the geneticssociety

www.nature.com/hdy

COMMENT

Population genetic concerns related to the interpretation of empirical outliers and the neglect of common evolutionary processes

Jeffrey D. Jensen[1] ✉

Heredity (2023) 130:109–110; https://doi.org/10.1038/s41437-022-00575-5

I wish to express a number of concerns related to the analyses and interpretations presented by Zhong et al. (2022). While my critique is contextualized around this specific publication, it simply serves as an illustrative example of issues common amongst many published analyses in evolutionary and ecological genomics; namely, the development of appropriate baseline models for evolutionary genomic analysis, and the interpretation of empirical outliers. Thus, the publication in question is certainly not uniquely problematic in these discussed regards.

In brief, using wild and domesticated soybean accessions, the authors first estimated a demographic history for these populations for use as a null expectation. They next calculated multiple summary statistics related to levels and patterns of within- and between-population variation, and interpreted the loci in the tails of those empirical distributions as being putatively positively selected. They concluded by interpreting these results in terms of the relative prevalence of hard versus soft selective sweeps during soybean domestication.

CONSTRUCTING AN APPROPRIATE EVOLUTIONARY BASELINE MODEL

The first primary concern relates to the construction of their demographic model, which was subsequently employed as a null for the performed selection scans. In order to estimate the timing and severity of demographic events, the authors applied the widely used PSMC approach to a subset of their data consisting of 18 accessions (9 wild and 9 landraces). Problematically, the authors neglected the contributions of purifying and background selection (BGS) in shaping observed levels and patterns of variation, despite the constant input of deleterious mutations in functional regions across the genome (for a review of the effects of selection at linked sites, see Charlesworth and Jensen 2021). While Zhong et al. did mention the possibility of BGS effects in their Supplementary Note 2, they determined based on examining haplotype distributions in genes relative to genes together with linked non-coding regions that "background selection can probably be ignored compared to positive selection during domestication." This is a peculiar statement given that BGS effects would be expected to be particularly pervasive in this predominantly selfing species (Barrett et al. 2014). In other words, the consistency that the authors interpret as an absence of BGS effects likely instead reflects widespread BGS effects.

Under their assumption of strict neutrality, the authors estimated a bottleneck followed by growth using PSMC, which they note to be a similar pattern to that observed in other cultivated plants. However, as demonstrated by Johri et al. (2021), even a constant population size model in the presence of BGS alone may generate this pattern; namely, that of an apparent ancestral size reduction, followed by population growth. Given that such demographic analyses using PSMC generally neglect BGS effects, this presents one possible and troubling interpretation of the observation that the resulting PSMC curves (as shown in Figure 1 of Zhong et al. 2022) tend to take a characteristic shape regardless of the species or population being analyzed. Specifically, other studies that have similarly neglected BGS effects have estimated a similar size-change history (e.g., in Yorubans, passenger pigeons, vervets, rice, grapevines, elephants, and Arabidopsis; as discussed in Johri et al. 2021).

While it is certainly reasonable to hypothesize the occurrence of one or multiple bottlenecks during soybean domestication, these results highlight that this demographic inference is strongly confounded by BGS when unaccounted for. Moreover, the findings of Johri et al. (2021) are but part of a growing list of problematic issues described for this demographic inference scheme—ranging from the inference of false size change in the presence of hidden population structure, to falsely inferred bottlenecks prior to population growth (e.g., Chikhi et al. 2018). Hence, their demographic model is likely mis-fit to the data owing (at least) to the neglect of purifying and BGS effects, meaning that the null expectation that they subsequently utilize when searching for selection is likely incorrect. In order to properly account for this oversight, it would instead be necessary to jointly infer the demographic history of this population together with the distribution of fitness effects charactering the contributions of purifying and BGS (e.g., as reviewed in Johri et al. 2022c). As noted above, this joint inference is particularly essential in this species given the expected strong genome-wide overlap in demographic and selective effects owing to selfing.

[1]School of Life Science, Arizona State University, Tempe, AZ, USA. Associate editor: Sara Goodacre. ✉email: jeffrey.d.jensen@asu.edu

Received: 19 July 2022 Revised: 27 October 2022 Accepted: 28 October 2022
Published online: 24 February 2023

INTERPRETING EMPIRICAL OUTLIERS WHEN SEARCHING FOR SELECTIVE SWEEPS

The second concern relates to the test statistics chosen, and to the exclusive use of an empirical outlier approach for determining significance. While their interpretation of empirical distributions is severely compromised by the likely mis-characterization of the demographic model noted above, there is in fact a more fundamental issue. Specifically, though scanning genomes for the types of patterns associated with selective sweeps initially described in the seminal work of Maynard Smith and Haigh (1974) (incorrectly cited in Zhong et al. 2022 as "Smith & Haigh 1974") is common in many fields, utilizing a simple empirical outlier approach to do so (as is also common) is highly inappropriate. That is, arbitrarily assigning importance to the 5% tails of an empirical distribution will identify 5% of loci as being putatively swept regardless of the true underlying fraction. Notably, this empirical outlier approach has been specifically advocated for by certain authors as an alternative to careful model fitting (e.g., Garud et al. 2021, though see the response of Johri et al. 2022a).

However, only by fitting a baseline model consisting of the underlying details of the population (demographic history, purifying selection and BGS effects, mutation rate variation, selfing, and so on) may one examine the statistical power to accurately differentiate and quantify hard and soft selective sweeps. In this way, it becomes possible to address: (a) whether swept loci are expected to even reside in the tails of the empirical distributions for the chosen summary statistics given the evolutionary details of the population (which is far from a given, particularly in the presence of population bottlenecks; Thornton and Jensen 2007), (b) whether the observed empirical outliers are of an unexpected severity given an appropriate baseline model, and (c) whether the baseline model with the addition of selective sweeps represents a significantly improved fit/likelihood relative to the baseline model alone (Johri et al. 2022b).

Turning to the specific outlier analyses performed by Zhong et al. as an example, the authors utilized five summary statistics. Following from their Methods section, the following criteria determined their list of sweep candidate genes: (a) the 5% of loci with the lowest $\pi_{cultivated}/\pi_{wild}$ values (identifying 2696 genes), (b) the 5% of loci with the most negative Tajima's D values in the cultivated but not wild populations (identifying 2697 genes), (c) loci with more than three SNPs that fell in the upper 2.5% tail of the F_{ST} distribution (identifying 5101 and 5250 genes for landraces and improved cultivars, and 1856 and 1833 genes when conditioning only on those with nonsynonymous SNPs, respectively), (d) loci located in the upper 5% tail of the H12 distribution (identifying 2698 and 2700 genes, respectively), and (e) loci containing haplotypes located in the 5% upper tail as assessed by EHH (identifying 6347 and 5542 genes, respectively).

Simply taking the sum of these expectations in cultivars suggests that the authors could identify a substantial fraction of genes as being potentially swept even in the complete absence of selective sweeps, given the defined outlier criteria. The authors indeed report that 33% of genes were identified by at least one test, with almost no overlap amongst all test statistics (0.39% of genes; and see their Figure S2). It is also important to add in this context that these statistics are not independent, but rather are correlated in complex ways that require description under the appropriate baseline model (i.e., π will by definition have an important relationship with Tajima's D and F_{ST}, H12 and EHH capture overlapping haplotype patterns, and so on). Because of these correlations, one would anticipate much more overlap between particular pairs of statistics than would be expected under independence, even under neutral models.

Moreover, the statistics chosen by the authors to identify putatively swept loci are themselves imprecise. For example, statistics such as Tajima's D are tests of the standard neutral model, not of a selective sweep model. As such, deviations in the statistic may owe to a wide variety of non-selective factors (Tajima 1989; Jensen 2009; Charlesworth and Jensen 2023). Furthermore, the H12 statistic has been shown to have poor power to differentiate neutrality from positive selection under a variety of demographic histories (Harris et al. 2018). Thus, for the reasons here discussed, explicitly quantifying the fit of the data to selective sweep expectations within the context of a population-specific baseline model is the more fruitful strategy (Johri et al. 2020). In this way, one importantly also allows for the possibility that no loci may be found to be uniquely consistent with a selective sweep—an outcome intrinsically excluded in empirical outlier approaches of this sort. However, despite these shortcomings, such approaches remain unfortunately common in the genomics literature.

REFERENCES

Barrett SCH, Arunkumar R, Wright SI (2014) The demography and population genomics of evolutionary transitions to self-fertilization in plants. Philos Trans R Soc B 369:20130344

Charlesworth B, Jensen JD (2021) Effects of selection at linked sites on patterns of genetic variability. Annu Rev Ecol Evol Syst 52:177–197

Charlesworth B, Jensen JD (2023) Population genetic considerations regarding evidence for biased mutation rates in *Arabidopsis thaliana*. Mol Biol Evol 40:msac275

Chikhi L, Rodriguez W, Grusea S, Santos P, Boitard S, Mazet O (2018) The IICR (inverse instantaneous coalescence rate) as a summary of genomic diversity: insights into demographic inference and model choice. Heredity 120:13–24

Garud N, Messer P, Petrov D (2021) Detection of hard and soft selective sweeps from *Drosophila melanogaster* population genomic data. PLoS Genet 17:e1009373

Harris RB, Sackman A, Jensen JD (2018) On the unfounded enthusiasm for soft selective sweeps II: examining recent evidence from humans, flies, and viruses. PLoS Genet 14:e1007859

Jensen JD (2009) On reconciling single and recurrent hitchhiking models. Gen Biol Evol 1:320–324

Johri P, Aquadro CF, Beaumont M, Charlesworth B, Excoffier L, Eyre-Walker A et al. (2022b) Recommendations for improving statistical inference in population genomics. PLoS Biol 20:e3001669

Johri P, Charlesworth B, Jensen JD (2020) Toward an evolutionarily appropriate null model: jointly inferring demography and purifying selection. Genetics 215:173–192

Johri P, Eyre-Walker A, Gutenkunst R, Lohmuller K, Jensen JD (2022c) On the prospect of achieving accurate joint estimation of selection with population history. Gen Biol Evol 14:evac088

Johri P, Riall K, Becher H, Excoffier L, Charlesworth B, Jensen JD (2021) The impact of purifying and background selection on the inference of population history: problems and prospects. Mol Biol Evol 38:2986–3003

Johri P, Stephan W, Jensen JD (2022a) Soft selective sweeps: addressing new definitions, evaluating competing models, and interpreting empirical outliers. PLoS Genet 18:e1010022

Maynard Smith J, Haigh J (1974) The hitch-hiking effect of a favourable gene. Genet Res 23:23–35

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585–595

Thornton KR, Jensen JD (2007) Controlling the false-positive rate in multilocus genome scans for selection. Genetics 175:737–750

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

ACKNOWLEDGEMENTS

This work was supported by the National Institutes of Health grant R35GM139383.

AUTHOR CONTRIBUTIONS

JDJ wrote the manuscript.

COMPETING INTERESTS

The author declares no competing interests.

ADDITIONAL INFORMATION

Correspondence and requests for materials should be addressed to Jeffrey D. Jensen.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.