Some complexities in interpreting apparent effects of hitchhiking: a commentary on Gompert, Feder & Nosil (2021)

Brian Charlesworth[1] and Jeffrey Jensen[2]

[1]University of Edinburgh
[2]Arizona State University

January 4, 2022

Abstract

We write to address recent claims by Gompert et al. (2021) about the potentially important and underappreciated phenomena of "indirect selection", the observation that neutral regions may be affected by natural selection. We argue both that this phenomenon – generally known as genetic hitchhiking – is neither new nor poorly studied, and that the patterns described by the authors have multiple alternative explanations.

Some complexities in interpreting apparent effects of hitchhiking: a commentary on Gompert, Feder & Nosil (2021)

Brian Charlesworth[1] and Jeffrey D. Jensen[2]

[1] Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, Edinburgh, UK

[2] School of Life Sciences, Arizona State University, Tempe AZ, USA

Corresponding author: Brian Charlesworth

Email: Brian.Charlesworth@ed.ac.uk

Keywords: epistatic selection, genetic hitchhiking, linkage disequilibrium, selective sweeps, demography

ABSTRACT

We write to address recent claims by Gompert *et al.* (2021) about the potentially important and underappreciated phenomena of "indirect selection", the observation that neutral regions may be affected by natural selection. We argue both that this phenomenon – generally known as genetic hitchhiking – is neither new nor poorly studied, and that the patterns described by the authors have multiple alternative explanations.

We wish to express a number of concerns about the recent paper by Gompert *et al* . (2021), who propose a "widespread indirect selection hypothesis", and assert that "functionally neutral genetic regions can be affected indirectly by natural selection, via their statistical association with genes under direct selection". They add that "the genomic extent of such indirect selection, particularly across loci not physically linked to those under direct selection, remains poorly understood". The authors present analyses of several datasets; they suggest that this process could be an important phenomenon over both short and long timescales, and may serve to make "aspects of evolution more predictable" given that "conditional on patterns of LD, indirect selection is a deterministic process." We have concerns about both their conceptual framework, and their interpretation of the experimental results.

First, the concept of "indirect selection" appears to be equivalent to genetic hitchhiking, which is only mentioned towards the end of their paper. The term hitchhiking is most often used in relation to the effects

of the spread of a beneficial mutation on variability at linked sites (a selective sweep: Maynard Smith & Haigh 1974), or to the similar effects of the elimination of deleterious mutations (background selection: Charlesworth *et al* . 1993; incorrectly attributed by Gompert *et al* . to Begun & Aquadro 1992). As noted in the recent review by Charlesworth & Jensen (2021), the various forms of hitchhiking can all be described in general terms by the Price-Robertson identity, which states that the change in mean of a trait is governed by the additive covariance between the trait and fitness (Robertson 1968; Price 1970). In this case, the trait is the allele frequency at the neutral locus, and the covariance is the product of the coefficient of linkage disequilibrium, D , and the additive fitness effect of the locus under selection (Santiago & Caballero 1995; Charlesworth & Jensen 2021). We note that variation maintained by long-term balancing selection or divergent selection between populations can also affect variability at linked sites, but this does not involve changes in allele frequencies that are noticeable over the short-timescales considered by Gompert *et al.* (2021), except in the initial phase of their establishment (see Zeng *et al* . 2021 for a recent theoretical analysis).

Thus, in one sense their remark about the deterministic nature of indirect selection is correct. However, the initial value of D is generated by various forms of chance associations between alleles at the neutral and selected loci, as is indicated by the title of their paper, so that there is inherently a stochastic element to hitchhiking. Furthermore, D is reduced by a factor of $1 - c$ each generation, where c is the frequency of recombination between the two loci concerned. (Charlesworth & Charlesworth 2010, p.381). Multiple theoretical analyses of hitchhiking have shown that significant effects of selection on linked neutral sites only occur if the ratio of $c$to the selection coefficient at the selected loci is small, or the neutral loci are embedded within a large number of loci that are simultaneously experiencing selection (Charlesworth & Jensen 2021). A neutral locus can only be affected by a selective sweep at an unlinked locus ($c = \frac{1}{2}$) if there is extremely strong selection, involving a much greater than two-fold selective advantage to the beneficial allele. Similarly, background selection caused by unlinked loci is likely to have only a minor effect on neutral variability (Charlesworth 2012). Detectable hitchhiking effects involving unlinked loci are therefore likely to be very infrequent.

Second, their assertion that widespread hitchhiking effects are "relatively untested" seems misplaced. One of the most consistent patterns to emerge from the study of DNA sequence variation over the past few decades is the generally positive correlation between neutral variability and the local rate of recombination (Begun & Aquadro 1992; Charlesworth & Jensen 2021). Genetic hitchhiking has long been discussed as a major factor in generating this genome-wide correlation, although in some species the mutagenic effect of recombination may contribute as well (Pratto *et al* . 2014; Arbeithuber *et al* . 2015). Indeed, the abundant evidence for pervasive background selection effects has prompted numerous authors to argue that it should be an essential component in any evolutionary null model when conducting population genomic studies (*e.g.,* Comeron 2017; Pouyet *et al* . 2018; Jensen *et al.* 2019).

Third, the authors suggest that, while hitchhiking effects have been studied over long periods of time, "it remains unclear whether indirect selection is pervasive on shorter time scales, such as generations or decades". Although it is not obvious how hitchhiking effects could operate over long time-scales without also acting at short time-scales, this remark reflects a more fundamental misunderstanding. As described above, recombination events between directly selected mutations and linked variants act to diminish hitchhiking effects, so that the footprints associated with recent hitchhiking events are naturally stronger than older ones. In fact, recombination, subsequent mutations, and genetic drift cause these patterns to decay so rapidly that most effects of individual selective sweeps are *only* detectable over relatively brief periods of time since their occurrence (Przeworski 2002; Kim & Stephan 2002). If there has been strong directional selection at certain loci in the experimental set-ups that are described by Gompert *et al.* (2021), the short time-scale involved is certainly favourable for the detection of their hitchhiking effects (although the statistical power associated with the relatively small experimental sample sizes and limited number of generations can be problematic: Barrett *et al.* 2019).

When considering the evidence for hitchhiking effects in the cases described by Gompert *et al* . (2021),

we focus specifically on their primary example - the stick insect *Tisema cristinae* . Previous work has documented the existence of an interesting colour polymorphism, controlled by a 10 megabase region, the *Mel-Stripe*locus, which appears to be potentially associated with a chromosomal inversion (Lindtke *et al.* 2017). Among approximately 7 million SNPS in their sample, Gompert *et al* . (2021) found 64 SNPs that had r^2 [?] 0.1 with *Mel-Stripe* and were also on different chromosomes, where r is the correlation coefficient between pairs of alleles at two loci (Charlesworth & Charlesworth 2010, p.373). The first question is whether this could be generated by the effect of random sampling of the 492 haploid genomes sequenced in their experimental population. Under the null hypothesis of no LD, an r^2 of 0.1 corresponds to a 1 d.f.χ^2 of 49.2, for which $p = 2.31 \text{ x } 10^{-12}$, using the incomplete gamma function with parameter 0.5, which is equivalent to $0.5\chi^2$(*https://keisan.casio.com/exec/system/1180573447*). The expected number of SNPS with r^2 [?] 0.1 is thus approximately $2.31 \text{ x } 10^{-12} \text{ x } 7 \text{ x } 10^6 = 1.62 \text{ x } 10^{-5}$. Not surprisingly, therefore, this explanation can be ruled out. Note that the expected value of r^2 generated by random sampling in the absence of true LD is 1/492 [?] 0.002, which is not far from the mean value of 0.004 for all pairs of unlinked SNPs reported by Gompert *et al* . (2021).

There appear to be at least six possible explanations for this unexpectedly large number of unlinked SNPs that are in fairly strong LD with *Mel-Stripe* :

(1) There exist technical errors that results in spurious cases of LD or incorrect assignment of the locations of SNPs; for example, so that SNPs that are actually in the *Mel-Stripe* region are placed on other chromosomes. The quality control details needed to evaluate this possibility were not presented by the authors.

(2) LD between neutral SNPs and the *Mel-Stripe* locus has been created by random genetic drift in a panmictic population over a long period of time. This seems improbable, as the expected value ofr^2 with no linkage is approximately $1/(2N_e)$ (Charlesworth & Charlesworth 2010, p.383), and the size of the *T. cristinae* population used in the experiment is said be of the order of thousands of individuals (Gompert*et al* . 2021).

(3) There has been very recent admixture from a genetically distinct population or populations, resulting in patches of LD with*Mel-Stripe* (Charlesworth & Charlesworth 2010, p.388). No structure/admixture modelling needed to evaluate this possibility appears to have been performed.

(4) There has been a recent and severe bottleneck in population size, generating random LD of a much higher magnitude than expected under a constant population size (Charlesworth & Charlesworth 2010, p.389), As with point 3, these signatures may be detectable using standard demographic modeling approaches (reviewed in Beichman *et al* . 2018); again, such analyses appear not to have been performed.

(5) The sample used in the experiment has captured an ongoing sweep involving very strong directional selection on *Mel-Stripe* ; this seems implausible, given the evidence that the *Mel-Stripe*variants represent a long-standing balanced polymorphism (Lindtke*et al.* 2017).

(6) There could be epistatic fitness interactions between the SNP loci and *Mel-Stripe* . This of course does not constitute "indirect selection" (*i.e* ., genetic hitchhiking) of the type proposed by Gompert *et al* . (2021), but is perhaps what they have in mind when they refer to "polygenic selection". However, it is stretching credulity that there could be tens of unlinked loci subject to very strong epistatic fitness interactions with *Mel-Stripe* , resulting in significant LD. Theoretical analyses of two-locus balanced polymorphisms have shown that substantial LD requires the measure of additive x additive epistasis for fitness to considerably exceed the recombination frequency (Charlesworth & Charlesworth 2010, pp.420-425).

For example, in the simple symmetric two-locus fitness model of Lewontin & Kojima (1960), LD is maintained at equilibrium with free recombination only if 2 [?] e , where e is the epistatic fitness parameter (fitnesses are here measured relative to the fitness of the double heterozygote). In this case, with equilibrium allele frequencies of 0.5 at both loci, $r^2 = 16D^2 = 1 - 2/e$. Very few convincing cases of LD maintained by epistatic selection among unlinked loci have been described. The classic case is that of the Australian grasshopper*Keyacris scurra* (formerly *Moraba scurra*), involving two inversion polymorphisms that show

a consistent pattern of LD across multiple populations, and substantial deviations from Hardy-Weinberg frequencies that indicate strong viability selection (Turner 1972). Even here, the magnitude of r^2 is small. In the sample that showed the highest value of D (*i.e.,* Royalla B 1958), the data in Table 3 of Turner (1972) give $D = 0.00116$ and$r^2 = 0.00349$.

In summary, the observation of many unlinked SNPs with substantial LD in *T. cristinae* seems to raise more questions than it answers, and cannot be taken as indicating widespread "indirect selection" without much more evidence, including a proper consideration of baseline expectations related to both technical (*e.g* ., data quality) and evolutionary (*e.g* ., population history) factors. Furthermore, even if these alternatives were to be ruled out by further analyses, genetic hitchhiking is neither a new nor poorly studied phenomenon.

REFERENCES

Arbeithuber, B., Betancourt, A.J., Ebner, T., & Tiemann-Boege, I. (2015). Crossovers are associated with mutation and biased gene conversion at recombination hotspots. *Proceedings of the National Academy of Sciences of the USA, 112* , 2109–14.

Barrett, R.D.H., Laurent, S., Mallarino, R., Pfeifer, S.P., Xu, C.C., Foll, M., ... Hoekstra, H.E. (2019). Linking a mutation to survival in wild mice. *Science* , *363* , 499–504.

Begun, D., & Aquadro, C.F. (1992). Levels of naturally occurring DNA polymorphism correlate with recombination rate in *Drosophila melanogaster* . *Nature* , *356* , 519–20.

Beichman, A.C., Huerta-Sanchez, E., & Lohmueller, K.E. (2018). Using genomic data to infer historic population dynamics of non-model organisms. *Annual Review of Ecology, Evolution and Systematics* ,*49* , 433-56.

Charlesworth, B. (2012). The effects of deleterious mutations on evolution at linked sites. Genetics, *190* , 5-22.

Charlesworth, B., & Charlesworth, D. (2010). *Elements of evolutionary genetics.* Greenwood Village, CO: Roberts and Company,

Charlesworth, B. & Jensen, J.D. (2021). Effects of selection at linked sites on patterns of genetic variability. *Annual Review of Ecology, Evolution and Systematics, 52* , 177-97.

Charlesworth, B., Morgan. M.T., & Charlesworth, D. (1993). The effect of deleterious mutations on neutral molecular variation. *Genetics, 134* , 1289–303.

Comeron, J.M. (2017). Background selection as null hypothesis in population genomics: insights and challenges from *Drosophila*studies. *Philosophical Transaction of the Royal Society, Series B, 372* , 20160471.

Gompert, Z., Feder, J.L., & Nosil, P. (2021). Natural selection drives genome-wide evolution via chance genetic associations. *Molecular Ecology,* 00, 1-15. https://doi.org/10.1111/mec.16247.

Jensen, J.D., Payseur, B.A., Stephan, W., Aquadro, C.F., Lynch, M., Aquadro, C.F., ...

Charlesworth, B. (2019). The importance of the Neutral Theory in 1968 and 50 years on: a response to Kern & Hahn 2018. *Evolution* ,*73* , 111-4.

Kim, Y., & Stephan, W. (2002). Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics, 160* , 765–77.

Lewontin, R. C., & Kojima, K.-I. (1960). The evolutionary dynamics of complex polymorphisms. *Evolution, 14* , 458-472.

Lindtke, D., Lucek, K., Soria-Carrasco, V., Villoutreix, R., Farkas, T. E., Riesch, R., . . . Nosil, P. (2017) Long-term balancing selection on chromosomal variants associated with crypsis in a stick insect.*Mol. Ecol., 26* , 6189-6205.

Maynard Smith, J., & Haigh, J. (1974). The hitch-hiking effect of a favourable gene. *Genetical Research, 23* , 23–35.

Pouyet, F., Aeschbacher, S., Thiery, A., & Excoffier, L. (2018). Background selection and biased gene conversion affect more than 95% of the human genome and bias demographic inferences. *eLife, 7* , e36317.

Pratto, F., Brick, K., Khil, P., Smagulova, F., Petukhova, G.V., & Camerini-Otero, R.D. (2014). DNA Recombination. Recombination initiation maps of individual human genomes. *Science, 346* , 1256442.

Price, G. R. (1970). Selection and covariance. *Nature,227* , 520-521.

Przeworski, M. (2002). The signature of positive selection at randomly chosen loci. *Genetics, 160* , 1179–89.

Robertson, A. (1968). The spectrum of genetic variation (ed. Lewontin R.C.), pp. 5-16. Syracuse, NY: Syracuse University Press.

Santiago, E., & Caballero, A. (1995) Effective size of populations under selection. *Genetics, 139* , 1013-1030.

Turner, J. R. G. (1972). Selection and stability in the complex polymorphism of *Moraba scurra* . *Evolution, 26* , 334-343.

Zeng, K., Charlesworth, B., & Hobolth, A. (2021). Studying models of balancing selection using phase-type theory. *Genetics ,218* , iyab055.