bioRxiv preprint doi: https://doi.org/10.1101/336503; this version posted June 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

A domestication history of dynamic adaptation and genomic deterioration in sorghum.

Oliver Smith[1,2], William V Nicholson[1,3], Logan Kistler[1,4], Emma Mace[5], Alan Clapham[1], Pamela Rose[6], Chris Stevens[7], Roselyn Ware[1], Siva Samavedam[1], Guy Barker[1], David Jordan[8], Dorian Q Fuller[7], Robin G Allaby[1*].

1. School of Life Sciences, University of Warwick, Coventry, CV4 7AL, United Kingdom.

2. Natural History Museum of Denmark, Øster Voldgade 5-7, 1350 København K, Denmark.

3. Warwick Medical School, University of Warwick, Coventry, CV4 7AL, United Kingdom.

4. Department of Anthropology,4. Smithsonian Institution, National Museum of Natural History, Washington, D.C. 20506, USA.

5. Department of Agriculture, Fisheries and Forestry Queensland (DAFFQ), Warwick, Queensland 4370, Australia.

6. The Austrian Archaeological Institute, Cairo Branch, Zamalek, Cairo, Egypt

7. Institute of Archaeology, UCL, London, United Kingdom

8. Queensland Alliance for Agriculture and Food Innovation, The University of Queensland, Warwick, Queensland 4370, Australia.

* Corresponding author

bioRxiv preprint doi: https://doi.org/10.1101/336503; this version posted June 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Abstract

The evolution of domesticated cereals was a complex interaction of shifting selection pressures and repeated introgressions. Genomes of archaeological crops have the potential to reveal these dynamics without being obscured by recent breeding or introgression. We report a temporal series of archaeogenomes of the crop sorghum (*Sorghum bicolor*) from a single locality in Egyptian Nubia. These data indicate no evidence for the effects of a domestication bottleneck but instead suggest a steady decline in genetic diversity over time coupled with an accumulating mutation load. Dynamic selection pressures acted sequentially on architectural and nutritional domestication traits, and adaptation to the local environment. Later introgression between sorghum races allowed exchange of adaptive traits and achieved mutual genomic rescue through an ameliorated mutation load. These results reveal a model of domestication in which genomic adaptation and deterioration was not focused on the initial stages of domestication but occurred throughout the history of cultivation.

Keywords

Ancient DNA, archaeobotany, bottleneck, introgression, genomic rescue

bioRxiv preprint doi: https://doi.org/10.1101/336503; this version posted June 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

The evolution of domesticated plant forms represents a major transition in human history that facilitated the rise of modern civilization. In recent years our understanding of the domestication process has become revised considerably (1). In the case of cereals it has been recognized that the selective forces that give rise to domestication syndrome traits such as the loss of seed shattering were generally weak and comparable to natural selection (2,3) and that the intensity of selection pressures changed over the course of time as human technology evolved (4). Furthermore, domesticated lineages have often been subjected to repeated introgressions from local wild populations that endowed adaptive traits and obscured historical signals in the genome (5). Such complexity obfuscates attempts to reconstruct the evolutionary history of domesticated species from modern plants. To counter these confounding factors in this study we directly tracked the evolutionary trajectory of a domesticated species, sorghum (*Sorghum bicolor* ssp. *bicolor* (L.) Moench.), through the archaeological record. This approach enabled the identification of selection pressures not clear today, and the tracking of the introgression process, revealing a domestication history which runs counter to the expectations of the current conventional model of domestication.

Sorghum is the world's fifth most important cereal crop and the most important crop of arid zones (6) used for food, animal feed, fibre and fuel. The evolution of sorghum has seen its transition from being a wild pluvial plant in north-eastern Africa (*S. bicolor* ssp. *verticilliflorum* (Steud.) De Wet ex Wiersema & Dahlberg, hereafter referred to as *S. verticilliflorum* for clarity) to the ancestral domesticated

bioRxiv preprint doi: https://doi.org/10.1101/336503; this version posted June 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

form *Sorghum bicolor* type bicolor in Central Eastern Sudan by around 5000 years ago, while cultivation is inferred to have begun by 6000 yrs BP (7)., Ultimately, four specialized agroclimatic adapted types evolved after domestication—durra, caudatum, guinea, and kafir (8-10). The derived types were likely founded on introgressions of the wild progenitor complex *Sorghum verticilliflorum* or closely related species into the ancestral bicolor type, endowing traits such as drought tolerance in the case of type durra (8,11). The evolutionary history of sorghum, replete with introgression, is difficult to reconstruct from modern datasets. However, a temporal series of archaeobotanical domesticated sorghums spanning back to 2100 before present (yrs BP) at the archaeological site of Qasr Ibrim, situated on the Nubian frontier of northern Africa, affords the opportunity to track this complex crop directly through time removing the obscuring effects of introgression (12). Prior to this wild sorghum is present at Qasr Ibrim from at least ca. 2800 yrs BP. Domesticated sorghum (race bicolor) appears at the site ca. 2100 yrs BP. After this time period phenotypically domesticated sorghum of the ancestral type bicolor occurs throughout all cultural periods until the site was abandoned 200 years ago. During the early Christian period at 1470 yrs BP, the oldest known drought-adapted, free-threshing durra type appears at the site and occurs there for the rest of the site's occupancy. The origins of the durra type are unclear. Current distributions in northern and eastern Africa and its dominance in the Near East and South Asia led to the proposal that durra originated on the Indian subcontinent (13) and returned to Africa at some point after 2000 yrs BP (14).

bioRxiv preprint doi: https://doi.org/10.1101/336503; this version posted June 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Results

Genetic diversity of sorghum over time

To gain a longitudinal insight into the evolutionary history of sorghum, we sequenced 9 archaeological genomes from different time points at Qasr Ibrim, including a wild phenotype from 1765 yrs BP and 8 domesticated phenotypes between 1805 and 450 yrs BP, a further 2 genomes from herbarium material, and 12 genomes of modern wild and cultivated sorghum types representing the varietal range (Table S1, S2).

We investigated how genetic diversity has changed through time by measuring within genome heterozygosity of 100 kbp genomic blocks that revealed a pattern of broad variation in heterozygosity in the wild progenitor *S. verticilliflorum* that became progressively narrower over time in the ancestral bicolor type, Figure S1. Interestingly, the wild phenotype of sorghum at Qasr Ibrim (sample A3) has a narrower variation in heterozygosity than the wild progenitor (represented by modern wild diversity), suggesting that it had been already been subject to genetic erosion. Conversely, the durra types all showed similar low levels of genomic variation in heterozygosity suggestive of genetic erosion prior to their appearance at Qasr Ibrim. Total genomic heterozygosity of bicolor over time confirmed that the 'wild' sorghum had already undergone considerable genetic erosion relative to the wild progenitor. To our surprise, the decreasing trend in heterozygosity over time fits a linear model (p values 0.0041 and 9.2×10^{-6} for parameters a and b respectively) better than an exponential model (p values 0.042 and 9.3×10^{-5}) as would be expected from an early initial

bioRxiv preprint doi: https://doi.org/10.1101/336503; this version posted June 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

loss of diversity through a domestication bottleneck (15), Figure 1, Table S3 suggesting that there was no measurable effect on genetic diversity attributable to a domestication bottleneck.

Mutation load over time in Sorghum

The apparent lack of a domestication bottleneck runs contrary to expectations for a domesticated crop. To investigate the apparent lack of a domestication bottleneck further, we considered the mutation load. An expected consequence of the bottleneck is a rise in mutation load as small populations incorporate deleterious mutations through strong-acting drift. High mutation loads have generally been observed in domesticated crops (16-18), which have been taken as a confirmation of the effects of the domestication bottleneck. We measured the mutation load over time in the archaeological sorghum using a genome evolutionary rate profiling (GERP) analysis considering the total number of potentially deleterious alleles (19) (see methods), Figure 2. As with other domesticated crops, modern sorghum has a higher mutation load than its wild progenitor under both recessive and additive models. In contrast to the expectations of a domestication bottleneck we did not observe an initial large increase in mutation load associated with domestication, but rather an overall increasing trend in mutation load over time to the present day suggesting a process of load accumulation combined with selective purging episodes. In this case the trend line is best described by a positive exponential model rather than a linear model, Table S3, indicating mutation load has become increasingly

bioRxiv preprint doi: https://doi.org/10.1101/336503; this version posted June 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

problematic in recent times. However, the p-values suggest that the coefficient for the time in each model is only weakly significant, which could be the result of multiple processes on going, such as a strong increase in the rate of mutation load accumulation in recent times. When we considered the number of sites containing deleterious alleles (dominant model) rather than total number of deleterious alleles we observe a decreasing trend over time (Figure S3). This pattern suggests that part of the rising mutation load in the bicolor type was due to the increased homozygosity over time causing fixation of deleterious alleles originating from the wild progenitor pool. There is variation over time in mutation load, most notably in 1805 year-old sorghum (sample A5) that shows a sharp increase due to the incorporation of strongly deleterious alleles, both in the total number of alleles and number of sites. Interestingly, we found that the durra types show a pattern that contrasts to the bicolor type with relatively little change in heterozygosity and a significant fall in mutation load over time, suggesting the purging of deleterious mutations either through selection or genomic rescue through hybridization (Figure S2, S3). The contrasting patterns in mutation load over time are also reflected in methylation state profiles, which can reflect the state of genome-wide stress (20), Figure S4.

Signals of selection in sorghum

We considered that episodes of selection could have contributed in part to the variation in mutation load observed over time, either through reducing population size due to the substitution load or through hitchhiking effects. Three approaches

bioRxiv preprint doi: https://doi.org/10.1101/336503; this version posted June 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

were used to identify candidate regions under selection. Firstly, we surveyed for wild/domestication heterozygosity to look for significant reduction in heterozygosity in domesticates, which revealed 30 peaks of genome-wide significance (denoted by prefix pk), Figure 3, Table S4, S5. We also specifically surveyed 38 known domestication loci and also found a significant reduction of heterozygosity in 15 of the 38 associated regions, Table S6. Secondly, we used a SweeD analysis to detect selective sweeps (21), which identified 11 peaks (denoted by prefix s), Figure 3, Table S7. In the third approach we utilized the temporal sequence of archaeogenomes to investigate episodes of selection intensification by considering the gradient of heterozygosity change over time (see methods). In this latter approach we tracked the gradient of change in the heterozygosity of regions identified in the first two approaches and assigned significance based on the gradient deviation from the genome average for each type. We considered multiple time sequences representing alternative possible routes through contemporaneous genomes over time within type bicolor and type durra respectively. This revealed a period of selection intensification associated with domestication loci prior to 1805 yrs BP, followed by oscillations in diversity, Table S8, Figures S5-S7.

Together, the selection identification approaches exploit a range of different types of signature left by selection, and reveal a complex and dynamic history of selection over time summarized in Figure 5. We generally found more evidence for selection in the bicolor type sorghum than the durra type. Despite its apparent wild phenotype, the wild sorghum (A3) at Qasr Ibrim from 1765 yrs BP shows

bioRxiv preprint doi: https://doi.org/10.1101/336503; this version posted June 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

evidence of selection at domestication loci concerned with architecture (*int1*, *tb1*), suggesting possible introgression with contemporaneous domesticated forms (represented by sample A5) that could have contributed to reduced heterozygosity. Interestingly, the intensification signals show some overlap between samples A3 and A5 (*int1* and *ae1*), with A5 showing further evidence for selection at shattering, dwarfing and sugar metabolism loci (*Sh3/Bt1*, *dw2* and *SPS5*) that would contribute to the domesticated phenotype of A5 relative to A3. Subsequent to this a period of intensification in selection is apparent both for dwarfing and sugar metabolism traits (710-715 years BP) in the bicolor type, with ten domestication loci showing significantly low levels of heterozygosity in this lineage by 710 yrs BP in A7. In the bicolor type, two of the sugar metabolism associated gene families show evidence of early selection controlling photosynthetic sucrose production first (*SPS*) and then an intensification of selection for breakdown (*SUS*). A third gene family (*SUT*) associated with sucrose transport appears to come under later selection in bicolor. In contrast, fewer domestication loci were found to show evidence of intensifying selection in the durra type, and none showed evidence of low heterozygosity. In this case we detected signals for an intensification of selection on tillering and maturity associated loci (*gt1*, *ma3*, the latter also being detected using SweeD s1 in the bicolor lineage). Significant heterozygosity reduction was identified in windows containing a large number of disease resistance loci (pk4, pk11, pk15, pk20, pk24, pk25) as well as sugar metabolism loci (pk14, pk18, pk19, pk22) in the bicolor type. One of SweeD peak (s2) was closely matched to pk5 on

bioRxiv preprint doi: https://doi.org/10.1101/336503; this version posted June 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

chromosome 2 in the 54.0 – 54.2 Mbp interval, possibly indicating signatures for the same selection process. This region shows a consistently low heterozygosity over time in the bicolor type with the notable exception of the 1805 year old sorghum (A5). The region contains the far-red impaired response genes (*FAR1*), as well as anther indehiscence 1 (*AI1*). The *FAR1* gene is associated with phytochrome A signal transduction (22), so is important in responses to far red light that divert resources away from tall growth to increase root and grain growth. The *AI1* gene regulates anther development (23), allowing earlier development. Either of these genes may be locally adapted to the Qasr Ibrim environment since they already appear to be under intense selection in the wild sorghum at this site (sample A3), but not apparently under as much constraint in modern sorghum type bicolor.

The dynamic selection over time detected with most intensification of selection occurring before 1805 years BP, appears to correlate with a sharp increase in mutation load in the bicolor type. In contrast, the durra type shows much less evidence of selection and on arrival at Qasr Ibrim shows initially similar levels of mutation load to the bicolor type that then decreased over time (Figure S3). To investigate whether loci of selection are associated with higher regions of mutation load we measured the maximum deviations between genomes in GERP load scores across the genome and compared those to the locations of selection peak candidates (Figure 3, Table S9). Selection signatures were highly significantly associated with regions of maximum deviation in mutation load with 30% of low heterozygosity peaks (p 8.04 x10^{-9}), 45% SweeD

bioRxiv preprint doi: https://doi.org/10.1101/336503; this version posted June 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

peaks (p 2.55x10^{-7}) and 26% domestication loci (p 5.03 x 10^{-5}) occurring in such regions. The intensification of selection is associated with increased mutation load and could explain the spike in mutation load observed in the 1805 year old sorghum (sample A5).

Genome rescue through hybridization

We considered that the decreasing mutation load observed in the durra type could be due to a genomic rescue caused by hybridization with the local bicolor type. To investigate for evidence of hybridization we first constructed a maximum likelihood phylogenetic tree of wild and cultivated total genomes (Figure S8), and individual trees for 970 sections across the genome (Supplementary data set 1). After accounting for biases introduced by ancient DNA modification, both the durra and bicolor type from Qasr Ibrim form a single clade to the exclusion of modern bicolor and durra types, suggesting they have indeed hybridized over time. D-statistic analysis for introgression (24) shows over time the durra type became increasingly similar to the local bicolor type, suggesting progressive introgression between the two types (Figure S9). We then compared the archaeological genomes against a global sorghum diversity panel (25,26) (Figure 4). The archaeological genomes are distributed along an axis of spread that has Asian durra types at the extremity. The oldest archaeological durra type (A11) sits between East African durra types and Asian durra types, whilst the wild phenotype sorghum, most closely aligned to the subsequent type bicolor, sits close to the center of the PCA, suggesting East African durras may have arisen

bioRxiv preprint doi: https://doi.org/10.1101/336503; this version posted June 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

from a hybridization between Asian durra and African bicolor. The oldest archaeological durra type in this study (sample A11) may represent one of the earliest of the east African durras. The younger archaeological genomes of the two types become progressively closer on the PCA supporting a process of ongoing hybridization between the two types over time.

Finally, we investigated whether the hybridization between the bicolor and durra types led to adaptive introgression or genomic rescue. Phylogenetic incongruence between the bicolor and durra type clades suggests that hybridization was frequent at loci under selection (Table S10). In agreement with previous studies (10) there is clear evidence for a donation of the dwarfing *dw1* allele from durra to bicolor with a single durra type sample sitting within the bicolor type clade in this region, but in most cases although the clades of bicolor and durra have become mixed, it is not sufficiently clear which is the more likely donor. Interestingly, seven of the nine sugar-metabolism associated loci potentially under selection in the bicolor type are also areas of introgression with durra. In all cases where identified was possible (*su*, *SUS1* and *SPS3*), durra was identified as the donor. However, in the case of *SPS5* in which we identified early intensification of selection in the bicolor type, no phylogenetic incongruence occurred. Conversely, at the maturity locus *ma3* containing region, the durra type A11 that was identified as potentially under selection sits within the bicolor clade suggesting a donation from bicolor to durra. The *FAR1/AI1* loci region, which appears to have been under strong selection in bicolor throughout, appears to have been donated from bicolor to durra.

bioRxiv preprint doi: https://doi.org/10.1101/336503; this version posted June 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Assuming that prior to the introduction of the durra type to Qasr Ibrim the two types, bicolor and durra, had accrued mutation loads independently, then hybridization would have afforded the opportunity for genomic rescue between the two types. We therefore considered all ancestor/descendent pairs of genomes within the bicolor and durra type lineages in the context of a third potential donor genome, and scanned all sites for comparative GERP load scores under the additive model. We calculated firstly the difference in GERP load scores between the ancestor and potential donor to give a 'total rescue value' that reflects a donor's potential to reduce mutation load across the entire genome, Figure 5. We secondly assessed the donor's potential to effect mutation load reduction specifically at only those sites in which there had been a reduction in GERP load score between the ancestor and descendent to give an 'on target rescue value'.

In the case of durra sample A11 (1470 yrs BP) as ancestor and A9 as descendent, bicolor samples A6 (715 yrs BP) and A7 (710 yrs BP) are intermediate in age and therefore potential donor genome types from the bicolor lineage. The analysis predicts that either A6 or A7 would have reduced load in the regions that were observed to be reduced in the descendent A9 (505 yrs BP), but an overall detrimental effect to genome wide load, which is in fact observed (Figure S3). However, had earlier bicolor types been available for introgression, such as A5 (1805 yrs BP), then hybridization would have been more beneficial for the durra types. Generally, there is strong rescue potential of bicolor types by durra on the sites that were observed to improve, however in most cases there is

bioRxiv preprint doi: https://doi.org/10.1101/336503; this version posted June 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

an expectation that the over all load would be increased from the transfer of durra specific load to bicolor. Notably, durra types in general are predicted to reduce the on target load and genome wide load in A7, which is observed (Figure S3).

Discussion

This study demonstrates that sorghum represents an alternative domestication history narrative in which the effects of a domestication bottleneck are not apparent, mutation load has accrued over time probably as a consequence of dynamic selection pressures rather than a domestication-associated collapse of diversity, and that genomic rescue from load occurred when two different agroclimatic types met.

The linear nature of the decreasing trend in diversity over time observed in sorghum in this study is surprising. An extreme bottleneck early in the history of would be expected to lead to a negative exponential trend as diversity is rapidly lost in the early stages of domestication. An alternative explanation for the trend could be that diversity has been lost steadily through drift over time. However, a simple drift model shows that such a ten-fold loss in diversity would also be associated with a negative exponential trend, Figure S11. It is possible that diversity loss could have been supplemented by gains through introgression from the wild over time, counteracting the trend made by drift. Sample A3 could be the result of a wild introgression event since there are older domesticate phenotypes in the archaeobotanical record, such as sample A5. Sorghum is known for its

bioRxiv preprint doi: https://doi.org/10.1101/336503; this version posted June 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

extensive introgression leading to a strong regional structure within cultivars (10), making continuous introgression seem like a plausible scenario for sorghum at Qasr Ibrim. Incorporation of three systems of introgression into the simple drift model in which introgression is either constant, diminishing or increasing over time still results in a non-linear trend, which become parabolic when introgression becomes very high over time (not shown), Figure S11. We therefore think it unlikely that a model of constant drift and introgression is causative of the apparent linear decrease in diversity over time observed in this study.

Such linear decreases in diversity have been observed in human populations with increasing geographic distance from Africa and are most robustly explained by sequential founder models (27). The annual cycle of crop sowing and harvesting also represents a serial founding event scenario. A simple model of founding events in which 25% of the harvest is set aside for sowing based on field experiments (28) demonstrates that loss of genetic diversity approximates a linear process as populations become large, Figure S12A, and that the gradient of diversity loss is highly correlated with the populations size, Figure S12B. On the basis of the gradient of diversity loss observed in sorghum, this model predicts a long-term population size of 289,407 for the sorghum in this study. This estimate is in excess of the effective population size estimated from the heterozygosity of wild sorghum, 135,823. It therefore seems plausible that in the case of sorghum diversity has likely been lost through a series of sequential founding episodes based on the cropping regime in a process that likely

bioRxiv preprint doi: https://doi.org/10.1101/336503; this version posted June 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

incorporated all the available wild genetic diversity at the outset rather than a substantial initial domestication bottleneck.

The deleterious effects of mutation load are becoming increasingly apparent and a major problem in modern crops such as the dysregulation of expression in maize (29). The study here demonstrates the potential immediacy of the problem in that mutation load may generally be a consequence of recent selection pressures leading to an exponentially rising trend rather than a legacy of the domestication process. While the general trend of the archaeogenomes is for the increase in the number of sites homozygous for deleterious variants (recessive model), the overall trend for the number of sites holding deleterious variants decreases (dominant model), which suggests a process of general purging of variants from the standing variation of the wild progenitor combined with the rise of homozygosity with decreasing diversity of the variant sites that remain. However, this is sharply contrasted by modern sorghum in which there is a leap in the number of sites holding deleterious mutations (dominant model). This process contributes to the accompanying jump in load under both the recessive and additive models in modern sorghum. This indicates a large influx of new deleterious variants within the last century giving the trend of mutation load accumulation an exponential shape. It is likely that this influx of mutation load is the product of recent breeding programs and the genetic bottlenecks associated with the Green Revolution. The accumulation of load has previously been associated with mutation meltdown and extinction of past populations (30) but it remains unclear whether crops could follow the same fate in the absence of

bioRxiv preprint doi: https://doi.org/10.1101/336503; this version posted June 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

rescue processes, or whether such episodes could have been involved with previous agricultural collapses when crops experienced extensive adaptive challenges (31,32). In the case of sorghum wild genetic resources may be valuable not only as a source of improved and environmentally adaptive traits, but also as a source for reparation of genome wide mutation load that may affect housekeeping and economic traits alike.

This represents the first plant archaeogenomic study that tracks multiple genomes to gain insight into changes in diversity over time directly. The trends revealed, based on a relatively low number of archaeological genomes, suggest a domestication history contrary to that typically expected for a cereal crop. Further archaeogenomes may establish whether this is a general trend for sorghum and other crops.

Methods

1. Sample Acquisition. Archaeological samples were sourced from A. Clapham from the archaeological site Qasr Ibrim, outlined in Table S1. For details on dating see section 1.3 below. Historical samples from the Snowden collection were sourced from Kew Gardens, Kew1: Tsang Wai Fak, collection no. 16366 Kew2: Tenayac, Mexico, collection assignation 's.n.'. Modern samples of *S. bicolor* ssp. *bicolor* type bicolor, durra, kafir, caudatum, drumondii and guinea were supplied through the USDA [accession numbers PI659985, PI562734, PI655976, PI509071, PI653734 and PI562938 respectively]. Wild sorghum samples *S. vertilliciliflorum, S. arundinaeum, and S. aethiopicum* were also

bioRxiv preprint doi: https://doi.org/10.1101/336503; this version posted June 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

obtained from the USDA [accession numbers PI520777, PI532564, PI535995], and wild *S. virgatum* was donated by D. Fuller. The outgroups *S. propinquum* and *S. halapense* were obtained from the USDA [accession numbers PI653737 and Grif 16307] respectively.

The genomes generated in this study were also compared to 1023 re-sequenced genomes taken from Thurber et al 2013 (26).

1.2 *A note on taxonomy.* The sorghum genus is complex with numerous taxonomic systems. After Morris *et al.*'s findings (10), we have elected not to describe the principal cultivar types as subspecies or races but rather simply 'types' to reflect the reality that there is evidence of considerable introgression between each of these forms. The wild progenitor of domesticated sorghum is a complex made up of four 'races' verticilliflorum, arundinaceum, aethiopicum and virgatum. However, the integrity of these races is also questioned, and the currently more accepted designation is one species, verticilliflorum, of which the other races are subtypes. For clarity and simplicity in this study we have used the race type as a variety designation.

1.3 A note on Qasr Ibrim and archaeological context of samples. Qasr Ibrim was a fortified hilltop site in the desert of Lower Nubia on the east bank of the Nile, about 200 km, south of Aswan in modern Egypt. It has been excavated over numerous field seasons, since 1963 by the Egyptian Exploration Society (UK). In recent years with higher Lake Nasser levels only upper parts of the site are

bioRxiv preprint doi: https://doi.org/10.1101/336503; this version posted June 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

preserved as an island (33,34). The desert conditions provided exceptional organic preservation by desiccation with exceptional preservation of a wide range of biomolecules (e.g. 35-37). Systematic sampling for plant remains was initiated in 1984 (38) and the first studies of these remains were carried out in the 1980s by Rowley-Conwy (39) and had continued by Alan Clapham (40,41). The exceptional plant preservation has previously allowed successful ancient genomic studies of barley (35) and cotton (36).

Qasr Ibrim was founded sometime before 3000 years BP. It had occupations associated the Napatan kings (Egyptian Dynasty 25: 747-656 BC), possible Hellenistic and Roman Egypt (3rd century BC to 1st c. AD), the Meroitic Kingdom (1st century to 4th century AD), and local post-Meroitic (AD 350-550) and Nubian Christian Kingdoms (AD 550-1300). Earlier periods are associated temples to Egyptian and Meroitic deities. After Christianity was introduced the site had a Cathedral. Later Islamic occupations finished with use as an Ottoman fortress. The site was abandoned in AD1812. The Sorghum material studied here comes from a range of different contexts from excavation seasons between 1984 and 2000. While the chronology of the site is well established by artefactual material, including texts in various scripts, several sorghum remains or associated crops, were submitted for direct AMS radiocarbon dating, as listed below in Table S2. For directly dated find the median of the 2-sigma calibrated age range has been used. Note that Radiocarbon calibration defines “the present” as AD 1950, and we have recalculated the median as before AD 2000, and assigned Snowden historical collections form the start of the 20th century as ca. 100 BP. For material

bioRxiv preprint doi: https://doi.org/10.1101/336503; this version posted June 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

not directly dated, sample A12 could be assigned based on associated pottery and finds, which have a well-established chronology through the Christian periods (42), A12 is associated with Islamic/Ottoman material (1500-1800 AD, ca. 400 BP)

2. *DNA extraction.* DNA was extracted from archaeological and historical samples in a dedicated ancient DNA facility physically isolated from other laboratories. All standard clean-lab procedures for working with ancient DNA were followed. Single seeds from each accession were ground to powder using a pestle & mortar and incubated in CTAB buffer (2% CTAB, 1%PVP, 0.1M Tris-HCl pH 8, 20mM EDTA, 1.4M NaCl) for 5 days at 37°C. The supernatant was then extracted once with an equal volume of 24:1 chloroform:isoamyl alcohol. DNA was then purified using a Qiagen plant Mini Kit with the following modifications: a) 5x binding buffer was used instead of 1.5x and incubated at room temperature for 2 hours before proceeding. b) After washing with AW2, columns were washed once with acetone and air-dried in a fume hood to prevent excessive G-forces associated with centrifugal drying. c) DNA was eluted twice in a total of 100µl elution buffer and quantified using a Qubit high sensitivity assay.

DNA from modern samples was extracted using a CTAB precipitation method due to excessive polysaccharide levels precluding column-based extractions. Briefly, seeds were ground to powder and incubated at 60°C for 1 hour in 750 ul CTAB buffer as previously described, with the addition of 1ul β-mercaptoethanol. Debris was centrifuged down and the supernatant was extracted once with an

bioRxiv preprint doi: https://doi.org/10.1101/336503; this version posted June 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

equal volume of 24:1 chloroform:isoamyl alcohol. The supernatant was then collected and mixed with 2x volumes precipitation buffer (1% CTAB, 50mM Tris-HCl, 20nM EDTA) and incubated at 4°C for 1 hour. DNA was precipitated at 6°C by centrifugation at 14,000 *g* for 15 minutes. The pellet was washed once with precipitation buffer and incubated at room temperature for 15 minutes before being centrifuged again under the same conditions. The pellet was dried and resuspended in 100µl high-salt TE buffer (10mM Tris-HCl, 1M NaCl) and incubated at 60°C for 30 minutes with 0.5µl RNase A. The DNA was then purified using Ampure XP SPRI beads.

3. *Library construction and genome sequencing.* Libraries for all samples were constructed using an Illumina TruSeq Nano kit, according to manufacturers' protocol. A uracil-intolerant polymerase (Phusion) was used to amplify the libraries, in order to eliminate the C to U deamination signal often observed in ancient DNA in favour of the 5' 5mC to T deamination signal. The purpose of this was to obtain epigenomic information after analysis using epiPaleomix (43). Consequently the data set was reduced for non-methylated cytosine deamination signals in the 5' end, but showed expected levels of G to A mismatches for ancient DNA (5-10%) in the 3' end and high levels of endogenous DNA content typical for samples from this site (Table S1). While this approach is thought to reduce library complexity by reducing the number of successfully amplified molecules, we considered this to be a worthwhile trade-off considering the exceptional preservation and endogenous DNA content of the Qasr Ibrim

bioRxiv preprint doi: https://doi.org/10.1101/336503; this version posted June 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

samples. We found no evidence to suggest insufficient library complexity after amplification. A minor modification was made to the protocol for ancient and historical samples: a column-based cleanup after end repair was used, in order to retain small fragments that would otherwise be lost under SPRI purifications as per the standard protocol. Genomes were sequenced on the Illumina HiSeq 2500 platform. Ancient and historical samples were sequenced on one lane each using SR100 chemistry and modern samples on 0.5 lanes each using PE100 chemistry.

4. *Preliminary Bioinformatics processing.* Illumina adapters were trimmed using cutadapt v1.11 using 10% mismatch parameters. Resulting FastQ files were mapped to the BTX623 genome (44) using bowtie2 v2.2.9 (46) under --sensitive parameters. SAM files containing mapped reads with a minimum mapping score of 20 were then converted to BAM files using samtools v1.14 (47). Variant calls format (VCF) files were then made from pileups constructed using samtools mpileup, and variant calls were made using bcftools v1.4 (47).

5. *Methylation analysis.* Since a uracil-intolerant polymerase was used for library generation, we analysed BAM files using epipaleomix (43) on the ancient samples. We then collated the number of identifiable 5mC sites globally for each sample. Epipaleomix is designed to characterise CpG islands typical to animal genomes and, is not suited to gene-specific analysis of plant genomes to due to their wider methylation states (CHH and CHG) (45). However when assessing

bioRxiv preprint doi: https://doi.org/10.1101/336503; this version posted June 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

relative overall genome methylation between individuals of the same species, CpG islands measured in this way provide a perfectly adequate proxy. We opted for global and windowed-measurements to determine relative methylation states between samples.

6. *Evolutionary and population analyses.* Two archaeological genomes (A8 and A12) were from phenotypes intermediate between bicolor and durra types. We found that sample A8 was predominantly of bicolor type and A12 predominantly of durra type. Given the uncertainty of these samples and their likely hybrid origins, we elected to leave them out of most analyses.

6.1 *Heterozygosity analysis* The number of heterozygous sites was measured for each 100 kbp window of genome aligned to the BTX_623 reference sequence (44). The frequency distribution of heterozygosity was then calculated by binning the windows in 1 heterozygous base site intervals. Ratios of wild:cultivated heterozygosity were calculated for each window using *S. verticilliflorum* as the wild progenitor. Ratios closely approximate a negative exponential distribution. Probabilities of observed heterozygosity ratios for each window were obtained from a negative exponential distribution with λ equal to $1/\mu$ for all ratios for each chromosome. A Bonferroni correction was applied by multiplying probability values by the number of windows on a chromosome in Figure 4. Locations of 38 known domestication syndrome loci (shown in Tables S5 and S7) were obtained by reference to the BTX_623 genome. Candidate domestication loci were

bioRxiv preprint doi: https://doi.org/10.1101/336503; this version posted June 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

obtained from the scans of Mace *et al* (25). In the genome-wide scan peaks were considered significant if $1/p > 100$ after Bonferroni correction.

We considered the possibility that the observed heterozygosity levels may be influenced by postmortem DNA damage. To explore this, we characterized the relationships between time, heterozygosity and postmortem deamination. As we previously described, C to U damage signals are eliminated at the 5' ends of sequence reads because of our choice of polymerase, so we therefore characterized damage profiles at the 3' ends only, using mapDamage output statistic '3pGtoA_freq' and taking a mean of the 25 reported positions for each ancient or historical sample. Unsurprisingly, we found that the accumulation of damage patterns is a function of time in a logistic growth model, assuming a zero-point intercept for both factors ($R^2 = 0.9$). 80% of damage capacity under this model is reached reasonable quickly, in 331.0 years. All the Qasr Ibrim samples are at least 400 years old, and so we re-fitted a linear regression model to these samples only so characterize these relationships in a true time-series. We found a negligible correlation between time and damage accumulation after 400 years ($R^2 = 0.15$, $p = 0.34$). Next, we characterized the relationship between age and heterozygosity under the same model (although without the assumption of a zero-point intercept, since even modern domesticate lines in this study show non-zero levels) and found a weak fit ($R^2 = 0.64$, $p = 0.14$). This relationship is however likely influenced by our central hypothesis, with 'less domesticated' samples being earlier in the archaeological record, and so a counter-argument should not be inferred from this analysis. Finally, we assessed the relationship

bioRxiv preprint doi: https://doi.org/10.1101/336503; this version posted June 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

between damage and heterozygosity by linear regression, assuming inappropriateness of a logistic model since both damage and heterozygosity factors are functions of time. We found a weak correlation when considering all samples (R^2 = 0.2, p = 0.2), and virtually no correlation when considering the Qasr Ibrim time series only (R^2 = 0.04, p = 0.61). Considering that the two historical Kew samples are ostensibly domesticates, and historical and geographic outliers to the rest of the dataset, we conclude that the observed levels of heterozygosity in the ancient samples are not influenced by postmortem damage patterns.

6.2 *Differential Temporal Heterozygosity Gradient Analysis.* Our rationale was to utilize the temporal sequence of genomes to identify time intervals associated with intensification of selection. To this end we designed an analysis to identify outliers in changing heterozygosity over time to the general genomic trend. We considered all possible historical paths between genomes given three pairs of samples were almost contemporaneous (A3/A5, A6/A7 and A9/A10), with wild *S. verticilliflorum* representative of the wild progenitor in the case of the bicolor lineages.

For each 100kbp window we calculated the gradient of change in heterozygosity between temporally sequential genome pairs by subtracting younger heterozygosity values from older and dividing through by the time interval between samples. Genome-wide gradient values for all 100kbp windows were used to construct a non-parametric distribution to obtain probability values

bioRxiv preprint doi: https://doi.org/10.1101/336503; this version posted June 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

of change over each time interval for a 100kbp window between a particular pair of samples. Peak regions identified by heterozygosity ratio, SweeD analysis and known domestication syndrome genes were then measured for gradient probability.

6.3 *SweeD analysis.* VCF files from our 23 ancient, historical and modern samples and also 9 samples from Mace et al (25) were combined using the GATK (52) program CombineVariants. Subsequently, the combined VCF file was filtered - using bcftools v1.4 (47) - to only include sites with 2 or more distinct alleles and at sites where samples have depth less than 5 or a variant calling quality score less than 20 to exclude those samples. Then a further filter was applied - using bcftools v1.4 - to exclude variant calls due to C->T and G->A transitions relative to the reference, which potentially represent post-mortem deamination which has a high rate in aDNA samples (48). SweeD (21) was run with options for multi-threading (to run with 64 threads) and to compute the likelihood on a grid with 500 positions for each chromosome.

6.4 *Genome Evolutionary Rate Profiling (GERP) analysis.* This analysis was carried out broadly following the methodology of Cooper *et al.* (19). We aligned the repeat-masked genomes of 27 plant taxa to the BTX_623 sorghum reference genome using last, and processed resulting maf files to form netted pairwise alignment fastas using kentUtils modules maf-convert, axtChain, chainPreNet, chainNet, netToAxt, axtToMaf, mafSplit, and maf2fasta. We forced all alignments

bioRxiv preprint doi: https://doi.org/10.1101/336503; this version posted June 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

into the frame of the sorghum reference using an expedient perl script, and built a 27-way fasta alignment excluding sorghum for GERP estimation. We created a fasta file of fourfold degenerate sites from chromosome 1 (347394 sites; NC_012870) with a perl script, and calculated a neutral rate model using phyloFit, assuming the HKY85 substitution model and the following tree:

(((((((((Trifolium_pratense,Medicago_truncatula),Glycine_max),Prunus_persica),(Populus_trichocarpa,Manihot_esculenta)),(((Arabidopsis_thaliana,Arabidopsis_lyrata),(Brassica_napus,Brassica_rapa)),Theobroma_cacao)),Vitis_vinifera),((Solanum_tuberosum,Solanum_lycopersicum),(Chenopodium_quinoa,Beta_vulgaris))),(((Zea_mays,Setaria_italica),(((Oryza_rufipogon,Oryza_longistaminata),Leersia_perrieri),(((Triticum_urartu,Aegilops_tauschii),Hordeum_vulgare),Brachypodium_distachyon))),Musa_acuminata)),Amborella_trichopoda)

We then calculated GERP rejected subsitutions (RS) scores using gerpcol with the default minimum three taxa represented for estimation. The mutation load for each genome was then assessed by scanning through their VCF files generated by alignment to BTX_623. Maize was used as an outgroup to judge the ancestral state, and only sites at which there was information from maize were incorporated into the analysis. Sites which differed to the ancestral state were scored based on the associated RS score for that site following the scheme of Wang *et al.* (18): 0, neutral, 0-2 slightly deleterious, 2-4, moderately deleterious, >4 seriously deleterious. We collected scores under three models, recessive,

bioRxiv preprint doi: https://doi.org/10.1101/336503; this version posted June 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

additive and dominant. Under the dominant model we counted each site once regardless of whether it had one or two alternative bases to the ancestor, so giving the total number of base sites containing at least one potentially deleterious allele. Under the additive model we counted the total number of alleles that were alternative to the ancestor such that each homozygous alternative site scored 2, but heterozygous sites scored 1. Under the recessive model only sites that were homozygous for potentially deleterious variants were counted.

To investigate the significance of overlap between regions significant GERP regions of difference (GROD) between taxa and signatures of selection we used a binomial test in which the null probability of selecting a GROD was equal to the total number of GRODS (193) divided by the total number of 100 kbp regions studied (6598), and N and *x* were the total number of selection signals and the number of selection signals occurring in a GROD respectively.

We used the GERP profiles to explore potential genomic rescue from mutation load accrued independently in the bicolor and durra lineages prior to hybridization between the two types. For the purposes of this analysis we used the wild sorghum genome A3 as a possible wild ancestor genome to the domesticated bicolor form A5 even though this wild sample is contemporaneous to that domesticated form. All possible ancestor descendent pairs were assembled within bicolor or durra types, and all 100 kbp windows were scanned for the relative additive model GERP load scores for ancestor, descendent and a third potential donor genome. The total potential for the donor genome to rescue

bioRxiv preprint doi: https://doi.org/10.1101/336503; this version posted June 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

the ancestral genome was scored summing the difference in GERP scores across all windows between the ancestor and donor. To better fit a scenario in which the donor genome was the causative agent of reduction GERP load score we identified windows that satisfied the condition ancestor GERP load score > descendent GERP load score, and summed up the difference in ancestor and potential donor scores to give an 'on target rescue' value.

6.5 *Phylogenetics* Maximum likelihood tress were constructed using exaML (49) firstly using whole genome sequences (Figure S7), and for 970 consecutive blocks across the genome (supplementary data set). Prior to computing phylogenetic trees, the VCF files were processed as described in section 6.3 (on the SweeD analysis) albeit with our 23 ancient, historical and modern samples only.

The maximum likelihood tree using the whole genome sequences was constructed as follows. Our own script created a multiple sequence alignment file by concatenating the variant calls in the VCF file and outputting the results in PHYLIP (50) format. The program parse-examl from the ExaML package (version 3.0.15) was run in order to convert the PHYLIP format file into ExaML's own binary format. Also, ExaML requires an initial starting tree which was obtained by running (on multiple threads) Parsimonator v1.0.2, a program available as part of the RaxML package (51) - developed by the same research group - for computing maximum parsimony trees. An ExaML executable

bioRxiv preprint doi: https://doi.org/10.1101/336503; this version posted June 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

(compiled to run using MPI) was run on multiple CPUs in order to compute the maximum likelihood tree.

The trees for 970 consecutive blocks across the genome were computed by essentially the same approach as described above for a single tree, after a script obtained the blocks from the input VCF file (for the combined samples) and output them in PHYLIP format.

To assess potential donation between genomes at candidate loci we examined trees spanning the corresponding100kbp windows. The tree topology was examined for congruence in the maintenance of bicolor and durra type groups within the Qasr Ibrim group of genomes. Instances of phylogenetic incongruence were interpreted as candidate regions of recombination between the two genome types, although identification of the donor and recipient genomes was not always clear. Simple cases in which a single genome from one sorghum type was found within the group of the other type were interpreted as possible genome donations from that group to the single genome. In the case of regions that scored highly in the SweeD analysis no phylogenetic congruence was attempted because the taxon in which selection has operated is not identified.

6.6 *Principal Component Analysis of global diversity set.* A subset of 1894 SNPs were used to find the principal axes of genetic variation for the 23 samples and an unpublished set of 1046 diverse sorghum lines spanning the racial and geographic diversity of the primary gene pool of cultivated sorghum. 580 of these

bioRxiv preprint doi: https://doi.org/10.1101/336503; this version posted June 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

diverse lines were described in Thurber et al (26). These lines were produced within the Sorghum Conversion Program which introgressed key height and phenology genes into exotic lines to enable them to be produced in sub-tropical environments. The introgressed regions spanned approximately 10% of the genome which were masked for the purposes of this analysis. Principal component analysis of the centered data matrix was performed in R (R core team, 2017) using the *prcomp* function in the base “stats” package.

6.7 D statistics. Patterson’s D-Statistic and modified F-statistic on Genome wide SNP data was used to infer patterns of introgression (24). D-statistic and fd-statistic for each of the 10 chromosomes was calculated using the R-package PopGenome. Variant Call Format (VCF) file, which is generated after mapping reads of an individual sample to the reference genome, was given as input to the readVCF() function of the package (52).

We used four R-language based S4 class methods from PopGenome package to carry out the introgression tests for every chromosome. First, we used the method set.population by providing 3 populations (2 sister taxa and an archaic group) viz., P1=BTX_623, P2=varying samples, P3=Most ancient S.bicolor A3. Second, using set.outgroup function, we set an outgroup (P4= S.halapense). Third, the method introgression.stats was employed to calculate the introgession tests. Finally, we used jack.knife.transform method (53) which transforms an existing object belonging to GENOME class into another object of the GENOME class with regions that corresponding to a Jackknife window.

bioRxiv preprint doi: https://doi.org/10.1101/336503; this version posted June 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Standard error was then calculated by eliminating one such window i.e., a single chromosome under study and calculation was applied to the union of all the other chromosomes.

We tested for admixture from the most ancient *S. bicolor* type bicolor sample (A3), assuming this represents a genome prior to the appearance of the durra type on the African continent. The BTX_623 sorghum reference genome was taken as P_1, sample A3 was taken as P_3 and *S. halapense* was taken as the out group P_4. *S. halapense* is native to southern Eurasia to east India and does not readily cross with *S. bicolor.* Samples were then tested at the P_2 position across all 100kbp windows, each chromosome tested separately. Negative values (indicating an excess of P_1/P_3 combinations) are expected when the BTX_623 genome is more similar to sample A3 than P_2. This is observed as expected for the durra types, although the value of D decreases over time consistent with either an increase in instances of P_2/P_3 or instances of P_1/P_2, both suggesting progressive introgression between the durra and bicolor types over time. Positive values (indicating an excess of P_2/P_3 combinations) suggest a close relationship between sample A3 and P_2, which is observed the Qasr Ibrim bicolor types (A5, A6 and A7).

6.8 Linear and exponential line fitting to heterozygosity and GERP score data.

A straight line was fit to the heterozygosity data in Figure 2 using the glm function (for generalized linear models) in R and also an exponential function was fit to the same data using the gnm package (for generalized non-linear models) in R

bioRxiv preprint doi: https://doi.org/10.1101/336503; this version posted June 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

obtaining the values for the parameters, standard errors, p and AIC shown in Table S3. (It was confirmed similar values were obtained for the parameters, standard errors, p and AIC by fitting the straight line model using the gnm package in R.)

6.9 Basic simulation of diversity loss through drift, introgression and serial founding events

To explore the effect on general trend line shape of introgression over time we used a basic simulation of drift loss using the standard equation:

$$\frac{H_t}{H_0} = \left(1 - \frac{1}{2N_{fo}}\right)\left(1 - \frac{1}{2N(\frac{N_e}{N})}\right)^{t-1}$$

where N_{fo}, N and N_e are the founding population size, census population size and effective population size respectively. For simplicity, we assumed in the case of our crop that all three population sizes were equal. To incorporate introgression we used a simulation to calculate and modify each generation by using the above equation to modify the diversity from the previous generation, and then adding a diversity value representative of gene flow. Gene flow was altered each generation by a power factor *f*, which was 1 in the case of constant introgression, 1.0001 in the case of diminishing introgression over time and 0.99995 in the case of increasing introgression over time, with an initial value for introgression as 0.000015, equating to the value of genetic diversity added to the population each generation. We used a founder population of 2000 for 6000 generations to

bioRxiv preprint doi: https://doi.org/10.1101/336503; this version posted June 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

recapitulate the observed 10-fold loss of diversity over this time frame in sorghum.

The serial founder event simulation was executed using the following model: To initialize, allele frequencies were randomly assigned for a defined number of alleles using a uniform distribution. We also applied a skewed distribution in which the first allele frequency generated as above was amplified to become a dominating allele frequency by a defined value. We found this made no difference to the simulation outcomes. N individuals were then randomly drawn using the allele frequencies, and the resultant frequency distribution calculated. Homozygosity was calculated as the sum of the squares of allele frequencies and subtracted from 1 for the heterozygosity. To convert these allele values to a per base site heterozygosity comparable to the sorghum data we divided by 1/He of the wild progenitor. A founder event was then generated by drawing Nb individuals from the allele frequency distribution, the new resultant allele frequency distribution calculated. N individuals were then drawn from this distribution, new frequencies calculated and the resultant heterozygosity calculated as above. The process was repeated for a defined number of cycles. We explored a scenario in which the founder population was based on setting aside 25% of the (seed) population each year, following classic experimental archaeology field trials (28). We explored several orders of magnitude of N (100, 1000, 10000, 100000), and assumed 5 alleles per gene, for 1000 founding events, equating to 1000 years of agriculture. Each trial was repeated 100 times, equating to 100 genes being simulated independently. While the overall

bioRxiv preprint doi: https://doi.org/10.1101/336503; this version posted June 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

distribution of diversity loss over time is exponential, seen more clearly at smaller population sizes, the trend approximates linear more closely with increasing population size (Figure 12A). We calculated the gradient of descent from the first 60 founding events and found the logs of the gradient and population size to be directly proportional (Figure 12B). We used linear regression of this relationship to predict the log of the population size associated with the log of the observed gradient of descent of diversity in sorghum. We independently calculated the effective population size associated with wild sorghum using the heterozygosity as an estimate of θ (using the relationship $\theta=4Ne\mu$), and an estimate of 5×10^{-9} subs/site/year for the neutral mutation rate (54).

Acknowledgements

The authors would like to thank M. Nesbitt for permitting the use of herbaria material from Kew. OS, WN, GB and RGA were supported by the NERC (NE/L006847/1) and LK was supported by NERC (NE/L012030/1). CJS and DQF work with archaeobotanical materials was supported by a European Research Council grant (no. 323842). Sequence data were deposited in the European Molecular Biology Laboratory European Bioinformatics Institute [project code PRJEB24962.].

bioRxiv preprint doi: https://doi.org/10.1101/336503; this version posted June 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

(1) Larson G, Piperno D, Allaby RG et al. Current perspectives and the future of domestication studies. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 6139-6146 (2014).

(2) Purugganan MD, Fuller DQ. The nature of selection during plant domestication. *Nature* **457**, 843-848 (2009).

(3) Fuller DQ, Denham T, Arroyo-Kalin M, Lucas L, Stephens C, Qin L, Allaby RG, Purugganan MD Convergent evolution and parallelism in plant domestication revealed by an expanding archaeological record. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 6147-6152 (2014).

(4) Allaby RG, Stevens S, Lucas L, Maeda O, Fuller DQ. Geographic mosaics and changing rates of cereal domestication. *Philosophical Transactions of the Royal Society B*. **372**, 20160429 (2017).

(5) Poets AM, Fang Z, Clegg MT, Morell PL. Barley landraces are characterized by geographically heterogeneous genomic origins. *Genome Biol*ogy **16**, 173 (2015).

(6) Food and Agriculture Organization of the United Nations. [http://www.fao.org/index_en.htm].

(7) Winchell F, Stevens CJ, Murphy C, Champion L, Fuller DQ. Evidence for Sorghum Domestication in Fourth Millennium BC Eastern Sudan: Spikelet Morphology from Ceramic Impressions of the Butana Group. *Current Anthropology* **58**, 673-683 (2017).

(8) Doggett H *Sorghum* 2nd Ed, Longman, Harlow (1988).

bioRxiv preprint doi: https://doi.org/10.1101/336503; this version posted June 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

(9) Brown PJ, Myles S, Kresowich S. Genetic support for a phenotype-based racial classification in sorghum. *Crop Sci.* **51**, 224-230 (2011).

(10) Morris G, *et al.* Population genomic and genome-wide association studies of agroclimatic traits in sorghum. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 453-458 (2013).

(11) Fuller, Dorian Q and Chris J. Stevens (n.d.) Sorghum Domestication and Diversification: A current archaeobotanical perspective. In: Anna Maria Mercuri, A. Catherine D'Andrea, Rita Fornaciari, Alexa Höhn (eds.) *Plants and People in Africa's Past. Progress in African Archaeobotany.* Springer (2017).

(12) Clapham AJ, Rowley-Conwy PA In Fields of Change–Progress in African Archaeobotany, Cappers R, ed. Groningen Archaeological Studies. Groningen, **5**, 157–164 (2007).

(13) de Wet JML, Harlan JR, Price EG Variability in *Sorghum bicolor*. In: Harlan JR, de Wet JMJ, Stemler ABL (eds) Origins of African plant domestication. Mouton Press, The Hague, p 453-463 (1976).

(14) Harlan JR, Stemler ABL The races of Sorghum in Africa. In: Harlan JR, de Wet J, Stemler ABL (eds) Origins of African plant domestication. Mouton Press, The Hague, p 465-478 (1976).

(15) Meyer R, Purugganan M. Evolution of crop species: genetics of domestication and diversification. *Nature Reviews Genetics* **14**, 840-852 (2013).

bioRxiv preprint doi: https://doi.org/10.1101/336503; this version posted June 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

(16) Liu Q, Zhou Y, Morrell PL, Gaut BS Deleterious variants in Asian rice and the potential cost of domestication. *Mol. Biol. Evol.* **34**(4), 908–924 (2017).

(17) Renault S, Rieseberg L. The accumulation of deleterious mutations as a consequence of domestication and improvement in sunflower and other Compositae crops. *Mol. Biol. Evol.* **32**(9), 2273–2283 (2015).

(18) Wang L, Beissinger TM, Lorant A, Ross-Ibarra C, Ross-Ibarra J, Hufford MB The interplay of demography and selection during maize domestication and expansion. *Genome Biol.* **18**, 215 (2017).

(19) Cooper GM *et al*. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.* **15**, 901-913 (2005).

(20) Smith O, Clapham A, Rose P, Liu Y, Wang J, Allaby RG. Genomic methylation patterns in archaeological barley show de-methylation as a time-dependent diagenetic process. *Scientific Reports* **4**, 5559 (2014).

(21) Pavlidis P, Živkovic D, Stamatakis A, Alachiotis N SweeD: Likelihood-based detection of selective sweeps in thousands of genomes. *Mol Biol Evol* **30**, 2224–2234 (2013).

(22) Hudson M, Ringli C, Boylan MT, Quail PH The *FAR1* locus encodes a novel nuclear protein specific to phytochrome A signaling. *Genes Devel.* **13**, 2017-2027 (1999).

(23) Zhu QH, Ramm K, Shivakkumar R, Dennis ES, Upadhyahya NM The *ANTHER INDEHISCENCE1* gene encoding a single MYB domain protein is involved in anther development in rice. *Plant Physiol.* **135**, 1514-1525 (2004).

bioRxiv preprint doi: https://doi.org/10.1101/336503; this version posted June 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

(24) Martin, S, Davey, J, Jiggins, C. Evaluating the Use of ABBA–BABA Statistics to Locate Introgressed Loci. *Mol. Biol. Evol.* **32**(1), 244-257 (2014).

(25) Mace ES *et al.* Whole genome sequencing reveals untapped genetic potential in Africa's indigenous cereal crop sorghum. *Nature Communications* **4**, 2320 (2013).

(26) Thurber CS, Ma JM, Higgins RH, Brown PJ Retrospective genomic analysis of sorghum adaptation to temperate-zone grain production. *Genome Biol.* **14**, R68 (2013).

(27) DeGiorgio M, Jakobsson M, Rosenberg N. Explaining worldwide patterns of human genetic variation using a coalescent-based serial founder model of migration outward from Africa. *Proc. Natl. Acad. Sci. U.S.A.* 106: 16057–16062 (2009).

(28) Hillman GC, Davies MS. Domestication rates in wild type wheats and barley under primitive cultivation. *B. J. Linn. Soc.* 39:39-78 (1990).

(29) Kremling KA *et al.* Dysregulation of expression correlates with rare-allele burden and fitness loss in maize. *Nature* 555:520-523 (2018).

(30) Rogers & Slatkin Excess defects in a woolly mammoth on Wrangel Island *PloS Genetics* 13(3): e1006601 (2017).

(31) Shennan S, Downey SS, Timpson A, et al. Regional population collapse followed initial agricultural booms in mid –Holocene Europe. *Nat Commun*. 4:2486 (2013).

bioRxiv preprint doi: https://doi.org/10.1101/336503; this version posted June 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

(32) Allaby RG, Kitchen JL, Fuller DQ (2016) Surprisingly low limits of selection in plant domestication. *Evolutionary Bioinformatics* 11:(S2) 41-51.

(33) Alexander, J The Saharan divide in the Nile Valley: the evidence from Qasr Ibrim. *The African Archaeological Review* **6**, 73-90 (1988).

(34) Rose P (2013) Qasr Ibrim In: Bagnall RS, Brodersen K, Champion CB, Erskine A, Huebner SR (eds.) *The Encyclopedia of Ancient History.* Oxford: Wiley. Pp. 5695-5697 DOI: 10.1002/9781444338386.wbeah15340.

(35) Palmer SA, Moore JD, Clapham AJ, Rose P, Allaby RG. Archaeogenetic evidence of ancient Nubian barley evolution from six to two-row indicates local adaptation. *PLoS One*, ***4***(7), e6301 (2009).

(36) Palmer SA, Clapham AJ, Rose P, Freitas F, Owen BD, Beresford-Jones D, Moore JD, Kitchen JL, Allaby RG. Archaeogenomic evidence of punctuated genome evolution in Gossypium. *Molecular biology and evolution*, ***29***(8), 2031-2038 (2012).

(37) O'Donoghue K, Clapham A, Evershed RP, Brown TA. Remarkable preservation of biomolecules in ancient radish seeds. Proc R Soc B Biol Sci. **263**, 541–547 1996.

(38) Alexander, J , Driskell B. Qasr Ibrim 1984. *Journal of Egyptian Archaeology* **71**, 12-26 (1985).

(39) Rowley-Conwy P Nubia AD 0-550 and the 'Islamic' agricultural revolution: Preliminary botanical evidence from Qasr Ibrim, Egyptian Nubia. *Archeologie du Nil Moyen* **3**, 131-138 (1989).

bioRxiv preprint doi: https://doi.org/10.1101/336503; this version posted June 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

(40) Clapham AJ, Rowley-Conwy PA (2007). New discoveries at Qasr Ibrim, Lower Nubia. Fields of change: progress in African archaeobotany. Barkhuis & Groningen University Library, Groningen, The Netherlands, 157-164.

(41) Clapham, A., & Rowley-Conwy, P. (2009). The archaeobotany of cotton (Gossypium sp. L.) in Egypt and Nubia with special reference to Qasr Ibrim, Egyptian Nubia. From foragers to farmers. Papers in honour of Gordon C. Hillman. Oxbow Books, Oxford, 244-253.

(42) Adams, WY (1986) *Ceramic Industries of Medieval Nubia.* University of Kentucky Press

(43) Hanghøj K, Seguin-Orlando A, Schubert M, Madsen T, Pedersen JS, Willerslev E, Orlando L. Fast, Accurate and Automatic Ancient Nucleosome and Methylation Maps with epiPALEOMIX. *Mol. Biol. Evol.* **33**, 3248-3298 (2016).

(44) Paterson AH *et al.* The Sorghum bicolor genome and the diversification of grasses. *Nature* **457**, 551-556 (2009).

(45) Bouyer et al. DNA methylation dynamics during early plant life. *Genome Biology* 18:179 (2017)

(46) Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nature Methods*. **9**(4), 357-359 (2012).

(47) Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G., Abecasis G., Durbin R. and 1000 Genome Project Data Processing Subgroup The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).

bioRxiv preprint doi: https://doi.org/10.1101/336503; this version posted June 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

(48) da Fonseca RR, Smith BD, Wales N, Cappellini E, Skoglund P, Fumagalli M, Samaniego JA, Carøe C, Ávila-Arcos MC, Hufnagel DE, Korneliussen TS, Vieira FG, Jakobsson M, Arriaza B, Willerslev E, Nielsen R, Hufford MB, Albrechtsen A, Ross-Ibarra J, Gilbert MT The origin and evolution of maize in the Southwestern United States. *Nature Plants* 1, 14003 (2015).

(49) Koslov AM, Aberer A, Alexandros S. ExaML version 3: a tool for phylogenomic analysis on supercomputers. *Bioinformatics* **31**, 2577-2579 (2015).

(50) Felsenstein J Evolutionary trees from DNA sequences: A maximum likelihood approach. *J Mol Evol* **17**(6), 368-376 (1981).

(51) Stamatakis A RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**(21), 2688-2690 (2006).

(52) Van der Auwera GA, Carneiro M, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, Banks E, Garimella K, Altshuler D, Gabriel S, DePristo M. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. *Current Protocols In Bioinformatics* **43**,11.10.1-11.10.33 (2013).

(53) Pfeifer B, Wittelsbürger U, Ramos-Onsins S, Lercher M. PopGenome: An Efficient Swiss Army Knife for Population Genomic Analyses in R. *Mol. Biol. Evol.* **31**(7):1929-1936 (2014).

(54) Wolfe KH, Sharpe PM, Li WH. Rates of synonymous substitution in plant nuclear genes. *J. Mol. Evol.* 29:208-211 (1989).

bioRxiv preprint doi: https://doi.org/10.1101/336503; this version posted June 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

bioRxiv preprint doi: https://doi.org/10.1101/336503; this version posted June 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Figure 1. Genomic heterozygosity over time in *S. bicolor* type bicolor.

Figure 2. Total recessive GERP load over time in *S. bicolor* type bicolor.

Figure 3. Selection signals across *S. bicolor* chromosomes 1 to 10. Heterozygosity ratio (wild/cultivated) inverted probabilities (Bonferroni corrected) shown in colours as described in key. Grey dashed line indicates 1% significance threshold after Bonferroni correction. SweeD values shown in red. Above: Locations of 38 known domestication genes shown in black. Locations of candidate domestication loci identified by Mace *et al* (24) shown in brown. Locations of GERP score regions of difference (grod) between genomes shown in green.

Figure 4. Principal Coordinate Analysis of 1894 SNPs from 23 genomes in this study and 1046 sorghum lines described in Thurber et al (25). Arrows indicate temporal movement of bicolor and durra type archaeogenomes in PCA.

Figure 5. Summary of selection signals over time in archaeogenomes. Red indicates selection intensification episodes, green indicates selection signals identified by low heterozygosity or SweeD analysis.

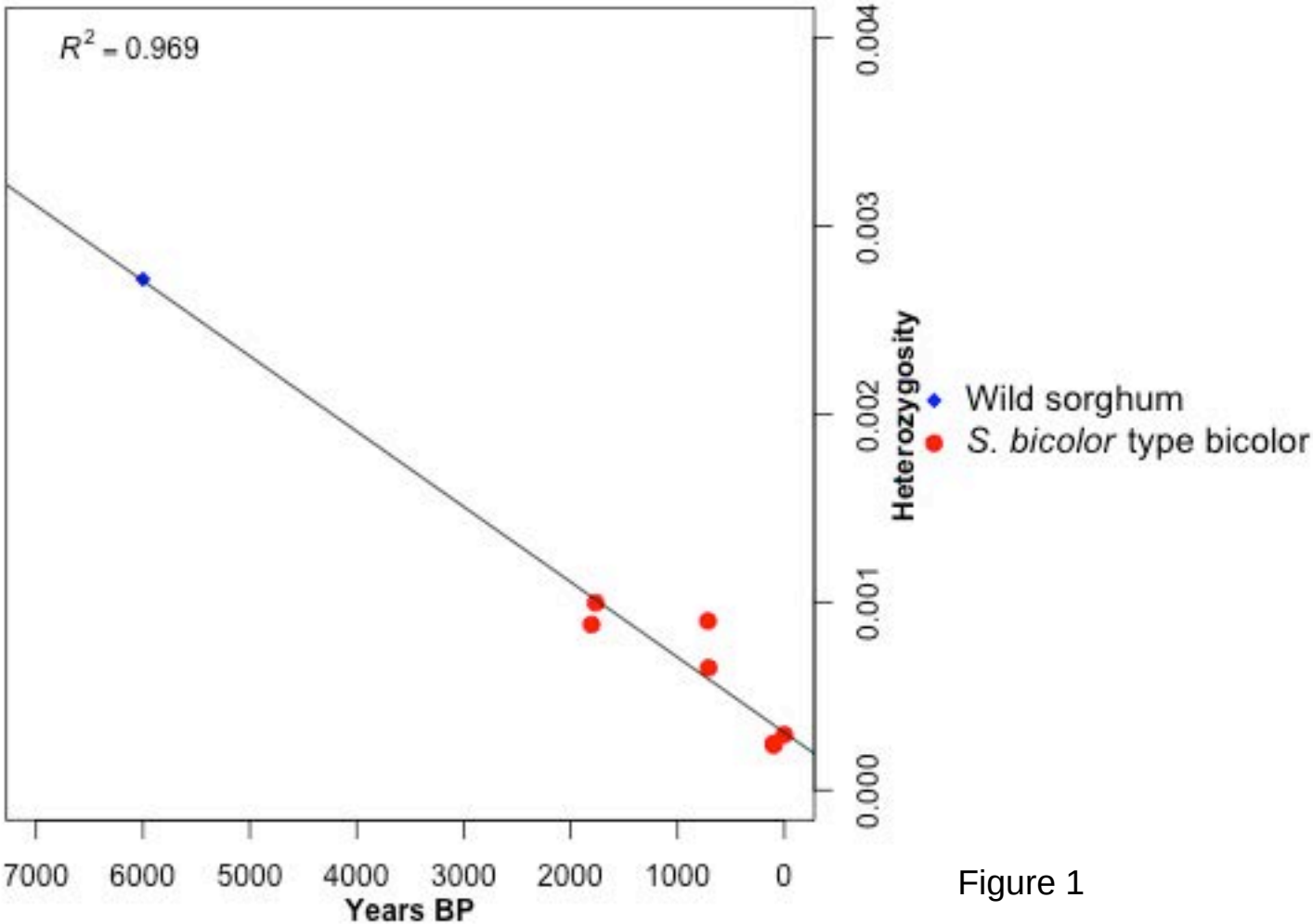

$R^2 = 0.969$
0.000
0.001
0.002
0.003
0.004
Heterozygosity
7000
6000
5000
4000
3000
2000
1000
0
Years BP
Wild sorghum
S. bicolor type bicolor

Figure 1

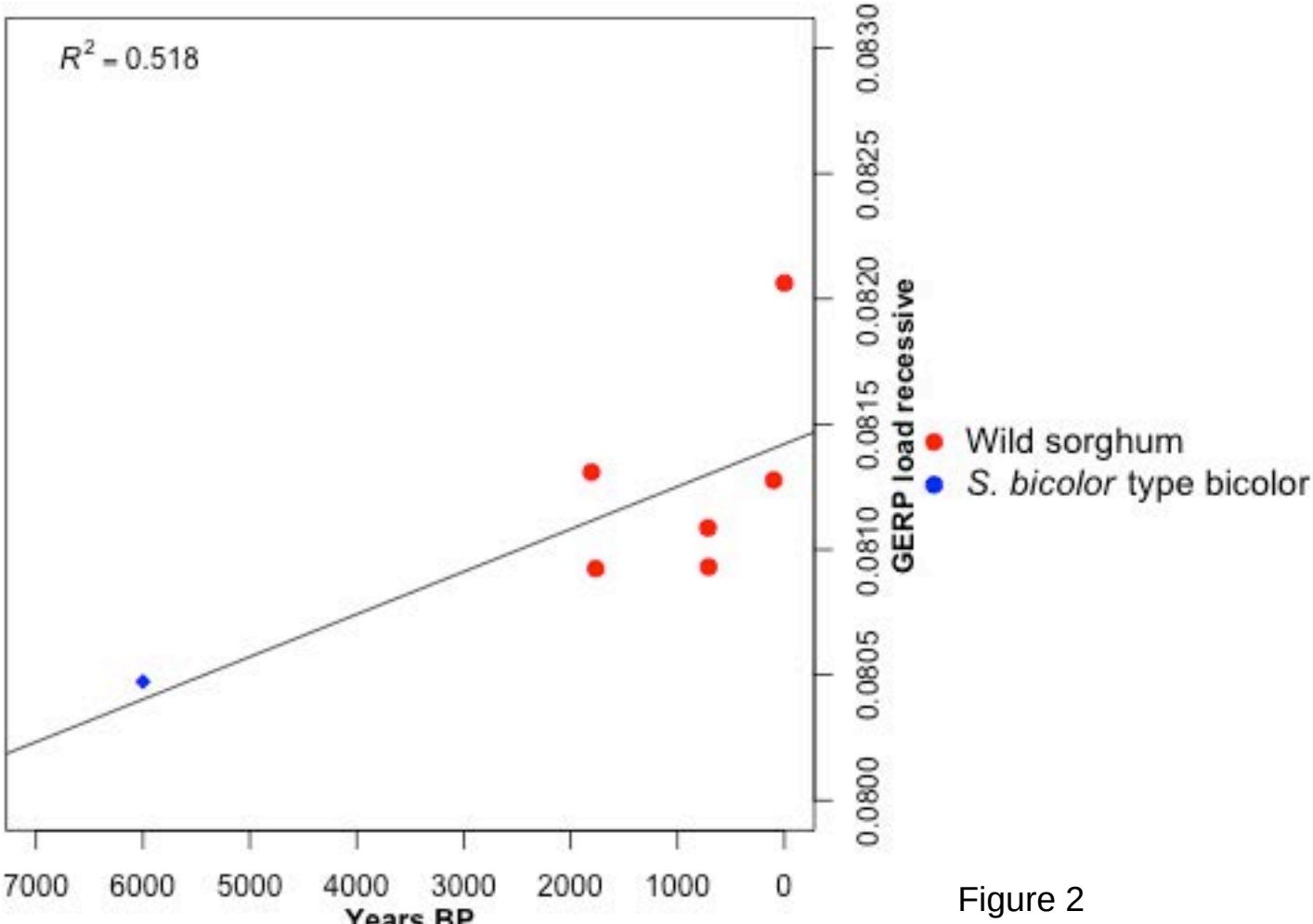
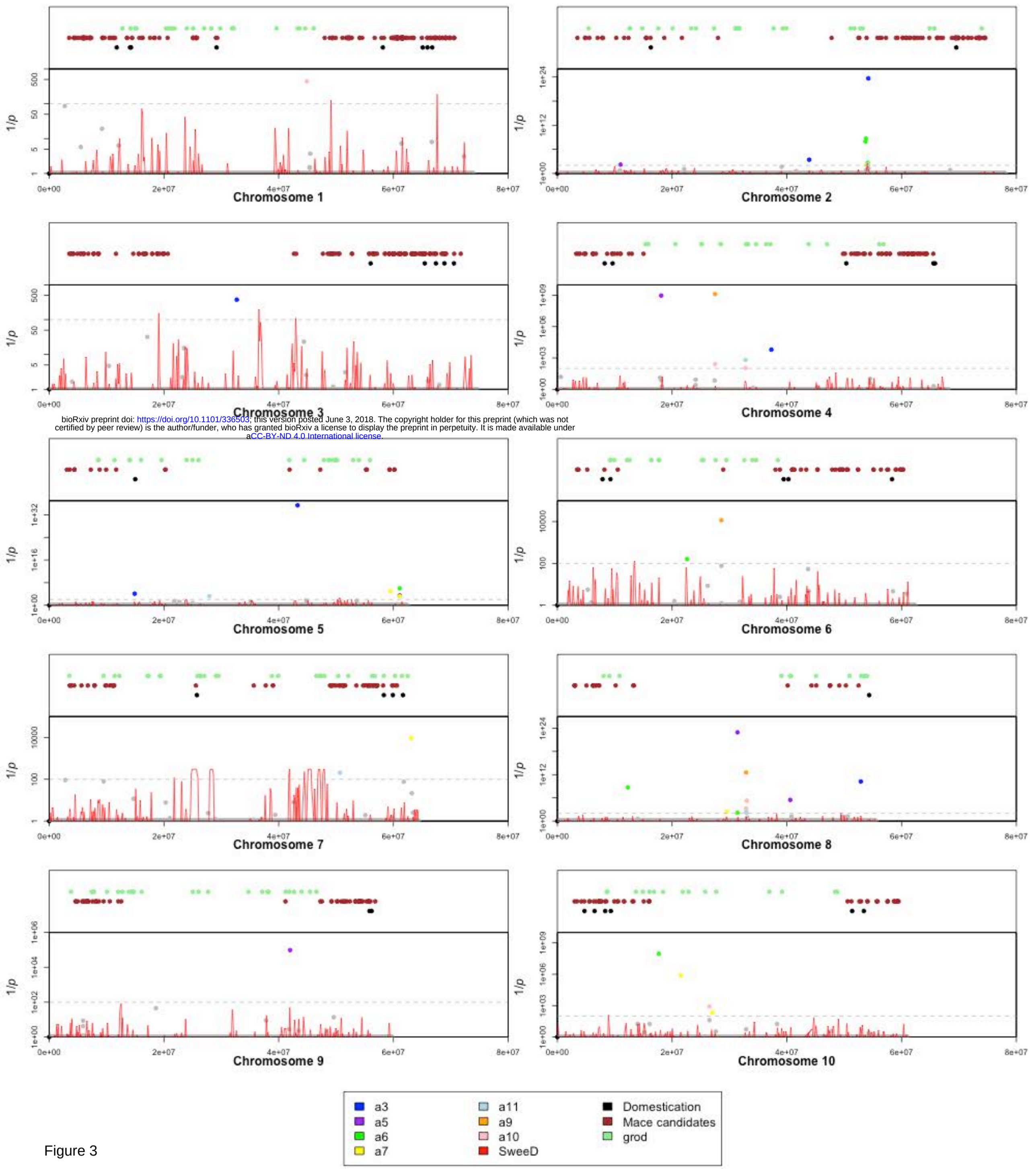

$R^2 = 0.518$
0.0800
0.0805
0.0810
0.0815
0.0820
0.0825
0.0830
GERP load recessive
7000
6000
5000
4000
3000
2000
1000
0
Years BP
Wild sorghum
S. bicolor type bicolor

Figure 2



bioRxiv preprint doi: https://doi.org/10.1101/336503; this version posted June 3, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Figure 3

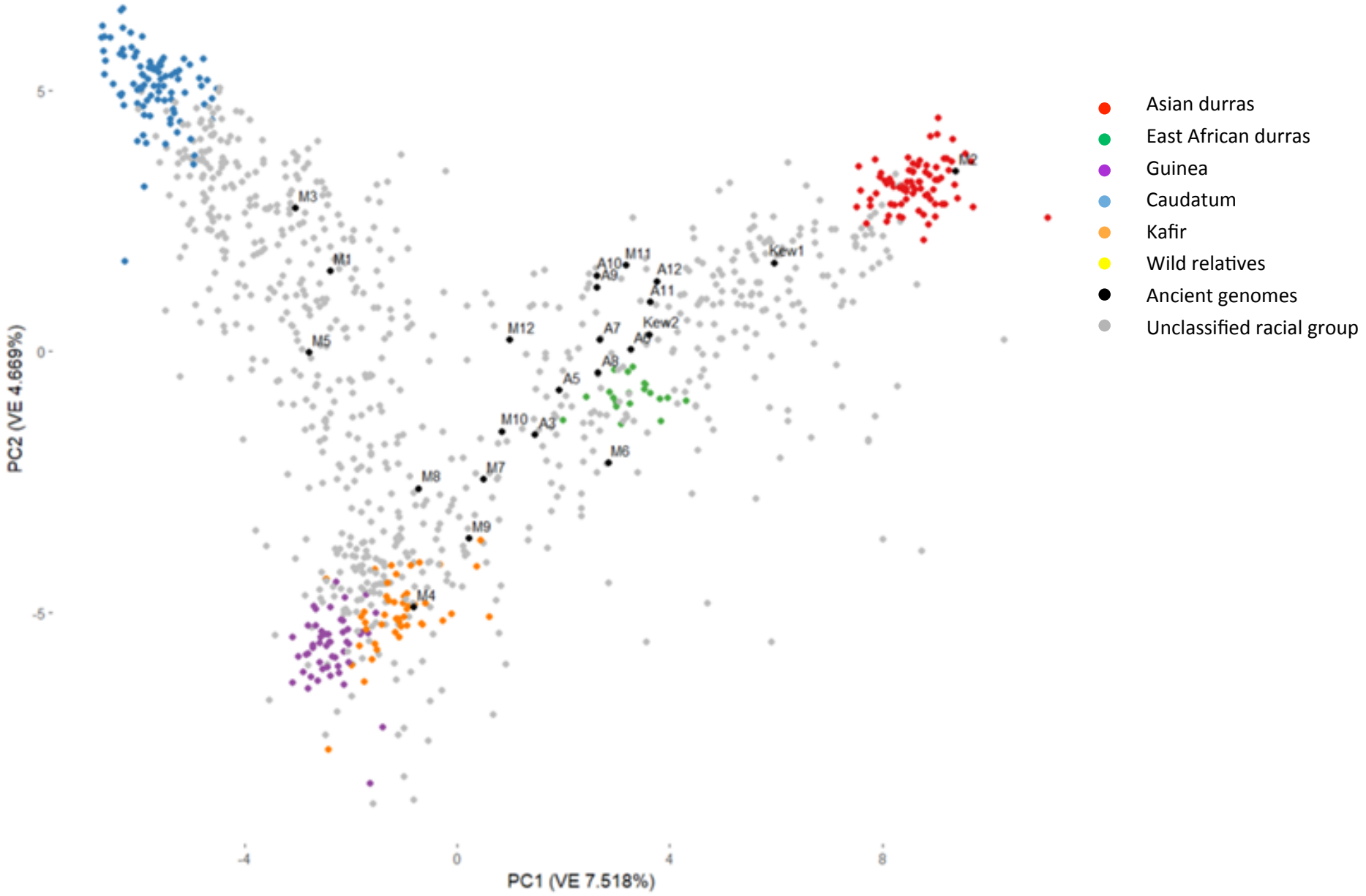

Asian durras
East African durras
Guinea
Caudatum
Kafir
Wild relatives
Ancient genomes
Unclassified racial group
PC1 (VE 7.518%)
PC2 (VE 4.669%)
-4
0
4
8
5
0
-5
M2
M3
M1
M5
M11
A10
A9
A12
A11
Kew1
Kew2
M12
A7
A6
A8
A5
M10
A3
M6
M7
M8
M9
M4

Figure 4

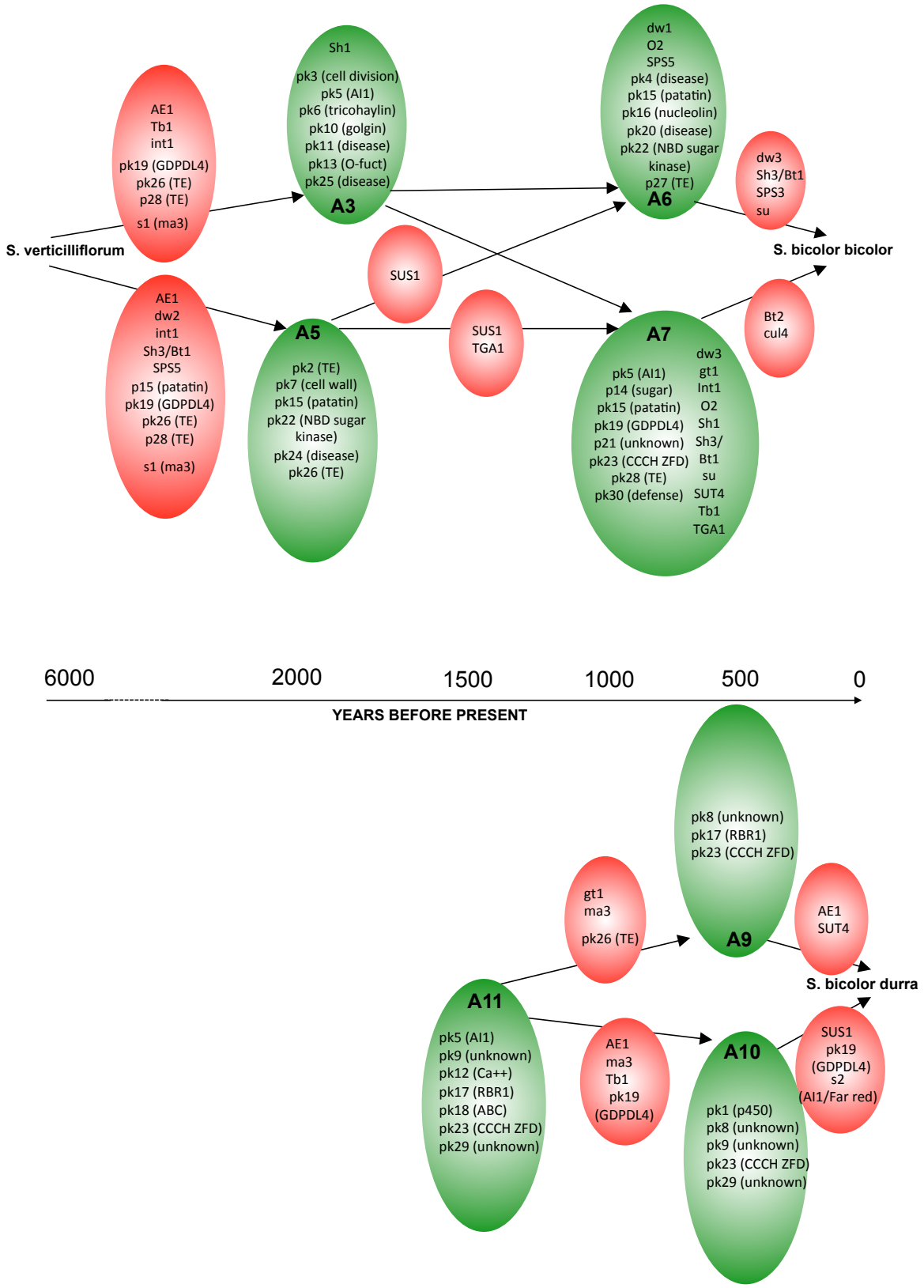

S. verticilliflorum
AE1
Tb1
int1
pk19 (GDPDL4)
pk26 (TE)
p28 (TE)
s1 (ma3)
AE1
dw2
int1
Sh3/Bt1
SPS5
p15 (patatin)
pk19 (GDPDL4)
pk26 (TE)
p28 (TE)
s1 (ma3)
A3
Sh1
pk3 (cell division)
pk5 (AI1)
pk6 (tricohaylin)
pk10 (golgin)
pk11 (disease)
pk13 (O-fuct)
pk25 (disease)
A5
pk2 (TE)
pk7 (cell wall)
pk15 (patatin)
pk22 (NBD sugar kinase)
pk24 (disease)
pk26 (TE)
SUS1
SUS1
TGA1
A6
dw1
O2
SPS5
pk4 (disease)
pk15 (patatin)
pk16 (nucleolin)
pk20 (disease)
pk22 (NBD sugar kinase)
p27 (TE)
A7
pk5 (AI1)
p14 (sugar)
pk15 (patatin)
pk19 (GDPDL4)
p21 (unknown)
pk23 (CCCH ZFD)
pk28 (TE)
pk30 (defense)
dw3
gt1
Int1
O2
Sh1
Sh3/
Bt1
su
SUT4
Tb1
TGA1
dw3
Sh3/Bt1
SPS3
su
Bt2
cul4
S. bicolor bicolor
6000
2000
1500
1000
500
0
YEARS BEFORE PRESENT
A11
pk5 (AI1)
pk9 (unknown)
pk12 (Ca++)
pk17 (RBR1)
pk18 (ABC)
pk23 (CCCH ZFD)
pk29 (unknown)
gt1
ma3
pk26 (TE)
AE1
ma3
Tb1
pk19
(GDPDL4)
A9
pk8 (unknown)
pk17 (RBR1)
pk23 (CCCH ZFD)
A10
pk1 (p450)
pk8 (unknown)
pk9 (unknown)
pk23 (CCCH ZFD)
pk29 (unknown)
AE1
SUT4
SUS1
pk19
(GDPDL4)
s2
(AI1/Far red)
S. bicolor durra

Figure 5

Table S1. Sample details of genomes sequenced in this study

Table S2. Radiocarbon dates on sorghum specimens or closely associated plant remains

Table S3 Summary statistics of fit to linear and exponential models of change in heterozygosity and GERP score in *S. bicolor bicolor.*

Table S4. Peaks of significant heterozygosity ratios (wild/cultivated) and associated annotated gene models. * refers to genes also found in Mace et al (24) candidate domestication loci gene set.

Table S5. *P* values of heterozygosity ratio peaks.

Table S6 *P* values of heterozygosity ratio peaks for known domestication loci.

Table S7 Peaks of high signal found with SweeD analysis and associated annotated gene models. * refers to genes also found in Mace et al (24) candidate domestication loci gene set.

Table S8 *P* values of gradient deviation in heterozygosity over time relative to genomic average.

Table S9 Genomic locations for 100kbp windows that differ in gerp load between genomes by more than two standard deviations.

Table S10 Phylogenetic congruence between type bicolor and type durra clades in selection candidate regions.

Figure S1 Frequency distributions of heterozygosity in genomes for 100 Kbp windows. A. *Sorghum verticilliflorum*. B. *S. bicolor* 'wild phenotype' (sample A3) 1765 yrs BP, C. *S. bicolor* type bicolor (sample A5) 1805 years BP, D. *S. bicolor* type bicolor (sample A6) 715 years BP, E. *S. bicolor* type bicolor (sample A7) 710 years BP, F. *S. bicolor* type bicolor BTX 623, G. *S. bicolor* type durra (sample A11) 1470 years BP, H *S. bicolor* type durra (sample A9) 505 years BP, I S. *bicolor* type durra (sample A10) 450 years BP, J *S bicolor* type durra modern.

Figure S2 Heterozygosity over time in *S bicolor* type durra

Figure S3 Additive, dominant and recessive model GERP load scores in *S bicolor* type bicolor and *S bicolor* type durra over time. Total GERP load calculated from variant sites with RS scores > 0, strongly deleterious GERP load calculated from variant sites with RS scores > 4. See methods for details on models.

Figure S4 Methylated site number in *S bicolor* type bicolor and *S bicolor* type durra over time.

Figure S5 Heterozygosity over time of regions containing genome-wide significant wild/cultivated ratios. Significant deviations from the genomic gradient of change over time shown only.

Figure S6 Heterozygosity over time of regions containing high SweeD scores. Significant deviations from the genomic gradient of change over time shown only.

Figure S7 Heterozygosity over time of regions containing domestication loci that have significantly reduced in heterogygosity relative to wild. Significant deviations from the genomic gradient of change over time shown only.

Figure S8. Maximum likelihood tree of whole genome sequence built in EXaML

Figure S9. D statistic analysis: P_1 = *S. bicolor* type bicolor BTX623, P_2 = sample displayed on X axis, P_3 = sample A3, P_4 halapense.

Figure S10 Potential genome rescue of descendents from ancestors by donors based on GERP scores. Red indicates the resultant change from combined score of ancestors and donors in regions of observed GERP load reduction in

descendents. Blue indicates the genome wide change in gerp score from combining ancestor and donor scores.

Figure S11. Standard model of loss of genetic diversity through drift combined with introgression over time. Arbitrary founding population of 2000 individuals simulated for 6000 generations to match the over all decrease observed in sorghum. Four models considered, no introgression (drift only), constant introgression (adding 0.000015 to the genetic diversity each generation). Dynamic introgression was defined where the gene flow (gf) contribution each generation is gf^{tf} , where t is the generation number and f the modification factor. Diminishing introgression, f is 1.0001, increasing introgression f is 0.99995. See methods for details of calculations.

Figure S12. Lost of heterozygosity through founder event model based on crop cycling. A. Sequential founding episodes based on 25% of harvest set aside for sowing (28) for various populations sizes (N). B. Gradients of diversity loss over time in the model and the observed gradient in sorghum with N obtained through linear regression of the model outputs.

Sample	Species	Age (median years cal. BP). BP= 2000 AD	source	Source id	Total reads LINES	Total reads/pairs	Reads mapped	% endogenous (genome)	Mean coverage at Q20
A3	*S. bicolor* type bicolor wild phenotype	1765	Qasr Ibrim	00/22008	896934328	224233582	200857980	89.57533399	9.56468
A5	*S. bicolor* type bicolor	1805	Qasr Ibrim	96/18082	932645300	233161325	204691251	87.78953842	6.62945
A6	*S. bicolor* type bicolor	*715*	Qasr Ibrim	86/128	889698924	222424731	182750309	82.16276498	8.13831
A7	*S. bicolor* type bicolor	710	Qasr Ibrim	84/155	839807884	209951971	188064725	89.57511763	7.7284
A8	*S. bicolor* intermediate durra/bicolor	890	Qasr Ibrim	84/142	1074259696	268564924	242003275	90.10978478	14.0648
A9	*S. bicolor* type durra	*505*	Qasr Ibrim	84/162	741640764	185410191	160463966	86.54538628	5.522
A10	*S. bicolor* type durra	450	Qasr Ibrim	84/55	486782488	121695622	108176150	88.89074909	4.02703
A11	*S. bicolor* type durra	1470	Qasr Ibrim	86/82	720632616	180158154	158338076	87.88837612	5.6919
A12	*S. bicolor* intermediate durra/bicolor	*450*	Qasr Ibrim	84/112	986445336	246611334	207934377	84.31663445	8.62576
Kew1	*S. bicolor* type bicolor	100	Kew Snowden Collection	Tsang Wai Fak 16366	1306192292	326548073	186573150	57.13497198	5.00975
Kew2	*S. bicolor* type bicolor	100	Kew Snowden Collection	Tenayac	1247071228	311767807	191726861	61.49668333	5.22517
M1	*S. bicolor* type caudatum	0	USDA	PI509071	492898852	123224713	217257730	88.15509678	18.3405
M2	*S. bicolor* type durra	0	USDA	PI562734	384496108	96124027	177009252	92.07336476	14.368
M3	*S. bicolor* type guinea	0	USDA	PI562938	425911276	106477819	196350280	92.20243326	16.0876
M4	*S. bicolor* type kafir	0	USDA	PI655976	385197696	96299424	177636336	92.23125571	16.0181
M5	*S. bicolor* type bicolor BTX623	0	USDA	PI659985	320880996	80220249	152004900	94.74222649	14.9171
M6	*S. verticilliflorum* var. *verticilliflorum*	0	USDA	PI520777	433237112	108309278	195610443	90.30179437	13.9573
M7	*S. verticilliflorum* var. *arundinaceum*	0	USDA	PI532564	403234860	100808715	167378436	83.01784027	12.0593
M8	*S. verticilliflorum* var. *aethiopicum*	0	USDA	PI535995	408447108	102111777	172471549	84.45232963	13.0077
M9	*S. propinquum*	0	USDA	PI653737	320834680	80208670	147017274	91.64674717	11.2134
M10	*S. verticilliflorum* var. *virgatum*	*75*	Vinall 11.7.1929 (UCL Archaeobotany)	*S. virgatum*	321596756	80399189	94103477	58.52265313	6.75004
M11	*S. bicolor* type drummondii	0	USDA	PI653734	674746364	168686591	293908369	87.1166959	23.2651
M12	*S. halapense*	0	USDA	Grif 16307	420386852	105096713	193426298	92.02300076	14.2121

Table S1 Sample details

Table S2. Radiocarbon dates on sorghum specimens or closely associated plant remains

Sample	Notes on date	Lab Code	Date BP	error	cal. BP Start*	cal. BP Finish*	Median years before AD 2000	Source id	Context
A3	direct date on sorghum	OxA-14892	1789	27	1780	1620	**1765**	00/22008	00/22008
A5	direct date on sorghum	OxA-14818	1818	32	1820	1710	**1805**	96/18082	96/18082
A6	direct date on sorghum	Beta-491610					**715**	86/128	86/128
A7	direct date on sorghum	Beta-491611					**710**	84/155	84/155
A8	date on *Vigna* (Room 8 pit 1028)	OxA-14757	906	27	910	780	**890**	84/142	House 785; Room 4; Level 8; in Floor 7
A9	direct date on sorghum	Beta-491612					**505**	84/162	84/162
A10	direct date on sorghum	Wk-21087	349	29	470	320	**450**	84/55	pit 932
A11	Direct date on sorghum?	OxA-1023	1440	50	1530	1310	**1470**	86/82	pit associated with X-horizon

*= Before AD 1950

Function fit	Parameter	Heterozygosity			GERP load recessive		
Straight line y=bx+a		value	standard error	p-value	value	standard error	p-value
	a	3.05E-04	6.80E-05	0.0041	8.10E-02	1.80E-04	1.07E-12
	b	4.00E-07	2.90E-08	9.20E-06	-1.70E-07	7.30E-08	0.0683
	AIC			-114			-87.039
	MSE			1.76E-08			9.89E-08
Exponential y=a.exp(b)		value	standard error	p-value	value	standard error	p-value
	a	0.00048	7.90E-05	0.042	8.10E-02	1.80E-04	1.08E-12
	b	0.00029	3.15E-05	9.30E-05	-2.10E-06	9.10E-07	0.0686
	AIC			-108			-87.047
	MSE			3.77E-08			9.88E-08

Table S3 - parameter values for curves fit to bicolor heterozygosity and GERP data (versus Years BP)

chromosome	window	peak	a3 (bic 1765)	a5 (bic 1805)	a6 (bic 715)	a7 (bic 710)	a11 (du 1470)	a9 (du 505)	a10 (du 450)
1	44900000	pk1	0.102805669	0.361722461	0.037909655	0.672815666	0.488542667	0.131199566	3.07E-06
2	11000000	pk2	0.41358704	8.09E-06	0.572475817	0.441985835	0.154849864	0.234619463	0.344378095
2	43900000	pk3	4.94E-07	0.068120687	0.113563673	0.312212401	0.323721596	0.140345085	0.127117347
2	53700000	pk4	0.344474996	0.401256338	1.52E-11	0.019124686	0.03962074	0.3485805	0.158021188
2	54200000	pk5	2.91E-27	0.420599337	0.000105802	8.86E-06	7.54E-05	0.798886327	0.77234906
3	32700000	pk6	3.58E-06	0.122276411	0.007236644	0.511002154	0.250304643	0.566516074	0.279493597
4	18100000	pk7	0.11396104	1.58E-12	0.088156494	0.000195191	0.471447038	0.41702393	0.372331665
4	27500000	pk8	0.261286593	0.205473343	0.14112128	0.46524292	0.003680721	1.14E-12	5.66E-06
4	32800000	pk9	0.24931856	0.362688509	0.058495099	0.299449624	2.21E-06	0.058438023	1.33E-05
4	37300000	pk10	2.31E-07	0.019809067	0.533628263	0.157409081	0.594011942	0.714415363	0.722213508
5	14900000	pk11	1.35E-07	0.25271921	0.337898946	0.350962433	0.503030099	0.353336892	0.538949648
5	27900000	pk12	0.713062847	0.225616734	0.073535584	0.137585802	9.34E-07	0.034101907	0.082247478
5	43300000	pk13	7.16E-39	0.227277619	0.353860854	0.42828102	0.5041582	0.495277017	0.482806499
5	59500000	pk14	0.51063073	0.361951397	0.624818174	1.73E-08	0.35261265	0.310154615	0.191120249
5	61100000	pk15	0.060786667	5.66E-07	1.74E-09	1.21E-06	0.106943639	0.083185043	0.216052513
6	22600000	pk16	0.485480771	0.293894564	1.00E-05	0.366941418	0.218434017	0.107525029	0.02336006
6	28600000	pk17	0.171185183	0.114355189	0.043428406	0.017284366	2.09E-05	1.38E-07	0.001290742
7	50700000	pk18	0.551471147	0.665910386	0.254773419	0.724932256	7.57E-06	0.800061835	0.36169038
7	63100000	pk19	0.081917263	0.178246552	0.677009647	1.68E-07	0.7213842	0.560745808	0.677619632
8	12300000	pk20	0.003378476	0.389234972	3.59E-12	0.689368891	0.602756238	0.032440655	0.147217192
8	29500000	pk21	0.464342822	0.268334064	0.550375439	7.12E-06	0.107161053	0.352865059	0.22445346
8	31400000	pk22	0.320457043	2.11E-26	1.23E-05	0.161024806	0.238387641	0.262500125	0.21551472
8	32900000	pk23	0.452847218	0.422118402	0.360210465	1.18E-06	1.08E-05	5.16E-16	1.05E-06
8	40600000	pk24	0.10546731	6.14E-09	0.004401777	0.286047016	0.234993339	0.293012606	0.454354348
8	52900000	pk25	1.07E-13	0.522375209	0.308411152	0.15312801	0.372493801	0.382384948	0.480322344
9	42000000	pk26	0.105550876	1.74E-08	0.206695471	0.206278078	0.497667977	0.00371231	0.006383025
10	17700000	pk27	0.47452556	0.248307926	1.82E-11	0.028057542	0.356069589	0.372543624	0.381636121
10	21500000	pk28	0.040683815	0.003257893	0.035047978	2.20E-09	0.049966079	0.055062808	0.033568321
10	26500000	pk29	0.340520232	0.109276198	0.501928453	0.028309324	4.03E-05	0.001987998	1.95E-06
10	27000000	pk30	0.459154643	0.23804888	0.069064252	7.95E-06	0.396138422	0.160478352	0.279727425

Table S4 *p* values for windows containing significant reduction in heterozygosity relative to *S. verticilliflorum*

chromosome	window	peak	start	stop	Uniparc code	Unparc code	Sb code	SORBI code	Gene description	
1	44900000	pk1	44830820	44831431	UPI0001A82246	UPI0001A82246	Sb01g026525	C5WP69_SORBI	unknown function (DUF1645)	
			44860037	44861635	UPI0001A82247	UPI0001A82247	Sb01g026530	C5WP70_SORBI	cytochrome p450	
2	11000000	pk2	10974343	10978599	UPI0001A83EC8	UPI0001A83EC8	Sb02g008271	C5X2V0_SORBI	reverse transcriptase	
			10981018	10981874	UPI0001A83EC9	UPI0001A83EC9	Sb02g008311	C5X2V1_SORBI	transposase	
2	43900000	pk3	43786194	43787234	UPI0001A842F4	UPI0001A842F4	Sb02g018043	C5X843_SORBI	chromsome segregation ATPase	
			43788873	43789595	UPI0001A842F5	UPI0001A842F5	Sb02g018110	C5X844_SORBI	transposase	
2	53700000	pk4	53484601	53494615	*UPI0001A83D56	UPI0001A83D56	Sb02g021535	C5X9S1_SORBI	RGA3 disease resistance protein (NB-ARC, LRR domain)	
			53668339	53669764	UPI0001A83D57	UPI0001A83D57	Sb02g021540	C5X9S2_SORBI	transposase	
			53671197	53673521	UPI0001A838F5	UPI0001A838F5	Sb02g021550	C5X9S3_SORBI	RGA3 disease resistance protein (NB-ARC domain)	
			53687315	53688481	UPI0001A838F6	UPI0001A838F6	Sb02g021560	C5X9S4_SORBI	RGA3 disease resistance protein (LRR domain)	
			53693445	53694484	UPI0001A838F7	UPI0001A838F7	Sb02g021570	C5X9S5_SORBI	RGA3 disease resistance protein (PKc domain)	
2	54200000	pk5	54156595	54156834	UPI0001A83D5D	UPI0001A83D5D	Sb02g021853	C5X9T6_SORBI	unknown	
			54169579	54170586	UPI0001A83D5E	UPI0001A83D5E	Sb02g021856	C5X9T7_SORBI	Polynucleotidyl transferase ribonuclease	
			54173272	54173781	UPI0001A83D5F	UPI0001A83D5F	Sb02g021860	C5X9T8_SORBI	unknown	
			54179475	54180150	UPI0001A83D60	UPI0001A83D60	Sb02g021911	C5X9T9_SORBI	Anther Indehiscence 1	
3	32700000	pk6	32649562..32651401		LOC110433684				trichohyalin-like	
4	18100000	pk7	17966716..17967603		LOC8155713				vegetative cell wall protein gp1	
			17982112..17982948		LOC110434760				serine/arginine repetitive matrix protein 1-like	
			18046591..18047799		LOC110434315				unknown function (DUF1668 domain)	
4	27500000	pk8	27402052	27402504	UPI0001A86024	UPI0001A86024	Sb04g014271	C5Y0N2_SORBI	unknown	
4	32800000	pk9	32660496	32661207	UPI0001A8598F	UPI0001A8598F	Sb04g014491	C5Y0N6_SORBI	unknown function	
4	37300000	pk10	37359135	37371246	UPI0001C80D84;Ontology_term=GO(		Sb04g016070		golgin a5 type protein	
5	14900000	pk11	14885427	14891094	UPI0001A863DE	UPI0001A863DE	Sb05g008160	C5Y1I6_SORBI	RPP-13 like disease resistance	
5	27900000	pk12	27791293	27795021	UPI0001A865C1	UPI0001A865C1	Sb05g013400	C5Y282_SORBI	GDT1 like protein	Ca transporter
			27797941	27798643	UPI0001A865C2	UPI0001A865C2	Sb05g013410	C5Y283_SORBI	Rho binding protein	Regulatory transcription inhibitor
5	43300000	pk13	43,247,187	43,269,799	LOC8075771				O-Fuct like	auxin independent growth promoter
5	59500000	pk14	59411298	59414857	UPI0001A8660D	UPI0001A8660D	Sb05g025890	C5Y7E1_SORBI	lipase	
			59416966	59417283	UPI0001A8660E	UPI0001A8660E	Sb05g025900	C5Y7E2_SORBI	glutaredoxin C10	
			59437739	59438068	UPI0001A8660F	UPI0001A8660F	Sb05g025910	C5Y7E3_SORBI	glutaredoxin C10	
			59452563	59457270	UPI0001A86610	UPI0001A86610	Sb05g025915	C5Y7E4_SORBI	galactose oxidase	
			59458058	59462321	UPI0001A86611	UPI0001A86611	Sb05g025920	C5Y7E5_SORBI	peptide chain release factor APG3	
			59463266	59463718	UPI0001A86639	UPI0001A86639	Sb05g025930	C5Y7E6_SORBI	RALF like protein	arrests root development
			59464960	59466737	UPI0001A86C54	UPI0001A86C54	Sb05g025940	C5Y7E7_SORBI	alkB	DNA repair
			59474035	59475117	UPI0001A86C55	UPI0001A86C55	Sb05g025945	C5Y7E8_SORBI	pollen extensin like	
			59475347	59477611	UPI0001A86C56	UPI0001A86C56	Sb05g025950	C5Y7E9_SORBI	pollen extensin like	
5	61100000	pk15	61009529	61010503	UPI0001A86440	UPI0001A86440	Sb05g026965	C5Y826_SORBI	RPP13 disease resistance protein NBS-LRR	
			61015798	61018316	UPI0001A86441	UPI0001A86441	Sb05g026970	C5Y827_SORBI	RPP13 disease resistance protein NBS-LRR	
			61028126	61029756	UPI0001A8643F	UPI0001A8643F	Sb05g026950	C5Y825_SORBI	dirigent protein	disease response involvig lignification
			61048282	61050689	UPI0001A86A3A	UPI0001A86A3A	Sb05g026990	C5Y829_SORBI	patatin	storage protein and fatty acid metabolism
			61057619	61058128	UPI0001A86A3B	UPI0001A86A3B	Sb05g026993	C5Y830_SORBI	isopentenyl transferase	
			61061641	61062092	UPI0001A86A3C	UPI0001A86A3C	Sb05g026996	C5Y831_SORBI	patatin	
			61082967	61094282	UPI0001A86442	UPI0001A86442	Sb05g027000	C5Y832_SORBI	RPP13 disease resistance protein NBS-LRR	
			61097301	61098179	UPI0001A86443	UPI0001A86443	Sb05g027005	C5Y833_SORBI	transposable element	
6	22600000	pk16	22725443..22735214		LOC110436433				nucleolin	
6	28600000	pk17	28027121	28028229	UPI0001A8715A	UPI0001A8715A	Sb06g010020	C5YDW7_SORBI	RBR1,"similar to Retinoblastoma related protein RBR1"	
7	50700000	pk18	50563493	50568895	UPI0001A87F06	UPI0001A87F06	Sb07g019540	C5YKN9_SORBI	ABC transporter	
			50772772	50778650	UPI0001A87F07	UPI0001A87F07	Sb07g019740	C5YKP0_SORBI	ABC transporter	
			50794332	50794661	UPI0001A87F08	UPI0001A87F08	Sb07g019745	C5YKP1_SORBI	transposable element	
			50896473	50902950	*UPI0001A87F09	UPI0001A87F09	Sb07g019750	C5YKP2_SORBI	ABC transporter	
7	63100000	pk19	63002738	63004800	UPI0001A87EBC	UPI0001A87EBC	Sb07g028040	C5YJ45_SORBI	MFS, putative peptide transporter	
			63007131	63012078	UPI0001A87A47	UPI0001A87A47	Sb07g028050	C5YJ46_SORBI	glycerophosphodiester phosphodiesterase GDPDL4	
			63019028	63020726	UPI00022071D4;b	UPI00022071D4	Sb07g028060		fibrous sheath CABYR-binding protein	
			63027930	63028675	UPI0001A87EBD	UPI0001A87EBD	Sb07g028065	C5YJ47_SORBI	unknown function	
			63029524	63033671	UPI0001A87A48	UPI0001A87A48	Sb07g028070	C5YJ48_SORBI	SWIB domain protain (p53 associated)	
			63059608	63062274	UPI0001A87EBE	UPI0001A87EBE	Sb07g028080	C5YJ49_SORBI	serine--glyoxylate aminotransferase	
			63062573	63062995	UPI0001A87A49	UPI0001A87A49	Sb07g028090	C5YJ50_SORBI	PEF family (apoptosis associated)	
			63073408	63073999	UPI0001A87A4A	UPI0001A87A4A	Sb07g028095	C5YJ51_SORBI	Peptidase M14 Succinylglutamate desuccinylase	
			63088404	63095390	UPI0001A87A4B	UPI0001A87A4B	Sb07g028100	C5YJ52_SORBI	wall-associated receptor kinase 5	
8	12300000	pk20	12228641	12230086	UPI0001A881AC	UPI0001A881AC	Sb08g007200	C5YTS4_SORBI	TNP1 like protein	
			12259170	12259505	UPI0001A881AD	UPI0001A881AD	Sb08g007210	C5YTS5_SORBI	TNP2 like protein	
			12279014	12286998	UPI0001A881AE	UPI0001A881AE	Sb08g007220	C5YTS6_SORBI	la-related 6B protein	
			12293716	12301627	UPI0001A880E7	UPI0001A880E7	Sb08g007230	C5YTS7_SORBI	outer envelope pore protein 37 chloroplastic	
			12351604	12355600	UPI0001A880E8	UPI0001A880E8	Sb08g007240	C5YTS8_SORBI	2 alkenal reductase	defense
			12367068	12370954	UPI0001A880E9	UPI0001A880E9	Sb08g007243	C5YTS9_SORBI	2 alkenal reductase	
			12372086	12372376	UPI0001A880EA	UPI0001A880EA	Sb08g007246	C5YTT0_SORBI	unknown	
			12385186	12386774	UPI0001A881AF	UPI0001A881AF	Sb08g007250	C5YTT1_SORBI	obtusifoliol 14 alpha demethylase	
			12388327	12390954	UPI0001A880EB	UPI0001A880EB	Sb08g007260	C5YTT2_SORBI	transposase (transposon)	

8	29500000 pk21	29483080	28483220	none					possible ncRNA
8	31400000 pk22	31451800	31452274	UPI0001A8836E	UPI0001A8836E	Sb08g012126	C5YNI7_SORBI		NBD sugar kinase HSP70
8	32900000 pk23	32307877	32329902	UPI0001A8824A	UPI0001A8824A	Sb08g012360	C5YNJ3_SORBI		zinc finger CCCH domain protein
8	40600000 pk24	40413311	40414000	UPI0001A8845B	UPI0001A8845B	Sb08g015335	C5YNU4_SORBI		RGA2 LRR disease resistance
		40414044	40416104	UPI0001A8845C	UPI0001A8845C	Sb08g015337	C5YNU5_SORBI		RGA2 NB-LRR disease resistance
		40426252	40430241	UPI0001A8845D	UPI0001A8845D	Sb08g015340	C5YNU6_SORBI		RGA2 NB-LRR disease resistance
		40602695	40619018	UPI0001A8845E	UPI0001A8845E	Sb08g015350	C5YNU7_SORBI		RGA2 NB-LRR disease resistance
		40625087	40634676	UPI0001A8845F	UPI0001A8845F	Sb08g015360	C5YNU8_SORBI		RAD-51 DNA repair
		40637341	40638570	UPI0001A882CF	UPI0001A882CF	Sb08g015370	C5YNU9_SORBI		methyl transferase
8	529000000 pk25	52813133	52813255	UPI0001A88235	UPI0001A88235	Sb08g021248	C5YRX2_SORBI		unknown function
		52821960	52823171	UPI0001A88236	UPI0001A88236	Sb08g021250	C5YRX3_SORBI		unknown function
		52832618	52849272	UPI0001A88103	UPI0001A88103	Sb08g021260	C5YRX4_SORBI		achilleol B synthase
		52877001	52880228	UPI0001A88104	UPI0001A88104	Sb08g021270	C5YRX5_SORBI		serine/threonine-protein kinase PBL13
		52886168	52889779	UPI0001A88105	UPI0001A88105	Sb08g021280	C5YRX6_SORBI		RGA2 NB-LRR disease resistance
		52899503	52904524	UPI0001A88106	UPI0001A88106	Sb08g021290	C5YRX7_SORBI		RGA3 disease resistance protein (NB-ARC domain)
9	42000000 pk26	41818067	41821596	UPI0001A88998	UPI0001A88998	Sb09g016555	C5YWB8_SORBI		unknown function
		41837290	41838894	UPI0001A88999	UPI0001A88999	Sb09g016560	C5YWB9_SORBI		transposase
		42022771	42023844	UPI0001A88C3B	UPI0001A88C3B	Sb09g016570	C5YWC0_SORBI		myb-related protein 330
		42026486	42027580	UPI0001A88C3C	UPI0001A88C3C	Sb09g016580	C5YWC1_SORBI		unknown function
		42093322	42093834	UPI0001A88C3D	UPI0001A88C3D	Sb09g016590	C5YWC2_SORBI		GRF zinc finger protein
		42094740	42097538	UPI0001A8899A	UPI0001A8899A	Sb09g016595	C5YWC3_SORBI		MuDR transposase
		42098013	42101984	UPI0001A8899B	UPI0001A8899B	Sb09g016600	C5YWC4_SORBI		RanBP1 (chromosome condensation)
10	17700000 pk27	17513261	17515058	UPI0001A895DC	UPI0001A895DC	Sb10g011850	C5Z1L6_SORBI		anthranilate O-methyltransferase 3
		17580569	17581168	UPI0001A895DD	UPI0001A895DD	Sb10g011916	C5Z1L7_SORBI		transposon protein
		17731600	17734506	UPI0001A88F2F	UPI0001A88F2F	Sb10g012050	C5Z1L9_SORBI		LRR receptor-like serine/threonine-protein kinase GSO2
10	21500000 pk28	21357175	21367692	UPI0001A8963D	UPI0001A8963D	Sb10g013495	C5Z2B4_SORBI		TNP2-like protein
		21385108	21386957	UPI0001A8963E	UPI0001A8963E	Sb10g013500	C5Z2B5_SORBI		putative receptor-like protein kinase
10	26500000 pk29	26207429	26207653	UPI0001A89698	UPI0001A89698	Sb10g015631	C5Z2G3_SORBI		unknown function
10	27000000 pk30	27060330	27070160	UPI0001A8902A		Sb10g015690	C5Z2G5_SORBI		U-box containing protein

Table S5 Regions of genome-wide significance in reduction of heterozygosity relative to *S. verticilliflorum* and associated genes

* indicates correspondance with the Mace et al (24) candidate domestication gene list

chromosome	window	gene	a3 (bic 1765)	a5 (bic 1805)	a6 (bic 715)	a7 (bic 710)	a11 (du 1470)	a9 (du 505)	a10 (du 450)
7	59800000	dw3	0.7169588	0.489916345	0.886234736	0.001029348	0.477686784	0.363034078	0.676461916
9	57100000	dw1	0.430326598	0.305130691	0.00230527	0.368279324	0.180977448	0.341941378	0.395721262
6	39400000	dw2	0.665201583	0.057426669	0.298814529	0.317679894	0.610788469	0.378785607	0.311719579
1	12100000	Sh1	0.001643632	0.297385872	0.028911337	0.000212513	0.354835471	0.07686203	0.06234129
3	57300000	Sh2	0.513833464	0.503304376	0.475469287	0.719246492	0.207224899	0.217029358	0.161350823
4	6900000	Sh3/Bt1	0.525863822	0.220438392	0.553620438	0.001634715	0.067144308	0.070605796	0.081976552
7	24600000	Bt2	0.582609529	0.501609592	0.412557788	0.801374592	0.321041498	0.356407175	0.075252275
1	12000000	SbWRKY	0.361395651	0.427090906	0.603085905	0.203227315	0.486529559	0.302235506	0.379554752
4	51200000	AE1	0.22551433	0.29453358	0.311256281	0.034073862	0.677054877	0.683438884	0.081427538
3	73000000	cul4	0.701952197	0.604183218	0.578502305	0.926511526	0.672861455	0.403217476	0.440427355
1	66700000	gt1	0.097794597	0.316054932	0.62376622	0.000168411	0.617927673	0.247660715	0.415874132
3	67300000	int1	0.24549123	0.384413387	0.494155025	0.004699184	0.127425119	0.01658232	0.062109557
6	40300000	ma1	0.617705339	0.532350457	0.590362723	0.593731789	0.571373012	0.553418355	0.594789672
1	68000000	ma3	0.246531398	0.434353232	0.517946331	0.037973221	0.833435067	0.489423682	0.412132698
6	6800000	ma6	0.15419183	0.357669236	0.101545909	0.555980974	0.807889865	0.583139814	0.679106223
10	52300000	Nud	0.450228785	0.446656636	0.010462224	0.483913261	0.340556622	0.437064507	0.319686294
6	5300000	O2	0.013429001	0.089494808	0.000287218	0.003409689	0.237697548	0.010100032	0.002282872
6	59800000	Pa1	0.198798653	0.422004202	0.50339961	0.769585741	0.180607303	0.49626172	0.482938499
3	69600000	SHP	0.576035163	0.485634537	0.442796826	0.682561457	0.083018346	0.07321182	0.094031087
3	71200000	SPS1	0.785464458	0.797677336	0.857555457	0.903142979	0.765446871	0.662016811	0.777044825
4	5700000	SPS2	0.695212274	0.348024273	0.56633189	0.056854355	0.085577796	0.1669642	0.175319527
5	13000000	SPS3	0.611808756	0.569310494	0.870647378	0.756971064	0.741684177	0.773642816	0.841701362
9	57500000	SPS4	0.208514324	0.331713243	0.558643711	0.574416919	0.309467581	0.3428043	0.361862021
10	54300000	SPS5	0.63860949	0.21920424	0.000651432	0.348786362	0.179840872	0.081184781	0.058967394
10	3800000	sss1	0.375041814	0.55845213	0.684414407	0.597448311	0.473798407	0.525436039	0.587421989
7	63400000	su	0.691383152	0.335374091	0.899667869	0.000604092	0.683502612	0.283731511	0.464984252
10	5800000	suc1	0.589443244	0.637865314	0.62481392	0.406182939	0.732509736	0.6082817	0.711589757
1	59600000	SUS1	0.102857534	0.245293066	0.0583269	0.555907938	0.176826781	0.293207289	0.352236403
4	67900000	SUS2	0.458986562	0.576702496	0.163889482	0.417223531	0.34109847	0.40643326	0.369316348
10	68700000	SUS3	0.098485418	0.238211851	0.397000681	0.470113482	0.223766466	0.311699977	0.214376648
1	68900000	SUT1	0.566041623	0.531485138	0.655006916	0.394999175	0.54577088	0.394546242	0.451672203
4	67600000	SUT2	0.567902772	0.53458443	0.474764929	0.063559061	0.28588962	0.292630349	0.45073245
1	28300000	SUT3	0.218145354	0.201505373	0.351179354	0.744264086	0.482446235	0.325694104	0.535576405
8	55400000	SUT4	0.317691189	0.517536698	0.103511984	0.007573251	0.433216268	0.743111943	0.58091866
1	9100000	TB1	0.014057578	0.573426704	0.093823381	0.003083594	0.511843222	0.260897205	0.158842179
7	61800000	TGA1	0.470174403	0.656829436	0.857173271	2.03E-05	0.399043914	0.321679462	0.582541775
2	71900000	vrs1	0.474901762	0.266734382	0.395277557	0.307403979	0.366941678	0.433615896	0.337698283
10	1900000	Wx	0.676418026	0.587602331	0.747884835	0.725536392	0.468105714	0.459480367	0.564016491
2	14400000	Wx_Chr2	0.73866598	0.67895781	0.66040409	0.570248694	0.348520009	0.502490469	0.536244356

Table S6 *p* values for reduction in heterozygosity in windows containing domestication loci observed in archaeological accessions relative to *S. verticilliflorum*.

chromosome	position	likelihood	name	start	end	Uniparc id	Uniparc id	Sb code	SORBI code	gene description
1	67625596	186.9738	s1	67527437	67528255	UPI0001A82E0A	UPI0001A82E0A	Sb01g044420	C5WUS5_SORBI	unknown function
				67532193	67537252	UPI0001A8295D	UPI0001A8295D	Sb01g044430	C5WUS6_SORBI	Pumilio RNA binding protein
				67538460	67540436	UPI0001C80BA9;Or	UPI0001C80BA9	Sb01g044440		ras-related protein RABH1b-like golgi trafficking
				67555746	67557224	UPI0001A8295E	UPI0001A8295E	Sb01g044450	C5WUS7_SORBI	WW domain protein
				67560187	67564611	UPI0001A8295F	UPI0001A8295F	Sb01g044460	C5WUS8_SORBI	conserved oligomeric Golgi complex subunit 8
				67568809	67573214	UPI0001C80BAA;Or	UPI0001C80BAA	Sb01g044470		mitochondrial adenine nucleotide transporter BTL3
				67576601	67583786	UPI0001A82960	UPI0001A82960	Sb01g044480	C5WUS9_SORBI	calmodulin-binding transcription activator 1 isoform X2
				67585849	67586537	UPI0001A82E0B	UPI0001A82E0B	Sb01g044485	C5WUT0_SORBI	reverse transcriptase
				67590805	67594920	UPI0001A82961	UPI0001A82961	Sb01g044490	C5WUT1_SORBI	unknown function
				67606065	67606893	UPI0001A82962	UPI0001A82962	Sb01g044500	C5WUT2_SORBI	trypsin like peptidase
				67612260	67612981	UPI0001A82E0C	UPI0001A82E0C	Sb01g044505	C5WUT3_SORBI	unknown function
				67623006	67629132	UPI0001A82963	UPI0001A82963	Sb01g044510	C5WUT4_SORBI	unknown function
				67631156	67632142	UPI0001A82964	UPI0001A82964	Sb01g044515	C5WUT5_SORBI	CCCH domain zinc finger proetin
				67653779	67664222	UPI0001A82965	UPI0001A82965	Sb01g044520	C5WUT6_SORBI	trypsin like peptidase
				67666492	67670699	UPI0001A829C6	UPI0001A829C6	Sb01g044530	C5WUT7_SORBI	ubiquitin carboxyl-terminal hydrolase 3
				67681955	67685954	UPI0001A829C7	UPI0001A829C7	Sb01g044540	C5WUT8_SORBI	LRR domain protein
				67687809	67690043	UPI0001A829C8	UPI0001A829C8	Sb01g044550	C5WUT9_SORBI	ras-related protein Rab11D
				67693174	67695434	UPI0001A829C9	UPI0001A829C9	Sb01g044560	C5WUU0_SORBI	2-carboxy-1,4-naphthoquinone phytyltransferase, chloroplastic
2	54037240	284.8075	s2	53900589	53917120	UPI0001A83D58	UPI0001A83D58	Sb02g021770	C5X9S6_SORBI	nudix hydrolase 20, chloroplastic
				53917736	53923682	UPI0001A83D59	UPI0001A83D59	Sb02g021780	C5X9S7_SORBI	UDP-glucose 4-epimerase
				53927463	53932982	UPI0001A83D5A	UPI0001A83D5A	Sb02g021790	C5X9S8_SORBI	E3 ubiquitin-protein ligase
				53972111	53976091	UPI0001C80E15;Or	UPI0001C80E15	Sb02g021800		receptor-like serine/threonine-protein kinase SD1-8
				53991898	53993342	UPI0001A838F8	UPI0001A838F8	Sb02g021810	C5X9S9_SORBI	protein FAR-RED IMPAIRED RESPONSE 1-like
				53994545	54014672	UPI0001A838F9	UPI0001A838F9	Sb02g021820	C5X9T0_SORBI	protein FAR1-RELATED SEQUENCE 5-like
				54016136	54017327	UPI0001A83D5B	UPI0001A83D5B	Sb02g021830	C5X9T1_SORBI	unknown function
				54020593	54022973	UPI0001A83D5C	UPI0001A83D5C	Sb02g021835	C5X9T2_SORBI	unknown function
				54032811	54033740	UPI0001A838FA	UPI0001A838FA	Sb02g021840	C5X9T3_SORBI	Polynucleotidyl transferase ribonuclease H-like superfamily protein
				54090301	54090747	UPI0001A838FC	UPI0001A838FC	Sb02g021850	C5X9T5_SORBI	NBD sugar kinase HSP70
				54156595	54156834	UPI0001A83D5D	UPI0001A83D5D	Sb02g021853	C5X9T6_SORBI	unknown
				54169579	54170586	UPI0001A83D5E	UPI0001A83D5E	Sb02g021856	C5X9T7_SORBI	Polynucleotidyl transferase ribonuclease
				54173272	54173781	UPI0001A83D5F	UPI0001A83D5F	Sb02g021860	C5X9T8_SORBI	unknown
				54179475	54180150	UPI0001A83D60	UPI0001A83D60	Sb02g021911	C5X9T9_SORBI	Anther Indehiscence 1
3	19095154	154.0901	s3	19004452	19005520	UPI0001A845AF	UPI0001A845AF	Sb03g014221	C5XJY4_SORBI	transposase
				19147217	19147932	UPI0001A851BF	UPI0001A851BF	Sb03g014261	C5XJY5_SORBI	transposase
				19148124	19148417	UPI0001A851C0	UPI0001A851C0	Sb03g014301	C5XJY6_SORBI	transposase
				19151124	19159094	UPI0001A851C1	UPI0001A851C1	Sb03g014340	C5XJY7_SORBI	thiamine pyrophosphokinase 2
				19161565	19162878	UPI0001A851C2	UPI0001A851C2	Sb03g014350	C5XJY8_SORBI	pollen-specific leucine-rich repeat extensin-like protein 1
				19166935	19167843	UPI0001A851C3	UPI0001A851C3	Sb03g014360	C5XJY9_SORBI	rapid alkalinization factor
				19173459	19178759	UPI0001C80BB1;Or	UPI0001C80BB1	Sb03g01437		protein LOW PSII ACCUMULATION 1
				19180099	19183534	UPI0001A845B0	UPI0001A845B0	Sb03g014380	C5XJZ0_SORBI	40S ribosomal protein S4
				19185886	19193629	UPI0001A851C4	UPI0001A851C4	Sb03g014390	C5XJZ1_SORBI	NB-LRR disease resistance protein
3	36549186	197.3805	s4	36474476..36475344		LOC110433364				ncRNA
				36469694..36470469		LOC110433369				ncRNA
5	50606510	428	s5	50633842	50666619	UPI0001A864C6	UPI0001A864C6	Sb05g020710	C5Y3T7_SORBI	reverse transcriptase
				50748352	50748630	UPI0001A8452A	UPI0001A8452A	Sb05g020712	C5XMS0_SORBI	reverse transcriptase
				50773752	50773961	UPI0001A864C7	UPI0001A864C7	Sb05g020715	C5Y3T9_SORBI	unknown function - similar to cadmium induced protein
7	25143622.2	6.96E+02	s6	24438987	24439399	UPI0001A878AD	UPI0001A878AD	Sb07g012310	C5YK24_SORBI	GDSL esterase/lipase
				24541299	24544032	UPI0001A87D1D	UPI0001A87D1D	Sb07g012315	C5YK25_SORBI	mucin-7-like
				24560864	24565790	*UPI000156629A	UPI000156629A	Sb07g012320	A5Y409_SORBI	Bt2
				25215441	25216665	UPI0001A87D1E	UPI0001A87D1E	Sb07g012421	C5YK26_SORBI	unknown function
				25266157	25267821	UPI0001A87D1F	UPI0001A87D1F	Sb07g012520	C5YK27_SORBI	transposase
				25696240	25700246	UPI0001A87D20	UPI0001A87D20	Sb07g012720	C5YK28_SORBI	pyrophosphate--fructose 6-phosphate 1-phosphotransferase subunit alpha-like
7	28238132.2	5.77E+02	s7	28360196..28363415		LOC8069849				ncRNA
7	41905551.1	3.44E+02	s8	42079162	42081307	UPI0001A87E0C	UPI0001A87E0C	Sb07g016970	C5YKE7_SORBI	exopolygalacturonase
7	45386874.8	8.52E+02	s9	44889340..44919902		LOC110436757				probable adenylate kinase 5, chloroplastic
7	46934129.7	5.98E+02	s10	47495562	47507543	UPI0001A87E83	UPI0001A87E83	Sb07g018430	C5YKG9_SORBI	alpha-soluble NSF attachment protein
7	47836695.1	3.47E+02	s11	47799636	47799914	UPI0001A879A6	UPI0001A879A6	Sb07g018531	C5YKH0_SORBI	TNP2-like
				47800384	47800926	UPI0001A879A7	UPI0001A879A7	Sb07g018630	C5YKH1_SORBI	unknown function
				47945333	47949221	UPI0001A879A8	UPI0001A879A8	Sb07g018640	C5YKH2_SORBI	TNP1-like

Table S7 Selective sweep regions identified with SweeD and associated genes

* indicates correspondance with the Mace et al (24) candidate domestication gene list

	w->A3	w->A5	A5->A6	A5->A7	A3->A6	A3->A7	A6->bicolor	A7->bicolor	A11->A9	A11->A10	A9->dur	A10->dur
pk1	0.35258451	0.48370471	0.15597999	0.44308019	0.46475671	0.16280127	0.1435501	0.31802334	0.09701379	0.31453691	0.44232227	0.31453691
pk2	0.33181749	0.23919964	0.07548886	0.24344399	0.18190086	0.47218433	0.29998484	0.37683796	0.30347127	0.44353494	0.34106412	0.44353494
pk3	0.23283311	0.3289374	0.4780961	0.47279066	0.28770653	0.20221313	0.23829013	0.21585569	0.20645748	0.41458239	0.36455965	0.41458239
pk4	0.47172957	0.30741246	0.1159618	0.25238745	0.35485827	0.49886312	0.0322874	0.03926027	0.15355465	0.19751402	0.23465211	0.19751402
pk5	0.19812036	0.49355768	0.06093679	0.11687131	0.36637866	0.25147794	0.00394119	0.00363802	0.0209186	0.00257693	0.03744126	0.00257693
pk6	0.08473549	0.11914507	0.29953009	0.22313173	0.37653479	0.1159618	0.16659087	0.29013188	0.06017887	0.47536759	0.07003183	0.47536759
pk7	0.47097165	0.44808246	0.31256632	0.3089283	0.40003032	0.43140822	0.22298014	0.06336213	0.28149159	0.27269971	0.29498257	0.27269971
pk8	0.42185842	0.43610732	0.48491739	0.31484008	0.4990147	0.34894649	0.32423829	0.41230863	0.30817038	0.23434895	0.19417917	0.23434895
pk9	0.45778384	0.44444444	0.28740337	0.43353039	0.45005305	0.41761407	0.05638927	0.05926936	0.27724723	0.0883735	0.09989389	0.0883735
pk10	0.48719115	0.34803699	0.37517053	0.11747764	0.1444596	0.2955889	0.29953009	0.1787176	0.03168107	0.12566318	0.0351675	0.12566318
pk11	0.2946794	0.43610732	0.48719115	0.49977262	0.25966348	0.19721085	0.23829013	0.14294376	0.1929665	0.38350765	0.37956647	0.38350765
pk12	0.18690314	0.38775201	0.39154161	0.49446718	0.09731696	0.14885554	0.104593	0.06790966	0.40427467	0.29210247	0.24465666	0.29210247
pk13	0.12657268	0.20645748	0.44217068	0.49446718	0.19114749	0.15810217	0.35622253	0.37683796	0.3659239	0.45990602	0.45793543	0.45990602
pk14	0.23283311	0.16552979	0.21903896	0.17811126	0.45929968	0.10413824	0.22131272	0.07230559	0.30817038	0.39972715	0.32408671	0.39972715
pk15	0.05714719	0.04259512	0.49446718	0.38623617	0.35015916	0.4780961	0.02940731	0.03061998	0.3659239	0.49598302	0.31711384	0.49598302
pk16	0.1226315	0.12611793	0.47157799	0.3783538	0.45005305	0.38866151	0.01879642	0.02243444	0.40230408	0.04683947	0.04289829	0.04683947
pk17	0.30695771	0.28103683	0.4780961	0.48613006	0.38032439	0.40958011	0.26026982	0.17371533	0.40230408	0.3022586	0.24465666	0.3022586
pk18	0.31817493	0.22586024	0.07867212	0.19433076	0.22858875	0.10792785	0.03956344	0.1787176	0.05138699	0.26390784	0.08382598	0.26390784
pk19	0.02667879	0.03440958	0.18326512	0.28452327	0.12869486	0.42595119	0.26421101	0.12066091	0.14749128	0.02213127	0.06684857	0.02213127
pk20	0.18523571	0.30013643	0.13475822	0.13172654	0.43171138	0.05623768	0.05381234	0.16689404	0.06169471	0.27269971	0.20190996	0.27269971
pk21	0.29816583	0.48370471	0.31969077	0.44474761	0.26421101	0.17401849	0.42337426	0.06336213	0.17750493	0.25162953	0.31711384	0.25162953
pk22	0.45778384	0.2557223	0.42140367	0.23222677	0.1444596	0.28816129	0.03698651	0.23434895	0.44308019	0.17265424	0.20190996	0.17265424
pk23	0.49310293	0.28952554	0.31332424	0.24511141	0.09519479	0.0779142	0.3745642	0.0551766	0.3659239	0.11005002	0.09110202	0.11005002
pk24	0.36243747	0.28952554	0.40791269	0.21145975	0.33121116	0.37501895	0.03350008	0.30726088	0.36774291	0.42837653	0.31711384	0.42837653
pk25	0.12323784	0.31772018	0.1435501	0.21934213	0.23162043	0.17614067	0.38153706	0.49840837	0.44308019	0.17598909	0.23601637	0.17598909
pk26	0.03471275	0.01667425	0.1121722	0.23995756	0.29922692	0.41458239	0.35758678	0.48825224	0.04517205	0.16628771	0.14324693	0.16628771
pk27	0.48279521	0.33909353	0.18887373	0.4144308	0.09458845	0.19220858	0.03956344	0.17371533	0.44308019	0.39972715	0.40093982	0.39972715
pk28	0.00045475	0.00045475	0.2946794	0.42079733	0.40427467	0.23722904	0.47748977	0.18826739	0.44308019	0.31453691	0.39502804	0.31453691
pk29	0.34227679	0.2557223	0.15597999	0.46127027	0.31484008	0.31362741	0.36198272	0.20221313	0.44308019	0.22343489	0.2176747	0.22343489
pk30	0.4600576	0.43610732	0.36167955	0.38987419	0.18887373	0.21494619	0.23025618	0.13430347	0.1606791	0.42686069	0.46869789	0.42686069
s1	0.04365621	0.01864484	0.47157799	0.4144308	0.1444596	0.1444596	0.4344399	0.473397	0.12141883	0.1502198	0.09004093	0.06533273
s2	0.12020615	0.30013643	0.06093679	0.12126724	0.45929968	0.33500076	0.00257693	0.00227376	0.0389571	0.00181901	0.05093224	0.00682128
s3	0.16189177	0.14658178	0.48719115	0.24344399	0.45293315	0.18447779	0.05381234	0.41382447	0.04683947	0.00879188	0.0807943	0.11899348
s4	0.02986206	0.08033955	0.39093527	0.29134455	0.38032439	0.49886312	0.1188419	0.12763377	0.13839624	0.28058208	0.45793543	0.37486736
s5	0.13112021	0.15249356	0.49446718	0.38623617	0.35273609	0.22207064	0.10186448	0.07821737	0.30347127	0.49113233	0.35061392	0.3022586
s6	0.10125815	0.08033955	0.39533121	0.29134455	0.35273609	0.22586024	0.11565863	0.12763377	0.192057	0.19781719	0.32954373	0.3260573
s7	0.10792785	0.08594816	0.42837653	0.29877217	0.36107322	0.21812945	0.13854782	0.21585569	0.33348492	0.34439897	0.35061392	0.34940124
s8	0.19887828	0.21327876	0.39154161	0.24602092	0.35485827	0.2955889	0.104593	0.39760497	0.40427467	0.47263908	0.03152948	0.03061998
s9	0.13354555	0.12157041	0.42837653	0.3089283	0.38153706	0.24207973	0.0627558	0.07230559	0.20782174	0.31256632	0.37956647	0.3260573
s10	0.13354555	0.12157041	0.4780961	0.31484008	0.45929968	0.25147794	0.05638927	0.13430347	0.27724723	0.49113233	0.16264969	0.12869486
s11	0.11702289	0.12157041	0.41397605	0.27724723	0.48810065	0.31165681	0.03501592	0.05017432	0.40427467	0.25420646	0.12081249	0.14703653
dw3	0.30847355	0.31772018	0.03547067	0.15340306	0.33227224	0.02561771	0.03865393	0.25132636	0.40427467	0.06684857	0.18008186	0.04577838
dw1	0.17280582	0.1226315	0.07260876	0.21312718	0.05775352	0.12445051	0.21464302	0.40381992	0.16416553	0.11095953	0.1891769	0.13339397
dw2	0.43686524	0.01394573	0.1255116	0.3089283	0.00257693	0.00106109	0.24359557	0.33894194	0.05214491	0.0389571	0.08215856	0.09261786
Sh1	0.06230105	0.13490981	0.09519479	0.15340306	0.38790359	0.36774291	0.22298014	0.15294831	0.08958618	0.09246627	0.40093982	0.36728816
Sh2	0.24768834	0.17462483	0.21464302	0.2737608	0.13930575	0.44156435	0.32620888	0.12217675	0.16416553	0.45308474	0.18008186	0.30786721
Sh3/Bt1	0.05487343	0.01167197	0.01106564	0.15340306	0.14491436	0.0209186	0.01212672	0.23434895	0.47991511	0.49113233	0.29361831	0.26178566
Bt2	0.35440352	0.38775201	0.19266333	0.02455662	0.18887373	0.03638017	0.18569047	0.04744581	0.25117478	0.34167046	0.44232227	0.27269971
SbWRKY	0.22010005	0.31772018	0.4629377	0.14521752	0.28770653	0.31362741	0.30923147	0.34106412	0.09701379	0.123541	0.22495074	0.12566318
AE1	0.02395028	0.03031681	0.28179476	0.14688495	0.20554798	0.25375171	0.20024253	0.32272245	0.48324996	0.01303623	0.01909959	0.47536759
cul4	0.15507049	0.27269971	0.31332424	0.01182356	0.11656814	0.03835077	0.49082916	0.01682583	0.1302107	0.12884645	0.16037593	0.15673791
gt1	0.07215401	0.13172654	0.24829468	0.13248446	0.08625133	0.43747158	0.18796423	0.15294831	0.0447173	0.0569956	0.1891769	0.05396392
int1	0.01015613	0.01167197	0.39533121	0.06639382	0.28361376	0.10944369	0.14067	0.18826739	0.1929665	0.37107776	0.32954373	0.42837653
ma1	0.26557526	0.47203274	0.25435804	0.34940124	0.28846445	0.11444596	0.39275428	0.12763377	0.30817038	0.37107776	0.34106412	0.36228589
ma3	0.34227679	0.48370471	0.31332424	0.1996362	0.31756859	0.45172048	0.49492193	0.18826739	0.01682583	0.01364256	0.12975595	0.15673791
ma6	0.44656662	0.48370471	0.34091254	0.44777929	0.47006215	0.31726542	0.21464302	0.46703047	0.04911323	0.05426709	0.22661816	0.34940124
Nud	0.33181749	0.30013643	0.07867212	0.3783538	0.09231469	0.48446263	0.20691223	0.20903441	0.192057	0.37107776	0.14188267	0.27679248
O2	0.05259967	0.0598757	0.30680612	0.40927694	0.43171138	0.47203274	0.11020161	0.0551766	0.11050477	0.1083826	0.17068364	0.16628771
Pa1	0.08473549	0.11156586	0.18508413	0.18038502	0.11535546	0.10050023	0.12990753	0.11232378	0.08276489	0.04517205	0.0807943	0.06139154
SHP	0.36243747	0.1862968	0.21464302	0.39396695	0.06593906	0.19433076	0.39987873	0.17341216	0.30347127	0.28058208	0.47112324	0.45990602
SPS1	0.04107928	0.08594816	0.36046688	0.25678339	0.4990147	0.34894649	0.29998484	0.19539184	0.33469759	0.19781719	0.25162953	0.11141428
SPS2	0.49310293	0.09170835	0.04714264	0.1653782	0.15522207	0.0075792	0.05093224	0.28634228	0.25117478	0.23419736	0.1891769	0.17598909
SPS3	0.4600576	0.47203274	0.01273306	0.42079733	0.02031226	0.37441261	0.0113688	0.24859785	0.20782174	0.02137335	0.23601637	0.03304532
SPS4	0.12020615	0.16552979	0.10989844	0.31529483	0.06123996	0.43140822	0.09140518	0.29998484	0.40230408	0.47263908	0.12581476	0.09943914
SPS5	0.27239654	0.01000455	0.08018796	0.2147946	0.0009095	0.00697287	0.23025618	0.1482492	0.1691678	0.15006821	0.41503714	0.45990602
sss1	0.42185842	0.43383356	0.19417917	0.36304381	0.10959527	0.16113385	0.14430802	0.1891769	0.27724723	0.18462938	0.46869789	0.45990602
su	0.48719115	0.1226315	0.00363802	0.23707746	0.09595271	0.01546157	0.00636653	0.23434895	0.0845839	0.0903441	0.40093982	0.32302562
suc1	0.21555252	0.19796877	0.45672275	0.32954373	0.48173412	0.3650144	0.41094437	0.37608004	0.1302107	0.24556617	0.1891769	0.09261786
SUS1	0.14961346	0.28952554	0.0075792	0.0218281	0.00045475	0.00045475	0.4344399	0.38562983	0.05366075	0.1083826	0.25162953	0.02728513
SUS2	0.45778384	0.40215249	0.10959527	0.2034258	0.3118084	0.38290132	0.35622253	0.44368652	0.33348492	0.49113233	0.42943762	0.47536759
SUS3	0.45778384	0.44444444	0.1435501	0.27785357	0.28846445	0.38896468	0.35015916	0.39760497	0.05729877	0.15855692	0.05366075	0.17598909
SUT1	0.34212521	0.29195089	0.39745339	0.23707746	0.48173412	0.29695316	0.47248749	0.25132636	0.1302107	0.17356374	0.16264969	0.21631044
SUT2	0.42185842	0.40912536	0.42837653	0.11141428	0.45005305	0.1130817	0.23919964	0.26769744	0.47991511	0.1502198	0.49658936	0.27679248
SUT3	0.07745945	0.08716083	0.42564802	0.1435501	0.47157799	0.10307716	0.03698651	0.29786266	0.08822192	0.47263908	0.05957253	0.25162953
SUT4	0.18523571	0.17053206	0.06427164	0.10807943	0.08503865	0.10944369	0.28027891	0.25132636	0.03592542	0.12748219	0.03077156	0.08321965
TB1	0.01455207	0.17462483	0.00424435	0.00939821	0.37153251	0.45596483	0.36228589	0.20221313	0.0560861	0.04532363	0.11444596	0.16674246
TGA1	0.14612703	0.34803699	0.27315446	0.02880097	0.05532818	0.10944369	0.03319691	0.20221313	0.27724723	0.08700925	0.46869789	0.29119297
vrs1	0.21252084	0.23071093	0.4981052	0.24511141	0.45929968	0.28543277	0.38077914	0.42974079	0.192057	0.31256632	0.11050477	0.11899348
Wx	0.17356374	0.27269971	0.15446415	0.18038502	0.36107322	0.43140822	0.15507049	0.15385781	0.47991511	0.25420646	0.42943762	0.30786721
Wx_Chr2	0.19251175	0.12157041	0.17841443	0.104593	0.47521601	0.2499621	0.38896468	0.39624072	0.20782174	0.12005457	0.32408671	0.20888283

Table S8 Probabilities of gradients for particular sample transitions against genomic average

chromosome	position	GERP score range	highest	lowest	associated selection signal peaks
1	10600000	0.18401937	M5	M2	
	12100000	0.818181818	A11	M2	Sh1
	13000000	0.666666667	A11	A10	
	13200000	0.27443609	A10	A3	
	18800000	0.524390244	A10	M2	
	19000000	0.22027972	A5	M2	
	19500000	0.442307692	M2	A10	
	19900000	0.214046823	A10	A6	
	20400000	0.380952381	Kew1	A11	
	22700000	0.243445693	M6	M2	
	23900000	0.305555556	A10	M2	
	26000000	0.195488722	A3	M6	
	27300000	0.188034188	A11	A9	
	29000000	0.236842105	M6	M2	
	31200000	0.209424084	A10	M6	
	31400000	0.186567164	M5	M6	
	31500000	0.188235294	A11	M6	
	39600000	0.962962963	A10	A5	
	43600000	0.222698073	M5	A10	
	43900000	0.605042017	A3	M2	
	45000000	0.5	Kew1	M2	
	46600000	0.25382263	M6	A5	
2	2700000	0.169745958	Kew1	A10	
	10400000	0.37295082	M5	M2	
	10500000	0.159751037	M6	M2	
	12800000	0.27173913	A3	M2	
	15800000	0.16	A10	A3	
	18500000	0.27027027	M2	A10	
	22000000	0.517241379	Kew1	A5	
	22900000	0.32238806	M6	A9	
	26300000	0.467532468	A6	A11	
	30300000	0.386554622	A10	M2	
	30400000	0.222222222	A10	M2	
	30800000	0.262365591	A5	M2	
	31000000	0.265060241	A10	Kew1	
	31100000	0.242038217	A3	M2	
	37500000	0.2734375	M6	A3	
	39200000	0.28125	A11	A10	
	39900000	0.195652174	A3	M2	
	52000000	0.862222222	A3	M2	
	52700000	0.208510638	A7	M2	
	53800000	0.2375	A10	A11	pk4
	54300000	0.418316832	M6	M2	pk5, s2
	66900000	0.153846154	M5	A9	
	67800000	0.153310105	A10	M2	
	76700000	0.466666667	A10	A7	
3	8100000	1.2	A3	A10	
	19500000	0.5	A11	A10	
	20500000	0.25	A11	A10	
	23200000	0.208566108	M2	A5	
	28200000	0.331818182	A3	A10	
	33400000	0.308539945	A6	M2	
	36300000	0.666666667	A10	M6	
	42000000	0.285714286	A10	M6	
	42500000	0.607260726	A9	M2	
	54900000	0.641025641	A7	M2	
	62100000	0.717948718	A11	A5	
	62200000	0.877005348	A3	A7	
4	13400000	0.666666667	A10	A3	
	13900000	0.333333333	A3	A10	
	19000000	0.75	A10	A5	
	23900000	0.471177945	Kew1	M2	
	24000000	0.214477212	A3	M2	
	27500000	0.244897959	Kew1	M2	pk8

	27600000	0.210023866	M6	M2	
	32300000	0.296296296	A10	M6	
	32700000	0.506849315	A3	M2	pk9
	34200000	0.210663199	A10	A7	
	36100000	0.196721311	A3	Kew1	
	36900000	0.425396825	A11	M2	
	44100000	0.354330709	M6	A9	
	47600000	0.189542484	A10	A6	
	57400000	1.125	A10	Kew1	
	58200000	0.376068376	A10	A3	
5	6000000	0.977777778	A5	M2	
	6100000	0.406015038	A10	M2	
	9100000	0.728323699	A6	A11	
	11900000	0.412280702	M6	A10	
	12000000	0.438202247	A6	A11	
	14100000	0.308300395	A10	M6	
	17900000	0.37254902	A10	M6	
	22700000	0.277777778	M6	M2	
	23800000	0.3	A10	A7	
	24900000	0.287356322	A10	M2	
	42000000	0.240469208	A3	A10	
	44800000	0.541176471	A3	M2	
	48500000	0.350553506	M6	A10	
	49400000	0.737777778	Kew1	M2	
	50000000	0.448717949	A10	M2	
	50600000	0.350515464	A5	M2	s5
	50700000	0.321782178	M6	M2	
	51000000	0.234332425	M6	M2	
	54000000	0.484018265	M5	A3	
	55100000	1.155555556	A9	A10	
	57200000	0.453333333	M5	Kew1	
6	6700000	0.23015873	M2	A10	
	6800000	1.04	A11	A7	
	7500000	0.168316832	Kew1	A7	
	9900000	0.213675214	A5	A9	
	10400000	0.157303371	A7	M2	
	14500000	0.171945701	A5	M2	
	15900000	0.289398281	M5	A10	
	16000000	0.584269663	A3	M2	
	24000000	0.174863388	A3	M2	
	24300000	0.339047619	M5	M2	
	26600000	0.152671756	A3	A10	
	28600000	0.298181818	M5	M2	pk17
	32000000	0.2	A9	M2	
	33500000	0.319767442	A3	M2	
	34200000	0.239043825	M5	M2	
	38300000	0.178571429	A10	A5	
7	600000	0.384937238	A5	A10	
	7000000	0.8	A10	Kew1	
	9100000	0.20441989	Kew1	A5	
	10000000	0.377135348	A5	M2	
	15300000	0.18344519	M5	A11	
	15500000	0.261538462	M5	A6	
	17600000	0.224137931	A10	M6	
	17800000	0.5	A3	M2	
	24700000	0.215246637	A10	A5	Bt2
	25100000	0.339285714	A10	M6	s6
	25200000	0.231974922	Kew1	M6	
	25300000	0.5	A10	M6	
	26000000	0.198300283	A10	M6	
	28200000	0.189189189	A10	A5	s7
	28700000	0.4375	A10	A3	
	38700000	0.256157635	Kew1	A3	
	39900000	0.208144796	A10	Kew1	
	47100000	0.38	A10	A3	
	47800000	0.295964126	A10	A5	s11
	48000000	0.217741935	A10	A3	

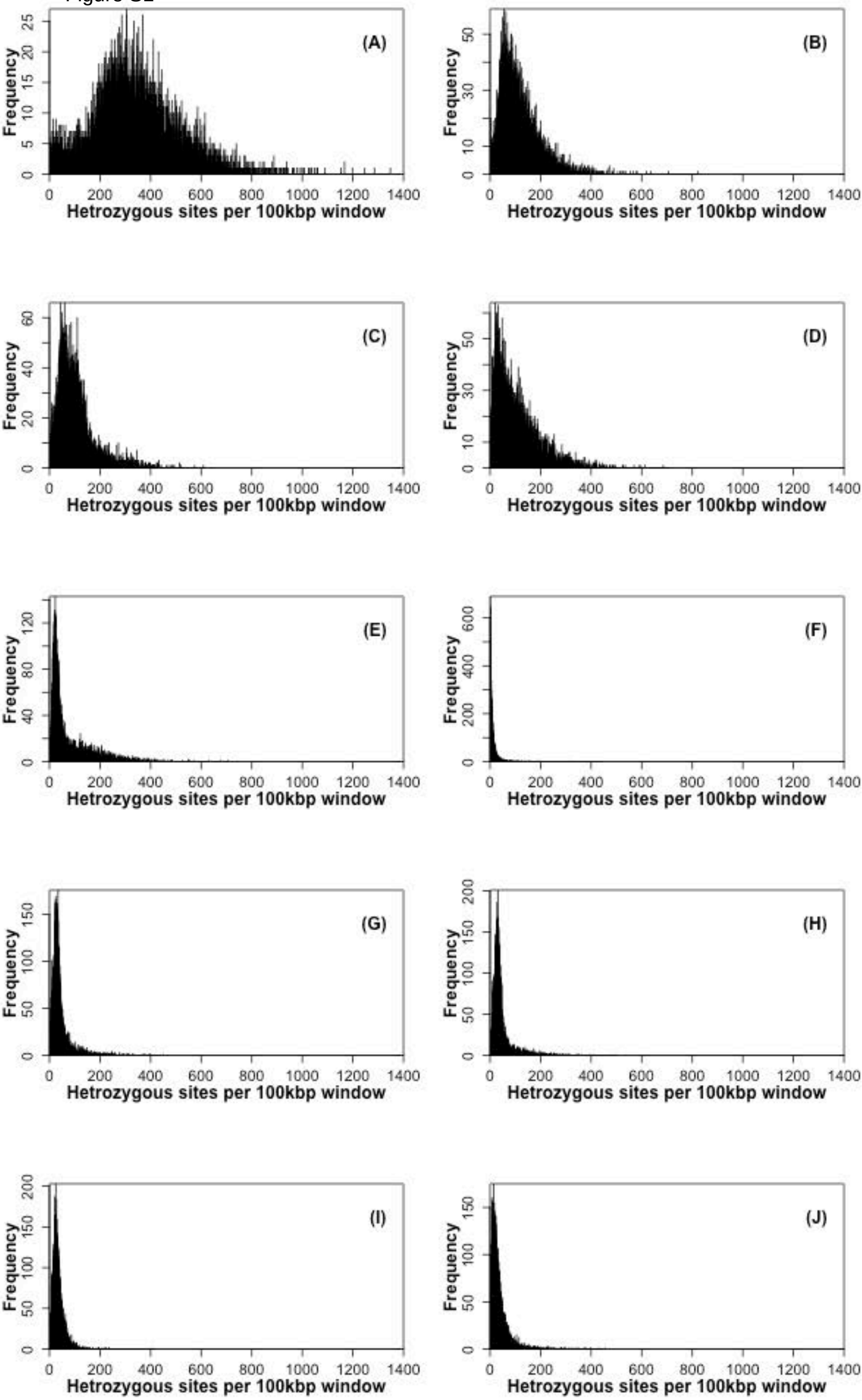
	48600000	0.302197802	A10	A3	
	51200000	0.2	A10	M2	
	53100000	0.202941176	A10	M2	
	57600000	0.260869565	A10	M2	
	58100000	0.649425287	A7	Kew1	
	59800000	0.227272727	A10	M6	dw3
	61900000	0.239669421	A10	A9	TGA1
	63200000	0.237541528	M6	A3	pk20
	64300000	0.404255319	A3	Kew1	
8	5500000	0.301754386	A5	M2	
	6600000	0.625	A10	A3	
	8500000	0.34893617	M5	Kew1	
	39000000	0.595555556	M5	M2	
	40600000	1.157894737	A10	M2	pk24
	40800000	0.27076412	A10	M2	
	45500000	0.5	A5	A10	
	51800000	0.753623188	A11	A5	
	54000000	0.575342466	A10	Kew1	
	54500000	0.865384615	A10	M2	
	54800000	1.368421053	A10	A3	
	55100000	0.29787234	A3	A10	
9	900000	0.145762712	M5	M2	
	4800000	0.181818182	A3	A10	
	5100000	0.195744681	M2	A7	
	5200000	0.212598425	A11	M2	
	7700000	0.117647059	A3	A10	
	9800000	0.146443515	A7	Kew1	
	10600000	0.375	A10	M6	
	11800000	0.123966942	A10	A3	
	12400000	0.202531646	M5	A10	
	12500000	0.1625	M5	A7	
	12700000	0.44015444	M5	M2	
	14200000	0.159509202	M6	A10	
	23800000	0.214511041	A3	A7	
	24900000	0.396694215	A3	M2	
	26700000	0.247619048	A10	M6	
	34300000	0.117647059	A5	A3	
	36900000	0.222222222	A3	M2	
	37900000	0.185661765	A5	A11	
	38100000	0.18522602	A11	A7	
	41300000	0.158878505	A10	A5	
	42000000	0.270967742	A10	A5	pk26
	42900000	0.17989418	A3	Kew1	
	44300000	0.191616766	A10	M2	
	45900000	0.140540541	M5	M2	
	47100000	0.236686391	A10	M2	
10	6100000	0.888888889	A3	A10	
	6200000	0.363636364	A10	A5	
	11600000	0.621848739	A3	M6	
	12900000	0.232142857	A9	M6	
	14100000	0.502617801	A11	M2	
	15000000	1.333333333	A10	A3	
	16700000	0.213675214	A6	A10	
	20400000	0.456410256	A3	A11	
	21500000	0.21978022	A5	A7	pk28
	24600000	0.259136213	A11	M6	
	26700000	0.238636364	A3	A5	
	36700000	0.583892617	A10	A5	
	39100000	0.264285714	A10	A5	
	49100000	0.410958904	M5	M2	
	49500000	0.242105263	A6	A3	

Table S9 Regions of GERP score deviation between genomes > 2 standard deviations

gene	window	tree number	selection	phylogenetic incongruence	potential donor identified
dw3	479000000	704	bicolor A3-A5, A7, durra A10	no	no
dw1	596200000	876	bicolor A6	yes	durra A11
dw2	396300000	582	bicolor A3-A5	no	no
su	482600000	709	bicolor A3-A5,A7	yes	durra A11
SPS2	232100000	341	bicolor A3-A5,A6-A7	no	no
SUS1	59600000	87	bicolor A3-A5, A5-A6	yes	durra A11,A9
SPS5	653100000	960	bicolor A3-A5	no	no
Sh3/Bt1	233300000	342	bicolor A7	yes	no
TB1	9100000	13	bicolor A5-A6,A7	no	no
O2	362200000	532	bicolor A6,A7	yes	durra A11, A10
SPS3	307500000	452	bicolor A6-A7	yes	durra A10
Sh1	12100000	17	bicolor A3	no	no
Ae1	277600000	408	bicolor wild-A3, A9-A10	no	no
int1	219200000	322	bicolor wild-A3,A7	yes	no
gt1	66700000	98	durra A11-A9, bicolor A7	yes	durra A11
SUT4	539000000	792	durra A9-A10, A7	yes	no
TGA1	481000000	707	bicolor A7	yes	durra A9, A10
ma3	68000000	99	durra A11-A9	yes	bicolor A3,A5
ma6	363700000	534	durra A11-A9	yes	bicolor A5
pk1	44900000	66	durra A10	no	no
pk2	84900000	124	bicolor A5	no	no
pk3	117800000	173	bicolor A3	no	no
pk4	127600000	187	bicolor A6	yes	no
pk5	128100000	188	bicolor A3,A5-A6,A7 durra A11	yes	bicolor A3
pk6	184600000	271	bicolor A3	no	no
pk7	244500000	359	bicolor A5	yes	durra A9
pk8	253900000	373	durra A9, A10	yes	bicolor A6
pk9	259200000	381	durra A11	no	no
pk10	263700000	387	bicolor A3	no	no
pk11	309400000	454	bicolor A3	yes	no
pk12	322400000	473	durra A11	no	no
pk13	337800000	496	bicolor A3	no	no
pk14	354000000	520	bicolor A7	yes	no
pk15	355600000	522	bicolor wild-A3,A5,A6,A7	yes	no
pk16	379500000	557	bicolor A6	no	no
pk17	385500000	566	durra A11	yes	no
pk18	469900000	690	durra A11	no	no
pk19	482300000	709	bicolor wild-A3, A7	yes	no
pk20	495900000	729	bicolor A6	yes	no
pk21	513100000	754	bicolor A7	no	no
pk22	515000000	757	bicolor A5,A6	yes	no
pk23	516500000	759	durra A11, A9, A10 bicolor A7	no	no
pk24	524200000	770	bicolor A5	no	no
pk25	536500000	788	bicolor A3	no	no
pk26	581100000	854	bicolor A5, durra A11-A9	yes	bicolor A5
pk27	616500000	906	bicolor A6	no	no
pk28	620300000	911	bicolor A7	no	no
pk29	635300000	933	durra A11, A10	yes	bicolor A6
pk30	625800000	920	bicolor A7	yes	no

Table S10 Phylogenetic congruence of regions containing significant reductions in heterozygosity

Figure S1


(A)
(B)
(C)
(D)
(E)
(F)
(G)
(H)
(I)
(J)
Frequency
Hetrozygous sites per 100kbp window

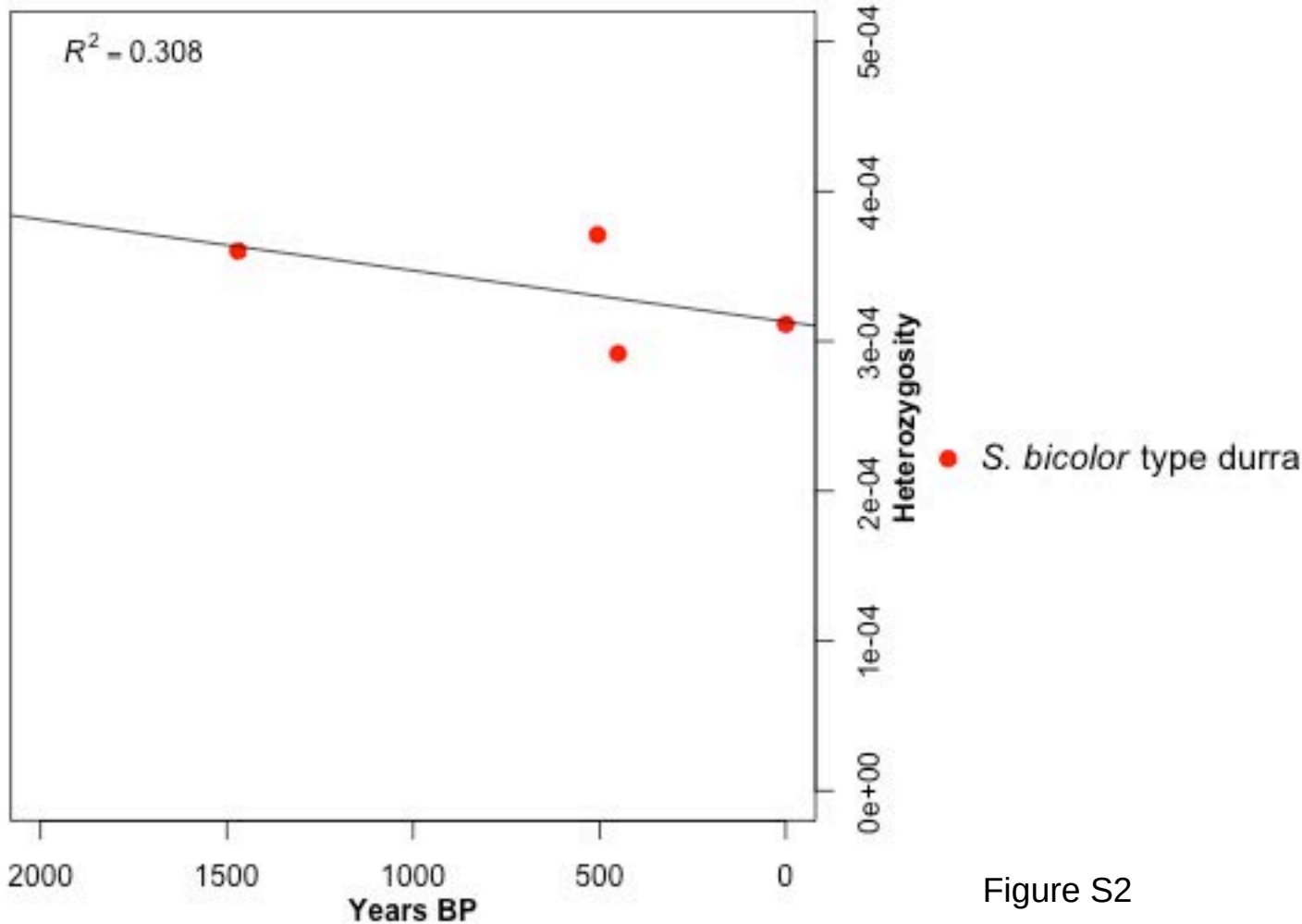

$R^2 = 0.308$
5e-04
4e-04
3e-04
2e-04
1e-04
0e+00
Heterozygosity
2000
1500
1000
500
0
Years BP
S. bicolor type durra

Figure S2

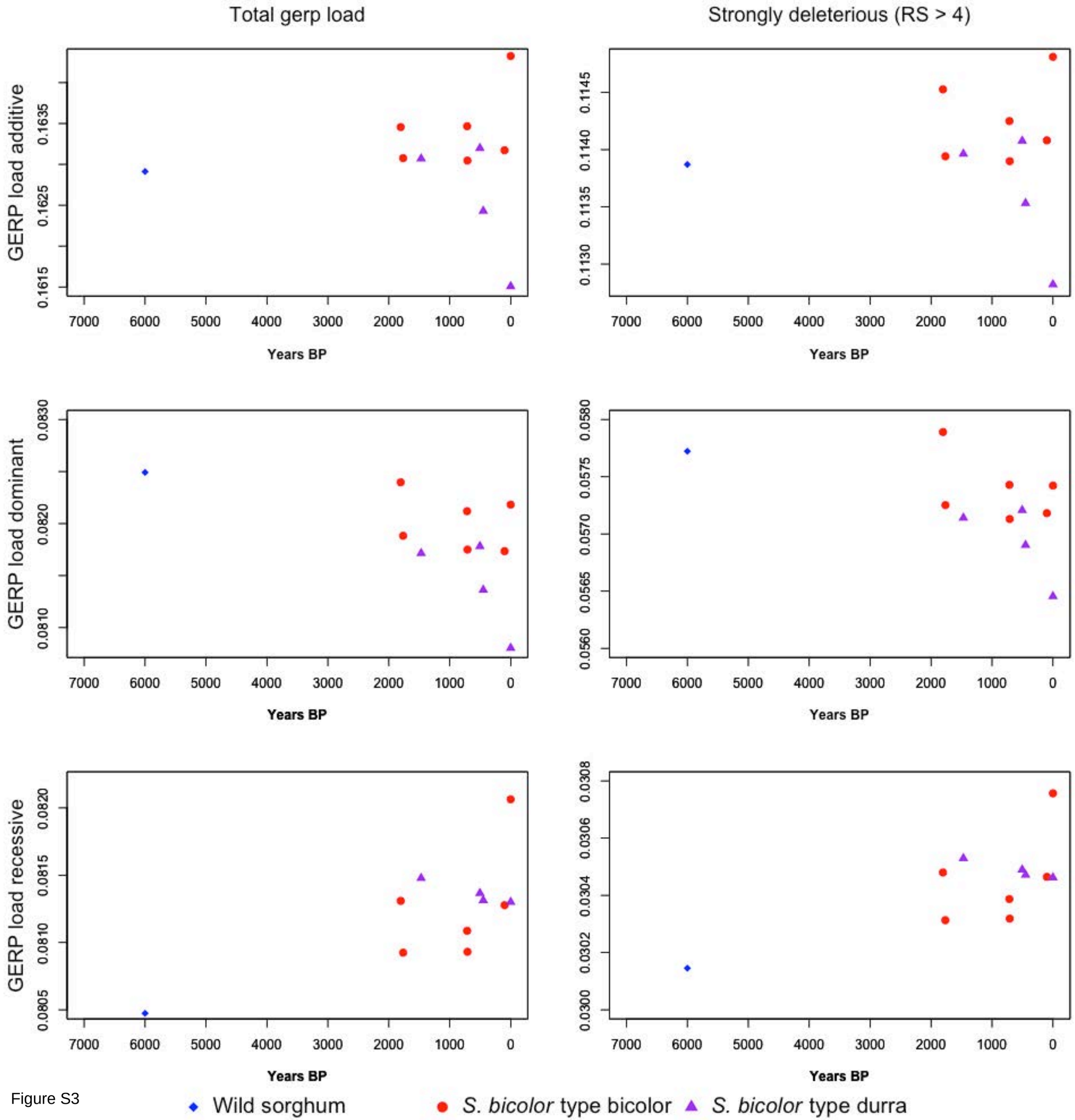

Total gerp load
Strongly deleterious (RS > 4)
GERP load additive
GERP load dominant
GERP load recessive
Years BP
Wild sorghum
S. bicolor type bicolor
S. bicolor type durra

Figure S3

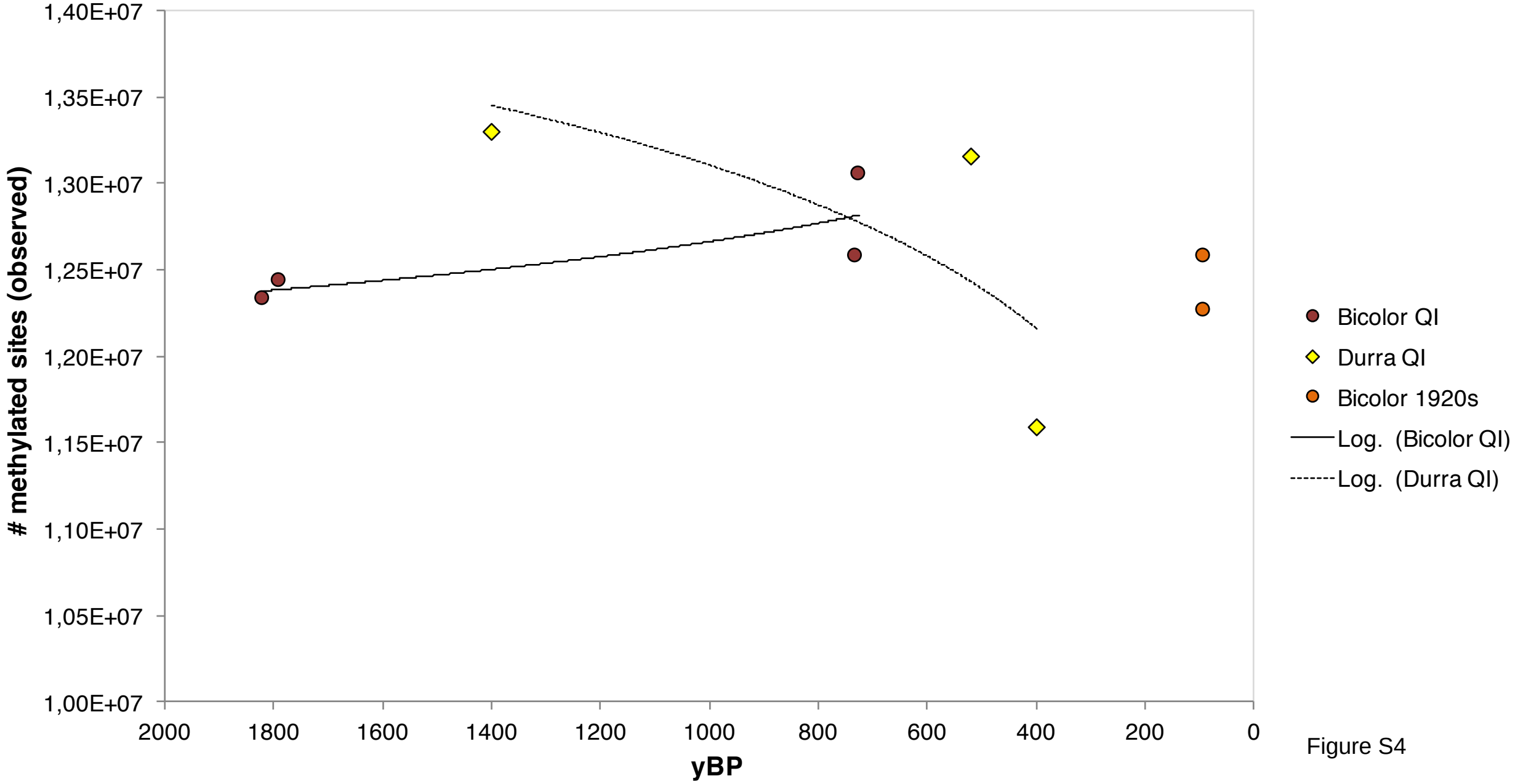
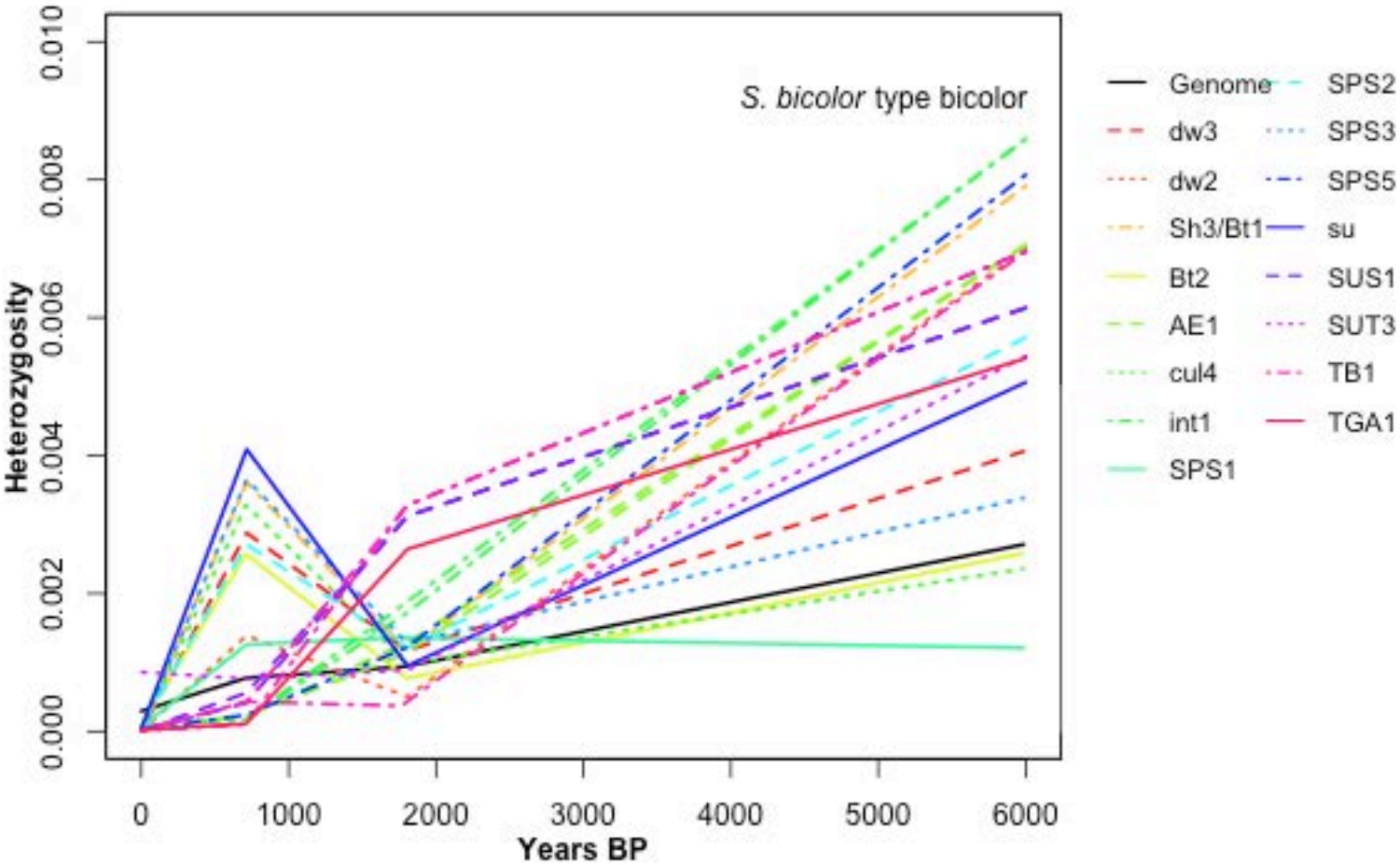

1,40E+07
1,35E+07
1,30E+07
1,25E+07
1,20E+07
1,15E+07
1,10E+07
1,05E+07
1,00E+07
methylated sites (observed)
2000
1800
1600
1400
1200
1000
800
600
400
200
0
yBP
Bicolor QI
Durra QI
Bicolor 1920s
Log. (Bicolor QI)
Log. (Durra QI)

Figure S4


S. bicolor type bicolor
Heterozygosity
0.000
0.002
0.004
0.006
0.008
0.010
0
1000
2000
3000
4000
5000
6000
Years BP
Genome
dw3
dw2
Sh3/Bt1
Bt2
AE1
cul4
int1
SPS1
SPS2
SPS3
SPS5
su
SUS1
SUT3
TB1
TGA1

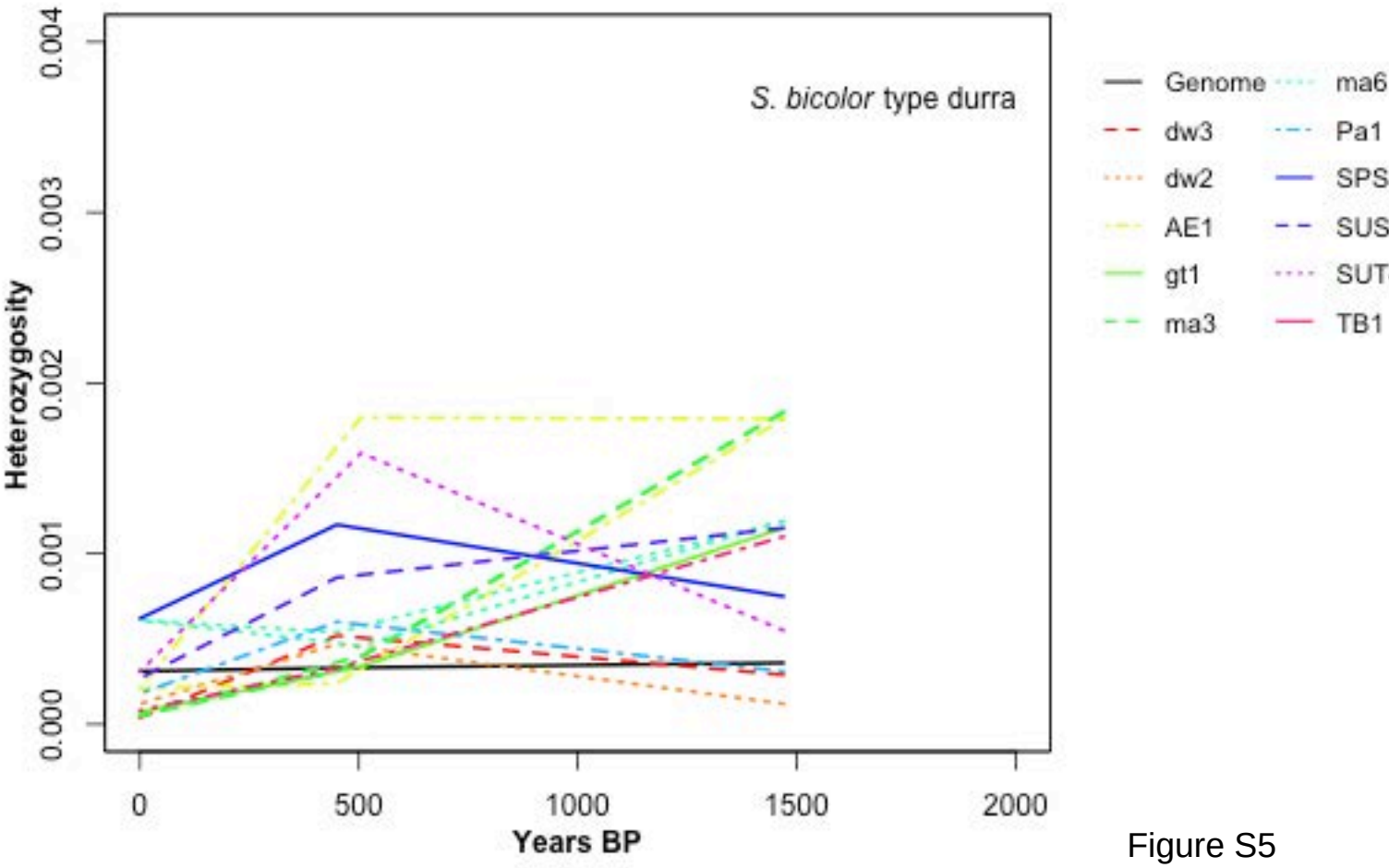
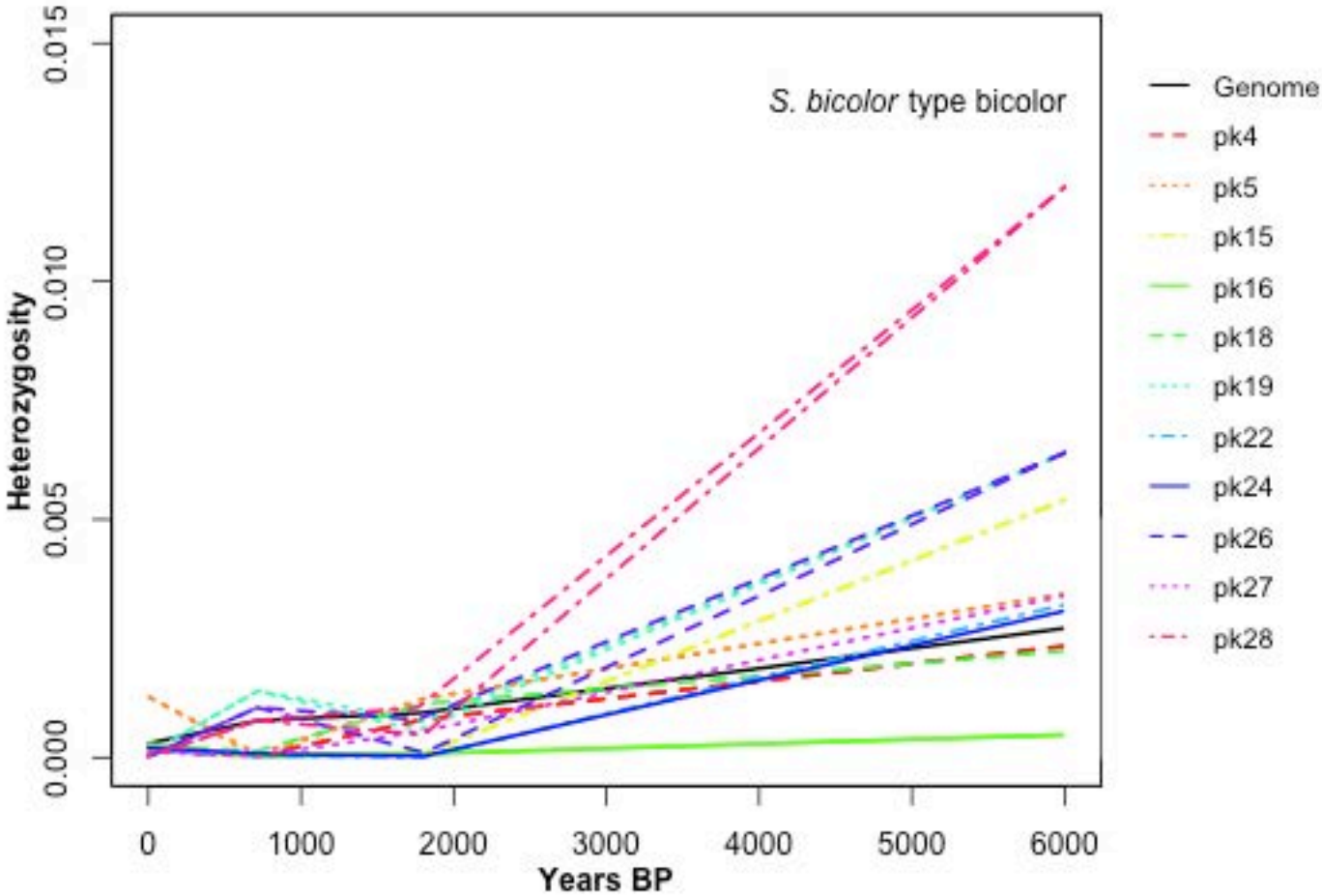

S. bicolor type durra
Heterozygosity
0.000
0.001
0.002
0.003
0.004
0
500
1000
1500
2000
Years BP
Genome
dw3
dw2
AE1
gt1
ma3
ma6
Pa1
SPS3
SUS1
SUT4
TB1

Figure S5


S. bicolor type bicolor
Heterozygosity
0.000
0.005
0.010
0.015
0
1000
2000
3000
4000
5000
6000
Years BP
Genome
pk4
pk5
pk15
pk16
pk18
pk19
pk22
pk24
pk26
pk27
pk28

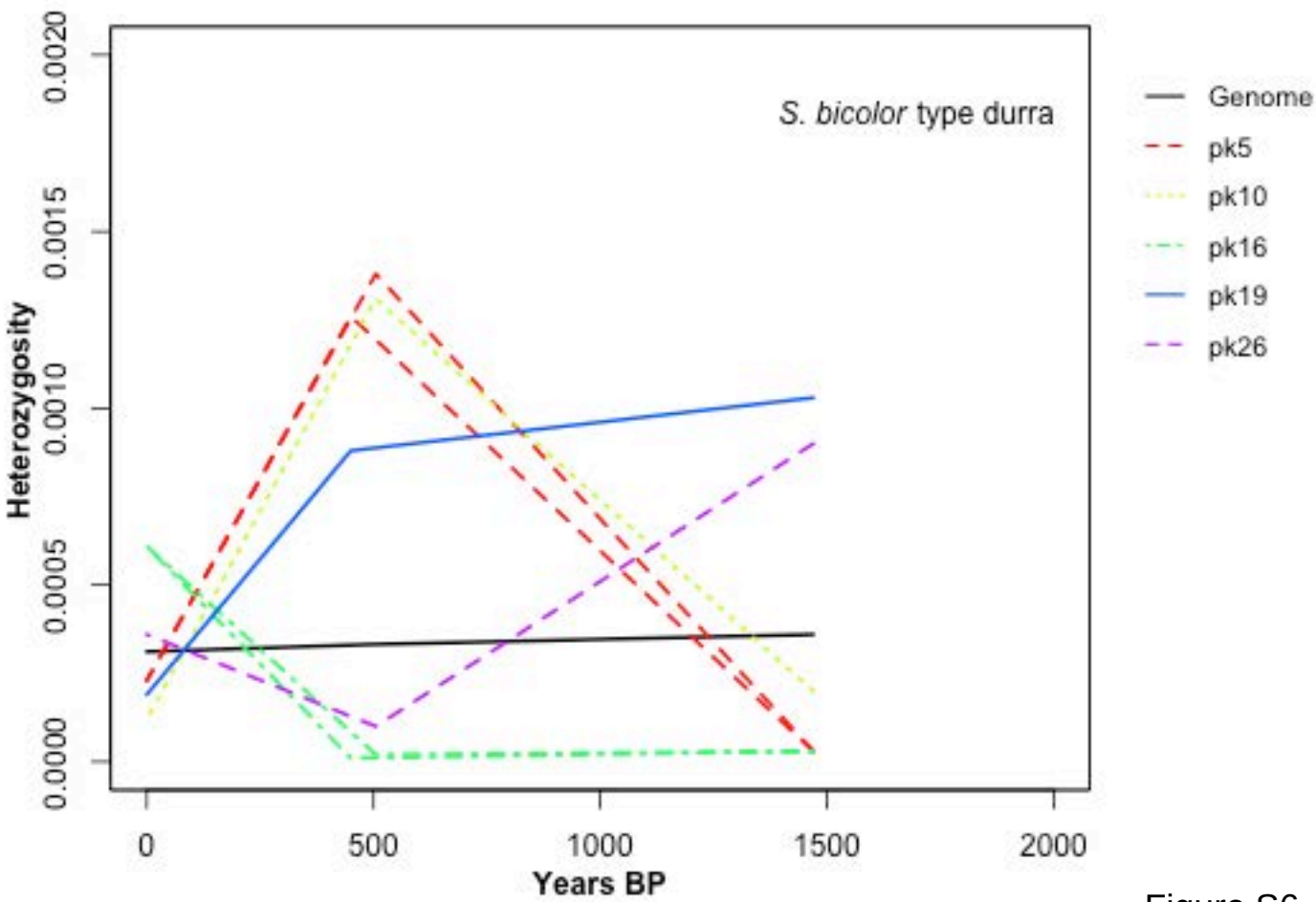
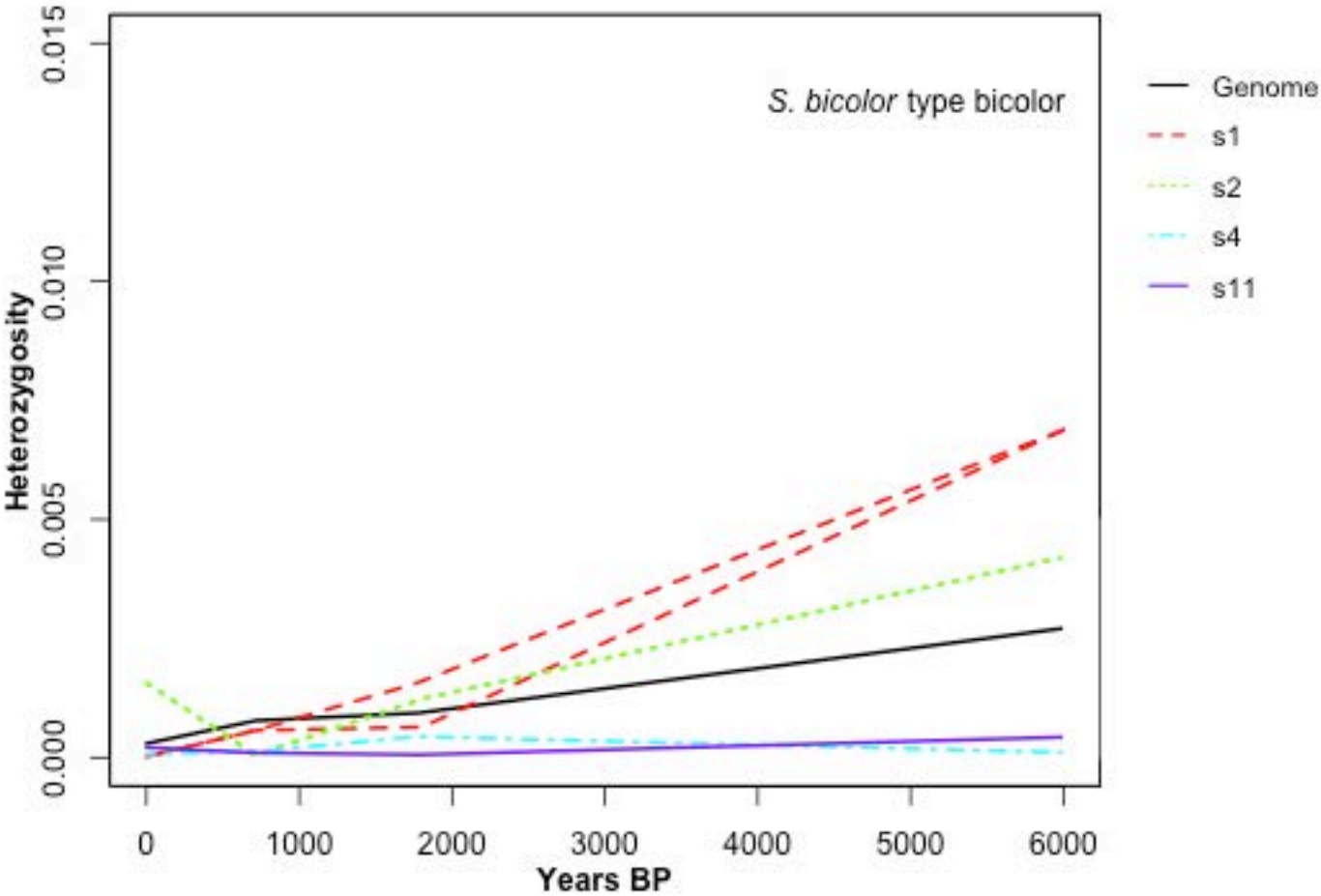

S. bicolor type durra
Heterozygosity
0.0000
0.0005
0.0010
0.0015
0.0020
0
500
1000
1500
2000
Years BP
Genome
pk5
pk10
pk16
pk19
pk26

Figure S6


S. bicolor type bicolor
Genome
s1
s2
s4
s11
Heterozygosity
0.000
0.005
0.010
0.015
0
1000
2000
3000
4000
5000
6000
Years BP

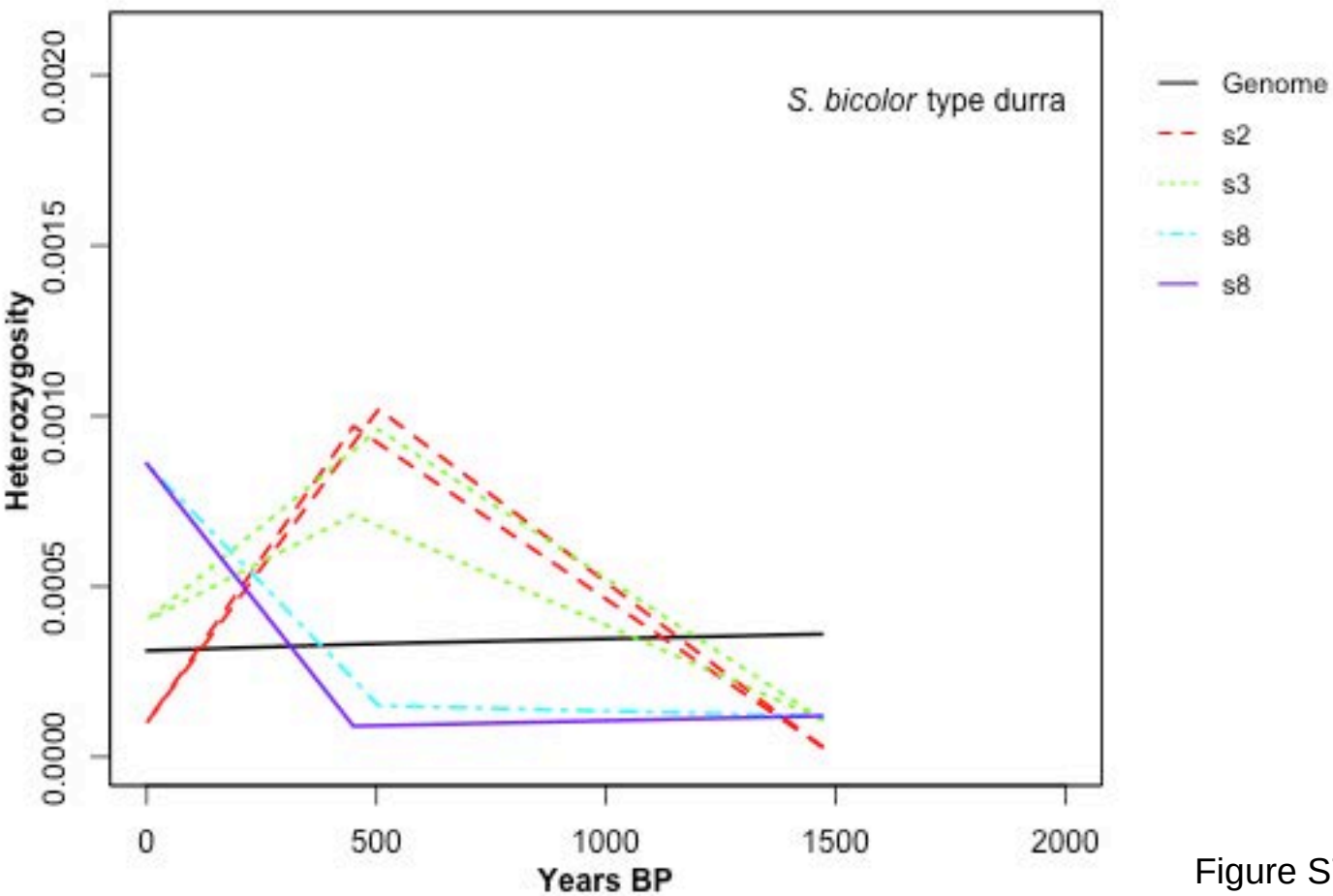

S. bicolor type durra
Genome
s2
s3
s8
s8
Heterozygosity
0.0000
0.0005
0.0010
0.0015
0.0020
0
500
1000
1500
2000
Years BP

Figure S7

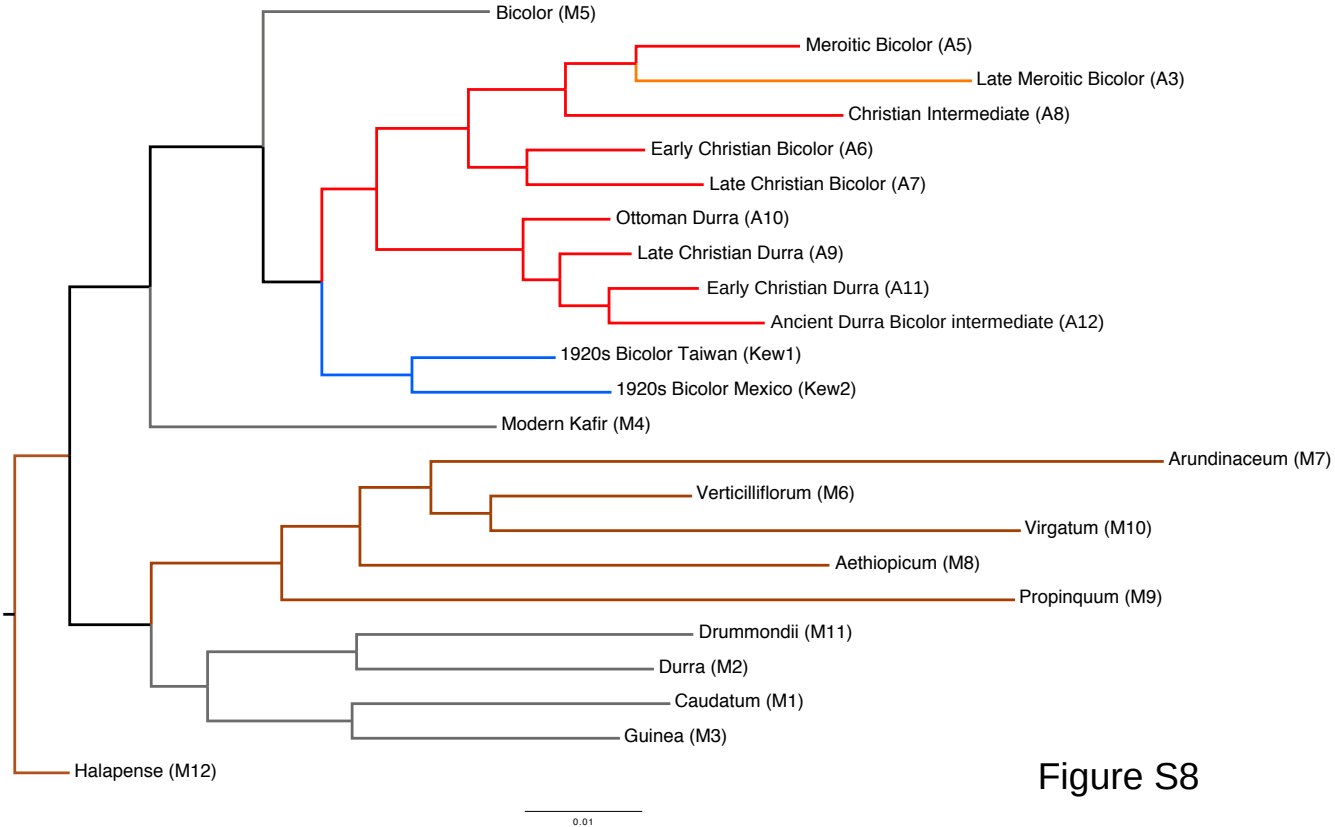

Bicolor (M5)
Meroitic Bicolor (A5)
Late Meroitic Bicolor (A3)
Christian Intermediate (A8)
Early Christian Bicolor (A6)
Late Christian Bicolor (A7)
Ottoman Durra (A10)
Late Christian Durra (A9)
Early Christian Durra (A11)
Ancient Durra Bicolor intermediate (A12)
1920s Bicolor Taiwan (Kew1)
1920s Bicolor Mexico (Kew2)
Modern Kafir (M4)
Arundinaceum (M7)
Verticilliflorum (M6)
Virgatum (M10)
Aethiopicum (M8)
Propinquum (M9)
Drummondii (M11)
Durra (M2)
Caudatum (M1)
Guinea (M3)
Halapense (M12)
0.01

Figure S8

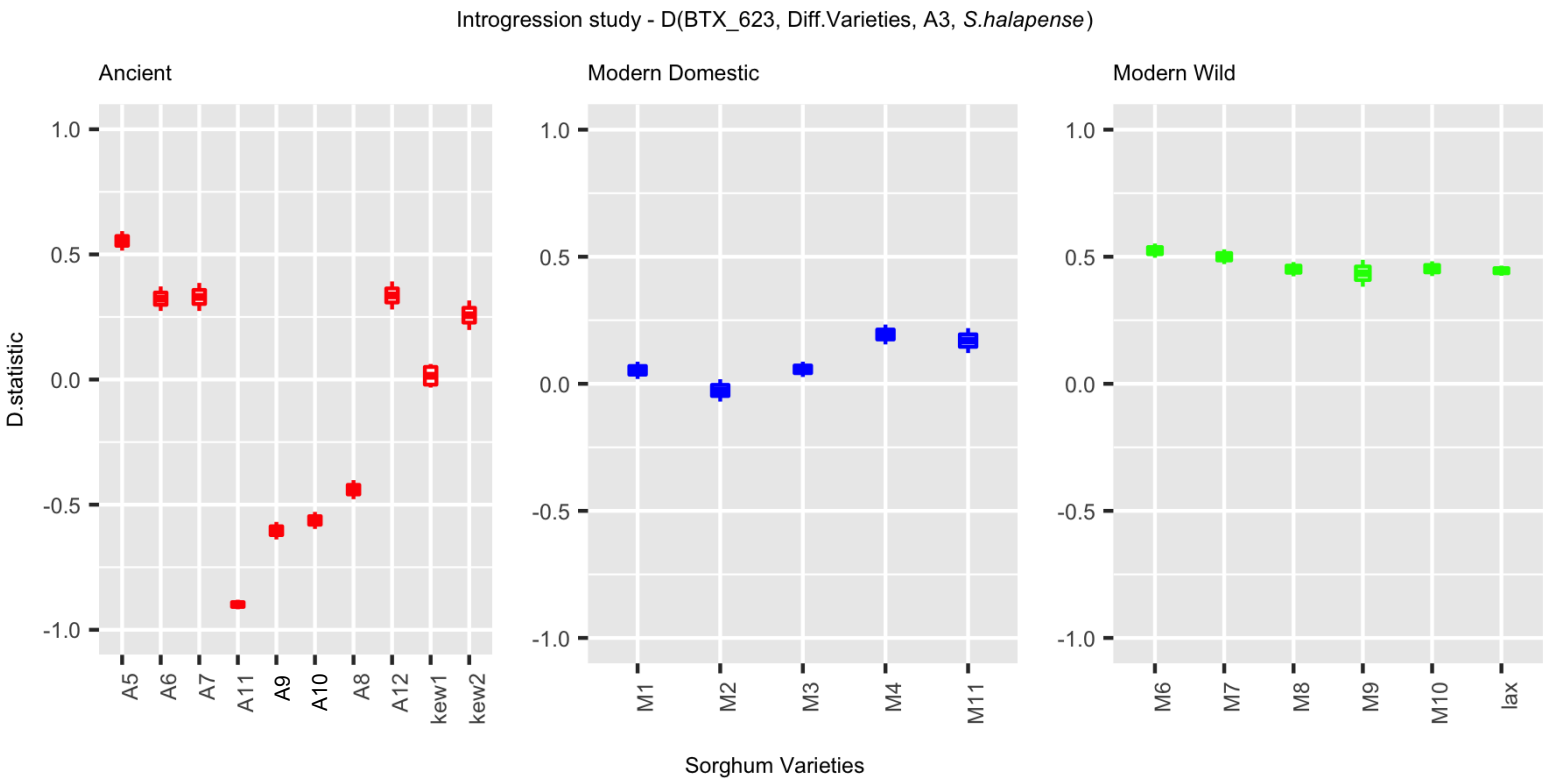

Introgression study - D(BTX_623, Diff.Varieties, A3, S.halapense)
Ancient
Modern Domestic
Modern Wild
D.statistic
1.0
0.5
0.0
-0.5
-1.0
A5
A6
A7
A11
A9
A10
A8
A12
kew1
kew2
M1
M2
M3
M4
M11
M6
M7
M8
M9
M10
lax
Sorghum Varieties

Figure S9

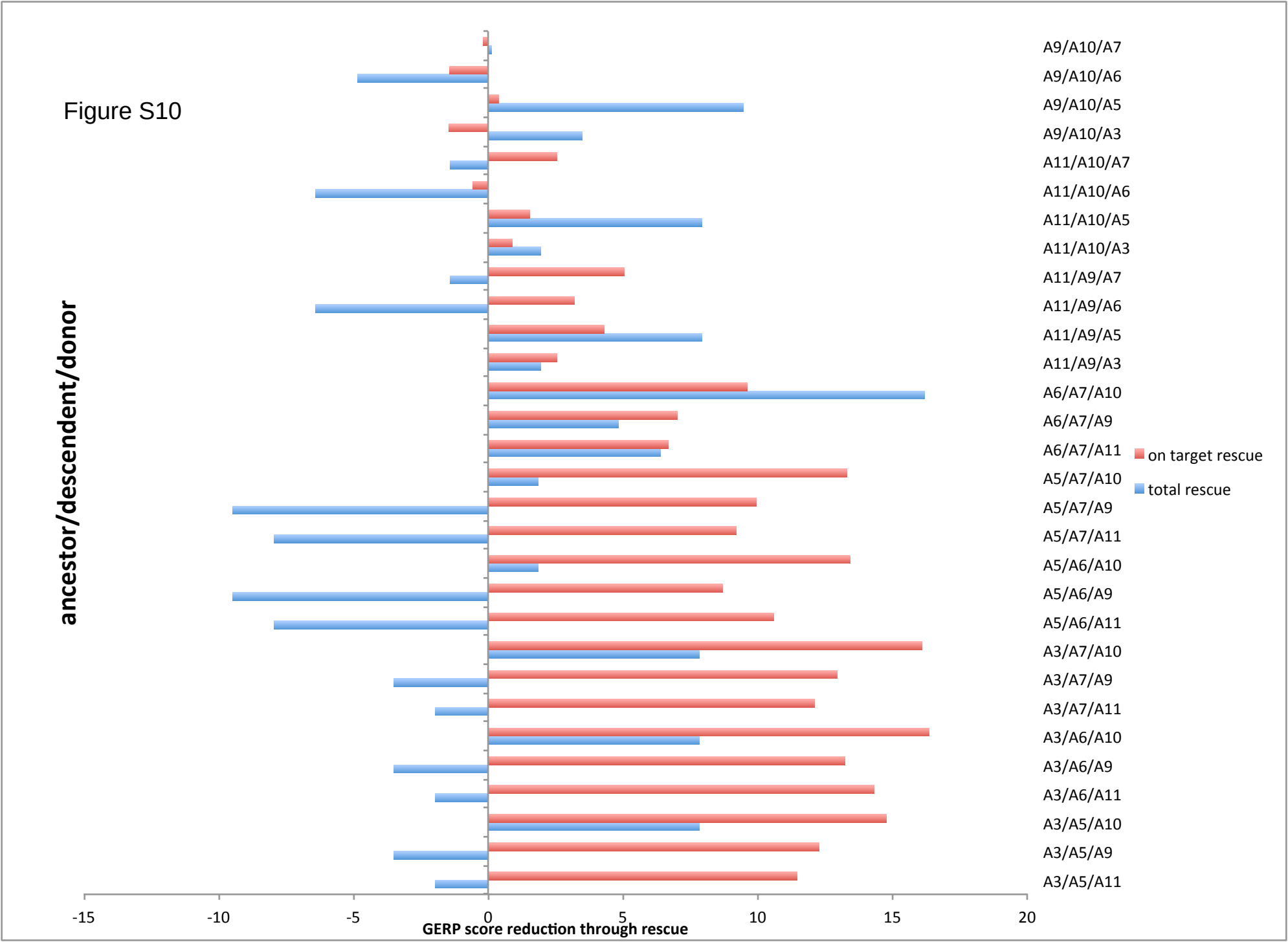


Figure S10
ancestor/descendent/donor
A9/A10/A7
A9/A10/A6
A9/A10/A5
A9/A10/A3
A11/A10/A7
A11/A10/A6
A11/A10/A5
A11/A10/A3
A11/A9/A7
A11/A9/A6
A11/A9/A5
A11/A9/A3
A6/A7/A10
A6/A7/A9
A6/A7/A11
A5/A7/A10
A5/A7/A9
A5/A7/A11
A5/A6/A10
A5/A6/A9
A5/A6/A11
A3/A7/A10
A3/A7/A9
A3/A7/A11
A3/A6/A10
A3/A6/A9
A3/A6/A11
A3/A5/A10
A3/A5/A9
A3/A5/A11
on target rescue
total rescue
-15
-10
-5
0
5
10
15
20
GERP score reduction through rescue

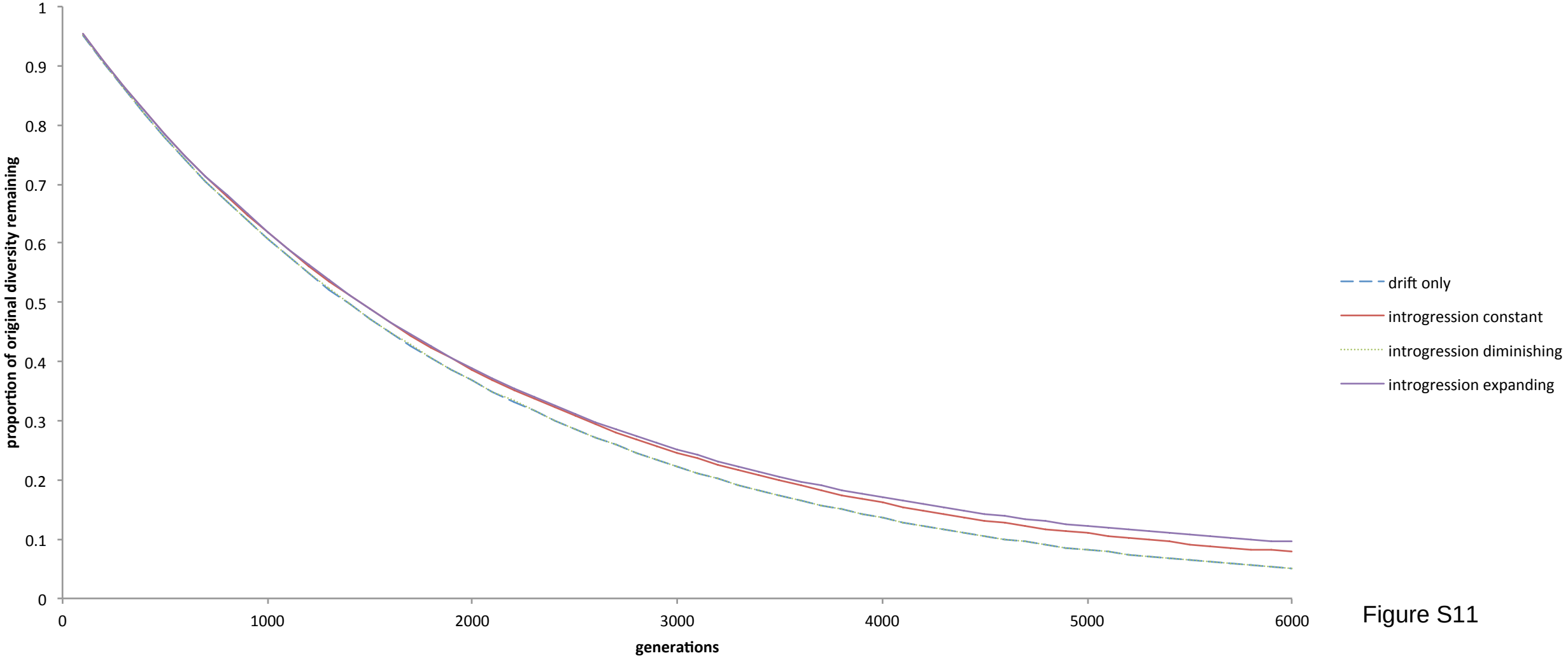
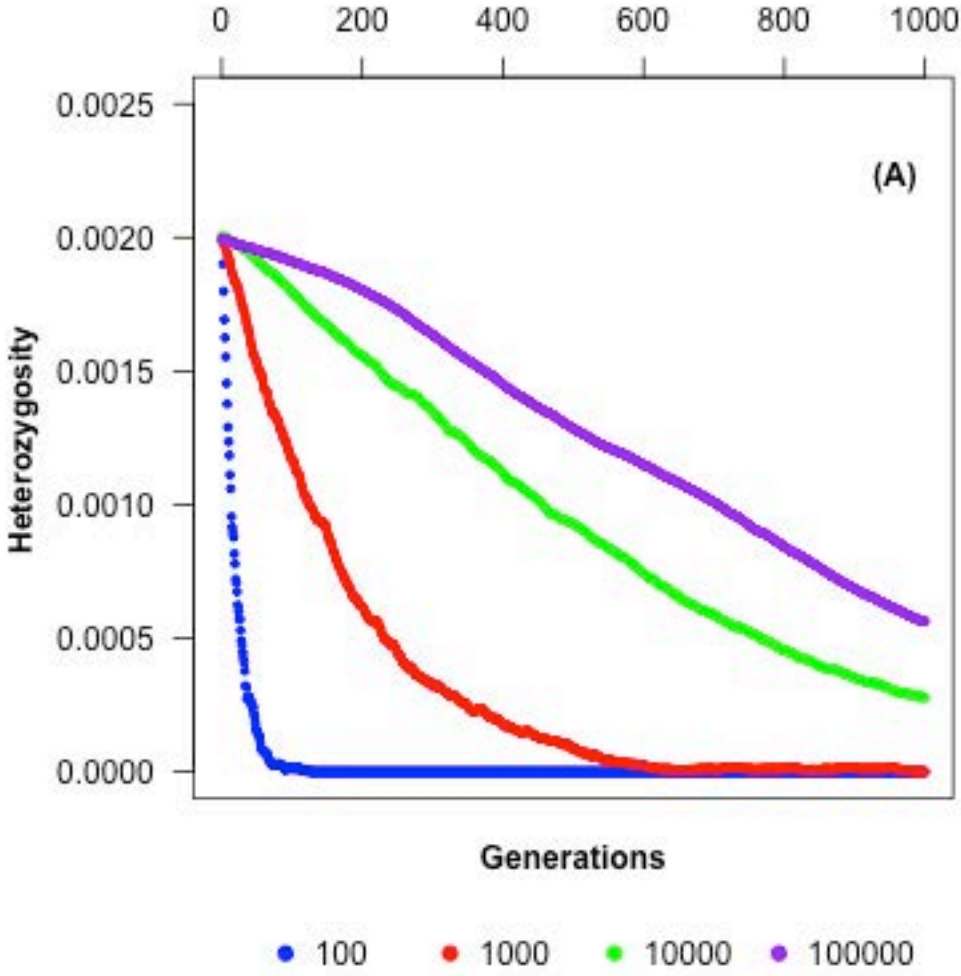

proportion of original diversity remaining
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
1000
2000
3000
4000
5000
6000
generations
drift only
introgression constant
introgression diminishing
introgression expanding

Figure S11


0
200
400
600
800
1000
0.0025
0.0020
0.0015
0.0010
0.0005
0.0000
(A)
Heterozygosity
Generations
100
1000
10000
100000

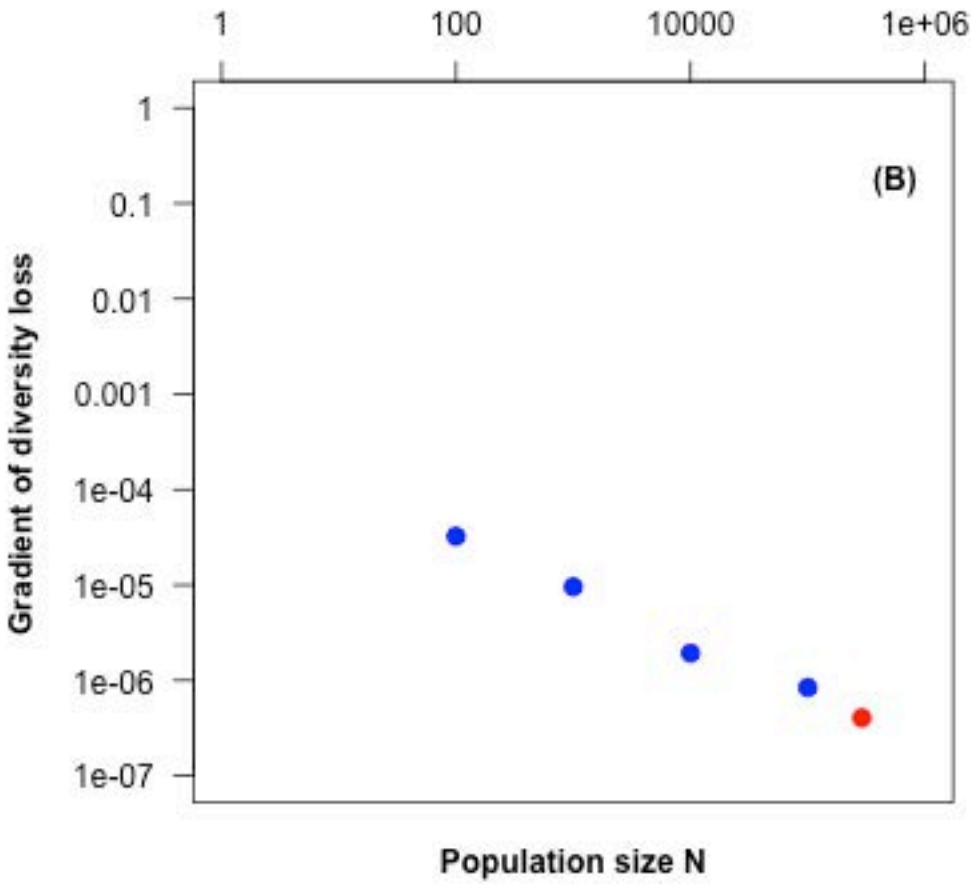

1
100
10000
1e+06
1
0.1
0.01
0.001
1e-04
1e-05
1e-06
1e-07
(B)
Gradient of diversity loss
Population size N
Model Gradient
Sorghum Gradient

Figure S12