MASTER´S PROJECT PRESENTATION

LUND
UNIVERSITY

Computational Simulations of Evolutionary Dynamics: The Fate of Deleterious Alleles

BY

SIMON CESARINI
Department of Biology, Lund University

SUPERVISOR: TORBJÖRN SÄLL

Computational Simulations of Evolutionary Dynamics: The Fate of Deleterious Alleles

Simon Cesarini
Department of Biology, Lund University, Lund, Sweden

Abstract: When a severe deleterious mutation appears in a population, it is expected to disappear through negative selection within a few generations. However, the variance of this number is significantly large to allow some deleterious mutations to exist for several generations. To extend the understanding about these dynamics can help to prevent and treat genetic disease in humans and other species. In order to understand this evolutionary process, computer simulations of deleterious mutations in populations have been performed. This will answer fundamental questions such as expected number of individuals affected by mutation, as well as number of generations until extinction. The approach of simulation will confirm results primarily calculated before, but will also outline completely new findings, such as how the average number of individuals in a mass of family lines with a deleterious mutation strives towards an equilibrium-like state, and how haplotype frequencies in a population can be used to find probable relationships between individuals with similar phenotype.

INTRODUCTION

The dynamics relating to how deleterious mutations can persist throughout a considerable number of generations has been studied through a number of mathematical models published during the 20th and 21st century. These have not been outright simulations but rather calculations, most notably the *'branching process'* based on the backward Kolmogorov equation and used by Fischer and Haldane, and later the *'diffusion process'* by Kimura and Otha (1969). Both these were further used by Li and Nei in studies with incomplete dominance (1972). In focus for these earlier studies has been the so called *persistence*, defined by Muller (1950) and is defined for how many generations a newly arisen deleterious mutated allele will exist before it is lost due to negative selection. The persistence is tightly coupled to what can be called the *pervasiveness*, stating how many individuals in total are carrying the allele over all generations.

From these models some descriptive properties, such as *variance of persistence*, can be derived with relative ease. Some properties are however neither easily accessible nor comprehensively understood using only mathematics.

Instead, this project will investigate the underlying evolutionary models through simulations with computer programs. The output from these programs will be samples from statistical populations of outcomes. There are several advantages with this approach, such as the convenience of only requiring mathematics of relatively low complexity. This offers transparency in how data is treated and processed, and also offers the ability to adjust or expand the model to accommodate for different variables. This study will focus on dominant alleles and recessive X-linked alleles, and the programming is done in the language *'R'*.

THE PROBLEMS UNDER STUDY

The problems in this study will be solved with simulations of adjusted *Wright–Fisher models*, were alleles in a generation are randomly drawn from the alleles from the former generation. In its simplest version, the assumptions are random mating, stable population size in *Hardy-Weinberg equilibrium* as well as non-overlapping generations.

An individual who carries a dominant allele that affect the overall fitness of the individual will therefore result in a different average number of offspring. This can of course depended on various factors in the real world, but will be summarized into a single value; *the selection coefficient*, greatly simplifying the work of simulating the model. Since the selection coefficient is a summary of the selective force, it is sometimes simply referred to as 'selection'. For the same reason, simulating a dominant allele is the natural first step when looking at deleterious mutations, since only individuals carrying the allele need to be considered in the simulation. When the diploid carrier produces offspring, the mate is assumed to be a non-carrier. The model for a recessive and X-linked allele is the next natural step and extends this study further. This is because it can be thought of as a special case of the dominant model. If a male is carrying the recessive deleterious allele, he will be hemizygote for the allele, lacking a 'healthy' X chromosome. The effect in this case will be as if the recessive deleterious mutant allele were in fact dominant. The females carrying the allele will not be affected for an allele that is completely recessive, but can pass it on to its offspring. With some modifications, it is also possible to simulate if the 'healthy' allele is not completely dominant over the new mutated allele.

In a population of a species that reproduces asexually, each individual has an average number of offspring equal to one if a stable population is to be maintained. For species that reproduces sexually, each pair should of course have an average number of offspring equal to two. However, for a single neutral allele, this means that the average number of alleles in the next generation should be one, regardless of the way the specie reproduces, and this is what will be simulated in the programs. That means that only offspring were the allele is present needs to be considered, and non-carrying offspring can be excluded from the simulation. Therefore, what is actually simulated is the propagation of an allele in a population over generations, not the entire family lines of individuals per se, although for this somewhat abstract study, it is not damaging for the data points to be considered as individuals and not alleles.

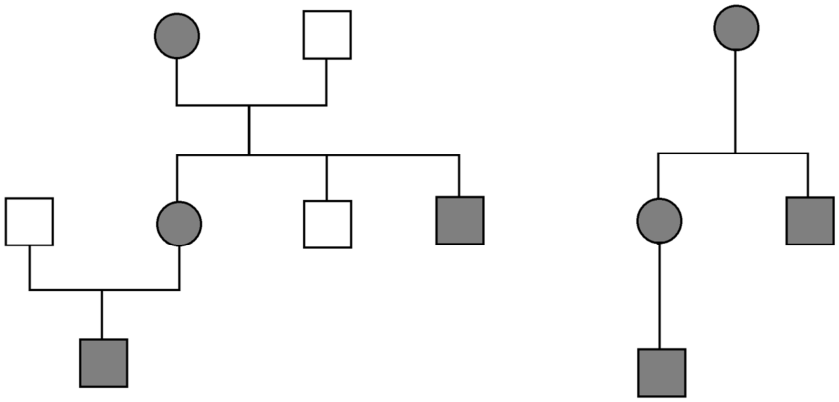

Figure 1: Grey color represents carriers of the mutant allele. To the left what a family tree could look like in the real world, and to the right the minimalistic way they are simulated in the programs (excluding the non-carriers).

In the programs, there are only heterozygotes (or hemizogots for males in the X-linked case). This is of course a simplification, but unless there is inbreeding, a newly arisen dominant mutation will almost never exist in homozygote form for many generations, and then the probability is still very low for an interbreeding population of considerable size.

In a population that fulfills the given assumptions (steady population size, et cetera); new mutations will be inserted at random over generations. Each type of hypothetical deleterious mutation will have its own mutation frequency; the average number of new mutated alleles of that type in each generation. Also, because of selection there will be a number of older mutated alleles of this type that will disappear from the population each generation. Therefore, the mutations will be introduced at a steady rate, but the number of mutations that will disappear each generation will be a fraction of the mutated alleles present in the population at that specific time. This will result in the *mutation-selection equilibrium*, where new alleles of a specific type will be introduced by mutation at the same rate as alleles of that type disappears by selection (Crow 1986).

What is considered a 'type' of mutation is not always clear and depends on the context. There is a distinction between *identical by type (IBT) and identical by decent (IBD)*. Two or more alleles are IBD if they share common ancestor from which the mutated allele has been transmitted. The probability for two individuals sharing an IBD allele will be simply referred to as '*identity*'. IBT are mutations that are attributed with very similar phenotypes and where the actual DNA change is in the same locus for the allele. Therefore, if two alleles are IBD, they also have to be IBT, unless they have undergone further divergent evolution. For example, all types of hemophilia A is considered to be caused by the same type of mutations, since they give rise to similar phenotypes (reduced blood clotting) *and* have the DNA damage in the same locus (*F8* gene). This is true even if the actual DNA changes can differ in different lineages. In contrast, hemophilia B is

considered to be caused be another type of mutation than A, even though the phenotypes can be similar. This is because it is caused by mutations in another locus (*F9* gene).

If we consider three individuals, there are three different constellations of IBD relationships: all of them share an IBD allele, two of them share it and the third one has a recurrent mutation, or all of them have recurrent mutations. Note that it is of course impossible for exactly two of the three individuals to have a recurrent mutation, since the third must have someone to have an IBD allele with.

In the populations generated by these programs, there will be mutated alleles of the same type. Some will have the same origin and have alleles that are IBD, and all will be IBT for each specific setting of the program. Individuals enter the programs with a newly mutated allele and no further mutations will be simulated (including back mutations) in order to simplify the simulations.

Aims

The programs used in this study will be utilized for a statistical approach to investigate and examine the population dynamics of deleterious alleles. Of interest is the *Persistence*, which is for how many generations a newly arisen deleterious mutated allele will exist before it is lost due to negative selection. The persistence is tightly coupled to the *Pervasiveness*, how many individuals in total are carrying the allele over all generations. In ***figure 1***, the persistence is 3 (generations), and the pervasiveness is 4 (individuals). In a population where deleterious mutations of a certain type appears and disappears all the time, there is an *Expected age* of that type of allele. That is, if a deleterious allele is drawn by random from the population of alleles of that type, what is the best guess of the number of generations since the mutation event for that allele? Also important in the context is *Lineage*, which will be used for describing the first individual carrying a mutation and its entire offspring in all future generations, living or not.

In this study, there is information about which individuals have an IBD allele and which do not. This is not always the case in the real world, where distant kinship is not always known. What can be done in the real world is to *haplotype* the individuals of interest, if they share the underlying haplotype surrounding the allele, thus increasing the probability of IBD rising. If the haplotype is very uncommon, it is unlikely for that haplotype to have muted in a similar way twice, but if the underlying haplotype is common in the population, it is still possible that there are two or more independent mutation events. Since this kind of information will be available in this study, the probabilities for different kinds of constellations of IBD mutations and recurrent mutations can be calculated. It will also be investigated how the frequency of the underlying haplotype changes this

probability. In this study, these probabilities for constellations of people in small groups, of two, three or four individuals, will be calculated.

METHODS & RESULTS

Programs & algorithms

There are two main algorithms with different structures for the programs, each with different applications. Both algorithms are in turn applied on a dominant and an x-linked recessive inheritance pattern. The purpose of the algorithms is to be as simple as possible, while still mimic the real life process.

The first algorithm: Double iteration loop

This algorithm begins with a single individual. It belongs to generation zero and is carrying the new mutation. Step two is reproduction. The number of offspring with the mutation (putative non-carriers are excluded from simulation) is Poisson distributed according to standard population genetics $\boldsymbol{X \sim Po(\lambda)}$, where X is the number of offspring with the new mutation and λ is the expected value. The full formula for the Poisson function is: $f(x) = \frac{\lambda^x e^{-\lambda}}{x!}$ were e is Euler's number. In the programs it is assumed that the expected value is equal to the fitness ω. The fitness will be equal to one minus the selection coefficient $\lambda=\omega=1-s$. Therefore: $\boldsymbol{X \sim Po(1-s)}$. Next, each offspring in generation 2 will each get the same probability to reproduce and so on. This means that in the world of the program, each generation of individuals are completely replaced by the next, in the manner a population of annual plants would behave over several years. This is of course not the way humans and many other organisms reproduce, but the model works surprisingly well for most breeding patterns, including that of humans.

As soon as all the offspring in a generation is equal to zero, the family lineage is gone. It could be that the linage only consisted of the first single individual who did not manage to get any offspring. Or it could be that the family linage existed for many generations and in total included many individuals, but finally came to an end. But no matter the genealogy, if a linage is gone the program leaves the generation loop and starts a new one with a new single individual. This process will go on for the specified number of iterations, maybe 10 000 or more.

For the X-linked recessive version some things are a bit different. As in the dominant version the simulation starts with a single individual. This individual will be female with a probability of two thirds and male with one third. This is due to the assumption that the probability of any given X chromosome to mutate is the same regardless of sex. Males will be hemi-zygotes for the mutation, but will not reproduce in the same manner as individuals in the dominant version. All female offspring from males will be

carriers. Since all male offspring from males will be healthy non-carriers, they will not be included in further simulation. Females will be heterozygotes and will have the heterozygote selective disadvantage *hs* when reproducing, were h is the coefficient of dominance. For example, if the selection coefficient is 0.2 and *h*=0.5, males will have a fitness of 0.8, and heterozygote females a fitness of 0.9. Therefore:

$\boldsymbol{X_{females} \sim Po(1 - hs)}$ and $\boldsymbol{X_{males} \sim Po(1 - s)}$.

Persistence

A linage from a mutated allele will persist for a certain number of generations and then be lost. Depending on the strength of the selection affecting the allele, a population of alleles of the same type is expected to persist to a certain generation with a specific frequency. Therefore, the unit for persistence is number of generations. Important to have in mind is that for all alleles with a selection disadvantage, the very first generation will always have the highest frequency of lost alleles. This is true even though most alleles will persist to further generations. This is because the number of linages lost will be a fraction of the linages from the previous generation. In time, the number of linages will decrease and so will the fraction of linages lost to the next generation.

Presented here are the outcomes for 10 000 individuals' linages with newly mutated alleles of the same type. Here and throughout the report it should be kept in mind that the number of iterations, in this case for 10 000 individuals, should not affect the frequency of e.g. persistence at a specific generation (provided there are sufficiently many to ensure a statistically significant output). On the other hand, the extreme value will be highly dependent of the number of iterations. The probability distribution of outcomes, that linages with different persistence are randomly "drawn" from, does not change depending on the number of outcomes drawn, but so does the probability of drawing a outcome with a very high persistence (or pervasiveness etc.).

Then value at each generation will be the relative frequency of linages that persisted to that specific generation, and not the cumulative frequency of linages still alive at that generation. Values of a relatively low selection disadvantage of 0.02 and a relatively high of 0.2 will be displayed. In addition for the x-linked case, an h value of 0.5 has been used.

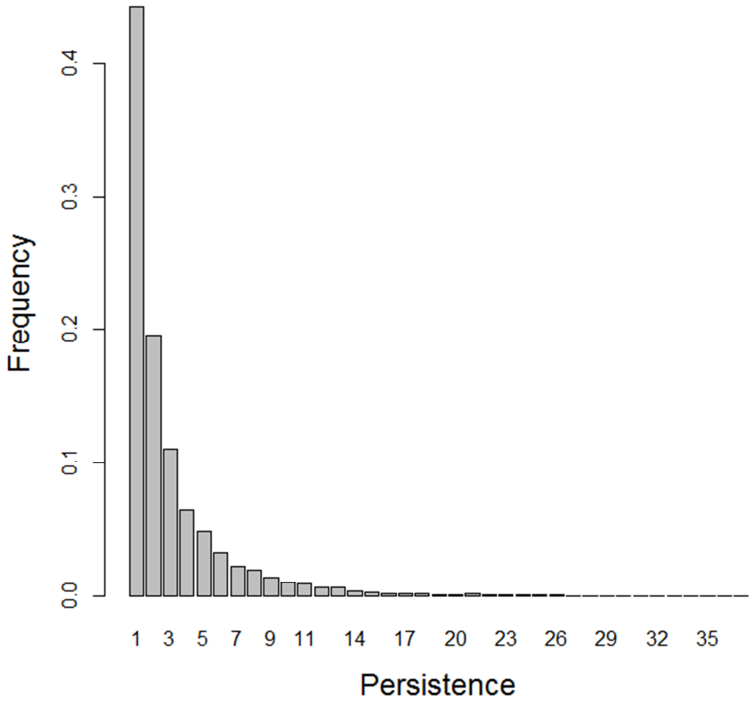

Figure 2a: Distribution of lineages with dominant mutation that persisted to a certain generation with a selection of disadvantage of *s=0.2*.

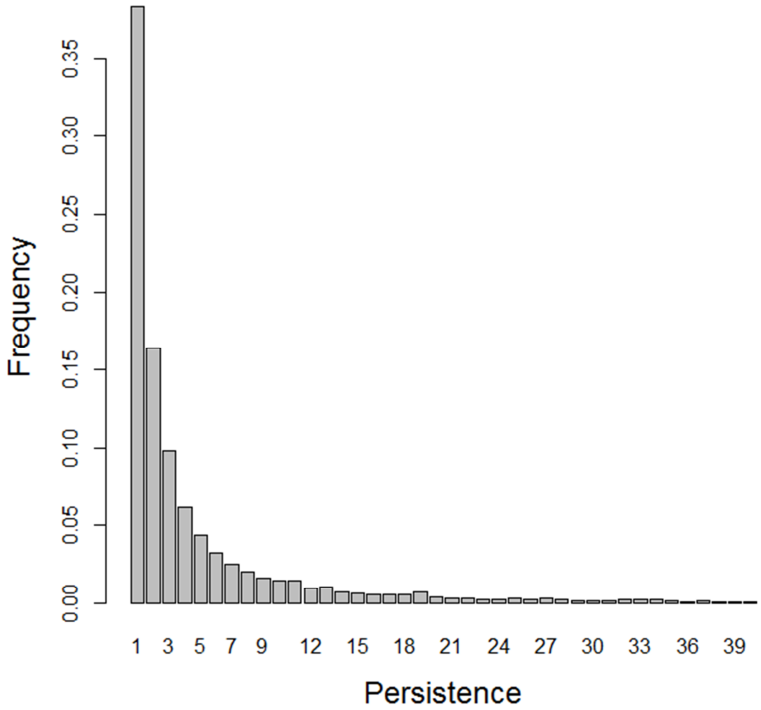

Figure 2b: Distribution of lineages with dominant mutation that persisted to a certain generation with a selection of disadvantage of s=*0.02*. *Note:* The x-axis has been cut, and the longest existing linage persisted for 327 generations.

The striking difference lies not in the relative frequencies of persistence for the first few generations. Instead, it is the presence of relatively high relative frequencies of persistence for the later generations for the low selection disadvantage.

Interestingly, but not surprisingly, for both the low and the high selective disadvantage, the number of mutations that only persisted for a single generation is about the same. Remember that the extreme values are very variable and therefore very dependent of the total number of iterations the program runs.

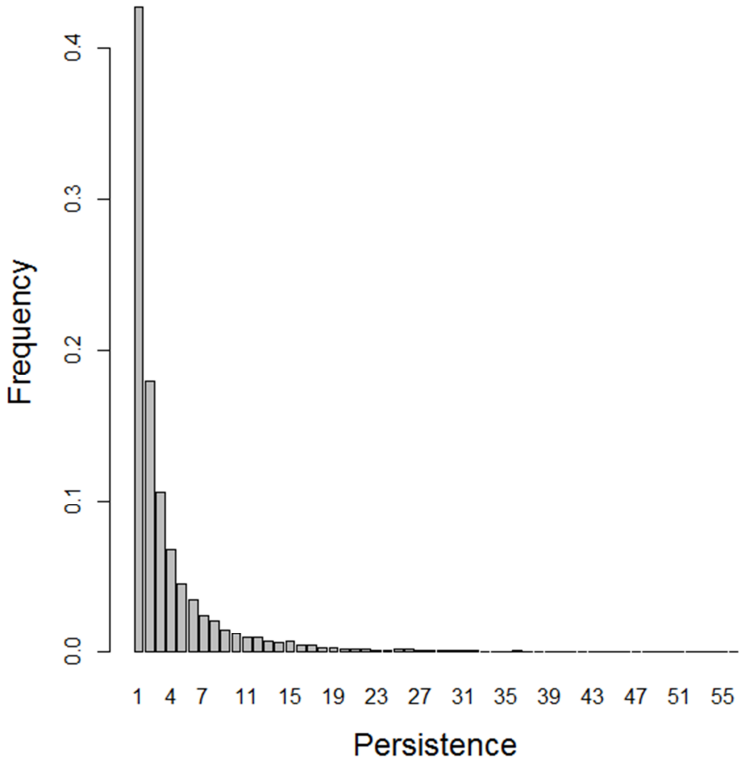

Figure 2c: Distribution of linages with x-linked mutation that persisted to a certain generation with a selection of disadvantage of *s=0.2* and *h=0.5*.

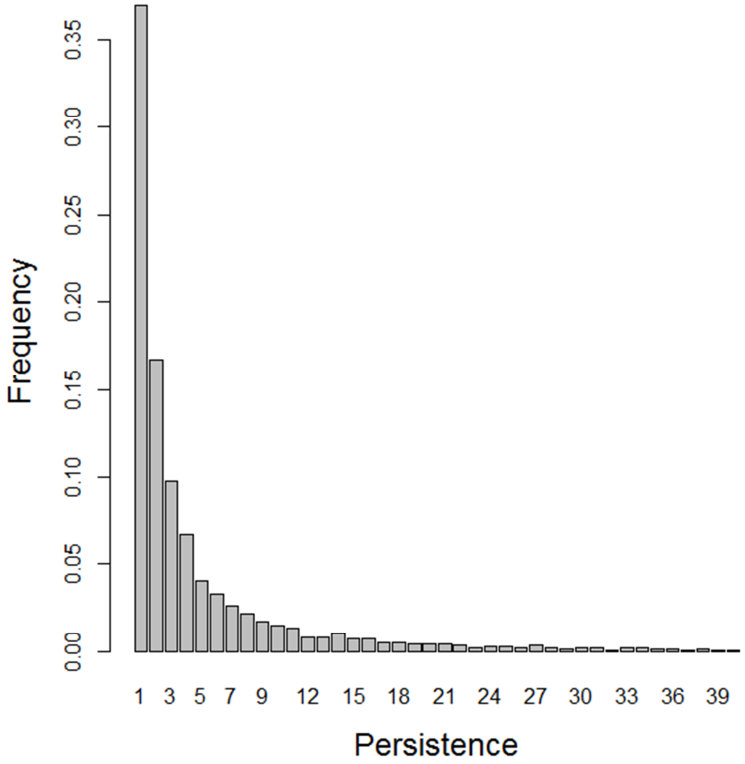

Figure 2d: Distribution of linages with x-linked mutation that persisted to a certain generation with a selection of disadvantage of *s=0.02* and *h=0.5*. *Note:* The x-axis has been cut, and the longest existing linage persisted for 455 generations.

For the x-linked case the scenario is similar, but since females in these examples are less affected by the negative selection (have higher fitness), the alleles generally persist longer in comparison with the same selection disadvantage in the dominant case.

For different degree of selection, there will be an expected value for the persistence, i.e. the mean persistence.

Table 1: Mean Persistence.

Persistence

Selection:	*Dominant*	*X-linked (h=0.0)*	*X-linked (h=0.5)*
0.01	8.48	10.28	9.26
0.02	6.95	9.04	7.93
0.03	6.28	8.38	6.91
0.04	5.69	7.56	6.29
0.06	4.98	6.80	5.59
0.08	4.30	6.29	5.11
0.10	4.12	5.76	4.68
0.20	2.91	4.51	3.52
0.30	2.40	3.96	2.94
0.40	1.99	3.35	2.52
0.50	1.74	2.98	2.21

Table 2: Variance of persistence

Variance of persistence

Selection:	*Dominant*	*X-linked (h=0.0)*	*X-linked (h=0.5)*
0.01	600.57	2215.48	943.27
0.02	236.61	907.27	409.76
0.03	160.24	668.70	278.99
0.04	113.48	398.23	162.73
0.06	68.27	251.28	110.04
0.08	38.20	183.17	74.27
0.10	33.46	129.49	52.84
0.20	9.73	46.53	19.01
0.30	5.05	29.21	10.50
0.40	2.61	16.16	6.06
0.50	1.53	10.81	3.92

There is not a big difference in persistence for the dominant case and the X-linked even when h=0 and the allele is completely recessive and carrying females are not affected at all. The variance of persistence differ more drastically, with much higher variance for the X-linked case, and even more so when h=0. This indicates that even when the mean persistence is similar, the distributions of persistence for the different alleles are very different for the two cases.

Pervasiveness

The total number of individuals carrying a new allele from the mutation event through the generations to the extinction of the allele in the population can be called the pervasiveness. This means that the unit of pervasiveness is number of individuals. If the persistence is one, the pervasiveness should also be one since a new allele always starts with a single copy in a single individual. Therefore, the relative frequency of pervasiveness for generation one should be the equal to the relative frequency of persistence for generation one. Also, the pervasiveness for a single allele must be at least as high as the persistence; there must be at least one individual in each generation. This means that the variance of pervasiveness will always be higher than the variance of persistence. For this very reason, when pervasiveness is simulated more iterations need to be performed to get an even distribution. In the results presented here, the number of iterations is 100 000 instead of the previously used 10 000.

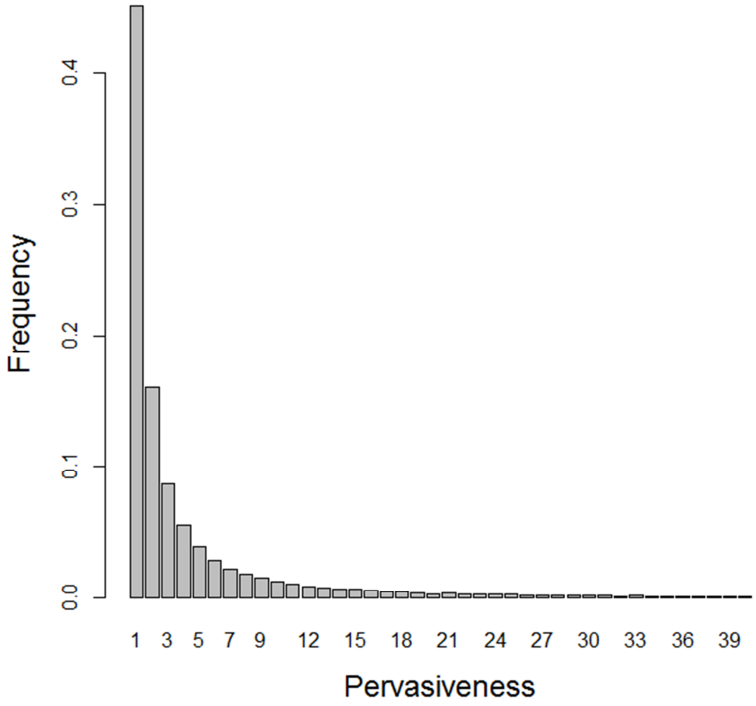

Figure 3a: Distribution of pervasiveness for dominant mutations with a selection disadvantage of *s=0.2*. *Note:* The x-axis has been cut, and the largest existing linage had a pervasiveness of 271 generations.

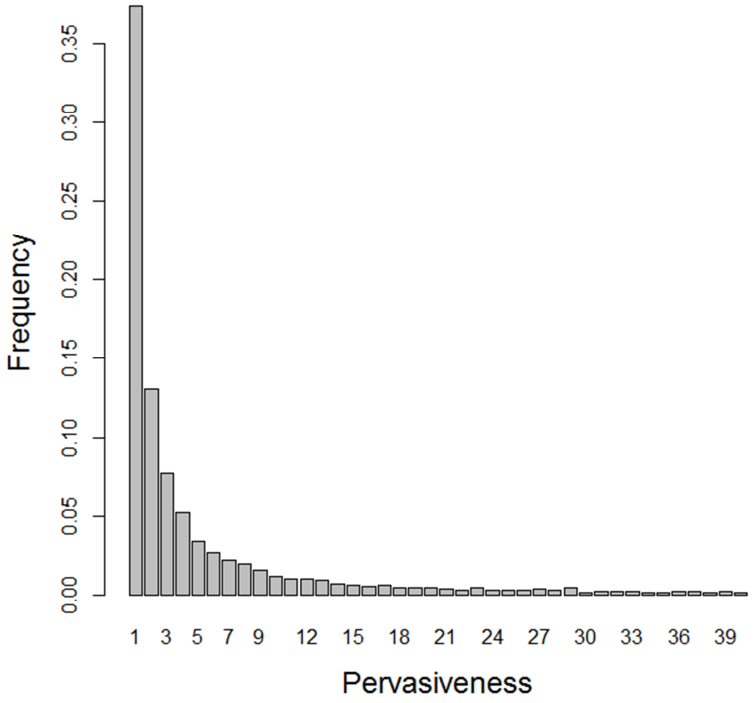

Figure 3b: Distribution of pervasiveness for a *dominant* mutations with a selection disadvantage *s=0.02*. *Note:* The x-axis has been cut, and the largest existing linage had a pervasiveness of 12 669 generations.

Most lineages only consisted of few individuals, but in some rare cases they could consist of many hundreds of individuals. As for the persistence, it is in the later generations that the main difference in relative frequency of pervasiveness lies.

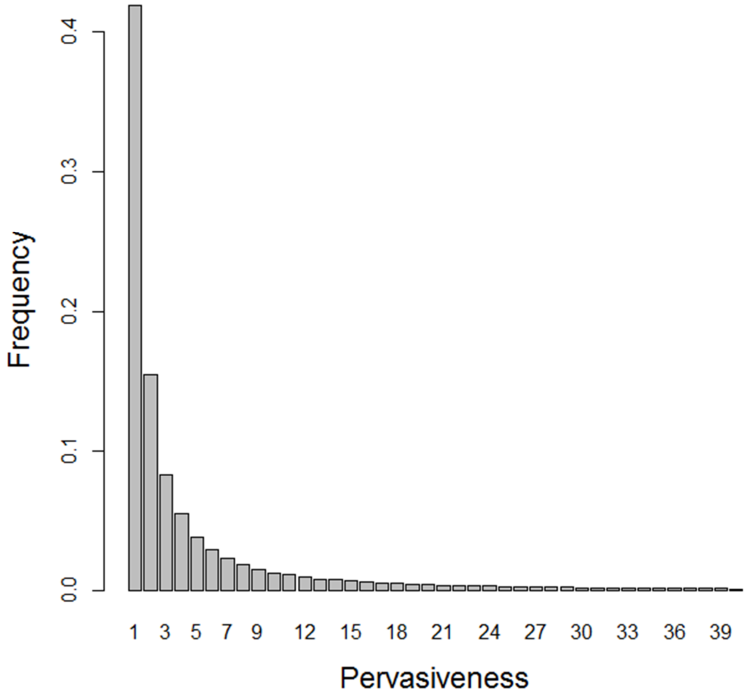

Figure 3c: Distribution of pervasiveness for *X-linked* mutations selection disadvantage of *s=0.2* and *h=0.5*. *Note:* The x-axis has been cut, and the largest existing linage had a pervasiveness of 582 individuals

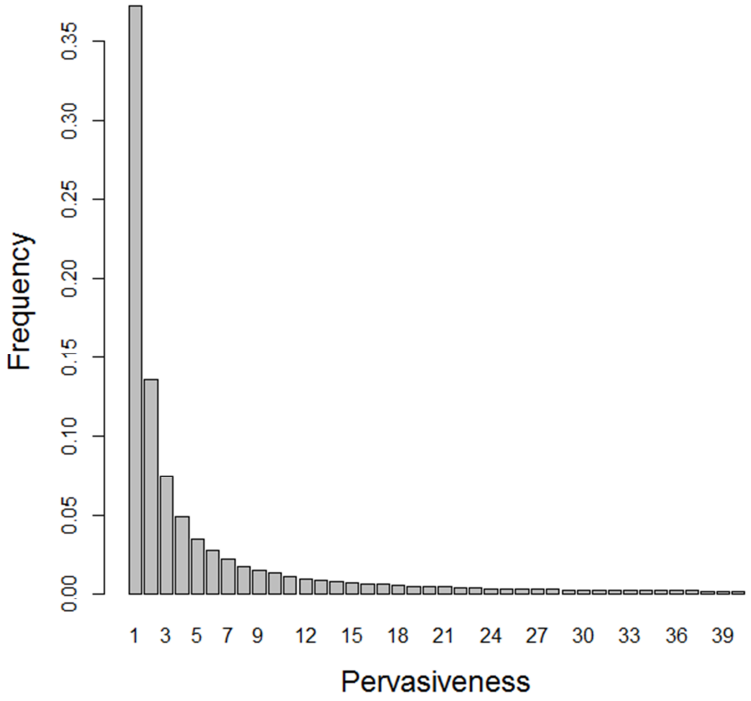

Figure 3d: Distribution of pervasiveness for *X-linked* mutations selection disadvantage of *s=0.02* and *h=0.5*. *Note:* The x-axis has been cut, and the largest existing linage had a pervasiveness of 22 372Individuals.

Again it is the length of the right tail of the distribution of relative frequencies that differ. Especially the extreme values are very far apart.

Table 3: Pervasiveness

Pervasiveness

Selection:	*Dominant*	*X-linked (h=0.0)*	*X-linked (h=0.5)*
0.01	110.29	322.89	181.22
0.02	50.69	160.44	76.63
0.03	35.09	110.76	53.93
0.04	26.17	72.66	35.82
0.06	17.62	50.36	25.65
0.08	12.28	37.88	19.03
0.10	10.56	28.80	14.70
0.20	4.87	13.79	7.36
0.30	3.42	9.76	5.03
0.40	2.46	6.71	3.72
0.50	2.00	5.25	2.95

Table 4: *Variance of* pervasiveness

Variance of Pervasiveness

Selection:	*Dominant*	*X-linked (h=0.0)*	*X-linked (h=0.5)*
0.01	1119107.00	22670817.00	3236206.00
0.02	119814.50	4052346.00	391489.70
0.03	32706.68	902410.70	135181.60
0.04	15309.92	401868.60	45501.76
0.06	3922.99	131180.30	13662.87
0.08	1742.87	46362.27	6976.90
0.10	885.21	23907.79	3138.31
0.20	100.02	2847.79	360.12
0.30	26.64	698.96	93.28
0.40	9.25	272.60	38.28
0.50	3.92	125.99	16.99

For high selection disadvantages, the difference in pervasiveness is not great, but for lower s-values, the pervasiveness becomes much higher for the lower s and h-values. This pattern is even clearer, with very high variance for the lower selection disadvantages.

Individuals per generation

Persistence and pervasiveness only describe the end result of mutations spread in a population. Here is shown the relative frequency of individuals living after a certain number of generations from when the mutation first occurred.

In other words, an individual with a mutated allele is picked at random, what is the probability that that mutation have existed for a given number of generations? Looking at all those relative frequencies at the same time represents the steady state age distribution that a certain type of mutated allele would have in a population. This gives an idea of how probable it is for a mutation to exist in a certain generation. Represented in this way, the relative frequency of number of individuals living in a certain generation may or may not have offspring living in later generations; and if so contributing to the relative frequencies in later generations. In the small example in figure 1, the relative frequencies would be 0.25 0.50 and 0.25 for the first, second and third generation. In the results presented here, the number of iterations is 100 000.

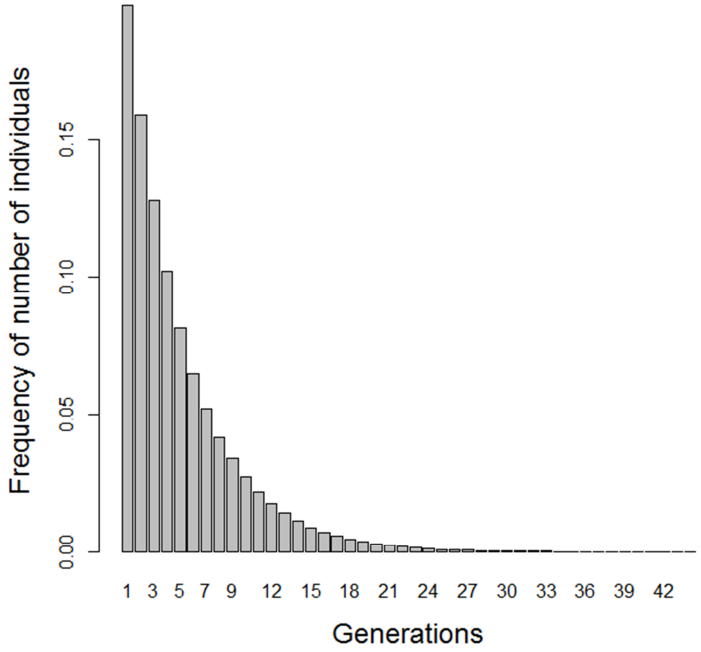

Figure 4a: Distribution of number of individuals (per generation) after a mutation occurred for a *dominant* mutation with a selection disadvantage of *s=0.2*.

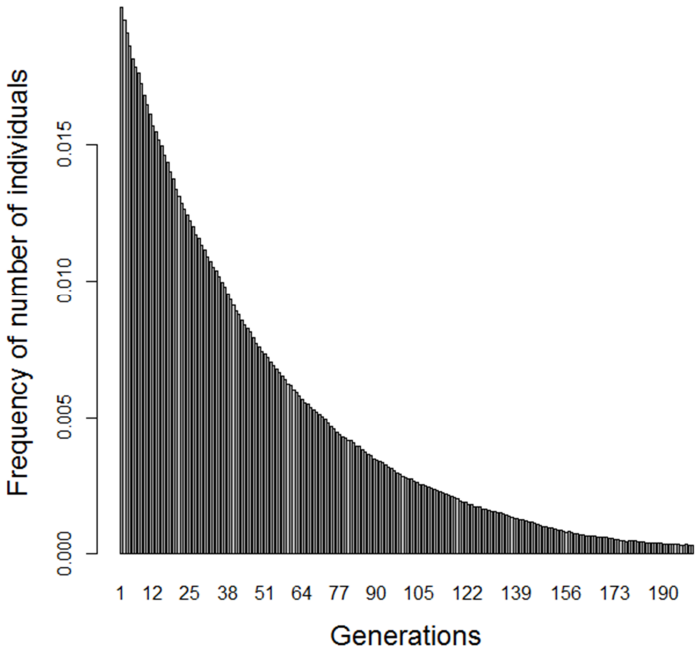

Figure 4b: Distribution of number of individuals (per generation) after a mutation occurred for a *dominant* mutation with a selection disadvantage of*s=0.02*. The x-axis has been cut, and the largest existing linage had a persistence of 458 generations.

Given the low relative frequency of each bar, the probability for an individual with a mutation to exist in a given generation after first occurrence is usually very low. The relative frequencies become more informative if compared with relative frequencies for persistence. For *s*=0.02 for example, there are very few lineages that persisted for more than 30 generations. During the same conditions, the relative frequency of individuals living in that generation is still relatively high. The decrease of individuals is not as drastic as the decrease in lineages. This means that in the surviving lineages, there are a considerable amount of individuals.

For the x-linked case, the relative frequencies are subdivided in relative frequencies of females and males. The total relative frequencies are therefore divided in females and males and their sums are represented by the total heights of the bars.

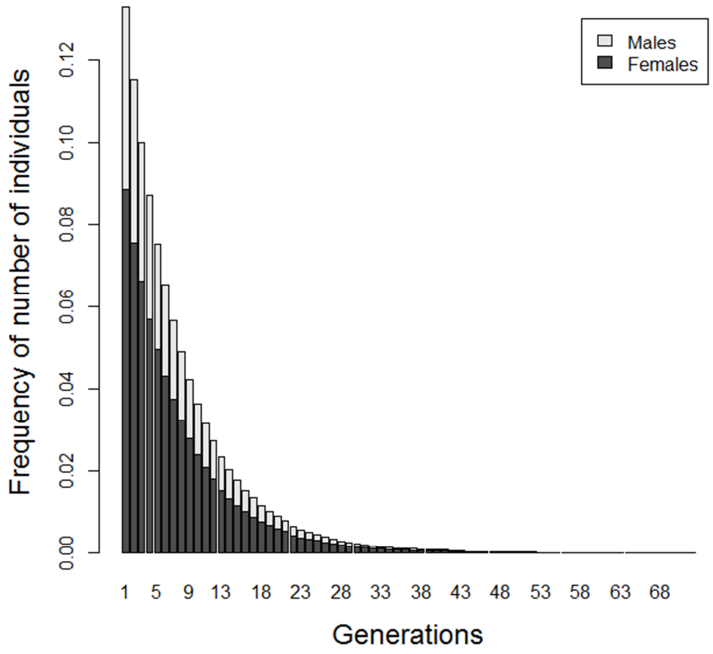

Figure 4c: Distribution of number of individuals (per generation) after a mutation occurred for an *X-linked* mutation with a selection disadvantage of *s=0.2* and *h=0.5*.

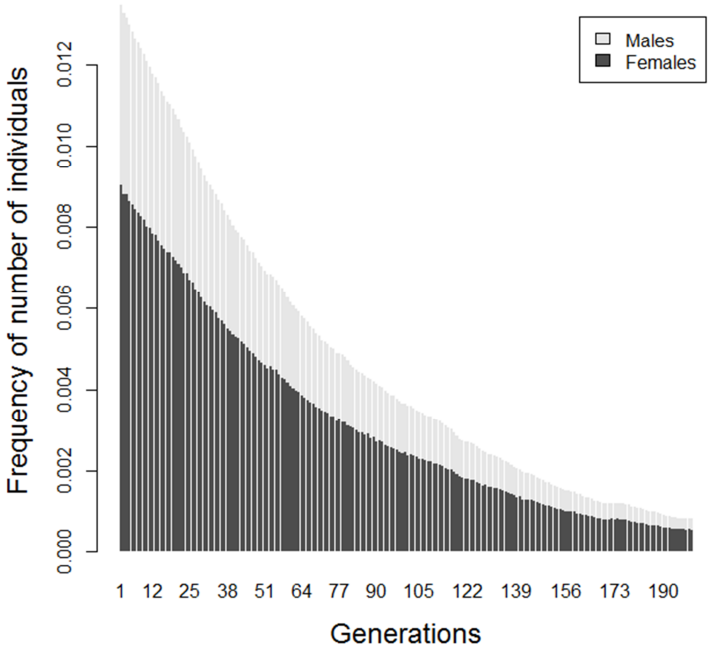

Figure 4d: Distribution of number of individuals (per generation) after a mutation occurred for an *X-linked* mutation with a selection disadvantage of *s=0.02* and *h=0.5*. The x-axis has been cut, and the largest existing linage had a persistence of 518 generations.

In the X-linked scenario there is additional information about the gender ratio. About two thirds in every generation are female. Apart from that, the general pattern of high relative frequencies of individuals in later generations is seen again. This is because males carrying the mutated allele only have female carriers in their offspring, but females have both males and females. Should one gender be randomly overrepresented in one generation, the ratio will go back in the next.

All this taken together, is possible to calculate the *expected age* of an allele. That is, if an allele from a population of mutated alleles, what is the best guess of the age in generations of that allele? This type of question still assumes that the age distribution of alleles does not change (much) from generation to generation in a continuous world. This can be calculated using the weighted arithmetic mean, where the expected value will be the expected age. Let G_1, G_2, G_3…G_n be the relative frequencies of the number of individuals with the mutated allele in generation 1, 2… et cetera. The first individual in a generation is from generation one, its offspring from two, et cetera.

$$\textbf{Expected age} = \sum_{i=1}^{n} G_i * i$$

Using frequencies from figure one as an example:

$$\text{Expected age} = (0.25 * 1) + (0.5 * 2) + (0.25 * 3) = 2$$

Table 5: Expected age of mutant alleles.

Expected age

Selection:	*Dominant*	*X-linked (h=0.0)*	*X-linked (h=0.5)*
0,01	98,23	285,80	146,26
0,02	50,11	147,87	74,91
0,03	35,24	103,42	50,50
0,04	24,71	74,91	37,79
0,06	16,51	52,04	24,39
0,08	12,60	34,57	19,03
0,10	10,03	29,58	14,84
0,20	4,99	14,17	7,50
0,30	3,33	9,20	4,96
0,40	2,51	6,82	3,68
0,50	2,01	5,16	2,94

The difference in expected generally big, especially for the lower selection disadvantages.

Individuals in lineages per generation

As mentioned, there seems to be a considerable amount of individuals with mutation living in later generation, even when most lineages are gone. The question is then, amongst the lineages that are not lost, how many individuals are there on average in each of those lineages? In the results presented here, the number of iterations is 1 000 000.

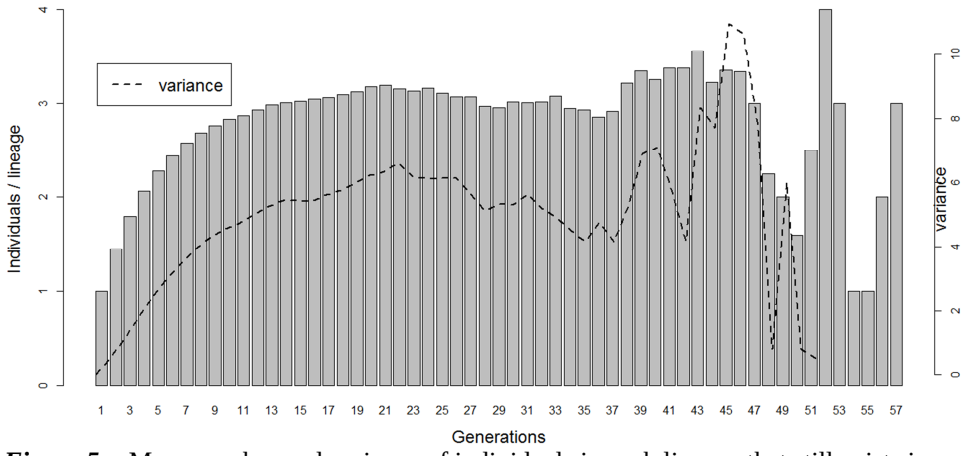

Figure 5a: Mean number and variance of individuals in each lineage that still exists in a generation. For dominant mutation with a selection disadvantage of *s=0.2.*

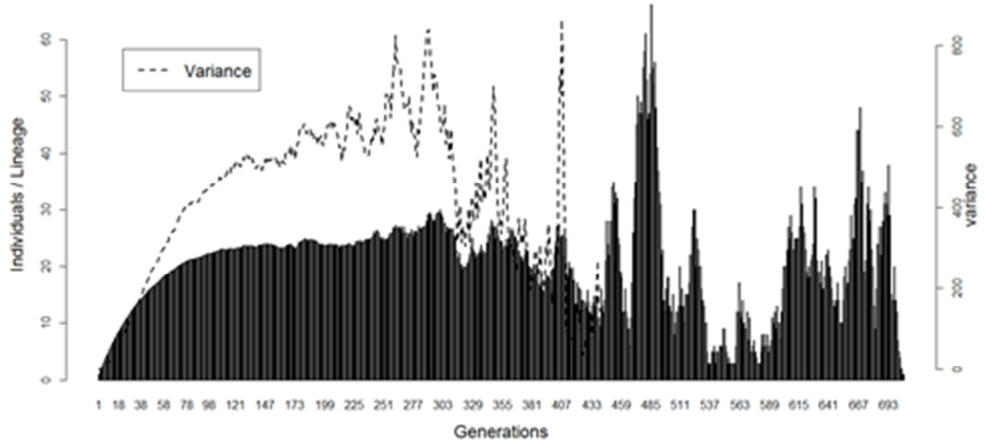

Figure 5b: Mean number and variance of individuals in each lineage that still exists in a generation. For a dominant mutation with selection disadvantage of *s=0.02*.

In the very first generation all lineages consist of a single individual and the mean will of course be 1 and the variance 0. In generation number 2, some lineages are already lost, but those that survived will consist of one person *or more*, and that is why the mean increases, even though the total number of individuals decreases. The mean continues to increase for later generations until what seems like a plateau is reached. At this point, there is a mixture of lineages on the verge of extinction with only a few (or one) individuals, and lineages on a temporary rise, with perhaps hundreds of individuals. The mean will then become unstable and fluctuate because there are too few lineages left to accurately represent the mean. In fact, already by generation 355 in the run with a dominant mutation and *s*=0.02 (***figure 5b***) there are only 214 of the original one million lineages still not lost. In the very last generations there will be only one lineage in existence and therefore there will be no variance. Also, there is a big difference in variance for the high and the low selection, seen by looking at the rightmost axis for variance in the different graphs.

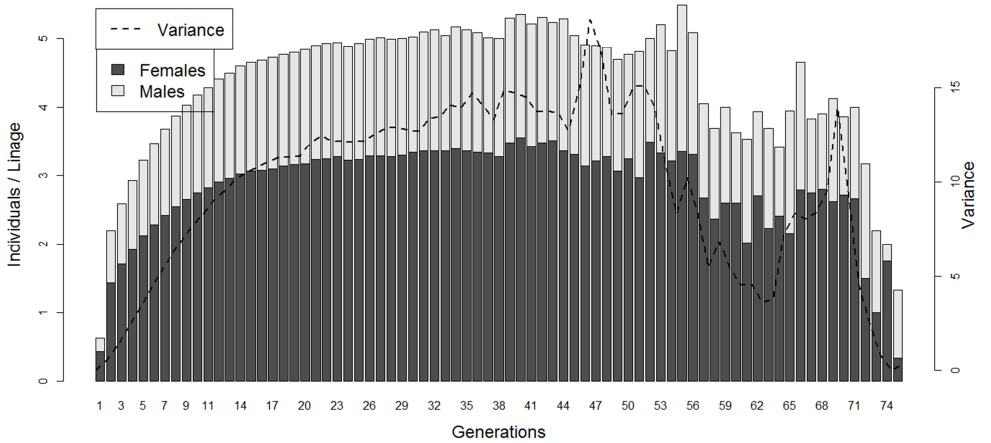

Figure 5c: Mean number and variance of individuals in each lineage that still exists in a generation. For an x-linked mutation with a selection disadvantage of *s=0.2* and *h=0.5*.

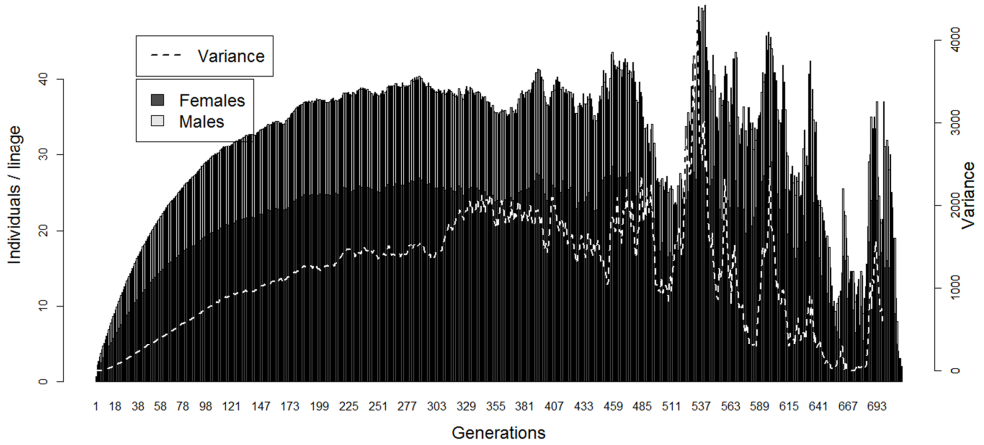

Figure 5d: Mean number and variance of individuals in each lineage that still exists in a generation. For an x-linked mutation with a selection disadvantage of *s=0.02* and *h=0.5*.

The x-linked scenario shows similar pattern.

Since the mean individuals per lineage and generation quickly become unstable, it is not entirely clear that there is indeed a plateau phase at all. It is possible that there would be a decline or increase of the values if there were enough lines left to contribute to a stable mean. In order to study the suspected plateau, the input of the program was changed to mimic a population which had already reached the plateau. Remember, in the original setup there were only one individual per lineage in the first generation. Instead, a population of lineages with a number of individuals Poisson distributed with the plateau value as the expected value was used as input.

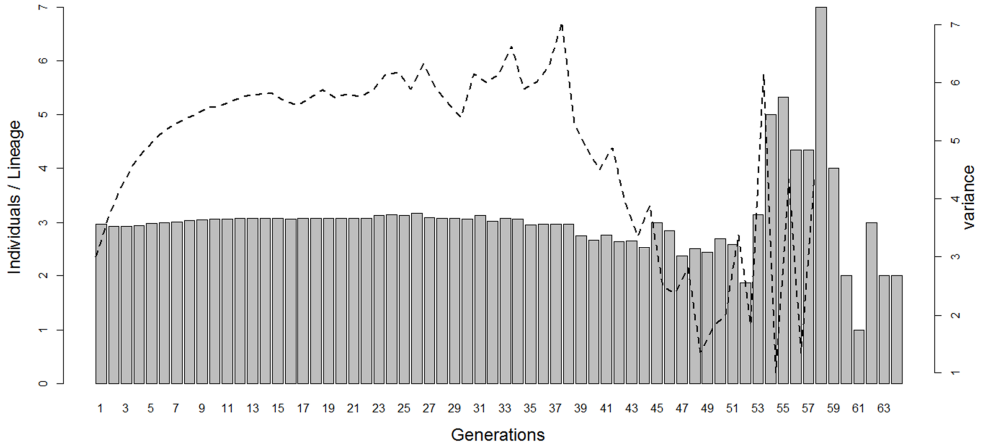

Figure 5e: Mean number and variance of individuals in each lineage that still exists in a generation. For an initially (*X~Po [plateau mean≈3]*1 000 000*) individuals with a *dominant* mutation with a selection disadvantage of *s=0.2*.

Now, the suspected plateau is stable for more generations before the mean starts to fluctuate. The variance initially increases and then stabilizes; indicating that the distribution has changed even though its mean has not.

At the plateau the mean of individuals per linage is at a steady state. Also the distribution of individuals will be steady at the plateau.

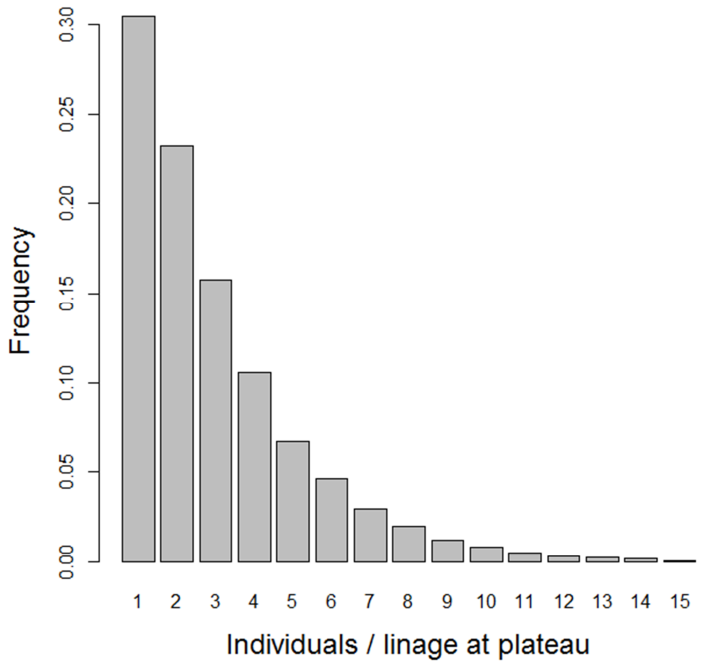

Figure 5f. Steady state distribution of individuals per linage at plateau (see *figure 5e*) with a selection disadvantage of *s=0.2*. Note: The x-axis has been cut, and the highest number individuals in any generation in any linage at the plateau were 29.

The distribution of individuals at plateau for s=0.02 is much skewed with almost a third of the linages with only a single individual.

The distribution of individuals at the plateau in these simulations is only dependent of the selection coefficient s.

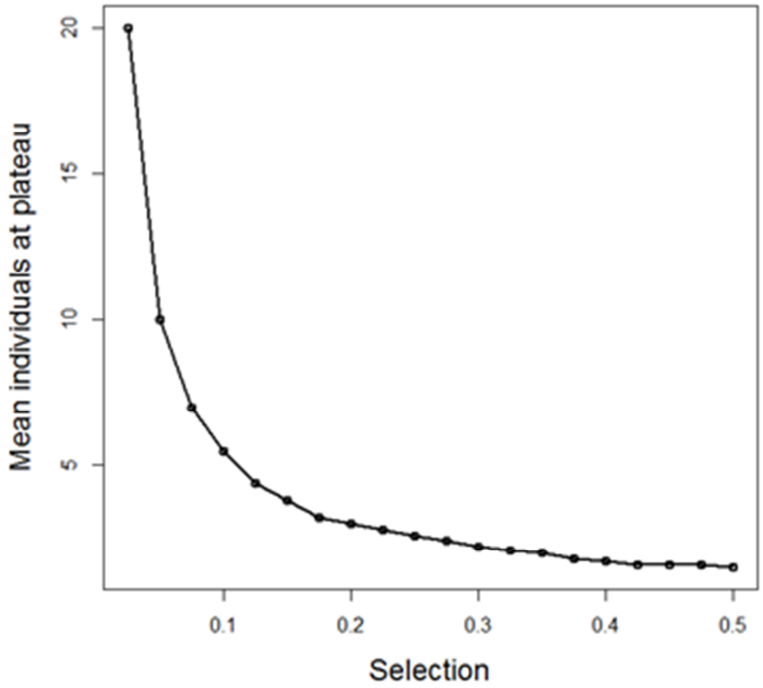

Figure 5g. Mean individuals at plateau for different selection coefficients.

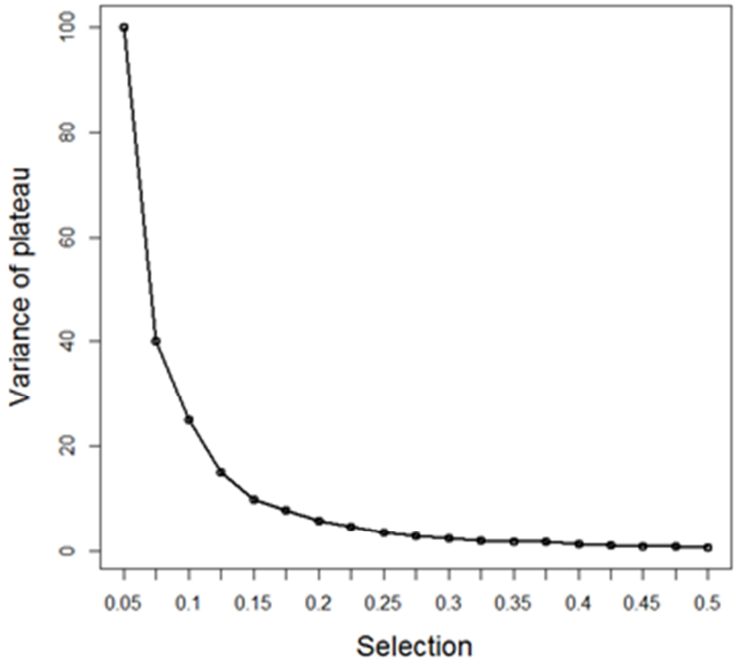

Figure 5h. Variance of individuals at different plateaus for different selection coefficients.

As the s values decreases, the plateau mean increases rapidly and approaches infinity as s approaches zero. At very high selection on the other hand, the mean approaches one. The mean can never be lower than one since there have to be at least one individual in a linage unless it will be lost. The variance does not have a simple relation with the selection coefficient

and neither has the standard deviation which relation have a similar appearance (not shown). But the multiplicative inverse of the standard deviation (1/SD) has a simple linear relation with the selection coefficient.

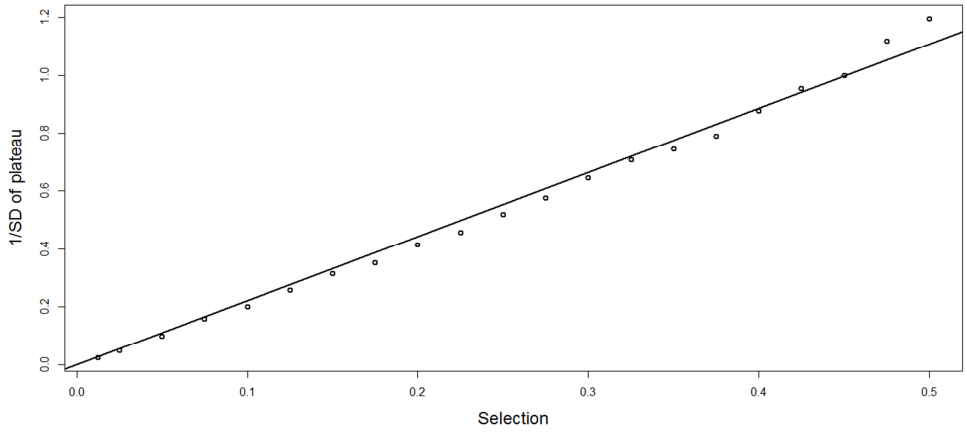

Figure 5j.1/SD for different selection coefficients with fitted line. Correlation coefficient= 0.997, p-value: < 2.2*10^-16, slope 2.213, forced through (0,0).

The second algorithm: Whole population equilibrium

In the first algorithm, the iteration loop could be thought of as parallel universes, each with its own outcome. In this second algorithm, there is only one universe and we begin with zero individuals carrying the mutation. For each generation a *Poisson distributed* number of new mutations occur and those individuals with the new mutation is added to the population. Also, for each new generation the individuals from the previous generation produce offspring in the same manner as in the first algorithm. This is of course more similar to a real world situation. In this algorithm, each lineage is treated as separate and therefore, number of individuals, identity and *expected age* can be calculated at any given moment in the simulation.

This scenario will result in the *mutation-selection equilibrium* mention earlier, where the frequency of individuals with mutation will fluctuate around a mean. This equilibrium is described by the well know population genetics formula:

$$f = \mu/s$$

Were μ is the mutation rate, s is the selection coefficient and f is the frequency of the mutant allele in a haploid population. For a rare dominant allele, this will approximately be the same as the frequency of individuals carrying the mutation in a diploid population. For convenience, we don't want to simulate an entire population, just the individuals carrying the mutant alleles. At any point the population of individuals with the mutant

allele will have a specific number. In the world of the program, there are no non carriers. To do this we have to convert the frequency of carriers into an actual number, called *number of individuals*. This frequency is of course just the total number of individuals carrying the mutation (here called *X*) divided by the total number of alleles in the population (*N*). Then: $X / N = \mu / s$. And if *N* shuffled to the right side:

$$\mathbf{X = N\mu / s.}$$

Now can we treat $N\mu$ as a single value since it is the product of two constants. This value will be the number of new mutations in the population in each generation, i.e. the lambda for the *Poisson distribution* used in the simulation. Accordingly, when the equilibrium is reached, the number of individuals with mutation in the simulation should fluctuate around $N\mu/s$.

To get a concept about the diversity of the origins of the alleles the *identity* is used. This is the probability that two individuals among those with the mutant allele have IBD alleles; the higher the identity, the lower the diversity.
Let L_1, L_2, $L_3 \ldots L_k$ be the relative frequency of the number of individuals in arbitrarily numbered lineages with the mutant allele. The probability for individuals in the population to share IBD alleles here referred to as identity, can be calculated as:

$$\mathbf{Identity} = \sum_{i=1}^{n} L_i^2$$

The expected age will be calculated as before. The main difference with this algorithm is that there will be a time axis. Number of individuals with the mutant allele and the expected age of that type of allele will change from generation to generation.

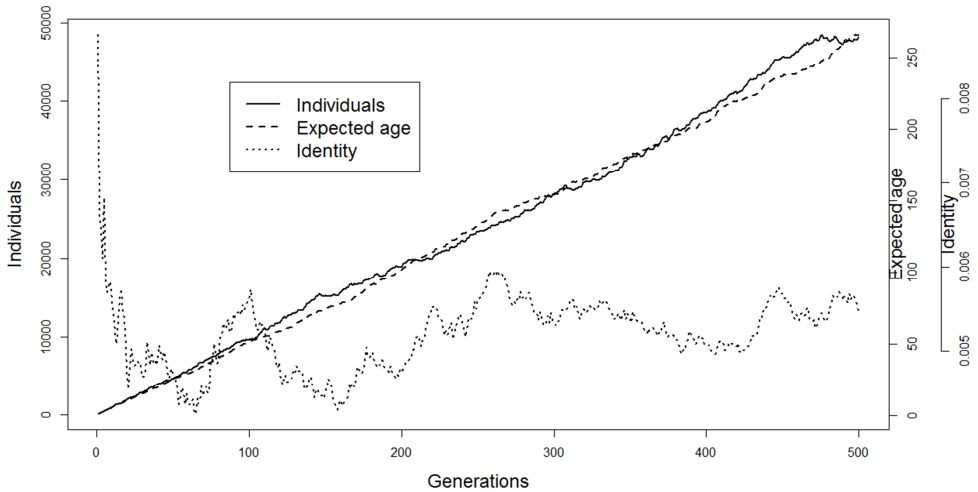

Figure 6a: Number of individuals with *no* mutation and their expected age and the identity by decent for the mutation, with *s=0.0* and with $N\mu = 100$ per generation.

To get a context of how this model works an introductory example with no selection is shown. This example can be thought of as an immigration of a new species to a new island. Each generation about a hundred new individuals immigrate to the island, while the population already immigrated on the island reproduces with a fitness of one. This scenario is not likely in the real world were the fitness usually are dependent of how close the population number are to the carrying capacity (Pianka 1970). If there is *no* selection, the number of individuals with mutant alleles will steadily increase as new individuals with mutant alleles are introduced in each generation. The identity for the alleles will start at $1/N\mu$ in the very first generation and then decrease. Because of genetic drift, the identity will not decrease indefinitely. Each generation the number of lost lineages increases (decreases identity) until the number approaches $N\mu$. The average number of new alleles will then be similar to the average number of lost alleles in each generation. In what first appears as a paradox, the *expected age* of an allele continues to increase even when the identity is momentarily increasing.

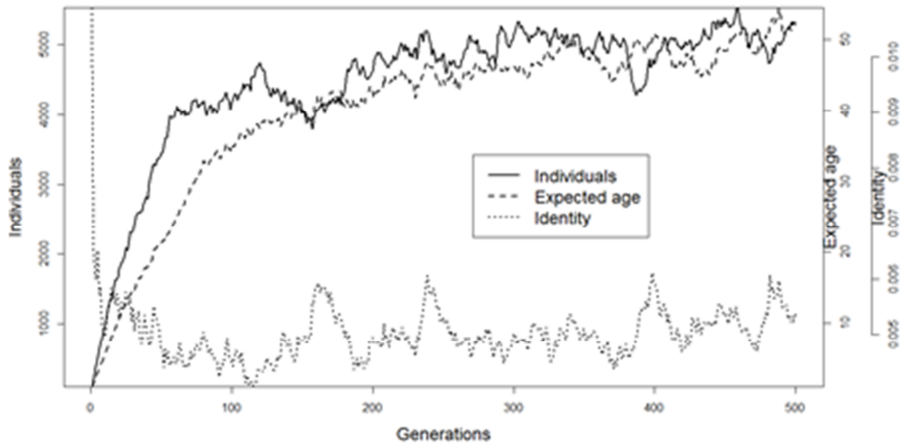

Figure 6b: Number of individuals with a *dominant mutation* and their expected age and the identity by decent for the mutation, with $s=0.02$ and with $N\mu=100$ per generation.

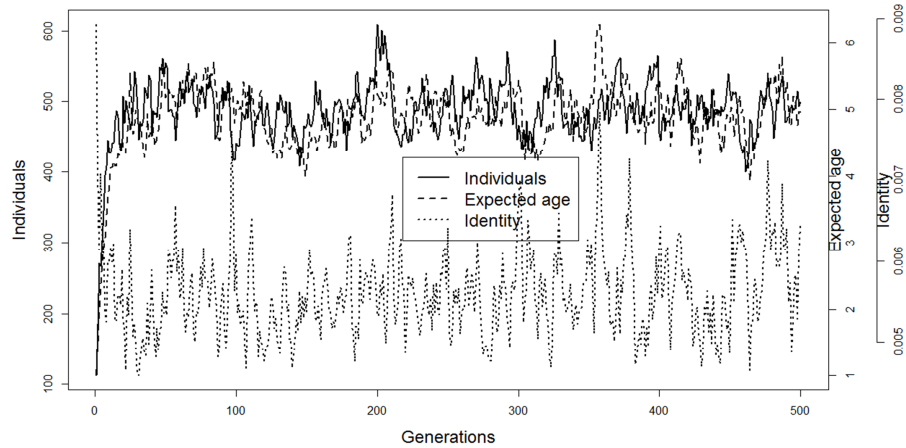

Figure 6c:Number of individuals with a *dominant mutation* and their expected age and the identity by decent for the mutation, with $s=0.2$ and with $N\mu=100$ per generation.

When there *is* selection, the total number of individuals will increase until the number reaches the *mutation-selection equilibrium*. At the equilibrium the number randomly fluctuates for the rest of the simulation. So does the *identity* as well as the *expected age* around their own equilibriums. Since all alleles behave independently of each other, the *expected age* at equilibrium will always be the same for certain selection strengths. If only the $N\mu$ is changed, but not the s, the expected age at equilibrium will be the same.

From this it is clear that there is some kind of dependence of these three factors. This is at least true for some extreme examples. Imagine a large population were only a single individual is carrying a certain mutated allele. The *identity* for that mutation would be 1, and the *expected age* would probably be very low. On the other hand, if all individuals in the same population carried the allele, the *identity* would be 1 and the *expected age* very high. The dependence could also be time lagged, were fore example the number of individuals compared to the expected age some generations later could be higher than a comparison for present numbers.

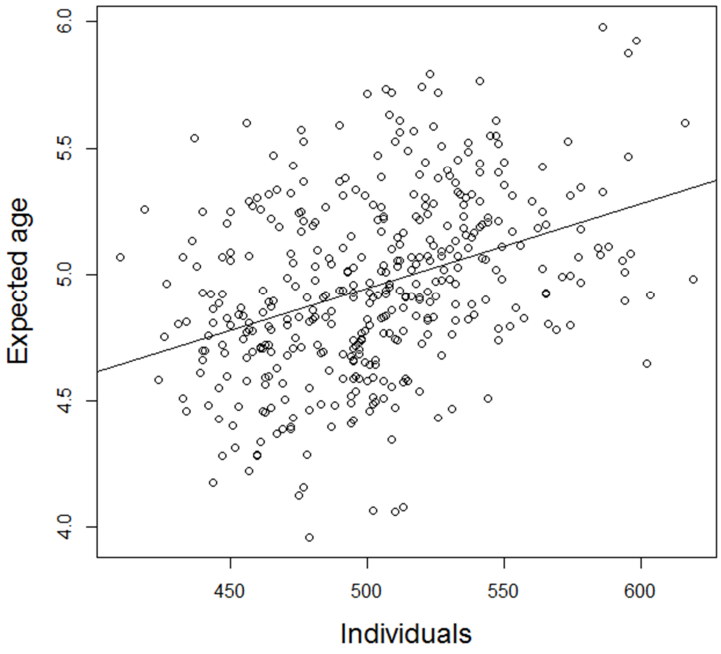

Figure 7a: Expected age, and number of individuals plotted against each other at 400 generations at mutant-selection equilibrium, with a dominant mutation and s=0.2 Nμ=100 per generation. Line fitted with linear regression has correlation coefficient 0.37.

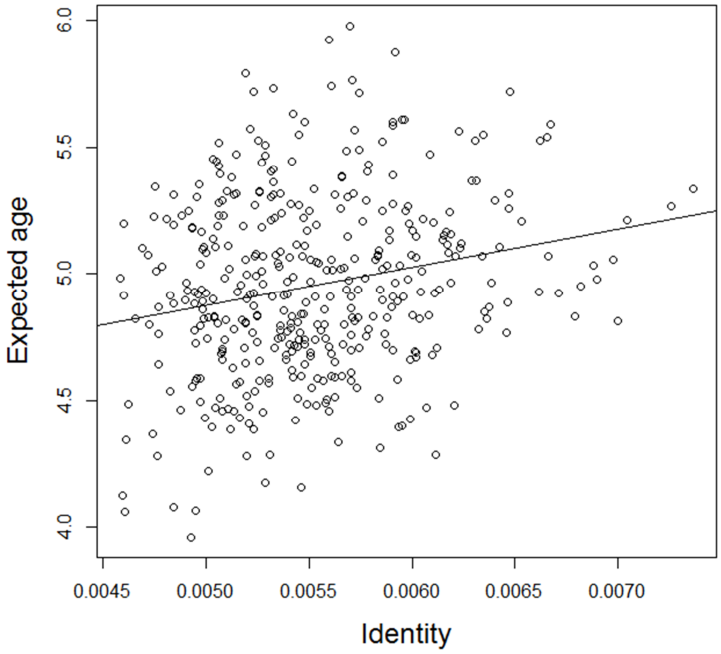

Figure 7a: Expected age, and identity plotted against each other at 400 generations at mutant-selection equilibrium, with a *dominant mutation* and $s=0.2$ $N\mu=100$ per generation. Line fitted with linear regression has correlation coefficient 0.21

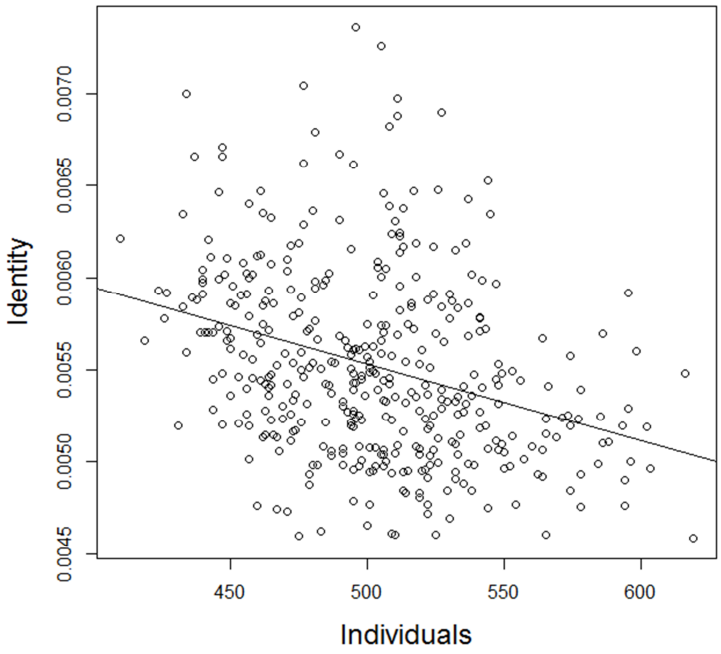

Figure 7a: Identity and number of individuals plotted against each other at 400 generations at mutant-selection equilibrium, with a *dominant mutation* and $s=0.2$ $N\mu=100$ per generation. Line fitted with linear regression has correlation coefficient -0.32.

Interestingly, compared two by two, there are surprisingly low correlations between these factors. But as expected, there is a positive, though small, correlation between *expected age* and the *number of individuals* as well as between *expected age* and *identities*, and a negative correlation between *identities* and the *number of individuals*. The plots are comparing traits generation for generation. If the time axis were lagged for up to 10 generations for any one of the traits, the correlations became poorer. This indicates that there is no lagged dependence for any of the traits.

A Bayesian approach

As mention, the identity (*I*) is the a priori probability of two individuals with a mutation of the same type to share an IBD allele. This probability changes if the underlying haplotype of the genomic area around the mutation is the same for both of the individuals. They cannot be IBD if the haplotype is different. So if the haplotypes are the same, the probability for IBD increases. It this case, the conditional probability that two individuals share an IBD allele if the haplotype is known to be the same, can be calculated using *Bayes theorem*.

Using just generic events *A* and *B*, *Bayes theorem* has a simple appearance:

$$\boldsymbol{P(A|B) = \frac{P(B|A)P(A)}{P(B)}}$$

It can be derived using the axiom that the probability for event A and B to happen together P(A and B), can be thought of an event of its own. The axiom states that $(A\ and\ B) = P(B)P(A|B)$. P(A|B) means the probability of A if it is known that B is true, or A *given* B. It's important to realize that even though that $P(A\ and\ B)$ and $P(A|B)$ represents the same event; their probabilities are different since they exist in different probability space. A simple example: we want to know the probability that a fruit that we pick at random in a fruit basket is a green banana. We cannot simply multiply the probability of a fruit being green with the probability of a fruit being a banana, the probabilities are obviously dependent. But if we know that the probability of randomly picking a banana in the basket P(banana), and we now the probability of a banana to be green P(green | banana). Then we know that the probability of picking a fruit that is green *and* is a banana is the product of the fraction of all fruits in the basket that are bananas and the fraction of all bananas that are green. P(banana) multiplied with P(green| banana) will be the fraction of all bananas that are green bananas, or P(green and banana) =P(banana) * P(green | banana) . If there are eight fruits in the basket and three of them are bananas and the probability of banana to be green is one third, then P(banana) * P(green | banana) = $\frac{3}{8} * \frac{1}{3} = \frac{1}{8}$, and that is the probability of picking a green banana among all the eight fruits.

This axiom can is also true for the reverse; P(A and B) = P(A)P(B|A). Since both are equal to P(A and B), then $P(B)P(A|B) = \mathrm{P(A)P(B|A)}$. Moving P(B) to the other side and we have Bayes theorem: $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$
The law of *total probability* allows the expansion the denominator; the probability for *B* is the sum of the probabilities when *B* happens in conjunction with other events, such as the event of *A*, or the event of *not A*, or *A*¬, therefore: $\mathrm{P(A|B)} = \frac{P(\mathrm{B|A})P(\mathrm{A})}{P(\mathrm{B|A})P(\mathrm{A}) + P(\mathrm{B}|\neg\mathrm{A})P(\neg\mathrm{A})}$. The theorem is used to calculate the *conditional probability* of event *A* under the condition that *B* is true. In this study the following notations will be used:

´2' = two alleles in one group, i.e. two alleles that are IBD.
I = identity
'1:1' = two alleles in two groups, i.e. two alleles that have a recurrent mutation
H = the specific underlying haplotype
*f(*H*)* = the frequency of H

Of interest is *P('2'|* H*)* which is the conditional probability of two alleles being IBD if it is known that they share the haplotype. *P('2')* is the non-conditional probability for the identity. Using the *Bayes theorem*, this will be calculated as following: $\mathrm{P('2'|H)} = \frac{P(\mathrm{H}|\prime 2\prime)P(\prime 2\prime)}{P(\mathrm{H}|\prime 2\prime)P(\prime 2\prime) + P(\mathrm{H}|\neg\prime 2\prime)P(\neg\prime 2\prime)}$.

The event $P(\mathrm{H}|\neg'2')$is in this case the same as $P(\mathrm{H}|1:1')$ since the event *'1:1'* is the complement to the event *'2'*. $P(\neg'2')$can then be replaced with *P('1:1')* and *P(*H | $\neg'2'$*)* with *P(*H |'1:1'*)* in the formula: $\mathrm{P('2'|H)} = \frac{P(\mathrm{H}|\prime 2\prime)P(\prime 2\prime)}{P(\mathrm{H}|\prime 2\prime)P(\prime 2\prime) + (\mathrm{H}|\ \prime 1{:}1\prime)\ P(\prime 1{:}1\prime)}$For the event $P(\mathrm{H}|'2')$to happen, the specific haplotype have only mutated once. If a mutation event takes place, the probability for that mutation to hit a specific haplotype is here assumed to be equal to the frequency of that haplotype. Therefore, if two mutated alleles are considered, the probability for those to be IBD is equal to *f(*H*).* Remembers that *(*¬H| *'2')* is of course not possible; two IBD alleles have to share haplotype. For this event *(*H|'1:1'*)* to take place, the haplotype must have mutated twice and this probability would then be *f(*H*)*^{2}. The probability for the event *P('1:1')* is one minus the complement *P('2'),* i.e. the total probability minus identity *(1-I).* All this taken together: $\mathrm{P('2'|H)} = \frac{f(\mathrm{H})*\mathrm{I}}{f(\mathrm{H})*\mathrm{I}+\mathrm{f(H)}^2*(1-\mathrm{I})}$.

Or simpler:

$$\mathbf{P('2'|H_i)} = \frac{\mathbf{I}}{\mathbf{I+f(H)*(1-I)}}.$$

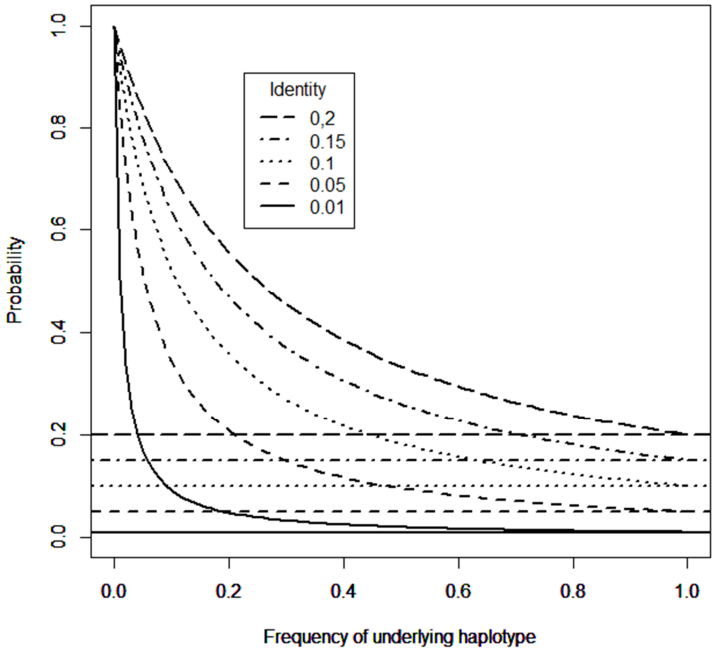

Figure 8: The probabilities for different IBD relationships for two individuals with mutant alleles of the same type, assuming the same underlying haplotype. *Identities used are examples; they are not from generated data*. The horizontal lines are the unconditional probabilities when the frequency of the underlying haplotype is one.

The frequency of the underlying haplotype becomes increasingly more informative when it is rare, especially for higher identities. For higher identities, the haplotype must be exceedingly rare to add change to the probability to any but low extent. As can be seen in previous results (***figures 6b-c***), the identities can be much lower than these examples.

Bayes theorem can also be used when more than two events are complimentary to each other, and the sum of the multiple probabilities equals to one. This can be used when multiple individuals have the same type of mutated allele, but the IBD relationship of their mutated alleles is not known.

If the generic event *A* in the original formula was a sum of multiple events $A_1, A_2, A_3 \ldots A_k$, *Bayes theorem* can be written as: $P(\mathrm{A_i}|B) = \frac{P(B|\mathrm{A_i})P(\mathrm{A_i})}{P(B)}$
And again using the law of *total probability* we can expand *P(B)* as the sum of all probabilities included:

$$\boldsymbol{P(A_i|B) = \frac{P(B|\mathrm{A_i})P(\mathrm{A_i})}{P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \cdots + P(B|A_k)P(A_k)}.}$$

This general formula can be used to find the probability of any number of individuals to share any possible number of IBD alleles. For three individuals, the notation is expanded:

'3' = the three mutated alleles are IBD
'2:1' = two of the three alleles are IBD (the third is not, i.e. recurrent mutation on the same haplotype)
'1:1:1' = all three alleles are recurrent mutation on the same haplotype

If we let the event A_i represent *either one* of these, the formula can be written as:

$$\mathbf{P(A_i|H)} = \frac{\boldsymbol{P(H|A_i)P(H)}}{\boldsymbol{P(H|'3')P('3')} + \boldsymbol{(H|'2{:}1')\,P('2{:}1')} + \boldsymbol{(H|'1{:}1{:}1')\,P('1{:}1{:}1')}}.$$

As mentioned, probability for identity is calculated by adding up all the probabilities for randomly picking two IBD alleles in the population of mutated alleles. In a similar manner, the probabilities for any constellation of IBD/non-IBD alleles can be calculated. For a group of two:

$$\text{Identity} = \text{P}('2') = \sum_{i=1}^{n} L_i^2$$

Or for:

$$\text{P}('1{:}\,1') = \sum_{\substack{i=1 \\ j\neq i}}^{n} \sum_{j=1}^{n} L_i * L_j.$$

Similarly for three:

$$\text{P}('3') = \sum_{i=1}^{n} L_i^3$$

$$\text{P}('2{:}\,1') = \sum_{\substack{i=1 \\ j\neq i}}^{n} \sum_{j=1}^{n} L_i^2 * L_j$$

$$\text{P}('1{:}\,1{:}\,1') = \sum_{\substack{i=1 \\ i\neq j\neq k}}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} L_i * L_j * L_k.$$

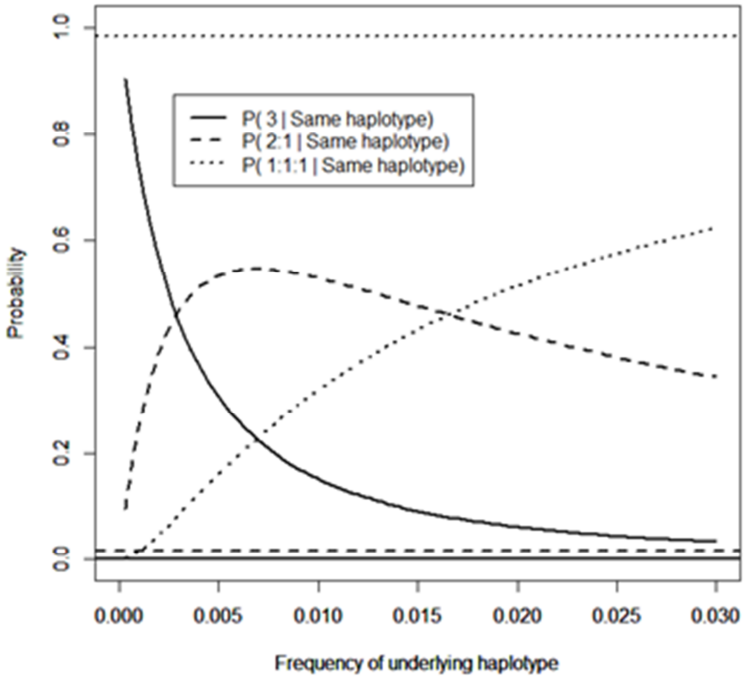

Figure 9a: The probabilities for different IBD relationships for three individuals with mutant alleles of the same type, assuming the same underlying haplotype. Horizontal lines indicate were the curves will converge when the underlying haplotype frequency is equal to one. Population at mutation-selection equilibrium with a *dominant* mutation with selection disadvantage of $s=0.2$ and $N\mu=100$ per generation. Only haplotype frequencies from 0 to 0.03 are shown.

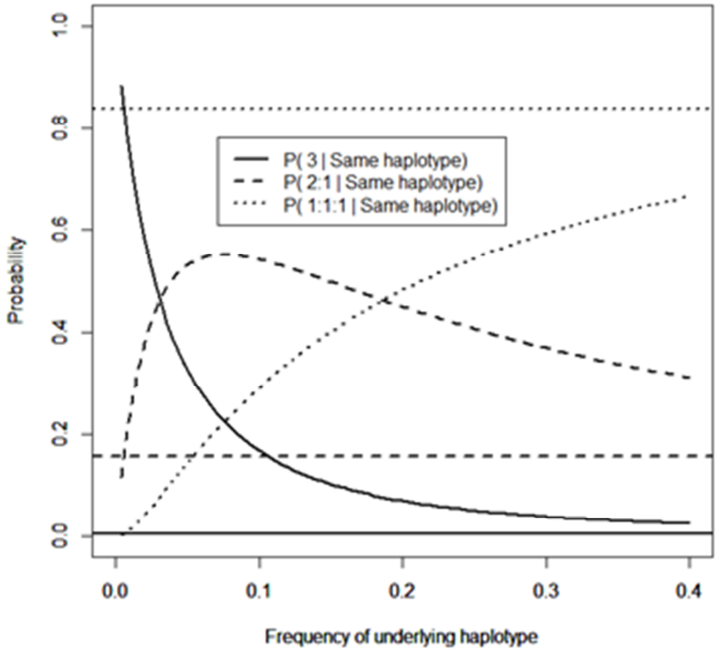

Figure 9b: The probabilities for different IBD relationships for three individuals with mutant alleles of the same type, assuming the same underlying haplotype. Horizontal lines indicate were the curves will converge when the underlying haplotype frequency is equal to one. Population at mutation-selection equilibrium with a *dominant* mutation with a selection disadvantage of $s=0.2$ and $N\mu=10$ per generation. Only haplotype frequencies from 0 to 0.4 are shown.

Again it can be seen that only when the haplotype is uncommon does the conditional probability differ much from the non-conditional probability. It is also evident that the information added by the *haplotype frequency* is much higher for the low $N\mu$ of 10, than for the higher 100 (***figures 9a-b***).Also noteworthy is that the *'2:1'*group's probability peaks at an intermediary haplotype frequency, though still at a very low frequency.

Using the general formula the same thing can be done for four individuals:
'4'= the four alleles are IBD
'3:1' = three out of four alleles are IBD
'2:2' = two separate groups of two who within the subgroup share IBD alleles
'2:1:1' = two of the four alleles are IBD
'1:1:1:1' = none of the four alleles are IBD

$$\mathbf{P}(\mathbf{A_i}|\mathbf{H}) = \frac{\boldsymbol{P}(\mathbf{H}|\mathbf{A_i})\boldsymbol{P}(\mathbf{H})}{\boldsymbol{P}(\mathbf{H}|'\mathbf{4}')\boldsymbol{P}('\mathbf{4}') + (\mathbf{H}|'\mathbf{3}:\mathbf{1}')\,\boldsymbol{P}('\mathbf{3}:\mathbf{1}') + (\mathbf{H}|'\mathbf{2}:\mathbf{1}:\mathbf{1}')\boldsymbol{P}('\mathbf{2}:\mathbf{1}:\mathbf{1}') + (\mathbf{H}|'\mathbf{1}:\mathbf{1}:\mathbf{1}:\mathbf{1}')\,\boldsymbol{P}('\mathbf{1}:\mathbf{1}:\mathbf{1}:\mathbf{1}')}$$

Where:

$$\mathrm{P}('4') = \sum_{i=1}^{n} L_i^4$$

$$\mathrm{P}('3:1') = \sum_{\substack{i=1 \\ j\neq i}}^{n}\sum_{j=1}^{n} L_i^3 * L_j$$

$$\mathrm{P}('2:2') = \sum_{\substack{i=1 \\ j\neq i}}^{n}\sum_{j=1}^{n} {L_i}^2 * L_j^2$$

$$\mathrm{P}('1:1:1:1') = \sum_{\substack{i=1 \\ i\neq j\neq k\neq l}}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n} L_i * L_j * L_k * L_l$$

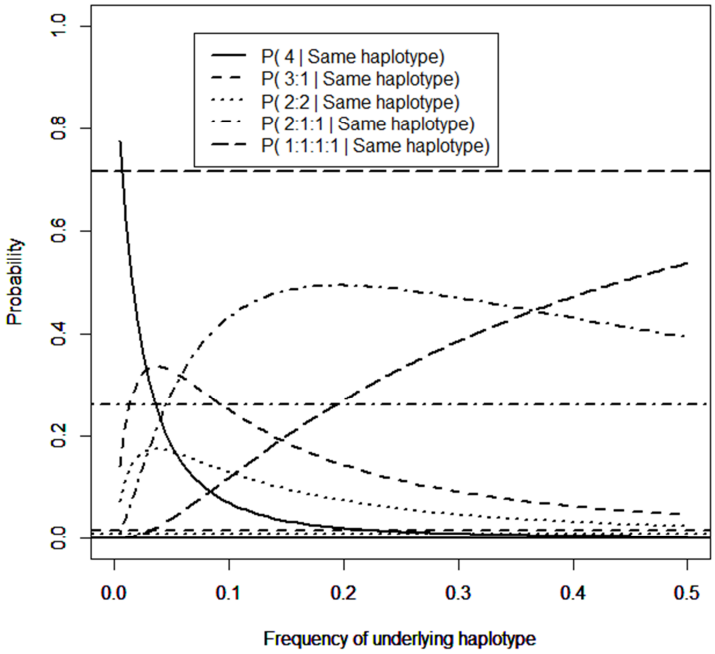

Figure 10: The probabilities for different IBD relationships for four individuals with mutant alleles of the same type, assuming the same underlying haplotype. Horizontal lines indicate were the curves will converge when the underlying haplotype frequency is equal to one. Population at mutation-selection equilibrium with a *dominant* mutation with a selection disadvantage of $s=0.2$and $N\mu=10$ per generation. Only haplotype frequencies from 0 to 0.5 are shown.

As more individuals are added, less probability is distributed to each outcome. Now, at somewhat higher frequencies of the underlying haplotype, the conditional probabilities are further away from the unconditional probabilities.

DISCUSSION

This study provides striking evidence that even a severely deleterious allele can exist for many generations, which is especially interesting when considering species with a long reproduction cycle, such as humans. 20, or even 100 generation may not sound very much, but in a human context it means hundreds or even thousands of years. And remember, this study only considers alleles in steady populations. It is not hard to imagine how the rapidly increasing human population that has been experienced since the industrialization will effectively make any small difference in general fitness almost negligible. For an evolutionist this both confirms previous assumptions, but can also call for re-evaluations. It is clear from the overall high variance for both persistence and pervasiveness that in individual cases, it is random factors rather than fitness that is the main player. There is a paradox in a steady population with an average fitness of one, where random fluctuations should rapidly change the number of individuals, and

even if the population sometimes could, or even should, become very large, eventually it would hit zero. The answer to this of course lies in ecology, where a steady environment favors a steady population. If a population should decrease in numbers, it is not hard to imagine that the overall fitness would increase, because of now further accessible habitats, et cetera. This reasoning puts reliability to the genetic models that are available, since the discrepancy from the real world is not as extensive as the axioms imply. It could then be argued that there is no such thing as a stable environment, but at least that relocates the problem from population genetics. The high variance also puts a time scale to the change of allele frequencies. If an allele attributed with positive selection has to 'compete' with an allele that will go extinct because of negative selection, but that momentarily increases in frequency by random fluctuations, how can alleles with positive selection ever go to fixation? The answer is of course time, where for every allele attributed with positive selection, there have probably been many which went extinct. Likewise, for every disease-causing allele that is known to be around in a family line for generations, there are many that just affected a single person.

Considering this, perhaps the most telling results from this study is the average individuals with deleterious mutation per linage and generation. It is shown that it is likely that this *mean* strives against equilibrium, not unlike the *mutation-selection equilibrium*. The selection is the obvious reason that prevents the 'individual per linage mean' to increase indefinitely. But what force is keeping this mean up at the level shown in the results? A pragmatic, but perhaps not so scientific, answer to this question is simple: it is luck. For a deleterious allele to persist to a late generation it has to have a history of 'luck'. Since luck is just randomness in subjective hindsight, if alleles are to persist to a later generation, the average number of alleles cannot be too few. If so, they would randomly go extinct much faster. On the other hand, if the alleles become many, the relentless force of negative fitness brings the number down again. For example, for 1 000 individuals with an average fitness of 0.8, their offspring would be very close to 800, and almost never above 1 000, but for just 10 of these individuals, it is not that unlikely that they would produce 10 or more offspring with some regularity. Hence, for a small group, randomness plays a much larger role than for big groups.

Even though this study focuses on deleterious alleles, the non-equilibrium situations with neutral alleles are shown as a pretext (***figure 6a***). The apparent paradox that the expected age of an allele continues to increase, even when the identity is momentarily increasing, has an explanation. It is because the total number of individuals from a certain generation is expected to be about the same for further generations in the near future, but they will stochastically coalesce to fewer and fewer origins because of drift. This causes a certain generation to increase the expected age as time passes, regardless of the composition of different origins. This is true when $N\mu$ is large in relation to the time scale. In the very long run, all the 'new' mutated alleles from a certain generation will all be gone in this model. This puts

emphasizes on the fact that this model works well for pragmatic purposes regarding some few alleles, but not so well for the evolution of an entire population. Alleles, as well as individuals, in one generation must have ancestors in all earlier generations in a real population. This connects to the *lottery paradox*; it is highly unlikely that any single ticket would be the winning one, but one must be the winning one. When there is selection on the other hand, the model becomes more appropriate for the context. Now the interesting thing is the low correlation between the identity, expected age and number of individuals with mutations of the same type. Intuitively, there should be a strong connection between these three. The random walk of these factors is relatively high compared to the expected causality between them.

The Bayesian approach to the IBD problem for small groups is only a first step to try and answer a very complex problem. The well-used *infinite-site model* for mutations presented by Kimura and Crow (1964) states that mutations affecting the same site in the DNA are so unusual that they can be neglected. Even though this is a very important and usually reliable model, its thinking can be counterproductive in this case. In this study, the very possibility that a single site has mutated twice must be included as a potential reality. The questions asked and answered in this report are informative, but perhaps a bit on the theoretical side. For example, it is not entirely clear what qualifies 'the same underlying haplotype', as a true shared IBD haplotype for two individuals can stretch for a different number of base pairs due to molecular recombination. Considering all this, the question still not answered is the probability that two or more identical mutations are not IBD, depending on the length of the shared underlying haplotype?

Attempts to validate the results in this report have been made. Of interest is of course to compare results from this study, using a simulating approach, with previous results were only mathematics were used. In the review article *'On the Persistence and Pervasiveness of a new Mutation'* (Garcia-Dorado A, Caballero A, Crow J, 2003) the results from various researches in the field were presented:

Table 6

hs	$\bar{p} = 1/hs$	$\bar{p} = 1/hs(1 + 1/8N_e hs)$	$\bar{t}$ (diffusion)	$\bar{t}$ (branching)
0.001	1,000	1012.5	13.28	12.64
0.01	100	101.2	8.67	8.13
0.05	20	20.2	5.45	5.16
0.1	10	10	4.06	4.00

Source -'On the Persistence and Pervasiveness of a new Mutation' *(Garcia-Dorado A, Caballero A, Crow J, 2003)*

The results are very similar. For example, if *h=1* and *hs=s*, then the two different values for persistence at *s=0.1* presented in the review article (***table 6***) are 4.06 and 4.00 respectively, and in the results presented in this study it was equal to 4.12 (***table 1***). The pervasiveness for the same *s* according to the article is 10, and in this study 10.56 (***table 4***).

Nei (1970) presents a probability distribution for extinction time for alleles using a diffusion equation:

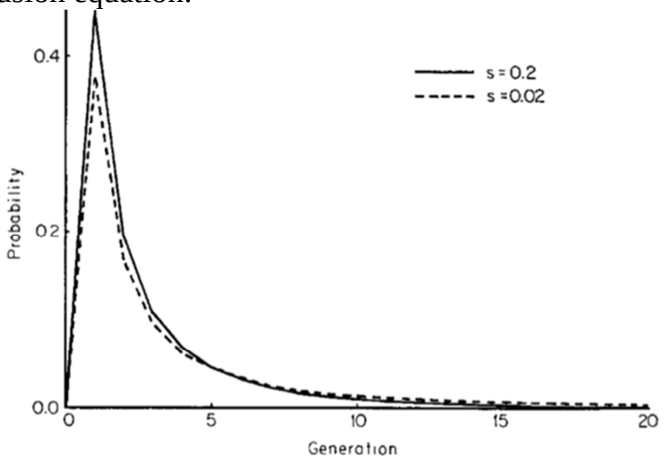

Figure 11: Probability distribution of extinction times for alleles using a diffusion equation. *Source – Nei M, 2003*

This probability distribution (***figure 11***) is remarkably similar to the actual frequency distribution presented in this report. Nei also points out the small difference in the appearance of the distribution for the two different selection coefficients, but that the differences in *mean* and *variance* are quite large. This can partially be explained by the nature of the *Poisson distribution*. Consider just the first generation after the mutations, and the fact that the probability of getting zero offspring does not change radically if *s* changes. The probability of getting zero offspring in the dominant version is *37%* if *s=0, 38%* if *s=0.02* and *45%* if *s=0.2*. This corresponds well to what can be seen in the results.

What is random?

The programming language *'R'* uses the *Mersenne-Twister algorithm* to produce pseudorandom numbers. Pseudorandom means that the numbers aren't actually random, but seems like it. From a statistical view this is of course troublesome, but in the algorithm, measures have been taken to avoid the obvious appearance of patterns. Pseudorandom numbers are thus widely used in statistical analysis and produce reliable results in most situations.

The results from this study often include a time dimension, and the situation in one generation is of course dependent on both the situation in the last generation, and the (pseudo-)random change.

This correspondence puts trust in the results presented in this report, even for the results that have not been presented in earlier studies, since the results are generated using the same algorithm: the *double iteration loop*.

Results from the second algorithm used in this study, *whole population equilibrium*, are easier to validate just using existing mutation selection equilibrium. $X=N\mu/f$ holds very well for the expected number of individuals carrying the mutation at equilibrium.

All in all, this study shows that the alternative approach of simulating rather than calculating probable outcomes serves its purposes. Though new insights have been made, several questions raised in this study still stand unanswered.

REFERENCES

1. Eric R. Pianka, *The American Naturalist*, Vol. 104, No. 940 (Nov-Dec., 1970), pp. 592-597, The University of Chicago Press

2. Fisher R.A 1930, *The Genetical Theory of Natural Science*, Clarendon Press, Oxford

3. Garcia-Doradoa A, Caballerob A, Crow J.F, 2003, On the Persistence and Pervasiveness of a New Mutation, *Evolution* 57(11):2644-2646. *Doi: http://dx.doi.org/10.1554/03-207*

4. Haldane J.B.S 1927, The Mathematical Theory of Natural and Artificial Selection - Part V: Selection and mutation, *Proc. Cambridge Philos. Soc.* 23, 838-844

5. J. F. Crow, 1986, *Basic concepts in population*, Quantitative and evolutionary genetics, New York: W.H. Freeman. p. 273.

6. Kimura M, Crow J, 1964, The Number of Alleles that can be Maintained in a Finite Population, *Genetics* 49: 725–738. *PMC: 1210609, PMID: 14156929*

7. Kimura M, Ohta T, 1969, The Average Number of Generations until Fixation of a Mutant Gene in a Finite Population, *Genetics* 61:763-771

8. Li W-H, Nei M. Total number of individuals affected by a single deleterious mutation in a finite population. *American Journal of Human Genetics* 1972;24(6 Pt 1):667-679.

9. Muller H. J. 1950, Our Load of Mutations, *AM. J. Hum. Genet*. 2:111-176

10. Nei M, 1971, Extinction time of Deleterious Mutant Genes in Large Populations, *Theoretical Population Biology* 2(4):419–425.

11. *Theoretical Population Biology*, Volume 2, Issue 4, December 1971, Pages 419–425