Open chromatin reveals the functional maize genome

Eli Rodgers-Melnick, Daniel L. Vera, Hank W. Bass, and Edward S. Buckler

Contributed by Edward S. Buckler, April 18, 2016 (sent for review December 21, 2015; reviewed by Paul Chomet, Sylvain Glémin, and Peter M. Visscher)

Cellular processes mediated through nuclear DNA must contend with chromatin. Chromatin structural assays can efficiently integrate information across diverse regulatory elements, revealing the functional noncoding genome. In this study, we use a differential nuclease sensitivity assay based on micrococcal nuclease (MNase) digestion to discover open chromatin regions in the maize genome. We find that maize MNase-hypersensitive (MNase HS) regions localize around active genes and within recombination hotspots, focusing biased gene conversion at their flanks. Although MNase HS regions map to less than 1% of the genome, they consistently explain a remarkably large amount (~40%) of heritable phenotypic variance in diverse complex traits. MNase HS regions are therefore on par with coding sequences as annotations that demarcate the functional parts of the maize genome. These results imply that less than 3% of the maize genome (coding and MNase HS regions) may give rise to the overwhelming majority of phenotypic variation, greatly narrowing the scope of the functional genome.

Significance

The maize genome, similar to those of most plant genomes, is 98% noncoding. Much of the remainder is a vast desert of repeats that remain repressed throughout the cell cycle. The plant cell orchestrates its complex activities by restricting access to functional regions with an open chromatin configuration. Here, we identify the small portion (~1%) of the maize genome residing in open chromatin. We demonstrate that open chromatin predicts molecular phenotypes such as gene expression and recombination. Furthermore, we show that genetic variation within open chromatin regions accounts for ~40% of phenotypic variation in agronomic traits. By greatly narrowing the scope of the functional maize genome, this study can help to accelerate the pace of crop improvement through highly focused genomic selection and genome editing.

Author contributions: E.R.-M., D.L.V., H.W.B., and E.S.B. wrote the paper.

E.S.B. wrote the paper.


The authors declare no conflict of interest.

Data deposition: The sequence reported in this paper has been deposited in the NCBI Sequence Read Archive (accession no. SRA02258).

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1525244113/-/DCSupplemental.

www.pnas.org/cgi/doi/10.1073/pnas.1525244113
maximum likelihood scaling of the neutral phylogenetic tree with GERP (9), such that scale values below 1 indicate higher conservation than expected under neutrality. Overall, sequence in MNase HS regions is significantly, albeit slightly, more conserved than 180 bp flanking regions everywhere except coding sequence (SI Appendix, Fig. S2), indicating that purifying selection in MNase HS regions is elevated with respect to the noncoding expectation, but remains diffuse. Although slight differences in sequence conservation occur between MNase HS sequences from different tissue types, differences in scaling factors are inconsistent between functional annotations and lower in magnitude than those between MNase HS and non-MNase HS flanking regions (SI Appendix, Fig. S3).

Given the known role of chromatin structure in gene function, we anticipated a close relationship between the distribution of MNase HS and genes. On a broad scale (1 Mb), MNase HS density positively correlates with gene density (Spearman rho = 0.64; $P < 1 \times 10^{-16}$) (Fig. 1B). Approximately 80% of MNase HS regions occur outside genes, but they are highly enriched in genic flanks (Fig. 1E). A second peak, ~100 kb, in the MNase HS to gene distance frequencies aligns with the overall genomic distribution (Fig. 1E), and these may represent regulatory elements and additional gene proximal elements for unannotated genes. Approximately 70% of all mapped MNase HS sequences occur within the small (11%) portion of the intergenic space that excludes transposable elements (TEs) (SI Appendix, Table S1). Within and surrounding genes, averaged MNase HS profiles show stereotypical patterns (Fig. 1D), with prominent hypersensitivity just upstream of gene model transcription start sites (TSSs). Approximately 25% of genes contain statistically significant peaks in shoot or root MNase HS (Materials and Methods).

A prominent but broad HS peak, covering and extending beyond the 3’ ends of transcript models, is evident, whereas coding and intronic regions show much lower MNase HS.

**MNase HS Is Associated with Gene Regulation.** Using microarrays spanning 1,688 genes, Vera et al., 2014 (8) showed that gene expression levels and MNase HS strongly covary at TSSs. To examine this relationship in a less-biased and genome-wide manner, we sorted the 36,441 maize genes by their steady-state mRNA levels and used heat map analysis to inspect MNase HS profiles at the TSSs (SI Appendix, Fig. S4). We detect similar numbers of expressed genes (reads per kilobase per million ≥ 0.1) in shoot ($n = 25,549$) and root ($n = 26,020$) (SI Appendix, Table S2). Gene expression levels and signal strength for MNase HS regions show a clear genome-wide positive correlation around TSSs. Tissue-specific expression patterns are recapitulated by MNase HS signals, confirming that TSS chromatin profiling is predictive of gene expression levels (Fig. 2). TSS-associated MNase HS signals also discriminate between the expression levels of paralogs from the most recent tetraploidization event 5–12 MYA (10). Although large portions of the duplicated genome are differentially expressed (11), the level of MNase HS matches gene expression levels (SI Appendix, Fig. S5). These observations establish the relative signals of promoter MNase HS, possibly reflecting the transcription rate, as one of the best epigenomic predictors of gene activity.

**MNase HS Regions Are Associated with DNA Hypomethylation and Recombination Hotspots.** Heterochromatin in plants is usually distinguished by constitutive hypermethylation of symmetric CpG and

---

**Fig. 1.** The distribution of MNase HS regions in maize. (A) The frequency of MNase HS bases across the genome in 1-Mb windows, along with the recombination frequency. (B) The relationship between gene density and MNase HS density in 1-Mb windows. (C) The total sizes of MNase HS regions in the root and shoot and the intersection of the two tissues. (D) MNase HS base frequency within and surrounding protein coding genes. Genic elements were binned according to percentage of total element size, whereas upstream and downstream regions were binned in units of 10 bp. (E) The distribution of distances to the nearest gene boundary for intergenic MNase HS bases (blue) and for all intergenic bases (green).
CHG DNA methylation motifs, where H refers to any nucleotide except G (12), so we expected localized symmetric methylation reductions in HS regions. Plants also contain methylation at asymmetric CHH motifs, which in maize are known to generally mark sites of RNA-dependent DNA methylation (13), so we anticipated elevated CHH methylation within TEs. However, we did not anticipate a dramatic localized relationship of MNase HS regions with CHH methylation. In support of reduced symmetric methylation surrounding MNase HS regions, CpG methylation outside TEs is reduced from a rate of 70% 2 kb away from the nearest MNase HS site to 5% within the MNase HS region (Fig. 3). A similar reduction occurs for CHG motifs. Likewise, a strong MNase HS-associated hypomethylation tendency is seen within TEs, although the difference in CpG and CHG methylation is accompanied by a four times increase in CHH methylation. Strikingly, CpG methylation in coding regions differs from the general pattern in noncoding regions. Coding CpG methylation is elevated within and directly surrounding MNase HS sites, and stronger hypomethylation occurs downstream, relative to upstream, of the MNase HS region.

In maize, meiotic recombination has a strong negative relationship with CpG and CHG methylation (14), linking chromatin status with crossover formation. Investigating this relationship, we find significant and genomic context-specific MNase HS enrichment at 1–30-kb recombination hotspots relative to comparable, adjacent, but nonhotspot regions. Although MNase HS is enriched threefold to fourfold within hotspot TEs, twofold within hotspot non-TE intergenic regions, and twofold within hotspot noncoding genic regions, it is not enriched within the coding regions of recombination hotspots (SI Appendix, Fig. S6A). The positive relationship between MNase HS and recombination frequencies extends to gene-distal regions, those >5 kb from the nearest gene, where we find a high association (Spearman rho = 0.35; P < 1 × 10^{-16}) for 1–10-kb bins (SI Appendix, Fig. S7). Mean crossover enrichment in these 1–10-kb bins is 16% higher (Wilcoxon rank sum test, P = 3.6 × 10^{-5}) for shoot-specific compared with root-specific MNase HS regions (SI Appendix, Fig. S8).

**Strong, Sustained GC-Biased Gene Conversion Surrounds MNase HS Regions.** In light of the association between seedling MNase HS regions and recombination hotspots, we tested whether MNase HS regions are also enriched for tracts of GC-biased gene conversion. Conversion tracts arise when base mismatches at recombination junctions resolve in favor of G+C nucleotides relative to A+T nucleotides (15, 16). We assessed the historical influence of GC-biased gene conversion on the maize lineage, using PHASTbias with an alignment of 12 monocots and eudicots (17), although limiting the alignment to only grasses had no significant effect (SI Appendix, Fig. S9). The mean probability of biased gene conversion increases fivefold in the 2 kb surrounding MNase HS regions (SI Appendix, Fig. S6B). Tract sizes are also positively associated with MNase HS frequency within the tract and the immediate flanking regions (Spearman rho = 0.107; P < 1 × 10^{-16}). We find that the highest historical conversion frequencies are located in 300 bp surrounding regions of HS tracts within hotspot coding sequence (mean probability = 0.34) (Fig. 4B and SI Appendix, Fig. S10B). Hotspot MNase HS regions in both coding and noncoding genic regions also show a 1.5- to twofold increase in historical conversion frequency compared with adjacent control regions. No similar increases occur within intergenic regions (Fig. 4B). Thus, although hypersensitivity is most enriched within hotspot noncoding regions, genic coding regions have the highest conversion rates (Fig. 4B and SI Appendix, Table S3), we find that crossover enrichment consistently contributes the most to explained variance (46%), followed by MNase HS frequency (32%). Thus, MNase HS regions may contribute to GC-biased gene conversion beyond their contribution to recombination frequency, perhaps through an increase in mismatch repair-associated conversion resulting from an elevated mutation rate. However, the errors associated with estimates of narrow crossover enrichments in 5,000 recombinant inbred lines hinder clear delineation of GC-biased gene conversion causes.

In general, the conversion pattern surrounding MNase HS regions mirrors the trend toward increasing GC composition in the flanking 1 kb (Fig. 4A and SI Appendix, Fig. S10A). Strikingly, the GC content of putatively neutral coding sites increases to nearly 80% in the few hundred base pairs flanking coding MNase HS regions, whereas GC content within coding MNase...
GC-biased gene conversion intensity (SI Appendix, Fig. S12). Intensity is greatest (B = 2.4) in the 100–200 bp of coding sequence surrounding MNase HS within or proximal to coding sequence, although conversion intensity within the interior of coding MNase HS is also high (B = 1.9). Within genic, noncoding sequence, conversion intensity is highest within the MNase HS interior (B = 0.88), and it is also centered on MNase HS tracts within gene distal regions of both TEs (B = 1.65) and non-TE (B = 0.45) regions. Extending beyond current population conversion rates, higher historical substitution rates of A/T to G/C vs. G/C to A/T in these same regions indicate localized GC-biased gene conversion has remained a consistent evolutionary force (SI Appendix, Fig. S13). Although the mechanistic basis for enhanced conversion frequencies is unclear, we find strong associations with epigenetic marks. In particular, the frequency of the histone modification H3K9me2 mirrors base content patterns in both coding and noncoding portions of genes (Fig. 4A). In contrast, H3K4me3 shows the same pattern in coding and non-coding regions. Strikingly, the absence of H3K9me2 clearly separates the high GC (≈80%) MNase HS-flanking regions from flanking regions with comparable GC content to the MNase HS interior (≈60%) (SI Appendix, Fig. S14). H3K4me3 has a much smaller association with GC content in the opposite direction, and both effects are significant in coding and noncoding segments.

MNase Hypersensitivity Marks Known Quantitative Trait Loci. Many cloned maize quantitative trait loci (QTL) fall within nongenic regulatory regions, prompting us to examine several cases of intergenic QTL for possible underlying explanatory MNase HS regions. The first case is the QTL associated with an expression increase in teosinte branched1 (tb1), a domestication gene with alleles for the single stalk form of modern field corn (19). Variation in tb1 expression levels maps to genetic variation in a region containing two TE insertions, Hopscotch and Tourist, located ~60 kb upstream of tb1. The presence of the proximal Hopscotch element causes increased tb1 expression, whereas the distal region containing Tourist represses expression (20). Intriguingly, we observe MNase HS in the 730 bp between these two TEs (Fig. S4), possibly localizing a distal TE-modulated enhancer as the underlying tb1 domestication QTL. We note that the Hopscotch element itself, too repetitive to measure, may also contain MNase HS sequences.
A second case involves the vegetative to generative transition 1 (vgt1) QTL, associated with altered expression of a maize flowering-time repressor gene, AP2 domain transcription factor, ZmRap2.7. Insertion of a miniature inverted-repeat transposable element (MITE) into conserved noncoding sequence in the QTL is associated with reduced expression of ZmRap2.7 and early flowering time (21, 22). Correspondingly, we observe a strong MNase HS signal overlapping the MITE insertion point (SI Appendix, Fig. S15A). Finally, a third case involves the prolificacy1.1 (prol1.1) QTL, found to contain several prominent MNase HS regions (SI Appendix, Fig. S15A). This QTL is presumed to contain a cis-regulatory element that increases expression of grassy tillers1 (gt1), a gene that acts to suppress secondary ear outgrowth (23). These examples reveal the potential for chromatin structural data to narrow the focus from ~100-kb windows to small intervals as a means to help identify actual enhancers near genes.

MNase HS Regions Explain Quantitative Traits. To more globally test the relationship between MNase HS and complex trait variation, we quantified the enrichment of genome-wide association study (GWAS) hits across 41 traits within 2 kb of MNase HS regions. We find twofold enrichment of GWAS hits in MNase HS proximal regions (95% credible interval, 1.73–2.16) compared with adjacent regions at least 2 kb away from the nearest MNase HS site (Fig. 5B). Furthermore, this enrichment is nearly unchanged as a function of distance to the nearest gene (Fig. 5B), revealing comparable MNase HS efficacy for mapping functionally significant genomic loci in gene-proximal versus gene-distal regions. To further investigate the role of open chromatin in complex organismal trait variation, we partitioned the heritable phenotypic variance into annotation-specific components using methods previously applied to human variation (24). We classified SNPs into coding (CDS), MNase HS, 5′ UTR, 3′ UTR, intronic, and intergenic regions, in that order of priority. A genetic relationship matrix was constructed for each SNP set, allowing for jointly estimated heritability explained by genetic variation within each annotation. We examined the United States nested association mapping (US-NAM) population, a set of 5,000 recombinant inbred lines from B73 × 25 diverse inbreds in controlled crosses. We found that the CDS explains the highest proportion of heritable variance (mean = 47.6%) for traits with moderate to high heritability (h² > 0.4), whereas MNase HS regions explain a remarkably large majority of remaining variance (mean = 39.3%) (Fig. 6C). This 18x enrichment of variation attributed to SNPs within MNase HS (Fig. 6D) is comparable to that (16x) for CDS SNPs. Similar results, 21x enrichment for MNase HS SNPs, are obtained from variance partitioning of flowering time and plant height phenotypes of a different mapping population, the Ames Diversity Panel (Fig. 6B and D). Splitting variance among MNase HS categories, we find that the largest contribution of MNase HS heritability is intergenic for both mapping panels (SI Appendix, Fig. S16).

Considering the MNase HS variance in shoots versus roots, we expected that the shoot chromatin profiles might better explain trait variation because of the aboveground nature of the measured phenotypes. Indeed, shoot-only MNase HS regions explain an average of 60% (US-NAM panel) or 75% (Ames panel) of the MNase HS heritable variance, whereas the root-only regions explain less than 20% (SI Appendix, Fig. S17 C and D). However, given the different sizes of the SNP sets within each tissue category, the level of enrichment varies. In US-NAM, the MNase HS regions common to both root and shoot show the most enrichment (24x) compared with those unique to shoot (19x) or root (16x). In the Ames panel, the MNase HS regions unique to shoot show the most enrichment (26x) compared with those either unique to root (23x) or common to both (8.3x) (SI Appendix, Fig. S17 A and B).

Several types of reliability testing for variance partitioning were performed. We note that SNP error rates do not appear to bias the observed results, as revealed by a lack of enhancement of SNP quality scores among MNase HS SNPs relative to SNPs from other categories (SI Appendix, Fig. S18D). In addition, little to no variance components are allocated to all SNP sets when phenotypes are permuted both within and between families,
another sign of unbiased categories (SI Appendix, Fig. S1C). As a test of robustness to patterns of LD, we tested partitioning with the separate category of intergenic SNPs 2–5 kb away from the nearest MNase HS region. This category explains little heritability in either population (SI Appendix, Fig. S19 A and B). We also reversed the order of annotation priority in SNP classes, but obtained essentially the same results for the Ames panel, but with more variance allocated to the UTRs within the US-NAM panel for several phenotypes (SI Appendix, Fig. S19 C and D).

Finally, we performed variance partitioning on 250 simulated phenotypes for each annotation, in which only SNPs within that annotation contributed to phenotypic variance. Using this method, for which we know causal locations, we find that most variance is unambiguously allocated to the correct SNP set in both populations (SI Appendix, Fig. S18 A and B).

Discussion

In this study, we demonstrate that relatively open chromatin structure, mapped with MNase HS profiling, marks functionally important regions that link genotype to phenotype. MNase HS regions delineate molecular phenotypes such as recombination breakpoints, enhancers, and other possible remote controllers of gene expression. We demonstrate how to use MNase HS regions as epigenomic annotations to resolve the variation underlying organism-level quantitative traits at high resolution.

We observe more MNase HS sites, in both number and total coverage, for the shoot compared with the root sample. The elevated signal in the shoot may reflect its greater overall tissue complexity with respect to cell type differentiation and development potential (reviewed in ref. 25). Moreover, the seedling harvest time coincided with major shoot developmental events, including the onset of autotrophy, the juvenile to adult vegetative phase change, and the genome-wide epigenetic change coupled to transposon expression and gene silencing pathways (26). Taken together, the detection of more open chromatin in shoots versus roots is likely reflective of more numerous genomic activities distributed across the genome, collectively measured by DNS-seq.

Regarding gene regulation, we find a prominent and consistent positive relationship between gene expression and chromatin accessibility surrounding the TSS. The strength of this relationship is so strong that one might be able to predict transcription rates directly from MNase HS site enriching transcript levels. However, three possible scenarios are considered in which the promoter DNS signals and transcript abundance could appear to be uncoupled, yet maintain the DNS–transcription rate relationship on a per cell basis. In one case, genes exhibiting low promoter MNase HS signals may produce high mRNA levels, a pattern that could result from a gene that is very highly expressed in a small proportion of cells. In a contrasting case, a gene with high promoter MNase HS signals may produce low mRNA levels, a pattern that could result from high, posttranscriptional mRNA turnover. However, another case could involve the mapping of MNase HS regions to the wrong gene, as might occur at promoters from nearby genes. Overall, however, most genes clearly show a robust positive relationship between RNA abundance and promoter MNase HS signals, likely related to transcriptional initiation and its dependence on open chromatin.

DNA methylation rates also closely mirror the chromatin accessibility landscape. Cytosine hypomethylation of open chromatin is common across eukaryotes, as reduced methylation within CpG contexts occurs in plants and animals (1, 4, 27). In contrast to the ubiquitous hypomethylation of cytosines in symmetric motifs, we observe a twofold to fourfold increase in CHH methylation in the region immediately surrounding and MNase HS regions in TEs. CHH hypermethylation within DNase1 HS regions was previously observed within Arabidopsis thaliana (27), and it co-occurs with accessible chromatin in the RNA-dependent DNA methylation–targeted regions of maize (13). Given that RNA-dependent DNA methylation requires active transcription of siRNAs, MNase HS regions may universally mark transcriptionally active DNA, including regions coupled to genomic silencing pathways.

Beyond gene expression, we find compelling evidence that open chromatin marks recombination hotspots. This relationship differs from that in humans, where PRDM9 marks the locations of recombination hotspots but without a strong relationship to open chromatin (28, 29). An unexpected finding with possible implications for mechanisms of recombination in plants is the relationship between open chromatin and GC-biased gene conversion. In coding sequence, where we observe the highest conversion rates, the allele frequencies of derived neutral alleles with conversion advantages are increased nearly 2.5-fold above those of disfavored alleles. Moreover, the patterns of substitution within the phylogeny indicate that GC-biased gene conversion has remained a consistent, localized force around MNase HS regions. This situation differs from that in humans, where the high conversion sites are inconsistent between populations and species (18), likely because of the rapid evolution of PRDM9 motifs (30). Because GC-biased gene conversion imposes a fitness-independent selective force, slightly deleterious alleles may become fixed if their selective disadvantages are sufficiently less extreme than their conversion advantages (31). Indeed, GC-biased gene conversion may increase disease burden in humans by up to 5% (32). However, high levels of GC-biased gene conversion may also favor increased recombination when deleterious mutations are highly recessive (33). Our estimates of the population conversion intensity indicate that GC-biased gene conversion is strong enough to overcome genetic drift, especially within coding regions. Therefore, this localized, nonadaptive force may substantially increase maize genetic load and contribute to the heterotic patterns observed in breeding germplasm through complementation of these deleterious variants.

The mechanistic basis of conversion rate differences in coding versus noncoding genomic sequence is unknown. One possibility is that optimal codon selection increases GC content in coding regions. However, codon selection does not explain the conversion bias localizing around MNase HS regions, specifically those lacking H3K9me2. If the same recombination-promoting mechanisms occur in coding and noncoding segments, the differences in base content may result from histone mark stability over deep evolutionary time within coding sequence. The support for any direct relationship between H3K9me2 and recombination is nonetheless tenuous. In A. thaliana, knockout of DNA polymerase α causes localized reduction of H3K9me2 and concomitant increases in recombination frequency (34). H3K4me3 also correlates with meiotic recombination hotspots in A. thaliana (35), and from our results, shows a slight positive relationship to recombination in maize. However, one or more chromatin marks correlated with those we assayed; for instance, H2A.Z, which is strongly associated with A. thaliana hotspots (35), could be the causative factor. The interrelationships among chromatin structure, epigenetic marks, and recombination control remain largely undefined, but their investigation bears on evolutionary paradigms and agricultural breeding strategies.

Relationship to Quantitative Traits. We observe consistent, robust relationships between genetic variation in MNase HS regions and complex trait variation, establishing an epigenetic framework for the discovery and analysis of enhancers and other genome-wide regulatory sites. As a group or genomic annotation, MNase HS regions represent a promising collection of gene regulatory candidates underlying quantitative loci, including longitude expression QTL. Several lines of evidence support this idea. First, MNase HS regions colocate with several well-studied intergenic QTL previously fine-mapped as regulatory regions: thl1, vgl1, and pro11.1. Notably, for these intergenic QTL, B73 has the high expression alleles. Alternate lines with structural variation
corresponding to the less-expressed alleles may lack the same open chromatin signals as a result of structural disruption or epigenetic modification such as heterochromatin spreading. Comparative genomics using MNase HS regions will become increasingly feasible as multiple other maize inbred lines (haplotypes) are sequenced. Second, the relationship of MNase HS regions to functional polymorphisms is supported by the enrichment of GWAS SNPs within the proximity of MNase HS regions. A previous study of maize quantitative traits found evidence for significant GWAS hit enrichment in gene-proximal intergenic regions (36). However, GWAS studies can only provide evidence for causal variants in LD with tagged SNPs, so nearby genetic causal variants cannot be dismissed. In contrast, our results show twofold GWAS hit enrichment near open chromatin, regardless of genetic proximity. This twofold enrichment is smaller than the ∼20-fold enrichment for explained variation from partitioning analysis. Two nonexclusive explanations may explain this discrepancy. First, the stochastic history of mutation and recombination will produce inconsistent GWAS resolution, blurring the true location of causal mutations. Second, the relationship of GWAS hit quantity to explained variation depends on the distribution of allelic effects, possibly skewed toward larger values in open chromatin.

Our findings show that SNPs within MNase HS regions explain ∼40% of the heritable variance of quantitative traits in multiple maize mapping populations. The scale and scope of this finding are remarkable, and thus worthy of considerable scrutiny, prompting us to examine the robustness and validity of our results. First, we note that the pattern of relative contributions to heritability is consistent across traits and populations. The US-NAM panel of 5,000 lines from controlled crosses between 26 diverse founders results in large linkage blocks that permit complete imputation of SNPs. The other population, a non-designed diversity panel, contains high amounts of missing data, but both populations benefit from the comparably rapid rate of LD decay in maize (37, 38). Using permutation testing and simulated phenotypes based on empirical genotypes, we show that the high variance apportioned to MNase HS and CDS regions is not explained by intrinsic bias toward these sites. SNPs outside of a given annotation class can inflate its contribution if they explain heritability and closely flank the annotation of interest. We find no experimental distortion when comparing intergenic MNase HS regions with those from a subset 2–5 kb from the nearest MNase HS region. Nonetheless, the considerable structural variation within maize invites the possibility that MNase HS-proximal indels, rather than genetic variants within the HS region, are often the causal polymorphisms behind the explained variance. Future studies of interline MNase HS variation will therefore focus on the relationship between chromatin state, structural variation, and QTL.

Although we observe a large phenotypic contribution from the MNase HS portion of the genome (40% of heritable variance), this value is only half the heritable variance explained by DNaSe1 HS regions for 11 human diseases (24). However, as a percentage of the genome, the MNase HS regions in the current study cover more than 50-fold less sequence than the human DNaSe1 regions. Although the plant genomes currently lack open chromatin profiling data as extensive as that for the human ENCODE project (39), the results of this study, along with previous genomic DNaSe1 HS profiles of rice and A. thaliana chromatin (3, 4, 27), suggest that the genome-wide extent of open chromatin in plants may not substantially scale with increasing genome size. Furthermore, DNaSe1 I HS profiles in animals are increasingly shifting from hypersensitivity to general sensitivity, further complicating the definition, let alone comparison, of the distributions of chromatin states across kingdoms. Given that our results support intergenic functional variation explained almost exclusively by open chromatin, these chromatin accessibility assays in plants are, a priori, at least as informative as protein-coding regions when defining the functional genome.

In summary, we show how DNS mapping can be used to delineate the functional portion of a large, complex genome, using maize as a model genetic system. Biochemical DNS footprints produced in situ are highly localized, and maize offers both a low LD and extensive structured mapping populations. Combined, these organismal and experimental attributes allowed us to measure the effects of local chromatin structure on heritable phenotypic variation at an unprecedented depth and breadth, using only seedling shoot and root tissues of one genotype. Furthermore, we illuminate the relationship of open chromatin to recombination, opening the door for future studies into the targeting of crossovers and the evolutionary consequences of strong, consistent GC-biased gene conversion. In agriculture, epigenomic profiling with DNS-seq can strategically guide the predictive accuracy of genomic selection, narrow candidate regions for experimentation with reverse genetics, and define the functions of intergenic chromatin toward organismal fitness. Overall, DNS profiling has multiple applications ranging from predicting transcription rates and recombination sites to defining enhancers and QTL candidates. As a genomic annotation, they bring an invaluable resource to bear on biological, agricultural, and societal problems, including contemporary and future challenges related to population growth and climate change.

Materials and Methods

Plant Material.

Seeds from the maize (Zea mays L., cultivar B73) were obtained from field-grown ears from the Buckler laboratory at Cornell University. Seeds were germinated in Fafard Seedling Mix in the greenhouse (Department of Biological Science, Florida State University). Tissue was harvested at 11:00 AM-12:00 PM, 9 d after planting. Shoot tissue was harvested by cutting and collecting the tissue just above the soil line, and root tissue was harvested by rinsing the root system in water and cutting off the kernel-attached roots. Harvested tissues were immediately flash frozen in liquid nitrogen and stored at −80°C.

Nucleus Isolation and Digestion.

As modified from Vera et al. (8), ten grams of tissue were ground under liquid nitrogen with a mortar and pestle and cross-linked by stirring for 10 min in 100 mL ice-cold fixation buffer (15 mM Pipes-NaOH at pH 6.8, 0.32 M sorbitol, 80 mM KCl, 20 mM NaCl, 0.5 mM EGTA, 2 mM EDTA, 1 mM DTT, 0.15 mM spermine, and 0.5 mM spermidine) containing 1% formaldehyde. Fixation was stopped by adding glycine to 125 mM. Nuclei were isolated by adding Triton X-100 to 1% final by addition of 0.1 vol of a 10% (vol/vol) Triton X-100 stock, followed by stirring for 10 min. The suspension was filtered through one layer of Miracloth (Calbiochem) and placed in 50-mL centrifuge tubes. In these centrifuge tubes, 50 mL of nuclear suspensions were underlaid with 15 mL Percoll cushion composed of 50% (vol/vol) Percoll (GE) in BFA. Nuclei suspensions were centrifuged at 3,000 × g for 15 min at 4°C. The nuclei at the Percoll interface were transferred to a 50-mL tube and diluted twofold with MNase digestion buffer (50 mM Tris-HCl at pH 7.5, 320 mM sucrose, 4 mM MgCl2, and 1 mM CaCl2). Nuclei suspensions were centrifuged at 2,000 × g for 10 min at 4°C, and nuclei pellets were resuspended in 2.5 mL MNase digestion buffer. Nuclei were aliquoted into 500-mL aliquots, flash frozen in liquid nitrogen, and stored at −80°C. Nuclei were thawed at room temperature and digested by adding MNase to 10 U/mL (light) or 100 U/mL (heavy), and incubated at room temperature for 5 min. Digestions were stopped with 10 mM EGTA. Nuclei were de-cross-linked by incubation overnight at 65°C in the presence of 1% SDS and 100 μg/mL proteinase K. DNA was extracted by phenol-chloroform extraction followed by EtOH precipitation. Digested DNA was resuspended in 40 μg/mL RNase A and electrophoresed in a 1% agarose gel. DNA fragments smaller than 200 bp were excised and gel extracted after ethidium bromide staining with the Qiaex II gel extraction kit (Qiagen), following the manufacturer’s instructions.

Library Preparation and Sequencing.

After nuclei isolation and digestion, gel-extracted DNA was used to prepare sequencing libraries using the NEBNext Ultra DNA Library Prep Kit (NEB), using manufacturer instructions. Indexed libraries were prepared and sequenced on eight Illumina HiSeq 2500 lanes with paired-end 50-cycle sequencing. Short-read data are deposited in the NCBI short read archive (SRP064243).
Read Assembly and Calling of Hypersensitive Sites. After the computational trimming of adaptor sequences using CutAdapt (40), paired-end reads were mapped to the maize B73 AGPv3 reference genome, using Bowtie2 with options "-n mixed," "-n disord," "-n unal," and "-dovetail" (41) for each replicate digest and for the genomic DNA. BED files were made from the resulting BAM files, using bedtools bamto BED, filtered for minimal alignment quality (≥10), and read coverage in 10-bp intervals was calculated using coverageBed (42). The DNS values were obtained by subtracting the mean normalized depth (in reads per million) of the heavy digest replicates from those of the light digest replicates. In this way, positive DNS values correspond to MNase hypersensitive footprints (as defined by ref. 8, and referred to here as MNase HS regions), whereas negative DNS values correspond to nucleosome hyper-resistant footprints (MRF, as per ref. 8). A Bayes factor criterion was used to classify as significantly hypersensitive.

ACKNOWLEDGMENTS. We thank Toby Kellogg from the Donald Danforth Center for providing samples of Vossia cuspidata and Coelorachis tuberculosa. This work was supported by National Science Foundation Grants IOS-0820619 and IOS-1238014 (to E.S.B.) and IOS-1025954 and IOS-1444532 (to H.W.B.) and the US Department of Agriculture-Agricultural Research Service.

Supplemental Methods

Bayes Factor Criterion

A Bayes factor criterion was used to call 10bp regions as significantly hypersensitive in order to produce discrete binary statements of MNase hypersensitive vs. non-hypersensitive regions for our analyses. This Bayes factor accounts for both the mean difference in read depths between light and heavy digests and the consistency between replicates of the digests. It compares the marginal likelihood of two models. In the null model, the relative depths – normalizing for total read depth – of all the heavy and light digest replicates and the genomic DNA sample are assumed equal with a Poisson($\lambda$) distribution. In the alternative model, the heavy digest samples are assumed to have a different distribution, Poisson($\lambda_H$), from the light digest and genomic DNA, Poisson($\lambda_L$). In each case, the parameter $\lambda$ is assumed to have a Gamma($\alpha, \beta$) prior, which we chose to be Gamma(10,1) based on the overall coverage of the maize genome. Thus, the marginal likelihood of the null model becomes:

$$f(x|\alpha, \beta) = \frac{1}{\beta^\alpha \Gamma(\alpha)} \prod_i x_i! \int_0^\infty \prod_i (c_i \lambda)^{x_i} e^{-c_i \lambda} \lambda^{\alpha-1} e^{-\lambda/\beta} d\lambda$$

where $c_i$ is the total number of reads in the $i$th sample divided by the minimum across all samples, and $x_i$ is the depth at the locus for sample $i$. Redefining as follows:

$$\alpha^* = \alpha + \sum_i x_i$$
$$\beta^* = \frac{\beta}{1 + \beta \sum_i c_i}$$

and evaluating the integral, the following marginal likelihood is obtained:

$$f(x|\alpha, \beta) = \frac{(\beta^*)^{\alpha^*} \Gamma(\alpha^*)}{\beta^\alpha \Gamma(\alpha)} \prod_i x_i! \prod_i c_i^{x_i}$$

The alternative marginal likelihood is then the product of two factors of this form, after the data is split into the appropriate partitions (heavy replicates vs. light replicates and genomic data). The Bayes factor is calculated as the quotient of the two marginal likelihoods, and we defined MNase HS sites as those regions with a positive mean difference between light and heavy digest normalized read depths and a ln(bayes factor) above 1.1, based on a suggested criterion for “positive” evidence against the null hypothesis (1).

RNA-seq

Total RNA was extracted from frozen, ground tissue with TriZol Reagent (Life Technologies) using manufacturer instructions. Libraries were then prepared from isolated RNA with the Illumina TruSeq RNA sample preparation kit (v2) according to manufacturer instructions. Libraries were subjected to single-end, 100-cycle sequencing on an Illumina HiSeq 2500. Reads were aligned to the maize B73 AGPv3 reference genome using TopHat (2). Gene and transcript models (v. 5b) were obtained from Ensembl. Transcript-level read counts of reads with an alignment quality greater than 10 were calculated using featureCounts (3) and samtools (4). All short reads are deposited in the NCBI short read archive (SRP063423).
Analyses of Base Content and Epigenetics

We evaluated the AT-content within and around MNase HS sites using bases from the maize B73 AGPv3 genome. AT-content was calculated as the relative frequency of A and T bases in 10bp windows at up to 2kb away from the MNase HS site, while each MNase HS region was divided into 4 equal-sized bins. Ambiguous nucleotides (e.g. N) were ignored during calculation. For several analyses, we controlled for the effect of amino-acid driven selection by dividing coding bases into invariant and neutral sites. These classifications were obtained based on SIFT 4G (siftdb.org). Invariant sites were classified as those with no non-reference SIFT scores above 0, while neutral sites were considered those with no non-reference SIFT scores below 1. For each classification, we limited ourselves to sites with scores based on at least 30 aligning proteins.

DNA methylation data from a whole-genome bisulfite sequencing study of B73(5) was used to calculate methylation levels for CpG, CHG, and CHH methylation using the same binning approach and calculating the relative frequency that the cytosines were in the methylated state relative to the total number of times a cytosine in a given methylation motif occurred. B73 H3K9me2 CHiP-seq reads were taken from a study of stalk tissue in 1-month-old plants (6). B73 H3K4me3 reads were taken from a study of 14-day-old shoots (7). For both sets of CHiP-seq reads, adapters were trimmed, and reads were quality-trimmed and aligned to AGPv3 according to the methods in (6). For our analyses, we considered genomic regions overlapping at least one of the appropriate CHiP-seq read as positive for the given histone modification.

Analysis of GC-biased Gene Conversion

Due to our discovery of elevated GC content within and around MNase HS regions and the association of MNase HS regions with recombination hotspots, we investigated whether GC-biased gene conversion might explain the base composition patterns. These analyses required the use of molecular phylogenetics. We constructed whole genome alignments of the maize AGPv3 genome with 12 angiosperms using the LASTz/MULTItz approach previously described for the alignment of 20 angiosperm genomes to the A. thaliana reference (8). Whole genome sequence was taken from Sorghum bicolor (v1.4), Oryza sativa ssp. Japonica (v7), Setaria italica (v2.1), Brachypodium distachyon (v1.0.21), Hordeum vulgare (v2.21), Musa acuminata (v1), Panicum virgatum (v1.1), Vitis vinifera (v3), Arabidopsis thaliana (v10.02), and Populus trichocarpa (v3). We also constructed de novo genome assemblies for the maize relatives Vossia cuspidata (Pasquet 1098) and Coelorachis tuberculosa (Layton & Zhong 169), provided by Elizabeth Kellogg. Sequencing was performed by pooling the 2 barcoded libraries and running on 4 lanes of an Illumina HiSeq 2500 with 200bp paired-end reads. Adaptors and low quality sequence (PHRED score < 25) were trimmed using Trimmomatic, and reads were assembled into de novo scaffolds using DISCOVAR de novo.

Following construction of the multiple sequence alignment, we estimated the topology of the neutral reference tree based on 4-fold degenerate sites using approximate maximum likelihood with FastTree2 (9). The neutral base substitution model (GTR) was then estimated using phyloFit in the PHAST package (10). We then estimated the posterior probability of GC-biased gene conversion on the branch leading to Z. mays using phastBias (11). When analyzing mean biased gene conversion probabilities, we restricted analysis to sites where at least 2 species could be aligned to maize. We inferred the extent of historical GC-biased gene conversion by estimating instantaneous rates of substitution from G/C to A/T alleles and from A/T to G/C alleles. In each case, we resampled 250 bootstrap samples and estimated substitution rates using phyloFit. In order to mitigate the effects of
reference bias, we weighted sites such that each mean base sampled base frequency in the reference sequence would be 0.25. We then masked the reference during estimation. We also estimated the extent of extant GC-biased gene conversion by comparing derived allele frequency for G/C->A/T and A/T->G/C alleles in maize. Polymorphism data was taken from HapMap 3.1 SNPs with quality scores of at least 30. Ancestral states of the most recent common ancestor of Z. mays and V. cuspidata were inferred using Prequel (12), following masking of the B73 reference genome to avoid reference bias. We then retained sites with maximum posterior probabilities of at least 0.95, in which the inferred ancestral base matched one of the segregating alleles.

We also estimated the population intensity of GC-biased conversion using the method of Glemin et al (13), which accounts for polarization errors. We again made use of polymorphism data from HapMap 3.1, at sites where the ancestral base could be determined with posterior probability of at least 0.95. In order to maximize efficiency and mitigate the effects of low coverage on individual HapMap 3.1 lines, we created 250 bootstrap replicates of 100-slot derived site frequency (DFS) spectra using a sampling procedure designed to account for varying amounts of missing data among loci. First, we sampled sites for each DFS category (i.e. no possible bias, S->W, W->S) with replacement. In order to account for varying uncertainty among sites due to missing genotype data, at each sampled site we sampled a “true” allele frequency from a beta distribution with a Beta(1,1) prior. In other words, frequencies were sampled from a Beta(1+N_D,1+N_A) distribution, where N_D is the number of genotyped derived alleles and N_A is the number of genotyped ancestral alleles. We then used a binomial distribution with the sampled “true” allele frequencies to obtain the number of derived alleles for each site in the 100-slot DFS. These bootstrapped DFSs were then used as input into the maximum likelihood algorithm, and we retained estimates of B from model M1*. For comparison, we also estimated null distributions for each set using the same approach but with sites randomly permuted between the 3 DFSs based on the relative numbers of polymorphisms in each.

Analysis of Recombination Enrichment

We obtained 555 recombination hotspots between 1kb and 30kb in size (median 10kb) from an earlier study of recombination patterns in 5000 recombinant inbred lines (RILs) anchored to B73 (14). We then defined 2 control regions for each hotspot – one on either side of the hotspot and directly adjacent to it. Each control region was defined to be half the size of the associated hotspot, such that the total size of the control regions equaled that of the hotspot. We then calculated posterior distributions for the relative frequency of 10bp MNase HS segments within both hotspots and controls, while controlling for the gene context – intergenic TE, intergenic non-TE, CDS – of both the hotspots and controls. Both posterior distributions were parameterized as beta-binomials with Beta(1,1) priors.

The same earlier study of recombination in NAM RILs (14) estimated crossover enrichment in narrow windows between genotyping by sequencing markers using a Bayesian approach. We used these narrow windows (1-10kb) to investigate the small-scale relationship between GC-biased gene conversion, MNase HS frequency, genic frequency, and recombination. We created a linear model with the mean PHASTbias-estimated posterior probability of GC-biased gene conversion in the window as the response. We included main effects of the crossover enrichment, MNase HS frequency, genic frequency, and TE frequency as predictors. In order to obtain normal, homoscedastic residuals, we imposed a logit transformation on the response and a log transformation of the crossover enrichment. We then estimated the relative importance of the variables using the
relaimpo package in R (15), with 1,000 bootstrap replicates used to calculate the ‘Img’ measure of \( R^2 \) contribution averaged over all possible models.

**Sequence Conservation**

In order to estimate sequence conversion on a site-by-site basis, we used Genomic Evolutionary Rate Profiling (GERP), which measures sequence conservation in terms of “rejected substitutions” (16). We use the same phylogeny described in the PHASTbias analysis for estimation of GERP scores. The GERP score depends both on the level of conservation and the ability to align orthologous sequences at a site. We accounted for variation in the set of aligned orthologous sequences in 2 ways. First, we calculated the scale factor \( (S=(R-G)/R) \) to facilitate comparison across sites with different sets of aligning species, where \( S \) is the scale factor, \( R \) is the length of the neutral tree, and \( G \) is the GERP score. Second, when calculating the mean scale factor for each set under a given functional annotation, we used bootstrapping with weights such that the probabilities of representation for sites of varying neutral tree length would be equal between sets.

**GWAS Enrichment**

As a primary means of testing for a significant relationship between MNase HS regions and quantitative traits, we quantified the enrichment of GWAS hits – relative to all SNPs – in proximity to MNase HS regions. We obtained 3,873 unique, high-confidence (resample model inclusion probability >= 0.05) GWAS hits from an earlier study of 41 phenotypes in the US-NAM population (17). These hits enabled us to classify each SNP in the HapMap2 SNP set as either “GWAS” or “non-GWAS”. We then classified genomic regions as MNase-proximal if they were within 2kb of the nearest MNase HS region, choosing 2kb as a conservative threshold based on the rate of LD decay within maize (18). We also defined control regions of half the MNase-proximal region size on both sides of the MNase-proximal region, 5kb away from each side, ensuring that no control region was within 2kb of the nearest MNase HS site. Finally, we calculated the posterior distribution of GWAS SNP frequency, relative to all SNPs, using a beta-binomial distribution with a Beta(1,1) prior. In order to control for the possible confounding effects of distance to the nearest gene, we estimated the same distributions after limiting both the MNase-proximal and control regions to lie within 5kb, more than 5kb, or more than 20kb from the nearest protein coding gene within the maize 5b filtered gene set.

**Variance Partitioning**

As a complement to the analysis of GWAS hits occurring near MNase HS regions, direct estimates were obtained for the heritable variance explained by SNPs within MNase HS regions using previously described techniques (19). Imputed SNP genotypes were obtained for all NAM RILs based on 50M HapMap2 SNPs (18), and for all Ames inbreds based on 50M HapMap 3.1 SNPs (20) imputed using previously described methods (21) and inbred lines (22, 23). Phenotypes were collected as previously described for the US-NAM lines (17) and for the Ames Diversity Panel (23, 24). Within each dataset, SNPs were hierarchically divided into disjoint annotation categories, in the following order of priority: CDS, MNase HS, 5’UTR, 3’UTR, Intron, Intergenic. For example, all SNPs positioned within CDS regions were annotated as CDS, while those that were not within CDS regions but were MNase HS were annotated as MNase HS. Kinship matrices were then constructed for each
SNP set based on the kinship estimation technique used in GCTA (25),

\[ A_{jk} = \frac{1}{N} \sum_{i=1}^{N} \frac{(x_{ij} - 2p_i)(x_{ik} - 2p_i)}{2p_i(1 - p_i)} \]

where \( x_{ij} \) is the genotype (0, 1, or 2) of the \( j \)th individual at locus \( I \) with allele frequency \( p_i \).

The background effect of each NAM family on each phenotype was accounted for by subtracting out family-specific random effects estimated using lmer in R prior to variance partitioning. However, removal of the family effect does not have a major impact on the final partitioning (Fig S20). Due to high coverage sequencing of all NAM founders, NAM RIL genotypes were able to be completely imputed. Because of the much higher haplotype diversity within the Ames diversity panel, many loci were missing data from a high percentage of the population. Data that was missing during the calculation of kinship was accounted for by setting \( p_i \) equal to the maximum likelihood estimate based on the observed data and setting the genotype of missing data equal to the estimated mean - \( 2p_i \) – thereby zeroing out terms in the sum containing missing genotypes. Any missing values in cells for the kinship matrix were accounted for by setting \( N \) equal to the number of loci at which neither taxon contained missing data. SNPs with more than 80% missing values or MAF below 0.025 were also eliminated.

Following the construction of kinship matrices, variance was partitioned among the annotation-specific SNP sets for each phenotype using the MultiBLUP method implemented in the LDAK package, version 4.5 (26).

**Simulation of Annotation-Specific Phenotypes**

In order to test the accuracy of variance partitioning on datasets with known genetic architectures and empirical patterns of linkage disequilibrium, we used a simulation approach. Similar to the approach of Gusev et al. (19), we simulated phenotypes in both the NAM and Ames panels wherein nonzero effects were only assigned to SNPs in a single annotation category. For each annotation category, we simulated a total of 250 phenotypes. Within each simulated phenotype, 0.5% of the SNPs were chosen to have nonzero effects, and SNP effects were drawn randomly from a normal distribution centered at zero with a standard deviation of 0.5. SNPs were assumed to act completely additively, and we imputed missing SNP genotypes to the mean. Following the simulation of genetic effects, we added random Gaussian noise such that the genetic effects accounted for 70% of the final variance. In the case of the NAM population, we subtracted out family-specific random effects in the same way as with the observed phenotype data.

**References**


Supplemental Tables

Table S1. Frequency of functional annotations in root and shoot MNase HS along with the entire B73 genome

<table>
<thead>
<tr>
<th>Annotation</th>
<th>Shoot MNase HS</th>
<th>Root MNase HS</th>
<th>Genome</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Shoot MNase HS</td>
<td>Root MNase HS</td>
<td>Genome</td>
</tr>
<tr>
<td>CDS</td>
<td>393,992 (3.54%)</td>
<td>132,388 (2.89%)</td>
<td>43,962,593 (2.13%)</td>
</tr>
<tr>
<td>5'UTR</td>
<td>773,225 (6.95%)</td>
<td>238,861 (5.22%)</td>
<td>17,055,306 (0.83%)</td>
</tr>
<tr>
<td>3'UTR</td>
<td>596,918 (5.37%)</td>
<td>217,650 (4.76%)</td>
<td>18,912,120 (0.92%)</td>
</tr>
<tr>
<td>Intron</td>
<td>646,371 (5.81%)</td>
<td>270,078 (5.90%)</td>
<td>80,950,046 (3.93%)</td>
</tr>
<tr>
<td>Intergenic, Non-TE</td>
<td>7,491,416 (67.4%)</td>
<td>3,098,033 (67.7%)</td>
<td>222,416,536 (10.8%)</td>
</tr>
<tr>
<td>Intergenic, TE</td>
<td>1,217,035 (10.9%)</td>
<td>619,075 (13.5%)</td>
<td>1,677,054,199 (81.4%)</td>
</tr>
</tbody>
</table>

Table S2. Number of genes with expression above or equal to a given RPKM threshold within root and shoot

<table>
<thead>
<tr>
<th>RPKM threshold</th>
<th>shoot</th>
<th>root</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>36441</td>
<td>36441</td>
</tr>
<tr>
<td>0.1</td>
<td>25549</td>
<td>26020</td>
</tr>
<tr>
<td>0.2</td>
<td>24398</td>
<td>24989</td>
</tr>
<tr>
<td>0.3</td>
<td>23691</td>
<td>24362</td>
</tr>
<tr>
<td>0.4</td>
<td>23121</td>
<td>23843</td>
</tr>
<tr>
<td>0.5</td>
<td>22644</td>
<td>23419</td>
</tr>
<tr>
<td>0.6</td>
<td>22180</td>
<td>23028</td>
</tr>
<tr>
<td>0.7</td>
<td>21823</td>
<td>22690</td>
</tr>
<tr>
<td>0.8</td>
<td>21498</td>
<td>22402</td>
</tr>
<tr>
<td>0.9</td>
<td>21203</td>
<td>22129</td>
</tr>
<tr>
<td>1</td>
<td>20919</td>
<td>21880</td>
</tr>
</tbody>
</table>

Table S3. Linear model for regression of the logit-transformed mean GC-biased gene conversion posterior probability in 1-10kb windows

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>SE</th>
<th>t-value</th>
<th>p-value</th>
<th>Relative Contribution*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-4.84</td>
<td>0.0559</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>log(crossover enrichment)</td>
<td>0.467</td>
<td>0.0161</td>
<td>29.040</td>
<td>&lt;2x10^-16</td>
<td>(0.417,0.502)</td>
</tr>
<tr>
<td>MNase HS Freq.</td>
<td>5.71</td>
<td>0.257</td>
<td>22.3</td>
<td>&lt;2x10^-16</td>
<td>(0.285,0.354)</td>
</tr>
<tr>
<td>Genic Freq.</td>
<td>-0.137</td>
<td>0.0533</td>
<td>-2.57</td>
<td>0.0101</td>
<td>(0.0918,0.136)</td>
</tr>
<tr>
<td>TE Freq.</td>
<td>1.74</td>
<td>0.104</td>
<td>16.7</td>
<td>&lt;2x10^-16</td>
<td>(0.0896,0.125)</td>
</tr>
</tbody>
</table>

Adjusted $R^2$=0.117

*95% CIs based on averaging $R^2$ over orders of models in 1,000 bootstrap replicates
Table S4. Traits used in variance portioning analyses and the PubMed identifiers of their citations (PMID)

<table>
<thead>
<tr>
<th>Trait</th>
<th>Population</th>
<th>Original Citation</th>
<th>PMID</th>
</tr>
</thead>
<tbody>
<tr>
<td>PGDD to Anthesis</td>
<td>US-NAM</td>
<td>22711828</td>
<td></td>
</tr>
<tr>
<td>100 Kernel Weight</td>
<td>US-NAM</td>
<td>25474422</td>
<td></td>
</tr>
<tr>
<td>PGDD to Silking</td>
<td>US-NAM</td>
<td>22711828</td>
<td></td>
</tr>
<tr>
<td>Avg. Internode Length Above Ear</td>
<td>US-NAM</td>
<td>24514905</td>
<td></td>
</tr>
<tr>
<td>Anthesis-Silking Interval</td>
<td>US-NAM</td>
<td>19661422</td>
<td></td>
</tr>
<tr>
<td>Ratio of Ear Hgt to Total Hgt</td>
<td>US-NAM</td>
<td>24514905</td>
<td></td>
</tr>
<tr>
<td>Hgt Above Ear</td>
<td>US-NAM</td>
<td>24514905</td>
<td></td>
</tr>
<tr>
<td>Nodes to Ear</td>
<td>US-NAM</td>
<td>24514905</td>
<td></td>
</tr>
<tr>
<td>Nodes Above Ear</td>
<td>US-NAM</td>
<td>24514905</td>
<td></td>
</tr>
<tr>
<td>Avg. Internode Length (Whole Plant)</td>
<td>US-NAM</td>
<td>24514905</td>
<td></td>
</tr>
<tr>
<td>Hgt per Day (Until Flowering)</td>
<td>US-NAM</td>
<td>24514905</td>
<td></td>
</tr>
<tr>
<td>Nodes per Plant</td>
<td>US-NAM</td>
<td>24514905</td>
<td></td>
</tr>
<tr>
<td>Tassel Length</td>
<td>US-NAM</td>
<td>22125498</td>
<td></td>
</tr>
<tr>
<td>Days to Anthesis</td>
<td>US-NAM</td>
<td>19661422</td>
<td></td>
</tr>
<tr>
<td>Leaf Length</td>
<td>US-NAM</td>
<td>21217756</td>
<td></td>
</tr>
<tr>
<td>Southern Leaf Blight</td>
<td>US-NAM</td>
<td>21217757</td>
<td></td>
</tr>
<tr>
<td>Ear Hgt</td>
<td>US-NAM</td>
<td>23840585</td>
<td></td>
</tr>
<tr>
<td>Cob Diameter</td>
<td>US-NAM</td>
<td>22125498</td>
<td></td>
</tr>
<tr>
<td>Boxcox-transformed Leaf Angle</td>
<td>US-NAM</td>
<td>21217756</td>
<td></td>
</tr>
<tr>
<td>Days to Silking</td>
<td>US-NAM</td>
<td>19661422</td>
<td></td>
</tr>
<tr>
<td>Plant Hgt</td>
<td>US-NAM</td>
<td>24514905</td>
<td></td>
</tr>
<tr>
<td>Ear Row Number</td>
<td>US-NAM</td>
<td>22125498</td>
<td></td>
</tr>
<tr>
<td>Leaf Width</td>
<td>US-NAM</td>
<td>21217756</td>
<td></td>
</tr>
<tr>
<td>Tassel Branch Number</td>
<td>US-NAM</td>
<td>22125498</td>
<td></td>
</tr>
<tr>
<td>Northern Leaf Blight</td>
<td>US-NAM</td>
<td>21482771</td>
<td></td>
</tr>
<tr>
<td>Growth Rate</td>
<td>Ames</td>
<td>24514905</td>
<td></td>
</tr>
<tr>
<td>GDD Anthesis-Silking Interval</td>
<td>Ames</td>
<td>24514905</td>
<td></td>
</tr>
<tr>
<td>Ear Hgt</td>
<td>Ames</td>
<td>24514905</td>
<td></td>
</tr>
<tr>
<td>Ear/Plant Hgt Ratio</td>
<td>Ames</td>
<td>24514905</td>
<td></td>
</tr>
<tr>
<td>Plant Hgt</td>
<td>Ames</td>
<td>24514905</td>
<td></td>
</tr>
<tr>
<td>Anthesis-Silking Interval</td>
<td>Ames</td>
<td>24514905</td>
<td></td>
</tr>
<tr>
<td>GDD to Silking</td>
<td>Ames</td>
<td>24514905</td>
<td></td>
</tr>
<tr>
<td>GDD to Anthesis</td>
<td>Ames</td>
<td>24514905</td>
<td></td>
</tr>
<tr>
<td>Days to Silking</td>
<td>Ames</td>
<td>24514905</td>
<td></td>
</tr>
<tr>
<td>Days to Anthesis</td>
<td>Ames</td>
<td>24514905</td>
<td></td>
</tr>
</tbody>
</table>
Figure S1. Recombination Frequency is strongly associated with MNase HS frequency. The log-log relationship of recombination frequency to MNase HS frequency within 1Mb bins.
Figure S2. Conservation of MNase HS sequence. The conservation of MNase HS sequence compared to the flanking 180bp regions. The scale factor indicates the maximum likelihood estimate factor by which the neutral phylogenetic tree is scaled to fit the observed substitution pattern at a site. The red line indicates the neutral value.
Figure S3. Conservation of MNase HS sequence between tissues. The conservation of MNase HS sequence compared among MNase HS present in the shoot-only, root-only, or within both root and shoot. The scale factor indicates the maximum likelihood estimate factor by which the neutral phylogenetic tree is scaled to fit the observed substitution pattern at a site. The red line indicates the neutral value.
Figure S4. MNase HS is Associated with Gene Regulation. Heatmaps of MNase hypersensitivity profiles for genes ordered by expression in root and shoot. The bottom shows the average hypersensitivity profiles for expression quintiles.
Figure S5. MNase HS explains relative expression of paralogs. Mean shoot MNase HS profiles of the maize1 and maize2 paralogs from the most recent whole genome duplication following division of
the duplicate pairs into 9 tertiles based on expression.

Figure S6. MNase HS Regions are Associated with Recombination Hotspots. (a) Frequency of MNase HS bases within recombination hotspots and control regions located adjacent to recombination hotspots (b) The mean genome-wide posterior probability of GC-biased gene conversion within and surrounding MNase HS regions.
Figure S7. MNase HS Associates with Recombination in Gene Distal Regions. The mean crossover enrichment measure in 1-10kb bins located at least 5kb from the nearest protein-coding gene at varying levels of MNase HS frequency. Error bars indicate 95% confidence intervals of the mean.
Figure S8. Shoot MNase HS Associates with Higher Recombination. (a) Density of crossover enrichment values in 1-10kb ranges containing either only shoot-specific or only root-specific MNase HS regions. The bar indicates the mean. (b) Distribution of differences in the means of crossover enrichment in the 2 categories of ranges (i.e. shoot-specific MNase HS only vs root-specific MNase HS only) based on 1,000 bootstrap replicates. The bar indicates the mean difference.

Figure S9. Relationship between estimated posterior probabilities of GC-biased gene conversion from PHASTbias with inclusion of the full phylogeny and only including the grass (Poaceae) species.
Figure S10. Intergenic regions show less sustained GC-biased gene conversion. (a) The frequency of A/T content and the histone modifications H3K9me2 and H3K4me3 surrounding and within MNase HS regions for non-TE and TE intergenic sites (b) Mean ranges of GC-biased gene conversion probabilities within MNase HS regions and regions 1-2kb from the nearest MNase HS region for both recombination hotspots and control regions flanking recombination hotspots. Ranges correspond to 95% credible intervals.
Figure S11. Derived allele frequencies in genes reflect ongoing GC-biased gene conversion.
(a) Mean derived allele frequencies within and at varying distances from MNase HS regions for polymorphisms with an S ancestral allele and a W derived allele, where S=G/C and W=A/T for coding and genic, noncoding polymorphic sites. (b) Mean derived allele frequencies for polymorphisms with a W ancestral allele and an S derived allele.
Figure S12. Estimates of the population GC-biased gene conversion intensity at varying distances from MNase HS regions within genic and gene-distal contexts. Blue plots indicate the estimates of B for the observed data based on 250 bootstrap replicates, while black plots indicate estimates for the null distribution based 250 permutations.
Figure S13. GC Biased Gene Conversion Surrounding MNase HS is a Sustained Force.
Distributions of substitution rate ratios of W->S to S->W, where S represents conversion-favored G/C alleles and W represent conversion disfavored A/T alleles. Densities denote the range calculated from 250 bootstrap samples, and horizontal bars denote the mean.
Figure S14. Lack of H3K9me2 marks GC-biased gene conversion targets. 95% credible ranges for mean AT frequency in and around MNase HS regions in the presence and absence of H3K4me3 and H3K9me2. Note that in the coding sequence, positions are limited to putative neutral sites in order to mitigate the effects of selection. The cartoon illustrates the trend in AT content for neutral noncoding sequence and the possibility that Spo11 may bind H3K9me2-negative regions adjacent to MNase HS sites.
Figure S15. MNase HS overlaps intergenic QTL. (a) MNase Hypersensitivity profile within the vgt1 locus and for the (b) gt1 control region. Regions with an MNase HS Bayes factor above 1 are shown in red, and the opacity is proportional to the significance of the difference. Regions without uniquely mapping reads are shown in grey. Regions covered by reads but without sufficient evidence for hypersensitivity are shown in blue.
Figure S16. Intergenic MNase HS Regions Explain Most of the Variance within MNase HS Regions. The average contributions of SNPs with the given annotations to heritable variance and the partitioning of variance to individual traits within (a) US-NAM and (b) Ames Diversity Panel following division of MNase HS regions into subcategories.
**Figure S17. Shoot MNase HS Explains most MNase HS Variance** (a and b) The enrichments of variance in functional categories within (a) US-NAM and (b) Ames Diversity Panel following the division of MNase HS regions into those exclusively within shoots, exclusively within roots, and common to both. (c and d) The average contributions of SNPs with the given annotations to heritable variance and the partitioning of variance to individual traits within (c) US-NAM and (d) Ames Diversity Panel.
**Figure S18. Variance partitioning is unbiased in maize.** Ranges of heritabilities obtained for 250 simulated phenotypes in (a) US-NAM and (b) Ames Panel. Bolded red categories contain the totality of the true heritability. (c) Boxplots of variance explained in several US-NAM phenotypes following 250 permutations of phenotypes within family. Note that the median for each case occurs at 0. (d) Densities of SNP quality scores for US-NAM, divided by category.
Variance partitioning is robust to annotation hierarchy and LD. Average contributions of MNase different annotation categories to heritability and the individual partitions within (a) US-NAM after inclusion of a category 2-5kb from the nearest MNase HS region (b) Ames Panel after inclusion of a category 2-5kb from the nearest MNase HS region (c) US-NAM following Photoperiod growing-degree days to anthesis 100 Kernel weight Photoperiod Growing-degree days to silk Average internode length (above ear) Anthesis-silking interval Ratio of ear height to total height Height above ear Nodes to ear Nodes above ear Ear row number Leaf length Southern leaf blight Ear height Cob diameter Leaf width Northern Leaf Blight Tassel branch number
reversal of the prioritization of SNP categories and (d) Ames panel following reversal of the prioritization of SNP categories.

Figure S20. NAM Variance Partitioning Not Largely Impacted by Family Correction. The average contributions of SNPs with the given annotations to heritable variance and the partitioning of variance to individual traits within US-NAM without inclusion of a family correction term.