A polygenic basis for late-onset disease

Alan Wright[1], Brian Charlesworth[2], Igor Rudan[3,4], Andrew Carothers[1] and Harry Campbell[3]

[1]MRC Human Genetics Unit, Western General Hospital, Crewe Road, Edinburgh EH4 2XU, UK
[2]Institute of Cell, Animal and Population Biology, King's Buildings, University of Edinburgh, Edinburgh EH9 3JT, UK
[3]Department of Public Health Sciences, Teviot Place, University of Edinburgh, Edinburgh EH8 9AG, UK
[4]Department of Epidemiology, University Medical School, Rockefellerova 4, Zagreb, Croatia

The biological basis of late-onset disease has been shaped by genetic factors subject to varying degrees of evolutionary constraint. Late-onset traits are not only more sensitive to environmental variation, owing to the breakdown of homeostatic mechanisms, but they also show higher levels of genetic variation than traits directly influencing reproductive fitness. The origin and nature of this variation suggests that current strategies are poorly suited to identifying genes involved in many complex diseases.

A major focus of current interest lies in the genetic variation underlying susceptibility to common, late-onset human diseases such as heart disease, diabetes and cancer. These diseases result from the cumulative breakdown of many quantitatively varying physiological systems over the course of decades of life. They are orders of magnitude more common than individual mendelian disorders and are typically more prevalent in post-reproductive life, which means that they may be less subject to SELECTIVE CONSTRAINTS (see Glossary). The precise mechanisms maintaining genetic variation in such traits are poorly understood, but three broad categories are identifiable [1]. First, variants that are deleterious in both early and later life, which are therefore efficiently screened by natural selection and held at low population frequencies. Second, variants that are selectively neutral in early life but show late deleterious effects; this means that they are subject only to weak selection and can reach higher frequencies. Third, variants that are favourable in early life but deleterious later on; these can be maintained by selection at intermediate frequencies. Strategies for identifying disease susceptibility genes depend both on the balance of common and rare variants maintained in the population, and on whether these occur at a limited (OLIGOGENIC) or a large (POLYGENIC) number of loci. In this article, evolutionary and population genetic arguments are used to examine these issues and to suggest that currently favoured strategies could be poorly suited to identifying disease susceptibility genes.

One strategy assumes that most disease susceptibility variants are common in the population (frequency >0.01) – the COMMON DISEASE/COMMON VARIANT (CD/CV) HYPOTHESIS [2]. This proposes that individuals with disease have an excess of common susceptibility alleles, and that these are potentially detectable in large-scale patient–control association studies. However, if late-onset diseases are due to large numbers of rare variants at many loci – the COMMON DISEASE/RARE VARIANT (CD/RV) HYPOTHESIS – this strategy would fail and the contribution of most individual variants would be too small to further our understanding of disease [3]. To evaluate these issues, we first examine our current knowledge of human genetic diversity.

Hidden genetic diversity

The human population is both evolutionarily young and genetically uniform, with less diversity than most other species, including other primates [4]. The most abundant differences between individuals are single nucleotide POLYMORPHISMS (SNPs), which account for most of the observed variability in typical sequence surveys [5]. The great majority of SNPs occur outside coding regions and their distribution is broadly consistent with SELECTIVE NEUTRALITY [6]. There are ~10 million predicted SNPs with allele frequencies above 0.01 [7]. Under the CD/CV hypothesis, these provide the major genetic substrate for common diseases. However, this picture may give a misleading impression of the genetic variation underlying the emergent diseases of modern civilizations.

The principal reason is that the vast majority of DNA sequence variants, including most of those with functional effects, are expected to be rare [8]. Genetic theory predicts that the distribution of neutral sites is heavily skewed towards low-frequency variants with as many below a frequency of 0.01 as above it [9]. But the proportion of rare variants is even higher for two reasons. First, most mutations with phenotypic effects are deleterious [10], so that their frequency is reduced by selection. Second, the human population has been expanding, generating large numbers of rare alleles by mutation [11] (Fig. 1). The overall pattern is therefore one of relatively few common SNPs and many individually rare single nucleotide variants.

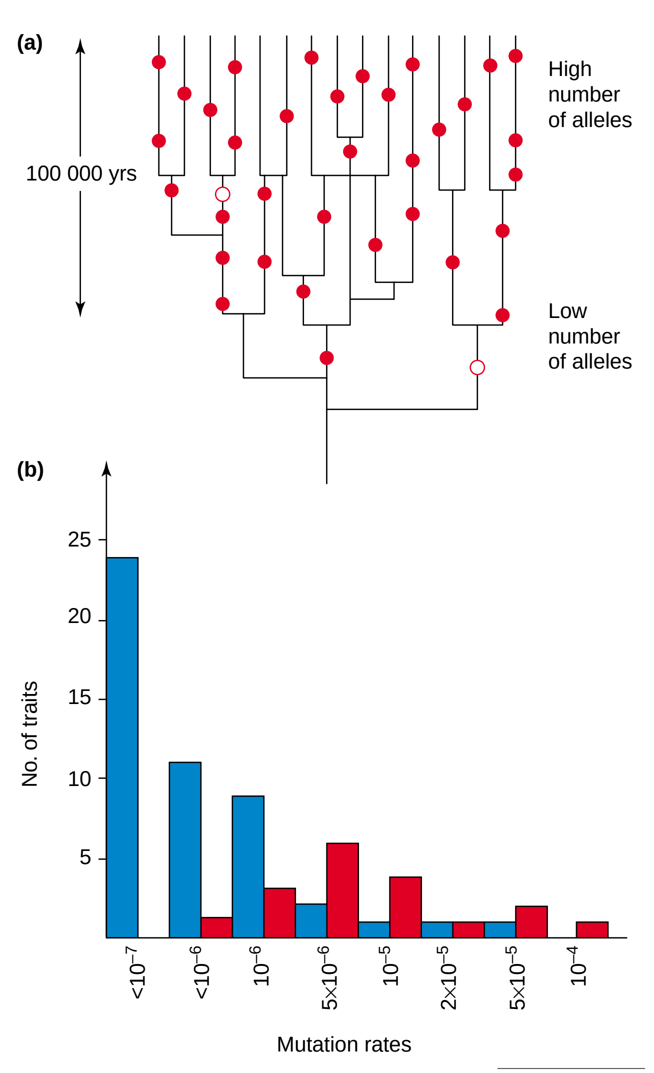
The majority of disease-causing alleles in early-onset mendelian disorders are recent, diverse and rare, resulting in extreme allelic heterogeneity. This is expected for deleterious alleles exposed to early selection, but is also found in later-onset diseases, including familial forms of cancer, coronary artery disease and Alzheimer dementia [12]. For example,

Corresponding author: Alan Wright (alan.wright@hgu.mrc.ac.uk).

http://tigs.trends.com 0168-9525/03/$ - see front matter © 2002 Elsevier Science Ltd. All rights reserved. PII: S0168-9525(02)00033-1

Glossary

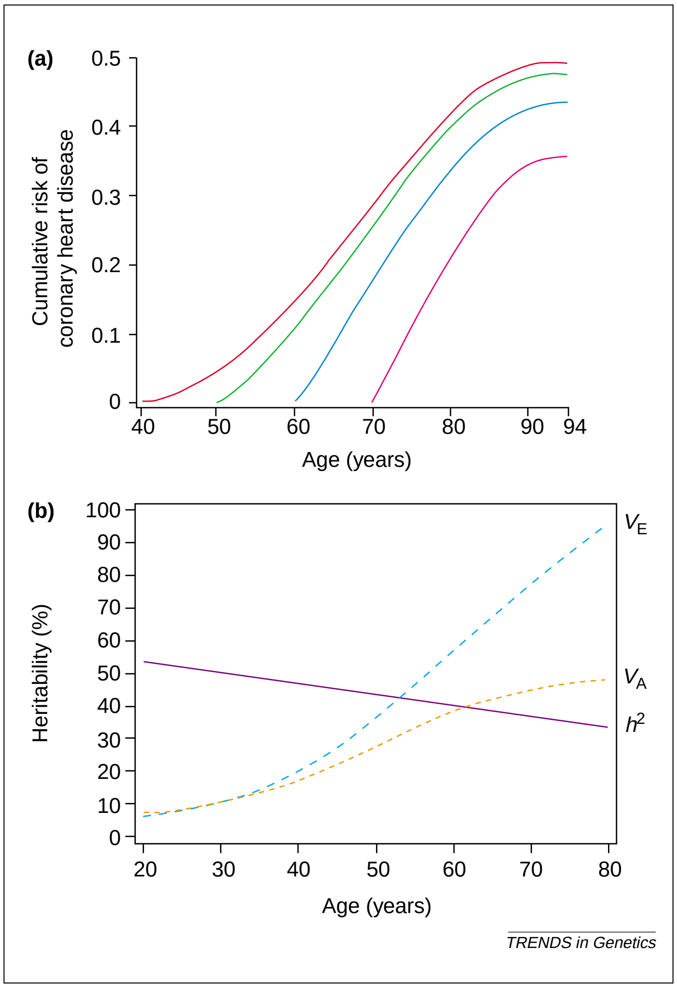
Antagonistic pleiotropy: see Trade-off model.
Balancing selection: selection that maintains more than one allele in the population at intermediate frequencies.
Common disease/common variant (CD/CV) hypothesis: susceptibility to common disease results from a small number of common polymorphic variants at one or more loci.
Common disease/fixed variant (CD/FV) hypothesis: susceptibility to common disease in a given population results from invariant sites at one or more loci. These can differ between populations and contribute to differences in disease susceptibility.
Common disease/rare variant (CD/RV) hypothesis: susceptibility to common disease results from numerous rare variants at many loci.
Directional dominance: dominance is directional when the value of the heterozygous effect (h) deviates from the expected intermediate value in heterozygotes (e.g. $h < 0.5$ for most loci causing inbreeding depression).
Dominance: nonadditive interactions between alleles at the same locus, which vary continuously from complete recessivity (heterozygous effect, h, is zero) through additivity ($h = 0.5$) to complete dominance ($h = 1$) (Box 1).
Effective population size (N_e): the number of individuals in a population contributing genes to succeeding generations, which predicts the rate of genetic drift.
Extreme concordant (discordant) sib pairs: pairs of siblings that are positively correlated (concordant) or negatively correlated (discordant) for a trait.
Fitness-related trait: a trait for which a change in value influences reproductive fitness.
Fixation: a state in which all members of a population are homozygous for a given allele (which is then said to be fixed, with an allele frequency of 1).
Genetic drift: random fluctuations in gene frequency arising from a finite effective population size (N_e).
Genome-wide mutation rate (U): the mean number of new deleterious mutant alleles arising per individual each generation.
Haplotype: a combination of linked variants on a single chromosome.
Identical-by-descent: alleles or genomic segments that are identical in one or more individuals as a result of inheritance from a common ancestor.
Inbreeding coefficient (F): the probability that both copies of an allele are inherited from a common ancestor (identical-by-descent).
Inbreeding depression: the detrimental effects of inbreeding, typically causing a reduction in the means of fitness-related traits, as a result of increased homozygosity.
Inbreeding load: the proportional reduction in the value of a fitness-related trait associated with a unit increase in the inbreeding coefficient.
Mutation accumulation model: a genetic model of senescence in which deleterious alleles with effects restricted to later life reach higher frequencies in the population than ones acting at an earlier age, assuming mutation–selection balance.
Mutational target: the proportion of the genome capable of influencing a trait as a result of *de novo* mutations.
Mutational variance: the genetic variance in a trait attributable to alleles maintained by mutation in opposition to selection.
Mutation–selection balance: a state of equilibrium between the input of genetic variants into a population by mutation and their elimination by natural selection.
Neutral alleles: see Selective neutrality.
Oligogenic: determined by a small number of genes of moderate effect.
Pleiotropy: an effect of a genetic variant on more than one trait.
Polygenic: determined by many genes of small effect.
Polymorphism: a variant allele with a frequency greater than 0.01.
Quantitative trait locus (QTL): any gene of small effect that contributes to quantitative variation in a trait.
Selection coefficient: the reduction in fitness of a given genotype, measured relative to the fitness of a standard genotype.
Selective constraints: the elimination of variants from a population as a result of natural selection.
Selective disadvantage: see Selection coefficient.
Selective neutrality: alleles with no effect on reproductive fitness.
Senescence: the decline with age in age-specific survival or other components of reproductive fitness.
Standing genetic variation: the naturally occurring genetic variation within a wild population.
Trade-off (antagonistic pleiotropy) model: a genetic model of senescence in which alleles show opposite pleiotropic effects on fitness-related traits in early and later life.



TRENDS in Genetics

Fig. 1. (a) Diagram of human population expansion illustrating (1) the large number of 'young' mutations (filled circles) compared with polymorphic variants (open circles), most of which are ancient (>100 000 years old) but some of which might have become common within recent times as a result of a selective advantage; and (2) the low number of alleles in human founder populations and high number of alleles in large modern populations [11]. (b) Mutation rates and disease. Mutation rates for human monogenic diseases vary from $<10^{-7}$ to 10^{-4} per locus per generation. There appear to be more loci with low than with high mutation rates resulting in disease. High mutation rates in genes causing Duchenne muscular dystrophy (5×10^{-5} to 10×10^{-5}) and neurofibromatosis type 1 (1×10^{-4}) account for their high incidence (at mutation–selection balance). Similarly, genes with high mutation rates are predicted to contribute disproportionately to the genetic variance underlying common diseases [62]. The red boxes indicate accurately estimated mutation rates for more common diseases, the blue boxes indicate rates for rarer diseases, which are only known within an order of magnitude. The data are taken, with permission, from [63,64] (© Springer, 1986).

according to the Breast Cancer Information Core (BIC) database (http://research.nhgri.nih.gov/bic) the familial breast cancer genes, *BRCA1* and *BRCA 2*, show at least 1200 and 1400 different mutations of large effect, each of which are rare, except in some founder populations. This is despite the fact that well over half of these tumours occur post-reproductively, so that causal alleles might reach higher frequencies. Most of these mutations are likely to be either of recent origin or to show direct or indirect deleterious effects on fitness. By contrast, out of the five



TRENDS in Genetics

Fig. 2. (a) Cumulative lifetime risk of coronary artery disease in males [37], which reaches almost one in two and shows a late-age levelling in disease risk, consistent with some of the genetic models discussed. (b) Heritability changes with disease age of onset. Possible changes in the additive component of genetic variance (V_A, orange) in relation to the environmental variance (V_E, blue) and heratibility (h^2, purple) with disease age of onset. Heritability is not an ideal measure of genetic variance in this context, because of the increase in V_E with age [1,27].

or six well-established common SNPs within *BRCA1* and *BRCA2* coding regions (coding SNPs or cSNPs), only one has been shown to exert a marginal (1.3-fold) increase in breast cancer risk [13]. A small increase in risk in many people might account for a large fraction of cases. But this is not so if such effects occur within highly interactive genetic networks, with many other variants of similar or opposite effect at varying frequencies in different populations, as expected under a CD/RV model. The pattern of thousands of recent and rare mutations, many with large effects, and a small number of ancient cSNPs is predictable [11], but here it is argued that most cSNPs are common precisely because they have little or no functional effect either on disease or on reproductive fitness.

If coding SNPs (comprising ~1.5% of all SNPs) are more likely to influence physiological function (hence disease) than noncoding ones, are they any less common? The subset of cSNPs that change an amino acid and are also predicted on structural grounds to be deleterious, occur at significantly lower frequencies than other SNPs, suggesting that they are indeed selected against [14,15]. Analyses of sequence divergence between humans and primates suggest that ~20% of all cSNPs are selectively neutral, most of which are common; of those predicted to be deleterious, over 80% are likely to be at frequencies below 0.01 (i.e. not truly polymorphic) [14].

In summary, the genetic variants that are most readily identified and studied in humans are SNPs, but most of these appear to have little or no effect either on reproductive fitness or on any sort of function. By contrast, the majority of deleterious variants, which are of most potential relevance to disease, are rare and accordingly difficult to study.

Genetic variation in late-onset traits

Is the pattern of genetic diversity likely to be different for variants influencing late-onset diseases? Unravelling the genetics of complex traits often requires indirect inferences about what are believed to be the many genes influencing them. Such 'polygenic' effects are thought to be too small and numerous to be measured individually, so their effects are measured collectively by partitioning the phenotypic variance into genetic and environmental components (Box 1).

How these components of genetic variance differ for late- versus early-onset traits has been examined in some detail theoretically. The intensity of selection on a gene with a late effect on fitness declines with the age at which it is expressed [16–19]. This implies that variants in such genes could reach higher frequencies, which would favour the CD/CV hypothesis. The 'mutation accumulation' (MA) model [17,19–21] extends this idea by suggesting that deleterious alleles with late effects accumulate in the genome, contributing to SENESCENCE and, by extension, to genetically influenced diseases that contribute to it. If these alleles have deleterious effects during reproductive life, they are still expected to be maintained at low frequencies, despite the diminishing force of selection with age. The 'TRADE-OFF' (TO) (or ANTAGONISTIC PLEIOTROPY) MODEL [18,19] provides an alternative, in which late-acting deleterious alleles can spread and even become universal in the population, if they also have favourable effects at an early age. Most genes are expressed before the end of reproductive life, and so are subject to selective scrutiny, but many show effects on different traits (PLEIOTROPY) at different times, with variable effects on fitness [19].

The expected higher frequencies of deleterious alleles with late-age effects are accompanied by increases in the components of genetic variance, because alleles with intermediate frequencies contribute more to these than do rare alleles [20,21]. Under both MA and TO models, the genetic variance components resulting from additive (V_A) and DOMINANCE (V_D) effects (Box 1) are expected to be larger for late- than for early-onset traits influencing fitness [20]. A late-age levelling in the rates of genetically influenced diseases, such as cancer, diabetes and cardiovascular disease, is also predicted by these models, much as observed (Fig. 2) [21,22]. Are the models supported by experimental data?

In *Drosophila melanogaster,* an increase in both V_A and V_D has been observed for several late-onset traits [20,23,24], suggesting that allele frequencies do indeed

Box 1. Variance components and inbreeding effects

Late-onset diseases can be considered to result when a threshold of quantitatively varying risk or liability is exceeded [a]. Liability results from the net effect of many quantitative traits (QT), which are influenced by genes and environment, often with small individual effects on risk, and hence are difficult to identify. These genetic effects can, however, be described collectively by analysing the components of genetic variance, which are estimated from the resemblance between relatives for disease or QT. The total genetic variance (V_G) of a complex trait can be partitioned into its components [a]:

- Additive genetic variance (V_A): the component of variance due to genetic effects that are directly transmissible from parent to offspring, and which are the main causes of resemblance between relatives.
- Dominance variance (V_D): the component of variance due to interactions (departures from additivity of effects) between alleles at the same locus, such as partial or complete dominance or recessivity.
- Epistatic variance (V_I): the component of variance due to interactions between alleles at different loci.

The total phenotypic variance (V_P) in a trait is the sum of V_G and any nongenetic (environmental) effects (V_E), together with effects of interactions between genotype and environment (V_{GE}). The (narrow-sense) heritability is the ratio of V_A to V_P.

These components can be estimated from correlations between relatives, such as parents and offspring or full-sibs. In practice, it is difficult to separate V_D and V_I, and they are often treated as a single component of nonadditive variance. Late-onset traits are predicted to show higher values of V_A and V_D [b,c].

Another source of information on genetic variation influencing disease is to measure the effects of inbreeding. Inbreeding can contribute to disease [d,e] and to inbreeding depression [f]. This results from increased homozygosity of trait alleles, which either show recessive effects on the trait acting in the same direction (DIRECTIONAL DOMINANCE; see Glossary) or show heterozygous advantage. *B*, the negative of the regression coefficient of trait mean on inbreeding coefficient *F*, provides a useful summary statistic for the genetic damage that would occur if all deleterious recessives were made homozygous ($F = 1$); it is often called the INBREEDING LOAD (see Glossary) [a,f].

If survival to adulthood is measured on a scale of natural logarithms, *B* provides an estimate of the number of deleterious mutations causing a genetic death (lethal equivalents). In one human study, *B* was 0.7 per gamete [e], suggesting that each (diploid) individual is heterozygous for 1.4 lethal equivalents affecting juvenile mortality.

The value of *B* for lethal mutations is similar in a variety of species. Recessive lethals contribute about half the inbreeding load for mortality in *Drosophila* up to the adult stage, the rest coming from mutations with minor effects (detrimentals) [f]. A typical fly carries one lethal mutation in the heterozygous state. A recent study of two fish species suggests that, despite their greater genome size, a similar value applies to vertebrates [g].

Theory shows that the value of *B* due to deleterious mutations depends only on the genome-wide mutation rate and the dominance of individual mutations [f,h]. If all mutations have the same effects, the value of *B* for a disease-related trait that is positively correlated with fitness is:

$$B = U\alpha\{(1/h) - 2\}$$

where *U* is the mutation rate per diploid individual to deleterious alleles affecting the trait; *h* is the extent to which fitness is reduced in a heterozygous mutation, relative to its effect in homozygotes; and α is a constant of proportionality relating the effect of a mutation on the trait to its effect on fitness.

Relating the measurement of variance components and inbreeding effects to the predictions of models of the maintenance of genetic variation provides an important means of testing the models [f,h,i]. Variance component analysis is also used in quantitative trait locus (QTL) mapping.

References

a Falconer, D.S. and Mackay, T.F.C. (1996) *Introduction to Quantitative Genetics*, 4th edn, Longman

b Partridge, L. and Gems, D. (2002) Mechanisms of ageing: public or private? *Nat. Rev. Genet.* 3, 165–175

c Charlesworth, B. and Hughes, K.A. (1996) Age-specific inbreeding depression and components of genetic variance in relation to the evolution of senescence. *Proc. Natl Acad. Sci. U.S.A.* 93, 6140–6145

d Rudan, I. *et al.* (2002) Inbreeding and the genetic complexity of human hypertension. *Genetics* in press

e Bittles, A.H. and Neel, J.V. (1994) The costs of human inbreeding and their implications for variations at the DNA level. *Nat. Genet.* 8, 117–121

f Crow, J.F. (1993) Mutation, mean fitness and genetic load. *Oxf. Surv. Evol. Biol.* 9, 3–42

g McCune, A.R. *et al.* (2002) A low genomic number of recessive lethals in natural populations of Bluefin Killifish and Zebrafish. *Science* 296, 2398–2401

h Charlesworth, B. and Hughes, K.A. (1999) The maintenance of genetic variation in life-history traits. *Evolutionary Genetics: From Molecules to Morphology* (Vol. 1) (Singh, R.S., Krimbas, C.B. eds), pp. 369–392, Cambridge University Press

i Barton, N.H. and Keightley, P.D. (2002) Understanding quantitative genetic variation. *Nat. Rev. Genet.* 3, 11–21

increase for late-onset traits. Other evidence comes from measuring the detrimental effects of inbreeding (INBREEDING DEPRESSION) (Box 1). Alleles with nonadditive effects (such as recessives) are expected to cause increased inbreeding depression for late- compared with early-onset traits, which is a unique prediction of the MA model. Age-related increases in V_D and in inbreeding depression have both been found for FITNESS-RELATED TRAITS in *Drosophila* [24]. If this is also true in humans, a significant fraction of the genetic variance underlying even diseases with small effects on fitness could be due to rare alleles [20]. However, the MA and TO models are not mutually exclusive and other *Drosophila* data support the trade-off model, implying the presence of alleles at higher frequencies as well [19]. The evidence from other species is scanty, but an increase with age in heritability (Box 1) of the human late-onset trait longevity [25] is consistent with the prediction of increased genetic variance underlying late-onset traits.

In short, the analysis of genetic variance components suggests that there is an increase in the frequencies of alleles influencing late-onset traits. But this does not imply that such variants are at high enough frequency to favour the CD/CV strategy; indeed, inbreeding effects suggest that many of them are at low frequencies. We now examine other evidence regarding the nature of such genetic variance.

Mutation and rare variants

Complex traits are influenced by many genes and so provide large MUTATIONAL TARGETS. Recent mutations provide a rich source of low-frequency variants, which account for a significant proportion of the STANDING GENETIC VARIATION in all organisms [10,26,27] (Box 2).

Box 2. The fate of new mutations

The average time that new mutations persist in a population depends on their selective disadvantage, their dominance effects and the effective population size (N_e) (see text) [a–c].

Most deleterious mutations are destined for ultimate loss from the population, within a time in generations of the order of the natural logarithm of N_e, unless their selection coefficients are smaller than the reciprocal of the effective population size [d]. Assuming an N_e in humans of ~10 000 [e,f] this gives a mean persistence time of around 10 generations, with a very wide distribution around the mean [d]. But the continual production of new mutations with each generation will lead to the maintenance of most deleterious variants at a given locus at low average frequencies close to those expected at equilibrium in an infinitely large population (mutation–selection balance) [c].

While they persist in the population, these alleles contribute genetic variance to the traits that they affect. The extent of the contribution depends on their mutation rates, their effects on the trait, the relationship between trait and fitness, and on N_e [c,g]. Most of these are not known with any precision.

Simulations suggest that loci with moderate to high *de novo* mutation rates contribute disproportionately to the genetic variance underlying human disease [h]. The over-representation of human disease loci with high mutation rates is readily seen in monogenic diseases (see Fig. 1b in main text) and this may be accentuated in more complex traits where the 'mutational target' is considerably larger, including noncoding and regulatory sequences, many with very small effects [i,j]. Neutral alleles, such as the majority of SNPs, may reach higher frequencies but are predicted to make a substantially smaller overall contribution to the disease variance, because they are likely to have very small effects on the trait.

References

a Slatkin, M. and Rannala, B. (2000) Estimating allele age. *Annu. Rev. Genomics Hum. Genet.* 1, 225–249

b Falconer, D.S. and Mackay, T.F.C. (1996) *Introduction to Quantitative Genetics*, 4th edn, Longman

c Bürger, R. (2000) *The Mathematical Theory of Selection, Recombination and Mutation*, John Wiley

d Kimura, M. (1983) *The Neutral Theory of Molecular Evolution*, Cambridge University Press

e Kaessmann, H. *et al.* (1999) Extensive nuclear DNA sequence diversity among chimpanzees. *Science* 286, 1159–1162

f Nachman, M.W. and Crowell, S.L. (2000) Estimate of the mutation rate per nucleotide in humans. *Genetics* 156, 297–304

g Charlesworth, B. and Hughes, K.A. (1999) The maintenance of genetic variation in life-history traits. *Evolutionary Genetics: From Molecules to Morphology* (Vol. 1) (Singh, R.S., Krimbas, C.B. eds), pp. 369–392, Cambridge University Press

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

i McKenzie, C.A. *et al.* (2001) Trans-ethnic fine mapping of a quantitative trait locus for circulating angiotensin I-converting enzyme (ACE). *Hum. Mol. Genet.* 10, 1077–1084

j Vafiadis, P. *et al.* (1997) Insulin expression in human thymus is modulated by *INS* VNTR alleles at the *IDDM2* locus. *Nat. Genet.* 15, 289–292

This is especially so for populations, such as humans, that have undergone large and recent expansions. This mutational variance is a function of mutation rates, mutational target size and population size. The effective population size (N_e) refers to the number of individuals that contribute genes to succeeding generations, and controls the rate of chance fluctuations in gene frequencies (genetic drift). Long-term N_e has been estimated to be ~10 000 in humans, based on sequence diversity. Population expansion means, however, that the number of rare variants of recent origin is greater than predicted for an equilibrium population with this value of N_e (Fig. 1). Such recent and rare mutations are clearly major contributors to human disease, and yet the inherent difficulty in detecting them tends to overemphasize variants, such as SNPs, that are readily detectable.

Most mutations are deleterious and destined to be lost within a few generations (Box 1). As a rule of thumb the frequency of a mutation that reduces fitness by at least $1/(4N_e)$ is mainly controlled by selection rather than genetic drift [28]. Therefore, even disease alleles with immeasurably small effects on fitness can be held at low population frequencies [20,28]. The frequencies of new mutations are therefore strongly influenced by their effects on fitness, which may be substantial for mutations with significant effects on disease, because the size of direct or indirect (pleiotropic) allelic effects on traits tends to be correlated with their effects on fitness [27,29,30]. For a particular disease susceptibility gene, the collective frequency of new deleterious mutations is therefore expected to be low, and close to that expected under mutation–selection balance (Box 2) [14–16].

The large size of the human genome results in a steady and substantial input of new mutations – estimated to be in the region of 175 per person per generation [31]. As a result, each of us carries a predicted 500–1200 deleterious mutations, most of which are rare (<0.01), in addition to those that are strongly deleterious and too rare to be detected in small sequence surveys [14].

Many variants with small phenotypic effects occur outside coding regions, so that genome-wide mutation rates to deleterious alleles [32] and the number of deleterious mutations each person carries could be considerably higher than estimates based only on coding regions. Noncoding regions might contribute at least 60% as much as coding regions to the genome-wide mutation rate [15]. There are examples of noncoding mutations with large effects on disease [33], but those with small effects should be even more common, because there are many more ways for these to subtly alter expression or function. Indeed, there is an increasing number of examples of small-effect quantitative trait locus (QTL) alleles located within noncoding regions, including ones relevant to human disease [34,35]. Mutations with essentially no effect on fitness can also influence late-onset traits, and should be at higher frequency. At the other extreme, only a small proportion of mutations show large phenotypic effects, but these tend to be over-represented in individuals with disease, as a result of selective ascertainment.

Summarizing the evidence

The possibility that a significant fraction of the genetic variation in complex traits is owing to rare alleles maintained by mutation–selection balance is supported

Box 3. Estimating the number and effect size of genes influencing a complex trait

In principle, it is possible to estimate the number of genetic loci influencing a complex trait. In practice, estimates often have large sampling variances and require assumptions that are not always met.

A classical method [a] is to cross inbred or natural populations that show a large difference in trait value and estimate the number of 'equivalent effect' genes accounting for the difference on the basis of the trait variances in the hybrid offspring (F1, F2 and backcross generations). This assumes that all 'high' alleles are fixed (homozygous) in one parental line and 'low' alleles in the other. Zeng [b] has eliminated bias in these estimates by accounting for recombination between loci and unequal gene effects. For tomato fruit weight, he found it difficult to estimate the total number of genes, which varied from 17 to 1540 depending on the shape of the gene effect distribution, but a minimum of 16 loci accounted for 95% of the genetic variance regardless of the effect distribution.

A second method depends on the systematic genetic mapping of trait loci distinguishing a pair of lines, yielding direct estimation of gene numbers [c]. This approach is constrained by lack of power to detect variants with small effects without unrealistically large samples but has been useful in showing the leptokurtic or L-shaped distribution of gene effect size using livestock and bristle number data [c,d].

Neither of these methods provides a direct estimate of the number of loci segregating within a population; if the lines concerned have been derived from a natural population, they at least provide a lower bound to the number of such loci.

A third method can be applied directly to population data, including humans, and is useful for estimating the number of recessive or partially recessive loci contributing to a trait that shows inbreeding depression [e,f]. The effect of inbreeding, B, and the dominance variance, V_D, are measured (Box 1). A lower bound to the number of genes, n, affecting the trait is provided by:

$$n \geq B^2/V_D$$

References

a Falconer, D.S. and Mackay, T.F.C. (1996) *Introduction to Quantitative Genetics*, 4th edn, Longman

b Zeng, Z.-B. (1992) Correcting the bias of Wright's estimates of the number of genes affecting a quantitative character: a further improved method. *Genetics* 131, 987–1001

c Barton, N.H. and Keightley, P.D. (2002) Understanding quantitative genetic variation. *Nat. Rev. Genet.* 3, 11–21

d Hayes, B. and Goddard, M.E. (2001) The distribution of the effects of genes affecting quantitative traits in livestock. *Genet. Sel. Evol.* 33, 209–230

e Charlesworth, B. and Hughes, K.A. (1999) The maintenance of genetic variation in life-history traits. *Evolutionary Genetics: From Molecules to Morphology* (Vol. 1) (Singh, R.S., Krimbas, C.B. eds), pp. 369–392, Cambridge University Press

f Rudan, I. *et al.* (2002) Inbreeding and the genetic complexity of human hypertension. *Genetics* in press

by 25 years of research into the genetics of quantitative traits (QTs) in *Drosophila*. A review of this extensive literature [27] suggests that deleterious alleles generated by mutation, and kept at low frequencies by selection, contribute between 33% and 67% of the genetic variation in a typical trait with at least some effect on reproductive fitness. These are likely to include many of the QTs that are major determinants of late-onset disease, whose effects on disease begin well within reproductive life. The remainder of the genetic variance appears to involve alleles at higher frequencies, maintained by some form of BALANCING SELECTION, such as heterozygote advantage or temporal, spatial or frequency-dependent selection [1,27].

Overall, allele frequencies might be higher for late- than for early-onset traits, but this is consistent with models favouring both low (MA model) and intermediate (TO model) frequency alleles. We currently lack any firm data on the size of these effects in humans, and on whether late-acting alleles that are at higher frequencies are the ones most likely to be important in understanding disease. On the contrary, the low predicted frequency of functionally deleterious alleles suggests that most clinically significant disease susceptibility alleles will not be 'high' in the sense implied by the CD/CV hypothesis. Strategies for identifying susceptibility alleles that are not robust in the face of a large and diverse group of rare alleles could therefore fail.

The number and size of allelic effects

An implicit assumption of most common disease mapping strategies, including the CD/CV hypothesis [11], is that complex traits are oligogenic, so that susceptibility loci are potentially detectable and informative for disease prediction or understanding. If, however, common diseases are truly polygenic, most individual effects will be too small to be useful. A limited number of loci appear able to exert strong monogenic effects on disease, suggesting that these influence rate-determining steps in disease pathways. The rarity of individual mutations under the CD/RV model, together with the large number of loci at which mutations can arise, means that the genetic basis of disease could vary greatly among individuals with the same disease.

The locus complexity of a trait depends on two factors. The first is its physiological complexity. Hirschsprung disease, for example, is a rare and highly specific lack of enteric ganglia, in which no more than eight loci account for most of the variance [36]. By contrast, coronary artery disease is a highly complex multidimensional trait with an exponentially increasing lifetime risk that is influenced by more than 280 risk factors [37,38] (Fig. 2a). The second factor is the distribution of gene effects. This is clearly important, because if most of the trait variance is determined by a handful of potentially detectable genes, the total number is irrelevant. An L-shaped or exponential distribution of mutation effect sizes has wide support [1,39], with many genes of small effect and fewer of large effect. However, even if a small number of genes account for the most extreme effects, the average patient with disease will not show large effects if there are many other determinants (Box 3).

Genetic models commonly assume a polygenic basis for complex traits, but the actual number of loci is hard to estimate [1]. Many loci with small effects will not be detectable using finite samples, and the effects of individual alleles are commonly overestimated. One way to estimate the number of loci is to combine estimates of dominance variance (V_D) and the INBREEDING LOAD (B) (Box 3). Because inbreeding depression results from increased homozygosity at loci influencing the trait, a regression of QT value (or

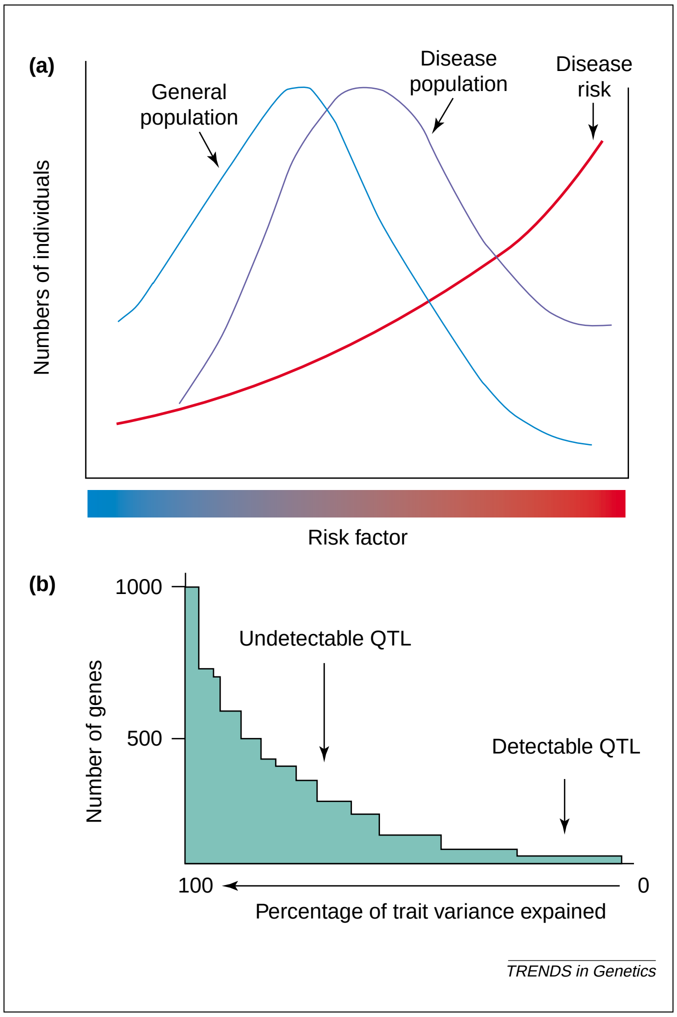


Fig. 3. Detection of clinically useful genetic effects in 'typical' patients with late-onset disease is likely to be impossible, but it may be feasible in those at the extremes of the disease or risk factor distribution. (a) Average patients have average genetic effects. The relationship between risk factor distributions in the general population and in a population with disease shows that most disease events occur in those within the normal range, because risk increases at all levels [41,42]. This emphasizes the need for genetic analysis of those at the extremes of the distribution. (b) Predicted L-shaped relationship between the number of quantitative trait loci (QTL) genes and the percentage of the trait variance explained. QTL contributing to the left of the distribution will often be undetectable, and those contributing to the tail of the distribution on the right show larger effects and are potentially detectable. The number of QTL might be too large for specific loci to be individually detectable in 'typical' patients, most of whom lie within the normal range for individual risk factors. However, gene identification could be both tractable and heuristically useful in those at the extremes of the distribution (e.g. extreme 10%) [40].

disease prevalence) on the INBREEDING COEFFICIENT can provide a lower limit for the number of segregating loci (Box 3). If the minimum number of loci required to explain most of the trait variance is large, then individual QTL effects might be too small to be detected with finite samples. Rudan *et al.* [40] came to exactly this conclusion in a study of inbreeding effects on blood pressure. Assuming an L-shaped distribution of allelic effect sizes and V_D/V_P of 0.33, the upper 25% of variation in systolic blood pressure could be explained by a minimum of 24–48 genes, and the upper 50% by at least 90–165 genes [40] (Fig. 3b). The total number of contributory loci must be even larger, because this estimate excludes those with purely additive effects. These results suggest that loci underlying a trait such as hypertension in a 'typical' patient are likely to be so numerous and of such small effect that gene identification will be either impossible or unhelpful. The problem is compounded by the fact that most risk factors in common diseases operate at virtually all levels, so that the 'average' patient typically has a trait value within the normal range, and most individual effects could be too small to detect [41,42] (Fig. 3a).

Global environmental change

To what extent does recent human environmental change affect our increasingly elderly populations? The shift in age structure is itself a major environmental change to which our genomes are poorly adapted. However, a wide range of common diseases have shown large changes even in age-standardized prevalence within the past 50 years, especially in western societies [43]. Environmental changes almost certainly account for this, raising questions about the overall significance of heritable components (Box 4). Gene–environment interactions can be seen with late-onset diseases, such as type 2 diabetes [44], in which recent dietary changes appear to be a major contributor to disease [45]. Ethnic variation in susceptibility to such diseases is often taken as support for the role of genes, but it is equally plausible that the most relevant genetic factors are essentially invariant in such populations as a result of strong selective advantages in the past [46] – the COMMON DISEASE/FIXED VARIANT (CD/FV) HYPOTHESIS. Conventional mapping studies would then fail to detect them. A striking example of the power of selection is the relatively recent FIXATION of the Duffy blood group $FY^{*}O$ allele in sub-Saharan Africans, probably because of increased resistance to *Plasmodium vivax* malaria [47].

Other genes appear to have been under strong selection in certain populations without leading to fixation, resulting in alleles at intermediate frequencies [48,49] (Fig. 1a). Variants such as those found in certain MHC class I and II HAPLOTYPES show functionally significant effects and are therefore good candidates for the CD/CV approach. Traits exposed to major environmental change, such as dietary alteration at the time of urbanization, immunity to disease, response to starvation and some cultural behaviours are promising in this respect.

Gene identification under a polygenic hypothesis

The scenario of a major class of deleterious but individually rare alleles of recent origin underlying the heritable component of late-onset diseases has been largely ignored until recently. It poses seemingly intractable statistical problems for gene mapping and is therefore an unattractive investment for grant-awarding agencies and biotechnology companies. More attractive alternatives, such as the CD/CV hypothesis, appear to be driving the research strategy in spite of, rather than because of, the science.

An example is the proposal to use high-frequency SNP haplotypes to identify common disease determinants by population association studies [50,51]. A recent mutational origin for the variants underlying much of the heritability in such diseases means that many will be superimposed on ancient core haplotypes. The majority of new variants are deleterious [26,27] and multiple variants

Box 4. Establishing a genetic basis for late-onset disorders

According to the website Online Mendelian Inheritance in Man (http://www.ncbi.nlm.nih.gov/entrez/Omim/mimstats.html), fewer than 10% of human transcribed sequences (including alternative transcripts) show mutations capable of major (monogenic) effects that are relatively independent of context. By contrast, most common diseases are highly dependent on both genetic and environmental context, although they generally show significant heritability (≥ 0.1) (see Fig. 2b in main text).

Heritability is assumed to be present if there is an increased trait correlation in relatives compared with unrelated individuals. Relatives (especially twins) are, however, exposed to more similar environments than random controls; also, genetic and environmental effects frequently interact and are correlated, so that in the absence of extensive half-sib or adoptive twin data, basic assumptions may not be met, leading to overestimation of heritability [a].

The magnitude of genetic effects in complex traits is frequently dwarfed by known environmental factors, leading some to question the current emphasis on genetics [b]. For example, diet, physical inactivity and tobacco have been proposed to account for 75% of new cases of cardiovascular disease [c]. In developed countries, tobacco smoking alone causes one-third of all cancer deaths [d], 80% of chronic bronchitis [e] and 13% of coronary artery disease [e]. In addition, recent environmental change has resulted in a global epidemic of type 2 diabetes, with uncertain consequences for heritability. In coronary artery disease, heritability declines with age of onset [f], implicating either increased environmental variance (V_E), reduced total genetic variance (V_G), or both, in later life (see Fig. 2b in main text).

The contribution of individual genetic factors to complex disease is, in most cases, seldom greater than a few percent of the trait variance and a small percentage of cases. For example, currently known breast cancer genes account for $<2\%$ of population risk [g]. The collective contribution of genetic variability, summarized by measures such as heritability or coefficients of genetic variation, is however both substantial and almost universal in complex traits, for reasons that are discussed in the text.

Even with the standardized environments and reduced diversity of laboratory mice, mapping of QTL for obesity-related traits suggests many genes of small effect [h]. Some QTL influencing human disease show larger effects, such as *APOE*E4*, which accounts for up to 10% of the variance in both plasma cholesterol and age of onset for Alzheimer disease, although some of these effects might result from interactions (e.g. with dietary fat [i]). Opportunities for disentangling interactions between genes and environment are often limited [j].

The identification of genes accounting for rare monogenic forms of common diseases (e.g. breast and colon cancer, Alzheimer disease) has made a substantial contribution towards elucidating disease mechanisms and has led to therapeutic progress (e.g. statins, secretase inhibitors). It is less clear whether the identification of variants with small effects on disease risk will have a similar impact. Predictions based on experimental organisms suggest that the lack of control of genotype or environment in human studies, together with context-dependence of QTL effects, mean that individual QTL effects will generally be very small [k].

References

a Kamin, L.J. and Goldberger, A.S. (2002) Twin studies in behavioral research: a skeptical view. *Theor. Popul. Biol.* 61, 83–95

b Holtzman, N.A. and Marteau, T.M. (2000) Will genetics revolutionize medicine. *N. Engl. J. Med.* 343, 141–144

c Beaglehole, R. (2001) Global cardiovascular disease prevention: time to get serious. *Lancet* 358, 661–663

d Peto, J. (2001) Cancer epidemiology in the last century and the next decade. *Nature* 411, 390–395

e Vineis, P. *et al.* (2001) Misconceptions about the use of genetic tests in populations. *Lancet* 357, 709–712

f Marenberg, M.E. *et al.* (1994) Genetic susceptibility to death from coronary heart disease in a study of twins. *N. Engl. J. Med.* 330, 1041–1046

g Pharoah, P.D. *et al.* (2002) Polygenic susceptibility to breast cancer and implications for prevention. *Nat. Genet.* 31, 33–36

h Cheverud, J.M. *et al.* (2001) Genetic architecture of adiposity in the cross of LG/J and SM/J inbred mice. *Mamm. Genome* 12, 3–12

i Tikkanen, M.J. (1997) Apolipoprotein E polymorphism and plasma cholesterol response to dietary change. *World Rev. Nutr. Diet.* 80, 15–21

j Wright, A.F. *et al.* (2002) Gene–environment interactions – the BioBank UK study. *Pharmacogenomics J.* 2, 75–82

k Mackay, T.F.C. (2001) Quantitative trait loci in *Drosophila*. *Nat. Rev. Genet.* 2, 11–20

can arise on the commonest haplotypes, so that association mapping is inefficient. Disease alleles might be missed that are not associated with ancient core haplotypes [52]. Haplotype frequencies will be poorly matched to disease allele frequencies, and disease associations will be hard to detect, especially if alleles of opposite effect arise on the same haplotype. But if susceptibility alleles are predominantly rare and numerous, are there any strategies that will work?

Most complex human disease alleles identified to date are partially recessive [12], as predicted from other organisms [26,27,53]. To maximize the 'detectance' [54] of such alleles, ascertainment for large phenotypic effects [12,54] and for populations showing reduced environmental and genetic variance [3,55], and the use of large sample sizes [56] will all help. The use of QTs rather than disease *per se* also increases informativeness [3,57]. Extreme CONCORDANT or DISCORDANT sibs with trait values (or disease age-of-onset) close to the extremes of the distribution could also be helpful [57]. Large endogamous populations might be especially valuable, because of their environmental uniformity and their ability to reveal identical-by-descent (IBD) segments based on kinship information (using relatively low-density genome scans) [3,36,55]. A search for homozygous IBD segments could help to detect rare recessive alleles of large effect. Many disease loci identified in this way could be 'local' to specific populations. If this provides new insights into disease pathways it will be invaluable, because new drug targets are in short supply. Finally, animal models, in which inbred or selected strains show differences in disease or trait value, are more tractable [58], because far fewer segregating loci are involved, although they represent a minute subset of the variation found in wild populations.

Genetic linkage is robust in the presence of allelic heterogeneity, whereas association studies are not, but both lack power in the face of a strongly polygenic basis for disease. The familial breast cancer genes were identified by linkage because some families showed large-effect alleles and locus heterogeneity was limited. The choice of study population can help to minimize disease complexity but the choice of disease-related phenotype is equally important [3,57].

Admixture mapping [59] has advantages, especially under a CD/FV model. Admixture between ethnic groups showing high and low disease prevalence, under currently similar but differing historical environments, could be the only way to identify moderate- or large-effect genes contributing to such differences. The diversity of hazards in different geographic and cultural environments would tend to exaggerate selective differences between populations, although demographic forces might have dispersed them throughout large urban populations.

In large continental populations, the high levels of locus and allelic heterogeneity can be turned to advantage by the relative ease of identifying either rare familial forms of disease or rare survivors [25]. This approach has produced disproportionate gains in understanding disease mechanisms in monogenic disorders [12]. The recruitment of sufficiently large series of sib trios or quartets, especially when concordant or discordant for QT or early-onset disease, might only be possible in large national studies [60].

The proposed complexity of late-onset disorders suggests that identifying genes with the largest effects, and which contribute most to the extremes of the disease or trait distribution, might be the most robust approach – only one step away from the methods so successfully applied to monogenic disorders. This will help to avoid the illusion of large QTL effects that are composite artefacts of numerous but clustered variants [1,61] or with individual effects so small as to be of little value in elucidating disease mechanisms. The difficulties encountered in identifying clinically relevant disease loci could reflect an overambitious goal of finding genes involved in the majority of patients with a disease, fuelled by commercial forces. The scenario of a truly polygenic basis for much of the heritability of complex traits might require both new approaches and a return to tried and tested ones.

Acknowledgements

I.R. and H.C. were supported in part by an International Research Development Award from The Wellcome Trust. B.C. is supported by the Royal Society. We thank A.S. Kondrashov, C.M. Steel, P.M. Visscher and N.D. Hastie for useful comments and suggestions, and S. Bruce for the artwork.

References

1 Barton, N.H. and Keightley, P.D. (2002) Understanding quantitative genetic variation. *Nat. Rev. Genet.* 3, 11–21
2 Lander, E.S. (1996) The new genomics: global views of biology. *Science* 274, 536–539
3 Wright, A.F. *et al.* (1999) Population choice in mapping genes for complex diseases. *Nat. Genet.* 23, 397–404
4 Kaessmann, H. *et al.* (1999) Extensive nuclear DNA sequence diversity among chimpanzees. *Science* 286, 1159–1162
5 Sachidanandam, R. *et al.* (2001) A map of human sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 409, 928–933
6 Przeworski, M. *et al.* (2000) Adjusting the focus on human variation. *Trends Genet.* 16, 296–302
7 Kruglyak, L. and Nickerson, D.A. (2001) Variation is the spice of life. *Nat. Genet.* 27, 234–236
8 Zwick, M.E. *et al.* (2000) Patterns of genetic variation in Mendelian and complex traits. *Annu. Rev. Genomics Hum. Genet.* 1, 387–407
9 Kimura, M. (1971) Theoretical foundations of population genetics at the molecular level. *Theor. Popul. Biol.* 2, 174–208
10 Lynch, M. *et al.* (1999) Perspective: spontaneous deleterious mutation. *Evolution* 53, 645–663
11 Reich, D.E. and Lander, E.S. (2001) On the allelic spectrum of human disease. *Trends Genet.* 17, 502–510
12 Wright, A.F. and Hastie, N.D. (2001) Complex genetic diseases: controversy over the Croesus code. *Genome Biol.* 2, 2007
13 Healey, C.S. *et al.* (2000) A common variant in *BRCA2* is associated with both breast cancer risk and prenatal viability. *Nat. Genet.* 26, 362–364
14 Fay, J.C. *et al.* (2001) Positive and negative selection on the human genome. *Genetics* 158, 1227–1234
15 Sunyaev, S. *et al.* (2001) Prediction of deleterious human alleles. *Hum. Mol. Genet.* 10, 591–597
16 Fisher, R.A. (1930) *The Genetical Theory of Natural Selection*, Oxford University Press
17 Medawar, P.B. (1946) Old age and natural death. *Mod. Quart.* 1, 30–56
18 Williams, G.C. (1957) Pleiotropy, natural selection and the evolution of senescence. *Evolution* 11, 398–411
19 Partridge, L. and Gems, D. (2002) Mechanisms of ageing: public or private? *Nat. Rev. Genet.* 3, 165–175
20 Charlesworth, B. and Hughes, K.A. (1996) Age-specific inbreeding depression and components of genetic variance in relation to the evolution of senescence. *Proc. Natl Acad. Sci. U.S.A.* 93, 6140–6145
21 Charlesworth, B. (2001) Patterns of age-specific means and genetic variances of mortality rates predicted by the mutation accumulation theory of ageing. *J. Theor. Biol.* 210, 47–65
22 Mueller, L.D. and Rose, M.R. (1996) Evolutionary theory predicts late-life mortality plateaus. *Proc. Natl Acad. Sci. U.S.A.* 93, 15249–15253
23 Shaw, F.H. *et al.* (1999) Toward reconciling inferences concerning genetic variation in senescence in *Drosophila melanogaster*. *Genetics* 152, 553–566
24 Hughes, K.A. *et al.* (2002) A test of evolutionary theories of aging. *Proc. Natl Acad. Sci. U.S.A.* 99, 14286–14291
25 Perls, T. *et al.* (2002) The genetics of aging. *Curr. Opin. Genet. Dev.* 12, 362–369
26 Simmons, M.J. and Crow, J.F. (1977) Mutations affecting fitness in *Drosophila* populations. *Annu. Rev. Genet.* 11, 49–78
27 Charlesworth, B. and Hughes, K.A. (1999) The maintenance of genetic variation in life-history traits. *Evolutionary Genetics: From Molecules to Morphology* (Vol. 1) (Singh, R.S., Krimbas, C.B. eds), pp. 369–392, Cambridge University Press
28 Kimura, M. (1983) *The Neutral Theory of Molecular Evolution*, Cambridge University Press
29 Keightley, P.D. and Hill, W.G. (1990) Variation maintained in quantitative traits with mutation–selection balance: pleiotropic side-effects on fitness traits. *Proc. R. Soc. Lond. Ser. B* B242, 95–100
30 Kruuk, L.E.B. *et al.* (2000) Heritability of fitness in a wild mammal population. *Proc. Natl Acad. Sci. U.S.A.* 97, 698–703
31 Nachman, M.W. and Crowell, S.L. (2000) Estimate of the mutation rate per nucleotide in humans. *Genetics* 156, 297–304
32 Kondrashov, A.S. (2001) Sex and *U*. *Trends Genet.* 17, 75–77
33 Kleinjan, D.A. *et al.* (2001) Aniridia-associated translocations, DNase hypersensitivity, sequence comparison and transgenic analysis redefine the functional domain of PAX6. *Hum. Mol. Genet.* 10, 2049–2059
34 McKenzie, C.A. *et al.* (2001) Trans-ethnic fine mapping of a quantitative trait locus for circulating angiotensin I-converting enzyme (ACE). *Hum. Mol. Genet.* 10, 1077–1084
35 Vafiadis, P. *et al.* (1997) Insulin expression in human thymus is modulated by *INS* VNTR alleles at the *IDDM2* locus. *Nat. Genet.* 15, 289–292
36 Gabriel, S.B. *et al.* (2002) Segregation at three loci explains familial and population risk in Hirschsprung disease. *Nat. Genet.* 31, 89–93
37 Lloyd-Jones, D.M. *et al.* (1999) Lifetime risk of developing coronary heart disease. *Lancet* 353, 89–92
38 Hopkins, P.N. and Williams, R.R. (1986) Identification and relative weight of cardiovascular risk factors. *Cardiol. Clin.* 4, 3–31
39 Hayes, B. and Goddard, M.E. (2001) The distribution of the effects of genes affecting quantitative traits in livestock. *Genet. Sel. Evol.* 33, 209–230
40 Rudan, I. *et al.* (2002) Inbreeding and the genetic complexity of human hypertension. *Genetics* in press
41 Rose, G. (1992) *The Strategy of Preventive Medicine*, Oxford University Press
42 Law, M.R. and Wald, N.J. (2002) Risk factor thresholds: their existence under scrutiny. *Br. Med. J.* 324, 1570–1576

43 Charlton, J. and Murphy, M. (1997) *The Health of Adult Britain 1841–1994*, HMSO

44 Ravussin, E. *et al.* (1994) Effects of a traditional lifestyle on obesity in Pima Indians. *Diabetes Care* 17, 1067–1074

45 Neel, J.V. (1962) Diabetes mellitus: a 'thrifty' genotype rendered detrimental by progress? *Am. J. Hum. Genet.* 14, 353–362

46 Björntorp, P. (2001) Thrifty genes and human obesity. Are we chasing ghosts? *Lancet* 358, 1006–1008

47 Hamblin, M.T. *et al.* (2002) Complex signatures of natural selection at the Duffy blood group locus. *Am. J. Hum. Genet.* 70, 369–383

48 Slatkin, M. and Rannala, B. (2000) Estimating allele age. *Annu. Rev. Genomics Hum. Genet.* 1, 225–249

49 Merriman, T.R. and Todd, J.A. (1995) Genetics of autoimmune disease. *Curr. Opin. Immunol.* 7, 786–792

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

51 Patil, N. *et al.* (2001) Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* 294, 1719–1723

52 Lee, C. (2002) Irresistible force meets immovable object: SNP mapping of complex diseases. *Trends Genet.* 18, 67–69

53 Crow, J.F. (1993) Mutation, mean fitness and genetic load. *Oxf. Surv. Evol. Biol.* 9, 3–42

54 Weiss, K.M. and Terwilliger, J.D. (2000) How many diseases does it take to map a gene with SNPs? *Nat. Genet.* 26, 151–157

55 Abney, M. *et al.* (2002) Quantitative-trait homozygosity and association mapping and empirical genomewide significance in large, complex pedigrees: fasting serum-insulin level in the Hutterites. *Am. J. Hum. Genet.* 70, 920–934

56 Dahlman, I. *et al.* (2002) Parameters for reliable results in genetic association studies in common disease. *Nat. Genet.* 30, 149–150

57 Gu, C. and Rao, D.C. (2001) Optimum study designs. *Adv. Genet.* 42, 439–457

58 Flint, J. and Mott, R. (2001) Finding the molecular basis of quantitative traits: successes and pitfalls. *Nat. Rev. Genet.* 2, 437–445

59 McKeigue, P.M. (1998) Mapping genes that underlie ethnic differences in disease risk: methods for detecting linkage in admixed populations, by conditioning on parental admixture. *Am. J. Hum. Genet.* 63, 241–251

60 Wright, A.F. *et al.* (2002) Gene–environment interactions – the BioBank UK study. *Pharmacogenomics J.* 2, 75–82

61 Visscher, P.M. and Haley, C.S. (1996) Detection of putative quantitative trait loci in line crosses under infinitesimal genetic models. *Theor. Appl. Genet.* 93, 691–702

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

63 Stevenson, A.C. and Kerr, C.B. (1967) On the distribution of frequencies of mutation in genes determining harmful traits in man. *Mutat. Res.* 4, 339–352

64 Vogel, F. and Motulsky, A.G. (1986) *Human Genetics*, Springer-Verlag