LETTER

On Reconciling Single and Recurrent Hitchhiking Models

Jeffrey D. Jensen

Program in Bioinformatics and Integrative Biology, University of Massachusetts Medical School

A major focus of modern population genetics involves using polymorphism data in order to identify regions impacted by recent positive selection (so-called genomic scans). Recently, methodology has been proposed not to identify individual loci, but rather to quantify genomic recurrent hitchhiking (RHH) parameters using this same type of polymorphism data. I here examine to what extent genomic scans for adaptively important loci may be informed by recently estimated RHH parameters (and vice versa). I find that published results are largely incompatible with one another, with approximately an order of magnitude more sweeps being empirically identified than would be predicted under RHH estimates. Results demonstrate that making this connection between SHH and RHH models is crucial for a more complete and accurate characterization of adaptive evolution.

Introduction

One of the most popular approaches for identifying loci recently impacted by positive selection is known as “hitchhiking mapping” (e.g., Harr et al. 2002). Broadly speaking, this approach involves scanning across a large number of regions in order to determine the average levels of variability that are characteristic of the genomic environment. Regions that show extreme values and fall in the tail of this observed empirical distribution are then subject to further investigation via resequencing—with the aim being the discernment of locus-specific adaptive effects from neutral genome-wide patterns of variation (e.g., Harr et al. 2002; Glinka et al. 2003; Tenaillon et al. 2004; Carlson et al. 2005; Haddrill et al. 2005; Nielsen 2005; Ometto et al. 2005; Williamson et al. 2005; Wright et al. 2005; Kelley et al. 2006).

Problematically, major assumptions about the underlying adaptive substitutions responsible for these patterns are made in such attempts to identify selected loci. Namely, as these scans rely on the impact of beneficial mutations upon closely linked neutral variability (i.e., the genetic hitchhiking effect; Maynard Smith and Haigh 1974), it is implicitly assumed that selection is strong enough to impact large genomic regions. Simultaneously, it is assumed that these selective events occur rarely enough that recently impacted regions will indeed uniquely reside in the tails of genomic distributions, and yet frequently enough to be detectable from patterns of variation. This suggests that the assumptions underlying genomic scans may correspond to a very specific parameter space. This disconnect between hitchhiking mapping and the true underlying rates and strengths of beneficial mutations (known as “recurrent hitchhiking” [RHH]) owes to the fact that the former relies upon a model of a single hitchhiking (SHH) event, in which a single adaptive fixation is assumed to have occurred immediately prior to sampling, whereas the latter considers a constant input of beneficial mutations, occurring at a given rate.

The first point of comparison between these two models comes from Wiehe and Stephan (1993), who predicted the expected level of reduction in variation at linked neutral sites under an RHH model, demonstrating that for $\lambda = $ constant (where $\lambda$ is the selection coefficient, and $\lambda$ is the rate of adaptive substitutions per site per generation), the mean reduction is identical among models. This result implies that regions of reduced variation may be consistent with models of rarely occurring but strongly advantageous, or commonly occurring but weakly advantageous, mutations. Recently, attempts have been made to estimate RHH parameters (i.e., $s$ and $\lambda$) directly from the same multilocus and genomic polymorphism data used in genomic scans (e.g., Kim 2006; Li and Stephan 2006; Andolfatto 2007; Macpherson et al. 2007; Jensen, Thornton, and Andolfatto 2008; and recently reviewed by Sella et al. 2009), in order to distinguish between these scenarios. Thus, rather than attempting to identify individual loci, these estimators attempt to quantify the average genomic strength and rate of adaptive evolution. As these recent estimators are fundamentally informed by the same underlying parameters as the hitchhiking mapping approach implemented in genomic scans, I here ask whether published results from both approaches are consistent with one another.

Relating Models of RHH to the Identification of Adaptive Loci

The ability to distinguish between models of weak and strong selection has significant implications for our ability to detect adaptively important regions of the genome. As shown in table 1 for a hypothetical 1-Mb region, the expected number of potentially identifiable sweeps differs strongly between models. For example, a 5% average reduction in variation implies that selection tends to be either weak or infrequent. Thus, strong selection (i.e., $s > 0.01$) would occur so rarely as to never be detectable, on average, from patterns of polymorphism. And although weaker selection occurs with an appreciable frequency, such that it may be detectable when scanning large genomic regions, there are still few sweeps, each resulting in a relatively small genomic impact. As such, any given marker would have an approximately 0.2% chance of falling within a swept region, necessitating an extremely dense screen in order to identify adaptively important loci.

In the other extreme, models positing a 90% reduction in variation are expected to have experienced a large number of recent sweeps at any given time of sampling. As such,
Table 1
Details of Recurrent Sweeps under Four Selection Coefficients and Four Levels of Reduction for a 1-Mb Region

| $s$   | $2N_s$ | Size of Sweep (in bp)$^a$ | $E$ (transit time in 4$N_s$ generation)$^b$ | $E$ (time between sweeps in 4$N_s$ generation) ($E_l$)$^c$ | $E$ (no. of sweeps) (no.)$^d$ | Fraction of Markers Swept (%)$^e$ | $E_l$ | No.$^d$ | %$^e$ | $E_l$ | No.$^d$ | %$^e$ | $E_l$ | No.$^d$ | %$^e$ |
|-------|--------|---------------------------|--------------------------------|--------------------------------|---------------------------|--------------------------------|------|--------|------|------|--------|------|------|--------|------|
| 1 x 10$^{-1}$ | 2 x 10$^5$ | 20,000 | 3.6 x 10$^{-5}$ | 1.1 | 1 | 0 | 0 | 27 x 10$^{-1}$ | 0 | 1.0 x 10$^{-1}$ | 0 | 0.02 | 8.3 x 10$^{-3}$ | 12 | 0.24 |
| 1 x 10$^{-2}$ | 2 x 10$^4$ | 2,000 | 3.6 x 10$^{-4}$ | 1.1 x 10$^{-1}$ | 1 | 0.002 | 2.7 x 10$^{-2}$ | 4 | 0.007 | 1.0 x 10$^{-2}$ | 10 | 0.02 | 8.3 x 10$^{-4}$ | 12-0 | 0.24 |
| 1 x 10$^{-3}$ | 2 x 10$^3$ | 200 | 3.6 x 10$^{-3}$ | 1.1 x 10$^{-2}$ | 8 | 0.002 | 2.7 x 10$^{-3}$ | 36 | 0.007 | 1.0 x 10$^{-3}$ | 100 | 0.02 | 8.3 x 10$^{-5}$ | 1,200 | 0.24 |
| 1 x 10$^{-4}$ | 2 x 10$^2$ | 20 | 3.6 x 10$^{-4}$ | 1.1 x 10$^{-3}$ | 80 | 0.002 | 2.7 x 10$^{-4}$ | 360 | 0.007 | 1.0 x 10$^{-4}$ | 1,000 | 0.02 | 8.3 x 10$^{-6}$ | 12,000 | 0.24 |

$^a$ The size of the region impacted by a given sweep, calculated as 0.01 base pairs (Kaplan et al. 1989), with $s = 5 x 10^{-8}$ per site generation (Charlesworth 1996; Andolfatto and Przeworski 2001).

$^b$ The expected transit time of a beneficial mutation, calculated as $-\log(1/2\gamma)$, in units of 4$N_s$ generations, where $\gamma = 1/2N_s$, $\gamma = 2N_s$, and $N = 10^6$.

$^c$ The expected time between beneficial fixations occurring within the region, calculated as $1/2MA$, in units of 4$N_s$ generations, where $A$ is the expected number of sweeps per recombination unit in the last 4$N_s$ generations and $M$ is the size of the region ($= 1$ Mb).

$^d$ The number of sweeps within the 1-Mb region that are recent enough to be detectable using polymorphism-based statistics, calculated as the average number of sweeps occurring within the last 0.1 4$N_s$ generations (Przeworski 2002); importantly, only a fraction of this number may be identifiable, as power has been shown to rarely exceed 50% for commonly used summary statistics (Przeworski 2002; Jensen, Thornton, and Aquadro 2008).

$^e$ The fraction of randomly placed markers across the 1 Mb under consideration that would fall within sweep regions, calculated by determining the proportion of the total region impacted by a recent sweep (e.g., if eight sweeps, each effecting -200 bp, are expected across the 1-Mb region, then the probability for an individual marker to fall in a sweep region is calculated as 0.001/1,000,000).
As in table 1, it is possible to calculate the expected number of sweeps occurring within these empirically scanned regions for given values of \( s \) and \( \lambda \) (table 2). For example, for the RHH values estimated by Andolfatto (2007), one may expect \( \sim 3,060 \) sweeps of \( s = 0.00001 \) to have occurred within the last 0.1 \( 4N \) generations across the 850-kb region examined by Harr et al. (2002). Despite estimating the same 20% reduction in genetic variation owing to RHH as Andolfatto (2007), Macpherson et al. (2007) estimate a much stronger \( s (=0.01) \), suggesting approximately three detectable sweeps on average across a region of this size. Given their relative strengths, both RHH estimators suggest that approximately 0.7% of markers should be impacted by a recent sweep. Using an SHH-based approach, Harr et al. (2002) identify 7% of their markers as being swept, and the combined scans of Bauer DuMont and Aquadro (2005) and Jensen et al. (2007), as well as Glinka et al. (2003), identify \( \sim 12\% \) of their markers as swept. Thus, the number of putatively swept markers identified empirically using SHH models far exceeds published RHH estimates, with roughly an order of magnitude more sweeps being detected than would be predicted (table 2).

Viewing these results graphically, figure 1 plots the reduction in genomic variation against the corresponding fraction of recently swept genomic regions, for both RHH- and SHH-based estimates. For genomic scan studies (grouped as “SHH model”), the expected reduction in variation is back-calculated based upon the empirically observed fraction of loci swept (i.e., what level of reduction is necessary in order for the identified number of loci to have experienced a sweep within the last 0.1 \( 4N \) generations). Conversely, for the RHH estimators (grouped as “RHH model”), the expected fraction of loci swept is calculated from the estimated reduction in variation (i.e., for the estimated rate, how many sweeps will have occurred within the last 0.1 \( 4N \) generations). The details of both calculations are given in table 2. As shown, RHH estimates as a whole suggest a less substantial reduction in variation, and thus a smaller fraction of swept loci. Interestingly, estimates strongly group by model—despite large differences among the estimators with regards to the type of data used, summary statistics utilized, and statistical framework—suggesting possible systematic biases in estimation under one, or possibly both, SHH model– and RHH model–based approaches. 

**Table 2**

Empirical Genomic Scan Results Compared with Expectations under Estimated RHH Models for *Drosophila*

| Region Size | No. of markers | Fraction Swept | \(-20\%\) Reduction | \(-50\%\) Reduction | \(E (s=1 \times 10^{-5})\) | \(E (s=2 \times 10^{-5})\) | \(E (s=3 \times 10^{-5})\) |
|-------------|----------------|----------------|---------------------|---------------------|-----------------|-----------------|-----------------|
| 256 kb      | 26             | 0.12           | 0.007               | 0.017               | \(-1\)          | \(-68\)         | \(-900\)        |
| 850 kb      | 28             | 0.07           | 0.007               | 0.017               | \(-3\)          | \(-225\)        | \(-3,060\)      |
| 17 Mb       | 105            | 0.12           | 0.007               | 0.017               | \(-61\)         | \(-4,620\)      | \(-61,200\)     |

a Total length of the region spanned by the scan.

b The number of scanned markers used in the study.

c The fraction of scanned markers proposed by the authors to be linked to selective sweeps.

d The expected fraction of markers that would fall in swept regions, for a \( \sim 20\% \) estimated reduction in variability (Andolfatto 2007; Macpherson et al. 2007).

e The expected fraction of markers that would fall in swept regions, for a \( \sim 50\% \) estimated reduction in variability (Jensen, Thornton, and Andolfatto 2008; Li and Stephan 2006).

f The expected number of sweeps that would fall in the sequenced regions within the last 0.1 \( 4N \) generations, for parameters estimated by Macpherson et al. (2007) for *Drosophila simulans*.

g The expected number of sweeps that would fall in the sequenced regions within the last 0.1 \( 4N \) generations, for parameters estimated by Li and Stephan (2006) for *Drosophila melanogaster*.

h The expected number of sweeps that would fall in the sequenced regions within the last 0.1 \( 4N \) generations, for parameters estimated by Andolfatto (2007) for *Drosophila melanogaster*.

i Only a fraction of this expected number may be identifiable, owing to the imperfect power of existing test statistics—see figure 2 (Przeworski 2002; Jensen, Thornton, and Aquadro 2008).

j The expected number of sweeps that would fall in the sequenced regions within the last 0.1 \( 4N \) generations, for parameters estimated by Jensen, Thornton, and Andolfatto (2008) for *Drosophila melanogaster*.

k The expected number of sweeps that would fall in the sequenced regions within the last 0.1 \( 4N \) generations, for parameters estimated by Harr et al. (2002) for an X-linked region of *Drosophila melanogaster*.

l From Glinka et al. (2003) for an X-linked region of *Drosophila melanogaster*.

m From Bauer DuMont and Aquadro (2005); Jensen et al. (2007) for an X-linked region of *Drosophila melanogaster*.
Evaluating Possible Explanations for the Observed SHH–RHH Discrepancy

One possible explanation for the discrepancy in SHH model– and RHH model–based analyses is that the true reduction in variation due to hitchhiking in *D. melanogaster* may be much more severe—a genomic reduction in variation of \(~79\%\) is necessary in order to accommodate the number of empirically identified sweep regions, compared with the maximum published RHH estimate of \(~50\%\)—and thus that existing RHH estimators are greatly underestimating the rate of adaptive evolution. Alternatively, the majority of the loci identified in genomic scans may be false positives. Recent studies have suggested that both demographic perturbations (e.g., Nielsen 2001; Przeworski 2002; Jensen et al. 2005; Nielsen et al. 2007) and ascertainment biases (Teshima et al. 2006; Thornton and Jensen 2007) likely contribute to a high rate of false inferences of selection in genomic scans. Along with this, it is additionally important to note that the expected number of sweeps in these calculations is not tantamount to the expected number of “identifiable” sweeps, as test statistics do not have perfect power. For example, examining the performance of three of the most common summary statistics (Tajima 1989), \(H\) [Fay and Wu 2000], and the composite likelihood ratio test [Kim and Stephan 2002] across a wide range of RHH parameters, Jensen, Thornton, and Aquadro (2008) found power to be less than 20% for RHH models of weak selection, and rarely in excess of 50% even under models of strong selection. As shown in figure 2, these factors may actually predict a pattern that is opposite to that which is observed—even if demography is properly modeled, fewer sweeps should be identified than have occurred, owing to this imperfect power. Thus, empirical observations appear more consistent with the scenario in which there is a large false-positive rate associated with genomic scans for selection, consistent with previous results (Teshima et al. 2006; Thornton and Jensen 2007).

Other possibilities exist as well. The impact of violations of both a constant-rate assumption on both SHH- and RHH-based approaches, and particularly systematic increases or decreases in the rate of adaptation, as well as the assumption that selection is largely acting only on new mutations (as opposed to segregating variation), remain as areas in need of further investigation. Additionally, under RHH models in which variation is strongly reduced, the approximations of Kaplan et al. (1989) and Stephan et al. (1992) are violated, owing to overlapping sweep patterns (Przeworski 2002). The impact of such a model on both SHH- and RHH-based estimation remains to be seen.

Conclusions

Comparison of a number of published studies in *D. melanogaster* suggests a lack of correspondence between SHH model– and RHH model–based analyses. Specifically, genomic scan results imply a much higher rate of adaptation, and thus a far greater level of reduction in genomic variation (\(~79\%\) reduction, whereas the mean RHH estimate \(~35\%)\). Given the significant differences among RHH estimators particularly, this result may suggest systematic biases associated with the methodologies themselves.

Although simulation results are suggestive of possible biases that may be inflating the number of loci identified in genomic scans, better disentangling these discrepancies has major implications. As RHH parameter estimates continue to come in to focus for natural populations of interest, it may become evident that searching for specific adaptive loci may be a difficult endeavor, owing to long expected waiting times between adaptive fixations. Alternatively, as putatively swept loci identified in genomic scans become functionally verified, it may appear more likely the case that existing RHH estimators are underestimating the true rate. Regardless of the species or population under consideration, these results highlight the need for future genomic studies to simultaneously consider and reconcile both classes of analyses in order to gain the most comprehensive and accurate understanding of the recent adaptive history of natural populations and suggest that SHH model– and RHH model–based approaches may indeed inform one another.

**Acknowledgments**

I thank Peter Andolfatto, Chip Aquadro, Doris Bachtrog, Anna-Sapfo Malaspinas, Bret Payseur, Nadia Singh, Kevin Thornton, and three anonymous reviewers for helpful comments and discussion.
Literature Cited

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

Andolfatto P, Przeworski M. 2001. Regions of lower crossing over harbor more rare variants in African populations of Drosophila melanogaster. Genetics. 158:657–665.

Andolfatto P. 2007. Hitchhiking effects of recurrent beneficial amino acid substitutions in the Drosophila melanogaster genome. Genom Res. 17:1755–1762.

Bauer DuMont V, Aquadro CF. 2005. Multiple signatures of positive selection downstream of notch on the X chromosome in Drosophila melanogaster. Genetics. 171:639–653.

Carlson CS, et al. 2005. Genomic regions exhibiting positive selection identified from dense genotype data. Genome Res. 15:1553–1565.

Charlesworth B. 1996. Background selection and patterns of genetic diversity in Drosophila melanogaster. Genet Res. 68:131–149.

Fay J, Wu C-I. 2000. Hitchhiking under positive Darwinian selection. Genetics. 155:1405–1413.

Glinka S, Ometto L, Mousset S, Stephan W, De Lorenzo D. 2003. Demography and natural selection have shaped genetic variation in Drosophila melanogaster: a multi-locus approach. Genetics. 165:1269–1278.

Haddrill PR, Thornton KR, Charlesworth B, Andolfatto P. 2005. Multilocus patterns of nucleotide variability and the demographic and selection history of Drosophila melanogaster populations. Genome Res. 15:790–799.

Harr B, Kauer M, Schlötterer C. 2002. Hitchhiking mapping: a population-based fine-mapping strategy for adaptive mutations in Drosophila melanogaster. Proc Natl Acad Sci USA. 99:12949–12954.

Jensen JD, Bauer DuMont V, Ashmore AB, Gutierrez A, Aquadro CF. 2007. Patterns of variability and divergence at the diminutive gene region of Drosophila melanogaster. Genetics. 177:832–840.

Jensen JD, Kim Y, Bauer DuMont V, Aquadro CF, Bustamante CD. 2005. Distinguishing between selective sweeps and demography using DNA polymorphism data. Genetics. 170:1401–1410.

Jensen JD, Thornton KR, Andolfatto P. 2008. An approximate Bayesian estimator suggests strong recurrent selective sweeps in Drosophila. PLoS Genet. 4.e1000198.

Jensen JD, Thornton KR, Aquadro CF. 2008. Inferring selection in partially sequenced regions. Mol Biol Evol. 25:438–446.

Kaplan NL, Hudson RR, Langley CH. 1989. The ‘hitchhiking effect’ revisited. Genetics. 123:887–899.

Kelley JL, Madey G, Calboum JC, Swanson W, Akey JM. 2006. Genomic signatures of positive selection in humans and the limits of outlier approaches. Genome Res. 16:980–989.

Kim Y. 2006. Allele frequency distribution under recurrent selective sweeps. Genetics. 172:1967–1978.

Kim Y, Stephan W. 2002. Detecting a local signature of genetic hitchhiking along a recombining chromosomes. Genetics. 160:765–777.

Li H, Stephan W. 2006. Inferring the demographic history and rate of adaptive substitutions in Drosophila. PLoS Genet. 2:e166.

Macpherson JM, Sella G, Davis JC, Petrov DA. 2007. Genome-wide spatial correspondence between nonsynonymous divergence and neutral polymorphism reveals extensive adaptation in Drosophila. Genetics. 177:2083–2099.

Maynard Smith JM, Haigh J. 1974. The hitchhiking effect of a favourable gene. Genet Res. 23:23–35.

Nielsen R. 2001. Statistical tests of selective neutrality in the age of genomics. Heredit. 86:641–647.

Nielsen R. 2005. Molecular signatures of natural selection. Annu Rev Genet. 39:197–218.

Nielsen RI, Hellmann M, Hubisz M, Bustamante CD, Clark AG. 2007. Recent and ongoing selection in the human genome. Nat Rev Genet. 8:857–868.

Ometto L, Glinka S, De Lorenzo D, Stephan W. 2005. Genomic scans for selective sweeps using SNP data. Mol Biol Evol. 22:2119–2130.

Przeworski M. 2002. The signature of positive selection at randomly chosen loci. Genetics. 160:265–279.

Sabeti PC, et al. 2006. Positive natural selection in the human lineage. Science. 312:1614–1620.

Sella G, Petrov DA, Przeworski M, Andolfatto P. 2009. Pervasive natural selection in the Drosophila genome. PLoS Genet. 5:e1000495.

Stephan W, Wiehe T, Lenz MW. 1992. The effect of strongly selected substitutions on neutral polymorphism: analytical results based on diffusion theory. Theor Popul Biol. 41: 237–254.

Tajima F. 1989. Statistical methods for testing the neutral mutation hypothesis. Genetics. 123:437–460.

Tenaillon ML, U’Ren J, Tenaillon O, Gaut BS. 2004. Selection versus demography: a multilocus investigation of the domestication process in maize. Mol Biol Evol. 21:1214–1225.

Teshima KM, Coop G, Przeworski M. 2006. How reliable are empirical genomic scans for selective sweeps? Genome Res. 16:702–712.

Thornton KR, Andolfatto P. 2006. Approximate Bayesian inference reveals evidence for a recent, severe bottleneck in a Netherlands population of Drosophila melanogaster. Genetics. 172:1607–1619.

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

Thornton KR, Jensen JD, Becquet C, Andolfatto P. 2007. Progress and prospects in mapping recent selection in the genome. Heredity. 98:340–348.

Wiehe TH, Stephan W. 1993. Analysis of a genetic hitchhiking model and its application to DNA polymorphism data from Drosophila melanogaster. Mol Biol Evol. 10:842–854.

Williamson SH, et al. 2005. Simultaneous inference of selection and population growth from patterns of variation in the human genome. Proc Natl Acad Sci USA. 102:7882–7887.

Wright SI, et al. 2005. The effects of artificial selection on the maize genome. Science. 308:1310–1314.

Wright SI, Gaut B. 2005. Molecular population genetics and the search for adaptive evolution in plants. Mol Biol Evol. 22:506–519.

Brandon Gaut, Associate Editor

Accepted August 28, 2009