Reviewer 1 (Anonymous)

Comments for the author

Populations of crop wild relatives often carry alleles associated with domestication traits at low to moderate frequencies. The presence of these alleles could potentially reflect post-domestication crop-to-wild gene flow or standing genetic variation that predates the domestication process. This study examines the frequency and phenotypic consequences of one of the best characterized crop domestication alleles, maize tb1, in populations of its wild ancestor, teosinte. The tb1 gene functions as a shade avoidance-mediated repressor of organ growth. In domesticated maize, a Hopscotch TE insertion in the tb1 promoter leads to enhanced gene expression and the loss of tillering and branching that characterizes domesticated maize. In this paper, the authors use PCR assays to survey for the presence of the TE insertion in a large sample of accessions, including maize, its wild ancestor teosinte, and highland teosinte (a separate subspecies less closely related to the domesticate). The study examines the potential ecological significance of the TE insertion allele through a Bayenv analysis, and population genetic analyses are used to confirm previous findings that the TE insertion predates maize domestication.

The paper is generally well written, and the conclusions are basically sound. Addressing the points below would improve the paper.

1) The terminology on accessions, individuals, and populations is confusing (e.g., 67-71). Please clarify the difference between individuals and accessions (e.g., line 70 is this referring to the number of individuals per accession? How are individuals of one accession related? Are these full sibs from the same maternal plant?..)
okay we can do this...
yep, pretty straight-forward

2) Is the Locality designation in Table S1 the same as population as used in the text? Please clarify and/or be consistent with this terminology.
Yes, can clarify

3) Fig. 1 should indicate the boundaries of the Mexican states that are used in describing population locations in the text (Jalisco, Colima, etc.). The four focal populations from Jalisco state should also be labeled in this figure (i.e., San Lorenzo, La Mesa, Ejutla A, Ejutla B).

I think this would make figure super busy and it's already verging on super busy. Maybe compromise on labelling the four pops within their pie charts.

if you still have access to the figure in ArcGIS, should be easy to add state boundaries and you can make the boundaries very light to limit busyness; if not, I'd just label the Jalisco pies

raster in R is not too hard either.

4) It would be helpful to include Fig. S1 in the main text of the paper and to label the two sequenced regions (Region 1 and 2) in the figure itself (not just in the caption).

Okay can do this

yep, easy to do and this is an online publication without figure charges so why not!

5) Fig. S2. Annotate this figure to show the size (bp) of each band in the 1 kb ladder. The caption should also define the abbreviations used in the labels.

this will be too much. I can label the brightest band (1kb) and the top and bottom bands

I'd label as many as it takes so that it's obvious what intervening band sizes are. We'll want to define the abbreviations too

6) What does No Hop/Pif indicate in Fig. S2? It doesn't have the three predicted bands of a TE heterozygote, and lane 6 seems to be empty altogether. Please clarify.

Maybe in the text I can clarify total bands for individuals. I.e. if an individual is Hop/Hop, it will have a total of two bands that are split between the two PCRs, whereas if you are a heterozygote you will have 3 split between two PCRs. I wonder if we should ditch this figure entirely and I can draw up a cartoon one?

since reviewer 2 is also concerned about the PCR and this is the only thing the editor really mentioned, I think we need to be pretty thorough here. Ideally we would have a Hop/Hop homozygote, a Hop/No Hop heterozygote, and a No Hop/No Hop homozygote in the figure. We don't need the Pif since this really isn't mentioned in the manuscript. Do you have gel figures showing both homozygotes and the heterozygote banding patterns? I'm afraid a cartoon would send the message

that we don't have clear PCR results

agree with the huff

7) Line 118. Figure S1 should be cited here, not Table S1.

Okay can fix

8) Line 150. Why were only 100 bootstrap replicates used for the NJ analysis? 1000 is more typical.

I do not know the answer to this. I don't think that 1000 bootstraps would give us really a difference.

if you still have the input file for this analysis, should be a quick fix and show we're doing everything we can to address reviewer concerns

agreed, should be easy to redo. while agree this is silly and i don't think 1000 will change anything, but since it's easy to do it's hard to argue against doing it.

9) Lines 164-175. It would be helpful to provide some more detail on the methods used in defining haplotypes in the STRUCTURE introgression analysis. If I understand correctly, SNPs within 5 kb windows were used to define haplotypes for that window. Were the haplotypes then used to define diploid genotypes for each individual in the STRUCTURE input file? If so, how was phasing performed in defining the haplotypes? Is the output in Fig. 3 showing maize population assignment (i.e., membership coefficient values) for adjacent 5 kb windows? Was the STRUCTURE analysis performed only for the 8 Mb region shown in Fig. 3, or do the results presented in Table 4 reflect the entire chromosome?

yes SNPs within 5KB window are a haplotype. I don't know about the diploid/phasing stuff. Yes to maize assignment. Point of figure is that low assignment to maize in populations with high frequency of Hopscotch. Am plotting the assigment value across a region of chromosome 1. I showed a region and not entire chromosome to be able to highlight the position of tb1 and so that you can see the difference in assignment values.

methods say that you used phased data from Tanja's manuscript. Did she give you phased data or did you phase this yourself with fastPHASE? Either way, probably best to mention the data were

phased with fastPHASE and add a citation to the methods paper for this

11) Line 186. How does the 1-foot spacing of plants in the phenotyping experiment compare to conditions used in the experiments of Lukens and Doebley that documented density-dependent tillering? It could be worth bringing this up in the discussion of the negative phenotyping results (i.e., Fig 4 and Discussion, p. 13).

No idea off the top of my head. Good point, can look up and mention

yep, good point to check on...I'd add some discussion on this, particularly if they saw very low tillering at this density

12) Line 223. Indicate here that these four populations are parviglumis and are from Jalisco state.

Okay can mention

13) Line 228. Why is Table S2 cited here? That table doesn't include any information on nucleotide diversity estimators.

will check

14) Table 2 and the corresponding text (p. 9) should indicate which Tajima's D values, if any, are statistically significant. If none are statistically significant, then this should be pointed out in the text.

Okay. I am not sure that we have significance values for these.

We don't and it's kind of a strange/silly request...Jeff and I will deal with this

i can do the simple but dumb calculation if you guys want for response to reviewers, but disagree we should include in manuscript. we mention they are at the tail of the empirical distribution from wright et al don't we?

15) Lines 231-232. Suggested rewording: 'in all populations except La Mesa, where a slightly negative value suggests a slight excess of low frequency variants'

Okay reword.

16) Lines 249-250. 'average r2 is slightly lower in the tb1 region..' Also, include these mean values as a row in Table 3.

Do you guys want to do this?

simple enough to make the reviewer happy here and just add one more row to the table

17) Fig. 3 could be moved to online supplementary data, since the key information with respect to the tb1 region is already presented in Table 4.

I have no strong feelings about this.
fine with me to move to supplement

18) Discussion, lines 306-8. 'The Hopscotch allele is more prevalent in parviglumis than in mexicana in our sample, suggesting a different history of the allele amongst teosinte subspecies.' This inference is contradicted by the complete lack of differentiation between the subspecies for the TE (Fct = 0; Results, lines 213-217). This should be addressed. I don't see any evidence presented that the TE differences between the subspecies are statistically significant. Given the unbalanced geographical sampling of the two subspecies (see Table S1), I would also question whether these samples are appropriate for comparing TE frequencies in this way.

Okay we can maybe take these statements out? It would seem that none of the reviewers have ever been happy with what we have had to say with respect to mex vs parv, or unequal sampling. we have discussed why sampling is unequal and why there was no phenotyping for mex.

my vote would be to remove the comparisons between *parv* and *mex* but to still note the high

frequency in Jalisco

ok either way here

19) There are some sloppy inferences about selection in the Discussion and Conclusion 'e.g., lines 345-354, 388-389: there's no evidence that the Tajima's D values cited here are actually statistically significant. The statement and citations in lines 349-351 suggests that they likely are, but this should be backed up with tests of statistical significance. See comment 14 above.

Okay. What do you guys think of this? I don't think we have tests of significance for tajd

Jeff and I will address this

sloppy my ass. ¡censored¿ reviewer!

Reviewer 2 (Anonymous)

Basic reporting

1) In the Introduction, the authors give some background about teosinte, but do so mainly as the precursor of maize, not so much as a wild species. For instance, they do not give much information about the ecological differences between parviglumis and mexicana, or about what is known for this species in terms of population structure, population density or ecological behavior. This should be added. In particular, the authors state that the tb1 gene is suspected to play a role in shade avoidance, so information on population density should be highlighted: Is it similar in all teosinte populations? Does it vary with altitude? With other ecological data? Does it vary between parviglumis and mexicana?

I don't think this is necessary. Again, everyone has seemed to want something different in the intro. Last person wanted more maize/shade avoidance/tb1. Perhaps we can add a couple of sentences in about mex vs parv or the population stuff. I would like to not have to rewrite the entire intro again.

I might mention something briefly about *parviglumis* being a ruderal species that can occur at high density (cite my *Molecular Ecology* manuscript), but, no, I would not add a whole paragraph on the ecologies of the teosintes.

this sounds OK. a wee bit of self-citation there hufford?

2) The authors refer several times to a previous study they performed: Pyhjrvi et al., 2013. This is fine, but it is not always easy to understand which results are new and which come from this previous study. For instance, on line 158, it is stated 'we had whole genome SNP data come from Pyhjrvi et al., 2013. Similarly, Table 1 on line 217 should be moved right after 'populations (0.23) on line 216 to avoid confusion.

Okay I can check to make sure citations and data stuff from tanja are clear

3) In the Results section, some titles like "Genotyping" "Sequencing" or "Phenotyping" are too vague, and sound more like Materials and Methods than to Results. They should be modified in order to highlight what question is addressed. I like our sections...
easy enough to make these more hypothesis oriented; good idea to do as many of these simple things as possible to show we're making a strong effort toward revisions

Paragraph "Neighbor joining trees" on lines 235-239 should be moved to the section "Evidence of introgression"

Experimental design

4) The biological question is to characterize whether the Hopscotch transposable element at tb1 (which enhances expression of tb1) plays a role in the ecology of teosinte, especially in high-density populations. To do so, they characterize the distribution of this element in parviglumis, mexicana and maize landraces and they examine the phenotypic effects of the insertion in parviglumis. The authors state that they sampled 1,110 individuals from 350 accessions, with between 1-18 and 1-43 individuals per population. However, they do not provide explanation on (i) how these populations were chosen and (ii) why a different number of individual was sampled per accession (visible also in suppl. Tables S1 and S2). This should be explained.

Simply put it is what we had available at the time. There was no specific choosing or decision about number of individuals per accession

we'll have to diplomatically state this...we sampled all teosinte available to us and while this didn't allow for even sample sizes across populations, it did allow us to calculate both frequency in a subset of populations and geographic breadth of the *Hopscotch* across a large number of independent sampling locations

5) In Suppl. Table S1, some of the parviglumis accessions are listed as "Breeders line". Can these be considered as natural populations? The authors should explain where they were originally collected. In Suppl. Table S2, the USDA (or other provider) ID corresponding to each accession should be provided, as it is done in suppl. Table S1.

Okay. Is there somewhere we can get all of this info.? I don't believe I have access to the database any more

Yes, we need to include this info.; that's a fair request

6) The authors write that the PCR amplification to investigate presence of the Hopscotch element leads to two amplification products: one for the entire element, (5kb) and one for amplification of part of the element only (1.1kb). Why is this? How do the authors explain the origin of the 1.1kb band? The sentence with "and" on line 83, suggests that the two bands are produced in one single homozygous plant. If some accessions really amplify two bands, is the primer located in the LTR? If it is a typo and the "and" on line 83 should be replaced by a "or" is the element truncated in some accessions, leading to a 1.1kb band instead of the 5kb one? This should be clarified.

Why what? Why two products? or why did we do it that way? I think it is clearly stated that we can test for presence/absence/hets and pcr bias with sometimes a large fragment being difficult to amplify. The location of the primers is all shown in the cartoon figure and stated in the text. two primers outside of LTR one primer within.

if there's any way to be even slightly more clear, I'd edit just so you can say here that you've edited for clarity

7) From Figure S1, the middle primer seems to be right in the middle of the element, which suggests it was not designed in the LTR region. Is this true? In which part of the retrotransposon

was the primer designed? The authors should add position of the retrotransposon LTRs and coding regions on this figure.

What? It is in the LTR!!!! Gah! That's the whole point!

8) In suppl. Figure S1, the names "HopF/HopR/HopIntR" should be added, as well as corresponding expected amplification sizes.

Okay can do this. maybe will clear up their misunderstanding detailed above

9) In suppl. Figure S2, band size of several bands (close in size to these amplified) should be indicated next to the ladder. Legend should explain better what the figure shows. It states "Genotypes are indicated at the top of the gel" What does this refer to? Numbers? "Hop/Hop" code? The figure should be completely relabeled, so that primer pair names and type of genotype (with presence of absence of Hopscotch) are clearly identified. What is "no Hop/Pif"? Why are there two bands on lane 5? Lane 6 has a weird smear. No primers are visible on lanes 5, 6 and 7.

I can relabel or clarify in the figure legend. I don't think that redoing a gel is possible at this point.

10) Bottom gel resolution is poor. On the right for low molecular weight bands, it is difficult to assess that there is a single band (lane 7). A 1% gel and 1kb ladder are clearly not adequate for 300bp band detection.

Redoing gel not possible at this point

even if you have a better resolution image to include, that might help. I'd also remove the pif lanes as possible as we never mention this

11) It is not stated whether PCR products were sequenced to check for correct amplification. Were some of the PCR products sequenced? If not, this should be done.

I think that we did initially to check. But all of the products match the appropriate and expected

size. The only issue we had was suddenly the different size with the Pif, which we did check with sequencing.

we need to check this. can you go back through emails or to your lab notebook to find whether this was sequenced?
i don't remember ever checking this. we checked the pif by PCRing stuff John had already sequenced in which i identified the pif

12) The authors write: "Environmental data represent average values for the last several decades (climatic data) or are likely stable over time (soil data)". Does this mean that soil data was not averaged over the last decades? How was it estimated? The fact that soil parameters are "likely" stable over time actually quite depends on what we consider as "soil". Does this include micro-organisms? A bit more detail should be given.

My guess is that we can look this up on WorldClim if you guys think it is necessary. My guess is that microorganisms were not included in analysis of soil characteristics, that it was things like what compounds are in soil, etc.

soil data are from the Harmonized world soil database that we cite and that Tanja also used in her manuscript. There will be further details on data collection there which we can briefly describe in our response and point the reviewer/editor to

13) Figure 1: This figure should show 37 populations of parviglumis and 4 populations of mexicana (see page 8, lines 211 to 213). But the figure legends indicates only parviglumis. It should include mexicana populations, and they should be differentiated by an appropriate color code. In some cases, several circles seem to derive from one circle, suggesting they correspond to several populations from the same location. On the left hand side, it is unexpected to globally have a null frequency of "No Hopscotch" while several populations have a large fraction of "Hopscotch". This means that, within 25 miles, the frequency of the "Hopscotch" allele varies greatly. The authors should discuss this point. Names of the parviglumis populations used for the rest of the study (La Mesa, San Lorenzo, Ejutla1 and Ejutla2) should be indicated.

i will check this but i believe mexicana not included because we didn't discuss them with respect to the high frequency stuff. I do not have arcgis any more and would prefer to not completely redo this figure. I am happy to put names on pops though. I don't understand comments about 'left hand side' stuff. Yes, some of the populations are close together when you are that zoomed out on a map. There is no way to have them not all coming from the same place in some cases unless the map was so gigantic to take up an entire page.

the reviewer is indicating they are surprised that there are such marked differences in Hop frequency in such a small area in the populations on the lefthand side of the map. You can cite my dissertation and the Tiffin and Moeller paper as finding that populations in this area were highly differentiated and that populations with Hop were much more closely related to each other than they were to pops without Hop...just some very different histories in these populations and little gene flow between them

Validity of the findings

14) Results of the PCR amplifications are not well enough described and of high enough quality to be able to conclude whether the retrotransposon detection is valid or not (see "Experimental Design"). The authors should improve labelling of suppl. Figure S2 and, if needed, run the PCR amplicons on a more concentrated gel and with appropriate ladder.

We can relabel but not redo gels at this point
the PCR results were conclusive, but I'd make an effort to make the figure more clear with labeling or a higher resolution image and to clarify the description of the banding patterns of the genotypes in the text; my guess is we'll need to do a good job of this in particular if we want this manuscript in *PeerJ*

15) Based on the association genetics study performed, the authors find an absence of association between the Hopscotch insertion and tillering index. At 40 days, they even find a weak but significant correlation, but in the unexpected direction (homozygotes for the Hopscotch insertion have a higher tillering). The authors discuss that this could be due to variation at other unlinked loci. However, they do not discuss on the effect of the environment on tillering. Considering the

experiment was performed in greenhouse conditions while teosintes grow in much different environmental conditions, the genotype x environment interaction could differ greatly to this obtained in teosinte natural environmental conditions. This should be discussed.

Sure greenhouse is different than outside, but we can't control for variation outside whereas we can in a greenhouse. I don't understand their point. Greenhouse is best we could do here since they wouldn't all fit in a growth chamber.

I'd point out that the effect of *tb1* on tillering has been verified a number of times in a number of environments (probably in a greenhouse setting) and cite studies like Briggs, etc...