Gene regulatory effects of a large chromosomal inversion in highland maize

Taylor Crow [1*], James Ta [1], Saghi Nojoomi [1], M. Rocío Aguilar-Rangel [2,3], Jorge Vladimir Torres Rodrǵuez [2], Daniel Gates [4], Ruben Rellan-Alvarez [2,5], Ruairidh Sawers [2,6], Daniel Runcie[1]

1 Department of Plant Sciences/University of California, Davis, CA, USA

2 Laboratorio Nacional de Genómica para la Biodiversidad/Unidad de Genómica Avanzada, Centro de Investigación y Estudios Avanzados (CINVESTAV- IPN), Irapuato CP 36821, Guanajuato, Mexico

3 Corteva Agriscience™, Agriculture Division of DowDuPont™, Tlajomulco, Jalisco, Mexico

4 Department of Evolution and Ecology/University of California, Davis, CA, USA **5** Department of Molecular and Structural Biochemistry, North Carolina State University, Raleigh, NC **6** Department of Plant Science, Pennsylvania State University, State College, PA

* tmcrow@ucdavis.edu

Abstract

Chromosomal inversions are frequently implicated in local adaptation. Inversions can capture multiple locally adaptive functional variants in a linked block by repressing recombination. However, this property makes it difficult to identify the genetic mechanisms that underlie an inversion's role in adaption. In this study, we explore how large-scale transcriptomic data can be used to dissect the functional importance of chromosomal inversions. Specifically, we study a 13 Mb inversion locus found almost exclusively in highland populations of maize (*Zea mays* ssp. *mays*) known as *Inv4m*. *Inv4m* is known to have introgressed into domesticated maize from a wild relative also present in the highlands of Mexico and is thought to be important for the adaptation of certain maize landraces to cultivation in highland environments. First, using a large publicly available association mapping panel, we confirmed that *Inv4m* is associated with locally adaptive agronomic phenotypes, but only in highland fields. Second, we created two Near Isogenic Line populations segregating for alternative alleles of *Inv4m*, and measured gene expression variation association with *Inv4m* across 9 tissues in two experimental conditions. With these data, we quantified both the global transcriptomic effects of the highland *Inv4m* allele, and the local cis-regulatory variation present within the locus. We found

diverse physiological effects of *Inv4m*, and speculate that the genetic basis of its effects on adaptive traits is distributed across many separate functional variants.

Introduction

Chromosomal inversions are structural rearrangements that form when a portion of a chromosome breaks in two places and reinserts in the opposite orientation. The reversed order of loci prevent recombination with the non-inverted homologous chromosome. This spontaneous, long-distance genetic linkage is important for speciation and local adaptation because it can capture multiple adaptive and potentially interacting loci in a single haplotype [1–3]. Inversions are common across taxa [4], often pre-date speciation events, and can spread through admixture [5, 6]. They have been linked to adaptive phenotypes and environmental clines [7–9], mating system evolution [10–12], social organization [13], and migratory phenotypes [14].

Chromosomal inversions were first discovered nearly a century ago in *Drosophila* [2, 15] by visualizing karyotypes, and can be identified based on their effects on recombination rates among nearby markers. However, both techniques are labor intensive and difficult to apply to large-scale population-level surveys within or among species. Modern genome-wide sequencing technologies provide the opportunity to identify inversions more rapidly and comprehensively, leading to the discovery of inversions across a wide range of species [4]. In fact, structural variants (e.g. insertions, deletions, duplications, translocations, fusions and inversions) have been shown to account for more variation by the number of base pairs than SNPs [16].

However, while whole-genome genotyping data can help rapidly discover inversion loci, measure their frequencies across populations, and test for associations with adaptation and speciation, there are still very few examples where the genetic mechanisms that underlie the adaptive value of any particular inversion is known. Are there are generally one or two loci of major effect within an adaptive inversion, or do inversions harbor many large and small-effect variants that combine to make inversion loci highly pleiotropic supergenes [17]? Because they suppress recombination across the whole locus, association mapping methods like QTL mapping and Genome-Wide Association Studies have little ability to resolve independent effects of different variants within the locus. This makes fine-mapping nearly impossible. Only in cases of very old inversion loci which have experienced rare recombinants or gene conversion events can association methods or population-genetic signatures of selection successfully identify causal loci within inversions [18].

RNA sequencing technologies may be a powerful tool for gaining rapid insight into the evolutionary role of inversion loci, particularly for loci for which the adaptive phenotypes regulated by the locus are unknown. RNA sequencing is a very high-throughput phenotyping technology that can simultaneously measure tens of

thousands of different traits on an organism. The expression of each gene responds to a different combination of transcription factors, gene networks, and cellular states, so measuring gene expression provides an indirect measurement of a wide range of cellular, developmental and physiological characteristics. Many of these traits may be important for adaptation, but are neglected in many studies because they are difficult to measure directly. At the same time, gene expression analysis can be used to scan across an inversion locus gene-by-gene to identify specific genes that have different cis-regulatory genetic control among alleles. If the functional variation captured by an inversion locus operates by directly altering the expression of genes in the inversion, we can identify these genes by their expression changes without relying on recombination. Together, these two types of gene expression analysis may greatly advance our understanding of inversion loci that are not feasible to study by other means.

In this study, we applied population genetic and gene expression analyses to study an inversion locus in maize. Maize is an important crop species worldwide and also a powerful model system for studying the mechanisms of recent and rapid local adaptation. Maize (*Zea mays* ssp. *mays*) was domesticated in the lowland Balsas river valley of southwestern Mexico from the narrowly distributed teosinte, *Zea mays* ssp. *parviglumus*, hereafter *parviglumus*. Since the domestication of maize approximately 9000 years ago [19, 20], populations of maize have been moved into high altitude environments, and landraces collected today show considerable local adaptation to their home elevation in a range of traits [21]. Interestingly, population genetic scans for loci associated with adaptation to elevation gradients have identified several loci common in highland landraces that have been introgressed from a different subspecies of teosinte, *Zea mays* ssp. *mexicana* (hereafter *mexicana*), which occurs in highland environments [22]. One of these introgressed regions, located at approximately 171.7 to 185.9 Mb of chromosome 4, is a chromosomal inversion known as *Inv4m* [23, 24]. *Inv4m* is observed in high altitude locations in Mexico, and is associated with a three day acceleration of flowering time, the largest effect flowering QTL in maize yet found [25]. *Inv4m* also overlaps with a quantitative trait locus (QTL) found in a previous study of leaf pigmentation and macrohairs in teosinte [26].

Are there are few loci of major effect that drive the principle selectively important phenotypes? Are there major functional variants distributed throughout the locus, making this a large *supergene* [17]? Or do loci within *Inv4m* have beneficial epistatic or pleiotropic interactions with other introgressed regions from *mexicana*?

We first comprehensively characterized the population-genetic context of the *Inv4m* locus and its association with key agronomic traits using dense whole-genome genotyping data. *Inv4m* is more closely associated with altitude in the center of maize diversity in Mexico than nearly any other locus in the maize

genome, and shows clear patterns of antagonistic pleiotropy indicative of a key role in local adaptation. Then, we dissected the function of the locus by broadly characterizing its effect on genome-wide gene expression across a panel of tissues. We isolated *Inv4m* from two separate donor sources into a common lowland-type maize background (B73), and used RNA sequencing of samples from nine different tissues and two developmental stages in two environments to identify molecular traits that were differentially regulated by the highland and lowland alleles of the inversion. These genes suggest a range of disparate biological processes affected by *Inv4m*, highlighting novel cell-biological or physiological traits that may be involved in the local adaptation. Finally, we scanned genes within *Inv4m* for associations with these gene expression traits and identified several outlier genes inside the *Inv4m* locus that are good candidates for further study.

Materials and methods

Population Genetics of *Inv4m*

We downloaded unimputed genotype-by-sequencing (GBS) data from 94,726 loci on chromosome 4 and 4,845 maize plants from the SeeD-maize GWAS panel [25, 27] and ran a principal components analysis on all positions within the *Inv4m* locus (between 168832447 and 182596678 in AGPv2 coordinates, [23]) with $< 25\%$ missing data and minor allele frequency > 0.05. PC1 explained 22% of the genetic variation among these plants in this interval. Scores on PC1 neatly divided plants into three groups, representing the two homozygous classes at *Inv4m* and their heterozygotes. We cross-referenced plants with landrace passport data from Germinate 3, the CIMMYT Maize Germinate Database `germinate.cimmyt.org/maize/` and extracted country of origin, latitude, longitude and elevation records. All but 7 of the plants containing the minor allele at *Inv4m* were from Mexico, so to study associations with elevation and to calculate other diversity statistics, we subsetted to only those plants collected in Mexico.

To calculate the association of *Inv4m* with elevation, we divided landraces into 100m bins, calculated the allele frequency of the minor allele in each bin, and fit a loess curve to the logit-transformed allele frequencies, weighted by the number of landraces in each bin. Based on this analysis, we labeled the minor allele in Mexico "High", and the major allele "Low". We used the R function *HWExact* from the *HardyWeinberg R* package to test genotype counts against the Hardy-Weinberg expectation. We calculated diversity statistics π and θ separately for plants homozygous for the "High" or "Low" alleles at *Inv4m* across all of chromosome 4 in 500 marker windows using TASSEL 5 [28]. Because many more plants were homozygous for the "Low" allele, we randomly sampled 371 to calculate the diversity statistics to make the sample sizes equal.

Association of *Inv4m* with agronomic traits

We re-analyzed phenotypic data from the F1 Association Mapping (FOAM) panel of Romero-Navarro *et al* [25] and Gates *et al* [29] to more fully characterize associations signatures of *Inv4m*. Full descriptions of this experiment and data access are described in those references. We downloaded BLUPs for each trait and line from Germinate 3, and subsetted to only those lines with GBS genotype data from Mexico. We fit a similar model to the GWAS model used by Gates *et al* [29] to estimate the effect of *Inv4m* genotype on the trait's intercept and slope on trial elevation, accounting for effects of tester ID in each field and genetic background and line effects on the trait intercept and slope using four independent random effects. We implemented this model in the *R* package *GridLMM* [30]. We extracted effect sizes and covariances conditional on the REML variance component estimates and used these to calculate standard errors for the total *Inv4m* effect as a function of elevation. To test whether the phenotypic effects of *Inv4m* on yield components could be explained as indirect effects via flowering time, we additionally re-fit each model using Days-To-Anthesis as a covariate with an independent effect in each trial.

Experimental material for isolating *Inv4m*

To directly assess additional phenotypic effects of the *Inv4m* locus, we selected two highland landraces which both carry the High allele of *Inv4m*, Palomero Toluqueno (PT) and Michoacán 21 (Mi21). These *Inv4m* donors were repeatedly backcrossed to introgress the locus into the B73 reference line. Both landrace accessions were obtained through the International Maize and Wheat Improvement Center (CIMMYT); PT came from accession mexi5 and Mi21 from accession Michoacán 21. B73 is a modern inbred from the United States, but carries the Low-type allele at the *Inv4m* locus. Both landraces were crossed with B73 and one resulting F1 individual from each cross was backcrossed to B73 for 4 generations, selecting on a diagnostic SNP for *Inv4m* each cycle with a cleaved amplified polymorphic sequence (CAPS) assay. DNA was extracted from leaf tissue using a Urea lysis buffer extraction protocol (https://github.com/RILAB/lab-docs/wiki/Wetlab-Protocols). Primers were designed to amplify the fragment of DNA carrying the diagnostic SNP (Forward: CTGAGCAGGAGATGATGGCCACTC; Reverse: GGAAAGGACATAAAAGAAAGGTGCA). Amplification consisted of 5 minute denaturation at 95°C, 35 cycles of 95-60-72°C for 30 seconds each, 7 minutes of final extension step at 72°C, followed by a 4°C hold. Amplified DNA was then digested with the *Hinf 1* enzyme for 1 hour at 37°C, and the resulting product was run out on a 1% agarose gel for genotyping.

Two of the resulting BC_5 nearly isogenic lines (NILs, also known as introgression lines) were

self-pollinated per population (which were heterozygous for *Inv4m*) to produce BC_5S_1 NILs segregating for *Inv4m* approximately at a 1:2:1 ratio.

Reciprocal transplant experiment to identify phenotypic effects of *Inv4m*

We planted seeds from the four introgression lines (two parents per *Inv4m* donor) in the UC Davis controlled environment facility growth chambers. Chambers were programmed to mimic temperatures in Mexican lowlands (22°C night, 32°C day, 12 hr light) and highlands (11°C night, 22°C day, 12 hr light). Kernels were soaked in distilled H_20 for 12 hours and planted in 10.2cm x 34.3cm nursery pots (Steuwe & Sons: CP413CH) in a soil mixture composed of a 3:1 ratio of Sungro Sunshine Mix #1 to sand. Pots were organized in racks with 9 pots per rack (Steuwe & Sons: tray10). Plants were watered every other day with a 1x Hoagland solution, and emergence was recorded daily. The experiment was replicated and growth chambers were switched to account for variation between instruments between replicates (See Figure S1 for a graphical workflow). The first replicate of the experiment began March, 2017 and the second replicate began April, 2017.

Two seeds were planted in each pot, one in the center and one near the corner, and a total of nine tissues were sampled from the two plants when they reached a specific developmental stage. The nine tissues were selected to maximize the diversity of gene expression profiles based on the transcription atlas of [31]. Plants were removed from the pot and sampled when the plant in the corner reached the V1 stage, and the plant in the center reached the V3 stage. Two tissue types were sampled from the V1 stage, and 7 tissue types were sampled from the V3 stage (Table 1). Sampling occurred between 2 and 4 hours after simulated sunrise. Plant tissue was placed in 2 ml centrifuge tubes, immediately flash frozen in liquid nitrogen, and stored at -70°C.

Table 1. Developmental stage and description of tissues sampled for gene expression analysis.

Developmental_Stage	Tissue	Description
V1	Root	Primary root
V3	Root	Primary root
V3	SAM	Stem Apical Meristem
V1	Leaf	Pooled leaf tissue
V3	Sheath	Leaf sheath
V3	Leaf base	5 cm of leaf base
V3	Leaf tip	5 cm of leaf tip
V3	S2 leaf base	5 cm of leaf base
V3	S2 leaf tip	5 cm of leaf tip

Genotyping and RNA sequencing

We used the same DNA extraction and CAPs genotyping methods as previously described to genotype the NILs for the *Inv4m* allele. We randomly sampled 3 biological replicates from each tissue, homozygous *Inv4m* allele (heterozygous plants were excluded to focus on additive effects of *Inv4m*), donor (Mi21 & PT), and temperature treatment from each experimental replicate for a total of 432 samples. Approximately 20 mg of tissue for each sample was placed in a 2ml centrifuge tube and flash-frozen in liquid nitrogen and ground using stainless steel beads in a SPEX Geno/Grinder (Metuchen, NJ, USA). mRNA was extracted using oligo $(dT)_{25}$ beads (DYNABEADS direct) to isolate polyadenylated mRNA using the double-elution protocol. We prepared randomly primed, strand specific, mRNA-seq libraries using the BRaD-seq [32] protocol with 14 PCR cycles. Samples underwent a single carboxyl bead clean-up, quantified using the Quant-iT^{TM} PicoGreen dsDNA kit, and normalized. We took 2ng per library and multiplexed 96 samples for sequencing. Each multiplexed library was sequenced on 1 lane of a Illumina HiSeq X platform, generating a mean of 4,241,500 reads per sample. Raw reads were quality checked using FastQC v.0.11.5 [33]. Adapter sequences, low quality reads (q<20), and sequences less than 25 bp were removed using Trimmomatic v.0.36 [34].

Effects of genotype at *Inv4m* on seedling emergence

The effect of *Inv4m* on seedling emergence were analyzed using the following random slope and intercept model for each donor and temperature treatment separately:

$$y_{ijk} \sim \mu + \beta_1 G_i + u_{ijk} + e_{ijk}$$

y_{ijk} is the emergence time for individual plant k in experimental replicate j in *Inv4m* genotype i. μ the model intercept, and $\mathbf{u} = \mathbf{u_{ijk}}$ is a random effects term for the experimental replicate, and e_{ijkl} is residual error. Variance components, coefficients and standard errors were estimated by REML using the R function *lmer* [35], and p-values were calculated using conditional F-tests [36].

Population characterization

BC_5S_1 plants are expected to contain $\sim 3\%$ residual DNA from the non-recurrent parent across the remaining 9 chromosomes. We used the RNAseq reads from each plant to comprehensively characterize all residual introgressed regions across the genome by calling variants in the expressed regions. Paired reads that passed filtering were aligned to the B73 reference genome version 4 [37] using hisat2 [38], and variant loci

were called using GATKv3 [39, 40]. We ran `MarkDuplicates`, `SplitNCigarReads` and `HaplotypeCaller` on every sample, including all 435 NIL samples and an additional 46 B73, 48 PT and 47 B73-PT F1 samples from plants run in parallel with the NILs but were not otherwise used in this experiment, and then ran `GenotypeGVCFs` on all samples jointly. We next used `SelectVariants` to extract SNPs, and `VariantFiltration` to remove SNPs with FS score < 30 and QD > 2.0. We further filtered for SNPs called homozygous-reference in all B73 samples, and which exhibited allele frequencies $> 1/8$ and $> 7/8$ in either the PT-NIL or Mi21-NIL populations (expected frequencies of each variant should be 0.25 or 0.5 depending on recombination between the two BC_5 parents of each population, but we allowed for some sampling error). We used this set of highly filtered SNPs for each population to genotype each of the NIL plants. For each plant, at each locus we first combined all genotype likelihoods across all RNA samples from the same plant. We then identified the approximate breakpoints of the introgressed regions by inspecting the density of variant sites. We identified 3 regions (on chromosomes 2, 4 and 5) in the PT-NIL population and 2 regions in the Mi21-NIL population (on chromosomes 3 and 4). Within these introgressed regions, we used `R/QTL` [41] to assign genotype probabilities across the *Inv4m* locus for each plant, allowing `error.prob = 0.2`. Finally, we observed that several genes outside these 5 introgressed regions each of which exhibited >= 2 SNPs relative to the reference. We hypothesized that these genes have probably a different chromosome location in the landraces relative to B73, and actually reside inside one of the 5 introgressed regions. We therefore assigned their genotype to the most common genotype among these variant loci.

RNA quantification

To quantify gene expression, we ran `kallisto` v.0.42.3 [42] separately on each sample using the B73 AGPv4.36 transcript models downloaded from the maize genome database [37]. We limited to only one bootstrap replicate, and then combined transcripts from the same locus into a total estimated transcript count per gene. Genes were retained where at a third of the samples had 10 or more reads. Gene counts were normalized using the weighted trimmed mean of M-values (TMM) with the *calcNormFactors* function in *edgeR* [43]. Normalization using TMM reduces bias of very highly and lowly expressed genes. The *voom* function [44] in the *limma* package [45] was used to convert normalized reads to log2-counts per million (log2CPM), estimate a mean-variance relationship, and assign each observation a weight based on its predicted variance. Observation-weights were used in downstream analyses to account for heteroscedasticity. We estimated batch effects using the *removeBatchEffects* function in limma using the experimental replicate as batch, which corrected the log2CPM expression values. Global patterns of gene expression across the

experiment were visualized with the *plotMDS* function from *edgeR*.

Analysis of *Inv4m* effects on gene expression

We divided genes into three groups to estimate the effects of *Inv4m* or other introgressed landrace alleles, based on whether each gene resided in a "clean" genomic region with only B73's allele present in the populations, inside the *Inv4m* locus itself, or if it resided within one of the genomic blocks containing residual landrace DNA but outside *Inv4m*. Each group of genes served a different purpose in the analysis of *Inv4m*. Genes in the "clean" region were used to assess the effects of *Inv4m* on global gene expression and indirectly assess the effects on development and physiology more broadly. Genes inside the *Inv4m* locus were scanned for candidate alleles underlying *Inv4m's* effects. Genes in the residual introgression blocks were used as controls to assess the similarity of PT and Mi21 alleles in other genomic loci, as well as compare effect size and expression correlation with *Inv4m*.

For genes that resided in "clean" genomic regions with only B73's allele present in the populations (approximately 89.8% of genes expressed in both donors), we estimated the effect of the *Inv4m* locus separately in each *Inv4m* donor, temperature treatment, and tissue using the linear model:

$$y_{ij} = \mu + \beta \mathrm{Inv4m}_i + e_{ij} \qquad e_{ij} \sim \mathrm{N}(0, \phi_{ij}\sigma^2),$$

where y_{ij} is a normalized, batch-corrected log2CPM value for a single gene in a single sample, μ is the intercept for that gene in the particular population, environment and tissue, β is the corresponding effect of the landrace *Inv4m* allele, and e_{ij} is the model residual, which is assumed to be independent of all other residuals and have variance proportion to ϕ_{ij}, the empirical weight factor calculated by `voom`. We fit this model to the whole set of "clean" genes using the *lmFit* function from *limma*, and extracted β and its standard error ($|\beta|/t$). We leveraged the correlations in effect sizes across tissues and environments to improve effect size estimates and identify a union set of genes regulated by *Inv4m* by combining results across tissues and environments using the *mash* method [46] implemented in the *mashr* R package. *mash* was run separately for the two NIL populations.

We also fit a separate model to test for interactions between *Inv4m* and the temperature environment, separately for each *Inv4m* donor and tissue:

$$y_{ijk} = \mu + \beta_1 \mathrm{Inv4m}_i + \beta_2 \mathrm{Temp}_j + \beta_3 \mathrm{Inv4m{:}Temp}_{ij} + e_{ijk} \qquad e_{ijk} \sim \mathrm{N}(0, \phi_{ij}\sigma^2).$$

This model adds β_2, the main effect of temperature environment on expression, and β_3, the interaction between *Inv4m* and temperature. However it is less flexible than the first model because the residual variance σ^2 is constrained to be equal for the two temperature environments. This model was also fit to each gene using *lmFit*, and the estimate of β_3 and its standard error were extracted. We again used *mashr* to identify the union set of genes affected by this interaction.

For genes residing inside the *Inv4m* locus, we fit the same two statistical models with the *lmFit* function (both sets of genes were analyzed jointly to leverage the empirical bayes shrinkage of standard errors). However, we did not include these genes in the multiple adaptive shrinkage analysis.

For genes residing outside *Inv4m*, but within one of the genomic blocks containing residual landrace DNA in both donors, we fit a slightly different statistical model:

$$y_{ij} = \mu + \beta \text{cis}_i + e_{ij} \qquad e_{ij} \sim \text{N}(0, \phi_{ij}\sigma^2),$$

where *cis* is the local genotype of the gene, and β is the associated effect. For genes in residual genomic blocks on chromosome 4, the *cis* genotypes were highly correlated with the *Inv4m*, so some of the *cis* effect may have been caused by *Inv4m*, but these effects were difficult to separate statistically. However for genes on other chromosomes, the two genotypes were largely uncorrelated.

Sequence divergence, and expression correlation between alleles within the High *Inv4m* allele and the residual introgressed regions

We estimated the sequence divergence between the two landrace donors and B73 for each gene within the chromosome 4 introgression containing *Inv4m* present in both populations. For each gene, we calculated the genetic similarity of the PT and Mi21 alleles relative to B73 by counting the number of shared SNPs divided by total number of observed SNPs within each gene window, using only the highly filtered SNP set described above. We calculated the correlation in effect sizes between donors for each tissue and temperature combination. T-tests were used to determine whether genetic similarity and expression correlation were higher within *Inv4m* relative to the shared introgressed region.

Gene ontology enrichment

Genes were assigned to gene ontology (GO) categories for functional annotation using an updated ontology annotation [47] which we expanded to include all ancestral terms for each gene with the *buildGOmap*

function of the R package clusterProfiler [48]. Genes in "clean" genomic regions which responded to the High *Inv4m* allele in the same direction in both donor populations were classified as *Inv4m-regulated* and tested as foreground genes in a GO enrichment analysis. Genes that were expressed in each tissue and temperature combination in both donor populations, but were not *Inv4m-regulated* were included in the set of background genes. We calculated the enrichment of each GO term using the *enricher* function in the clusterProfiler R package. We selected GO terms with a false discovery rate less than 1% after a Benjamini-Hochberg multiple test correction. We then calculated the percent of genes in each GO terms that were *Inv4m-regulated*, and using the highest enriched GO term across conditions and ranked all GO terms by their maximum enrichment. We then selected the highest enriched GO term among terms that had a semantic-similarity >0.5.

Candidate gene pathway assessment among *Inv4m-regulated* genes

We inspected two additional candidate gene sets: genes known to regulate flowering time in maize [49], and genes regulated by the microRNA miR172. miR172 is a highly conserved micro-RNA across the plant kingdom that regulates development and flowering time. These lists were selected, as *Inv4m* has previously been associated with flowering [25]. For miR172 targets, we found the mature sequence for zma-miR172c: "AGAAUCUUGAUGAUGCUGCA" from miRBase `http://www.mirbase.org`, and used this as a query of the Plant Small RNA Target Analysis Server (psRNATarget, `http://plantgrn.noble.org/psRNATarget`), and collected all predicted target genes. We also used TAPIR's pre-computed target genes for *zma-miR172a-b-c-d*. These two categories of genes were inspected by hand for evidence of regulation by *Inv4m*.

Candidate genes within *Inv4m*

We used the intersection of two separate methods to identify candidate adaptive genes within *Inv4m*. First, we quantified the proportion of conditions (tissue:temperature combinations) that each *Inv4m* gene was differentially expressed in. To be considered differentially expressed, the gene needed to be differentially expressed in both *Inv4m* donor populations and in the same direction. Genes where at least one donor was not expressed were removed from this analysis per condition. The second approach we used was to quantify the proportion of "clean" *Inv4m-regulated* genes, that each gene within *Inv4m* was significantly correlated with in the same direction in both *Inv4m* donors. For this analysis, we used the *lm* function in R to implement a linear model with the *Inv4m-regulated* gene expression as the response variable, and the *Inv4m* gene's expression and *Inv4m* genotype as predictors, and included an error term.

De-novo assembly of novel genes

We collected all un-mapped reads from the samples homozygous for the highland (PT or Mi21) alleles at *Inv4m*, and used Trinity v2.4 [50] to assemble un-annotated transcripts using default settings. We then used Kallisto to quantify the expression of each of these novel transcripts using the un-mapped reads from each RNAseq sample. To search for candidate "novel" genes in the highland allele, we filtered for Trinity genes that had zero estimated counts in any of the samples that were homozygous for the B73 allele of *Inv4m*, but had non-zero estimated counts in at least 2 samples homozygous for the PT or Mi21 alleles in each NIL population (to exclude genes that may reside in either PT or Mi21).

Results

To confirm the population genetic signature of local adaptation at *Inv4m*, we genotyped 4845 maize plants from the SeeD-maize GWAS panel for the *Inv4m* inversion using published unimputed genotype-by-sequencing data. Of these lines, 707 were homozygous for the minor allele, and 351 were heterozygous for the locus. Of the 585 plants carrying at least one of the minor alleles and complete with geographic information, all but 7 were from Mexico, with the majority collected from the central highlands (Fig 1A). Therefore, to assess evidence of local adaptation, we assessed the association of *Inv4m* genotype with elevation among the 1757 Mexican plants. In Mexico, 1186 and 381 plants were homozygous for the alternate alleles, and 190 were heterozygous, a distribution that significantly differs from Hardy-Weinberg expectations (D=-252.0, p = 1.91e-197). Genotypes at *Inv4m* were strongly associated with elevation, as previously reported [25] Fig 1B. The highland allele at *Inv4m* had much lower genetic diversity across the locus; however diversity measures rebounded immediately outside of the published boundaries of the inversion. Plants carrying the highland allele did show a slightly lower proportion of segregating sites (θ) across chromosome 4 (Fig 1C). Diversity estimates were relatively constant across the "High" allele at the *Inv4m* locus, with little evidence of large-scale introgression of lowland alleles; only 5% of markers had segregating variation in common (MAF > 0.05) in both haplotypes at the locus, and some of these may have been caused by cryptic paralogous loci which is common in GBS [51]. The lower diversity estimates are unlikely to be caused entirely by mapping biases against this divergent allele. Of the 9201 GBS markers within the locus, $\sim 80\%$ were successfully genotyped in $> 5\%$ of the highland individuals, and of these markers, 14% were segregating in the highlands and 33% were segregating in the lowlands. Among these markers, rates of missing genotypes were similar between the two alleles. If all of the un-scored markers were actually present and variable in the highland individuals, there would still be fewer segregating positions

among highland individuals than lowland individuals.

Fig 1

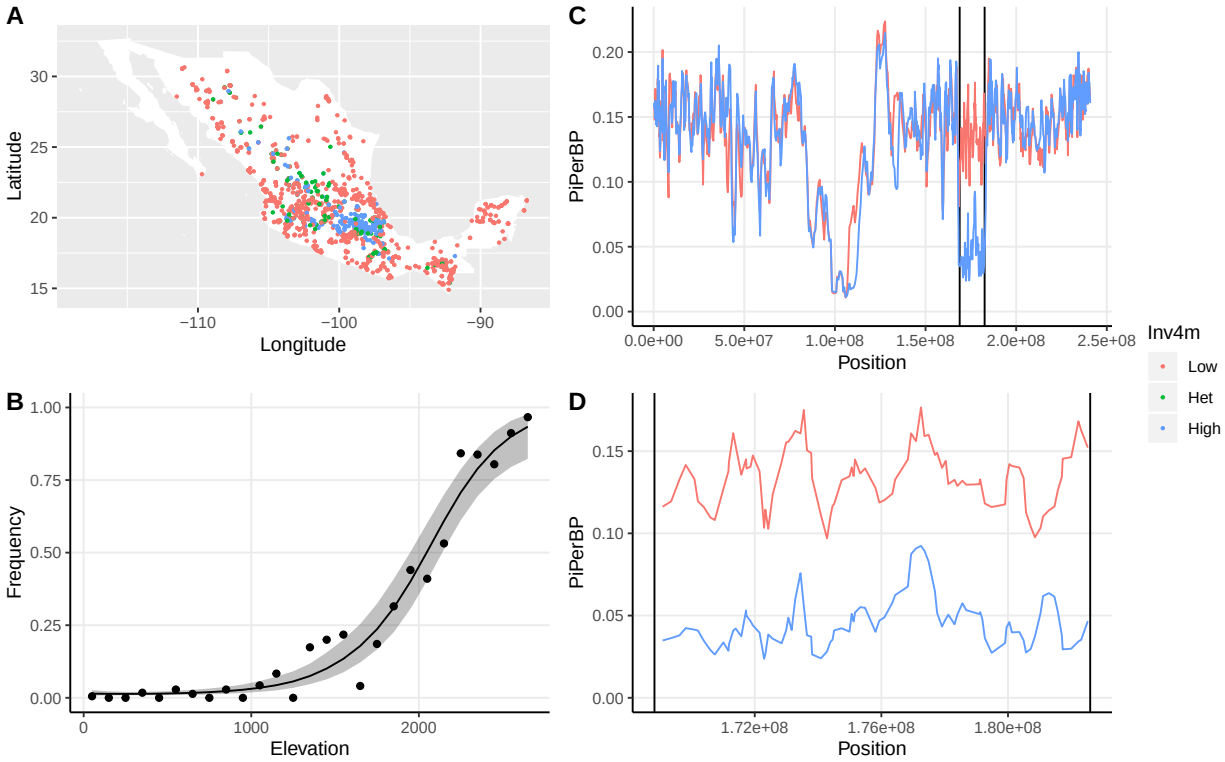

Fig 1. Association of *Inv4m* genotype with environmental factors and agronomic traits A. Geographic locations for each of the 1757 Mexican plants genotyped by GBS, colored by their imputed genotypes at *Inv4m*. **B**. Association of *Inv4m* and elevation. Each point shows the mean frequency of the "High" allele at the *Inv4m* locus among plants from landraces collected in each 100m bin. The ribbon shows a loess fit ($\pm$ 2SE) to the logit-transformed frequencies weighted by the number of landraces in each elevation bin. Bins with fewer than 10 landraces were excluded (those with elevation >2700m). **C** and **D**. Diversity estimates π and θ for 371 sampled plants homozygous for the "Low" allele and 371 plants homozygous for the "High" allele at *Inv4m*, along chromosome 4. The boundaries of *Inv4m* from [23] are denoted by vertical lines.

Romero-Navarro *et al* [25] identified *Inv4m* as a large-effect QTL in a large multi-environment trial of landrace hybrids grown in multiple field sites in Mexico. Gates *et al* [29] analyzed five additional agronomic traits from some of these field trials and found evidence for effects of the *Inv4m* locus on several traits. However, these studies did not explicitly show effect sizes for *Inv4m* across trials or traits, and none of the individual SNP markers in these studies was perfectly associated with our *Inv4m* genotype. Therefore, we re-analyzed the phenotype dataset focusing specifically on estimating the effect of *Inv4m*, and how this effect changed across the elevations of the trials.

The highland allele of *Inv4m* was significantly associated with Days-to-Anthesis, the Anthesis-Silking Interval (ASI), Grain Weight-per-hectare, Bare cob weight, and Field Weight, but in each case, the effect size

changed across elevations in the direction consistent with local adaptation: earlier flowering, reduced ASI, and increased yield components in highland trials, while the opposite in lowland trials (Fig 2). The highland allele was weakly associated with greater plant height, but the relationship was not significant (p=0.11 for the main effect). The relationship between *Inv4m* genotype and these traits was not simply an indirect effect of the change in flowering time; each relationship remained qualitatively the same even after accounting for the the effect of Days-to-Anthesis separately within each trial. However, even though we attempted to account for population structure in our analyses, it is still possible that some of these results remain confounded due to the relatedness among individuals; *Inv4m* is among the markers best-correlated with the elevation of origin, and the first eigenvector of genetic variation across the remainder of the genome (excluding chromosome 4) is also strongly correlated with elevation. Therefore, these results are also consistent with a polygenic basis for the divergence of each of these traits along elevational gradients. Functionally validating the association of traits with *Inv4m* therefore requires experimentally breaking the association between *Inv4m* and the rest of the genome through experimental crosses.

Fig 2

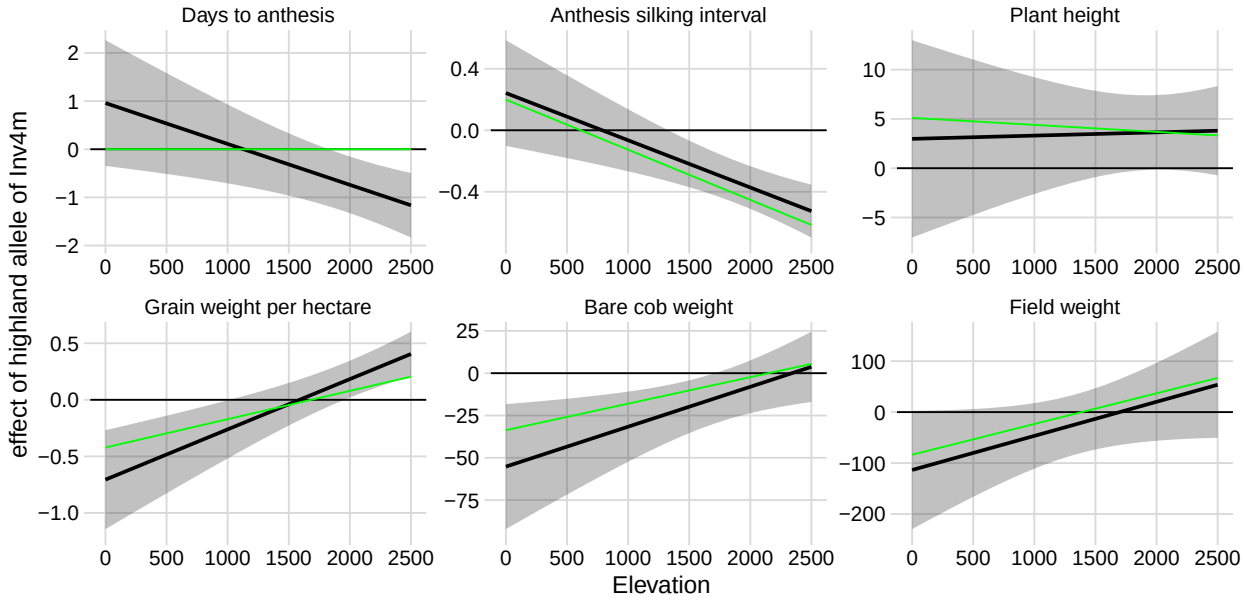

Fig 2. Association of *Inv4m* genotype with agronomic traits across depended on trial elevation We modeled each trait as a function of *Inv4m* genotype, trial elevation, and tester line, with controls for main effects and responses to elevation of the genomic background. Black lines and ribbons show estimates of the effect of the highland allele of *Inv4m* as a function of trial elevation ± 2SE, based on conditional F-tests at the REML solutions of the random effect variance components. Green lines show estimates of the *Inv4m* effect in a model that additionally included effects of Days-to-Anthesis on the focal trait within each trial.

To directly measure phenotypic effects of the *Inv4m* locus, we created two Near Isogenic Line (NIL)

populations by introgressing *Inv4m* alleles from two highland *Inv4m* donors (PT and Mi21) into B73, a genetic reference maize line. We grew a total of 456 plants from the BC_5S_1 generations in a replicated growth chamber experiment under two temperature treatments (warm: 32C/22C and cold: 22C/11C), and harvested nine tissues for gene expression analysis of the global and local effects of the alternative *Inv4m* alleles.

Genotyping of NILs sampled from the growth chamber experiment was successful for 364 plants, while 92 were not resolved either due to failed DNA extraction or during the CAPs genotyping methods. The Low, Heterozygote, and High *Inv4m* alleles were segregating in the PT NILs at a 42:78:45 ratio, within Hardy-Weinberg equilibrium (HWE; D= -2.24, p-value=0.53). The inversion was segregating in the Mi21 NILs at a 59:97:43 ratio, also within HWE (D= -0.928392, p-value=0.7768964). We next checked if the High *Inv4m* allele had an effect on time to seedling emergence using a linear mixed effect model, with *Inv4m* genotype as the fixed effect, and experimental replicate as the random effect. We ran separate models by donor and temperature, and found that in the cold chamber the average time to emergence was approximately 9 days, and the High *Inv4m* allele had a -0.75 and -0.35 effect size on emergence time in PT and Mi21 introgression lines respectively (Table S1). In the warm chamber, the average emergence time was approximately 4 days, and the High *Inv4m* allele had no effect (Table S1).

Based on genotype calls from the RNA-seq reads, both populations had extensive landrace introgressions flanking the *Inv4m* locus despite the five generations of backcrossing to B73, extending 57Mb on either side in the PT-NIL population and 18Mb in the Mi21-NIL population (Fig 3A). Additionally, the PT-NIL population segregates for a large paracentric introgression on chromosome 5 and a small introgression on the right end of chromosome 2, and the Mi21-NIL population segregates for a large paracentric residual introgression on chromosome 3 (Figure S2). Beyond these large contiguous blocks, we identified another 821 and 52 genes in the PT-NIL and Mi21-NIL populations, respectively, that harbored high-confidence SNPs in the RNAseq data, yet were not contiguous with any of the large residual introgression regions. It is unlikely that there was sufficient recombination in the BC_5 populations to generate these independent blocks; rather these genes likely have moved genomic coordinates in the landraces relative to B73, and actually reside inside one of the large introgression regions [52,53]. However, none of these genes had genotypes that were perfectly correlated with genotypes at the *Inv4m* locus, so we excluded them all from further analysis.

Only 355 of the 7,236 or 4,095 genes with PT or Mi21 alleles actually reside inside *Inv4m*. Of genes with genotypes strongly correlated with genotypes at the *Inv4m* locus ($\rho > 0.5$), only 14% and 29% in the PT-NIL and Mi21-NIL populations, respectively resided inside *Inv4m*. Therefore, many associations of traits with genotypes at the *Inv4m* locus that we observe in one population may be caused by functional variants in regions flaking the inversion, rather than functional variants inside the inversion itself. However, only variants

actually inside the inversion (or in the breakpoint regions) are likely to be causal for the locus's effect. On the other hand, 23% of genes with landrace alleles in both populations are inside the inversion, and for those that carry alleles from both donors, the two donor alleles are more similar (relative to B73) inside *Inv4m* than outside, as expected based on the population-level variation statistics from the SeeD-maize GWAS panel. The proportion of shared SNPs between PT and Mi21 alleles within *Inv4m* was significantly higher than in the flanking introgressed regions (Fig 3B, t = -10.773, df = 357.8, p-value < 2.2e-16). The same pattern occurs for gene expression variation attributed to landrace genotypes for each gene: the landrace genotype's effect on genes inside *Inv4m* are more correlated than landrace genotype effects in flanking regions (Fig 3C, t = -2.8869, df = 27.768, p-value=0.007) across the 18 tissue:temperature combinations in our experiment. Therefore, phenotype effects associated with *Inv4m* that replicate across the two donor populations are likely caused by functional variation inside *Inv4m* rather than in residual landrace DNA present in each population (Fig 3). We also quantified the correlation between genetic and expression divergence between *Inv4m* sources, and found that they were positively correlated (r=0.79, p=2.2e-16, df=76, r^2=0.6357).

Of the 432 tissue samples collected (nine tissues × two temperature treatments × two NIL populations × two *Inv4m* arrangements × three biological replicates × two experimental replicates), we excluded 53 samples with fewer than 100,000 reads, leaving a total of 379 samples. We detected a total of 23,428 unique genes with an average of at least 10 transcript counts per sample in at least one of the tissue:treatment combinations (with an average of 17,016 genes per tissue) for a total of more than 306,000 gene expression traits. We visualized the overall transcriptome variation among samples using multi-dimensional scaling (Figure S3). Samples clustered predominantly by tissue, with and additional slight separation by temperature treatment, but no visible separation by NIL population or genotype at the *Inv4m* locus. This was expected because only $\sim 7\%$ of the genomes differed among samples. Among tissues, all leaf samples except the S2 leaf base formed one major cluster and the two root samples formed a second cluster, and the stem and SAM and S2 leaf base tissues formed separate individual clusters.

To assess the effects of highland alleles at *Inv4m* on global gene expression and plant development and physiology, we focused on genes with expression associated with *Inv4m* genotype that resided in "clean" genomic regions, so that the local genotype of each gene could not affect the association results. Overall, we identified 11,842 unique genes associated with genotype at *Inv4m* in the PT-NIL population and 12,482 genes in the Mi21-NIL population, both using a 5% local false sign rate (*lfsr*) threshold for significance (Table 1). The number of associated genes varied by tissue and temperature treatment with a range of 1,932-5,753 and 1,018-6,000 genes identified per tissue in the PT-NIL and Mi21-NIL populations respectively. Of these genes, 285-1646 per tissue replicated across both NIL populations, where replication required *lfsr*

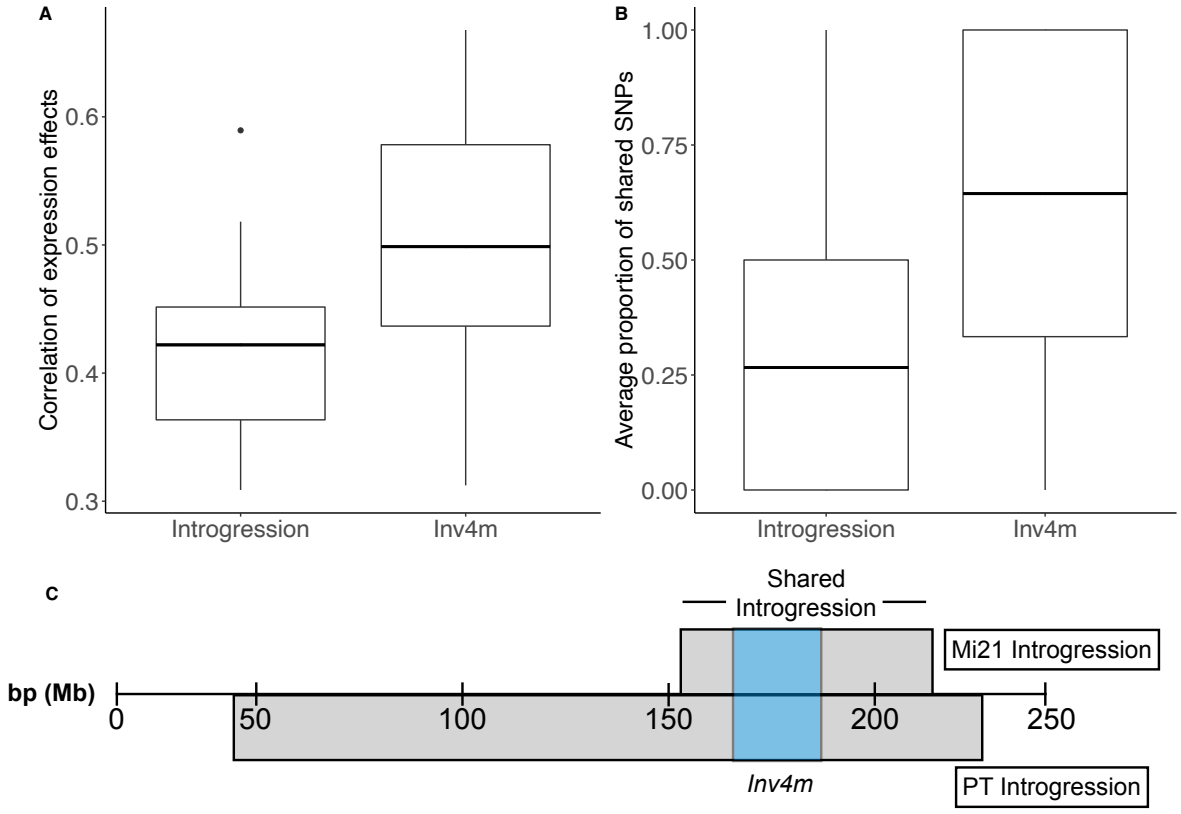

Fig 3. Barplots comparing the similarity of expression and genetic divergence between landraces relative to B73. A) The expression effects correlation between donors for each tissue and temperature combination and B) the average proportion of shared SNPs per gene. C) diagram of PT and Mi21 introgressions containing *Inv4m* on chromosome 4.

$< 5\%$ and the effect in the same direction in both populations in each tissue:temperature where the effect was significant. This reduced list of genes constitutes *Inv4m-regulated* candidate genes, and constituted $8 - 41\%$ of the differentially expressed genes in the PT-NIL population, and $11 - 38\%$ of the differentially expressed genes in the Mi21-NIL population. Candidate *Inv4m-regulated* genes were distributed across the genome, with no visible clustering by chromosome (data not shown).

In contrast, we detected only 38-413 genes per tissue in the PT-NIL population and 435-2398 genes per tissue in the Mi21-NIL population with significant genotype-treatment interactions at a 5% *lfsr* threshold. Of these, 4-23 genes per tissue were shared across the two populations (Table 2).

We extracted the Gene Ontology (GO) terms associated with *Inv4m*-regulated genes and tested for enriched terms. For each tissue:treatment experiment, we separately tested for enrichments among the genes

Table 2. Number of differentially expressed genes in each population and their overlap using a local false sign rate threshold of 5% across tissues and temperature treatments.

	Cold			Warm			G × E		
Tissue	PT	Mi21	Shared	PT	Mi21	Shared	PT	Mi21	Shared
V1 Leaf tissue	2415	1534	285	2107	2330	453	105	575	6
V1 Primary root	1932	3516	495	3106	2759	649	84	1377	8
V3 leaf base	2550	3785	789	2809	3761	932	73	843	6
V3 leaf sheath	4047	5126	1646	3321	4049	1062	469	742	26
V3 Primary root	5753	3143	1206	2144	2039	413	42	455	NA
V3 S2 leaf base	3527	1018	268	NA	5665	NA	NA	2573	NA
V3 S2 leaf tip	2806	3054	721	4170	3190	1249	136	1527	19
V3 Stem and SAM	3118	6000	1269	2569	2961	752	190	1509	24
V3 Leaf tip	2207	NA	NA	4225	2453	997	362	NA	NA
Totals	12574	12485	2649	11468	12233(20.7%)	2542	1306	5006	85

up-regulated, down-regulated, or both by *Inv4m*. We identified 596 enriched categories overall, with 0-152 categories enriched per tissue and temperature treatment at a 5% FDR. These GO terms provide candidate descriptors of the global effects of *Inv4m*. To reduce the number of categories found across the tissue:temperature treatments, we collapsed terms into clusters by semantic similarity and selected the most-enriched term across all tissue:temperature treatments in each cluster. After filtering for redundancy, we report twenty-two GO terms across the three main GO ontologies: two cellular component terms, two molecular function terms, and eighteen biological processes terms (Table 3).

Table 3. Gene ontology (GO) terms remaining after final filtering. A universal enrichment analysis was conducted on each tissue and temperature and directional (up-regulated, down-regulated, or both) combination for *Inv4m*-regulated genes. Terms were then ranked by enrichment score and grouped by a semantic similarity score of higher than 0.5. The top term in each semantic similarity group was then selected.

ID	Ontology	Term	minq	maxRatio
GO:0051169	BP	nuclear transport	1.47e-06	9.09e-02
GO:0006364	BP	rRNA processing	1.04e-04	9.02e-02
GO:2000241	BP	regulation of reproductive process	1.77e-03	8.80e-02
GO:0017038	BP	protein import	1.29e-05	8.61e-02
GO:0006195	BP	purine nucleotide catabolic process	4.14e-03	8.33e-02
GO:0048831	BP	regulation of shoot system development	1.77e-03	7.87e-02
GO:0009886	BP	post-embryonic animal morphogenesis	5.48e-03	7.41e-02
GO:1903047	BP	mitotic cell cycle process	6.25e-03	7.41e-02
GO:0016458	BP	gene silencing	6.92e-03	7.41e-02
GO:0016571	BP	histone methylation	1.92e-03	6.94e-02
GO:0030258	BP	lipid modification	5.44e-03	6.85e-02
GO:0072524	BP	pyridine-containing compound metabolic process	1.77e-03	6.49e-02
GO:0006310	BP	DNA recombination	9.73e-03	6.22e-02
GO:0009629	BP	response to gravity	5.10e-03	6.09e-02
GO:0009561	BP	megagametogenesis	4.85e-04	5.74e-02
GO:0019682	BP	glyceraldehyde-3-phosphate metabolic process	1.60e-03	5.59e-02
GO:0048498	BP	establishment of petal orientation	4.42e-04	5.56e-02
GO:0007034	BP	vacuolar transport	1.13e-04	5.11e-02
GO:0019899	MF	enzyme binding	7.80e-03	7.87e-02
GO:0004386	MF	helicase activity	1.77e-03	6.48e-02
GO:0044451	CC	nucleoplasm part	9.96e-03	7.66e-02
GO:0030684	CC	preribosome	1.29e-06	5.74e-02

To identify candidate gene sets within the *Inv4m*, we measured the *Inv4m* genotype effect (i.e. *cis* effect) on each of the annotated genes located inside the *Inv4m* locus, and the correlation between the expression of

these *Inv4m*-genes and all *Inv4m-regulated* genes (a total of 4642 unique genes, with a range of 89-713 genes per tissue; Figure4). We repeated this analysis in each of the 18 tissue:temperature treatments. For *cis*-genotype effects to be counted, we required the effects to be significant ($lfsr < 5\%$) and in the same direction between the two donor populations.

Overall, of 355 annotated genes within the boundaries of *Inv4m*, 224 were expressed in both donors. Of those, 155 were differentially expressed in the same direction in both donors in at least one treatment, and 89 of those were significantly correlated with at least one *Inv4m-regulated* genes located on other chromosomes used to measure the global *Inv4m effect*, even after accounting for the effect of *Inv4m* itself (Figure 4). Of these, 6 genes in particular stood out as being correlated with a large number ($> 3\%$) of the reporter genes: Zm00001d051908, Zm00001d051998, Zm00001d052075, Zm00001d052079, Zm00001d052153, Zm00001d052259, and 9 were differentially expressed between *Inv4m* alleles in 90% or more of the tissue:temperature combinations. These were: Zm00001d051872, Zm00001d051882, Zm00001d051987, Zm00001d052051, Zm00001d052136, Zm00001d052210, Zm00001d052242, Zm00001d052245, Zm00001d052269. Finally, three of the 89 candidate *Inv4m* genes were transcription factors: Zm00001d051879, Zm00001d052180, Zm00001d052229. These genes are reported in Table S1, with associated description, GO term(s), and previous GWAS trait associations.

Since the phenotypic effect of *Inv4m* on flowering time is so large, we tested specifically for effects of the locus on the expression of genes with known roles in regulating flowering. Of the list of 48 flowering genes in Dong et. al. 2012 [49], 5 were consistently regulated by *Inv4m* in both populations in the same direction (Table S3). The 5 different flowering genes were *Inv4m*-regulated in the V1 and V3 primary root, S2 leaf tip, and stem and SAM tissues.

Another *a priori* candidate gene for the effect of the inversion is the microRNA MIR172c, which is located inside *Inv4m* at coordinates: 4:174154928-174155050. MIR172 has known roles in regulating developmental transitions as well the development of pigmentation and macrohairs in maize [54]. The expression of the pre-miRNA 172c was too low to assess differential expression in all but the leaf sheath tissue, and was not associated with *Inv4m* genotype in that tissue. However, of 31 genes that are predicted targets of miR172, 9 were consistently regulated by *Inv4m* in both populations in one-or-more tissues. Of these, six were down-regulated by the High allele of *Inv4m*. These genes and their associated descriptions are found in Table S3.

Finally, we used the RNAseq reads that did not map to the B73 genome to search for genes that may be present in the High allele of *Inv4m* but not in the Low allele, and thus may have been missed by our analysis. We assembled all un-mapped reads from samples carrying the PT or Mi21 alleles at *Inv4m* using Trinity, and

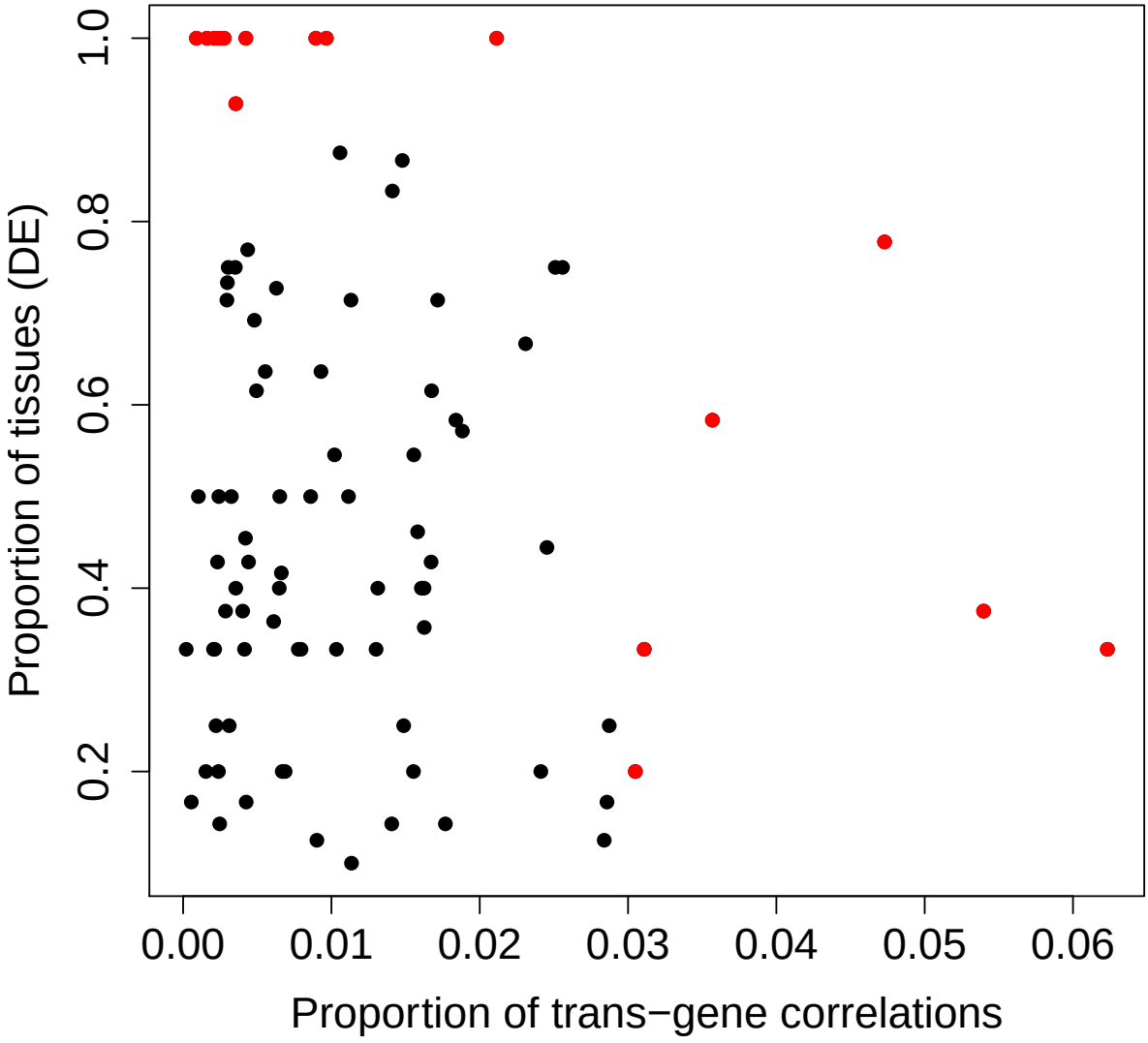

Fig 4. Scatter plot of candidate gene scans within *Inv4m*. Each point represents one of the 89 expressed genes inside the *Inv4m* locus which were also correlated with at least one *Inv4m-regulated* gene. The x-axis represents the proportion of the set of genes regulated by *Inv4m* in *trans* correlated with the *Inv4m*-gene (p-value$<$ 0.05). The y-axis is the proportion of tissues in which the *Inv4m*-gene is differentially expressed (lfsr$<$ 0.05 in both populations, and regulated in the same direction in both populations.)

searched for de-novo transcripts with evidence for expression only in the samples carrying the High allele. We found 772 candidate transcripts. However, all were very lowly expressed. Only five had estimated transcript counts of at least 10 summed across all *PT* or *Mi21* samples, and none had estimated transcript counts of at least 10 in two or more samples per NIL population. Therefore, we found no evidence that the

High allele of *Inv4m* carries high-expressed genes that are not present in the Low allele.

Discussion

We combined three methodological approaches to study the function of the chromosomal inversion, *Inv4m*, found in highland populations of domesticated maize (*Zea mays* ssp. *mays*). First, we used a large maize diversity panel designed for Genome-wide association studies (GWAS) to identify agronomic traits associated with *Inv4m*. Second, we created a two-population Nested Association Mapping (NAM) panel consisting of two sets of Near Isogenic Lines (NILs) segregating for both arrangements of the $\sim 13Mb$ inversion and grew them reciprocally in two temperature treatments to measure gene expression effects of *Inv4m* across nine tissues from early developmental plants. In total, we inspected *Inv4m* effects on more than 300,000 traits and employed statistical methods that leveraged the replication of the effects across conditions [46] to broadly characterize the effects of this locus. Finally, we inspected cis-regulatory variation on genes within the inversion to identify candidate genes responsible for the *Inv4m* effects.

0.1 Association analysis of inversion loci

GWASs in diversity panels and QTL mapping in biparental populations or NILs are complimentary methods for locating genetic loci associated with phenotypic variation. GWASs can pinpoint the location of functional variants more precisely than bi-parental mapping populations because they use diverse association panels that encompass a large number of ancestral recombination events [55]. However, GWAS approaches suffer from confounds due to population structure and non-equal relatedness leading to high frequencies of false positives [56]. QTL mapping populations, on the other hand, are limited in resolution due to the requirement for new recombination events between candidate loci.

Multi-parent populations like NAM [57] or MAGIC [58, 59] capture advantages of both methods, approaching the resolution of GWAS without the confounds of population structure. Our two NIL populations shared a common parent (B73), and thus compose a two-population NAM panel. NAM-QTL mapping can have high precision when the donor parents share the same alleles at a causal locus and do not share alleles at other loci. In our case, the two donor parents PT and Mi21 likely share many alleles other than *Inv4m* due to shared highland ancestry throughout the genome, so NILs from both parents will overlap at other causal alleles linked to *Inv4m*. However, our results show that the genetic similarity between PT and Mi21 is much higher inside *Inv4m* than in the remaining ~50Mb of shared introgression between the two populations. Therefore, while we cannot unambiguously attribute shared phenotypic effects across the two

populations to *Inv4m*, the majority of effects of *Inv4m* that we identified are likely caused by functional variation within this locus itself, rather than due to other shared alleles between the two donors.

Our GWAS study of diverse maize landraces confirmed earlier results [25, 29] that highland alleles of *Inv4m* are strongly associated with agronomically important phenotypes including flowering time, the Anthesis-Silking interval, and several measures of yield. In each case, the allelic effect of the highland allele was more beneficial when grown in highland environments than in lowland environments, consistent with this single locus causing antagonistic pleiotropy [60] across elevation environments. This may explain the strong evidence for selection at this locus: its divergence in allele frequencies across elevations and the evidence that this locus introgressed into highland maize from the wild relative *mexicana* [22]. That the locus appears to independently control both flowering and yield traits is consistent with the idea that *Inv4m* contains multiple important variants that each contribute to phenotypic differences between lowland and highland maize [25, 29]. Because they are inherited as a single unit, inversions are thought to contribute to the prevalence of local adaptation despite high gene flow [61–63] by linking these multiple variants into a single *supergene* [64].

However, while compelling, due to the strong population structure in the diversity panel used above, we caution that the associations of *Inv4m* with the agronomic traits is still preliminary. *Inv4m* genotypes are highly correlated to overall genetic ancestry (as measured by PC1 calculated using all other chromosomes except chromosome 4) within Mexico. We corrected for ancesetry using genome-wide kinship (again excluding chromosome 4), but if this correction was incomplete, it may have lead to false-positive associations of phenotypic traits with *Inv4m*. An alternative explanation for the phenotypic associations above is that each has a polygenic basis, with small-effect loci distributed throughout the genome, each of which has subtle allele-frequency differences across elevations. Unfortunately, our NILs grow poorly in high-elevation fields, so it has not yet been possible to test the effect of *Inv4m* more directly.

0.2 Progress towards inversion fine-mapping

Whether or not *Inv4m* directly controls flowering time and yield, neither GWAS not NAM-QTL mapping themselves can provide direct insight into which functional variant(s) *inside* the locus are responsible for these phenotypic effects. And unlike a typical GWAS peak that may cover dozens of possible variants in high LD with a tested marker, there may be hundreds of thousands of variants between the two alleles of *Inv4m* within this $\sim$ 13Mb region, any of which could be responsible for some of the phenotypic effect of the locus. Neither GWAS nor QTL mapping itself can prioritize any of these variants, because they remain in

near-perfect LD in either type of panel.

Ultimately, identifying specific functional loci within *Inv4m* will require experimental mutagensis or other genetic perturbations within the locus. However, we aimed to begin to characterize the diversity of functional variants in this locus using gene expression analysis. We used gene expression in three distinct ways to dissect the functional variation captured by *Inv4m*: 1) by analyzing genome-wide gene expression responses to *Inv4m* to mine for phenotypic effects across >300,000 traits; 2) by analyzing local *cis*-regulatory effects of the locus on the 355 genes within *Inv4m*; and 3) by analyzing the co-expression between *Inv4m* genes and the rest of the genome. The first analysis provided an exceptionally detailed phenotypic dissection of the total effects of the *Inv4m*, and showed that there are likely many distinct components to the cellular and physiological effects of *Inv4m*. The second analysis provided an estimate of the density of functional variants within the *Inv4m* locus: we detected likely *cis*-regulatory variation affecting 155 genes. The third analysis showed that many of these cis-regulatory variants may have functional consequences beyond the immediate genes they regulate.

By studying gene expression effects of *Inv4m* in two relevant environmental contexts (hot and cool temperatures) and across nine distinct tissues, we aimed to maximize our ability to discover developmental and phenotypic effects of the locus. It is certainly possible that we missed important phenotypic effects of *Inv4m* by sampling only tissues on young plants - effects on pathways specific to reproductive tissues were likely missed. However, we selected the nine tissues based on the published maize gene expression atlas [31] so as to capture as much variation as possible in expression profiles, given the experimental constraints on how large we could let the plants grow in our growth chambers.

Overall, our expression results show that *Inv4m* affects many disparate biological processes in young maize tissues. The strongest Gene Ontology enrichment signals among *Inv4m*-regulated genes were in terms related to mRNA and protein processing around the nucleus (nuclear transport and import, and the pre-ribosome). We also found evidence of effects on epigenetic regulation, cell-cycle processes, metabolism, and development. None of these results provide clear explanations for the effects of *Inv4m* on flowering time and yield. However both flowering and yield are highly complex traits that are affected by many aspects of development, physiology, and stress responses, and so the mechanistic links among these traits may not be obvious [57]. We looked more specifically at *a priori* candidate genes for flowering and yield traits both inside and outside *Inv4m* and found possible effects on several of these genes, but no strong enrichment of *Inv4m* effects on either class.

Among the genes within *Inv4m*, nearly 70% of those expressed high enough to measure showed evidence of *cis*-regulatory variation among alleles. While some of these genes may share regulatory elements, its likely

that the majority of these genes are affected by independent genetic variants. This suggests that the two alleles of *Inv4m* harbor a large number of functionally relevant genetic differences. However, does *Inv4m* harbor more functional variants than any other similarly sized introgression among maize landraces? To test this, we compared the number of genes (genome-wide) correlated with *Inv4m* in each NIL population to the number of genes that show similar expression in both NIL populations. The latter genes are those we believe are truly affected by *Inv4m*, while the remainder are likely regulated by PT or Mi21 alleles that reside in introgressed genomic outside of *Inv4m* in each population. In both populations, the proportion of *Inv4m* candidates among all *Inv4m*-correlated genes is roughly similar to the relative sizes of *Inv4m* to the whole chromosome 4 introgression in each population. This suggests that introgressing any region from PT or Mi21 into B73 will cause diverse effects on gene expression, and that *Inv4m* is not exceptional in the magnitude of these perturbations.

Together, these results imply that the *Inv4m* locus has many effects on corn development and physiology, and therefore it's contribution to local adaptation is complex and not simply a change to major flowering or yield-related genes. The incorporation of $\sim$ 13Mb of the genome of *mexicana* likely brought with it a large number of functional variants that have both positive and negative effects on many molecular traits, most of which are not directly visible, but may impact performance in different conditions.

Conclusions

This study represents a broad characterization of an adaptive chromosomal inversion. Our results give insight into the role of this inversion in adapting to high altitude environments. GWAS results show that *Inv4m* is associated with faster flowering and higher yield in highland common gardens. The molecular roles of genes within the inversion are summarised by the phenotypic effects of *Inv4m*-regulated genes (Table 2) and enriched GO terms (Table 3), and the candidate gene set within *Inv4m* (Table S2). Fine-mapping in this region is required to further dissect the functional role of loci within *Inv4m*, but will have additional challenges due to suppressed recombination between heterokaryotypes. Novel genomic technologies, such as a CRISPR/CAS system [65] that can reverse the orientation of the High *Inv4m* allele could be used to induce recombination across the newly collinear genomic region, allowing the localization of specific effects of the different variants linked in this locus.

Acknowledgments

We would like to thank Brittney Gillespie, Luis Avila, and Po-Kai Huang for help sampling tissue, and to the staff at the controlled environment facility for helping us set up and maintain light and temperature regimes in the growth chambers.

References

1. White MJD. Animal cytology and evolution. CUP Archive; 1977.
2. Dobzhansky T, Dobzhansky TG. Genetics of the evolutionary process. vol. 139. Columbia University Press; 1970.
3. Kirkpatrick M, Barton N. Chromosome inversions, local adaptation and speciation. Genetics. 2006;173(1):419–434.
4. Wellenreuther M, Bernatchez L. Eco-evolutionary genomics of chromosomal inversions. Trends in ecology & evolution. 2018;.
5. Donnelly MP, Paschou P, Grigorenko E, Gurwitz D, Mehdi SQ, Kajuna SL, et al. The distribution and most recent common ancestor of the 17q21 inversion in humans. The American Journal of Human Genetics. 2010;86(2):161–171.
6. Stefansson H, Helgason A, Thorleifsson G, Steinthorsdottir V, Masson G, Barnard J, et al. A common inversion under selection in Europeans. Nature genetics. 2005;37(2):129.
7. Krimbas CB, Powell JR. Drosophila inversion polymorphism. CRC press; 1992.
8. Anderson AR, Hoffmann AA, McKechnie SW, Umina PA, Weeks AR. The latitudinal cline in the In (3R) Payne inversion polymorphism has shifted in the last 20 years in Australian *Drosophila melanogaster* populations. Molecular Ecology. 2005;14(3):851–858.
9. Kapun M, Fabian DK, Goudet J, Flatt T. Genomic evidence for adaptive inversion clines in *Drosophila melanogaster*. Molecular biology and evolution. 2016;33(5):1317–1336.
10. Horton BM, Moore IT, Maney DL. New insights into the hormonal and behavioural correlates of polymorphism in white-throated sparrows, *Zonotrichia albicollis*. Animal behaviour. 2014;93:207–219.

11. Kunte K, Zhang W, Tenger-Trolander A, Palmer D, Martin A, Reed R, et al. Doublesex is a mimicry supergene. Nature. 2014;507(7491):229.

12. Joron M, Papa R, Beltrán M, Chamberlain N, Mavárez J, Baxter S, et al. A conserved supergene locus controls colour pattern diversity in *Heliconius* butterflies. PLoS biology. 2006;4(10):e303.

13. Wang J, Wurm Y, Nipitwattanaphon M, Riba-Grognuz O, Huang YC, Shoemaker D, et al. A Y-like social chromosome causes alternative colony organization in fire ants. Nature. 2013;493(7434):664.

14. Pearse DE, Miller MR, Abadía-Cardoso A, Garza JC. Rapid parallel evolution of standing variation in a single, complex, genomic region is associated with life history in steelhead/rainbow trout. Proceedings of the Royal Society of London B: Biological Sciences. 2014;281(1783):20140012.

15. Sturtevant A. A case of rearrangement of genes in *Drosophila*. Proceedings of the National Academy of Sciences. 1921;7(8):235–237.

16. Catanach A, Crowhurst R, Deng C, David C, Bernatchez L, Wellenreuther M. The genomic pool of standing structural variation outnumbers single nucleotide polymorphism by threefold in the marine teleost *Chrysophrys auratus*. Molecular ecology. 2019;28(6):1210–1223.

17. Thompson M, Jiggins C. Supergenes and their role in evolution. Heredity. 2014;113(1):1–8.

18. Ayala D, Zhang S, Chateau M, Fouet C, Morlais I, Costantini C, et al. Association mapping desiccation resistance within chromosomal inversions in the African malaria vector *Anopheles gambiae*. Molecular ecology. 2019;28(6):1333–1342.

19. Matsuoka Y, Vigouroux Y, Goodman MM, Sanchez J, Buckler E, Doebley J. A single domestication for maize shown by multilocus microsatellite genotyping. Proceedings of the National Academy of Sciences. 2002;99(9):6080–6084.

20. Van Heerwaarden J, Doebley J, Briggs WH, Glaubitz JC, Goodman MM, Gonzalez JdJS, et al. Genetic signals of origin, spread, and introgression in a large sample of maize landraces. Proceedings of the National Academy of Sciences. 2011;108(3):1088–1092.

21. Mercer KL, Perales H. Structure of local adaptation across the landscape: flowering time and fitness in Mexican maize (*Zea mays* L. subsp. *mays*) landraces. Genetic Resources and Crop Evolution. 2019;66(1):27–45.

22. Hufford MB, Martínez-Meyer E, Gaut BS, Eguiarte LE, Tenaillon MI. Inferences from the historical distribution of wild and domesticated maize provide ecological and evolutionary insight. PLoS One. 2012;7(11):e47659.

23. Pyhäjärvi T, Hufford MB, Mezmouk S, Ross-Ibarra J. Complex patterns of local adaptation in teosinte. Genome biology and evolution. 2013;5(9):1594–1609.

24. Hufford MB, Lubinksy P, Pyhäjärvi T, Devengenzo MT, Ellstrand NC, Ross-Ibarra J. The genomic signature of crop-wild introgression in maize. PLoS Genetics. 2013;9(5):e1003477.

25. Navarro JAR, Willcox M, Burgueño J, Romay C, Swarts K, Trachsel S, et al. A study of allelic diversity underlying flowering-time adaptation in maize landraces. Nature genetics. 2017;49(3):476.

26. Lauter N, Gustus C, Westerbergh A, Doebley J. The inheritance and evolution of leaf pigmentation and pubescence in teosinte. Genetics. 2004;167(4):1949–1959.

27. Hearne CBEMS Sarah; Chen. Unimputed GbS derived SNPs for maize landrace accessions represented in the SeeD-maize GWAS panel. CIMMYT Research Data and Software Repository Network, V5. 2014;.

28. Glaubitz JC, Casstevens TM, Lu F, Harriman J, Elshire RJ, Sun Q, et al. TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. PloS one. 2014;9(2):e90346.

29. Gates DJ, Runcie D, Janzen GM, Navarro AR, Willcox M, Sonder K, et al. Single-gene resolution of locally adaptive genetic variation in Mexican maize. bioRxiv. 2019; p. 706739.

30. Runcie DE, Crawford L. Fast and flexible linear mixed models for genome-wide genetics. PLOS Genetics. 2019;15(2):1–24. doi:10.1371/journal.pgen.1007978.

31. Sekhon RS, Lin H, Childs KL, Hansey CN, Buell CR, de Leon N, et al. Genome-wide atlas of transcription during maize development. The Plant Journal. 2011;66(4):553–563.

32. Townsley BT, Covington MF, Ichihashi Y, Zumstein K, Sinha NR. BrAD-seq: Breath Adapter Directional sequencing: a streamlined, ultra-simple and fast library preparation protocol for strand specific mRNA library construction. Frontiers in plant science. 2015;6.

33. Andrews S. FastQC v0.11.5. A quality control tool for high throughput sequence data. Retrieved from: http://wwwbioinformatic sbabrahamacuk/projects/fastqc/. 2015;.

34. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114–2120.

35. Bates D, Maechler M, Bolker B, Walker S. lme4: Linear mixed-effects models using Eigen and S4. R package version. 2014;1(7):1–23.

36. Halekoh U, Højsgaard S, et al. A kenward-roger approximation and parametric bootstrap methods for tests in linear mixed models–the R package pbkrtest. Journal of Statistical Software. 2014;59(9):1–30.

37. Portwood JL, Woodhouse MR, Cannon EK, Gardiner JM, Harper LC, Schaeffer ML, et al. MaizeGDB 2018: the maize multi-genome genetics and genomics database. Nucleic acids research. 2018;47(D1):D1146–D1154.

38. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nature methods. 2015;12(4):357.

39. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research. 2010;20(9):1297–1303.

40. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature genetics. 2011;43(5):491.

41. Broman KW, Wu H, Sen Ś, Churchill GA. R/qtl: QTL mapping in experimental crosses. Bioinformatics. 2003;19(7):889–890.

42. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nature biotechnology. 2016;34(5):525.

43. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139–140.

44. Law CW, Chen Y, Shi W, Smyth GK. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. Genome biology. 2014;15(2):R29.

45. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic acids research. 2015;43(7):e47–e47.

46. Urbut SM, Wang G, Carbonetto P, Stephens M. Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. Nature Publishing Group; 2018.

47. Wimalanathan K, Friedberg I, Andorf CM, Lawrence-Dill CJ. Maize GO Annotation—Methods, Evaluation, and Review (maize-GAMER). Plant Direct. 2018;2(4):e00052.

48. Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. Omics: a journal of integrative biology. 2012;16(5):284–287.

49. Dong Z, Danilevskaya O, Abadie T, Messina C, Coles N, Cooper M. A gene regulatory network model for floral transition of the shoot apex in maize and its dynamic modeling. PLoS One. 2012;7(8):e43450.

50. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature biotechnology. 2011;29(7):644.

51. Torkamaneh D, Laroche J, Belzile F. Genome-Wide SNP Calling from Genotyping by Sequencing (GBS) Data: A Comparison of Seven Pipelines and Two Sequencing Technologies. PLOS ONE. 2016;11(8):1–14. doi:10.1371/journal.pone.0161333.

52. Eichten SR, Foerster JM, de Leon N, Kai Y, Yeh CT, Liu S, et al. B73-Mo17 Near-Isogenic Lines Demonstrate Dispersed Structural Variation in Maize. Plant Physiology. 2011;156(4):1679–1690. doi:10.1104/pp.111.174748.

53. Sun S, Zhou Y, Chen J, Shi J, Zhao H, Zhao H, et al. Extensive intraspecific gene order and gene structural variations between Mo17 and other maize genomes. Nature Genetics. 2018;50(9):1289–1295.

54. Lauter N, Kampani A, Carlson S, Goebel M, Moose SP. microRNA172 down-regulates glossy15 to promote vegetative phase change in maize. Proceedings of the National Academy of Sciences. 2005;102(26):9412–9417.

55. Hirschhorn JN, Daly MJ. Genome-wide association studies for common diseases and complex traits. Nature Reviews Genetics. 2005;6(2):95.

56. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nature genetics. 2006;38(8):904.

57. Buckler ES, Holland JB, Bradbury PJ, Acharya CB, Brown PJ, Browne C, et al. The genetic architecture of maize flowering time. Science. 2009;325(5941):714–718.

58. Kover PX, Valdar W, Trakalo J, Scarcelli N, Ehrenreich IM, Purugganan MD, et al. A multiparent advanced generation inter-cross to fine-map quantitative traits in *Arabidopsis thaliana*. PLoS genetics. 2009;5(7):e1000551.

59. Dell'Acqua M, Gatti DM, Pea G, Cattonaro F, Coppens F, Magris G, et al. Genetic properties of the MAGIC maize population: a new platform for high definition QTL mapping in Zea mays. Genome biology. 2015;16(1):167.

60. Hall M, Lowry D, Willis J. Is local adaptation in Mimulus guttatus caused by trade-offs at individual loci? Molecular Ecology. 2010;19(13):2739–2753.

61. Twyford AD, Friedman J. Adaptive divergence in the monkey flower *Mimulus guttatus* is maintained by a chromosomal inversion. Evolution. 2015;69(6):1476–1486.

62. Gould BA, Chen Y, Lowry DB. Pooled ecotype sequencing reveals candidate genetic mechanisms for adaptive differentiation and reproductive isolation. Molecular ecology. 2017;26(1):163–177.

63. Faria R, Johannesson K, Butlin RK, Westram AM. Evolving inversions. Trends in ecology & evolution. 2019;.

64. Jay P, Whibley A, Frézal L, de Cara MÁR, Nowell RW, Mallet J, et al. Supergene evolution triggered by the introgression of a chromosomal inversion. Current Biology. 2018;28(11):1839–1845.

65. Schmidt C, Pacher M, Puchta H. Efficient induction of heritable inversions in plant genomes using the CRISPR/Cas system. The Plant Journal. 2019;98(4):577–589.