Evaluating candidate agents of selective pressure for cystic fibrosis

Eric M. Poolman* and Alison P. Galvani

Department of Epidemiology and Public Health, Yale University School of Medicine, 60 College Street, Room 147, New Haven, CT 06520, USA

Cystic fibrosis is the most common lethal single-gene mutation in people of European descent, with a carrier frequency upwards of 2%. Based upon molecular research, resistances in the heterozygote to cholera and typhoid fever have been proposed to explain the persistence of the mutation. Using a population genetic model parameterized with historical demographic and epidemiological data, we show that neither cholera nor typhoid fever provided enough historical selective pressure to produce the modern incidence of cystic fibrosis. However, we demonstrate that the European tuberculosis pandemic beginning in the seventeenth century would have provided sufficient historical, geographically appropriate selective pressure under conservative assumptions. Tuberculosis has been underappreciated as a possible selective agent in producing cystic fibrosis but has clinical, molecular and now historical, geographical and epidemiological support. Implications for the future trajectory of cystic fibrosis are discussed. Our result supports the importance of novel investigations into the role of arylsulphatase B deficiency in cystic fibrosis and tuberculosis.

Keywords: cystic fibrosis; tuberculosis; heterozygote advantage; cholera; typhoid fever

1. INTRODUCTION

Cystic fibrosis (CF) is one of the most common lethal single-gene genetic diseases in populations of European descent, with more than 1400 catalogued mutations and a carrier frequency upwards of 2% (Cystic Fibrosis Genetic Analysis Consortium 2005). The age of the most prevalent mutation, ΔF508, has been estimated to be approximately 600 generations (Wiuf 2001; Morris *et al.* 2002). Many hypotheses were originally forwarded to explain the high incidence of CF, including combinations of a high mutation rate, a founder effect, genetic drift and heterozygote advantage (Bertranpetit & Calafell 1996). Hypotheses that assume no heterozygote advantage have been found inconsistent with microsatellite and haplotype data (Romeo *et al.* 1989; Sereth *et al.* 1993; Wiuf 2001), the occurrence of CF in multiple countries in Europe (Serre *et al.* 1990), the population history of Europe (Bertranpetit & Calafell 1996) and the relative absence of CF outside of Europe (Bertranpetit & Calafell 1996).

By contrast, heterozygote advantage against an infectious disease has been repeatedly supported as a tenable explanation for the high level of allele persistence (Anderson *et al.* 1967; Hansson 1988; Bertranpetit & Calafell 1996; Welsh *et al.* 2001). The prevailing candidates for the selective agents are enteric diseases supported by molecular research: typhoid fever (Pier *et al.* 1998), and secretory diarrhoea including cholera (SDC) (Chao *et al.* 1994; Cuthbert *et al.* 1994; Gabriel

et al. 1994). Tuberculosis has also been proposed originally on the basis of clinical observations (Anderson *et al.* 1967; Crawfurd 1972).

The correct selective agent must be plausible according to three lines of evidence: molecular; clinical; and historical–geographical (Galvani & Slatkin 2003). Firstly, molecular and cellular studies should demonstrate a mechanism through which heterozygosity provides resistance. Secondly, clinical studies should indicate correlations between host genotype and disease morbidity/mortality. Finally, on a historical–geographical level, the temporal and spatial distribution of the selective agent must be consistent with the observed allele frequency and incidence of the genetic disease in geographical regions of both high and low incidences. This third means of evaluation has been neglected. We develop a model that integrates demographic and population genetic dynamics which we parameterize using historical demographic and epidemiological data. This interdisciplinary approach provides a means to assess three diseases that have been considered candidates: typhoid fever; SDC; and tuberculosis. By evaluating the upper bounds of mortality and selective impact for all three diseases, we conclude that only tuberculosis resistance would have been a sufficient source of selective pressure to result in modern CF incidences.

The isolation of the CF gene product, the CF transmembrane conductance regulator (CFTR) (Kerem *et al.* 1989; Riordan *et al.* 1989; Rommens *et al.* 1989), has allowed numerous experiments to explore molecular mechanisms for the hypothesized heterozygote advantage. The discovery that the secretory toxin of

*Author for correspondence (eric.poolman@yale.edu).

enterotoxigenic *Escherichia coli* (ETEC) acts through the CFTR increased the interest in the idea that secretory diarrhoea played a role in the emergence of CF (Chao *et al.* 1994; Cuthbert *et al.* 1994; Goldstein *et al.* 1994). A similar toxin is produced in cholera, which also causes secretory diarrhoea (Gabriel *et al.* 1994). Experimental results regarding a difference in response to secretory toxins between wild-type and heterozygous mice have been contradictory: one study found a difference (Gabriel *et al.* 1994), while another has shown no difference (Cuthbert *et al.* 1995). Moreover, results in a study of human heterozygotes showed no differences in chloride secretion (Hogenauer *et al.* 2000). No study has incorporated clinical outcomes such as hospitalizations. Thus, clinical evidence for resistance to SDC is at best inconclusive, and there is no evidence of resistance to diarrhoea without a secretory toxin (e.g. amoebic, rotaviral, other bacterial diarrhoea).

The causative agent of typhoid, *Salmonella enterica* serovar Typhi, requires wild-type CFTR to enter mouse gastrointestinal submucosa (Pier *et al.* 1998). Intact CFTR is markedly reduced in mice heterozygous for ΔF508, suggesting a role for CF mutations in preventing typhoid fever (Pier *et al.* 1998). However, a recent clinical study in Indonesia, an area highly endemic for typhoid fever, found no ΔF508 mutations in their sample of 775 individuals (van de Vosse *et al.* 2005). Thus, clinical evidence supporting a role of CFTR mutations in typhoid resistance is lacking.

Resistance to tuberculosis in CF patients was recently postulated to arise from diminished activity of the enzyme arylsulphatase B (Tobacman 2003). *Mycobacterium tuberculosis* lacks intrinsic arylsulphatase activity, but the bacterium has an arylsulphotransferase that may require sulphate freed by host arylsulphatase to construct its cell wall (Rivera-Marrero *et al.* 2002). Cell wall sulphate content is in turn correlated with virulence (Goren *et al.* 1976). Thus, dysfunction of arylsulphatase activity in CF is postulated to protect against tuberculosis by depriving the bacteria of a necessary nutrient (Tobacman 2003).

There is also clinical evidence of an association between CF and resistance to tuberculosis. A lower tuberculosis mortality rate has been observed among heterozygotes versus controls (Crawfurd 1972). Additionally, the low incidence of tuberculosis in CF patients, particularly in relation to non-tubercular mycobacteria, has been noted by a number of investigators (Anderson *et al.* 1967; Smith *et al.* 1984; Hjelte *et al.* 1990; Kilby *et al.* 1992). Thus, tuberculosis is supported by both molecular and clinical evidence.

While molecular and clinical facets of the hypotheses of heterozygote advantage for the CFTR locus have been repeatedly addressed, quantitative assessment based upon historical or epidemiological data has been neglected. Calculations of selection coefficients for heterozygote advantage in CF (Meindl 1987; Bertranpetit & Calafell 1996) have neither used historical data, nor compared particular selective agents within a population genetic framework. Each candidate that we examine here has been responsible for notable historic mortality (figure 1). However, in addition to the magnitude of overall mortality, the age
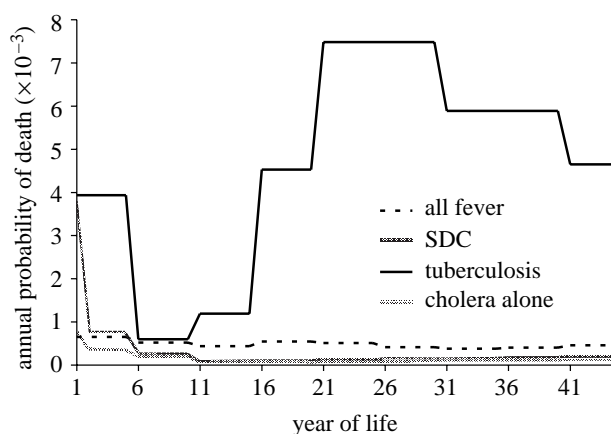

Figure 1. Age-specific probability of death owing to candidate diseases from historical records (Woods & Shelton 1997). Tuberculosis curve represents the peak of the European epidemic during 1600–1900, with 20% of all-cause mortality; the rates we used prior to 1600 were fractions of the same curve. SDC refers to secretory diarrhoea including cholera, and hence includes diarrhoea due to *E. coli*.

distribution of mortality as well as human demography are important determinants of selection (Samuelson 1978; Crow 1979; Galvani & Slatkin 2003; Galvani & Slatkin 2004).

Cystic fibrosis follows a well-known geographical pattern, with an incidence in newborns of European descent of approximately 1 in 2500 (Welsh *et al.* 2001), lower incidences in the Middle East of 1 in 15 000 (Frossard *et al.* 1998), down to 1 in 40 000 in Indian populations (Powers *et al.* 1996) and vanishingly small incidences in African and East Asian populations (Yamashiro *et al.* 1997; Welsh *et al.* 2001). This is in stark contrast to geographical clines for typhoid fever and cholera, which are more common in tropical regions (Barua 1992; Crump *et al.* 2004). The historical–geographical distribution of tuberculosis more closely matches that of CF, i.e. while tuberculosis has been globally endemic for at least 15 000 years (Morse *et al.* 1964; Kapur *et al.* 1994; Rothschild *et al.* 2001; Hughes *et al.* 2002), a tuberculosis epidemic of record mortality broke out in Europe in the early seventeenth century (Twelfth Annual Report 1854; Grigg 1958; Heberden & Percival 1973; Bates & Stead 1993; Kipple 1993; Davis 2000). The epidemic did not spread to India, Africa and other parts of the world until the late nineteenth century (Bates & Stead 1993). Our model incorporates this historical outbreak and tests the correspondence of the historical–geographical course of the tuberculosis epidemic to the modern distribution of CF.

2. MATERIAL AND METHODS

We used British mortality records from 1861 to 1870 to determine age-structured mortality rates owing to SDC (Woods & Shelton 1997). The original registers separately tabulated deaths from 'diarrhoea and dysentery' and 'cholera'. We multiplied the mortality rate of the former by a parameter, ε, representing the proportion of diarrhoea that is secretory, so as to include only diarrhoea for which molecular evidence of

CF resistance exists. A modern estimate of 18.5% for ε comes from Bangladesh, an area highly endemic for secretory diarrhoea (Albert *et al.* 1999). For our sensitivity analysis, we varied ε from half to twice this value, producing an estimate of 2% of deaths owing to SDC. As contemporary mortality rates are higher in urban environments, and the population became increasingly urbanized, we probably overestimated earlier mortality owing to SDC. It is debatable whether true cholera existed in Europe prior to 1817 (Barua 1992); we conservatively allow that it did and apply the record of 1861–1870 cholera mortality across all time.

For typhoid fever, a recent study showed that clinical diagnosis underestimates disease incidence among infants and children (Sinha *et al.* 1999). Therefore, we used recorded typhoid fever mortality as the lower bound in our sensitivity analysis (Woods & Shelton 1997). For our upper bound, we used the mortality rate from all fevers (Creighton 1965). As this includes death to typhus—which itself was often blamed for more than half of deaths owing to fever—and other fevers, it is necessarily an overestimate of death to typhoid fever. We used the higher mortality rate in determining the age-structured mortality, as it showed more deaths among the young; the lower figures produced on an average 39% of this mortality, or approximately 2% of all-cause mortality. As with SDC, higher and increasing rates of typhoid fever in urban areas suggest that earlier times suffered less mortality from this disease than that we modelled.

For tuberculosis, numerous sources ascribe to it 20–25% of all deaths from the sixteenth century to the early twentieth century, with peak death rates in excess of 750 per 100 000 annually (Twelfth Annual Report 1854; Grigg 1958; Heberden & Percival 1973; Bates & Stead 1993; Kipple 1993; Davis 2000). We used the lower figure of 20% as an average and the shorter term of 1600–1900. We conservatively assigned zero mortality to tuberculosis after 1900. Prior to 1600, we assigned tuberculosis responsibility for a variable proportion, $\tau_{\mathrm{euro}} < 20\%$, of all deaths. We additionally modelled resistance to tuberculosis in India, where the modern epidemic started only several hundred years later (Bates & Stead 1993). For India, we used a peak share of mortality of 20% from 1800 to 1950 to reflect the shift in the start of the epidemic there. To construct the age distribution of deaths, we used figures from 1849 to 1853 in the United States (1854). The age-specific mortality rates used for the model are shown for comparison in figure 1.

We use an age-structured population to incorporate the different shapes of mortality curves from the candidate diseases, with $n_{t,a,z}$ females of age a and genotype z in the population at time t, and with $n_{t,a}$ total females across all genotypes (Crow 1979; Galvani & Slatkin 2003; Galvani & Slatkin 2004). Females bear female children at age-specific rate m_a and suffer age- and time-specific all-cause mortality at rate $\mu_{t,a}$. The mortality from the candidate disease is $d_{t,a}$. The wild-type homozygote experiences the full mortality, whereas the heterozygote experiences mortality reduced by their fractional resistance, ρ. For initial modelling of all three candidate diseases, we assume that heterozygotes enjoy

perfect resistance to the candidate diseases, $\rho = 1$. The CF homozygote individuals experience 100% mortality prior to reproduction, which is consistent with the life-history curves for such patients even up to the beginning of the last century (Warwick & Monson 1967). Baseline mortality and fecundity rates were taken from data derived from English parish records from 1550 to 1900 (Wrigley & Schofield 1981). For our sensitivity analysis, we varied mean age at first childbirth from 23.4 to 26.0, the range for the period of 1600–1849 (Wrigley & Schofield 1981). The stable population distribution used to initiate model realizations was constructed from these rates (Preston *et al.* 2001).

Baseline mortality rates from the source life tables must necessarily include deaths owing to endemic diseases. For each disease, we adjusted the baseline mortality downwards from μ to μ' so as not to double-count deaths,

$$\mu'_{t,a} = 1 - \frac{1 - \mu_{t,a}}{1 - d_{t,a}}. \tag{2.1}$$

Using these estimates for fecundity, baseline mortality and disease mortality, we evolve the population according to Hardy–Weinberg ratios, with $p_{t,a}$ representing the frequency of the wild-type allele at time t in age group a, and assuming that each female chooses a partner from a pool with allele frequency identical to that of her age group. We use 45 age classes of 1 year, consistent with the age-specific fertility rates (Wrigley & Schofield 1981). We ignore post-reproductive adults. The number of newly born wild-type homozygous (wt), heterozygous (het) and CF homozygous (cf) individuals is governed by the following difference equations:

$$\left.\begin{aligned}
p_{t,a} &= \frac{2n_{t,a,\mathrm{wt}} + n_{t,a,\mathrm{het}}}{2n_{t,a}}, \\
n_{t+1,1,\mathrm{wt}} &= \sum_{a=1}^{45} p_{t,a}^2 n_{t,a} m_a, \\
n_{t+1,1,\mathrm{het}} &= \sum_{a=1}^{45} 2p_{t,a}(1 - p_{t,a}) n_{t,a} m_a, \\
n_{t+1,1,\mathrm{cf}} &= \sum_{a=1}^{45} (1 - p_{t,a})^2 n_{t,a} m_a.
\end{aligned}\right\} \tag{2.2}$$

And the existing population ages according to

$$\left.\begin{aligned}
n_{t+1,a+1,\mathrm{wt}} &= (1 - \mu'_{t,a})(1 - d_{t,a}) n_{t,a,\mathrm{wt}}, \\
n_{t+1,a+1,\mathrm{het}} &= (1 - \mu'_{t,a})(1 - (1 - \rho)d_{t,a}) n_{t,a,\mathrm{het}}, \\
n_{t+1,2,\mathrm{cf}} &= 0.
\end{aligned}\right\} \tag{2.3}$$

To address the hypotheses that ΔF508's high prevalence means it alone bestows heterozygote advantage, and that non-ΔF508 mutations are merely detrimental, we formulated an extended version of the above model with three alleles (e.g. wild-type, ΔF508 and non-ΔF508 mutation) and the six resultant genotypes. We then modelled a population in which a single ΔF508 mutation is introduced into the population with an equilibrium level of non-ΔF508 mutations (Wirth *et al.* 1997; Freeman & Herron 2004), assuming the former is protective against

tuberculosis in the heterozygote and that both the mutant alleles produce CF in the homozygotes and mixed heterozygote. We tracked the ratio of ΔF508 to non-ΔF508 mutations over time, assuming that non-ΔF508 mutations continue to arise *de novo* at a rate of 6.7×10^{-7} (Wirth *et al.* 1997; Freeman & Herron 2004).

Our model of a freely mixing population facilitates the increase of an overdominant lethal recessive; hence, if any disease is to provide sufficient selective pressure, it must do so under these conditions (Nishino & Tajima 2005). Moreover, the well-studied ΔF508 mutation has a unique origin (Morral *et al.* 1994), indicating pan-European mixing of CFTR alleles. Each model realization follows the change in allele frequencies in a single region; thus, differences across Europe can be due to differences in selective pressure (e.g. force of infection of candidate selective agent.)

Target CF incidence in Europe, i_{euro}, was 1 in 3000 births (Welsh *et al.* 2001). We used the CF incidence from a study of Indian immigrants to the United Kingdoms, 1 in 40 000 births, for the target outside of Europe, i_{asia}, as India experiences all three candidate diseases (Powers *et al.* 1996). Historic continuation of childbearing into the fifth decade of life produced mean generation times in our model from 31.9 to 32.8 years (Preston *et al.* 2001).

The model was initiated from 50 000 years before the present (yrBP) and run through to equilibria for both SDC and typhoid fever, affording these diseases maximal opportunity for selection. Model realizations for tuberculosis were initiated from a more conservative 18 000 yrBP, consistent with an intermediate estimate of the age of the most common CF mutation (Morris *et al.* 2002). Moreover, low endemic levels of tuberculosis mortality were applied until AD 1600, with the higher epidemic level from 1600 to 1900, and no tuberculosis mortality after AD 1900. This is consistent with both the estimate of 35 000 years since the origin of *M. tuberculosis* (Hughes *et al.* 2002) and historical records of the tuberculosis pandemic's origin in Europe.

3. RESULTS

Neither SDC nor typhoid fever is able to produce the observed European incidence, given perfect resistance in the heterozygote and record disease mortality extended over all time. Perfect resistance to SDC, $\rho_{dia} = 1$, produces an equilibrium CF incidence of 1 in 10 800 live births (1 in 20 100 to 1 in 4700 from sensitivity analysis), significantly below observed frequencies of greater than 1 in 3000. Perfect resistance to typhoid fever, $\rho_{typh} = 1$, produces an equilibrium incidence of 1 in 7600 live births (1 in 29 200 to 1 in 4100 from sensitivity analysis), again considerably below observed modern incidences (figure 2). At this incidence, the mortality owing to CF balances the mortality prevented owing to putative heterozygote advantage.

The tuberculosis pandemic originating in Europe in the seventeenth century provides adequate, appropriately located selective pressure to produce the modern CF incidence in Europe, greater than 1 in 3000, assuming that endemic tuberculosis prior to the
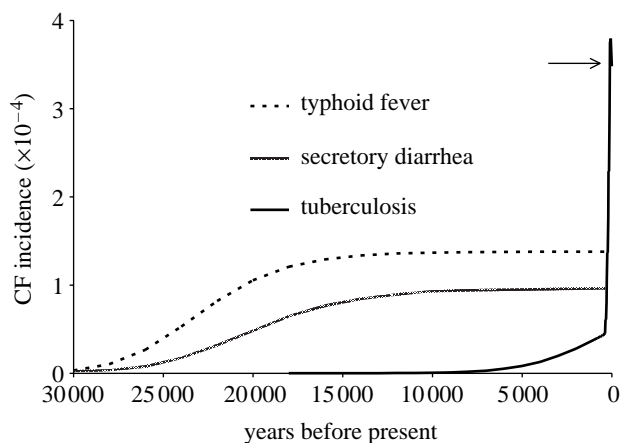

Figure 2. Model realizations of CF incidence over time for candidate diseases, given 100% resistance ($\rho = 1$) to the specified disease. Tuberculosis has been assigned responsibility for a fraction, $\tau = 1.6\%$, of mortality prior to 1600. All the populations are initiated with 1 in 20 000 CF allele frequencies. Typhoid fever and SDC were initiated 50 000 yrBP. Tuberculosis was given the more restrictive start time of 18 000 yrBP. Under these conditions, tuberculosis resistance produces an observed modern European CF incidence of 1 in 3000 (arrow). The inflection point in the tuberculosis curve is due to the conservative assumption that tuberculosis rates jumped to their historically documented levels in 1600, rather than increasing gradually prior to 1600.

seventeenth century was responsible for less than 2% of all-cause mortality, which is consistent with historical records, and only one-tenth of its mortality during the recent European epidemic (Bates & Stead 1993). Partial resistance to tuberculosis in the heterozygote, down to $\rho_{TB} = 15\%$, can similarly produce the modern CF incidence, given higher levels of endemic tuberculosis prior to the seventeenth century. Thus, tuberculosis provides sufficient selective pressure over a wide range of resistances and mortality levels (figure 3).

We additionally modelled the CF incidence in India. If tuberculosis mortality in India had been approximately 70% of that in Europe, it would produce the observed Indian CF incidence. This lower level of tuberculosis in India is historically plausible (Bates & Stead 1993).

When we modelled two types of CFTR mutations, those with and without heterozygote advantage, the advantageous category of alleles rises to represent 98% of CFTR mutant alleles in the modern day. This rise occurs despite *de novo* disadvantageous mutations at a rate of 6.7×10^{-7} per generation (Wirth *et al.* 1997; Freeman & Herron 2004). This projected 98% contribution is significantly higher than the 66% modern frequency of ΔF508 among CFTR mutations (Bobadilla *et al.* 2002), suggesting that the heterozygote advantage in CF is not associated solely with ΔF508.

Tuberculosis remains a major cause of mortality worldwide (Nunn *et al.* 2005), but from a historical perspective, it is relatively well controlled in most developed countries. In situations where tuberculosis is no longer responsible for significant mortality among individuals of reproductive age, our model demonstrates that the incidence of CF will fall. The decline
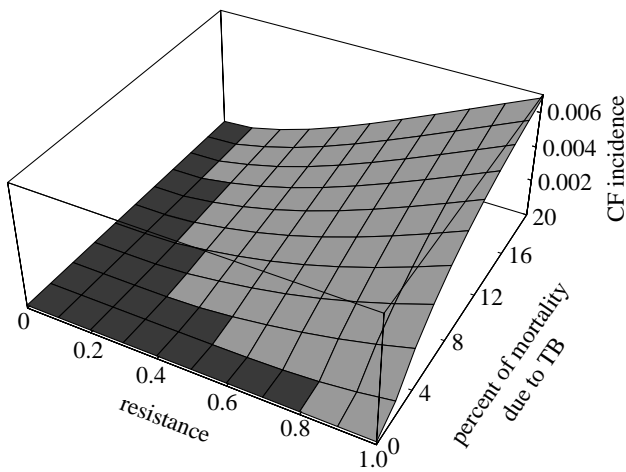
Figure 3. Range of CF incidences given resistance to tuberculosis in the heterozygote. The independent variables are fraction of resistance to tuberculosis in the CF carrier and percent of all deaths owing to tuberculosis prior to the pandemic that began in the 1600s. From 1600 until 1900, 20% of all deaths are ascribed to tuberculosis, as it is consistent with historical records. Light grey regions indicate combinations of parameters that could have produced the observed CF incidence (greater than 1 in 3000) in Europe. Tuberculosis resistance is a sufficient cause of CF across a broad range of parameters. In contrast, neither SDC nor typhoid fever produces sufficient CF incidence in any region of their parameter space, even allowing peak mortality levels extending back indefinitely into the past—their entire graphs are dark grey (not shown).

amounts to a 0.1% decrease in CF incidence annually over the next 100 years. To reduce the incidence by half requires approximately 20 generations.

4. DISCUSSION

The tuberculosis pandemic, the 'white plague' that claimed over 20% of all European lives beginning in the 1600s, swept across the world over subsequent centuries (Bates & Stead 1993). However, its origin and duration in Europe meant that the greatest selective pressure was applied there—enough so that, according to our results, tuberculosis resistance in the CF heterozygote can explain the modern gradient in CF incidence between Europe and the rest of the world.

Our quantitative analysis revealed that secretory diarrhoea mortality—including both cholera and ETEC—was insufficient to account for the present CF frequencies in Europe, given perfect resistance over all time and liberal estimates of disease-associated mortality. Likewise, typhoid fever did not provide adequate mortality under similar assumptions. We were generous to these diseases in numerous aspects of our evaluations. Firstly, our estimates of the incidences of cholera, typhoid and ETEC are likely to be high; as all these diseases tend to increase with increasing population density, the urbanizing communities of the 1800s from which the mortality estimates were drawn would suffer the highest mortality rates in history from these diseases (Woods & Shelton 1997). In the case of typhoid fever, we broadly included all deaths by fever in our parameterization to avoid difficulties in retrospective

diagnosis. Secondly, we used a high estimate for the mean age at first childbirth; lower estimates predict even lower CF incidences owing to these diseases, as the differential resistance in heterozygotes has less time over which to act. Thirdly, many regions in Europe have CF incidences higher than our target incidence, up to 1 in 1700 (Welsh *et al.* 2001); hence, we set a lenient standard for determining whether typhoid fever or SDC could produce European incidences. In contrast, tuberculosis is able to match even the higher target incidences (data not shown). Finally, neither typhoid fever nor SDC shows a historical or geographical distribution consistent with the modern distribution of CF.

Abundant literature attests to the ancient, persistent, worldwide presence of tuberculosis (Morse *et al.* 1964; Davis 2000; Rothschild *et al.* 2001). A tuberculosis contribution of approximately 2% of all mortality during endemic periods is adequate in our model and higher contributions are consistent if resistance is imperfect. Tuberculosis can also explain geographical differences in CF incidence, i.e. tuberculosis produced the observed CF incidence for India with tuberculosis mortality levels about 70% of that in Europe, consistent with the historical data (Bates & Stead 1993). Comparisons to Africa and Asia are more difficult, primarily because clear data on the incidence of CF are unavailable. Furthermore, other polymorphisms that lead to incomplete forms of CF, e.g. congenital bilateral absence of the vas deferens or chronic pancreatitis (Sharer *et al.* 1998), are common in parts of Asia and complicate the analysis there (Cuppens *et al.* 1998; Fujiki *et al.* 2004). While these and other factors may contribute to the difference in CF incidence between Europe and elsewhere, our results show that the timing and duration of the most recent tuberculosis pandemic provides a parsimonious explanation.

Tuberculosis supplies greater selective pressure than either typhoid fever or SDC primarily because it has been associated with mortality rates routinely of an order of magnitude higher than those owing to the other candidate diseases and secondarily because mortality closer to the onset of reproductive age removes relatively more reproductive potential per death (figure 1).

A founder effect has been proposed to explain the high incidence of CF (Klinger 1983). We modelled a putative bottleneck population consisting solely of heterozygotes. In such a population, CF incidence begins at 1 in 4 births and decays below the observed European incidence in 1800 years. Thus, a founder effect could not produce existing levels of CF, as there has been no such dramatic bottleneck for the entire European population within the past 1800 years.

The particular prevalence of the $\Delta F508$ mutation has led to the suggestion that it alone provides heterozygote advantage (Welsh *et al.* 2001; Wiuf 2001) and that other CFTR mutations are purely detrimental. However, our results demonstrate that selective pressure sufficient to drive CF up to its present incidence would also drive the class of purely detrimental CFTR alleles down to less than 2% of all CFTR mutations. This is significantly lower than the 34% of all CFTR mutations that are not $\Delta F508$ (Bobadilla *et al.* 2002), suggesting that $\Delta F508$ is neither the sole beneficial CFTR allele nor a genetic

hitchhiker (Macek *et al.* 1997). Additionally, the historical presence of tuberculosis provides a ready explanation for the range of ages suggested for ΔF508 (Wiuf 2001; Morris *et al.* 2002); it would have provided modest selective advantage against the endemic disease even prior to the seventeenth century epidemic.

As our model relates CF to tuberculosis, it predicts that the incidence of CF will fall in regions of the world where tuberculosis is well controlled at a rate of approximately 0.1% per year. In parts of the world where tuberculosis incidence is increasing (as owing to HIV and limited public health resources), the equilibrium incidence of CF may be higher or lower than present levels; prediction in such regions is outside the scope of this paper.

We were only able to use fertility and mortality data from one portion of Europe, and that assembled and compared across multiple sources. Moreover, we had to extrapolate relatively recent demographic parameters back in time to earlier periods with markedly different living conditions. Retrospective diagnosis of the causes of death is notoriously error-prone, but we took a generous approach of ascribing deaths to diarrhoea and typhoid fever. Even if our exact parameters are not correct, our basic finding holds: tuberculosis resistance is sufficient to explain CF across a wide range of parameters, while SDC and typhoid fever appear much more limited in the range of feasible parameters (in our case, none was found consistent with the historical data).

Our model assumes random mixing across the entire population. It is also deterministic rather than stochastic. Further, we include neither spatial nor temporal heterogeneity in the force of infection of the candidate diseases. All these factors could conspire to raise or lower the incidence of CF within a community, irrespective of selective pressures. However, we believe that the magnitude of these effects is insufficient to produce the broad incidence of CF across so many European populations.

We demonstrate that the two agents commonly invoked to explain the persistence of CF—secretory diarrhoea (including cholera) and typhoid fever—have not imposed sufficient historical selection. In contrast, tuberculosis combines ancient endemicity with a recent epidemic of European origin and severe mortality, particularly among young adults at the peak of their reproductive potential. Consequently, tuberculosis would provide sufficient selective pressure in the appropriate parts of the world. Additionally, tuberculosis is a plausible candidate based upon clinical observations and recent molecular hypotheses. Thus, investigation of the putative link between tuberculosis and CF may be an important area for inquiry, both through genetic studies of tuberculosis patients and examination of the proposed role of arylsulphatase B deficiency in both tuberculosis and CF.

REFERENCES

Albert, M. J., Faruque, A. S., Faruque, S. M., Sack, R. B. & Mahalanabis, D. 1999 Case-control study of enteropathogens associated with childhood diarrhea in Dhaka, Bangladesh. *J. Clin. Microbiol.* **37**, 3458–3464.

Anderson, C. M., Allan, J. & Johansen, P. G. 1967 Comments on the possible existence and nature of a heterozygote advantage in cystic fibrosis. *Bibl. Paediatr.* **86**, 381–387.

Barua, D. 1992 History of Cholera. In *Cholera* (ed. D. Barua & W. Greenough), pp. 1–36. New York, NY: Plenum Medical Book Company.

Bates, J. H. & Stead, W. W. 1993 The history of tuberculosis as a global epidemic. In *The medical clinics of North America: tuberculosis*, vol. 77 (ed. J. B. Bass), pp. 1205–1217. Harcourt Brace & Company: Philadelphia.

Bertranpetit, J. & Calafell, F. 1996 Genetic and geographical variability in cystic fibrosis: evolutionary considerations. *Ciba Found. Symp.* **197**, 97–114 discussion 114–8

Bobadilla, J. L., Macek Jr, M., Fine, J. P. & Farrell, P. M. 2002 Cystic fibrosis: a worldwide analysis of CFTR mutations—correlation with incidence data and application to screening. *Hum. Mutat.* **19**, 575–606. (doi:10. 1002/humu.10041)

Chao, A. C., de Sauvage, F. J., Dong, Y. J., Wagner, J. A., Goeddel, D. V. & Gardner, P. 1994 Activation of intestinal CFTR Cl-channel by heat-stable enterotoxin and guanylin via cAMP-dependent protein kinase. *EMBO J.* **13**, 1065–1072.

Crawfurd, M. A., 1972 A genetic study, including evidence for heterosis, of cystic fibrosis of the pancreas. In *168th Meeting of the Genetical Society of Great Britain*, pp. 126. University of Leeds.

Creighton, C. 1965 *A history of epidemics in Britain.* London, UK: Cass.

Crow, J. F. 1979 Gene frequency and fitness change in an age-structured population. *Ann. Hum. Genet.* **42**, 355–370.

Crump, J. A., Luby, S. P. & Mintz, E. D. 2004 The global burden of typhoid fever. *Bull. World Health Organ.* **82**, 346–353.

Cuppens, H. *et al.* 1998 Polyvariant mutant cystic fibrosis transmembrane conductance regulator genes. The polymorphic (Tg)m locus explains the partial penetrance of the T5 polymorphism as a disease mutation. *J. Clin. Invest.* **101**, 487–496.

Cuthbert, A. W., Halstead, J., Ratcliff, R., Colledge, W. H. & Evans, M. J. 1995 The genetic advantage hypothesis in cystic fibrosis heterozygotes: a murine study. *J. Physiol.* **482**, (Pt 2) 449–454.

Cuthbert, A. W., Hickman, M. E., MacVinish, L. J., Evans, M. J., Colledge, W. H., Ratcliff, R., Seale, P. W. & Humphrey, P. P. 1994 Chloride secretion in response to guanylin in colonic epithelial from normal and transgenic cystic fibrosis mice. *Br. J. Pharmacol.* **112**, 31–36.

Cystic Fibrosis Genetic Analysis Consortium 2005 *Cystic Fibrosis Mutation Database.* See http://www.genet. sickkids.on.ca/cftr/.

Davis, A. L. 2000 History of Tuberculosis. In *Tuberculosis: a comprehensive international approach* (ed. L. B. Reichman & E. S. Hershfield), pp. 3–54. New York, NY: Marcel Dekker, Inc.

Freeman, S. & Herron, J. C. 2004 *Evolutionary analysis.* Upper Saddle River, NJ: Pearson Prentice Hall.

Frossard, P. M., Girodon, E., Dawson, K. P., Ghanem, N., Plassa, F., Lestringant, G. G. & Goossens, M. 1998 Identification of cystic fibrosis mutations in the United Arab Emirates. Mutations in brief no. 133. Online. *Hum. Mutat.* **11**, 412–413. (doi:10.1002/(SICI)1098-1004(1998) 11:5<412::AID-HUMU15>3.0.CO;2-O)

Fujiki, K. *et al.* 2004 Genetic evidence for CFTR dysfunction in Japanese: background for chronic pancreatitis. *J. Med. Genet.* **41**, e55. (doi:10.1136/jmg.2003.014456)

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

Galvani, A. P. & Slatkin, M. 2003 Evaluating plague and smallpox as historical selective pressures for the CCR5-Delta 32 HIV-resistance allele. *Proc. Natl Acad. Sci. USA* **100**, 15 276–15 279. (doi:10.1073/pnas.2435085100)

Galvani, A. P. & Slatkin, M. 2004 Intense selection in an age-structured population. *Proc. R. Soc. B* **271**, 171–176. (doi:10.1098/rspb.2003.2573)

Goldstein, J. L., Sahi, J., Bhuva, M., Layden, T. J. & Rao, M. C. 1994 *Escherichia coli* heat-stable enterotoxin-mediated colonic Cl-secretion is absent in cystic fibrosis. *Gastroenterology* **107**, 950–956.

Goren, M. B., D'Arcy Hart, P., Young, M. R. & Armstrong, J. A. 1976 Prevention of phagosome–lysosome fusion in cultured macrophages by sulfatides of *Mycobacterium tuberculosis*. *Proc. Natl Acad. Sci. USA* **73**, 2510–2514. (doi:10.1073/pnas.73.7.2510)

Grigg, E. R. 1958 The arcana of tuberculosis with a brief epidemiologic history of the disease in the U.S.A. *Am. Rev. Tuberc.* **78**, 151–172.

Hansson, G. C. 1988 Cystic fibrosis and chloride-secreting diarrhoea. *Nature* **333**, 711. (doi:10.1038/333711c0)

Heberden, W. & Percival, T. 1973 *Population and disease in early industrial England. Pioneers of demography.* Farnborough, UK: Gregg.

Hjelte, L., Petrini, B., Kallenius, G. & Strandvik, B. 1990 Prospective study of mycobacterial infections in patients with cystic fibrosis. *Thorax* **45**, 397–400.

Hogenauer, C., Santa Ana, C. A., Porter, J. L., Millard, M., Gelfand, A., Rosenblatt, R. L., Prestidge, C. B. & Fordtran, J. S. 2000 Active intestinal chloride secretion in human carriers of cystic fibrosis mutations: an evaluation of the hypothesis that heterozygotes have subnormal active intestinal chloride secretion. *Am. J. Hum. Genet.* **67**, 1422–1427. (doi:10.1086/316911)

Hughes, A. L., Friedman, R. & Murray, M. 2002 Genomewide pattern of synonymous nucleotide substitution in two complete genomes of *Mycobacterium tuberculosis*. *Emerg. Infect. Dis.* **8**, 1342–1346.

Kapur, V., Whittam, T. S. & Musser, J. M. 1994 Is *Mycobacterium tuberculosis* 15 000 years old? *J. Infect. Dis.* **170**, 1348–1349.

Kerem, B., 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.

Kilby, J. M., Gilligan, P. H., Yankaskas, J. R., Highsmith Jr, W. E., Edwards, L. J. & Knowles, M. R. 1992 Nontuberculous mycobacteria in adult patients with cystic fibrosis. *Chest* **102**, 70–75.

Kipple, K. 1993 *The Cambridge world history of human disease.* Cambridge, UK: Cambridge University Press.

Klinger, K. W. 1983 Cystic fibrosis in the Ohio Amish: gene frequency and founder effect. *Hum. Genet.* **65**, 94–98. (doi:10.1007/BF00286641)

Macek Jr, M. *et al.* 1997 Possible association of the allele status of the CS.7/HhaI polymorphism 5′ of the CFTR gene with postnatal female survival. *Hum. Genet.* **99**, 565–572. (doi:10.1007/s004390050407)

Meindl, R. S. 1987 Hypothesis: a selective advantage for cystic fibrosis heterozygotes. *Am. J. Phys. Anthropol.* **74**, 39–45. (doi:10.1002/ajpa.1330740104)

Morral, N. *et al.* 1994 The origin of the major cystic fibrosis mutation (delta F508) in European populations. *Nat. Genet.* **7**, 169–175. (doi:10.1038/ng0694-169)

Morris, A. P., Whittaker, J. C. & Balding, D. J. 2002 Fine-scale mapping of disease loci via shattered coalescent modeling of genealogies. *Am. J. Hum. Genet.* **70**, 686–707. (doi:10.1086/339271)

Morse, D., Brothwell, D. R. & Ucko, P. J. 1964 Tuberculosis in Ancient Egypt. *Am. Rev. Respir. Dis.* **90**, 524–541.

Nishino, J. & Tajima, F. 2005 Effect of population structure on the amount of polymorphism and the fixation probability under overdominant selection. *Genes Genet. Syst.* **80**, 287–295. (doi:10.1266/ggs.80.287)

Nunn, P., Williams, B., Floyd, K., Dye, C., Elzinga, G. & Raviglione, M. 2005 Tuberculosis control in the era of HIV. *Nat. Rev. Immunol.* **5**, 819–826. (doi:10.1038/nri1704)

Pier, G. B., Grout, M., Zaidi, T., Meluleni, G., Mueschenborn, S. S., Banting, G., Ratcliff, R., Evans, M. J. & Colledge, W. H. 1998 *Salmonella typhi* uses CFTR to enter intestinal epithelial cells. *Nature* **393**, 79–82. (doi:10.1038/30006)

Powers, C. A., Potter, E. M., Wessel, H. U. & Lloyd-Still, J. D. 1996 Cystic fibrosis in Asian Indians. *Arch. Pediatr. Adolesc. Med.* **150**, 554–555.

Preston, S. H., Heuveline, P. & Guillot, M. 2001 *Demography: measuring and modeling population processes.* Oxford, UK: Blackwell Publishers Ltd.

Riordan, J. R. *et al.* 1989 Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. *Science* **245**, 1066–1073.

Rivera-Marrero, C. A., Ritzenthaler, J. D., Newburn, S. A., Roman, J. & Cummings, R. D. 2002 Molecular cloning and expression of a novel glycolipid sulfotransferase in *Mycobacterium tuberculosis*. *Microbiology* **148**, 783–792.

Romeo, G., Devoto, M. & Galietta, L. J. 1989 Why is the cystic fibrosis gene so frequent? *Hum. Genet.* **84**, 1–5. (doi:10.1007/BF00210660)

Rommens, J. M. *et al.* 1989 Identification of the cystic fibrosis gene: chromosome walking and jumping. *Science* **245**, 1059–1065.

Rothschild, B. M., Martin, L. D., Lev, G., Bercovier, H., Bar-Gal, G. K., Greenblatt, C., Donoghue, H., Spigelman, M. & Brittain, D. 2001 *Mycobacterium tuberculosis* complex DNA from an extinct bison dated 17 000 years before the present. *Clin. Infect. Dis.* **33**, 305–311. (doi:10.1086/321886)

Samuelson, P. A. 1978 Generalizing Fisher's "reproductive value": overlapping and nonoverlapping generations with competing genotypes. *Proc. Natl Acad. Sci. USA* **75**, 4062–4066. (doi:10.1073/pnas.75.8.4062)

Sereth, H., Shoshani, T., Bashan, N. & Kerem, B. S. 1993 Extended haplotype analysis of cystic fibrosis mutations and its implications for the selective advantage hypothesis. *Hum. Genet.* **92**, 289–295. (doi:10.1007/BF00244474)

Serre, J. L., Simon-Bouy, B., Mornet, E., Jaume-Roig, B., Balassopoulou, A., Schwartz, M., Taillandier, A., Boue, J. & Boue, A. 1990 Studies of RFLP closely linked to the cystic fibrosis locus throughout Europe lead to new considerations in populations genetics. *Hum. Genet.* **84**, 449–454. (doi:10.1007/BF00195818)

Sharer, N., Schwarz, M., Malone, G., Howarth, A., Painter, J., Super, M. & Braganza, J. 1998 Mutations of the cystic fibrosis gene in patients with chronic pancreatitis. *N. Engl. J. Med.* **339**, 645–652. (doi:10.1056/NEJM199809033391001)

Sinha, A. *et al.* 1999 Typhoid fever in children aged less than 5 years. *Lancet* **354**, 734–737. (doi:10.1016/S0140-6736(98)09001-1)

Smith, M. J., Efthimiou, J., Hodson, M. E. & Batten, J. C. 1984 Mycobacterial isolations in young adults with cystic fibrosis. *Thorax* **39**, 369–375.

Tobacman, J. K. 2003 Does deficiency of arylsulfatase B have a role in cystic fibrosis? *Chest* **123**, 2130–2139. (doi:10.1378/chest.123.6.2130)

Twelfth Annual Report 1854 Twelfth annual report relating to the registry of births, marriages, and deaths in Massachusetts. In *From consumption to tuberculosis: a documentary history* (ed. B. Gutmann Rosenkrantz). New York, NY: Garland.

van de Vosse, E., Ali, S., Visser, A. W., Surjadi, C., Widjaja, S., Vollaard, A. M. & Dissel, J. T. 2005 Susceptibility to typhoid fever is associated with a polymorphism in the cystic fibrosis transmembrane conductance regulator (CFTR). *Hum. Genet.*, 1–3.

Warwick, W. J. & Monson, S. 1967 Life table studies of mortality. *Bibl. Paediatr.* **86**, 353–367.

Welsh, M. J., Ramsey, B. W., Accurso, F. & Cutting, G. R. 2001 Cystic fibrosis. In *The metabolic and molecular bases of inherited disease* (ed. C. R. Scriver, A. L. Beaudet, D. Valle, W. S. Sly, B. Vogelstein, B. Childs & K. W. Kinzler). New York, NY: McGraw-Hill.

Wirth, B., Schmidt, T., Hahnen, E., Rudnik-Schoneborn, S., Krawczak, M., Muller-Myhsok, B., Schonling, J. & Zerres, K. 1997 *De novo* rearrangements found in 2% of index patients with spinal muscular atrophy: mutational mechanisms, parental origin, mutation rate, and implications for genetic counseling. *Am. J. Hum. Genet.* **61**, 1102–1111. (doi:10.1086/301608)

Wiuf, C. 2001 Do delta F508 heterozygotes have a selective advantage? *Genet. Res.* **78**, 41–47. (doi:10.1017/S0016672301005195)

Woods, R. & Shelton, N. 1997 *An atlas of Victorian mortality.* Liverpool, UK: Liverpool University Press.

Wrigley, E. A. & Schofield, R. S. 1981 *The population history of England 1541–1871.* Cambridge, MA: Harvard University Press.

Yamashiro, Y., Shimizu, T., Oguchi, S., Shioya, T., Nagata, S. & Ohtsuka, Y. 1997 The estimated incidence of cystic fibrosis in Japan. *J. Pediatr. Gastroenterol. Nutr.* **24**, 544–547. (doi:10.1097/00005176-199705000-00010)