Response to Reviewers

Dear Jun Lyu (Associate Editor, Nature Plants),

Thank you for considering our manuscript, “Recent demography drives changes in linked selection across the maize genome”. We have carefully considered your suggestions, as well as those of the three reviewers who read our work. Below, we provide a point-by-point response addressing these comments and highlighting the corresponding changes we have made to the manuscript. The word for word comments are below in plain text, and our responses are set in bold typeface.

Thank you kindly,
Tim Beissinger and Jeff Ross-Ibarra

Editorial comments:

The current version is formatted for Nature Genetics. Although we agree it's an interesting idea, our referees have complained about it in their remarks to editor. As such, we need you to provide a plain Word version without pre-formatting for the following processing of the manuscript.

The reason for the Nature Genetics formatting is that the manuscript was transferred from Nature Genetics directly to Nature Plants. We have revised the manuscript to a plain LaTex format, consistent with that specified in the instructions for authors.

Reviewers comments:

Reviewer #1 (Remarks to the Author):

The authors analyzed 23 maize and 13 teosinte genomes to investigate the forces shaping the genetic diversity. They observed a reduced genetic diversity around genic regions. They excluded prominent effects by hitchhiking or soft sweeps, and concluded that background selection/purifying selection may be the major cause. An important conclusion is that demography affects the efficiency of purifying selection. The topic and the conclusions are really interesting. However, some analyses in this paper have some caveats that I would suggest the authors to address:

1. Page 2, section "Patterns of diversity differ between genic and inter-genic regions of the genome".

The authors should shortly discuss whether any cryptic population structure among the 23 maize samples or not. Tajima's D and the demographic inference with dadi are valid only in a random mating population without population structure.

We very much agree with the reviewer that cryptic populations structure among our samples is a concern. We include discussion on how population structure in teosinte may influence our estimate of maize-teosinte divergence time, and we performed an evaluation of population structure using principal component analysis in Figure S8. This analysis shows no discernible population structure among our maize samples, but as described in the methods we did remove four teosinte individuals because of concerns about structure stemming from this principal component analysis.

2. Page 2, right Col, line 1-4 "we utilized genotyping data from more than 4,000 maize landraces to estimate the modern maize effective population size using low frequency variants informative of population expansion. This analysis yields a much higher estimate of the modern maize effective population size at Nm ≈ 993, 000.":
Details of the inference approach is needed here. What are the inferred values of ancestral population and initial growth time (or the growth rate)? The authors may also try the method from a recent paper in MBE (Chen, Hey and Chen (2015)), which also infers recent demography from large samples.

We apologize for the lack of clarity in the main text. We have modified the text to better reflect the methods, specifying that we use Fu and Li's 1997 estimator of theta and a published estimate of the mutation rate to arrive at this estimate. This is also detailed more fully in the methods section. The results now read:

> **...we investigated two alternative approaches for demographic inference. First, we utilized genotyping data from more than 4,000 maize landraces [30] to estimate the modern maize effective population size. Because rare variants provide the best information about recent effective population sizes [31] , we estimate N_e using a singleton-based estimator [32] of the population mutation rate $\theta = 4N_e\mu$ and published values of the mutation rate [33] (see online methods for details). This yields a much higher estimate of the modern maize effective population size at $N_m \approx 993, 000$.**

3. "Finally, we applied a model-free coalescent approach (30) using a subset of our samples..."

How many samples were used by msmc (ref 30) to estimate the parameters? Sample size affects the performance of msmc.

We have now clarified the results to include mention of the sample size:

“Finally, we employed a model-free coalescent approach [34] to estimate population size change using a subset of six genomes each of maize and teosinte.” Additional details of the MSMC approach are included in the methods.

Furthermore, msmc provides divergence time for the two species, which was not mentioned in the main text. Is it similar to dadi's estimate? The divergence time by dadi is ≈ 15, 000 years ago (page 5, left col, line 5), which is quite different from archeological record. Does msmc support this number too?

We have run an additional MSMC analysis to estimate cross-coalescent rates between maize and teosinte. These should be 0 for completely isolated populations, and large for a single panmictic population. This results (now in supplementary Figure S2), shows rates of cross-coalescence beginning to plateau at 10-15K years, roughly consistent with the timing found using dadi. That these values do not plateau at a value of 1 suggests some level of ancestral population structure, consistent with our explanation for why our estimate of 15K is older than the archaeological record. This section of the results now reads,

> **“Though this analysis suggests non-equilibrium dynamics for teosinte not included in our initial model, it is nonetheless broadly consistent with the other approaches, identifying population isolation beginning between 10,000 and 15,000 generations ago, a clear domestication bottleneck, and ultimately rapid population expansion in maize to an extremely large extant size of ≈ 10^9 (Figure S2).”**

4. The authors chose a two population model when running dadi (Fig. 2), in which teosinte population size is set to be constant and equal to ancestral population size. This may not be an ideal model for reality. The Tajima's D value of non-coding regions is "-0.087 in maize and -0.25 in teosinte" (Page 2), which indicates teosinte has some non-equilibrium recent demographic history, e.g., some more severe exponential growth (?). A fitted model for teosinte is very important for the analysis in later sections: the author’s conclusion on efficiency of purifying selection was made based on comparing diversity patterns between two populations, and by assuming that teosinte is of constant size.

The reviewer makes an excellent and important point that highlights a technical limitation of our dadi analysis: restricting teosinte to a constant population size was a necessity we had to impose in dadi in order to reduce the number of parameters being estimated. Our other analyses also suggest the reviewer is correct, with e.g. MSMC (figure S2) showing that teosinte N_e has increased in the recent past. We have now clarified this in the methods: “To minimize the number of parameters estimated, we employed a simple demographic model which posits a teosinte population of constant effective size N_a. At time T_b generations in the past, this population gave rise to a maize population of size

N_b which grew exponentially to size N_m in the present (Figure 2).” We have also added a sentence to the discussion clarifying this could be a point for future work: “This model paves the way for future work on the demography of domestication, evaluating for example the significance of differences in gene flow estimated here or removing assumptions about demographic history in teosinte”

Nonetheless, our main conclusion, that recent growth in maize has qualitatively changed the effects of linked selection, is robust to these deviations from the model estimated in dadi. In figure S5 we show that the difference between maize singleton diversity and pairwise diversity is much greater than the difference between these measures of diversity in teosinte. Therefore, though teosinte N_e has likely increased, it has not grown to match the modern N_e of maize.

As a final validation of our conclusion, we have also added new simulation results to the supplement. Briefly, these simulations show that, under a model very similar to the one we estimate for maize, differences between mean pairwise nucleotide diversity and singleton diversity should be apparent between maize and teosinte for weakly selected mutations.

We have modified the text in the purifying selection section of the manuscript to now read “While direct comparison of pairwise and singleton diversity within taxa is consistent with non-equilibrium dynamics in teosinte, these too reveal much stronger differences in maize (Figure S5) and mirror results from simulations of purifying selection (Figure S6).

5. The average of Tajima's D for genic regions of maize is 0.4. Is it due to one or several outlier genes? If not, the authors may check why many snps in the genic regions are in high frequency.

Histograms shown in Figure 1 show that differences in Tajima’s D in genic regions are not due to a few outliers. We argue that the shift in Tajima’s D towards fewer rare variants in genes is a direct consequence of the change in Ne and resulting changes in the efficacy of linked selection. This is best shown in Fig. 4 where we show decreased singleton diversity in maize with respect to teosinte.

Reviewer #2 (Remarks to the Author):

This manuscript explores the interaction between demography and selection across the maize and teosinte genomes. The MS is well written and easy to read, and the authors achieved a good balance between length and complexity of the subject matter. The most interesting results of this paper are as follows: 1) the apparent lack of 'hard' selective sweeps in the maize genome (potentially as a result of genome size and trait complexity), and 2) the increased efficacy of purifying selection in maize following population expansion post-domestication. I think that the

latter finding is particularly interesting, and that it has not (to my knowledge) been demonstrated in another domesticated plant. This finding may well change some of the research conducted on domesticated plants, where the effects of bottlenecks and associated drift tend to be emphasized (along with the effects of artificial selection), but where the effects of population expansion are not often considered. Overall, this is a valuable paper for the field of population genetics and plant domestication.

Before publication, I would ask that the authors address a few things about the manuscript.
1) Can you add a section to the discussion explaining why it is that the increased population size of maize has not yet lead to a more pronounced decrease in π for genic regions in maize? That is, why does purifying selection only seem to act on the younger polymorphisms in maize and not more broadly across the genome (especially given the massive Ne for maize)? Why hasn't it 'caught up' with teosinte since the bottleneck? Do you expect that this point might be reached given sufficient time (at least theoretically), or would it be impossible for this to occur for some reason?

This is an important aspect of the study, and we thank the reviewer for noting that it needs to be explained more clearly. To more explicitly address the reviewer's questions, we have re-written portions of the "Population expansion leads to stronger purifying selection in modern maize" section of the text. Additionally, we have added subheadings to the discussion section where we elaborate on the implications of this observation, including sections called "Demography influences the efficiency of purifying selection" and "Rapid changes in linked selection"

2) The second sentence of the abstract seems out of place; it addresses drift, which is not really a focus of this paper. If it was meant as a contrast to the paper's focus on selection, there should be a better transition to the third sentence.

We thank the reviewer for identifying a lack of clarity regarding the manuscript's focus. Drift is a major focus of this paper, in the sense that demographic history is akin to changes in population size, and the magnitude of drift is controlled by population size. To strengthen this connection and develop a more natural transition, we changed the sentence in question to read, " The impact of genetic drift in a population is largely determined by its demographic history, typically summarized by its long-term effective population size (N_e). Rapidly changing population demographics complicate this relationship, however."

3) Finally, in the first full paragraph on page 5, you state that "the estimated timing of domestication at ≈ 15, 000 years before present." While you go on to soften this statement by saying that what you are actually measuring is the split between teosinte and maize, the first sentence is still an exaggeration, and one of the type that tends to lead to animosity between the archeological and genetic communities. The isolation of a cultivated population from wild congeners may be a first step towards eventual domestication, but it is not equivalent to the

many genetic and phenotypic changes required for domestication. Can you please modify this first sentence to indicate what is actually being measured?

This is an excellent caution, as we do not mean to suggest that ancient farmers began domesticating maize 15,000 years ago. We have changed this sentence to read, "One surprising result from our model is the estimated divergence time of maize and teosinte approximately 15,000 generations before present."

Reviewer #3 (Remarks to the Author):

I have read the work of Beissinger and colleagues on the genome wide comparison and estimate of demographic parameters in wild teosinthe and domesticated maize.

I see no "major" flaws in the manuscript. Ideas are sound (although some are debatable), analyses are properly conducted and the arguments are nicely laid out in the discussion. As such, it is a paper worthy of a publication in a good journal. Is it "good enough" for NaturePlants? That's hard to say. As I read the manuscript, I went back and forth between yes and no. Partly because NaturePlants is a new journal and therefore I am not sure where it sits in the Impact Factor ladder. Partly also for the reasons I explain below.

I'll start with the weaknesses:. The data are not novel and this manuscript essentially presents new analyses on previously published data. Some of the results are also fairly basic, not all that surprising (e.g. Fig1, and the fact that there is little difference documented between wild teosinthe and domesticated maize). The study also documents extremely broad genome-wide patterns that surely miss locus-specific effects. On a positive side, this is a topic of great interest (maize domestication, and population expansion/demography in general), the article is readable for a wide audience, and therefore should attract a faire bit of attention. The reconstruction of the demographic history from whole genome data is I believe fairly robust and novel. The analysis of "younger - polymorphisms" (singletons) showing patterns distinct from pairwise diversity" is also quite interesting, even though I think it could be better introduced / explained.
Overall I tend to lean towards accepting
the manuscript at NaturePlants.

We appreciate the reviewer's recognition of merit in our work and agree that demography is an important consideration for understanding evolutionary processes during crop domestication. We too feel that comparison of singleton diversity (younger variation) to pairwise diversity (older variation) reveals novel, previously overlooked dynamics during the process of crop domestication and subsequent expansion. We did not use new sequencing data because of the cost of sequencing additional genomes and the fact that the full genome data for teosinte and maize landraces published in the maize HapMap2 project were entirely sufficient for the goals of this project. These data, combined with a simple, well understood demographic history and an excellent genetic map make *Zea* the

ideal study system for characterizing the complex interplay between demography and selection. While Reviewer 3 may not find some of our results surprising, we concur with Reviewer 2 that our study represents the first demonstration of increased efficiency in purifying selection during crop expansion following domestication. This should prove to be a valuable and lasting contribution of our manuscript. Finally, while Reviewer 3 hoped to see locus-specific results, the effects of demography are more clearly understood through a genome-wide approach since historical change in population size affect diversity across the genome.

###Does the manuscript have flaws which should prohibit its publication?
no. Methods are sound and stat tests are appropriate. Raw data was already available, Code used for analysis is freely available on github.

###Specific Comments:
I found the title: "linked selection" and how you present the idea of "linked selection" a little odd. I know what you mean, but you appear to present this concept as a new, different selective force, where in the end, all you mean is genetic linkage among sites.

We apologize if we seem to be presenting this idea as new. The reviewer's interpretation of it referring to linkage between neutral and selected sites is spot-on. The phrase "linked selection" is relatively widespread in population genetics: Google Scholar shows nearly 300 uses of it in articles since 2012. Below are a few references in several species, most of which we cite in the text, demonstrating that the term is frequently applied, including in manuscript titles. The first time we use the phrase in the introduction, we have modified the text to better describe it. It now reads, "Linked selection, which refers to the effects of selection at one site on diversity at linked sites [8],...". We would like to keep it in the title so that authors searching for similar results are more likely to find our work. Relevant references include:

1. **The impact of <u>linked selection</u> on plant genomic variation, Slotte, 2014, Briefings in Functional Genomics.**
2. **<u>Linked selection</u> and recombination rate variation drive the evolution of the genomic landscape of differentiation across the speciation continuum of Ficedula flycatchers, Burri et al, 2015, Genome Research**
3. **Genomic signatures of selection at linked sites: unifying the disparity among species, Cutter and Payseur, 2013, Nature Reviews Genetics**
4. **A genomic map of the effects of <u>linked selection</u> in drosophila., Elyashiv et al, 2014, arXiv**
5. **Natural Selection Constrains Neutral Diversity across A Wide Range of Species, Corbett-Detig et al, 2015, PLoS Biology**

Intro doesn't talk about mutation and recombination rate which is odd also because mutation rate is ultimately responsible for genetic diversity and recomb. rate strongly influences "linked selection".

While it is certainly the case that genetic diversity is the result of the combined effects of selection, drift, recombination, and mutation, we chose to focus on the first two in the introduction since space is limited and they are of primary concern to the manuscript. We have no evidence for differences in mutation rate between maize and teosinte or among regions in maize, so the impact of mutation rate on linked selection (in any organism) is unknown and difficult to quantify. We do discuss linkage in the introduction, which is the relevant result of recombination rate, and throughout the rest of the text we refer regularly to both recombination, as recombination rate heavily influences the effects of linked selection, and mutation, since our demographic inference depends on an estimate of mutation rate. We also discuss mutation in the context of singletons in maize -- new mutations that have arisen after domestication experience a different effective population size because of growth and thus experience more efficient selection.

I was a little confused about your results and explanation behind the "singleton polymorphism". The paper analyses genome wide data, but then it feels like all of a sudden it shifts focus to genotyping data from 4000 landraces, without much details. It seems like the results and implications could be explained better.

Thank you for identifying this section, which was unclear and also identified by reviewer #1. The relevant section now includes a more natural set-up and transition, as well as a reference to the methods section to explicitly point out where details can be found. It now reads:

> **“Because our modest sample size of fully sequenced individuals has limited power to infer recent population expansion, we investigated two alternative approaches for demographic inference. First, we utilized genotyping data from more than 4,000 maize landraces [30] to estimate the modern maize effective population size. Because rare variants provide the best information about recent effective population sizes [31] , we estimate N_e using a singleton-based estimator [32] of the population mutation rate $\theta = 4N_e\mu$ and published values of the mutation rate [33] (see online methods for details). This yields a much higher estimate of the modern maize effective population size at $N_m \approx 993,000$.”**

Page 2:"These observations suggest that diversity in genes is not evolving neutrally, but instead is reduced by the impacts of selection on linked sites."
Again this idea of selection on "linked sites" which I think is oddly formulated.

We have chosen to keep this phrase because it is commonly used in the population genetic literature as discussed above.

Page2: "After establishing that genic sites are not evolving neutrally, we used sites > 5kb from genes to estimate the parameters of a simple domestication bottleneck model (Figure 2)." Here you subtly imply that non-genic sites are evolving neutrally. Sure, you've established that selection acts on genic sites, but that does not imply that there is no selection on non-genic. A few words or a sentence would help to clear this out.

This is an excellent point, and the reviewer is correct that we cannot assume neutrality. We have rephrased the sentence to be more explicit: "To minimize the impact of selection on our estimates, we only included sites >5kb from genes". Even though some non-genic sites are almost certainly under selection, removing sites in and near genes will almost certainly ameliorate to a great extent the confounding of selection in our estimates of demography.

Page 2: "...Mtm = 1.1 × 10−5 × Na migrants per generation from teosinte to maize and Mmt = 1.4 × 10−5 × Na migrants from maize to teosinte.". Can you say something regarding whether or not these migration rates are significantly different (i.e. is there asymmetric gene flow and why)?

Our demographic model is likely an oversimplification (see responses to reviewer #1 about population structure and growth in teosinte above), and therefore we are not confident that our estimates of migration are extremely accurate. Because ignoring gene flow would seem an inappropriate assumption, we included migration in our model. Nonetheless, the exact values of gene flow are not important to our main goal of documenting population size change in maize, and we thus leave precise estimation of differences in gene flow for further studies. To clarify these points we have added text to the results:

> **Although our model provides only a rough approximation of migration rates, we included migration parameters during demographic inference because omitting these could bias our population size estimates. We observe that maize and teosinte have continued to exchange migrants after the population split. We estimate that gene flow from teosinte to maize was M tm = 1.1 × 10 −5 × N_a migrants per generation, and from maize to teosinte we estimate M mt = 1.4 × 10 −5 × N a migrants per generation.**

We have also added a sentence to the discussion to highlight this limitation: "This model paves the way for future work on the demography of domestication, evaluating for example the significance of differences in gene flow estimated here or removing assumptions about demographic history in teosinte."

Page 2: "This analysis yields a much higher estimate of the modern maize effective population size at Nm ≈ 993, 000." Is this because landraces harbour more genetic diversity than the

modern elite lines of maize and therefore they do not represent the same "modern maize"? This should be made clear.

Our study does not include any elite maize lines and therefore cannot address the question. We have changed "effective size of modern maize" to "modern effective size of maize" to hopefully clarify we are still evaluating landraces and not modern industrialized maize agriculture. We also describe in the "Plant Materials" section of the Online Methods that both sets of maize individuals (23 sequenced and 4000 genotyped) were collected from landraces.

Instead, the reason for this difference is because the sample of 4000 individuals has more power to identify rare alleles, and therefore is more sensitive to recent population expansion. As described above, we have re-written this section for clarity. It now reads,

> **"we utilized genotyping data from more than 4,000 maize landraces [30] to estimate the modern maize effective population size. Because rare variants provide the best information about recent effective population sizes [31] , we estimate N_e using a singleton-based estimator [32] of the population mutation rate θ = 4N_eμ and published values of the mutation rate [33] (see online methods for details). This yields a much higher estimate of the modern maize effective population size at N_m ≈ 993, 000."**

Page 5: "Among well-characterized domestication loci, only the gene tga1 shows evidence of a hard sweep on a missense mutation, while several loci are consistent with "soft sweeps" from standing variation or multiple mutations." Can you show the data for these "several loci"?

We apologize if this was confusing. We are not referring to our own data, but instead simply citing results from several published studies of cloned loci which show evidence of standing variation or multiple mutations (i.e. "soft sweeps"). We have modified this slightly to now read:

"Among well-characterized domestication loci, only the gene *tga1* shows evidence of a hard sweep on a missense mutation [36], while published data for several loci are consistent with soft sweeps from standing variation [47,48] or multiple mutations [49]. Moreover, genome-wide studies of domestication [28] , local adaptation [50] and modern breeding [51,52] all support the importance of standing variation as primary sources of adaptive variation."

minor:
Page 2. Typo. last sentence. "in the the maize"

Thank you, this has been corrected.

Figure 1:change "while C and D depict and Tajimas D" to "while C and D depict Tajima's D". Also not clear which graph is maize, which is teosinthe

Thank you, this has been corrected.

I linked Fig S2 as it shows how Ne changes with time. I wonder if it should be moved to the main text. (but fix labeling of x-axis: Years (generations?) since last common ancestor?)

We have modified the x-axis to read, “years before present”. Our method to convert from generations to years is described in the legend: “Time is estimated assuming an annual generation time”. In the interest of space, we prefer to leave this figure in the supplement.