Genome Biology

CrossMark

The interplay of demography and selection during maize domestication and expansion

Li Wang[1,2], Timothy M. Beissinger[3,4,5], Anne Lorant[3], Claudia Ross-Ibarra[3], Jeffrey Ross-Ibarra[3,6*] and Matthew B. Hufford[1*]

Abstract

Background: The history of maize has been characterized by major demographic events, including population size changes associated with domestication and range expansion, and gene flow with wild relatives. The interplay between demographic history and selection has shaped diversity across maize populations and genomes.

Results: We investigate these processes using high-depth resequencing data from 31 maize landraces spanning the pre-Columbian distribution of maize, and four wild teosinte individuals (*Zea mays* ssp. *parviglumis*). Genome-wide demographic analyses reveal that maize experienced pronounced declines in effective population size due to both a protracted domestication bottleneck and serial founder effects during post-domestication spread, while *parviglumis* in the Balsas River Valley experienced population growth. The domestication bottleneck and subsequent spread led to an increase in deleterious alleles in the domesticate compared to the wild progenitor. This cost is particularly pronounced in Andean maize, which has experienced a more dramatic founder event compared to other maize populations. Additionally, we detect introgression from the wild teosinte *Zea mays* ssp. *mexicana* into maize in the highlands of Mexico, Guatemala, and the southwestern USA, which reduces the prevalence of deleterious alleles likely due to the higher long-term effective population size of teosinte.

Conclusions: These findings underscore the strong interaction between historical demography and the efficiency of selection and illustrate how domesticated species are particularly useful for understanding these processes. The landscape of deleterious alleles and therefore evolutionary potential is clearly influenced by recent demography, a factor that could bear importantly on many species that have experienced recent demographic shifts.

Keywords: Demography, Domestication, Genetic load, Maize

Background

Genomes are shaped over the course of their evolutionary history through a complex interaction of demography and selection. Neutral processes that comprise a species' demographic history, such as stochastic changes in population size and migration events, influence both the pool of diversity upon which selection can act and its efficiency. Selection and genetic drift then jointly determine the fate of this diversity.

After the development of agriculture, both crops and humans have experienced profound demographic shifts that left clear signatures in genome-wide patterns of diversity [1, 2]. Early agriculturalists sampled a subset of the diversity present in crop wild relatives, resulting in an initial demographic bottleneck for many domesticates [3]. Subsequent to domestication, humans and their crops experienced a process of global expansion facilitated by the rise of agriculture [4]. In many cases expansion was accompanied by gene flow with close relatives, a demographic process that further altered patterns of diversity [5, 6].

Recent interest in the effects of demography on functional variation has led to a growing body of theory that

*Correspondence: rossibarra@ucdavis.edu; mhufford@iastate.edu
[3]Department of Plant Sciences, University of California, Davis, USA
[1]Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, USA
Full list of author information is available at the end of the article

Wang *et al. Genome Biology* (2017) 18:215

Page 2 of 13

is increasingly supported by empirical examples. To date, the relationship between demography and selection has been most thoroughly explored in the context of deleterious alleles. While theory suggests mutation load (i.e., the reduction in mean fitness caused by the presence of deleterious alleles) may be insensitive to demography over long periods [7, 8], empirical results are consistent with load being shaped by demography over shorter timescales [9–13]. For example, evidence in both plant and animal species has revealed increased mutation load in populations that have undergone recent, sudden declines in effective population size (N_e) [10–12, 14]. Similarly, in geographically expanding populations, repeated sub-sampling of diversity (i.e., serial founder effects) can occur during migration away from a center of origin [15, 16], a phenomenon shown to have decreased genetic diversity and increased counts of deleterious alleles in human populations more distant from Africa [17, 18]. Finally, gene flow may also affect genome-wide patterns of deleterious variants, particularly when occurring between populations with starkly contrasting N_e. For instance, during the Out-of-Africa migration, modern humans inter-mated with the Neanderthal species, a close relative with substantially lower N_e and higher mutation load [9]. The higher mutation load in Neanderthals presented a cost of gene flow, and subsequent purifying selection appears to have limited the amount of Neanderthal introgression near genes in the modern human genome [9, 19].

The domesticated plant maize (*Zea mays* ssp. *mays*) has a history of profound demographic shifts accompanied by selection for agronomic performance and adaptation to novel environments, making it an ideal system in which to study the interaction between demography and selection. Maize was domesticated in a narrow region of southwest Mexico from the wild plant teosinte (*Zea mays* ssp. *parviglumis*; hereafter, *parviglumis* [20–22]) and experienced an associated genetic bottleneck that removed a substantial proportion of the diversity found in its progenitor [23, 24]. Archaeological evidence suggests that after initial domestication, maize spread across the Americas, reaching the southwestern USA by approximately 4500 years before the present (BP) [25] and coastal South America as early as 6700 years BP [26]. Gene flow into maize from multiple teosinte species has been documented in geographical regions outside of its center of origin [5, 27]. To date, genetic studies of demography and selection in maize have primarily focused on initial domestication [28], only broadly considering the effects of subsequent change in population size on diversity [2] and largely disregarding the spatial effects of geographic expansion and gene flow (but see [29]). Furthermore, the effect of maize demography on the prevalence of deleterious alleles has yet to receive in-depth attention.
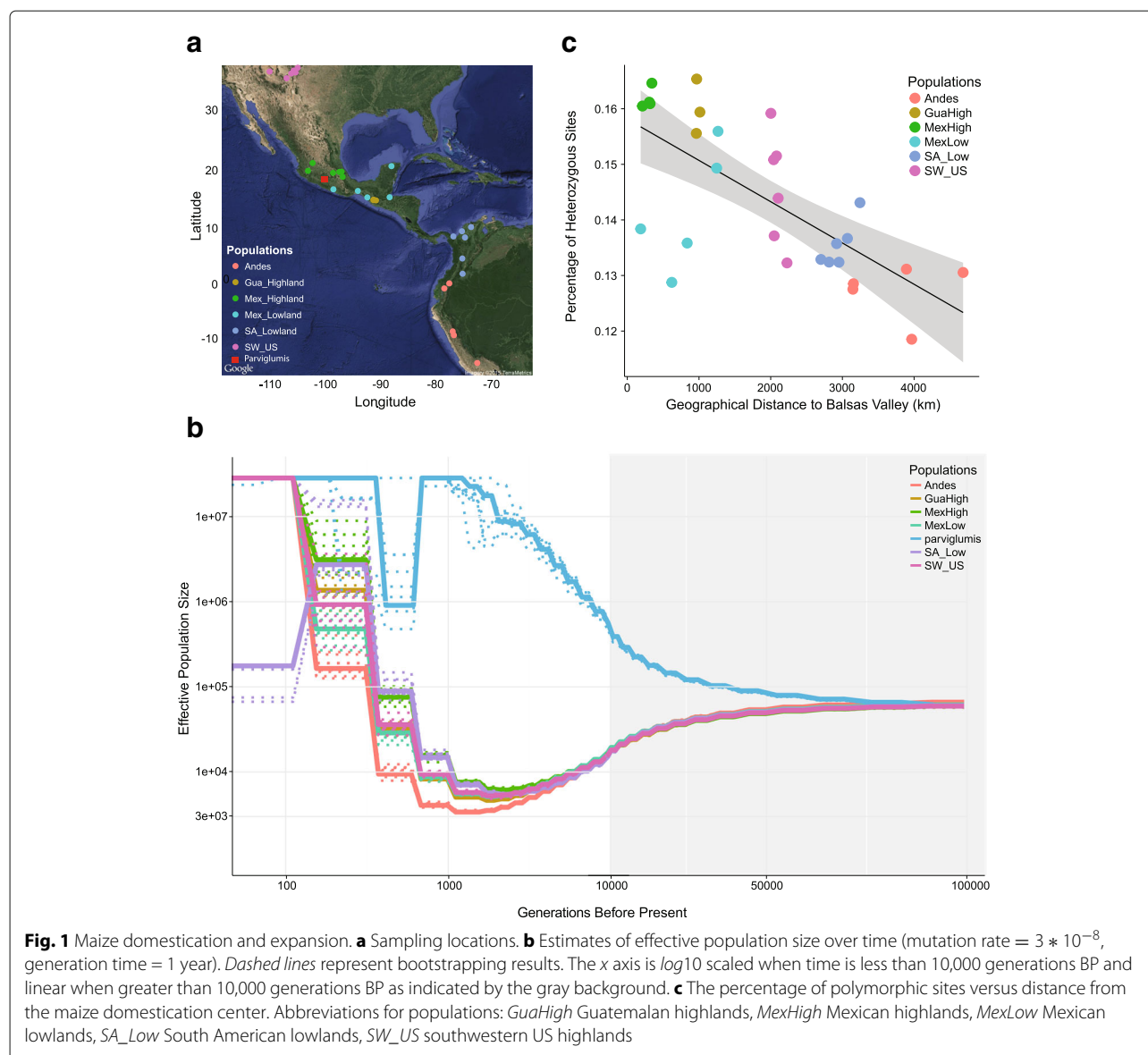
Here, we investigate the genome-wide effects of demographic change in maize during domestication and subsequent expansion using high-depth resequencing data from a panel of maize landraces. We present evidence for a protracted domestication bottleneck, further loss of diversity during crop expansion, and gene flow between maize and its wild relatives outside of its center of origin. We then explore how this demographic history has shaped genome-wide patterns of deleterious alleles.

Results

Maize population size change during domestication and expansion

We resequenced 31 maize individuals, each from one open-pollinated landrace, representing six geographical regions that span the pre-Columbian range of maize cultivation (southwestern US highlands, 6 individuals; Central Mexican Plateau, 6 individuals; Mexican lowlands, 5 individuals; Guatemalan highlands, 3 individuals; South American lowlands, 6 individuals; Andes, 5 individuals). In addition, we resequenced four wild *parviglumis* individuals from a single population located in the Balsas River Valley in Mexico (Fig. 1a). The median sequencing depth was 29X, with a range of 24–53X, resulting in a data set consisting of 49,508,640 single nucleotide polymorphisms (SNPs). Landrace accessions were selected to broadly reflect the diversity of maize in the Americas and to be representative of defined ecogeographic regions based on consultation with experts on landrace germplasm (Major Goodman, personal communication) and on descriptions in the Races of Maize handbooks [30].

We first estimated historical changes in effective population size (N_e) of maize and *parviglumis* using the multiple sequentially Markovian coalescent (MSMC) [31]. Consistent with archaeological evidence [21], we find that the demographic histories of the various maize populations begin to diverge from one another approximately 10,000 years BP (Fig. 1b). Surprisingly, our single population of *parviglumis* diverges from maize much earlier, around 75,000 years BP. All maize populations show a gradual decline in diversity concomitant with divergence from *parviglumis*, but the slope becomes more pronounced around the time of domestication. This period of declining N_e continues until the recent past ($\approx 1100 - 2400$ years BP) and is followed by extremely rapid population growth, suggesting recovery from domestication post-dated expansion of maize across the Americas. In contrast to our results in maize, *parviglumis* shows an increase in N_e which also lasts until the recent past ($\approx 1200 - 1800$ years BP). To determine if linked selection associated with domestication could bias estimates of N_e in maize (see [32]), we masked previously identified domestication candidates [24] and observed nearly identical results (Additional file 1: Figure S1A).

Wang *et al. Genome Biology* (2017) 18:215

Page 3 of 13

Fig. 1 Maize domestication and expansion. **a** Sampling locations. **b** Estimates of effective population size over time (mutation rate $= 3 * 10^{-8}$, generation time = 1 year). *Dashed lines* represent bootstrapping results. The *x* axis is *log*10 scaled when time is less than 10,000 generations BP and linear when greater than 10,000 generations BP as indicated by the gray background. **c** The percentage of polymorphic sites versus distance from the maize domestication center. Abbreviations for populations: *GuaHigh* Guatemalan highlands, *MexHigh* Mexican highlands, *MexLow* Mexican lowlands, *SA_Low* South American lowlands, *SW_US* southwestern US highlands

One explanation for the prolonged population size reduction in maize following the onset of domestication would be repeated colonization bottlenecks during the spread of maize across the Americas. Genome-wide levels of heterozygosity across our maize samples are consistent with this idea, showing a strong negative correlation ($R^2 = 0.3636, p = 0.0004$; Fig. 1c) with distance from the center of maize domestication in the Balsas River Basin. To confirm this trend, we performed a similar analysis with a much larger sample of published genotyping data ($n = 3520$; Additional file 1: Figure S1B) [33] and observed similar results.

While the gradual decrease in genetic diversity seen with distance from the Balsas indicates serial founder effects, our analyses also point to a more extreme founder event in the Andean highlands of South America. Andean landraces show a deeper bottleneck in our MSMC analysis (Fig. 1b), have the lowest overall diversity (Additional file 1: Figure S2), and show both a distinct reduction of low frequency alleles and a greater proportion of derived homozygous alleles compared to other populations (Additional file 1: Figure S2). To shed light on the timing of this extreme founder event, we assessed evidence for recent inbreeding. Inbreeding coefficients in Andean samples were quite low and not statistically different from other populations (all $F < 0.002$ and $p > 0.05$ based on a Wilcoxon test). Likewise, no significant difference could be found across populations in the number of runs of homozygosity (ROHs) longer than 1 cM ($p > 0.05$ in all cases, Wilcoxon test). Using simple

Wang *et al. Genome Biology* (2017) 18:215

Page 4 of 13

conversions between generations and the genetic length of an inherited region in the genome [34], these results provide further evidence for limited recent (< 50 generations) inbreeding in the Andes. However, when ROHs were limited to those shorter than $0.05cM$ and longer than $0.005cM$ (inbreeding from approximately 1000–10,000 generations in the past), Andean samples demonstrated significantly greater cumulative ROHs compared to all ($p < 0.05$, Wilcoxon test) but the South American lowland population ($p = 0.165$, Wilcoxon test; Additional file 1: Figure S3). Together, these lines of evidence are consistent with an unusually strong founder event during colonization of the Andes.

Introgression from wild maize in highland populations

Adaptive introgression from the wild teosinte taxon *Zea mays* ssp. *mexicana* (hereafter, *mexicana*) has previously been observed in maize in the highlands of Mexico [5]. Our broad sampling allowed us to investigate whether introgressed *mexicana* haplotypes have spread to highland maize populations outside of Mexico, potentially playing a role in adaptation in other regions. In order to test this hypothesis, we calculated Patterson's D statistic [35] across all maize populations. All individuals from both the Mexican and Guatemalan highlands exhibited highly significant evidence for shared ancestry with *mexicana* (Additional file 1: Figure S4). Maize from the southwestern USA also showed more limited evidence of introgression, consistent with findings from ancient DNA suggesting this region was originally colonized by admixed maize from the highlands of Mexico [36]. In contrast, the distribution of z-scores for South American populations overlapped zero, providing no evidence for substantial spread of *mexicana* haplotypes to this region.

We localized introgression to chromosomal regions through genome-wide calculation of the $\hat{f}_d$ statistic [37]. Megabase-scale regions of introgression were identified in both Mexican and Guatemalan highland populations that correspond to those reported by [5] on chromosomes 4 and 6 (Fig. 2; Additional file 1: Figure S5). On chromosome 3 (at around $75 - 90$ Mb), a large, previously unidentified region of introgression can be found in the Mexican and southwestern US highlands (Fig. 2; Additional file 1: Figure S5). This region overlaps a putative chromosomal inversion associated with flowering time in maize landraces [38] and in the maize nested association mapping population [39] and may be an example of *mexicana* contribution to modern maize lines.

### The influence of demography on accumulation of deleterious alleles

Population-specific changes in historical N_e should influence the efficiency of purifying selection and alter genome-wide patterns of deleterious variants [10].

Introgression from a species with substantially different N_e may also influence the abundance and distribution of deleterious alleles in the genome [9, 19]. Below we evaluate the effects of major demographic events during the pre-Columbian history of maize on patterns of deleterious alleles.

Domestication and deleterious alleles

We first compared counts of deleterious alleles in Mexican lowland maize individuals to four *parviglumis* individuals from a single population in the Balsas River Valley. Maize from the Mexican lowlands has not experienced substantial introgression from wild relatives and is near the center of maize origin [22], and thus best reflects the effects of domestication alone. After identifying putatively deleterious mutations using Genomic Evolutionary Rate Profiling (GERP) [40], we calculated the number of derived deleterious alleles per genome under both an additive and a recessive model across four levels of mutation severity (see Methods for details). Maize showed significantly more deleterious alleles than teosinte under both additive (< 10% more; $p = 0.0079$, Wilcoxon test; Additional file 1: Figure S6) and recessive (< $20 - -30$% more; $p = 0.0079$; Fig. 3) models across all categories (Additional file 1: Figure S7). Additionally, maize contained more than twice as many fixed deleterious alleles than teosinte (57,881 versus 26,947) and 10% fewer segregating deleterious alleles (429,837 versus 478,594), effects expected under a domestication bottleneck (Fig. 3c; [7]). GERP load (GERP score $\times$ frequency of deleterious alleles), a more direct proxy of mutation load quantified at the population level, revealed a similar trend (additive model: maize median = 23.635, teosinte median = 22.791, $p = 0.008$, Wilcoxon test; recessive model: maize median = 14.922, teosinte median = 12.231, $p = 0.008$). Maize, like other domesticates [12, 14, 41, 42], thus appears to have a higher mutation load compared to its wild progenitor *parviglumis*.

While the elevated mutation load we observe in maize relative to *parviglumis* may be driven primarily by the domestication bottleneck, positive selection on causal variants underlying domestication phenotypes may also fix nearby deleterious variants through genetic hitchhiking, which would result in a higher number of deleterious alleles in regions linked to domestication loci [41, 43]. To test this hypothesis, we first confirmed that 420 previously identified domestication candidates [24] showed evidence of selection in our data (Additional file 1: Figure S8), and then assessed the distribution of deleterious alleles in and near (5 kb upstream and downstream) these genes by calculating the number of deleterious alleles per base pair under both recessive and additive models. No significant difference was found in the prevalence of deleterious alleles near domestication and random sets of genes (Additional file 1: Figure S9), suggesting the increased

Wang *et al. Genome Biology* (2017) 18:215

Page 5 of 13
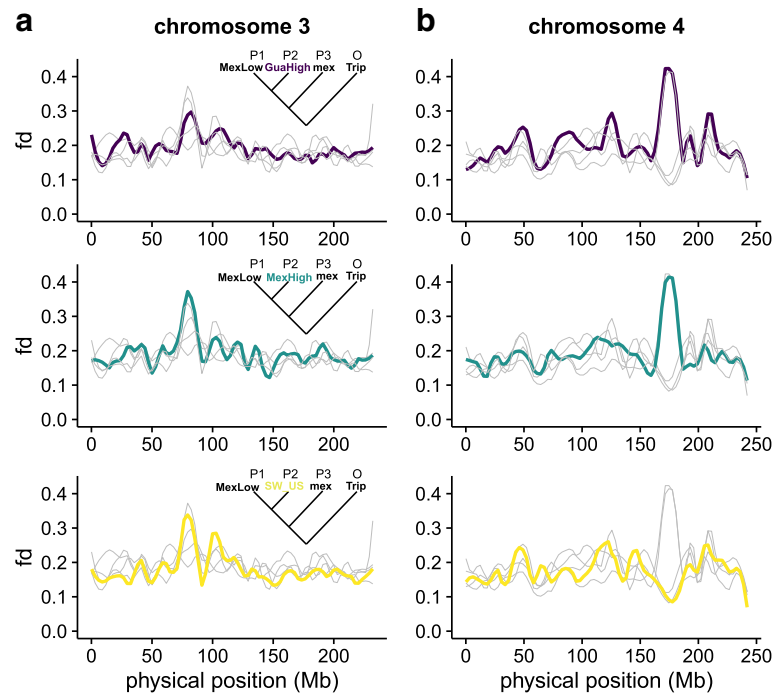
Fig. 2 Introgression from *mexicana* into maize landraces. Loess regression of $\hat{f}_d$ is plotted for all five populations on **a** chromosome 3 and **b** chromosome 4. Each plot highlights a single population, with other populations shown in *gray*. The Mexican lowlands population is used as a reference and thus not plotted. No significant introgression was detected in the South American lowlands or the Andes, and loess regressions for these populations are only shown as *gray lines*. The statistic $\hat{f}_d$ was calculated based on the tree in which P2 is varied across populations. *mex mexicana*, *Trip Tripsacum*

mutation load we observe in maize has been driven primarily by the genome-wide effects of the domestication bottleneck rather than linkage associated with selection on specific genes.

### The effect of the Andean founder event on deleterious alleles

The extreme founder event observed in the Andes could potentially alter genome-wide patterns of deleterious variants beyond the effects of domestication alone. Under a recessive model, maize from the Andes contains significantly more deleterious alleles than any other population (Fig. 3b; Additional file 1: Figure S7; all *p* values < 0.02, Wilcoxon test), and this difference becomes more extreme when considering more severe (i.e., higher GERP score) mutations (Additional file 1: Figure S7). In contrast, we observe no significant difference under an additive model (Additional file 1: Figure S6; Additional file 1: Figure S7). The Andean founder event therefore appears to have resulted in higher mutation load than seen in other maize populations. This result is further supported by a higher proportion of fixed deleterious alleles within the Andes and fewer segregating deleterious alleles (Additional file 1: Figure S10; Fig. 3d), a result comparable to the differences observed between maize and *parviglumis*.

Introgression decreases the prevalence of deleterious alleles

Highly variable rates of *mexicana* introgression were detected across our landrace populations (Fig. 2; Additional file 1: Figure S4; Additional file 1: Figure S5). To explore the potential effects of introgression on the genomic distribution of deleterious alleles, we fit a linear model in which the number of deleterious sites is predicted by introgression (represented by $\hat{f}_d$) and gene density (exonic base pairs per centimorgan) in 10-kb non-overlapping windows in the Mexican highland population where we found the strongest evidence for *mexicana* introgression. Gene density was included as a predictor in the regression to control for the positive correlation observed between gene density and both introgression ($p = 3.48e - 08$) and deleterious alleles ($p \approx 0$). When identifying deleterious alleles under both additive and recessive models, we found a strong negative correlation with introgression (i.e., fewer deleterious alleles in introgressed regions; $p \approx 0$ under both models). These findings likely reflect the larger ancestral N_e and more efficient purifying selection in *mexicana*.

Discussion

Demographic studies in domesticated species have focused largely on identifying progenitor population(s)

Wang *et al. Genome Biology* (2017) 18:215

Page 6 of 13
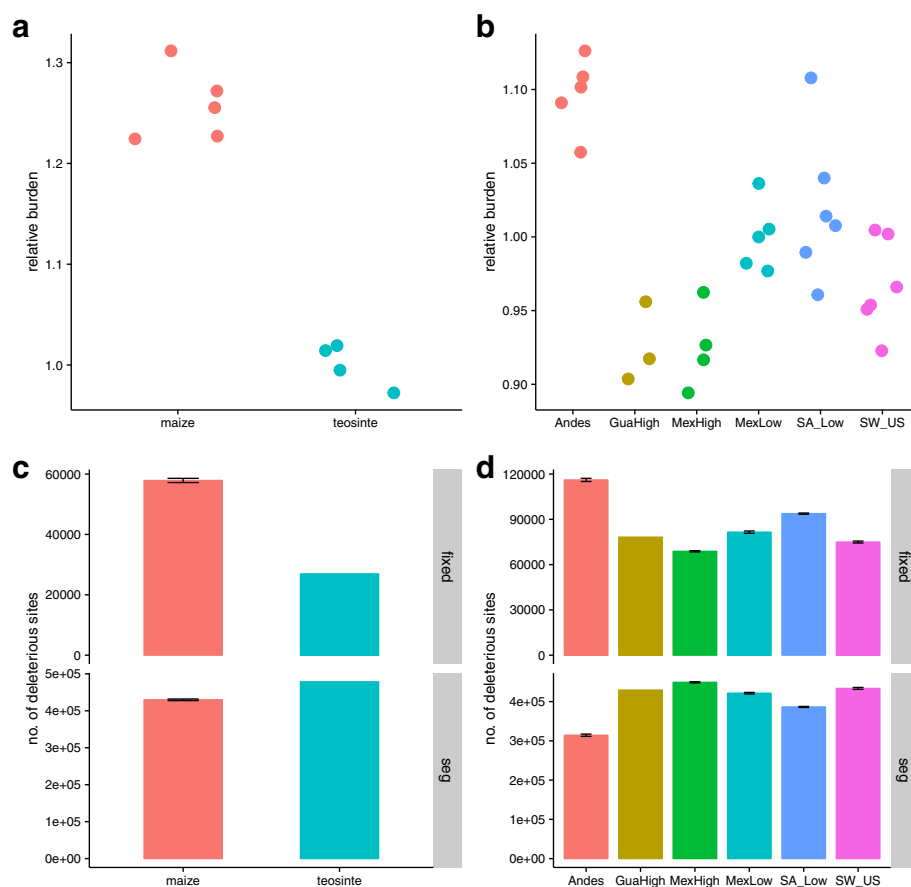
Fig. 3 Burden of deleterious mutations during maize domestication and expansion. Comparison of counts of deleterious alleles at the individual level **a** between *parviglumis* and maize (mean value in *parviglumis* population was used as the standard to calculate the relative burden) and **b** among maize populations (mean value in Mexican lowland population was utilized as the standard to calculate the relative burden) under a recessive model. Comparison of fixed versus segregating (seg) deleterious alleles at the population level **c** between *parviglumis* and maize and **d** among maize populations. A jackknife sub-sampling approach ($n = 4$) was utilized for maize in **c** and for individual maize populations ($n = 3$) in **d**

and quantifying the effect of the domestication bottleneck on genetic diversity [24, 44, 45]. It is likely, however, that the demographic history of domesticates is generally more complex than a simple bottleneck followed by recovery [46, 47]. Many crops and domesticated animals have expanded from defined centers of origin to global distributions, experiencing population size changes and gene flow from closely related taxa throughout their histories [48]. With this in mind, we have characterized maize demography from domestication through initial expansion in order to provide a more complete assessment of the influence of demography on deleterious variants.

Historical changes in maize population size

Early models of maize demography suggested the ratio of the domestication bottleneck size and duration was between $\approx 2.5 : 1$ and $\approx 5 : 1$, but little statistical support was found for specific estimates of these individual parameters [23, 28, 49]. Most recently, Beissinger et al. [2] fit a model assuming a bottleneck followed by instantaneous exponential recovery. While our results concur with the most recent model in finding a similar bottleneck size ($\approx 10\%$ compared to $\approx 5\%$ in Beissinger et al.) and that the modern N_e of maize is quite large, the flexibility of MSMC also allowed us to estimate the duration of the bottleneck. We find that the domestication bottleneck may have lasted much longer than previously believed, spanning ≈ 9000 generations and only beginning to recover in the recent past (Fig. 1b). Analysis of bottlenecks during African rice and grape domestication have also suggested a duration of several thousand generations [46, 47], indicating that demographic bottlenecks during crop evolution may have generally occurred over substantial periods of time. Previous work has suggested population structure can generate spurious signals of population size change in methods like MSMC [50, 51], such that individuals sampled from a

Wang *et al. Genome Biology*   (2017) 18:215

Page 7 of 13

single deme of a highly structured population can falsely demonstrate signatures of a population bottleneck similar to what we observe in maize [51]. Given that our maize landraces are sampled from broad ecogeographic regions, however, this effect should be minimal. Moreover, a similar analysis in an Americas-wide sample of maize landraces demonstrated qualitatively similar results [2].

In addition to a strong bottleneck during domestication, our finding that levels of diversity decline in populations increasingly distant from the center of maize domestication are suggestive of serial founder effects during the spread of maize across the Americas (Fig. 1c; Additional file 1: Figure S1). Serial founder effects are the result of multiple sampling events during which small founder populations are repeatedly drawn from ancestral pools, leading to a stepwise increase in genetic drift and a concomitant decrease in genetic diversity. During maize range expansion, serial founder effects would have occurred if seed carried to each successive colonized region was limited to a sample of whole ears that contained a fraction of the diversity found in the source population [29]. Movement of entire ears involves a collective transfer of seeds that are either full or maternal half-siblings and could lead to more substantial founder effects than would be seen if dispersal were truly random. Such "kin-structured" migration, which is common in nature, has theoretically been demonstrated to increase inbreeding due to a reduction in the number of effective colonists [52]. Consistent with serial founder effects, other researchers have found a correlation between geographic and genetic distance in maize landraces [22, 53], though this was previously attributed to limited migration across the species range leading to isolation by distance (IBD). Neutral expectations of allele frequencies across populations under serial founder effects differ substantially from those predicted under equilibrium conditions [15]. For example, Slatkin and Excoffier [15] have demonstrated that allele frequency clines previously attributed to adaptation could be generated entirely by neutral processes under expansion. Many of the world's crops have experienced such histories of expansion, and studies attempting to identify loci underlying crop adaptation during post-domestication spread to new environments may most accurately compare empirical data to neutral expectations under a serial founder effects demography [15].

While a history of serial founder effects partially explains the variation in diversity across maize landraces, there are deviations from this model. For example, our combined results showing increased ROHs (Additional file 1: Figure S3), lower nucleotide diversity (Additional file 1: Figure S2), and smaller effective population size (Fig. 1) in the Andes all suggest a pronounced, ancient founder event and are in agreement with previous work modeling demography in this region [54]. The founder

event in the Andes may reflect initially limited cultivation due to the poor performance of maize in this region relative to established root and tuber staples [55]; maize cultivation may have only become widespread after an initial lag period necessary for adaptation. Additionally, we observe somewhat higher than expected nucleotide diversity in maize landraces from the highlands of Mexico and Guatemala (Fig. 1c), which may be linked to the introgression we have detected from *mexicana*.

In striking contrast to the bottleneck we observe in maize, the effective population size in *parviglumis* increases steadily from the time of initial maize domestication until the recent past. Multiple population genetic studies have reported negative genome-wide values of Tajima's D in *parviglumis* from the Balsas River Valley [2, 23, 56], findings characteristic of an expanding population. Likewise, analyses of pollen content in sediment cores from Mexico suggest herbaceous vegetation and grasslands have expanded over the last 10,000 years due to changing environmental conditions during the Holocene and human management of vegetation with fire [57, 58]. While our *parviglumis* samples are drawn from a single population in the Balsas, these data collectively suggest *parviglumis* from this region has experienced expansion over the last several millennia.

Consistent with archaeological evidence of the timing of initial maize domestication [21], we find that maize demographies begin to diverge $\approx 10,000$ generations BP, a time that appears to coincide with a steeper decline in maize N_e as well. In contrast, we estimate the timing of the split between maize and our single population of *parviglumis* to be $\approx 75,000$ generations BP, potentially reflecting population structure in *parviglumis*. Beissinger et al. [2], using samples from additional populations, also find an estimate of maize-*parviglumis* divergence older than the probable onset of domestication, suggesting that currently available sequences of *parviglumis* may not sample well from the populations directly ancestral to domesticated maize.

The prevalence of gene flow during maize diffusion

Increasingly, range-wide analyses of crops and their wild relatives have identified evidence of gene flow during post-domestication expansion from newly sympatric populations of their progenitor taxa and closely related species [59–61]. Consistent with previous results from genotyping data [5, 22, 62], we find strong support for introgression from *mexicana* to maize in the highlands of Mexico. While *mexicana* is not currently found in the highlands of Guatemala, we also find strong evidence for *mexicana* introgression in maize from this region, suggesting either *mexicana* was at one time more broadly distributed, or, perhaps more likely, that highland maize from Mexico was introduced to the Guatemalan highlands.

Wang *et al. Genome Biology* (2017) 18:215

Page 8 of 13

Support is also found for *mexicana* introgression in the southwestern USA at specific chromosomal regions such as a putative inversion polymorphism on chromosome 3 (Fig. 2). These results confirm previous findings suggesting maize from the highlands of Mexico originally colonized the southwestern USA [36]. The more limited signal of *mexicana* introgression here may be due to subsequent gene flow from lowland maize as suggested by [36]. Very little evidence is found for *mexicana* haplotypes extending into South America, as highland-adapted haplotypes would likely have been maladaptive and removed by selection as maize traversed the lowland regions of Central America [54].

Impacts of demography on accumulation of deleterious variants

Previous work in maize has characterized genome-wide trends in deleterious alleles across modern inbred maize lines, revealing that inbreeding during the formation of modern lines has likely purged many recessive deleterious variants [63] and that complementation of deleterious alleles likely underlies the heterosis observed in hybrid breeding programs [63, 64]. Additionally, [2] revealed that purifying selection has removed a greater extent of pairwise diversity (θ_π) near genes in *parviglumis* than in maize due to the higher historical N_e in *parviglumis*, but that this trend is reversed when considering younger alleles due to the recent dramatic expansion in maize population size. To date, however, few links have been made between the historical demography of maize domestication and expansion and the prevalence of deleterious alleles. Our analysis reveals, for the first time, that demography has played a pivotal role in determining both the geographic and genomic landscapes of deleterious alleles in maize.

Population size and deleterious variants

Previous studies have suggested a "cost of domestication" in which a higher burden of deleterious alleles is found in domesticates compared to their wild progenitors [12, 41, 43, 65, 66]. Consistent with these results, we detect an excess of deleterious alleles in maize relative to *parviglumis* (Fig. 3; Additional file 1: Figure S6; Additional file 1: Figure S7), which could be caused by two potential factors. First, reduced population size during the domestication bottleneck could result in deleterious alleles drifting to higher allele frequency. Second, hitchhiking caused by strong positive selection on domestication genes could cause linked deleterious alleles to rise in frequency [12, 65]. While we find support for the former in maize, we see little evidence of the latter. Recent studies have reported contrasting results regarding the effect of selective sweeps in patterning the distribution of deleterious alleles. For example, putative selective sweeps in cassava showed a paucity of deleterious alleles, a result

that was attributed to purifying selection [67]. Sweep regions in grape exhibited an overall decrease in the number of deleterious alleles but an increase in the ratio of deleterious mutations to synonymous variants, a pattern suggesting deleterious alleles may have hitchhiked along with the targets of positive directional selection [46]. Finally, selective sweeps in Asian rice contained a roughly equivalent ratio of deleterious mutations to synonymous mutations when compared to neutral regions [68]. Clearly, further exploration is warranted to clarify the effect of selection on the distribution of deleterious mutations. In addition to the cost of domestication, we find a cost of geographic expansion that is likely driven by serial founder effects. The increase in deleterious alleles during expansion is most pronounced in the Andes and may be symptomatic of the extreme founder event we propose above.

Differences in the number of deleterious alleles between maize and *parviglumis* and non-Andean and Andean maize are more dramatic under a recessive model than an additive model. This trend may indicate that the bulk of deleterious alleles in maize are at least partially recessive, such that heterozygous sites contribute less to a reduction in individual fitness. Previous work in human populations has shown that the majority of deleterious mutations are recessive or partially recessive [69], and data from knock-out mutations in yeast have revealed that large-effect mutations tend to be more recessive [70]. Likewise, both theory and empirical evaluation across a number of organisms suggest that mildly deleterious mutations are likely to be partially recessive [71]. In maize, Yang et al. [63] have found that most deleterious alleles are at least partially recessive and note a negative correlation between the severity of a deleterious variant and its dominance. Our results thus match nicely both with previous empirical data and theoretical expectations showing that recent population bottlenecks should only show strong differences in load under a recessive model [7].

Introgression and deleterious variants

Very few studies have investigated the effects of introgression from a taxon with substantially different N_e on the genomic landscape of deleterious variants. The best example is found in the human literature, where confirmation has been found that introgression from Neanderthals with low ancestral N_e increased the overall mutation load in modern humans [9, 19]. We report here the opposite pattern in maize, as introgression appears to have reduced the proportion of deleterious variants. Nonetheless, the underlying interpretation is similar: the taxon donating alleles *mexicana* has had a larger ancestral N_e than maize [27], and introgressed haplotypes have thus experienced more efficient long-term purging of deleterious alleles.

Wang *et al. Genome Biology* (2017) 18:215

Page 9 of 13

Conclusions

We have demonstrated that demography during the domestication and expansion of maize across the Americas has profoundly influenced putative functional variation across populations and within individual genomes. More generally, we have learned that population size changes and gene flow from close relatives with contrasting effective population size will influence the distribution of deleterious alleles in species undergoing rapid shifts in demography. The significance of our results extends far beyond maize. For example, invasive species that have recently experienced founder events followed by expansion, endangered species subject to precipitous declines in N_e, species with a history of post-glacial expansion, and new species expanding their range will all likely show clear genetic signals of the interplay between demography and selection. This interaction bears importantly on the adaptive potential of both individual populations and species. By fully characterizing this relationship, we can better understand how the current evolutionary trajectory of a species has been influenced by its history.

Methods

Samples, whole genome resequencing, and read mapping

A total of 31 maize landrace accessions were obtained from the US Department of Agriculture (USDA)'s National Plant Germplasm System and through collaborators (Additional file 2: Table S1). Samples were chosen from four highland populations (Andes, Mexican highlands, Guatemalan highlands, and southwestern US highlands) and two lowland populations (Mexican and South American lowlands) (Fig. 1a). In addition, four open-pollinated *parviglumis* samples were selected from a single population in the Balsas River Valley in Mexico. DNA was extracted from leaves using a standard cetyltrimethyl ammonium bromide (CTAB) protocol [72]. Library preparation and Illumina HiSeq 2000 sequencing (100-bp paired-end) were conducted by BGI (Shenzhen, China) following their established protocols. the Burrows-Wheeler Aligner (BWA) v.0.7.5.a [73] was used to map reads to the maize B73 reference genome v3 (GenBank BioProject PRJNA72137) [74] with default settings. The duplicate molecules in the realigned bam files were removed with MarkDuplicates in Picardtools v.1.106 (http://broadinstitute.github.io/picard), and indels were realigned with the Genome Analysis Toolkit (GATK) v.3.3-0 [75]. Sites with mapping quality less than 30 and base quality less than 20 were removed, and only uniquely mapped reads were included in downstream analyses.

Demography of maize domestication and diffusion

The MSMC method [31], which models ancestral relationships under recombination and mutation and has been used in several plant species [46, 47], was utilized to infer effective population size changes in both *parviglumis* and maize. SNPs were called via HaplotypeCaller and filtered via VariantFiltration in GATK [75] across all samples. SNPs with the following metrics were excluded from the analysis: QD <2.0; FS >60.0; MQ <40.0; MQRankSum < −12.5; ReadPosRankSum < −8.0. Vcftools v.0.1.12 [76] was used to further filter SNPs to include only bi-allelic sites. Following these data filtering steps, our data set consisted of 49 million SNPs. SNPs were phased using BEAGLE v.4.0 [77] with SNP data from the maize HapMap2 panel [78] used as a reference. Only sites with depth between half and twice of the mean depth were included in analyses. In addition, the software SNPable (http://lh3lh3.users.sourceforge.net/snpable.shtml) was used to mask genomic regions in which reads were not uniquely mapped. The mappability mask file for MSMC was generated by stepping in 1-bp increments across the maize genome to generate 100-bp single-end reads, which were then mapped back to the maize B73 reference genome [74]. Sites with the majority of overlapping 100-mers mapped uniquely without mismatch were determined to be "SNPable" sites and used for the MSMC analyses. For effective population size inference in MSMC, we used $5 \times 4 + 25 \times 2 + 5 \times 4$ as the pattern parameter, and the value m was set as half of the heterozygosity in *parviglumis* and maize populations, respectively.

In order to explore the trend of genetic diversity away from the domestication center, the correlation between the percentage of polymorphic sites and the geographic distance to the Balsas River Valley (latitude 18.099138, longitude −100.243303) was examined via linear regression. Geographical distance in kilometers was calculated based on great circle distance using the haversine transformation [17]. The correlation between percentage of heterozygous sites and distance away from domestication center was also explored in the SeeDs data set. SNPs with more than 50% missing samples and samples with more than 50% missing genotypes were removed from the SeeDs data set.

Population structure, genetic diversity, and inbreeding coefficients

We first evaluated population structure using principal component analysis (PCA) with ngsCovar [79] in ngsTools [80] based on the matrix of posterior probabilities of SNP genotypes produced in Analysis of Next Generation Sequencing Data (ANGSD) v.0.614 [81], and then utilized NGSadmix v.32 [82] to investigate the admixture proportion of each accession. The NGSadmix analysis was conducted based on genotype likelihoods for all individuals, which were generated with ANGSD (options -GL 2 -doGlf 2 -SNP_pval $1e − 6$), and K from 2 to 10 was set to run the analysis for sites present in a minimum of 77%

Wang *et al. Genome Biology* (2017) 18:215

Page 10 of 13

of all individuals (24 in 31). A clear outlier in the Mexican highland population was detected, RIMMA0677, a sample from relatively low altitude, which was suspected to contain a divergent haplotype. A neighbor-joining tree of SNPs within an inversion polymorphism on chromosome 4 that includes a diagnostic highland haplotype [5] was constructed with the R package phangorn [83]. The sample RIMMA0677 was not clustered with other highland samples, but embedded within lowland haplotypes (Additional file 1: Figure S11), so it was removed from further analyses.

The genetic diversity measures Watterson's θ and θ_π were calculated in ANGSD [81] for each population. The neutrality test statistic Tajima's D was calculated with an empirical Bayes approach [84] implemented in ANGSD by first estimating a global site frequency spectrum (SFS) then calculating posterior sample allele frequencies using the global SFS as a prior. The three statistics were summarized across the genome using 10-kb non-overlapping sliding windows.

Inbreeding coefficients for each individual were estimated with ngsF [85] with initial values of F_{IS} set to be uniform at 0.01 with an epsilon value of $1e-5$.

In addition, SNPs were polarized using the *Tripsacum dactyloides* genome to assess the frequency of derived homozygous sites in each maize landrace population. *T. dactyloides* short reads were downloaded from the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) database (SRR447804–SRR447807), mapped to the B73 reference genome v3 [74] with BWA [73], and incorporated into SNP calling as described above.

Runs of homozygosity
SNPs were down-sampled to contain one SNP in a 2-kb window to identify segments representing homozygosity by descent (i.e., autozygosity) rather than by chance. PLINK v.1.07 [86] was applied to identify segments of ROHs in a window containing 20 SNPs, among which the number of the maximum missing SNPs was set to 2 and the number of the maximum heterozygous sites was set to 1. The shortest length of final ROHs was set to be 300 kb. Physical distances were converted into genetic distances based on a recent genetic map [87].

Detection of introgression
To assess per-genome evidence of population admixture between maize landraces and teosinte, we calculated the D statistic using ANGSD [81]. The statistic was calculated using trees of the form $(((X, \text{low}), mexicana), T. dactyloides)$. One accession from the Mexican lowland population was randomly sampled as the "low" taxon, and each sample from all other populations except the Mexican lowland was set as "X". The *mexicana* accession

TIL25 from the maize HapMap2 project [78] was treated as the third column species. The D statistic was calculated in a 1-kb block, and then jackknife bootstrapping was conducted to estimate significance.

In addition, the $\hat{f}_d$ statistic [37] was calculated based on a similar tree form $(((P_1, P_2), P_3), O)$, but using allele frequencies across multiple individuals for each position on the tree. We fixed P_1 as the Mexican lowland population, P_3 as two lines of *mexicana* (TIL08 and TIL25), and *T. dactyloides* as the outgroup. P_2 was set to each of the four highland populations and the South American lowland population.

The $\hat{f}_d$ statistic was calculated in 10-kb non-overlapping windows across the genome with the python script egglib_sliding_windows.py (https://github.com/johnomics/Martin_Davey_Jiggins_evaluating_introgression_statistics), which makes use of the EggLib library [88]. The input file was generated by first identifying genotypes using ANGSD (-doMajorMinor 1 -doMaf 1 -GL 2 -doGeno 4 -doPost 1 -postCutoff 0.95 -SNP_pval $1e-6$) followed by format adjustments with a custom script (see "Availability of data and materials"). Outliers were detected by setting the 95% quantile of the $\hat{f}_d$ distribution in the South American lowland population as the cutoff.

Estimating burden of deleterious mutations
We estimated the individual burden of deleterious alleles based on GERP scores [89] for each site in the maize genome, which reflects the strength of purifying selection based on constraint in a whole genome alignment of 13 plant species [90]. The alignment and estimated GERP scores are available at iplant (https://doi.org/10.7946/P2WS60). Scores above 0 may be interpreted as historically subject to purifying selection, and mutations at such sites are likely deleterious. We identified *Sorghum bicolor* alleles in the multiple species alignment as ancestral and defined the non-*Sorghum* allele as the deleterious allele. Only biallelic sites were included for our evaluation. Inclusion of the maize B73 reference genome when calculating GERP scores [90] introduces a bias toward underestimation of the burden of deleterious alleles in maize versus teosinte populations. Therefore, we corrected the GERP scores of sites where the B73 allele is derived following [7]. Briefly, we divided SNPs where the B73 allele is ancestral into bins of 1% derived allele frequency based on maize HapMap3 [91] and used this frequency distribution to estimate the posterior probability of GERP scores for SNPs where the B73 allele is derived.

The sum of GERP scores multiplied by deleterious allele frequency for each SNP site was used as a proxy of individual burden of deleterious alleles under an additive model ($HET * 0.5 + HOM * 1$). This burden was calculated under a recessive model as the sum of GERP scores multiplied

Wang *et al. Genome Biology* (2017) 18:215

Page 11 of 13

by one for each deleterious homozygous site ($HOM * 1$). For a better understanding of the variation of individual burden among sites under varied selection strength, we partitioned the deleterious SNPs into four categories ($-2 <$ GERP ≤ 0, nearly neutral; $0 <$ GERP ≤ 2, slightly deleterious; $2 <$ GERP ≤ 4, moderately deleterious; GERP > 4, strongly deleterious) and recapitulated the preceding statistics.

Additional files

Additional file 1: Figure S1. Demography of maize populations. A. MSMC results before and after masking candidate regions under selection during domestication. B. Percentage of heterozygous sites versus distance from the Balsas Valley in 3520 samples from the SeeDs data set. **Figure S2.** Boxplot of multiple population genetic statistics. Watterson's *theta* (A), θ_π (B), and Tajima's *D* (C) are based on values in 10-kb non-overlapping windows across the genome. Percentage of derived homozygous sites was calculated for each individual and reported per population (D). **Figure S3.** Cumulative length of ROHs in cM across populations. *Lines* indicate median values in each population. *ROH* runs of homozygosity. **Figure S4.** Calculation of *D* statistic across populations. Evidence of introgression from *mexicana* into Mexican highland, Guatemalan highland, and southwestern US highland maize populations. The *dashed lines* correspond to *Z*-scores equal to −10 and 10. **Figure S5.** $\hat{f}_d$ statistic results. Loess regression of $\hat{f}_d$ in 10-kb nonoverlapping windows across all chromosomes. **Figure S6.** Relative burden of deleterious alleles under additive model between maize and teosinte (A; mean value in teosinte population was used as the standard to calculate the relative burden) and among maize populations (B; mean value in Mexican lowland population was utilized as the standard to calculate the relative burden). **Figure S7.** Relative burden of deleterious alleles under both additive and recessive models with different Genomic Evolutionary Rate Profiling (GERP) partitions between maize and teosinte (A; mean value in teosinte population was used as the standard to calculate the relative burden) and among maize populations (B; mean value in Mexican lowland population was utilized as the standard to calculate the relative burden). **Figure S8.** Domestication candidate genes exhibited lower θ_π ratio between maize and teosinte, a signal of selection in these genes. Distribution of ratio of θ_π between maize and teosinte in 420 domestication candidate genes (mean value is indicated with *red line*) against 10,000 replicates of genome-wide sampling of 420 random genes. **Figure S9.** Distribution of number of deleterious sites per bp in 420 domestication candidate genes (indicated with *blue line*) compared to genome-wide random samples under an (A) additive model and (B) recessive model. **Figure S10.** Site frequency spectrum (SFS) of deleterious SNPs in five populations. GuaHigh is not included since the small sampling limited power for the SFS. **Figure S11.** Neighbor-joining tree of SNPs from an inversion on chromosome 4 with a diagnostic haplotype for highland Mexican material. (PDF 877 kb).

Additional file 2: Table S1. Basic information regarding the sampled maize landrace accessions. *NM* New Mexico. (XLSX 11 kb).

[92]. The whole genome sequencing (WGS) raw reads have been deposited in NCBI SRA (SRP065483) [93].

Authors' contributions
AL and CRI performed the experiments. LW, JRI, and TMB conducted the analyses. LW, MBH, and JRI wrote the manuscript. MBH and JRI planned the study. All authors read and approved the final manuscript.

Author details
[1]Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, USA. [2]Genome Informatics Facility, Iowa State University, Ames, USA. [3]Department of Plant Sciences, University of California, Davis, USA. [4]USDA-ARS Plant Genetics Research Unit, Columbia, USA. [5]Divisions of Plant and Biological Sciences, University of Missouri, Columbia, USA. [6]Genome Center and Center for Population Biology, University of California, Davis, USA.

References
1. Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, et al. Worldwide human relationships inferred from genome-wide patterns of variation. Science. 2008;319:1100–4.
2. Beissinger TM, Wang L, Crosby K, Durvasula A, Hufford MB, Ross-Ibarra J. Recent demography drives changes in linked selection across the maize genome. Nat Plants. 2016;2:16084.
3. Doebley JF, Gaut BS, Smith BD. The molecular genetics of crop domestication. Cell. 2006;127:1309–21.
4. Gignoux CR, Henn BM, Mountain JL. Rapid, global demographic expansions after the origins of agriculture. Proc Natl Acad Sci. 2011;108: 6044–9.
5. Hufford MB, Lubinksy P, Pyhäjärvi T, Devengenzo MT, Ellstrand NC, Ross-Ibarra J. The genomic signature of crop-wild introgression in maize. PLoS Genet. 2013;9:e1003477.
6. Prufer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, et al. The complete genome sequence of a Neanderthal from the Altai Mountains. Nature. 2014;505:43–9.
7. Simons YB, Turchin MC, Pritchard JK, Sella G. The deleterious mutation load is insensitive to recent population history. Nat Genet. 2014;46:220–4.
8. Do R, Balick D, Li H, Adzhubei I, Sunyaev S, Reich D. No evidence that selection has been less effective at removing deleterious mutations in Europeans than in Africans. Nat Genet. 2015;47:126–31.
9. Harris K, Nielsen R. The genetic cost of Neanderthal introgression. Genetics. 2016;203:881–91.
10. Fu W, Gittelman RM, Bamshad MJ, Akey JM. Characteristics of neutral and deleterious protein-coding variation among individuals and populations. Am J Human Genet. 2014;95:421–36.
11. Zhang M, Zhou L, Bawa R, Suren H, Holliday J. Recombination rate variation, hitchhiking, and demographic history shape deleterious load in poplar. Mol Biol Evol. 2016;33:2899–910.
12. Marsden CD, Ortega-Del Vecchyo D, O'Brien DP, Taylor JF, Ramirez O, Vilà C, et al. Bottlenecks and selective sweeps during domestication have increased deleterious genetic variation in dogs. Proc Natl Acad Sci. 2016;113:152–7.
13. Simons YB, Sella G. The impact of recent population history on the deleterious mutation load in humans and close evolutionary relatives. Curr Opin Genet Dev. 2016;41:150–8.
14. Liu Q, Zhou Y, Morrell PL, Gaut BS. Deleterious variants in Asian rice and the potential cost of domestication. Mol Biol Evol. 2017;34:908–924.
15. Slatkin M, Excoffier L. Serial founder effects during range expansion: a spatial analog of genetic drift. Genetics. 2012;191:171–81.

Wang *et al. Genome Biology*   (2017) 18:215

Page 12 of 13

16. Austerlitz F, Jung-Muller B, Godelle B, Gouyon PH. Evolution of coalescence times, genetic diversity and structure during colonization. Theor Popul Biol. 1997;51:148–64.

17. Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proc Natl Acad Sci of the U S A. 2005;102:15942–7.

18. Henn BM, Botigué LR, Bustamante CD, Clark AG, Gravel S. Estimating the mutation load in human genomes. Nat Rev Genet. 2015;16:333–43.

19. Juric I, Aeschbacher S, Coop G. The strength of selection against Neanderthal introgression. PLoS Genet. 2016;12:1–25.

20. Matsuoka Y, Vigouroux Y, Goodman MM, Sanchez GJ, Buckler E, Doebley J. A single domestication for maize shown by multilocus microsatellite genotyping. Proc Natl Acad Sci. 2002;99:6080–4.

21. Piperno DR, Ranere AJ, Holst I, Iriarte J, Dickau R. Starch grain and phytolith evidence for early ninth millennium B.P. maize from the Central Balsas River Valley. Mexico. Proc Natl Acad Sci. 2009;106:5019–24.

22. van Heerwaarden J, Doebley J, Briggs WH, Glaubitz JC, Goodman MM, Gonzalez JdJS, et al. Genetic signals of origin, spread, and introgression in a large sample of maize landraces. Proc Natl Acad Sci. 2011;108:1088–92.

23. Wright SI, Bi IV, Schroeder SG, Yamasaki M, Doebley JF, McMullen MD, et al. The effects of artificial selection on the maize genome. Science. 2005;308:1310–4.

24. Hufford MB, Xu X, Van Heerwaarden J, Pyhäjärvi T, Chia JM, Cartwright RA, et al. Comparative population genomics of maize domestication and improvement. Nat Genet. 2012;44:808–11.

25. Merrill WL, Hard RJ, Mabry JB, Fritz GJ, Adams KR, Roney JR, et al. The diffusion of maize to the southwestern United States and its impact. Proc Natl Acad Sci. 2009;106:21019–26.

26. Grobman A, Bonavia D, Dillehay TD, Piperno DR, Iriarte J, Holst I. Preceramic maize from Paredones and Huaca Prieta, Peru. Proc Natl Acad Sci. 2012;109:1755–9.

27. Ross-Ibarra J, Tenaillon M, Gaut BS. Historical divergence and gene flow in the genus Zea. Genetics. 2009;181:1399–413.

28. Tenaillon MI, U'Ren J, Tenaillon O, Gaut BS. Selection versus demography: a multilocus investigation of the domestication process in maize. Mol Biol Evol. 2004;21:1214–25.

29. Van Etten J, Hijmans RJ. A geospatial modelling approach integrating archaeobotany and genetics to trace the origin and dispersal of domesticated plants. PLoS One. 2010;e12060:5.

30. Races of maize. Available from:https://www.ars.usda.gov/midwest-area/ames/plant-introduction-research/docs/races-of-maize/. Accessed 15 Apr 2015.

31. Schiffels S, Durbin R. Inferring human population size and separation history from multiple genome sequences. Nat Genet. 2014;46:919–25.

32. Schrider DR, Shanku AG, Kern AD. Effects of linked selective sweeps on demographic inference and model selection. Genetics. 2016;204:1207–23.

33. Hearne S, Chen C, Buckler E, Mitchell S. Unimputed GBS derived SNPs for maize landrace accessions represented in the SeeD-maize GWAS panel. CIMMYT. 2014. http://data.cimmyt.org/dvn/dv/seedsofdiscoverydvn/faces/study/StudyPage.xhtml?studyId=21&tab=files.

34. Thompson EA. Identity by descent: variation in meiosis, across genomes, and in populations. Genetics. 2013;194(2):301–26.

35. Durand EY, Patterson N, Reich D, Slatkin M. Testing for ancient admixture between closely related populations. Mol Biol Evol. 2011;28:2239–52.

36. da Fonseca RR, Smith BD, Wales N, Cappellini E, Skoglund P, Fumagalli M, et al. The origin and evolution of maize in the American Southwest. Nat Plants. 2015;1:14003.

37. Martin SH, Davey JW, Jiggins CD. Evaluating the use of ABBA–BABA statistics to locate introgressed loci. Mol Biol Evol. 2015;32:244–57.

38. Navarro JAR, Willcox M, Burgueño J, Romay C, Swarts K, Trachsel S, et al. A study of allelic diversity underlying flowering-time adaptation in maize landraces. Nat Genet. 2017;49:476–80.

39. Buckler ES, Holland JB, Bradbury PJ, Acharya CB, Brown PJ, Browne C, et al. The genetic architecture of maize flowering time. Science. 2009;325(5941):714–8.

40. Cooper GM, Stone EA, Asimenos G, Green ED, Batzoglou S, Sidow A. Distribution and intensity of constraint in mammalian genomic sequence. Genome Res. 2005;15:901–13.

41. Renaut S, Rieseberg LH. The accumulation of deleterious mutations as a consequence of domestication and improvement in sunflowers and other Compositae crops. Mol Biol Evol. 2015;32:2273–83.

42. Günther T, Schmid KJ. Deleterious amino acid polymorphisms in Arabidopsis thaliana and rice. Theor Appl Genet. 2010;121:157–68.

43. Kono TJ, Fu F, Mohammadi M, Hoffman PJ, Liu C, Stupar RM, et al. The role of deleterious substitutions in crop genomes. Mol Biol Evol. 2016;33:2307–17.

44. Zhu Q, Zheng X, Luo J, Gaut BS, Ge S. Multilocus analysis of nucleotide variation of Oryza sativa and its wild relatives: severe bottleneck during domestication of rice. Mol Biol Evol. 2007;24:875–88.

45. Lam HM, Xu X, Liu X, Chen W, Yang G, Wong FL, et al. Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. Nat Genet. 2010;42:1053–9.

46. Zhou Y, Massonnet M, Sanjak J, Cantu D, Gaut BS. Evolutionary genomics of grape (Vitis vinifera ssp. vinifera) domestication. Proc Natl Acad Sci. 2017;114:11715–11720. doi:10.1073/pnas.1709257114.

47. Meyer RS, Choi JY, Sanches M, Plessis A, Flowers JM, Amas J, et al. Domestication history and geographical adaptation inferred from a SNP map of African rice. Nat Genet. 2016;48:1083–8.

48. Gaut BS, Díez CM, Morrell PL. Genomics and the contrasting dynamics of annual and perennial domestication. Trends Genet. 2015;31:709–19.

49. Eyre-Walker A, Gaut RL, Hilton H, Feldman DL, Gaut BS. Investigation of the bottleneck leading to the domestication of maize. Proc Natl Acad Sci. 1998;95:4441–6.

50. Nielsen R, Beaumont MA. Statistical inferences in phylogeography. Mol Ecol. 2009;18:1034–47.

51. Mazet O, Rodriguez W, Grusea S, Boitard S, Chikhi L. On the importance of being structured: instantaneous coalescence rates and human evolution—lessons for ancestral population size inference? Heredity. 2016;116:362.

52. Whitlock MC, McCauley DE. Some population genetic consequences of colony formation and extinction: genetic correlations within founding groups. Evolution. 1990;44(7):1717–24.

53. Vigouroux Y, Glaubitz JC, Matsuoka Y, Goodman MM, Sánchez J, Doebley J. Population structure and genetic diversity of New World maize races assessed by DNA microsatellites. Am J Bot. 2008;95:1240–53.

54. Takuno S, Ralph P, Swarts K, Elshire RJ, Glaubitz JC, Buckler ES, et al. Independent molecular basis of convergent highland adaptation in maize. Genetics. 2015;200:1297–312.

55. Pearsall, DM. Plant domestication and the shift to agriculture in the Andes In: Silverman H, Isbell WH, editors. The handbook of South American archaeology. New York: Springer; 2008. p. 105–20.

56. Ross-Ibarra J, Tenaillon M, Gaut BS. Historical divergence and gene flow in the genus Zea. Genetics. 2009;181:1399–413.

57. Piperno DR, Moreno JE, Iriarte J, Holst I, Lachniet M, Jones JG, et al. Late Pleistocene and Holocene environmental history of the Iguala valley, central Balsas watershed of Mexico. Proc Natl Acad Sci. 2007;104:11874–81.

58. Correa-Metrio A, Lozano-García S, Xelhuantzi-López S, Sosa-Nájera S, Metcalfe SE. Vegetation in western Central Mexico during the last 50000 years: modern analogs and climate in the Zacapu Basin. J Quat Sci. 2012;27:509–18.

59. Poets AM, Fang Z, Clegg MT, Morrell PL. Barley landraces are characterized by geographically heterogeneous genomic origins. Genome Biol. 2015;16:1.

60. Bredeson JV, Lyons JB, Prochnik SE, Wu GA, Ha CM, Edsinger-Gonzales E, et al. Sequencing wild and cultivated cassava and related species reveals extensive interspecific hybridization and genetic diversity. Nat Biotechnol. 2016;34:562–70.

61. Miao B, Wang Z, Li Y. Genomic analysis reveals hypoxia adaptation in the Tibetan Mastiff by introgression of the grey wolf from the Tibetan Plateau. Mol Biol Evol. 2016;34:734–43.

62. Doebley J, Goodman MM, Stuber CW. Patterns of isozyme variation between maize and Mexican annual teosinte. Econ Bot. 1987;41(2):234–46.

63. Yang J, Mezmouk S, Baumgarten A, Buckler ES, Guill KE, McMullen MD, et al. Incomplete dominance of deleterious alleles contributes substantially to trait variation and heterosis in maize. PLoS Genetics. 2017;13:e1007019.

64. Gerke JP, Edwards JW, Guill KE, Ross-Ibarra J, McMullen MD. The genomic impacts of drift and selection for hybrid performance in maize. Genetics. 2015;201(3):1201–11.

65. Lu J, Tang T, Tang H, Huang J, Shi S, Wu CI. The accumulation of deleterious mutations in rice genomes: a hypothesis on the cost of domestication. Trends Genet. 2006;22:126–31.

Wang *et al. Genome Biology*   (2017) 18:215

Page 13 of 13

66. Schubert M, Jónsson H, Chang D, Der Sarkissian C, Ermini L, Ginolhac A, et al. Prehistoric genomes reveal the genetic foundation and cost of horse domestication. Proc Natl Acad Sci. 2014;111:E5661—9.
67. Ramu P, Esuma W, Kawuki R, Rabbi IY, Egesi C, Bredeson JV, et al. Cassava haplotype map highlights fixation of deleterious mutations during clonal propagation. Nat Genet. 2017;49:959–63.
68. Liu Q, Zhou Y, Morrell PL, Gaut BS. Deleterious variants in Asian rice and the potential cost of domestication. Mol Biol Evol. 2017;34:908–24.
69. McQuillan R, Eklund N, Pirastu N, Kuningas M, McEvoy BP, Esko T, et al. Evidence of inbreeding depression on human height. PLoS Genet. 2012;8: e1002655.
70. Agrawal AF, Whitlock MC. Inferences about the distribution of dominance drawn from yeast gene knockout data. Genetics. 2011;187:553–66.
71. Manna F, Martin G, Lenormand T. Fitness landscapes: an alternative theory for the dominance of mutation. Genetics. 2011;189:923–37.
72. Doyle JJ. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochem Bull. 1987;19:11–15.
73. Li H, Durbin R. Fast and accurate long-read alignment with Burrows–Wheeler transform. Bioinformatics. 2010;26:589–95.
74. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, et al. The B73 maize genome: complexity, diversity, and dynamics. Science. 2009;326:1112–5.
75. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43:491–8.
76. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27:2156–8.
77. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet. 2007;81:1084–97.
78. Chia JM, Song C, Bradbury PJ, Costich D, de Leon N, Doebley J, et al. Maize HapMap2 identifies extant variation from a genome in flux. Nat Genet. 2012;44:803–7.
79. Fumagalli M, Vieira FG, Korneliussen TS, Linderoth T, Huerta-Sánchez E, Albrechtsen A, et al. Quantifying population genetic differentiation from next-generation sequencing data. Genetics. 2013;195:979–92.
80. Fumagalli M, Vieira FG, Linderoth T, Nielsen R. ngsTools: methods for population genetics analyses from next-generation sequencing data. Bioinformatics. 2014;30:1486–7.
81. Korneliussen TS, Albrechtsen A, Nielsen R. ANGSD: analysis of next generation sequencing data. BMC Bioinformatics. 2014;15:356.
82. Skotte L, Korneliussen TS, Albrechtsen A. Estimating individual admixture proportions from next generation sequencing data. Genetics. 2013;195: 693–702.
83. Schliep KP. phangorn: phylogenetic analysis in R. Bioinformatics. 2011;27: 592–3.
84. Korneliussen TS, Moltke I, Albrechtsen A, Nielsen R. Calculation of Tajima's D and other neutrality test statistics from low depth next-generation sequencing data. BMC Bioinformatics. 2013;14:289.
85. Vieira FG, Fumagalli M, Albrechtsen A, Nielsen R. Estimating inbreeding coefficients from NGS data: impact on genotype calling and allele frequency estimation. Genome Res. 2013;23:1852–61.
86. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81:559–75.
87. Ogut F, Bian Y, Bradbury PJ, Holland JB. Joint-multiple family linkage analysis predicts within-family variation better than single-family analysis of the maize nested association mapping population. Heredity. 2015;114: 552–63.
88. De Mita S, Siol M. EggLib: processing, analysis and simulation tools for population genetics and genomics. BMC Genet. 2012;13:27.
89. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput Biol. 2010;6:e1001025.
90. Rodgers-Melnick E, Bradbury PJ, Elshire RJ, Glaubitz JC, Acharya CB, Mitchell SE, et al. Recombination in diverse maize is stable, predictable, and associated with genetic load. Proc Natl Acad Sci. 2015;112:3823–8.
91. Bukowski R, Guo X, Lu Y, Zou C, He B, Rong Z, et al. Construction of the third generation Zea mays haplotype map. bioRxiv. 2015;026963.
92. Wang L, Beissinger TM, Lorant A, Ross-Ibarra C, Ross-Ibarra J, Hufford M. The interplay of demography and selection during maize domestication and expansion. github. 2017. Available from: https://doi.org/10.5281/ zenodo.1013415. Accessed 15 Oct 2017.
93. Wang L, Beissinger TM, Lorant A, Ross-Ibarra C, Ross-Ibarra J, Hufford M. The interplay of demography and selection during maize domestication and expansion. NCBI SRA; 2017. BioProject Accession: PRJNA300309; SRP065483. Available from: https://www.ncbi.nlm.nih.gov/bioproject/ 300309?log$=activity. Accessed 15 Oct 2017.