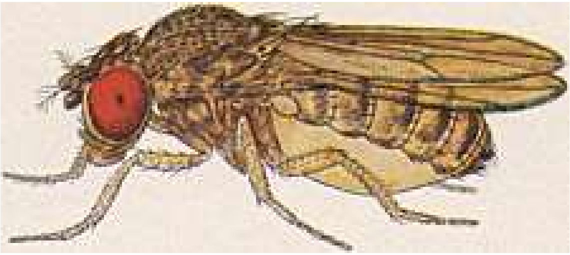
Understanding the decay of heterozygosity in Drosophila



Genetics Honours Project Report

B122250

Word Count: 4998

Table of contents

Abstract

The effect of inbreeding on the decay of heterozygosity is influenced by a population's mutational load. Crossovers during meiosis, as well as selection results in heterozygosity being present in tracts following several full-sib-mating generations. Inbred lines with greater mutational loads experience a slower heterozygosity decay, as recessive lethal balancing systems are set up that retain heterozygosity levels. Chromosomal inversions that are present in a heterozygous state can also inhibit or reduce this decay in heterozygosity. In this project, the observed decay in heterozygosity of 3 different allopatric *D.mojavensis* races is examined following a sib-mating experiment. The heterozygosity proportions, as well as their distribution in tracts is compared against various simulation models, which simulate recessive lethal balancing systems, as well as chromosomal inversions. I find that the heterozygosity decay is slower than the neutrality predictions, as well as identifying different regions with abnormal heterozygous tracts. I identify an overlapping tract in chromosome 2 of the Sonora lines, as well as a large tract in chromosome 3 of one of the Baja lines, on the same position as previously reported polymorphic chromosomal inversions.

1. Introduction

The fitness of an individual depends on both the heterozygous fitness effect of an allele, as well as its homozygous effect. Thus, the way selection acts on an allele depends on its dominance coefficient, and therefore will have a stronger effect on dominant alleles(Haldane, JBS., 1927). It is therefore expected for a wild population to contain recessive deleterious alleles, whose effect is unmasked in those individuals which are most inbred and have greater homozygosity proportions.

Because of recombination events, sequences that are identical by descent/homozygous will come in tracts, with junctions representing recombination breakpoints. At the end of a generation, tracts that do not coalesce in previous generations are heterozygous. Fisher, R. A, 1954 first studies these tracts, and their junctions, to mathematically predict their expected number and lengths under neutrality after n generations of sib-mating(**Figure 1**).

Furthermore, in many cases, complete homozygosity cannot be achieved because of deleterious or lethal recessive alleles that lead to inbreeding depression when present in a homozygous state(Charlesworth B, Charlesworth D, 1999).

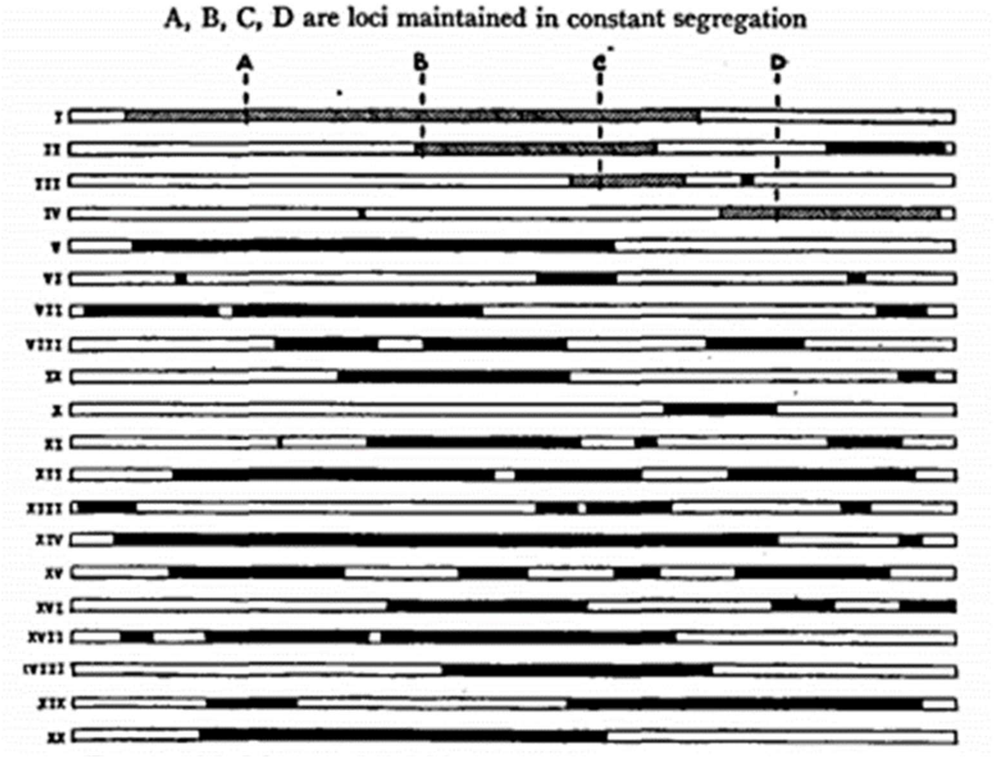


Figure 1. Model map of heterozygous tracts (black) of an organism with 20 chromosomes after 5 generations of inbreeding. From Fisher, 1949.

A single recessive lethal allele is easily acted upon by selection and is quickly eliminated from an inbred line. However, if a different lethal is present on the opposite homologue, it would balance the first one and increase the possibility of survival for both(Gowen et al., 1946, Bennet J., 1956). Under this scenario, heterozygous tracts can remain for many generations despite subsequent sib-mating, as the genetic load that results lethal when homozygosed must be kept in a heterozygous state. A similar, but more extreme pattern can be observed when a chromosomal inversion appears in a heterozygous state, as it can drastically reduce crossover frequencies, first observed by Sturtevant AH., 1917 in *D.melanogaster*.

Chromosomal inversions are mutations in which a section of a chromosome is inverted. An inversion that is present in a heterozygous state, can supress recombination as it creates genetically unbalanced gametes and therefore unviable offspring(Kirkpatrick, M., 2010). A genetic load in an inversion can easily form a balancing system. However, in rare occasions, recombination can still occur in the shape of double cross-overs, or as gene conversion(Andolfatto, P. et al., 2001).

D.mojavensis is a well-known member of the Drosophila lineage, with a well characterised, and readily available genome(Clark, AG. et al., 2007). Individuals belonging to this species thrive under the extreme conditions present in the deserts of North-western México, and Southern California/Arizona. These insects rely on dead tissue from cacti, where they feed and develop as larvae(Allan, CW. & Matzkin, LM., 2019). There are different host races with known differences in ecology, as well as genetic differences, where they differ in fixed and polymorphic chromosomal inversions. The original population, found in Baja California is one the races studied in this project, together with the Sonora race, and a race from the Santa Catalina island. The Santa Catalina race became isolated and suffered from a bottleneck, leaving them with the lowest diversity(Benowitz, KM., 2019). The Sonora race split from the Baja California race during the Pleistocene and has not had any significant contact with the Baja race since(Smith, G. et al., 2012). These factors contribute towards creating a strong genetic structure within this species, making them the subject of ongoing divergence and speciation studies.

This study uses sequencing data produced for a different study on how inversions play a role in speciation(Lohse, K., et al., 2015). For this study, different lines corresponding to the *D.mojavensis* races described above were sib-mated to minimise heterozygosity. Large tracts of heterozygosity were observed, sparking the idea that the genetic load on each line might be impeding the decay in heterozygosity. Here, I use data from different simulation models, on the size and frequency of heterozygous tracts. Comparing this against the size and frequency of the heterozygous tracts observed on the different lines, we can better understand the deviation in heterozygosity proportions from the neutral predictions.

3. Materials and methods

3.1 Raw sequenced data

3.1.1 Samples

This experiment includes data for five different lines of Drosophila mojavensis: SC05, A900, A975, A976, PO88. Lines A975 and A976 were collected from Baja California, A900 and PO88 from the Sonora desert, and SC05 samples from the small island of Santa Catalina(Figure S1 and Table 1, in appendix). These lines were maintained in large population sizes and fed with banana food until the time of the experiment, as explained by Lohse K. et al., 2015. Each line was sib-mated to maximize homozygous proportions. Surviving full-sib individuals were chosen for the next generation, for 10 consecutive generations. It was unclear if this meant one mating of a randomly chosen founding pair flies, followed by 9 generations of full sib-mating, or followed by 10 generations of sib-mating. To ensure the reduction in heterozygosity for the different lines was not overestimated, it was assumed that there had been 9 generations of sib-mating, which is conservative. After the sib-mating, one pool of 12 females for each line was sequenced to a 24- 29-fold coverage using Illumina 100bp paired-end reads(for details see Lohse K, et al., 2015).

3.1.2 Data processing previous to the start of the project

The trimmed reads for each inbred line were obtained from Lohse, K. et al., 2015. The reads were mapped to Version 1.04 of the D. mojavensis reference genome(Gramates LS., 2017) using the BWA-MEM algorithm (Li H., 2013).

Variants were called using FreeBayes(Garrison, E. & Marth, G., 2012) and the data subsequently filtered using the pre-processing module of gIMble(Laetch D.R.), to generate a VCF file. A minimum depth of 8 reads was required to produce high quality variant calls. Regions with a read depth greater than twice the standard deviation were also filtered out. Regions of excessive coverage can arise due to unresolved gene duplications, which could exaggerate the estimates of heterozygosity.

Finally, only regions that were covered across all lines were retained using Bfctools (Li H., 2011) on the VFC file to obtain the genotype calls for those regions. This resulted in a BED file that could be used to calculate the coverage, together with a text file containing all the heterozygous calls.

3.1.3 Visualising heterozygosity

Heterozygosity in each line was visualized in 10 kb windows along every autosome. For each window, heterozygosity was calculated by dividing the number of heterozygous sites, by 2*the number of sites that passed the filtering, ensuring that any coverage variations were accounted for.

3.1.4 Identifying heterozygosity tracts.

Drosophila males are achaismatic, with female Drosophila experiencing one obligate cross-over for every pair of bivalent chromosomes, leading to chromosomes of length 50cM (Lenormand T., 2016).

Assuming on average one cross-over per female generation and chromosome, residual heterozygosity is expected to occur in large tracts. Heterozygous tracts in each inbred line were determined by applying a simple threshold of heterozygosity for a given window.

Using sliding windows of 150 kb, with an offset of 30kb, reduced the variance in heterozygosity that could arise due to gaps in mapping, which would lead to inaccurate measures of heterozygosity. A range of heterozygosity thresholds were explored, and the proportion of heterozygosity remaining was calculated for each threshold. The variation in heterozygous tract numbers was also noted at different thresholds.

Once an appropriate threshold was determined, heterozygosity was measured for every 150kb sliding window, with the maximum heterozygosity chosen for the overlapping segments. Windows with a heterozygosity above the threshold were labelled as heterozygous, whereas those below were labelled homozygous. Multiple consecutive heterozygous windows were combined into single tracts.

3.1.5 Tract overlap

The overlap in heterozygosity tracts was calculated between pairs from the same host race (Baja California or Mainland Sonora) and between host races. Using a heterozygosity threshold of 0.0015, and sliding windows of 150kb, the tracts were identified on each chromosome, for each line. Since line SC05, was the only line from Santa Catalina, this sequence was excluded from the analyses of overlap. All chromosome lengths were normalised to 1 so that comparison between different chromosomes was possible. A bootstrap analysis was performed: each chromosome was compared against all other chromosomes to obtain a null distribution of pairwise overlap in heterozygosity. Comparing the overlap in heterozygosity for any pair of lines and chromosomes to this distribution allows testing whether there was a significant more overlap in heterozygosity tracts between closely related lines and unrelated lines than expected by chance.

3.2 SLiM simulations

The sib-mating experiment was simulated forwards in time under different scenarios, neutral and selective. For the simulations, version 3.6 of the evolutionary simulation framework SLiM, developed in the Messer Lab was used(Haller B.C. & Messer P.W., 2019). Simulations were set up using SLiM's own scripting language called Eidos, a hybrid between C and R programming languages.

3.2.1 Neutrality Simulations

Key features of Drosophila meiosis conditions were implemented into the simulations. To match Drosophila biology SLiM simulations were performed without male cross-overs and assuming a recombination rate of an average of 0.5 cross-overs per female gamete. In the first set of simulations, inbreeding due to repeated full-sib-mating was simulated under neutrality, without any selection, or mutations(Neutral). 100 000 replicate simulations, each consisting of a single chromosome were performed.

3.2.2 Comparing simulations against neutrality predictions

To ensure the code was running correctly, and that the data produced was accurate, different measures of the neutral simulations were checked. The decrease in average heterozygosity was compared against the analytic prediction(Lerner I. M., 1950). The distribution of total cross-overs per simulation was also checked against the expected cross-overs for that recombination rate and number of generations. This was done by adding up the number of breakpoints in each generation. With recombination supressed in male individuals, and a recombination rate that leads to an average of half a cross-over per female gamete, an average of 4.5 cross-over events are expected over 9 generations, which would average 90 crossover events in replicates of 20 chromosomes.

3.2.3 Heterozygous-lethal and recessive-lethal simulations

The next step was to implement an extreme case of heterozygote advantage into the models, whereby fitness decreases to zero if heterozygosity is lost. A similar simulation was set up, with the same crossing over constraints. However, to simulate this scenario, a mutation was added in the middle of the chromosome, in every individual. At the end of each generation, the mutation was checked, and if it were not present in a heterozygous state, the simulation would restart(Sim 2). This simulated the breakpoint on a chromosomal inversion, which is forced to remain heterozygous.

Simulations were also set up to test if individuals could escape the presence of recessive lethal mutations. This differs from the above scenario in that the mutation can be lost without losing fitness, or remain heterozygous, but if the mutation is present in a homozygous state, it results lethal. Sim 3 model tested the presence of a single recessive mutation present in one homologue. Sim 4 tested the effect of two recessive lethals at the positions representing either end of the large heterozygous tract seen in chromosome 2 of the Sonora lines. The mutations were added on opposite homologues to test the strength of a balancing system(**Figure 2**).

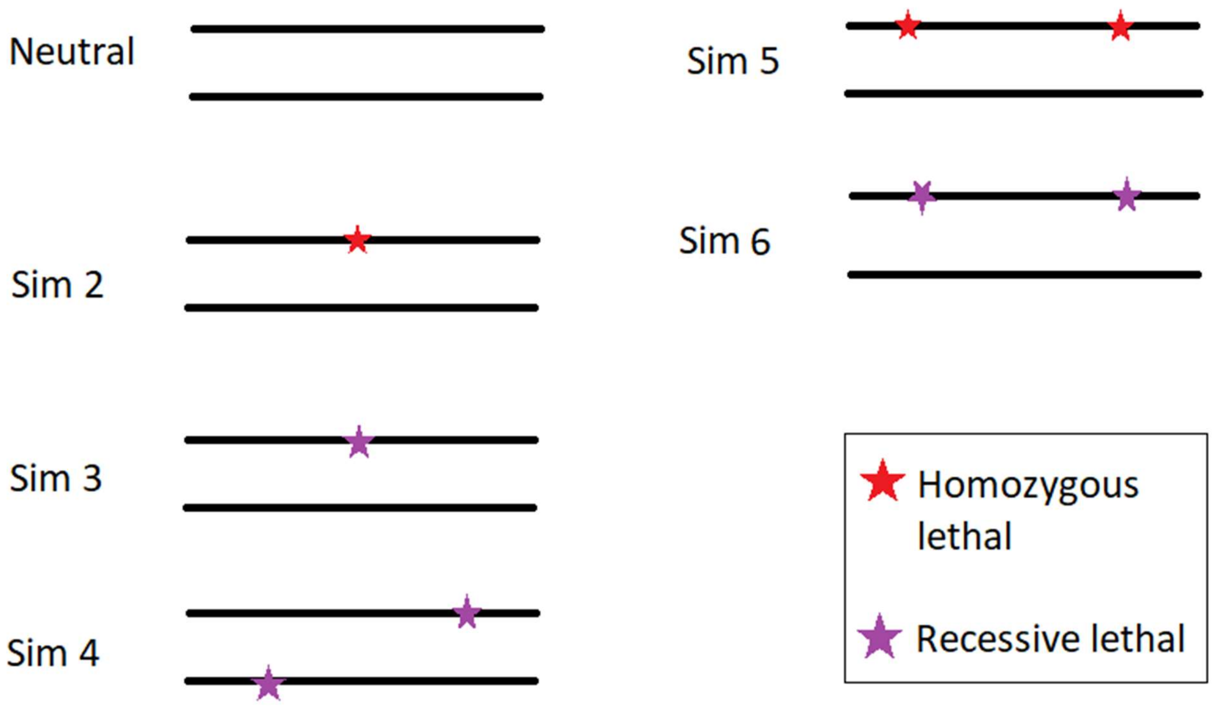


Figure 2. Simple schematic representation of the simulation models implemented in this project. Each haplotype is represented with a black line, with homozygous and recessive lethal mutations distinguished with either a red or a purple star. Homozygous lethal mutations differ from recessive lethal mutations in that they cannot be eliminated, the ancestral state with no mutations is as lethal as the mutation being present in a homozygous state, and therefore must remain heterozygous.

3.2.4 Calculating Heterozygous proportions

Samples of 20 simulations were used to record the distribution of remaining heterozygosity, as well as the distribution for the number and lengths of heterozygous tracts. Sample sizes of 20 simulations were chosen for the different analyses, as the raw data consisted of 4 chromosomes for 5 different lines.

The simulations were exported as tree-sequence files, which enables tracking the local ancestry at any given position in the genome. Each simulation was analysed using the msprime coalescent simulation package(Kelleher J., 2016). Sections of the genome that coalesced during the simulation were considered homozygous, whereas tracts that did not coalesce during the simulation were labelled as being heterozygous. Tree-sequence files record the generation at which each breakpoint occurred. This was used to measure the remaining proportion of heterozygosity on each generation. The simulation data was summarised in terms of average heterozygous fraction, and in number of heterozygous tracts.

3.2.5 Tract overlap

10 000 simulations were divided into groups of 2, to mimic the tract overlap comparison done by geographical location for the real data. The distribution of tract overlap between all simulations was recorded, for neutral simulations, simulations implementing the homozygous lethal models, and models containing recessive lethal alleles. The mean overlap was calculated and compared to the overlap seen on related lines.

4. Results

4.1 Neutral simulations match expectations

To ensure their reliability, two measures of the neutral simulations were compared against their expectations: the rate of loss of heterozygosity, and the expected number of cross-over events throughout the whole simulation.

Derived from Lerner I. M., 1950 is the expected rate of loss of heterozygosity under full-sib-mating:

$$f_n = \frac{1}{4}(1+2\, f_{n-1} + f_{n-2})$$

Figure 3 shows the recursive evaluation of this equation, together with the normalised plot for the proportion of remaining heterozygosity observed in SLiM, which shows nearly identical rates of heterozygosity loss.

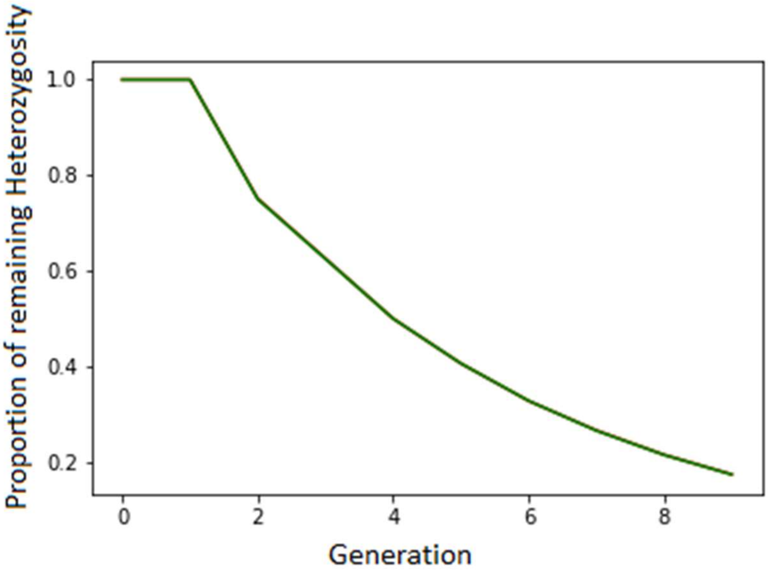


Figure 3. Recursive evaluation of the predicted loss in heterozygosity after n generations of sib-mating by Lerner I. M., 1950 (red, behind green line), and observed loss in heterozygosity on the neutral simulations (green). The prediction assumes that the initial individuals were completely unrelated and works independently of the recombination rate. This equation ignores new mutations, genetic linkage, and selection.

The number of tracts and therefore their overlap depends on the recombination rate, the second measure from the neutral simulations that was compared against expectations. The average number of recombination events was close to 90, as could be predicted with this recombination rate(**Figure S2**). This number is expected to be greater than the number of observed tracts, as crossover events can occur on already homozygous tracts.

Fisher, 1954, gives the following expression for the number of heterozygous tracts after n rounds of sib-mating:

$$\{1.45(n\text{-}3.19)L+v\}Z\epsilon^{n}$$

Where Z stands for:

$$\frac{1}{10}(5+16\epsilon)$$

ϵ stands for:

$$\frac{1}{4}(\sqrt{5}+1)$$

and L and v stand for the total map length in Morgans, and the haploid number of chromosomes, respectively. As written by Fisher, this equation is unreliable when n < 10 generations, so the number of tracts could not be checked under this model. A simulation was set up with 20 generations, and recombination was enabled in male individuals, as this would affect the observed number of tracts. The average number of tracts observed for 100 000 replicates was 0.033, which is analogous to the predicted number of tracts under this scenario (0.032).

4.2 Heterozygous proportions and number of tracts deviate from neutrality

Visualising the heterozygosity in windows of 10kb revealed tracts of conserved heterozygosity on top of a mostly homozygous background(**Figure 4**).

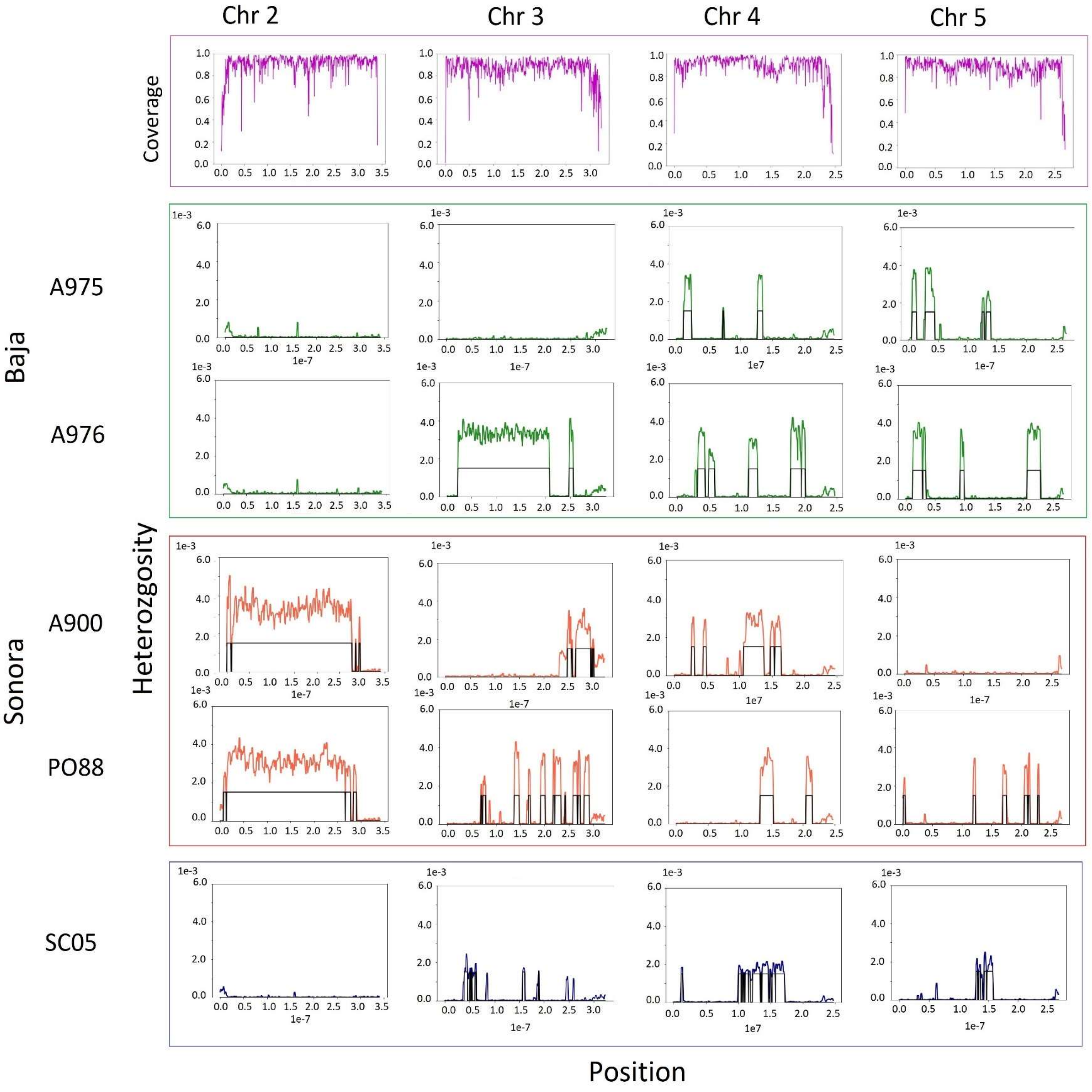


Figure 4. Heterozygosity levels and coverage throughout the *D.mojavensis* lines. Coverage is visualised for each chromosome (purple), and heterozygosity is visualised with a different colour for each host race (green for lines from Baja California, red for lines from the Sonora desert, and blue for the line from Santa Catalina island). Black lines denote the heterozygous tracts identified using a heterozygosity threshold of 0.0015.

These tracts were present on most chromosomes in all lines, although the heterozygous proportion differed between lines and chromosomes. When residual heterozygosity is retained, it is expected that there will be a plateau of remaining heterozygosity when examining remaining heterozygosity at different thresholds. This occurs as the heterozygous tracts are detected at the heterozygosity levels at the start of the experiment, allowing us to rule out different sources of noise such as mapping errors or areas with reduced coverage. The ideal threshold should lie within this plateau, where tracts that remained heterozygous throughout the sib-mating experiment can be detected. This plateau was observed, indicating that the ideal heterozygosity threshold for this data lied between 0.0005 and 0.003(**Figure 5**).

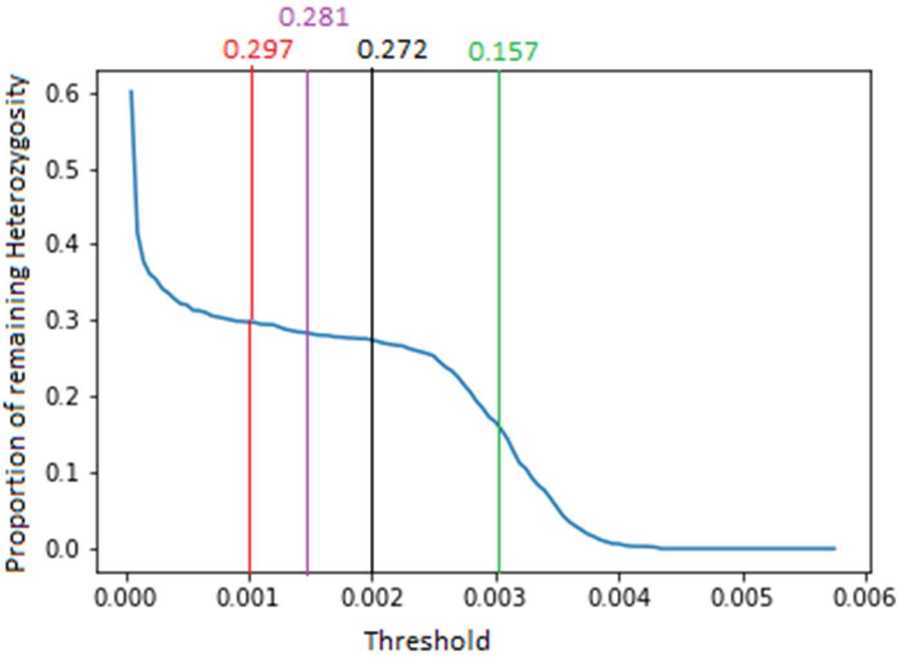


Figure 5. Proportion of remaining heterozygosity at different heterozygosity levels, calculated with 150kb sliding windows. Vertical lines show the proportion of heterozygosity remaining at different thresholds.

This chart was not sufficient to determine an ideal threshold. Too high of a threshold, and “real” tracts will split into smaller tracts due to variation in heterozygosity across windows, that might arise because of mapping errors. This creates an apparent rise in the number of tracts with increasing thresholds before it drops back down as the tracts are not detected, because the threshold is too high. Too low of a threshold, and the number of tracts will also start increasing, as weaker signals such as mutations arising during the experiment, or mapping errors are detected. This “trough” was observed when analysing the number of tracts at a given threshold(**Figure S3**). A threshold of 0.0015 was chosen to calculate the remaining heterozygosity after the inbreeding experiment.

With a heterozygosity threshold of 0.0015, 28.1% of all the chromosomes in all the lines remained heterozygous(**Figure 5**). This lies outside the 95th percentile observed in the neutral simulations (27.9%)(**Figure 6A**), which had an average heterozygosity of 17%.

Sim 2 scenario resulted in a significant increase in the proportion of conserved heterozygosity throughout 20 replicates (53%)(**Figure 6B**), one homozygous lethal mutation in every replicate could not explain the observed heterozygosity proportions. This model could best explain the observed heterozygosity when using 9 Sim 2 replicates out of 20 (28%), equivalent to 1.8 homozygous lethal mutations per genome, ignoring the sex chromosomes, which were not analysed in this experiment(**Figure 6E**).

Sim 3 scenario had similar heterozygous proportions to the neutral model(**Figures 6C**). Sim 4, with one recessive lethal allele on opposing haplotypes retained greater heterozygosity proportions (24%)(**Figure 6D**), but only resembled the data when applied to all 20 replicates.

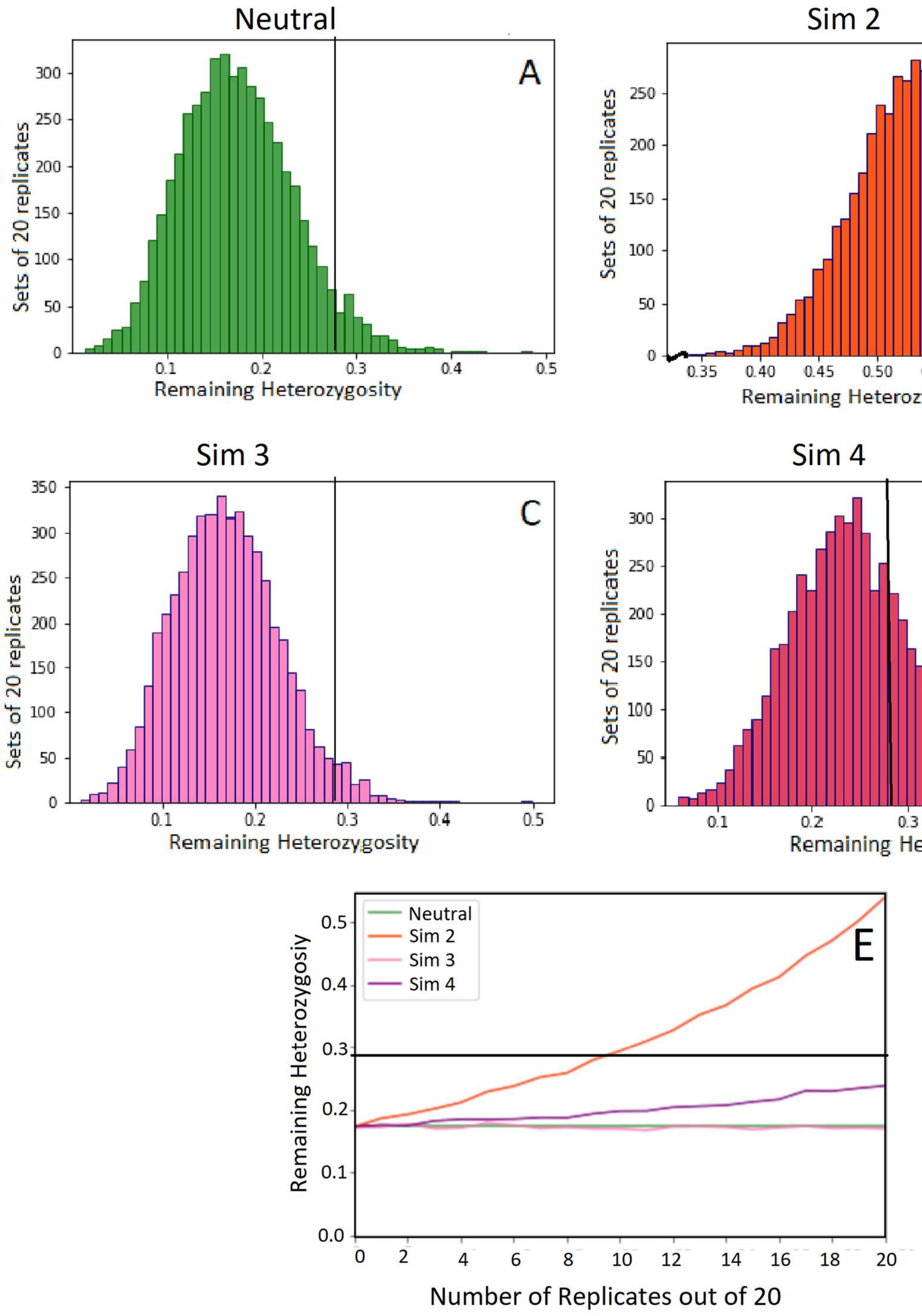


Figure 6. Simulated distribution of heterozygosity proportions in sets of 20 replicates (**A** to **D**). **E** shows the proportion of remaining heterozygosity in sets of 20 replicates, with increasing replicates of a given simulation scenario in a background of neutral replicates. This replicates across all lines (four chromosomes were analysed for 5 lines), and assuming that the genome is comprised of the four autosomes analysed in this experiment. Horizontal and vertical black lines denote the remaining heterozygosity proportions in the real data.

There was an average number of 12 tracts per replicate of 20 chromosomes under neutrality. The number of tracts observed in the *D.mojavensis* lines was 77(**Figure S3**), which lied completely outside of this distribution, so could not be explained by it, and neither could any of the other models simulated in this project(**Figure S4**). This could mean that some tracts had been split, as the chosen threshold was conservative so that heterozygosity was not overestimated, or gene conversion, which could split a heterozygous tract and was not implemented in the simulations. A good example of this could be the tracts observed in chromosome 4 of the SC05 line, where a tract appears to be split into many tracts because of small variations in heterozygosity levels, as the background diversity of this line is lower than the rest(Benowitz KM., 2019)(**Figure 4**).

4.3 Tract overlap

Both Sonora lines contained a large heterozygous tract in chromosome two, spanning more than 80% of the chromosome in both cases(**Figure 4**), visibly overlapping. It was hypothesised that this could happen because of contamination with lines from Baja California, as they have an inversion on chromosome two that could have acted as a recombination barrier. If this were the case, the tract breakpoints on both Sonora lines would be similar. The breakpoints were not the same, differing for around 500kb. This possibility remained unlikely, but would explain the observed pattern, as Baja California lines have fixed and segregating polymorphic inversions on chromosome two(Ruiz, A., et al., 1990).

To test whether this overlap was significant, overlaps between Sonora and Baja lines were compared against the simulations, and against each other. Comparing all possible overlaps against the distribution on the simulations could lead to a false positive because of multiple testing, so they were also compared against each other. Chromosome two in the Sonora lines overlapped in heterozygosity tracts for 79% of the chromosome, which completely deviates from the neutral scenario, where the 95th percentile of simulations overlapped for 17% of the chromosome, with an average overlap of only 3%(**Figure S5**).

Sim 2 scenario guaranteed that at least a small amount of heterozygosity was conserved in each simulation. This model could not explain the observed overlap, as the 95th percentile overlap consisted of a 31% overlap(**Figure S5**).

Applying the recessive models could not explain the observed overlap. Sim 4 retained significantly more heterozygosity than Sim 3 and had greater overlapping proportions. A recessive lethal balancing system was established but was still unable to explain the observation.

The model that best suited the observed overlap was an extreme overdominance model containing one homozygous lethal mutation on either end of the tract, Sim 5. Here, an average overlap of 67% was observed, as both ends of the tracts were forced to remain heterozygous (**Figure S5**).

Running a pairwise comparison between all chromosomes revealed that lines from the same geographical area are more likely to contain overlapping heterozygous tracts. The average proportion of the genome that overlapped in heterozygous tracts between lines from the same area was 0.12(**Table 3**), which was more than twice of the overlap that was observed between all the lines, 0.05(Data not shown). This pattern was mostly driven by chromosome 2 of the Sonora lines.

5. Discussion

Selection enforced retention of heterozygosity, has been subject of extensive study(Crow J. F., 1952, Bennet J., 1956, Dobzhansky T. et al., 1963). Here, I show how this has occurred in D.mojavensis inbred lines, as well as providing various concepts and models that might explain this phenomenon. Several factors can contribute to reduce genetic exchange between haplotypes and therefore lead to a greater retention of heterozygosity. These include chromosomal rearrangements, recessive lethal genes on their own, or involved in a balancing recessive lethal system, and can be as simple as clustered alleles with negative selection coefficients.

There are several points that emerge from this experiment:

5.1. Maintenance of heterozygosity in D.mojavensis

All lines remained heterozygous at some points throughout their genome after the sib-mating experiment. Individuals that were collected from the Sonora desert retained the highest levels of heterozygosity, with individuals from Santa Catalina retaining the least amount of heterozygosity, even though it was a line from Baja California that had the lowest heterozygosity proportions(A975). This was expected, as Sonora and Baja populations differ substantially in their effective population size(N_e), with the Sonora race having a lower N_e, which contributes to increase the genetic load of a line(Smith G. et al., 2012).

Models including one homozygous lethal mutation in every pair of chromosomes, Sim 2, led to greater proportions of conserved heterozygosity which were too high to explain the observations. However, suppression of recombination in every chromosome is an unrealistic scenario. An average of 1.8 homozygous lethal mutations per genome yielded similar heterozygosity levels as the observed data. This can be related to the frequency of segregating recessive lethal alleles on a single individual, with estimates lying around 1.6 recessive lethal alleles segregating on one genome at any point in a *D.melanogaster* genome(Gao Z. Y., et al., 2015). Homozygous lethal mutations are more extreme than recessive lethal mutations, and if they were the only force acting on retaining heterozygosity in this experiment, we would expect a number lower than 1.6.

However, interactions between alleles with varying distributions of fitness effects(DFE) can lead to greater proportions of conserved heterozygosity. The nature of the distribution of deleterious mutational effects on fitness is central to population genetics studies and is extremely difficult to calculate. It has been shown that most segregating amino acid substitutions are deleterious, with most of them having small selection coefficients(Loewe L, Charlesworth B, 2006).

This interaction can be seen in Sim 4, where the recessive lethal alleles on opposing haplotypes can interact to establish a balancing system. I was therefore able to conclude that the heterozygosity proportions could not be seen under neutrality, and that selective forces were acting to enforce the retention of heterozygosity at multiple loci, and that this was probably because of alleles with varying DFEs.

5.2. Possible chromosomal inversion in the A976 line

Inversions in the D.mojavensis lineage have been extensively studied and characterised (Dobzhansky T. et al., 1963, Lohse et al., 2015, Delprat A., et al., 2019).

Chromosome 3 on the A976 line harboured a large heterozygous tract, representing 62% of its length. Polymorphic chromosomal inversions in chromosome 3 have previously been reported to be segregating in lines from Baja California(Ruiz A., et al., 1990). This inversion can be the subject of balancing selection, whereby fitness is greatest for polymorphic individuals, maintaining the tract in a heterozygous state(Fuller ZL., et al., 2018).

5.3. Tract overlap in Chromosome 2 of the Sonora lines

Chromosomes belonging to the same race were more likely to overlap than chromosomes from different races, supported by strong genetic structure observed between races.

The large overlapping tracts in chromosome two of the Sonora lines was a very interesting find, that could have occurred through multiple mechanisms:

Firstly, even though it would be very unlikely, this result could be product of contamination with a fly from Baja California. Baja lines are polymorphic for multiple inversions in chromosome two(Ruiz A., et al., 1990) and contamination from one of these flies could introduce an inversion on both lines that was sequenced to be heterozygous.

Second, this overlap could represent a polymorphic chromosomal inversion on both lines. Literature is more abundant for fixed inversions and polymorphic inversions in the Baja lines, but an inversion in chromosome two of the Sonora lines, segregating at low frequencies has been reported(Ruiz A., et al., 1990). Considering the low frequency(5%) of this inversion, it is unlikely that both lines were heterozygous for it. However, this hypothesis cannot be neglected, as it is still possible.

Finally, this overlap could occur because of multiple deleterious loci with low dominance values on both haplotypes. This genetic load on both chromosomes could have resulted lethal if homozygosed, forcing the tracts to remain heterozygous. The strongest model that implemented recessive lethal alleles was the one that included one recessive lethal allele on each haplotype, each one on either the start or the end of the observed tract overlap. This model could not explain the overlap observed, so the selective force acting on both lines must have been greater, with either more recessive lethal alleles along the tract, or a combination of deleterious alleles that add together and balance each other, to produce a tract that gives rise to unviable offspring if homozygous. Alleles in this tract are not necessarily lethal but could harbour multiple alleles with deleterious selection coefficients that act together to enforce heterozygosity.

The models with homozygous lethal alleles were a better fit for the observations. However, these models are unrealistic, as they represent regions that cannot be homozygosed under any circumstance, which is rarely observed. However, this means that the overall effect of the genetic load within this tract must have been similar, completely preventing homozygosing of the region.

5.4. Possible improvements

This project occurred as a by-product of a population genetics study. It is a widespread technique in this field to sib-mate a line to remove the heterozygosity, which makes sequence analysis easier. There are two additions that could have been made to the methods before the start of this project that could have strengthened the results.

Firstly, the heterozygosity of each line could have been calculated the start of the experiment. Here, I assume that the initial flies were completely heterozygous, and no inbreeding had occurred prior to the sib-mating. However, the lines were collected and stored in containers, in some cases for more than a decade(**Table 2**). The population size was meant to be large enough to maintain wild type heterozygosity levels, but some inbreeding might have occurred. This would have translated to lower heterozygosity levels at the start, and therefore an apparent greater reduction in heterozygosity than what would have been observed if the initial parents were completely heterozygous.

Secondly, the fitness of each generation could have been recorded, and maybe even sequence the genome of those too unfit to reproduce. This would have given us a better insight on the nature of the heterozygosity tracts, giving further evidence on the tracts not being present by chance.

Given that the different lines contained different levels of background diversity, it would be informative to apply line specific thresholds and see how that changes the observations.

Finally, this approach only encapsulated the most extreme kinds of recessive deleterious mutations, as well as the homozygous lethal mutations. However, the DFE has been estimated for other *Drosophila* species, with widely distributed selection coefficients (s), and even though this approach is revealing, it remains unrealistic(Keightley PD, Eyre-Walker A, 2007). Allowing for a range of s values and implementing gene conversion in the simulations would reveal more about the genetic load in *Drosophila.*

6. Conclusion and future prospects

To understand how species diverge and eventually speciate, it is essential to understand how gene exchange is limited, or in some cases completely restricted. Here I show how the retention of heterozygosity is limited to certain loci and apply various models to shed light on the possible mechanisms behind it.

For future study, it would be interesting to investigate whether retention of heterozygosity in one chromosome can have a positive effect on increasing the heterozygosity levels on different chromosomes. It would also be interesting to calculate the current frequency of the inversion present in chromosome two in the Sonora lines. It was reported to be 5% in 1990 (Ruiz, A., et al., 1990) but calculating how that frequency could have changed would provide information of how the frequency of polymorphic inversions can drift over time.

7. References

1. Allan CW, Matzkin LM (2019) Genomic analysis of the four ecologically distinct cactus host populations of Drosophila mojavensis. Bmc Genomics 20: 13
2. Andolfatto P, Depaulis F, Navarro A (2001) Inversion polymorphisms and nucleotide variability in Drosophila. Genetics Research 77: 1-8
3. Benowitz KM, Coleman JM, Matzkin LM (2019) Assessing the Architecture of Drosophila mojavensis Locomotor Evolution with Bulk Segregant Analysis. G3-Genes Genomes Genetics 9: 1767-1775
4. Bennett, J. (1956). Lethal genes in inbred lines. Heredity 10, 263–270 (1956). https://doi.org/10.1038/hdy.23
5. Charlesworth B, Charlesworth D (1999) The genetic basis of inbreeding depression. Genetical Research 74: 329-340
6. Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, Markow TA, Kaufman TC, Kellis M, Gelbart W, Iyer VN et al (2007) Evolution of genes and genomes on the Drosophila phylogeny. Nature 450: 203-218
7. Crow, J. F., 1952 Dominance and overdominance. pp. 28S-297. Heterosis. Iowa State College
8. Delprat A, Guillen Y, Ruiz A (2019) Computational Sequence Analysis of Inversion Breakpoint Regions in the Cactophilic Drosophila mojavensis Lineage. Journal of Heredity 110: 102-117
9. Dobzhansky T, Spassky B, Tidwell T (1963). Genetics of natural populations. 32. Inbreeding and the mutational and balanced genetic loads in natural populations of Drosophila pseudoobscura. Genetics.48(3):361-73. PMID: 14028273; PMCID: PMC1210477.
10. Fisher RA (1949) Theory of inbreeding. Heredity 3: 123-123
11. Fisher RA (1954) A fuller theory of junctions in inbreeding. Heredity 8: 187-197
12. Fuller ZL, Leonard CJ, Young RE, Schaeffer SW, Phadnis N (2018) Ancestral polymorphisms explain the role o chromosomal inversions in speciation. Plos Genetics 14: 26
13. Gao ZY, Waggoner D, Stephens M, Ober C, Przeworski M (2015) An Estimate of the Average Number of Recessive Lethal Mutations Carried by Humans. Genetics 199: 1243-1254
14. Garrison E, Marth G. (2012). Haplotype-based variant detection from short-read sequencing. arXiv preprint arXiv:1207.3907 [q-bio.GN]
15. Gowen JW, Stadler J, Johnson LE (1946) On the mechanism of heterosis - The chromosomal or cytoplasmic basis for heterosis in Drosophila-melanogaster. American Naturalist 80: 506-531

16. Gramates LS, Marygold SJ, dos Santos G, Urbano JM, Antonazzo G, Matthews BB, Rey AJ, Tabone CJ, Crosby MA, Emmert DB et al (2017) FlyBase at 25: looking to the future. Nucleic Acids Research 45: D663-D671

17. Haldane JBS (1927) A mathematical theory of natural and artificial selection, Part V: Selection and mutation. *Proceedings of the Cambridge Philosophical Society* 23: 838-844

18. Haller BC, Messer PW (2019) SLiM 3: Forward Genetic Simulations Beyond the Wright-Fisher Model. Molecular Biology and Evolution 36: 632-637

19. Keightley PD, Eyre-Walker A (2007) Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. Genetics 177: 2251-2261

20. Kelleher J, Etheridge AM, McVean G (2016) Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. Plos Computational Biology 12: 22

21. Kirkpatrick M (2010) How and Why Chromosome Inversions Evolve. Plos Biology 8: 5

22. Laetch, D.R. A genome-wide IM blockwise likelihood estimation toolkit. https://github.com/DRL/gIMble

23. Lenormand T, Engelstadter J, Johnston SE, Wijnker E, Haag CR (2016) Evolutionary mysteries in meiosis. Philosophical Transactions of the Royal Society B-Biological Sciences 371: 14

24. Lerner, I. M. (1950). Population Genetics and Animal Improvement. Cambridge University Press, p. 102.

25. Li H (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27: 2987-2993

26. Li H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997v2 [q-bio.GN].

27. Loewe L, Charlesworth B (2006) Inferring the distribution of mutational effects on fitness in Drosophila. Biology Letters 2: 426-430

28. Lohse K, Clarke M, Ritchie MG, Etges WJ (2015) Genome-wide tests for introgression between cactophilic Drosophila implicate a role of inversions during speciation. Evolution 69: 1178-1190

29. Ruiz A, Heed WB, Wasserman M (1990) EVOLUTION OF THE MOJAVENSIS CLUSTER OF CACTOPHILIC DROSOPHILA WITH DESCRIPTIONS OF 2 NEW SPECIES. Journal of Heredity 81: 30-42

30. Smith G, Lohse K, Etges WJ, Ritchie MG (2012) Model-based comparisons of phylogeographic scenarios resolve the intraspecific divergence of cactophilic Drosophila mojavensis. Molecular Ecology 21: 3293-3307

31. Sturtevant AH (1917) Genetic factors affecting the strength of linkage in Drosophila. Proceedings of the National Academy of Sciences of the United States of America 3: 555-558

8. Acknowledgements

I am extremely thankful to the members of the Lohse lab for making my stay very enjoyable and for making me always feel welcome.

I want to thank Konrad Lohse for providing me with hard-to-find literature and guiding me through this project, offering plentiful feedback and advice.

Derek Setter was always available for any issues that came up and always aiding me with python issues, which has substantially helped me improve my programming skills.

I want to thank Gertjan Bisschop for being available for any SLiM queries that came up.

Finally, I want to acknowledge NERC, for funding the sequencing that took place for this project.

9. Appendix

9.1 Supplementary Figures

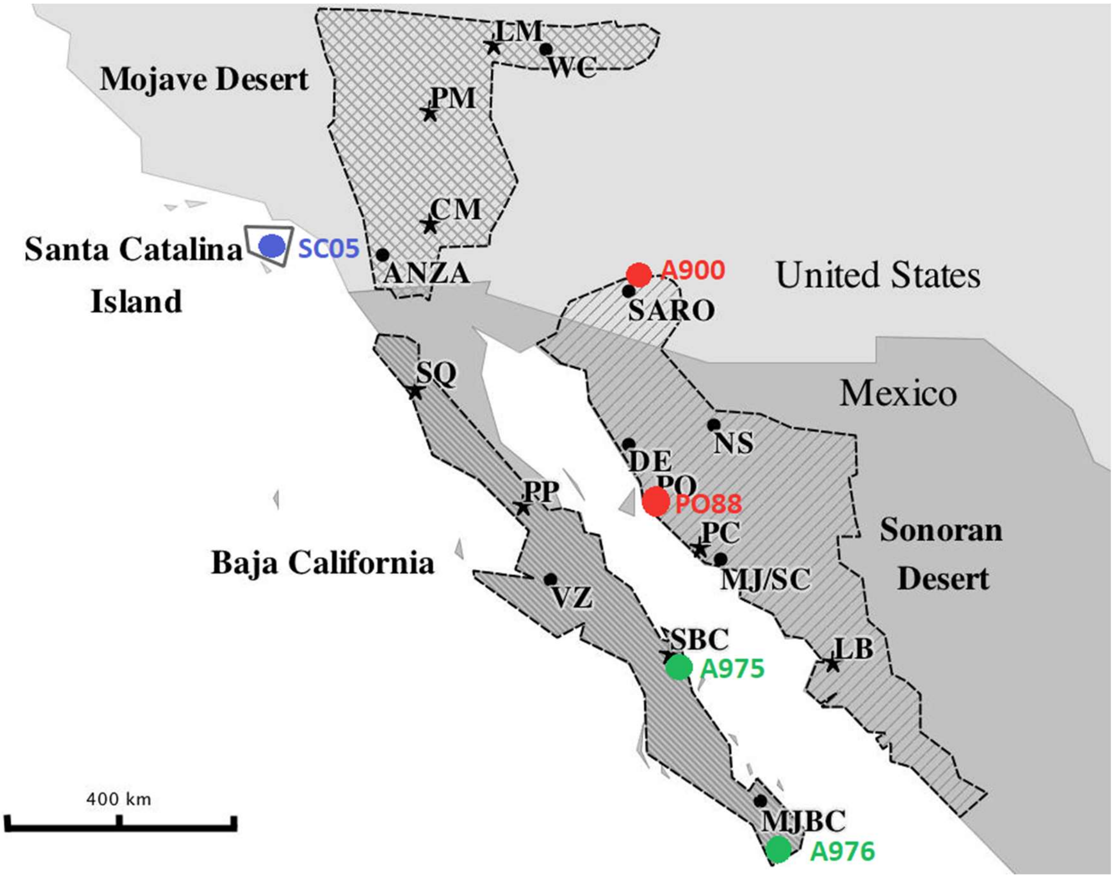


Figure S1. Map from Baja California and the Sonoran Desert with the locations where the lines were collected.

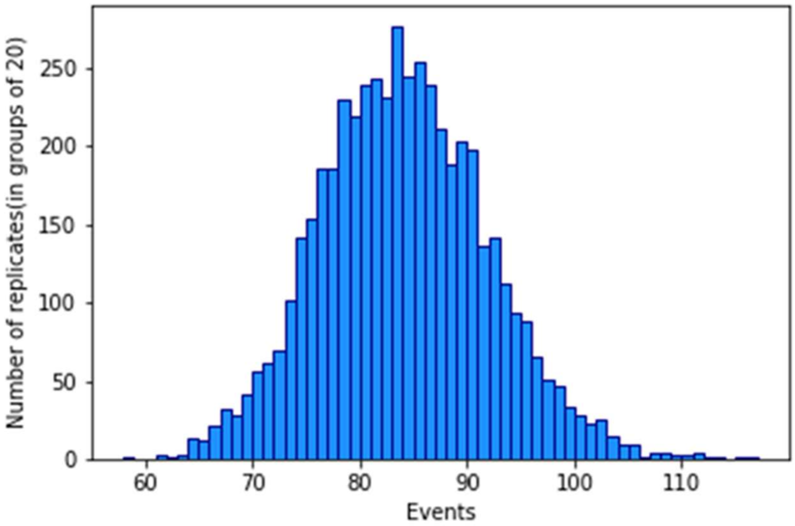


Figure S2. Distribution of recombination events in sets of 20 replicates.

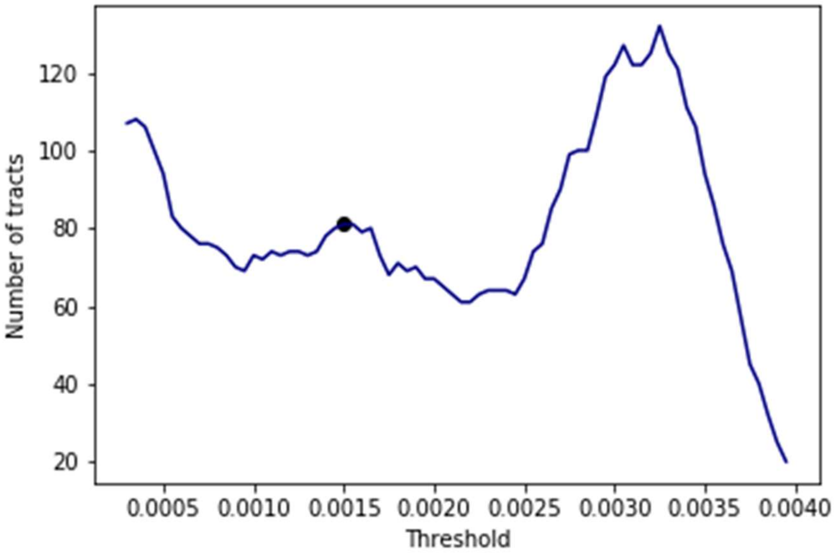


Figure S3. Number of tracts at different heterozygosity thresholds. The black dot represents the number of tracts for the chosen threshold, 0.0015

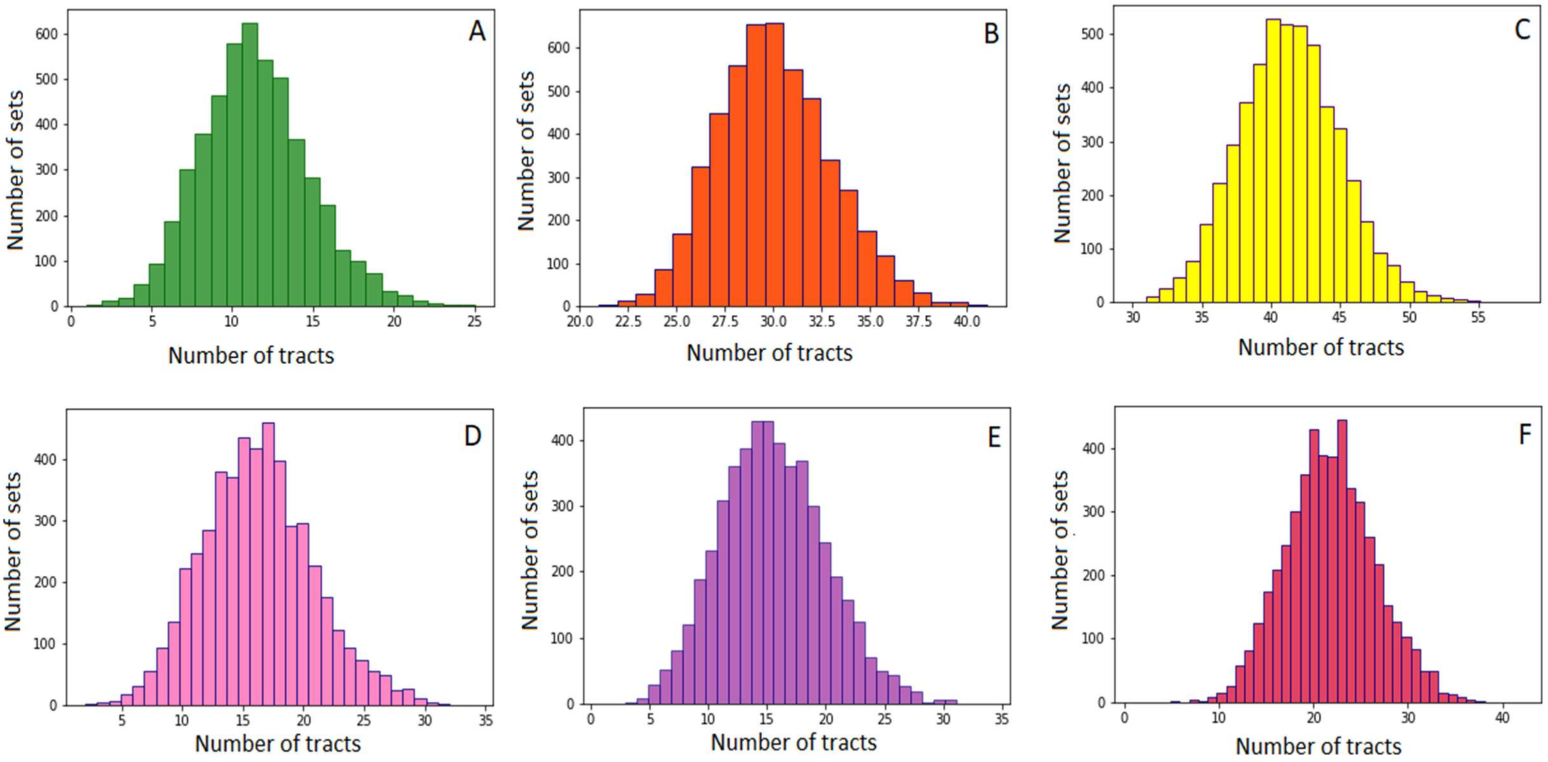


Figure S4. Simulated distribution of heterozygous tracts in sets of 20 replicates. **A)** Neutral model. **B)** Sim 2 model **C)** Model with two homozygous lethal alleles on the same haplotype. **D)** Sim 3 model **E)** Model with two recessive lethal alleles on the same haplotype. **F)** Sim 4 model

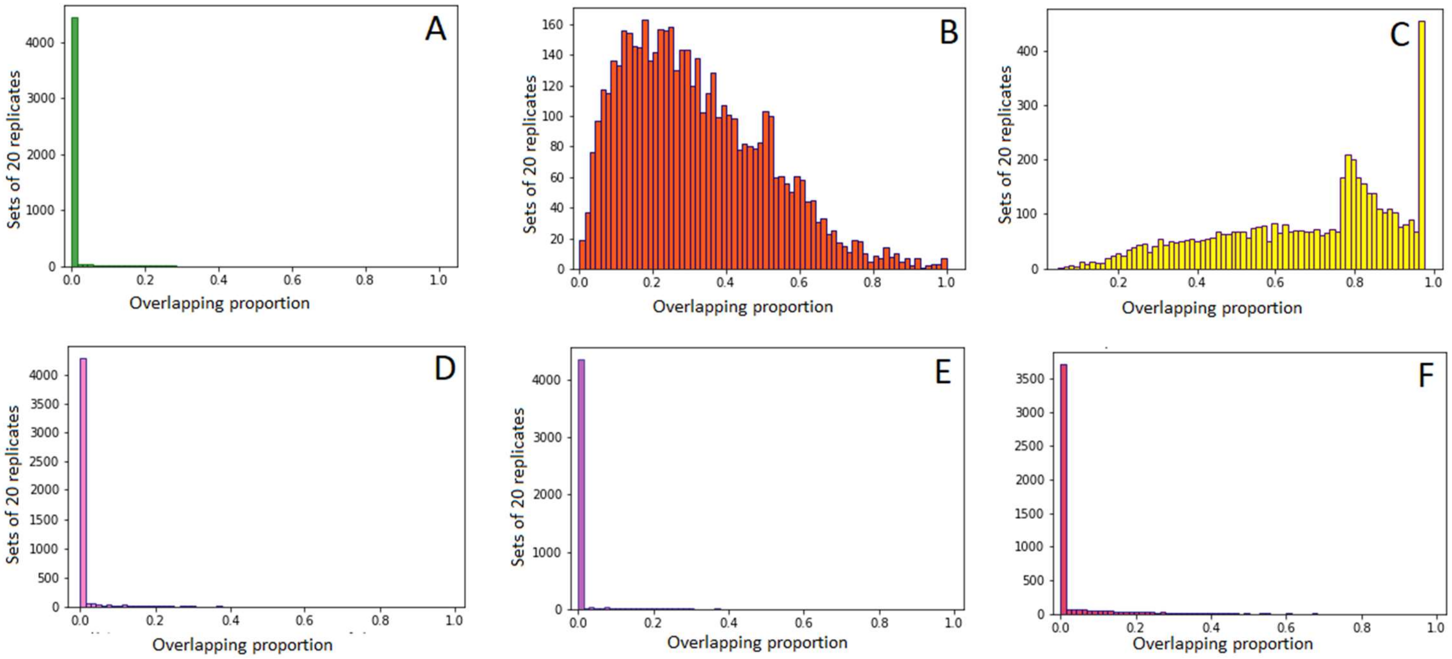


Figure S5. Simulated overlap of heterozygous tracts in pairs of 2 replicates. **A)** Neutral model. **B)** Sim 2 model **C)** Model with two homozygous lethal alleles on the same haplotype. **D)** Sim 3 model **E)** Model with two recessive lethal alleles on the same haplotype. **F)** Sim 4 model

9.2 Tables

Line name	**Population**	**Latitude**	**Longitude**
A975	Bahía de Concepción, Baja C.	26°32’03.72’’N	111°44’03.72’’ W
A976	Santiago, Baja C. Sur	23°28’07.91’’N	109°40’45.8’’ W
A900	Santa Rosa Mountains	*33*°65’45.31’’N	*116*°46’67.83’’W
PO88	Punta Onah, Sonora	29°5 ‘23.15’’N	112°10’15.59’’ W
SC05	Isla de Santa Catalina	*33*°39’17.87’’N	*118*°41’74.94’’ W

Table 1. Adapted from Lohse et al., 2015. Locations of the 5 *D.mojavensis* lines used in this project

Line name	***Year***
A975	1986
A976	1996
A900	1996
PO88	1988
SC05	2005

Table 2. Year in which each line used in this study was collected. Lines were collected over banana baits and maintained on banana food at room temperature as described in Lohse et al., 2015.

	Chr 2	***Chr 3***	***Chr 4***	***Chr5***
Sonora	*0.79*	*0.06*	*0.04*	*0.00*
Baja	*0.00*	*0.00*	*0.00*	*0.03*

Table 3. Tract overlap between lines from the same host races.