the geneticssociety

www.nature.com/hdy

COMMENT

Wild progenitors provide a sound baseline model for evolutionary analysis of domesticated crop species

Limei Zhong [1]✉, Youlin Zhu [1]✉ and Kenneth M. Olsen [2]✉

Heredity (2023) 130:111–113; https://doi.org/10.1038/s41437-023-00605-w

We write to respond to Jensen's comments on our recently published paper entitled "Hard versus soft selective sweeps during domestication and improvement in soybean" (Zhong et al. 2022). The author has raised two main criticisms of our selection analyses. First, he argued that the effects of purifying and background selection (BGS) were not taken into account when constructing the null demographic model, which could confound our inferences about hard and soft sweeps; second, he questioned the applicability of the test statistics and the empirical outlier approach we used for selection scans. He also noted that these issues are common among many published analyses in evolutionary and ecological genomics, and our paper simply serves as an illustrative example. Here we provide a rebuttal to the critique of our study, although we do not guarantee it is applicable to other publications that were not explicitly indicated.

ON THE QUESTION OF PURIFYING AND BACKGROUND SELECTION

The first criticism put forward by Jensen was that we neglected to consider the effects of purifying selection and BGS in our demographic model, which could affect the interpretation of subsequent selection analysis. The demographic model in our paper was indeed constructed under a neutral assumption; it served as a reference for examining the likelihood that genes would be found with H12 values above the significance thresholds even under neutral evolution. The thresholds were shown to be greater than 99.5% of the simulated neutral loci, so in the context of our study the essence of this question should be: how likely is it that the significant deviations we observed for our selection candidates were in fact generated by purifying and background selection?

We would like to start from the domestication system we studied to address this question. Two important reasons why crop domestication has long been considered an excellent model for studying evolution are: (1) the existence of wild progenitor species, which can serve as living representatives of the crop species as it existed prior to domestication, and (2) the very short evolutionary time frame during which domestication has occurred (typically < 10,000 years). Because crop species and their wild progenitors share the vast majority of their evolutionary history, the influence of shared evolutionary factors can be largely controlled by directly comparing the variation pattern of the crop species with the wild progenitor. In our soybean analyses, we observed a contrasting pattern of H12 & H2/H1 between domesticated and wild populations; this is unlikely to be caused by BGS as BGS would have acted similarly on the wild and domesticated soybean genomes for all but the recent evolutionary past. Moreover, for domesticated species such as soybean, factors including inbreeding and the domestication bottleneck (see below) are expected to reduce the efficacy of purifying selection (Glémin 2007; Charlesworth 2009; Petit and Barbadilla 2009; Makino et al. 2018; Bosse et al. 2019); this phenomenon is referred to as 'cost of domestication' (Moyers et al. 2018). In sum, given biologically realistic rates at which deleterious mutations would arise, the short evolutionary time frame since domesticated soybean diverged from its wild progenitor, and the diminished efficiency of purifying selection in crop lineages, it is unlikely that BGS and purifying selection would be major factors in shaping the evolution of the crop genome, or that any such effects would seriously confound selective sweep inferences based on the H12 and H2/H1 statistics.

Although we did not introduce fitness parameters characterizing purifying selection into our demographic model, we did examine the effect of BGS using another approach. Jensen argued that it is problematic to say background selection can probably be ignored compared to positive selection during domestication 'based on examining haplotype distributions in genes relative to genes together with linked non-coding regions.' However, this is not what we did—nor is it what we described doing in our Supplementary Note 2 that Jensen cites. What we did do was to compare selection mode inferred from analysis of the whole genome to selection inferences specifically for genes with linked non-coding regions; this allowed us to understand the effect of background selection on patterns of selection inferred for the protein-coding genes. If the effect of BGS were significant, the results from gene and genome-based analyses would be expected to be different because gene regions are highly likely to evolve under pervasive BGS compared to the intergenic regions. However, there is no obvious difference between gene and genome-based analyses in both landrace and improved soybean. A reasonable explanation is that artificial positive selection has been much stronger than background selection during the period of soybean domestication. This is not groundless in the domestication context, as many non-synonymous substitutions that are generally deleterious in wild settings but strongly favored under domestication have reached a high frequency in

[1]Key Laboratory of Molecular Biology and Gene Engineering in Jiangxi, School of Life Sciences, Nanchang University, Nanchang, China. [2]Department of Biology, Washington University in St. Louis, St. Louis, MO, USA. Associate editor: Sara Goodacre. ✉email: zhonglm@ncu.edu.cn; ylzhu@ncu.edu.cn; kolsen@wustl.edu

Received: 2 December 2022 Revised: 9 February 2023 Accepted: 10 February 2023
Published online: 24 February 2023

domesticated populations. A classic example comes from the *sh4* gene for shattering in rice; the mutation responsible for reduced shattering is nearly fixed in cultivated rice populations while reduced shattering is a maladaptive trait in wild (Li et al. 2006). The presence of strong positive selection was also evident in the soybean genome, as we discussed in our paper.

The PSMC method developed by Li and Durbin (2011) uses hidden Markov models and coalescent theory to infer the history of population size from a single genome. We agree that the PSMC curves in theory cannot always reliably be interpreted as plots of population-size changes, as purifying selection tends to remove genetic variations, and which would also lead to a reduction in population size. However, in our specific case, whole genome sequences instead of coding sequences or gene regions were used as input so that the confounding effects of selection on demographic inference would be greatly mitigated. Also, Jensen questioned that 'the resulting PSMC curves tend to take a characteristic shape regardless of the species or population being analyzed'. This is not a convincing argument, even judging from our article alone. We observed four types of shape in our analyses, including consistent N_e patterns for the 36 pseudogenomes created from 9 landrace accessions, contrasting trajectories for the pseudogenomes generated by wild individuals from the same or different subgroups, and continuous decline without fluctuation for the raw 9 wild genotypes and 9 landraces (data not shown). These observations reflect the sensitivity of PSMC to selfing mating systems and population structure (Li and Durbin 2011; Orozco-Terwengel 2016). Finally, as noted above, we point to the advantage of domestication systems for these analyses, given the ability to sample their extant wild relatives. For the specific case of soybean, a domestication bottleneck has been well documented in previous work through comparisons of genetic diversity or demographic models between domesticated population and their wild progenitors (Hyten et al. 2006; Guo et al. 2010). With this prior information, we therefore believe that the gradual reduction pattern of PSMC curves starting from about 10,000 years ago (corresponding to the time for the rise of agricultural civilization) (Doebley et al. 2006), is much more likely to reflect the real changes of population size caused by domestication bottleneck and human activities in the protracted domestication process (Purugganan 2019).

ON THE QUESTIONS OF TEST STATISTICS AND THE EMPIRICAL OUTLIER APPROACH

The second criticism put forward by Jensen is on the test statistics and the empirical outlier approach we (and many others) have used for selection detection. He argued that utilizing an empirical outlier approach to perform selection scans is highly inappropriate because assigning the 5% tails of an empirical distribution will identify 5% of loci as being putatively swept regardless of the true underlying fraction. As a typical population genomics study design for detecting positive selection (Akey 2009), we believe that the outlier approach is effective for those systems where positive selection is evident (*e.g.*, domesticated crops), although false positives are inevitable. For our soybean study system, this approach was strongly supported by the successful detection of the causative SNPs of the five known hard-sweep genes as F_{ST} outliers in our case (See Figure 4c in Zhong et al. 2022). It should also be noted that we did not focus on the raw candidates identified by the five statistics individually, and we clearly pointed out the presence of considerable false-positive genes among them. An important use of those selection analyses was to help us feature the population characteristics of the six well-known domestication-related genes that we used as positive controls, and understand how they were detected by the five statistics in the genomic context. To limit the number of false positives, we used common signatures from the six known selection targets to inform identification of other candidate loci with similar features and identified 348 candidate genes. Compared to fitting a baseline model consisting of the underlying details of the population suggested by Jensen, we believe that this practice-based strategy is more intuitive and less assumption-dependent.

We also refute Jensen's interpretation that the low overlap amongst all test statistics provides evidence for absence of selective sweeps. Because the level and pattern of selection signature depends on selection strength, selection mode, and how long ago it occurred, even truly selected genes do not necessarily satisfy all the selection criteria unless the sweeps are young and strong enough (e.g., only *Tof12* among the six positive-control genes could be detected by all of the five tests). While these statistics are indeed correlated in some ways, they were calculated based on different population characteristics and designed for different selection situations — i.e., hard sweeps, soft sweeps, and incomplete sweeps of varying ages. The most powerful approach would be an integrated one that has proven ability to capture the real selection signatures and detect as many known genes as possible. This is the strategy we used to detect the 348 genes we identified, which together account for just 0.6% of genes in the soybean genome but include all five of the hard sweep positive control genes.

In summary, we argue against judging the methods of analysis used in an empirical study without considering the specific biological context of the study system (in our case, domesticated soybean and its wild ancestor) and prior information from previous published papers on that study system. For domestication systems in particular, the variation pattern of extant wild ancestors instead of a sophisticated model with many assumptions can best provide a baseline model for evolutionary analysis of selection under domestication.

REFERENCES

Akey JM (2009) Constructing genomic maps of positive selection in humans: where do we go from here? Genome Res 19:711–722

Bosse M, Megens HJ, Derks MFL, de Cara ÁMR, Groenen MAM (2019) Deleterious alleles in the context of domestication, inbreeding, and selection. Evol Appl 12:6–17

Charlesworth B (2009) Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. Nat Rev Genet 10:195–205

Doebley JF, Gaut BS, Smith BD (2006) The molecular genetics of crop domestication. Cell 127:1309–1321

Glémin S (2007) Mating systems and the efficacy of selection at the molecular level. Genetics 177:905–916

Guo J, Wang Y, Song C, Zhou J, Qiu L, Huang H et al. (2010) A single origin and moderate bottleneck during domestication of soybean (*Glycine max*): implications from microsatellites and nucleotide sequences. Ann Bot 106:505–514

Hyten DL, Song Q, Zhu Y, Choi I-Y, Nelson RL, Costa JM et al. (2006) Impacts of genetic bottlenecks on soybean genome diversity. Proc Natl Acad Sci USA 103:16666–16671

Li C, Zhou A, Sang T (2006) Rice domestication by reducing shattering. Science 311:1936–1939

Li H, Durbin R (2011) Inference of human population history from individual whole-genome sequences. Nature 475:493–496

Makino T, Rubin CJ, Carneiro M, Axelsson E, Andersson L, Webster MT (2018) Elevated proportions of deleterious genetic variation in domestic animals and plants. Genome Biol Evol 10:276–290

Moyers BT, Morrell PL, McKay JK (2018) Genetic costs of domestication and improvement. J Hered 109:103–116

Orozco-Terwengel P (2016) The devil is in the details: The effect of population structure on demographic inference. Heredity 116:349–350

Petit N, Barbadilla A (2009) Selection efficiency and effective population size in *Drosophila* species. J Evol Biol 22:515–526

Purugganan MD (2019) Evolutionary insights into the nature of plant domestication. Curr Biol 29:R705–R714

Zhong L, Zhu Y, Olsen KM (2022) Hard versus soft selective sweeps during domestication and improvement in soybean. Mol Ecol 31:3137–3153

ACKNOWLEDGEMENTS

We would like to thank the two anonymous reviewers for their helpful suggestions and comments during the review process. This work was supported by the National Natural Science Foundation of China (grant nos. 31801050, 31960433) and China Scholarship Council (grant no. 201806820031).

AUTHOR CONTRIBUTIONS

LZ wrote the manuscript draft, YZ and KMO commented and revised the manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Correspondence and requests for materials should be addressed to Limei Zhong, Youlin Zhu or Kenneth M. Olsen.