

ORIGINAL ARTICLE

A remark on rare variants

Konrad Oexle

The genetic architecture of a disease determines the epidemiological methods for its examination. Recently, Bodmer and Bonilla suggested that moderately strong, moderately rare variants contribute substantially to the genetic population attributable risk (PAR) of common diseases. In the first part of this communication, I provide a concise reconstruction of their deliberation. Variants contributing to human disease can be identified by linkage or by association tests. Risch and Merikangas analyzed the power of these tests by comparing the affected sib-pair linkage test (ASP) and the transmission disequilibrium association test (TDT). In the second part of this paper, I give an accessible reconstruction of this comparison and derive simple approximations in the low allele frequency range, directly showing that the linkage test is much more sensitive to a decrease of frequency or effect size. In the third part, I analyze a disease model whose genetic architecture is proportional to Kimura's infinite sites model. The relation between a variant's selection coefficient and its effect size in disease generation is assumed to be simple, and the number of contributing genetic variants is determined by the sum of their approximative PAR contributions. An association test (TDT) is finally applied to this disease model. For different ranges of effect size and allele frequency, I derive the minimal sample size necessary to detect at least one contributing variant. It turns out that, although the majority of contributing variants is not accessible with realistic sample sizes, a minimum of sample size may be given for moderately strong variants in the 1% frequency range.

Journal of Human Genetics (2010) 55, 219–226; doi:10.1038/jhg.2010.9; published online 5 March 2010

Keywords: association test; genetic architecture; genetic variants; infinite sites model; linkage analysis; mutation rate; population attributable risk; power analysis

Happy families are all alike; every unhappy family is unhappy in its own way.

(Tolstoy¹ 1875)

INTRODUCTION

Different assumptions on the nature of genetic factors accompany the history of genetics since its very beginning; that is, since Darwin² in 1859 who pictured a gradual originating of species as a sequence of small effects, and Mendel³ in 1866 who examined traits determined by a single factor. Little more than 100 years later, genome-wide investigations in human genetics have identified rare variants with strong effects by linkage analyses of monogenic disorders and common variants with weak effects by association analyses of multifactorial traits. However, the genetic influence on phenotypes cannot be accounted for by these two extreme types of variants only. Weak-effect variants probably occur in all frequency ranges. Moreover, there may be a considerable number of intermediate variants with moderately strong effects that have moderately low allele frequencies.

In this paper, I first pick up a reflection of Bodmer and Bonilla⁴ who recently called attention to the contribution of moderately rare variants with moderately strong effects to the incidence of multifactorial diseases. Frequency and effect size of a variant determine the power of tests used for its detection. Comparative power analyses of

genome-wide linkage and association tests have been performed by Risch and Merikangas.⁵ I reconstruct these analyses in a dense but complete and comprehensible manner, and provide simple approximations for small frequencies. The approximations allow for a direct comparison of the influences of frequency and effect size on the power (that is, necessary sample sizes) of linkage and association tests in detecting disease-related variants.

Bodmer and Bonilla⁴ also made assumptions on the relation between number, frequency and effect size of variants involved in a multifactorial disease. However, to some extent these assumptions on the genetic architecture were merely postulated. I construct a disease model whose effectors are distributed in proportion to the infinite sites model of Kimura⁶ with the disease being the relevant selective force. This disease model is then used for a general consideration of association testing. On the basis of this consideration, I discuss strategies to increase test efficiency, such as candidate space reduction, formation of equivalence classes of mutations or selection of probands (controls) with increased (decreased) mutation probability.

BODMER AND BONILLA ON RARE VARIANTS IN COMMON DISEASE

Bodmer and Bonilla⁴ provided a quantitative estimate of the role of moderately rare (0.1–3%) variants in the susceptibility to complex

disease. They derived a simple relation between the population attributable risk (PAR) of a variant, its risk allele frequency and the associated odds ratio (OR; see Box 1 of their paper). However, their derivation is difficult to comprehend. Here, I provide a reconstruction. At first they defined the PAR as the fractional contribution of a risk allele A to the incidence $P(D)$ of a disease D ,

$$\text{PAR} = (P(D) - P(D|a))/P(D) \quad (1)$$

where $P(D|a)$ is the background incidence of D in the absence of A (see their equation 1 that lacks the bracketing of the numerator, however). They also introduced the 'additional penetrance, f , due the presence of a dominant variant' and stated without derivation,

$$f = P(D|a)(\text{OR} - 1)/(1 + P(D|a)(\text{OR} - 1)) \quad (2)$$

If f is the penetrance above background,

$$f = P(D|A) - P(D|a) \quad (3)$$

it is indeed possible to approximate Equation (2) from the definition of the odds ratio, $\text{OR} = P(D|A)/(1 - P(D|A)) / (P(D|a)/(1 - P(D|a)))$: Replacing $P(D|A)$ by $f + P(D|a)$ according to Equation (3) yields $f = (\text{OR} - 1)(P(D|a)(1 - f) - P(D|a)^2)$ that results in Equation (2), if $P(D|a)$ is small enough so that $P(D|a)^2$ can be ignored. Assuming a small background incidence or a small effect size (that is, $P(D|a)(\text{OR} - 1) \ll 1$), Bodmer and Bonilla⁴ further simplified Equation (2) to

$$f \approx P(D|a)(\text{OR} - 1) \quad \text{if } P(D|a)(\text{OR} - 1) \ll 1 \quad (4)$$

by which they restated the fact that the OR approximates the relative risk (RR) if it is small or if the incidence of the disease is small: With Equations (3) and (4), $\text{RR} = P(D|A)/P(D|a) = f/P(D|a) + 1 \approx \text{OR}$.

Bodmer and Bonilla⁴ then claimed that $\text{PAR} \approx 2fP(A)/P(D)$. With a Hardy-Weinberg assumption on the disease incidence, $P(D) = P(D|AA)P(A)^2 + 2P(D|A)P(A)(1 - P(A)) + P(D|a)(1 - P(A))^2$, this claim can be examined: It is true if the variant is rare enough so that the $P(A)^2$ terms can be neglected. With Equation (3),

$$\begin{aligned} P(D) &\approx 2P(D|A)P(A) + P(D|a)(1 - 2P(A)) \\ &\approx 2fP(A) + P(D|a) \quad \text{if } P(A) \ll 1 \end{aligned} \quad (5)$$

Thus, with Equations (1), (4) and (5),

$$\begin{aligned} \text{PAR} &\approx 2fP(A)/P(D) && \text{if } P(A) \ll 1 \\ &\approx 2(\text{OR} - 1)P(A)P(D|a)/P(D) && \text{if } P(A) \ll 1, f \ll 1 \\ &\approx 2(\text{OR} - 1)P(A) && \text{if } P(A) \ll 1, f \ll 1 \end{aligned} \quad (6)$$

Bodmer and Bonilla⁴ applied Equation (6) both to rare and to common variants in spite of the condition $P(A) \ll 1$. Moreover, they assumed additivity of the PAR, which also must be regarded as a simplifying approximation. Their purpose was to compare

the cumulated PAR contributions of variants in different frequency ranges,

$$\begin{aligned} \text{PAR}_p &\approx 2n_p f_p p / P(D) \approx 2n_p p (\text{OR}_p - 1) P(D|a) / P(D) \\ &\approx 2n_p p (\text{OR}_p - 1) \end{aligned} \quad (7)$$

with OR_p , f_p and n_p as OR, penetrance and number, respectively, of the variants in the frequency class with $P(A) = p$. Bodmer and Bonilla⁴ compared a small set of common variants associated with small effects ($n_p = 10$, $p = 0.5$, $\text{OR}_p = 1.4$) and a larger set of moderately rare variants associated with stronger effects ($n_p = 200$, $p = 0.002$, $\text{OR}_p = 3.5$; see Box 1 in their paper). The inverse relation of p and OR_p was derived from a survey of published association studies. The assumptions on n_p were rather speculative. From these two sets, they calculated 'relative PAR contributions', $2n_p p (\text{OR}_p - 1)$, of 4 and 2, respectively, suggesting that moderately rare variants contribute substantially to the incidence of multifactorial disease. Moreover (see Box 2 in their paper), using the example of breast cancer (incidence $P(D|a) \approx 0.1$), they compared the moderately rare, moderately strong variants at a single locus (assuming $n_p = 1000$, $p = 0.002$, $\text{OR}_p = 2$) with the set ($n_p = 1000$) of very strong, that is, highly penetrant ($f_p = 1$) mutations involved in the Mendelian subtype of the disease. These variants are very rare ($p = 5 \times 10^{-7}$) as the authors derived by assuming a balance between mutation pressure m and selective loss; that is, $(1 - p)m = -ps$, with $m = 5 \times 10^{-8}$ and $s = -0.1$. Again they found a comparatively large PAR contribution of the moderately rare variants (Table 1).

Since the disease incidence $P(D)$ and background incidence $P(D|a)$ do not differ substantially if OR or p is small, the 'relative PAR contribution' $2n_p p (\text{OR}_p - 1)$ actually equals the absolute contribution of the variants from the p -frequency class (see Equation (6)). As such, it cannot really have values of 2 or 4. This indicates a limitation of the examples selected by Bodmer and Bonilla⁴.

RISCH AND MERIKANGAS ON LINKAGE AND ASSOCIATION TESTING IN THE DETECTION OF RARE VARIANTS

Risch and Merikangas⁵ compared the power of linkage and association analyses in detecting causative genes of complex disorders. As an example of linkage analysis they used the test for disproportional allele sharing in affected sib-pairs (ASP^{7,8}), as an example of an association analysis they used the transmission/disequilibrium test (TDT⁹) that enquires for preferential allele transmission from parents to affected children. They simulated these tests on a disease model with genotype frequencies $P(A) = p$ and $P(a) = q = 1 - p$ and multiplicative genotype relative risks γ and γ^2 for Aa and AA subjects, respectively.

For that purpose, Risch and Merikangas⁵ considered a random variable B with discrete values (for example, b and $-b$) that indicates the positive or negative outcome of each of M tested instances. The expected value and variance of B are given as μ and σ^2 , respectively. The sum $S = \sum B_i$ is examined to identify a genetic effect. If M is large

Table 1 Cumulative population attributable risk (PAR contribution) of variants in different frequency classes (PAR_p) according to Bodmer and Bonilla⁴ as calculated with a background disease frequency $P(D|a) \approx P(D) \approx 10\%$

Type of variant	Risk allele frequency (p)	Variants involved (n_p)	Odds ratio (OR_p)	Estimated PAR_p	Variants per individual
Common	0.5	10	1.4	4	10
Moderately rare	0.002	200	3.5	2	0.8
Moderately rare*	0.002	1000	2	4	4
Very rare*	5×10^{-7}	1000	∞ ($f_p = 1$)	0.01	0.001

Variants marked by (*) refer to the single locus example. PAR_p is estimated as $2n_p(\text{OR}_p - 1)p$ in case of common and moderately rare variants whereas in case of the very rare variants with complete penetrance ($f_p = 1$, $\text{OR}_p \rightarrow \infty$), it is estimated as $2n_p f_p p / P(D)$.

enough, S has a normal distribution with a mean of $M\mu$ and a variance of $M\sigma^2$. The minimal necessary separation of the distributions of S under the null (μ_0, σ_0) and the alternative (μ, σ) hypotheses is determined by the acceptable error rates as expressed by the intended significance α and the projected power $1-\beta$; that is,

$$M\mu + M^{1/2}\sigma Z_\beta = M\mu_0 + M^{1/2}\sigma_0 Z_{1-\alpha} \quad (8)$$

where $Z_{1-\alpha} = -Z_\alpha$ and $Z_\beta = -Z_{1-\beta}$ are quantiles of a standard normal distribution corresponding to α and β , respectively. The necessary number of instances can be calculated from Equation (8). If the test-specific definition of B (see below) is chosen such as to imply $\mu_0=0$ and $\sigma_0^2=1$ for the null hypothesis, Equation (8) yields

$$M = (Z_\alpha - \sigma Z_{1-\beta})^2 / \mu^2 \quad (9)$$

Notably, the cumulative rate of false-positives in genome-wide analyses depends on the number of tested markers. As compared to association analysis, the number of markers can be small in linkage analysis because interfering recombinations are rare within families. Risch and Merikangas⁵ applied a Bonferroni correction and arrived at commonly accepted levels of $\alpha=10^{-4}$ ($Z_{\alpha,\text{link}}=-3.72$) in linkage analysis and $\alpha=5 \times 10^{-8}$ ($Z_{\alpha,\text{assoc}}=-5.33$) in association analysis. For the power $(1-\beta)$ they chose the usual level of 80%. Despite the difference in Z_γ , they showed that for small effect size γ the necessary M is much larger in linkage analysis because the allele sharing between two affected sibs is already close to random whereas the bias in disease allele transmission to an affected still is substantial (see below).

Examining the linkage test (ASP), Risch and Merikangas⁵ assumed fully informative allele characterization and scored $B=1$ if the sib-pair shared an allele identical by descent from a parent and $B=-1$ if not. They derived the expected proportion Y of identical by descent-shared alleles to calculate $\mu=(1)Y+(-1)(1-Y)$ and $\sigma^2=(1-\mu)^2Y+(-1-\mu)^2(1-Y)$. Y was calculated from the probabilities of an ASP to share j (0, 1 or 2) alleles identical by descent: $Y=\sum_j j P(j|\text{sib_pair_aff})/2$ (see their footnote 2; watch for typing errors, however). These probabilities are given as

$$P(j|\text{sib_pair_aff}) = z_j = P(j)P(\text{sib_pair_aff}|j)/P(\text{sib_pair_aff}) \quad (10)$$

After specification of the genetic model (multiplicative; see above) and assuming random mating, the z_j can be calculated from the weighted list of genotypic combinations. For example, the list of parental mating types with all possible genotypes in two children yields $P(\text{sib_pair_aff})$. Less tediously, Risch and Merikangas⁵ (also see earlier publications of Risch) followed James¹⁰ and used a partitioning of the variance in genotypic disease frequency for calculating the correlation between relatives. Hence, $P(\text{sib_pair_aff}|0)=c^2K^2$, $P(\text{sib_pair_aff}|1)=c^2(K^2+0.5V_a)$, $P(\text{sib_pair_aff}|2)=c^2(K^2+0.5V_a+0.25V_d)$ with the mean disease frequency $cK=p^2c^2\gamma^2+2pq\gamma+q^2c=c(p\gamma+q)^2$ and the variance partitioning $c^2V=(\text{mean of squares})-(\text{square of mean})=p^2c^2\gamma^4+2pq\gamma^2+q^2c^2-c^2K^2=c^2(p\gamma^2+q)^2-c^2K^2=c^2(p(1-p)\gamma^2+p^2\gamma^2+2pq\gamma-2pq\gamma+q(1-q)+q^2)^2-c^2K^2=c^2(K+pq(\gamma-1))^2-c^2K^2=c^22pq(\gamma-1)^2K+c^2p^2q^2(\gamma-1)^4=c^2V_a+c^2V_d$, where c is a constant. With $P(j=0)=0.25$, $P(j=1)=0.5$, $P(j=2)=0.25$ and $z_2=1-z_0-z_1$, the z_j and $Y=\sum_j j z(j)/2$ can be calculated from Equation (10). The ASP analysis thus resulted in

$$\mu = 2Y - 1, \quad \sigma^2 = 4Y(1 - Y) \text{ with } Y = (1+w)/(2+w)$$

$$\text{where } w = pq(\gamma - 1)^2 / (p\gamma + q)^2 \quad (11)$$

For the examination of the association test (TDT), Risch and Merikangas⁵ used the probability h that a parent of an affected child is a heterozygote and scored $B=1/\sqrt{h}$ if the parent is heterozygous and transmitted the risk allele A , $B=-1/\sqrt{h}$ if the parent is hetero-

zygous and transmitted allele a , and $B=0$ if the parent is homozygous. (The weighting factor $1/\sqrt{h}$ has not been explained by the authors, but it can be understood in analogy to Penrose's square root principle of representative voting:¹¹ Transmissions at different loci should have the same *a priori* influence on the TDT. The influence of a transmission can be quantified as its probability to be deciding on a split result at the respective locus. This probability is proportional to the square root of the number of informative transmissions at that locus as can be derived by calculating the height of (that is, infinitesimal area under) the peak of a binomial distribution using Stirling's approximation. Of note, Penrose⁸ had encountered the problem of weighting according to informativeness in genetics already.)

To calculate μ and σ^2 as $\mu=hG(1/\sqrt{h})+h(1-G)(-1/\sqrt{h})+0=(2G-1)/\sqrt{h}$ and $\sigma^2=hG(1/\sqrt{h}-\mu)^2+h(1-G)(-1/\sqrt{h}-\mu)^2+(1-h)(0-\mu)^2=1-(2G-1)^2h$, respectively, Risch and Merikangas⁵ used the probability G for an affected child of a heterozygote parent to have received the risk allele from this parent. Under the multiplicative model with random mating, Bayes' theorem yielded $G=P(\text{transaff}_{\text{het_par}})=P(\text{trans})P(\text{aff}_{\text{het_par}}|\text{trans})/P(\text{aff}_{\text{het_par}})=0.5\gamma(p\gamma+q)c/(0.5\gamma(p\gamma+q)c+0.5(p\gamma+q)c)=\gamma/(\gamma+1)$, where c is a constant. Bayes' theorem also yielded $h=P(\text{het_par}|\text{aff})=2pq[0.5(\gamma+1)p\gamma+0.5(\gamma+1)q]/(p\gamma+q)^2=pq(\gamma+1)/(p\gamma+q)$ (see their footnote 4 that contains a typing error of the [...] bracketing, however). Thus, for the TDT on affected singletons,

$$\mu = \sqrt{h(\gamma - 1)/(\gamma + 1)},$$

$$\sigma^2 = 1 - h(\gamma - 1)^2/(\gamma + 1)^2 \quad (12)$$

with $h_{\text{singletons}} = pq(\gamma + 1)/(p\gamma + q)$

In case of the null hypothesis with $\gamma_0=1$, mean and variance are given as $\mu_0=0$ and $\sigma_0^2=1$, respectively.

For the TDT on families with affected sib-pairs (TDT_{sibs}), h has to be adapted. Risch and Merikangas⁵ indicated the formula $h_{\text{sibs}}=pq(\gamma+1)^2/(2(p\gamma+q)^2+pq(\gamma-1)^2)$ without further explanation. $P(\text{het_par}|\text{sib_pair_aff})=h_{\text{sibs}}$ can be derived using Bayes' theorem and the list of parental matings types, which implies $P(\text{sib_pair_aff}|\text{het_par})=0.25c^2(\gamma+1)^2(K+0.5pq(\gamma-1)^2)$ and $P(\text{sib_pair_aff})=c^2(K+0.5pq(\gamma-1)^2)^2$, where $K=(p\gamma+q)^2$ and c is a constant (compare ASP analysis above).

For rare variants ($p \rightarrow 0$), Equations (11) and (12) can be simplified substantially. Because with Equation (11) $w \approx p(\gamma-1)^2 \ll 1$ for $p \rightarrow 0$, we get $\mu \approx w/(2+w) \approx w/2$ and $\sigma^2 \approx 4(1+w)/(2+w)^2 \approx 1$. Thus, with Equation (9) the number of families $N_{\text{ASP}}=M/2$ is

$$N_{\text{ASP}} \approx 2(Z_{\alpha,\text{link}} - Z_{1-\beta})^2 / (p^2(\gamma - 1)^4) \quad \text{if } p \ll 1 \quad (13)$$

For the TDT on singletons, Equation (12) yields $h \approx p(\gamma+1)$ if $p \rightarrow 0$. Therefore, $\mu \approx (\gamma-1)(p/(\gamma+1))^{1/2}$ and $\sigma^2 \approx 1-p(\gamma-1)^2/(\gamma+1) \approx 1-\mu^2 \approx 1$, because $\mu^2 \ll 1$ for small p . Hence, with Equation (9), the number of families (each with two independently transmitting parents) can be approximated as

$$N_{\text{TDT,singletons}} \approx (Z_{\alpha,\text{assoc}} - Z_{1-\beta})^2 / (2p(\gamma - 1)^2/(\gamma + 1)) \quad \text{if } p \ll 1 \quad (14)$$

where—as Risch and Merikangas⁵ apparently also have done—it is assumed that the risk allele can be identified *a priori* (that is, one-sided test). In case of rare variants this assumption is warranted.

The approximation for the TDT on sib-pairs implies $h \approx p(\gamma+1)^2/2$, and consequently, $\mu \approx (\gamma-1)(p/2)^{1/2}$, $\sigma^2 \approx 1$. Considering that two

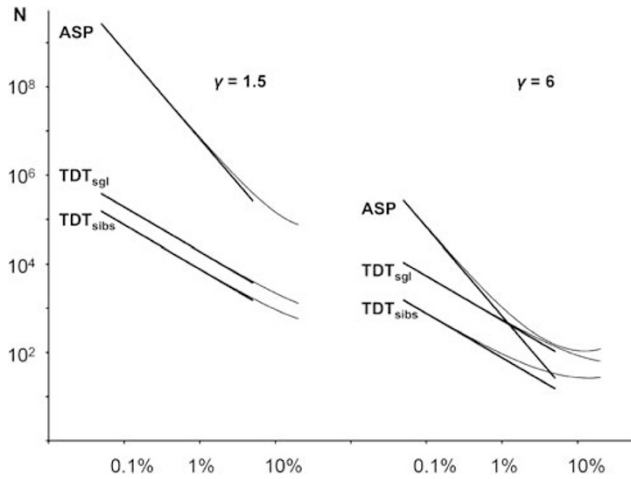


Figure 1 Necessary sample sizes (number of families) in a linkage test on affected sib-pairs (ASP) and in an association test (transmission disequilibrium test, TDT, with singletons or sibs) as a function of the allele frequency according to Risch and Merikangas⁵. Sample sizes were calculated for two different genotypic relative risks using a multiplicative disease model. Calculations were carried out as approximations for rare variants (thick lines; see text) and as exact solutions (thin lines). Also see Risch and Merikangas²⁶ for a correction of their ASP calculation.

parents transmit to two cases each and ignoring the small nonrandomness error due to the sibs' relatedness,

$$N_{TDT,sibs} \approx (Z_{\alpha,assoc} - Z_{1-\beta})^2 / (2p(\gamma - 1)^2) \quad \text{if } p \ll 1 \quad (15)$$

Equations (13–15), that is, the powers of their denominators, directly show the conclusion of Risch and Merikangas⁵ that in linkage analysis the necessary sample size is much more sensitive to a decline of risk allele frequency p or effect size γ (also see Figure 1). An allele's association with a trait (that is, preferential transmission in case of an affected child) may still be noticeable when the correlation of its transmissions to two or more affected relatives is already obscured by noise (that is, by incomplete penetrance, genocopies and phenocopies).

FREQUENCY DISTRIBUTION OF GENETIC VARIANTS

Mutation rate, selection and random drift determine the frequency distribution of genetic variants. Generalizing previous results of Wright, Kimura⁶ derived the number of heterozygous sites due to the steady flux of mutations from a diffusion theoretical approach. The number of sites with a variant of frequency p and selection coefficient s is given as

$$\Phi(s, p) = 4N_e v(s) / (p(1 - p)) (e^{-2N_e s(1-p)} - 1) / (e^{-2N_e s} - 1) \quad (16)$$

where N_e is the 'variance effective population size' and $v(s)$ the rate per gamete and generation of mutations with selection coefficient s . In this 'infinite sites model', each novel mutation occurs at a site in which a previous mutation is not still segregating. Moreover, the mutated sites are assumed to segregate independently of each other. Equation (16) was derived as stationary solution of the Kolmogorov backward equation modeling the temporal change in allele frequency as a result of selection and gamete sampling variance (drift). With Equation (16), the number of binary sites (for example, single-nucleotide polymorphisms, SNPs) is

$$\Psi(s, p) = \Phi(s, p) + \Phi(s, 1 - p) \quad (17)$$

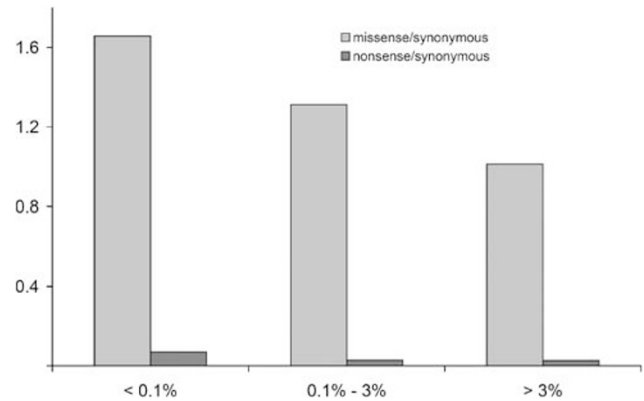


Figure 2 Fraction of missense and nonsense mutations as compared to synonymous base exchanges in different ranges of the minor allele frequency. dbSNPs (build 129) were downloaded from the UCSC browser using the corresponding filters for class (single), average heterozygosity and function. All pairwise comparisons (cross-tab test) between the ranges were highly significant except for the comparison of nonsense and synonymous mutations between the 0.1–3%-range and the > 3%-range.

Handling of Equation (16) is tedious but some of its characteristics are detected easily. With the first-order approximation $e^{-ks} \approx 1 - ks$, the frequency of neutral mutants ($s \rightarrow 0$) simplifies to

$$\Phi(0, p) = 4N_e v(0) / p; \quad \Psi(0, p) = 4N_e v(0) / (p(1 - p)) \quad (18)$$

$\Psi(0, p)$ is minimal at $p=0.5$. The relative number $\Psi(s, p) / \Psi(0, p)$ also has a single extremum at $p=0.5$; its approximation for $p \rightarrow 0$ is $v(s) / v(0) ((1 + 2N_e s p) / (1 - e^{-2N_e s}))$. These values predict the frequency relation of variants with different selection coefficients; for example, for $v(s_1) \leq v(s_2)$ and $s_1 < s_2 \leq 0$ the frequency distribution of the more deleterious s_1 variants is shifted toward lower frequencies as compared to the s_2 variants. In accordance with this prediction, Figure 2 displays the analysis of a recently updated SNP database showing that the relation of nonsynonymous (that is, more or less deleterious) SNPs to synonymous (that is, mostly neutral) SNPs increases with declining allele frequency.^{12,13} Moreover, nonsense mutations appear to rise in a lower frequency interval as compared to missense mutations, fitting the expectation that, on average, nonsense mutations should be more deleterious than missense mutations and, therefore, less likely to reach higher allele frequencies.

The distribution of variants depends on the effects of drift and selection but also on the distribution of new mutations $v(s)$. Kimura¹⁴ argued that advantageous mutations are very rare whereas nearly neutral mutations predominate. Therefore, he modeled $v(s)$ as a reflected gamma distribution that can accommodate enough probability mass in the region of nearly neutral mutations. Eyre-Walker and Keightley¹⁵ recommended to model $v(s)$ by the composition of more than one gamma function to account for the differences in selective relevance of different genome regions (for example, coding versus noncoding).

For the further analysis (see below), $v(s)$ is assumed to be composed of two reflected gamma functions with shape parameters derived from the available evidence as outlined in the legend of Figure 3.

ASSOCIATION ANALYSIS OF A DISEASE MODEL BASED ON THE INFINITE SITES MODEL

I now construct a disease model caused by variants with a distribution proportional to infinite sites model of Kimura⁶ (see above). In reality,

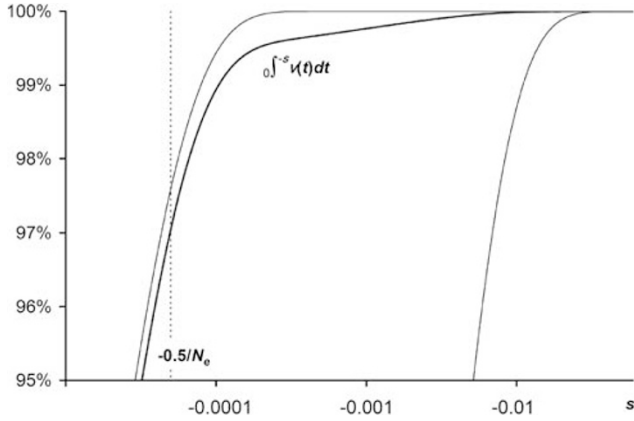


Figure 3 Semiempirical cumulative distribution (thick line) of new mutations according to their selection coefficient s , ignoring the small fraction of advantageous mutations with $s \geq 0$. The distribution density $v(s)$ was modeled by the weighted sum of two reflected gamma distributions (thin lines), $v(s) = 0.01g_1(s) + 0.99g_2(s)$, with $g_i(s) = b^b(-s)^{b-1}e^{-bs}/\Gamma(b)$. Weighting followed the percentage (1%) of coding sequence in the genome. The shape parameter ($b_1=0.2$) of $g_1(s)$ was chosen according to the analyses of Eyre-Walker and Keightley.^{15,27} Choosing an average of $\hat{s}_1 = -0.001$, only 40% of the mutations under $g_1(s)$ were neutral ($s > -1/(2N_e)$) accounting for the fact that a large fraction of nonsynonymous mutations are mildly deleterious.^{27,28} For $g_2(s)$ an even more leptokurtic shape parameter ($b_2=0.1$) and an effectively neutral average ($\hat{s}_2 = -0.1/(2N_e)$) were chosen to account for the overwhelming part of the genome that does not appear to be relevant for fitness. With these parameters, 3% of the mutations under $v(s)$ were deleterious (in keeping with Eyre-Walker and Keightley²⁹). Of these 3%, 80% belong to $g_2(s)$ but almost all of the strongly deleterious mutations belong to $g_1(s)$. The total number of new mutations per gamete was estimated to be 50 by indirect and direct analyses^{30–32}; that is, $v(s) = 50v(s)^*$.

of course, the relation between the evolutionary effect and the disease-causing potential of an allele is not that simple. Deleterious alleles may affect fertility without causing a disease and disease-causing alleles do not necessarily inflict a selective disadvantage. Nonetheless, in general, a positive correlation may be assumed. The stronger a disease-causing effect, the more likely is an early onset of the disease with an impact on fertility. Moreover, even if the average age of onset is late, disease susceptibility genes may be associated with a selective disadvantage if the variance in the age of onset is large enough. The latter explains why causative alleles in late-onset diseases, such as certain familial forms of cancer, coronary artery diseases and Alzheimer dementia, are typically recent and rare.¹⁶

To construct the genetic architecture of the disease model, a relation is needed between the selection coefficient s and the effect size. I derive a simple version of this relation using some approximations; that is, $P(D|aa) \approx P(D)$ (prevalence in the absence of the variant is close to the general prevalence of the multifactorial disorder) and $OR \approx RR$ (OR is close to RR). In keeping with Kimura^{6,17}, dominance effects are neglected, that is, the selective advantage, $W_{AA} - W_{aa}$, of homozygote mutants over the preexisting form is considered to be twice as large as the corresponding advantage, $W_{Aa} - W_{aa}$, of heterozygotes

$$\begin{aligned} W_{AA} - W_{aa} &= s \\ W_{Aa} - W_{aa} &= s/2 \end{aligned} \quad (19)$$

With frequency normalization after selection, $p^2W_{AA} + 2p(1-p)W_{Aa} + (1-p)^2W_{aa} = 1$, Equation (19) yields $W_{aa} = 1 - ps$, and the change in frequency of the variant is $\Delta p = (2 \text{freq}(AA) \text{ after selection} + \text{freq}(Aa) \text{ after selection})/2 - p = p(pW_{AA} + (1-p)W_{Aa} - 1) = 0.5sp(1-p)$ as indicated by Kimura⁶ (his equation 2 with $h=1/2$).

Assuming that selection behaves by reduced fertility of those who have the disease, the genotypic fitness parameters may be approximated as

$$\begin{aligned} W_{AA} &= 1 - P(D|AA)(1 - k) \\ W_{Aa} &= 1 - P(D|Aa)(1 - k) \\ W_{aa} &= 1 - P(D|aa)(1 - k) \end{aligned} \quad (20)$$

$$\approx 1 - P(D)(1 - k) \quad \text{if } P(D|aa) \approx P(D)$$

with k indicating the relative fertility of those who have the disease ($0 \leq k \leq 1$). Comparison of Equations (19) and (20) yields the homozygote and heterozygote OR

$$\begin{aligned} OR_{\text{hom}} &\approx P(D|AA)/P(D) \approx 1 - s/(P(D)(1 - k)) \\ &\text{if } P(D|aa) \approx P(D) \end{aligned} \quad (21)$$

$$\begin{aligned} OR_{\text{het}} &\approx P(D|Aa)/P(D) \approx 1 - s/(2P(D)(1 - k)) \\ &\text{if } P(D|aa) \approx P(D) \end{aligned} \quad (22)$$

For small values of s , Equations (21) and (22) are compatible with $OR_{\text{hom}} \approx (OR_{\text{het}})^2$ as can be shown by first-order expansion using $dOR_{\text{hom}}/ds = 2OR_{\text{het}} dOR_{\text{het}}/ds = -1/(P(D)(1 - k))$ for $s=0$.

In the following delineation, we will use only the heterozygote OR, that is, $OR \approx OR_{\text{het}}$, because we perform an approximation that neglects homozygote states in keeping with Bodmer and Bonilla.⁴ Moreover, we will calculate a specific example with disease prevalence $P(D)=1\%$ and fertility reduction in affecteds by 50% (that is, $k=1/2$). Thus,

$$OR \approx 1 - s/(2P(D)(1 - k)) = 1 - 100s \quad (23)$$

which is larger than 1 for $s < 0$. Equation (23) can now be used to calculate the PAR according to Equation (7) as

$$\begin{aligned} PAR(s, p) &\approx 2c\Psi(s, p)(OR(s) - 1)p \\ &\text{with } s < -1/(2N_e) \end{aligned} \quad (24)$$

where c is a proportionality constant and n_p has been replaced by $c\Psi(s, p)$ according to Equations (16) and (17) with a mutation rate as outlined in Figure 3. Thus, we finally arrive at the genetic architecture of the disease model by assuming that the frequency spectrum of the variants involved in the disease is proportional to the frequency spectrum of the set of all deleterious variants. The normalization condition is

$$1 = \int_0^{0.5} \int_{-1}^{-1/(2N_e)} PAR(s, p) ds dp \quad (25)$$

where the integration runs up to the limit of effective neutrality, that is, $s < 1/(2N_e)$, and across the whole frequency spectrum, $0 \leq p = p(A) \leq 0.5$.

Of course, this model is simplistic because it assumes a simple relation between the selective effect of a variant and its contribution to the disease. Moreover, following Bodmer and Bonilla⁴, it assumes only two alleles per locus, absence of linkage between loci, calculation and summation of the PAR across the whole frequency spectrum as in Equations (24) and (25), respectively, and neglectability of homozygosity. Nonetheless, it may allow for some general considerations. As an example, let us assume a disease frequency $P(D)=1\%$, a relative fertility $k=0.5$ and a constant effective population size of $N_e=10000$. With these values, the numeric integration of Equation (25) yields the normalization factor $c=1/492.9$. We now form classes of variants with different effect size, that is, a class with very weak effect ($-1/20000 \geq s > -1/2000$), a class with weak effect ($-1/2000 \geq s > -1/200$, that is, $1.05 < OR < 1.5$), a class with moderately strong effect ($-1/200 \geq s > -1/20$, that is, $1.5 < OR < 6$) and a class with strong effect

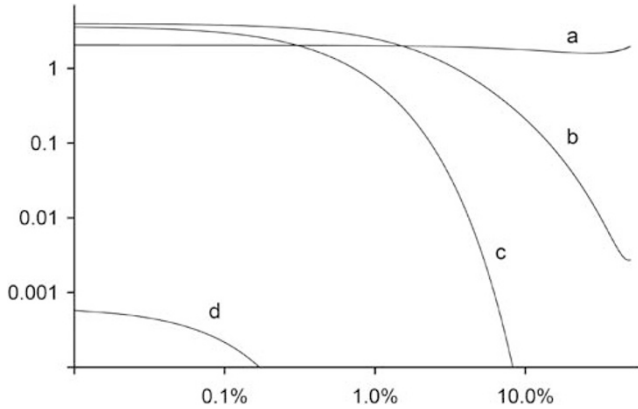


Figure 4 Density of the cumulative population attributable risk (PAR) (“PAR contribution”) of variants as a function of the allele frequency for four different groups of effect size; that is, variants with very weak (a, 1.005 < OR < 1.05), weak (b, 1.05 < OR < 1.5), moderately strong (c, 1.5 < OR < 6) and strong (d, OR > 6) effects. Calculation followed the example delineated in the main text assuming a distribution of new mutations as described in Figure 3. The largest contribution (86%) is provided by the very weak variants whereas strong effect variants contribute little to the PAR. Among the rare variants, however, the moderately strong variants (c) may reach a PAR contribution that equals those of the weak (b) and very weak (a) variants.

($-1/20 \geq s$). Figure 4 shows the contributions, $\int_{s_1}^{s_2} PAR(s,p)ds$, of these variants to the PAR density distribution as a function of their frequency p . The largest part of the PAR sum is contributed by very weak effect variants.

In the next step, the frequency range is also subdivided into several groups; that is, from common to rare. The analysis shown in Figure 5 comprised seven frequency segments crossed with seven groups of variants of different effect sizes. For each of the $7 \times 7 = 49$ cells the number of variants, their average effect and their average frequency were determined. These numbers were used to calculate the sample size $N_{TDT, \text{singletons}}$ that is required for an association analysis of the disease model according to Equations (9), (12) or (14; for an approximation). The number of contributing variants in the cell delimited by p_i, p_j, s_m and s_n is given as

$$\Omega_{ijmn} = \int_{p_i}^{p_j} \int_{s_m}^{s_n} c\Psi(s, P) ds dp \quad (26)$$

according to Equations (16) and (17) and which is derived by numeric integration. Also by numeric integration, the average frequency and the average selection coefficient in each cell were derived as harmonic means (p_h, s_h)

$$p_{h_{ijmn}} = \Omega_{ijmn}^{-1} \left(\int_{p_i}^{p_j} \int_{s_m}^{s_n} c\Psi(s, p) / p ds dp \right)^{-1} \quad (27)$$

$$s_{h_{ijmn}} = \Omega_{ijmn}^{-1} \left(\int_{p_i}^{p_j} \int_{s_m}^{s_n} c\Psi(s, p) / s ds dp \right)^{-1}$$

(The use of arithmetic means leads to qualitatively similar results.)

The genotypic relative risk γ used in Equations (9), (12) and (14) was approximated by the OR (\approx RR) calculated from the average selection coefficient s_h according to Equation (23):

$$\gamma(s_h) = 1 - 0.5s_h/P(D)(1 - k) \quad (28)$$

The aim of the association analysis was set to identify at least one contributing variant with a power of 80%; that is, the probability that all truly associated variants test negative had to be less than 20%. Thus, besides the significance level $\alpha=0.05$ that was adapted to the

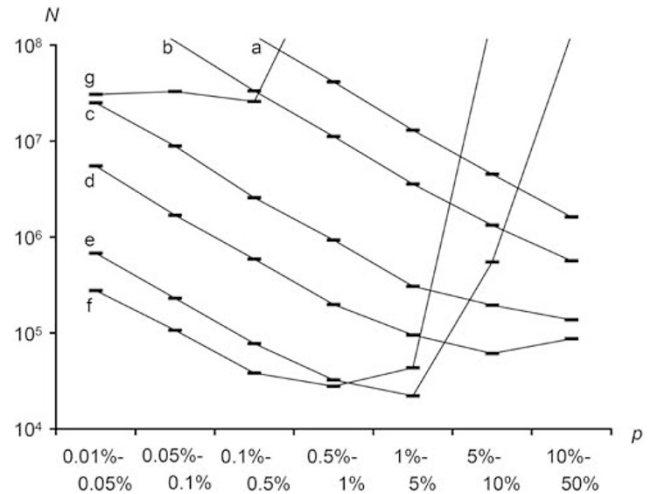


Figure 5 Sample size N necessary for a transmission disequilibrium (TDT) association analysis of a disease model to detect at least one contributing variant with a genome-wide significance of 0.05 and a power of 0.8 in different segments of allele frequency p . Variants are assumed to be distributed according to the infinite sites model (see Figure 3 and main text for more details). The graphs a–g indicate classes of variants with different effect sizes as in Figure 4 but with a finer subdivision (a, 1.005 < OR < 1.01; b, 1.01 < OR < 1.05; c, 1.05 < OR < 1.1; d, 1.1 < OR < 1.5; e, 1.5 < OR < 2; f, 2 < OR < 6; g, 6 < OR). The minimum is reached by moderately strong variants (graphs e and f) in the 1% frequency range. Up to this range, at least about one of such variants can be expected to contribute to the disease. Above this range, the expected number of contributing variants is low ($\ll 1$) as is revealed by the steep increase of the necessary sample size. Strong variants (g) do not reach a sufficiently large number at frequencies above 0.01% due to strong selective repression.

genome-wide multiple testing scheme, the power $1 - \beta = 0.8$ was adapted to the fact that there is not exactly one true positive. For correction of α the total number A_{ij} of variants in the frequency range $[i, j]$ was calculated. It can be approximated from above using Equation (18) by assuming that all novel mutations are neutral

$$\Lambda_{ij} = \int_{p_i}^{p_j} \int_{-1}^0 \Psi(s, p) ds dp \approx 4N_e v_n \int_{p_i}^{p_j} 1/(p(1 - p)) dp \quad (29)$$

with $v_n = \int_{-1}^0 v(s) ds$. The Bonferroni–Sidak correction of α then was given as

$$\alpha \rightarrow \alpha_{ij}, \alpha = 1 - (1 - \alpha_{ij})^{A_{ij}} \quad (30)$$

with $\alpha_{ij} \approx \alpha/A_{ij}$. Analogously, the power level was corrected using the number of contributing variants as calculated in Equation (26),

$$\beta \rightarrow \beta_{ijmn}, \beta = (\beta_{ijmn})^{\Omega_{ijmn}} \quad (31)$$

Here, $(1 - \alpha_{ij})^{A_{ij}}$ is the probability that no test is false positive, and $(\beta_{ijmn})^{\Omega_{ijmn}}$ is the probability that none of the true positives is detected in the analysis (if there is no statistical interference among them).

Thus, the necessary sample size $N_{TDT, \text{singletons}}$ for an association analysis on variants in the effect size interval $[s_m, s_n]$ and the frequency segment $[p_i, p_j]$ can be approximated with Equations (14), (28), (30) and (31):

$$N_{ijmn} \approx (Z_{\alpha_{ij}} - Z_{1 - \beta_{ijmn}})^2 / (2p_{h_{ijmn}}(\gamma(s_{h_{ijmn}}) - 1)^2 / (\gamma(s_{h_{ijmn}}) + 1)) \quad (32)$$

For the analysis shown in Figure 5, the precise solutions were used as indicated by Equations (9) and (12). Figure 5 shows that the

necessary sample sizes are smallest for variants with moderately strong effect in the 1% allele frequency range.

DISCUSSION

Detecting rare variants demands large samples, a well-known fact that is made obvious once more in the approximations for the association (TDT) and linkage (ASP) tests provided here (Equations (13–15), Figure 1). Less well-known, perhaps, is the fact that the rareness of the variants itself inflates the sample size if the analysis is extended to ever more rare variants, and not the multiple testing correction due to the increasing number of variants. This is evident for linkage analysis where the number of tests has an upper limit due to linkage disequilibrium across long genome segments, but it is also true for association tests with linkage disequilibrium being ignored: Because almost all new mutations are neutral or nearly neutral (cf. Figure 3), the number of variants can be approximated with Equation (29) as $A(p) \approx \int_p^{0.5} 4N_e v_n / (p(1-p)) dp = 4N_e v_n \ln(1/p) = 2 \times 10^6 \ln(1/p)$, where the number v_n of new mutations is approximated by 50 neutral mutations per gamete. The α level in an association test is thus corrected to $\alpha'(p) \approx 0.05/A(p) = 0.25 \times 10^{-7} \ln(1/p)^{-1}$. Because this is very small, the quantile $Z_{\alpha'}$ (see Equations (8) and (9)) can be approximated¹⁸ as $Z_{\alpha'(p)} \approx -(2 \ln(1/\alpha'))^{1/2} = (35 + 2 \ln(\ln(1/p)))^{1/2}$ being in the range of 6.2–6.4 for $10^{-2} < p < 10^{-10}$ and, thus, almost constant for all realistic levels of rare variants. Candidate space reduction strategies must be ever more stringent, therefore, to substantially ameliorate the sample size problem, and genomic enrichment schemes such as exome sequencing may need to imply further measures such as the formation of equivalence classes of variants, for example.

The approximations (Equations (13–15), Figure 1) also show that the inflation due to variants' rareness is much stronger in case of linkage tests such as the ASP. The sample size is proportional to $(1/p)^2$ in case of the ASP while it is proportional to $1/p$ in case of the association test (TDT). Similarly, linkage analysis is much more sensitive to a decline in effect size γ . This has led to the assumption that linkage analysis cannot be helpful in localizing causative mutations if their penetrance is incomplete as in case of variants that contribute to common multifactorial traits. However, this rule should be applied carefully: (1) Intermediate phenotypes that contribute to a multifactorial trait may individually be determined by factors that are detectable by linkage analysis. Smirnov *et al.*,¹⁹ for instance, identified linkage signals (mostly *in trans*) with genome-wide significance for one third of 3280 molecular phenotypes defined as 1.5-fold change of gene expression upon radiation. (2) Allelic series of mutations may affect the same gene so that linkage analysis of variants with high penetrance localizes genes that also comprise low effect variants.²⁰ (3) Linkage analysis of a recessive phenotype may unravel a variant of low penetrance in the heterozygote state (for example, the C282Y mutation of the hemochromatosis gene *HFE*).^{21,22}

The genetic architecture of traits and diseases thus governs the method and the expenditure to be used for their analyses. General assumptions on the architecture of multifactorial diseases have been made before. Bodmer and Bonilla⁴, for instance, inferred that moderately rare variants with moderately strong effect sum up to a substantial PAR. The latter—which they used as a measure of the contribution to the disease generation (PAR contribution)—was comparable to the contribution by common variants (see Table 1). However, their study implied somewhat arbitrary assumptions on the numbers of contributing variants in different effect size classes. Here, I analyzed a disease model with a distribution of such contributing variants proportional to the evolutionary distribution of

non-neutral variants and with effect sizes being related in a simple manner to the selection coefficients. In this model, the PAR contribution declined substantially with the effect size, with the largest contribution being made by a large number of very weak variants distributed over all frequency ranges (Figure 4).

With rare variants being included in the association analysis (for example, by next-generation sequencing), the identification of a moderately strong variant may be achievable with the comparatively smallest sample size: My investigation indicated a sample size minimum for this effect size class in the frequency range of about 1% (see Figure 5). This minimum is due to a balance between frequency and number of contributing variants. Figure 5 suggests that most contributing variants are not accessible with reasonable sample sizes whereas some harvesting of moderately rare, moderately strong variants may be possible.

The genetic architecture of real disorders probably is not as simple and uniform as delineated in the model presented here. The influence of selection may vary substantially between traits. Moreover, the dissection of complex traits in animal models using chromosome substitution strains has shown that the number of strong effects may be larger than previously assumed and that substantial epistatic interactions account for the subadditivity of these effects.²³

Nonetheless, the detection by association analysis of more than a few effects contributing to a phenotype will demand very large samples. To limit the sample size, several strategies have been proposed. One, of course, is the formation of equivalence classes of variants; for example, by treating different variants from the same gene or the same pathway as units for the association test.²⁴ Enrichment strategies in proband selection might also be successful: Crow²⁵ proposed to examine patients with old fathers because, as compared to common variants or environmental factors, new mutations (as such being of the rare and possibly strong type) may have a more prominent function in these cases than in cases with young fathers. Moreover, environmental factors might be of reduced relevance in young cases because, in general, they need time to operate. Finally, choosing controls and cases from subpopulations with and without environmental risk factors, respectively, (for example, nonsmoking lung cancer cases versus smoking controls) might also enrich for causative and protective variants, respectively, in a way that increases the power of an association analysis.

ACKNOWLEDGEMENTS

I thank Professor Thomas Meitinger for an ongoing series of most stimulating and instructive discussions on problems in genetics. I also thank two anonymous reviewers for their useful comments.

- 1 Tolstoy, L. *Anna Karenina* (English translation by Garnett, C.) 5 (Barnes & Noble, New York, 2003).
- 2 Darwin, C. R. *On the Origin of Species By Means of Natural Selection, or the Preservation of Favoured Races in the Struggle For Life* (John Murray, London, 1859).
- 3 Mendel, J. G. Versuche (ber Pflanzen-Hybriden). *Verhandlungen des naturforschenden Vereines in Brünn* **4**, 3–47 (1866).
- 4 Bodmer, W. & Bonilla, C. Common and rare variants in multifactorial susceptibility to common diseases. *Nat. Genet.* **40**, 695–701 (2008).
- 5 Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517 (1996).
- 6 Kimura, M. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics* **61**, 893–903 (1969).
- 7 Penrose, L. S. The detection of autosomal linkage in data which consists of pairs of brothers and sisters of unspecified parentage. *Ann. Eugen.* **6**, 133–138 (1935).
- 8 Penrose, L. S. A further note on the sib-pair linkage method. *Ann. Eugen.* **13**, 25–29 (1946).

- 9 Spielman, R. S., McGinnis, R. E. & Ewens, W. J. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am. J. Hum. Genet.* **52**, 506–516 (1993).
- 10 James, J. W. Frequency in relatives for an all-or-none trait. *Ann. Hum. Genet. (Lond.)* **35**, 47–49 (1971).
- 11 Penrose, L. S. Elementary statistics of majority voting. *J. R. Stat. Soc.* **109**, 53–57 (1949).
- 12 Cargill, M., Altshuler, D., Ireland, J., Sklar, P., Ardlie, K., Patil, N. *et al.* Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat. Genet.* **22**, 231–238 (1999).
- 13 Halushka, M. K., Fan, J. B., Bentley, K., Hsie, L., Shen, N., Weder, A. *et al.* Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nat. Genet.* **22**, 239–247 (1999).
- 14 Kimura, M. Model of effectively neutral mutations in which selective constraint is incorporated. *Proc. Natl Acad. Sci. USA* **76**, 3440–3444 (1979).
- 15 Eyre-Walker, A. & Keightley, P. D. The distribution of fitness effects of new mutations. *Nat. Rev. Genet.* **8**, 610–618 (2007).
- 16 Pavard, S. & Metcalf, C. J. Negative selection on *BRCA1* susceptibility alleles sheds light on the population genetics of late-onset diseases and aging theory. *PLoS ONE* **21**, e1206 (2007).
- 17 Kimura, M. Change of gene frequencies by natural selection under population number regulation. *Proc. Natl Acad. Sci. USA* **75**, 1934–1937 (1978).
- 18 Hastings, C. *Approximations for Digital Computers* (Princeton University Press, Princeton, 1995).
- 19 Smirnov, D. A., Morley, M., Shin, E., Spielman, R. S. & Cheung, V. G. Genetic analysis of radiation-induced changes in human gene expression. *Nature* **459**, 587–591 (2009).
- 20 Pfeufer, A., Sanna, S., Arking, D. E., Müller, M., Gateva, V., Fuchsberger, C. *et al.* Common variants at ten loci modulate the QT interval duration in the QTSCD Study. *Nat. Genet.* **41**, 407–414 (2009).
- 21 Simon, M., Le Mignon, L., Fauchet, R., Yaouanq, J., David, V., Edan, G. *et al.* A study of 609 HLA haplotypes marking for the hemochromatosis gene: (1) mapping of the gene near the HLA-A locus and characters required to define a heterozygous population and (2) hypothesis concerning the underlying cause of hemochromatosis-HLA association. *Am. J. Hum. Genet.* **41**, 89–105 (1987).
- 22 Whitfield, J. B., Cullen, L. M., Jazwinska, E. C., Powell, L. W., Heath, A. C., Zhu, G. *et al.* Effects of HFE C282Y and H63D polymorphisms and polygenic background on iron stores in a large community sample of twins. *Am. J. Hum. Genet.* **66**, 1246–1258 (2000).
- 23 Shao, H., Burrage, L. C., Sinasac, D. S., Hill, A. E., Ernest, S. R., O'Brien, W. *et al.* Genetic architecture of complex traits: large phenotypic effects and pervasive epistasis. *Proc. Natl Acad. Sci. USA* **105**, 19910–19914 (2008).
- 24 Kryukov, G. V., Shpunt, A., Stamatoyannopoulos, J. A. & Sunyaev, S. R. Power of deep, all-exon resequencing for discovery of human trait genes. *Proc. Natl Acad. Sci. USA* **106**, 3871–3876 (2009).
- 25 Crow, J. F. The origins, patterns and implications of human spontaneous mutation. *Nat. Rev. Genet.* **1**, 40–47 (2000).
- 26 Risch, N. & Merikangas, K. (Reply to comments). *Science* **275**, 1329–1330 (1997).
- 27 Eyre-Walker, A., Woolfit, M. & Phelps, T. The distribution of fitness effects of new deleterious amino acid mutations in humans. *Genetics* **173**, 891–900 (2006).
- 28 Kryukov, G. V., Pennacchio, L. A. & Sunyaev, S. R. Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *Am. J. Hum. Genet.* **80**, 727–739 (2007).
- 29 Eyre-Walker, A. & Keightley, P. D. High genomic deleterious mutation rates in hominids. *Nature* **397**, 344–347 (1999).
- 30 Vogel, F. & Rathenberg, R. Spontaneous mutation in man. *Adv. Hum. Genet.* **5**, 223–318 (1975).
- 31 Kondrashov, A. S. Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases. *Hum. Mutat.* **21**, 12–27 (2002).
- 32 Xue, Y., Wang, Q., Long, Q., Ng, B. L., Swerdlow, H., Burton, J. *et al.* Human Y chromosome base-substitution mutation rate measured by direct sequencing in a deep-rooting pedigree. *Curr. Biol.* **19**, 1453–1457 (2009).