# On the well-founded enthusiasm for soft sweeps in humans: a reply to Harris, Sackman, and Jensen

Daniel R. Schrider[1] and Andrew D. Kern[2]

[1]Department of Genetics, University of North Carolina, Chapel Hill; email: drs@unc.edu

[2]Institute of Ecology and Evolution, University of Oregon; email: adkern@uoregon.edu

**ABSTRACT**

A challenging but central question in population genetics is the detection of genomic regions underpinning recent adaptation. To this end, we recently devised a machine learning method, termed S/HIC, which detects both "hard" selective sweeps on *de novo* mutations and "soft" sweeps on standing genetic variation with high sensitivity and specificity, while being exceptionally robust to demographic model misspecification. We previously applied S/HIC to human population genomic data and uncovered evidence of a large number of recent selective sweeps across the genome, most of which we classified as soft sweeps. A critique of recent efforts to detect soft sweeps, including our own, has made the argument that S/HIC is in fact so vulnerable to demographic misspecification that our analyses with it should be completely discounted. Through a careful consideration of the claims of this critique, we argue that the impact of such misspecification on our analysis in humans is minimal with respect to our conclusions. The critique in question also argued that our false discovery rate in humans was essentially 100%; however we show that this inaccurate claim is due to a regrettable error on the part of its authors. We argue that our scan for selection has produced several interesting observations on recent adaptation in humans that are highly concordant with independent efforts to detect signatures of more ancient positive selection. We conclude that the evidence for the utility of S/HIC, and the validity of our application of it to human data, is highly compelling, and that strictly demographic explanations for our results are clearly untenable.

A recent preprint by Harris et al. (2018) levels a series of critiques at the work done by us as well as other groups on finding soft selective sweeps in genomes. While we will leave the defense of our colleagues' work to them, we here respond to the criticisms raised by Harris et al. of the work presented in Schrider and Kern (2016) and Schrider and Kern (2017). In the first of these two papers we describe a supervised machine learning approach for detecting both hard and soft selective sweeps, termed S/HIC. In the second, we apply S/HIC to six human population samples where we classify the majority of sweeps as soft. Harris et al. argue that 1) the results from our scan for selection in humans are essentially all false positives, and 2) our scan does not accurately discriminate between hard and soft selective sweeps. In short, Harris et al.'s criticisms of our work fall into one of two categories: statistical errors on the part of those authors, or highlighting issues related to model misspecification that loom generally for population genomic inference, and whose impact on S/HIC we have characterized previously. We address the specifics of each of their claims below.

By way of background, detecting selective sweeps is an exceedingly difficult problem, largely due to the propensity of non-equilibrium demographic processes to mimic the effects of selection (Simonsen *et al.* 1995; Jensen *et al.* 2005). Indeed, it was our awareness of the need for a method that is both sensitive to selective sweeps and also far more robust to non-equilibrium demography than previously existing methods that motivated us to develop S/HIC in the first place.

Harris et al. argue that Schrider and Kern (2017) present results that are nearly all false positives. Their main assertion is summarized in Table 1 of Harris et al. in which they present an analysis of our false discovery rate (FDR) to suggest that the number of "significant tests" (i.e. genomic windows classified as sweeps) could be completely explained by the false positive rate that we reported. Unfortunately, Harris et al. have used the wrong numbers. Harris et al have apparently used data from Table 1 of Schrider and Kern (2017), even though this table does not give the raw number of windows classified as sweeps. As we explain in the methods section of the paper, those numbers instead represented the "merged" set of window classifications, such that physically neighboring sweep classifications were collapsed to a single call in an effort to arrive

at a more conservative estimate of the number of distinct sweep candidate regions. The total uncollapsed number of windows classified as sweeps is roughly 35% higher than what we report in our merged set of sweep candidate regions. Indeed the total number of windows that we classified as hard and soft sweeps in each population was given in the supplementary table S2 of our original paper and even plotted by Harris et al in their Fig S2. For clarity, we show each population sample's correct false positive rate, total number of sweep classifications, FDR-corrected number of sweep classifications, and number of windows expected to be misclassified as sweeps given our FDR in Table 1 in this paper. This table contains the total number of windows that we examined in our analysis (i.e. the total number of "tests"). In Schrider and Kern (2017) we performed downstream analysis on a set of windows that were filtered of low-recombining regions; an equivalent table for those windows is shown in Table 2. For the interested reader, BED files listing each window classification have been and still are available in the S/HIC GitHub repository (https://github.com/kern-lab/shIC/tree/master/humanScanResults).

Because Harris et al. used the number of merged sweep candidate regions instead of the total number of windows classified as sweeps, the denominator in their FDR calculation is underestimated, leading them to overestimate our FDR. Rather than the 100% FDR as reported by Harris et al., the correct FDR averaging across each population is approximately 65% for both hard and soft sweeps taken together or 67% for soft sweeps alone. This means that even after one accounts for the high FDR inherent with our classifier, hundreds of significant true positive sweep windows are detected (~500 in each population on average), with the overwhelming majority of them classified as soft sweeps. Harris et al. also claim that the posterior probability estimates produced by S/HIC imply very few high-confidence sweep candidates. However, these estimates are often miscalibrated unless a very large number of training examples are used. As we discuss in Schrider et al. (2018), these estimates are useful for ranking candidates, but false positive rates assessed on an independent test set (as we have done for the FDR estimate above) are far more reliable.

Harris et al. also make the case for increasing posterior probability thresholds in such a way as to strictly control the false positive rate (they plot our analysis of this in their Fig S2). There is of

course a well-known tradeoff between type 1 and type 2 error rates, so one sacrifices the false negative rate for an extremely low false positive rate. For experimentalists, who might invest much time and money in functionally characterizing an individual sweep candidate, such an emphasis on reduced FDRs would make sense. However in population genomics we are often more interested in describing broader-scale patterns of evolution. Thus in Schrider and Kern (2017), we chose to cast a wide net in an attempt to capture as many sweeps as possible, while cognizant of the fact that any individual candidate has a high probability of being a false positive (though substantially lower than Harris et al. argue).

While the erroneously inflated FDR estimate was a fundamental misunderstanding in Harris et al.'s criticism of our work, their piece also reiterates our own concerns about model misspecification. This is a general problem for population genetic inference and one that we take quite seriously. Indeed these were the reasons we presented a scenario of catastrophic model misspecification of the underlying population model in the original S/HIC paper (i.e. Figures 6, 7, S9, S10, S11, and S12 of Schrider and Kern 2016). Specifically, we presented results of a classifier trained on equilibrium population simulations and then tested on data simulated under the exponential growth dynamics estimated from African and non-African human populations, as well as simpler bottleneck scenarios. Harris et al. reiterate our original point—that completely ignoring population size changes may cause a large fraction of hard sweeps to be misclassified as soft. Importantly, as they point out and as we showed in Schrider and Kern (2016), in this case accuracy to discriminate between sweeps and unselected regions of the genome, the primary goal of any sweep scan, is relatively unaffected. However, such extreme model misspecification can reduce accuracy in discriminating between hard and soft sweeps. Thus under an unrealistically pessimistic scenario of demographic misspecification, S/HIC suffers no drop in its ability to distinguish between selected and unselected regions of the genome, although misclassification rates between selected classes increases. This was all clearly shown in Schrider and Kern (2016).

As an aside, the authors point out that our detailed results of applying a S/HIC classifier trained under equilibrium demography but to data simulated under a demographic model estimated by Tennessen et al. (2012) from African exome data is relegated to a supplementary figure that is not described in detail in the main text. The authors neglect to mention that we instead

prominently feature results from an even more challenging scenario wherein the equilibrium-trained S/HIC classifier is applied to the European model from Tennessen et al. In this latter scenario ~50% hard sweeps may be misclassified as soft.

So to what extent might misclassification caused by demographic misspecification affect the conclusion of Schrider and Kern (2017) that soft sweeps are the predominant mode of adaptation in six human subpopulations? If 50% of hard sweeps are misclassified as soft, as under the overly pessimistic scenario of demographic misspecification described above, then we expect 48 soft sweep windows on average in each population to actually represent hard sweeps. Given that after correcting for our false discovery rates we expect on the order of 500 windows to be classified as soft sweeps on average across the six population samples, our conclusion that the vast majority of selective sweeps in humans are soft is unaffected.

Harris et al. also explore misspecification of the distribution of selection coefficients used in sweep simulations. We originally simulated moderate to strong selection coefficients ($s$ ranging from 0.005 to 0.1) as such sweeps are more likely to reach fixation. However, we cannot rule out the possibility that some completed selective sweeps have lower $s$ in practice. While that is so, estimates of selection coefficients in well characterized completed selective sweeps are fairly high (e.g. ranging from 0.023 to 0.14 in the loci examined by Peter *et al.* 2012), but this is probably a biased subset of sweeps and more work in estimating selection coefficients is needed. Under a scenario of misspecification where S/HIC is applied to data with weaker selection than it was trained upon, Harris et al. report a misclassification rate of hard sweeps to soft sweeps at 47%. We show a similar result in Figure S5 from Schrider and Kern (2016)—when using training data encompassing a very wide range of selection coefficients but applying the classifier to test data with weak selection, S/HIC accurately detects sweeps, though the mode of selection may be misinferred. Harris et al. have simply confirmed this result. Again, given the numbers of hard and soft sweeps that we report in Schrider and Kern (2017), this would mean that 1) there are significantly more hard sweeps that we estimate and 2) that there would still be many more soft sweeps than hard. So while such misspecification is important, and well worth being wary of, it does not affect the qualitative conclusions of Schrider and Kern (2017).

One type of model misspecification that Harris et al. examine that we did not address previously is that of population structure. As one might expect, the authors find that cryptic population subdivision or unmodeled migration can reduce accuracy. This is an important analysis, and population structure may contribute to our reduced accuracy in the admixed PEL population of the 1000 Genomes Project relative to the other population samples we examined (supplementary fig. S1 Schrider and Kern 2017). It is important to note however that there is no reason to believe that well-studied population samples such as CEU or JPT harbor sufficient cryptic population structure or have received enough recent ancestry from other, genetically distant populations so as to bias our results. Indeed, all five populations we examined besides PEL appear to be far more homogenous (Auton *et al.* 2015). Given that our results demonstrate that soft sweeps are more frequent in every population, this form of misspecification seems unlikely to have affected our conclusions. In summary, neither of Harris et al.'s primary claims—that our detected sweeps candidates are false positives and that our inferences about the mode of selection are incorrect—are consistent with a careful examination of our results.

Beyond defending our work from the claims of Harris et al. on the basis of S/HIC itself, we would like to highlight our biological results that point to coherent, interpretable clusters of loci, often echoing what has been reported in previous (and orthogonal) selection scans of the human genome. In Schrider and Kern (2017) we reported functional enrichments of sweep loci—soft sweeps in particular—through the use of a careful permutation procedure. Our permutation test accounts for: 1) the number and spatial arrangement of the genomic windows tests, 2) the autocorrelation of S/HIC's classifications, 3) potential biases in the annotation (e.g. gene lengths), and 4) the number of terms tested (using an FDR correction). This analysis revealed numerous enrichments including sperm-egg recognition, spermatogenesis, cancer mutations, and virus-interacting proteins. Each of these annotations has been implicated previously as having experienced recurrent positive selection in the human genome on the basis of divergence (i.e. $d_N/d_S$) or McDonald-Kreitman tests (Dorus *et al.* 2004; Nielsen *et al.* 2005; Kosiol *et al.* 2008; Enard *et al.* 2016). In addition, we found strong evidence for enrichments of sweeps in interacting genes, a result that also has been reported previously using different signatures of selection (Qian *et al.* 2015). That our results should align so well with results based on McDonald-Kreitman tests or $d_N/d_S$ ratios but in reality be false positives seems extremely

unlikely. Moreover, it is difficult to imagine a scenario in which demographic misspecification would result in a large excess of false positives from particular sets of genes, such as immune related proteins that are known to evolve under positive selection in numerous taxa. Thus, Harris et al.'s claim that our scan for positive selection in humans produced no signal of sweeps is not only a misinterpretation of our results, it does not seem consistent with a more detailed inspection of the data.

In conclusion, we are well aware that S/HIC will not be completely immune to all forms of model misspecification, and our previous work is careful to point this out. As we (Schrider and Kern 2016) and Harris et al. (2018) have shown, it is possible to devise scenarios of model misspecification that can impact the accuracy of S/HIC (and probably any method), sometimes severely. It is paramount to evaluate methods under a wide variety of selective and demographic scenarios in order to better inform users of these potential pitfalls and we welcome Harris et al.'s efforts in this respect. However, the mere existence of such confounding scenarios should not be taken as evidence that efforts to detect positive selection are futile. Rather, a more fruitful question to ask is whether, in spite of the presence of potential confounders, selection scans are informative about evolutionary processes in practice. In our case, we can ask whether the application of S/HIC to human data sets (Schrider and Kern 2017) has proved illuminating about the landscape of recent positive selection across the genome. As we argue above, the answer to this question is a resounding "yes".

**ACKNOWLEDGEMENTS**

We thank Peter Ralph, Matt Hahn, Jeff Adrion, CJ Battey, Bernard Kim, and Murillo Rodrigues for comments and suggestions that improved this paper.

## REFERENCES


Auton, A., L. D. Brooks, R. M. Durbin, E. P. Garrison, H. M. Kang *et al.*, 2015 A global reference for human genetic variation. *Nature* **526:** 68-74.

Dorus, S., E. J. Vallender, P. D. Evans, J. R. Anderson, S. L. Gilbert *et al.*, 2004 Accelerated evolution of nervous system genes in the origin of Homo sapiens. *Cell* **119:** 1027-1040.

Enard, D., L. Cai, C. Gwennap and D. A. Petrov, 2016 Viruses are a dominant driver of protein adaptation in mammals. *eLife* **5:** e12469.

Harris, R., A. Sackman and J. D. Jensen, 2018 On the unfounded enthusiasm for soft selective sweeps II: examining recent evidence from humans, flies, and viruses. *bioRxiv***:** 443051.

Jensen, J. D., Y. Kim, V. B. DuMont, C. F. Aquadro and C. D. Bustamante, 2005 Distinguishing between selective sweeps and demography using DNA polymorphism data. *Genetics* **170:** 1401-1410.

Kosiol, C., T. Vinař, R. R. da Fonseca, M. J. Hubisz, C. D. Bustamante *et al.*, 2008 Patterns of positive selection in six mammalian genomes. *PLoS Genet.* **4:** e1000144.

Nielsen, R., C. Bustamante, A. G. Clark, S. Glanowski, T. B. Sackton *et al.*, 2005 A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol.* **3:** e170.

Peter, B. M., E. Huerta-Sanchez and R. Nielsen, 2012 Distinguishing between selective sweeps from standing variation and from a *de novo* mutation. *PLoS Genet.* **8:** e1003011.

Qian, W., H. Zhou and K. Tang, 2015 Recent Coselection in Human Populations Revealed by Protein–Protein Interaction Network. *Genome Biol. Evol.* **7:** 136-153.

Schrider, D., J. Ayroles, D. R. Matute and A. D. Kern, 2018 Supervised machine learning reveals introgressed loci in the genomes of *Drosophila simulans* and *D. sechellia*. *PLoS Genet.* **14:** e1007341.

Schrider, D. R., and A. D. Kern, 2016 S/HIC: Robust Identification of Soft and Hard Sweeps Using Machine Learning. *PLoS Genet.* **12:** e1005928.

Schrider, D. R., and A. D. Kern, 2017 Soft sweeps are the dominant mode of adaptation in the human genome. *Mol. Biol. Evol.* **34:** 1863–1877.

Simonsen, K. L., G. A. Churchill and C. F. Aquadro, 1995 Properties of statistical tests of neutrality for DNA polymorphism data. *Genetics* **141:** 413-429.

Tennessen, J. A., A. W. Bigham, T. D. O'Connor, W. Fu, E. E. Kenny *et al.*, 2012 Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337:** 64-69.

**Table 1. False positive rates, total number of sweep classifications, and expected number of true sweep windows from Schrider and Kern (2017).**

| | Proportion of neutral windows expected to be mis-classified as sweeps | | Number of genome-wide sweep windows reported by Schrider and Kern (2017) | | FDR-corrected number of true-positive sweep windows | | Number of windows expected to be mis-classified as sweeps | |
|---|---|---|---|---|---|---|---|---|
| Population | Soft | Hard | Soft | Hard | Soft | Hard | Soft | Hard |
| CEU | 0.066 | 0.001 | 1579 | 110 | 518.7 | 93.9 | 1060.3 | 16.1 |
| GWD | 0.041 | 0 | 1254 | 15 | 547.1 | 15.0 | 706.9 | 0 |
| JPT | 0.074 | 0 | 1643 | 123 | 454.2 | 123.0 | 1188.8 | 0 |
| YRI | 0.044 | 0 | 1284 | 29 | 577.1 | 29.0 | 706.8 | 0 |
| PEL | 0.062 | 0.002 | 1016 | 65 | 20.0 | 32.9 | 996.0 | 32.1 |
| LWK | 0.045 | 0.001 | 1286 | 11 | 563.1 | 0* | 722.9 | 16.1 |

*rounded up from -5.0.

**Table 2. False positive rates, total number of sweep classifications, and expected number of true sweep windows from Schrider and Kern (2017), after filtering regions with low recombination rates.**

| | Proportion of neutral windows expected to be mis-classified as sweeps | | Number of genome-wide sweep windows reported by Schrider and Kern (2017) | | FDR-corrected number of true-positive sweep windows | | Number of windows expected to be mis-classified as sweeps | |
|---|---|---|---|---|---|---|---|---|
| Population | Soft | Hard | Soft | Hard | Soft | Hard | Soft | Hard |
| CEU | 0.066 | 0.001 | 1207 | 81 | 285.2 | 67.0 | 921.8 | 14.0 |
| GWD | 0.041 | 0 | 938 | 5 | 323.4 | 5.0 | 614.6 | 0 |
| JPT | 0.074 | 0 | 1282 | 71 | 248.4 | 71.0 | 1033.5 | 0 |
| YRI | 0.044 | 0 | 964 | 14 | 349.4 | 14.0 | 614.6 | 0 |
| PEL | 0.062 | 0.002 | 780 | 37 | 0* | 9.0 | 865.9 | 28.0 |
| LWK | 0.045 | 0.001 | 964 | 3 | 335.4 | 0** | 628.5 | 14.0 |

*rounded up from -86.0.
**rounded up from -11.0.