Int. J. Epidemiol. Advance Access published July 25, 2005

Published by Oxford University Press 2005

International Journal of Epidemiology
doi:10.1093/ije/dyi130

How many genes underlie the occurrence of common complex diseases in the population?

Quanhe Yang,[1]* Muin J Khoury,[2] JM Friedman,[3] Julian Little[4] and W Dana Flanders[5]

Accepted 31 May 2005

Background Most common human diseases are due to complex interactions among multiple genetic variants and environmental risk factors. There is debate over whether variants of a relatively small number of genes, each with weak or modest individual effects, account for a large proportion of common diseases in the population, or whether a large number of rare variants with large effects underlie genetic susceptibility to these diseases. It is not clear how many genes are necessary to account for an appreciable population-attributable fraction of these diseases.

Methods In this analysis, we estimated the number of disease susceptibility genes needed to account for varying population attributable fractions of a common complex disease, taking into account the genotype prevalence, risk ratios for individual genes, and the model of gene–gene interactions (additive or multiplicative).

Results Very large numbers of rare genotypes (e.g. those with frequencies of 1 per 5000 or less) are needed to explain 50% of a common disease in the population, even if the individual risk ratios are large (RR = 10–20). On the other hand, only ~20 genes are usually needed to explain 50% of the burden of a disease in the population if the predisposing genotypes are common (⩾25%), even if the individual risk ratios are relatively small (RR = 1.2–1.5).

Conclusions Our results suggest that a limited number of disease susceptibility genes with common variants can explain a major proportion of common complex diseases in the population. Our findings should help focus the search for common genetic variants that provide the most important predispositions to complex human diseases.

Keywords Epidemiology, aetiology, genes, population attributable fraction

The rapid identification of genes that are associated with human diseases has revolutionized the field of medicine, providing more accurate diagnosis, prevention opportunities, and the potential for improved treatments.[1] The development of human genome research has been accompanied by a shift of attention from the classical model of discovering loci involved in single-gene disorders (Mendelian traits) to elucidation of multiple genetic factors of small effect involved in common complex diseases.

Most common diseases occur as a result of complex interactions among multiple genetic and environmental predisposing factors.[2–5] The present study provides a general framework for estimating the number of genes needed to account for an appreciable proportion of a disease in the population. The common-disease–common-variant hypothesis holds that the genetic predisposition to common diseases results from multiple, relatively common genetic variants with small or modest effects.[5,6] An alternative, the heterogeneity hypothesis, maintains that the genetic predisposition to common diseases is caused by many different rare genetic variants, with a relatively

[1] National Center on Birth Defects and Developmental Disabilities, Centers for Disease Control and Prevention (CDC), Atlanta, GA 30333, USA.

[2] Office of Genomics and Disease Prevention, Centers for Disease Control and Prevention (CDC), Atlanta, GA 30333, USA.

[3] Department of Medical Genetics, University of British Columbia, Vancouver, BC, Canada.

[4] Department of Epidemiology and Community Medicine, University of Ottawa, Ottawa, Canada.

[5] Department of Epidemiology, School of Public Health, Emory University, Atlanta, GA, USA.

* Corresponding author. National Center on Birth Defects and Developmental Disabilities, Centers for Disease Control and Prevention (CDC), 1600 Clifton Road, Mail-Stop E-86, Atlanta, GA 30333, USA. E-mail: qay0@cdc.gov

large effect produced by each allele.[4] We do not yet know which of these hypotheses is more often correct, or the extent to which some combination of rare alleles with relatively large effects and common variants with small effects may be important.

Identification of these disease-associated genetic variants represents a high public health priority because of the contribution these common conditions make to the total burden of disease in the population.[7] However, it is not clear how many genetic variants are needed to produce an appreciable population attributable fraction (PAF) of these diseases. The PAF may be defined as the proportion of disease cases in a population that result from the effects of a risk factor, in this case the genetic predisposition. In this analysis, we estimated the number of disease susceptibility genes needed to account for varying PAFs for a common complex disease, taking into account the number of genes involved, the genotype prevalence, the risk ratios for individual genetic variants, and the model of gene–gene interactions.

Methods

For simplicity of illustration, we consider N independent biallelic disease susceptibility loci; each disease-predisposing gene is assumed to have the same prevalence and risk ratio. We also assume that there is only one at-risk genotype for each disease susceptibility locus. We use 'disease susceptibility gene' and 'genetic variant' interchangeably, and both terms refer to the one at-risk genotype for each locus. The effect of each susceptibility genotype may be dominant or recessive, so it is important to note that G is the frequency of the at-risk genotype, not the allele.

Let G be the population prevalence of the susceptibility genotype at each locus (0 = variant absent and 1 = variant present) and R_g be the lifetime risk ratio for disease for genotype = 1 compared with genotype = 0 at one locus. We assume no confounding or competing risks for R_g. We also assume that the lifetime risk of disease reflects the joint effects of measured genetic variants at N unlinked loci, along with other unmeasured factors. In reality, many genes/loci, environmental exposures and gene–gene/gene–environment interactions are probably involved in common diseases. We assume that the effects of additional genes and exposures are not directly measured here as part of the risk characterization equations.

For N independent disease susceptibility genotypes, the population can be partitioned into 2^N strata, with a different genotype prevalence and disease risk associated with each stratum. The lifetime risk of disease (D) in the population as a whole is a function of the size and disease risk associated with each stratum.[8] We assume that interactions among multiple disease susceptibility genes may occur for most common diseases, but we do not know how these joint effects operate. We recognize that any set of predisposing genotypes may interact in a variety of different ways, but for simplicity we consider these joint effects on either a purely additive or purely multiplicative scale in this analysis.[9] We also assume that these interactive effects are of the same magnitude for all genotypes involved.

Additive effects model

To illustrate the additive effects model, we consider two disease susceptibility genes, G1 and G2. Let R_{g11}, R_{g10} and R_{g01} be risk ratios of people having both genes, gene 1 only (G1) or gene 2 only (G2), respectively. The state of no interaction on an additive scale is given as: $(R_{g11} - 1) = (R_{g10} - 1) + (R_{g01} - 1)$ or $R_{g11} = R_{g10} + R_{g01} - 1$.

Assuming additive joint effects of multiple disease susceptibility genes, the lifetime risk in the population as a whole of a common disease (D) involving N genes can be modelled as:

$$\begin{aligned} D = I \Bigg[& (1-G)^N + \binom{N}{1} G(1-G)^{(N-1)} R_g + \binom{N}{2} G^2 (1-G)^{(N-2)} \\ & \times 2(R_g - 1) + \cdots + \binom{N}{j} G^j (1-G)^{(N-j)} (jR_g - (j-1)) \\ & + \binom{N}{N} G^N (NR_g - (N-1)) \Bigg] \\ = I \sum_{j=0}^{N} & \frac{N!}{j!(N-j)!} G^j (1-G)^{(N-j)} (jR_g - (j-1)), \end{aligned} \tag{1}$$

where I is the background risk of disease in the absence of the N susceptibility genotypes and J ($j = 0, 1, 2, \ldots, N$) indicates the number of disease susceptibility genotypes.

Multiplicative effects model

The state of no interaction on a multiplicative scale is given as: $R_{g11} = R_{g10} * R_{g01}$. Assuming multiplicative joint effects of multiple disease susceptibility genes, the lifetime population risk of a common disease (D) involving N genes might be modelled as:

$$\begin{aligned} D = I \Bigg[& (1-G)^N + \left(\frac{1}{N}\right) G(1-G)^{(N-1)} R_g + \left(\frac{2}{N}\right) G^2 (1-G)^{(N-2)} \\ & \times R_g^2 + \cdots + \left(\frac{j}{N}\right) G^j (1-G)^{(N-j)} R_g^j + \left(\frac{N}{N}\right) G^N R_g^N \Bigg] \\ = I \sum_{j=0}^{N} & \frac{N!}{j!(N-j)!} G^j (1-G)^{(N-j)} R_g^j, \end{aligned} \tag{2}$$

where I and J are defined as in the Equation (1).

Estimating N

For a given lifetime risk (D), genotype prevalence (G), number of susceptibility genes (N) and risk ratio (R_g), we can solve Equation (1) or (2) for I. In a hypothetical population with multiple disease susceptibility loci, there will be some number of genes N, given any particular combination of background disease risk I and risk ratio R_g that satisfies $I * [N * R_g - (N-1)] > 1$ for the additive model or $I * R_g^N > 1$ for the multiplicative model, i.e. the background risk I multiplied by stratum-specific risk for disease exceeds 100%. For values of j, for which $I * [j * R_g - (j-1)] \geqslant 1$ for the additive model (or $I * R_g^j \geqslant 1$ for the multiplicative model), we define the risk to be 1.

Population attributable fraction

In epidemiological research, PAF (also called attributable risk or aetiologic fraction) is usually defined as the proportion of disease cases in a population that would be prevented if an exposure were eliminated, assuming the exposure to be causal. In applying this concept to genetic predispositions to disease, we recognize that genetic risk factors cannot be removed, but interventions could be developed on the basis of information about the genotype. Therefore, we define PAF for a genetic predisposition to disease as the proportion of disease cases in a population that would not occur if interventions prevented the occurrence of

adverse effects of the genetic variants in that population. For the disease risk model we used, we can define the PAF as:

$$\mathrm{PAF} = \frac{D - I}{D}, \tag{3}$$

where D is the population lifetime risk of disease and I is the background risk of disease in the absence of the genetic susceptibility variants.[10]

To estimate the number of genes needed to achieve a particular PAF with varying genotype prevalence G and risk ratios R_g, we need to solve Equation (1) or (2) for N. We have not found a closed form for N that corresponds to the PAF, so we developed a simple computing algorithm to estimate N for any given PAF. For example, to estimate the number of genes needed to achieve a PAF of 30% (target PAF) for a lifetime disease risk of 5% ($D = 5\%$), genotype prevalence 10% ($G = 10\%$), and risk ratio 1.5 ($R_g = 1.5$) using a multiplicative model [Equation (2)], we start with one gene and solve Equation (2) for I, then use Equation (3) to calculate the PAF and check if the calculated PAF is less than, equal to, or greater than the target PAF (30% in this example). If the calculated PAF is less than the target PAF, we increase N to two genes, solve Equation (2) for I, then recalculate the PAF to see if it is less than, equal to, or greater than the target PAF. We repeat this process until the calculated PAF is equal to or greater than the target PAF. In this example, nine or more genes are needed to produce a PAF >30%.

The PAF calculated for the number of genes determined by this algorithm often is greater than the target PAF, especially if the genotype is common (prevalence ⩾30%). For example, the PAF calculated for nine genes in the example given above is actually 32.3% (not 30% exactly). If the genotype prevalence in this example were 50% instead of 10%, the estimated PAF would be 36% for three genes.

Results

For common diseases involving multiple susceptibility genetic variants with weak to moderate effects ($R_g = 1.2$–1.5),[11–13] the genotype prevalence plays a dominant role in determining the number of genes needed to account for an appreciable PAF. For genotype frequencies of 10%, the number of genes needed to explain 50% of the burden of disease in the population ranges from 15 to 50 (Table 1). For very common genotypes ($G \geqslant 30\%$), only 10–20 genes are needed to achieve a PAF of 50%, even if the effect size for each gene is weak ($R_g = 1.2$), and regardless of whether the genes exert additive or multiplicative joint effects.

As few as five disease susceptibility genes with risk ratios in the range of 1.01–2.00 will often produce a PAF of ⩾30% if the genotype prevalence is very common ($G \geqslant 30\%$, $D = 5\%$) (Figure 1). For 10 genes, the expected PAF for a disease with a population risk of 5% is almost always ⩾30% when calculated using our model (Figure 2), but individual genotype prevalences of ⩽1% predict that people who have all or even most of these 10 susceptible genotypes will probably never be observed. For example, if there are 10 susceptibility genotypes, each with a population prevalence of 1%, the expected frequency of people with all 10 susceptibility genotypes would be 10^{-20}. In reality, most people would have various subsets of the 10 susceptibility genotypes.[14]

Table 1 Number of genes needed to achieve a given population attributable fraction (PAF) for a complex disease with lifetime risk of 5% in the population and high genotype prevalences and low risk ratios for each gene

	PAF			
Genotype prevalence	5%	10%	30%	50%
Additive model				
Risk ratio = 1.2				
1%	27	56	215[a]	500[a]
5%	6	12	43	100
10%	3	6	22	50
20%	2	3	11	25
30%	1	2	8	17
50%	1	2	5	10
Risk ratio = 1.5				
1%	11	23	86[a]	200[a]
5%	3	5	18	40
10%	2	3	9	20
20%	1	2	5	10
30%	1	1	3	7
50%	1	1	2	4
Multiplicative model				
Risk ratio = 1.2				
1%	26[b]	53[b]	179[b]	347[b]
5%	6	11	36[b]	70[b]
10%	3	6	19	36[b]
20%	2	3	10	18
30%	1	2	7	12
50%	1	2	4	8
Risk ratio = 1.5				
1%	11[b]	22[b]	72[b]	140[b]
5%	3	5	15	29[b]
10%	2	3	8	15[b]
20%	1	2	4	8
30%	1	1	3	5
50%	1	1	2	4

[a] Indicates that $I * [j * R_g - (j - 1)] \geqslant 1$ for some values of j (i.e. lifetime disease risk is >100%); for these values of j we set risk = 1.

[b] Denotes that $I * R_g^N \geqslant 1$ for some values of j (i.e. lifetime disease risk is >100%); for these values of j we set risk = 1.

If the susceptibility genotypes are rare (e.g. 1 per 5000), many genes ($N = 183$–556) are needed to explain 50% of a common disease in the population, even with large individual risk ratios ($R_g = 10$–20) (Table 2). Many of the combinations of G, RR, and PAF given in Table 2 predict a lifetime risk for disease >100% among individuals with all of the susceptibility genotypes, indicating an inappropriate assumption about the joint effects. For example, assuming a population lifetime disease risk of 0.1% and rare susceptibility genotypes, most of the estimates for number of genes needed to achieve an appreciable PAF (PAF ⩾10%) for the multiplicative model have $I * R_g^N \geqslant 1$, i.e. the risk is >100% to develop the disease. We used a much lower

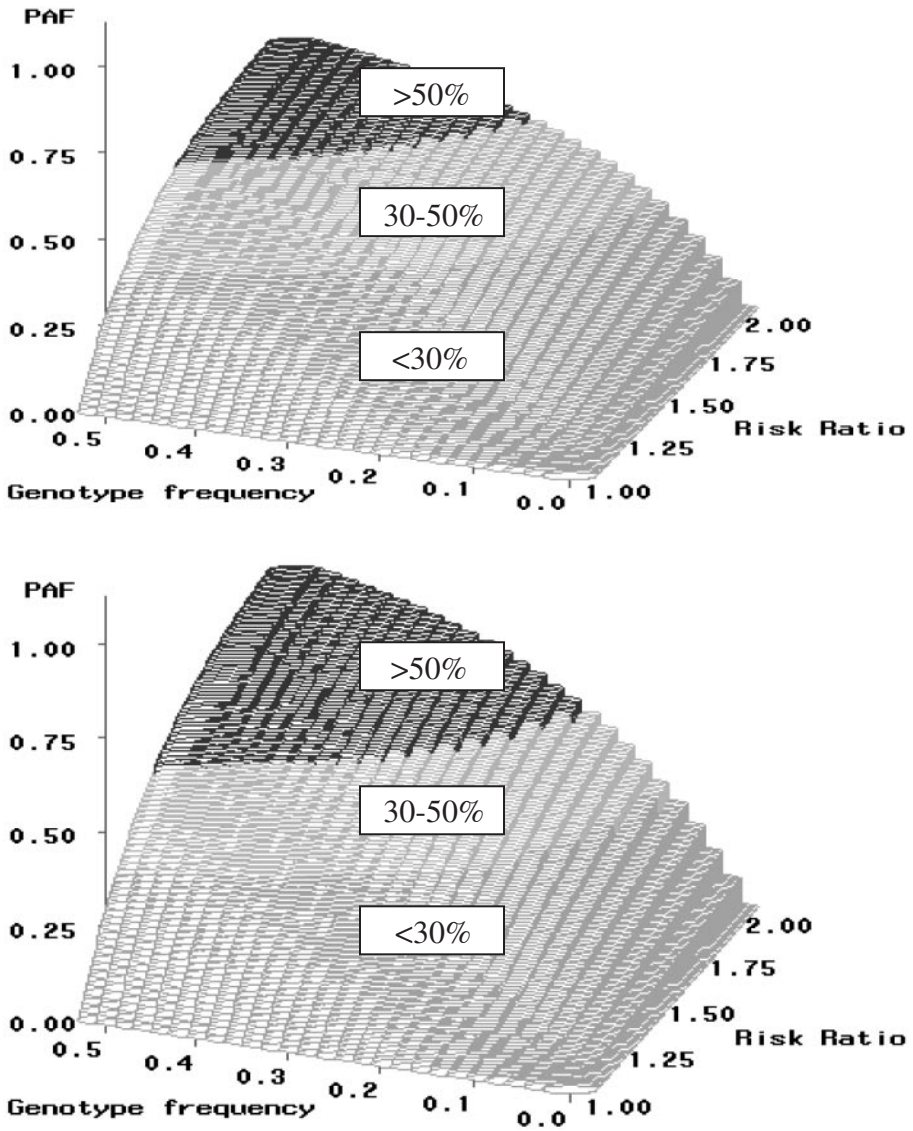

Figure 1 Expected population-attributable fraction (PAF) of five disease susceptibility genes with varying genotype prevalences and risk ratios assuming an additive (top) or multiplicative (bottom) joint effects model and lifetime risk of disease = 5.0%

lifetime risk for disease in Table 2 than in Table 1 (0.1% vs 5%) because the predicted risk for disease for almost all scenarios would be >100% if a common lifetime risk, e.g. 5%, were used, although the estimated number genes needed for any PAF would remain unchanged. This demonstrates a limitation of our model in the situation of multiple rare alleles with high risk ratios.

Example

In the United States, colorectal cancer is the fourth most common cancer, with an estimated annual incidence of 55.1 per 100 000 population and a lifetime risk of 5.7% in 2000.[15] A meta-analysis examined 30 genetic variants in 20 different genes for colorectal cancer susceptibility.[16] The study suggested that seven genetic variants were associated with the risk of colorectal cancer. We excluded the aldehyde dehydrogenase 2 (*ALDH2*) gene, which is only prevalent among Asians, and the tumour protein p53 (*TP53*) gene, with which an association had only been found in one study.[16] We included the remaining five colorectal cancer-associated genetic variants in our example: c-Ha-ras1 proto-oncogene (*HRAS*, rare allele), glutathione *S*-transferase theta 1 (*GSTT1*, null allele), tumour necrosis factor alpha-chain (*TNF*-α, a2 allele), *N*-acetyl transferase-2 (*NAT2*; fast acetylation phenotype) and 5,10-methylenetetrahydrofolate reductase gene (*MTHFR*, lack of C677T variant).

As shown in Table 3, the prevalence of the four genotypes and one phenotype (*NAT2*) considered ranges from 4.0 to 60.0%, the odds ratios range from 1.4 to 2.7 and the PAF for each genetic variant considered alone ranges from 6.3 to 29.1%. Assuming that the effects of these five genetic variants are independent, the population can be partitioned into 32 strata (Appendix Table A1). The lifetime risk of colorectal cancer (D) in the population as a whole is a function of each stratum's size (G) and associated risk (R_g) [Equation (1) or (2)]. Solving Equation (1) or (2) for I and using Equation (3) to calculate the PAF for the five genetic variants together, we estimate a combined PAF of 53.9% (95% CI 28.7–68.9%) assuming

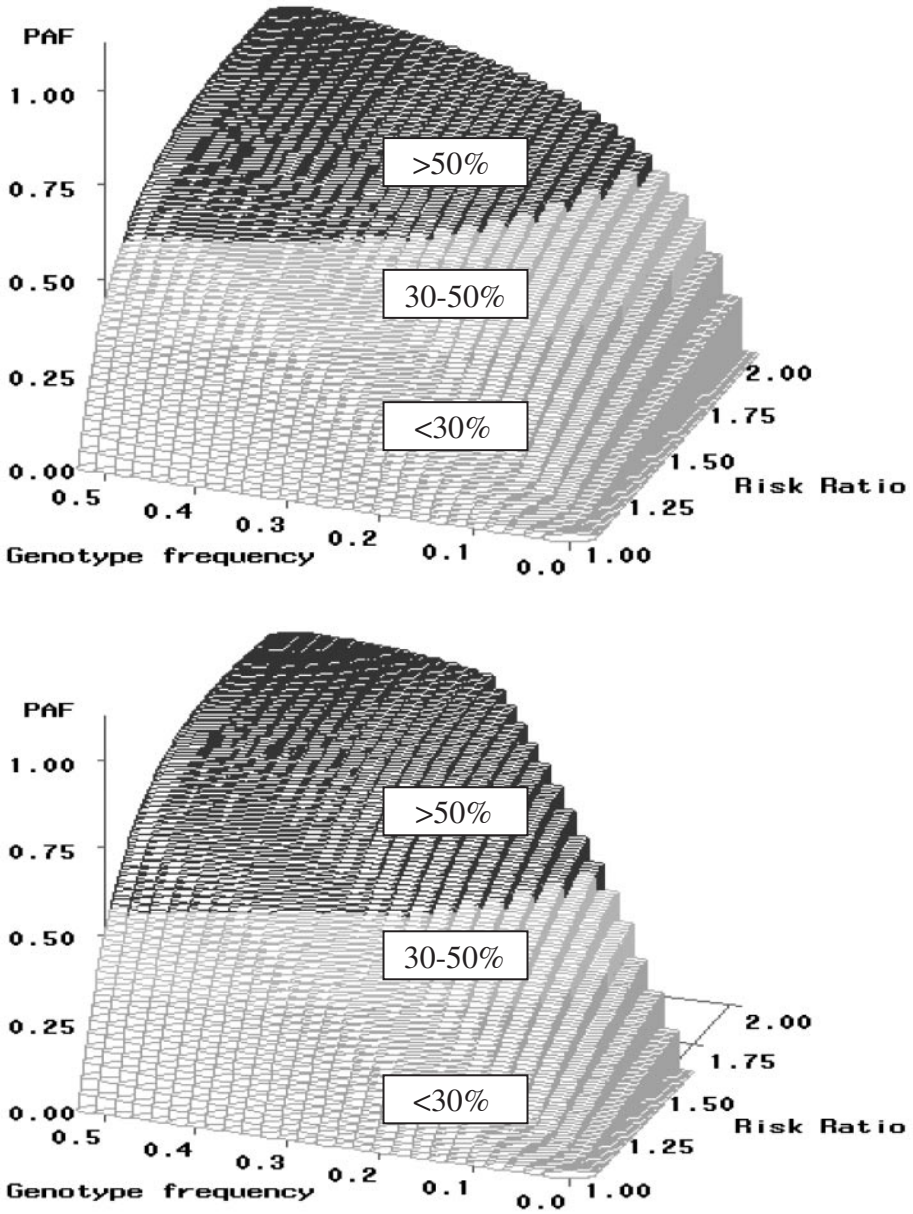

Figure 2 Expected population-attributable fraction (PAF) of 10 disease susceptibility genes with varying genotype prevalences and risk ratios assuming an additive (top) or multiplicative (bottom) joint effects model and lifetime risk of disease = 5.0%. PAF is not plotted if $I * [N * R_g - (j - 1)] \geqslant 1$ (additive model) or $I * R_g^N \geqslant 1$ (multiplicative model), i.e. lifetime disease risk is >100%

additive joint effects, and a combined PAF of 63.9% (95% CI 31.5–82.5%) assuming multiplicative joint effects (Table 3). If we exclude the *NAT2* gene (since the phenotype, not the genotype, is associated with increased risk for colorectal cancer), the estimated PAFs are 43.2% (95% CI 25.1–57.1%) and 49.1% (95% CI 27.0–67.0%) for additive and multiplicative joint effects, respectively.

Discussion

Identification of genes associated with common complex diseases is accorded a high public health priority because of the large contribution these conditions make to the total burden of disease in the population. Measurement of PAF provides a public health dimension to the appraisal of risks and creates an important link between disease causality and public health action.[17] In this paper, we have explored hypothetical scenarios in which causality is assumed to follow straightforward polygenic models with simple forms of gene–gene interaction. In the real world, causality has to be established on the basis of appraisal of the entire body of evidence,[18] and such simple models of gene action are very unlikely to be encountered.

There is a substantial difference in interpretation of the PAF related to the genetic contribution to a common complex disease and interpretation of a conventional attributable fraction calculated for a single exposure (risk factor) in an epidemiological study. The PAF is generally considered to be the fraction of disease cases that could be prevented by eliminating

Table 2 Number of genes needed to achieve a given population attributable fraction (PAF) for a complex disease with lifetime risk of 0.1% in the population and varying genotype prevalences and risk ratios

	PAF			
Genotype prevalence	5%	10%	30%	50%
Additive model				
Risk ratio = 10				
1/10,000	59	121[a]	477[a]	>1000[a]
2/10,000	30	62	239[a]	556[a]
1/1,000	6	13	48	112
2/1,000	3	7	24	56
Risk ratio = 20				
1/10,000	28	59[a]	226[a]	527[a]
2/10,000	14	30	113[a]	264[a]
1/1,000	3	6	23	53
2/1,000	2	3	12	27
Multiplicative model				
Risk ratio = 10				
1/10,000	58[b]	118[b]	397[b]	771[b]
2/10,000	29[b]	59[b]	199[b]	386[b]
1/1,000	6	12[b]	40[b]	78[b]
2/1,000	3	6	20[b]	39[b]
Risk ratio = 20				
1/10,000	28[b]	56[b]	188[b]	366[b]
2/10,000	14[b]	28[b]	95[b]	183[b]
1/1,000	3	6[b]	20[b]	37[b]
2/1,000	2	3	10[b]	19[b]

[a] Indicates that $I * [j * R_g - (j - 1)] \geqslant 1$ for some values of j (i.e. lifetime disease risk is >100%); for these values of j we set risk = 1.

[b] Denotes that $I * R_g^N \geqslant 1$ for some values of j (i.e. lifetime disease risk is >100%); for these values of j we set risk = 1.

a causal exposure. It has been argued that PAF is a meaningless concept in genetics because genetic risk factors cannot be removed.[19] While it is certainly true that one cannot eliminate the genetic risk factors an individual has inherited from his or her parents, we believe that the concept of PAF for genetic susceptibility provides a useful metric of the potential impact of interventions that may be developed on the basis of information about that genotype. Perhaps the most intuitive potential applications are genotype-specific screening and targeted interventions. Although targeted interventions are not available for the majority of mutations that have been identified so far,[19] examples such as neonatal screening indicate the potential importance of such an approach for public health.[20] More generally, knowledge that a group of genetic variants accounts for a substantial PAF could enhance understanding of disease pathogenesis and thereby aid in identifying interventions relevant to the general population. In particular, Mendelian randomization has been proposed as a means of obtaining estimates of the effects of environmental exposures in association studies of functional genetic variants.[21–23]

A further issue regarding the application of the concept of PAF to genetic variants is that the genotypes may have simultaneous effects on many different diseases.[19,24] PAF is disease-specific, but genetic predispositions to common diseases such as cancer or autoimmunity, for example, often are not. The attributable community risk (ACR), a measure recently reintroduced by Wacholder,[25] is particularly useful for comparing the potential population impact of complex genotypes on several different diseases. The ACR is the proportion of the population that develops disease that is attributable to an exposure or, in the current context, a disease susceptibility genotype. The ACR is related to PAF:

$$\text{ACR} = \text{PAF} * D,$$

where D is the lifetime risk of disease. For example, for a disease for which the lifetime risk is 5% and estimated PAF 50%, the corresponding ACR is 2.5%. For a disease with a lifetime risk of 1% and PAF 50%, the corresponding ACR is 0.05%.

Table 3 Prevalence, risk (95% CI), and population attributable fraction (PAF) (95% CI) of five genetic variants for colorectal cancer susceptibility

Genetic variants	Risk group	Genotype prevalence (%)	Odds ratio (95% CI)	PAF % (95% CI)[a]
HRAS1	Rare allele vs others	4.0	2.67 (1.47–4.85)	6.3 (1.9–13.3)
GSTT1	Null vs others	37.6	1.37 (1.17–1.60)	12.2 (6.0–18.4)
TNF-α	α2 allele vs others	39.2	2.02 (1.51–2.71)	28.6 (10.0–40.1)
NAT2 [imputed from phenotype]	Fast acetylation vs others	[60.3]	1.68 (1.11–2.46)	29.1 (6.2–46.8)
MTHFR	Wild-type vs variant (C677T)	42.3	1.35 (1.12–1.64)	12.9 (4.8–21.3)
Five genes combined				
Additive model	–	–	–	53.9 (28.7–68.9)
Multiplicative model	–	–	–	63.9 (31.5–82.5)

HRAS1, c-Ha-ras1 proto-oncogene; *GSTT1*, glutathione *S*-transferase theta 1; *TNF*-α, tumor necrosis factor alpha-chain; *NAT2*, *N*-acetyl transferase-2 gene; *MTHFR*, 5,10-methylenetetrahydrofolate reductase gene.

[a] The lower and upper 95% CIs of the PAF were calculated by taking lower and upper 95% odds ratio estimates, respectively.

There is controversy over whether the genetic basis for susceptibility to common diseases results from a relatively small number of common alleles that each produce only a modest predisposition or a larger number of rare variants each of which has a much greater predisposing effect[4,5,26–28] Our analysis suggests that genetic information would have a much greater public health impact if the first scenario, the common-disease–common-variant hypothesis were correct. From a public health perspective, the important issue is to identify common genetic factors that lead to strategies for the prevention or improved treatment of common complex disease in these predisposed individuals. Depending on the nature of the gene–environment interactions involved, pharmacological, dietary, or lifestyle interventions among genetically predisposed individuals may have a disproportionately large effect on the associated morbidity and mortality in the population as a whole. This targeted approach to prevention may be especially useful in the case of diseases with major environmental components, such as type 2 diabetes and cardiovascular diseases.[7,29]

Most common human diseases are due to complex interactions among multiple genetic variants and environmental risk factors.[2,3] The present study provides a general framework for estimating the number of genes needed to account for an appreciable proportion of a disease in the population, but different methods are required to estimate the separate effects of genes and environmental exposures or of gene–environment interactions.[30] Our model can be extended to include gene–environment interactions by adding additional terms in each of the 2^N strata. Persons exposed to environmental risk factors may be considered to be a higher-risk subgroup within each stratum. Considering both genetic and environmental risk factors together may permit assessment of the relative benefits and feasibility of eliminating environmental risk factors within genetically predisposed groups as opposed to eliminating these exposures in the population as a whole. Although eliminating environmental risk factors and promoting a healthy lifestyle are usually recommended for the population as a whole regardless of genetic predisposition, there are situations in which targeted intervention may be more cost-effective.[31]

For most common diseases, we do not understand the nature of the joint effects among predisposing genes. We considered only the simplest additive and multiplicative models and assumed that the multiple disease susceptibility genes are unlinked. In reality, the gene–gene and gene–environment interactions for common diseases are likely to be much more complex, and a model combining both additive and multiplicative interactions might more accurately reflect biological reality.[32]

Furthermore, the models we considered have some limitations. For a fixed lifetime risk of disease D in the population as a whole, the maximum number of genes attainable (keeping risk of the disease $\leqslant$100%) is

$$N \leqslant \frac{1}{I[R_g - (j-1)]}$$

for the additive model and

$$N \leqslant \frac{1}{I * R_g^N}$$

for the multiplicative model. As the lifetime risk of disease becomes more common, the background risk I increases, and the number of genes that satisfy the conditions of the model becomes smaller. Alternatively, for the multiplicative effect model, one may use the odds ratio instead of the risk ratio in a logistic risk model to estimate the population risk of a common disease D involving N genes. The logistic risk model is free from the constraint of the background risk I multiplied by stratum-specific risk for disease >100%.

We employed an epidemiological approach to estimate the number of genes needed to account for an appreciable PAF for common diseases. Another commonly used approach is the mutifactorial-threshold model, which postulates a continuously distributed latent trait, liability, that causes the disease.[33] Two additive, normally-distributed components underlie liability—a genetic component produced by numerous small, additive (polygenic) effects, and a random environmental component. An individual is affected by the disease when her or his liability exceeds a particular threshold. Risch has discussed the multifactorial-threshold model and its relationship to epidemiological attributable risk in common forms of cancer.[34] He introduced the concept of PAF related to genetic factors and pointed out its dependence on the combined effect of all susceptibility alleles at disease-predisposing loci.

Our findings have a potential impact on narrowing the search for disease susceptibility genes for complex human diseases. For common genetic variants ($G \geqslant 10\%$), only a limited number of genes are needed to produce an appreciable PAF, even if the disease risk associated with each gene is moderate or weak (e.g. $R_g \leqslant 1.5$). From a public health point of view, identification of these genes should receive high priority. On the other hand, the PAF associated with rare genetic variants ($G \leqslant 1/1000$) tends to be small, and a large number of genes ($N > 150$) are needed to produce an appreciable PAF, even if the risk associated with each gene is strong ($R_g \geqslant 10$). These patterns suggest that greater public health importance is likely to be associated with common disease-predisposing genetic variants than with rare variants, even if the rare variants each produce a higher relative risk.

KEY MESSAGES

- Most common human diseases result from complex interactions among multiple genetic variants and environmental risk factors.
- Variants of as few as 20 susceptibility genes, each of which has weak to moderate individual effects, may account for ⩾50% of the burden of most common complex diseases if each variant is common in the population.
- Identifying these common variants is potentially of great public health importance because their recognition may provide opportunities for screening and targeted reduction of modifiable environmental risk factors.

References

[1] Collins FS, Green ED, Guttmacher AE, Guyer MS. A vision for the future of genomics research. *Nature* 2003;**422:**835–47.

[2] Glazier AM, Nadeau JH, Aitman TJ. Finding genes that underlie complex traits. *Science* 2002;**298:**2345–49.

[3] Guttmacher AE, Collins FS. Genomic medicine—a primer. *N Engl J Med* 2002;**347:**1512–20.

[4] Wang WY, Barratt BJ, Clayton DG, Todd JA. Genome-wide association studies: theoretical and practical concerns. *Nat Rev Genet* 2005;**6:**109–18.

[5] Hirschhorn JN, Daly MJ. Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet* 2005;**6:**95–108.

[6] Reich DE, Lander ES. On the allelic spectrum of human disease. *Trends Genet* 2001;**17:**502–10.

[7] Merikangas KR, Risch N. Genomic priorities and public health. *Science* 2003;**302:**599–601.

[8] Khoury MJ, Yang Q, Gwinn M, Little J, Dana Flanders W. An epidemiologic assessment of genomic profiling for measuring susceptibility to common diseases and targeting interventions. *Genet Med* 2004;**6:**38–47.

[9] Rothman KJ, Greenland S. *Modern Epidemiology*. Philadelphia: Lippincott-Raven, 1998.

[10] Kleinbaum DG, Morgenstern H, Kupper LL. Epidemiologic research: principles and quantitative methods. Belmont, CA: Lifetime Learning Publications, 1982.

[11] Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K. A comprehensive review of genetic association studies. *Genet Med* 2002;**4:**45–61.

[12] Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat Genet* 2003;**33:**177–82.

[13] Ioannidis JP, Ntzani EE, Trikalinos TA, Contopoulos-Ioannidis DG. Replication validity of genetic association studies. *Nat Genet* 2001; **29:**306–9.

[14] Janssens AC, Pardo MC, Steyerberg EW, van Duijn CM. Revisiting the clinical validity of multiplex genetic testing in complex diseases. *Am J Hum Genet* 2004;**74:**585–88.

[15] Ries LAG, Eisner MP, Kosary CL *et al. SEER Cancer Statistics Review, 1975–2001*. National Cancer Institute. Bethesda, MD: National Cancer Institute, 2004. Available at: http://seer.cancer.gov/csr/1975_2001/

[16] de Jong MM, Nolte IM, te Meerman GJ *et al*. Low-penetrance genes and their involvement in colorectal cancer susceptibility. *Cancer Epidemiol Biomarkers Prev* 2002;**11:**1332–52.

[17] Northridge ME. Public health methods—attributable risk as a link between causality and public health action. *Am J Public Health* 1995;**85:**1202–4.

[18] Little J, Bradley L, Bray MS *et al*. Reporting, appraising, and integrating data on genotype prevalence and gene-disease associations. *Am J Epidemiol* 2002;**156:**300–10.

[19] Terwilliger JD, Weiss KM. Confounding, ascertainment bias, and the blind quest for a genetic 'fountain of youth'. *Ann Med* 2003; **35:**532–44.

[20] Hannon HW, Henderson OL, Bell CJ. Newborn Screening Quality Assurance. In: Khoury MJ, Burke W, Thomson EJ (eds). Genetics and public health in the 21st century: using genetic information to improve health and prevent disease. Oxford; New York: Oxford University Press, 2000, p. 639.

[21] Davey Smith G, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int J Epidemiol* 2003;**32:**1–22.

[22] Little J, Khoury MJ. Mendelian randomisation: a new spin or real progress? *Lancet* 2003;**362:**930–31.

[23] Davey Smith G, Ebrahim S. Mendelian randomization: prospects, potentials, and limitations. *Int J Epidemiol* 2004;**33:**30–42.

[24] Vineis P, Schulte P, McMichael AJ. Misconceptions about the use of genetic tests in populations. *Lancet* 2001;**357:**709–12.

[25] Wacholder S. The impact of a prevention effort on the community. *Epidemiology* 2005;**16:**1–3.

[26] Pritchard JK. Are rare variants responsible for susceptibility to complex diseases? *Am J Hum Genet* 2001;**69:**124–37.

[27] Wright AF, Hastie ND. Complex genetic diseases: controversy over the Croesus code. *Genome Biol* 2001;**2:**COMMENT2007.

[28] Becker KG. The common variants/multiple disease hypothesis of common complex genetic disorders. *Med Hypotheses* 2004; **62:**309–17.

[29] Willett WC. Balancing life-style and genomics research for disease prevention. *Science* 2002;**296:**695–98.

[30] Clayton D, McKeigue PM. Epidemiological methods for studying genes and environmental factors in complex diseases. *Lancet* 2001;**358:**1356–60.

[31] Nallamothu BK, Fendrick AM, Rubenfire M, Saint S, Bandekar RR, Omenn GS. Potential clinical and economic effects of homocyst(e)ine lowering. *Arch Intern Med* 2000;**160:**3406–12.

[32] Khoury MJ, Beaty TH, Cohen BH. *Fundamentals of genetic epidemiology. Monographs in Epidemiology and Biostatistics v. 22*. New York: Oxford University Press, 1993. Vol. vi, p. 383.

[33] Hartl DL, Clark AG. Principles of population genetics. Sunderland, MA: Sinauer Associates, 1997. Vol. xiii, p. 542.

[34] Risch N. The genetic epidemiology of cancer: interpreting family and twin studies and their implications for molecular genetic approaches. *Cancer Epidemiol Biomarkers Prev* 2001;**10:**733–41.

Appendix

Table A1 Expected population prevalence, risk and associated population attributable fraction (PAF) of selected genetic variants susceptibility to colorectal cancer

Genetic variants[a]					**Expected population prevalence%**[b]	**Risk**		**PAF %**	
HRAS1	*GSTT1*	*TNF-α*	*NAT2*	*MTHFR*		Additive	Multiplicative	Additive	Multiplicative
0	0	0	0	0	8.1	1.0	1.0	–	–
1	0	0	0	0	0.3	2.7	2.7	0.3	0.2
0	1	0	0	0	4.9	1.4	1.4	0.8	0.7
0	0	1	0	0	5.2	2.0	2.0	2.4	1.9
0	0	0	1	0	12.3	1.7	1.7	3.8	3.0
0	0	0	0	1	6.4	1.4	1.4	1.0	0.8
1	1	0	0	0	0.2	3.0	3.7	0.2	0.2
1	0	1	0	0	0.2	3.7	5.4	0.3	0.3
1	0	0	1	0	0.5	3.4	4.5	0.6	0.6
1	0	0	0	1	0.3	3.0	3.6	0.3	0.3
0	1	1	0	0	3.1	2.4	2.8	2.0	2.0
0	1	0	1	0	7.4	2.1	2.3	3.6	3.5
0	1	0	0	1	3.9	1.7	1.9	1.3	0.2
0	0	1	1	0	7.9	2.7	3.4	6.2	6.8
0	0	1	0	1	4.1	2.4	2.7	2.6	2.6
0	0	0	1	1	9.7	2.0	2.3	4.6	4.4
1	1	1	0	0	0.1	4.1	7.4	0.2	0.3
1	1	0	1	0	0.3	3.7	6.1	0.4	0.6
1	1	0	0	1	0.2	3.4	4.9	0.2	0.2
1	0	1	1	0	0.3	4.4	9.1	0.5	1.0
1	0	1	0	1	0.2	4.0	7.3	0.2	0.4
1	0	0	1	1	0.4	3.7	6.1	0.5	0.7
0	1	1	1	0	4.8	3.1	4.6	4.5	6.3
0	1	1	0	1	2.5	2.7	3.7	2.0	2.5
0	1	0	1	1	5.8	2.4	3.1	3.8	4.5
0	0	1	1	1	6.3	3.1	4.6	5.9	8.1
1	1	1	1	0	0.2	4.7	12.4	0.3	0.8
1	1	1	0	1	0.1	4.4	10.0	0.2	0.3
1	1	0	1	1	0.2	4.1	8.3	0.3	0.6
1	0	1	1	1	0.3	4.7	12.2	0.5	1.1
0	1	1	1	1	3.8	3.4	6.3	4.2	7.2
1	1	1	1	1	0.2	5.1	16.8	0.3	0.9

[a] *HRAS1*, c-Ha-ras1 proto-oncogene: rare allele vs others; *GSTT1*, glutathione *S*-transferase theta 1: null vs others; *TNF-α*, tumor necrosis factor alpha-chain gene: a2 allele vs others; *NAT2, N*-acetyl transferase-2 gene (imputed phenotype): fast-acetylation vs others; *MTHFR*, 5,10-methylenetetrahydrofolate reductase gene: wild-type vs C677T variant. 1 indicates the present of the genetic variants and 0 indicates the absence.

[b] We assume the independent assortment of multiple genetic variants in the population.