Mistaken Identity: Another Bias in the Use of Relative Genetic Divergence Measures for Detecting Interspecies Introgression

by

Kathryn R. Ritz

University Program in Genetics and Genomics
Department of Biology
Duke University

Date:_____

Approved:

_____
Mohamed A. F. Noor, Supervisor

_____
John Willis, Chair

_____
Kathleen Donohue

Thesis submitted in partial fulfillment of
the requirements for the degree of Master of Science in the
University Program in Genetics and Genomics
in the Graduate School of Duke University

2016

ABSTRACT

Mistaken Identity: Another Bias in the Use of Relative Genetic Divergence Measures for Detecting Interspecies Introgression

by

Kathryn R. Ritz

University Program in Genetics and Genomics
Department of Biology
Duke University

Date:_____
Approved:

_____
Mohamed A. F. Noor, Supervisor

_____
John Willis, Chair

_____
Kathleen Donohue

An abstract of a thesis submitted in partial fulfillment
of the requirements for the degree of Master of Science in the
University Program in Genetics and Genomics
in the Graduate School of Duke University

2016

Abstract

Measures of genetic divergence have long been used to identify evolutionary processes operating within and between species. However, recent reviews have described a bias in the use of relative divergence measures towards incorrectly identifying genomic regions that are seemingly immune to introgression. Here, we present a novel and opposite bias of relative divergence measures: misidentifying regions of introgression between sister species. We examine two distinct haplotypes of intermediate frequency within Drosophila pseudoobscura at the DPSX009 locus. One of these haplotypes had lower relative divergence than another to sister species D. persimilis. Although we and others initially presumed one haplotype have spread via introgression between D. pseudoobscura and D. persimilis, absolute divergence measures and individual sequence analysis suggest that haplotype structuring occurred as the result of within-species processes. The potential for this type of misinference may occur with any haplotype that recently spread within a species. We conclude that absolute measures of genetic divergence are necessary for confirming putative regions of introgression.

Contents

Abstract iv

List of Tables vi

List of Figures vii

Acknowledgements viii

1. Introduction 1

2. Results and Discussion 4

    2.1 Haplotype structure and comparison to sister species 4

    2.2 Tests for introgression 10

    2.3 Spread of a single haplotype can mimic observed pattern 11

3. Synopsis 14

4. Materials and Methods 15

    4.1 Sequences and Fly Stocks 15

    4.2 DNA Prep, Amplification, and Genotyping 15

    4.3 Haplotype Classification and Data Analysis 16

    4.4 In silico psA Haplotype Expansions 17

Supplemental Information 19

References 23

List of Tables

Table 1: Summary of population genetic measures at the *DPSX009 locus* 8

Table 2: Support for models explaining haplotype structuring observed at *DPSX009* . 8

List of Figures

Figure 1: Significant linkage disequilibrium across the *DPSX009* locus in *D. pseudoobscura* .. 5

Acknowledgements

I would like to thank my advisor, Mohamed Noor. While he is well known at Duke for his perpetual optimism, willingness to take on more responsibilities than he is required, and for always being photographed with a "thumbs up" sign, I can say with confidence his greatest quality is his devotion as an advisor. He continuously puts the needs of his students before the success of the lab, and this is a rare quality in at a top tier research institution. Thank you Mohamed for putting your faith in me from day one. I'll never be able to put into words how grateful I am to have been a part of your lab, where I have grown immensely as a scientist and person. Here, not only did I learn skills applicable to a variety of professions, but how to determine which skills I actually enjoy. I learned that it is possible to be happy in your job (as you are a great example of that). Thank you for supporting my decision and for what I anticipate to be a lifelong friendship.

Thank you to my lab mate Katharine Korunes, who makes me laugh when I am tempted to throw things out the window and has been a rock for me at Duke since our first moments here. We have been through so many trials and tribulations together – and as a result I've found a lifelong friend.

Thank you to the rest of my lab and our lab manager Brenda, who have supported me when lab has been frustrating and difficult, and celebrated with me in my successes. They also laugh at my jokes even when they are not that funny, and for that I am grateful.

Finally, I would like to thank my family and friends for their endless support and love over the past few years. There are far too many of you to list individually, but I am grateful to each and every one of you for standing by my side not only when I've needed you most, but every single day. I am lucky to be so well loved.

This is a modified version of my work published as:

> Ritz, KR and MAF Noor. 2016. Mistaken Identity: Another Bias in the Use of Relative Genetic Divergence Measures for Detecting Interspecies Introgression. PLoS One, in press.

1. Introduction

One of the most surprising findings since the implementation of molecular evolutionary genetics is that a very large number of species hybridize and successfully exchange genes (See review in [1]). With the increase of high-throughput sequencing availability and more sensitive molecular techniques, a plethora of approaches have been developed to analyze and test for introgression at particular loci. However, investigators often employ the simple and classic approach of examining haplotypes for segregating polymorphisms vs. fixed differences between the species. Several studies specifically report the appearance of introgression based on observed distinct haplotypes, which may appear to be more similar to haplotypes found in the sister taxon [2–4].

Several recent reviews [5–7] have emphasized that the use of relative divergence measures that compare within- to between-species variation, such as F_{ST} or Nei's D_a (average nucleotide divergence between species corrected for average divergence within species [8]), may be misleading with respect to testing for gene flow, particularly in regions of low recombination. Those reviews highlighted several empirical studies that

perhaps erroneously attributed variation among regions of the genome in interspecies divergence to interspecies gene flow because they used relative divergence measures. Selective sweeps and background selection reduce variation within species particularly in regions of low recombination, such that the partitioned variation by measures such as F_{ST} and D_a exhibits an excess of divergence between species. These low recombination regions then appear to be "islands of divergence" relative to the remainder of the genome, which then incorrectly appears to have experienced gene exchange (the mirage referred to by Noor & Bennett [5]). Absolute divergence measures do not suffer this particular problem, but they can have other confounding issues.

Recently, Gredler *et al.* [9] reported within- and between-species variation at various loci in *Drosophila pseudoobscura* and *D. persimilis* across time, and the haplotype structure at one locus (*DPSX009*) showed signals of introgression between these species. At this locus, *D. pseudoobscura* harbors two intermediate-frequency haplotypes with several linked differences between them, and one of these haplotypes appears to share more variation with *D. persimilis* than with other *D. pseudoobscura* haplotypes. Previous work by another group [10] identified this same structuring at *DPSX009*, and noted, "the sequence data suggest the occurrence of gene flow and possibly recent introgression at *X009." The case for introgression is especially strong because recombination is low around*

DPSX009 [11], *and thus hitchhiking and background selection around this region should make it unlikely to retain intermediate frequency variation.* However, this purported introgression is curious because the mapping studies have shown that *DPSX009* is linked to factor(s) conferring reproductive isolation [12] which should impede introgression, though it is possible factor(s) have evolved more recently than introgression occurred.

Here, we report a more detailed examination of the haplotype structure at this locus. While previous studies have interpreted the pattern of divergence at this locus as introgression [9,10], closer examination and use of absolute divergence measures suggest the haplotype structuring is instead likely a consequence of within-species processes. Our study demonstrates that relative measures of divergence may not only inflate measurements of divergence in areas of low genetic diversity [5,6], but also inflate measurements of introgression in areas that are currently undergoing within-species processes, such as selective sweeps. We do not argue that this particular bias is necessarily common, however we observe it within our case study, and other studies have inappropriately relied heavily upon relative divergence measures in their interpretations [5–7], so we stress that caution is warranted.

2. Results and Discussion

2.1 Haplotype structure and comparison to sister species

D. pseudoobscura DPSX009 sequences from 1997 and 2013 revealed the appearance of a distinct haplotype structure spanning approximately 300 bases [9]. This haplotype contains 7 SNPs and two 15bp indels in complete linkage disequilibrium, an extremely unusual observation in *D. pseudoobscura* where LD typically decays to >10% of background level within approximately 20bp [13]. Other SNPs in this region also exhibited strong but incomplete LD extending further out (up to 700bp, see Fig 1). The presence of two haplotype groups at intermediate frequency in the 2013 samples drove the estimate of Tajima's D to be strongly positive (0.73806; [9]), which contrasts most other loci studied in the species (e.g., [10,14,15]).

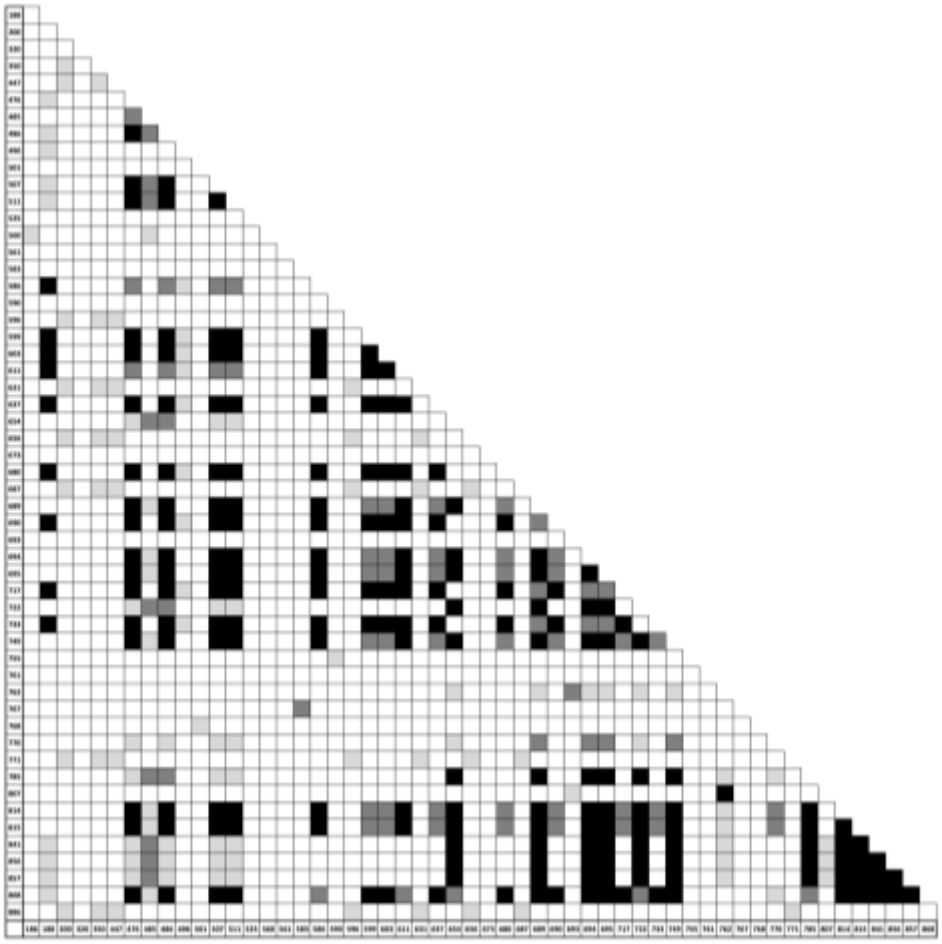

Figure 1: Significant linkage disequilibrium across the *DPSX009* locus in *D. pseudoobscura*. Shading corresponds to significant values (light grey: $p<0.05$, dark grey: $p<0.005$, and black: $p<0.0005$). 55 polymorphic sites across 51 sequences were compared using a Fisher's exact test. Bases range from nucleotide 168 to 886 across the locus (see axes).

Three processes might explain such an observation. First, reduced levels of recombination, perhaps through a microinversion, can cause haplotype structuring within a population. Indeed, previous studies have found that this region of the *D. pseudoobscura* genome has unusually low recombination rates relative to the rest of the

genome (data in [11,13]). Second, recent introgression from a hybridizing species may result in the observed pattern. *D. pseudoobscura* is known to hybridize with its sibling species *D. persimilis* in the wild [16] and many studies have documented evidence of gene exchange between them [10,17–19]. The presence of long haplotypes often indicates introgression between species [20,21], and Machado *et al*. [10] report shared partial haplotypes between these two species. Third, within-species processes such a selective sweep in progress or a new balanced polymorphism can cause rapid spread of a previously rare haplotype. This mechanism is potentially consistent with Machado *et al*.'s [10] report of population structuring at this locus, as well as the suggestion of a possible shift in abundance of one of the haplotypes between 1997 and 2013 [9] causing a sign change in Tajima's D to positive. We note that these processes are not mutually exclusive, and that combinations of these evolutionary events may occur simultaneously or sequentially. Here we interpret analyses of the *DPSX009* locus to better understand the pattern observed.

We compared both *D. pseudoobscura* haplotype groups to *D. persimilis* at the *DPSX009* locus, where the haplotype groups are defined by alleles at two 15-bp indels that are in perfect linkage disequilibrium. One *D. pseudoobscura* haplotype (hereafter, haplotype psA) has 1 fixed difference from *D. persimilis* and 9 shared polymorphisms, while the other *D. pseudoobscura* haplotype (psB) has 10 fixed differences and 2 shared

polymorphisms (Fisher's exact test, two-tailed p=0.019, Table 1), where shared polymorphisms are defined as polymorphic sites with alleles present in both groups [22]. As a consequence of defining haplogroups psA and psB by their genotype at the two 15bp indels, we exclude these indels from further nucleotide diversity analysis, though we acknowledge that excluding them deflates slightly the absolute differentiation between psB and *D. persimilis*. Nonetheless, indels are routinely excluded in many population genomic analyses, and coding these indels as SNPs would have created circularity in that we would be comparing divergence between pairs of groups when one group was explicitly defined so as to have greater divergence. We note that the definition of haplotypes from polymorphism data in cases like this is subjective; factors such as window size, number of fixed differences chosen, and recombination may influence classification of samples. We keep this potential for bias in mind throughout our analysis.

Quantifying the nucleotide differences between species relative to variation within species using Nei's D_a [8], we find that D_a between psA and *D. persimilis* is 0.0100, while D_a between psB and *D. persimilis* is 0.0157. Such a pattern, where one of two haplogroups found within a species is more similar by relative divergence measures to an abundant haplotype found in another species, could erroneously be interpreted as potential evidence of introgression.

Table 1: Summary of population genetic measures at the *DPSX009 locus*. Observed haplotype structure within *D. pseudoobscura*, divergence measures relative to *D. persimilis*, and nucleotide diversity of considered haplotypes.

	Fixed Differences from *D. persimilis*		**Shared Polymorphisms with *D. persimilis***		**D_a relative to *D. persimilis***	**D_{xy} relative to *D. persimilis***	**Nucleotide Diversity (π)**
	SNPs	**indels**	**SNPs**	**indels**			
psA	1	0	7	2	0.0100	0.0243	0.0165
psB	7	3	0	2	0.0157	0.0226	0.0029
D. persimilis	N/A	N/A	N/A	N/A	N/A	N/A	0.0119

Machado *et al*. [10] noted geographical structuring of variation at *DPSX009* with a small number of samples. We explored this further by genotyping 50 isolates derived from four diverse geographic areas using one of the indels differentiating psA and psB. Both haplotypes psA and psB were present in each population at an intermediate frequency (data in S1 Table), and no significant structuring was apparent (Fisher's exact test, p=0.36). Consequently, we test a suite of predictions for whether the observed pattern of divergence results from a recent introgression or within-species processes (Table 2).

Table 2: Support for models explaining haplotype structuring observed at *DPSX009*.

Introgression. If the haplotype structuring at this locus is the result of introgression of psA into *D. psuedoobscura* from *D. persimilis*, we predict the following:		
Hypothesis	**Observed**	**Support for introgression**

		model?
D_{xy} between psA and *D. persimilis* is **less than** D_{xy} between psB and *D. persimilis*.	D_{xy}(psA-*D. persimilis*) = 0.0243, D_{xy}(psB-*D. persimilis*) = 0.0226	NO
π for psA is **less than** π for *D. persimilis*.	π(psA) = 0.0165, π(*D. persimilis*) = 0.0119	NO
π for psA is **less than** π for psB	π(psA) = 0.0165, π(psB) = 0.00285	NO
D_a between *in silico* expanded haplotypes and *D. persimilis* is **less than** D_a between psA and *D. persimilis*.	Mean *in silico*-*D. persimilis* D_a = 0.0183, D_a(psA-*D. persimilis*) = 0.0100	NO
Within species processes. If the haplotype structuring at this locus is the result of within species processes, we predict the following:		
Hypothesis	**Observed**	**Support for within species processes model?**
D_{xy} between psA and psB should be 1) **less than** D_{xy} between psA and *D. persimilis* and 2) **less than** D_{xy} between psB and *D. persimilis*.*	D_{xy}(psA-psB) = 0.0217, D_{xy}(psA-*D. persimilis*) = 0.0243, and D_{xy}(pB-*D. persimilis*) = 0.0226 Comparison 1: $p<0.000001$ Comparison 2: $p=0.1204$	YES[†]
Fixed differences between psB and *D. persimilis* will be derived in psB. They will not be shared with outgroup species *D. miranda*.	10 of 10 fixed differences between psB and *D. persimilis* are not shared with *D. miranda*, and thus are derived (see S1 Fig)	YES

*These patterns may not be strong because of extensive shared variation between *D. pseudoobscura* and *D. persimilis* given their recent divergence.

[†]Comparison 1 was statistically significant, and comparison 2 was not significant but was in the expected direction. However, the application of a statistical analysis in this instance is subject to a pseudoreplication bias, which we discuss in Materials and Methods.

2.2 Tests for introgression

Previous studies have suggested potential artifacts associated with the use of relative measures of divergence in testing for introgression [5,6]. As such, we examined variation at the *DPSX009* locus using an absolute measure of divergence, Nei's D_{xy} [8]. If the psA haplotype resembles *D. persimilis* due to introgression between these species, we predict D_{xy} should be lower between psA and *D. persimilis* than between psB and *D. persimilis* (Table 2). In contrast, if one of the two *D. pseudoobscura* haplotypes arose and spread as a result of within-species processes, psA and psB should more closely resemble each other than either resembles *D. persimilis* (Table 2). We observe that D_{xy} is slightly higher between psA and *D. persimilis* than between psB and *D. persimilis*, and we find lower sequence difference between psA and psB than between psA and *D. persimilis* or between psB and *D. persimilis*, thus failing to support the introgression hypothesis prediction and providing some support for the within-species processes hypothesis prediction (summary in Table 2).

Additionally, if the psA haplotype introgressed recently from *D. persimilis* into *D. pseudoobscura*, we predict that there should be less nucleotide diversity within the newer psA haplotype than the potentially ancestral psB haplotype. The opposite was true (see Table 2), however these predictions potentially assume directionality and a single introgression event rather than multiple introgression events. Further, introgression of

psA would likely result in less nucleotide diversity in psA than in a *D. persimilis* progenitor haplotype. Again, the opposite was true (see Table 2).

Finally, if the less polymorphic haplotype psB was originally a single haplotype that has spread within *D. pseudoobscura* through within-species processes rather than introgression, we predict that most of the differences between psB and *D. persimilis* would be derived in psB relative to an outgroup species, whereas the introgression hypothesis does not necessarily make such a prediction. We compared the 10 differences fixed between psB and *D. persimilis* to outgroup species *Drosophila miranda*. We found all 10 sites in *D. persimilis* shared the *D. miranda* allele, consistent with the hypothesis that psB is a newer haplotype whose spread within *D. pseudoobscura* resulted from within-species processes (see S1 Fig). The phylogenetic relationship between the *D. pseudoobscura*, *D. persimilis*, and *D. miranda* sequences are depicted in S2 Fig.

2.3 Spread of a single haplotype can mimic observed pattern

The above lines of evidence suggest that the presence of two haplotype groups at the *DPSX009* locus in *D. pseudoobscura* may be that the haplogroup with less nucleotide diversity (psB) stems from a single, derived *D. pseudoobscura* haplotype that recently became more abundant via selection or other within-species processes. Intermediate frequency variation in *D. pseudoobscura* is rare, and most molecular variation is present

as singleton alleles [10]. Linked singleton differences across a single progenitor psB haplotype may spread in a population and result in a positive Tajima's D overall and a high D_a between the new haplogroup and *D. persimilis* because of the low nucleotide diversity within psB. To test this hypothesis, we examined how divergence from *D. persimilis* is reflected if selected haplotypes within the hypothesized ancestral haplogroup (psA) were to suddenly become abundant (data in S2 Table). We predicted that, if psA is the ancestral haplogroup and prevalence of psB is the result of a within-species spread, individual psA haplotypes spread *in silico* would exhibit a pattern of differentiation similar to that observed for psB: a D_a similar to psB and higher than psA relative to *D. persimilis*.

Consistent with this prediction, D_a from the artificial haplogroup ranged from 0.01262 to 0.0251 (median = 0.01544), with each artificial spread having a D_a to *D. persimilis* higher than the original psA haplogroup (psA-*D. persimilis* D_a = 0.0100). The median D_a to *D. persimilis* across the artificial haplogroups was also highly similar to D_a from psB (psB-*D.persimilis* D_a = 0.0157). We see a similar pattern when indels are coded in the *in silico* analysis as SNPs, where D_a ranges from 0.0127 to 0.0256 (median = 0.0163, data in S3 Table).

We can further compare the pattern of divergence observed for psB to the artificial haplogroups by examining the number of fixed differences between each group and *D. persimilis*. We see that the *in silico* expansions have between 7 and 17 fixed differences from *D. persimilis* (mean = 11.76), a range consistent with that observed for psB (10 fixed differences). Again, this result is consistent with our prediction that the apparent divergence between psB and psA is the result of within-species processes.

We also found that one artificial spread resulted in a highly positive Tajima's D in the overall population including the expanded haplogroup and the original psA haplogroup (Tajima's D = 1.5878, data in S2 Table). This observation is similar to the aforementioned pattern seen by Gredler *et al*. (2015) [9]. Our simulations of the expansion of single haplotypes create a pattern of genetic diversity nearly identical to what was observed in natural populations: two intermediate frequency haplotype groups of which one (the ancestral group, psA) appears to be more similar to *D. persimilis* by relative divergence measures. From these results, we suggest that the pattern of intermediate haplotype frequency observed at *DPSX009* is likely the result of the spread of a single progenitor haplotype.

3. Synopsis

Together, these data indicate that the patterns of variation at *DPSX009* in *D. pseudoobscura* are entirely consistent with the action of within-species processes such as selection, and we find an absence of evidence for interspecies introgression. The inference of potential introgression at this locus described by Machado *et al*. [10] and initially hypothesized by us here reflects a bias associated with patterns of intraspecific variation misleading interspecies comparisons, particularly those using relative divergence measures like F_{ST} or D_a. The exact nature of the within-species processes at work is unclear: rather than a selective sweep in progress, it could represent a new balanced polymorphism or extensive gene flow from an isolated population that experienced a local bottleneck.

The previously identified bias associated with relative divergence measures [5,6] focused on the misleading appearance of high divergence and immunity to introgression in regions of low recombination. Here, we present a distinct and opposite bias: a region of low recombination incorrectly appearing to have introgressed between species. As with the previously reported bias, use of absolute divergence measures can help to determine the strength of evidence for introgression between species at particular loci.

4. Materials and Methods

4.1 Sequences and Fly Stocks

DPSX009 sequences for 37 *D. pseudoobscura* and 20 *D. persimilis* individuals originating from Mt. Saint Helena, California, USA were downloaded from GenBank (Gredler *et al.* 2015 [9]). Additional *D. pseudoobscura* populations were surveyed using flies previously collected by M.A.F. Noor and stored in at -80°C. These populations include 18 isolines from western Washington (Roslyn, Goldendale, and Easton) and 15 isolines from eastern Washington (Cheney), collected in 1996, 8 isolines from American Fork Canyon in American Fork, Utah, collected in 1997, and 9 isolines from Flagstaff, Arizona, collected in 1997. No permits were required for the described study, which complied with all relevant regulations.

4.2 DNA Prep, Amplification, and Genotyping

Genomic DNA from one frozen male individual per isoline was extracted using a single fly squish protocol [23]. Although the lines were slightly inbred, males were chosen to prevent amplification of heterozygous individuals, as *DPSX009* is on the X-chromosome. Primers were designed to span a 15bp insertion within the psA haplotype at *DPSX009*. Individuals were genotyped following PCR and imaging on a 2% agarose in TBE gel.

4.3 Haplotype Classification and Data Analysis

MSH *D. pseudoobscura* sequences provided by J. Gredler were aligned using ClustalW in BioEdit 7.0.9 [24] and fixed differences between samples were manually documented and compared across the *DPSX009* locus. A 9-bp microinversion was removed from all sequences to prevent biases in genetic diversity calculations. Samples were then divided into groups by haplotype (psA and psB), and compared to *D. persimilis* across the locus. 13 *D. pseudoobscura* sequences were classified as psA by the presence of 7 SNPs and two 15-bp indels in complete linkage disequilibrium. The remaining 24 *D. pseudoobscura* sequences were classified as psB by the absence of some or all of the 9 diagnostic psA traits.

DNAsp 5.10.01 [25] was used to calculate Nei's (1987) D_a, Nei's D_{xy}, and π at *DPSX009* between psA (n=13) and *D. persimilis* (n=20) and again between psB (n=24) and *D. persimilis* (n=20). DNAsp calculates Nei's D_{xy} and Nei's D_a using Nei 1987 equations 10.20 and 10.21, respectively. Fixed differences and shared polymorphisms across compared groups were confirmed manually (data in Table 1). Shared polymorphisms are defined as sites that are polymorphic in both groups. To test for the statistical significance of difference in divergence between each D. pseudoobscura haplogroup to the other vs. to D. persimilis, we created a matrix indicating percent sequence difference between every pair of haplotypes. We then bootstrapped (with replacement) the matrix

1,000,000 times and assessed how often average difference between D. persimilis to either psA or psB was equal to or greater than the difference between psA to psB. The difference between psA-psB vs. psA-D. persimilis appeared to be highly statistically significant (no resamplings exhibited equal or greater divergence) while that between psA-psB and psB-D. persimilis was not so (120,416 bootstraps exhibited equal or greater divergence). However, we do not emphasize these outcomes as test statistics because of the issue of pseudoreplication due to what appears to be a recent shared coalescent event within psA. Instead, the weight of evidence from multiple separate comparisons (Table 2) argues for our interpretation of the results.

4.4 In silico psA Haplotype Expansions

Each of the 13 psA-classified *DPSX009* sequences was expanded *in silico* to simulate an increase in haplotype abundance. These sequences were expanded to the same population size as the available sequences for the psA group (n=13). Specifically, each sequence with the psA haplotype was copied 13 times to represent a theoretical haplogroup resulting from selection or within-species processes. To prevent exclusion of indels by DNAsp, we repeat analyses with indels coded as SNPs in these sequences. Expanded haplotypes were then compared to *D. persimilis* (n=20) in DNAsp, where D_a and D_{xy} were calculated. Fixed differences and shared polymorphisms across each expanded haplotype and *D. persimilis* were confirmed manually. Mean D_a and D_{xy}

values were calculated across the 13 simulations (data in S2 and S3 Tables). We do not discuss the D_{xy} values within our analysis, since this measurement is uninformative here. D_{xy} is always equal to D_a plus a constant factor (0.0059, half of $\pi_{D.\ persimilis}$) because there is no variation within the artificial haplogroups.

Supplemental Information

	Location	138	447	470	507	511	586	611	631	638	654	657	699	700	701	785	807	815	817	896
	Description	7bp indel	SNP	SNP	SNP	SNP	SNP	SNP	SNP	1bp indel	SNP	SNP	SNP & 1bp indel	SNP & 1bp indel	15bp indel	SNP	SNP	SNP	15bp indel	1bp indel
D. persimilis	D. persimilis_2013-78	-	T	G	G	T	C	C	G	-	G	T	G	G	+	C	A	C	-	-
	D. persimilis_2013-45	+	T	G	C	T	C	C	T	+	A	T	T	G	+	C	A	G	-	-
	D. persimilis_2013-48	+	T	G	G	T	C	C	T	+	A	T	T	G	+	T	A	G	-	-
	D. persimilis_2013-24	+	T	G	G	T	C	C	G	-	G	T	G	G	+	C	A	C	-	-
	D. persimilis_2013-20	+	T	G	G	T	C	C	G	-	A	T	T	G	+	C	A	G	-	-
	D. persimilis_2013-53	+	T	G	C	T	C	G	G	-	A	T	G	G	+	T	A	G	-	-
	D. persimilis_2013-26	+	G	G	C	T	C	C	G	-	G	T	G	G	+	T	A	G	-	-
	D. persimilis_2013-42	+	T	G	C	T	C	G	G	-	G	T	G	G	+	T	A	G	-	-
	D. persimilis_2013-63	+	T	G	G	T	C	C	G	-	A	T	T	G	+	T	A	G	-	-
	D. persimilis_2013-17	+	T	G	G	T	C	C	T	+	A	T	T	G	+	T	A	G	-	-
	D. persimilis_2013-5	+	A	G	G	T	C	G	G	-	A	T	T	G	+	T	A	G	-	-
	D. persimilis_2013-64	-	T	G	A	T	C	C	G	-	G	T	G	G	+	C	A	G	-	+
	D. persimilis_2013-33	+	T	G	G	T	C	G	G	-	A	T	G	G	+	T	A	G	-	-
	D. persimilis_2013-86	-	A	G	G	T	C	C	T	+	A	T	T	G	+	C	A	G	-	-
	D. persimilis_2013-93	+	T	G	C	T	C	C	G	-	A	T	T	G	+	T	A	G	-	-
	D. persimilis_1997-42	+	T	G	C	T	C	G	G	-	A	T	T	G	+	C	A	G	-	-
	D. persimilis_1997-1	+	T	G	G	T	C	C	T	+	A	T	T	G	+	T	A	C	-	-
	D. persimilis_1997-7	+	T	G	A	T	C	C	G	-	G	T	G	G	+	C	A	G	-	-
	D. persimilis_1997-3	+	T	G	G	T	C	C	G	-	G	T	G	G	+	T	A	G	-	-
	D. persimilis_1997-26	+	T	G	G	T	C	C	T	+	A	T	G	G	+	C	A	C	-	-
D. pseudoobscura psA	D. pseudoobscura_1997-30	+	T	A	G	C	G	C	G	-	G	T	G	G	+	C	G	C	-	-
	D. pseudoobscura_1997-91	+	A	A	G	C	C	G	T	+	A	C	T	G	+	T	A	C	-	-
	D. pseudoobscura_1997-4	+	T	A	C	T	C	G	G	+	A	T	T	G	+	T	G	C	-	-
	D. pseudoobscura_2013-85	-	T	A	G	C	G	C	G	-	G	T	G	A	+	C	G	C	-	-
	D. pseudoobscura_2013-15	+	T	A	C	T	G	C	G	-	G	T	G	G	+	C	G	C	-	-
	D. pseudoobscura_2013-79	+	T	A	C	T	G	C	G	-	G	T	G	G	+	C	G	C	-	-
	D. pseudoobscura_2013-7	+	T	A	C	T	G	C	G	-	G	T	G	G	+	C	G	C	-	-
	D. pseudoobscura_2013-37	-	T	A	C	T	G	C	G	-	G	T	G	G	+	C	G	C	-	-
	D. pseudoobscura_2013-1	-	T	A	C	T	G	C	G	-	G	T	G	G	+	C	G	C	-	-
	D. pseudoobscura_2013-4	+	T	A	C	T	C	G	G	+	A	T	T	G	+	T	G	C	-	-
	D. pseudoobscura_2013-35	+	T	A	C	T	C	G	G	+	A	T	T	G	+	T	G	C	-	-
	D. pseudoobscura_2013-60	-	T	A	C	T	C	G	G	+	A	T	T	G	+	T	G	C	-	-
	D. pseudoobscura_2013-76	+	T	A	C	T	C	G	G	+	A	T	T	G	+	T	G	C	-	-
D. pseudoobscura psB	D. pseudoobscura_1997-16	+	T	A	G	C	G	C	G	-	A	G	-	-	-	T	G	T	+	-
	D. pseudoobscura_1997-9	+	T	A	G	C	G	C	G	-	A	C	-	-	-	T	G	T	+	-
	D. pseudoobscura_1997-13	+	T	A	G	C	G	C	G	-	A	C	-	-	-	T	G	T	+	-
	D. pseudoobscura_1997-10	+	T	A	G	C	G	C	G	-	A	C	-	-	-	T	G	T	+	-
	D. pseudoobscura_1997-24	+	T	A	G	C	G	C	G	-	A	C	-	-	-	T	G	T	+	-
	D. pseudoobscura_1997-31	-	T	A	G	C	G	C	G	-	A	C	-	-	-	T	G	T	+	-
	D. pseudoobscura_1997-2	+	T	A	G	C	G	C	G	-	A	C	-	-	-	T	G	T	+	-
	D. pseudoobscura_1997-37	-	T	A	G	C	G	C	G	-	A	C	-	-	-	T	G	T	+	-
	D. pseudoobscura_1997-32	-	T	A	G	C	G	C	G	-	A	C	-	-	-	T	G	T	+	-
	D. pseudoobscura_2013-36	+	T	A	G	C	G	C	G	-	A	G	-	-	-	T	G	T	+	-
	D. pseudoobscura_2013-11	+	T	A	G	C	G	C	G	-	A	G	-	-	-	T	G	T	+	-
	D. pseudoobscura_2013-65	+	T	A	G	C	G	C	G	-	A	C	-	-	-	T	G	T	+	-
	D. pseudoobscura_2013-83	+	T	A	G	C	G	C	G	-	A	C	-	-	-	T	G	T	+	-
	D. pseudoobscura_2013-67	+	T	A	G	C	G	C	G	-	A	C	-	-	-	T	G	T	+	-
	D. pseudoobscura_2013-3	+	T	A	G	C	G	C	G	-	A	C	-	-	-	T	G	T	+	-
	D. pseudoobscura_2013-6	+	T	A	G	C	G	C	G	-	A	C	-	-	-	T	G	T	+	-
	D. pseudoobscura_2013-52	+	T	A	G	C	G	C	G	-	A	C	-	-	-	T	G	T	+	-
	D. pseudoobscura_2013-84	+	T	A	G	C	G	C	G	-	A	C	-	-	-	T	G	T	+	-
	D. pseudoobscura_2013-81	+	T	A	G	C	G	C	G	-	A	C	-	-	-	T	G	T	+	+
	D. pseudoobscura_2013-68	-	T	A	G	C	G	C	G	-	A	C	-	-	-	T	G	T	+	-
	D. pseudoobscura_2013-29	-	T	A	G	C	G	C	G	-	A	C	-	-	-	T	G	T	+	+
	D. pseudoobscura_2013-77	-	T	A	G	C	G	C	G	-	A	C	-	-	-	T	G	T	+	-
	D. pseudoobscura_2013-30	+	T	A	G	C	G	C	G	-	A	C	-	-	-	T	G	T	+	-
	D. pseudoobscura_2013-18	-	T	A	G	C	G	C	G	-	A	C	-	-	-	T	G	T	+	-
	D. miranda	-	A	G	G	T	C	C	G	+	A	T	T	G	+	C	A	C	-	-

	Analysis																				
	psA to *D. persimilis*																				***Totals***
	Fixed Differences			1																	1
	Shared Polymorphisms	1	1		1			1	1	1	1		1			1					9
	psB to *D. persimilis*																				
	Fixed Differences			1		1	1					1	1	1	1		1	1	1		10
	Shared Polymorphisms	1																		1	2
	psB to *D. miranda*																				
	Derived fixed differences between psB and *D. persimilis*			1		1	1					1	1	1	1		1	1	1		10
	% differences derived																				100

Figure S 1: Variable sites across *DPSX009* haplotypes. Summary of fixed differences and shared polymorphisms between psA, psB, and *D. persimilis*. Additional alignment to *D. miranda* and comparison to fixed differences between psB and *D. persimilis* supports hypothesis that psB is derived and results from within-species processes.

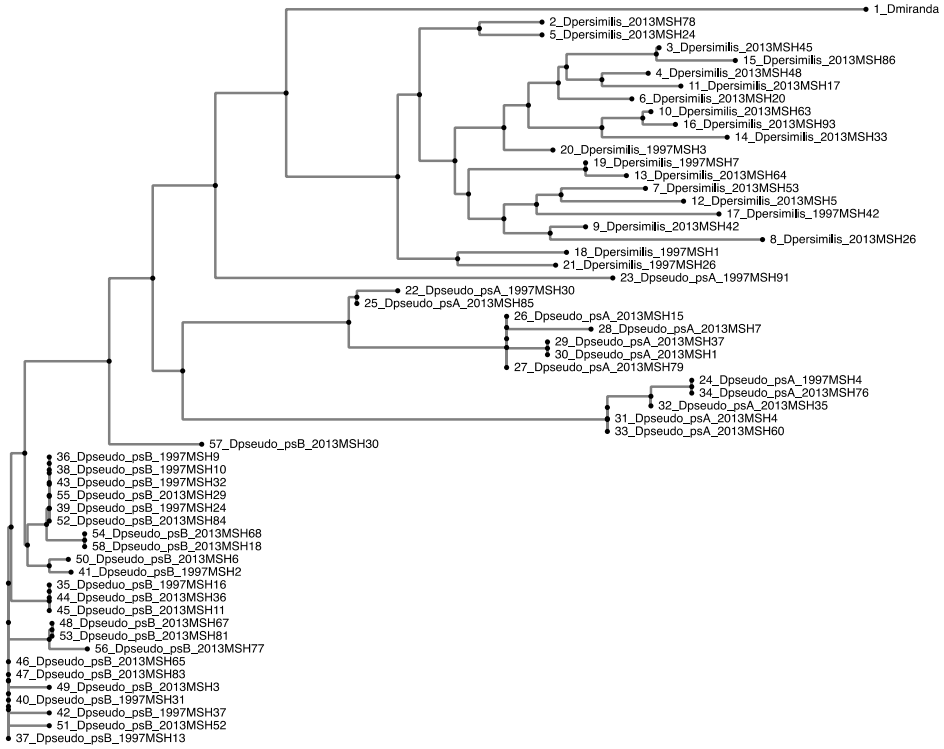

Figure S 2: Phylogeny of *DPSX009* sequences shows clustering of psA and psB haplotypes. Independent clustering of psB and of psA with *D. persimilis* sequences can be seen in this neighbor joining phylogeny generated using phylo.io [26]. *D. miranda* is used as an outgroup.

Table S 1: Presence of psA and psB haplotype structure in additional populations of *D. pseudoobscura*.

	psA	psB	Total
Roslyn/Easton/Goldendale, WA	10	8	18
Cheney, WA	7	8	15
American Fork Canyon, UT	2	6	8
Flagstaff, AZ	6	6	9

Table S 2: *In silico* expansion of individual psA haplotypes and comparison to *D. persimilis*, with indels excluded.

Expanded Sequence (psA-type)	Fixed Differences		Tajima's D*	D_a relative to *D. persimilis*	D_{xy} relative to *D. persimilis*
	SNPs	indels			
1997_MSH_pse30	7	0	-0.24378	0.01262	0.01855
1997_MSH_pse91	10	3	1.5878	0.0179	0.02387
1997_MSH_pse4	17	0	-0.0563	0.0251	0.03102
2013_MSH_pse1	9	0	-0.57549	0.01541	0.02133
2013_MSH_pse4	15	0	-0.28705	0.02265	0.02857
2013_MSH_pse7	9	0	-0.48896	0.01541	0.02133
2013_MSH_pse15	8	0	-0.80625	0.0142	0.02012
2013_MSH_pse35	16	0	-0.1861	0.02387	0.0298
2013_MSH_pse37	9	0	-0.57549	0.01544	0.02138
2013_MSH_pse60	15	0	-0.28705	0.02265	0.02857
2013_MSH_pse76	17	0	-0.0563	0.0251	0.03102
2013_MSH_pse79	8	1	-0.80625	0.0142	0.02012
2013_MSH_pse85	8	0	-0.14283	0.01383	0.01976
Mean	11.6923		-0.22492	0.01833	0.02426
Median	9		-0.28705	0.01544	0.02138
Standard Deviation	3.8384		0.60211	0.00476	0.00476
Maximum	17		1.5878	0.0251	0.03102
Minimum	7		-0.80625	0.01262	0.01855
Range	10		2.39405	0.01248	0.01247

Table S 3: *In silico* expansion of individual psA haplotypes and comparison to *D. persimilis*, with indels included.

Expanded Sequence (psA-type)	Fixed Differences	Tajima's D*	D_a relative to *D. persimilis*	D_{xy} relative to *D. persimilis*
1997_MSH_pse30	7	-0.63866	0.01267	0.01928
1997_MSH_pse91	14	-0.00019	0.02307	0.02968
1997_MSH_pse4	17	0.05897	0.02557	0.03219
2013_MSH_pse1	9	-0.37605	0.01628	0.02289
2013_MSH_pse4	15	-0.05012	0.02313	0.02974

2013_MSH_pse7	9	-0.46097	0.01544	0.02205
2013_MSH_pse15	8	-0.52991	0.01423	0.02085
2013_MSH_pse35	16	0.0052	0.02435	0.03096
2013_MSH_pse37	9	-0.37605	0.01632	0.02295
2013_MSH_pse60	15	0.02749	0.02398	0.0306
2013_MSH_pse76	17	0.05897	0.02557	0.03219
2013_MSH_pse79	9	-0.46097	0.01544	0.02205
2013_MSH_pse85	8	-0.48061	0.01471	0.02133
Mean	11.76923077	-0.247915385	0.019289231	0.025904615
Median	9	-0.37605	0.01632	0.02295
Maximum	7	0.05897	0.02557	0.03219
Minimum	17	-0.63866	0.01267	0.01928
Range	10	0.69763	0.0129	0.01291

References

1. Mallet J. Hybridization as an invasion of the genome. Trends Ecol Evol. 2005;20: 229–237. doi:10.1016/j.tree.2005.02.010

2. Staubach F, Lorenc A, Messer PW, Tang K, Petrov DA, Tautz D. Genome Patterns of Selection and Introgression of Haplotypes in Natural Populations of the House Mouse (Mus musculus). PLoS Genet. 2012;8. doi:10.1371/journal.pgen.1002891

3. Llopart A, Herrig D, Brud E, Stecklein Z. Sequential adaptive introgression of the mitochondrial genome in Drosophila yakuba and Drosophila santomea. Mol Ecol. 2014;23: 1124–1136. doi:10.1111/mec.12678

4. Huerta-Sánchez E, Jin X, Asan, Bianba Z, Peter BM, Vinckenbosch N, et al. Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. Nature. 2014;512: 194–197. doi:10.1038/nature13408

5. Noor MAF, Bennett SM. Islands of speciation or mirages in the desert? Examining the role of restricted recombination in maintaining species. Heredity (Edinb). Nature Publishing Group; 2009;103: 439–44. doi:10.1038/hdy.2009.151

6. Cruickshank TE, Hahn MW. Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. Mol Ecol. 2014;23: 3133–3157. doi:10.1111/mec.12796

7. Charlesworth B. Measures of divergence between populations and the effect of forces that reduce variability. Mol Biol Evol. 1998;15: 538–543. doi:10.1093/oxfordjournals.molbev.a025953

8. Nei M. Molecular Evolutionary Genetics. Columbia university press; 1987.

9. Gredler JN, Hish AJ, Noor MAF. Temporal stability of molecular diversity measures in natural populations of drosophila pseudoobscura and drosophila persimilis. J Hered. 2015;106: 407–411. doi:10.1093/jhered/esv027

10. Machado CA, Kliman RM, Markert JA, Hey J. Inferring the History of Speciation from Multilocus DNA Sequence Data: The Case of Drosophila pseudoobscura and Close Relatives. Mol Biol Evol. 2002;19: 472–488. doi:10.1093/oxfordjournals.molbev.a004103

11. McGaugh SE, Heil CSS, Manzano-Winkler B, Loewe L, Goldstein S, Himmel TL, et al. Recombination Modulates How Selection Affects Linked Sites in Drosophila.

PLoS Biol. 2012;10. doi:10.1371/journal.pbio.1001422

12. Noor MAF, Johnson NA, Hey J. Gene flow between Drosophila pseudoobscura and D-persimilis. Evolution (N Y). 2000;54: 2174–2175. Available: <Go to ISI>://000166682500032

13. Heil CSS, Ellison C, Dubin M, Noor MAF. Recombining without hotspots: A comprehensive evolutionary portrait of recombination in two closely related species of Drosophila. Genome Biol Evol. 2015;October 1,: 1–32. doi:10.1093/gbe/evv182

14. Kovacevic M, Schaeffer SW. Molecular population genetics of X-linked genes in *Drosophila pseudoobscura*. Genetics. 2000;156: 155–172. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1461252&tool=pmcentrez&rendertype=abstract

15. Wallace AG, Detweiler D, Schaeffer SW. Molecular Population Genetics of Inversion Breakpoint Regions in Drosophila pseudoobscura. G3 Gene Genomes Genet. 2013;3: 1151–1163. doi:10.1534/g3.113.006122

16. Dobzhansky T. Is there gene exchange between Drosophila pseudoobscura and Drosophila persimilis in their natural habitats? Am Nat. 1973;107: 312–314. doi:10.1086/521238

17. Hey J, Nielsen R. Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of Drosophila pseudoobscura and D. persimilis. Genetics. 2004;167: 747–760. doi:10.1534/genetics.103.024182

18. Wang RL, Wakeley J, Hey J. Gene flow and natural selection in the origin of Drosophila pseudoobscura and close relatives. Genetics. 1997;147: 1091–1106.

19. McGaugh SE, Noor MAF. Genomic impacts of chromosomal inversions in parapatric Drosophila species. Philos Trans R Soc Lond B Biol Sci. 2012;367: 422–9. doi:10.1098/rstb.2011.0250

20. Racimo F, Sankararaman S, Nielsen R, Huerta-Sánchez E. Evidence for archaic adaptive introgression in humans. Nat Rev Genet. Nature Publishing Group; 2015;16: 359–371. doi:10.1038/nrg3936

21. Vernot B, Akey JM. Resurrecting Surviving Neandeltal Linages from Modern Human Genomes. Science (80-). 2014;343: 1017–1021.

doi:10.5061/dryad.5t110.Supplementary

22. Clark AG. Neutral behavior of shared polymorphism. Proc Natl Acad Sci U S A. 1997;94: 7730–7734. doi:10.1073/pnas.94.15.7730

23. Gloor G, Engels WR. Single-fly DNA preps for PCR. Drosoph Inf Serv. 1992;71: 148–149.

24. Hall T. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT [Internet]. Nucleic Acids Symposium Series. 1999. pp. 95–98. doi:citeulike-article-id:691774

25. Rozas J, Sánchez-DelBarrio JC, Messeguer X, Rozas R. DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics. 2003;19: 2496–2497. doi:10.1093/bioinformatics/btg359

26. Robinson O, Dylus D, Dessimoz C. Phylo.io : interactive viewing and comparison of large phylogenetic trees on the web. Mol Biol Evol. 2016;33: msw080. doi:10.1093/molbev/msw080