# On the allelic spectrum of human disease

David E. Reich and Eric S. Lander

**Human disease genes show enormous variation in their allelic spectra; that is, in the number and population frequency of the disease-predisposing alleles at the loci. For some genes, there are a few predominant disease alleles. For others, there is a wide range of disease alleles, each relatively rare. The allelic spectrum is important: disease genes with only a few deleterious alleles can be more readily identified and are more amenable to clinical testing. Here, we weave together strands from the human mutation and population genetics literature to provide a framework for understanding and predicting the allelic spectra of disease genes. The theory does a reasonable job for diseases where the genetic etiology is well understood. It also has bearing on the Common Disease/Common Variants (CD/CV) hypothesis, predicting that at loci where the total frequency of disease alleles is not too small, disease loci will have relatively simple spectra.**

**David E. Reich***
**Eric S. Lander‡**
The Whitehead Institute/MIT Center for Genome Research, Nine Cambidge Center, Cambridge, MA 02142, USA.
*e-mail: reich@genome.wi.mit.edu
‡e-mail: lander@wi.mit.edu.

For human disease genes, the number and frequency of the individual disease-predisposing alleles (the allelic spectrum) varies enormously from locus to locus. Rare diseases are often caused by a panoply of different mutations, with each mutation constituting only a small fraction of the total class of disease alleles in the population. Some disorders, such as cystic fibrosis, are associated with a few relatively high-frequency alleles set against a background of many very rare alleles[1]. Others, such as β-thalassemia[2] or early-onset breast cancer[3,4], are diverse in the general population, but have only a few genetic causes in some isolated groups. Much less is known about the allelic spectrum for genes underlying common disorders such as diabetes, coronary artery disease or asthma.

The Common Disease/Common Variant (CD/CV) hypothesis, proposed several years ago[5–7] predicts that the genetic risk for common diseases will often be due to disease-predisposing alleles with relatively high frequencies – that is, there will be one or a few predominating disease alleles at each of the major underlying disease loci. There is currently not enough empirical evidence to either prove or disprove the CD/CV hypothesis. However, a few prototypical examples of such common variants are known, including the APOE ε4 allele in Alzheimer's disease[8], Factor V$^{Leiden}$ in deep venous thrombosis[9], and PPARγ Pro12Ala in type II diabetes[10].

The allelic spectrum of disease in general, and the CD/CV hypothesis in particular, has important consequences for both research and clinical practice. Population-based methods of gene discovery, such as association and linkage disequilibrium studies, depend on the existence of a relatively simple allelic spectrum in the study population[11]. Clinical characterization and design of diagnostic and therapeutic interventions are also substantially easier when the allelic spectrum is simple.

The purpose of this paper is to explore the following questions: Why is there the wide range of allelic spectra in human disease genes? To what extent is the spectrum of a disease predictable from its genetic properties? Weaving together strands from the human mutation and population genetics literature, we provide a framework for predicting allelic spectra that applies equally well to monogenic and polygenic disease.

Classical population genetics theory has focused primarily on populations of constant size, for which there is a strong prediction (see below) that allelic diversity will be similar for common and for rare diseases[12,13]. In real human populations, by contrast, rare diseases tend to have a much more diverse spectrum than is predicted by the constant-size theory (Table 1), and the deviation indicates that demography must be important in determining allelic diversity[14,15]. Genetic and archaeological evidence suggests that the human population experienced a dramatic expansion from a small, ancestral population, with an effective size on the order of 10 000 individuals, to the modern, large population size, in an epic growth that began 18 000–150 000 years ago[16–18]. To understand the modern-day allelic spectrum, we need to understand the diversity of disease-causing alleles in the ancestral population, and how it changed after the rapid expansion.

## A simple model

We outline a simple model for making predictions about disease allele diversity.

### *Population history*

We shall assume that the ancestral population was freely interbreeding and constant in size at $N = 10\,000$ until it expanded nearly instantaneously to the modern size of $N = 6 \times 10^9$. Of course, the human population expansion was not instantaneous. However, it turns out that our results are largely insensitive to the details of the model provided that the expansion was reasonably rapid and large (see Appendix).

### *Disease-susceptibility locus*

We consider a single disease locus that contributes to a monogenic or polygenic disease. The locus has disease alleles $d_1$, $d_2$, $d_3$, etc., assumed to be selectively equivalent. Let $f$ be the total frequency of the set of disease alleles in the current population; that is, the sum of the frequencies of the individual alleles. We are interested to know whether the disease alleles at the locus have a simple spectrum

 0168–9525/01/$ – see front matter  PII: S0168-9525(01)02410-6

**Table 1. Selected genetic disorders and their allelic spectra**

| Disorder (gene) (inheritance pattern)[a] | Lifetime incidence | Population | *F* | Percent of disease class due to most common allele | Observed $\varphi$[b] | Predicted $\varphi$ (predicted half-life of disease class in years)[c] | Comments | Refs |
|---|---|---|---|---|---|---|---|---|
| Retinoblastoma (*RB1*) (AD) | 1/20 000 | US | 1/40 000 | | < 0.01 | 0 (38) | $\mu \approx 1/90\,000$ | 31,32 |
| Aniridia (*PAX6*) (AD) | 1/100 000 | US | 1/200 000 | | < 0.01 | 0 (51) | $\mu \approx 1/600\,000$ | 33 |
| Tuberous sclerosis (*TSC2*) (AD) | 1/10 000 | US | 1/40 000 | | < 0.01 | 0 (23) | $\mu \approx 1/55\,000$ Can also occur due to *TSC1* mutations. | 34 |
| Familial hypercholesterolemia (*LDL*) (AD) | 1/500<br>1/67 | UK<br>S. African Jews | 1/1000<br>1/135 | 12%<br>80% | 0.02<br>0.62 | 0 (5000) | Heterozygotes have a milder phenotype. | 35–37 |
| Duchenne muscular dystrophy (Dystrophin) (XLR) | 1/4200 males | Netherlands | 1/4200 | | < 0.01 | 0 (49) | $\mu \approx 1/12\,000$ | 38 |
| Hemophilia A (*F8C*) (XLR) | 1/5000 males | Germany | 1/5000 | | < 0.01 | 0 (61) | $\mu \approx 1/18\,000$ | 20,39 |
| G6PD deficiency (*G6PD*) (XLR) | 1/9 males | US Blacks | 1/9 | 90% | 0.81 | 0.77 (660 000) | Mutations may confer resistance to malaria. | 40–44 |
| | 1/28 males | South China | 1/28 | 50% | 0.3 | 0.53 (200 000) | Different alleles are predominant in different populations. | |
| | 1/33 males | Greece | 1/33 | 79% | 0.62 | 0.48 (170 000) | | |
| | 1/25 males | Mexico | 1/25 | 70% | 0.55 | 0.56 (220 000) | | |
| Cystic fibrosis (*CFTR*) (AR) | 1/2000 | Europe-wide | 1/45 | 67% | 0.45 | 0.38 (120 000) | Europeans, with ΔF508 mutation, have highest incidence. | 1,20 |
| Gaucher disease (β-glucosidase) (AR) | 1/50 000<br>1/850 | US non-Jews<br>Ashkenazi Jews | 1/220<br>1/30 | 30%<br>76% | 0.18<br>0.6 | 0.01 (24 000) | | 45 |
| Tay-Sachs disease(*HEX A*) (AR) | 1/110 000 | US non-Jews | 1/335 | 16% | 0.03 | 0 (16 000) | | |
| | 1/3600 | Ashkenazi Jews | 1/60 | 73% | 0.56 | | | |
| Phenylketonuria (PAH) (AR) | 1/10 000 | Germany | 1/100 | 26% | 0.1 | 0.13 (54 000) | Different alleles are predominant in the UK and China. | 46–49 |
| | 1/16 500 | China | 1/128 | 20% | 0.09 | 0.08 (42 000) | | |
| Wilson Disease (*ATP7B*) (AR) | 1/30 000 | UK | 1/173 | 17% | 0.04 | 0.03 (31 000) | | 46,50 |
| | 1/7000 | Sardinia | 1/84 | 61% | 0.39 | | | |
| Hemochromatosis (*HLA-H*) (AR) | 1/400 | US | 1/20 | 85% | 0.7 | 0.62 (280 000) | | 51 |
| β-thalassemia (β-globin) (AR) | 1/900 | Delhi, India | 1/30 | 35% | 0.17 | 0.51 (180 000) | Mutations may confer resistance to malaria. | 52–54 |
| | 1/1600 | Lebanon | 1/40 | 40% | 0.2 | 0.42 (140 000) | These populations generally have different alleles in their spectra, indicating heterogeneous origins. | |
| | 1/200 | Sardinia | 1/14 | 95% | 0.9 | | | |

[a]Abbreviations for inheritance patterns: AD, autosomal dominant; XLR, X-linked recessive; AR, autosomal recessive.

[b]For some diseases, the expected allelic identity ($\varphi$) is calculated on the basis of an incomplete allelic spectrum; that is, there are some mutations causing the disease at the gene that have failed to be identified. However, these missed alleles are in general quite rare and should only slightly raise the estimate of $\varphi$ (which is calculated as the sum of the squared allele frequencies).

[c]The predicted half-life of the disease class (in years) and the expected allelic identity $\varphi$ are obtained from Equations 3 and 4 for the outbred populations (calculations are not made for the Sardinian and Jewish populations, which have experienced recent bottlenecks followed by expansions). $\mu$ is assumed to be $3.2 \times 10^{-6}$ per generation unless it can be estimated directly, $f_0$ is estimated as $f$, and a generation is assumed to be 25 years. For the prediction of $\varphi$ we assume that the ancestral human population size was stable at $N_0 = 10{,}000$ and that the expansion to a stable size of 6 billion occurred 3000 generations ago. These assumptions, though simplistic, result in surprisingly good predictions of disease allele diversity in outbred human populations (Fig. 2).

or a diverse spectrum. A simple allelic spectrum means that a handful of disease alleles account for a large proportion of the overall set of disease alleles (the disease class); a diverse allelic spectrum means that there are many alleles with none accounting for a large proportion of the overall set of disease alleles. Let $f_0$ be the equilibrium frequency for the class of disease alleles – that is, the frequency

expected under the balance between mutation and selection. Classical population genetics provides formulae for $f_0$ in the case of autosomal dominant, recessive and X-linked diseases (see below), in terms of the mutation rate, $\mu$ (of wild-type alleles into the disease alleles) and the selection coefficient, $s$, (the reduced reproductive fitness owing to inheritance of a disease allele[19]). Also, let $f_{exp}$ be the frequency for the class of disease alleles just before the population expansion.

*Measuring allelic diversity*

A simple measure of allelic diversity is the probability that two randomly chosen alleles are identical. This is called the 'expected allelic identity' (denoted $\varphi$), and its reciprocal is the 'effective number of alleles' (denoted $n$). A classic population genetic formula states that $\varphi$ at a neutral locus (not subject to selection) in a population that is freely mixing and stable (of constant size for a long period) should be:

$$\text{Expected } \varphi \text{ at a neutral locus} = \frac{1}{1+4N\mu} \quad (1)$$

where $N$ is the effective population size and $\mu$ is the mutation rate at the locus[19].

What is less well known is that a similar formula applies to describe the diversity of disease alleles at a locus (provided that they are equivalent from the standpoint of selection – if not, the theory needs to be applied separately to each selectively equivalent class[12]). If the size of a disease class does not fluctuate far from its equilibrium $f_0$, the expected allelic identity in a freely mixing, stable population is (see Appendix and Refs 12,13):

$$\text{Expected } \varphi_{disease} \text{ (that is, } \varphi \text{ among disease alleles)} \approx \frac{1}{1+4N\mu(1-f_0)} \quad (2)$$

This is now the probability that two alleles *within* the disease class are the same, with $N$ still denoting the effective overall population size and $\mu$ now the probability that a non-disease allele will mutate into a disease allele. (Mutation rates for genes almost always fall in the range $10^{-7}$–$10^{-4}$ and usually in the range $10^{-6}$ to$10^{-5}$ per generation[20]. Below, we will typically use the geometric mean: $\mu = 3.2 \times 10^{-6}$ per generation.)

The CD/CV hypothesis can thus be rephrased as the prediction that $\varphi_{disease}$ is high for the disease loci responsible for most of the population risk for common diseases.

**A tale of two loci**

Consider two hypothetical monogenic disorders with the same underlying mutation rate, but different overall frequency of disease alleles in the population: a rare disease with $f_0 = 0.001$ and a common disease with a much larger $f_0 = 0.2$. As noted above, the frequency of the disease class, $f_0$, is determined by the balance between mutation and selection. In our example, the mutation rates will be assumed to be equal at $\mu = 3.2 \times 10^{-6}$ per generation. Hence, the frequency difference must reflect different degrees of selection. The rare disease might be subject to intense selection (for example, it might be reproductive lethal), whereas the common disease might be subject to only mild selection (as might occur with type II diabetes or hypertension, which act later in life) or even be associated with a heterozygote advantage. For simplicity at the beginning of our discussion, we consider the case where selection pressures have been constant over time and the class of disease alleles has remained fixed at its equilibrium size ($f_{exp} = f = f_0$).

*In the ancestral population (N = 10 000), all disease loci had a simple spectrum*

With a 'typical' mutation rate of $\mu = 3.2 \times 10^{-6}$ per generation, the effective number of disease alleles should have been about $n = 1.1$ for both the rare and common diseases, because $1 - f_0$ is close to 1 in both cases (Equation 2). Hence, both rare and common diseases should have had simple allelic spectra in the ancient population, consisting of a single predominant disease-causing allele accounting for 90% or more of the disease class.

*In a modern-sized population at equilibrium, all disease loci should have a complex spectrum*

Now, consider the extreme case of a nearly instantaneous population expansion, as discussed above. The effective number of alleles, $n$, should remain at about 1.1 immediately after the expansion, because the initial effect of a rapid expansion is merely to increase (amplify) the number of chromosomes (both disease causing and normal) without giving rise to new alleles. However, as new mutations occur and accumulate, the diversity would be expected to increase to a much higher level, reaching the new 'mutation-drift equilibrium' predicted by Equation 2. The effective number of alleles should grow to about 77 000 for the rare disease and 61 000 for the common disease.

*Kinetics differ for rare and common diseases*

The initial and final states are thus approximately the same for the two diseases, with low diversity in the initial population and high diversity in the final large population. But there is a crucial difference in the kinetics of the process (Fig. 1). The increase in allelic diversity will occur rapidly for the rare disease, but will require millions of years after the expansion for the common disease. This is easy to explain by focusing either on the loss of old disease alleles by selection or the gain of new ones by mutation (at equilibrium, these two processes balance). Selection against the rare disease class is more intense than against the common disease class, explaining the more rapid turnover. Equivalently from the mutation perspective, the rate of newly

arising disease mutations per generation is about the same for the rare and common disease, but the pool of disease chromosomes they enter is much smaller for the rare disease, resulting in a greater proportional effect on that class per generation and an increased turnover rate.

These arguments can be made quantitatively precise. The number of alleles expected to arise through new mutations per generation is $2N(1-f_0)\mu$, the product of the number of normal alleles existing in the population and the mutation rate. Thus, a fraction $2N(1-f_0)\mu/2Nf_0 = (1-f_0)\mu/f_0$ of the alleles in the disease class are expected to be replaced every generation: 0.3% for the rare disease, and 0.0013% for the common diseases. The proportion of original alleles that are expected to remain after $t$ generations, therefore, follows a simple exponential decay: $e^{-(1-f_0)\mu t/f_0}$. The half-life of allelic replacement, the time that should elapse before the ancestral disease class is half-replaced by the modern spectrum, is then $(\ln 2)f_0/\mu(1-f_0)$ generations. With 25 years per generation and $1-f_0 \approx 1$:

$$\text{Half-life of ancestral replacement (in years)} \approx \frac{17 f_0}{\mu(1-f_0)} \approx 17 f_0/\mu \qquad (3)$$

The half-life for the ancestral allelic spectrum is thus about 5000 years for the rare disease and 1.3 million years for the common disease. Figure 1a shows the proportion of ancestral alleles remaining for various values of $f_0$ as a function of the time since the expansion. The ancestral allelic spectrum clearly decays rapidly for rare diseases, but persists for a long time for more common diseases. Figure 1b shows how much time is expected to elapse before disease classes of original size $f_0$ decay to half of their original size. Figure 1c shows the effective number of alleles in the population at various times following the expansion. Although the ancestral spectrum decays rapidly, it nonetheless requires a very long time to establish the new equilibrium level of allelic diversity for a population of $6 \times 10^9$. The graph for a small final population size (e.g. 20 million) would be nearly the same, except that it would reach an asymptote at a smaller effective number of alleles ($n$). Whatever the case, the new equilibrium will have been approached only for extremely rare diseases.

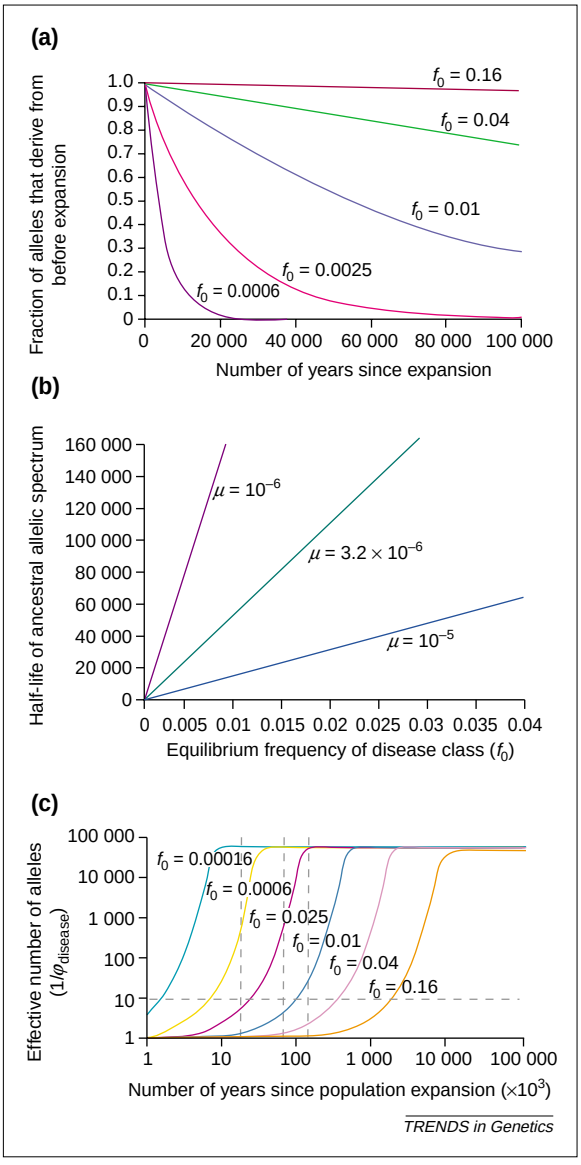


**Fig. 1.** The kinetics of change in allelic diversity depends on $f_0$, the overall fraction of the alleles at the locus that predispose to the disease. We assume a sudden expansion from 10 000 to $6 \times 10^9$ individuals and mutation rate $\mu = 3.2 \times 10^{-6}$ per generation. (a) The kinetics of the disappearance of the ancestral disease class, which is the driving force behind the increase in diversity after an expansion. (b) The expected time until the ancestral disease class decays to half of its original value (Equation 3). (c) The change in the effective number of alleles (based on Equation 4) as a function of time. The time range for human population expansion (18 000–150 000 years ago) and a specific estimate of 75 000 years ago are indicated by the vertical dashed lines. A horizontal line indicates the cutoff for a 'simple' allelic spectrum, $1/\varphi_{disease} = 10$, as described in the text.

### Implications for disease mapping

The success of an association study to identify a disease-susceptibility locus depends on the detection of an increased frequency of specific disease alleles in affected individuals. This requires that the locus have a relatively simple allelic spectrum: that is, a few predominant alleles. The analysis above shows that these conditions should hold – and thus association studies should be feasible – for loci at which the total frequency $f$ of disease alleles is above some threshold.

The threshold depends somewhat on the time of the large population expansion and on the degree of allelic complexity that is tolerable for gene mapping. We shall assume that the expansion occurred about 75 000 years ago (estimates range from 18 000 to 150 000 years[16–18], and our estimate is therefore not likely to be too far off). We shall also assume that a 'simple' allelic spectrum is one where $1/\varphi_{disease} < 10$, which corresponds to ancestral alleles comprising

30% or more of the modern spectrum of disease-causing alleles. The threshold is thus about $f = 0.9\%$ (from Appendix, Equation 4, or by examining Fig. 1c). For roundness, we will use $f = 1\%$ as a reasonable threshold.

For populations that experienced more recent expansions (following founding bottlenecks[21]), or loci with lower-than-average mutation rates, the threshold should be even lower.

**Two caveats**

Our simple model predicts that for disease loci with large $f_0$ and typical values of $\mu$, modern allelic spectra should be simple because of a slow decay of the ancestral disease class. We now introduce two caveats due to oversimplifications of the model. The first is a mechanism that could yield a more diverse spectrum than expected. The second is an additional mechanism by which a simpler spectrum could arise.

*Fluctuations in allele frequency at the time of expansion*

We assumed above that the allele frequency at the time of expansion was equal to the equilibrium value, $f_{exp} = f_0$. However, genetic drift can result in fluctuations away from the equilibrium value. It is straightforward to adapt Equation 3 to show that the time until half of the allelic spectrum has decayed to modern alleles is in this case multiplied by a factor of approximately $C = \log_2(1+f_{exp}/f_0)$ (see Appendix for details). When $f_{exp} = f_0$, the factor $C = 1$. In the extreme case that genetic drift transiently eliminated all disease alleles, we have $f_{exp} = 0$ and the factor $C = 0$. The half-time for their 'decay' is 0 in this case because there are no ancestral alleles. (In general, for $f_{exp}$ far below $f_0$, the modern allelic spectrum should be more diverse than expected by our theory, because an unusually large fraction of alleles today are descended from ones that arose after population expansion). On average, excursions below $f_0$ have a greater proportional impact on the half-time than similar excursions above $f_0$, leading to a downward bias in half-time in the presence of a fluctuating disease class size.

Although loci with low values of $f_{exp}$ will have more complex modern allelic spectra than expected (based on $f_0$), these loci will also tend to have lower-than-expected frequency $f$ in the modern population and might therefore make a diminished contribution to the risk of disease. The lower-than-expected modern frequency is due to the fact that in the approximately 3000 generations (75 000 years assuming 25 years per generation) that have passed since expansion, not enough time has elapsed to repopulate an under-populated class of disease alleles. (Assuming $f_{exp} = 0\%$ and $\mu = 3.2 \times 10^{-6}$ per generation as the maximum filling rate for the disease class, 3000 generations is just barely enough time to repopulate the disease alleles for a locus with $f_0 = 1\%$ and too short for a locus with much larger $f_0$.)

*Population substructure, changing selection*

The human population is not panmictic (randomly interbreeding), but rather has substructure. In particular, population bottlenecks (such as might have occurred during the great human migrations) have the potential to greatly reduce disease allele frequency, with the most striking effect on small disease classes (see Appendix). This mechanism could generate simpler allelic spectra, and it could also account for why there could be different predominant alleles in different populations (owing to different alleles surviving passage through various bottlenecks).

This effect can be exacerbated by changes in selection over time. To illustrate the point, consider an isolated population having a simple allelic spectrum owing to a bottleneck. Suppose that environmental change affecting the population (e.g. the appearance of a pathogen) causes disease alleles to gain a strong heterozygote advantage and leads to a tenfold increase in $f$. The common disease allele in the isolated population will then attain substantial frequency. Subsequent gene flow can then cause this allele to dominate the allelic spectrum in neighboring populations. Such a mechanism might explain how certain alleles in the Ashkenazi Jewish population came to dominate and subsequently spread, leading to a simpler allelic spectrum in surrounding non-Jewish populations (see below).

**Real diseases**

How well do these insights explain the data for various monogenic disorders?

*Rare autosomal dominant and X-linked diseases*

For autosomal dominant and X-linked monogenic diseases, the total frequency $f_0$ of disease-causing alleles in the population tends to be extremely low because the alleles are constantly subject to intense selection. Classical population genetics predicts a total frequency $f_0 \approx \mu/s$ for a dominant disease and $f_0 \approx 3\mu/s$ for an X-linked recessive disease. Because of intense selection, these diseases usually have tiny disease classes and consequently extremely diverse allelic spectra, with most alleles being lost rapidly to selection and new alleles replenishing the disease class up to the level of mutation–selection balance. Examples (Table 1) include aniridia, retinoblastoma, tuberous sclerosis, and Duchenne Muscular Dystrophy (DMD), with the last being among the most common such disorder with $f \approx 1/4000$.

Some autosomal dominant and X-linked disorders are not under extreme negative selection, and, as a result, have larger disease classes and more homogenous allelic spectra in the modern population. Glucose-6-phosphate dehydrogenase (G6PD) deficiency, for example, has a large disease class ($f \approx 11\%$ in African Americans) and a simple

spectrum[22]. Its prevalence is due to its mild phenotype – hemolytic anemia only among people who have certain environmental exposures – and the fact that it can even confer a selective advantage in malaria-endemic areas.

*'Rare' autosomal recessive diseases*

Autosomal recessive diseases often have much larger disease classes than autosomal dominant and X-linked recessive diseases because selection acts only on homozygotes; classical population genetics[19] predicts that the proportion of alleles in the disease class should be $f_0 \approx \sqrt{\mu/s}$. However, the absolute sizes of the disease classes are still in general expected to be relatively small, with $f_0$ typically between 0.001–0.01 for autosomal recessive lethal diseases. This would imply a half-life of the disease class of 5000–50 000 years (using Equation 3 and $\mu = 3.2 \times 10^{-6}$ per generation). Our prediction is therefore that the allelic spectra should be moderately diverse for most autosomal recessive diseases, but simpler if the disease class is large or the mutation rate is low. Table 1 and Fig. 2 show that empirical data tend to fit these theoretical expectations.

Wilson's disease, for example, has a very diverse allelic spectrum, with the most common mutation in Britain accounting for only 17% of the allelic spectrum, and an experimentally observed allelic identity, $\varphi$, of approximately 0.04. Phenylketonuria is similar: in Germany and China $\varphi$ is about 0.1 and 0.09, respectively, with (interestingly) almost completely different disease mutations in the two populations corresponding to mutations having arisen in high frequency recently, after the historical divergence of the two populations (Table 1). This difference in mutation could be due to historical population contractions that caused different alleles to rise to high frequency in the ancestral populations of the Germans and Chinese, respectively.

Gaucher and Tay–Sachs diseases are autosomal recessive diseases in which the simple version of the theory breaks down due to the coupled effects of population bottlenecks and population mixing. These diseases are relatively prevalent in Ashkenazi Jews, where population bottlenecks have fairly recently driven up their frequencies. However, even among non-Jewish Americans[23,24], where the diseases are more rare, the spectra are simpler than predicted by our theory, probably because of gene flow from Ashkenazis to the other groups (Table 1, Fig. 2).

In general, autosomal recessive diseases can have simple or complex allelic spectra. Their disease classes sizes from 0.001–0.01 are so close to the range where simple allelic spectra are expected (Fig. 1b) that in the case of an unusually low $\mu$, or a recent history of founding bottlenecks, simple allelic spectra can easily arise.

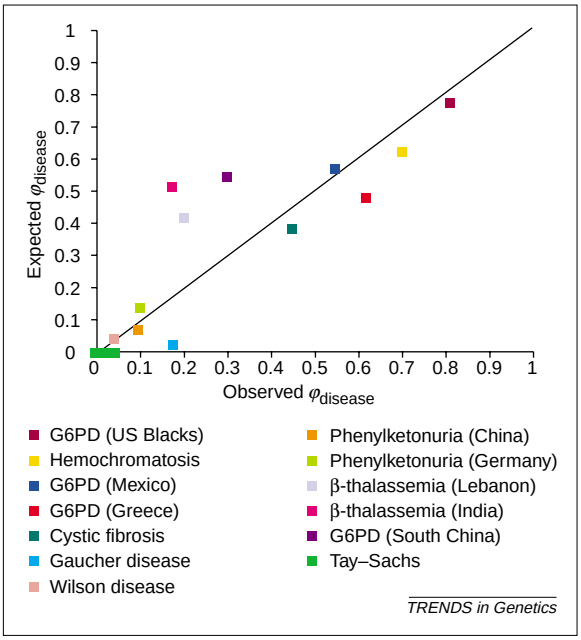


**Fig. 2.** Reality versus theory. This figure, based on the data points specified in Table 1, compares the observed allelic identity $\varphi$ based on the actual spectrum of disease alleles in an outbred population, to the predicted value of $\varphi$ based on theory (Equation 4). We assume $N_0 = 10\,000$, $N_1 = 6 \times 10^9$, and $t = 3000$ generations. Unless a direct estimate of the mutation rate is available, we assume $\mu = 3.2 \times 10^{-5}$ per generation. Although these model parameters are only rough estimates, the high correlation between observed and predicted $\varphi$ shown in the figure suggests that the theory does a reasonable job of predicting the degree of allele diversity among disease-predisposing alleles. The following diseases are not shown as they were too near the origin to be discerned: Retinoblastoma, Aniridia, Tuberous sclerosis, Duchenne muscular dystrophy, Hemophilia A, Familial hypercholestrolemia.

*'Common' autosomal recessive diseases*

Some recessive diseases could become common because there is only mild selection against disease alleles or because heterozygotes enjoy a selective advantage. Cystic fibrosis is a prototypical example, with disease alleles constituting a fraction $f = 2.3\%$ of the total population of all cystic fibrosis alleles transmembrane conductance regulator (*CFTR*) in individuals of European descent. It has been suggested that the large disease class is due to resistance to *Salmonella typhi* among heterozygous individuals[25]. Suppose that, whatever the reason, the size of the disease class was similarly large in the ancestral population. In that case, the simple ancestral allelic spectrum would likely have a half-life for replacement of 180 000 years using the 'typical' value of $\mu = 3.2 \times 10^{-6}$ (39 000–390 000 years, if we allow $\mu$ to vary from $10^{-5}$ to $10^{-6}$ per generation). Our theory thus predicts that the ancestral alleles should predominate in the modern spectrum, tens of thousands of years after the great population expansions, which agrees well with the observation of a single allele accounting for 67% of cystic fibrosis alleles in the Caucasian population. The most recent common ancestor for the mutant allele has been estimated by haplotype studies to

be at least 52 000 years old[26], consistent with an origin before the great population expansions. A similar analysis helps explain the simple allelic spectrum of hemochromatosis, which has a disease class of size $f = 0.05$ and a single allele accounting for 85% of the spectrum.

*Special cases*

Sickle-cell anemia is an example of a disease in which the allelic spectrum is simple because only a single mutation, in this case the Glu6Val mutation in β-globin, can give rise to the disease. Achondroplasia is another special case because it is usually due to recurrent mutation, in this case at the hypervariable nucleotide 1138 in the *FGFR3* gene[27]. Mutations occur at this site at a rate of once approximately every $10^5$ meioses (about three orders of magnitude faster than at a typical nucleotide). The allelic spectrum is not as diverse as predicted by the model because this mutation predominates.

G6PD deficiency and β-thalassemia are also special cases. The high prevalence of these diseases in Sub-Saharan Africa, the Mediterranean and East Asia is thought to be due primarily to the fact that these populations have been exposed to malaria recently, rather than that the disease classes have been historically large. Mutations in these genes confer resistance to the malaria pathogen, and probably rose to high frequency only after the advent of malaria, subsequent to the historical divergence of the populations – as confirmed by the fact that different alleles predominate in each population. The simple spectra now observed at these loci are not likely to be due to historically large disease classes, but instead could be a result of the alternative mechanism of generating allelic simplicity – bottlenecks followed by selection for disease alleles – described above.

Table 1 provides a more complete list of rare and common genetic diseases, indicating that the theory is not too bad at predicting allelic diversity (see also Fig. 2).

## Discussion

The simple theory above aims to predict the allelic spectrum in the human population for a class of selectively equivalent alleles at a single locus, as a function of the overall frequency $f_0$ of the class[14,15,28]. At equilibrium, the allelic variation should be almost independent of $f_0$. However, the human population is far from equilibrium, as it has been growing dramatically during the past 100 000 years. The consequences of this growth, in terms of genetic diversity, will require millions of years to play out fully. The approach to a modern diverse spectrum is expected to occur much more rapidly for smaller values of $f_0$, so that, at present, only a few thousand generations after the beginning of population growth, allelic diversity is much greater for rare than for common diseases.

The theory provides useful predictions of disease allele diversity for monogenic disorders, including rare autosomal dominant diseases, moderately rare recessive disorders and relatively common recessive diseases. In particular, it implies that the presence of predominant disease alleles for cystic fibrosis (ΔF508) and hemochromatosis (C282Y) require no special selective advantage of these alleles compared with other disease-causing alleles, but merely the accident that they were in the right place at the right time.

What does the theory suggest about the CD/CV hypothesis – the conjecture that the genes responsible for most of the risk of common diseases, such as hypertension, heart disease and asthma, have relatively simple allelic spectra (high values of $\varphi_{disease}$)? The CD/CV hypothesis would have important consequences for medical genetics, implying that the causes of diseases could be found by association studies using common gene variants.

We have shown that a high value of $\varphi_{disease}$ should occur if the total frequency of disease alleles is not too low, e.g. $f_0 > 0.9\%$ at loci responsible for most of the risk for the disease. Thus, we have reformulated the CD/CV hypothesis so that instead of focusing on the frequencies of individual disease alleles at a locus, it focuses on the total size of the disease class, and we have come up with a hard prediction that if the loci contributing to common disease have even moderate-sized disease classes, the allelic spectra should be simple.

We do not currently understand the genetic architecture of complex disease, because we lack empirical data. It is possible, in principle, that the risk for some common diseases is due to a very large number of loci, with each having a low frequency of disease-predisposing alleles. For example, a disease with 10% incidence in the population might reflect 100 independent monogenic diseases each with high penetrance and an incidence of 0.001. However, this would imply a much higher relative risk to family members than is actually seen for most complex diseases. In fact, the observed fall-off in the relative risk to family members suggests (but does not prove) that most of the risk could be attributable to a modest number of loci with a higher frequency of disease predisposing alleles[29], as is seen for APOE ε4 allele in Alzheimer's disease[8], Factor $V^{Leiden}$ in deep venous thrombosis[9] and PPARγ Pro12Ala in type II diabetes[10].

It will be interesting to explore the extent to which the population characteristics of known common diseases constrain the frequency of the class of disease-predisposing alleles at a locus accounting for a substantial proportion of disease risk.

Overall, our results lend support to the CD/CV hypothesis by showing that a high overall frequency of disease alleles implies a high frequency of some individual disease alleles. This provides encouragement

to the search for common alleles responsible for the important common human diseases.

## Appendix

*Gradual population growth*

The analysis above assumed that the human population expanded massively and suddenly, but of course the demographic details were more complicated. If the expansion were slower than we assumed, the growth in diversity would have been slower, although the effect turns out to be rather small as long as the rate of growth is moderate and the final population size is moderately large. We compared various scenarios to our simple assumption of an instantaneous 600 000-fold expansion, using both analytical formulae and computer simulations. The time for the ancestral allelic spectrum to decay to comprising only a third of all disease-predisposing alleles is no more than 10% greater (compared with the result for a massive, instantaneous expansion) provided that the expansion is at least 75-fold and that the doubling time for population growth substantially exceeds the predicted half-life for decay of the ancient allelic spectrum given in Equation 3. These conditions are quite mild unless the disease class is very small. Under a wide range of scenarios, it is therefore reasonable to predict allelic diversity based on modeling human history as a sudden, massive population expansion[30].

*Ancestral population might not have been constant in size*

We assumed that the ancestral population was freely mixing and of constant size long enough for $\varphi$ to be at equilibrium. In fact, the ancestral population might have fluctuated in size or had significant substructure. Size fluctuations would have had to be huge to substantially influence disease allele diversity: increasing the effective number of alleles from 1.1 to 2 requires increasing the population from approximately 10 000 to 100 000 individuals.

*Expression for the expected allelic identity φ (Equation 2)*

Equation 2 can be derived from Equation 1 by considering the set of selectively equivalent disease alleles to be a diploid population of (approximately) constant size $N' = Nf_0$, with an effective mutation rate of $\mu' = (1 - f_0)\mu/f_0$. The effective mutation rate[12] is obtained by dividing the number of new disease alleles arising per generation, $2N(1–f_0)\mu$, by the number of alleles already existing in the disease class, $2Nf_0$. Within the class of disease alleles, there is no selection favoring one allele over any other, and hence it is appropriate to apply Equation 1. Sawyer[13] derives much more extensive results along these lines, showing that the detailed distribution of allele frequencies within a disease class is also similar to the theoretical expectation for a neutral locus.

*The kinetics of change of diversity*

Suppose that a population has size $N(t)$ at generation $t$. The reduction in the expected allelic identity $\varphi$ due to new mutation in that generation is equal to the kinship coefficient in the last generation multiplied by the probability that either of the two alleles will be replaced by a new mutation entering the disease class: $\varphi \times 2 \times \mu(1 - f_0)/f_0$. The increase in $\varphi$ due to genetic drift is equal to the probability of no relatedness in the last generation, multiplied by the probability that two alleles will share a common parent in a disease class of size $2N(t)f_0$: $(1–\varphi)/2N(t)f_0$. Hence, the expected change of $\varphi$ per generation is:

$$\Delta\varphi = \frac{-2\mu(1-f_0)\varphi}{f_0} + \frac{1-\varphi}{2N(t)f_0}$$

Consider the case of a sudden expansion to size $N_1$. Replacing $\Delta\varphi$ with $d\varphi/dt$, and solving the differential equation for the initial condition $\varphi(0) = 1/(1 + 4N_0\mu(1 - f_0))$ (Equation 2) leads to the following expected allelic identity $t$ generations following the expansion:

$$\varphi(t) = \frac{1}{1+4N_1\mu(1-f_0)} + \left(\frac{1}{1+4N_0\mu(1-f_0)} - \frac{1}{1+4N_1\mu(1-f_0)}\right) e^{-t_{exp}\left(\frac{1+4N_1\mu(1-f_0)}{2N_1 f_0}\right)} \quad (4)$$

*Half-life of the ancestral disease class when $f_{exp} \neq f_0$*

Let $A(t)$ be the component of the disease class due to ancestral alleles, and $B(t)$ be the component of the disease class due to disease mutation that occurred after the expansion. The ancestral spectrum $A(t)$ decays exponentially from its original size at a rate determined by the selection coefficient: $f_{exp}e^{-(1-f_0)\mu t/f_0}$. $B(t)$ grows approximately as $f_0(1–e^{-(1-f_0)\mu t/f_0})$, reflecting the filling of the disease class with new mutations every generation until it asymptotically approaches its equilibrium value $f_0$. The expected time until the ancestral disease class decays to a fraction $Q$ of the total class of disease alleles is then $A(t)/(A(t) + B(t)) = Q$. With some algebra, this reduces to $t_Q = \ln(1+f_{exp}/f_0(1/Q\text{-}1)) \times f_0/\mu(1 - f_0)$ and $t_{½} = \ln(1+f_{exp}/f_0) \times f_0/\mu(1 - f_0)$. Equation 3 is thus multiplied by $\ln(1 + f_{exp}/f_0)/\ln 2 = \log_2(1+f_{exp}/f_0)$ to obtain the time until only half of the allelic spectrum is ancestral. Note that the decay of the ancestral disease class is not exactly exponential when $f_{exp} \neq f_0$; in particular, $t_{¼}$ is not twice $t_{½}$.

*The effect of a bottleneck on disease allele diversity*

A simple rule of thumb for defining a 'severe' bottleneck that will substantially reduce disease allele diversity can be derived as follows. The probability that two random alleles in a disease class share a common ancestor during a bottleneck – the inbreeding coefficient of the bottleneck – can be calculated by noting that at any generation during the bottleneck, the probability that two

**Acknowledgements**
We thank Haninah Levine for assistance with computer simulations, and Edward Byrne and Kirk Lohmueller for help in researching Table 1. We thank David Altshuler, Michele Cargill, David Goldstein and Joel Hirschhorn for comments and discussions.

alleles do *not* share a common parent is $1 - 1/2N_b f$. Hence, the probability that they share a common ancestor during any of the $t_b$ generations of the bottleneck is $F = 1-(1-1/2N_b f)^{t_b} \approx 1-e^{-t_b 2N_b f}$. Defining a severe bottleneck (arbitrarily) as one in which $F > 1/3$, then if a population falls to a size $N_b$ for a period of $t_b$ generations and re-expands rapidly, there will be reduction in diversity in the spectrum of disease alleles when $f < 1.3t_b/N_b$. For example, a population that drops to 1000 for 10 generations or 5000 for 50 generations will experience a dramatic decline in diversity at loci with $f < 0.013$.

**References**


1 Estivill, X. *et al.* (1997) Geographic distribution and regional origin of 272 cystic fibrosis mutations in European populations. *Hum. Mut.* 10, 135–154

2 Kazazian, H.H., Jr and Boehm, C.D. (1988) Molecular basis and prenatal diagnosis of β-thalassemia. *Blood* 72, 1107–1116

3 Roa, B.B. *et al.* (1996) Ashkenazi Jewish population frequencies for common mutations in *BRCA1* and *BRCA2*. *Nat. Genet.* 14, 185–187

4 Dunning, A.M. *et al.* (1997) Common BRCA1 variants and susceptibility to breast and ovarian cancer in the general population. *Hum. Mol. Genet.* 6, 285–289

5 Lander, E.S. (1996) The new genomics: global views of biology. *Science* 274, 536–539

6 Cargill, M. *et al.* (1999) Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat. Genet.* 22, 231–238

7 Chakravarti, A. (1999) Population genetics – making sense out of sequence. *Nat. Genet.* 22, 56–60

8 Corder, E.H. *et al.* (1993) Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science* 261, 921–923

9 Bertina, R.M. *et al.* (1994) Mutation in blood coagulation factor V associated with resistance to activated protein C. *Nature* 369, 64–67

10 Altshuler, D. *et al.* (2000) The common *PPARγ* Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. *Nat. Genet.* 26, 76–80

11 Risch, N. and Merikangas, K. (1996) The future of genetic studies of complex human diseases, *Science* 273, 1516–1517

12 Hartl, D.L. and Campbell, R.B. (1982) Allelic multiplicity in simple Mendelian disorders. *Am. J. Hum. Genet.* 34, 866–873

13 Sawyer, S. (1983) A stability property of the Ewens sampling formula. *J. Appl. Prob.* 20, 449–459

14 Thompson, E.A. and Neel, J.V. (1997) Allelic disequilibrium and allele frequency distribution as a function of social and demographic history. *Am. J. Hum. Genet.* 60, 197–204

15 Lange, K. and Fan, R.Z. (1997) Branching process models for mutant genes in nonstationary populations. *Theor. Popul. Biol.* 51, 118–133

16 Pritchard, J.K. *et al.* (1999) Population growth of human Y chromosomes: A study of Y chromosome microsatellites. *Mol. Biol. Evol.* 16, 1791–1798

17 Reich, D.E. and Goldstein, D.B. (1998) Genetic evidence for a Paleolithic human population expansion in Africa. *Proc. Natl. Acad. Sci. U. S. A.* 95, 8119–8123

18 Rogers, A.R. and Harpending, H. (1992) Population growth makes waves in the distribution of pairwise genetic differences. *Mol. Biol. Evol.* 9, 552–569

19 Crow, J.F. and Kimura, M. (1970) *An Introduction to Population Genetics Theory*, Harper and Row

20 Sakanaranarayanan, K. (1998) Ionizing radiation and genetic risks IX. Estimates of the frequencies of mendelian diseases and spontaneous mutation rates in human populations: a 1998 perspective. *Mut. Res.* 411, 129–178

21 Reich, D.E. *et al.* (2001). Linkage disequilibrium in the human genome. *Nature* 411, 199–204

22 Beutler, E. (1996) *G6PD*: Population genetics and clinical manifestations. *Blood Rev.* 10, 45–52

23 Balicki, D. and Beutler, E. (1995) Reviews in molecular medicine: Gaucher disease. *Medicine* (*Baltimore*), 74, 305–323

24 Paw, B.H. *et al.* (1990) Frequency of three *Hex A* mutant alleles among Jewish and non-Jewish carriers identified in a Tay-Sachs screening program. *Am. J. Hum. Genet.* 47, 698–705

25 Pier, G.B. *et al.* (1998) *Salmonella typhi* uses CFTR to enter intestinal epithelial cells. *Nature* 393, 79–82

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

27 Bellus, G.A. *et al.* (1995) Achondroplasia is defined by recurrent G380R mutations of FGFR3. *Am. J. Hum. Genet.* 56, 368–373

28 Pritchard, J.K. (2001) Are rare variants responsible for susceptibility to complex diseases? *Am. J. Hum. Genet.* 69, 124–137

29 Risch, N. (1990) Linkage strategies for genetically complex traits. I. Multilocus models. *Am. J. Hum. Genet.* 46, 222–228

30 Reich, D.E. *et al.* (1999) Statistical properties of two tests that use multilocus data sets to detect population expansion. *Mol. Biol. Evol.* 16, 453–466

31 Lohmann, D.R. *et al.* (1996) The spectrum of *RB1* germ-line mutations in hereditary retinoblastoma. *Am. J. Hum. Genet.* 58, 940–949

32 Shields, J.A. and Shields, C.L. (1992) *Intraocular tumors: A text and atlas.* W.B. Saunders

33 Chao, L-Y. *et al.* (2000) Mutation in the *PAX6* gene in twenty patients with aniridia. *Hum. Mut.* 15, 332–339

34 Niida, Y. *et al.* (1999) Analysis of both *TSC1* and *TSC2* for germline mutations in 126 unrelated patients with tuberous sclerosis. *Hum. Mut.* 14, 412–422

35 Goldstein, J.L. *et al.* (1995) Familial hypercholesterolemia. In *The Metabolic and Molecular Bases of Inherited Disease* (7th Edn), (Scriver, C.R *et al.*, eds), pp. 2863–2913 McGraw–Hill

36 Mandelshtam, M. *et al.* (1998) Prevalence of Lithuanian mutation among St. Petersburg Jews with familial hypercholesterolemia. *Hum. Mut.* 12, 255–258

37 Webb, J.C. *et al.* (1996) Characterization of mutations in the low density lipoprotein (LDL)-receptor gene in patients with homozygous familial hypercholesterolemia, and frequency of these mutations in FH patients in the United Kingdom. *J. Lipid Res.* 37, 368–381

38 van Essen, A.J. *et al.* (1992) Birth and population prevalence of Duchenne muscular dystrophy in the Netherlands. *Hum. Genet.* 88, 258–266

39 Tavassoli, K. *et al.* (1998) Molecular diagnostics of 15 Hemophilia A patients: Characterization of eight novel mutations in the Factor VIII gene, two of which result in exon skipping. *Hum. Mut.* 12, 301–303

40 Luzatto, L. and Mehta, A. (1995) Glucose 6-phosphate dehydrogenase deficiency. In *The Metabolic and Molecular Bases of Inherited Disease* (7th Edn), (Scriver, C.R. *et al.*, eds), pp. 4517–4553, McGraw–Hill

41 Medina, M.D. *et al.* (1997) Molecular genetics of glucose-6-phosphate dehydrogenase deficiency in Mexico. *Blood Cells Mol. Dis.* 23, 88–94

42 Xu, W. *et al.* (1995) Glucose-6 phosphate dehydrogenase mutations and haplotypes in various ethnic groups. *Blood* 85, 257–263

43 Kay, A.C. *et al.* (1992) The origin of glucose-6-phosphate-dehydrogenase (*G6PD*) polymorphisms in African-Americans. *Am. J. Hum. Genet.* 50, 394–398

44 Beutler, E. *et al.* (1989) Molecular heterogeneity of glucose-6-phosphate dehydrogenase $A^-$. *Blood* 74, 2550–2555

45 Cormand, B. *et al.* (1998) Molecular analysis and clinical findings in the Spanish Gaucher disease population: Putative haplotype of the N370S ancestral chromosome. *Hum. Mut.* 11, 295–305

46 Lo, W.H.Y. *et al.* (1993) Molecular basis of PKU in China. *Chinese Med. Sci. J.* 8, 180–185

47 Eisensmith, R.C. *et al.* (1992) Multiple origins of phenylketonuria in Europe. *Am. J. Hum. Genet.* 51, 1355–1365

48 Scriver, C.R. *et al.* (1995) The hyperphenylalaninemias. In *The Metabolic and Molecular Bases of Inherited Disease* (7th Edn), (Scriver, C.R. *et al.*, eds), pp. 1857–1895 McGraw–Hill

49 Curtis, D. *et al.* (1999) A study of Wilson disease mutations in Britain. *Hum. Mut.* 14, 304–311

50 Loudianos, G. *et al.* (1999) Molecular characterization of Wilson disease in the Sardinian population – evidence of a founder effect. *Hum. Mut.* 14, 294–303

51 Feder, J.N. *et al.* (1996) A novel MHC class I-like gene is mutated in patients with hereditary haemochromatosis. *Nat. Genet.* 13, 399–408

52 Zahed, L. *et al.* (1997) The spectrum of beta-thalassemia mutations in the Lebanon. *Hum. Hered.* 47, 241–249

53 Madan, N. *et al.* (1998) Beta-thalassemia mutations in northern India (Delhi). *Indian J. Med. Res.* 107, 134–141

54 Pirastu, M. *et al.* (1983) Prenatal diagnosis of β-thalassemia. *New Engl. J. Med.* 309, 284–287