Inferring non-additive multi-locus selection in introgressed populations using hidden Markov models

Nicolas Ayala

Advised by Dr. Russell Corbett-Detig

University of California Santa Cruz Department of Biomolecular Engineering and Bioinformatics

Abstract

Admixture is a phenomenon where genetic material from potentially disparate source populations combines and is thought to be a major source of adaptive novelty. As such, multi-locus and non-additive selection on introgressing mutations is potentially common in natural admixed populations. However, existing tools for inferring adaptive introgression only account for additive selection at a single site, overlooking phenomena such as interference among selected loci that are located proximally along a chromosome and dominance between alleles on sister chromosomes. Furthermore, most existing applications and methods assume that the landscape of local ancestry along the genome can be inferred prior to searching for selection, ignoring the fact that the local ancestry landscape is shaped by natural selection. To meet this important need, we present AHMM-MLS, a hidden Markov model based tool for inferring and identifying multiple selected sites on a chromosome. This tool numerically calculates the expected local ancestry landscapes in an admixed population for a given MLS model, and then optimizes the model to fit the data. It uses read pileup data in an introgressed population to identify selected sites and estimate a multi-locus selection model. In applying our method to a suite of simulated admixed populations, we find that the estimated strength of selection can be affected by ignoring the contributions of other sites and that our method can often identify the number of selected sites and their dominance coefficients. In applying our method to real data from admixed populations of *Drosophila melanogaster* we find that the selection coefficients of some selected sites have been overestimated in the past, and that some selected sites show evidence for dominance. This method will enable more accurate and detailed analyses of selection in admixed populations than has been possible previously.

Acknowledgements

I would like to thank my advisor Russel Corbett-Detig for making this project possible and providing invaluable support. I would also like to thank Jesper Svedberg (Post doc in the Corbett-Detig lab) for his helpful suggestions and preparations of the *Drosophila* data, as well as every other member of the Corbett-Detig lab.

Table of contents

1 INTRODUCTION

Admixture is one of the primary sources of adaptive variation in natural populations. Several recent compelling examples demonstrate the potential for introgression to drive adaptive outcomes. In *Helianthus* sunflowers, admixture provided the raw materials to enhance herbivore resistance at a number of loci (Whitney et al. 2006). In *Fundulus grandis* fish, a recent adaptive introgression allows resistance to extreme pollution and environmental change (Oziolor et al. 2019). In humans, introgression from archaic hominids is thought to have facilitated adaptation to a range of novel environments (*e.g.,* Racimo et al. 2016). Although the importance of adaptive introgression is increasingly appreciated, generalized methods to accurately detect and quantify the impacts of adaptive introgression from genome sequence data are in their infancy.

Computational methods for detecting adaptive introgression have expanded substantially in recent years, but important challenges remain for developing generalized frameworks. Many applications search for local ancestry outliers after applying tools that assume a neutral and uniform admixture process (*e.g.,* Guan 2014; Zhou et al. 2016; Sankararaman et al. 2008). However, selection in admixed populations *in itself* shapes the landscape of local ancestry, potentially introducing important biases for these approaches. Recently developed methods resolve some of these difficulties by explicitly modeling selection during admixture in read pileup data rather than genotypes. Using this approach it is possible to fit a model with a single locus under additive selection pressure (Svedberg et al. 2021). This is useful for finding evidence of selection at a locus, but it fails to distinguish between dominant and additive selection. It also fails to account for cases where two selected sites are near each other, and may affect each other through interference (Hill and Robertson 1966). This may cause existing methods to overestimate the selection coefficients of these sites.

A relevant example of a population whose evolutionary history is affected by adaptive introgression is *Drosophila Melanogaster*. A specific population in South Africa is thought to be the result of a single pulse admixture event (Medina et al. 2018), with the introgressing population showing evidence of loci under selection (Svedberg et al. 2021). Of particular interest is chromosome 3R of this population, which potentially has many nearby selected sites, close enough for interference to affect their evolutionary dynamics since the initial admixture event. We do not think that this is a rare phenomenon, as admixture has the potential to introduce many selected mutations simultaneously and those mutations will be in strong linkage disequilibrium. This population of *D. melanogaster* is potentially a model of a very general phenomenon.

In this paper we introduce a novel method of adaptive introgression modeling called AHMM-MLS (Ancestry_HMM Multi Locus Selection). Our method can calculate the effects that multiple selected loci with unrestricted selection have on an introgressed population. This enables us to test the possibility that a site of selection is caused by a dominant allele, or that a

signal of selection is actually caused by two nearby selected sites. We tested our method under various simulated scenarios of adaptive introgression with either two sites under additive selection or a single site under dominant selection. We found that our method can accurately predict the presence of two nearby selected sites, as well as determine their location and estimate their selection coefficients. Our method can also accurately predict the selection coefficients of a site under dominant selective pressure, and can predict the presence of a dominant site in certain population conditions. We applied AHMM-MLS to chromosome 3R of *D. Melanogaster,* and found that a model with two nearby selected sites produced a better fit than a model with a single selected site, and previous methods may have overestimated the selection coefficients of closely linked selected positions.

2 RESULTS AND DISCUSSION

AHMM-MLS is an extension of Ancestry_HMM (Corbett-Detig and Nielsen 2017) that can model the effects of multiple selected sites with arbitrary selection. In AHMM-MLS, we assume that the admixed population is the result of a single admixture event that took place t generations prior to the samples being collected. The underlying statistical model is a hidden markov model, in which the hidden states are the local ancestry states along the chromosome, and the emissions are the observed allele frequencies. As found in our previous work, the emission probabilities of the ancestry states in a hidden Markov model are not affected by selection, but the transition probabilities are influenced by the effect of selection at sites along the genome (Svedberg et al. 2021). AHMM-MLS captures the effects of multiple selected sites on the transition probabilities, allowing for the optimization of the likelihood of various models of selection. For details on how the emission probabilities are calculated, and how the likelihood of the HMM is calculated, we refer readers to our previous work (Corbett-Detig and Nielsen 2017).

2.1 Generating Transition Probabilities

The key innovation of this program is the way in which we generate the transition probabilities for the HMM. To capture the effects of multiple selected sites on the transition probabilities between local ancestry types, we created a novel numerical method that tracks the changes in the expected distribution of haplotypes after t generations. The haplotypes we are tracking are the local ancestries of all the selected sites and two neutral sites. If there are s selected sites, then we track $s + 2$ sites, leading to 2^{s+2} haplotypes. The haplotype distribution is modeled by a row vector H, which undergoes a transformation in each generation, producing a

sequence of vectors $H^0 \ldots H^t$. The transformation transforms the vector H^g to the vector H^{g+1}, and represents the change in the haplotype distribution from generation g to $g + 1$. At generation 0, only two haplotypes, $H^0{}_1$ and $H^0{}_{2^{s+2}}$, have a non zero value. These are the two haplotypes where all sites are of the same ancestry, and their initial values are dictated by the admixture proportion m.

$$Let\ h\ = 2^{s+2}$$

$$H^0{}_1 = m$$

$$H^0{}_h = 1 - m$$

$$H^0{}_i = 0 \quad if \quad i \neq 1\ and\ i \neq h$$

In each generation, a row vector D is made, which corresponds to the expected diploid genotype distribution in an infinite population with random mating.

$$D^g{}_{i*h+j} = H^g{}_i * H^g{}_j$$

Now, to go from this diploid genotype distribution to the haplotype distribution of the next generation, the row vector D^g is matrix multiplied with a matrix $\boldsymbol{M}$, the result of which is H'^{g+1}. H'^{g+1} is then normalized to give H^{g+1}, whose elements sum to 1.

$$H'^{g+1} = D^g \boldsymbol{M}$$

$$H^{g+1}{}_j = \frac{H'^{g+1}{}_j}{\sum_{i=1}^{h} H'^{g+1}{}_i}$$

We call this matrix $\boldsymbol{M}$ the diplotype to haplotype transformation, as it converts the diplotypes of one generation to the haplotypes of the next, accounting for recombination and natural selection. The rows of $\boldsymbol{M}$ correspond to the diploid genotypes, and the columns correspond to the haplotypes. A specific entry, such as $\boldsymbol{M}_{i,j}$, is the contribution of the diploid genotype i on the haplotype j, assuming an infinite population, random mating, and that there is no selection on the gametes during reproduction. To calculate the values for a particular row of $\boldsymbol{M}$, such as $\boldsymbol{M}_i$, we take every possible haplotype that may result from a recombination of diplotype i, and fill its corresponding entry with the rate of that recombination, multiplied by the fitness of diplotype i. By reducing the most computationally expensive parts of the numerical simulation to matrix multiplications, we are able to use libraries which specialize in speeding up matrix multiplications to quickly compute the transition probabilities.

After iterating this process for t generations, we are left with the haplotype distribution that is expected after that number of generations under our model of selection. With this haplotype distribution, we wish to calculate the transition probabilities between the local ancestries of the neutral sites. We directly calculate this by iterating through the haplotype

distribution and recording the rates of the four possible ancestry state combinations for the two neutral sites.

Once we have the transition rates for a particular model, we use the forward equations to compute a likelihood for this model. We optimize the parameters of the model to maximize its likelihood using a direct search simplex optimization algorithm (Nelder and Mead 1965). For a single site with unrestricted selection, three parameters are optimized. These are the location of the site, and the fitnesses of the two homozygous genotypes relative to the heterozygous genotype. If the selection is restricted to be additive or purely dominant, then there is only one selection parameter to optimize because the other is fixed. To optimize the parameters of a model, we start with neutral selection, and a guess of the location of the selected site, which is supplied by the user. We then center several simplexes of different sizes around this guess, and apply the Nelder Mead algorithm to these simplexes, until a stopping condition is met.

2.2 Evaluation of AHMM-MLS Over Simulated Data Sets.

To evaluate the hypothesis testing ability of AHMM-MLS we ran extensive forward simulations of plausible scenarios of adaptive introgression. We simulated diploid populations with a single admixture event, in which the admixing population had either a single allele with dominant selection, or two nearby selected alleles with additive selection. We ran both two-site hypothesis tests and dominant hypothesis tests on the locus, and evaluated the likelihoods and optimized parameters found by AHMM-MLS.

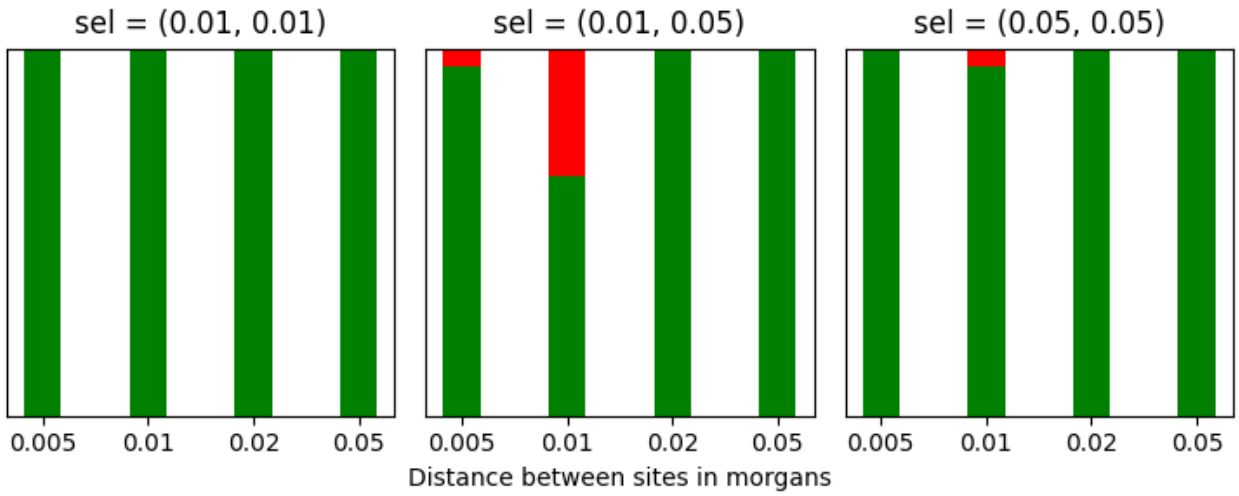

Figure 1. AHMM-MLS two site hypothesis testing was validated over a variety of adaptive scenarios with two selected sites. We ran a total of 240 simulations of adaptive introgression with two nearby sites under additive selection. The simulations span 12 different adaptive introgression scenarios, varying the strength of selection of the two sites and the distance between them. Each bar represents a different adaptive introgression scenario, with each graph representing a different set of selection coefficients for the two selected sites, and the distance in morgans between the two sites varying along the x-axis. The bars highlight in green the cases where a two-site model fits better than a one-site model.

AHMM-MLS can accurately reject the single site hypothesis in favor of the two site hypothesis over a variety of simulated populations with two nearby selected sites. AHMM-MLS can also accurately estimate the location of the two selected sites. We simulated populations which admixed 500 generations prior to sampling, with an admixture proportion of 0.2. Each simulation had two nearby selected sites under additive selection. We simulated three variations of the selection coefficients of these two sites: one in which both selection coefficients were 0.01, one in which both were 0.05, and one in which one site had 0.01 and the other had 0.05. On top of these selective variations, we also varied the distance between the sites from half a centimorgan to five centimorgans, making a total of 12 parameter variations. We simulated 20 introgressions for each parameter variation. We performed a two site hypothesis test between the two selected sites, which is where one would most likely see a signal of selection. The two site hypothesis test produces a likelihood for both an optimized one site model, and an optimized two site model, both with additive selection. AHMM-MLS was able to reject the single site hypothesis at a rate of at least 95% in 11 out the 12 parameter variations (Fig. 1). Our method did especially well in cases where the selection coefficient for both sites was 0.01, rejecting the single site hypothesis in all 80 cases. Our method failed in the case where the two sites were a centimorgan away and one of the sites had a high selection coefficient. Over these simulations, AHMM-MLS was able to accurately infer the location of both selected sites in cases where the two selection coefficients were similar, or the two sites were at least 2 centimorgans away (Fig. 2). In the cases where the selection coefficients differed and the two sites were closer than 2 centimorgans, our method would sometimes misplace the site with lesser selection.

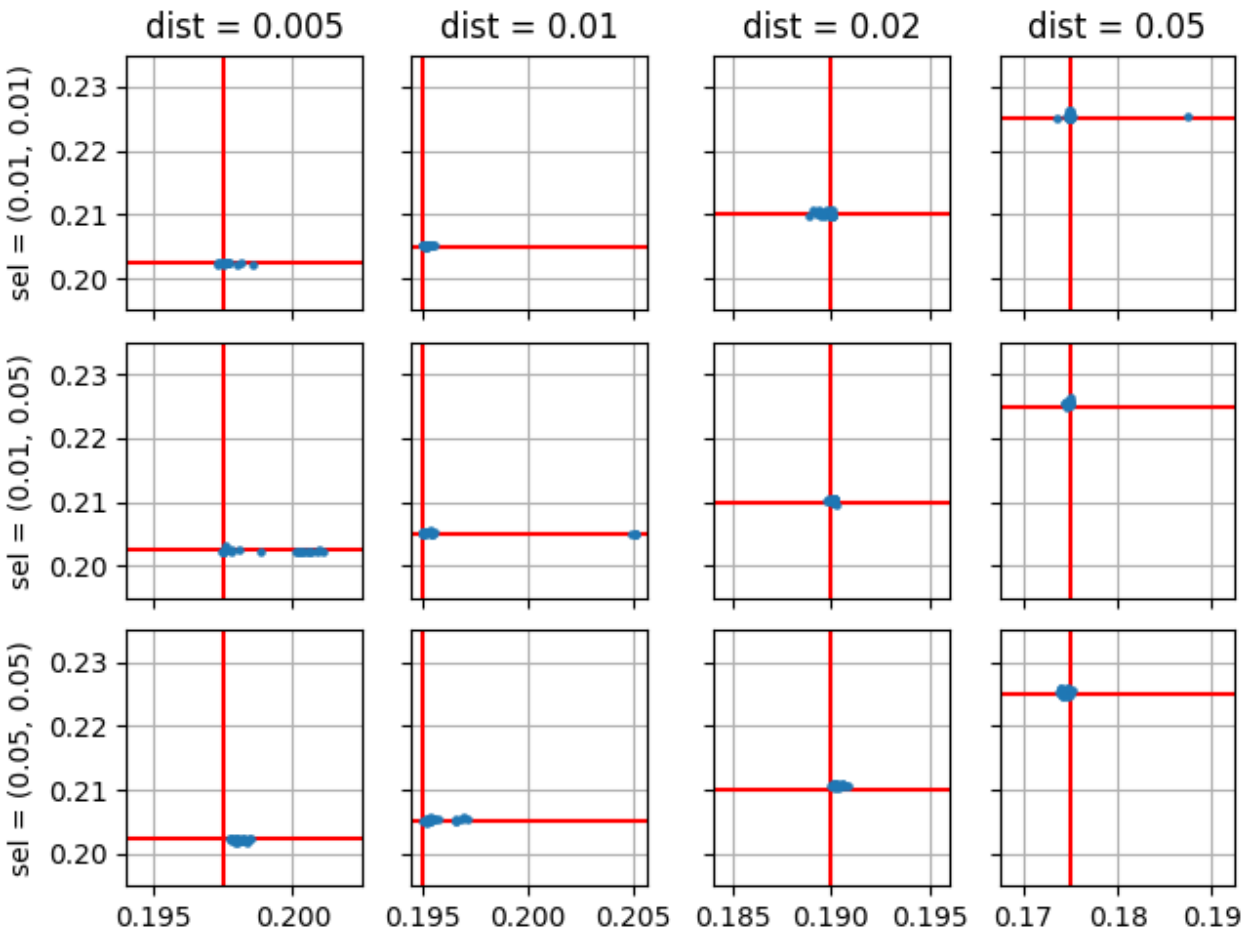

Figure 2. The ability of AHMM-MLS to infer the location of two nearby selected sites was evaluated for a variety of adaptive scenarios. Each graph shows a different adaptive scenario, with the x and y axis being the locations of the selected sites in morgans. In the cases where the selection coefficients of the two sites differed, the site placed below 0.2 morgans had the lower selection coefficient. Each blue dot within a graph corresponds to running AHMM-MLS on a distinct simulation, and shows where AHMM-MLS inferred the sites to be located. The red lines show the true locations of the selected sites.

Optimizing for a single selected site when the true population had two nearby selected sites causes an overestimation of the selection coefficients (Fig. 3). This effect is especially prominent in cases where the sites are at most 2 centimorgans away. This effect disappears when the selected sites are 5 centimorgans away, unless both sites are strongly selected. While the two site model did not always accurately estimate the strength of selection for both selected sites, it did generally perform better than when we fit a model with a single site.

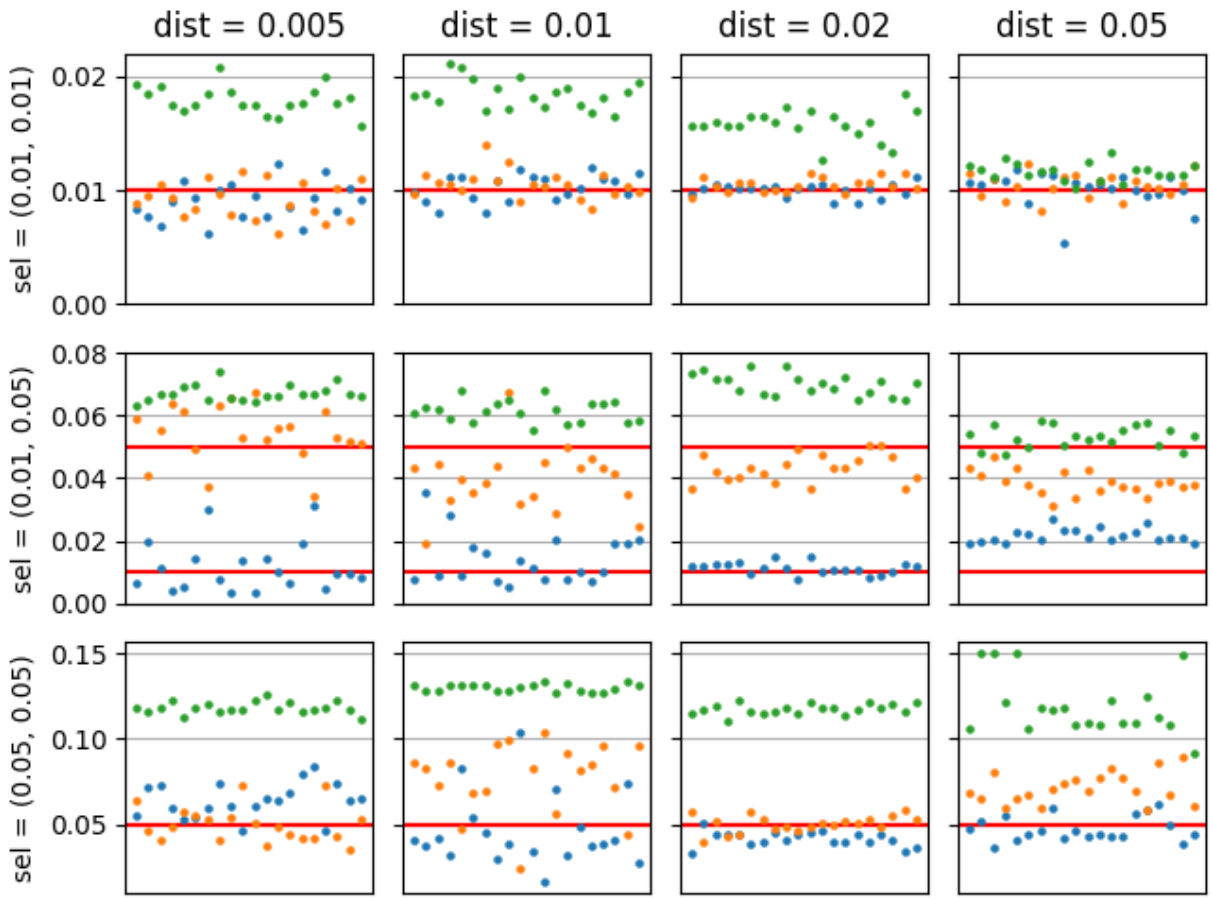

Figure 3. Comparison of the inferred selection coefficients when optimizing for two selected sites versus a single selected site. The orange and blue dots are the inferred selection coefficients of the two sites when AHMM-MLS is optimizing for both sites. The green dots show the inferred selection coefficient of the single site when AHMM-MLS is optimizing for a single site. The red lines are the true selection coefficients of the sites in the simulation.

AHMM-MLS can accurately find the selective strength of a site under dominant selection over a variety of scenarios of adaptive introgression. We simulated a range of admixed populations in which the introgressing population carried a dominantly selected allele. The populations vary in the selection coefficient of the dominant site, in the admixture proportion, and in the number of generations since admixture. We varied the selection coefficient from 0.005 to 0.05, the admixture proportion from 0.05 to 0.5, and the generations since admixture from 100 to 1000. These simulations spanned a total of 64 dominant introgression scenarios, with 20 simulations for each scenario. We optimized for a site with dominant selection in each of the 1280 simulations, and we examined the inferred selection coefficient at this site. AHMM-MLS was able to accurately infer the selection coefficient for most of the admixture conditions (Fig. 4). Our method generally did better in the populations where the selective strengths were stronger, the admixture proportion was lower, and the number of generations since admixture was higher.

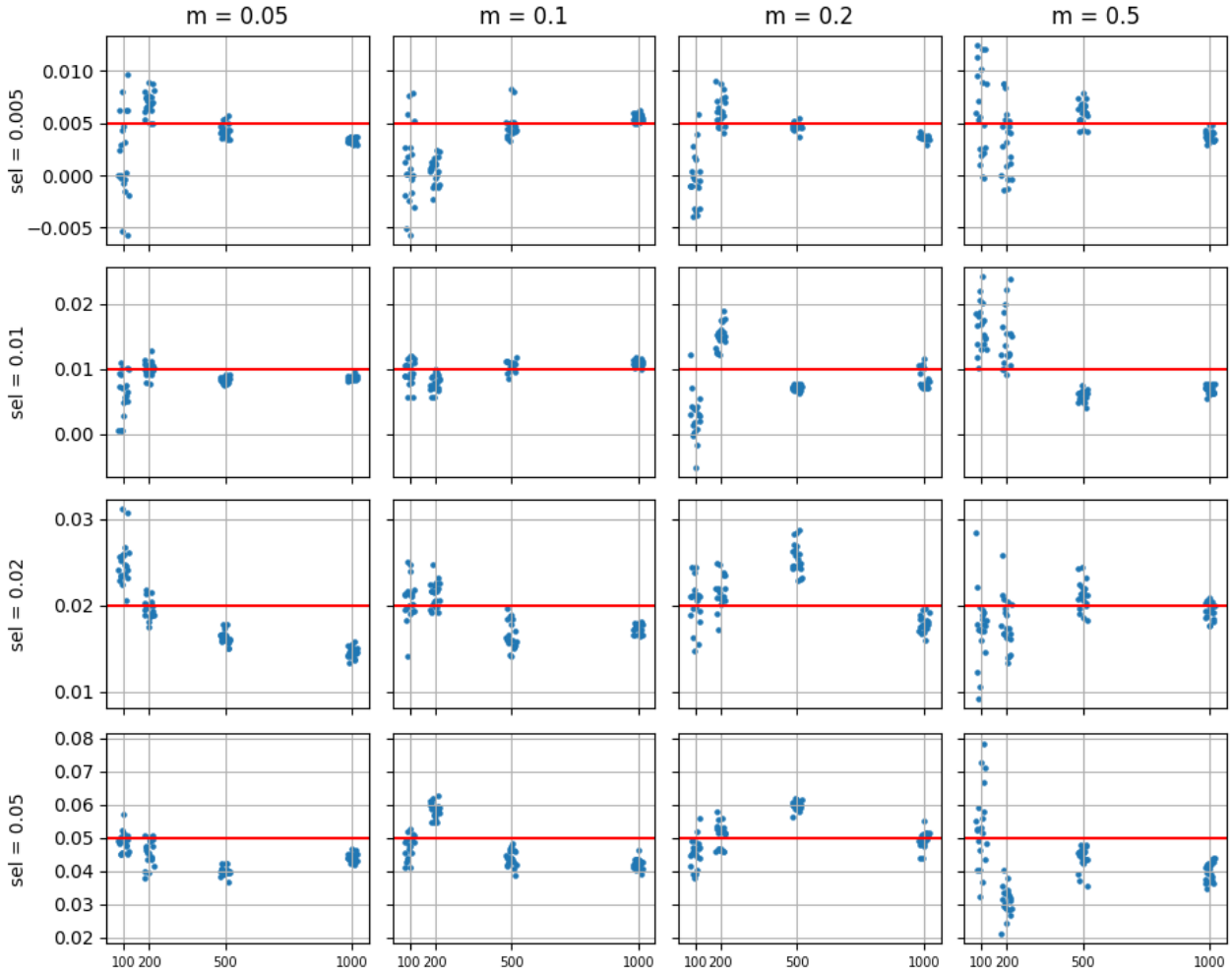

Figure 4. AHMM-MLS is able to infer the selection coefficient of a locus under dominant selection over a variety of adaptive introgression scenarios. For each of these tests, we simulated a single site under dominant selection, and used AHMM-MLS to optimize for its strength of selection. Each column of graphs represents a different admixture proportion, and each row represents a different selection coefficient for the dominant site. Within each graph, the x-axis shows the number of generations since admixture and the y-axis represents the selection coefficient of the dominant site. Each blue dot represents the selection coefficient found by one of the 1280 runs of AHMM-MLS, and the red line marks the true selection coefficient of the simulation.

We ran a dominance hypothesis test at the dominant site for each of the 1280 simulations, and found that our method requires certain population conditions in order to accurately reject the additive hypothesis (Fig. 5). AHMM-MLS produced likelihoods that justify rejecting the additive hypothesis at a rate of at least 95% in only 20 of the 64 dominant admixture scenarios. Our method did poorly in cases where the selective strength was weak, the admixture proportion was low, or the number of generations since admixture was low.

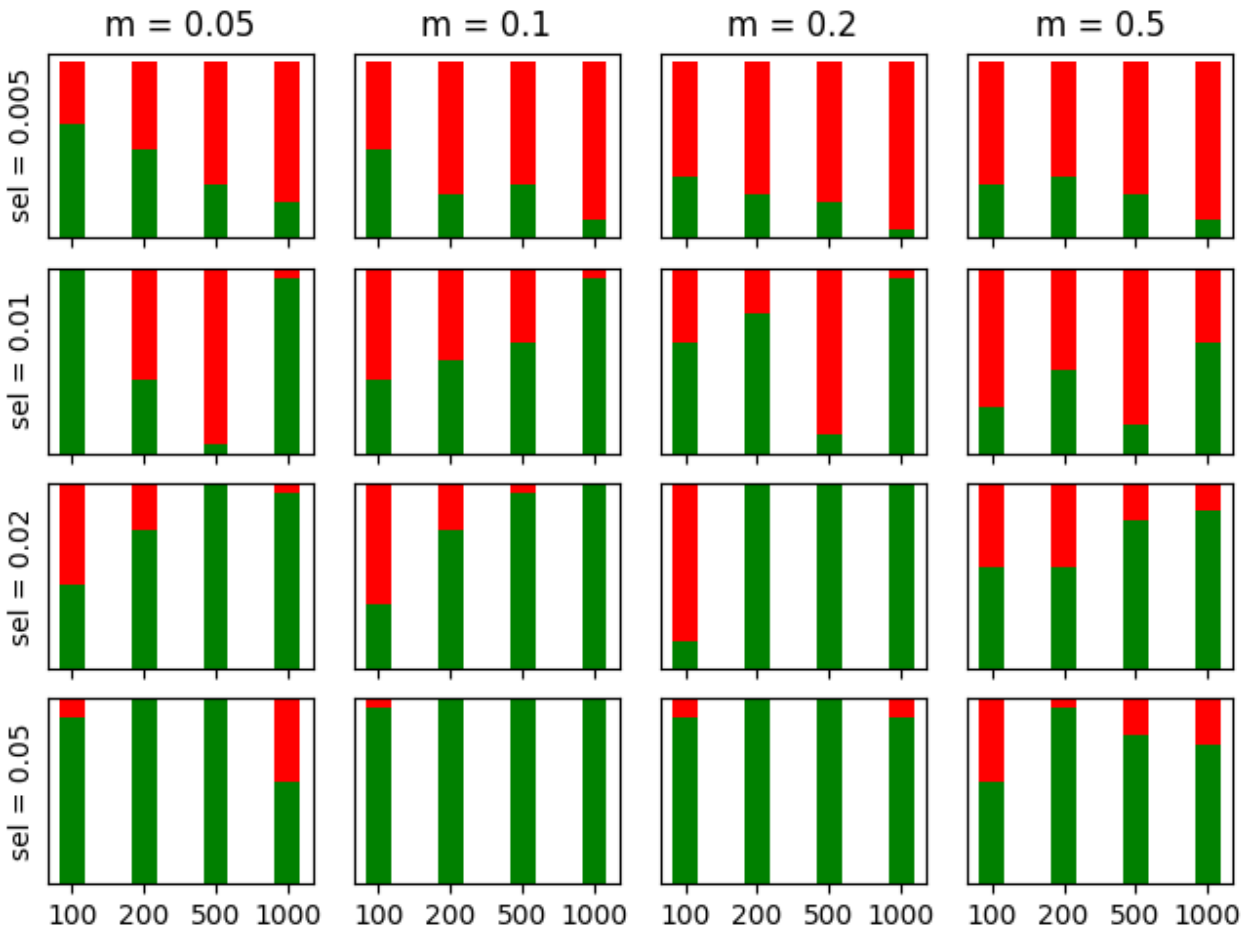

Figure 5. We performed additive vs dominance hypothesis tests over the same variety of scenarios as Figure 4. Each bar represents a single adaptive introgression scenario. For each simulation, we used AHMM-MLS to optimize for the likelihood while restricting to dominant selection, and to optimize for the likelihood while restricting to additive selection. We show in green the cases where AHMM-MLS produced likelihoods that justify rejecting the additive hypothesis, and show in red the cases where it failed to do so.

2.3 Applying AHMM-MLS to Chromosome 3R of D. Melanogaster

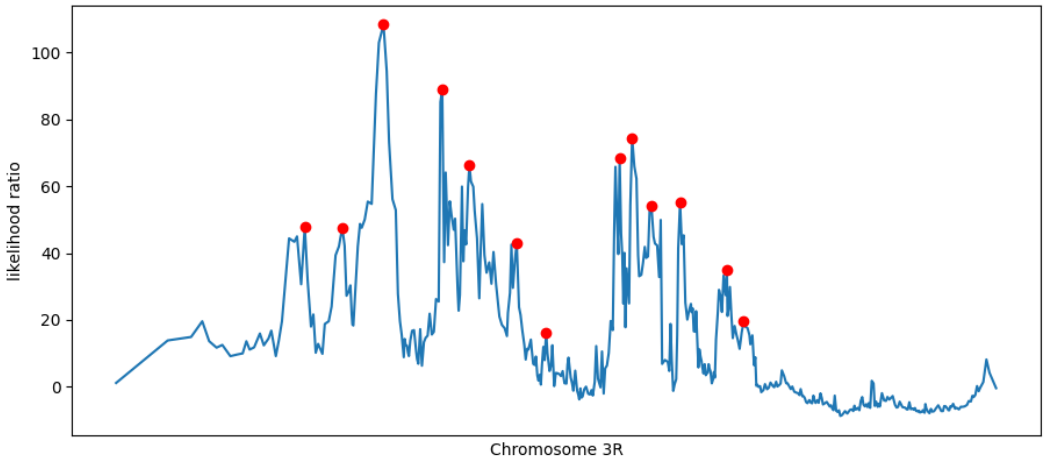

Figure 6. Chromosome 3R of *D. Melanogaster* shows signs of many nearby selected sites. Here we graph the likelihood ratio outputs of AHMM-S, which test each site for additive selection. Using a simple peak finding algorithm, we can identify 13 loci that show evidence of selection.

In order to validate AHMM-MLS on real data, we apply multiple two site hypothesis tests and dominance tests to chromosome 3R of *D. Melanogaster*. The population that we ran the tests on has shown signals of admixture in previous studies (Lack et al. 2015; Corbett-Detig and Nielsen 2017; Medina et al. 2018). The admixture history is consistent with a one-pulse model, with admixture parameters ($m = 0.17, t = 430$) that make this population suitable for our program (Corbett-Detig and Nielsen 2017). In a previous study, we found evidence that suggests chromosome 3R may have multiple nearby loci under selection (Svedberg et al. 2021). This study found 13 putative loci under selection on 3R, most of which were less than 5 centimorgans away from another selected site (Fig. 6). We have shown that estimating the selection coefficient of a single site which is less than 5 centimorgans away from another leads to an overestimation (Fig. 3). By fitting two nearby sites, we hope to correct for this overestimation and provide a more accurate selection coefficient.

After applying multiple two-site hypothesis tests to chromosome 3R, we have found that a two-site model is a significant improvement over a single-site model, as determined by a likelihood ratio test (Neyman and Pearson 1928). We have also found that estimations of the selection coefficients of the sites on 3R using our previous method, AHMM-S, may be overestimated because they only fit a single site (Svedberg et al. 2021). The selection coefficients estimated by AHMM-S were 25% to 67% higher than those found by our method. In Table 1 we compare our inferred selection coefficients with those inferred using AHMM-S, along with showing the approximate location of the sites our method found. We ran five two site hypothesis tests on 3R, centering our site guesses on locations that appeared to have multiple

nearby selected sites. The results of three of the five tests are shown below, each test inferring two additional sites. Of the two tests not shown, one of them found the same sites as another test, and the other only found a single site.

Table 1.

Approximate Position of selected site (bp)	Selection Coefficient estimated by AHMM-MLS	Selection Coefficient estimated by AHMM-S	Log Likelihood difference vs single site optimization	P-value of two site hypothesis test
12,136,402	0.0052	0.0073	6.588	0.0014
13,186,732	0.0068	0.0113	6.588	0.0014
15,047,725	0.0072	0.0106	36.311	< 0.00001
15,822,119	0.0059	0.0085	36.311	< 0.00001
20,399,330	0.0070	0.0101	12.247	< 0.00001
21,257,078	0.0068	0.0085	12.247	< 0.00001

After applying multiple dominance hypothesis tests on 3R, we found three sites that may have dominance. These three sites were each relatively close to one of the six sites found by the two-site tests. As with the two-site tests, we ran five dominance hypothesis tests over 3R, each on a location that showed interesting behavior of selection. Only three of the tests produced a significant likelihood ratio in favor of a dominance model over an additive model. We have shown that a two-site model can accurately predict both locations of nearby selected sites (Fig. 2), so the locations of these putatively dominant positions would be better estimated by looking for the nearest site found by the two-site hypothesis tests.

Table 2.

Approximate Position of selected site (bp)	Log Likelihood difference between dominance fitting and additive fitting	Relative likelihood of dominance model over additive model
13,173,785	3.905155	49.6578
15,047,725	20.155896	> 1,000
20,823,315	9.908762	> 1,000

3 METHODS

A diplotype to haplotype transformation matrix, $\boldsymbol{M}$, is generated every time the transition rates between two neutral sites need to be calculated, and is used for every generation of the numerical computation of those transition rates. The matrix depends on the location of the two neutral sites, the location of every selected site, and the fitness coefficients of those selected sites. The location is represented as a single value, l, the position on the chromosome in morgans. The fitness is represented by two values, f and e, where f is the fitness of the first homozygote, relative to the heterozygote, and e is the fitness of the second homozygote, relative to the heterozygote. This would make the selection coefficients f, 1, e. To simplify calculations, we treat the neutral sites as selected sites where $f = e = 1$.

To generate $\boldsymbol{M}$, we iterate through all possible diploids. For a particular diploid i, it has an associated fitness S, which is computed by taking the product of the relevant selection coefficients for each site. For this diploid i, we iterated through each region where a recombination event may occur. If there are n tracked sites, then there are $n + 1$ regions. Each of these regions has a corresponding recombination rate r, which is calculated by taking the difference of the morgan positions of the two ends of the regions. A recombination in a specific region would produce two haplotypes, k and l. So if there were a recombination event in this region with this diplotype, then there would be a contribution proportional to $D_i * S * r$ to both haplotypes. This contribution is reflected in $\boldsymbol{M}$ by adding $S * r$ to the $\boldsymbol{M}_{i,k}$ and $\boldsymbol{M}_{i,l}$ entries. $\boldsymbol{M}$ is fully computed after we have iterated through all possible diploids, and added their contributions to each haploid that they may produce in a recombination event.

3.1 Nelder-Mead Optimization

We optimize the log likelihoods of a selective model using the Nelder Mead algorithm (Nelder and Mead 1965). We used a reflection constant of 1, a contraction constant of 0.5, an expansion constant of 2, and a shrinkage constant of 0.5. Each Nelder Mead optimization begins with a simplex centered around a starting point. By default each point in the simplex is placed equally far from the starting point, but this may be altered so that points have a different distance from the starting point in a particular dimension. This allows a search to have a greater extent in a particular dimension. A Nelder Mead search is terminated when the range of log likelihoods of the simplex points drops below a certain threshold. It is also terminated if four shrinkage transformations occur in a row, to prevent cases where many expensive shrinkages happen in a row.

To reduce the time taken to optimize, we don't calculate the effects of a selected site across the whole chromosome. We only calculate the effects of selected sites on transition rates between two neutral sites if those neutral sites are less than five centimorgans away from a selected site. This speeds up the calculation of the likelihood of a specific model of selection. Once a model of selection is optimized, we calculate the effects of the selected sites across the entire chromosomes to produce the likelihoods that the program will output.

3.2 Multi Level Optimizations

For any log likelihood optimization, guesses as to where selected sites may be located must be supplied by the user. Each instance of log likelihood optimization is done in two parts, that we term ‘bottlenecks’. Each bottleneck consists of two search stages, which have different conditions for the initial simplexes and for terminating the searches. A particular search stage is parameterized by the ‘height’ and ‘width’ of the starting simplexes, the ‘depth’ of the searches, and the number of individual searches to be performed with those parameters. The height of a simplex determines its extent in the selection coefficient space. A larger height means that the search stage will sample selection coefficients which deviate further from neutral. The width of a simplex is its extent in the location dimension, measured in morgans. A wider width means that the search stage will sample site locations which deviate further along the chromosome from the initial guess. The depth of a search determines when a particular search will terminate. A search will terminate when the range of log likelihoods for the points in the simplex drop below the depth threshold. To ensure that each search samples new points, the simplexes of individual searches within a search stage are reflected around the simplex center for each dimension with a 50% probability.

The first bottleneck is the shallow bottleneck, in which shallow searches are performed with the Nelder Mead simplexes centered around the sites of interest with neutral selections. We term the two stages of the shallow bottleneck as ‘shallow short’ and ‘shallow tall’. Both shallow search stages have a width of 0.01, and a depth of 20. Shallow short has a height of 0.01, and shallow tall has a height of 0.05. Having these two separate stages allows the optimization to span a wider breadth of selective strengths. The second bottleneck is the deep bottleneck, with simplexes centered around the best point found in the entire shallow bottleneck. Like the shallow bottleneck, the deep bottleneck has a short stage and a tall stage. The short stage has a height of 0.005 while the tall stage has a height of 0.01. Both deep stages have a width of 0.005 and a depth of 5. Each stage in both bottlenecks consists of five individual Nelder Mead searches. The best point found in the deep bottleneck is the final result of the optimization.

3.3 Dominance vs Additive Hypothesis Testing

When testing the hypothesis that a particular point is under selection with dominance, we perform three distinct multi-level optimizations. These three optimizations correspond to restricting the selection space to either the first allele being dominant, the second allele being dominant, or additive selection. After these three optimizations, the resulting log likelihoods are compared to determine which model best fits the data. Neither of these models are nested with another, so an AIC test must be performed to justify rejecting a model in favor of another one (Akaike 1973).

3.4 Two Site vs Single Site Hypothesis Testing

When testing the hypothesis that a signal of selection is caused by two nearby sites rather than a single site, we start with a guess as to the general location of this signal, and then we perform two multi-level optimizations. The first optimization centers its initial simplexes around two selected sites a centimorgan away from the initial guess on both sides, this is the model where the signal is caused by two nearby selected sites. The second optimization centers its initial simplexes around a single selected site at the initial guess. AHMM-MLS allows for two types of two site hypothesis testing, one which restricts the selection space to additive selection, and one which leaves the selection space unrestricted. The two site model and the single site model are nested, so a likelihood ratio test can be used to compare the fit of each model.

3.5 Robustness of Dominance Testing to Parameter Misspecification

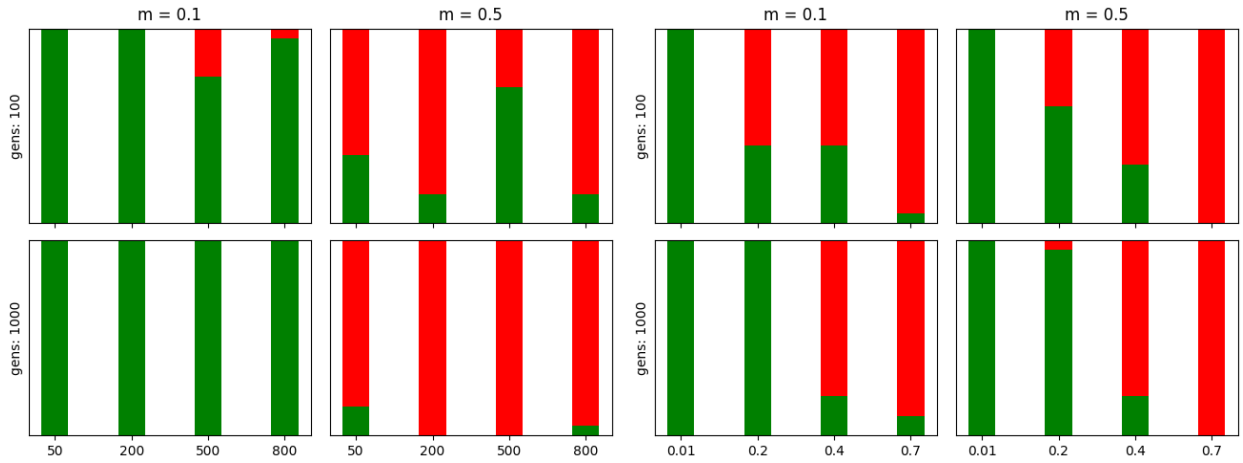

Figure 7. We performed additive vs dominance hypothesis testing while misspecifying certain admixture parameters. We simulated 20 instances of four scenarios of introgression with a single locus under dominant selection. For these scenarios we varied the admixture proportion from 0.1 to 0.5, and the number of generations since admixture from 100 to 1000, while keeping the selection coefficient constant at 0.02. For each scenario of introgression, we ran an additive vs dominance hypothesis test while misspecifying the admixture parameters in eight different ways. On the left, we misspecified the number of generations since admixture. On the right, we misspecified the initial admixture proportion.

AHMM-MLS dominance hypothesis testing is severely hindered when misspecifying the admixture proportion, and is hindered by misspecifying the time since admixture when the admixture proportion is high (Fig. 6). We ran 80 forward simulations of admixture in which the introgressing population carried a dominantly selected allele. The populations had an admixture proportion of either 0.1 or 0.5, and the generations since admixture was either 100 or 1000. The dominant site in each population had a selection coefficient of 0.02, making a total of four different admixture scenarios. We ran 8 dominance hypothesis tests on each of the 80 simulations, in which we either misspecified the admixture proportion or the time since admixture. The admixture misspecifications ranged from 0.01 to 0.7, and the generations since admixture misspecifications ranged from 50 to 800.

3.6 Drosophila Data and External Tools and Libraries

We used publicly available datasets of *D. Melanogaster* collected from South Africa (Lack et al. 2016). In a previous study by our lab, the data was prepared so that it could be analyzed by the AHMM programs (Svedberg et al. 2021). This included removing the

chromosomal inversions found on some of the chromosome arms (Corbett-Detig and Hartl 2012). We used a publicly available fine-scale recombination map of chromosome 3R both for our program, and to locate the approximate base pair location of the detected selected sites (Comeron et al. 2012).

We used the Armadillo c++ linear algebra library to speed up the computation of matrix multiplications (Sanderson and Curtin 2016; Sanderson and Curtin 2020). We used GNU parallel to run many batches of simulations at once (Tange 2018). We used the FlyBase sequence coordinates converter to convert assembly 5 base pair coordinates to assembly 6 (Larkin et al. 2021).

4 GITHUB

The code used to generate the simulated data, as well graph the results, can be found here:

https://github.com/genicos/ahmmmls_sim_analysis

Ancestry_HMM-MLS can be downloaded here:

https://github.com/genicos/ahmm_mls

5 REFERENCES

1. Corbett-Detig, R. & Nielsen, R. *A hidden Markov model approach for simultaneously estimating local ancestry and admixture time using next generation sequence data in samples of arbitrary ploidy*. http://biorxiv.org/lookup/doi/10.1101/064238 (2016) doi:10.1101/064238.
2. Nelder, J. A. & Mead, R. A Simplex Method for Function Minimization. *The Computer Journal* **7**, 308–313 (1965).
3. Lack, J. B., Lange, J. D., Tang, A. D., Corbett-Detig, R. B. & Pool, J. E. A Thousand Fly Genomes: An Expanded *Drosophila* Genome Nexus. *Mol Biol Evol* **33**, 3308–3313 (2016).

4. Oziolor, E. M. *et al.* Adaptive introgression enables evolutionary rescue from extreme environmental pollution. *Science* **364**, 455–457 (2019).
5. Whitney, K. D., Randell, R. A. & Rieseberg, L. H. Adaptive Introgression of Herbivore Resistance Traits in the Weedy Sunflower *Helianthus annuus*. *The American Naturalist* **167**, 794–807 (2006).
6. Sanderson, C. & Curtin, R. An Adaptive Solver for Systems of Linear Equations. in *2020 14th International Conference on Signal Processing and Communication Systems (ICSPCS)* 1–6 (IEEE, 2020). doi:10.1109/ICSPCS50536.2020.9309998.
7. Racimo, F. *et al.* Archaic adaptive introgression in *TBX15/WARS2*. *Mol Biol Evol* msw283 (2016) doi:10.1093/molbev/msw283.
8. Sanderson, C. & Curtin, R. Armadillo: a template-based C++ library for linear algebra. *JOSS* **1**, 26 (2016).
9. Guan, Y. Detecting Structure of Haplotypes and Local Ancestry. *Genetics* **196**, 625–642 (2014).
10. Sankararaman, S., Sridhar, S., Kimmel, G. & Halperin, E. Estimating Local Ancestry in Admixed Populations. *The American Journal of Human Genetics* **82**, 290–303 (2008).
11. Medina, P., Thornlow, B., Nielsen, R. & Corbett-Detig, R. Estimating the Timing of Multiple Admixture Pulses During Local Ancestry Inference. *Genetics* **210**, 1089–1107 (2018).
12. Larkin, A. *et al.* FlyBase: updates to the *Drosophila melanogaster* knowledge base. *Nucleic Acids Research* **49**, D899–D907 (2021).
13. Tange, O. *Gnu Parallel 2018*. (Zenodo, 2018). doi:10.5281/ZENODO.1146014.
14. Svedberg, J., Shchur, V., Reinman, S., Nielsen, R. & Corbett-Detig, R. *Inferring Adaptive Introgression Using Hidden Markov Models*.

http://biorxiv.org/lookup/doi/10.1101/2020.08.02.232934 (2020) doi:10.1101/2020.08.02.232934.

15. Akaike, H. Information Theory and an Extension of the Maximum Likelihood Principle. in *Selected Papers of Hirotugu Akaike* (eds. Parzen, E., Tanabe, K. & Kitagawa, G.) 199–213 (Springer New York, 1998). doi:10.1007/978-1-4612-1694-0_15.
16. Neyman, J. & Pearson, E. S. ON THE USE AND INTERPRETATION OF CERTAIN TEST CRITERIA FOR PURPOSES OF STATISTICAL INFERENCE PART I. *Biometrika* **20A**, 175–240 (1928).
17. Corbett-Detig, R. B. & Hartl, D. L. Population Genomics of Inversion Polymorphisms in Drosophila melanogaster. *PLoS Genet* **8**, e1003056 (2012).
18. Zhou, Q., Zhao, L. & Guan, Y. Strong Selection at MHC in Mexicans since Admixture. *PLoS Genet* **12**, e1005847 (2016).
19. Lack, J. B. *et al.* The *Drosophila* Genome Nexus: A Population Genomic Resource of 623 *Drosophila melanogaster* Genomes, Including 197 from a Single Ancestral Range Population. *Genetics* **199**, 1229–1241 (2015).
20. Hill, W. G. & Robertson, A. The effect of linkage on limits to artificial selection. *Genet. Res.* **8**, 269–294 (1966).
21. Comeron, J. M., Ratnappan, R. & Bailin, S. The Many Landscapes of Recombination in Drosophila melanogaster. *PLoS Genet* **8**, e1002905 (2012).