*Genet. Res., Camb.* (2001), **78**, *pp.* 41–47. *With 2 figures.* © 2001 Cambridge University Press
DOI: 10.1017/S0016672301005195 Printed in the United Kingdom


# Do Δ*F*508 heterozygotes have a selective advantage?

CARSTEN WIUF*
*Department of Statistics, University of Oxford, 1 South Parks Road, Oxford OX1 3TG, UK*

(*Received 28 November 2000 and in revised form 21 March 2001*)

**Summary**

In this paper the fitness of the Δ*F*508 heterozygote is assessed and the age of the Δ*F*508 mutation in the cystic fibrosis locus is estimated. Data from three microsatellite loci are applied. The analysis is performed conditional on the present-day frequency of the Δ*F*508 mutation and based on assumptions about the demographic history of the European population and the mutation rate in the three microsatellite loci. It is shown that the data gives evidence of positive selection (up to 2–3% per Δ*F*508 heterozygote), but also that data could be explained by negative selection of roughly the same order of magnitude. The age of the Δ*F*508 mutation is subsequently estimated; it is found that the mutation is at least 580 generations old, but could be much older depending on the microsatellite mutation rate and the exact number of substitutions experienced in the history of the three microsatellite loci.

## 1. Introduction

Cystic fibrosis (CF) is the most common fatal recessive single-gene disorder of the European population affecting 1 in 2000–4000 individuals (Boat *et al.*, 1989). More than 400 disease-causing mutations have been found in the Cystic Fibrosis Transmembrane Conductance Regulator (CFTR) gene responsible for CF. Most of these occur very infrequently but one, the Δ*F*508 mutation, accounts for about 70% of all CF chromosomes (Kerem *et al.*, 1989; Tsui, 1992).

The high frequency of CF chromosomes has led to the assumption that CF heterozygotes have some advantage over 'non-CF' homozygotes (Serre *et al.*, 1990; Morral *et al.*, 1994). This assumption has also been supported by experimental evidence; for example Rodman & Zamudio (1991) argue that CF heterozygotes have an advantage in surviving cholera, and Schroeder *et al.* (1995) argue that the Δ*F*508 mutation offers some protection against bronchial asthma. Further, Rodman & Zamudio's (1991) hypothesis has been supported by Gabriel *et al.* (1994) who, in a mouse study, suggest that a deletion similar to Δ*F*508 also has a selective advantage in surviving cholera.

* Tel: +44 (0)1865 27 28 70. Fax: +44 (0)1865 27 25 95. e-mail: wiuf@stats.ox.ac.uk

It is the aim of the present paper to assess the selective advantage of Δ*F*508 mutation in the CF locus and to estimate the age of allele carrying the Δ*F*508 mutation. Microsatellite data, obtained by Morral *et al.* (1994), from 1705 individuals sampled throughout Europe are used together with the frequency, $q$, of the Δ*F*508 allele in the European population. Provided some assumptions are made about demographic history, both the frequency and the variability within the Δ*F*508-allelic class carry information about age and selection. Two different expansion scenarios of the European population have been chosen: one corresponding to a Paleolithic expansion, the other to a Neolithic expansion.

The age of the Δ*F*508 mutation has previously been estimated using various techniques. Serre *et al.* (1990) found the age to be about 200 generations. Their estimate was based on linkage disequilibrium patterns observed in markers closely linked to the CF locus. Morral *et al.* (1994) estimated the age from intra-allelic variability in three microsatellites in the CF locus. They suggested that the Δ*F*508 mutation is at least 2600 generations old. This was challenged by Kaplan *et al.* (1994), who re-evaluated Morral *et al.*'s (1994) conclusions and found the age to be less than 900 generations old. All these estimates are estimates of the time, $T_2$, until the most recent common ancestor

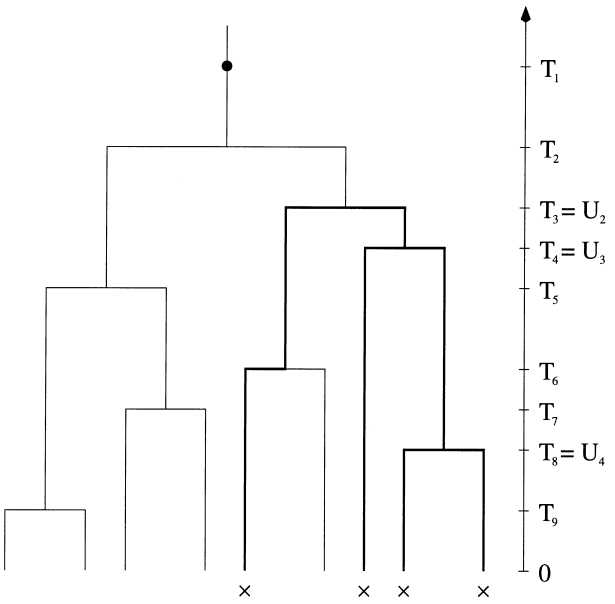


Fig. 1. An example of a genealogy of $\mathscr{D}$ and a sample, $\mathscr{D}_0$, embedded in $\mathscr{D}$. The sample $\mathscr{D}_0$ is shown with crosses and the genealogy relating $\mathscr{D}_0$ is represented by heavy lines. The size of $\mathscr{D}$ is 9 and that of $\mathscr{D}_0$ is $k = 4$. The variable $T_j$ denotes the time while there are at least $j$ ancestors of $\mathscr{D}$; in particular we let $T_1$ denote the time the mutation arose in the population (here marked with a black dot). Coalescence events in $\mathscr{D}_0$ happen at times $U_j$, $j = 2,\ldots, k$ (here $k = 4$) and each $U_j$ coincides with one $T_i$ for some $i$. We have $T_1 > T_j > 0$ for $j > 1$ and $U_2 > U_j > U_k > 0$ for $k > j > 2$. The time, $T_1 - T_2$, from the MRCA of $\mathscr{D}$ until the mutation arose is called the stick.

(MRCA) of the sample and not of the time, $T_1$, at which the mutation arose in the population (see Fig. 1). The distinction between $T_1$ and $T_2$ has been discussed by Slatkin & Rannala (1997), who also incorporated assumptions about selection and demography into their analysis. Depending on what is assumed about selection and demography they estimated the age, $T_1$, to be less than 500 generations old.

As mentioned above, the issue of heterozygote advantage has been addressed in previous studies (e.g. Serre *et al.* 1990; Morral *et al.* 1994), but an estimate of the strength of selection has not been given. If selection operates in favour of CF heterozygotes, the age of a CF causing mutation is expected to be different from the age of a mutation under neutrality (given that the current frequency of the mutation is the same). Various models to explore the relationship between the frequency and the age have been put forward, but no general agreement on this issue has been reached. Cavalli-Sforza & Bodmer (1971) consider a two-allele deterministic model where one homozygote (say, $A_1A_1$) is lethal and find the approximate number, $t$, of generations it takes $A_1$ to rise from a low frequency, $q_0$, to a given frequency $q_t > q_0$. If heterozygotes have a disadvantage the frequency of $A_1$ cannot rise to $q_t$ at all, and if heterozygotes have an advantage $t$ decreases with increasing advantage. Cavalli-Sforza & Bodmer's (1971) arguments were applied by Serre *et al.* (1990), among others. Slatkin & Rannala (1997) adopt an approach based on birth–death processes and estimate the age of a rare allele in terms of the heterozygote fitness and the frequency of the allele; the allele is expected to be older if heterozygotes are disadvantageous than if they are advantageous (Slatkin & Rannala, 1997, formula 8). Both approaches have been opposed by Maruyama (1974), who found that the mean age depends on the strength of selection, only through the absolute value of the fitness, not the direction. Recently, this has been supported by Wiuf (2001). In the papers by Maruyama (1974) and Wiuf (2001) the age is described conditional on $q$, whereas one obtains the likelihood of $q$ in Slatkin & Rannala (1997).

Here, theory developed in Wiuf (2000, 2001) is applied to describe the distribution of the age and the genealogical structure of the subpopulation, $\mathscr{D}$, carrying the $\Delta F508$ mutation and a sample $\mathscr{D}_0$ taken from $\mathscr{D}$. This theory is based on an analysis of the coalescent structure of a subpopulation of rare alleles (say, $q < 10\,\%$), conditioned on the frequency $q$, and the assumption that the mutation causing the rare allele has only happened once in the history of the entire population. In the case of the $\Delta F508$ mutation this is likely to be true (Serre *et al.*, 1990; Morral *et al.*, 1994). The genealogical structure is essentially described by two parameters: one parameter relates to the expansion rate of the entire population and the other describes the fitness of $\Delta F508$ heterozygotes. Because $q$ is low, selection against $\Delta F508$ homozygote can be ignored. The basic notation is introduced in Fig. 1.

At least two points are worth mentioning in connection with the previous approaches. Firstly, in all approaches the time at which the mutation arose (or the time until the MRCA of the sample) is treated as a parameter, a fixed quantity in the model. Wiuf & Donnelly (1999) (see also Wiuf 2000, 2001) found that the age should be considered a stochastic variable (that takes different values with different probabilities) based on stochastic models with parameters describing the population history and the mutation process. Effects of both the demography and the mutation process are not appropriately accounted for when the age is considered to be a parameter. For example, the mutation process is not modelled in any of the above-mentioned approaches. Secondly, the stick $W_1 = T_1 - T_2$, i.e. the time from the MRCA to the time the mutation arose, is not stochastically independent of $T_2$ but positively correlated with $T_2$. Thus, observed intra-allelic data carry information not just about $T_2$ but also about $T_1 - T_2$.

Given that the growth rate of the population is

https://doi.org/10.1017/S0016672301005195 Published online by Cambridge University Press

known, the fitness of ΔF508 heterozygotes can be inferred, using the method of maximum likelihood, from the genealogical structure of the observed sample and the intra-allelic variability in the sample. Subsequently, the distribution of the age, conditional on the observed intra-allelic variability, can be estimated.

## 2. Theory

The following is from Wiuf (2001). A model of exponential growth of the entire population is adopted. The population consists of two types of alleles, $A_1$ and $A_2$, and allele $A_1$ carries the mutation. Let $N$ be the total present-day population size (and thus $2N$ is the number of chromosomes) and $q$ the frequency of $A_1$. The model contains four parameters: (1) the number $2Nq$ of $A_1$ alleles, (2) the scaled growth rate $\rho = 2Nqr$, where $r$ is the growth rate per generation, (3) the scaled selection coefficient $\sigma = 2Nqs$, where $s$ is the fitness of the heterozygote $A_1A_2$ (selection for homozygotes $A_1A_1$ can be ignored because $q$ is assumed to be small) and (4) the scaled mutation rate $\theta = 4Nqu$, where $u$ is the intra-allelic mutation rate per locus per generation. Let $\alpha = \rho + \sigma$. Note that $\rho$, $\sigma$ and $\theta$ depend on $N$ and $q$ only through $Nq$.

Time is measured in units of $2Nq$ generations and $T_j, j \geqslant 1$, is the time while there are at least $j$ ancestors of the subpopulation $\mathscr{D}$, conditioned on the frequency, $q$, of $A_1$ today. In particular, the variable $T_1$ denotes the time at which the mutation arose in the population, $T_2$ is the time until the MRCA of $\mathscr{D}$, and $W_j = T_j - T_{j+1}$, $j \geqslant 2$, are the times between successive coalescence events (see Fig. 1). There exist independent variables $X_j$, $j \geqslant 1$, such that

$$T_j = \frac{1}{|\alpha|} \log\left(1 + \frac{2|\alpha|}{X_1 + \ldots + X_j}\right), \tag{1}$$

$$X_1 \sim C \frac{x^{\rho/|\alpha|}}{(2|\alpha| + x)^{\rho/|\alpha|}} e^{-x}, \tag{2}$$

with $C$ a normalizing constant and

$$X_j \sim \text{Exp}(1), \tag{3}$$

for $j \geqslant 2$. Here, Exp($\lambda$) denotes an exponential density with parameter $\lambda$ and $|x|$ denotes the absolute value of $x$. If $\alpha = 0$ equation (1) becomes $T_j = 2/(X_1 + \ldots + X_j)$ and equation (2) becomes $X_1 \sim e^{-x-2\rho/x}$. The variables $X_j$, $j \geqslant 1$, are for simulation purposes only.

Below a number of interesting points that derive from (1)–(3) are listed. Let $M(\rho, \sigma)$ denote the model with parameters $\rho$ and $\sigma$.

- For $\rho$ and $\sigma$ fixed, $q$ affects time only through the linear scaling factor $2Nq$. For instance, if $q$ is doubled all times are doubled (when measured in units of generations).
- The effects of selection and growth are not additive; i.e. equation (1) depends on both $|\alpha|$ and $\rho$, and not just the sum, $\alpha$, of $\rho$ and $\sigma$. This non-additivity is most profound when $\rho$ and $\sigma$ both are small and vanishes if they are large and $\sigma$ positive. For large $\rho$ and $\sigma$, $X_1$ is approximately gamma-distributed with parameters $\phi = \rho/|\alpha| + 1$ and 1; $X_1 \sim \Gamma(\phi, 1)$. Further, if $|\alpha|$ is large, the dominating term in (1) is of order $\log(|\alpha|)/|\alpha|$.
- The waiting times $W_j = T_j - T_{j+1}$ between successive coalescence events are not independent, but positively correlated. In particular, the stick $W_1$ correlates positively with $T_2 = \sum_{j \geqslant 2} W_j$. In the standard coalescent model the corresponding waiting times are independent.
- Two models $M(\rho, \sigma)$ and $M(\rho', \sigma')$ cannot share distributions (1)–(3) unless $\rho = \rho'$. Furthermore one cannot, from the shape of the genealogy alone, distinguish a model with growth $\rho$ and selection $\sigma$ from one with growth $\rho$ and selection $\sigma' = -2\rho - \sigma$. Either $\sigma \geqslant 0$ and $\sigma' \leqslant 0$ or $\sigma, \sigma' \leqslant 0$; both cannot be positive.

The last point has an unexpected implication, namely the following:

- Assume the demography (that is $\rho$) is known. Then, conditioned on the frequency $q$, data observed at the present time can always be explained assuming that heterozygotes are selectively disadvantageous.

Formulas (1)–(3) give the distribution of the genealogy of the entire subpopulation, $\mathscr{D}$. In practice, only a subsample, $\mathscr{D}_0$, of size $k$ taken randomly from $\mathscr{D}$ is considered. For instance, in the data of Morral *et al.* (1994), $\mathscr{D}_0$ has size $k = 1705$. The distribution of the genealogy of a sample $\mathscr{D}_0$ of size $k$ can be found applying (1)–(3) and theory in Saunders *et al.* (1984). This is also reviewed in Wiuf (2000), and details will not be given here. Let $U_j, j = 2, \ldots, k$, denote the time while there are at least $j$ ancestors of $\mathscr{D}_0$ (see Fig. 1), and let $B_k$ denote the total branch length of the genealogy of $\mathscr{D}_0$,

$$B_k = U_2 + \sum_{j=2}^{k} U_j. \tag{4}$$

As is standard, the number of mutations, $S_k$, in the history of the sample $\mathscr{D}_0$ is assumed to have a Poisson distribution, $Po(\theta B_k/2)$, i.e.

$$P(S_k = n | B_k) = \frac{(\theta B_k)^n}{n!\, 2^n} e^{-\theta B_k/2}. \tag{5}$$

## 3. Methods

Morral *et al.* (1994) examined $k = 1705$ copies of ΔF508 from European individuals with respect to intra-allelic variability in three microsatellite loci. A parsimony analysis suggested that the number of mutations $S_{1705}$ in the three microsatellite loci was

https://doi.org/10.1017/S0016672301005195 Published online by Cambridge University Press

Table 1. *Parameter values applied in estimation of* $\sigma = 2Nqs$

| | | $N$ | | |
|---|---|---|---|---|
| $q = 0.02$ | | $1 \times 10^8$ | $3 \times 10^8$ | $9 \times 10^8$ |
| $\theta = 4Nqu$ | $u = 10^{-5}$ | $8.0 \times 10^1$ | $2.4 \times 10^2$ | $7.2 \times 10^2$ |
| | $u = 10^{-4}$ | $8.0 \times 10^2$ | $2.4 \times 10^3$ | $7.2 \times 10^3$ |
| | $u = 10^{-3}$ | $8.0 \times 10^3$ | $2.4 \times 10^4$ | $7.2 \times 10^4$ |
| $r$ | Rec Exp | $1.9 \times 10^{-2}$ | $2.1 \times 10^{-2}$ | $2.3 \times 10^{-2}$ |
| | Old Exp | $3.7 \times 10^{-3}$ | $4.1 \times 10^{-3}$ | $4.6 \times 10^{-3}$ |
| $\rho = 2Nqr$ | Rec Exp | $7.4 \times 10^4$ | $2.5 \times 10^5$ | $8.3 \times 10^5$ |
| | Old Exp | $1.5 \times 10^4$ | $4.9 \times 10^4$ | $1.7 \times 10^5$ |

Rec Exp denotes a recent expansion beginning 10000 years ago from a population of size 10000 and growing exponentially until the present size, $N$. Old Exp denotes an expansion beginning 50000 years ago, but otherwise similar to Rec Exp. One generation is counted as 20 years.

about 46. The actual number is assumed to be $S_{1705} = 46$, but an alternative $S_{1705} = 92$ (following Slatkin & Rannala, 1997) is also considered.

The strength of selection acting on the $\Delta F508$ mutation is estimated assuming the scaled growth rate $\rho = 2Nqr$ and the scaled mutation rate $\theta = 4Nqu$ are known (defined in Section 2). This leaves the scaled fitness $\sigma = 2Nqs$ (defined in Section 2) to be estimated from the data $S_{1705}$. Here this is accomplished using maximum likelihood estimation, though other approaches could be applied as well. The estimator, $\hat{\sigma}$, of $\sigma$ is the set of $\sigma$'s that solve the equation

$$L(\hat{\sigma}) = \max_{\sigma} L(\sigma), \tag{6}$$

where $L(\sigma)$ is the likelihood of the data, $L(\sigma) = P(S_{1705} = n \mid \sigma)$ and $n$ is the observed number of mutations, $n = 46$ or $n = 92$. As pointed out previously there is not just one $\sigma$ that solves (6) but two (see Section 2).

Similar values of $N$, $q$, $\theta$ and $\rho$ as in Slatkin & Rannala (1997) are adopted. These are described below and shown in Table 1. The $\Delta F508$ mutation is the most frequent amongst mutations causing CF and comprises about 70% of all CF-causing mutations. Combined with an overall frequency of CF-causing mutations of 3% in European populations this gives $q \approx 0.7 \times 0.03 \approx 0.02$. Morral *et al.* (1994) gave a lower and an upper bound for the combined mutation rate, $u$, at all three microsatellite sites, $u \in (0, 10^{-3})$, based on an experimental study of 3000 meioses. Here, in addition to $u = 10^{-3}$ the possibilities $u = 10^{-4}$ and $u = 10^{-5}$ are considered.

## 4. Results

Table 2 shows the estimator of $\sigma$ (only the larger of the two possible estimators is shown) together with estimated 95% confidence intervals (CIs). The mutation rate is $u = 10^{-4}$. Initial investigations suggest that the $-2\log Q$ statistic (where $Q = L(\sigma)/L(\hat{\sigma})$ is the likelihood ratio) is approximately $\chi^2(1)$-distributed and this is applied in calculations of CIs. As an example, Fig. 2 shows the log likelihood curves for $N = 3 \times 10^8$ and $u = 10^{-4}$ under Old Exp (Old Expansion model, see Table 1).

It transpires that under Rec Exp (Recent Expansion, Table 1) (and $S_{1705} = 46$) the hypothesis that $\Delta F508$ is selectively neutral cannot be rejected whereas neutrality is easily rejected under Old Exp. Here the selective advantage reaches high levels (Table 2). However, it also transpires that $|\hat{\sigma} + \rho|$ is approximately constant; e.g. if $N = 3 \times 10^8$ then $\hat{\sigma}$ and $\rho$ are linearly related $\hat{\sigma} = 2.5 \times 10^5 - \rho$.

Qualitatively different results are obtained with other mutation rates. If $u = 10^{-3}$ and $\sigma = 0$, the expected number of mutations in the history of a sample of size 1705 is extraordinarily high for all reasonable demographic scenarios and strong selection is needed to counterbalance expectation and explain why few mutations are observed. The level of selection is found to be far higher than realistically possible (which might be about the 2–3% estimated in Table 2). Under Rec Exp, $\hat{\sigma}$ is 5–15 times higher than $\rho$ and under Old Exp, $\hat{\sigma}$ is at least 30 times higher than $\rho$, depending on $N$ and $S_{1705}$. Under both scenarios, this corresponds to $\hat{s}$ being larger than 0.1. The above suggests that $u = 10^{-3}$ is unrealistically high.

If $u = 10^{-5}$ there is evidence in favour of selection, but the data cannot be explained by positive selection. For all investigated values of $N$ and $r$ we find that the larger of the two $\hat{\sigma}$'s is less than $-0.70\rho$ and in most cases close to $-\rho$. If $u$ is even lower than $10^{-5}$, the observed number of mutations is far larger than would be expected for any value of $\sigma$, and $\hat{\sigma} \approx -\rho$. These results contradict the experimental studies of Rodman & Zamudio (1991) and Schroeder *et al.* (1995) who argue for heterozygote advantage, and suggests that $u = 10^{-5}$ is unrealistically low.

https://doi.org/10.1017/S0016672301005195 Published online by Cambridge University Press

Table 2. *Estimates of σ and confidence intervals*

| | | Rec Exp | | | Old Exp | | |
|---|---|---|---|---|---|---|---|
| $S_{1705}$ | $N$ | $\hat{\sigma}/\rho$ | 95% CI | $\hat{s}$ | $\hat{\sigma}/\rho$ | 95% CI | $\hat{s}$ |
| 46 | $1\times10^8$ | −0.14 | (−0.42, 0.26) | $-3.4\times10^{-3}$ | 3.20 | (1.90, 5.25) | $1.2\times10^{-2}$ |
| | $3\times10^8$ | 0.00 | (−0.30, 0.44) | $0.0\times10^{-3}$ | 4.10 | (2.60, 6.35) | $1.7\times10^{-2}$ |
| | $9\times10^8$ | 0.10 | (−0.20, 0.58) | $2.4\times10^{-3}$ | 4.40 | (2.90, 6.70) | $2.0\times10^{-2}$ |
| 92 | $1\times10^8$ | −0.66 | (−0.76, −0.54) | $-1.3\times10^{-2}$ | 0.70 | (0.25, 1.25) | $2.6\times10^{-3}$ |
| | $3\times10^8$ | −0.58 | (−0.68, −0.44) | $-1.3\times10^{-2}$ | 1.15 | (0.65, 1.85) | $4.7\times10^{-3}$ |
| | $9\times10^8$ | −0.52 | (−0.62, −0.36) | $-1.2\times10^{-2}$ | 1.40 | (0.85, 2.10) | $6.4\times10^{-3}$ |

The mutation rate is assumed to be $u = 10^{-4}$, and only the larger of the two possible estimators of $\sigma$ is shown. The other solution, $\tilde{\sigma}$, to the likelihood equation is given by $\sigma/\rho = -2-\hat{\sigma}/\rho$. The parameter $\sigma/\rho$ was varied in jumps of 0.02 under Rec Exp and in jumps of 0.05 under Old Exp. $\hat{s} = \hat{\sigma}/(2Nq)$ is the estimated selective (dis)advantage per generation. One thousand simulations were performed for each value of $S_{1705}$, $N$ and $\sigma/\rho$.

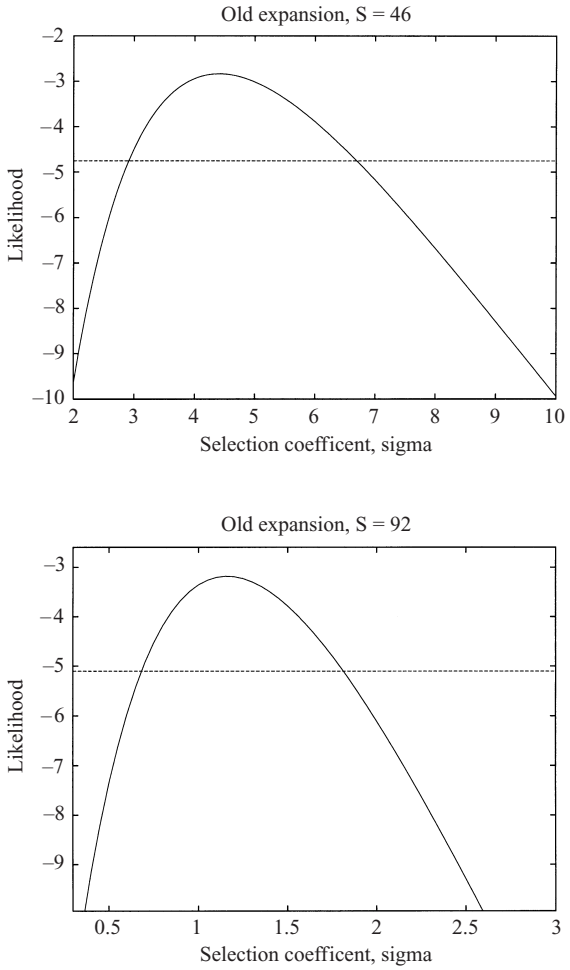


Fig. 2. The figure shows the log likelihood curves as function of $\sigma$ (shown in units of $\rho$) for $N = 3\times10^8$ and $u = 10^{-4}$ under Old Exp. If log $L(\sigma)$ is above the dotted line, $\sigma$ is in the 95% CI.

The distribution of the age of the $\Delta F508$ mutation, conditional on $S_{1705}$, is found using (1)–(5) and applying the estimated value of $s$ (or $\sigma$) together with prefixed values of $N$, $r$ and $u$ (or $\rho$ and $\theta$). The simulation scheme in Tavaré *et al.* (1996) is applied. Mainly the case where $u = 10^{-4}$ is studied and further, because $|\hat{\sigma}+\rho|$ is approximately constant and large, only Old Exp is considered. In other words, the estimated age is roughly the same under the two scenarios Rec Exp and Old Exp (see equations (1) and (2)). Table 3 shows summary statistics of the distribution of the age (in generations). One generation is counted as 20 years (which is also assumed in previous studies).

Table 3. *Age of the mutation in generations*

| $N$ | $S_{1705} = 46$ | | $S_{1705} = 92$ | |
|---|---|---|---|---|
| | Mean | SD | Mean | SD |
| $1\times10^8$ | 770 | 71 | 1700 | 140 |
| $3\times10^8$ | 650 | 54 | 1300 | 120 |
| $9\times10^8$ | 580 | 38 | 1200 | 85 |

The mutation rate is assumed to be $u = 10^{-4}$. Because $|\rho+\hat{\sigma}|$ is almost constant and large there are essentially no differences between the results obtained under Rec Exp and Old Exp (results under Old Exp are shown in the table). 'Mean' refers to the mean of the distribution of the age of the $\Delta F508$ mutation, conditional on $S_{1705}$, and 'SD' to the standard deviation of the same distribution. One thousand simulations were performed to obtain the distribution of the age for each value of $S_{1705}$ and $N$.

First note that the age is more sensitive to the observed number of mutations $S_{1705}$ than to $N$: the age increases by (at least) a factor 2 when going from $S_{1705} = 46$ to $S_{1705} = 92$, whereas the age is reduced less than 30% when $N$ is increased by a factor of 9. Second, the high ages found with $S_{1705} = 92$ predate the start of the recent expansion (10000 years ago; 1200 generations is about 24000 years). Under Rec Exp the age would be even older than shown in Table 3 because in the pre-expansion period the population

https://doi.org/10.1017/S0016672301005195 Published online by Cambridge University Press

size is constant (or nearly constant) and not decreasing exponentially (as the model assumes). Note the low standard deviations shown in Table 3. In all cases they are less than 10% of the mean age. If uncertainty in the estimate of $\sigma$ is taken into account higher standard deviations would be found.

If $u = 10^{-3}$, the $\Delta F508$ mutation is very young (50–130 generations old, depending on the various parameters) and this does not seem consistent with the widespread geographical distribution of $\Delta F508$ (Morral *et al*., 1994). If $u = 10^{-5}$ the $\Delta F508$ mutation is more than 10000 generations old and this is far greater than is realistically possible.

## 5. Discussion

Theory in Wiuf (2000, 2001) has been applied to assess the fitness of the $\Delta F508$ heterozygote in the CF locus and to estimate the age of the $\Delta F508$ mutation. The amount of selection depends strongly on what is assumed about the demographic history of the population and the microsatellite mutation rate. Morral *et al*. (1994) found an upper bound, $u = 10^{-3}$, to the mutation rate per generation in three microsatellites in the CF locus. If $u = 10^{-3}$, the $\Delta F508$ mutation is very young and this conflicts with the geographical spread of the mutation. If the mutation rate is very low, say $u \leqslant 10^{-5}$, the age of the mutation predates the 'out of Africa'-expansion. This does not seem likely. If $u = 10^{-4}$ and the European population has grown since 50000 years ago, the data of Morral *et al*. (1994) do support the hypothesis that heterozygotes have a selective (dis)advantage. If the expansion is more recent (starting 10000 years ago) there is no or little evidence for selection in the data. In either case the age of the $\Delta F508$ mutation is estimated to be between 11000 and 16000 years old if $S_{1705} = 46$ (Table 3) and at least 24000 years if $S_{1705} = 92$. Since $S_{1705} = 46$ is a lower bound to the possible number of mutations, 11000 years (or 580 generations) gives a lower bound to the age of the mutation. The onset of European population expansion was probably somewhere 'between' the two scenarios used here: either a Neolithic or a Paleolithic expansion. (The 'true' demography could have been more complicated than a simple exponential expansion.) Thus, the best explanation of the data may be to conclude that $\Delta F508$ heterozygotes have a selective advantage.

Selection at rate $\sigma$ and selection at rate $-2\rho-\sigma$ affect the genealogy in identical ways (conditioned on the present-day frequency, $q$, of the mutation). Thus, it can be alternatively concluded that $\Delta F508$ heterozygotes have a selective disadvantage. However, the probability that the mutation is found in frequency $q$ today is different under the two scenarios, rate $\sigma$ and rate $-2\rho-\sigma$. In the case of CF it is unlikely that the $\Delta F508$ allele could have risen to the present-day high frequency (about 2%) if the allele was under negative selection. The chance of seeing the allele at frequency 2% today is higher if $\sigma > 0$ than if $\sigma < 0$. The evidence in favour of selection is thus taken to be evidence in favour of a selective advantage. According to Table 2 this advantage per generation per allele could be as high as 3%.

Uncertainty about the expansion rate could be circumvented in at least two different ways. A Bayesian view could be adopted and a prior distribution put on $\rho$ and $\sigma$ (or $r$, $s$ and $N$). (The same procedure could be used to deal with uncertainty in assigning a value to $u$.) A prior distribution on $\rho$ could be constructed from former demographic studies and $\sigma$ could be assigned a flat prior based on the poor knowledge available about selection at the CF locus. The posterior distribution of $\sigma$ given data could thus be estimated and the amount of selection inferred from the shape of the distribution. Alternatively, the data could be extended with samples of selectively neutral and unlinked markers taken from the same population. If an estimate of mutation rate in each marker locus could be achieved, both $\rho$ and $\sigma$ could be estimated based on an analysis similar to the one performed here. However, as demonstrated, uncertainty in mutation rates can have very profound effects.

This study suggests a lower bound to the age of the $\Delta F508$ mutation (about 580 generations). Thus, the age predates a Neolithic expansion and is consistent with the finding of $\Delta F508$ mutations in Pakistani patients (Malone *et al*., 2000). Several previous studies are not consistent with this finding. Under a Neolithic expansion there was no evidence in Morral *et al*.'s (1994) data for selection. It is therefore not surprising that a debate has been going on as to whether there has been selection or not; only if expansion is slow is selection inferred.

R. Harding is thanked for reading and commenting on several versions of the manuscript. Her comments improved the presentation considerably. The Mathematical Genetics Group at the department and M. Schierup are thanked for helpful comments. The author was supported by grant BBSRC 43/MMI09788 and by the Carlsberg Foundation, Denmark.

## References

Boat, T. F., Welsh, M. J. & Beaudet, A. L. (1989). Cystic fibrosis. In *Basis of Inherited Disease*, 6th edn (ed. C. L. Scriver, A. L. Beaudet, W. S. Sly & D. Valle), pp. 2649–2680. New York: McGraw-Hill.

Cavalli-Sforza, L. L. & Bodmer, W. F. (1971). *The Genetics of Human Populations*. San Francisco: Freeman.

Gabriel, S. H., Brigman, K. N., Koller, B. H., Boucher, R. C. & Jackson Stutts, M. (1994). Cystic fibrosis heterozygote resistance to cholera toxin in the cystic fibrosis mouse model. *Science* **266**, 107–109.

https://doi.org/10.1017/S0016672301005195 Published online by Cambridge University Press

Kaplan, N. L., Lewis, P. O. & Weir, B. S. (1994). Age of the Δ*F*508 cystic fibrosis mutation. *Nature Genetics* **8**, 216–218.

Kerem, B. S., Rommens, J. M., Buchanan, J. A., Markiewicz, D., Cox, T. K., Chakravarti, A., Buchwald, M. & Tsui, L.-C. (1989). Identification of the cystic fibrosis gene: genetic analysis. *Science* **245**, 1073–1080.

Malone, G., Haworth, A., Schwartz, M. J., Cuppens, H. & Super, M. (1998). Detection of five novel mutations of the Cystic Fibrosis Transmembrane Regulator (CFTR) gene in Pakastani patients with cystic fibrosis: Y569D, Q98X, 296 + 12(T > C), 1161delC and 621 + 2(T > C). *Human Mutation* **11**, 152–157.

Maruyama, T. (1974). The age of a rare mutant gene in a large population. *American Journal of Human Genetics* **26**, 669–673.

Morral, N., Bertranpetit, J., Estivill, X., Nunes, V., Casals, T., Giménez, J. & Reis, A. *et al*. (1994). The origin of the major cystic fibrosis mutation (Δ*F*508) in European populations. *Nature Genetics* **7**, 169–175.

Rodman, D. M. & Zamudio, S. (1991). The cystic fibrosis heterozygote: advantage in surviving cholera? *Medical Hypotheses* **36**, 253–258.

Saunders, I. W., Tavaré, S. & Watterson, G. A. (1984). On the genealogy of nested subsamples from a haploid population. *Advances in Applied Probability* **16**, 471–491.

Schroeder, S. A., Gaughan, D. M. & Swift, M. (1995). Protection against bronchial-asthma by CFTR Δ*F*508 mutation: a heterozygote advantage in cystic fibrosis. *Nature Medicine* **1**, 703–705.

Serre, J. L., Simon-Bouy, B., Mornet, E., Jaume-Roig, B., Balassopoulou, A., Schwartz, M., Taillandier, A., Boué, J. & Boué, A. (1990). Studies of RFLP closely linked to the cystic fibrosis locus throughout Europe lead to new considerations in population genetics. *American Journal of Human Genetics* **84**, 449–454.

Slatkin, M. & Rannala, B. (1997). Estimating the age of alleles by use of intraallelic variability. *American Journal of Human Genetics* **60**, 447–458.

Tavaré, S., Balding, D. J., Griffiths, R. C. & Donnelly, P. (1996). Inferring coalescence times from DNA sequence data. *Genetics* **145**, 505–518.

Tsui, L.-C. (1992). Mutations and sequence variations detected in the cystic fibrosis conductance regulator (CFTR) gene: a report from the Cystic Fibrosis Genetic Analysis Consortium. *Human Mutation* **1**, 197–203.

Wiuf, C. (2000). On the genealogy of a sample of neutral rare alleles. *Theoretical Population Biology* **58**, 61–75.

Wiuf, C. (2001). Rare alleles and selection. *Theoretical Population Biology* (In press).

Wiuf, C. & Donnelly, P. (1999). Conditional genealogies and the age of a neutral mutant. *Theoretical Population Biology* **56**, 183–201.