Natural variation in teosinte at the domestication locus *teosinte branched1* (*tb1*)

Laura Vann[1], Thomas Kono[1,2], Tanja Pyhäjärvi[1,3], Matthew B. Hufford[*1,4], and Jeffrey Ross-Ibarra[†1,5]

[1]Department of Plant Sciences, University of California Davis

[2]Department of Agronomy and Plant Genetics, University of Minnesota Twin Cities

[3]Department of Biology, University of Oulu

[4]Department of Ecology, Evolution, and Organismal Biology, Iowa State University

[5]Center for Population Biology and Genome Center, University of California Davis

May 15, 2014

[*]mhufford@iastate.edu

[†]rossibarra@ucdavis.edu

Abstract

The *teosinte branched1* (*tb1*) gene, a repressor of lateral organ growth, is a major QTL involved in branching differences between maize and its wild progenitor, teosinte. Further studies have shown that the insertion of a transposable element (Hopscotch) upstream of *tb1* enhances its expression, causing the reduction in branching observed in domesticated maize. Observations of the maize *tb1* allele in teosinte individuals, coupled with estimates of the age of insertion of the Hopscotch element, led us to investigate the role of *tb1* in teosinte. Results from genotyping across many natural populations suggest that the Hopscotch element is segregating at a higher than expected frequency in a number of populations of both subspecies *Zea mays* ssp. *parviglumis* and subspecies *Zea mays* ssp. *mexicana*. Analysis of linkage disequilibrium between the Hopscotch element and variation in surrounding regions does not support a hypothesis of recent introgression from maize into teosinte, and we find no evidence of environmental correlations that might suggest recent selection. Finally, two greenhouse experiments fail to find an important role for *tb1* in controlling tillering in natural populations of *parviglumis*. Our findings suggest that the role of the Hopscotch in tillering in teosinte is not as straightforward as is in domesticated maize, and that other loci may play a role in observed variation in tillering in teosinte.

Check citation format for AJB. I don't think it's numbers. adjust bibliography style accordingly I would rather send to an open access journal. I don't like that no everyone would be able to have access to paper AJB allows open access for $500 which is cheaper than PLoS ONE

Introduction

Domesticated crops and their wild progenitors provide an excellent system in which to study adaptation and genomic changes associated with human-mediated selection (**?**). Perhaps the central focus of the study of domestication has been the identification of genetic variation underlying agronomically important traits such as fruit size and plant architecture (**?**). Additionally, many domesticates show reduced genetic diversity when compared to their wild progenitors, and an understanding of the distribution of diversity in the wild and its phenotypic effects has become increasingly useful to crop improvement (**?**). But while some effort has been invested into understanding how wild alleles behave in their domesticated relatives (**?**), very little is known about the role that alleles found most commonly in domesticates play in natural populations of their wild progenitors. (**?**).

Maize (*Zea mays* ssp. *mays*) was domesticated from the teosinte *Zea mays* ssp. *parviglumis* (hereafter, *parviglumis*) roughly 9,000 B.P. in southwest Mexico (**??**). Domesticated maize and the teosintes are an attractive system in which to study domestication due to the abundance of genetic tools developed for maize and well-characterized domestication loci (**???**). Additionally, large naturally occurring populations of both *Zea mays* ssp. *parviglumis* (the wild progenitor of maize) and *Zea mays* ssp. *mexicana* (highland teosinte; hereafter *mexicana*) can be found throughout Mexico (**??**), and genetic diversity and of these taxa is estimated to be high (**?**).

Many morphological changes are associated with the domestication of maize, and understanding the genetic basis of these changes has been a focus of maize research for a number of years(**?**). One of the most dramatic changes is found in plant architecture: domesticated maize is characterized by a central stalk with few tillers

and lateral branches terminating in a female inflorescence, while teosinte is highly tillered and bears tassels (male inflorescences) at the end of its lateral branches. The *teosinte branched1* (*tb1*) gene, a repressor of organ growth, was identified as a major QTL involved in branching differences between maize and teosinte (**???**). *an odd assortment of cites. why not the 1997 nature paper?* Further work showed that the insertion of a 4.9 kb retrotransposon (Hopscotch) in the upstream control region of *tb1* led to increased expression of this gene, causing the reduction in branching observed in domesticated maize **?**. The effects of this insertion have been observed in tiller number in maize, but little is known about its role, if any, in teosinte (**?**). Dating of this element has suggested that its insertion predates the domestication of maize, leading to the hypothesis that it was segregating as standing variation in ancient populations of teosinte and increased to high frequency in maize due to selection during domestication (**?**). Furthermore, (**?**) investigated the phenotypic effects of 9 teosinte *tb1* alleles in an isogenic maize background and found that the introgressions sort into three distinct phenotypic classes, suggesting that variation at the *tb1* locus may play a functional role in teosinte.

Tillering, or more specifically lack of tillers in teosinte, may provide an ecological advantage to plants. In species such as teosinte, where plants grow densely packed together, if plants invest more resources into growing taller and fewer resources into tillering, plants may be able to out-compete neighbors for limiting resources such as light. *there must be some citations in ecology lit to support this?* In natural populations of teosinte, variation at the *(*tb1) locus may then play a role in the ecology of teosinte. In this study we aim to characterize the distribution of the Hopscotch insertion in *parviglumis*, *mexicana*, and landrace maize, and to examine the phenotypic effects of the insertion in *parviglumis*. We use a combination of PCR genotyping for the Hopscotch element in our full panel and sequencing of two small regions upstream

of *tb1* in a subset of teosinte populations to explore patterns of genetic variation at this locus. Finally, we test for an association between the Hopscotch element and tillering phenotypes in a population of *parviglumis*.

Methods

Sampling and Genotyping

We sampled 1,110 individuals from 350 accessions (247 maize landraces, 17 *Zea mays* ssp. *mexicana* populations, and 86 *parviglumis* populations) and assessed the presence or absence of the Hopscotch insertion (**??**, **??**). DNA was extracted from leaf tissue using a modified CTAB approach (**??**). We designed primers using PRIMER3 (**?**) implemented in Geneious **?** to amplify the entire Hopscotch element, as well as an internal primer allowing us to simultaneously check for possible PCR bias between presence (5 kb amplification product) and absence (300 bp amplification product) of the Hopscotch insertion. Two PCRs were performed for each individual, one with primers flanking the Hopscotch (HopF/HopR) and one with a flanking primer and an internal primer (HopF/HopIntR). Primer sequences are HopF, 5'-TCGTTGATGCTTTGATGGATGG-3'; Hop R, 5'-AACAGTATGATTTCATGGGACCG-3'; and HopIntR, 5'-CCTCCACCTCTCATGAGATCC-3' (). Homozygotes show a single band for either the Hopscotch element (5 kb) or the absence of the element (300 bp), while heterozygotes are three-banded, showing a band for both the presence and absence of the Hopscotch element as well as a band for the internal primer set (). *fig references still need to say "Fig" etc., all the ref function does is set the number automaticall* . When only one PCR resolved well, we scored one allele for the individual, which explains the odd number of alleles included in our analyses. We used Phusion High Fidelity

Enzyme (Finnzymes, Inc.) and the following conditions for amplifications: 98°C for 3 min, 30 cycles of 98°C for 15 s, 65°C for 30 s, and 72°C for 3 min 30 s, with a final extension of 72°C for 10 min. PCR products were visualized on a 1% agarose gel and scored for presence/absence of the Hopscotch based on band size.

Sequencing

In addition to genotyping, we chose a subset of *parviglumis* individuals for sequencing. We chose twelve individuals from each of four populations from Jalisco state, Mexico (San Lorenzo, La Mesa, Ejutla A, and Ejutla B). For amplification and sequencing, we selected two regions approximately 600bp in size from within the 5' UTR of *tb1* (sequenced region 1) and from 1,235 bp upstream of the start of the Hopscotch and 66,169 bp upstream from the start of the *tb1* ORF (sequenced region 2). The 5' UTR (containing sequenced region 1) has been shown to have elevated diversity in ssp. *parviglumis* with respect to maize, while the the area upstream from the Hopscotch (sequenced region 2) has been shown to be critical in determining basal branching and ear architecture (**??**). *is this true? the area upstream of the hopscotch? those papers just found that something upstream was the causal variant, no?* We designed the following primers using PRIMER3 (**?**): for the 5' UTR, 5' GGATAATGTGCACCAGGTGT 3' and 5' GCGTGCTAGAGACACYTGTTGCT 3'; for the 50 kb upstream region, 5' TGTCCTCGCCGCAACTC 3' and 5' TGTACGCCCGCCCCTCATCA 3' (). We used Taq Polymerase (New England Biolabs) and the following thermal cycler conditions to amplify fragments: 94°C for 3 min, 30 cycles of 92°C for 40 s, annealing for 1 min, 72°C for 40 s, and a final 10 min extension at 72°C. Annealing temperatures for sequenced region 1 and sequenced region 2 were 59.7°C and 58.8°C, respectively. To clean excess primer and dNTPs we added two units of Exonuclease1 and 2.5 units of

Antarctic Phosphatase to 8.0 μ L of amplification product. This mix was placed on a thermal cycler with the following program: 37°C for 30 min, 80°C for 15 min, and a final cool-down step to 4°C.

We cloned cleaned fragments into a TOPO-TA vector (Invitrogen, Carlsbad) using OneShot TOP10 chemically competent *E. coli* cells, with an extended ligation time of 30 min for a complex target fragment. We plated cells on LB agar plates containing kanamycin, and screened colonies using vector primers M13 Forward and M13 Reverse under the following conditions: 96°C for 5 min; then 35 cycles at 96°C for 30 s, 53°C for 30 s, 72°C for two min; and a final extension at 72°C for 4 min. We visualized amplification products for incorporation of our insert on a 1% agarose TAE gel.

Amplification products with successful incorporation of our insert were cleaned using Exonuclease 1 and Antarctic Phosphatase following the procedures detailed above, and sequenced with vector primers M13 Forward and M13 Reverse using Sanger sequencing at the College of Agriculture and Environmental Sciences (CAES) sequencing center at UC Davis. We aligned and trimmed primer sequences from resulting sequences using the software Geneious (**?**). Following alignment, we verified singleton SNPs by sequencing an additional one to four colonies from each clone. If the singleton was not present in these additional sequences it was considered an amplification or cloning error, and we replaced the base with the base of the additional sequences. If the singleton appeared in at least one of the additional sequences we considered it a real variant and kept it for further analyses.

Genotyping Analysis

please drop scripts for bayenv and STRUCTURE into github which scripts are you talking about? I used matt's for making structure chunks, and then just command line to run structure, same with BayEnv – it was Tanjas covariance matrix and then just the basic bayenv commands matt's scritp for structure should be up there. and even a readme documenting the commandline for bayenv and including the covariance matrix. basically whatever someone would need to be able to literally repeat what you did.

We examined discrepancies between observed and expected genotype frequencies by calculating Hardy-Weinberg Equilibrium (HWE). To calculate differentiation between populations (F_{ST}) and subspecies (F_{CT}) we used HierFstat (**?**). These analyses only included populations in which 8 or more individuals were sampled. To test the hypothesis that the Hopscotch insertion may be adaptive under certain environmental conditions, we looked for significant associations between the Hopscotch frequency and environmental variables using BayEnv (**?**). BayEnv creates a covariance matrix of relatedness between populations, and then tests a null model that allele frequencies in populations are determined by the covariance matrix of relatedness alone against the alternative model that allele frequencies are determined by a combination of the covariance matrix and an environmental variable, producing a posterior probability (Bayes Factor)(**?**). We used genotyping and covariance data from **?** for BayEnv, with the Hopscotch insertion coded as an additional SNP (). Environmental data were obtained from `www.worldclim.org`, the Harmonized World Soil Database and `www.harvestchoice.org`, and summarized by principle component analysis (**?**).

Sequence Analysis

For population genetic analyses of sequenced region 1 and sequenced region 2 we used the analysis package of Libsequence (**?**) to calculate pairwise F_{ST} between popula-

tions, and to calculate standard diversity statistics (number of haplotypes; haplotype diversity; Watterson's estimator $\hat{\theta}_W$; pairwise nucleotide diversity $\hat{\theta}_\pi$; and Tajima's D). To produce a visual representation of differentiation between sequences and to examine patterns in sequence clustering by Hopscotch genotype we used Phylip (`http://evolution.genetics.washington.edu/phylip.html`) to create neighbor-joining trees with bootstrapping (100 repetitions) to examine the support of nodes in our trees. For creation of trees we also included homologous sequence data from teosinte inbred lines (TILs), some of which are known to be homozygous for the Hopscotch insertion (TIL03, TIL17, TIL09), as well as 59 lines of domesticated maize and landraces (data from Maize HapMapV2, (**?**)).

Introgression Analysis

In order to assess patterns of linkage disequilibrium (LD) around the Hopscotch element in the context of chromosomal patterns of LD we used Tassel (**?**) and calculated LD between SNPs across chromosome 1 using previously published data from twelve plants each of the Ejutla A (EjuA), Ejutla B (EjuB), San Lorenzo (SLO), and La Mesa (MSA) populations (**?**). We chose these populations because we had both genotyping data for the Hopscotch as well as chromosome-wide SNP data for chromosome 1. For each population we filtered the initial set of 5,897 SNPs on chromosome 1 to accept only SNPs with a minor allele frequency of at least 0.1, resulting in 1,671, 3,023, 3,122, and 2,167 SNPs for SLO, EjuB, EjuA, and MSA, respectively. We then used Tassel (**?**) to calculate linkage disequilibrium (r^2) across chromosome 1 for each population.

We examined evidence of introgression on chromosome 1 in these same four populations (EjuA, EjuB, MSA, SLO) using STRUCTURE (**?**) and the same phased 55K

SNP data from (**?**) that we used for LD analysis, combined with the corresponding SNP data from a diverse panel of 282 maize lines (**?**). SNPs were anchored in a modified version of the IBM genetic map ((**?**), `http://arxiv.org/abs/1307.7313`). We created haplotype blocks using a custom Perl script that grouped SNPs separated by less than 5kb into haplotypes. We ran STRUCTURE at K=2 under the linkage model, performing 3 replicates with an MCMC burn-in of 10,000 steps and 50,000 steps post burn-in. *i'd like this perl script on github, maybe in this repo or as a gist. also structure input file too. all the stuff we'd need to redo this. See above note..not all of this or the BayEnv was script'ified sure, even command line info should be included where possible. idea is to maximize reproducibility – either other people or subsequent students. for example, matt has a student who wants to work on tb1 in natural pops, and she might want to try/redo some of these analyses with the same or new data. okay so I should just put my command line stuff with good commenting as to what is what in repository?*

Phenotyping of *Zea mays.* ssp. *parviglumis*

To investigate the phenotypic effects of the Hopscotch insertion in teosinte, we conducted an initial phenotyping trial (Phenotyping 1). We germinated 250 seeds of *parviglumis* collected in Jalisco state, Mexico (population San Lorenzo) (**?**) where the Hopscotch is segregating at highest frequency (0.44) in our initial genotyping sample set. In order to maximize the likelihood of finding the Hopscotch in our association population we selected seeds from sites where genotyped individuals were homozygous or heterozygous for the insertion. We chose between 10-13 seeds from each of 23 sampling sites. We treated seeds with fungicide and germinated them in petri dishes with filter paper. Following germination, 206 successful germinations were then planted into one gallon size pots with potting soil and randomly spaced one foot apart on greenhouse benches. Plants were watered three times a day with

an automatic drip containing 10-20-10 fertilizer. *it ended up being a combination of drip and hand watering because they dried out so much and did better when they had water on the leaves as well as in the soil*

To investigate the phenotypic effects of the Hopscotch insertion in teosinte, we conducted an initial phenotyping trial (Phenotyping 1). We germinated 250 seeds of *parviglumis* collected in Jalisco state, Mexico (population San Lorenzo) (**?**) where the Hopscotch is segregating at highest frequency (0.44) in our initial genotyping sample set. In order to maximize the likelihood of finding the Hopscotch in our association population we selected seeds from sites where genotyped individuals were homozygous or heterozygous for the insertion. We chose between 10-13 seeds from each of 23 sampling sites. We treated seeds with fungicide and germinated them in petri dishes with filter paper. Successful germinations (206 individuals) were then planted into one gallon size pots with potting soil and randomly spaced one foot apart on greenhouse benches. Plants were watered three times a day.

Starting on day 15, we measured tillering index, the ratio of the sum of tiller lengths to the height of the plant (**?**). Tillering index has been shown to be the most effective way to observe the phenotypic effects of the Hopscotch insertion on plant architecture in maize (**?**). Following initial measurements, we phenotyped plants for tillering index every 5 days through day 40, and then on day 50 and day 60. On day 65 we measured culm diameter between the third and fourth nodes of each plant. Culm diameter is not believed to be correlated with tillering index, or variation at *tb1* (e.g. Hopscotch genotype). Following phenotyping we extracted DNA from all plants using a modified SDS extraction protocol (http://www.ars.usda.gov). We genotyped individuals for the Hopscotch insertion following the protocols listed above. Based on these initial data, we conducted a post hoc power analysis using data from day 40 of phenotyping 1, indicating that a minimum of 71 individuals in each genotype class are needed to detect the observed effect of the Hopscotch on tillering index. *do*

you still have these posthoc calculations? I believe I do, otherwise they are likely in a lab meeting slide on Dropbox – do you want them in here? would be good to include in a github, again so we could go back and reassess how we do things in case, for example, we decide to give the greenhouse experiment a 3rd try (yes, i'm a masochist) I would do it a 3rd time, I'm convinced something went wrong....third time with expression analyses :)

We performed a second phenotyping experiment (phenotyping 2) in which we germinated 372 seeds of *parviglumis*, choosing equally between sites previously determined to have or not have the Hopscotch insertion. Seeds were germinated and planted on day 7 post fruit-case removal into 2 gallon pots. Plants were watered twice daily, alternating between fertilized and non-fertilized water. We began phenotyping successful germinations (302) for tillering index on day 15 post fruit case removal, and phenotyped every five days until day 50. At day 50 we measured culm diameter between the third and fourth nodes. We extracted DNA and genotyped plants following the same guidelines as in phenotyping 1.

Resulting tillering index data for each genotype class did not meet the criteria for a repeated measures ANOVA, so we transformed the data using a Box-Cox transformation ($\alpha = 0$) implemented in the car package in R (**?**) to improve the normality and homogeneity of variance among genotype classes. We analyzed relationships between genotype and tillering index and tiller number using a repeated measures ANOVA through a general linear model function implemented in SAS v.9.3 (SAS Institute Inc., Cary, NC, USA). Additionally, in order to compare any association between Hopscotch genotype and tillering and associations at other presumably unrelated traits, we performed an ANOVA between culm diameter and genotype using the same general linear model in SAS. *please add SAS scripts/code to a gist or something.*

Results

Genotyping

Genotype of the Hopscotch insertion was confirmed with two PCRs for 837 individuals. Among the 247 maize landrace accessions genotyped, all but 8 were homozygous for the presence of the insertion (**??**, **??**). *please fix table/figure references to say table/figure* Within our *parviglumis* and *mexicana* samples we found the Hopscotch insertion segregating in 37 and 4 populations *this is confusing as the map shows ¡37 populations. we should be consistent about reporting for all pops, or all pops with n¿8, etc. or at least be explicit what cutoff we are using for each result* , respectively, and at highest frequency in the states of Jalisco, Colima, and Michoacán in central-western Mexico in both subspecies (1) *the map only shows parviglumis. can we add mexicana? the text makes it seem as if both should be on the map* . We examined Hardy-Weinberg equilibrium in a total of 14 populations (10 *parviglumis* and 4 *mexicana*) with more than 8 individuals sampled per population. Three populations (RIMPA0073, RIMPA0093, and RIMPA0158) show evidence of deviations from expected genotype frequencies under the assumptions of HWE (p¡0.05). *in what direction? too many hets? what's the F? they weren't all in the same direction, do you still want me to list out? no need to list all, but if there were sig. deviations in multiple directions in different pops, that is worth saying. maybe could list them all out in supp. table?*

please fix whitespace and black border on figure

Using our Hopscotch genotyping data, we calculated differentiation between populations (F_{ST}) and subspecies (F_{CT}) for populations in which we sampled 8 or more alleles. F_{CT} is 0 within our dataset, and we found similar levels of F_{ST} among populations within each subspecies (0.22) *is this an average? this comes from libsequence?* and among all populations (0.23), to those reported in genome-wide estimates from previous

studies **?** (1).

Table 1: Pairwise F_{CT} values from sequence and Hopscotch genotyping data

Comparison	**Seq. Region 2**	**Seq. Region 1**	**Hopscotch**
EjuA & EJuB	0	0	0
EjuA & MSA	0.328	0.326	0.186
EjuA & SLO	0.258	0.416	0.28
EjuB & MSA	0.365	0.397	0.188
EjuB & SLO	0.29	0.512	0.28
MSA & SLO	0	0.007	0.016

Although we found large variation in Hopscotch allele frequency among our populations, BayEnv analysis did not indicate a correlation between the Hopscotch insertion and environmental variables (all Bayes Factors < 1;).

Sequencing

To investigate patterns of sequence diversity and linkage disequilibrium (LD) in the *tb1* region, we sequenced two small (<1kb) regions upstream of the tb1 ORF in four populations. After alignment and singleton checking we recovered 40 and 48 segregating sites for the 50kb upstream region and the 5' UTR region, respectively. For region 1, Ejutla A has the highest values of haplotype diversity, and θ_π, while Ejutla B and La Mesa have comparable values of these summary statistics, and San Lorenzo has much lower values. Additionally, Tajima's D is strongly negative in the two Ejutla populations and La Mesa, but is more positive in San Lorenzo (2). *can drop theta W from table and text. we show pi and D, which is sufficient. fix caption in table. also add Hopscotch allele frequencies to table or list somewhere – i think it's useful for comparison of seq stats the frequencies are*

in supplemental table 1, do you want them put in this table too, or just me to refer people to supp table 1? yeah refer to table is fine

Table 2: Add caption

Population	# Haplotypes	Hap. Diversity	$\hat{\theta}_W$	$\hat{\theta}_\pi$	Tajima's D
Seq. region 2 (50kb upstream)					
EJUA	8	0.89394	0.01548	0.01763	0.6231
EJUB	8	0.89394	0.01493	0.01591	0.29504
MSA	3	0.68182	0.01111	0.01055	-0.22212
SLO	4	0.74242	0.01167	0.01413	0.93185
Seq. region 1(5? UTR)					
EJUA	8	0.85897	0.00874	0.00527	-1.64955
EJUB	5	0.70909	0.00663	0.00378	-1.83123
MSA	6	0.68182	0.00646	0.00373	-1.75506
SLO	3	0.31818	0.00176	0.00137	-0.72873

For region 2, haplotype diversity, θ_W, and θ_π, are similar for Ejutla A and Ejutla B, while La Mesa and San Lorenzo have slightly lower values for these statistics (2). Tajima's D is positive in all populations except San Lorenzo, indicating an excess of low frequency variants in this population (2). Pairwise values of F_{ST} within population pairs Ejutla A/Ejutla B and San Lorenzo/La Mesa are 0 for both sequenced regions as well as for the Hopscotch, while they are high for other population pairs (1). Neighbor joining trees of our sequence data and data from the teosinte inbred lines (TILs; data from Maize HapMapV2, (**?**)) do not reveal any clear clustering pattern with respect to population or Hopscotch genotype (**??**); individuals within our sample that have the Hopscotch insertion do not group with the teosinte inbred lines or the lines of domesticated maize that have the Hopscotch insertion.

Evidence of introgression

The teosinte populations *which?* with the highest frequency of the Hopscotch insertion in this study were sympatric with cultivated maize. Our initial hypothesis was that the high frequency of the Hopscotch element in these populations could be attributed to introgression from maize into teosinte. To investigate this possibility we examined overall patterns of linkage disequilibrium across chromosome one, and specifically in the *tb1* region. If the Hopscotch is found in these populations due to recent introgression we would expect to find large blocks of linked markers near this element. We find no evidence of elevated linkage disequilibrium between the Hopscotch and SNPs surrounding the *tb1* region in our resequenced populations (2), and r^2 in the *tb1* region does not differ significantly between populations with (average r^2 of 0.085) and without the Hopscotch genotype (average $r^2 = 0.082$). In fact, average r^2 is lower in the *tb1* region ($r^2 = 0.056$) than across the rest of chromosome 1 ($r^2 = 0.083$) (3).

table is too wide, need to round numbers, and column headers are messed up.

Table 3: r^2 values between SNPs in the *tb1* region (positions 264,596,664-265,891,456 on chromosome 1 of the maize AGPv2 genome) and the rest of chromosome 1, within the 5' UTR (Sequenced region 1), and within the 66,169 bp upstream region (Sequenced region 2).

Population	Chromosome 1	tb1 region	Seq. region 1	Seq. region 1
Ejutla A	0.095426101	0.050304	0.747295	0.214933
Ejutla B	0.068681837	0.051295	0.660354	0.186395
La Mesa	0.069500533	0.053306	0.914286	0.766234
San Lorenzo	0.100536784	0.067251	0.912281	0.636364

The lack of clustering of Hopscotch genotypes in our NJ tree as well as the lack of LD around *tb1* does not support the hypothesis that the Hopscotch insertion in these populations of *parviglumis* is the result of recent introgression. However, to further explore this hypothesis we performed a STRUCTURE analysis using Illumina MaizeSNP50 data from four of our *parviglumis* populations (EjuA, EjuB, MSA, and SLO) and the maize 282 diversity panel (**??**). The linkage model implemented in STRUCTURE can be used to identify ancestry of blocks of linked variants, which would arise as a result of recent admixture between populations. If the Hopscotch insertion is present in populations of *parviglumis* as a result of recent admixture with domesticated maize, we would expect the insertion and linked variants in surrounding sites to be assigned to the "maize" cluster in our STRUCTURE runs, not the "teosinte" cluster. In all runs, assignment to maize in the *tb1* region across all four *parviglumis* populations is low (average 0.017) *is this really 0.017 or 0.17? Yes really 0.017 assignment to maize in the tb1 region, and avg assignment across chr1 is 0.2 I also have a table of assignment values for SLO individuals based on genotype. Though we had decided this wasn't super informative because sample size was low* and much below the chromosome-wide average (0.20; 3).

please put figures in the text rather than at the end. I can't figure out why the figures are going at the end. I did them following the format in Sofiane's 282 paper

Phenotyping

Phenotyping

To assess the contribution of *tb1* to phenotypic variation in tillering in a natural population, we grew plants from seed sampled from the San Lorenzo population of *parviglumis*, which had a high mean frequency (0.44) of the Hopscotch insertion from our initial genotyping. We measured tillering index (TI), the ratio of the sum of tiller

lengths to plant height, for 216 plants from within the San Lorenzo population, and genotyped plants for the Hopscotch insertion. We found the Hopscotch segregating at a frequency of 0.65 with no significant deviations from expected frequencies under Hardy-Weinberg equilibrium. After performing a repeated measures ANOVA between our transformed tillering index data and Hopscotch genotype we find a weak positive correlation between presence of the Hopscotch and tillering index on day 40 (p=0.0848), but no correlation between tillering index and genotype on any other day (4). Additionally we find no significant correlation between tiller number and Hopscotch genotype, or culm diameter and Hopscotch genotype in phenotyping 1. *shouldn't we expect a negative correlation betweeh Hop and TI on day 40? need to have an A and B in the figure and explain one is for pheno1 and one is for pheno2. please explain whiskers and dots on figure too. sure, I mean, presumably we would expect things with Hop to have a smaller TI yup, but we should mention that the expectation is negative/*

lots of white space in fig 4 and fig. s1 too. We performed a second grow-out of teosinte to assess whether lighting conditions or sample size may have affected our ability to detect and effect of *tb1*. For the second grow-out we measured tillering index every five days through day 50 for 302 plants. We found the Hopscotch allele segregating at a frequency of 0.69, *is it in HWE in this pop?* with a 0.6 frequency of Hopscotch homozygotes, and a 0.2 frequency of both heterozygotes and homozygotes for the teosinte allele. We found similar patterns, with a weak positive correlation between tillering index and Hopscotch genotype at day 40 (p¡0.0611), with no significant correlation on any day. Similarly, relationships between Hopscotch genotype and tiller number, and hopscotch genotype and culm diameter are not significant.

Discussion

Adaptation occurs either due to selection on standing variation or on *de novo* mutations. Adaptation as a result of selection on standing variation has been well-described in a number of systems, for example, selection for lactose tolerance in humans (**??**); variation at the *Eda* locus in three-spined stickleback (**??**); and pupal diapause in the Apple Maggot fly (**?**). Although the role of standing variation with respect to adaptation has been described in many systems, its importance to domestication is not as well studied.

In maize, alleles at important domestication loci (*RAMOSA1*, (**?**); *barren stalk1*, (**?**); and *grassy tillers1*, (**?**)) have been shown to have been selected from standing variation, suggesting that diversity already present in teosinte may have played an important role in the domestication of maize. The *teosinte branched1* gene has long been a central focus of research concerning maize domestication, and, while previous studies have suggested that differences in plant architecture between domesticated maize and teosinte are a result of selection on standing variation, little is known about variation at this locus in teosinte (**??**). **?** genotyped 90 accessions of teosinte (inbred and outbred), providing the first evidence that the Hopscotch insertion is segregating in teosinte (**?**).

Given that the Hopscotch insertion has been estimated to predate the domestication of maize, it is not surprising that it can be found segregating in populations of teosinte. However, in sampling numerous individuals from many teosinte populations our study provides greater insight into the distribution and prevalence of the Hopscotch in teosinte. While our findings are consistent with a previous study by **?** in that we identified the Hopscotch allele segregating in teosinte, we find it at higher frequency than previously suggested (**?**). Many of our populations with high

frequency of the Hopscotch allele fall in the Jalisco cluster identified by **?**, possibly suggesting a different history of the *tb1* locus than in the Balsas region where maize was domesticated (**?**). While gene flow from crops into their wild relatives is well-known, ((**????????**)), our results are more consistent with **?** who found resistance to introgression from maize into teosinte (**?**). Furthermore, **?** showed that domestication loci, such as *tb1*, are particularly resistant to introgression in both directions of gene flow (i.e., maize to teosinte and teosinte to maize) (**?**).

We find no evidence of recent introgression in our analyses. Clustering patterns in our NJ trees do not reflect a pattern expected if maize alleles at the *tb1* locus had introgressed into populations of teosinte. Moreover, analysis of linkage in the *tb1* region does not reveal patterns of high LD relative to the rest of chromosome 1, and assignment to maize in this region in our STRUCTURE analysis is lower than the average across chromosome 1 (3, 4). Together, these data point to an explanation other than recent introgression for the high observed frequency of Hopscotch in some of our *parviglumis* populations.

Table 4: Assignments to maize and teosinte in the *tb1* and chromosome 1 regions from STRUCTURE

	tb1 region		Chr 1	
Population	Maize assignment	Teosinte assignment	Maize assignment	Teosinte assignment
Ejutla A	0.02158681	0.9784132	0.2026814	0.7973186
Ejutla B	0.01888194	0.9811181	0.1872131	0.8127869
La Mesa	0.0118675	0.9881333	0.8068998	0.1931017
San Lorenzo	0.01551389	0.9844861	0.2048252	0.7951748

Although recent introgression seems unlikely, we cannot rule out ancient introgression as an explanation for the presence of the Hopscotch in these populations.

If the Hopscotch allele was introgressed in the distant past, they could have been sufficient recombination to break up any initial LD, leading to observations similar to those obtained here. We find this scenario less plausible, however, as there is no reason why gene flow should have been high in the past but absent in present-day sympatric populations. In fact, early generation maize-teosinte hybrids are easy to find in these populations today (MB Hufford, pers. observation), and genetic data support ongoing gene flow between domesticated maize and both *Zea mays* ssp. *mexicana* and *Zea mays* ssp. *parviglumis* in a number of sympatric populations (**???**).

Other explanation for differential frequencies of the Hopscotch among teosinte populations include both drift and natural selection. Previous studies using both SSRs and genome-wide SNP data have found evidence for a population bottleneck in the San Lorenzo population (**??**), and the lower levels of sequence diversity in the 5' UTR region and the more positive values of Tajima's D we present here are consistent with these findings. *deviations from HWE may be consistent too if we see excess of homozygotes. do we?* . This bottleneck, however, does not explain differences in Hopscotch allele frequency among populations, and the available information on diversity and population structure among these populations (**??**) is not suggestive of colonization or other demographic events that might predict a high frequency of the allele in multiple populations. *here we need a few sentences on selection. the 5' UTR has much more negative D than the upstream. do we know the Hop genotype for sequenced lines? can we separate the sequences into hop/no hop and look for differences? it wasn't until we did this that gt1 stuff really popped out. we should know for some of them, i will check*

The phenotypic effects of the Hopscotch insertion in domesticated maize have been well documented (**??**), and **?** have described its effects in partially inbred lines of teosinte (**?**) *i don't think these were inbred.please doublecheck.* . Our study is the first to

explicitly examine the phenotypic effects of the Hopscotch insertion in individuals sampled from a natural population of teosinte. *isn't this what weber did?? for 70+ populations!?* However, we found no significant effect of the Hopscotch on tillering index or tiller number in our phenotyping experiments, and the effect of the Hopscotch insertion in teosinte is discordant with that of maize. The lack of correlation between Hopscotch genotype and tillering index or tiller number is surprising given its effects in maize. It is certainly possible that even though previous data demonstrate an effect of the Hopscotch on tillering in maize (**?**), that the effect of the Hopscotch in teosinte is more complicated and may be more difficult to observe. Moreover, *tb1* is a single gene in a complex pathway that affects branching and tillering traits, and perhaps in combination with alleles at other loci the phenotypic effects of the Hopscotch on tillering may not be consistent. *thsi section still needs work. i think we have to do more here. weber shows an association between SNPs in tb1 and branch length. we need to discuss that!*

MBH todo Variation at *tb1* has also been shown to contribute to phenotypes other than tillering (**?**), and a recent study by **?** examined the possibility of an allelic series at the *tb1* locus in teosinte. **?** introgressed 9 separate teosinte segments (one from *Zea diploperennis*, and four from both *Zea mays* ssp. *mexicana* and *Zea mays* ssp. *parviglumis*) spanning the *tb1* locus into an isogenic maize background and investigated their effects on previously associated phenotypes. They found that plants with teosinte chromosomal segments had greater tillering than their maize isogenic lines, and that different chromosomal segments of *tb1* confer different amounts of tillering, suggesting that there are multiple genetic factors in this region that affect tillering. However, in addition to elucidating variance in tillering among *tb1* teosinte segments, **?** found significant variance among W22 control lines, suggesting that there are other genetic factors aside from alleles at the *tb1* locus that affect tillering in maize. **??** first attempted to map QTL controlling many of the phenotypic differences between

domesticated maize and teosinte, and demonstrated the existence of numerous QTL that contribute to the differences in branching architecture between the two. Many of these loci (*grassy tillers*, *gt1*; *tassel-replaces-upper-ears1*, *tru1*; *terminal ear1*, *ter1*) have been shown to interact with *tb1* (**??**), and both *tru1* and *ter1* have been shown to affect the same phenotypic traits as *tb1* (**?**). *tassel-replaces-upper-ears1* (*tru1*), for example, has been shown to act either epistatically or downstream of *tb1*, affecting both branching architecture (decreased apical dominance) and tassel phenotypes (shortened tassel and shank length and reduced tassel number) (**?**). It seems plausible that variation in some of these other loci could have affected tillering in our greenhouse population, and contributed to the lack of correlation we see between Hopscotch genotype and tillering.

In summary, our findings demonstrate that the Hopscotch allele is more widespread in populations of *parviglumis* and *mexicana* than previously thought. Analysis of linkage using SNPs from across chromosome 1 does not suggest that the Hopscotch allele is present in these populations due to recent introgression; however, it seems unlikely that it would have drifted to high frequency in multiple populations and there may be another explanation for the high frequency we observe in some of our populations. The Hopscotch does not appear to have a strong effect reducing tillering in teosinte as it does in maize, and other loci involved in branching architecture may play also play roles in the regulation of tillering in teosinte. Finally, although we see no clear evidence of recent strong selection, the high frequency of the Hopscotch insertion in a number of populations continues to suggest to us that it plays an ecological role in teosinte. In the future, additional experiments will be needed to examine expression levels of *tb1* and additional loci involved in branching architecture (e.g. *gt1*, *tru1*, and *ter1*) in conjunction with a more exhaustive phenotyping and genotyping assay. *why not Phyb and phya?* *Are they necessary to include? I'd had them in before*

in a paragraph but had been voted out I'd ditch gt1 tru1 ter1 and maybe just cite some people including phyb etc.

please check format of supp figs and tables; some are running off the page. you can use "longtable" to fix that (ask Paul for example). check fig/table references, bibliography, etc. what does "rotation" mean in supp. table 3? it isn't mentioned in methods. please check that all the tables and figs (including supplement) are referenced in the text.

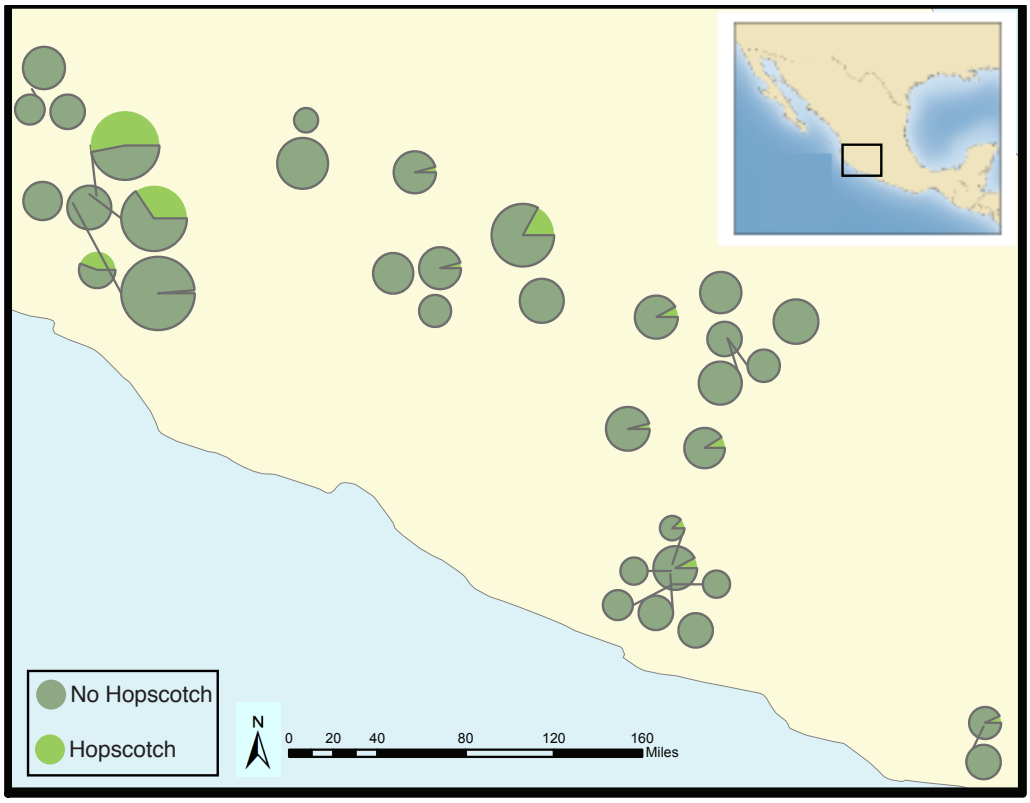


Figure 1: Map showing the frequency of the Hopscotch allele in populations of *Zea mays* ssp. *parviglumis* where we sampled more than 6 individuals. Size of circles reflects number of alleles sampled.

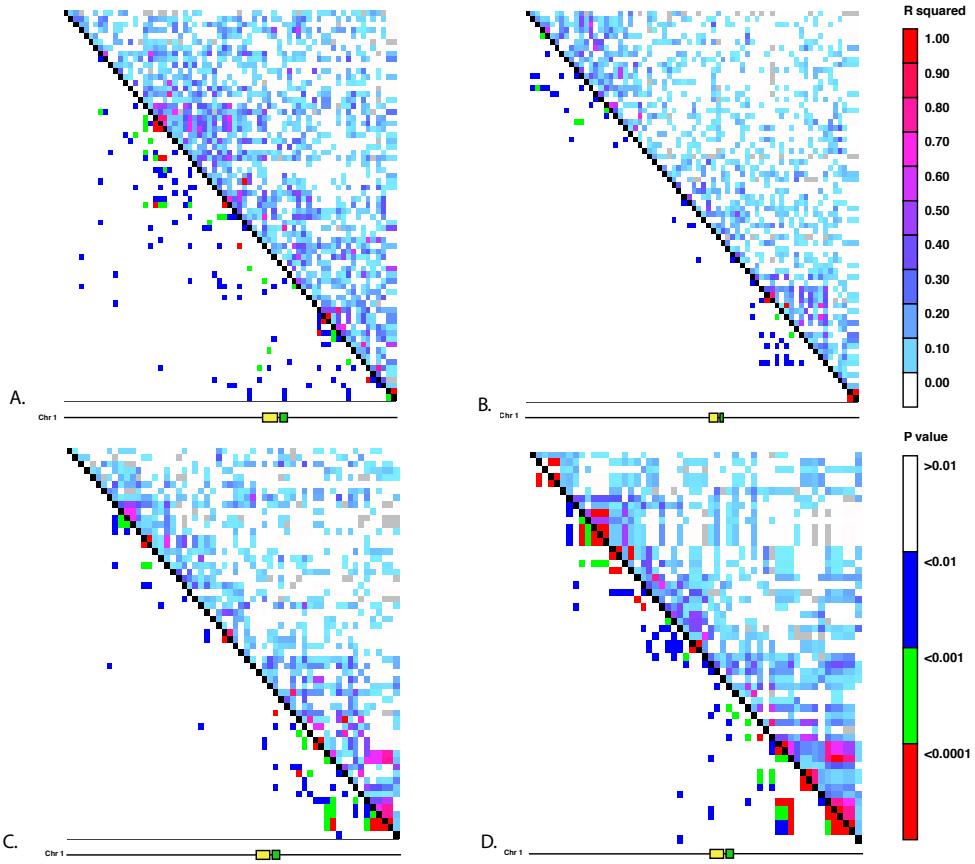


Figure 2: Linkage disequilibrium for SNPs in Mb 261-268 on chromosome 1. The yellow rectangle indicates the location of the Hopscotch insertion and the green represents the *tb1* ORF. A) Ejutla A; B) Ejutla B; C) La Mesa; D). San Lorenzo

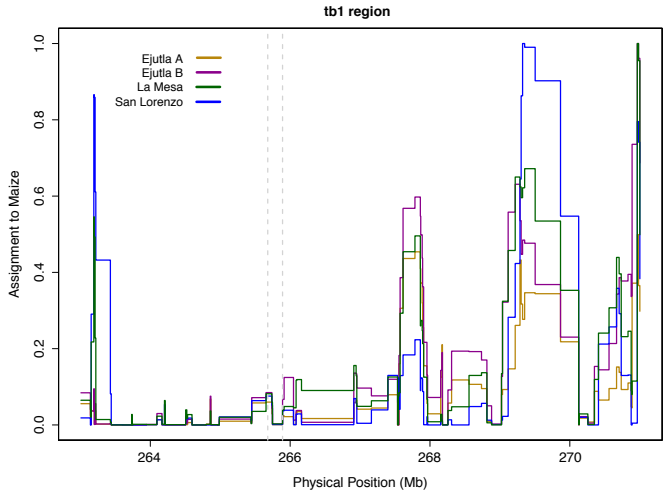


Figure 3: STRUCTURE assignment to maize across a section of chromosome 1. The dotted lines mark the beginning of the sequenced region 50kb upstream (Sequenced region 2) and the end of the *tb1* ORF.

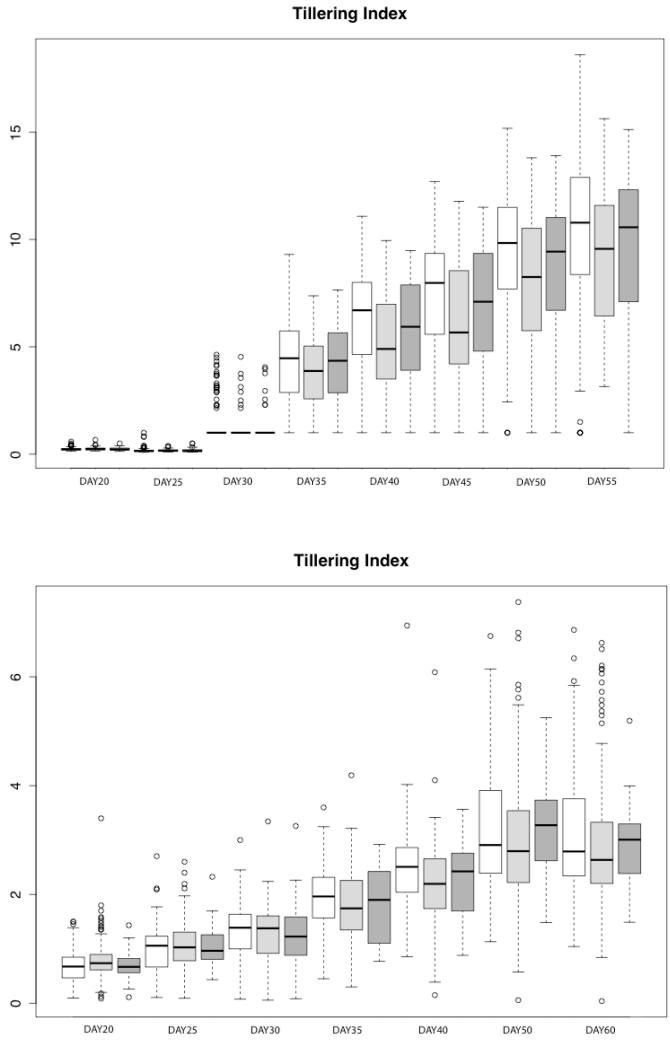
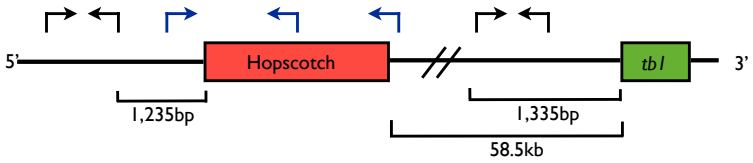
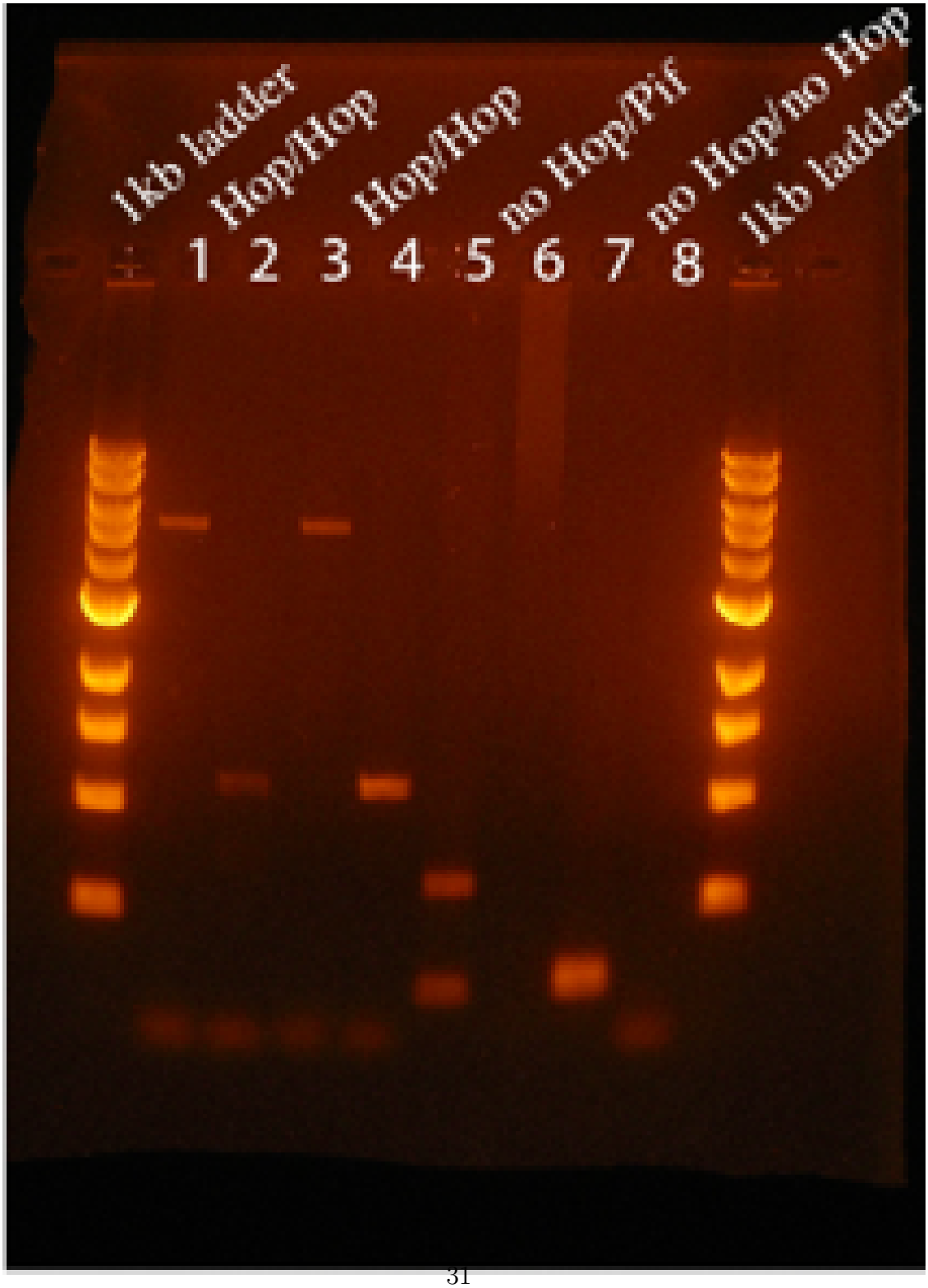


Figure 4: Box-plot showing tillering index in our greenhouse population from day 20-60. White indicates individuals homozygous for the Hopscotch, light grey represents heterozygotes, and dark grey represents homozygotes for the teosinte (No Hopscotch) allele. Within boxes, dark black lines represent the median, and the edges of the boxes are the first and third quartiles.

Supplementary Materials


5'
Hopscotch
tbl
3'
1,235bp
1,335bp
58.5kb


1kb ladder
Hop/Hop
Hop/Hop
no Hop/Pif
no Hop/no Hop
1kb ladder
1 2 3 4 5 6 7 8


EjuB41
EjuA1
EjuB44
EjuA13
EjuA18
TIL08
TIL04
TIL16
EjuB40
Maize reference
SLO23
MSA50
MSA52
EjuB6
EjuB7
EjuB9
EjuB16
EjuA16
EjuA23
EjuA29
EjuA31
EjuA36
SLO1
SLO47
EjuB21
EjuB22
EjuB23
EjuB24
EjuB29
EjuA28
EjuA41
EjuA47
EjuA52
TIL15
MS71
OH43
BKN022
KI3
TIL09
TZI8
W64A
CML277
CML247
M37W
TDD39103
CAU5003
BKN019
HP301
NC358
TX303
W22
CML333
BKN014
P39
BKN011
NC350
TIL10
CAU478
TIL07
TIL05
CAUCHANG72
BKN032
BKN034
BKN009
BKN029
TIL03
IL14H
CML322
TIL06
BKN026
TIL11
CAUZHENG58
BKN030
BKN040
BKN016
CML228
MO17
BKN018
BKN031
B73
BKN025
BKN020
TIL17
KI11
BKN015
CML103
OH7B
BKN027
BKN033
KY21
ANC
TIL14
CAUMO17
BKN035
BKN023
BKN017
MO18W
CML69
M162W
CAU178
BKN010
CML52
B97
TIL01
TIL12
TIL25
MSA53
MSA43
MSA41
MSA35
MSA10
MSA1
SLO48
SLO37
SLO30
SLO7
MSA39
MSA16
MSA12
MSA2
SLO50
SLO49
SLO21
SLO6
SLO20
EjuB9
EjuA28
CML322
TIL04
Maize reference
EjuB7
EjuB16
EjuB40
EjuA16
EjuB41
EjuA23
EjuA29
EjuA36
MSA35
EjuB21
EjuB22
EjuB23
EjuB44
EjuA41
EjuA47
EjuA52
TIL15
MS71
OH43
BKN022
KI3
TIL09
TZI8
W64A
CML277
CML247
M37W
TDD39103
CAU5003
BKN019
NC358
TX303
W22
CML333
BKN014
P39
BKN011
NC350
TIL10
CAU478
TIL05
CAUCHANG72
BKN032
BKN034
BKN009
BKN029
TIL03
IL14H
TIL08
TIL06
BKN026
TIL11
TIL01
CAUZHENG58
BKN030
BKN040
BKN016
CML228
MO17
BKN018
BKN031
B73
BKN025
BKN020
TIL17
TIL16
KI11
BKN015
CML103
TIL12
OH7B
BKN027
BKN033
KY21
ANC
TIL14
CAUMO17
BKN035
BKN023
TIL25
BKN017
MO18W
CML69
CAU178
M162W
BKN010
CML52
B97
SLO37
SLO47
EjuA18
EjuB24
EjuB6
EjuA3
HP301
EjuA31
MSA43
EjuB29
EjuA13
TIL07
SLO30
SLO48
MSA1
MSA10
MSA41
SLO7
SLO1
MSA52
MSA50
MSA39
MSA16
MSA12
MSA2
SLO50
SLO49
SLO23
SLO21
SLO20
SLO6
MSA53

Accession	USDA Accession ID	Locality	Number alleles sampled	Hop
RIHY0009	N/A	N/A	2	
RIMME0006	566673	Durango, Mexico	2	
RIMME0007	566680	Guanajuato, Mexico	2	
RIMME0008	566681	Michoacan, Mexico	2	
RIMME0009	566682	Distrito Federal, Mexico	2	
RIMME0011	566685	Mexico, Mexico	2	
RIMME0014	714151	Breeders line; Puga: 11066	6	
RIMME0017	699874	Ayotlan, Mexico	8	
RIMME0021	N/A	El Porvenir, Mexico	69	
RIMME0026	N/A	Opopeo, Mexico	42	
RIMME0028	N/A	Puruandiro, Mexico	28	
RIMME0029	N/A	Ixtlan, Mexico	35	
RIMME0030	N/A	San Pedro, Mexico	27	
RIMME0031	N/A	Tenango del Aire, Mexico	25	
RIMME0032	N/A	Nabogame, Mexico	24	
RIMME0033	N/A	Puerta Encantada, Mexico	25	
RIMME0034	N/A	Santa Clara, Mexico	23	
RIMME0035	N/A	Xochimilco, Mexico	25	
RIMPA0001	87168	El Salado, Mexico	4	
RIMPA0003	87171	Mazatlan, Mexico	8	
RIMPA0017	87200	N/A	4	
RIMPA0019	87213	El Salado, Mexico	2	
RIMPA0029	87244	N/A	2	
RIMPA0031	87249	N/A	2	
RIMPA0035	87288	Jalisco, Mexico	4	
RIMPA0040	288185	Mexico, Mexico	4	
RIMPA0042	288187	Guerrero, Mexico	4	
RIMPA0043	288188	Guerrero, Mexico	4	
RIMPA0045	288193	Guerrero, Mexico	4	
RIMPA0055	714152	Breeders line	2	
RIMPA0056	714153	Breeders line	2	

Accession	Number of alleles sampled		Hopscotch Frequency
RIMMA0066	2		1
RIMMA0075	2		1
RIMMA0077	2		1
RIMMA0079	2		1
RIMMA0081	2		1
RIMMA0084	2		1
RIMMA0086	2		1
RIMMA0088	2		1
RIMMA0089	2		1
RIMMA0090	2		1
RIMMA0092	4		1
RIMMA0094	4		1
RIMMA0097	2		1
RIMMA0099	2		1
RIMMA0100	2		1
RIMMA0101	2		1
RIMMA0104	2		1
RIMMA0108	2		1
RIMMA0111	6		1
RIMMA0115	2		1
RIMMA0117	2		1
RIMMA0130	2		1
RIMMA0133	2		1
RIMMA0134	2		1
RIMMA0135	2		1
RIMMA0142	2	34	0.5
RIMMA0143	4		1
RIMMA0146	4		1
RIMMA0149	2		1
RIMMA0152	2		1
RIMMA0153	2		1

PC1		PC2		PC3		PC4		PC5	
Var	**Rot**	**Var**	**Rot**	**Var**	**Rot**	**Var**	**Rot**	**Var**	**Rot**
bio1	0.146	bio4	0.244	prec7	0.287	ts_clay	0.41	bio2	0.38
tmean11	0.146	bio3	0.241	prec8	0.276	v_mod	0.359	sq4	0.328
tmean12	0.145	bio7	0.241	prec11	0.262	ts_sand	0.329	ts_loam	0.289
bio11	0.145	prec6	0.237	bio13	0.247	bio15	0.272	ts_sand	0.266
tmax12	0.145	sq7	0.218	prec1	0.246	prec4	0.259	sq7	0.231
tmin5	0.145	prec9	0.217	bio16	0.242	x_mod	0.244	bio18	0.213
tmean1	0.145	sq3	0.207	prec12	0.24	prec3	0.226	bio13	0.207
tmean2	0.145	prec12	0.207	bio19	0.238	sq3	0.21	prec11	0.183
tmin4	0.145	bio12	0.204	bio12	0.231	prec5	0.21	bio7	0.17
tmax1	0.145	bio19	0.196	prec2	0.222	prec7	0.19	bio16	0.163
tmean4	0.145	prec2	0.188	bio18	0.221	sq4	0.186	bio4	0.157
tmin11	0.144	prec1	0.185	sq4	0.2	bio3	0.185	bio12	0.156
tmax11	0.144	prec10	0.184	prec9	0.18	bio18	0.178	bio3	0.155
tmin12	0.144	bio16	0.183	prec10	0.171	sq7	0.132	prec6	0.154
tmin2	0.144	prec8	0.17	prec5	0.161	bio14	0.116	x_mod	0.152
tmean5	0.144	prec5	0.165	prec4	0.154	bio13	0.099	prec9	0.144
tmean10	0.144	bio14	0.158	sq3	0.147	bio16	0.095	prec8	0.143
bio6	0.144	bio13	0.151	bio2	0.143	prec8	0.09	v_mod	0.142
tmax2	0.144	bio17	0.149	bio17	0.129	bio7	0.077	bio15	0.136
tmean3	0.144	prec3	0.144	ts_loam	0.127	bio4	0.075	prec7	0.112
tmin1	0.143	ts_clay	0.141	v_mod	0.123	bio2	0.074	prec4	0.108
tmin10	0.143	bio2	0.129	prec3	0.113	prec2	0.074	bio14	0.096
Altitude	0.143	prec7	0.108	x_mod	0.111	bio19	0.068	tmax7	0.093
bio9	0.143	tmax6	0.107	bio14	0.099	prec12	0.056	tmax8	0.092
tmin3	0.143	x_mod	0.106	bio4	0.07	ts_loam	0.053	prec1	0.091
bio10	0.142	bio15	0.098	tmax3	0.067	tmax12	0.047	prec2	0.086
tmax10	0.142	ts_loam	0.088	ts_clay	0.065	bio17	0.047	tmin11	0.086
tmax3	0.142	tmean6	0.085	bio15	0.056	bio9	0.043	prec5	0.082
tmax4	0.142	tmin7	0.082	tmax2	0.055	tmax8	0.042	bio17	0.082
tmin6	0.142	bio5	0.082	tmean3	0.052	tmax1	0.041	tmin12	0.08

Ejutla A	4	0.15217	0.11902	0.76191
Ejutla B	5	0.15258	0.14877	0.07412
La Mesa	3	0.12802	0.08926	1.09209
San Lorenzo	3	0.09098	0.08926	0.04845