Genetic linkage to explain genetic variation

J.L. Mijangos[a], C.E. Holleley[b,f], R.A. Nichols[b,c], I.N. Towers[a], Z. Jovanoski[a], H.S. Sidhu[a], S. Watt[a], A.T. Adamack[d,e] and W.B. Sherwin[b]

[a] School of Physical, Environmental and Mathematical Sciences, University of New South Wales, Canberra, Australia.
[b] Evolution and Ecology Research Centre, School of Biological, Earth and Environmental Sciences, University of New South Wales, Sydney NSW 2052, Australia.
[c] School of Biological and Chemical Sciences, Queen Mary University of London, Mile End Road, London E1 4NS, UK.
[d] Institute for Applied Ecology, University of Canberra, Bruce, Australia.
[e] Department of Biology, Virginia Commonwealth University, Richmond, VA 23284, USA.
[f] Australian National Wildlife Collection, National Research Collections Australia, CSIRO, Australia.
Email: luis.mijangos@unsw.edu.au

Abstract: Population genetics theory has been used to develop models to inform biodiversity conservation. The implementation of these models for the decision-making, monitoring and evaluating conservation efforts has improved their efficacy. Many of these models are based on neutral loci, which are assumed to not have any effect on the survival or reproduction of organisms. However, when neutral loci are linked to loci that are under selection, they do not follow the expectations predicted by theoretical neutral models. The term linked selection has been used to refer to this phenomenon.

Linked selection might accelerate the rate of loss of genetic diversity, with respect expectations under neutral models. This acceleration is typically expected to occur under two different scenarios: selective sweeps and background selection. A selective sweep occurs when an advantageous allele is spread across the population along with the alleles that are linked to it. Conversely, background selection occurs when a deleterious allele is eliminated from the population along with the alleles linked to it. A third scenario of linked selection has been hypothesised to occur in small populations: associative overdominance. This scenario rather than an acceleration involves a retardation of the loss rate of genetic diversity. The proposed mechanism in operation is a type of natural selection that maintains two or more alleles in the population (*i.e.* balancing selection). Ultimately, linked selection will bias the conclusions obtained by neutral models.

To investigate the mechanisms by which linked selection alters genetic diversity, we built a population genetics model incorporating the main factors involved in the occurrence of linked selection. We modelled the following factors: recombination/linkage disequilibrium, fitness/selection, population size and dominance. Our overall aims are twofold:

- Build a model that serves as a base to develop more complex models to test competing hypotheses of linked selection.
- Implement the model in a programming language and validate it against theoretical expectations

The model was implemented in the programing language R and was extensively tested. The program produces outputs that are consistent with predictions from population genetics theory. By using this model as a starting point, we will be able to investigate potential mechanisms by which linked selection affects genetic diversity and its derived consequences. Our research may highlight the importance of the need to adjust neutral models.

Keywords: *Associative overdominance, population genetics, linkage disequilibrium, linked selection*

1. INTRODUCTION

Genomes consist of sequences of deoxyribonucleic acid (DNA) that encode most of an organism's traits. DNA is packaged within structures called chromosomes. Diploid organisms carry two copies of each chromosome with one copy inherited from each of its parents. The location of a specific DNA sequence within the genome is called a "locus" ("loci" in plural), and the different variants of DNA sequences at a locus are called "alleles". The word "gene" is often ambiguously used to make reference to both "locus" and "allele"; therefore, we avoid the use of gene in this text. An individual carrying different alleles at a particular locus is said to be heterozygote at that locus, while an individual carrying the same allele at that locus is a homozygote at that locus.

Population geneticists have developed theoretical models to understand and predict evolutionary processes. Many core models in population genetics are based on neutral loci, called neutral because they do not affect **fitness**[1] (*i.e.* survival or reproduction of living organisms). Therefore, neutral loci are not affected by **natural selection**, but by genetic drift (variation in allele frequencies across generations due to random chance) and gene flow (exchanges of genetic material between populations). Neutral loci models have been developed to use our knowledge about genetic drift and gene flow to derive measures and parameters that can be used to inform the management of natural populations. For instance, such measures can help us to determine levels of landscape connectivity, population size, genetic divergence, population origin of individuals, delineation of populations and infer past and present population trajectories (Frankham *et al.* 2009). The effective population size (*Ne*) is a major concept in population genetics and is particularly important for informing the conservation and restoration of biodiversity. *Ne* predicts the effects of genetic drift such as rates of loss of neutral genetic diversity, and the interactions of genetic drift and natural selection such as fixation of deleterious and favourable alleles (England *et al.* 2006). *Ne* is defined as the size of an idealised population that would have the same amount of genetic drift as the population under consideration (Kimura & Crow 1963).

Neutral models frequently assume that neutral loci are inherited by offspring as independent units from their parents. However, loci that are close to each other on the same chromosome are often inherited together in clusters of different loci termed haplotypes. This non-independent inheritance of loci will cause a non-random association between their alleles causing a linkage disequilibrium (LD) between loci. Conversely, the process of recombination rearranges loci within haplotypes and hence dilutes the amount of LD between loci. Recombination occurs during **meiosis**, when **homologous chromosomes** temporarily fuse and exchange loci between each other (Fig. 1). As a result of recombination, haplotypes with different combinations of alleles arise, which are then inherited by offspring.

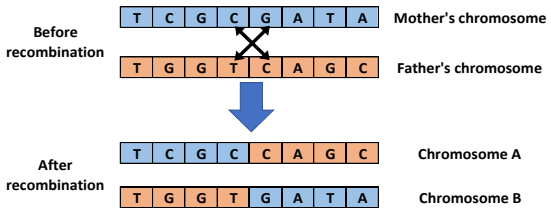

Figure 1. Diagram depicting the recombination process. Upper DNA sequences represent homologous chromosomes of an individual (its mother's and its father's chromosomes). Black arrows represent the location where recombination is occurring within a chromosome. Lower DNA sequences represent the resulting chromosomes (A and B) that may be transmitted to the individual's offspring.

When LD occurs between two loci, one neutral and one under selection, theoretical expectations for neutral loci break down, and sometimes neutral loci appear as if they are under selection ("linked selection"). As a consequence, this phenomenon biases the conclusions derived from neutral models.

Our objective is to develop a population genetics model that recreates the mechanisms occurring in linked selection. This model will serve as a base to develop more complex models that allow us to investigate factors that may be influencing linked selection, such as dispersal, different types of selection and multiple loci under linked selection. This work will highlight the need to adjust neutral models when linked selection occurs and will provide insights into how linked selection may influence models used in conservation and restoration.

[1] Please note that all the words or phrases in **bold** in this manuscript are defined in the glossary section.

In this manuscript, we present the factors involved with linked selection, and then outline how these factors were incorporated in the model. We then ran test simulations to validate the proposed model. We used R v3.3.3 (R Core Team 2017) to implement the model and run the simulations.

2. HAPLOTYPES WITH LINKED SELECTION

Linkage disequilibrium between nearby loci, within a haplotype, will cause loss of genetic variation if one or more loci within the haplotype is under selection. Loss of genetic variation via linked selection is expected to occur under two different scenarios. An allele under positive selection (i.e. advantageous) will be spread across the population along with the alleles that are linked to it (Fig. 2A). This scenario was first described by Smith and Haigh (1974) and termed a "selective sweep". Conversely, an allele under negative selection (*i.e.* deleterious) will be eliminated from the population along with the alleles linked to it (Fig. 2B). This scenario was first described by Charlesworth (1994) and termed "background selection". Ultimately, selective sweeps and background selection provoke the loss of genetic diversity at a higher rate than expected by neutral models.

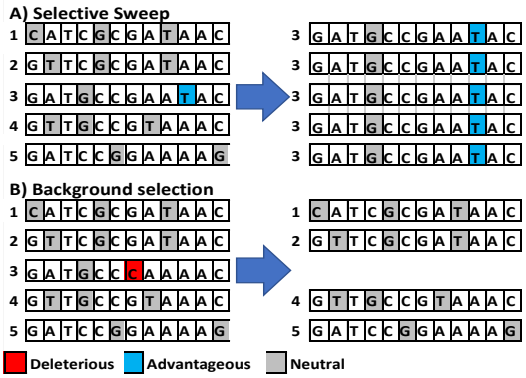

Figure 2. Hypothetical scenarios showing the consequences of selective sweeps and background selection on genetic diversity after several generations. Five haplotypes are present in the population (1 to 5) and colored cells depict alleles (red, deleterious; blue, advantageous; gray, neutral). In the selective sweep scenario (A), the advantageous allele "T" has a higher fitness than the alternative allele "A" at the same position in other haplotypes, resulting in the spread of the advantageous allele across the population along with the alleles that are linked to it (within the same haplotype). In the background selection scenario (B), the deleterious allele "C" has lower fitness than the alternative "G", resulting in the elimination of the deleterious allele from the population along with the alleles linked to it (within the same haplotype). Modified from Charlesworth and Charlesworth (2010).

2.1. Associative overdominance

In some circumstances linked selection may result in the maintenance of genetic variation rather than its loss (Rumball *et al.* 1994). The current hypothesis explaining this phenomenon has been named "associative overdominance". Overdominance arises when heterozygotes have a higher fitness than either homozygote. This is a type of balancing selection, where selection maintains two or more alleles in the population. Two different scenarios have been proposed to explain the conditions under which associative overdominance occurs. The first scenario occurs when there is LD between a neutral locus and a locus under selection in favour of heterozygotes (*i.e.* overdominance). Consider the following example: an overdominant locus named A (with alleles A1 and A2) is in LD with a neutral locus named Z (with alleles Z1 and Z2). Allele A1 is in LD with allele Z1, and A2 is in LD with Z2. Balancing (overdominant) selection on locus A results in retention of the A1 and A2 alleles, and because of LD, Z1 and Z2 alleles are also retained. Eventually, this system will break up due to recombination between loci A and Z.

The second scenario occurs when a neutral locus is located between two or more **recessive** deleterious alleles, and these loci are close enough, so LD exists between them. This scenario involves directional selection, where selection replaces one allelic type by another that is more advantageous in terms of fitness. Consider the following example: a "+" sign denotes the higher fitness of the fitter allele, and a "-" sign denotes the lower fitness of the deleterious allele. Locus A (with fitter allele A1+ and deleterious allele A2-) and locus B (with fitter allele B1+ and deleterious allele B2-) are under directional selection and in LD. Two haplotypes are present in the population: A1-B1+ and A2+B2-. Neutral locus Z (with alleles Z1 and Z2) is linked to these haplotypes as follows, Z1 with A1-B1+ and Z2 with A2+B2-. Since each haplotype has one fitter and one deleterious allele, neither haplotype is fitter than the other by itself. However, a homozygote of either haplotype will be less fit than a heterozygote. This results in a mechanism that is similar to the previous scenario, where due to an apparent overdominance between haplotypes, alleles Z1 and Z2 are retained. Eventually, this will break up due to recombination between the three loci (A, B and Z).

2.2. Factor involved with linked selection in general

Recombination

For linked selection to occur, it is necessary for neutral loci be located within a specific distance of the locus under selection, so recombination will be relatively slow to break up the LD between these loci. The genetic distance between two loci within a chromosome is measured in units called Centimorgans. One centimorgan corresponds to one percent of probability that the two loci will be separated by recombination.

2.3. Factors affecting associative overdominance

While an acceleration of the loss of genetic diversity, with respect to neutral theory, is observed in scenarios of background selection and selective sweeps, a retardation is observed in associative overdominance. To discern the occurrence of these opposing scenarios, we describe below the factors required for associative overdominance to occur.

***Population size and natural selection*:** The fate of alleles (whether to be lost or become fixed) in a population is determined by the joint effects of natural selection and genetic drift. The effects of genetic drift depend on *Ne*. The smaller *Ne* is, the stronger the effects of genetic drift. The effects of natural selection depend on **selection coefficients**. The higher the selection coefficient, the stronger the effect of natural selection. It has been hypothesised that associative overdominance only occurs when populations are of a particular size. In very small populations, the necessary LD arises by chance frequently, but selection is very ineffective relative to random processes. Thus, there is probably an optimum population size for associative overdominance

***Population size and linkage disequilibrium*:** Haplotypes, just as alleles, are lost more rapidly in small populations than in large populations due to the increased effects of genetic drift. As a result, recombination will be less efficient in creating new combinations of alleles, as there are fewer available haplotypes for which recombination could act. In the long term, reductions in the number of haplotypes will lead to an increase in LD between loci.

***Dominance*:** Whether an allele is expressed in the **phenotype** depends on its level of dominance (whether dominant or recessive). Recessive alleles are expressed only when they are present in a homozygote. In contrast, dominant alleles are expressed when an individual is homozygous or heterozygous for the dominant allele. Background selection and selective sweeps will have different efficiency depending upon whether the alleles concerned are dominant or recessive, resulting in accelerated rates of loss of genetic diversity. In the case of associative overdominance due to LD with deleterious alleles it is necessary that these alleles to be recessive, so that they can be expressed in homozygotes, but not in heterozygotes and thus allowing an apparent overdominance.

3. MODEL DESCRIPTION

The main aim of this paper is to build a model to simulate loci that are neutral or under selection and are transmitted from one generation to the next within a population using R v3.3.3 (R Core Team 2017). We modeled diploid organisms where each individual contains one locus under selection (A) and one neutral locus (Z), each locus has two alleles (A1, A2 and Z1, Z2 respectively).

The following parameters are established at the beginning of the simulation: Population size (*Ntot*), number of offspring per mating (*x*), recombination probability (*c*), proportion of each allele (*p* and *q*), selection coefficient for homozygote A1A1 (s_1), selection coefficient for homozygote A2A2 (s_2), number of generations (*g*) and number of replicates (*r*).

3.1. Initial population

The population consists of *Ntot* individuals, with half of the individuals being male and half female. For each simulated individual, the program tracks its sex (male or female), its neutral locus, its locus under selection, and in the case of individuals created following the initial generation, its two parents. The sex of each individual is determined randomly with equal probability of being male or female. Initial allele proportions are determined separately for each locus by randomly drawing genotypes from a binomial distribution with the probability of an allele being equal to its proportion (*p* and *q*). For subsequent generations, alleles assigned to individual offspring depended on the alleles carried by their parents, using Mendelian inheritance with random sampling of one of the two haplotypes in each parent, after any recombination had occurred. To generate new offspring, *Ntot/2* pairs of males and females are sampled without replacement from the previous generation. For each pairing, *x* offspring are produced with half being male and half female.

3.2. Recombination

To determine each parent's haplotypic contribution to an individual offspring, the program first determined for each parent separately whether recombination had occurred by generating a uniform random deviate between 0 and 1. If the deviate was less than *c*, the probability of recombination, then the program recombined that parent's haplotype for that offspring (e.g. Fig. 1). *E.g.* if a parent had haplotype A1Z1 and A2Z2 and recombination occurred, haplotypes available to the offspring would be A1Z2 and A2Z1. If recombination did not occur then the initial haplotypes from the parent were available to the offspring. Once the available haplotypes from each parent were determined, one haplotype from each parent was assigned randomly to the offspring. This process was repeated separately for each offspring produced meaning that some potential offspring from a pairing could have recombined haplotypes while others might not.

3.3. Natural selection

If a resulting offspring is homozygote for the alleles under selection (A1A1 or A2A2) a fitness test was performed: a uniform random deviate between 0 and 1 was drawn. If that number was less than s_1 or s_2 (depending on the homozygote variant), the individual was removed from the offspring pool from which the next generation of parents would be drawn (Table 1). After all of the offspring had been generated and selection had been applied, a subset of the offspring pool equal to *Ntot* was randomly selected to make up the next generation of parents, with half being male and half being female. This process was repeated for *g* generations, and *r* replicates.

4. MODEL VALIDATION

Table 1. Example for the calculation of fitness in the case where number offspring per mating = 10

	Homozygote A1A1	**Heterozygote A1A2**	**Homozygote A2A2**
Selection coefficient (s)	$s_1 = 0.2$	1	$s_2 = 0.4$
Fitness equation (w=1-s)	$w_{11} = 1 - s_1$	$w_{12} = 1$	$w_{22} = 1 - s_2$
Fitness value	0.8	1	0.6
Offspring contribution to next generation	8	10	6

Our objective in building this model is to use it as a starting point for developing more complex models that will allow us to test linked selection hypotheses. Therefore, this model should be general enough that it can accommodate more complex scenarios, but more importantly, the model should be in agreement with existing population genetics theory. The model was validated by testing the simulations against theoretical expectations of the cases that are relevant to our model. The validation process was partially based on the validation report for the simulation program **EvolGenius** v6.1 (Kliman 2014). We tested the following theoretical predictions:

Heterozygote advantage (overdominance). If the heterozygote has the highest fitness, allele proportion should reach an equilibrium, as follows: allele proportion at equilibrium $\hat{p} = s_2 / (s_1 + s_2)$, where s_1 and s_2 are the selection coefficients for the respective homozygous genotypes.

The model was tested using the following values: *Ntot* = 1,000, $s_2 = 0.4$, $s_1 = 0.1$. Simulations were run for 100 generations and 100 replicates. Observed values were in agreement with values predicted by theory: mean observed $\bar{\hat{p}} = 0.79$ (+/- 0.0098) and expected $\hat{p} = 0.8$.

Fixation time for a neutral allele. Following Kimura & Ohta (1969), the probability that a given allele will ultimately fix its starting proportion (p). The average number of generations until fixation is equal to: $\bar{t}(p) = (-4)\,(Ne)\,(q)\log_n(q) / p$, where p = initial proportion of allele A1, q = initial proportion of allele A2 and *Ne* is population size. The model was tested using the following values: *Ntot* = 100, $p = 0.5$, $q = 0.5$. Simulations were run for 1,000 generations and 100 replicates. Observed values were in agreement with theory: mean observed $\bar{t}(0.5) = 278$ (+/-150) generations and expected $\bar{t}(0.5) = 277$ generations.

Directional selection. The recurrence equation for allele proportion is: $p_{(t)} = \dfrac{p^2 w_{11} + pq w_{12}}{p^2 w_{11} + 2pq w_{12} + q^2 w_{22}}$,

where t is a measure of time in generations, $p_{(t)}$ is allele proportion of allele A1 at generation t, p = proportion of allele A1 at t-1, q = proportion of allele A2 at t-1, w_{11}= fitness homozygote A1A1, w_{12} = fitness heterozygote and w_{22} = fitness homozygote A2A2. The model was tested for a recessive deleterious allele with the following values: $w_{11} = 1$, $w_{12} = 1$, $w_{22} = 0.9$, $t = 20$, $p = 0.1$, $q = 0.9$, *Ntot* = 1,000. Simulations were run for 100 replicates. Observed values were in agreement with theory: mean observed $p_{(20)} = 0.36$ and expected $p_{(20)} = 0.36$.

Recombination. Decay of linkage disequilibrium (D) as a function of recombination rate (c) can be expressed as follows: $D_t = (1\text{-}c)^t\,(D_0)$, where t is a measure of time in generations, D_t is LD at generation t, D_0 is initial LD. The model was tested using the following values: $c = 0.01$, $t = 50$, $D_0 = 0.25$, *Ntot* = 1,000. Simulations were run for 100 replicates. Observed values were in agreement with theory: mean observed $D_{50} = 0.15$ and expected $D_{50} = 0.15$.

4. DISCUSSION AND CONCLUSIONS

The use of genetics models to guide efforts to conserve natural populations is now widespread. To draw correct conclusions from these models is crucial and therefore it is necessary to recognise under which conditions linked selection might occur, and the degree to which linked selection influences the predictions of genetic models. Previous research performed by Rumball *et al.* (1994) using the vinegar fly (*Drosophila melanogaster*) as a study species, found that the rate of loss of genetic diversity was 20% slower than was predicted by neutral models. Recent research also suggests that linked selection might have a more significant role in structuring genetic diversity than previously thought (Elyashiv *et al.* 2016). Further research is needed to understand the mechanisms and consequences of linked selection. Some research questions that warrant further investigation are:

- What is the role of population size in linked selection under new evidence suggesting that genetic drift does not overwhelm natural selection more in small than in large natural populations (Wood *et al.* 2016)?
- How new measures of genetic diversity based on information theory could give us insights about how to correct genetic models when linked selection occur (Sherwin 2015)?
- How differences in recombination spots and rates between individuals might influence linked selection (Hunter *et al.* 2016)?
- How pervasive is linked selection within the genome and between species?

In this paper, we have developed a model based on linked selection. This model was extensively tested and the results were found to be consistent with predictions from population genetics theory. By using this model as a starting point, we believe that we are able to develop further models that will allow us to investigate potential mechanisms by which linked selection alters genetic diversity and its derived consequences, such as

the above research questions. Our research may highlight the importance of the need to modify neutral models.

GLOSSARY

Dominance. Relationship between alleles of one locus, in heterozygotes, where the effect on the phenotype of one allele masks the contribution of a second allele. The first allele is dominant and the second allele is recessive.
Fitness. Describes individual survival and reproductive success. In genetics, the hard meaning of fitness of an individual is its capacity of transmitting its alleles to the next generation.
Genotype. Genetic makeup of an organism. It describes an organism's complete set of alleles. Diploid organisms have two copies of the genome, one copy inherited from each parent
Homologous chromosomes. Pair of chromosomes encoding the same traits and thus containing the same loci, however, they may contain different alleles. One of the chromosomes is inherited from the father and one from the mother.
Meiosis. Cell division that reduces the chromosome number by half, creating four haploid cells, each genetically distinct from the parent cell that gave rise to them.
Natural selection. Process whereby organisms better adapted to their environment tend to survive and produce more offspring. Alleles carried by such organisms become more frequent.
Phenotype. Observable physical properties of an organism, *e.g.* appearance, development or behavior. An organism's phenotype is determined by interaction between its genotype and the environment.
Recessive allele. Alleles that are expressed in the phenotype only in homozygotes.
Selection coefficient. For a particular genotype, the selection coefficient expresses the reduction of a contribution of offspring to the next generation with respect to the fittest genotype.

REFERENCES

Charlesworth, B. (1994). The effect of background selection against deleterious mutations on weakly selected, linked variants. *Genetical Research*, 63, 213-227.
Charlesworth, B., D. Charlesworth (2010) *Elements of Evolutionary Genetics* Roberts and Company Publishers.
Elyashiv, E., S. Sattath, T.T. Hu, A. Strutsovsky, G. McVicker, P. Andolfatto, G. Coop, G. Sellla (2016). A genomic map of the effects of linked selection in Drosophila. *PLoS genetics*, 12, e1006130.
England, P.R., J.M. Cornuet, P. Berthier, D.A. Tallmon, G. Luikart (2006). Estimating effective population size from linkage disequilibrium: severe bias in small samples. *Conservation Genetics*, 7, 303-308.
Frankham, R., J.D. Ballou, D.A. Briscoe (2009) *Introduction to conservation genetics*, 2nd edn. Cambridge University Press, Cambridge; New York.
Hunter, C.M., W. Huang, T.F.C. Mackay, N.D. Singh (2016). The genetic architecture of natural variation in recombination rate in *Drosophila melanogaster*. *PLoS genetics*, 12, e1005951.
Kimura, M., J.F. Crow (1963). The measurement of effective population number. *Evolution*, 17, 279-288.
Kimura, M., T. Ohta (1969). The average number of generations until fixation of a mutant gene in a finite population. *Genetics*, 61, 763.
Kliman, R.M. (2014). *EvolGenius 6.1 program validation.* http://www2.cedarcrest.edu/academic/bio/rkliman/EG_Validation.pdf
R Core Team (2017). R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria.*, URL https://www.r-project.org/.
Rumball, W., I.R. Franklin, R. Frankham, B.L. Sheldon (1994). Decline in heterozygosity under full-sib and double first-cousin inbreeding in *Drosophila melanogaster*. *Genetics*, 136, 1039-1049.
Sherwin, W.B. (2015). Genes are information, so information theory is coming to the aid of evolutionary biology. *Molecular Ecology Resources*, 15, 1259-1261.
Smith, J.M., J. Haigh (1974). The hitch-hiking effect of a favourable gene. *Genetical Research*, 23, 23-35.
Wood, J.L.A., M.C. Yates, D.J. Fraser (2016). Are heritability and selection related to population size in nature? Meta-analysis and conservation implications. *Evolutionary Applications*, 9, 640-657.