Genetic Load and Efficacy of Selection on Admixed Populations

Candidate: Jônatas Eduardo da Silva César[1]

Supervisor: Diogo Meyer[1]

[1] *D*epartment of Genetics and Evolutionary Biology,
Institute of Biosciences, University of São Paulo.

BEPE Supervisor: John Novembre[2]

[2] Department of Human Genetics, University of Chicago.

Abstract

To understand the interplay between demographic and selective processes in shaping genetic variation, two key quantities of interest are genetic load and the efficacy of selection, which is the contribution of selection to the rate of change of the genetic load. Demographic processes such as population splits, bottlenecks and population growth can bring about changes in the magnitude of genetic load and efficacy of selection. For humans, an emerging consensus is that among population differences in genetic load are likely to be small, while differences in the efficacy of selection may be large, with African populations having the highest estimated efficacy of selection. Admixture processes, on the other hand, can increase the genetic load and efficacy of selection above naive expectations and are still not completely understood.

Given the increasing importance of admixture in shaping the genetic variation of future human generations, we propose to theoretically and computationally investigate the genetic load and efficacy of selection of admixed populations. We aim to extend existing analytical treatments of load and efficacy of selection to models involving arbitrary numbers of populations and rates of admixture. We will also quantify the magnitude of load and efficacy of selection in various admixed populations, including those present in the 1000 Genomes and a dataset of 1320 admixed Brazilian individuals of the AbraOM project. This study will be carried out in collaboration with Prof. John Novembre who is a member of the Department of Human Genetics of the University of Chicago and an expert in the fields of genetic load and admixture of human populations.

1 Introduction

Genetic load results from the accumulation of deleterious variants in a population. According to the seminal work of Kimura et al. (1963), genetic load at equilibrium is expected to be higher in populations with smaller effective size. The intuition behind this result is that as the intensity of drift increases, there is a reduced efficacy in removing slightly deleterious variants, which segregate in the population and thus contribute to polymorphism (and thus to the load). In larger populations, where drift is a weaker force, the deterministic selective processes are responsible for the fate of most mutations, so deleterious variants are more efficiently purged from the populations and as a consequence the resulting load is lower.

Using this reasoning, Lohmueller et al. (2008) analyzed one of the first next generation sequencing (NGS) datasets, and argued that the higher proportion of non-synonymous over synonymous variants (P_N/P_S) in Europeans in comparison with Africans was evidence of increased load in Europeans caused by a decrease in the efficacy of selection during the Out-of-Africa bottleneck.

More recently, Simons et al. (2014) and Do et al. (2015), using large NGS datasets, claimed that there are no significant differences in the average number of deleterious alleles per individual when Europeans and Africans are compared. Both studies concluded that the population growth after the Out-of-Africa event results in an increase in the efficacy of selection, canceling out the effect of the population bottleneck in creating increased genetic load.

If such a cancellation due to demography occurred, why did (Lohmueller et al., 2008) identify an excess of load among non-Africans? According to Koch and Novembre (2017) as well as Simons et al. (2014) and Do et al. (2015), the apparent contradiction between their results is due to the method used to quantify load. The Lohmueller et al. (2008) study quantified load using P_N/P_S (i.e. the ratio of number of segregating non-synonymous to synonymous variants), which is a statistic that is only expected to document increased load in equilibrium conditions. This occurs when populations have maintained a constant population size for a long span of time, so that the opposing effects of mutation (introducing deleterious variants), selection (removing them) and drift (interfering with the efficacy of selection) reach a steady state. In non-equilibrium conditions (e.g., when populations recently expanded) the expectations of increase P_N/P_S for populations with smaller effective population sizes are no longer valid (Koch and Novembre, 2017).

Quantifying Genetic Load. To better understand the current knowledge on genetic load we will follow the discussion and terminology introduced by Gravel (2016). Given alleles a and A in a locus i the Malthusian fitness values of the genotypes aa, aA and AA can be written as 1, $1 + h_i s_i$ and $1 + s_i$. We assume A to be the least favored allele ($s_i < 0$) and $0 \leq h_i \leq 1$ the dominance coefficient. If the allele A is present at a frequency x_i in a random mating population it will add a value of $\delta\omega_i = s_i(2h_i x_i + (1 - 2h_i)x_i^2)$ to the mean fitness, relative to the fitness of the most favorable allele.

In the case of multiple independent loci of small effect, the net fitness is evaluated by $\omega = \prod_i (1 + \delta\omega_i) \approx 1 + \sum_i \delta\omega_i$. Then, using the definition of the genetic load as the relative fitness reduction compared to the optimal genotype, $L = (\omega_{max} - \omega)/\omega_{max}$ and since by our definition $\omega_{max} = 1$ it follows that,

$$L = -\sum_i s_i(2h_i x_i + (1 - 2h_i)x_i^2). \tag{1}$$

Notice from equation (1) that in order to properly measure differences in genetic load between populations we ideally need to know which variants in the genome are deleterious so as to estimate their allelic frequencies, fitness effects and dominance coefficients. Given that none of these variables are known with great confidence, disagreement among studies are often due disagreement regarding the assumptions on how to measure these variables (Henn et al., 2015a; Gravel, 2016). One common strategy that was used by Simons et al. (2014); Do et al. (2015) is to assume equal and constant additive fitness effects ($h_i = 0.5$, $s_i = s$) over all loci, making the estimate of genetic load proportional to the average number of putatively deleterious alleles per individual ($L_{add} = -s \sum_i x_i$).

In fact, the average number of deleterious variants per individual is a robust summary statistic to test differences in load, since it always increases with the population genetic load. Nevertheless, this statistic still depends on the prediction of the deleterious phenotype for alleles and depending on the tool which is chosen to predict deleterious effects, different conclusions can be reached. For example, in contrast to the results of Simons et al. (2014), Fu et al. (2014) used the software PhyloP (Pollard et al., 2010) to impute the non-synonymous deleterious variation (instead of Polyphen-2) (Adzhubei et al., 2010), and in an analysis of 6500 exomes of African Americans and European Americans found that there is a small but statistically significant difference in the number of deleterious mutations between these populations.

Both studies backed their conclusions with simulations based on the European Out-of-Africa demographic history, with the major difference being that Fu et al. (2014) used selection parameters based on Eyre-walker and Keightley (2007)'s distribution of fitness effects while Simons et al. (2014) explored several selection scenarios with constant fitness coefficients. Moreover, as pointed out by Henn et al. (2015b), large differences in genetic load between human populations can be measured if estimates for dominance (h_i) and fitness coefficients (s_i) are used to estimate load according to (1).

Efficacy of Selection. Despite the fact that large differences in load between human populations are still in dispute, Gravel (2016) showed that the present day efficacy of selection significantly differs between modern human populations. The efficacy of selection is defined as the rate at which the genetic load changes over time and can be quantified using the allele frequency of deleterious variants (more details in subsection 4.2).

Differences in efficacy of selection between populations have strong implications for human evolution, implying that large differences in genetic load may emerge even if the present day differences are small. Specifically, Gravel (2016) used the 1000 Genomes data and showed that the efficacy of selection is significantly higher for African populations than for Asians. Strikingly, in a result that was not theoretically explored by the author, Gravel (2016) showed that the efficacy of selection of North American admixed populations was greater than expected with respect to the weighted average of the efficacy of selection in the parental populations. These results suggest that non-linear phenomena result from the admixture processes and are important to determine the genetic load of future human generations.

2 Objectives

Population admixture has had a significant role into shaping human genetic variability (Wall and Brandt, 2016). Nevertheless, few studies using realistic simulations investigated the fate of deleterious mutations within admixed populations (Harris and Nielsen, 2016; Kim et al., 2017). For example, motivated by the evidence provided by (Sankararaman et al., 2014) that, in Europeans, the Neanderthal ancestry near conserved regions of the genome appears to be depleted, Harris and Nielsen (2016) showed that similar patterns could be replicated by realistic simulations of Human-Neanderthal split and admixture processes. In a subsequent study, (Kim et al., 2017) explored other demographic parameters in which such depletion of

Neanderthal ancestry of deleterious alleles could be obtained.

Given the increasing importance of the admixture process into shaping the genetic variation of future human generations we propose to investigate the following question:

How does the interplay between the relative contributions of source populations, dominance coefficients and the strength of selection determine the load and efficacy of selection of an admixed population ?

To investigate this question we will develop the theory of genetic load and efficacy of selection for admixed populations using Kimura's diffusion equations . This will be done by adapting the calculation done by Gravel (2016) (detailed in the subsection 4.2 for the case of a single population) with the inclusion of the appropriate terms that account for admixture events, similarly to those introduced by Jouganous et al. (2017). Furthermore, our theoretical predictions will be tested against simulated datasets as well against whole-genome sequencing data of two admixed populations.

3 Work plan

In a first phase, we will carry out theoretical studies of the processes of admixture using diffusion theory in order to dissect the contributions of the effective population size, the proportion of source populations, the intensity of selection, and the dominance coefficients to the genetic load and efficacy of selection of the admixed population. In order to test our predictions, we will perform a series of simple evolutionary simulations of admixture processes with two source populations. By checking for consistency between the analytical treatment and the simulations we will validate both of these approaches.

In a second phase, we will test the predictions of the theoretical modeling of genetic load and efficacy of selection against two genomic data sets of admixture population. The first dataset is composed of the North American admixed population of the 1000 Genomes project **?**. The analysis of this data will be followed by extensive simulations based on the demographic history inferred by (Gravel et al., 2013). In this step we will be able compare the estimation of load and efficacy of selection for the real data with the expected values of simulations carried out using different values of dominance and selection coefficients.

The second data analysis will be done in collaboration with the *"Centro de Estudos do*

Genoma Humano" from the University of São Paulo. Here, we will study the genetic load and efficacy of selection of Brazilian admixed population using a whole-exome sequencing data of 1320 individuals of the AbraOM project (Naslavsky et al., 2017) which is funded under FAPESP process #2014/50931-3. We will restrict our analyses of this data by using only the subset of coding variants that passes the same quality control pipeline proposed in Naslavsky et al. (2017) which consists in the use of GATK quality control filters.

4 Materials and Methods

4.1 Simulations

There are several tools in the literature that can be used to simulate evolutionary processes with complex demography and to thoroughly explore the parameter space and nuances of the theoretical predictions. We propose to employ two layers of simulations. The first will use the *ms* software (Hudson, 2002) to carry out neutral coalescent simulations of the studied populations, incorporating realistic demographic histories. The coalescent neutral simulations are computationally efficient and will provide the null distributions for the genetic variability since the load and efficacy of selection statistics are equal to zero by definition.

For the second layer, we will use software such as *moments* (Jouganous et al., 2017) and *SLiM-2* (Haller and Messer, 2016) to simulate the site frequency spectrum under several combinations of dominance and selections coefficients. The software *moments* implements the evolution in time of the Kimura's diffusion equation for complex demographic scenarios including admixture process with two or more populations. The software *SLiM-2* implements a realistic simulations using the Wright-Fisher model that takes into account the position of selected variants, recombination and dominance parameters as well the distribution of fitness effects.

With the proposed set of simulations we will be able to test the dependency of load and efficacy of selection in admixed population with the proportion and effective sizes of source populations, and to evaluate the effect of different combinations of dominance and selection regimes.

4.2 Efficacy of selection derivation

Differences in genetic load between populations can be understood in the light of the concept of efficacy of selection, which is the capacity of natural selection to increase the population fitness by removing deleterious variation. This process counters the effects of drift and mutation which are forces that can decrease the population fitness by increasing the frequency of deleterious mutation and by introducing new deleterious mutations in the population, respectively.

Nevertheless, only recently did Gravel (2016) explore a precise statistical test to measure differences in efficacy of selection based on the temporal variation of genetic load and allele frequencies in large populations, using the Kimura's diffusion equation (Kimura, 1955a,b).

Let the allele frequency moments be defined by $\mu_k \equiv \int_0^1 x^k \phi(x,t)dx$, the fitness in the diploid case can be written as

$$W = s[2h\mu_1 + (1-2h)\mu_2] \tag{2}$$

Then the rate of change in fitness will depend on the derivatives of the first and second moment which can be calculate using the diffuion equation for $\phi(x)$.

The diffusion equation defines the temporal evolution of the site frequency spectrum function $\phi(x,t)$, which gives the probability density that one allele is at a frequency x at a time t in a randomly mating population of size $N = N(t) \gg 1$. A modern version of the diffusion equation, as defined by Ewens (2004), is given by,

$$\begin{aligned} \frac{\partial}{\partial t}\phi(x,t) \approx & \frac{1}{4N}\frac{\partial^2}{\partial x^2}x(1-x)\phi(x,t) \\ & - s\frac{\partial}{\partial x}(h+(1-2h)x)x(1-x)\phi(x,t) \\ & + 2Nu\delta(x-\frac{1}{2N}), \end{aligned} \tag{3}$$

where u is the mutation rate, δ is Dirac's delta function, and s and h are constant over time. From this equation we can calculate the evolution for the allele frequency moments.

In the infinite allele model there is pontentially a infinity number of alleles ate frequency 0 and 1 so it is convinient to write

$$\mu_k \equiv \int_0^1 dx x^k \phi(x,t)dx = \int_{0^+}^{1^-} dx x^k \phi(x,t) + K_0\delta_{0,k} + K_1$$

where K_0 and K_1 are the number of sites at frequency 0 and 1 and $\delta_{0,k}$ is the Kronecker's delta.

$$\begin{aligned}\dot{\mu}_k &= \int_{0^+}^{1^-} dx x^k \frac{\partial}{\partial t}\phi(x,t) + \dot{K}_0\delta_{0,k} + \dot{K}_1 \\ &= \frac{\partial}{\partial t}\int_{0^+}^{1^-} dx x^k \phi(x,t) + \dot{K}_0\delta_{0,k} + \dot{K}_1 \end{aligned} \tag{4}$$

Writing the diffusion equation as

$$\frac{\partial}{\partial t}\phi(x) = \frac{1}{4N}\frac{\partial^2}{\partial x^2}f(x)\phi(x) - s\frac{\partial}{\partial x}g(x)\phi(x) + 2Nu\delta(x - \frac{1}{2N})$$

We can write

$$\begin{aligned}\dot{\mu}_k = &\left.\frac{1}{4N}\int_{0^+}^{1^-} dx x^k \frac{\partial^2}{\partial x^2}[f(x)\phi(x,t)] \qquad\right\} I_1 \\ &\left. - s\int_{0^+}^{1^-} dx x^k \frac{\partial}{\partial x}[g(x)\phi(x,t)] \qquad\right\} I_2 \\ &+ \frac{u}{(2N)^{k-1}} + \dot{K}_0\delta_{0,k} + \dot{K}_1 \end{aligned} \tag{5}$$

For $k = 0$, it follows

$$\begin{aligned}\dot{\mu}_0 &= \frac{1}{4N}\frac{d}{dx}[f(x)\phi(x)]_{0^+}^{1^-} - s[g(x)\phi(x)]_{0^+}^{1^-} + 2Nu + \dot{K}_0 + \dot{K}_1 \\ &= -\frac{\phi(0^+) + \phi(1^-)}{4N} + 2Nu + \dot{K}_0 + \dot{K}_1 \end{aligned} \tag{6}$$

Because the number of sites is constant $\dot{\mu}_0 = 0$ then we can require
for $k = 0$ it follows

$$\dot{K}_0 = \frac{\phi(0^+)}{4N} - 2Nu \tag{7}$$

$$\dot{K}_1 = \frac{\phi(1^-)}{4N} \tag{8}$$

For $k \geq 1$ we can solve by parts the integrals (5) as follows

$$\begin{aligned}
I_2 &= -s\{[g(x)\phi(x)x^k]_{0^+}^{1^-} - k\int_{0^+}^{1^-} dxx^{k-1}[g(x)\phi(x,t)]\} \\
&= -sk\int_{0^+}^{1^-} dxx^{k-1}(h+(1-2h)x)x(1-x)\phi(x,t) \\
&= sk\int_{0^+}^{1^-} dx[h(x^{k+1}-x^k)+(1-2h)(x^{k+2}-x^{k+1})]\phi(x,t) \\
&= sk[h(\mu_{k+1}-\mu_k)+(1-2h)(\mu_{k+2}-\mu_{k+1})] && (9) \\
I_1 &= \frac{1}{4N}\{[\frac{d}{dx}(f(x)\phi(x)x^k]_{0^+}^{1^-})] - k\int_{0^+}^{1^-} dxx^{k-1}\frac{d}{dx}[f(x)\phi(x,t)]\} \\
&= \frac{1}{4N}\{[\frac{d}{dx}(f(x)\phi(x))x^k]_{0^+}^{1^-})] - k[f(x)\phi(x)x^{k-1}]_{0^+}^{1^-})] \\
&+ k(k-1)\int_{0^+}^{1^-} dxx^{k-2}f(x)\phi(x,t)\} \\
&= \frac{1}{4N}\{\phi(1^-)+k(k-1)\int_{0^+}^{1^-} dx(x^{k-1}-x^k)\phi(x,t)\} \\
&= \dot{K}_1 - \frac{k(k-1)}{4N}(\mu_k-\mu_{k-1}) && (10)
\end{aligned}$$

It follows

$$\dot{\mu_k} = \frac{k(k-1)}{8N}\pi_{k-1} + \frac{sk}{4}\Gamma_{k,h} + \frac{1}{(2N)^{k-1}}u, \tag{11}$$

where $\pi_k = 2(\mu_k - \mu_{k+1})$ and $\Gamma_{k,h} = 2[h\pi_k + (1-2h)\pi_{k+1}]$.

The expected change in genetic load can be written as three components representing the instantaneous contributions of selection, mutation and drift:

$$\dot{W} = \dot{W}_s + \dot{W}_u + \dot{W}_N,$$

where

$$\begin{aligned}
\dot{W}_s &= s\left[\frac{s}{1}(h\Gamma_{1,h}+)1-2h)\Gamma_{2,h})\right], && (12) \\
\dot{W}_u &= s(2hu), \\
\dot{W}_N &= s\left[\frac{1-2h}{4N}\pi_1\right].
\end{aligned}$$

When two populations diverge over time the mutation component, $\dot{W}_u$, does not contribute to load differences between the populations since it is assumed to be constant over time and shared among populations. The drift term, $\dot{W}_N$, is explicitly dependent of population size and leads to increased differentiation between populations over time. Nevertheless, in the case of variants of additive effects, the drift component is equal to zero. More importantly, the selection component, $\dot{W}_s$, defines the efficacy of selection and is the main contribution to differences in load between large populations over time. It is important to emphasize that equation (12) can be easily estimated from real data since it only depends on estimation of allele frequency moments.

The function $\dot{W}_s$ is called the "FIT" efficacy of selection given its contribution to the fitness increase theorem (Gravel, 2016; Ewens, 2004). Under an additive scenario $\dot{W}_s$ is called the Morton efficacy of selection and is equivalent to the measurement of the rate of change in the average number of deleterious variants in a population.

Equation (12) can be used as a summary statistic to measure differences of efficacy between two populations. Nevertheless, when analyzing admixture events, this formula is not informative about the dependence of the efficacy of selection in the admixed population with the parameters of the source population. We therefore propose to develop the formulation of the efficacy of selection for admixed population using a similar approach as the one proposed by Jouganous et al. (2017), numerically evaluating the time dynamic of the function $\phi(x,t)$ for multiple populations under complex demographic scenarios including admixture events.

5 Efficacy of Selection of Admixed Population

In a single pulse admxiture model we can write the

$$\phi(x_1,x_2,x_3) = \phi(x_1,x_2)\delta(x_3-\alpha_1 x_1-\alpha_2 x_2)$$

In the infinite allele model there is pontentially a infinity number of alleles ate frequency 0 and 1 so it is convinient to write

$$\mu_{3,k} \equiv \int_0^1 dx_3 x_3^k \phi(x_3,t)dx = \int_{0^+}^{1^-} dx_3 x_3^k \phi(x_3,t) + K_{3,0}\delta_{0,k} + K_{3,1}$$

where K_0 and K_1 are the number of sites at frequency 0 and 1 and $\delta_{0,k}$ is the Kronecker's delta.

then we can calculate the moments of the admix population as

$$\begin{aligned}
\mu_{3,k} &= \int_0^1 \int_0^1 \int_0^1 dx_1 dx_2 dx_3 x_3^k \phi(x_1, x_2) \delta(x_3 - \alpha_1 x_1 - \alpha_2 x_2) \\
&= \int_0^1 \int_0^1 dx_1 dx_2 (\alpha_1 x_1 + \alpha_2 x_2)^k \phi(x_1, x_2) \\
&= \int_0^1 \int_0^1 dx_1 dx_2 \sum_{j=1}^{k} \binom{j}{k} (\alpha_1 x_1)^j (\alpha_2 x_2)^{k-j} \phi(x_1, x_2) \\
&= \sum_{j=1}^{k} \binom{j}{k} \alpha_1^j \alpha_2^{k-j} \int_0^1 \int_0^1 dx_1 dx_2 x_1^j x_2^{k-j} \phi(x_1, x_2) \\
&= \sum_{j=1}^{k} \binom{j}{k} \alpha_1^j \alpha_2^{k-j} \mu_{j,k-j}^{1,2} \qquad (13)
\end{aligned}$$

where $\mu_{j,k-j}^{1,2} = \int_0^1 \int_0^1 dx_1 dx_2 x_1^j x_2^{k-j} \phi(x_1, x_2)$
The change in time of the moments of the admix populations is

$$\begin{aligned}
\dot{\mu}_{3,k} &= \int_{0^+}^{1^-} \int_{0^+}^{1^-} \int_{0^+}^{1^-} dx_1 dx_2 dx_3 x_3^k \frac{\partial}{\partial t} \phi(x_1, x_2, x_3, t) + \dot{K}_{3,0} \delta_{0,k} + \dot{K}_{3,1} \\
&= \frac{\partial}{\partial t} \int_{0^+}^{1^-} \int_{0^+}^{1^-} \int_{0^+}^{1^-} dx_1 dx_2 dx_3 x_3^k \phi(x_1, x_2, x_3, t) + \dot{K}_{3,0} \delta_{0,k} + \dot{K}_{3,1} \\
&= \sum_{j=1}^{k} \binom{j}{k} \alpha_1^j \alpha_2^{k-j} \frac{\partial}{\partial t} \mu_{j,k-j}^{1,2} \qquad (14)
\end{aligned}$$

The diffusion equation for 2 populations is given by

$$\begin{aligned}
\frac{\partial}{\partial t} \phi(x_1, x_2, t) \approx{}& \frac{1}{4N_1} \frac{\partial^2}{\partial x_1^2} x_1 (1 - x_1) \phi(x_1, x_2, t) + \frac{1}{4N_2} \frac{\partial^2}{\partial x_2^2} x_2 (1 - x_2) \phi(x_1, x_2, t) \qquad (15) \\
&- s \frac{\partial}{\partial x_1} (h + (1 - 2h) x_1) x_1 (1 - x_1) \phi(x_1, x_2, t) \\
&- s \frac{\partial}{\partial x_2} (h + (1 - 2h) x_2) x_2 (1 - x_2) \phi(x_1, x_2, t) \\
&+ 2N_1 u \delta(x_1 - \frac{1}{2N_1}) + 2N_2 u \delta(x_2 - \frac{1}{2N_2})
\end{aligned}$$

Writing the diffusion equation as

$$\begin{aligned}\frac{\partial}{\partial t}\phi(x_1,x_2) = \frac{1}{4N_1}\frac{\partial^2}{{\partial x^2}_1}f_1(x_1)\phi(x_1,x_2) - s\frac{\partial}{{\partial x}_1}g_1(x_1)\phi(x_1)\\ \frac{1}{4N_2}\frac{\partial^2}{{\partial x^2}_2}f_2(x)\phi(x_1,x_2) - s\frac{\partial}{{\partial x}_2}g_2(x_2)\phi(x)\\ +2N_1u\delta(x_1-\frac{1}{2N_1}) + 2N_2u\delta(x_2-\frac{1}{2N_2})\end{aligned} \tag{16}$$

We can write

$$\begin{aligned}\dot{\mu}^{12}_{kj} &= \frac{1}{4N_1}\int_{0^+}^{1^-}dx_1x_1^k\frac{\partial^2}{{\partial x_1}^2}[f_1(x_1)\psi_1^j(x_1,t)] && \Big\}I_{1,1}\\ &+\frac{1}{4N_2}\int_{0^+}^{1^-}dx_2x_2^j\frac{\partial^2}{{\partial x_2}^2}[f_2(x_2)\psi_2^k(x_2,t)] && \Big\}I_{1,2}\\ &-s\int_{0^+}^{1^-}dx_1x_1^k\frac{\partial}{{\partial x}_1}[g_1(x_1)\psi_1^j(x_1,t)] && \Big\}I_{2,1}\\ &-s\int_{0^+}^{1^-}dx_2x_2^j\frac{\partial}{{\partial x}_2}[g_2(x_2)\psi_2^k(x_2,t)] && \Big\}I_{2,2}\\ &+\frac{u}{(2N_1)^{k-1}}+\frac{u}{(2N_2)^{j-1}}+\dot{K}_0^{12}\delta_{0,k}+\dot{K}_1^{12} && \Big\}I_0\end{aligned} \tag{17}$$

where $\psi_1^j(x_1) = \int_{0^+}^{1^-}dx_2x_2^j\phi(x_1,x_2)$.

For $k=0$, it follows

$$\begin{aligned}\dot{\mu}^{1,2}_{0,0} &= \frac{1}{4N_1}\frac{d}{{dx}_1}[f_1(x_1)\psi_1^0(x_1)]_{0^+}^{1^-} - s[g_1(x_1)\psi_1^0(x_1)]_{0^+}^{1^-}\\ &\quad+\frac{1}{4N_2}\frac{d}{{dx}_2}[f_2(x_2)\psi_1^0(x_2)]_{0^+}^{1^-} - s[g_2(x_1)\psi_2^0(x_2)]_{0^+}^{1^-}\\ &\quad+2N_1u+2N_2u+\dot{K}_{3,0}+\dot{K}_{3,1}\\ &= -\frac{1}{4N_1}[\int_{0+}^{1-}dx_2\phi(0^+,x_2)+\int_{0+}^{1-}dx_2\phi(1^-,x_2)]\\ &\quad-\frac{1}{4N_2}[\int_{0+}^{1-}dx_1\phi(x_1,0^+)+\int_{0+}^{1-}dx_1\phi(x_1,1^-)]\\ &\quad+2N_1u+2N_2u+\dot{K}_{3,0}+\dot{K}_{3,1}\end{aligned} \tag{18}$$

Because the number of sites is constant $\dot{\mu}_0 = 0$ then we can require

for $k = 0$ it follows

$$\begin{aligned}
\dot{K}_{3,0} &= \frac{1}{4N_1}\int_{0+}^{1-} dx_2\phi(0^+, x_2) + \frac{1}{4N_2}\int_{0+}^{1-} dx_1\phi(x_1, 0^+) - 2u(N_1 + N_2) && (19)\\
\dot{K}_{3,1} &= \frac{1}{4N_1}\int_{0+}^{1-} dx_2\phi(1^-, x_2) + \frac{1}{4N_2}\int_{0+}^{1-} dx_1\phi(x_1, 1^-) && (20)
\end{aligned}$$

For $k \geq 1$ we can solve by parts the integrals (17) as follows

$$\begin{aligned}
I_{2,1} &= -s\{[g_1(x_1)\psi_1^j(x_1)x_1^k]_{0^+}^{1^-} - k\int_{0^+}^{1^-} dx_1 x_1^{k-1}[g_1(x_1)\psi_1^j(x_1,t)]\}\\
&= -sk\int_{0^+}^{1^-} dx_1dx_2x_1^{k-1}x_2^j(h + (1-2h)x_1)x_1(1-x_1)\phi(x_1,x_2,t)\\
&= sk\int_{0^+}^{1^-} dx_1dx_2[h(x_1^{k+1} - x_1^k) + (1-2h)(x_1^{k+2} - x_1^{k+1})]x_2^j\phi(x_1,x_2,t)\\
&= sk[h(\mu^{12}_{k+1,j} - \mu^{12}_{k,j}) + (1-2h)(\mu^{12}_{k+2,j} - \mu^{12}_{k+1,j})] && (21)\\
I_{1,1} &= \frac{1}{4N_1}\{[\frac{d}{dx}_1(f_1(x_1)\psi_1(x_1)x_1^k]_{0^+}^{1^-})] - k\int_{0^+}^{1^-} dx_1x_1^{k-1}\frac{d}{dx}_1[f_1(x_1)\psi_1(x_1,t)]\}\\
&= \frac{1}{4N_1}\{[\frac{d}{dx}_1(f_1(x_1)\psi(x_1))x_1^k]_{0^+}^{1^-})] - k[f_1(x_1)\psi(x)x_1^{k-1}]_{0^+}^{1^-})]\\
&+ k(k-1)\int_{0^+}^{1^-} dx_1dx_2x^{k-2}x_2^jf_1(x_1)\phi(x_1,x_2,t)\}\\
&= \frac{1}{4N_1}\{\int_{0^+}^{1^-} dx_2x_2^j\phi(1^-,x_2) + k(k-1)\int_{0^+}^{1^-} dx(x^{k-1} - x^k)x_2^j\phi(x,t)\}\\
&= \frac{1}{4N_1}\int_{0^+}^{1^-} dx_2x_2^j\phi(1^-,x_2) - \frac{k(k-1)}{4N}(\mu^{12}_{k,j} - \mu^{12}_{k-1,j}) && (22)
\end{aligned}$$

Finally

$$\begin{aligned}
\dot{\mu}^{12}_{kj} &= I_{1,1} + I_{1,2} + I_{2,1} + I_{2,2} + I_0\\
&= \frac{1}{4N_1}\int_{0^+}^{1^-} dx_2x_2^j\phi(1^-,x_2) + \frac{k(k-1)}{4N_1}(\mu^{12}_{k-1,j} - \mu^{12}_{k,j})\\
&+ \frac{1}{4N_2}\int_{0^+}^{1^-} dx_1x_1^k\phi(x_1,1^-) + \frac{j(j-1)}{4N_2}(\mu^{12}_{k,j-1} - \mu^{12}_{k,j})\\
&+ sk[h(\mu^{12}_{k,j} - \mu^{12}_{k+1,j}) + (1-2h)(\mu^{12}_{k+1,j} - \mu^{12}_{k+2,j})]\\
&+ sj[h(\mu^{12}_{k,j} - \mu^{12}_{k,j+1}) + (1-2h)(\mu^{12}_{k,j+1} - \mu^{12}_{k,j+2})] && (23)
\end{aligned}$$

References

I. a. Adzhubei, S. Schmidt, L. Peshkin, et al. A method and server for predicting damaging missense mutations. *Nature methods*, 7(4):248–249, 2010. ISSN 1548-7091.

R. Do, D. Balick, H. Li, and I. Adzhubei. No evidence that natural selection has been less effective at removing deleterious mutations in Europeans than in West Africans. *Nature Genetics*, 47(2), 2015. ISSN 1061-4036.

W. Ewens. *Mathematial population genetics. I. Theoretical introduction, 2nd edition.* 2004.

A. Eyre-walker and P. D. Keightley. The distribution of fitness effects of new mutations. *Nature Reviews Genetics*, 8(August):610–618, 2007.

W. Fu, R. M. Gittelman, M. J. Bamshad, and J. M. Akey. Characteristics of Neutral and Deleterious Protein-Coding Variation among Individuals and Populations. *The American Journal of Human Genetics*, 95 (4):421–436, 2014. ISSN 0002-9297.

S. Gravel. When is selection effective? *Genetics*, XXX (September):1–23, 2016. ISSN 0016-6731.

S. Gravel, F. Zakharia, A. Moreno-estrada, et al. Reconstructing Native American Migrations from Whole-Genome and Whole-Exome Data. *PLoS genetics*, 9 (12), 2013.

B. C. Haller and P. W. Messer. SLiM 2: Flexible, Interactive Forward Genetic Simulations. *Molecular Biology and Evolution*, 34(November):230–240, 2016.

K. Harris and R. Nielsen. The Genetic Cost of Neanderthal Introgression. *Genetics*, 203(June):881–891, 2016.

B. M. Henn, L. R. Botigué, C. D. Bustamante, A. G. Clark, and S. Gravel. Estimating the mutation load in human genomes. *Nature Reviews Genetics*, 16 (June), 2015a. ISSN 1471-0056.

B. M. Henn, L. R. Botigué, S. Peischl, et al. Distance from sub-Saharan Africa predicts mutational load in diverse human genomes. *Proceedings of the National Academy of Sciences*, page 201510805, 2015b. ISSN 0027-8424.

R. R. Hudson. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics (Oxford, England)*, 18(2):337–338, 2002. ISSN 1367-4803.

J. Jouganous, W. Long, A. P. Ragsdale, and S. Gravel. Inferring the Joint Demographic History of Multiple Populations: Beyond the Diffusion Approximation. *Genetics*, XXX(May), 2017.

B. Y. Kim, C. D. Huber, and K. E. Lohmueller. Deleterious variation mimics signatures of genomic incompatibility and adaptive introgression. *arXiv*, pages 1–29, 2017.

Kimura. Random genetic drift in multi-allelic locus. *Evolution*, 9(4):419–435, 1955a.

M. Kimura. Solution of a process of random genetic drift with a continuous model. *Pnas*, 2(2):144–150, 1955b.

M. Kimura, T. Marayama, and J. F. Crow. The mutation load in small populations. *Evolution*, (918): 1303–1312, 1963.

E. Koch and J. Novembre. A Temporal Perspective on the Interplay of Demography and Selection on Deleterious Variation in Humans. *G3 (Bethesda, Md.)*, 7 (March):g3.117.039651, 2017. ISSN 2160-1836.

K. E. Lohmueller, A. R. Indap, S. Schmidt, et al. Proportionally more deleterious genetic variation in European than in African populations. *Nature*, 451(7181): 994–997, 2008. ISSN 0028-0836.

M. S. Naslavsky, D. Bozoklian, M. Lazar, et al. Exomic variants of an elderly cohort of Brazilians in the ABraOM database. *Human Mutation*, 38(December 2016):751–763, 2017.

K. S. Pollard, M. J. Hubisz, K. R. Rosenbloom, and A. Siepel. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research*, 20:110–121, 2010.

S. Sankararaman, S. Mallick, M. Dannemann, et al. The genomic landscape of Neanderthal ancestry in present-day humans. *Nature*, 507:354–357, 2014.

Y. B. Simons, M. C. Turchin, J. K. Pritchard, and G. Sella. The deleterious mutation load is insensitive to recent population history. *Nature genetics*, 46(3):220–4, 2014. ISSN 1546-1718.

J. D. Wall and D. Y. C. Brandt. Archaic admixture in human history. *Current Opinion in Genetics & Development*, 41:93–97, 2016. ISSN 0959-437X.