

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 Mace⁵, Alan Clapham¹, Pamela Rose⁶, Chris Stevens⁷, Roselyn Ware¹, Siva Samavedam¹, Guy Barker¹, David Jordan⁸, Dorian Q Fuller⁷, 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, Smithsonian Institution, National Museum of Natural History, Washington, D.C. 20560, 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

Abstract

The evolution of domesticated cereals was a complex interaction of shifting selection pressures and repeated episodes of introgression. 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 to shape architectural and nutritional domestication traits, and to facilitate 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

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—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 (5-7). 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 (8) 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 form *Sorghum bicolor* type bicolor in Central Eastern Sudan by around 5000 years ago, while cultivation is inferred to have begun by 6000 years before present (yrs BP) (9). Ultimately, four specialized agroclimatic adapted types evolved after domestication—durra, caudatum, guinea, and kafir (10-12). 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 (10,13). 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 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 (14). Prior to this, apparently 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 (10,15). It is now thought that durra arose after an African bicolor

type reached the subcontinent 3.8 Kyr BP and hybridized with an indigenous wild species, with durra spreading to Africa at some point after 1800 yrs BP (13,16).

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 heterozygosity within genomes. First, we validated the use of genomic heterozygosity as a measure of genetic diversity, since heterozygosity can become depressed through excessive inbreeding, for example in the maintenance of germplasm. To do this, we determined the effective mating strategy of modern, wild, and ancient genomes based on their relative heterozygosity and comparable nucleotide diversity values (see methods), Figure S1, Table S3. Wild sorghum is expected to have a mating strategy of around 40% inbreeding while land races are expected to be closer to 80% inbreeding (17). One of our four wild genomes (race *verticilliflorum*) showed an appropriate level of effective mating strategy and so passed our validation, which we used as representative of sorghum wild progenitor diversity. The modern bicolor reference genome (BTX623) has a high effective mating strategy

of between 90-95% indicative of an elite inbred line. The ancient genomes showed intermediate mating strategies, suggesting that the switch from 40% to 80% inbreeding may have been a gradual process.

The heterozygosity of 100 kbp genomic blocks across the genome 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 S2. The remaining wild genomes all showed narrow and low ranges of heterozygosity indicative of inbreeding, as suggested by our validation, Figure S3.

The wild phenotype 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, and it post-dates the earlier bicolor type sample A5. Together, this evidence suggests that sample A3 is likely a feral form of sorghum resulting from introgression between wild and cultivated populations. 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, Figure S2, S4. Total genomic heterozygosity of bicolor over time confirmed that the feral 'wild' ancient 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 loss of diversity

through a domestication bottleneck (18), Figure 1, Table S4 suggesting that there was no measurable effect on genetic diversity attributable to a domestication bottleneck.

We further examined the pairwise sequentially Markovian coalescent (PSMC) profile of the bicolor type genomes (19) to generate a profile of past population size from the genomic diversity, Figure S5. This produced an estimate of a long-term trend of decreasing population size over the past 100 Kyr, a pattern similar to that previously observed in maize (20) and rice (21). This pattern is produced spuriously when sampling a single deme from a larger metapopulation that has received gene flow over time (22), which is likely to be the case for sorghum. All genomes also produce a strong signal of population expansion after 2000 years ago, which can also be produced by hybrid genomes between two populations (22,23). While there is no clear signal in these data for a domestication bottleneck, the profiles are indicative of long-term gene flow followed by a hybridization event between two populations that closely matches the arrival of the durra type sorghum at Qasr Ibrim.

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 (20,24,25), which have been taken as a confirmation of the effects of the domestication bottleneck, although other causes are possible such as hitchhiking effects during selection. 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 (26) (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 (20). 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 well described by either a positive exponential model that approaches a straight line (p values 1.08×10^{-12} and 0.0683 for parameters a and b respectively) or a linear model (p values 1.07×10^{-12} and 0.0686), Table S4, suggesting mutation load may have become increasingly 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 ongoing, 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 S6). 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. This notion is supported by the observation that 0.43-0.48 of genomic blocks associated with a decrease in the number of deleterious sites over time are associated with an increase in homozygous deleterious sites, Figure S7, Table S5. However, there is also evidence for expansions in deleterious sites in both sample A5 and BTX623 that are not associated with increased homozygosity indicating new influxes of mutation load not originating from the wild progenitor, Table S5. 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 S4, S6). The contrasting patterns in mutation load over time are also reflected in methylation state profiles, which can reflect the state of genome-wide stress (27), Figure S8.

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

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 S6, S7. We also specifically surveyed 38 known domestication loci and also found a significant reduction of heterozygosity in 15 of the 38 associated regions, Table S8. We further characterized the domestication loci in their pairwise nucleotide distance to the modern domestic alleles in BTX623, Figure S9. Secondly, we used a SweeD analysis to detect selective sweeps (28), which identified 11 peaks (denoted by prefix s), Figure 3, Table S9. 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, with oscillations in diversity discernible in the data after this time, Table S10, Figures S10-S12.

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 S13. 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 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 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 (*A11*). The *FAR1* gene is associated with phytochrome A signal transduction (29), so is important in responses to far red light that divert resources away from tall growth to increase root and grain growth. The *A11* gene regulates anther development (30), 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 S6). 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 S10). Selection

signatures were highly significantly associated with regions of maximum deviation in mutation load with 30% of low heterozygosity peaks ($p = 8.04 \times 10^{-9}$), 45% SweeD peaks ($p = 2.55 \times 10^{-7}$) and 26% domestication loci ($p = 5.03 \times 10^{-5}$) 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 S14), and individual trees for 970 sections across the genome (Supplementary data set 1). After accounting for biases introduced by ancient DNA modification (see methods), 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 (31) shows over time the durra type moved from a genomic background highly differentiated to that of bicolor and *S. verticilliflorum*, possibly relating to the putatively hybrid origin of durra (13,16), to becoming increasingly similar to the local bicolor type, suggesting progressive introgression between the two types (Figure S15). To estimate the proportion of the durra genome attributable to the hybrid origin background we examined f_d statistics. These show that in the case of the oldest durra (A11) about 25% of the

genome is of the highly differentiated background, while in the younger A10 this proportion has been reduced, Figure S16. We then compared the archaeological genomes against a global sorghum diversity panel (32,33) (Figure 4, Supplementary video). 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 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 S12). In agreement with previous studies (12) there is incongruence at the dwarfing *dw1* allele between durra and bicolor, with a single durra type sample sitting within the bicolor type clade in this region. Interestingly, seven of the nine sugar-metabolism associated loci potentially under selection in the bicolor type are also areas of incongruence with durra. 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. The *FAR1/A11* loci region, which appears to have been under strong selection in bicolor throughout, also is a region of incongruence with durra.

Assuming that the two types, bicolor and durra, had accrued mutation loads independently for the 2 Kyr (13) prior to the introduction of the durra type to Qasr Ibrim, 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 an 'unbiased rescue value' that reflects a donor's potential to reduce mutation load across the entire genome, Figure 5a. 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 most ancestor/descendent/donor combinations the reduction of mutation load in the descendent could be explained by a reduced load value in the donor, but often the unbiased exchange of load across the genome would be expected to result in an overall increase in mutation load. We considered whether if any of these introgressions occurred there was a bias towards genetic exchange that reduces mutation load. To do this we considered the expected proportion of instances in which both the descendent and donor

would both have a lower or higher load than the ancestor by chance relative to the observed values, Figure 5b, Table S13. In general, there was a highly significant threefold increase in the co-occurrence of lower load in the descendants and donors, but only a marginal increase in the co-occurrence of higher load. This is strong evidence that one or more of these introgression scenarios occurred, and that there was a bias in genetic exchange in regions of lower load in the donor. In the case of bicolor, the most significant scenario is A5/A6/A11 for ancestor/descendent/donor respectively, and in the case of durra A11/A9/A7. Both of these scenarios make chronological sense and indicate that a process of mutual rescue between durra and bicolor occurred.

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 sorghum would be expected to lead to a strong 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 S17. 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 extensive introgression leading to a strong regional structure within cultivars (12), 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 S13. 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.

The transition in mating system from 40% to 80% inbreeding over time in sorghum would be expected to reduce heterozygosity, and a model of constant change in mating system that is supported by the observations in this study would be expected to produce a linear decrease. However, we calculate that the ten-fold reduction in heterozygosity observed would require a 90-95% change in inbreeding, much wider than that observed in sorghum, making this unlikely to be a complete explanation.

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 (34), and a similar observation has been made in maize (20). The annual cycle of crop sowing and harvesting in sorghum also represents a serial founding event scenario. Ethnobotanical evidence shows that only 1% of the sorghum population contributes to the next generation as farmers set aside a small number of plants to be the parents of the next generation, in contrast to practices with other cereals such as wheat and barley where seeds are sampled from across the whole population (35). We explored five population simulation models in which diversity was lost over time. Two models included a 1% annual founding event, one with a switch from 40% to 80% inbreeding in a single step at the point of domestication (model 1), the other with a gradual shift in mating systems from 40% to 80% inbreeding over 6000 years (model 2). The remaining three models lacked founding events, but respectively had either a gradual switch in mating system (model 3), a single step mating system switch at domestication (model 4), or no change in mating system and a continued 40% inbreeding level (model 5).

For each model we determined the population size associated with the gradient of heterozygosity loss observed in our data from the linear regression of gradients of loss associated with differing orders of magnitude of population size (Figure S18, Table S14). We then simulated the models at their respective population sizes over 6000 generations, Table S15. All models with the exception of model 2 had a better fit to an exponential rather than a linear trend. Model 2

was the closest fit to a linear descent in diversity ($R^2 = 0.97$), model 3 gave a weaker fit ($R^2 = 0.90$). The population size associated with model 3 was small ($N=15955$), while model 2 was considerably larger ($N=727778$). Given sorghum planting densities (35), model 3 equates to a 4-hectare system, while model 2 equates to 204 hectares. Subsistence farmers typically utilize a single one-hectare field of sorghum (35), reflecting highly localized and regional levels of diversity for models 3 and 2 respectively. Sorghum genetic diversity is structured on the regional scale in Africa within small language groups (36) in agreement with the regional scale represented in model 2. Together this evidence supports a scenario in which genetic diversity was lost in sorghum at a regional scale due to the founding effects of the agricultural practices involved. This may explain why similar large losses of diversity are not observed in other cereals such as maize and barley (37). The process likely 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 (38). 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 number of sites holding deleterious variants decreases

(dominant model). This trend suggests a 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 (39) but it remains unclear whether crops could follow the same fate in the absence of rescue processes, or whether such episodes could have been involved with previous agricultural collapses when crops experienced extensive adaptive challenges (40,41). 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 related 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 assignment '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. verticilliflorum*, *S. arundinaeum*, and *S. aethiopicum* were also 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 (33).

1.2 A note on taxonomy. The sorghum genus is complex with numerous taxonomic systems. After Morris *et al.*'s findings (12), 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. The subtype we used in the analysis as representative of the wild sorghum progenitor was race verticilliflorum for two reasons. Firstly, the remaining three genome types we sequenced all showed low levels of heterozygosity compatible with a high degree of germ plasm inbreeding leading to effective mating strategy estimates of 85% and 95% (see main text), inappropriate for wild sorghum which is expected to be closer to 40%. The second reason was that race types verticilliflorum and aethiopicum would be geographically more appropriate wild ancestors relative to the other two subraces being distributed across northern Africa.

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 preserved as an island (42,43). The desert conditions provided exceptional organic preservation by desiccation with exceptional preservation of a wide range

of biomolecules (e.g. 44-46). Systematic sampling for plant remains was initiated in 1984 (47) and the first studies of these remains were carried out in the 1980s by Rowley-Conwy (48) and had continued by Alan Clapham (49,50). The exceptional plant preservation has previously allowed successful ancient genomic studies of barley (44) and cotton (45).

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 AD 1812. 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 from the start of the 20th century as ca. 100 BP. For material not directly dated, sample A12 could be assigned based on associated pottery and finds, which have a well-established chronology through the Christian

periods (51), 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 µl CTAB buffer as previously described, with the addition of 1µl β-mercaptoethanol. Debris was centrifuged down and the supernatant was extracted once with an 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 (52). 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). We compensated for G to A mismatches and 5mC to T mismatches for later analyses, depending on potential biases that could be introduced (see sections 4 and 5). 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 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 (53) using bowtie2 v2.2.9 (54) 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 (55). To eliminate the possibility of downstream biases resulting from damage patterns typical of ancient DNA, deamination patterns were then identified and masked from BAM alignments using a Perl script of LK's design. The overall genome coverage of all archaeological samples was sufficient to ensure no low-coverage biases occurred. Variant calls format (VCF) files were then made from pileups constructed using samtools mpileup, and variant calls were made using bcftools v1.4 (55).

5. *Methylation analysis.* Since a uracil-intolerant polymerase was used for library generation, we analysed BAM files, without the masking treatment as described above, using epipaleomix (52) 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 due to their wider methylation states (CHH and CHG) (56). However, when assessing 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 (53). The frequency distribution of heterozygosity was then calculated by binning the windows in 1 heterozygous base site intervals. Genomic heterozygosity can become depressed under conditions of inbreeding, such as through the

maintenance of germplasm. Therefore, we validated heterozygosity as a measure of genetic diversity by determining the extent to which inbreeding had reduced diversity relative to the population, and related this to the mating strategy. To do this we calculated the expected reduction in heterozygosity due to inbreeding from values of nucleotide diversity by determining the inbreeding coefficient F for a given mating strategy using the estimate F function of the *selection time* program (57). We then used the nucleotide diversity value for wild sorghum of 0.0038 (32) as a comparison to the genomic heterozygosity values obtained from the wild sorghum samples in this study. This yielded estimates of 40% inbreeding for our race *verticilliflorum*, which is close to the expected value for wild sorghum (17), Figure S1. The remaining wild sample genomes (*aethiopicum*, *arundinaceum* and *virgatum*) showed effective mating strategies of 85%, 95% and 95% respectively, indicative of extensive inbreeding. We therefore used the race *verticilliflorum* as our representative genome of wild sorghum. In the case of the ancient samples, we estimated a nucleotide diversity value by calculating the average pairwise distance between the A5, A6 and A7 genomes for bicolor types, and A11, A9 and A10 genomes for durra types, Table S3. Ancient bicolor showed an effective mating strategy of 60% inbreeding, while ancient durra was 50% inbreeding, Figure S1. In the case of A3, which was a feral type, there was no comparable nucleotide diversity to use for validation. Modern bicolor was compared to the elite line nucleotide estimate of Mace et al (32) of 0.0023, yielding an estimate of 90-95% inbreeding indicative of an inbred line.

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 S8 and S10) were obtained by reference to the BTX_623 genome. Candidate domestication loci were obtained from the scans of Mace *et al* (32). 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 domesticated 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 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 PSMC estimates

A Pairwise Sequential Markovian Coalescent (PSMC) approach was taken to infer population history for all ancient and historical samples, using the PSMC package (19). Input parameters were defined by the following: average coverage

depth per sample was calculated using the samtools depth function, and applied to `-d` (at one third) and `-D` (at 2x) parameters of the `vcf2fq` function of `vcfutils.pl` to generate diploid consensus sequences with bins of 100bp from soft-masked BAM files. The PSMC utility `fq2psmcfa` was then applied with a minimum mapping quality of 20. The PSMC algorithm itself was then run using the default parameters. The PSMC output was then rescaled and plotted using `psmc_plot.pl`, assuming a generation time of 1 year for an annual plant such as sorghum, and mutation rate of 3.0×10^{-8} substitutions per site per generation, being the closest estimate of mutation rates available from domesticates of the Panicoideae subfamily (58).

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 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 (32) were combined using the GATK (59) program CombineVariants. Subsequently, the combined VCF file was filtered - using bcftools v1.4 (55) - 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 (60). SweeD (28) 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.* (26). 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 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 substitutions (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.* (20): 0, neutral, 0-2 slightly deleterious, 2-4, moderately deleterious,

>4 seriously deleterious. We collected scores under three models, recessive, 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 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 ($p1$). We assessed whether the instances of on target rescue occurred significantly more frequently than would be expected by chance given the descendent and potential donor GERP profiles, by comparing the observed proportion of windows satisfying on target rescue with the expect proportion given the frequency of the component conditions of descendent < ancestor and donor < ancestor ($e1$). The number of windows satisfying a set of criterion out of a set number of windows follows a Binomial distribution, so we used the posterior Beta distribution to derive p values which is a better fit to the data than the Chi squared distribution in this case. Beta distributions were generated for both the observed and expected number of events, and the p values taken from the overlap of the two distributions representing the probability that both values could have been drawn from the same distribution. We repeated this analysis to compare the incidences in which both the descendent and the potential donor exceeded the GERP load score of the ancestor ($p2$), and their corresponding expected values ($e2$).

6.5 *Phylogenetics* Maximum likelihood trees were constructed using exaML (61) firstly using whole genome sequences (Figure S14), 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 (62) 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 (63) - developed by the same research group - for computing maximum parsimony trees. An ExaML executable (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 recombination between genomes at candidate loci we examined trees spanning the corresponding 100kbp 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 possible using this approach. Simple cases in which a single genome from one sorghum type was found within the group of the other type were identified as incongruent taxa. 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 common SNPs were identified between two separate data sets; the resequencing data of the 23 samples and GBS data from an unpublished set of 1046 diverse sorghum lines, which span the racial and geographic diversity of the primary gene pool of cultivated sorghum. The subset of SNPs were selected based on a minor allele frequency of >0.05 and frequency of missing data points <0.2 . 580 of these 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 (31). 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 (64).

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=*S. verticilliflorum*. Second, using set.outgroup function, we set an outgroup (P4= *S. halapense*). Third, the method introgression.stats was employed to calculate the introgression tests. Finally, we used jack.knife.transform method (64) which transforms an existing object belonging to GENOME class into another object of the GENOME class with regions that corresponding to a Jackknife window. 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 modern wild *S. verticilliflorum*, 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 P₁, *S. verticilliflorum* was taken as P₃ and *S. halapense* was taken as the out group P₄. *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 *S. verticilliflorum* than P_2 . This is observed as expected for the durra types since they are thought to have genomic introgression from wild sorghum species on the South Asian subcontinent which is expected to unite P_1 and P_3 in regions where durra is different to both bicolor and its wild ancestor. 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. We affirmed this scenario by placing the oldest and youngest durra samples (A11 and A10 respectively) as P_1 . While A11 consistently shows highly positive values, indicating P_2/P_3 combinations, A10 shows a reduced positive D score with bicolors, and highly negative D scores with A11, indicating introgression from the *verticilliflorum* background, most likely directly from bicolors.

Positive values (indicating an excess of P_2/P_3 combinations) suggest a close relationship between *S. verticilliflorum* and P_2 , which is observed the Qasr Ibrim bicolor types (A5, A6 and A7). This is expected if these ancient bicolors share diversity with the wild progenitor that has been subsequently lost in modern *S. bicolor*.

Since the f_d statistic cannot be meaningfully interpreted from negative D statistic values, to investigate the extent to which the durra genomes are made up of an unknown genome putatively of South Asian origin we repeated the D

statistic analysis using the durra samples A10 and A11 at position P1 and examined the resulting fd statistics (Figure S16). Bicolor types group with verticilliflorum to the exclusion of durra (P2/P3 combinations) are in excess by 25%. These values lead us to conclude that about 0.25 of the durra genome is constituted of the unknown genome.

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

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 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:

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

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

The serial founder event simulation was executed using the program *founderv6.pl* available for download at :

(<https://warwick.ac.uk/fac/sci/lifesci/research/archaeobotany/downloads/founderv6>) following model:

To initialize, allele frequencies were randomly assigned for a defined number of alleles, and the associated heterozygosity calculated for a given breeding system for which the inbreeding coefficient F was calculated using the estimate F function of the *selection time* program (57). Frequencies were then adjusted randomly in a Markov Chain until a user defined starting heterozygosity was reached. In this case all simulations were initiated at a starting heterozygosity

equivalent to 0.003 and 40% inbreeding to reflect wild sorghum. The genotype frequency distribution was then calculated for the given inbreeding coefficient F and allele frequencies. Individuals were drawn from the population by randomly selecting two parent genotypes, gametes were either drawn from one or both parents with a probability determined by the mating system. Gametes were mutated to generate new alleles based on a general plant rate of 5×10^{-9} substitutions/site/year (65). N individuals were randomly drawn in this way, and the resultant allele frequency distribution calculated. A founder event was then generated by drawing N_b individuals from the allele frequency distribution, following the same methodology as above. The process was repeated for a defined number of cycles, with the inbreeding coefficient being recalculated each generation in simulations where the mating system changed over time.

We explored five model scenarios using this simulation, over several orders of magnitude of N (100, 1000, 10000, 100000, 1000000), for 1000 founding events, equating to 1000 years of agriculture. Each trial was repeated 100 times, equating to 100 genes being simulated independently. Model 1: 80% inbreeding population from the point of domestication with a 1% bottleneck each generation; model 2: dynamic mating system switch from 40% inbreeding which would reach 80% inbreeding over 6000 years with a 1% bottleneck each generation; model 3 dynamic mating system switch from 40% inbreeding which would reach 80% inbreeding over 6000 years with no bottlenecks; model 4 80% inbreeding population from the point of domestication with no bottlenecks; model 5 static 40% inbreeding with no bottlenecks.

The log of the initial gradient of loss of diversity plotted against the log of the population size for each model gave a linear relationship from which a linear regression was derived. The linear regression was then used to derive the population size for each model that corresponded to the observed gradient of diversity loss observed in the real archaeogenomic data, Table S14. Simulations were then carried out under the five different models and their respective predicted population sizes, each repeated ten times.

Data availability

Sequence data were deposited in the European Molecular Biology Laboratory European Bioinformatics Institute [project code PRJEB24962.].

Code availability

The serial founder event simulation was executed using the program

founderv6.pl available for download at :

(<https://warwick.ac.uk/fac/sci/lifesci/research/archaeobotany/downloads/founderv6>).

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).

Competing interests

The authors declare no competing interests.

- (1) Larson G, Piperno D, Allaby RG et al. Current perspectives and the future of domestication studies. *Proceedings of the National Academy of Sciences 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. *Proceedings of the National Academy of Sciences 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 Biology* **16**, 173 (2015).
- (6) Hardigan MA, Laimbeer FPE, Newton L, Crisovan E, Hamilton JP, Vaillancourt B, Wiegert-Rininger K, Wood JC, Douches DS, Farré EM et al. 2017. Genome diversity of tuber-bearing solanum uncovers complex evolutionary history and targets of domestication in the cultivated potato. *Proceedings of the National Academy of Sciences*, 114: E9999–E10008.

- (7) Hufford M, Lubinsky P, Pyhäjärvi T, Devengenzo M, Ellstrand N, Ross-Ibarra J. 2013. The genomic signature of crop-wild introgression in maize. *PLoS Genetics*, 9: e1003477
- (8) Food and Agriculture Organization of the United Nations. [http://www.fao.org/index_en.htm].
- (9) 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).
- (10) Doggett H *Sorghum* 2nd Ed, Longman, Harlow (1988).
- (11) Brown PJ, Myles S, Kresowich S. Genetic support for a phenotype-based racial classification in sorghum. *Crop Sci.* **51**, 224-230 (2011).
- (12) Morris G, *et al.* Population genomic and genome-wide association studies of agroclimatic traits in sorghum. *Proceedings of the National Academy of Sciences U.S.A.* **110**, 453-458 (2013).
- (13) 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).
- (14) Clapham AJ, Rowley-Conwy PA In Fields of Change—Progress in African Archaeobotany, Cappers R, ed. Groningen Archaeological Studies. Groningen, **5**, 157–164 (2007).

- (15) 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).
- (16) 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).
- (17) Ohadi S, Hodnett G, Rooney W, [Bagavathiannan](#), M. Gene Flow and its consequences in Sorghum spp. *Crit RevCritical Reviews in Plant SciSciences* 36:367-385 (2017).
- (18) Meyer R, Purugganan M. Evolution of crop species: genetics of domestication and diversification. *Nature Reviews Genetics* **14**, 840-852 (2013).
- (19) Li H, Durbin R. Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493–496 (2011).
- (20) 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).
- (21) Meyer RS *et al.* (2016) Domestication history and geographical adaptation inferred from a SNP map of African rice. *Nature Genetics* 48:1083-1088.
- (22) Mazet O, Rodríguez W, Grusea S, Boitard S, Chikhi L 2016. On the importance of being structured: instantaneous coalescence rates and human evolution-- lessons for ancestral population size inference. *Heredity* 116: 362-371 (2016).

- (23) Orozco-terWengel P. The devil is in the details: the effect of population structure on demographic inference. *Heredity* 116: 349-350. (2016).
- (24) 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).
- (25) 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).
- (26) Cooper GM *et al.* Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.* **15**, 901-913 (2005).
- (27) 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).
- (28) 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).
- (29) 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).
- (30) 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).

- (31) 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).
- (32) Mace ES *et al.* Whole genome sequencing reveals untapped genetic potential in Africa's indigenous cereal crop sorghum. *Nature Communications* **4**, 2320 (2013).
- (33) Thurber CS, Ma JM, Higgins RH, Brown PJ Retrospective genomic analysis of sorghum adaptation to temperate-zone grain production. *Genome Biol.* **14**, R68 (2013).
- (34) 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).
- (35) Alvarez N, Garine E, Khasah C, Dounias E, Hossaert-McKey M, McKey D). Farmers' practices, metapopulation dynamics, and conservation of agricultural biodiversity on-farm: a case study of sorghum among the Duupa in sub-sahelian Cameroon. *Biological Conservation* 121:533-543. (2005).
- (36) Westengen OT, Okongo MA, Onek L, Berg T, Upadhyaya H, Birkeland S, Khalsa SDK, Ring KH, Stenseth NC, Brysting AK. Ethnolinguistic structuring of sorghum genetic diversity in Africa and the role of local seed systems. *Proc. Natl. Acad. Sci. USA* 111:14100-14105. (2014).

- (37) Allaby RG, Ware R, Kistler L. A re-evaluation of the domestication bottleneck from archaeogenomic evidence. *Evolutionary Applications* doi.org/10.1111/eva.12680 (2018).
- (38) Kremling KA *et al.* Dysregulation of expression correlates with rare-allele burden and fitness loss in maize. *Nature* 555:520-523 (2018).
- (39) Rogers & Slatkin Excess defects in a woolly mammoth on Wrangel Island *PLoS Genetics* 13(3): e1006601 (2017).
- (40) Shennan S, Downey SS, Timpson A, *et al.* Regional population collapse followed initial agricultural booms in mid-Holocene Europe. *Nat Commun.* 4:2486 (2013).
- (41) Allaby RG, Kitchen JL, Fuller DQ. Surprisingly low limits of selection in plant domestication. *Evolutionary Bioinformatics* 11:(S2) 41-51 (2016).
- (42) Alexander, J The Saharan divide in the Nile Valley: the evidence from Qasr Ibrim. *The African Archaeological Review* 6, 73-90 (1988).
- (43) Rose P. 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 (2013).
- (44) 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).
- (45) 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).

- (46) 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).
- (47) Alexander, J , Driskell B. Qasr Ibrim 1984. *Journal of Egyptian Archaeology* **71**, 12-26 (1985).
- (48) 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).
- (49) Clapham AJ, Rowley-Conwy PA. New discoveries at Qasr Ibrim, Lower Nubia. *Fields of change: progress in African archaeobotany*. Barkhuis & Groningen University Library, Groningen, The Netherlands, 157-164 (2007).
- (50) Clapham, A., & Rowley-Conwy, P. 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 (2009).
- (51) Adams, WY. *Ceramic Industries of Medieval Nubia*. University of Kentucky Press (1986)
- (52) 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).

- (53) Paterson AH *et al.* The Sorghum bicolor genome and the diversification of grasses. *Nature* **457**, 551-556 (2009).
- (54) Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nature Methods*. **9**(4), 357-359 (2012).
- (55) 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).
- (56) Bouyer et al. DNA methylation dynamics during early plant life. *Genome Biology* 18:179 (2017)
- (57) Allaby, R. G., Stevens, S., Lucas, L., Maeda, O., & Fuller, D. Q. Geographic mosaics and changing rates of cereal domestication. *Philosophical Transactions of the Royal Society B*, 372, 20160429. (2017). Clark, R.M., Tavaré, S., & Doebley, J. Estimating a Nucleotide Substitution Rate for Maize from Polymorphism at a Major Domestication Locus. *MBE* 22 (11): 2304-2312 (2005).
- (58) 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).
- (59) 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).

- (60) Koslov AM, Aberer A, Alexandros S. ExaML version 3: a tool for phylogenomic analysis on supercomputers. *Bioinformatics* **31**, 2577-2579 (2015).
- (61) Felsenstein J Evolutionary trees from DNA sequences: A maximum likelihood approach. *J Mol Evol* **17**(6), 368-376 (1981).
- (62) Stamatakis A RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**(21), 2688-2690 (2006).
- (63) 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).
- (64) Wolfe KH, Sharpe PM, Li WH. Rates of synonymous substitution in plant nuclear genes. *J. Mol. Evol.* 29:208-211 (1989).

Figure 1. Genomic heterozygosity over time in *S. bicolor* type bicolor (N=7).

Figure 2. Total recessive GERP load over time in *S. bicolor* type bicolor (N=7).

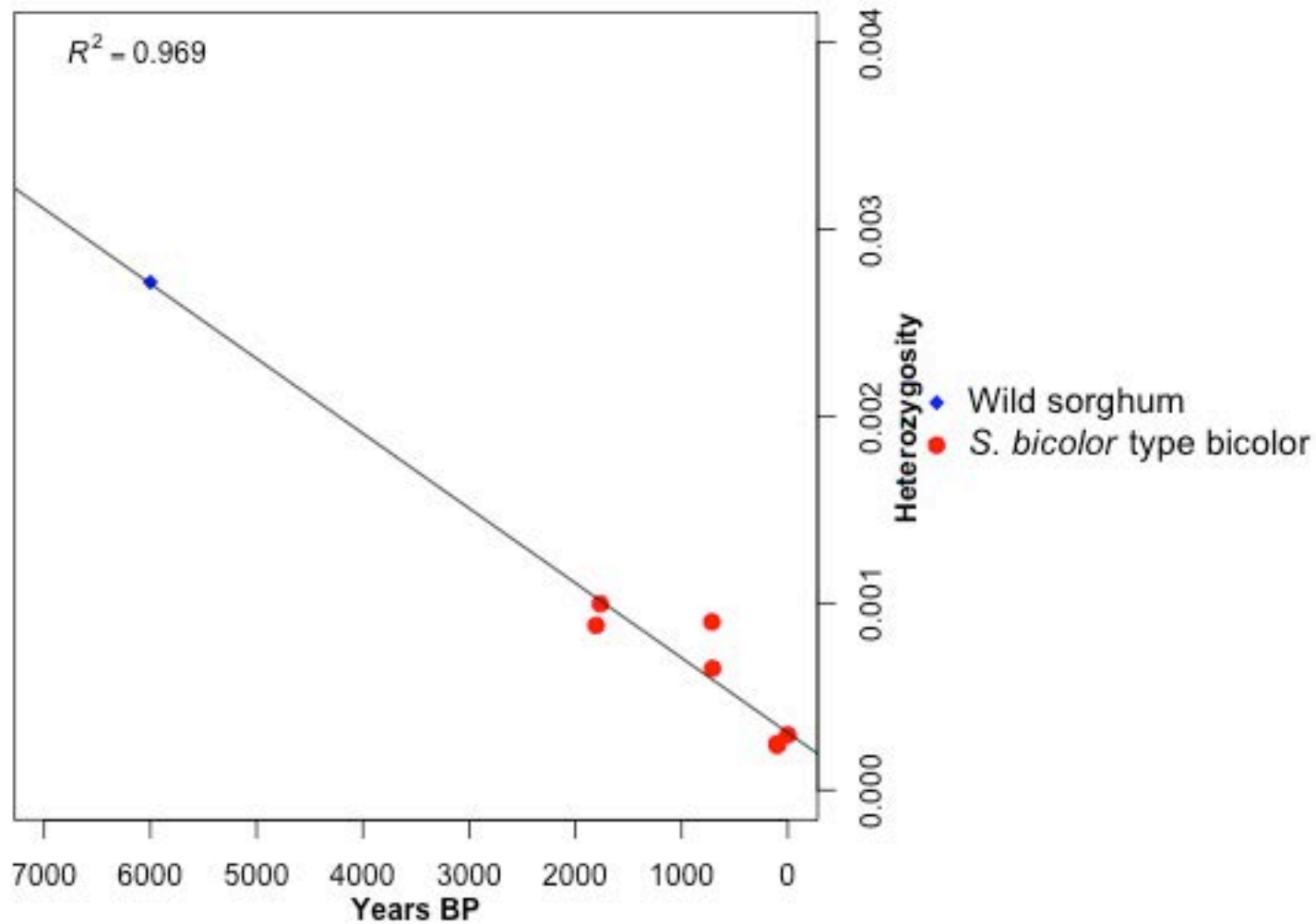
Figure 3. Selection signals across *S. bicolor* chromosomes 1 to 10.

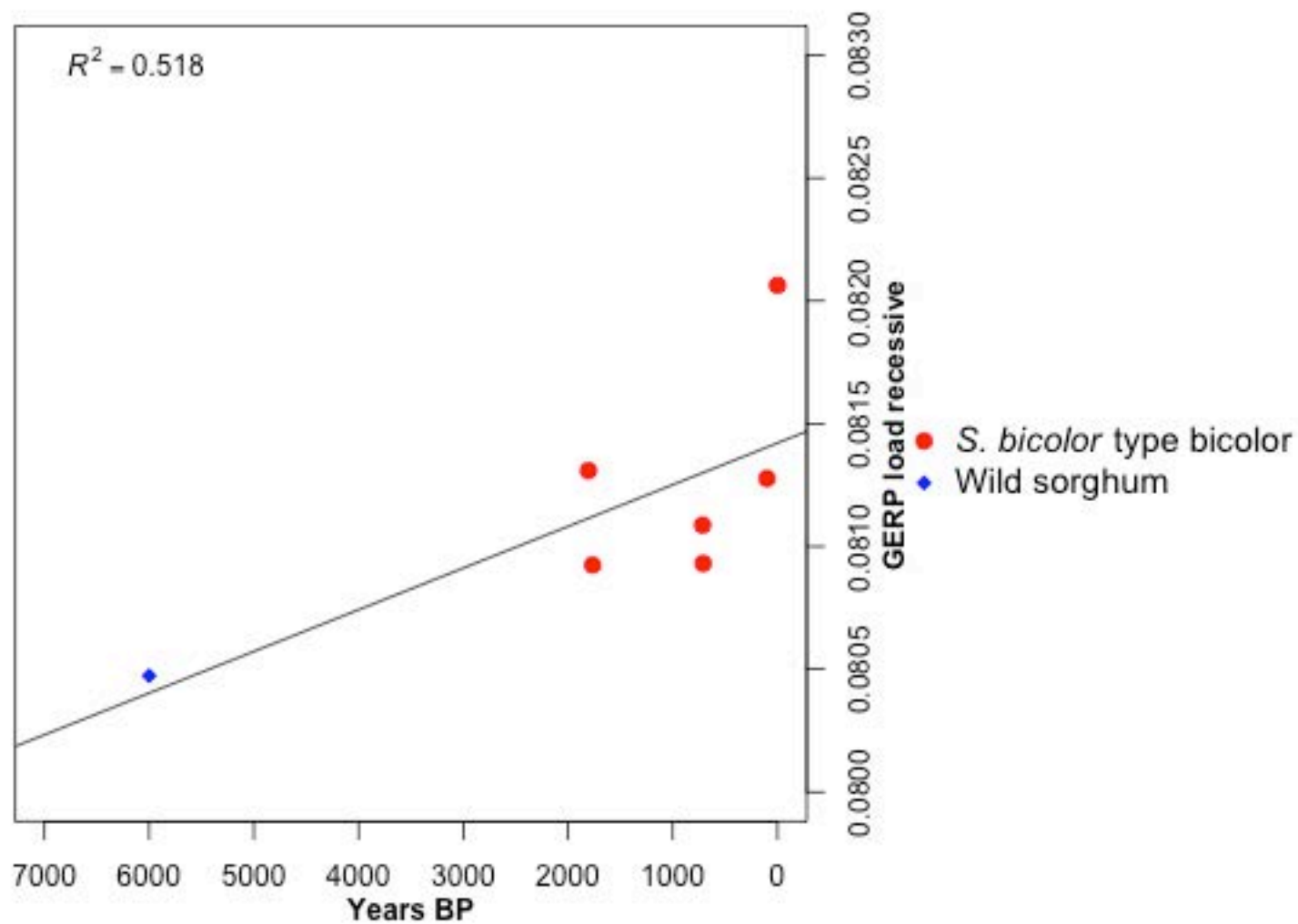
Heterozygosity ratio (wild/cultivated) inverted probabilities (Bonferroni corrected) shown in colours as described in key. Grey dashed line indicates 1% significance threshold after Bonferroni correction. SweeD values shown in red. Above: Locations of 38 known domestication genes shown in black. Locations 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.

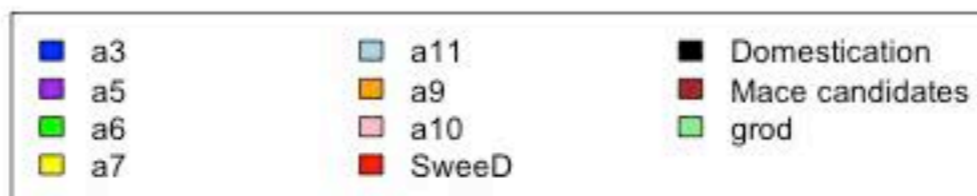
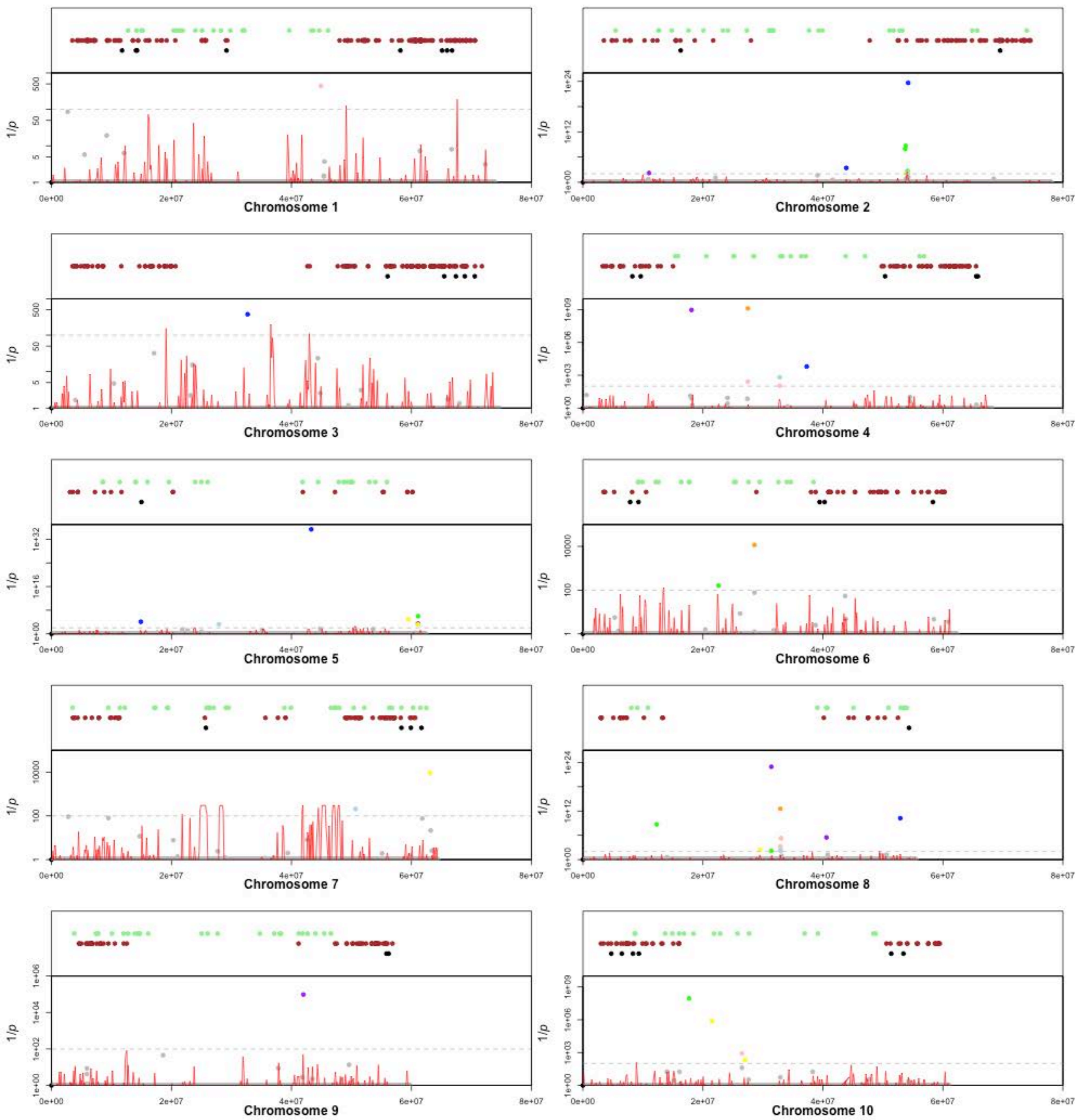
Figure 4. Principal Coordinate Analysis of 1894 SNPs from 23 genomes in this study and 1046 sorghum lines described in Thurber *et al* (25).

Figure 5. Genome rescue between bicolor and durra lineages. A. 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 gerpGERP score from combining ancestor and donor scores. B. *P* values for the extent of overlap in genomic regions between descendents and potential donors that are lower (*p*1) or higher (*p*2) in

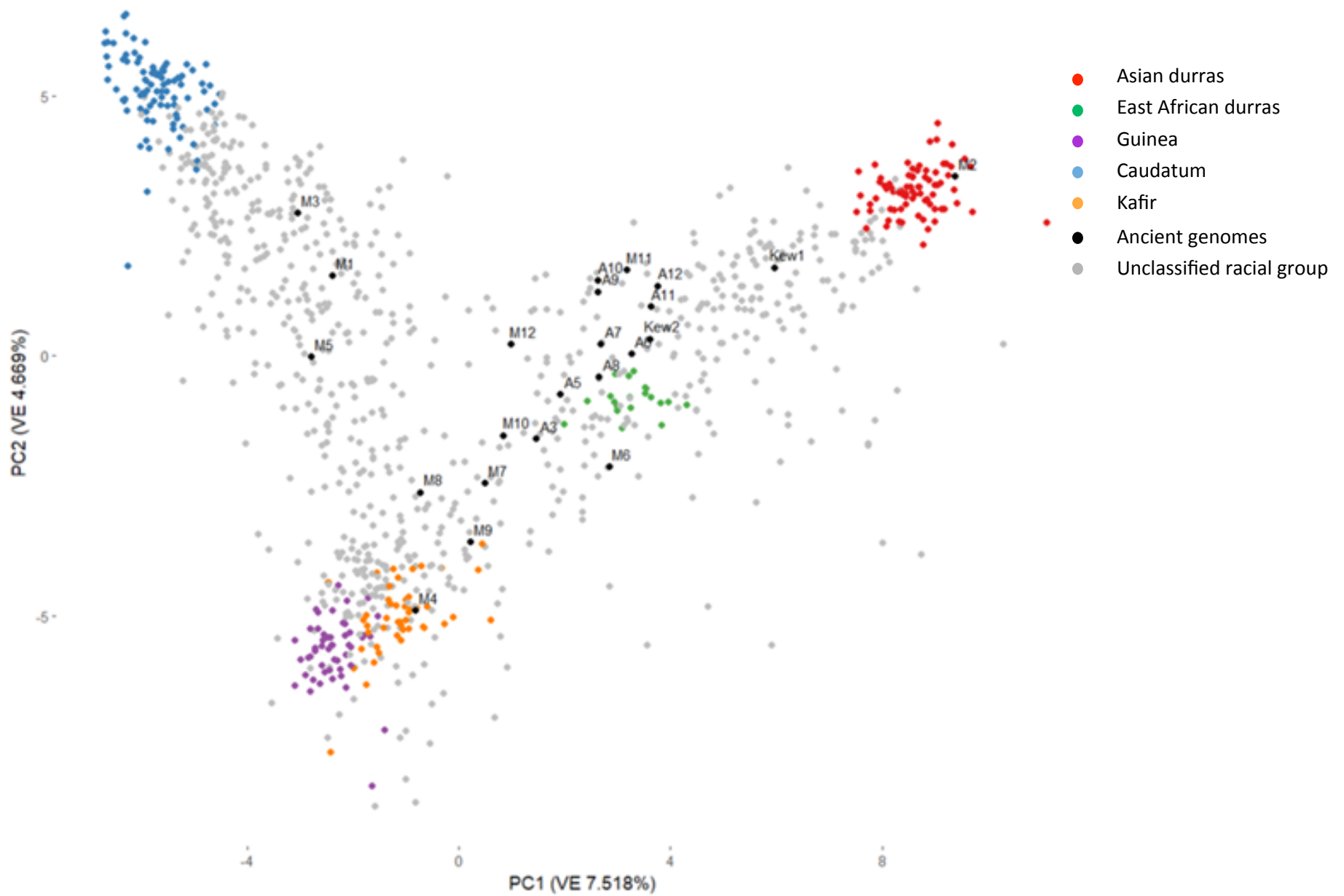
GERP score than the ancestor. P values calculated from Beta Distribution overlap of observed and expected proportions of coincidence in mutation load states relative to ancestor (1-tailed test), see methods for details.



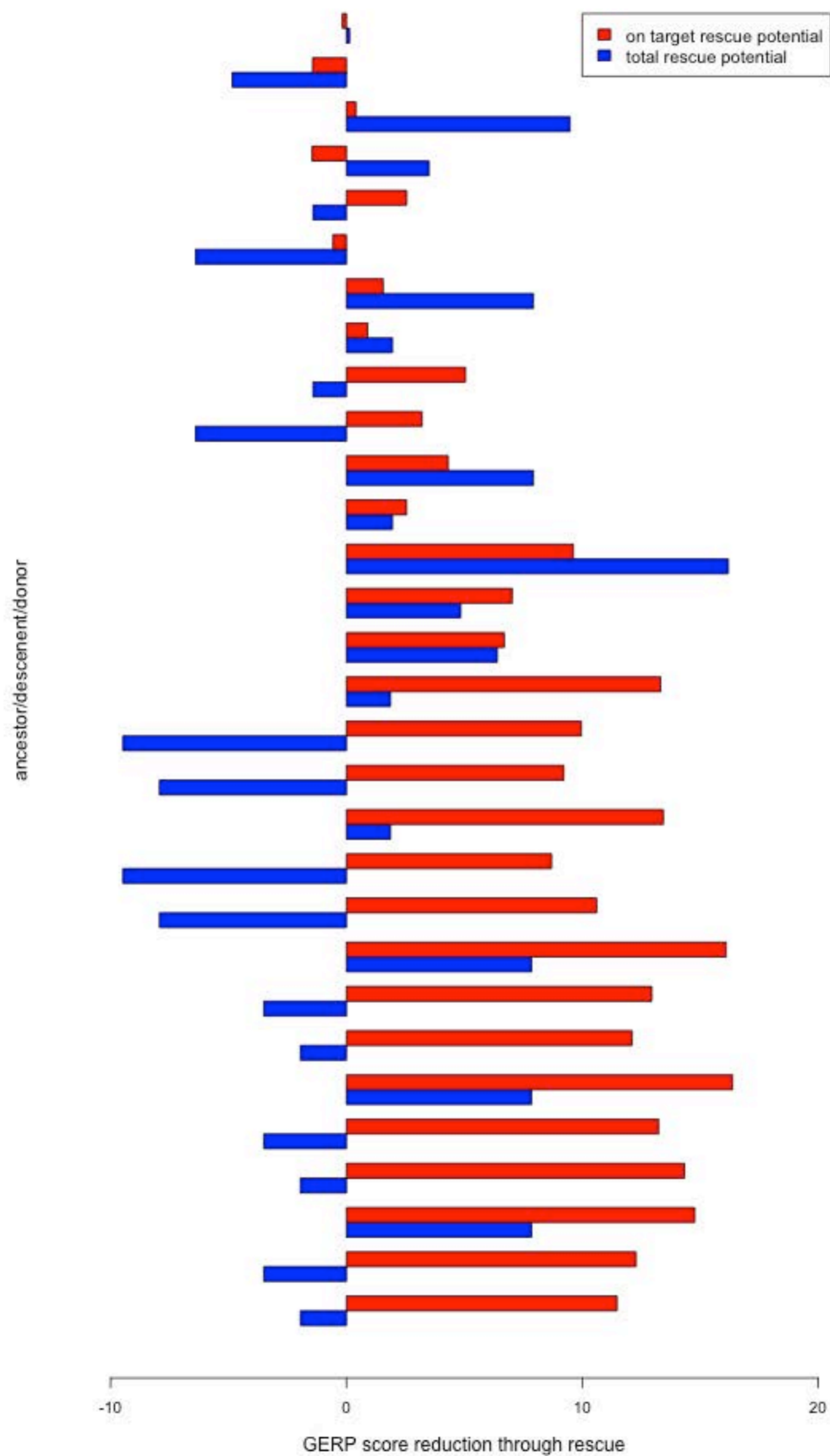




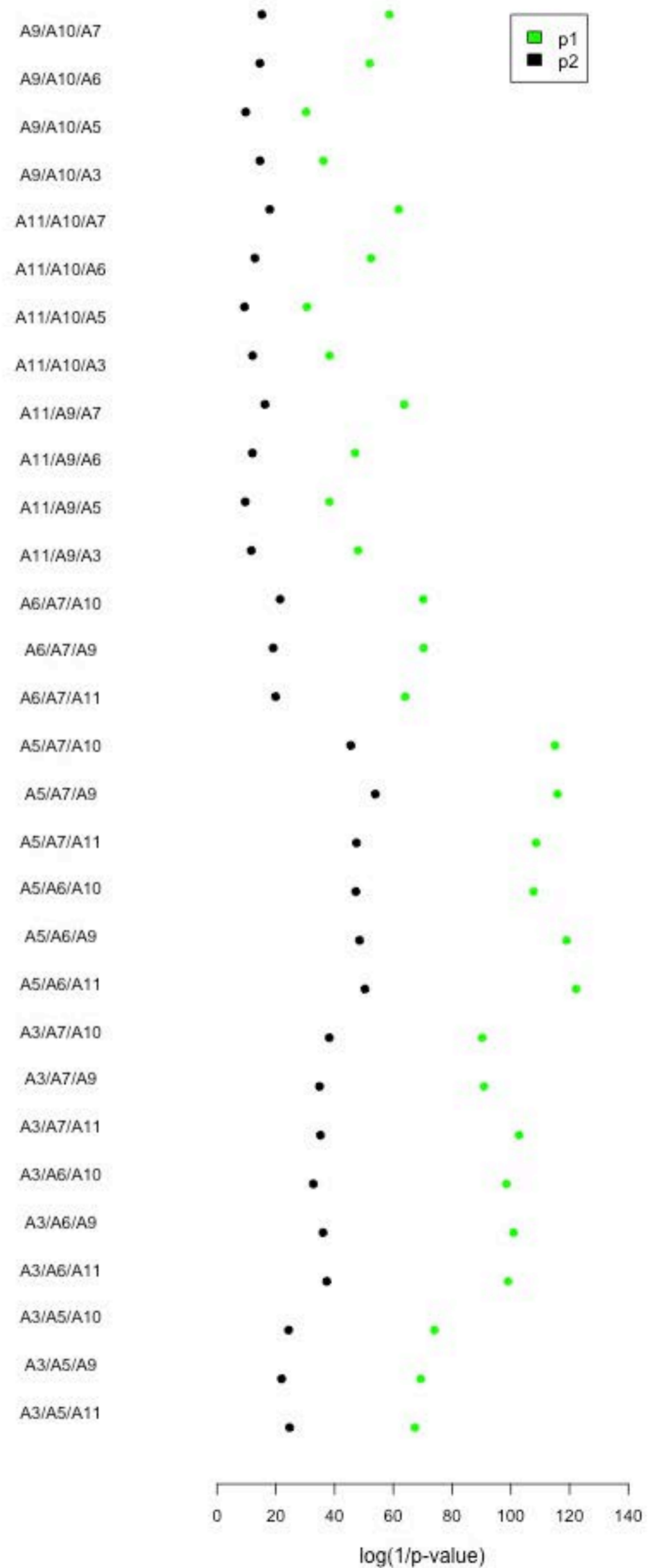
PCA via singular value decomposition of 1046 individuals using 1894 markers.



(A)



(B)



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

Supplementary Figures and Tables

Table S1. Sample details of genomes sequenced in this study

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

Table S3. Genome wide pairwise differences between ancient genomes

Table S4 Summary statistics of fit to linear and exponential models of change in heterozygosity and GERP score in *S. bicolor bicolor*.

Table S5. Changes in deleterious sites and recessive model sites per block between wild and cultivated bicolor.

Table S6. *p* values for windows containing significant reduction in heterozygosity relative to *S. verticilliflorum*.

Table S7. 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

Table S8 *p* values for reduction in heterozygosity in windows containing domestication loci observed in archaeological accessions relative to *S. verticilliflorum*. *P* values generated from an exponential distribution based on 739, 780, 745, 681, 624, 623,644, 555, 597 and 610 windows for chromosomes 1-10 respectively (one-tailed test), see methods.

Table S9 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 S10 *P* values of gradient deviation in heterozygosity over time relative to genomic average. *P* values calculated from a non-parametric distribution based on gradients of heterozygosity change over time from 739, 780, 745, 681, 624, 623,644, 555, 597 and 610 window pairs between dated samples for chromosomes 1-10 respectively, see methods.

Table S11 Genomic locations for 100kbp windows that differ in *gerp* load between genomes by more than two standard deviations.

Table S12 Phylogenetic congruence between type bicolor and type durra clades in selection candidate regions.

Table S13 Observed and expected genomic proportions of overlap in descendent and donor GERP scores relative to ancestor. ¶ p1 below ancestor score, § above ancestor score.

Table S14 Linear regression parameters for lnN vs gradient of heterozygosity loss in simulations ¶N value for gradient 4.03614E-07

Table S15 Regression fit values for simulations using N values from Table S14

Figure S1 Genomic heterozygosity against population nucleotide diversity (π).

Figure S2 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 S3 Frequency distributions of heterozygosity in genomes for 100 Kbp windows in wild sorghum races (*aethiopicum*, *arundinaceum*, *verticilliflorum*, *virgatum*), agroclimatic cultivars (*caudatum*, *kafir*, *guinea*) and related species (*halapense*, *propinquum*).

Figure S4 Heterozygosity over time in *S bicolor* type durra (N=4).

Figure S5 PSMC profiles of ancient sorghum genomes.

Figure S6 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 S7 θ dominant model (deleterious sites total) sites against θ recessive model (homozygous deleterious sites) between wild sorghum and bicolor types. A. wild/sample A3. B wild/sample A5. C. wild/sample A6. D wild/sample A7. E. wild/ BTX623.

Figure S8 Methylated site number in *S bicolor* type bicolor and *S bicolor* type durra over time.

Figure S9 Pairwise nucleotide difference heatmap between samples and BTX623 at known domestication loci.

Figure S10 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 S11 Heterozygosity over time of regions containing high SweeD scores. Significant deviations from the genomic gradient of change over time shown only.

Figure S12 Heterozygosity over time of regions containing domestication loci that have significantly reduced in heterozygosity relative to wild. Significant deviations from the genomic gradient of change over time shown only.

Figure S13 Summary of selection signals over time in archaeogenomes. Red indicates selection intensification episodes, green indicates selection signals identified by low heterozygosity or SweeD analysis.

Figure S14. Maximum likelihood tree of whole genome sequence built in EXaML

Figure S15. D statistic analysis: A. $P_1 = S. bicolor$ type bicolor BTX623, $P_2 =$ sample displayed on X axis, $P_3 =$ sample A3, P_4 halapense. B $P_1 =$ ancient durra sample A10, P_2 - P_4 as for A. C $P_1 =$ ancient durra sample A11, P_2 - P_4 as for A. D

statistics based $N = 6598$ 100 kbp windows across the genome. Box plots indicate one (box) and two (stick) standard errors respectively.

Figure S16 fd statistics. A P_1 = ancient durra sample A10, P_2 = sample displayed on X axis, P_3 = sample A3, P_4 halapense. B P_1 = ancient durra sample A11, P_2 - P_4 as for A. Statistics based $N = 6598$ 100 kbp windows across the genome. Box plots indicate one (box) and two (stick) standard errors respectively.

Figure S17. 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^t , 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 S18. Lost of heterozygosity through forward simulations of 5 models over varying population sizes. A. Model 1: 40-80% inbreeding switch at domestication, 1% bottleneck each generation. B Model 2: Gradual switch from 40-80% inbreeding over 6000 generations, 1% bottleneck each generation. C Model 3: Gradual switch from 40-80% inbreeding over 6000 generations. D Model 4: 40-80% inbreeding switch at domestication. E Model 5 40% inbreeding over 6000 generations. See methods for simulation details.

Supplementary video. Principal Coordinate Analysis of 1894 SNPs from 23 genomes in this study and 1046 sorghum lines described in Thurber et al (25), 3 d movie file. Genomes from this study (red), durra (blue), caudatum (yellow), guinea (orange), kafir (green).

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

	Archaic			Bicolor				Durra			Intermediate	
	A3	A5	A6	A7	Kew1	Kew2	A9	A10	A11	A8	A12	
Archaic	A3											
	A5	0.1795%										
	A6	0.1764%	0.1685%									
Bicolor	A7	0.1526%	0.1460%	0.1334%								
	Kew1	0.1258%	0.1126%	0.1132%	0.0874%							
	Kew2	0.1214%	0.1092%	0.1112%	0.0841%	0.0419%						
Durra	A9	0.1350%	0.1231%	0.1211%	0.0948%	0.0579%						
	A10	0.1278%	0.1160%	0.1147%	0.0895%	0.0505%	0.0542%					
	A11	0.1342%	0.1220%	0.1193%	0.0948%	0.0567%	0.0622%	0.0561%				
Intermed	A8	0.1488%	0.1407%	0.1375%	0.1158%	0.0856%	0.0946%	0.0874%	0.0932%			
	A12	0.1208%	0.1094%	0.1081%	0.0816%	0.0430%	0.0485%	0.0425%	0.0472%	0.0797%		

Table S3 Pairwise differences between sample genomes

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 S4 - Summary statistics of fit to linear and exponential models of change in heterozygosity and GERP score in *S. bicolor bicolor* (N=7).

ancestor/descendent	blocks increased in deleterious sites	reduced deleterious site blocks increased in recessive model sites	blocks reduced in deleterious sites and increased in recessive model sites	blocks increased in deleterious sites and increased in recessive model sites
wild/a5	0.692319369	0.47804637	0.14708561	0.15528233
wild/a3	0.514420158	0.43857455	0.21296296	0.26958106
wild/a6	0.547814208	0.47264183	0.21372192	0.23527626
wild/a7	0.493472981	0.46239137	0.23421372	0.26958106
wild/BTX623	0.57483303	0.6836844	0.29068002	0.11900425

Table S5 Changes in deleterious sites and recessive model sites per block between wild and cultivated bicolor

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 S6 *p* values for windows containing significant reduction in heterozygosity relative to *S. verticilliflorum*. *P* values generated from an exponential distribution based on 739, 780, 745, 681, 624, 623,644, 555, 597 and 610 windows for chromosomes 1-10 respectively (one-tailed test), see methods .

chromosome	window	peak	start	stop	Uniparc code	Uniparc code	Sb code	SORBI code	Gene description	
1	44900000	pk1	44830820 44860037	44831431 44861635	UPI0001A82246 UPI0001A82247	UPI0001A82246 UPI0001A82247	Sb01g026525 Sb01g026530	C5WP69_SORBI C5WP70_SORBI	unknown function (DUF1645) cytochrome p450	
2	11000000	pk2	10974343 10981018	10978599 10981874	UPI0001A83EC8 UPI0001A83EC9	UPI0001A83EC8 UPI0001A83EC9	Sb02g008271 Sb02g008311	C5X2V0_SORBI C5X2V1_SORBI	reverse transcriptase transposase	
2	43900000	pk3	43786194 43788873	43787234 43789595	UPI0001A842F4 UPI0001A842F5	UPI0001A842F4 UPI0001A842F5	Sb02g018043 Sb02g018110	C5X843_SORBI C5X844_SORBI	chromosome segregation ATPase transposase	
2	53700000	pk4	53484601 53668339 53671197 53687315 53693445	53494615 53669764 53673521 53688481 53694484	*UPI0001A83D56 UPI0001A83D57 UPI0001A838F5 UPI0001A838F6 UPI0001A838F7	UPI0001A83D56 UPI0001A83D57 UPI0001A838F5 UPI0001A838F6 UPI0001A838F7	Sb02g021535 Sb02g021540 Sb02g021550 Sb02g021560 Sb02g021570	C5X9S1_SORBI C5X9S2_SORBI C5X9S3_SORBI C5X9S4_SORBI C5X9S5_SORBI	RGA3 disease resistance protein (NB-ARC, LRR domain) transposase RGA3 disease resistance protein (NB-ARC domain) RGA3 disease resistance protein (LRR domain) RGA3 disease resistance protein (PKc domain)	
2	54200000	pk5	54156595 54169579 54173272 54179475	54156834 54170586 54173781 54180150	UPI0001A83D5D UPI0001A83D5E UPI0001A83D5F UPI0001A83D60	UPI0001A83D5D UPI0001A83D5E UPI0001A83D5F UPI0001A83D60	Sb02g021853 Sb02g021856 Sb02g021860 Sb02g021911	C5X9T6_SORBI C5X9T7_SORBI C5X9T8_SORBI C5X9T9_SORBI	unknown Polynucleotidyl transferase ribonuclease unknown Anther Indehiscence 1	
3	32700000	pk6	32649562..32651401	32649562..32651401	LOC110433684				trichohyalin-like	
4	18100000	pk7	17966716..17967603 17982112..17982948 18046591..18047799	17966716..17967603 17982112..17982948 18046591..18047799	LOC8155713 LOC110434760 LOC110434315				vegetative cell wall protein gp1 serine/arginine repetitive matrix protein 1-like 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 27797941	27795021 27798643	UPI0001A865C1 UPI0001A865C2	UPI0001A865C1 UPI0001A865C2	Sb05g013400 Sb05g013410	C5Y282_SORBI C5Y283_SORBI	GDT1 like protein Rho binding protein	Ca transporter Regulatory transcription inhibitor
5	43300000	pk13	43,247,187	43,269,799	LOC8075771				O-Fuct like	auxin independent growth promoter
5	59500000	pk14	59411298 59419666 59437739 59452563 59458058 59463266 59464960 59474035 59475347	59414857 59417283 59438068 59457270 59462321 59463718 59466737 59475117 59477611	UPI0001A8660D UPI0001A8660E UPI0001A8660F UPI0001A86610 UPI0001A86611 UPI0001A86639 UPI0001A86C54 UPI0001A86C55 UPI0001A86C56	UPI0001A8660D UPI0001A8660E UPI0001A8660F UPI0001A86610 UPI0001A86611 UPI0001A86639 UPI0001A86C54 UPI0001A86C55 UPI0001A86C56	Sb05g025890 Sb05g025900 Sb05g025910 Sb05g025915 Sb05g025920 Sb05g025930 Sb05g025940 Sb05g025945 Sb05g025950	C5Y7E1_SORBI C5Y7E2_SORBI C5Y7E3_SORBI C5Y7E4_SORBI C5Y7E5_SORBI C5Y7E6_SORBI C5Y7E7_SORBI C5Y7E8_SORBI C5Y7E9_SORBI	lipase glutaredoxin C10 glutaredoxin C10 galactose oxidase peptide chain release factor APG3 RALF like protein alkB pollen extensin like pollen extensin like	arrests root development DNA repair
5	61100000	pk15	61009529 61015798 61028126 61048282 61057619 61061641 61082967 61097301	61010503 61018316 61029756 61050689 61058128 61062092 61094282 61098179	UPI0001A86440 UPI0001A86441 UPI0001A8643F UPI0001A86A3A UPI0001A86A3B UPI0001A86A3C UPI0001A86442 UPI0001A86443	UPI0001A86440 UPI0001A86441 UPI0001A8643F UPI0001A86A3A UPI0001A86A3B UPI0001A86A3C UPI0001A86442 UPI0001A86443	Sb05g026965 Sb05g026970 Sb05g026950 Sb05g026990 Sb05g026993 Sb05g026996 Sb05g027000 Sb05g027005	C5Y826_SORBI C5Y827_SORBI C5Y825_SORBI C5Y829_SORBI C5Y830_SORBI C5Y831_SORBI C5Y832_SORBI C5Y833_SORBI	RPP13 disease resistance protein NBS-LRR RPP13 disease resistance protein NBS-LRR dirigent protein patatin isopentenyl transferase patatin RPP13 disease resistance protein NBS-LRR transposable element	disease response involg lignification storage protein and fatty acid metabolism
6	22600000	pk16	22725443..22735214	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 50772772 50794332 50896473	50568895 50778650 50794661 50902950	UPI0001A87F06 UPI0001A87F07 UPI0001A87F08 *UPI0001A87F09	UPI0001A87F06 UPI0001A87F07 UPI0001A87F08 UPI0001A87F09	Sb07g019540 Sb07g019740 Sb07g019745 Sb07g019750	C5YKN9_SORBI C5YKP0_SORBI C5YKP1_SORBI C5YKP2_SORBI	ABC transporter ABC transporter transposable element ABC transporter	
7	63100000	pk19	63002738 63007131 63019028 63027930 63029524 63059608 63062573 63073408 63088404	63004800 63012078 63020726 63028675 63033671 63062274 63062995 63073999 63095390	UPI0001A87EBC UPI0001A87A47 UPI00022071D4;b UPI0001A87EBD UPI0001A87A48 UPI0001A87EBE UPI0001A87A49 UPI0001A87A4A UPI0001A87A4B	UPI0001A87EBC UPI0001A87A47 UPI00022071D4 UPI0001A87EBD UPI0001A87A48 UPI0001A87EBE UPI0001A87A49 UPI0001A87A4A UPI0001A87A4B	Sb07g028040 Sb07g028050 Sb07g028060 Sb07g028065 Sb07g028070 Sb07g028080 Sb07g028090 Sb07g028095 Sb07g028100	C5YJ45_SORBI C5YJ46_SORBI C5YJ47_SORBI C5YJ48_SORBI C5YJ49_SORBI C5YJ50_SORBI C5YJ51_SORBI C5YJ52_SORBI	MFS, putative peptide transporter glycerophosphodiester phosphodiesterase GDPDL4 fibrous sheath CABYR-binding protein unknown function SWIB domain protein (p53 associated) serine-glyoxylate aminotransferase PEF family (apoptosis associated) Peptidase M14 Succinylglutamate desuccinylase wall-associated receptor kinase 5	
8	12300000	pk20	12228641 12259170 12279014 12293716 12351604 12367068 12372086 12385186 12388327	12230086 12259505 12286998 12301627 12355600 12370954 12372376 12386774 12390954	UPI0001A881AC UPI0001A881AD UPI0001A881AE UPI0001A880E7 UPI0001A880E8 UPI0001A880E9 UPI0001A880EA UPI0001A881AF UPI0001A880EB	UPI0001A881AC UPI0001A881AD UPI0001A881AE UPI0001A880E7 UPI0001A880E8 UPI0001A880E9 UPI0001A880EA UPI0001A881AF UPI0001A880EB	Sb08g007200 Sb08g007210 Sb08g007220 Sb08g007230 Sb08g007240 Sb08g007243 Sb08g007246 Sb08g007250 Sb08g007260	C5YTS4_SORBI C5YTS5_SORBI C5YTS6_SORBI C5YTS7_SORBI C5YTS8_SORBI C5YTS9_SORBI C5YTT0_SORBI C5YTT1_SORBI C5YTT2_SORBI	TNP1 like protein TNP2 like protein la-related 6B protein outer envelope pore protein 37 chloroplastic 2 alkenal reductase 2 alkenal reductase unknown obtusifolliol 14 alpha demethylase transposase (transposon)	defense

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	C5YNU3_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	52900000 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 S7 Regions of genome-wide significance in reduction of heterozygosity relative to *S. verticilliflorum* and associated genes
 * indicates correspondence 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	38000000	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	vrsl	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_Ch2	0.73866598	0.67895781	0.66040409	0.570248694	0.348520009	0.502490469	0.536244356

Table S8 *p* values for reduction in heterozygosity in windows containing domestication loci observed in archaeological accessions relative to *S. verticilliflorum*. *P* values generated from an exponential distribution based on 739, 780, 745, 681, 624, 623,644, 555, 597 and 610 windows for chromosomes 1-10 respectively (one-tailed test), see methods.

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	C5WU5_SORBI	unknown function				
				67532193	67537252	UPI0001A8295D	UPI0001A8295D	Sb01g044430	C5WU6_SORBI	Pumilio RNA binding protein				
				67538460	67540436	UPI0001C80BA9;Or	UPI0001C80BA9	Sb01g044440		ras-related protein RABH1b-like	golgi trafficking			
				67555746	67557224	UPI0001A8295E	UPI0001A8295E	Sb01g044450	C5WU7_SORBI	WW domain protein				
				67560187	67564611	UPI0001A8295F	UPI0001A8295F	Sb01g044460	C5WU8_SORBI	conserved oligomeric Golgi complex subunit 8				
				67568809	67573214	UPI0001C80BAA;Or	UPI0001C80BAA	Sb01g044470		mitochondrial adenine nucleotide transporter BTL3				
				67576601	67583786	UPI0001A82960	UPI0001A82960	Sb01g044480	C5WU9_SORBI	calmodulin-binding transcription activator 1 isoform X2				
				67585849	67586537	UPI0001A82E0B	UPI0001A82E0B	Sb01g044485	C5WU0_SORBI	reverse transcriptase				
				67590805	67594920	UPI0001A82961	UPI0001A82961	Sb01g044490	C5WU1_SORBI	unknown function				
				67606065	67606893	UPI0001A82962	UPI0001A82962	Sb01g044500	C5WU2_SORBI	trypsin like peptidase				
				67612260	67612981	UPI0001A82E0C	UPI0001A82E0C	Sb01g044505	C5WU3_SORBI	unknown function				
				67623006	67629132	UPI0001A82963	UPI0001A82963	Sb01g044510	C5WU4_SORBI	unknown function				
				67631156	67632142	UPI0001A82964	UPI0001A82964	Sb01g044515	C5WU5_SORBI	CCH domain zinc finger proetin				
				67653779	67664222	UPI0001A82965	UPI0001A82965	Sb01g044520	C5WU6_SORBI	trypsin like peptidase				
				67666492	67670699	UPI0001A829C6	UPI0001A829C6	Sb01g044530	C5WU7_SORBI	ubiquitin carboxyl-terminal hydrolase 3				
				67681955	67685954	UPI0001A829C7	UPI0001A829C7	Sb01g044540	C5WU8_SORBI	LRR domain protein				
				67687809	67690043	UPI0001A829C8	UPI0001A829C8	Sb01g044550	C5WU9_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	C5X956_SORBI	nudix hydrolase 20, chloroplastic
								53917736	53923682	UPI0001A83D59	UPI0001A83D59	Sb02g021780	C5X957_SORBI	UDP-glucose 4-epimerase
								53927463	53932982	UPI0001A83D5A	UPI0001A83D5A	Sb02g021790	C5X958_SORBI	E3 ubiquitin-protein ligase
								53972111	53976091	UPI0001C80E15;Or	UPI0001C80E15	Sb02g021800		receptor-like serine/threonine-protein kinase SD1-8
								53991898	53993342	UPI0001A838F8	UPI0001A838F8	Sb02g021810	C5X959_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	C5X1Y4_SORBI	transposase
								19147217	19147932	UPI0001A851BF	UPI0001A851BF	Sb03g014261	C5X1Y5_SORBI	transposase
				19148124	19148417	UPI0001A851CO	UPI0001A851CO	Sb03g014301	C5X1Y6_SORBI	transposase				
				19151124	19159094	UPI0001A851C1	UPI0001A851C1	Sb03g014340	C5X1Y7_SORBI	thiamine pyrophosphokinase 2				
				19161565	19162878	UPI0001A851C2	UPI0001A851C2	Sb03g014350	C5X1Y8_SORBI	pollen-specific leucine-rich repeat extensin-like protein 1				
				19166935	19167843	UPI0001A851C3	UPI0001A851C3	Sb03g014360	C5X1Y9_SORBI	rapid alkalization factor				
				19173459	19178759	UPI0001C808B1;Or	UPI0001C808B1	Sb03g01437		protein LOW PSII ACCUMULATION 1				
				19180099	19183534	UPI0001A845B0	UPI0001A845B0	Sb03g014380	C5X1Z0_SORBI	40S ribosomal protein S4				
				19185886	19193629	UPI0001A851C4	UPI0001A851C4	Sb03g014390	C5X1Z1_SORBI	NB-LRR disease resistance protein				
				3	36549186	197.3805	s4	36474476..36475344	LOC110433364					ncRNA
								36469694..36470469	LOC110433369					
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	AS4Y09_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	exopolysaccharonase				
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				

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.

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	6000000	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	9000000	0.145762712	M5	M2	
	48000000	0.181818182	A3	A10	
	51000000	0.195744681	M2	A7	
	52000000	0.212598425	A11	M2	
	77000000	0.117647059	A3	A10	
	98000000	0.146443515	A7	Kew1	
	106000000	0.375	A10	M6	
	118000000	0.123966942	A10	A3	
	124000000	0.202531646	M5	A10	
	125000000	0.1625	M5	A7	
	127000000	0.44015444	M5	M2	
	142000000	0.159509202	M6	A10	
	238000000	0.214511041	A3	A7	
	249000000	0.396694215	A3	M2	
	267000000	0.247619048	A10	M6	
	343000000	0.117647059	A5	A3	
	369000000	0.222222222	A3	M2	
	379000000	0.185661765	A5	A11	
	381000000	0.18522602	A11	A7	
	413000000	0.158878505	A10	A5	
	420000000	0.270967742	A10	A5	pk26
	429000000	0.17989418	A3	Kew1	
	443000000	0.191616766	A10	M2	
	459000000	0.140540541	M5	M2	
	471000000	0.236686391	A10	M2	
10	61000000	0.888888889	A3	A10	
	62000000	0.363636364	A10	A5	
	116000000	0.621848739	A3	M6	
	129000000	0.232142857	A9	M6	
	141000000	0.502617801	A11	M2	
	150000000	1.333333333	A10	A3	
	167000000	0.213675214	A6	A10	
	204000000	0.456410256	A3	A11	
	215000000	0.21978022	A5	A7	pk28
	246000000	0.259136213	A11	M6	
	267000000	0.238636364	A3	A5	
	367000000	0.583892617	A10	A5	
	391000000	0.264285714	A10	A5	
	491000000	0.410958904	M5	M2	
	495000000	0.242105263	A6	A3	

Table S11 Genomic locations for 100kb windows that differ in gerp load between genomes by more than two standard deviations.

gene	window	tree number	selection	phylogenetic incongruence	incongruent taxa 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	596000000	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	121000000	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	667000000	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	680000000	99	durra A11-A9	yes	bicolor A3,A5
ma6	363700000	534	durra A11-A9	yes	bicolor A5
pk1	449000000	66	durra A10	no	no
pk2	849000000	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 S12 Phylogenetic congruence between type bicolor and type durra clades in selection candidate regions.

Ancestor/ Descendent /Donor	<i>p1</i> ¶	<i>p1/e1</i>	<i>p</i>	<i>p2</i> §	<i>p2/e2</i>	<i>p</i>
A3/A5/A11	0.12155199	2.64469648	5.8763E-30	0.69051228	1.12450546	1.9038E-11
A3/A5/A9	0.11791452	2.69670569	7.644E-31	0.69793877	1.12051977	2.819E-10
A3/A5/A10	0.12412852	2.45985131	7.1375E-33	0.67080934	1.12492625	2.6914E-11
A3/A6/A11	0.14352834	3.07669044	9.5919E-44	0.70945741	1.15977176	6.106E-17
A3/A6/A9	0.13837527	3.11786341	1.5243E-44	0.71536829	1.15289013	2.1618E-16
A3/A6/A10	0.14337678	2.79929355	1.6192E-43	0.68702637	1.15652332	5.8762E-15
A3/A7/A11	0.14458927	2.92640078	2.3813E-45	0.69839345	1.15940288	5.1741E-16
A3/A7/A9	0.14110336	3.00184037	3.5021E-40	0.70597151	1.15540282	7.043E-16
A3/A7/A10	0.14959079	2.75756696	6.3427E-40	0.68111549	1.16436666	2.5628E-17
A5/A6/A11	0.17762958	2.83657946	8.1044E-54	0.67550773	1.20681418	1.3799E-22
A5/A6/A9	0.17384056	2.84517906	2.1861E-52	0.67793271	1.20109649	8.5402E-22
A5/A6/A10	0.1826311	2.66106267	1.633E-47	0.65595635	1.21195375	3.1351E-21
A5/A7/A11	0.17929676	2.73613744	6.6906E-48	0.66580782	1.20769885	2.5085E-21
A5/A7/A9	0.17869051	2.79476833	4.8515E-51	0.67141558	1.2077649	4.0287E-24
A5/A7/A10	0.18505608	2.57673373	1.1173E-50	0.64701425	1.2137371	1.6968E-20
A6/A7/A11	0.1039709	2.77924477	1.5305E-28	0.71551985	1.10414303	2.2181E-09
A6/A7/A9	0.10063656	2.80838489	2.8377E-31	0.72052137	1.10040285	5.1362E-09
A6/A7/A10	0.11170052	2.56548664	3.2299E-31	0.6908154	1.11103425	4.834E-10
A11/A9/A3	0.07850864	2.42081621	1.4843E-21	0.70960897	1.07098748	8.7416E-06
A11/A9/A5	0.06911185	2.43095658	2.5082E-17	0.72734162	1.06078734	6.8639E-05
A11/A9/A6	0.07820552	2.90136164	3.98E-21	0.74643831	1.07528873	5.9851E-06
A11/A9/A7	0.08320703	2.97587377	2.1999E-28	0.74461958	1.08171316	8.395E-08
A11/A10/A3	0.0957866	2.17385475	2.3179E-17	0.67399212	1.0846485	5.5997E-06
A11/A10/A5	0.07941801	2.0560091	5.1775E-14	0.68475296	1.06486048	8.953E-05
A11/A10/A6	0.0923007	2.52029188	1.7203E-23	0.70763868	1.08695518	2.5518E-06
A11/A10/A7	0.09821158	2.58522622	1.4773E-27	0.70672931	1.09471147	1.6377E-08
A9/A10/A3	0.09836314	2.19152175	1.9468E-16	0.67187026	1.08801131	4.5288E-07
A9/A10/A5	0.08047893	2.07791673	6.9254E-14	0.6853592	1.06638479	5.8694E-05
A9/A10/A6	0.09214914	2.48011376	2.6969E-23	0.70506214	1.08614926	4.835E-07
A9/A10/A7	0.09669597	2.4928044	3.3639E-26	0.70127311	1.09155912	2.3992E-07

Table S13 Observed and expected genomic proportion of overlaps in descendent and donor GERP scores

¶ *p1* below ancestor score

§ *p2* above ancestor score

Simulation			
model	a	b	Nη
model 1	-2.6958356	-0.9746723	227359
model 2	-2.6718894	-0.8928092	727778
model 3	-4.949024	-1.0099473	15955
model 4	-2.6185686	-1.4762804	3637
model 5	-4.1812295	-1.1742583	7920

Table S14 Linear regression parameters for $\ln N$ vs gradient of heterozygosity loss in simulations ηN value for gradient 4.03614E-07

simulation model	regression model	R2	AIC	MSE
model 1	linear	0.739	-66181.21	3.83E-09
model 2	linear	0.973	-67039.6	3.09E-09
model 3	linear	0.9004	-59531.76	2.02E-08
model 4	linear	0.6512	-66018.2	3.99E-09
model 5	linear	0.733	-58635.12	2.52E-08
model 1	exponential	0.7675	-4347.816	0.01972124
model 2	exponential	0.9631	-13775.5	0.00186893
model 3	exponential	0.9608	-9937.214	0.00487781
model 4	exponential	0.5859	3556.878	0.1422201
model 5	exponential	0.7648	-7385.665	0.00922968

Table S15 Regression fit values for simulations using N values from Table S14.

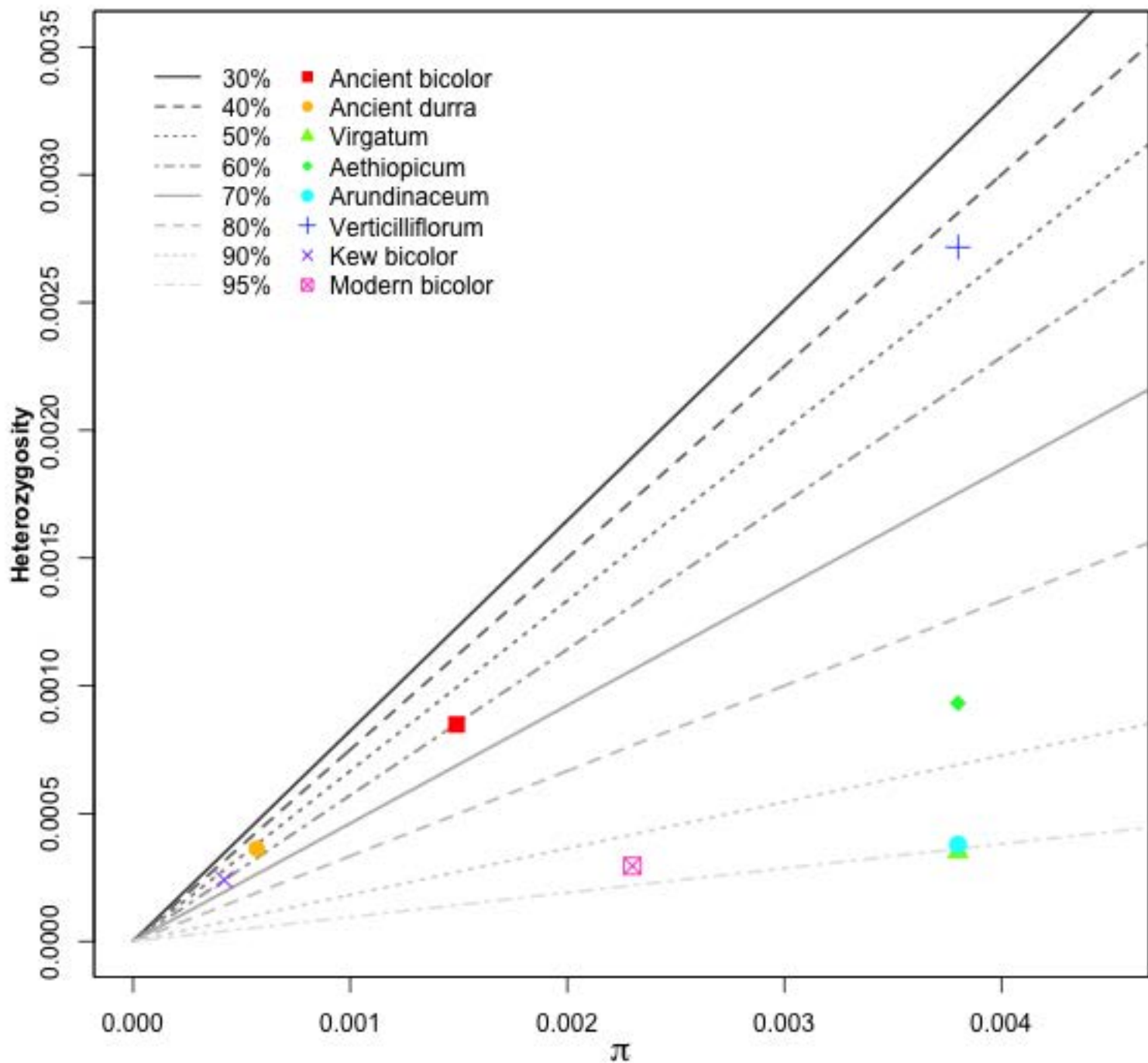
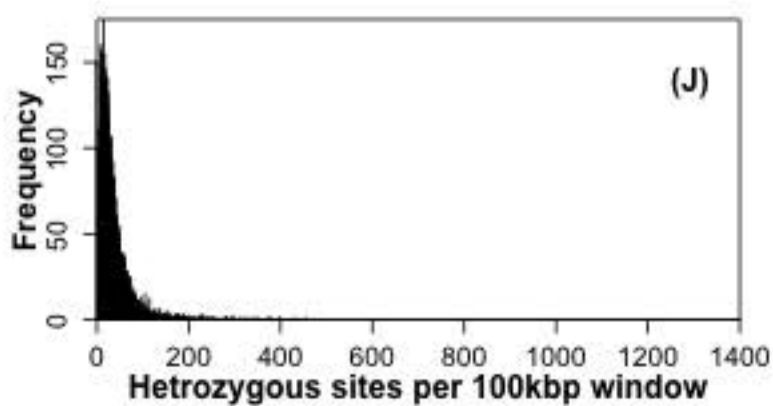
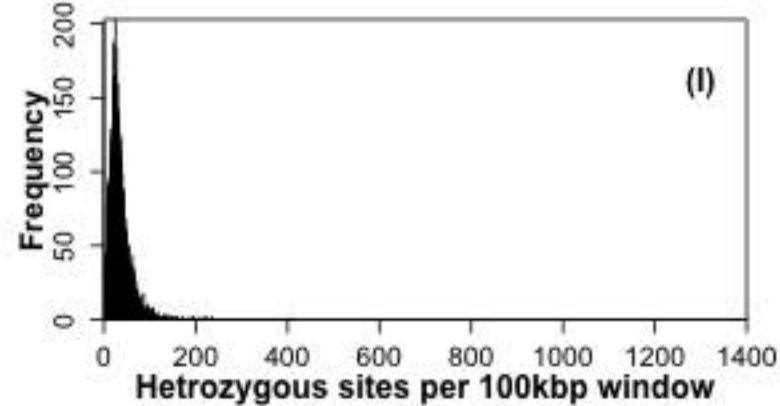
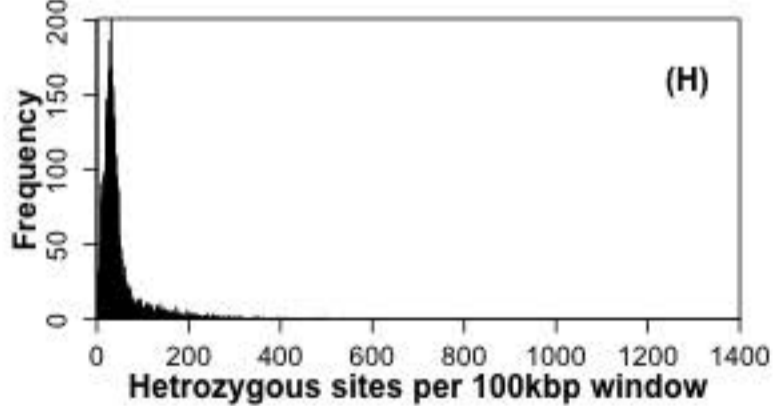
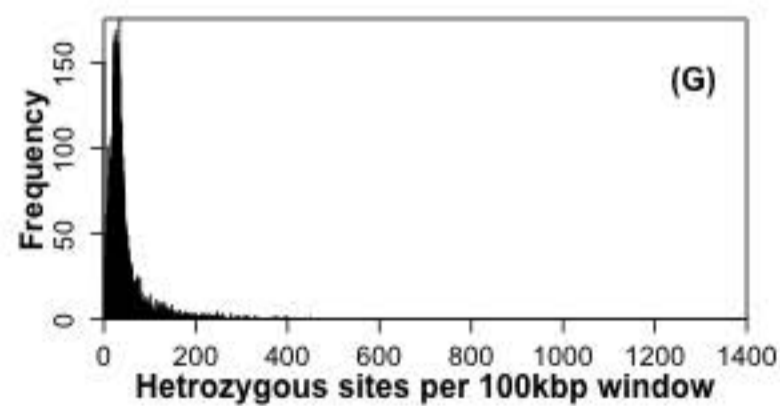
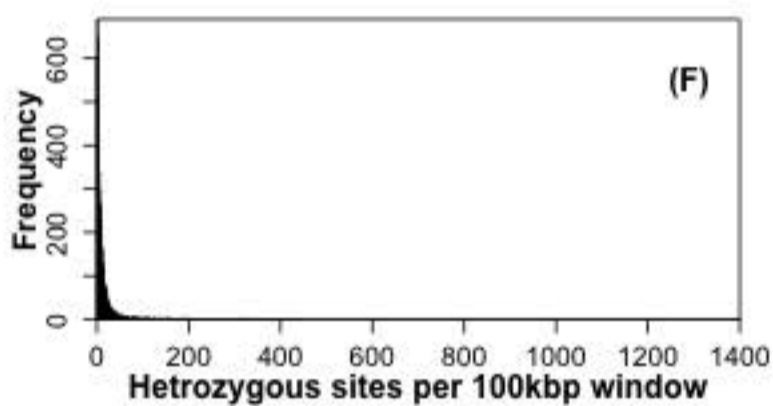
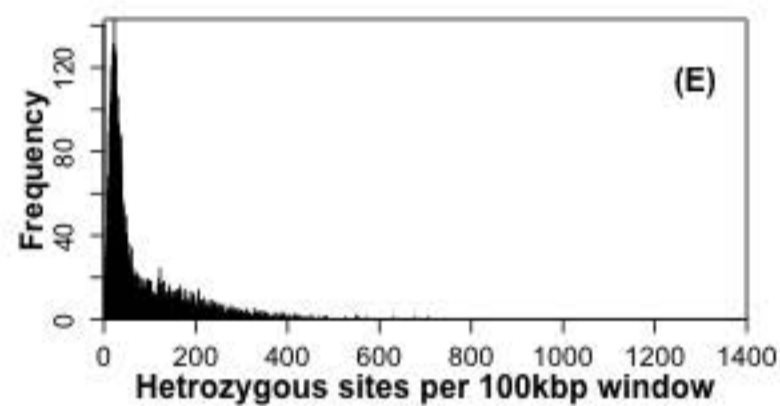
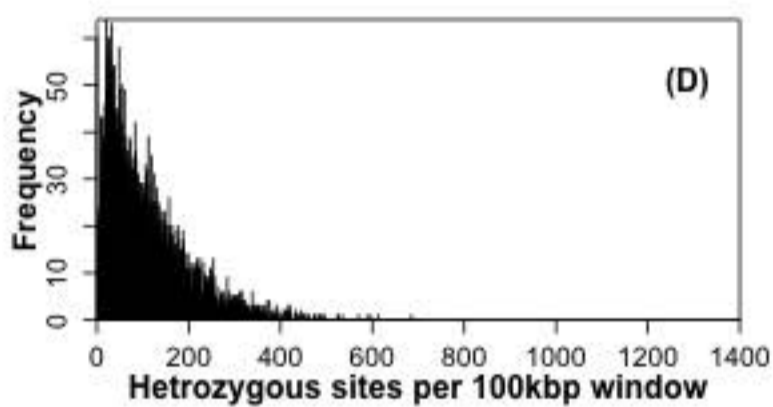
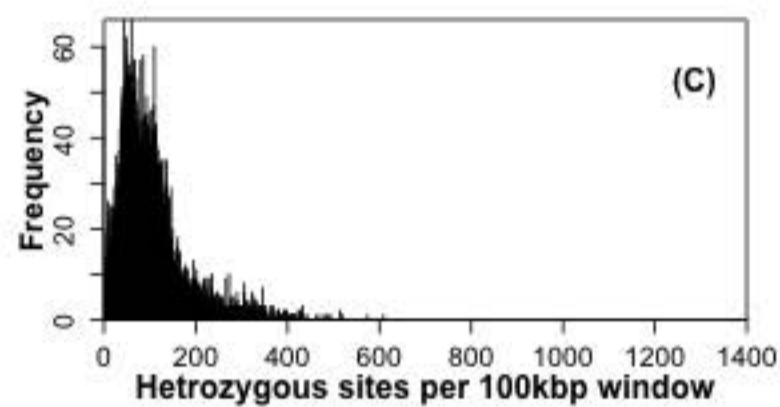
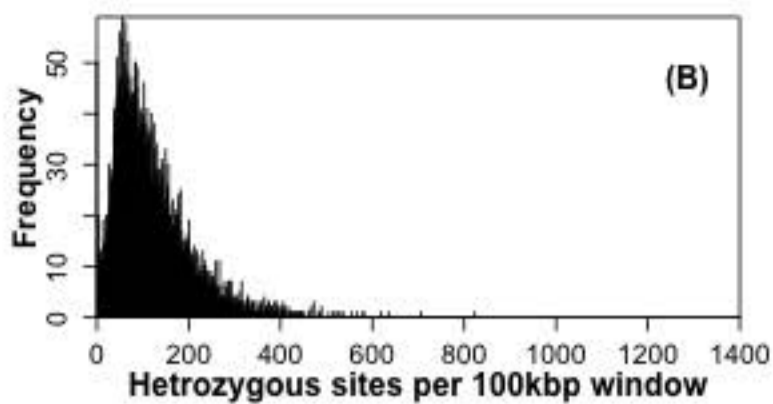
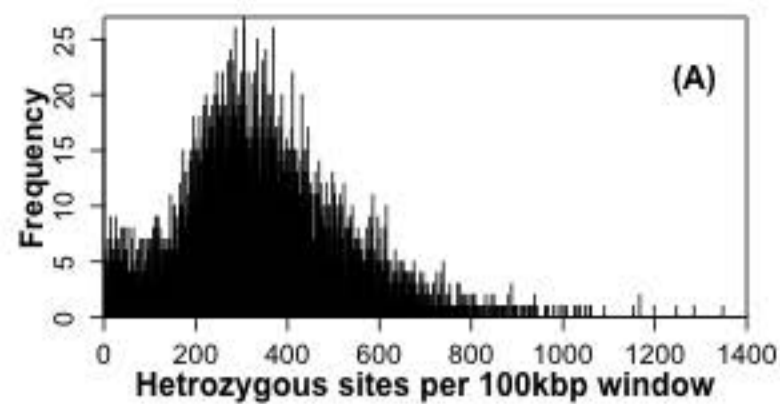


Fig S1

Fig S2



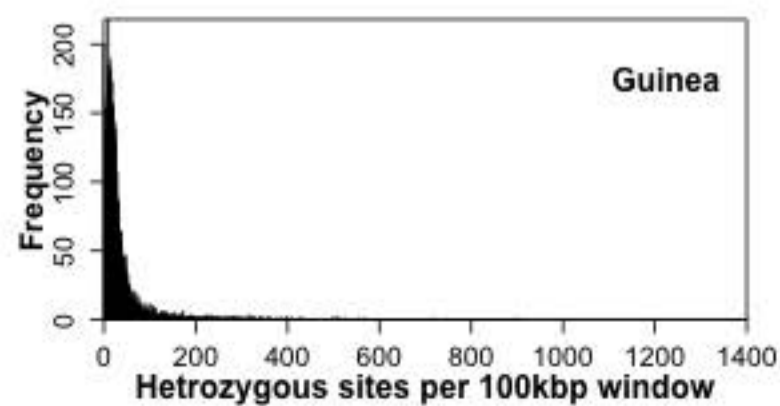
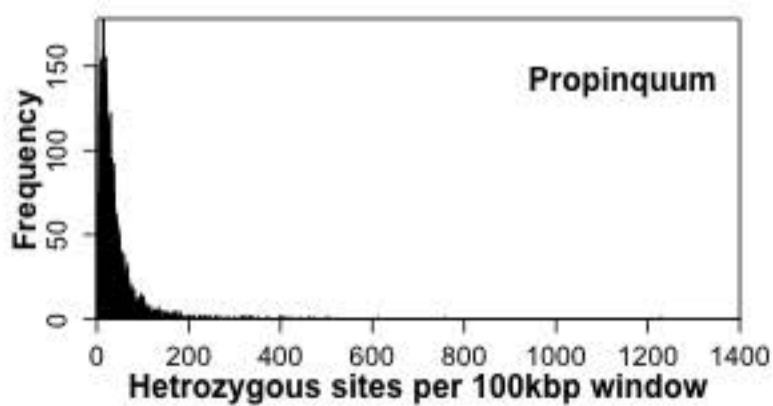
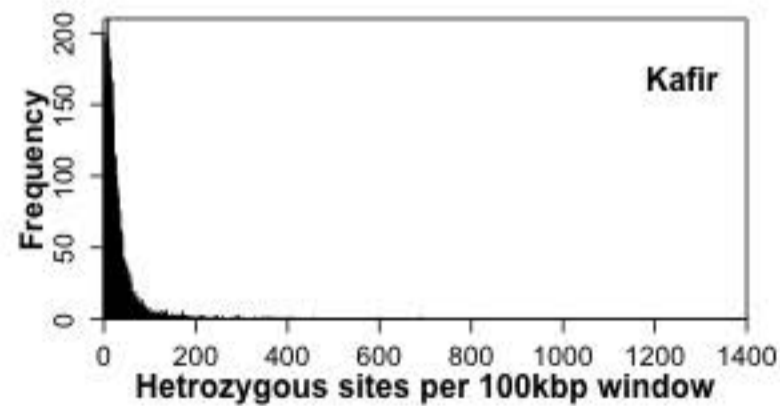
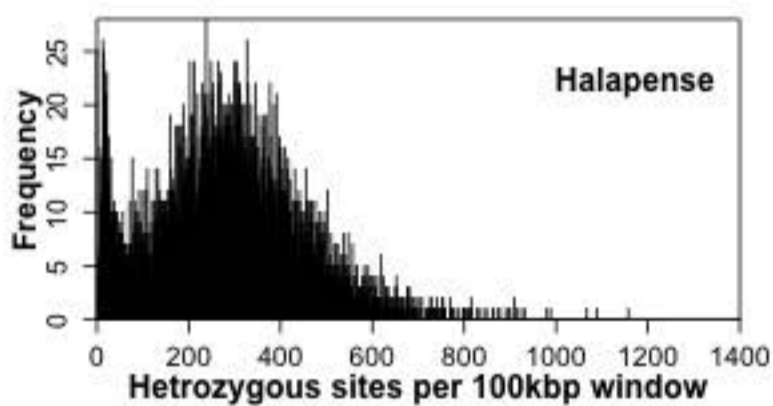
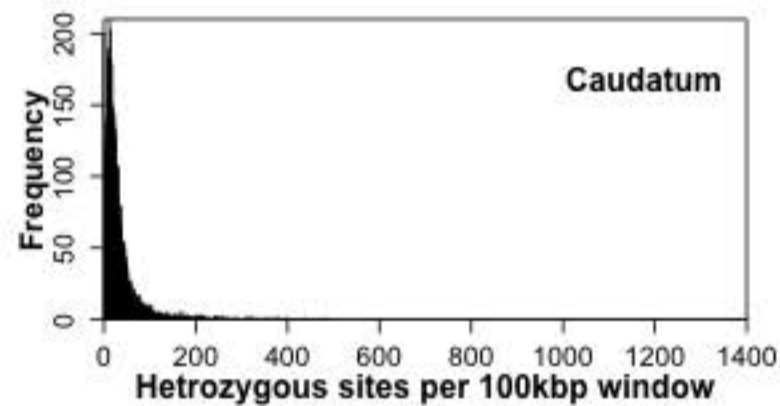
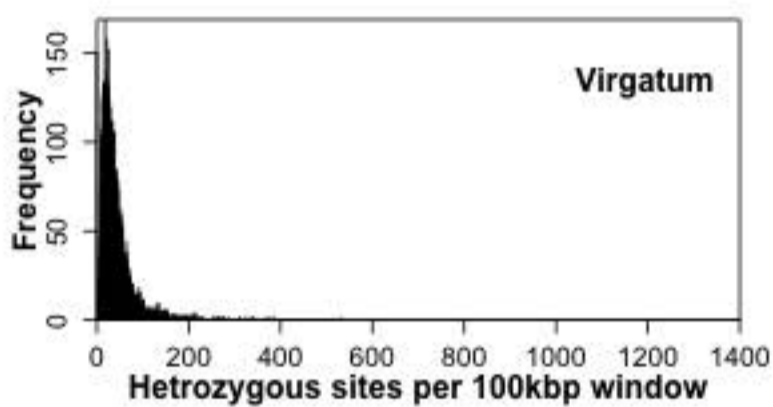
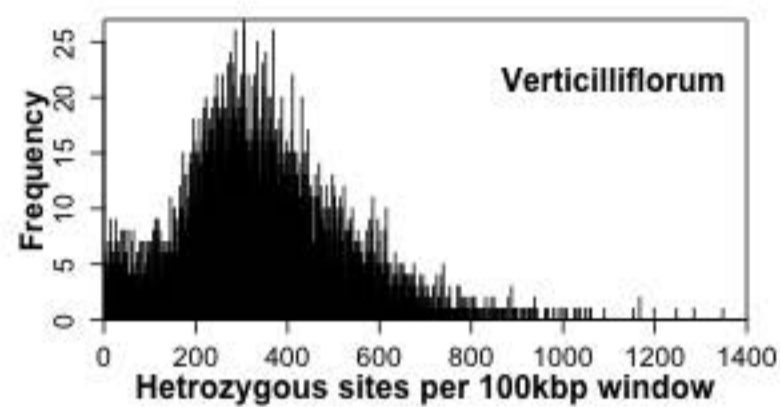
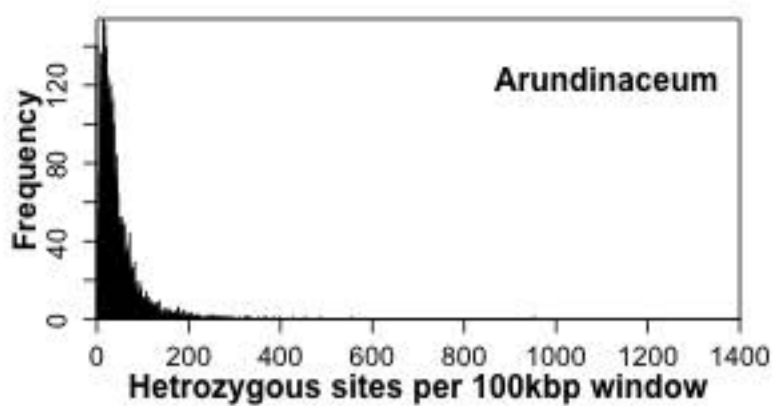
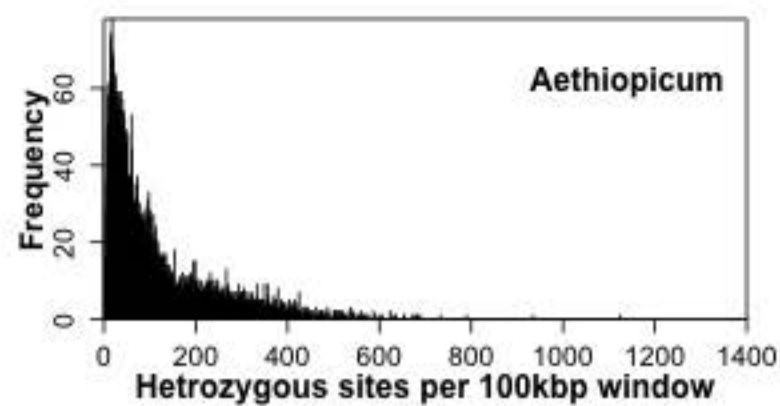


Fig S3

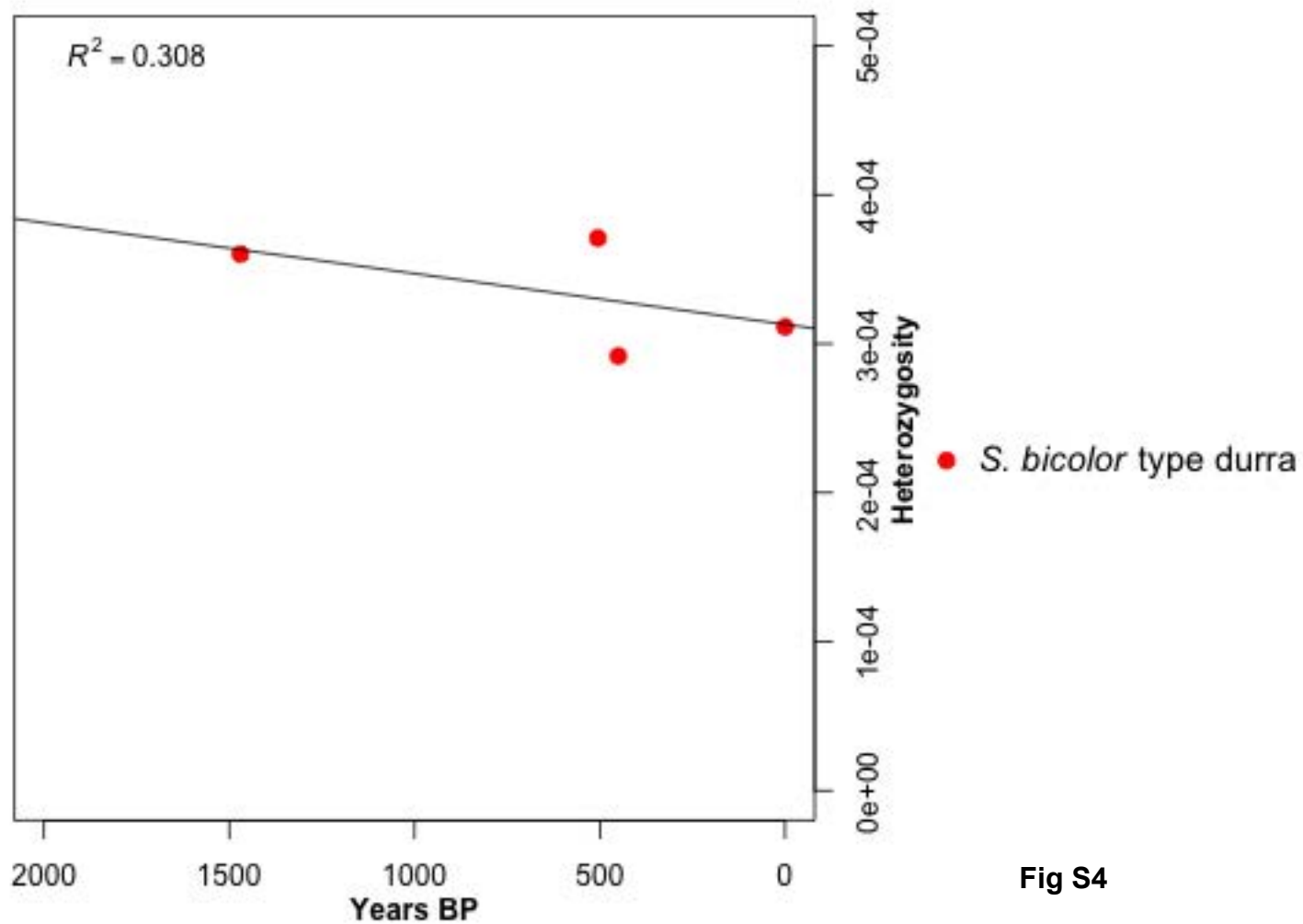


Fig S4

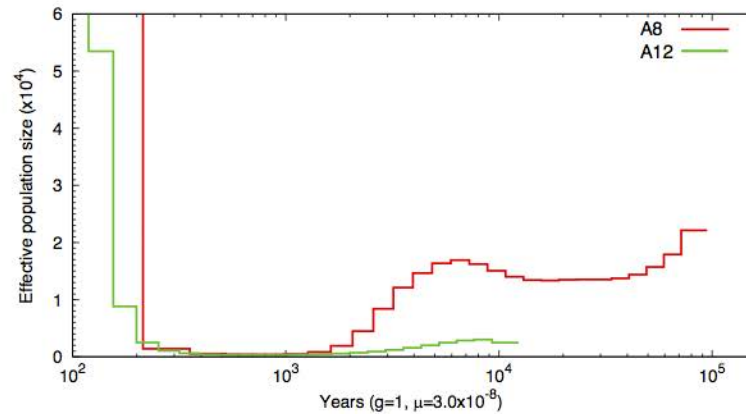
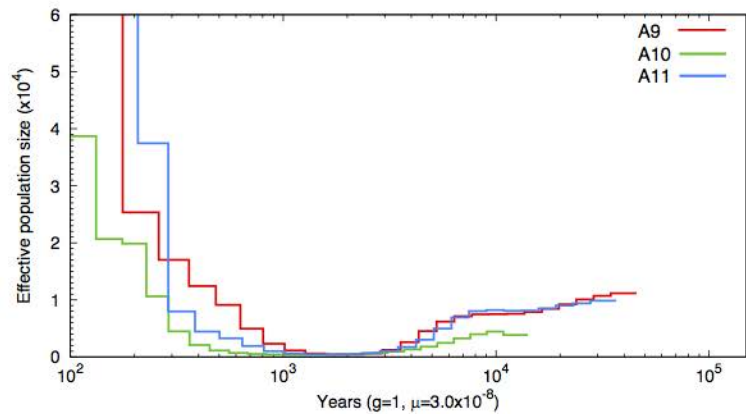
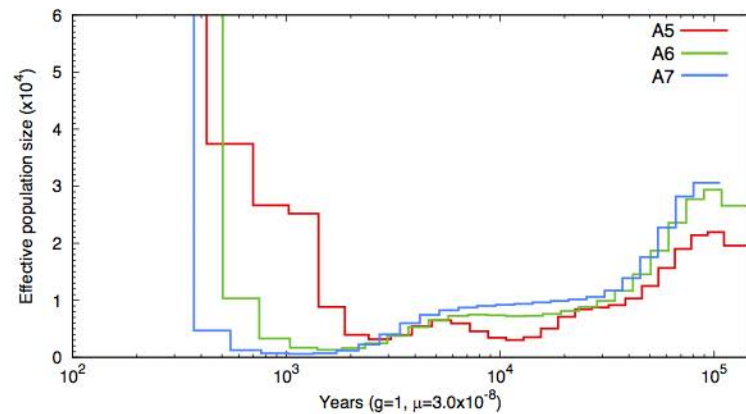
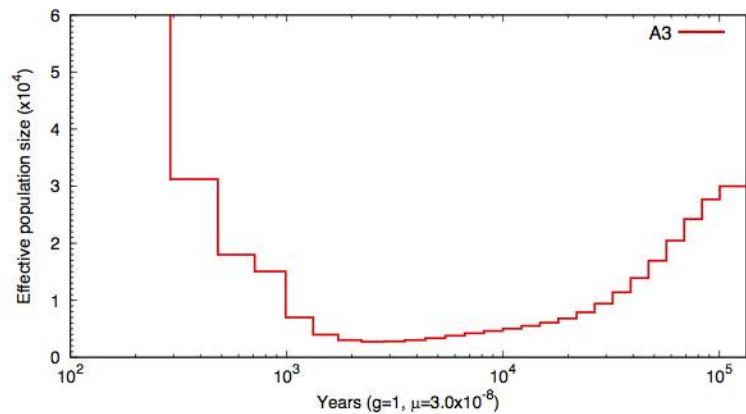


Fig S5

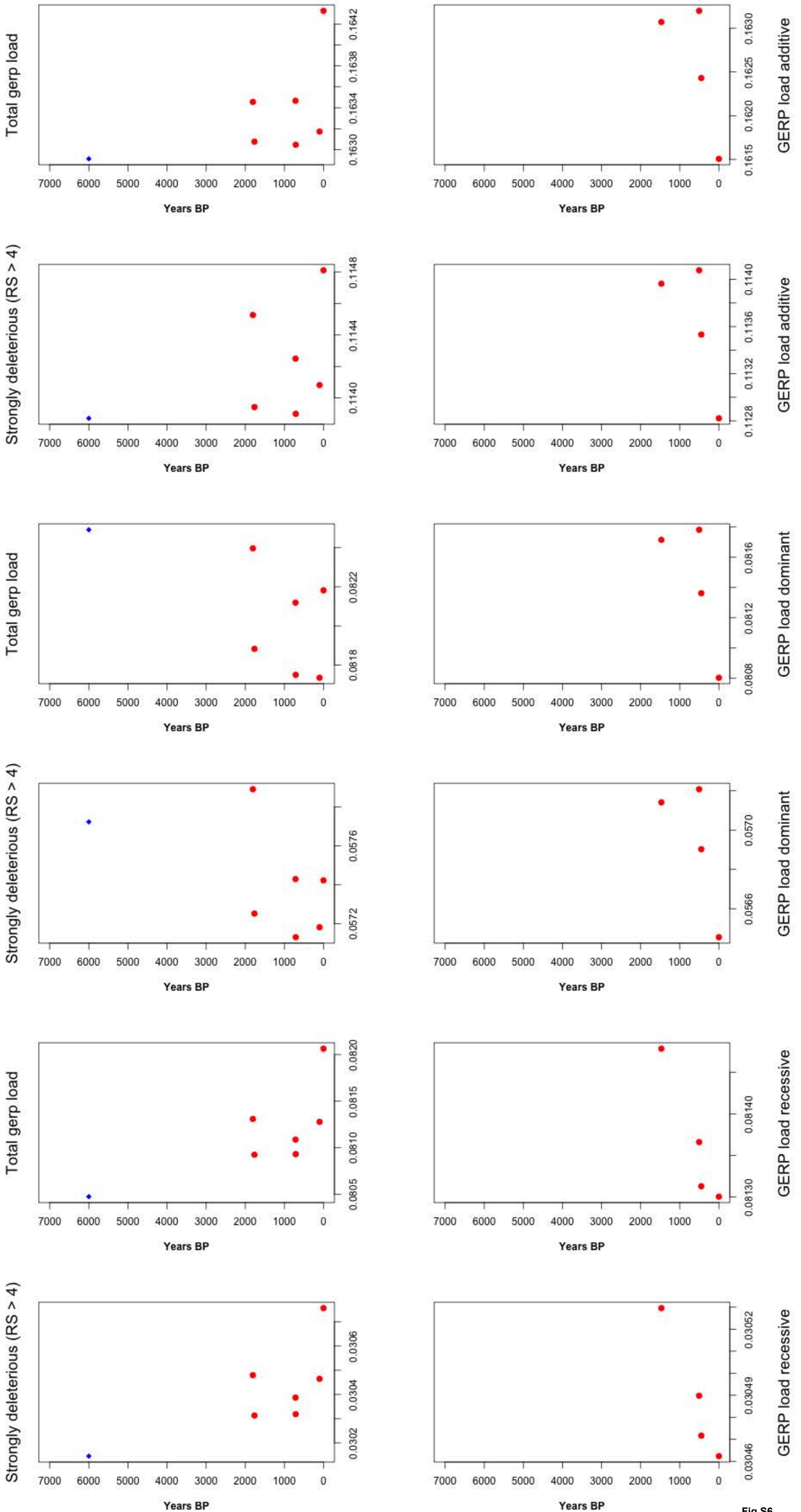
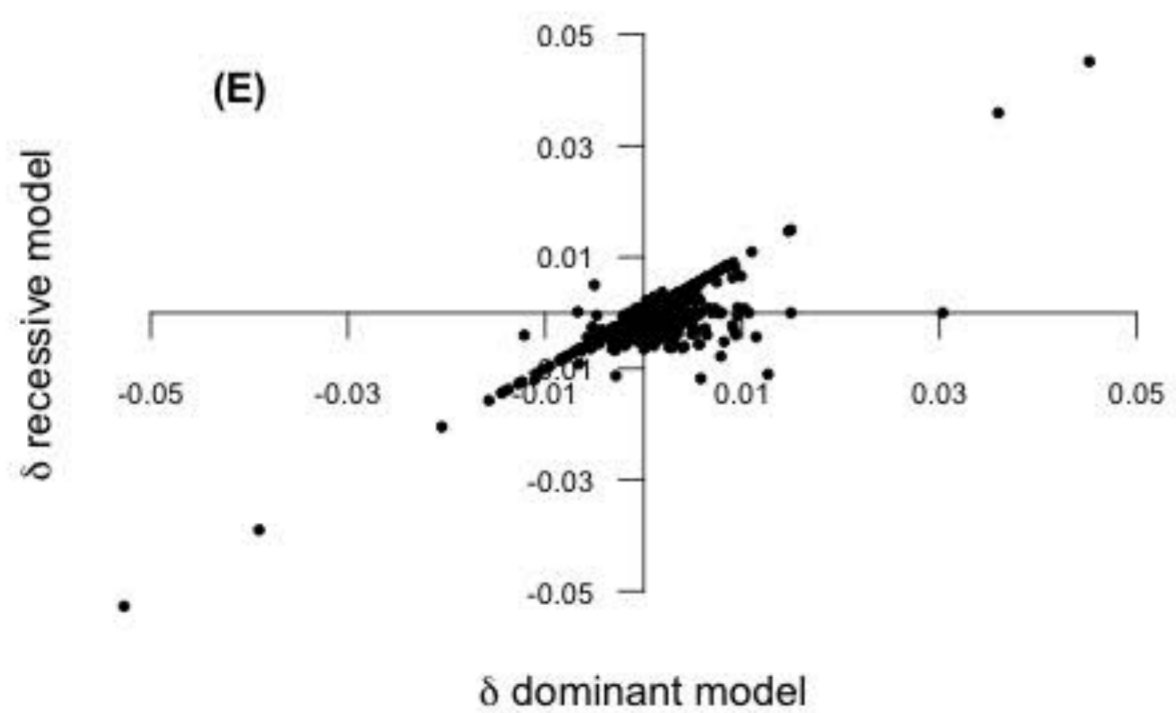
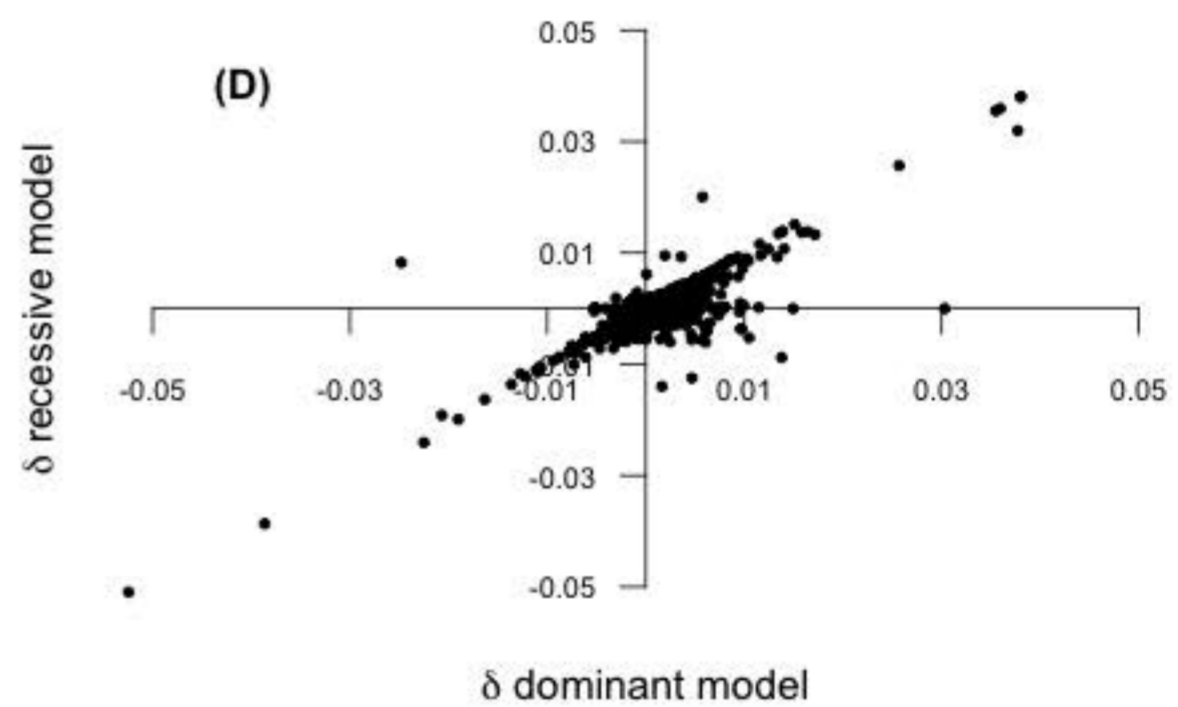
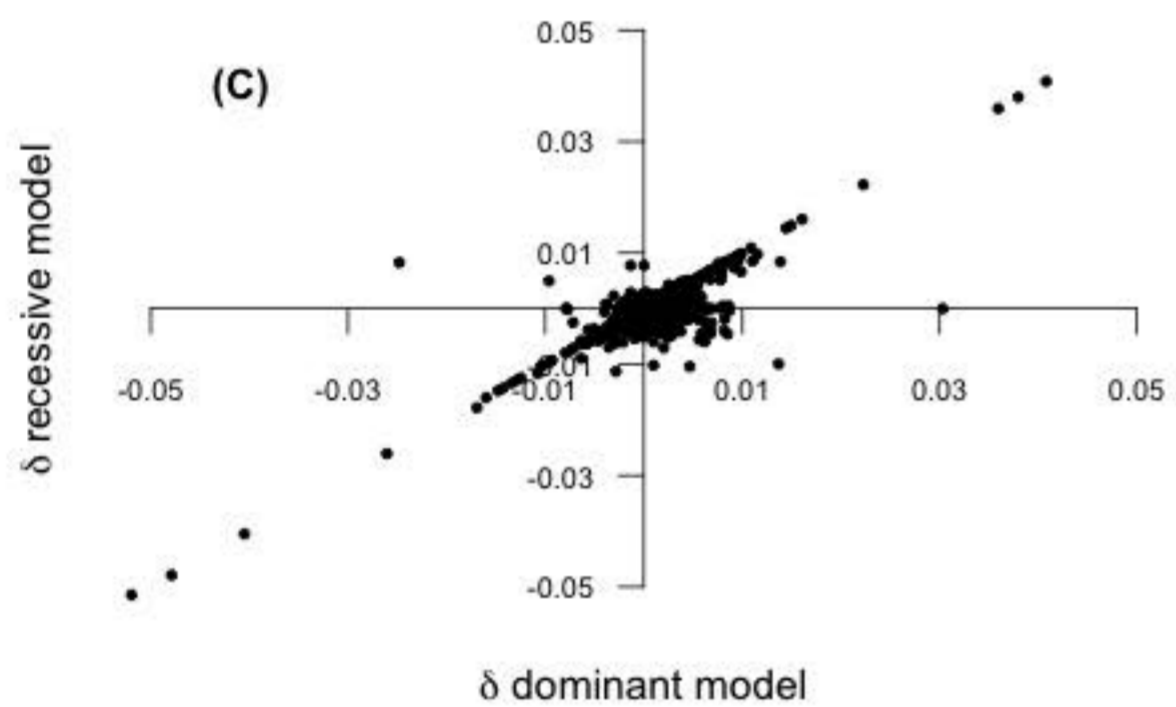
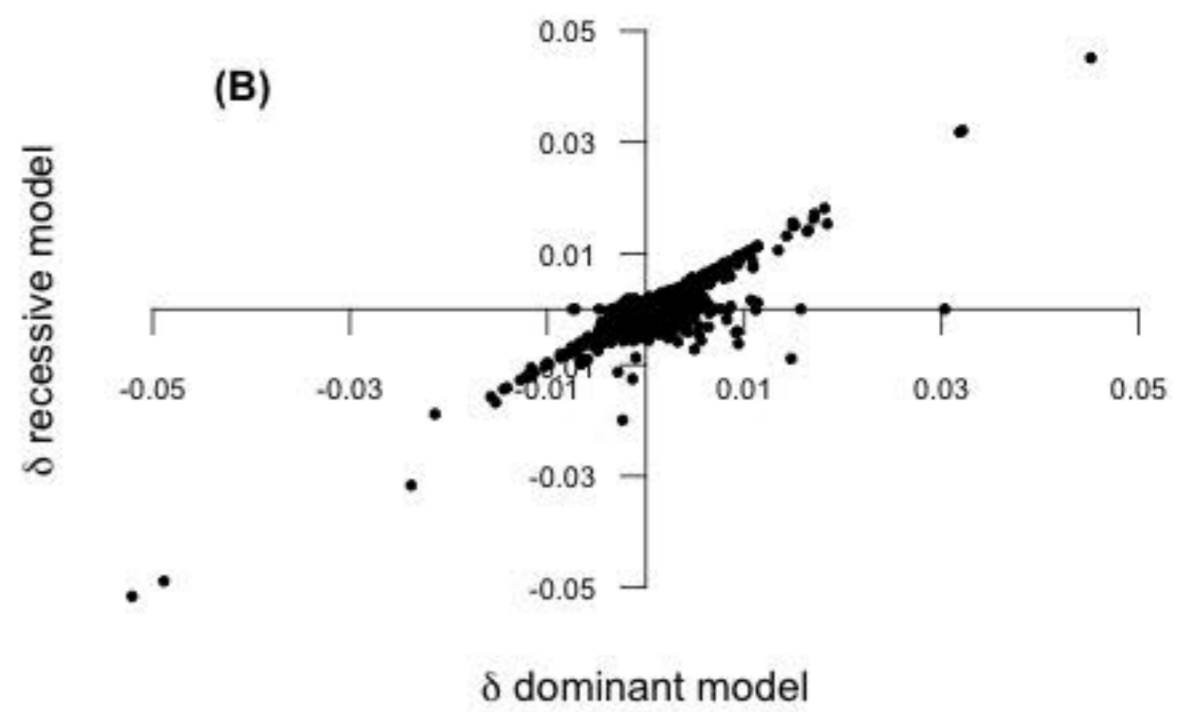
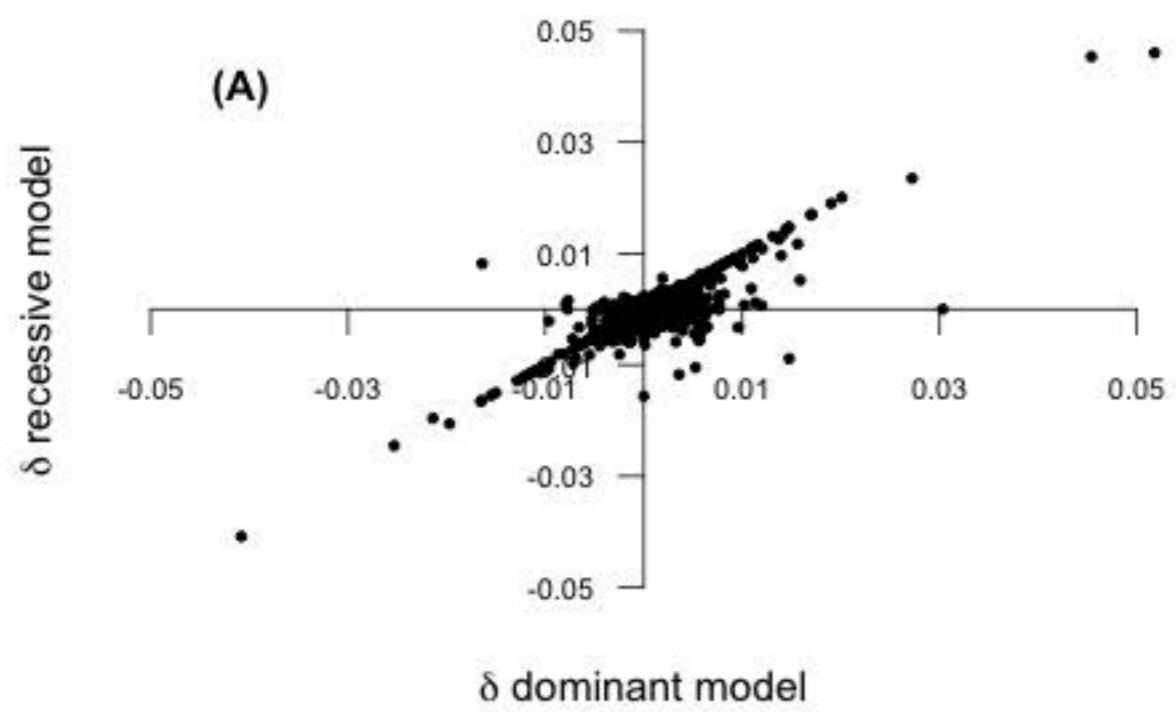


Fig S6



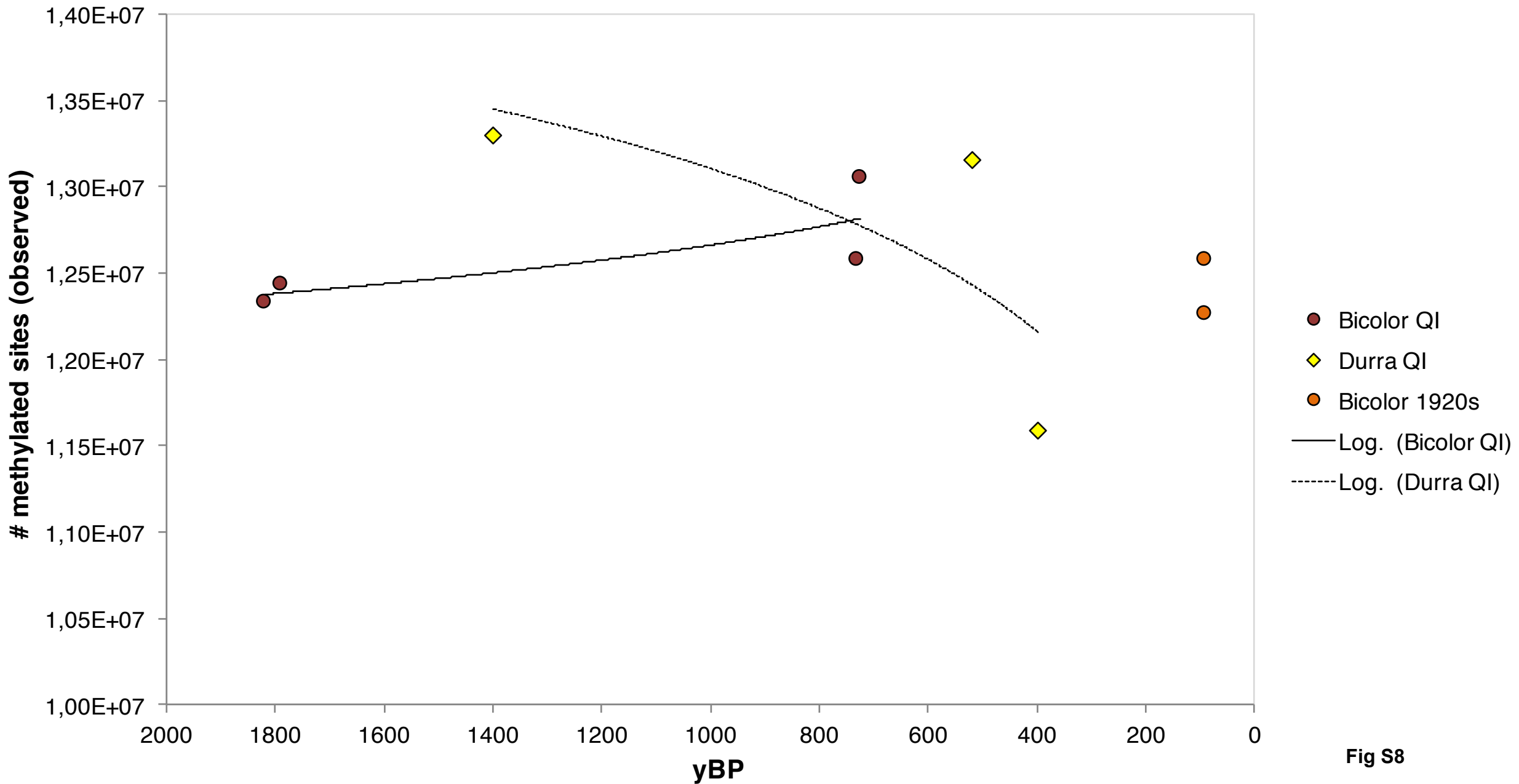


Fig S8

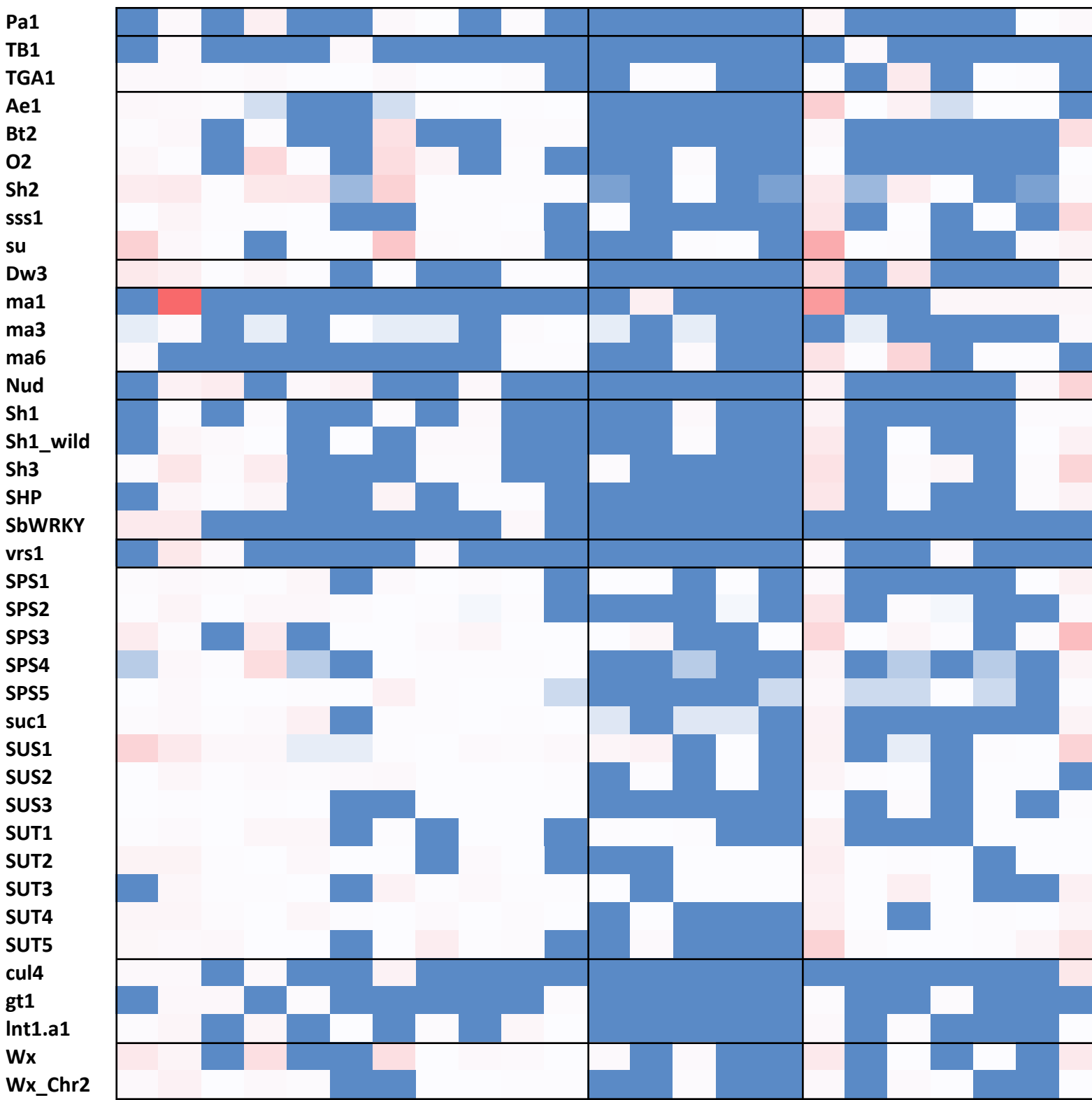
Gene A3 A5 A11 A6 A8 A12 A7 A9 A10 K1 K2 M1 M2 M3 M4 M5 M6 M7 M8 M9 M10 M11 M12

Archaeological

Modern domesticates

Modern wild-types

Fig S9



Compact / loose panicle

Glume architecture

Grain quality

Internode length

Maturity

Naked seed

Shattering

Spikelet architecture

Sugar production

Tillering

Waxy endosperm

Min

Max

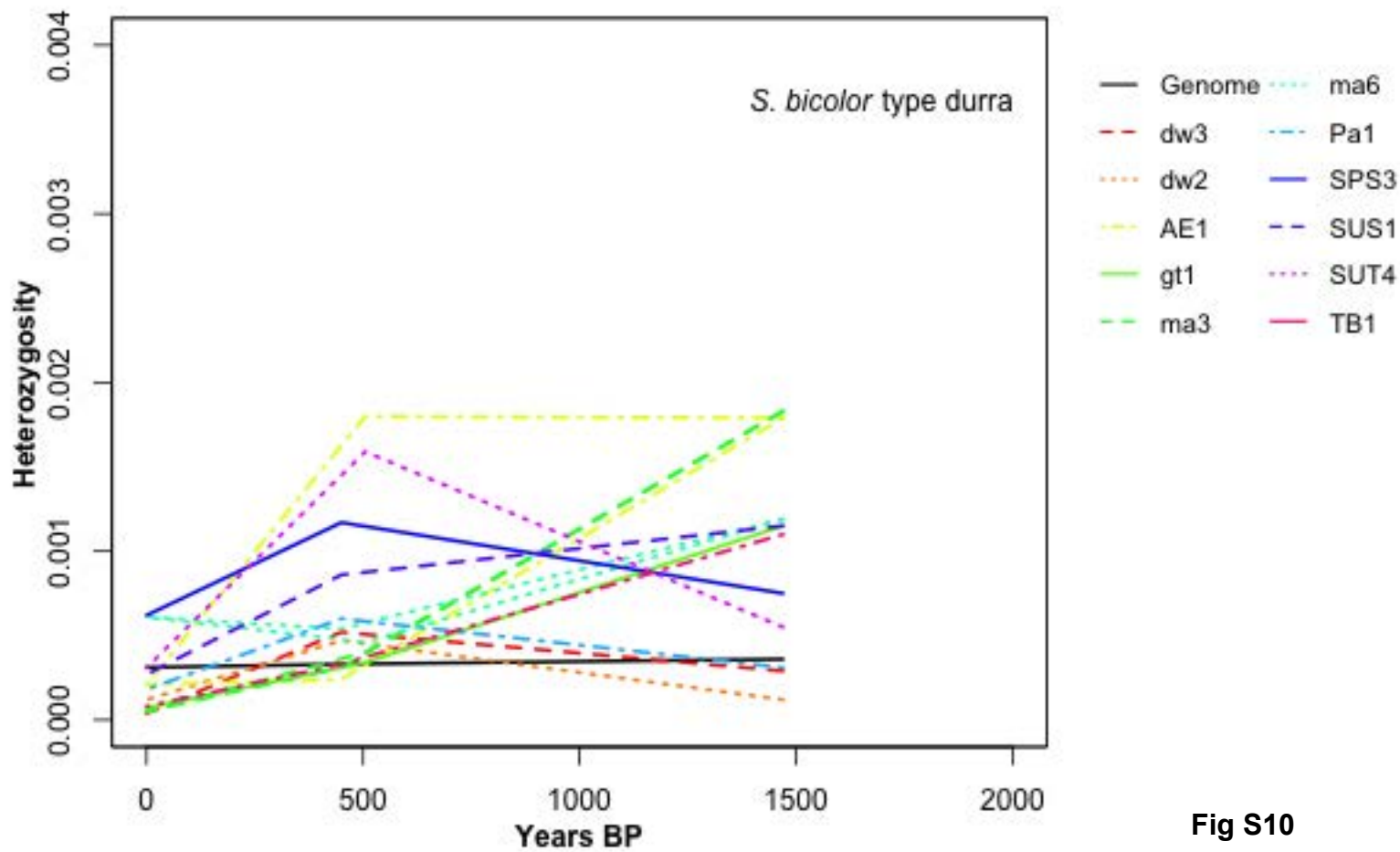
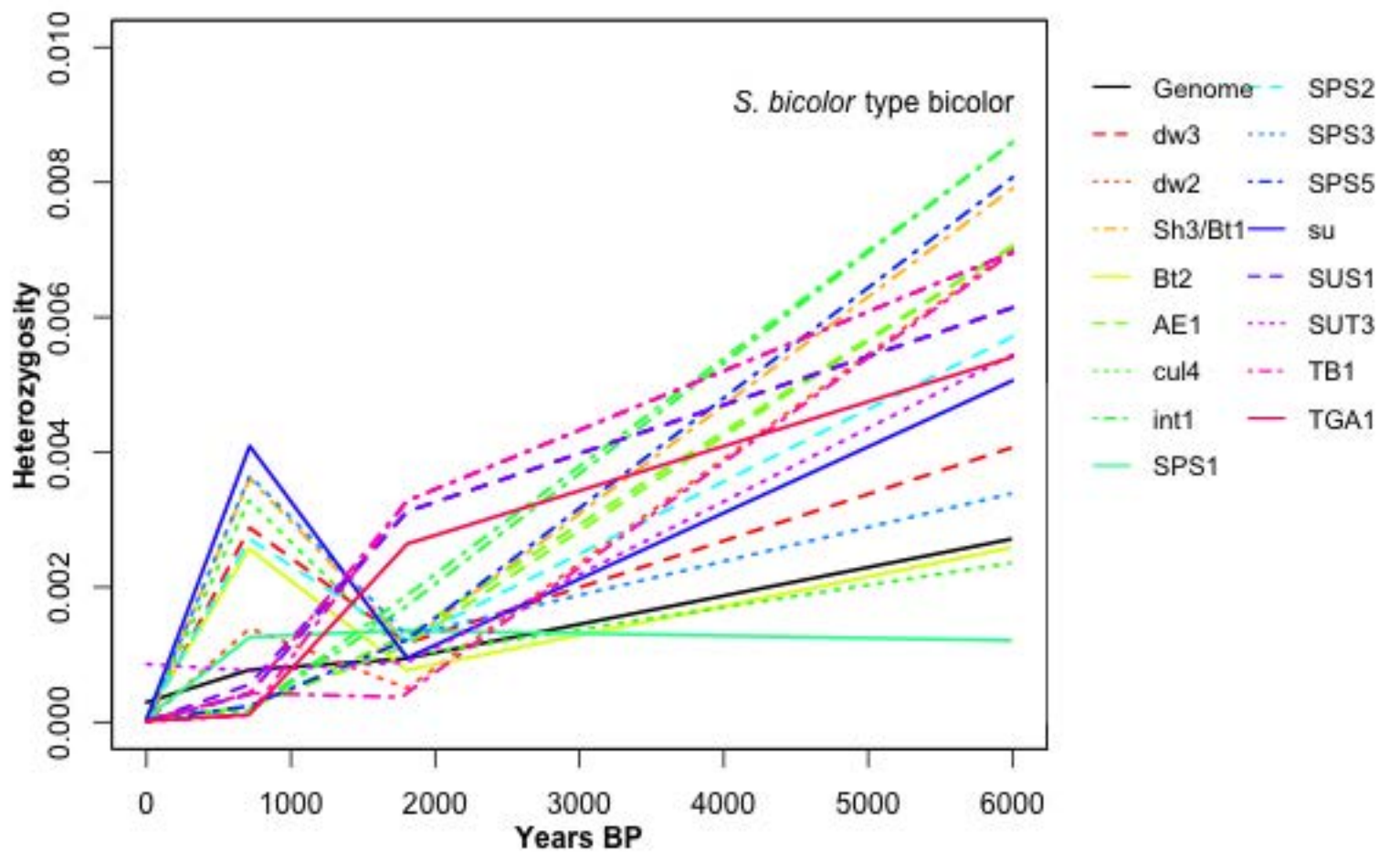


Fig S10

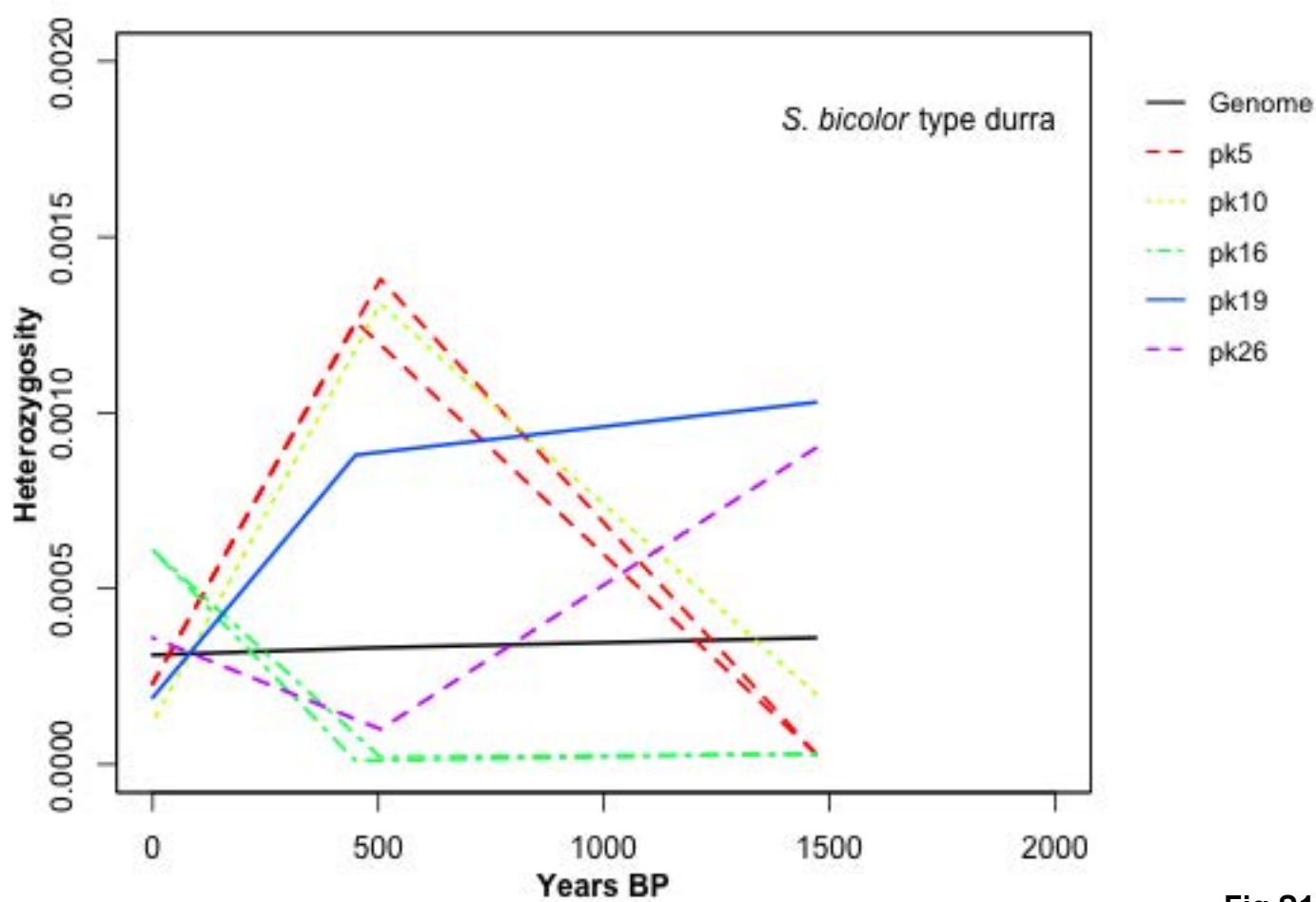
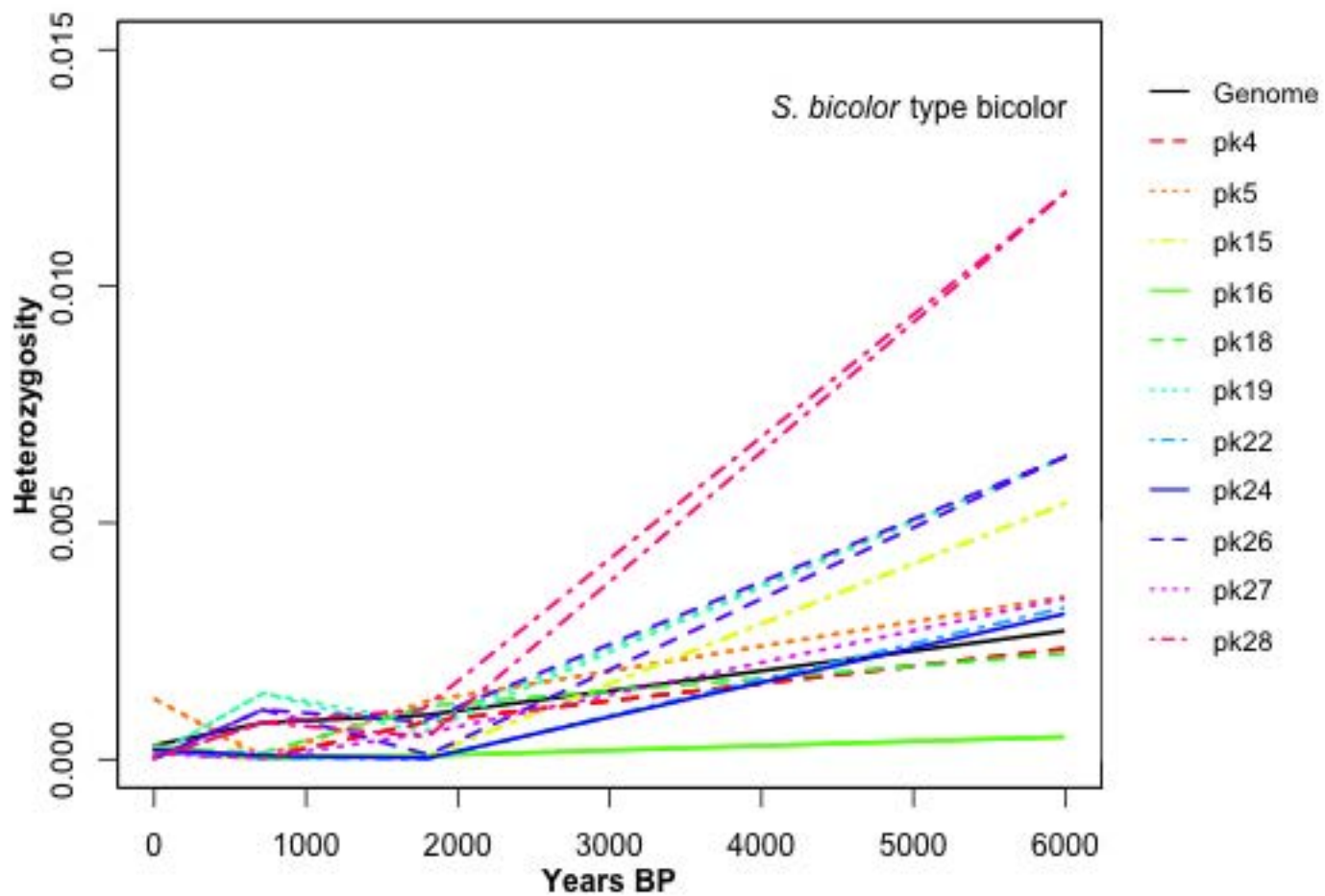


Fig S11

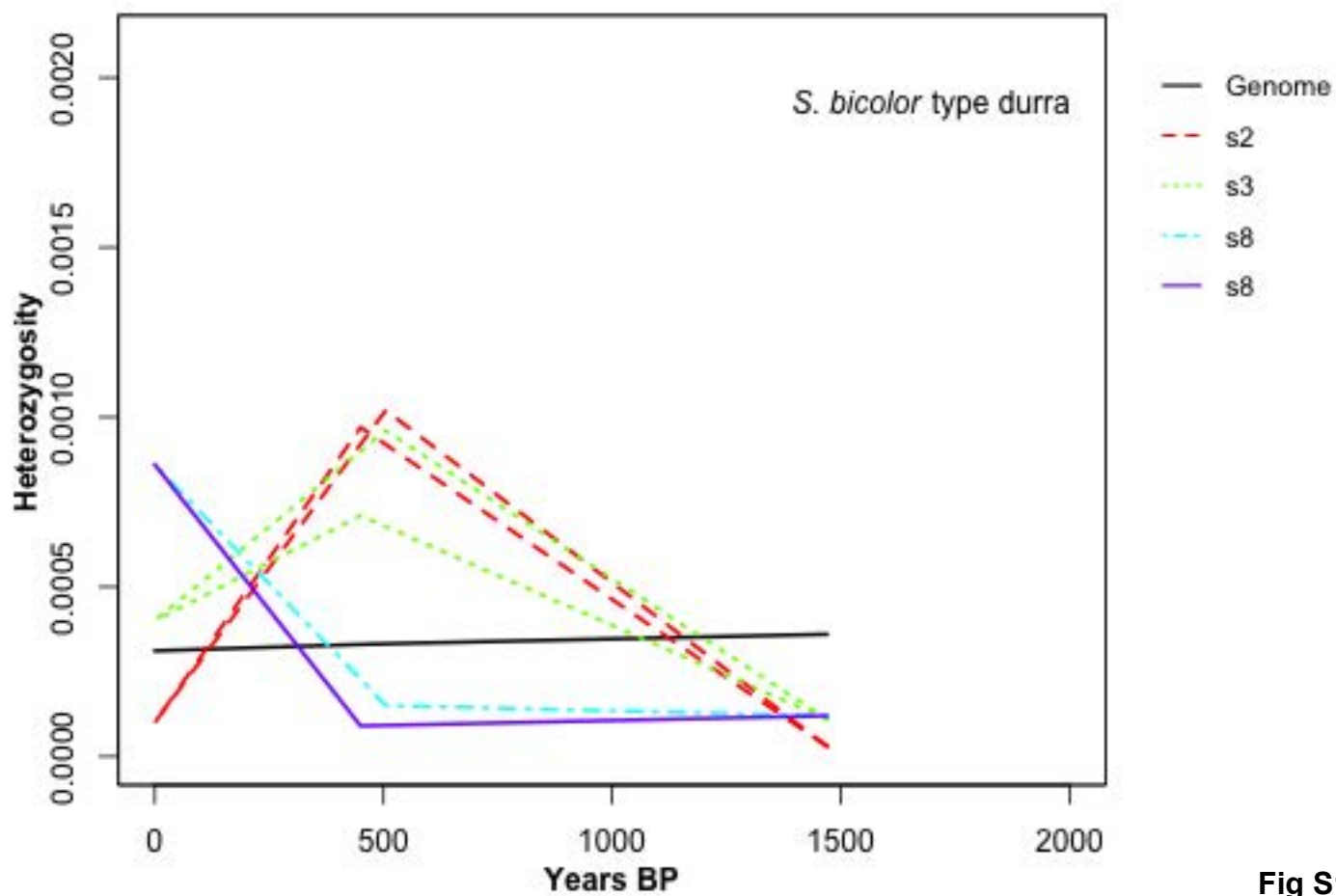
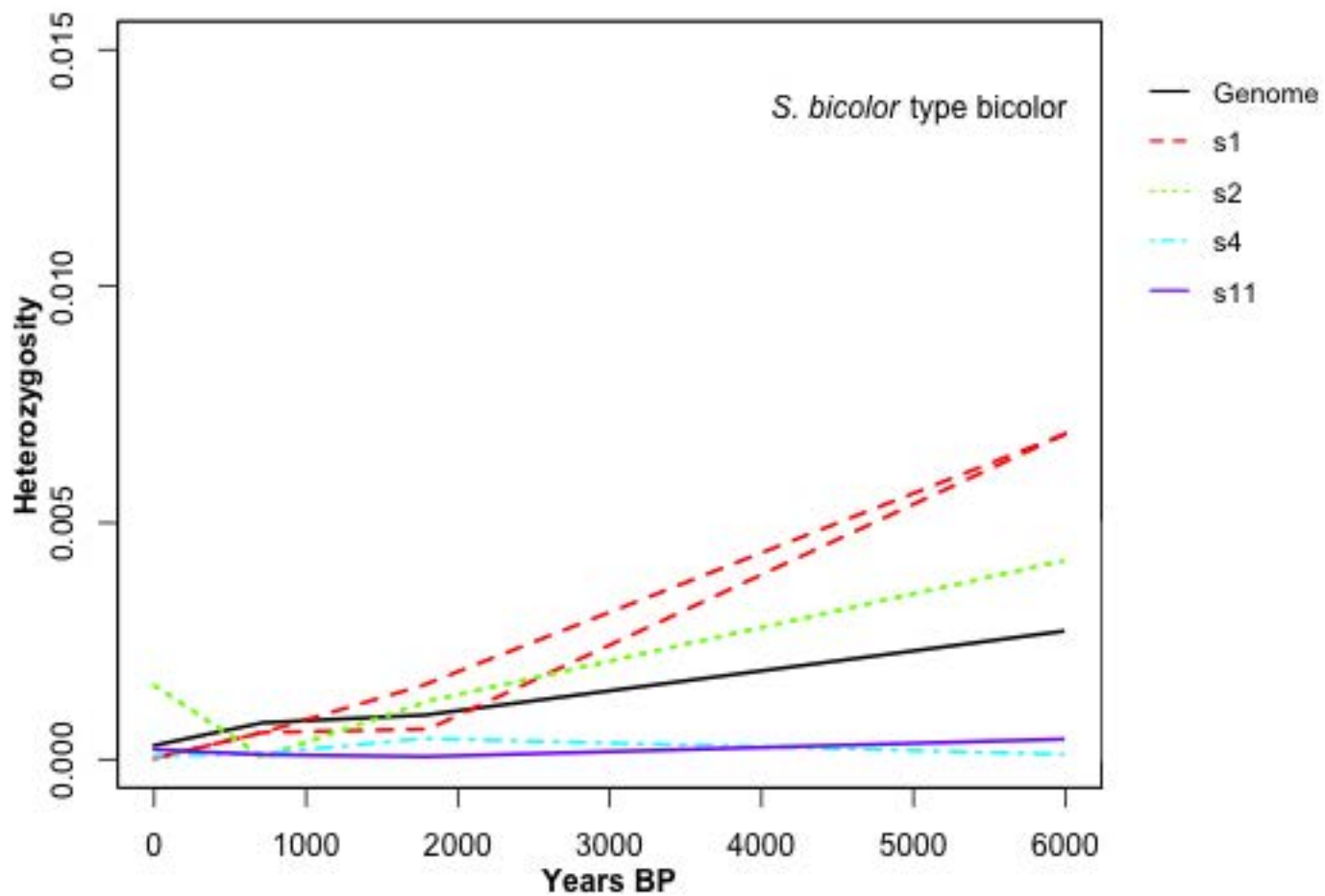


Fig S12

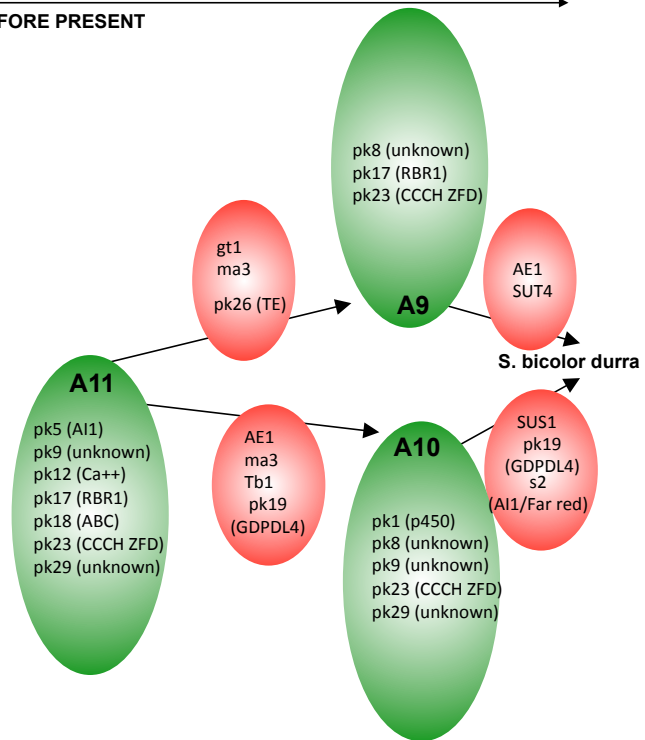
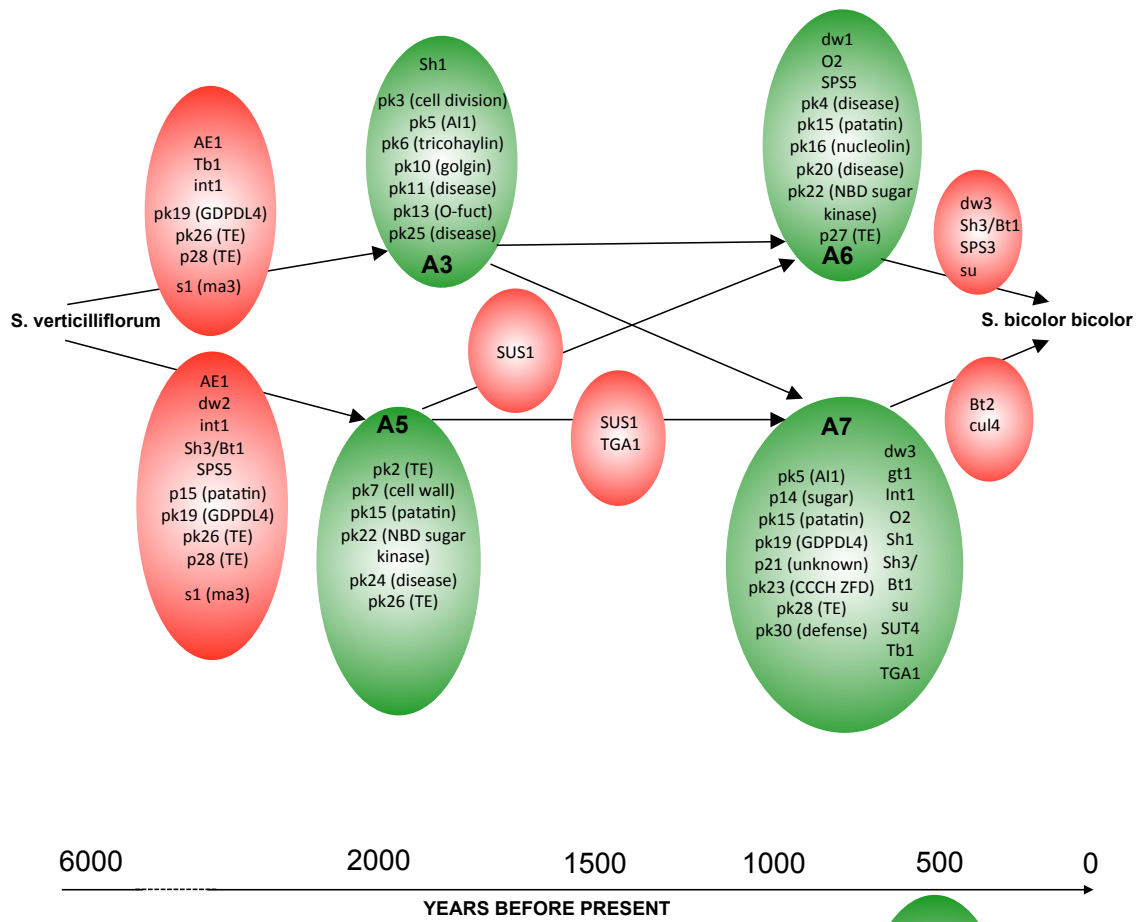


Fig S13

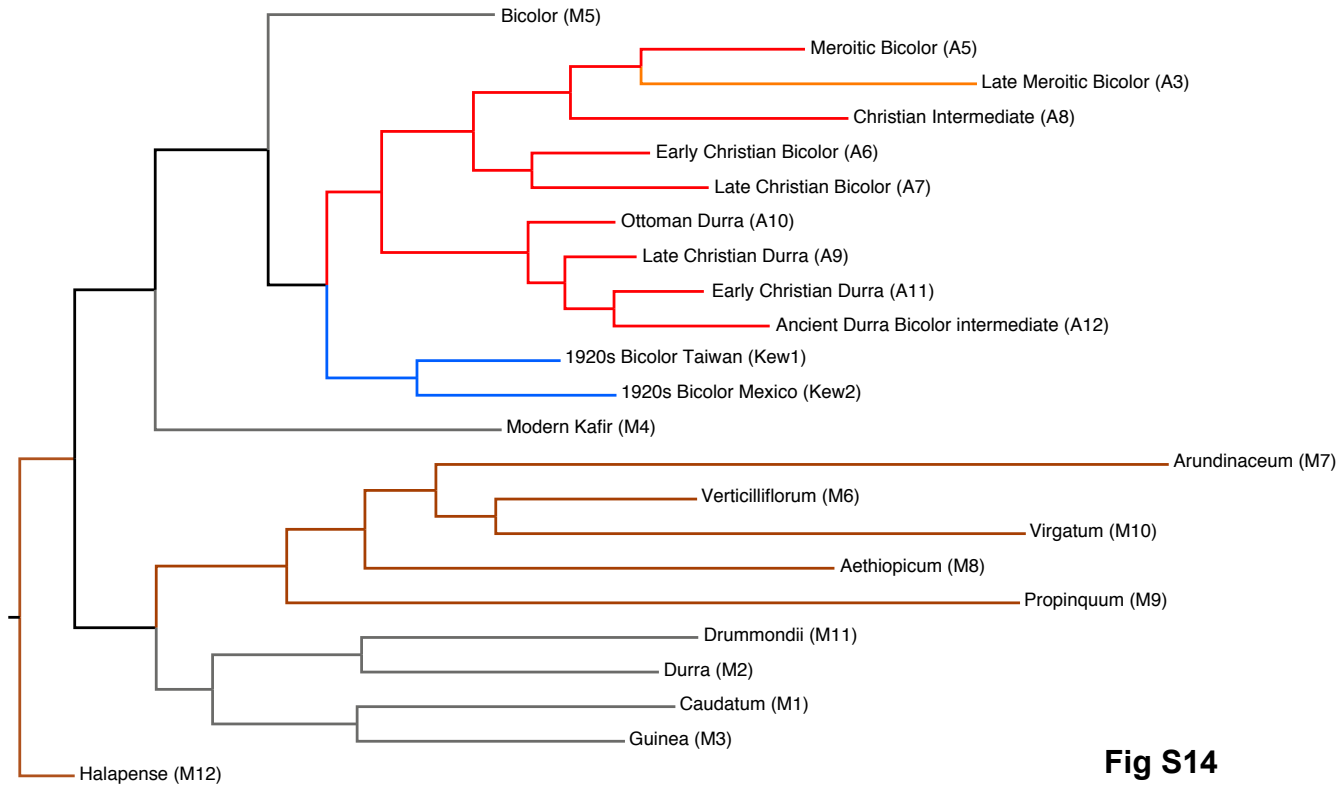
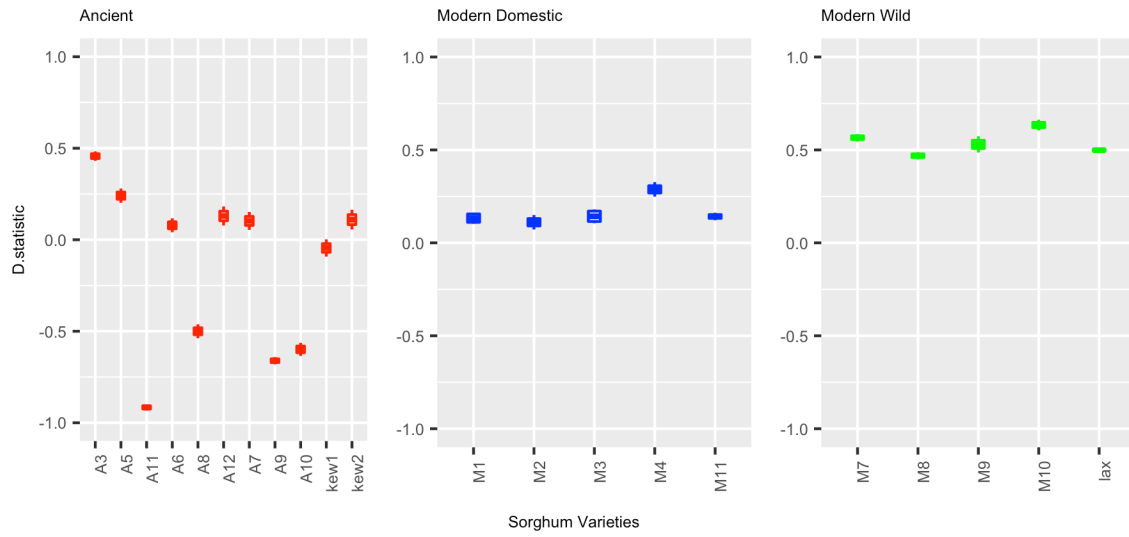


Fig S14

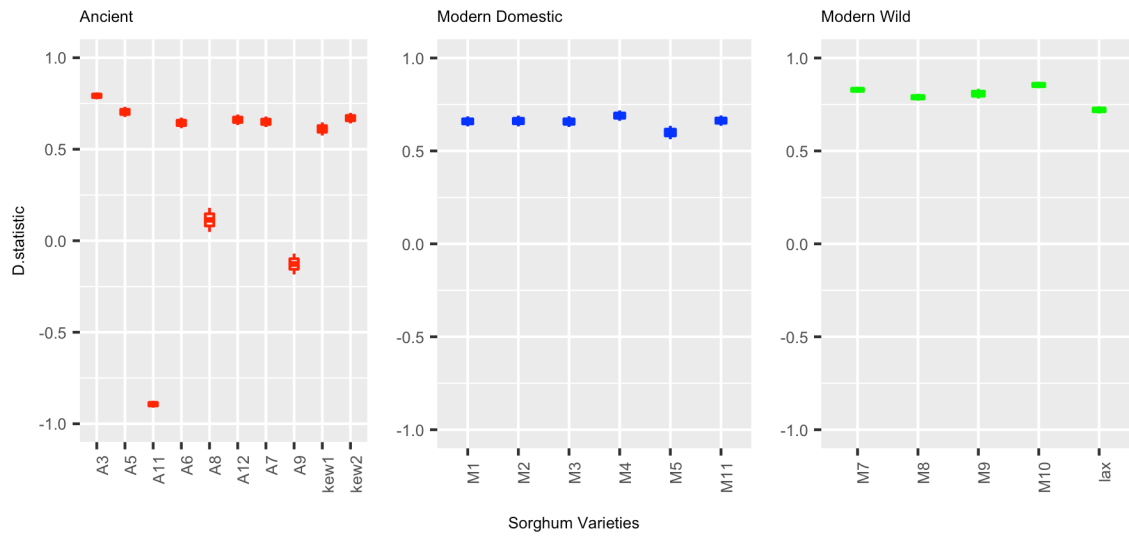
Introgression study - D(BTX623, Diff.Varieties,*S.verticilliflorum*,*S.halapense*)

A



Introgression study - D(A10, Diff.Varieties,*S.verticilliflorum*,*S.halapense*)

B



Introgression study - D(A11, Diff.Varieties,*S.verticilliflorum*,*S.halapense*)

C

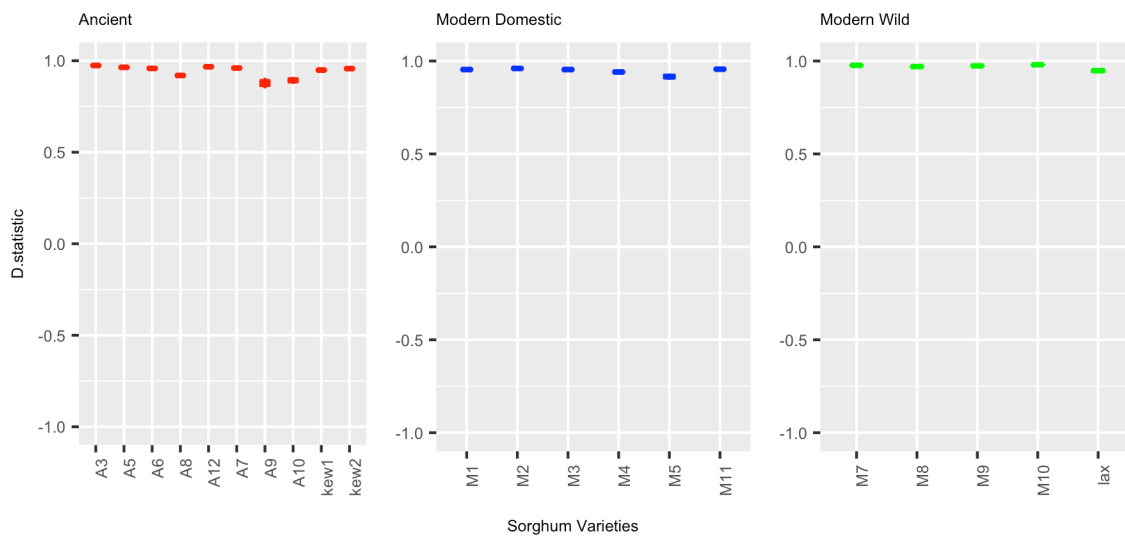
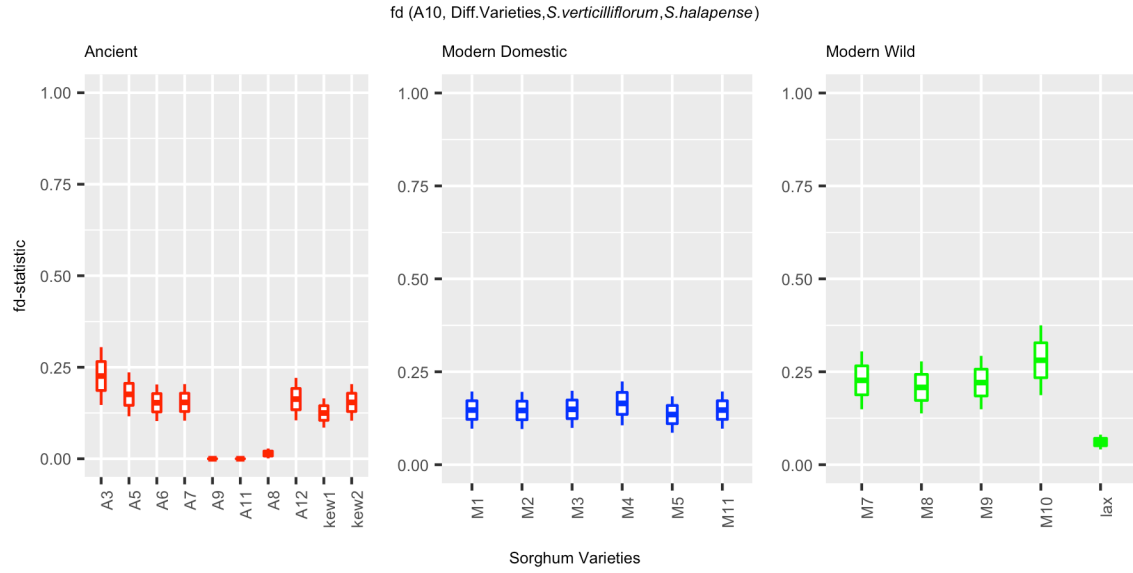
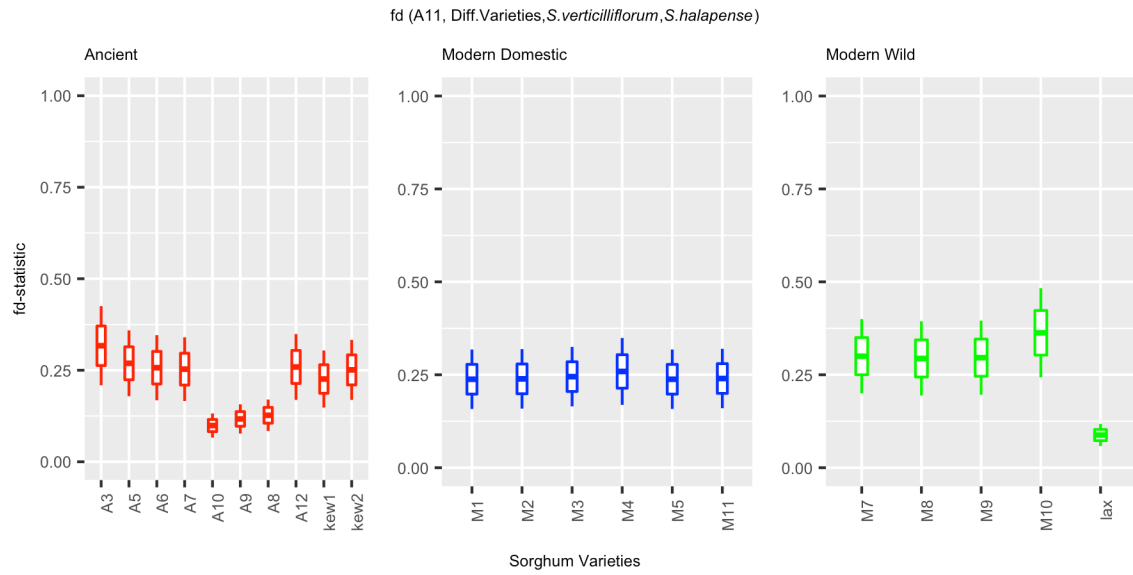


Fig S15

A**B****Fig S16**

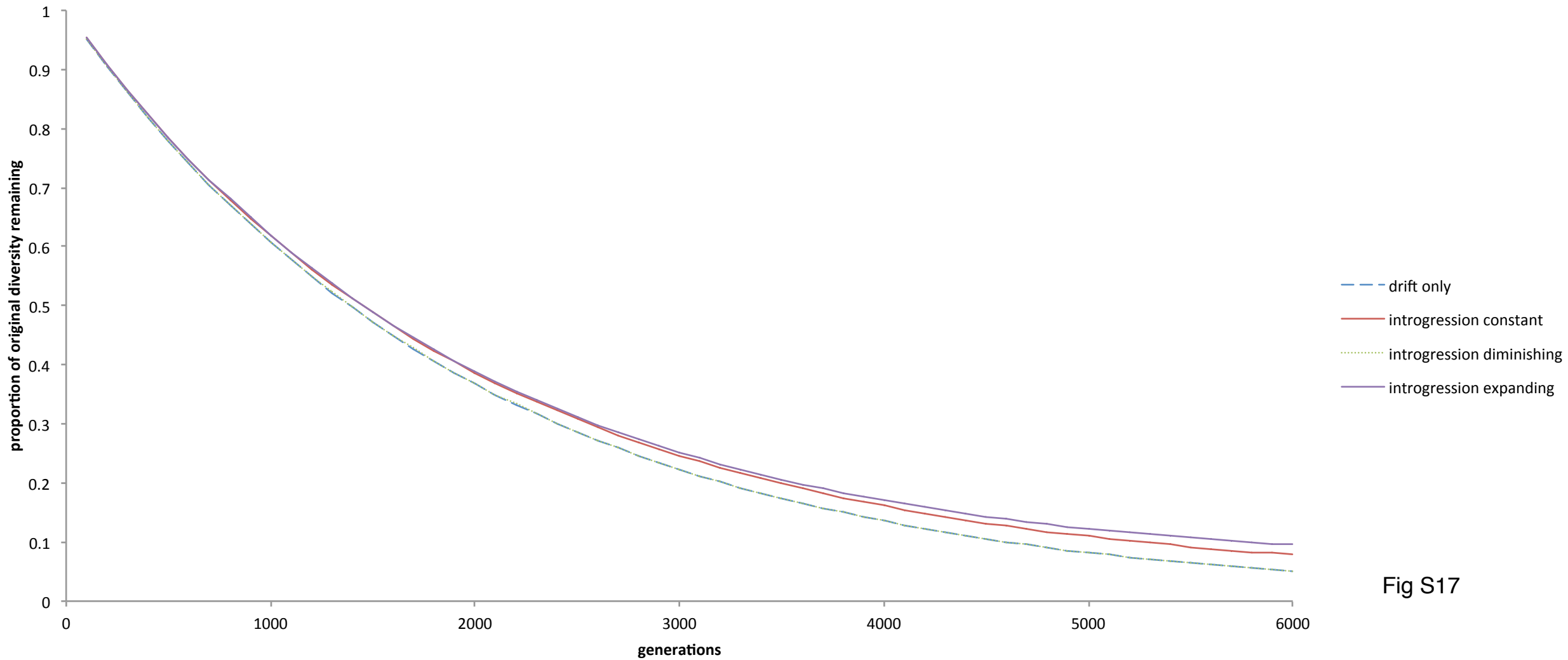


Fig S17

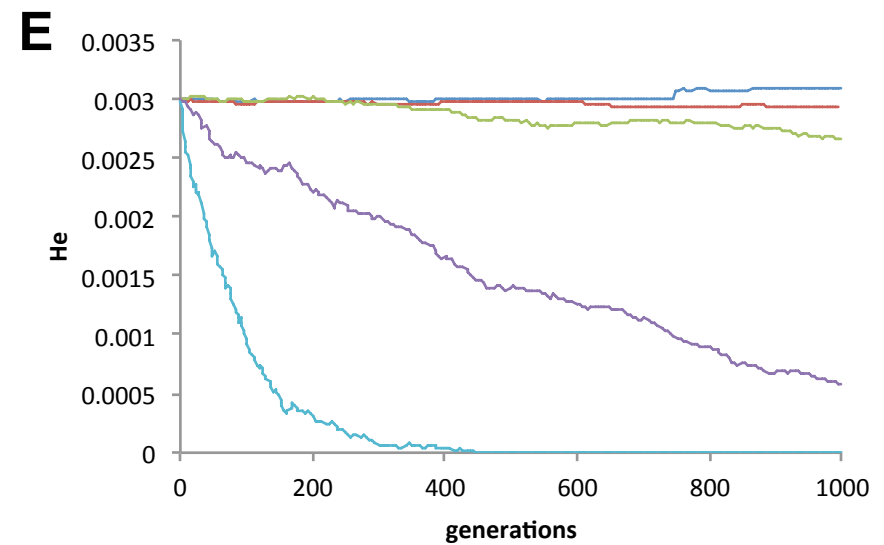
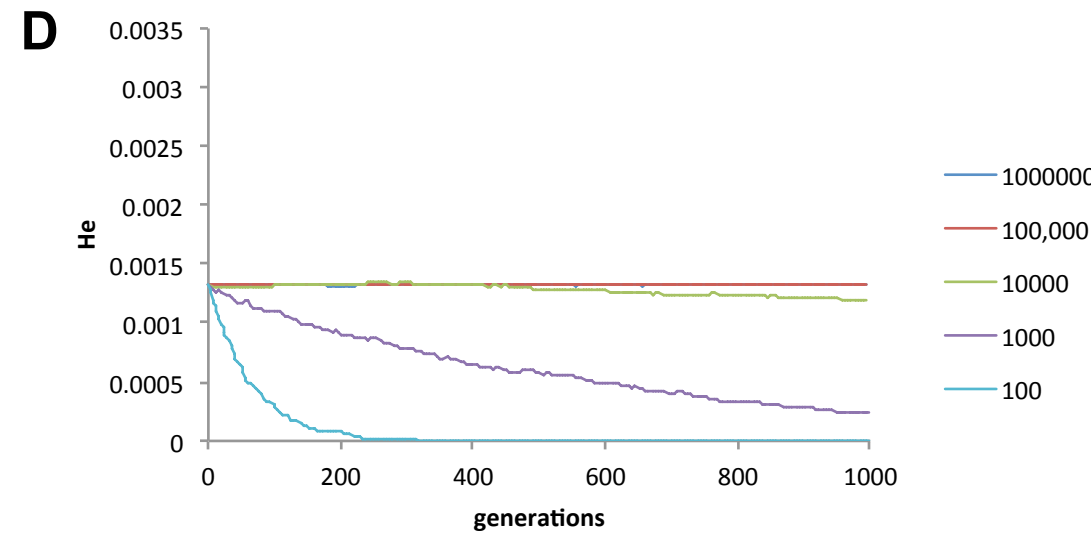
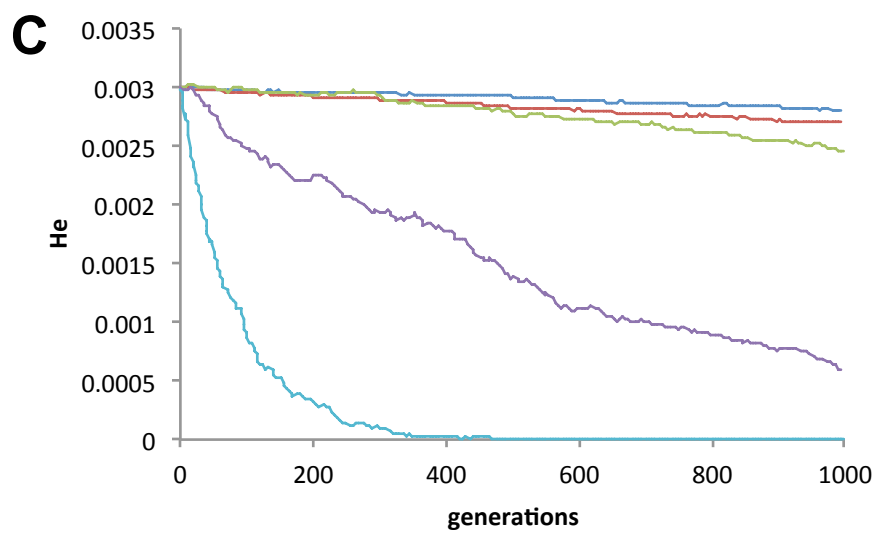
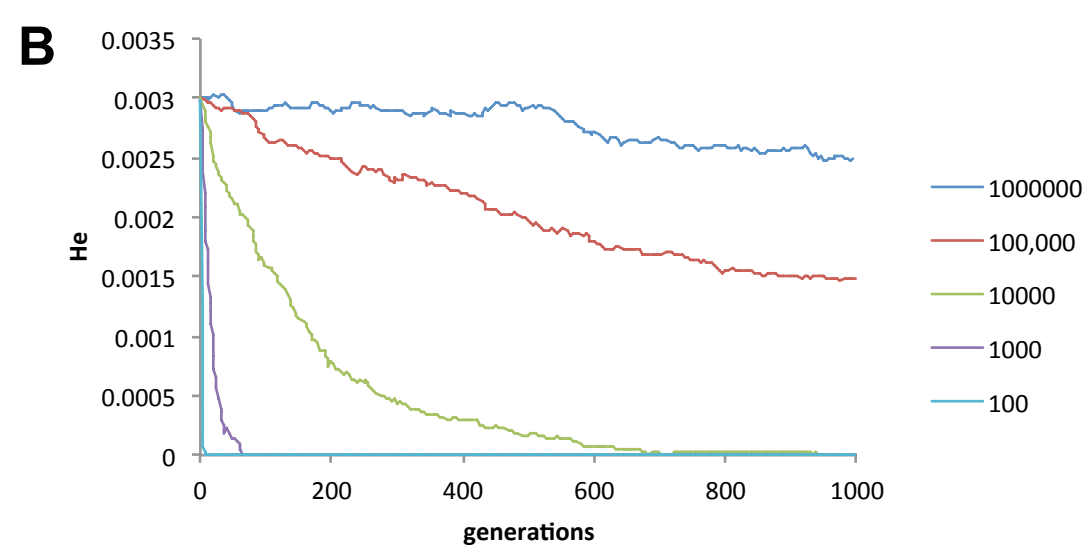
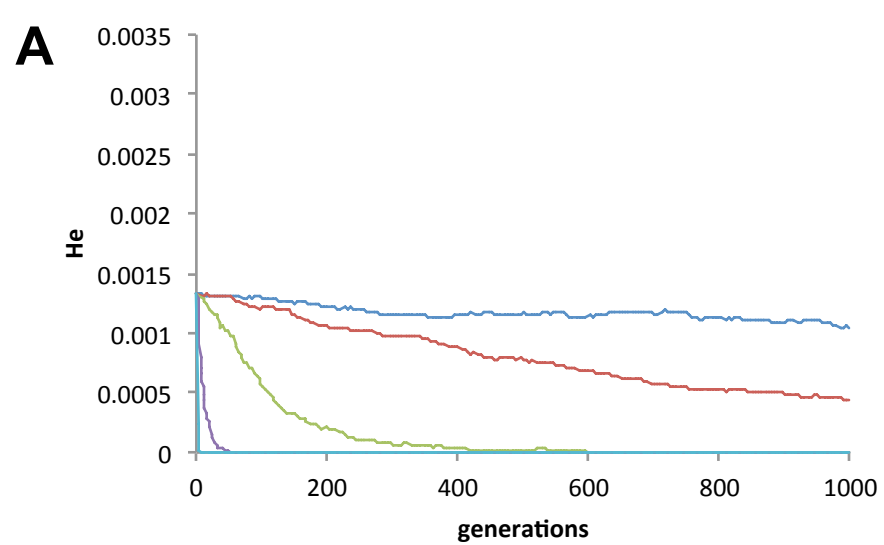


Fig S18