Why recessive lethal alleles have not disappeared?

Jorma Jormakka
jorma.o.jormakka@gmail.com

Lethal recessive alleles are gene alleles, which either are lethal for a homozygote, or were so in past centuries. They cause rare serious diseases including Cystic Fibrosis (carrier frequency 1/24 in Northern Europeans, see [1]), Tay-Sachs disease (1/30 Ashkenazi Jews), Gaucher disease (1/15 Ashkenazi Jews), α-Thalessemia (1/25 Chinese and SE Asians), β-Thalessemia (1/30 Greeks and Italians). Most of these diseases can be caused by several different mutations, but the disease is expressed by a homozygote of a single mutated allele. In the past a homozygote of a lethal allele died before reaching the reproductive age, thus two mutated alleles carried by a homozygote were removed from the gene pool in every generation. We would expect that such deadly diseases became less frequent in each generation and they would vanish after a certain time and consequently the lethal alleles we now can observe must have been created relatively recently. Yet this is not the case: the age of the most common allele causing Cystic Fibrosis is estimated as 52,000 years. There must be some mechanism keeping these extremely harmful alleles in the gene pool.

One proposal is that a heterozygote of the mutated allele has a selective advantage. The classical case of a heterozygote advantage is the Sickle Cell disease. This disease is not quite lethal for a homozygote but causes a serious illness. The disease has the carrier frequency 1/12 in African Americans and relatively higher carrier frequencies in areas where malaria occurs than in areas without this infective disease. It has been demonstrated that a heterozygote of the Sickle Cell disease has partial immunity towards malaria. However, heterozygote advantage has not been sufficiently well demonstrated for any of the mentioned recessive lethal alleles, though it has been suggested, for instance, that Cystic Fibrosis gives partial protection against cholera [2] and the Ashkenazi Jewish diseases may offer a cognitive advantage for a heterozygote [3]. The explanation of the persistence of recessive lethal alleles by a heterozygote advantage is weak, and in this analysis it will be shown that this explanation cannot be correct since it would lead to a different ratio between the disease prevalence and the carrier frequency than what is observed.

The second proposal for an explanation is a founder effect followed by a genetic drift. It is of course possible that among a small number of founders several have the same rare disease and in this way the mutated alleles become enriched in the population. A genetic drift, especially in small populations, can still increase the frequency of mutated alleles. The problem with this explanation is that such a process would be very unlikely e.g. in the case of the main allele of Cystic Fibrosis (CF). It will be shown in this analysis that a recessive lethal allele would vanish from the population in 50 generations unless there is a mechanism keeping it in the population. A generation is about 30 years. The main allele of CF is 52,000 years old. That is about 30 times longer than the time for the allele to disappear. We should assume that a founder effect occurred some 30 times. As such a founder effect must be a quite rare event it cannot have a probability very close to 1. This probability, what-ever it is, raised to power 30 gives a number very close to zero. There is a nano-scale chance that a sequence of 30 founder effects could be the correct explanation why the CF allele still is there.

After discarding these two common proposals a "new" mechanism is proposed. There is nothing especially new in this mechanism as such: it is just that certain family lineages tend to have many children and this alone can in about 8-10 generation produce observed carrier frequencies for recessive lethal alleles in the population. Still in the present context the proposal seems to be new as the only mechanisms that usually are suggested are heterozygote advantage or founder effect with a genetic drift.

Genetics of such a system is easy: every mutated allele can be considered separately as a system of two alleles: the original allele A and the mutated allele a. If the mutated allele brings neither selective benefits not disadvantages, the relation between the frequency of homozygote (aa) of allele a and heterozygote (aA) can be calculated from Hardy-Weinberg equilibrium [4][5]: assuming that the probability of allele a is x, then the probability of A is $(1-x)$. Thus the probability of two alleles a is x^2 and it is the probability of a homozygote aa. In a similar way the probability of a homozygote AA is $(1-x)^2$ and consequently the probability of a heterozygote is $1-x^2-(1-x)^2 = 2x(1-x)$. Denoting the probability of a homozygote by q and of heterozygote by p we get $q = x^2$, $p = 2x(1-x)$ and the probability of homozygotes AA is $1-p-q = (1-x)^2$. Eliminating x yields $p = 2\sqrt{q}\sqrt{1-p-q}$ whence

$$q^2 - (1-p)q + \frac{1}{4}p^2 = 0,\ q = \frac{1}{2}(1-p) - \frac{1}{2}\sqrt{1-2p} = \frac{1}{4}p^2 + \frac{1}{4}p^2 + O(p^4). \qquad (1)$$

This is the steady state solution of a two allele system assuming that the mutated allele gives neither advantage not disadvantage, but with a lethal recessive allele homozygote aa naturally have a major disadvantage. We can model a system, which is not in a steady state by using recursion formulas. Let p_n and q_n be the frequencies of heterozygote aA and homozygote aa respectively in a generation n, thus p_n is the carrier frequency and q_n is the disease prevalence as a function of time given as number n of generations from the beginning.

Two AA parents will have only AA children. This occurs with probability $(1-p_n-q_n)^2$ as each parent comes from a pool of AA homozygotes, which has the probability $(1-p_n-q_n)$. Similarly, two aa parents have only aa children and this event has the probability q_n^2. The probability of the case of aA having children with AA is $2p_n(1-p_n-q_n)$. Half of the children will be AA and half aA. Similarly aA-aa has the probability $2p_nq_n$. Half of the children are aA, half aa. Two aA parents producing children has the probability p_n^2. Half of these children will be aA, one fourth AA and one fourth aa. The final case is AA-aa. This event has the probability $2(1-p_n-q_n)q_n$ and all children are heterozygotes aA.

Let us insert two nonnegative parameters α and β to describe heterozygote advantage and homozygote disadvantage respectively. These parameters increase or decrease the number of children for a couple having certain combination of alleles a and A. For a lethal allele a no heterozygote aa can have children. Thus aa-aa have children of type aa with the probability βq_n^2 in the generation $n+1$ where $\beta = 0$. Likewise, aa-AA have children, all aA, with the probability $2\beta(1-p_n-q_n)q_n$ in the generation $n+1$ with $\beta = 0$, and aA-aa have children with the probability $2\beta p_nq_n$ in the generation $n+1$ with $\beta = 0$.

There are three combinations aA-aA, aA-AA and aA-aa, where the heterozygote aA appears but we give the heterozygote advantage only to the case aA-aA. This is done by modifying the original model so that if both parents are aA, then they produce more children. We define that the children of aA-aA have the probability αp_n^2 in the generation $n+1$, that is, these couples produce α times as many children than AA-AA couples. Half of these children will be aA, one fourth AA and one fourth aa.

The reason for not giving a heterozygote advantage to the case aA-aa is that as we are mostly interested in lethal allele a, β is zero and aA-aa have no children. It does not matter if we multiply zero by any α. The reason why we do not give a heterozygote advantage to aA-

AA is that it is not possible to find a steady state solution for small p if we do so and the carrier frequency p for recessive lethal alleles is on the range of 1/25. Consider what would happen if we multiply the number of children of aA-AA couples by α. The leading term of p for the number of heterozygote children is obtained from children of the couples aA-AA and it would be αp as the leading term of $\frac{1}{2}\alpha 2p(1-p-q)$, here always $q \propto p^2$. The two leading terms of p for the number of AA children are obtained from AA-AA and aA-AA couples and the terms would be $1-(2-\alpha)p$ from $(1-p-q)^2+\frac{1}{2}\alpha 2p(1-p-q)$. If the system is in a steady state, then ratio of aA children to AA children must be the same as the ration of aA parents to AA parents. Thus $\frac{p}{1-p} \cong \frac{\alpha p}{1-(2-\alpha)p}$ where $\cong$ indicates that we ignored $O(p^2)$ terms. This equation can only be satisfied if p is close to ½. So, for recessive lethal alleles we cannot give recessive advantage to the case aA-AA.

It is rather natural to give a recessive advantage to the case aA-aA. In the case aA-AA no children is a homozygote of the type aa and in the past the parents cannot have known that one of them is a carrier, but for the case aA-aA the situation is different: one fourth of their children die young. The parents may have tried to compensate this situation by having more children. This is mathematically a heterozygote advantage even though a heterozygote has no real gain from one copy of the lethal allele. However, as will be seen, here is a surprise. We may initially think that if aA-aA parents simply have 4/3 times as many children as AA-AA parents, then they have effectively compensated to the lethal allele, but this is not so. The steady state requires that the number of aA heterozygotes stays the same from parents to children and a lethal allele removes the children of AA-aa, which is the second largest term of p contribution to the number of aA children and has the leading term $2q \propto p^2$ from $2q(1-p-q)$. Indeed, to compensate $\beta = 0$ it will be seen that we need $\alpha \approx 3$.

It sounds unrealistic that aA-aA couples would have had in the past three times as many children as AA-AA couples, but that is what the following calculation shows for a steady state solution and the idea in recessive advantage is that it is a steady state solution: loss of heterozygotes aA because of the lethal homozygote is compensated by more children because of recessive advantage.

If $\alpha = \beta = 1$ we have the original system and we get the Hardy-Weinberg equilibrium. If $\alpha = 1,\ \beta = 0$ the system cannot be in a steady state. The allele a is decreasing in each generation and we can estimate how fast the allele is removed from the population to undetectable frequencies. If $\beta = 0$, there is a value α which gives a steady state solution. Then α is greater than 1 and it is the heterozygote advantage. Naturally we could select β greater than one and study homozygote advantage, but this is not done in the present analysis. The scaled recursion equations from generation n to generation $n+1$ are:

$$q_{n+1} = \frac{1}{s}\left(\frac{1}{4}\alpha p_n^2 + \beta\left(q_n^2 + p_n q_n\right)\right) \quad \text{and} \tag{2}$$

$$p_{n+1} = \frac{1}{s}\left(p_n - p_n^2 - p_n q_n + \frac{1}{2}\alpha p_n^2 + \beta\left(2q_n - 2q_n^2 - p_n q_n\right)\right)$$

where the scaling factor s is the total probability

$$s = 1 + (\alpha - 1)p_n^2 - 2q_n + q_n^2 + \beta\left(2q_n - q_n^2\right). \tag{3}$$

Dividing by s assures that the total probability of equations (2), which is 1 in the generation n, stays as 1 in the generation $n+1$. Assigning $\alpha = \beta = 1$ yields $s = 1$ and

$$q_n^2 - q_{n+1} + p_n q_n + \frac{1}{4} p_n^2 = 0 .$$

We can see that the system has a steady state solution $p_{n+1} = p_n = p$, $q_{n+1} = q_n = q$ giving the Hardy-Weinberg equilibrium (1). If $\alpha = 1$, $\beta = 0$ there is no steady state solution. We will make a simple approximation of the solution. To make it simple, we will not scale the equations as in (2) by dividing the equations with s. Then for $\alpha = 1$, $\beta = 0$ they are

$$q_{n+1} = \frac{1}{4} p_n^2 \quad \text{and} \tag{4}$$

$$p_{n+1} = \frac{1}{2} p_n^2 + (1 - p_n - q_n) p_n .$$

If p_n is small to start with at $n = 0$, scaling by dividing with s makes little difference. We will assume that p_n is so small that $O(p_n^3)$ terms can be ignored. The equation for p_{n+1} reduces to

$$p_{n+1} = p_n - \frac{1}{2} p_n^2 .$$

This equation is approximately solved by

$$p_n = \frac{1}{N}\left(C + \frac{1}{2}\frac{n}{N} \right)^{-1} \tag{5}$$

which satisfies

$$p_{n+1} = p_n - \frac{1}{2} p_n^2 + O\left(N^{-3}\right).$$

In order to use this solution, N, the total number of generations must be so large that N^{-3} is ignorable. As the carrier frequency p_n is on the range of 1/25 for most recessive lethal alleles, $N \geq 25$ should be enough. The constant C in (5) is fixed by the initial value for p_n. Indeed, from (5) follows that

$$p_n = \frac{2}{2p_0^{-1} + n} \qquad , \qquad p_0 = \frac{2}{2p_n^{-1} - n} .$$

Necessarily $2p_n^{-1} - n$ must be larger than zero in the second equation. It is just stating the condition that p_n is decreasing in each generation and cannot have been higher than one in generation zero. Thus, if the allele a is still detectable at carrier frequency p_n in the generation n, there is a maximum number of generations it can have been decreasing. For Cystic Fibrosis p_n has the value 1/24 today. Consequently n must be smaller than 48. That means some 1450 years. The solution is approximation and cannot give precise values. Yet the age 52,000 years for the main allele of CF is too much at odds with this approximation. As promised in the beginning, this mathematical argument shows that some kind of mechanism must keep recessive lethal alleles in the population. Else we would only see relatively recent lethal mutations.

No more elaborated argument against a founder effect and genetic drift than was given in the beginning will be offered, but the possibility of heterozygote advantage will be analyzed. This advantage means that for $\beta = 0$ there is $\alpha \geq 1$ that keeps the lethal allele in the population. Over all these generations the lethal allele has not replaced the healthy allele but has a rather small carrier frequency. This means that the system must be in a steady state and

frequencies of p_n and q_n do not any more depend on n. We set $p_{n+1} = p_n = p$, $q_{n+1} = q_n = q$ in (2) and solve α from both equations:

$$\alpha = \frac{p^3 + pq - pq^2 - p^2 - \beta\left(3pq - pq^2 - 2q + 2q^2\right)}{p^2\left(p - \frac{1}{2}\right)}, \tag{6}$$

$$\alpha = \frac{q - qp^2 - 2q^2 + q^3 + \beta\left(q^2 + q^3 - pq\right)}{p^2\left(\frac{1}{4} - q\right)}.$$

Eliminating α gives an third order equation of q with parameters β and p

$$\frac{1}{2}(1+3\beta)q^3 - \left(1 + 2\beta - 3\beta p - \frac{3}{4}(1-\beta)p\right)q^2 \tag{7}$$
$$+\left(\left(\frac{1}{2} + \beta\right)p^2 - \frac{1}{4}(3+5\beta) + \frac{1}{2}(\beta+1)\right)q + \frac{1}{4}p^3 - \frac{1}{4}p^2 = 0.$$

A third order equation has an exact solution, but it is inconvenient. Assuming that p is small, $q \propto p^2$ is so small that the term q^3 can be ignored and as an approximation, we get a second order equation for q. The solution has a square root, which can be expanded as a power series of p, which is assumed small. In order to see what size of an error we are introducing by dropping the third order term q^3, we can do this approximation when $\beta = 1$. When $\beta = 1$ equation (7) reduces to (1) and yields the exact solution of Hardy and Weinberg. In our approximation we get by setting $\beta = 1$ in (7) the third order equation

$$2q^3 + q^2(3p-3) + q\left(\frac{3}{2}p^2 - 2p + 1\right) + \frac{1}{4}p^3 - \frac{1}{4}p^2 = 0$$

which we approximate with a second order equation, solve it and expand to a power series of p. The result is

$$q = \frac{1}{4}p^2\left(1 + p + \frac{3}{4}p^2\right) + O(p^5)$$

while the Hardy-Weinberg solution expanded as a power series yields

$$q = \frac{1}{2}(1-p) - \frac{1}{2}\sqrt{1-2p} = \frac{1}{4}p^2\left(1 + p + \frac{5}{4}p^2\right) + O(p^5).$$

The difference is $O(p^4)$ and ignorable for realistic values if *p* for recessive lethal alleles. Thus, the approximation is sufficiently good for our purposes, but if p is larger, this method must be used with care. For arbitrary β the approximation gives (to $O(p^4)$, which is the highest precision we can get)

$$q = \frac{1}{4}p^2\left(1 - \frac{2A}{1+2\beta}\right)\left\{1 + \left(\frac{3\beta + \frac{3}{4}(1-\beta)}{1+2\beta} - \frac{B}{\frac{1}{2} + \beta - A}\right)p\right\} + O(p^4) \tag{8}$$

where

$$A = \frac{1}{(\beta+1)^4}\left(-1 - 3\beta - \beta^2 + 3\beta^2 + 2\beta^4\right),$$

$$B = \frac{1}{(\beta+1)^6}\left(-1 - 6\beta - 13\beta^2 - 16\beta^3 - 5\beta^4 + 31\beta^5 + 10\beta^6\right).$$

Let us evaluate the solution (8) for some values of β:

If $\beta = 1$, then $A = B = 0$, $q = \frac{1}{4}p^2 + \frac{1}{4}p^3 + O(p^4) = 0.25p^2 + 0.25p^3 + O(p^4)$.

If $\beta = 0$, then $A = B = -1$, $q = \frac{3}{4}p^2 + \frac{17}{16}p^3 + O(p^4) = 0.75p^2 + 1.06p^3 + O(p^4)$.

If $\beta = 2$, then $A = \frac{45}{81}$, $B = \frac{1487}{729}$, $q = 0.1575\,p^2 + 0.11p^3 + O(p^4)$.

Heterozygote advantage α can be solved by inserting β and the approximation of q into (6). For $q = c_1 p^2 + c_2 p^3 + O(p^3)$ equation (6) gives $\alpha = 4c_1 + 4(c_2 - \beta c_1)p + O(p^2)$:

If $\beta = 1$, $\alpha = 1 + O(p^2)$ (in this case the exact solution for α is 1.)

If $\beta = 0$, $\alpha = 3 + 4.24\,p + O(p^2)$.

If $\beta = 2$, $\alpha = 0.63 - 0.82\,p + O(p^2)$.

We see that in all cases $q \propto p^2$, but the coefficient is different. For all recessive lethal alleles the values announced in literature for the disease prevalence q are related to the carrier frequency p in the way that is very close to the Hardy-Weinberg equilibrium

$$q \cong \frac{1}{4}p^2 + \frac{1}{4}p^3.$$

In some cases this may be a result of measuring only the disease prevalence and calculating the carrier frequency from the Hardy-Weinberg formula, but at least [1] contains direct measurements of carrier frequencies and announces also how many homozygote cases the test sample contained. The sample in [1] is sufficiently large for measuring the carrier frequency, while it may be too small for estimating the disease prevalence in the sample. There fortunately are better values for the disease prevalence. There are certain problems arising from the composition of the samples in [1], but the results seem to fit to the Hardy-Weinberg equilibrium, or to the exact non-steady state un-scaled solution $q = \frac{1}{4}p^2$ when $\beta = 0, \alpha = 1$. The difference between these solutions and the steady state scaled solution for $\beta = 0, \alpha \approx 3$ with $q \cong 0.75\,p^2 + 1.06\,p^3$ is so large that it should be seen in the sample of [1]. Consequently, the explanation of persistence of recessive lethal alleles because of heterozygote advantage must be discarded. It can also be questioned if recessive advantage is the only mechanism in the Sickle Cell disease. In that disease recessive advantage is a likely cause, but not necessarily the only cause for the observed carrier frequency.

The final contribution of this analysis is a proposal of a mechanism that can explain why recessive lethal alleles do not disappear. The argument is based on a simple model, which is not in every respect realistic, but illustrates the mechanism sufficiently well. The idea is that many family lineages tended in the past to have about the same number of children over several generations. Thus, there were family lineages where most women had a large number of children, and the number could be higher than what was customary in the general population. This is still the case in some religious sects, which do not practice birth control and consider children as gifts from God. The proportion of people in a population originating from these family lineages grows over generations and if one such lineage included carriers of rare diseases, the carrier frequency of the population grows.

The model is a Markov model, which is constructed to be easy to analyze. Let $s_{n,j}$ be the fraction of the population of generation n being born into a family of j girls who grow old enough to reproduce. This implies that the state (n,j), which has the state probability $s_{n,j}$,

contains also women born into a family where both the mother and father were heterozygotes and more than j girls were born but homozygotes died before reaching the reproductive age. Naturally, the recessive disease is not the only reason why children die before reaching the reproductive age. Most families faced this situation.

For the model we take a birth and death process:

$$s_{n+1,j} = \frac{1}{s_n}\left\{(1-(j-1)\lambda-(j+1)\mu)s_{n,j} + (j-1)\lambda s_{n,j-1} + (j+1)\mu s_{n,j+1}\right\}.$$

Here s_n is the scaling factor to get the total probability to remain at one. The term $(j-1)\lambda s_{n,j-1}$ describes women, who were born into a family of $j-1$ daughters who grow up to reproduce, but who themselves have j daughters. The parameter λ describes the probability of having one daughter more than the mother, while the multiplier $j-1$ indicates that all $j-1$ daughters have this decision to make. In a similar way, the term $(j+1)\mu s_{n,j+1}$ describes women, who were born into a family of $j+1$ daughters who grow up to reproduce, but who themselves have j daughters. The parameter μ describes the probability of having one daughter less than the mother, while the multiplier $j+1$ indicates that all $j+1$ daughters have this decision to make. The remaining term $(1-(j-1)\lambda-(j+1)\mu)s_{n,j}$ describes those women who have the same number of daughters as their mother.

Assuming that the system is in a steady state, the flow in and out of state (n,j) are equal:

$$(j-1)\lambda s_{n,j-1} = j\mu s_{n,j}$$

yielding the solution

$$s_{n,j} = \frac{(j-1)}{j}\frac{\lambda}{\mu}s_{n,j-1} = \frac{(j-1)}{j}\frac{(j-2)}{(j-1)}\ldots\frac{1}{2}\sigma^{j-1}s_{n,1} = \frac{1}{j}\sigma^{j-1}s_{n,1}\ ,\ \sigma = \frac{\lambda}{\mu}. \qquad (9)$$

The scaling factor s_n is the sum of the state probabilities:

$$\frac{s_n}{s_{n,1}} = \sum_{j=1}^{\infty}\frac{1}{j}\sigma^{j-1} = \sigma^{-1}\int\sum_{j=1}^{\infty}\sigma^{j-1}d\sigma = \sigma^{-1}\int\frac{1}{1-\sigma}d\sigma = -\sigma^{-1}\ln(1-\sigma).$$

Setting $s_n = 1$ fixes $s_{n,1} = -\dfrac{\sigma}{\ln(1-\sigma)} = \dfrac{\sigma}{\ln(1/(1-\sigma))}$. (10)

As $s_n = 1$ the recursion simplifies to

$$s_{n+1,j} = (1-(j-1)\lambda-(j+1)\mu)s_{n,j} + (j-1)\lambda s_{n,j-1} + (j+1)\mu s_{n,j+1}. \qquad (11)$$

The average number of girls in generation n (women in generation $n+1$) is

$$Av = \sum_{j=1}^{\infty} j s_{n,j} = \sum_{j=1}^{\infty}\sigma^{j-1}s_{n,1} = \frac{1}{1-\sigma}\frac{\sigma}{\ln(1/(1-\sigma))}.$$

In distant history human population growth was very small. In this simple model $Av = 1$ corresponds to a population with zero growth, that is, one daughter implies two children in average. This value for Av gives a small σ and we can determine σ from a power series expansion of the logarithm:

$$1 = \frac{1}{1-\sigma}\frac{1}{\sigma + \frac{1}{2}\sigma^2 + O(\sigma^3)} = 1 + \frac{1}{2}\sigma + O(\sigma^2),$$

so if $Av = 1$, then $\sigma = 0$, but we cannot select $\sigma = 0$ because then the system does not reach the steady state solution that was calculated before. In order to reach it, λ and μ must be positive. Let us set $\sigma = 0.1$. Then $Av = 1.055$, which is very close to zero growth. The total

fertility rate is the double of Av, 2.1, and it is very close to the minimum for sustaining a population. We can also notice that the value $\sigma = 0.72$ gives $Av = 2.02$ implying about 4 children per woman, that is 2.3% annual growth and 30 years (=one generation) doubling time. Before modern times such growth rates were a rarity.

Selecting σ does not fix λ and μ, only their relation, and the absolute values of λ and μ are important for determining the average number of descendants in the nth generation for a family, which started with j children at the generation zero. This is obviously so because if $\lambda = \mu = 0$, all daughters of the family lineage will have j daughters reaching the productive age. Then the number of women grows as j^n and the population as $2j^n$. We cannot select $\lambda = \mu = 0$, because that implies that after some time the whole population grows as $2j^n$, where j is the highest number of daughters any woman of the zero generation had. However, we can set λ and μ to small positive values. Doing so, we can estimate the number of female descendants of a single woman of the zero generation having j daughters, who reach the reproductive age.

If λ and μ are small, it is sufficient to calculate only one or two state changes from one j value to another in the whole run of generations from 0 to n. In the beginning all probability is in the state (0,j), i.e., $s_{0,j} = 1$. No state changes gives the following contribution to (n,j):

$$s_{n,j,0} = (1-(j-1)\lambda-(j+1)\mu)^n s_{0,j} = (1-(j-1)\lambda-(j+1)\mu)^n . \qquad (12)$$

If there is only one state change in the run, there is no contribution to (n,j), while from two state changes there are. The state can change from (m,j) to (m+1,j ± 1) and back from (r,j ± 1) to (r+1,j) giving second order contributions to

$$s_{n,j,2a} = j(j-1)\lambda\mu\sum_{m=0}^{n}(1-(j-1)\lambda-(j+1)\mu)^{m+n-r}\sum_{r=m+1}^{n-1}(1-(j-2)\lambda-j\mu)^{r-m} ,$$

$$s_{n,j,2b} = j(j+1)\lambda\mu\sum_{m=0}^{n}(1-(j-1)\lambda-(j+1)\mu)^{m+n-r}\sum_{r=m+1}^{n-1}(1-j\lambda-(j+2)\mu)^{r-m} .$$

For simplicity we ignore from now on the second order contributions. Thus, (12) is the approximation of the state probability of $s_{n,j}$. The (first order) approximations of the state probabilities $s_{n,j-1}$ and $s_{n,j+1}$ are respectively

$$s_{n,j-1} = j\mu\sum_{m=0}^{n}(1-(j-1)\lambda-(j+1)\mu)^m(1-(j-2)\lambda-j\mu)^{n-m} ,$$

$$s_{n,j+1} = j\lambda\sum_{m=0}^{n}(1-(j-1)\lambda-(j+1)\mu)^m(1-j\lambda-(j+2)\mu)^{n-m} .$$

The number of female descendants with j daughters of the one woman in the zero generation is approximated by

$$Num_j = j^n(1-(j-1)\lambda-(j+1)\mu)^n .$$

If λ and μ are very small, we can ignore even the first order terms and keep only this term. It is essentially j^n. Including male descendants, the woman has $2j^n$ descendants and if the woman was a carrier, half of the descendants are carriers of the lethal allele. As the total population has negligible growth, the carrier frequency of the population reaches relatively high levels because of this exponential growth. This exponential growth does not continue infinitely. It stops when n is on the range of $(j(\lambda+\mu))^{-1}$. We may estimate that this n could be about 10 by the following reasoning.

Human female has a upper limit for number of children probably around 16, but very large families, where children grow up to have their own children, must have been rare. We probably can ignore families with more than 4 daughters. For $j = 4$ daughters $n = 10$ generations of growth approximated by $j^n(1 - nj(\lambda + \mu))$ gives about 1 million female descendants to the generation n, that is two million people. If $\lambda + \mu = \lambda(1 + \sigma^{-1}) = 11\lambda$ (we have selected $\sigma = 0.1$) is sufficiently much smaller than $(nj)^{-1} = 1/40$, the growth is almost exponential. That means that $\lambda \le 1/440$, which is small but not necessarily impossible in this simple model. It very much depends on the value selected for σ.

Two millions is 4%, that is 1/25, of 50 million, which in the past was a large population. We see that in ten generations observed carrier frequencies can be reached by a family lineage which has 8 children per woman. For $j = 3$ ten generations of growth produces 118,000 people. It is about 4% of 3 million. That is a more typical size that a population, which today has diseases caused by recessive lethal alleles, may have had 300 years (=10 generations) ago. For $j = 2$ we get 2000 people in 10 generations, for $j = 1$ the number stays at one and the case $j = 0$ there are no daughters.

Assuming that the woman, who starts this family line at the generation zero, is a typical member of the larger population, the probability for her to have j daughters living to a reproductive age is

$$\frac{1}{j}\sigma^{j-1}\frac{\sigma}{\ln(1/(1-\sigma))} \cong \frac{1}{j}\sigma^{j-1}$$

Multiplying this probability by the number of descendants in the generation $n = 10$ and summing over the values $j = 0,...,4$ yields the average number of descendants:

$$Descendants = 0 + 1 + \frac{1}{2}0.1 \cdot 2000 + \frac{1}{3}0.1^2 \cdot 118{,}000 + \frac{1}{4}0.1^3 \cdot 2{,}000{,}000 = 994.$$

The woman was a heterozygote for a lethal allele having the carrier frequency p with the probability p. We may assume that her husband mostly was AA, as a is a rare allele. Thus, half of her children were carriers. Half of the children are boys and we may assume for simplicity that all boys were AA and all girls aA. This way it is not necessary to track the boys. All daughters in all generations are therefore carriers in this calculation. If the woman of the zero generation was a carrier, she produced 994 carriers to the 10th generation.

At the same time the number of heterozygotes aA decreases by a considerable factor. Since $\beta = 0$, the couples AA-aa do not exist and they do not produce $2q(1 - p - q) \approx 2q$ carriers. The couples aA-aa also do not exist, but their contribution is of the order $O(p^3)$. From (2) we can see what is missing if $\alpha = 1,\ \beta = 0$ to the steady state solution $\alpha = \beta = 1$. The carrier frequency decreases in each generation by a fraction

$$1 - \frac{1}{2}p_n.$$

As 10 generations is a short time, p_n does not change very much and we can estimate that the change is about

$$p \to p\left(1 - 10\frac{1}{2}p\right).$$

If p is originally about 1/25=0.04, it decreases to about 0.032. At the same time we get 993 new carriers. In order 0.8% of the population (0.04-0.032) to be 993, the population size should be 124,125.

This is of course a very simple conceptual model and cannot be fully realistic. Yet it shows that if family lineages have a practice of getting the same number of children to adulthood as their parents did, which is about the same as making the same number of children, then it creates a pump, which increases the number of heterozygotes and can balance the loss of heterozygotes due to the death of homozygotes. Is there any reason to think that there were such practices in the past? The age of the main allele of Cystic Fibrosis, unless the dating will be revised, takes us back to the Stone Age. Hunter-gatherer societies usually have few children because many children restrict the mobility of women. As Hunter-Gatherer women get pregnant in a normal way, such societies practice infanticide: only one child, who cannot walk alone, can be nursed by a woman. This implies that the time between children is typically 3-5 years. As a woman reaches maturity at around 15 years and the life length around 35 years, a woman could raise 4-6 children, but as such societies tend to be violent, few lived long. Usually only tribal chiefs had more wives, often 2-4, and consequently more children, but that does not increase female fertility. It seems that there was no possibility for families with a large number of children, which is required by the mechanism proposed here. However, this may be a too fast judgment. There could have been areas and times when food was abundant, women could be semi-sedentary and nurse the children, or something else.

From the time of sedentary habitation, first in the Levant already before agriculture, family sizes could grow and families with 6 to 8 children were more like a rule in agricultural societies. Still the population grew very slowly, much below 1% annually. These facts can be combined by an assumption that most children did not reach the reproductive age, or that many adults died young, were widowed, taken to slavery, or for some other reason did not raise a large family. Some family lineages did and the gene pool probably was all the time changing with more fertile lineages replacing less fertile ones. Some may see here a place for natural selection, some only a play of chance. Such a situation explains why some religions gained support much easier than lends support to the forces of natural selection. Fertility cults and later patriarchic religions, which forbade infanticide, created family lineages which produced many children. Such lineages grew to represent the majority of the society.

Can this be a better explanation to the puzzle of Ashkenazi Jewish intelligence, pondered in [3]. Probably not for intelligence, but it may explain their collection of rare genetic diseases. Ashkenazi Jews had for about 800 years population growth rate about 1.4% annually. It was much higher than in the host society. A high growth rate implies large families and while 1.4% per year (50 years doubling time) means only doubling in two generations (setting the female generation to 25 years for simplicity), which is 2.8 children per woman, some family lineages almost certainly grew much faster. The pump mechanism described here could have contributed to keeping recessive lethal alleles in the population. It could also make non-lethal, even advantageous, alleles more common, but this is not the topic of the present analysis. In any case, the mechanism was not simply recessive advantage if understood in a simple way that heterozygotes had more children grown to the reproductive age.

A positive side in this is that as family sizes today are small in developed countries, such a pump mechanism cannot work. Recessive lethal alleles would be purged out of the population, unless modern medicine makes them non-lethal and the removal mechanism is blocked. That this can be so may be shown by Finns not having Cystic Fibrosis even though Finns have a large portion of genes from European Western Hunter Gatherers, who presumably had this disease as it is common in Northern Europe. If Finns had smaller families, the disease was purged out. Interestingly, the same mechanism may slow down evolution, but that topic I will leave to another time.

References:

[1] Gabriel A. et al, “An empirical estimate of carrier frequencies for 400+ causal Mendelian variants: results from an ethnically diverse clinical sample of 23,453 individuals,” (2013), Genetics in Medicine volume 15, pages 178–186.
https://www.nature.com/articles/gim2012114
[2] Rodman DM and Zamudio S., “The cystic fibrosis heterozygote--advantage in surviving cholera?” Med Hypotheses. (Nov 1991);36(3):253-8.
https://www.ncbi.nlm.nih.gov/pubmed/1724059
[3] Gregory Cochran, Jason Hardy, and Henry Harpending. (2006). “Natural history of Ashkenazi Intelligence”, Journal of Biosocial Science 38:659-693:1-35.
[4] Hardy, G. H. (Jul 1908). *"Mendelian Proportions in a Mixed Population"*, Science. 28 (706): 49–50.
[5] Weinberg, W. (1908). "Über den Nachweis der Vererbung beim Menschen". Jahreshefte des Vereins für vaterländische Naturkunde in Württemberg. 64: 368–382.