of genes and thereby traits, and the affinity of TF binding is mostly determined by

specific local sequences (*cis*-elements) (White *et al.* 2013; Levo and Segal 2014). Identifying *cis*-elements for individual TFs via approaches such as ChIP-seq is time-consuming, not strictly quantitative, limited in scope, and often provides relatively low resolution of functional regions. In contrast, MNase-defined cistrome Occupancy Analysis (MOA-seq) identifies putative TF binding sites globally, in a single experiment, with relative high resolution yielding footprint regions typically <100bp in size (Savadel *et al.* 2021). In maize, MOA-seq identified ~100,000 occupied loci, including about 70% of the sequences identified in more than 100 ChIP-seq experiments (Tu *et al.* 2020; Savadel *et al.* 2021). Notably, many of MOA-footprint regions were previously uncharacterized, with only 35% identified in previous ATAC-seq data (Savadel *et al.* 2021). Similarly, an analysis of small MNase-defined fragments from *Arabidopsis* seedlings revealed more than 15,000 accessible chromatin regions missed by ATAC- or DNase-seq (Zhao *et al.* 2020).

Here, we quantified haplotype-specific TF footprints pan-genome-wide with MOA-seq, utilizing F1 hybrids that share a common reference to minimize biological, technical, and trans-effect variation between the haplotypes. We defined a maize leaf pan-cistrome and identified ~175,000 genetic variants linked to haplotype-specific variation in MOA coverage at *cis*-element loci, we term binding quantitative trait loci (QTLs). bQTLs explained the majority of heritable trait variation in ~70% of the tested traits in the nested association mapping (NAM) panel while representing less than 0.01% of the genome. Haplotype-specific TF footprints coincided with causative loci known to affect leaf angle, branching, and flowering time traits, and identified more than 3500 drought-response *cis*-regulatory loci, including near *ZmTINY* as candidates loci for future genome editing.

Results

Quantification of functional *cis*-variation

Beside genetic variation at TF bound loci, trans-effects, including variation in the amount of active TFs, as well as biological and technical variation between assays can all contribute to differences in TF footprints between inbreds. We hypothesized that both trans-effect and biological/technical variation can be minimized by quantifying TF footprint specific to each haplotype in F1 hybrids with a shared reference parent (B73). To validate our approach, we applied MOA-seq to nuclei of both the B73 and Mo17 F1 hybrid and its inbred parents (Fig. 1a), two founders of key maize breeding populations whose hybrid has been extensively studied (Springer and Stupar 2007; Pressoir *et al.* 2009; Hartwig *et al.* 2023). MOA footprints were determined by mapping sequencing reads to a concatenated hybrid genome, and retaining reads that mapped uniquely or once to each genome (Supplemental Table S1, see methods). We detected 327,029 MOA footprints (at a false discovery rate of 5%) with strong correlation across biological replicates (Pearson's correlation coefficient > 0.95, Additional file 1, Fig. S1). The footprints harbored 325,933 MOA sequence polymorphisms (MPs) and were flanked by 53,220 nearby (5kb up- and 1kb downstream) genes in the hybrid, representing 67.9% of the F1 coding sequences (B73 and Mo17 annotated genes; Supplemental Table S2). We analyzed the ratio of MOA reads at biallelic SNPs in the F1, designating the B73 allele as the reference and the other allele as the variant. A total of 48,935 allele-specific MPs (AMPs) deviated significantly from the expected 1:1 allelic ratio (binomial test with 1% false discovery rate; see Methods for details, Additional file 1, Fig. S2).

Notwithstanding trans-effects, we assumed our F1 MOA-seq results to be independently reproducible in a comparison of individual inbred parents. Consistent with this expectation, about 88% (194594/221187) of all MPs showed no significant difference in their allelic bias

comparing F1 and B73 vs. Mo17 inbred alleles, which suggests that 12% ($p<0.05$, 4.5% $p<0.01$) of loci vary either due to trans-effects between the parents or technical variation (Fig. 1b). Furthermore, approximately 90% (32251/36001) of AMP sites in the B73xMo17 F1 showed bias towards the same allele as in the inbred parents, with fewer than 1% (227/36001) showing bias in the opposite direction (Fig. 1c). The high accordance of the MOA-seq data between the F1 and its inbred parents not only establishes the reproducibility of the assay but also indicates that the majority of AMPs are coupled to genotypic differences in *cis* at the binding site, rather than owing to *trans*-acting or *cis*-by-*trans* interaction effects.

To independently validate haplotype-specific, MOA-defined TF footprints in B73xMo17, we compared AMPs to recently published allele-specific ChIP-seq data of the major brassinosteroid TF ZmBZR1 in the same F1 (Hartwig *et al.* 2023). More than 70% of AMPs overlapping with ZmBZR1 binding sites showed allelic bias in the same direction in both studies (Fig. 1d). More than 20% of AMPs show no bias in the ChIP-seq data, likely due to the lower resolution of haplotype-specific ChIP-seq (~500 bp fragments compared to ~65 bp for MOA-seq) leading to a substantially larger region being designated as differential bound by a single TF. Only 8% of AMPs show bias to the opposing allele between studies, potentially reflecting biological differences due to differences in analyzed tissues (meristem and leaf vs. leaf) or ectopic BZR1 activity due to exogenous brassinosteroid treatment (Hartwig *et al.* 2023).

If haplotype-specific MOA-seq can detect relevant variation in TF binding, we expected it to coincide with potential *cis*-variation in the B73xMo17 F1. We previously reported the allele-specific binding of the major Brassinosteroid TF ZmBZR1 upstream of the TSS of *POLYGALACTURONASE-INHIBITING PROTEIN2* (*ZmPGIP2*, Zm00001eb034870, Fig. 1e). PGIP2 is a cell wall protein and a candidate gene for both northern and southern leaf blight resistance (Balint-Kurti *et al.* 2007,). The overlapping MOA and BZR1 ChIP binding peaks were both significantly higher for the B73 allele compared to Mo17, exhibiting a 5-fold ($p<0.01$) higher

MOA coverage (Fig.1e-f). Each of the B73 AMP alleles overlapped with known BZR1 motifs (CGTGTG, CACGTG, and CACGTT, respectively). In contrast, the Mo17 allele contained both a SNP in the BRRE motif as well as a HIP-superfamily helitron insertion between the BRRE and G-Box motifs. In another example, we found a significantly higher (1.96-fold; p=0.016) MOA occupancy for the B73 allele of a region upstream of *BARREN INFLORESCENCE 2* (*ZmBIF2,* Zm00001eb031760) (Fig. 1g-h), along with codirectional changes in transcript abundance (Additional file 1, Fig S3). Mutant *bif2* maize plants exhibit reduced formation of all axillary structures, including tassel branches, spikelets, and ear shoots (McSteen and Hake 2001). Interestingly, previous work linked natural variation at a small 80 bp region in the *BIF2* proximal promoter between B73 and Mo17 (in Fig. 1g) to variation in traits such as tassel branch zone length, plant height, and leaf width and length (Pressoir *et al.* 2009). Together, the examples illustrate the power of MOA-seq to annotate candidate *cis*-regulatory elements with quantitative chromatin footprint data that connects *cis*-variation to biases in *cis*-element occupancy.

Defining functional sites in a maize pan-cistrome

Although the binding sites of TF can be defined by existing methods such ChIP-seq or DAP-seq, they lack the desired combination of scalability, resolution, and the ability to define *cis*-regulatory variation quantitatively *in vivo*. To define a leaf pan-cistrome of maize, we analyzed a population of 25 F1 hybrids using haplotype-specific MOA-seq (Fig. 1a, Supplemental Table S3). The hybrid population, created by crossing to 25 inbred lines with high quality genome assemblies (Lin *et al.* 2021; Hufford *et al.* 2021) to the reference genome line B73, represents a diverse set of maize including many of the parents of an important mapping population and several important genetic stocks (Supplemental Table S4).

We analyzed allele-specific TF occupancy and mRNA abundance in leaf blades for three biological replicates of each hybrid (Supplemental Table S5). By aligning MOA-seq and RNA-seq reads to concatenated dual-reference genomes, rather than a single reference, our approach resolves issues of reference bias that confound most allele-specific analyses (Fig. 2a, see methods, (Castel *et al.* 2015)). We identified an average of 362,000 MOA peaks (at a false discovery rate of 5%) per F1, covering approximately 2% (around 80 Mbp) of each hybrid genome (Additional file 1, Fig. S4). On average 19.9% (14-30%) of MPs showed a significant allelic bias (binomial test, false discovery rate of 1%, Supplemental Table S6) with an overall even split between the parental alleles (51% B73 and 49% diverse parents, Additional file 1, Fig. S5, Supplemental Table S7). In total, AMPs covered 36.1% of the B73 footprint peaks (Fig. 2b), and plots of the identified MOA peaks and cumulative bp suggest our sample is near saturation and has identified the majority of the B73 cistrome (Additional file 1, Fig. S6).

We next sought to identify variants associated with differences in MOA-detected TF occupancy between parents, or binding quantitative trait loci (bQTLs), in our population. We first verified that lines sharing an AMP haplotype share similar patterns of allelic bias (Fig. 2c), suggesting that our SNPs were in sufficient linkage disequilibrium (LD) with causal differences to perform association analysis. Because differences in methylation between parental alleles can affect TF binding affinity (O'Malley *et al.* 2016; Hartwig *et al.* 2023), we added previously published methylation data for 24 of our parental lines (Noshay, Liang, *et al.* 2021; Lin *et al.* 2021; Hufford *et al.* 2021). Mixed linear modeling identified 132,468 bQTLs significantly (FDR<0.05) associated with variation in MOA occupancy (Supplemental Table S8). SNPs alone explained 42% of the observed differences in MOA coverage, whereas SNPs and DNA methylation combined explained 57% of the MOA differences (Additional file 1, Fig. S7).

bQTLs coincide with numerous known, causative regulatory loci

Detailed analyses of regulatory variation for a number of maize genes provide an opportunity to compare bQTLs to previously identified causal variation. Flowering time is a well-studied trait with a number of known, non-coding, regulatory loci, which we hypothesized could overlap with bQTLs. Indeed, bQTLs flanked the causative transposon insertions at *ZmCCT9* and *ZmCCT10* (Huang *et al.* 2018) and an INDEL in the promoter of the FT-like *ZmZCN8* (Meng *et al.* 2011), as well as multiple GWAS hits for flowering time (Additional file 1, Fig. S8). In addition to identifying bQTL in both of the known distal regulatory regions of the key flowering time locus *ZmRAP2.7* (vgt1 and vgt1-DMR) (Salvi *et al.* 2007; Castelletti *et al.* 2014; Xu *et al.* 2020) (Fig. 2d), our analysis finds a novel, third enhancer more than 100 kb upstream, which we termed vgt1-MOA. HiC long-range interaction data (Ricci *et al.* 2019) confirmed that vgt1-MOA physically interacts with both *ZmRAP2.7* and vgt1 (Fig. 2d).

Transposable element (TE) insertion plays important roles in natural variation of gene regulation (Noshay, Marand, *et al.* 2021). While the repetitive nature of transposons makes overlapping footprint analysis challenging, we find that bQTLs flank transposons at *cis*-element insertion sites. In addition to bQTL that coincide with flowering loci, we observed bQTLs that colocalize with the regulatory variation upstream *teosinte branched1* (*TB1*) and *grassy tillers1* (*GT1*), which form a regulatory module involved in bud dormancy and growth repression (Dong *et al.* 2019). bQTLs flank both causal transposon insertions at *TB1* (Studer *et al.* 2011), and coincide with the TE-associated causal regulatory region (prol1.1) upstream of *gt1* (Wills *et al.* 2013) (Additional file 1, Fig. S9).

MOA-seq combined with the diversity present in the NAM parents provides an opportunity to evaluate not only overlap with known regulatory sites, but how variation at these sites compares to changes at known regulatory motifs. For example, the causative C2C2 motif

(AGTGTG) underlying the *Upright Plant Architecture2* (UPA2) QTL involved in leaf angle determination contained a MOA bQTL (Additional file 1, Fig. S10). Compared to teosinte *UPA2*, the maize *UPA2* element is bound by the YABBY TF drooping leaf1 (DRL1) with lower affinity, due to a deletion of one of the TG sites (Tian *et al.* 2019). While none of our F1s contained the teosinte *UPA2* allele, all lines (17/17, including B73) which contained the canonical maize allele showed a similar MOA occupancy. In contrast, most (6/8) alleles which contained an additional deletion, eliminating both TG sites, showed lower MOA occupancy (Additional file 1, Fig. S10). The *UPA2* region is also bound by other TFs (Tu *et al.* 2020; Hartwig *et al.* 2023), including BZR1, which may explain the unchanged MOA occupancy in two of the TGTG-deletion alleles. In another example, an INDEL in the *TREHALASE1* promoter (*TRE1*) has been associated with both trehalose and *TRE1* transcript levels in maize (Wen *et al.* 2018). We observed haplotype-specific footprints, both at the previously reported 8 bp Insertion (Wen *et al.* 2018) and an additional SNP 29 bp upstream, which coincided with a bQTL (Additional file 1, Fig. S11). Notably, while the 8-bp insertion creates an ABI motif (TGCCACAC), the *TRE1*-bQTL overlaps with a DOF binding motif (AAAAGGTG). Indeed previously published ChIP-seq results show that the *TRE1*-bQTL site is targeted by ZmDOF17 (Tu *et al.* 2020). Furthermore, all alleles (6/6) in our F1 population without the 8 bp insertion and with the non-canonical DOF motif (TC/TG) at the bQTL site, showed concomitant low MOA signal and allele-specific *TRE1* mRNA levels (Fig. 2e-f, Additional file 1, Fig. S11).

Variation in MOA occupancy is correlated to differential transcript accumulation

If variation in MOA coverage accurately captures TF binding affinity, we expect to see associations between haplotype-specific MOA coverage and transcript abundance in our F1s. Indeed, we find that the promoters (within 3 kb upstream of the TSS) of genes with significant

($p<0.05$) haplotype-specific transcripts abundance (see methods) were ~35% and ~75% enriched for the presence of AMPs compared to all expressed and non-differentially ($p>0.95$) expressed genes, respectively, in both well-watered and drought conditions (Fig. 2g, Supplemental Table S9). MOA bQTLs were also substantially more likely than matched background (bg) SNPs (33.8%, 1649/1232 for bQTLs and bgSNPs, respectively) to be in high linkage disequilibrium (> 0.6) with nearby *cis*-expression QTLs in a panel of 340 maize genotype (Sun *et al.* 2023) (Fig. 2h). These broad patterns are reflected at the level of individual genes as well. For example, all of the NAM parents showing greater MOA occupancy at the two bQTLs upstream of phospo-glycerate mutase (*ZmPGM1*, Zm00001eb196320) showed increased abundance of the NAM transcript, whereas lines without polymorphism showed no haplotype-specific transcript levels (Fig. 2i-k). Two NAM parents, Ki3 and CML69, showed drastically reduced *ZmPGM1* mRNA levels compared to B73 in the F1, despite similar MOA coverage. The MOA peaks for the Ki3/CML69 haplotypes, however, were shifted by approximately 300 bp compared to B73 due to the insertion of a PIF/Harbinger transposon resulting in hypermethylation of the DNA between the MOA peak and the TSS, a pattern not observed for alleles with the transposon (Fig. 2i-k, Additional file 1, Fig. S12). Taken together, the results suggest that variation at bQTLs contributes to differences in gene regulation.

Variation in DNA-methylation can predict MOA occupancy

The vast majority of TFs in *Arabidopsis* have been shown, *in vitro*, to have higher binding affinity to hypomethylated DNA (O'Malley *et al.* 2016). We explored this association in our data, focusing on mCG/mCHG variation, as they accounted for >99.8% of methylation differences at MOA sites. Significant DNA methylation differences (following (Regulski *et al.* 2013), one allele <10% methylated and the other >70%) overlapped with 14.8% of MPs in F1s. At AMP loci, haplotype-specific mCG/mCHG overlap increased by 2.6-fold (38.1%) and reached more than

half (51.5%) for AMPs with a strong haplotype-bias (>85% to one allele) (Fig. 3a). We observed a very strong correlation between a higher footprint occupancy and the hypomethylated allele (Fig. 3b) with only ~2% of AMPs showing higher MOA coverage at the hypermethylated allele. Notably, nearly half of the 2% AMPs biased towards hypermethylation alleles had no Cs in the 10 bp window surrounding the AMP, suggesting there is no actual methylation difference at the occupied site despite hypermethylation of the surrounding region (+/- 20 bp) (Additional file 1, Fig. S13). On average, the vast majority (71.2%) of F1s that shared differentially methylated alleles at a given site also shared haplotype-specific MOA footprints, compared to 42.9% of F1s that shared equally methylated alleles at the same site (Fig. 3c). The observed strong correlation between differential CG and CHG methylation and haplotype-specific MOA occupancy confirm an important role for DNA methylation in determining TF binding in maize.

MOA bQTLs explain a large portion of heritable variation

Regulatory variation is thought to underlie a significant proportion of phenotypic variation in maize (Wallace *et al.* 2014). To assess the relationship between MOA bQTLs and complex trait variation, we first quantified the enrichment of bQTLs surrounding GWAS hits (+/- 100 bp) across two curated datasets of 41 and 279 traits (Wallace *et al.* 2014; Tian *et al.* 2020). In both cases, MOA bQTLs were approximately twofold enriched for co-localization with GWAS hits, compared to matched background SNPs (bgSNPs, 100 permutations, see methods) with similar allele frequency and distance to the nearest gene (Additional file 1, Fig. S14). This enrichment remained stable as a function of distance to the nearest gene, indicating comparable efficacy of bQTLs to mark functionally significant loci genome-wide (Fig. 3d). Next, to explore the degree to which bQTLs can more broadly capture the genetic variation underlying phenotypic diversity, we partitioned heritable trait variance for 143 traits in the NAM population (see methods, (Rodgers-Melnick *et al.* 2016; Hartwig *et al.* 2023)). We modeled additive genetic variation for

traits as a function of genomic relatedness matrices estimated from bQTLs, matched bgSNPs, and SNPs from the rest of the genome. Variances estimated this way for trait data simulated from our observed matrices accurately reflected the proportional contributions of each SNP set (Additional file 1, Fig. S15). Across a large majority of phenotypes in the NAM panel (98 of 143 or ~70%), bQTL explained the majority of the total additive genetic variance captured by SNPs Fig. 3e, Additional file 1, Fig. S16, Supplemental Table S10). Consistent with previous findings that open chromatin and TF binding play a key role in trait variation ((Rodgers-Melnick *et al.* 2016; Hartwig *et al.* 2023)), our matched bgSNPs often accounted for more additive genetic variation than SNPs from the rest of the genome (120 out of 143 traits), but MOA bQTLs nonetheless outperformed bgSNPs for most traits (103 traits; Fig. 3e). MOA bQTLs also outperformed randomly sampled SNPs within MOA peaks (Additional file 1, Fig. S17). Traits where bQTLs explained the largest portion of genetic variance included plant height, leaf size or shape, and disease-resistance, while almost all traits related to e.g. vitamin E production were best explained by bgSNPs or SNPs from the rest of the genome (Fig. 3f), likely because of the oligogenic nature of the vitamin E traits and that MOA bQTLs identified in leaf tissue may not be representative of regulatory patterns in genes specifically expressed in kernels (Diepenbrock *et al.* 2017).

Characterization of a drought-responsive cistrome

To assess the relationship between variations of MOA occupancy induced by a response to changes in environmental conditions, we compared the morphological and molecular response of our F1 population under well-watered (WW) and drought (DS) conditions. We observed diverse responses in the V4-stage plants subjected to 86hr of drought stress, with the reduction of relative leaf water content ranging from 10-32% and remaining soil water content ranging from 6.3 % to 25.6 %, depending on the F1 (Additional file 1, Fig. S18). With the

exception of B73xOh7b all F1s showed leaf rolling/wilting compared to their WW controls, and differences in relative water content mirrored morphological changes (Fig. 4a-b, Additional file 1, Fig. S19). To evaluate drought-induced differences in *cis*-element regulation, we performed allele-specific MOA- and RNA-seq. The number of MOA peaks showing significant drought-induced increases or decreases ($p<0.05$) in occupancy varied substantially among F1s, ranging from around 23k to 83k and 37k to 134k, respectively (Supplemental Table S11). Local association mapping identified 114,809 DS-bQTLs under drought conditions (for a total of 174,819 unique bQTLs) (Supplemental Table S12). To identify a set of candidate drought-response loci, we selected bQTLs with drought-responsive occupancy ($p<0.05$) near genes (5kb upstream/1kb downstream) that displayed both allele-specific and drought-responsive transcript accumulation, resulting in 917 (370 genes) and 2675 (1068 genes) up- and down-regulated bQTLs/genes, respectively. Further integration with GWAS and *cis*-expression QTL hits for drought-response traits (Li *et al.* 2016; Shikha *et al.* 2017; Wu *et al.* 2021) resulted in high-confidence candidates (Supplemental Table S13). Notably, the candidate list includes the known drought-response variation in the proximal promoter of *ZmNAC111*. A DS-bQTL flanked the causative 82 bp MITE transposon insertion site, which reduces both *ZmNAC111* expression and drought tolerance in maize seedlings, likely through RNA-directed DNA methylation (Mao *et al.* 2015) (Additional file 1, Fig. S20).

A further identified candidate is ZmTINY, the maize homolog of the *Arabidopsis* drought-inducible AP2/ERF TF TINY. The AtTINY family regulates the interplay between plant growth and drought response (Xie *et al.* 2019). Over-expression of *AtTINY* increases drought tolerance, at the cost of severely stunted growth, a limitation often observed with drought-related TFs (Xie *et al.* 2019). The maize homolog of *TINY* is a candidate gene for drought and leaf size variation (Shikha *et al.* 2017). DS-bQTLs and drought-responsive MOA footprint were observed surrounding *ZmTINY*, which were shared in more DS-tolerant alleles,

e.g. Oh43 and CML333, but not found in B73 or DS-susceptible lines such as Mo18W (Fig. 4c). The variation in MOA footprints also correlated with allele-specific transcript levels at *ZmTINY*. Under DS, but not WW, conditions Oh43 and CML333 mRNA accumulated 18- and 84-fold higher than B73 transcripts, respectively, while Mo18W transcripts were only 4-fold higher than B73 (Fig. 4d). Differences in methylation patterns reflected the observed mRNA differences. The downstream MOA peaks with high occupancy were hypomethylated in Oh43 and CML333, but hypermethylated in the lines with low occupancy, e.g. Mo18W (Fig. 4d). To validate the differences observed in MOA occupancy, we performed dual-luciferase transient expression assays, comparing the 1.5kb upstream promoter of $TINY^{B73}$ and $TINY^{CML333}$ with and without abscisic acid (ABA) treatment to simulate DS. Both promoter fragments exhibited significantly higher LUC and REN ratio than the vector control. However, exogenous application of 1 and 10 μM ABA further increased the LUC/REN ratio by 41.4% and 60.3%, respectively, for maize protoplasts harboring the $TINY^{CML333}$ but not $TINY^{B73}$ promoter fragment (Fig. 4e).

Lastly, we evaluated whether the amount of MOA occupancy or direction of allelic-bias change between drought-responsive AMP loci. The results showed that approximately 75% of drought-responsive AMPs changed MOA occupancy levels, either from no detectable MOA signal to allele-specific binding (35%) or a change in MOA coverage between WW and DS conditions with no change in bias (40%) (groups I and II in Fig. 4f). In contrast, only about 25% AMPs changed bias, either from no significant bias to an allelic bias (~15%), or changing the direction of bias (~10%). (Fig. 4f). Based on the results we propose that the majority of drought-responsive AMPs change the amount of MOA occupancy rather than the direction of allelic bias between WW and drought conditions.

Conclusions

We present a robust method for identifying functional, genetic variants linked to trait variation in plants. By integrating haplotype-specific TF footprints and transcript abundance, nested F1 hybrids, and local association mapping at putative *cis*-element loci, we defined a pan-cistrome of the maize leaf under well-watered and drought conditions. Utilization of concatenated dual-reference genomes and F1 hybrid analysis resolved issues of reference bias, haplotype-specific trans-effects, and technical variation that commonly compromise haplotype-specific quantitation. Our analysis demonstrates a high level of variation in *cis*-regulatory networks among 25 diverse maize genotypes, and provides a high resolution map of regulatory elements underpinning the function of nearly two-hundred thousand putative *cis*-element loci in the maize leaf. Finally, we highlight the relevance of bQTL loci for understanding phenotypic diversity in maize, demonstrating that haplotype-specific MOA-seq allowed us to capture the majority of additive genetic variation for most tested phenotypes in maize in less than 0.01% of the genome.

Author Contributions

T.H. conceived the research project. J.E., S.J.S., J.R.I., D.E.R., M.B.H., B.S., M.St., W.B.F., and T.H. designed experiments. J.E., S.J.S., T.K., V.C.S., A.S., J.V.TR, A.K., and T.H., performed experiments. J.E., S.J.S., A.S., A.K., M-K.B., M.S., J.R.I., D.E.R., G.S., V.C.S, F.F., A.S., A.S.S., J.V.TR, J.C.S., M.B. and T.H. analyzed the data. J.E., S.J.S, A.S.S., and T.H. developed the hybrid mapping pipeline and J.E., S.J.S., J.R.I., D.E.R., J.E., J.V.TR., S.J.S., M.S., A.K., M.B. and T.H. developed analysis scripts. J.E., J.R.I., and T.H. wrote the manuscript, and W.B.F., H.W.B., F.F., S.J.S., and B.S. revised it.

Acknowledgement

We are grateful to Marty Sachs for germplasm support and the Maize COOP. We thank Martin Krist and Florian Esser for field support. We thank S. Pophaly and M. H. Thoben at MPIPZ for IT and Martin Krist for field support. We thank Christian Becker, Kerstin Becker, and Bruno Huettel for library and sequencing support. We are grateful to Sergius Weitzel, Max Blank, Michail D. Mueller, Ram S. Kumar, Guy Koentjes, Isabelle Tol, Sohini Mukherjee, Nafiseh Sargheini, Duong Doan, Luca Watschke, Viola Rajan and the MPIPZ and HHU greenhouse teams for greenhouse / experiment support. We would like to thank Felix Andrews for suggesting a mascot for this work, but international restrictions on the transport of fossil bird specimens prohibited its use. We also like to thank Nathan Springer for his critical discussion of the manuscript.

Funding

This study was supported by funding from the "Deutsche Forschungsgemeinschaft" (DFG, ID:458854361 to T.H.) as part of DFG Sequencing call 3 and the European Union's Horizon Europe program BOOSTER (ID: 101081770 to T.H.). B.S., W.B.F., and T.H. received funding by the DFG, under Germany's Excellence Strategy (CEPLAS Cluster, EXC 2048/1, Project ID: 390686111). J.R.I. was supported by funding from the US National Science Foundation (ID:1822330) and US Dept. of Agriculture (Hatch project CA-D-PLS-2066-H 548). J.C.S & J.V.RT's work on this project was supported by the U.S. Department of Energy (Grant no. DE-SC0020355). S.S. was supported by funding from the US National Science Foundation through the Graduate Research Fellowship Program (DGE#1744592). This work was supported, and some of the NGS analysis performed by, the DFG Research Infrastructure West German Genome Center (407493903) as part of the Next Generation Sequencing Competence Network (project 423957469). Some of the results reported in this paper were partially

supported by the HPC@ISU equipment at Iowa State University, some of which has been purchased through funding provided by NSF under MRI grants number 1726447 and MRI2018594.

Conflict of interests

The authors declare that they have no competing interests.

Data availability statement

The raw and processed MOA-seq and RNA-seq data used in this work have been deposited to the NCBI SRA. The BioProject ID and BioSamples IDs will be added upon acceptance.

References:

Balint-Kurti PJ, Zwonitzer JC, Wisser RJ, *et al.*2007. Precise Mapping of Quantitative Trait Loci for Resistance to Southern Leaf Blight, Caused by Cochliobolus heterostrophus Race O, and Flowering Time Using Advanced Intercross Maize Lines. *Genetics* **176**: 645–657.
Bian Y, Yang Q, Balint-Kurti PJ, Wisser RJ, Holland JB. **2014**. Limits on the reproducibility of marker associations with southern leaf blight resistance in the maize nested association mapping population. *BMC Genomics* **15**: 1068.
Bolger AM, Lohse M, Usadel B. **2014**. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114–2120.
Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. **2007**. TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* **23**: 2633–2635.
Castel SE, Levy-Moonshine A, Mohammadi P, Banks E, Lappalainen T. **2015**. Tools and best practices for data processing in allelic expression analysis. *Genome Biology* **16**: 195.
Castelletti S, Tuberosa R, Pindo M, Salvi S. **2014**. A MITE Transposon Insertion Is Associated with Differential Methylation at the Maize Flowering Time QTL Vgt1. *G3: Genes, Genomes, Genetics* **4**: 805–812.
Chia J-M, Song C, Bradbury PJ, *et al.*2012. Maize HapMap2 identifies extant variation from a genome in flux. *Nature Genetics* **44**: 803–807.
Diepenbrock CH, Kandianis CB, Lipka AE, *et al.*2017. Novel Loci Underlie Natural Variation in Vitamin E Levels in Maize Grain. *The Plant Cell* **29**: 2374–2392.
Dobin A, Davis CA, Schlesinger F, *et al.*2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15–21.
Dong Z, Xiao Y, Govindarajulu R, *et al.*2019. The regulatory landscape of a core maize domestication module controlling bud dormancy and growth repression. *Nature Communications* **10**: 3810.
Flint-Garcia SA, Thuillet A-C, Yu J, *et al.*2005. Maize association population: a high-resolution platform for quantitative trait locus dissection. *The Plant Journal* **44**: 1054–1064.
Hartwig T, Banf M, Prietsch GP, *et al.*2023. Hybrid allele-specific ChIP-seq analysis identifies variation in brassinosteroid-responsive transcription factor binding linked to traits in maize. *Genome Biology* **24**: 108.
Howe KL, Contreras-Moreira B, De Silva N, *et al.*2020. Ensembl Genomes 2020—enabling non-vertebrate genomic research. *Nucleic Acids Research* **48**: D689–D695.
Huang C, Sun H, Xu D, *et al.*2018. ZmCCT9 enhances maize adaptation to higher latitudes. *Proceedings of the National Academy of Sciences* **115**: E334–E341.
Hufford MB, Seetharam AS, Woodhouse MR, *et al.*2021. De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. *Science* **373**: 655–662.
Hung H-Y, Browne C, Guill K, *et al.*2012. The relationship between parental genetic or phenotypic divergence and progeny variation in the maize nested association mapping population. *Heredity* **108**: 490–499.
Jiao Y, Peluso P, Shi J, *et al.*2017. Improved maize reference genome with single-molecule technologies. *Nature* **546**: 524–527.
Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D. **2010**. BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics* **26**: 2204–2207.
Kump KL, Bradbury PJ, Wisser RJ, *et al.*2011. Genome-wide association study of quantitative resistance to southern leaf blight in the maize nested association mapping population. *Nature Genetics* **43**: 163–168.
Lane AK, Niederhuth CE, Ji L, Schmitz RJ. **2014**. pENCODE: A Plant Encyclopedia of DNA Elements. *Annual Review of Genetics* **48**: 49–70.
Levo M, Segal E. **2014**. In pursuit of design principles of regulatory sequences. *Nature*

Reviews. Genetics **15**: 453–468.
Li H, Durbin R. **2009**. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**: 1754–1760.
Li H, Handsaker B, Wysoker A, *et al.*2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)* **25**: 2078–2079.
Li C, Sun B, Li Yongxiang, *et al.*2016. Numerous genetic loci identified for drought tolerance in the maize nested association mapping populations. *BMC Genomics* **17**: 894.
Lin G, He C, Zheng J, *et al.*2021. Chromosome-level genome assembly of a regenerable maize inbred line A188. *Genome Biology* **22**: 175.
Lorant A, Ross-Ibarra J, Tenaillon M. **2020**. Genomics of Long- and Short-Term Adaptation in Maize and Teosintes In: Dutheil JY, ed. Methods in Molecular Biology. *Statistical Population Genomics*. New York, NY: Springer US, 289–311.
Love MI, Huber W, Anders S. **2014**. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**: 550.
Mao H, Wang H, Liu S, *et al.*2015. A transposable element in a NAC gene is associated with drought tolerance in maize seedlings. *Nature Communications* **6**: 8326.
Marand AP, Eveland AL, Kaufmann K, Springer NM. **2023**. cis-Regulatory Elements in Plant Development, Adaptation, and Evolution. *Annual Review of Plant Biology* **74**: 111–137.
McSteen P, Hake S. **2001**. barren inflorescence2 regulates axillary meristem development in the maize inflorescence. *Development* **128**: 2881–2891.
Meng X, Muszynski MG, Danilevskaya ON. **2011**. The FT-like ZCN8 Gene Functions as a Floral Activator and Is Involved in Photoperiod Sensitivity in Maize. *The Plant Cell* **23**: 942–960.
Noshay JM, Liang Z, Zhou P, *et al.*2021. Stability of DNA methylation and chromatin accessibility in structurally diverse maize genomes. *G3: Genes|Genomes|Genetics* **11**: jkab190.
Noshay JM, Marand AP, Anderson SN, *et al.*2021. Assessing the regulatory potential of transposable elements using chromatin accessibility profiles of maize transposons (K Bomblies, Ed.). *Genetics* **217**: 1–13.
O'Malley RC, Huang SC, Song L, *et al.*2016. Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. *Cell* **165**: 1280–1292.
Peiffer JA, Romay MC, Gore MA, *et al.*2014. The Genetic Architecture Of Maize Height. *Genetics* **196**: 1337–1356.
Pressoir G, Brown PJ, Zhu W, *et al.*2009. Natural variation in maize architecture is mediated by allelic differences at the PINOID co-ortholog barren inflorescence2. *The Plant Journal* **58**: 618–628.
Purcell S, Neale B, Todd-Brown K, *et al.*2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics* **81**: 559–575.
Quinlan AR, Hall IM. **2010**. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842.
Regulski M, Lu Z, Kendall J, *et al.*2013. The maize methylome influences mRNA splice sites and reveals widespread paramutation-like switches guided by small RNA. *Genome Research* **23**: 1651–1662.
Ricci WA, Lu Z, Ji L, *et al.*2019. Widespread long-range cis-regulatory elements in the maize genome. *Nature Plants* **5**: 1237–1249.
Rodgers-Melnick E, Vera DL, Bass HW, Buckler ES. **2016**. Open chromatin reveals the functional maize genome. *Proceedings of the National Academy of Sciences of the United States of America* **113**: E3177-3184.
Salvi S, Sponza G, Morgante M, *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**: 11376–11381.
Savadel SD, Hartwig T, Turpin ZM, *et al.*2021. The native cistrome and sequence motif families of the maize ear. *PLOS Genetics* **17**: e1009689.

Sharon E, Chen S-AA, Khosla NM, Smith JD, Pritchard JK, Fraser HB. **2018**. Functional Genetic Variants Revealed by Massively Parallel Precise Genome Editing. *Cell* **175**: 544-557.e16.
Shikha M, Kanika A, Rao AR, Mallikarjuna MG, Gupta HS, Nepolean T. **2017**. Genomic Selection for Drought Tolerance Using Genome-Wide SNPs in Maize. *Frontiers in Plant Science* **8**: 550.
Song B, Buckler ES, Wang H, *et al.*2021. Conserved noncoding sequences provide insights into regulatory sequence and loss of gene expression in maize. *Genome Research*: gr.266528.120.
Speed D, Hemani G, Johnson MR, Balding DJ. **2012**. Improved Heritability Estimation from Genome-wide SNPs. *American Journal of Human Genetics* **91**: 1011–1021.
Springer NM, Anderson SN, Andorf CM, *et al.*2018. The maize W22 genome provides a foundation for functional genomics and transposon biology. *Nature Genetics* **50**: 1282–1288.
Springer NM, Stupar RM. **2007**. Allele-Specific Expression Patterns Reveal Biases and Embryo-Specific Parent-of-Origin Effects in Hybrid Maize. *The Plant Cell* **19**: 2391–2402.
Studer A, Zhao Q, Ross-Ibarra J, Doebley J. **2011**. Identification of a functional transposon insertion in the maize domestication gene tb1. *Nature Genetics* **43**: 1160–1163.
Sturm M, Schroeder C, Bauer P. **2016**. SeqPurge: highly-sensitive adapter trimming for paired-end NGS data. *BMC Bioinformatics* **17**: 208.
Sun G, Yu H, Wang P, *et al.*2023. A role for heritable transcriptomic variation in maize adaptation to temperate environments. *Genome Biology* **24**: 55.
Tian D, Wang P, Tang B, *et al.*2020. GWAS Atlas: a curated resource of genome-wide variant-trait associations in plants and animals. *Nucleic Acids Research* **48**: D927–D932.
Tian J, Wang C, Xia J, *et al.*2019. Teosinte ligule allele narrows plant architecture and enhances high-density maize yields. *Science* **365**: 658–664.
Tu X, Mejía-Guerra MK, Valdes Franco JA, *et al.*2020. Reconstructing the maize leaf regulatory network using ChIP-seq data of 104 transcription factors. *Nature Communications* **11**: 5089.
Wallace JG, Bradbury PJ, Zhang N, Gibon Y, Stitt M, Buckler ES. **2014**. Association Mapping across Numerous Traits Reveals Patterns of Functional Variation in Maize. *PLOS Genetics* **10**: e1004845.
Wen W, Jin M, Li K, *et al.*2018. An integrated multi-layered analysis of the metabolic networks of different tissues uncovers key genetic components of primary metabolism in maize. *The Plant Journal* **93**: 1116–1128.
White MA, Myers CA, Corbo JC, Cohen BA. **2013**. Massively parallel in vivo enhancer assay reveals that highly local features determine the cis-regulatory function of ChIP-seq peaks. *Proceedings of the National Academy of Sciences of the United States of America* **110**: 11952–11957.
Wills DM, Whipple CJ, Takuno S, *et al.*2013. From Many, One: Genetic Control of Prolificacy during Maize Domestication. *PLOS Genetics* **9**: e1003604.
Wolter F, Puchta H. **2018**. Application of CRISPR/Cas to Understand Cis- and Trans-Regulatory Elements in Plants. *Methods in molecular biology (Clifton, N.J.)* **1830**: 23–40.
Wu X, Feng H, Wu D, *et al.*2021. Using high-throughput multiple optical phenotyping to decipher the genetic architecture of maize drought tolerance. *Genome Biology* **22**: 185.
Xie Z, Nolan T, Jiang H, *et al.*2019. The AP2/ERF Transcription Factor TINY Modulates Brassinosteroid-Regulated Plant Growth and Drought Responses in Arabidopsis. *The Plant Cell* **31**: 1788–1806.
Xu G, Lyu J, Li Q, *et al.*2020. Evolutionary and functional genomics of DNA methylation in maize domestication and improvement. *Nature Communications* **11**: 5539.
Zhao H, Sun Z, Wang J, Huang H, Kocher J-P, Wang L. **2014**. CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics* **30**: 1006–1007.

Zhao H, Zhang W, Zhang T, *et al.*2020. Genome-wide MNase hypersensitivity assay unveils distinct classes of open chromatin associated with H3K27me3 and DNA methylation in Arabidopsis thaliana. *Genome biology*.

Materials and Methods

Plant materials

B73, Mo17, A619, W23, W22, A188, and US-NAM seeds were supplied by the GRIN National Agricultural Library. Seeds were pre-germinated for 48h at 28-30 ℃. Each pot contained soil equalized by volume and four seedlings. Plants were grown side by side in greenhouses, under long-day conditions (16h day/8h night, 28-30 ℃) for approximately 26 days until 75% of the plants per pot showed the formation of the leaf 4 auricle. Plants were then randomized and 12 pots per treatment were grown either with or without periodic watering through a bottom drench system for 86h. At this time point, plants treated without periodic watering were scored for drought related leaf-rolling and wilting phenotypes using a scoring from 1-5 (Additional file 1, Fig. S21). Plants were then harvested and the leaf blades of the oldest leaf without an auricle were immediately frozen in liquid nitrogen.

Relative water content measurements

At harvest, a 3 cm leaf sheath section below the third leaf ligule was removed for each plant. Four sheath sections (one of each plant in the pot) were processed together to yield one RWC value per pot. First, the fresh weight (FW) of the pieces was determined, followed by determination of the turgor weight (TW) after 24h in water at 4 degrees. The dry weight (DW) was measured after several days at 60 degrees and the relative water content (RWC) was calculated as ((FW-DW)/(TW-DW))*100.

MOA-seq and RNA-seq sample preparation and sequencing

MOA-seq sample and library preparation was performed using 1 g of frozen, finely ground leaf blade tissue following the procedure described in (Savadel *et al.* 2021). RNA-seq of hybrid leaf blade tissue was performed as described in (Hartwig *et al.* 2023).

MOA-seq data analysis

Reads were filtered using SeqPurge (v2022-07-15) with parameters “-min_len 20 -qcut 0” (Sturm *et al.* 2016). Due to the short fragment length in MOA, read pairs completely overlapped. MOA-seq paired-end reads were merged into single-end reads, including base quality score correction, using NGmerge (v0.3) (Gaspar 2018) with parameters “-p 0.2 -m 15 -d -e 30 -z -v”. Diploid genomes were created by concatenating the B73 V5 genome with the respective paternal genome (NAM v1/2 genomes (Hufford *et al.* 2021), Mo17 CAU v1, W22 v2 (Springer *et al.* 2018), A188 v1 (Lin *et al.* 2021), and A619 see separate entry). Reads were mapped to the diploid genome (or the separate genomes for inbred data) using STAR (v2.7.7a). As STAR is designed to map RNA, we set the flag --alignIntronMax 1 for DNA (no introns allowed) as well as parameters “--outSAMmultNmax 2, --winAnchorMultimapNmax 100, and -outBAMsortingBinsN 5 (Dobin *et al.* 2013).

Bam files were converted to bed format using bamToBed (Quinlan and Hall 2010). Format conversion and the calculation of the average mapped fragment length (AMFL) was done using SAMtools (v1.9) (Li *et al.* 2009). The effective genome size was calculated using unique-kmers.py (https://github.com/dib-lab/khmer/) with AFML and respective genome fasta as inputs.

To generate fragment-center tracks, each read was shortened to 20 bp centered around the middle of the read. Read shortening was performed using awk: for reads with uneven number of bases, the middle base was taken and then read extended 10 bp to each site. For reads with even numbers of bases, one of the two middle bases was chosen randomly and the reads were extended 10 bp to each site. The function genomeCoverageBed of Bedtools (v2.29.0) was then used to convert the bed files to bedgraph, scaled by the quotient of the effective genome size

and the number of uniquely mapped reads (reads per genome coverage, RPGC). BigWig files for visualization were generated using bedGraphToBigWig (Kent *et al.* 2010).

MOA-seq WGS control

WGS reads for all lines were downloaded from NCBI SRA (Jiao *et al.* 2017; Springer *et al.* 2018; Lin *et al.* 2021; Hufford *et al.* 2021). All reads were quality trimmed using seqtk (https://github.com/lh3/seqtk) trimfq (v1.3 r106), then cropped to 65 bp length using trimmomatic (v0.39) (Bolger *et al.* 2014) and 600 million reads per haplotype (except B73xIL14H with 300 million reads each) were extracted at random using seqtk sample (seed -s 100). Per F1, B73 and parental reads were merged and mapped as single-end reads to the F1 genome using STAR (2.7.7a) with the same parameters as for MOA reads above. Bam files were converted to bigwig/bedgraph using BamCoverage (v) with the same parameters as for MOA reads above.

Personalized A619 genome

The maize inbred A619 does not have a reference assembly. Of the inbred maize lines with high quality assemblies, A619 is most closely related to Oh43. (Flint-Garcia *et al.* 2005). We downloaded approximately 30x coverage of A619 WGS data from NCBI SRA (SRR8997919, SRR10127976, SRR8907067, SRR5725670, SRR5663982, and SRR5663981) and then mapped it to the Oh43 genome using bwa-mem2 (v2.2.1) (Li and Durbin 2009). Only uniquely mapping reads (q30) were retained. The GATK4 pipeline in combination with the “best practices workflow” (https://gatk.broadinstitute.org/hc/en-us/sections/360007226651) was used to identify initial A619/Oh43 SNPs and INDELs with the hard filtering settings: “QD < 2.0”, "QUAL < 100.0", "SOR > 3.0", "FS > 60.0", and "MQ < 40.0". We used these initial variants for base- and

variant-recalibration, followed by two additional rounds of training based on convolutional neural networks (CNN) to optimize filtration settings (training input setting: “prior=15.0" and “tranche=99”) and, ultimately selected a refined set of biallelic, homozygous A619/Oh43 SNPs and INDELs. G2Gtools (v. 0.2.7, https://churchill-lab.github.io/g2gtools/) was employed to integrate SNPs and INDELs into the Oh43 genome and generate a A619 pseudo-reference fasta and gff3 file. The average SNP mismatch rate of respective MOA reads from the F1 to the B73xA619 genome was 0.1%, only twice that of the average high-quality B73xNAM genomes (0.04-0.05%), and significantly lower than mapping to either B73 or the A619 psuedo-reference alone (~0.54%).

Peak calling

For peak calling MOA bam files were used with MACS 3 (v3.0.0a7, https://github.com/macs3-project/MACS) using the following parameter: -s and --min-length “AMFL”, --max-gap “2x AMFL”, –nomodel, –extsize “AMFL”, --keep-dup all, -g “effective genome size”.

Treatment-specific peak calling

MOA-alignment bam files were converted to bed format using bedtools bamToBed (v2.29.0). The genomeCoverage function of bedtools was used to convert pooled replicated bed files to bedgraph with the reads per million (RPM) scaling factor. The RPM-normalized coverage difference between treatments was calculated with use of intersect and subtract functions of bedtools. The resulting difference coverage counts for each treatment were used to create an unbinned (1 bp resolution) bed file to produce a bigwig coverage track and as input for MACS 3

(v3.0.0a7) peak calling using parameters optimized for our data: --min-length 60, --max-gap 30, –nomodel, –extsize 1, --keep-dup all, -g “effective genome size”, -q .01. To further statistically validate the MOA difference peaks, we applied a Welch t-test to the individual bio replicate coverages. The coverage function of bedtools was used to compute replicate reads at each peak and were scaled with RPM factor. With three replicates for each treatment, a Welch t-test was done to compare the average normalized read count at each peak. Peaks with a p-value < 0.05 were retained.

DNA methylation analysis

Parental DNA methylation data of the NAM lines (Hufford *et al.* 2021) was obtained from iplant (/iplant/home/maizegdb/maizegdb/NAM_PROJECT_JBROWSE_AND_ANALYSES). Methylation data for non-NAM lines (Noshay, Liang, *et al.* 2021; Lin *et al.* 2021) were obtained as SRA archives (Bioprojects PRJNA657677 and PRJNA635654) and processed as described in Hartwig *et al.* (2023). B73 x Mo17 hybrid methylation data was previously published in Hartwig *et al.* (2023). Context specific methylation around AMPs/MPs was determined as described in Hartwig et al. (2023), significant differences in DNA methylation were determined following following Regulski *et al.* (2013) (one allele <10% methylated and the other >70%). Sites for testing consistency of DNA methylation/ haplotype specific binding relations were selected based on having at least two F1 lines differentially methylated, at least two F1 lines equally methylated, and at least one F1 line AMP at the given site. In this analysis (Fig. 3c), a more stringent definition of equal methylation (as opposed to not being differentially methylated) was employed: Equal methylation was defined as both alleles <10% or both >70% methylated.

Local association mapping to map bQTLs

The binding ratio of the MOA peaks, CH, CHG and CHH binding, and the read depth were collected separately for all hybrid lines for the well-watered and drought-stressed conditions. The binding frequency at loci with no reads were set to NA. Genotyping information and the methylation ratio information for $_{m}$CG, $_{m}$CHG, and $_{m}$CHH were used together with the read depth to conduct local association studies. All MPs with the respective haplotype-specific MOA coverage (binding frequency) and average surrounding (+/- 20 bp) methylation ratios were sampled and used to run five different mixed linear models for each MP, where:

$$MOApeak = Marker_i + CG + CHG + CHH + Readdepth$$

$$MOApeak = Marker_i + CHH + Readdepth$$

$$MOApeak = Marker_i + CG + Readdepth$$

$$MOApeak = Marker_i + CHG + Readdepth$$

$$MOApeak = Marker_i + Readdepth$$

All factors, except the read depth, were used as fixed effects. Associated MPs at a false discovery rate of 5% were selected to calculate the explained variance for all five models separately. The explained variance and probability values for each MP were compared across the five models genome-wide to identify any superior prediction ability associated with incorporating any methylation information. The analyses were performed in Julia 1.8.1 and R 4.1.2. The built-in linear model in R and the mixed linear model from the R package lme4 were utilized to estimate predictions of the MOA peaks.

RNA-seq analysis

RNA-seq data was mapped to the concatenated diploid genomes using STAR (v2.7.7a), with options --outSAMmultNmax 1, --outFilterMultimapNmax 1, --winAnchorMultimapNmax 100, --twopassMode Basic, --outFilterIntronMotifs RemoveNoncanonical, --outFilterType BySJout, --quantMode GeneCounts and using concatenated gff3 file containing gene models from both parents. To determine allele-specific transcript abundance, for each line, B73 and corresponding paternal positions for all SNPs determined by halSnps were generated via halLiftover. Of the resulting position pairs, ambiguous ones were removed. Each SNP was then assigned the B73 genes it overlaps with. The respective NAM gene info was added using a Pan-gene file (downloaded from MaizGDB), retaining strand information in both cases. Mapped read information was converted into read bed files using bamToBed and each SNP was assigned all reads overlapping with it in B73 and at the parental genome coordinates (strand-specific, separately for the three replicates). Only SNPs carrying reads in both alleles were retained to ensure that the SNP was truly located within the gene in both alleles. Afterwards, reads for each gene were counted per replicate (reads which had two or more SNPs were counted only once) and allele. For A188, where no Pan-gene entries were available, SNPs were also mapped onto the A188 gff and gene pairs were generated based on shared SNP positions. This way, B73 reads and paternal reads could be compared for differential transcript abundance analysis in DEseq2 (Love *et al.* 2014). Genes with a false discovery rate corrected p-value < 0.05 were considered allele-specific in their transcript abundance.

Variance Component Analysis Pipeline (VCAP)

To run the Variance Component Analysis Pipeline (VCAP), we required three datasets: (1) genome-wide markers across the Nested Association Mapping (NAM) population Recombinant Inbred Lines (RILs), (2) trait values across NAM RILs for each trait analyzed, and (3) coordinates for MOA peaks or bQTL SNPs across founder lines to partition each component.

For the genome-wide markers, we used publicly available resequencing SNPs from the NAM founders (Hufford *et al.* 2021) that had been projected onto the NAM RILs (Cyverse:/iplant/home/shared/NAM/Misc/NAM-SV-projected-V8/). Trait data collected from the NAM RILs (n=143) were curated from publications (Supplemental file NAM_phenotype_metadata.xlsx). We used two sampling schemes to create our MOA partitions. First, we randomly sampled 1 SNP from each MOA peak called from plants under well-watered conditions, resulting in ~1 million SNPs. Second, only the bQTL SNPs were used to represent MOA, resulting in ~100K SNPs. Any SNP outside of a MOA peak created the non-MOA pool from which the background SNPs were drawn. Because of the non-random distribution of MOA peaks throughout the genome, we also included a matched background component: each MOA SNP was matched to a random non-MOA SNP by allele frequency (number of lines containing the alt allele/total lines without missing data at that position, 0.1 bin size) and distance from the nearest gene (TSS or TTS).

This sampling and matching of MOA and background SNPs created the set of SNPs for estimating a set of kinship matrices for a single VCAP run. Kinship matrices were created for the MOA SNPs, BKGD SNPs, and the rest of the genome (remaining non-MOA and non-BKGD SNPs) using Tassel 5 (Bradbury *et al.* 2007). Under each sampling method (1 SNP per peak or bQTL strict), we sampled 100 times, creating 100 permutations of kinship matrix sets. To calculate heritabilities of all 143 traits, each set of kinship matrices and traits were run through a REML model using LDAK (Speed *et al.* 2012). Thus, the permutations gave us a range of heritability estimates that could result from these particular components, traits, and the population (Figure 3). For the strictly bQTL SNP runs, the same bQTL SNPs were used in every permutation while the background SNPs differed across permutations.

To evaluate the reliability of our heritability estimation method, we simulated traits with defined contributions from specific sets of kinship matrices and compared estimates of the heritabilities generated by the above VCAP protocol. We used one of our previously generated kinship

matrix sets (1 SNP per peak sampling) to simulate traits assigned certain heritabilities for each component (four sets of heritabilities, 10 traits per set). We simulated traits as the sum of four normally distributed random vectors, each with zero mean and covariance equal to one of the three kinship matrices or the identity matrix (for residual variation) multiplied by a specific heritability value. The simulated traits and kinship matrices were used in the REML modeling step to estimate heritabilities. Estimated heritabilities were then compared to known heritabilities. All scripts written for the analyses in the study were deposited at https://github.com/Snodgras/MOA_Analysis.

MOA bQTL and eQTL linkage analysis

Linkage disequilibrium was calculated between the binding QTL reported in this study and a set of 10,618 cis-eQTL identified based on expression data the roots of 340 maize genotypes (Sun *et al.* 2023). Genomic coordinates of the 98,383 binding QTL on the B73_RefGen_V5 maize genome were converted to B73_RefGen_V4 positions using CrossMap as implemented in EnsemblPlants (Zhao *et al.* 2014; Howe *et al.* 2020). 99.5% of binding QTL positions could be successfully converted to B73_RefGen_V4 positions and, of these, 36,777 were present in a set of 12,191,984 genetic markers segregating in the population of 340 maize lines used to conduct eQTL analysis with a minor allele frequency ≥ 0.05 and less than 2% of genotypes exhibiting heterozygous genotype calls. Linkage disequilibrium was calculated between bQTL markers and cis-eQTL markers in all cases where a cis-eQTL and a bQTL were located within ten kilobases of each other using genotype calls from the 340 maize varieties (Purcell *et al.* 2007; Sun *et al.* 2023). To assemble a control set of genetic markers with the same properties as the bQTLs, bQTL that were successfully converted to B73_RefGen_v4 and matched to Sun et al. markers were divided into ten bins based on their distance from the closest annotated transcription start site (0-1 kilobase, 1-2 kilobases, and so on), plus two additional categories for

intragenic SNPs and SNPs > 10 kilobases from the nearest annotated gene. A random subset of 2 million B73_RefGen_v5 SNPs used to detect bQTL were also converted to B73_RefGen_v4 and matched to segregating markers from Sun et al., as described above. These markers were subsampled to create a second set of 36,777 control markers with representation in each of the twelve bins equal to the levels observed for the real bQTLs.

Further data processing

To obtain a narrowed down list of drought regulation related bQTLs, all MOA bQTLs (415766 ungrouped associated SNPs), were filtered for overlap with AMPs located in significantly elevated or reduced MOA occupancy regions in DS compared to WW conditions in at least two F1s. With this filter 2683 drought elevated and 10877 drought reduced region associated bQTLs were identified.

Comparisons and calculations of lists were either performed in bedtools intersect or custom awk and bash scripts. Hypergeometric tests for over/under-representation and data visualization were performed in R.

Figure legends

Fig 1. Quantitative *cis*-element occupancy analysis in B73xMo17 hybrids. a) Haplotype-specific MOA-seq flowchart: 1) Nuclei purified from diverse nested (B73 common mother) F1s are analyzed by MOA-seq, producing small, non-nucleosomal, protein-DNA interaction footprints. 2) SNPs between alleles at MOA sites allow the identification, quantification, and, in a population, association of variants coupled to occupancy of putative *cis*-elements. Allele-specific MOA footprints can be compared between treatments, e.g., well-watered versus drought. 3) Allele-specific mRNA-seq allows further characterization of functional variants associated with gene regulation. **b**) Correlation of haplotype-specific MOA-seq data at all MP loci in nuclei from B73 vs. Mo17 inbreds (X axis) vs. those from the F1 (Y axis) (Pearson correlation 0.78). MPs with (red, $p<0.05$, expected *trans*) and without (black, expected *cis*) significant differences between F1 and parental alleles. **c**) Genome-wide comparison of allelic bias (50-66.7% to one allele considered no bias, >66.7% considered biased) at B73xMo17 F1 AMP sites to inbred B73 vs. Mo17 data. **d**) Genome-wide directionality analysis, comparing biased SNPs (AMPs) detected by MOA-seq to ChIP-seq data of a single TF BZR1 (Hartwig *et al.* 2023) in the B73xMo17 hybrid. In **c**) and **d**) MOA occupancy was largely consistent (red circle) between either F1 and parents or ChIP-seq, respectively, in showing parental bias towards B73 (green) or Mo17 (blue), with a smaller fraction of MOA sites showing no bias (gray) in ChIP-seq, or bias to the opposite parent/allele (B73 purple and Mo17 yellow). **e-h**) Examples of B73xMo17 MOA-seq (fragment center data, see methods) and allele-specific ChIP-seq. **e**) Average, normalized, allele-specific ChIP-seq of the ZmBZR1 TF (top two rows, black) and MOA-seq (bottom two rows, B73 green, Mo17 blue) shown upstream of *ZmPGIP2*. **g**) Average, normalized, allele-specific differences in MOA coverage upstream of *ZmBIF2* overlap with a known, “hypervariable” (Hv) *cis*-regulatory region. **f, h**) Normalized, average MOA coverage of three biological replicas at orange arrow positions in **e** and **g**. RPGC = number of

reads per bin (1 bp) / scaling factor (total number of mapped reads * fragment length) / effective genome size).

Fig 2. A Pan-cistrome of the maize leaf

a) Mapping strategies comparison for the haplotype-specific MOA analysis. B73xHP301 data was analyzed using a single B73 reference (single ref), an SNP-replaced pseudo-genome (SNP replaced) with HP301 SNPs replaced by Ns, and our dual-parent mapping strategy (dual ref) using a concatenated B73 x HP301 genome. For A619 F1 without an assembled genome, our "reference-guided" strategy (A619 dual-ref) showed similar AMP-balanced haplotypes without reference bias. **b**) Percent of B73 MOA peak covered by MPs (black) and AMPs (red) relative to the number of F1s. **c**) Density of mean MOA binding frequencies at SNP positions where at least one, two, three or four lines had AMPs, compared to a control with randomized binding frequencies. **d**) Overview of bQTLs (red arrows), MOA coverage (blue), and HiC interaction sites (black lines) near the classical flowering repressor RAP2.7. MOA bQTLs overlap both known enhancers, *vgt1* and *vgt1-DMR* (orange bars), associated with *RAP2.7* expression. An additional MOA bQTL, termed *vgt1-MOA* (purple), also interacts with *vgt1* and the RAP2-7 promoter. **e**) Haplotype-specific footprint occupancy (%B73) at the *TRE1* (Zm00001eb021270) proximal promoter bQTL. B73 and paternal occupancy in green and blue, respectively. **f**) Fold change (FC) of mRNA abundance (NAM/B73) for *TRE1* (Zm00001eb021270). **e-f**) All F1s for which both haplotype-specific MOA and mRNA could be determined shown from left to right: CML277, Ki3, Ms71, NC358, Oh7b, Tx303, CML103, CML247, CML322, HP301, IL14H, Mo17, Mo18W, P39, and W22. **g**) ASE and non-ASE genes are significantly more and less enriched for AMPs in their 3kb promoter upstream of the TSS, respectively. **h**) MOA bQTLs are more often in linkage disequilibrium with *cis*-expression QTLs, identified in roots of 340 recombinant inbred lines than matched bgSNPs. **i**) Average, normalized MOA coverage (CPM) for B73 (top) and NAM (bottom) alleles of B73xMo18W, B73xKy21, and B73xCML69 upstream of phosphoglycerate mutase (PGM1, Zm00001eb196320). Compared to B73, the Mo18W and Ky21 alleles showed significantly higher MOA occupancy upstream of PGM1 (green panels). In

contrast, the CML69 allele showed similar MOA coverage compared to B73, yet peaks were shifted ~300 bp due to a mite TE insertion (orange panel). **k**) Allele-specific mRNA levels (log2 FC) of PGM1 in the different F1 hybrids. mRNA categories: NAM > B73 (green), B73 = NAM (purple), and B73 < NAM (orange).

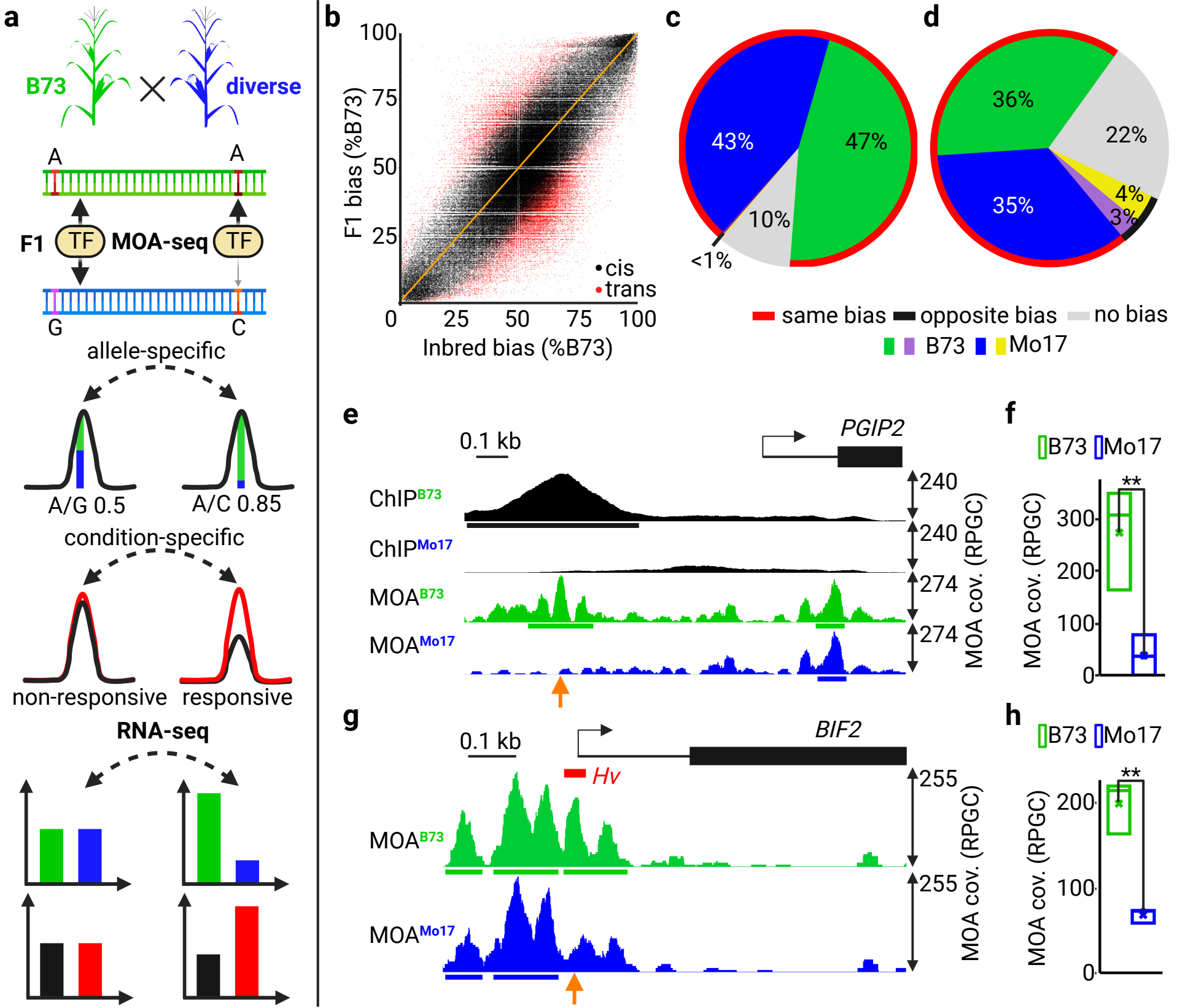
Fig 3. bQTLs are linked to variation in DNA methylation and traits

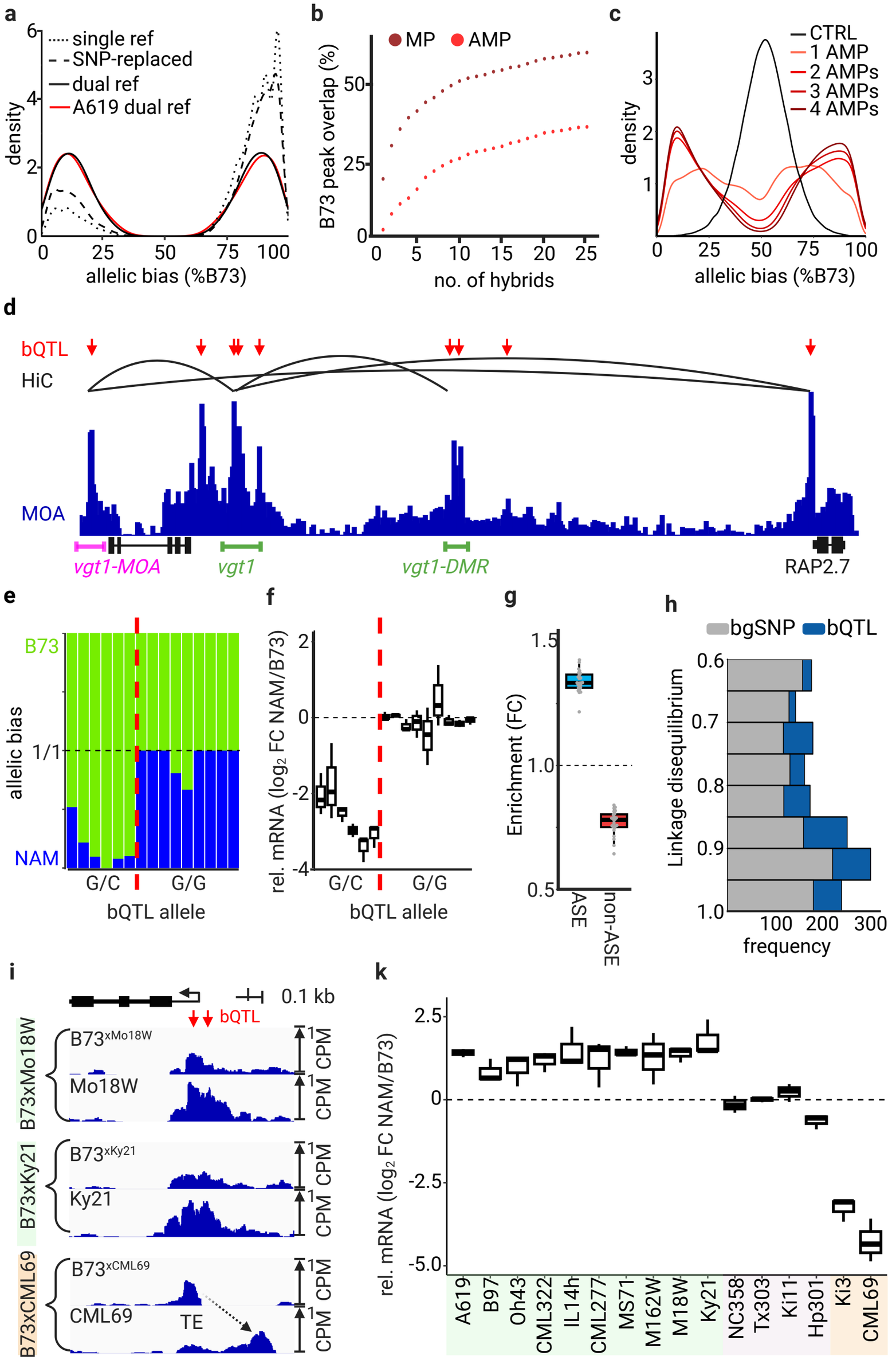
a) Genome-wide overlap of differentially CG methylated DNA regions with MPs, AMPs, and AMPs that showed a strong (>85%) allelic bias, across the 24 F1s. **b**) Correlation of differentially CG methylated DNA across the allelic bias for MPs in the 24 F1 hybrids. MP methylation categories: equal (em, blue), B73 hyper- vs. NAM hypo- (red), or NAM hypo- vs. B73 hyper- (blue) methylated. **c**) Correlation between MOA footprint bias and differential methylation at loci which varied both in allele-specific footprint occupancy (≥1 F1s with AMP) and CG methylation (≥2 F1s with and without allele-specific methylation difference) between the 24 F1s. Box and violin plots of percentage distribution of F1s with haplotype-specific binding (AMPs) at these positions, partitioned into F1s with either differential allelic CG methylation (red) or equal CG methylation (blue). **d**) Association of ~42k curated GWAS hits (+/− 100 bp, (Tian *et al.* 2020)) with bQTLs or matched bgSNPs at distances ranging from intra-genic to >20kb to the nearest gene. **e**) Estimated additive genetic variance organized by 143 traits. Colored ridges show the estimated additive genetic variance across 100 permutations for either MOA bQTLs (green), bgSNPs (orange), and rest (purple) components. Black symbols represent the mean estimated value across permutations. Traits arranged by bQTL mean variance estimates and color coded according to general trait groupings: vitamin E metabolites = navy blue, metabolites = purple, stalk strength = light blue, flowering time = gold, plant architecture = red, disease = green, tassel architecture = pink, ear architecture = orange, misc. = grey. **f**) A subset of traits (y-axis) and their estimated percent additive genetic variance (x-axis) shown as colored boxplots instead of ridges. PH=plant height (Hung *et al.* 2012; Peiffer *et al.* 2014), LeafL=leaf length (Hung *et al.* 2012), DTA=days to anthesis (Hung *et al.* 2012; Peiffer *et al.* 2014), DTS=days to silking (Hung *et al.* 2012), SLB=southern leaf blight (Kump *et al.* 2011; Bian *et al.* 2014), and delta-tocopherol concentration = vitamin E biosynthesis (Diepenbrock *et al.* 2017).

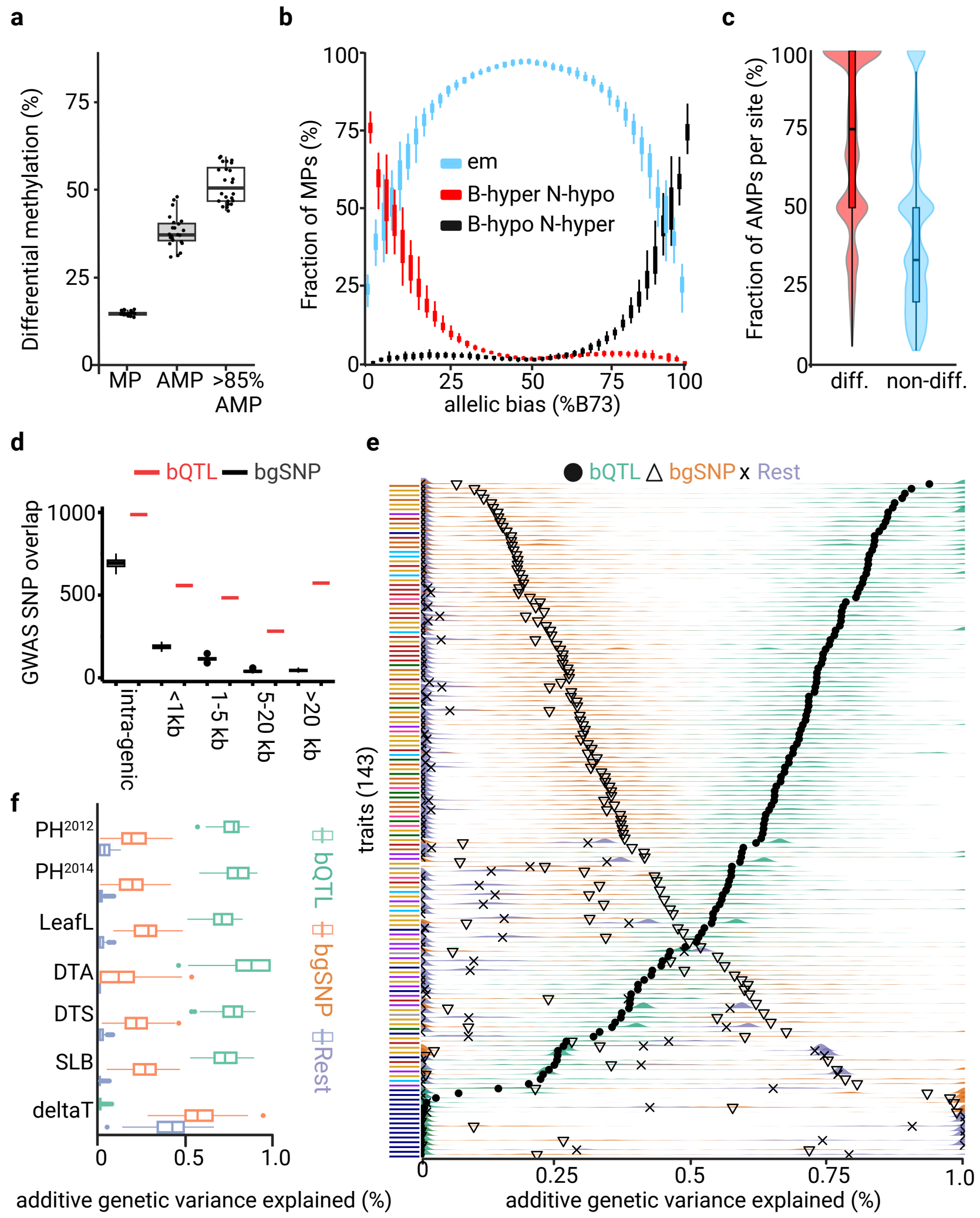
Fig 4. MOA-seq detects drought-responsive *cis*-regulaotry loci

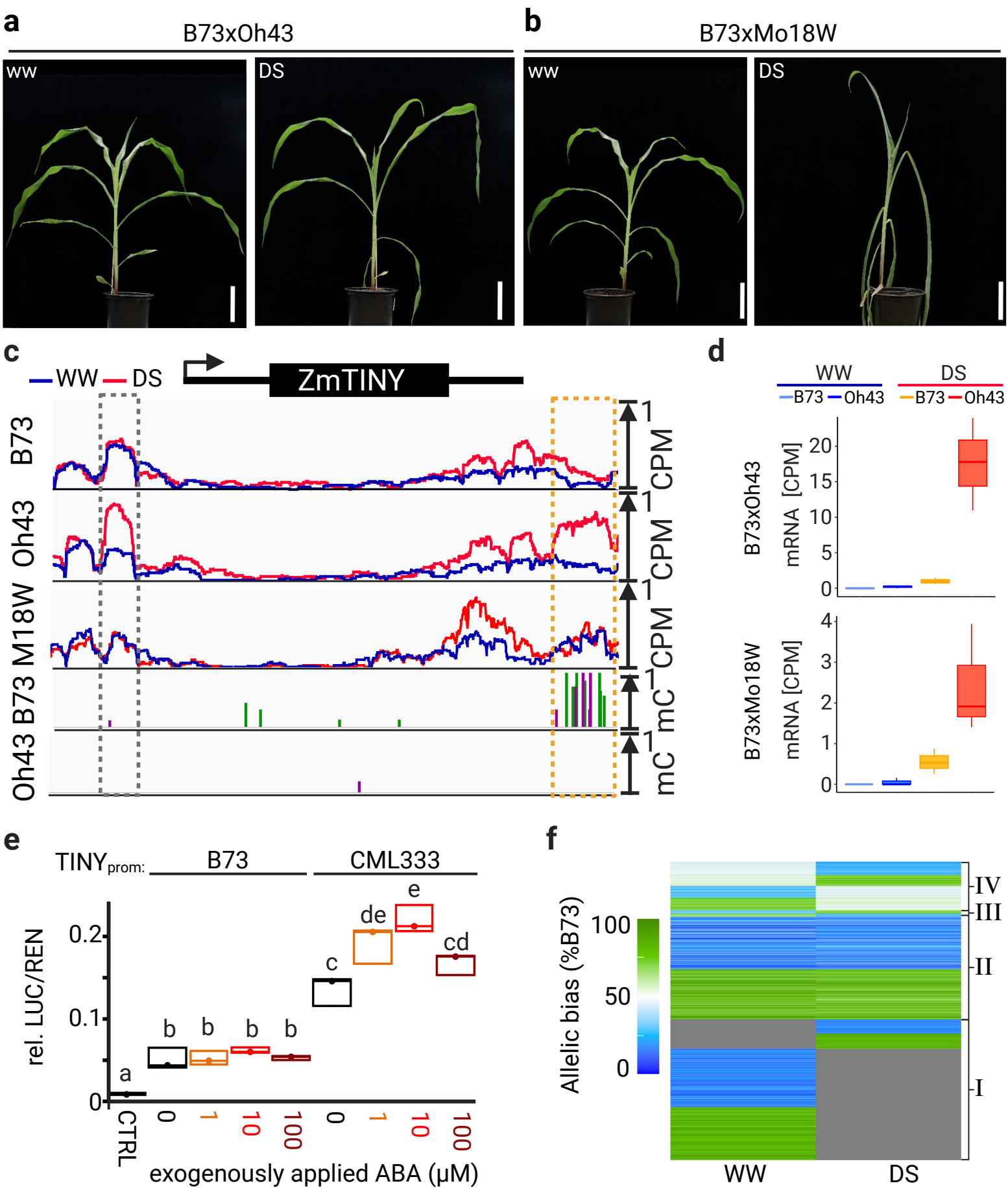
a-b) Morphological phenotypes of the more tolerant and susceptible B73xOh43 (**a**) and B73xMo18W (**b**) F1 lines, respectively, grown under well-watered (WW) and drought (DS) conditions. **c**) Average, normalized MOA coverage (CPM) for B73xOh43 and B73xMo18W F1s under WW (blue line) and DS (red line) conditions, surrounding *ZmTINY* (Zm00001eb120590). In response to the drought conditions, the Oh43 allele showed drought-induced increased MOA coverage not shared by B73 or Mo18W alleles (dashed boxes). Notably, the downstream drought-response MOA footprints overlap with hypomethylation in Oh43 and CML333, but hypermethylation lines with low MOA coverage, such as B73 and Mo18W (bottom two tracks). CG (green) and CHG (purple) methylated sites. **d**) Normalized (weighted FPKM) AS transcript levels of *ZmTINY* for B73xOh43 (top panel) and B73xMo18W (bottom panel) under WW and DS conditions. The mRNA levels of the Oh43 and Mo18W alleles were 34 and 4 fold higher under DS, respectively, when compared to their B73 allele. **e**) Relative luciferase levels in maize leaf protoplasts harboring either 1.5kb of the B73 or CML333 proximal promoter allele upstream of the TSS of *ZmTINY* with and without 1-100 μM ABA treatment. **f**) Heatmap of MOA allelic bias at AMP loci under well watered and drought stress conditions. Color scale ranging from green (100% bias towards B73) to blue (100% bias to Oh43), gray represents no binding (below threshold for calling MPs). Only MPs that are AMPs in at least one of the conditions and are located in regions of significant change in MOA occupancy between WW and DS conditions are displayed here for the drought resistant F1 B73xOh43. Four main categories can be observed: allele-specific binding in one condition and no binding in the other (I), allele-specific binding with a binding bias in the same direction in both conditions (II), allele-specific binding with bias in opposite direction in the two conditions (III) and allele-specific binding in one condition and

equal binding (BF>0.4 and <0.6) in the other (IV). Similar patterns of bias were observed for all lines and all AMPs (Additional File 1, Fig. 22).


a
B73
diverse
A
A
F1
TF
MOA-seq
TF
G
C
allele-specific
A/G 0.5
A/C 0.85
condition-specific
non-responsive
responsive
RNA-seq
b
F1 bias (%B73)
Inbred bias (%B73)
cis
trans
c
43%
47%
10%
<1%
d
36%
22%
35%
4%
3%
same bias
opposite bias
no bias
B73
Mo17
e
0.1 kb
PGIP2
ChIP^B73
ChIP^Mo17
MOA^B73
MOA^Mo17
MOA cov. (RPGC)
f
B73
Mo17
MOA cov. (RPGC)
g
0.1 kb
BIF2
Hv
MOA^B73
MOA^Mo17
MOA cov. (RPGC)
h
B73
Mo17
MOA cov. (RPGC)


a
single ref
SNP-replaced
dual ref
A619 dual ref
density
allelic bias (%B73)
b
MP
AMP
B73 peak overlap (%)
no. of hybrids
c
CTRL
1 AMP
2 AMPs
3 AMPs
4 AMPs
density
allelic bias (%B73)
d
bQTL
HiC
MOA
vgt1-MOA
vgt1
vgt1-DMR
RAP2.7
e
B73
NAM
allelic bias
1/1
G/C
G/G
bQTL allele
f
rel. mRNA (log2 FC NAM/B73)
G/C
G/G
bQTL allele
g
Enrichment (FC)
ASE
non-ASE
h
bgSNP
bQTL
Linkage disequilibrium
frequency
i
0.1 kb
bQTL
B73xMo18W
B73xKy21
B73xCML69
CPM
TE
k
rel. mRNA (log2 FC NAM/B73)
A619
B97
Oh43
CML322
IL14h
CML277
MS71
M162W
M18W
Ky21
NC358
Tx303
Ki11
Hp301
Ki3
CML69


a
Differential methylation (%)
0
25
50
75
MP
AMP
>85% AMP
b
Fraction of MPs (%)
0
25
50
75
100
em
B-hyper N-hypo
B-hypo N-hyper
allelic bias (%B73)
c
Fraction of AMPs per site (%)
diff.
non-diff.
d
bQTL
bgSNP
GWAS SNP overlap
1000
500
intra-genic
<1kb
1-5 kb
5-20 kb
>20 kb
e
bQTL
bgSNP
Rest
traits (143)
0
0.25
0.5
0.75
1.0
additive genetic variance explained (%)
f
PH2012
PH2014
LeafL
DTA
DTS
SLB
deltaT
bQTL
bgSNP
Rest
0
0.5
1.0
additive genetic variance explained (%)


a
B73xOh43
ww
DS
b
B73xMo18W
ww
DS
c
WW
DS
ZmTINY
B73
Oh43
M18W
B73
Oh43
CPM
mC
d
WW
DS
B73
Oh43
B73
Oh43
B73xOh43
B73xMo18W
mRNA [CPM]
e
TINYprom:
B73
CML333
rel. LUC/REN
CTRL
exogenously applied ABA (µM)
f
Allelic bias (%B73)
WW
DS
IV
III
II
I