The Rise and Fall of the Common Disease–Common Variant (CD–CV) Hypothesis: How the Sickle Cell Disease Paradigm Led Us All Astray (Or Did It?)

Oliver Mayo

CSIRO Livestock Industries, Adelaide, South Australia, Australia

The common disease–common variant (CD–CV) hypothesis requires an explanation for the origin of the variation observed, since substantial neutral, but not deleterious, variation, that is, several alleles each at moderate to high frequency, can be maintained at any gene/locus by mutation. It is argued here that the guiding principle, not always stated, has been balancing selection, influenced by the well-established cases of deleterious alleles maintained through heterozygous advantage in the face of strong malarial selection against normal alleles. It is further argued that, although balanced polymorphisms have indeed arisen and reduced population loss through infectious disease, the history of balance in other contexts should have prevented acceptance of any hypothesis that generalized such a specific mechanism. Finally, it is suggested that in the present state of knowledge no single hypothesis for the genetical contribution to common disorders is justifiable.

The common disease–common variant (CD–CV) hypothesis states that the genetical component in the causation of common diseases is likely to arise from relatively common alleles of a relatively small number of genes/loci (see Reich & Lander, 2001, for an important discussion and partial quantification of the hypothesis). This hypothesis is to be contrasted with the common disease–rare variant (CD–RV) hypothesis, whereby common diseases are genetically influenced by rare alleles of a large number of different genes. Pritchard (2001) has argued persuasively that the CD–RV hypothesis is at least partially supported by plausible models and much of the data.

Established theory gives us confidence that we understand the maintenance by mutation-selection balance of rare deleterious alleles and the maintenance by mutation of common neutral alleles. Common disorders present more complex problems. First, they are generally non-Mendelian in their mode of inheritance. Second, they frequently have clear environmental contributions. Finally, they are common and yet deleterious. They therefore seem to require some different explanation. Several are possible. First, many of them, unlike malaria, have little effect on reproduction and hence are not subjected to strong stabilising or purifying selection, so it could be argued that the relevant deleterious alleles are neutral or even advantageous before reproduction. Second, there could be other balancing selection with a similar time spectrum to the common disease. Third, they could be frequent by chance, after a population bottleneck and recent very rapid population expansion.

Balancing selection, as mentioned, and as will be discussed extensively below, has been established as a reality. Selection acting in opposite directions at different stages of the lifetime is mainly a theoretical possibility. Bottlenecks are real, but yield a different explanation in every case and have been hard to quantify until the advent of extensive genomic information.

The case of maintenance of sickle-cell haemoglobin, presented in detail below, is so well established, and so convincing, that it appeared to provide a guiding model for the action of selection, even if less intense, for much of human disease. It will be argued that this case has guided many in a wrong direction, even if it still has lessons to teach.

Historical Origins of the Hypothesis

Many human disorders show a strong genetical component in their causation, but are not inherited in a Mendelian fashion, that is, they do not result from segregation at one or a few genes in predictable ways. Some of these conditions are quite frequent — prevalence for bipolar disorder, type 1 diabetes, schizophrenia, ischaemic heart disease, peptic ulcer, depression and endometriosis have been estimated at 0.1%, 0.5%, 1%, 3%, 4%, 5% to 17%, and 8% to 10% respectively (see Keller & Miller, 2006). Since

Received 11 September, 2007; accepted 19 September, 2007.

Address for correspondence: Oliver Mayo, CSIRO Livestock Industries, Adelaide, South Australia. E-mail: Oliver.Mayo@csiro.au

https://doi.org/10.1375/twin.10.6.793 Published online by Cambridge University Press

many of these conditions are extremely debilitating, or worse, and can lead to reduced fertility, as well as to morbidity and death, some scientists have found their population prevalence oddly high, and have sought an explanation in evolutionary genetics.

The explanation has been that these traits arise from homozygosity, or occasionally heterozygosity, for relatively common alleles, to be found at moderate frequencies in many populations. If there is a requirement for homozygosity, the homozygotes must come from segregation at loci where heterozygotes are fitter than homozygotes, or where in recent evolutionary history this has been the case. For example, where incursion into a population of a novel infectious disease has led to the increase in frequency of alleles that protect against that disease, but which are deleterious in the homozygote. An example of such an allele is *APOE*4*, an apolipoprotein E allele associated with increased risk of coronary artery disease and Alzheimer's disease, for which Corbo and Scacchi (1999) have hypothesised an advantage in the past, of the kind of 'thriftiness' first suggested by Neel (1962). This hypothesis was an ancestor of the CD–CV hypothesis, and it has guided the thinking of many investigators.

Examples of such common variants include a variant of *TCF7L2* on chromosome 10q which, in heterozygotes, increases relative risk to type 2 diabetes to 1.45 against non-carrier homozygotes, and in homozygotes to 2.41 (Grant et al., 2006). Taking into account the frequencies of the two alleles in the Scandinavian populations considered, the authors held that 21% of the risk of type 2 diabetes was attributable to the 'risk allele'. The nature of the selection against the 'normal' allele, if it exists, is unknown.

Another hypothesis, regarded as antithetical to the CD–CV, is the common disease-rare variant (CD–RV) hypothesis: homozygosity for rare variants of many different genes contributes to the disease. In this case, the homozygotes come from segregation at loci where variation is maintained by a balance between mutation and selection against rare deleterious recessive alleles. This second hypothesis is analogous to the established fact that there are many rare monogenic disorders, and that these disorders appear for the most part to be in approximate mutation-selection balance. It is generally considered that this hypothesis can explain the observed prevalences and patterns of inheritance without recourse to an assessment of evolutionary history that can only be supported by indirect evidence. It is therefore of interest to analyse the origin of the CD–CV hypothesis and determine whether the current consensus represents an advance in thinking.

Fisher (1922) wrote in his first exposition of the dynamics of gene frequency in populations:

> Cases where the heterozygote is favoured by selection in preference to both homozygous forms are of additional interest, as, when the selection is intense, these cases may form the basis upon which a system of balanced lethal factors is built up. Muller (1918) has shown that such systems will tend to develop when selection strongly favours the heterozygote, and has explained how, in the light of such systems, the majority of the phenomena, including the 'mutations' of *Oenothera*, find a genetic explanation.

This was the first description of balance in a population-genetical context; Muller had the view that an 'ideal' type was possible in a species, and considered the balance which he had discovered to be rare or pathological. As Fisher, Haldane, Wright and others developed the fundamental models for population genetics, they all considered that mutation (broadly defined to include major changes like duplication and polyploidization, as well as changes of unknown nature to a gene) was the source of new variability, and they showed that the loss of variability depended on population size.

Thus, the level of genetical variability in a population could depend on the unknown size of the genome, the rates of production of new mutations, and the size of the species or population. Fisher showed how balancing selection could maintain variability. Indeed, in a large population, such variability was almost impossible to eliminate. Other variability would be maintained by balance among several factors, selective and otherwise: population size (where small size leads to loss of variability), mutation, and selection for rare advantageous, and against common disadvantageous, new mutations.

The concept of balancing selection proved very attractive to scientists, since it appeared to provide an explanation for persistent deleterious variation, particularly observable in human populations, which directional selection would otherwise be expected to eliminate. The case of variation in human hemoglobin was very important and will be discussed in detail below; here, I simply note that balanced polymorphism was discovered and shown to be important in one case of a major human disease thought to be fairly recently harmful in human populations, and this appeared to be very strong support for the idea that a common variant, deleterious in the homozygote, could be maintained at high frequency by an external selective agent. It was argued that other common diseases whose cause was not clear-cut, like malaria's, could be maintained by selective balance.

From this line of reasoning arose the hypothesis that much human deleterious variation is the manifestation of segregation at a modest number of polymorphic loci, whereby homozygotes were at a disadvantage to heterozygotes. A completely distinct line of reasoning also made the hypothesis attractive, namely, that based on the knowledge that inbreeding led to decreased fitness which could be restored by outbreeding: the implication from the hypothesis of overdominance was that, if many loci were maintained in this way, then inbreeding would lead to fixation of either homozygote and a loss of fitness.

https://doi.org/10.1375/twin.10.6.793 Published online by Cambridge University Press

However, if overdominance were rare, but deleterious recessives maintained by a balance of mutation and selection were not, then loss of fitness on inbreeding would result from chance fixation of some slightly deleterious alleles at many loci. At the present time, evidence from experimental organisms favours the latter hypothesis; see Bürger (2000) and Leach and Mayo (2005) for discussion. It is noteworthy that the dominance hypothesis held sway as an explanation for hybrid vigour in maize from the beginnings of that industry until about 1945, after which overdominance was invoked for some thirty years because of experimental results which dominance could not then explain. In fact, strong linkage disequilibrium (see below) resulting from intense directional selection had mimicked overdominance. Crow (1998) discusses the critical experiments in maize that settled the issue.

Increased fitness can be positively associated with genetic variability per se (e.g., Allendorf & Leary, 1986; Lesbarrères et al. 2007), although as Fisher's (1930) Fundamental Theorem of Natural Selection shows, it is the rate of increase in fitness that will be associated with a non-negative attribute of population variation in fitness. An inference often drawn was that variability would be beneficial in itself, contrary to the Mullerian view mentioned earlier. The problem then became how to explain a persistent high level of deleterious variation, such as that relating to common diseases in human populations.

In this review I consider, first, balanced polymorphism, then the level of variation in human populations, then the evidence for selection associated with polymorphisms, and finally where the CVCD hypothesis stands.

The Hypothesis of Balanced Polymorphism

Fisher had shown that, in a large population, a diallelic locus with the following selective values would have stable gene frequencies from generation to generation:

Genotype	A_1A_1	A_1A_2	A_2A_2
Selective value	$1-s$	1	$1-t$
Genotype frequency	p^2	$2pq$	q^2

Fisher showed that there is a stable equilibrium with $p = t/(s + t)$ and $q = s/(s + t)$ for $0 < s, t \leq 1$. This advantage of the heterozygote over both homozygotes was sometimes termed overdominance.

Fisher and others identified strong selective forces acting on polymorphisms. For example, Fisher (1939) showed $s = 0.07$, $t = 0.05$ for a number of different closely linked polymorphisms in *Paratettix texanus*. He found further, however, that double heterozygotes were at a great disadvantage in nature to single heterozygotes.

Despite this latter important finding, it was thought, as more polymorphisms were discovered, and some were shown to be influenced by strong selection, that balancing selection through heterozygous advantage could be an important explanation for much observed natural variation in populations. Ford (1940) enunciated the rule that balanced polymorphism was '… the occurrence together in the same locality of two or more discontinuous forms of a species in such proportions that the rarest of them cannot be maintained merely by recurrent mutation' (Ford, 1975).

As was frequently noted (e.g., Mayo, 1972), Ford's definition assumed what it had to prove, that recurrent mutation alone could not maintain large numbers of different allelic forms of a given gene, although Fisher (1922) had shown that 'the rate of mutation required [to balance the loss through selection of deleterious alleles] diminishes as the number of individuals increases', and the ability of mutation alone to maintain neutral and nearly neutral alleles in large populations was implicit in Fisher's (1930) analysis. Indeed, the later controversy between Fisher and Wright over the number of gametophytic self-incompatibility (SI) alleles that could be maintained by mutation in a small population clarified the fact that in a large population, mutation could readily maintain many different alleles in the absence of selection (see Leach and Mayo, 2005 for references). SI is in fact a type of polymorphism, which is maintained as a multiallelic balanced polymorphism through the effective lethality of homozygotes, like Muller's original case.

Extensions to the Basic Model

In the discussion above, it has already been recognised that strong, constant selection acting on the three phenotypes of a single diallelic locus in a large, stable population in a constant environment is not necessarily the norm in nature. Indeed, any aspect of this model could be changed, and has been investigated: variable selection, multiple alleles, multiple loci, variable environment, and variable population size.

Through investigation of natural phenomena such as mimicry, the idea of frequency-dependent selection was developed: the intensity of selection on a phenotype would depend in some way on the frequency of that phenotype. A simple example would be as follows:

Genotype	A_1A_1	A_1A_2	A_2A_2
Selective value	$1-sp$	1	$1-tq$
Genotype frequency	p^2	$2pq$	q^2

In this case, there is a stable equilibrium with $p = t^{1/2}/(s^{1/2}+ t^{1/2})$ and $q = s^{1/2}/(s^{1/2}+ t^{1/2})$ for $0 < s, t \leq 1$. SI manifests frequency-dependent selection, though not of this type; each of the many alleles is advantageous solely according to its rarity. Single diallelic loci manifesting frequency dependence have indeed been discovered; a recent one is a locus influencing foraging behaviour in *Drosophila melanogaster* (Fitzpatrick et al., 2007). In this case, selection is against one of two phenotypes, one of which contains one homozygote and the heterozygote, so there is no heterozygous advantage. In the same way, selection alternating in direction in space or in time could maintain variation, but would not produce balanced polymorphism of the classical type.

https://doi.org/10.1375/twin.10.6.793 Published online by Cambridge University Press

Multiallelic systems were identified very early in the development of genetics following the rediscovery of Mendel's work in 1900. Indeed, one of the first was the human ABO blood group system (Landsteiner, 1900), though the genetics were not completely elucidated for some time. A multiallelic polymorphism can be maintained by balancing selection, but it is unlikely that any particular multiallelic polymorphism examined at a particular time will be in equilibrium (Bürger, 2000; Mandel, 1959).

Multilocus polymorphism began to be investigated soon after single locus polymorphism. The simplest case, two diallelic loci, immediately required so many more parameters that solutions could always be found, but were likely not to be unique:

Genotype	A_1A_1	A_1A_2	A_2A_2	
	Frequency	p_A^2	$2p_Aq_A$	q_A^2
B_1B_1	p_B^2	α	β	γ
B_1B_2	$2p_Bq_B$	δ	ε	ζ
B_2B_2	q_B^2	η	θ	κ

Few, if any, good examples of multilocus polymorphism were found in the era before biochemical genetics, apart from 'supergenes', chromosomal regions concerned with many aspects of a given complex trait, such as heteromorphic self-incompatibility in flowering plants (see Leach and Mayo, 2005). Indeed, influences on fitness of interactions among alleles of two loci have rarely been detected (for a recent example, see Martin et al., 2007). However, the model above provided the basis for consideration of another population phenomenon, so-called linkage disequilibrium (gametic or allelic association would have been better names). If the frequencies of the gametes A_1B_1, A_1B_2, A_2B_1 and A_2B_2 are g_1, g_2, g_3 and g_4 respectively, then $D = g_1.g_4 - g_2.g_3$ is termed linkage disequilibrium (LD), since at equilibrium $D = 0$, both terms being equal to $p_Ap_Bq_Aq_B$. The approach to equilibrium takes place at a rate dependent on the frequency of recombination between the two loci, but is also affected by population size. For example, Sved (1968) suggested that $D^2 = 1/(16(4Nr + 1))$ at equilibrium in a population of size N with recombination at a rate r between A and B. If there are 3.10^9 base pairs of DNA and 10^6 polymorphisms, then LD must be substantial in any small region. This means that association without causation is to be expected between any phenotype and any marker in a region close to an actual causal gene. As Sved (1968) put it, 'in populations of small size the major factor responsible for keeping any locus polymorphic is not the selective values of the genes at that locus, but the selective values at surrounding loci'. It should be noted that although present-day human populations are large, it is a short evolutionary time since they were very small. Furthermore, many human populations are likely to have undergone bottlenecking, which also generates substantial LD (see Laurie et al., 2007, for a clear analysis of such situations in wild mice.)

Even ignoring linkage, equilibrium would be a most unusual state. In 1931, Fisher wrote to Sewall Wright about equilibria involving many genes, in connexion with Wright's concept of an adaptive topography with multiple equilibria: 'In one dimension a curve gives a series of alternate maxima and minima, but in two dimensions two inequalities must be satisfied for a strict maximum, and I suppose that only about 1/4 of the stationary points will satisfy both. Roughly I should guess that with n factors only about 2^{-n} of the stationary points would be stable for all types of displacement, and any new mutation will have a half chance of destroying the stability. This suggests that true stability in the case of many interacting genes may be of rare occurrence, though its consequences when it does occur are especially interesting and important.' (Bennett, 1983, p. 280) Fisher's surmise is in general correct; see Bürger (2000) for a review. To reach Bürger's conclusions, many important arguments were developed: multiple independent cases of heterozygous advantage were unlikely, because of the extreme fitness disadvantages that would result; the actual contributions to fitness variance observed in experiments did not match those expected if overdominance were common (as in the case of hybrid maize discussed earlier). Furthermore, as the true extent of polymorphism became evident through genomics, to the extent that there are millions of human DNA polymorphisms, single locus heterozygous advantage collapsed under its own weight as a potential explanation of the maintenance of genetical variation.

Population size had been shown to be important in the maintenance of variability by Fisher, Wright, and others in foundational work in the 1920s. In the classical balanced polymorphism, however, Robertson (1962) discovered an unexpected effect. If s and t were widely divergent, and if the population size N were small, variability would be lost through fixation of one or other allele (generally that with lower selection against it in the homozygote), and the loss of variability could occur more rapidly than for a pair of neutral alleles.

Although the Robertson effect and some other exceptions made the balanced polymorphism a less than universal expectation, nevertheless it appeared to be a possible general explanation for the persistence of some variation, particularly where strong selection existed.

Evidence From Human Populations

Variation

Several types of variation were identified long before any workable theory of inheritance had been developed. First, there were severe heritable diseases. Although diseases like hemophilia were rare, they were detected, identified, and described in several populations hundreds or thousands of years ago. Second, there were conditions like achondroplasia, which permitted a relatively normal life, yet reduced reproductive fitness very substantially

https://doi.org/10.1375/twin.10.6.793 Published online by Cambridge University Press

(Mørch, 1941). Third, there were variations, which appeared to be unrelated to fitness, such as skin, hair, or eye colour.

In the Mendelian era, research identified increasing numbers of representatives of all types of variation. In addition, quantitative traits, which manifested both normal and rare deleterious variation, such as measured intelligence and height/size, were shown to be strongly influenced by genes.

Because humans are interested in themselves (*Homo sum; humani nil a me alienum puto*: I am a man; I count nothing human indifferent to me — Terence, ca 150BC), knowledge of human variation has increased rapidly, the more deleterious or important, the more rapidly, so that the knowledge acquired was by no means either systematic or random in its growth.

Evidence is slowly accumulating to the effect that *Homo sapiens* is a relatively youthful species, in terms of generations (10,000–50,000) as well as years (< 1M), and is in consequence not as variable as many other species. For example, chimpanzees and humans are thought to have shared common ancestors about 5M years ago, yet chimpanzees differ at about twice as many DNA sites as humans (e.g., Ebersberger et al., 2002). This aspect of human variation does not appear to be germane to the present topic, however, so will be discussed no further.

Many genes damaging to normal human mental activity have been discovered, but so far none of substantial effect in the opposite direction. It is instructive to consider some genome-wide searches for genes influencing measured intelligence.

A genome-wide association scan for a qualitative trait begins with the assay of a set of DNA polymorphisms spaced out across the entire genome in a set of affected individuals (and possibly their families) and in an unrelated control set of unaffected individuals. The sets can then be compared by a range of standard statistical techniques to determine which alleles of which polymorphic genes/sites differ in their frequencies in the two groups. At the same time, or subsequently, family studies show whether, or to what extent, the elevated-frequency alleles segregate with the trait. Consequently chromosomal sites can be mapped which contribute to the risk of the trait. These are termed QTL (quantitative trait loci), though of course the trait, if it is a disease, is discrete insofar as the diagnosis of the disease is concerned.

In the case of a quantitative trait like a measure of intelligence, it is the magnitude of the measure in each individual which is associated with the relevant allele of each segregating gene locus. A number of scans have been published, with the results shown in Table 1. Even the major study of Butcher et al. (2007) gives only limited confidence that anything important or indeed useful has yet been discovered, beyond perhaps confirmation that measured intelligence in the normal range is a classical quantitative trait, influenced by very many genes of individually tiny effect. We should also note that any trait which is not defined relatively simply in physical terms, and which has high heritability, is likely to be similarly influenced. See, for example, the discussion of alcoholism by Rodd et al. (2007). Normal variation in many genes contributes, on their hypothesis, to 'physical and physiological robustness' to withstand the debilitating effects of alcohol, and separately to 'reactivity to its rewarding effects', thereby producing a phenotype likely to develop addiction.

Variation Associated With Malaria

A mutant of the hemoglobin β chain, 'sickle hemoglobin', was discovered to be widespread in African populations exposed to endemic malaria. Research showed that heterozygotes *HBBA/HBBS* were protected against malaria as compared with normal homozygotes *HBBA/HBBA*, whereas homozygotes for the 'sickle' allele *HBBS/HBBS* suffered from often severe anaemia. Allison (1955) was able to estimate fitnesses as follows:

Genotype	*HBBA/HBBA*	*HBBA/HBBS*	*HBBS/HBBS*
Selective agent	Malaria	—	Hemolytic anaemia
Fitness	0.80	1	0.25

Table 1

Chromosomal Regions/DNA Polymorphisms/Genes Associated With Measured Intelligence

Study size	No. DNA markers	No. regions discovered	Regions	Other associations	Authors
> 600 twin families	761 (micro-satellites)	2	2q24.1–31.1 (*D2S142– D2S2364*) *6p25.3–22.3*	Autism	Posthuma et al. 2005
> 320 twin families	795	1	*D2S2313–D2S335*		Luciano et al. 2006
200 + individuals at distributional extreme	1842 (simple sequence repeat)	0			Plomin et al. 2001
7000 individuals	500,000 SNPs	5	3q22.1, 6p24.1, 7q32.1, 11q12.3, 16p13.3		Butcher et al. 2007

https://doi.org/10.1375/twin.10.6.793 Published online by Cambridge University Press

Haldane (1949) had earlier suggested that thalassaemia, another blood disease shown to be associated with a series of hemoglobin variants, might be associated with malaria in some protective fashion, and this has indeed been shown to be the case. 'The picture that is emerging, therefore, is of multiple red cell polymorphisms which protect against different forms of malaria by different mechanisms' (Weatherall 2007). They all protect the individual in youth, when malarial death and morbidity are particularly severe.

The genes of which alleles have been confirmed to provide protection against malaria include, besides Hbα and Hbβ, the Duffy blood group, glucose-6-phosphate dehydrogenase (G6PDH), tissue necrosis factor (TNF), and HLA (Kwiatowski, 2005). Investigation of these genes' relationship with health in general, and malaria in particular, has yielded many remarkable insights, such as the relationship between selection and recombination for an allele, *HBBC*, under strong selection (Wood et al., 2005). In this case, it is also possible to estimate the approximate age of the mutation as just a few centuries, on reasonable assumptions about population size and other factors.

In the case of HLA, Hill et al. (1991) showed that the antigens HLA-Bw53 and HLA-DRB1*1302-DQB1*0501 are independently protective against malaria in a region of West Africa, providing more protection than the hemoglobin alleles because of their population frequencies. This is a case of special interest as it involves both a single allele of HLA class I and a haplotype of class II, and if selection has been constant (an unrealistic assumption), a selection duration of 2000 to 7000 years, not nearly long enough to induce LD found in the haplotype or between the two classes. The major histocompatibility complex is also involved in reducing the parasite burden in other primates, such as the fat-tailed dwarf lemur *Cheirogaleus medius* and the Malagasy mouse lemur *Microcebus murinus* (Schwensow et al., 2007)

Quite a different matter is whether these strong selective interactions should ever have guided thinking about common diseases, which usually cause post-reproductive morbidity and mortality. As other examples of past strong selection contributing to present-day genomes arise (e.g., the proposal of Kaiser et al. [2007] that resistance to an extinct retrovirus has left humans vulnerable to HIV infection), they will provide opportunities to test the hypothesis directly, but few of the genes associated with malaria appear to be relevant in complex, common human disorders.

Association With Disease

The original evidence for balanced polymorphism came mostly from studies like that of Fisher (1939), involving strong natural selection on visible traits in insects and other organisms that could be studied experimentally. Human populations could only provide indirect evidence, though this was forthcoming in substantial quantities. For example, associations between particular human diseases and particular polymorphisms were detected and verified some fifty years ago. Table 2 displays a number of these associations. The work tended to be conducted because the polymorphisms were assayed for other reasons than investigation of disease association (e.g. blood groups of blood donors), so that the studies, individually well conducted, were not systematic or in many cases followed up. In this context, it is noteworthy that some established before the genomics era have stood the test of time, and have led to other investigations of association, linkage and causation. In these cases, it is likely that there are causal factors in the chromosomal regions marked by the polymorphism in question, even if, as opponents of this research used to claim, the associations were not clinically useful.

Genome-wide scans using very large numbers of DNA markers, particularly single nucleotide polymorphisms (SNPs), are now being conducted systematically, and will certainly lead to the discovery of genes significant in disease causation. They could also assist in disproving the CV-CD hypothesis, despite being based by definition on assay of common

Table 2

Associations Between Human Polymorphism (Other Than DNA Polymorphism) and Disease

Polymorphism	Disease	Phenotypes compared	Relative risk	Reference
ABO blood groups	Duodenal ulcer	O:A	1.90	McConnell 1969
	Gastric ulcer	O:A	1.19	McConnell 1969
	Colorectal carcinoma	O:A	0.90	Vogel 1970
			0.42	Guleria et al. 2005
	Mammary cancer	O:A	0.92	Vogel 1970
			0.31	Guleria et al. 2005
Secretor system	Duodenal ulcer	se:Se	1.46	McConnell 1969
	Gastric ulcer	se:Se	1.19	McConnell 1969
Major histocompatibility complex (HLA) A locus	Psoriasis vulgaris	A13: not-A13	4.30	Svejgaard et al. 1975
Major histocompatibility complex (HLA) B locus	Ankylosing spondylitis	B27: not-B27	120.90	Svejgaard et al. 1975
α_1-antitrypsin	Emphysema	normal: deficient	27.67	Tarkoff et al. 1968

https://doi.org/10.1375/twin.10.6.793 Published online by Cambridge University Press

Table 3

Results of One Multi-Disease Genome Scan (Wellcome Trust Case Control Consortium 2007)

Disease	No. genes detected at very high significance level	Genes previously identified as associated
Bipolar disorder	1	
Coronary artery	1	*APOE*
Hypertension	0	
Crohn's	9	
Rheumatoid arthritis	2	
Diabetes type I	7	*HLA-DRBI*
Diabetes type II	3	*HLA-DRBI*

variants, that is, polymorphisms. It is in this context that LD comes to the fore: substantial LD over short intervals arises because a selectively neutral polymorphism arises by mutation and sampling close to a causal gene, and the association with the polymorphism arises because of LD.

Table 3 shows some of the results of one very large genome-wide scan for seven diseases, using more than 300,000 SNPs and 2,000 cases for each disease with 3,000 unaffected controls for all diseases.

Table 4 shows some of the results of a number of other DNA based genome scans or studies, several published as followups to that in Table 3.

It is noteworthy that many studies have been conducted of asthma traits, and many regions have been identified in different studies, but few regions have been consistently associated with the traits. In other words, in different samples or populations, either different regions have been genuinely associated with the traits, or many of the associations have been spurious. Both explanations are sure to have some validity. However, if we consider the original CD–CV hypothesis, genuine but different associations in different populations imply that different common variants are causal in different populations, or that different causal variants are common in the different populations.

It should also be noted that the predictive power of these associations is usually modest, even when the P-value associated with the result is very low indeed. For example, consider the association of colorectal cancer (CRC) susceptibility with alternative SNP alleles in a region of Chromosome 8q24. The odds ratio is only about 1.20, that is, persons carrying one allele are only about 20% more likely to suffer from CRC than persons carrying the other allele. Tomlinson et al. (2007) concluded as follows: 'Given that the allele frequency of rs6983267 [the 'susceptibility allele'] is ~50% in the European population, the risk associated with hetero- and homozygosity implies that this variant underlies ~20% of CRCs. Although it may account for only ~1.1% of the excess familial risk of CRC, under a multiplicative model of susceptibility, the locus has the potential to have a substantial effect on an individual's risk by acting in concert with other low-penetrance alleles. These data provide the strongest evidence to date for common CRC susceptibility alleles'. However, the study has in fact not shown that a rare, highly penetrant allele in strong LD with rs6983267 is not the causal agent, unlikely though this may be.

Interaction

All of the analyses considered above have been based on the hypothesis of independent gene action. It is certain that this hypothesis is false, in the majority of cases, however effective it may have been in practice (especially when analyses have been undertaken as effectively marginal; Ewens &Thomson, 1977). Methods of multivariate analysis of fitness are being developed (e.g. Jones et al., 2007), but they are not yet applicable to human data. At the same time, cases of strong interaction are being detected as the ability to analyse metabolic pathways genomically increases, such as the role of histone demethylases in X-linked mental retardation (Tahiliani et al., 2007). However, the extent of interaction involving a human trait or disorder has not yet been assayed in the same way as individual loci affecting a trait. This is not the case in other species, though of course similar scans of individual loci in other animals are conducted on the same basis as in humans, with the possibility of experimental verification of what has been discovered (e.g. see the analysis of feed conversion efficiency in cattle by Barendse et al., 2007). In the plant kingdom, experiments can readily be designed to detect interactions.

Luo et al. (2001) conducted a cross between a japonica rice variety and an indica variety and assessed heterosis and the contribution to heterosis of epistasis. They had over 250 recombinant inbred lines from the F10. Epistasis was taken to be a digenic phenomenon. They used the estimation procedure of Wang et al. (1999), and they grew the experiment at two locations, which allowed a ready comparison of consistency. They found 30 QTL, seven, fifteen, and eight for panicles/plant, grains/panicle, and 1000-grain weight respectively, and only one was consistent between the two environments. They found eleven, twelve, and twelve digenic epistatic interactions for these three components of yield, and none was consistent between the two environments. Furthermore, only eight of the 70 possible main effects associated with these interactions were significant at the authors' chosen .1% significance level. Taken together, these results imply very high levels of genotype–environment interaction. They also show that heterosis is not the mirror-image of inbreeding depression.

Whether these results, from a normally inbreeding species, have any relevance to humans is uncertain, apart from suggesting the likelihood of discovering very many interactions, if not genotype-environment interactions. Use of the tool PLINK (http://pngu.mgh.harvard.edu/~purcell/plink/) may change this situ-

https://doi.org/10.1375/twin.10.6.793 Published online by Cambridge University Press

Table 4

Results of Genome Scans

Disease or condition	No. DNA polymorphisms	No. subjects, families, and so forth	No. regions detected at very high significance level	Genes identified	Confirmation of regions previously identified as associated	Chromosomes involved	Reference
Asthma traits	200-1500		0 (1 at moderate level)		Incomplete	(20q13)	Ferreira et al. 2005
Childhood asthma	> 300,000 SNPs	1000 cases, 1200 controls	1	*ORMDL3*		17q21	Moffatt et al. 2007
Endometriosis	400		0 (1 at moderate level)		Confirmation of one 'suggestive' finding	(10q26)	Treloar et al. 2005
Coeliac disease		Follow-up to study in Table 2		*TENR, KIAA1109, IL2, IL21*		4q27	van Heel et al. 2007
Crohn's disease		Follow-up to study in Table 2		*IRGM*		5q33.1 1q, 5p13	Parkes et al. 2007
Type 1 diabetes		Follow-up to study in Table 2				6p21	
	> 500,000 SNPs	3 patient groups > 500, control groups	10	*KIAA0350*	*HLA-DRB1, BTNL2, PTPN22, INS*	16p13	Hakonarson et al. 2007
Type 2 diabetes	> 300,000 SNPs	8 case-control groups	1	*TCF2* (HNF1β)			Gudmunsson et al.2007
Breast cancer	10,000 SNPs	14 families	0 (1 at moderate level)	*BRCA1/2* not involved in these families		0q23.32-q25.3	Bergman et al. 2007
	> 400 microsatellite	14 families	0 (1 at moderate level)	*BRCA1/2* not involved in these families	No confirmation of 13q21	2q32	Huusko et al. 2004
Prostate cancer	> 400 microsatellite	175 families	0 (1 at moderate level)	Close to *BRCA1*	No confirmation of 1, 20 or X	17q	Lange et al. 2003
	> 300,000 SNPs	4 patient groups > 500, control groups	2		17q12	17q12, 17q24.3	Gudmunsson et al. 2007
Colorectal cancer	100,000 SNPs	4 patient groups > 500 (>7,000 patients), > 7,000 controls	1	Also prostate cancer association		8q24	Zanke et al. 2007, Haiman et al. 2007
	> 500,000 SNPs	1000 patients, 1,000 controls	1			8q24.21	Tomlinson et al. 2007

ation, but it is unlikely to simplify what is known. (Use of PLINK has confirmed at least one case of epistasis; see Hakonarson et al., 2007.) As noted earlier just two diallelic loci require estimation of genotypic effects for nine genotypes, so without any assumptions about gene action, this requires estimation of at least nine parameters. Sepúlveda et al. (2007) have presented a model and analysis for a pair of diallelic loci interacting in a way that can be analysed as variable penetrance, thereby potentially reducing the number of parameters to be estimated, and have applied it to mouse backcross data on susceptibility to *Plasmodium berghei ANKA*. Such methods cannot be applied realistically when many interactions have been identified.

What is interpreted as variable penetrance may be related to other phenomena which have only recently been clearly identified, but which are now being intensively investigated. Epigenetic regulation is a prime example. X-chromosomal inactivation was the first human case to be well characterised, but imprinting and other types of epigenetics have been established in a number of cases. Tsankova et al. (2007) have listed nine disorders, all involving mental impairment, which arise from chromatin remodelling, in most cases hypo- or hypermethylation. It is noteworthy than many of these syndromes, though clearly coherent and likely to be genetical in origin, showed no familiality when originally defined (e.g. Angelman's syndrome, Mayo et al., 1973). Experiments on laboratory animals show that chromatin remodelling can occur developmentally or through long-term treatment, and this seems likely to apply more widely, so that disorders of late onset with risk factors such as chronic over-nutrition or lack of exercise or smoking would be candidates.

Conclusion

As noted earlier, the sheer number of polymorphisms identified by DNA analysis ruled out universal heterozygous advantage as a possible hypothesis to explain widespread variation of any kind. This conclusion applies even after allowing for forms of interaction other than multiplicative (Sved et al., 1967).

Table 4 allows another conclusion to be drawn: common variants are indeed associated with disease, as first established half a century ago, but the frequent lack of consistency among samples or populations shows that the causal relationship is not clear, even if the common variants are not themselves associated through linkage disequilibrium with different rare variants. Furthermore, because selection is largely postreproduction, causal hypotheses must build on very weak selection.

To go from an association to a causal relationship requires several steps. First, the association must be verified, either in fresh samples from the original population or in samples from other populations. This step has been taken in a number of cases, as illustrated in Tables 2 to 4. Second, other causes of the association (such as population stratification through introgression, for example, large groups of immigrants from one population into two others, the association then being detected in the two recipient populations, e.g. Köhler & Bickeböller, 2005) must be ruled out. Third, the problem mentioned in the previous paragraph must be addressed. That is, LD analysis must show definitively that the association is with the gene (structural gene, regulatory element etc.) itself, not with another in LD with it. An example of an appropriate analysis is given by Lowe et al. (2007). Fourth, the nature of the causal relationship must be elucidated. In the case of complex disorders of medium to late onset, the problem is particularly difficult: the cause of a disease may be a combination of environmental stress with moderate genetical propensity. For example, some cardiovascular disease may result from sedentary habit plus excessive food intake plus genetically determined high blood pressure. Is the latter then an outcome of inadequate level of production of some gene product, so that a functional pathway adequate over a short time-span is inadequate over a longer time span, or is it a slightly less effective gene product?

In some cases, where an environmental trigger is well understood or at least known, the fourth requirement can be met. For example, consider some of the polymorphisms associated with lung cancer. Glutathione S-transferases are likely to be involved in the catabolism of many of the compounds introduced into the lung in smoking, and the very frequent null allele of *GSTM1* has a relative risk of at least 1.3 compared with the normal allele (Kihara et al., 1994). Hung et al. (2003) showed, additionally, that cytochrome P450 1A1 (*CYP1A1*) interacted with GSTM1 in protecting/predisposing towards lung cancer. Here, we probably have real examples of CD–CV causation, but we have returned also to a very strong and well identified environmental selective agent, smoking, just as we began with malaria.

The CD–RV hypothesis, which is not really antithetical to CD–CV since they lie at the ends of a continuous spectrum of frequency and influence, is, as noted earlier, analogous to the case of rare monogenic disorders maintained at low frequency by mutation and selection. It has long been known that many of these disorders are much more frequent in one population than in others, such as Huntington's chorea (1.7 in 10,000 in Tasmania), adult type Gaucher's disease (4/10,000 in Israel), porphyria variegata (30/10,000 South African whites), adrenogenital syndrome (20/10,000 Yupik Eskimos; Brock, 1972). These high frequencies may have arisen through drift or bottlenecking or causally, but they illustrate the fact that deleterious genes can increase in frequency in isolated, small, or derivative populations, such as all human populations were once. Furthermore, the genes involved in different populations are themselves different. The simplest explanation for this fact is chance. See Pritchard (2001) for further discussion and modelling.

https://doi.org/10.1375/twin.10.6.793 Published online by Cambridge University Press

If we bring these facts together, and consider a complex disorder influenced by many genes, we have the following likely causal circumstances. A complex disorder is the outcome of the interaction of a number of genes with the environment over many years. The genetical component of the causal complex may involve alleles of many of the genes in the networks that influence the physiology underlying the disorder. A particular allele of a particular gene may be identified as causal in one population because it is frequent enough for its familial outcomes to be identified and measured, yet in another population it may be so rare as not to be detectable because it does not contribute significantly to the assessed genetical variation. Thus, both CV and RV contribute to complex disorders, depending how one assesses them, and we also establish a simple explanation for the frequent failures of replication of association between allele and disorder illustrated in Table 4.

References

Allendorf, F. W., & Leary, R. F. (1986). Heterozygosity and fitness in natural populations of animals. In M. E. Soulé (Ed.), *Conservation biology: The science of scarcity and diversity* (pp. 57–76). New York: Sinauer Associates.

Allison, A. (1955). Aspects of polymorphism in man. *Cold Spring Harbor Symposia on Quantitative Biology, 20,* 239–255.

Barendse, W., Reverter, A., Bunch, R. J., Harrison, B. E., Barris, W., & Thomas, M. B. (2007). A validated whole-genome association study of efficient food conversion in cattle. *Genetics, 176*, 1893–1905.

Bennett, J. H. (Ed.). (1983). *Natural selection, heredity, and eugenics including selected correspondence of R. A. Fisher with Leonard Darwin and others*. Oxford: Oxford University Press.

Bergman, A., Karlsson, P., Berggren, J., Martinsson, T., Björk, K., Nilsson, S., Wahlström, J., Wallgren, A., & Nordling, M. (2007). Genome-wide linkage scan for breast cancer susceptibility loci in Swedish hereditary non-BRCA1/2 families: Suggestive linkage to 10q23.32-q25.3. *Genes, Chromosomes and Cancer, 46*, 302–309.

Brock, D. J. H. (1972). Inborn errors of metabolism. In D. J. H. Brock. & O. Mayo (Eds.), *The biochemical genetics of man* (pp. 385–476). London: Academic Press.

Bürger, R. (2000). *The mathematical theory of selection, recombination, and mutation*. Chichester and New York: John Wiley.

Butcher, L. M., Davis, O. S. P., Craig, I. W., & Plomin, R. (2007). *Genomewide QTL association scan of general cognitive ability using pooled DNA and 500k SNP microarrays*. Manuscript submitted for publication.

Corbo, R. M., & Scacchi, R. (1999). Apolipoprotein E (APO E) allele distribution in the world. Is APOE*4 a 'thrifty' allele? *Annals of Human Genetics, 63*, 301–310.

Crow, J. F. (1998). 90 years ago: The beginning of hybrid maize. *Genetics, 148*, 923–928.

Ebersberger, I., Metzler, D., Schwarz, C. & Pääbo, S. (2002). Genomewide comparison of DNA sequences between humans and chimpanzees. *American Journal of Human Genetics, 70*, 1490–1497.

Ewens, W. J., & Thomson, G. (1977). Properties of equilibria in multi-locus genetic systems. *Genetics, 87*, 807–819.

Ferreira, M. A. R., O'Gorman, L., Le Souëf, P., Burton, P. R., Toelle, B. G., Robertson, C. F., Visscher, P. M., Martin, N. G., & Duffy, D. L. (2005). Robust estimation of experimentwise P values applied to a genome scan of multiple asthma traits identifies a new region of significant linkage on chromosome 20q13. *American Journal of Human Genetics, 77*, 1075–1085.

Fisher, R. A. (1922). On the dominance ratio. *Proceedings of the Royal Society of Edinburgh, 42,* 321–341.

Fisher, R. A. (1930). *The genetical theory of natural selection. Oxford: Clarendon Press.*

Fisher, R. A. (1939). Selective forces in wild populations of *Paratettix texanus*. *Annals of Eugenics, 9*, 109–122.

Fisher, R. A. (1941). Average excess and average effect of a gene substitution. *Annals of Eugenics, 11,* 53–63.

Fisher, R. A., & Ford, E. B. (1947). The spread of a gene in natural conditions in a colony of the moth Panaxia dominula L. *Heredity, 1*, 143–174.

Fisher, R. A., & Ford, E. B. (1950). The 'Sewall Wright' effect. *Heredity, 4*, 117–119.

Fitzpatrick, M. J., Feder, E., Rowe, L. & Sokolowski, M. B. (2007). Maintaining a behaviour polymorphism by frequency-dependent selection on a single gene. *Nature, 447*, 210–213.

Ford, E. B. (1940). Genetic research in the Lepidoptera. *Annals of Eugenics, 10*, 227–252.

Ford, E. B. (1975). *Ecological genetics*. (4th ed., ch. 7) London: Chapman & Hall.

Grant, S. F., et al. (2006). Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. *Nature Genetics, 38*, 320–323.

Guleria, K., Singh, H. P., Kaur, H., & Sambyal, V. (2005). ABO blood groups in gastrointestinal tract (GIT) and breast carcinoma patients. *Anthropologist, 7*, 189–192.

Haiman, C. A., Le Marchand, L., Yamamoto, J., Stram, D. O., Sheng, X., Kolonel, L. N., Wu, A. H., Reich, D., & Henderson, B. E. (2007). A common genetic risk factor for colorectal and prostate cancer. *Nature Genetics, 39*, 954–956.

Hakonarson, H., et al. (2007). A genome-wide association study identifies KIAA0350 as a type 1 diabetes gene. *Nature, 448*, 591–595.

Haldane, J. B. S. (1949) The rate of mutation of human genes. Hereditas Supplement, 35, 267–273.

Hill, A. V. S., Allsopp, C. E. M., Kwiatowski, D., Anstey, N. M., Twumasi, P., Rowe, P. A., Bennett, S., Brewster, D., McMichael, A. J., & Greenwood, B. M. (1991). Common West African HLA antigens are associated with protection from severe malaria. *Nature, 352*, 595–600.

Hung, R. J., Boffetta, P., Brockmöller, J., Butkiewicz, D., Cascorbi, D., Clapper, M. L., Garte, S., Haugen, A., Hirvonen, A., Anttila, S., Kalina, I., Le Marchand, L., London, S. J., Rannug, A., Romkes, M., Salagovic, J., Schoket, B., Gaspari, L., & Taioli, E. (2003). CYP1A1 and GSTM1 genetic polymorphisms and lung cancer risk in Caucasian non-smokers: A pooled analysis. *Carcinogenesis, 24*, 875–882.

Huusko, P., et al. (2004). Genome-wide scanning for linkage in Finnish breast cancer families. *European Journal of Human Genetics, 12*, 98–104.

Jones, A. G., Arnold, S. J., & Bürger, R. (2007). The mutation matrix and the evolution of evolvability. *Evolution, 61*, 727–745.

Kaiser, S. M., Malik, H. S., & Emerman, M. (2007). Restriction of an extinct retrovirus by the human TRIM5a protein. *Science, 316*, 1756–1758.

Keller, M. C., & Miller, G. (2006). Resolving the paradox of common, harmful, heritable mental disorders: Which evolutionary genetic models work best? *Behavioral and Brain Sciences, 29*, 385–452.

Kihara, M., Kihara, M., & Noda K. (1994). Lung cancer risk of GSTM1 null genotype is dependent on the extent of tobacco smoke exposure. *Carcinogenesis, 15*, 415–418.

Köhler, K., & Bickeböller, H. (2005). Case-control association tests correcting for population stratification. *Annals of Human Genetics, 69*, 98–115.

Kwiatowski, D. P. (2005). How malaria has affected the human genome and what human genetics can teach us about malaria. *American Journal of Human Genetics, 77*, 171–192.

Landsteiner, K. (1900). Zur Kenntnis der antifermentativen, lytischen und agglutinierenden Wirkungen des Blutserums und der Lymphe. *Zentralblatt fur Bakteriologie, Parasitenkunde, Infektionskrankheiten und Hygiene, 27*, 357–362.

Lange, E. M., Gillanders, E. M., Davis, C. C., Brown, W. M., Campbell, J. K., Jones, M., Gildea, D., Riedesel, E., Albertus, J., Feas-Lutz, D., Markey, C., Giri, V., Dimmer, J. B., Montie, J. E., Trent, J. M., & Cooney, K. A. (2003). Genome-wide scan for prostate cancer susceptibility using families from the University of Michigan prostate cancer genetics project finds evidence for linkage on chromosome 17 near BRCA1. *The Prostate, 57*, 326–334.

Laurie C. C., Nickerson D. A., Anderson, A. D., Weir, B. S., Livingston, R. J., Dean M. D., Smith, K. L., Schadt, E. E., & Nachman, M. W. (2007). Linkage disequilibrium in wild mice. *PLoS Genetics, 9*, e144, 1–9.

Leach, C. R., & Mayo, O. (2005). *Outbreeding mechanisms in flowering plants an evolutionary perspective from Darwin onwards*. Stuttgart: Gebrüder Borntraeger Verlagsbuchhandlung.

Lesbarrères, D., Schmeller, D. S., Primmer, C. R., & Merilä, J. (2007). Genetic variability predicts common frog (Rana temporaria) size at metamorphosis in the wild. *Heredity, 99*, 41–48.

Lowe, C. E., Cooper, J. D., Brusko, T., Walker, N. M., Smyth, D. J., Bailey, R., Bourget, K., Plagno, V., Field, S., Atkinson, M., Clayton, D. G., Wicker, L. S., & Todd, J. A. (2007). Large-scale genetic fine mapping and genotype phenotype associations implicate polymorphism in the IL2RA region in type 1 diabetes. *Nature Genetics, 39*, 1074–1082.

Luciano, M., Wright M. J., Duffy, D. L., Wainwright, M. A.,G. Zhu, G., Evans, D. M., Geffen G. M., Montgomery G. W., & Martin, N. G. (2006). Genome-wide scan of IQ finds significant linkage to a quantitative trait locus on 2q. *Behavior Genetics, 36*, 45–56.

Luo, L. J., Li, Z.-K., Mei, H. W., Shu, Q. Y., Tabien, R., Zhong, D. B., Ying, C. S., Stansel, J. W., Khush, G. S., & Paterson, A. H. (2001). Overdominant epistatic loci are the primary genetic basis of inbreeding depression and heterosis in rice. II. Grain yield components. *Genetics, 158*, 1755–1771.

Mandel, S. P. H. (1959). The stability of a multiple allelic system. *Heredity, 13*, 289–302.

Martin, M. P., Qi, Y., Gao, X., Yamada, E., Martin, J. N., Pereyra, F., Colombo, S., Brown, E. E., Shupert, W. L., Phair, J., Goedert, J.J., Buchbinder, S., Kirk, G. D., Telenti, A., Connors, M., O'Brien, S. J., Walker, B. D., Parham, P., Deeks, S. G., McVicar, D. W., & Carrington, M. (2007). Innate partnership of HLA-B and KIR3DL1 subtypes against HIV-1. *Nature Genetics, 39*, 733–740.

Mayo, O. (1972). Polymorphism, selection and evolution. In D. J. H. Brock & O. Mayo (Eds.), *The biochemical genetics of man* (pp. 131–155). London: Academic Press.

Mayo, O., Nelson, M. M. & Townsend, H. R. A. (1973). Three more 'happy puppets'. *Developmental Medicine and Child Neurology, 15*, 6374.

McConnell, R. B. (1969). *The genetics of gastro-intestinal disorders*. Oxford:Oxford University Press.

Moffatt, M. F., et al. (2007). Genetic variants regulating *ORMDL3* expression contribute to the risk of childhood asthma. *Nature, 448*, 470–474.

Mørch, E. T. (1941). *Chondrodystrophic dwarfs in Denmark*. Copenhagen: Munksgaard.

Muller, H. J. (1918). Genetic variability: Twin hybrids and constant hybrids in a case of balanced lethal factors. *Genetics, 3*, 422–499.

https://doi.org/10.1375/twin.10.6.793 Published online by Cambridge University Press

Neel, J. V. (1962). Diabetes mellitus: A 'thrifty' genotype rendered detrimental by 'progress'? *American Journal of Human Genetics, 14,* 353–362.

Parkes, M., et al. (2007). Sequence variants in the autophagy IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility. *Nature Genetics, 39*, 830–832.

Plomin, R., Hill, L., Craig, I. W., McGuffin, P., Purcell, S., Sham P., Lubinski, D., Thompson, L. A., Fisher, P. J., Turic, D., & Owen, M. J. (2001). A genome-wide scan of 1842 DNA markers for allelic associations with general cognitive ability: A five-stage design using DNA pooling and extreme selected groups. *Behavior Genetics, 31*, 497–509.

Posthuma, D., Luciano, M., de Geus, E. J. C., Wright, M. J., Slagboom, P. E., Montgomery, G. W., Boomsma, D. I., & Martin, N. G. (2005). A genomewide scan for intelligence identifies quantitative trait loci on 2q and 6p. *American Journal of Human Genetics, 77*, 318–326.

Pritchard, J. K. (2001). Are rare variants responsible for susceptibility to complex diseases? *American Journal of Human Genetics, 69*, 124–137.

Reich, D. E. & Lander, E. S. (2001). On the allelic spectrum of human disease. *Trends in Genetics, 17,* 502–510.

Robertson, A. (1962). Selection for heterozygotes in small populations. *Genetics, 47*, 1291–1300.

Rodd, Z. A., Bertsch, B. A., Strother, W., N., Le-Niculescu, H., Balaraman, Y., Hayden, E., Jerome, R. E., Lumeng, L., Nurnberger, J. I., Edenberg, H. J., Mcbride, W. J., & Niculescu, A. B. (2007). Candidate genes, pathways and mechanisms for alcoholism: An expanded convergent functional genomics approach. *Pharmacogenomics Journal, 7*, 222–256.

Schwensow, N., Fietz, J., Dausmann, K. H. & Sommer, S. (2007). Neutral versus adaptive genetic variation in parasite resistance: Importance of major histocompatibility complex supertypes in a free-ranging primate. *Heredity, 99,* 265–277.

Sepúlveda, N., Paulino, C. D., Carneiro, J., & Penha-Gonçalves, C. (2007). Allelic penetrance approach as a tool to model two-locus interaction in complex binary traits. *Heredity, 99*, 173–184.

Tahiliani, M., Mei, P., Fang, R., Leonor, T., Rutenberg, M., Shimizu, F., Li, J., Rao, A., & Shi, Y. (2007). The histone H3K4 demethylase SMCX links REST target genes to X-linked mental retardation. *Nature, 447*, 601–606.

Sved, J. A., Reed, T. E., & Bodmer, W. F. (1967). The number of balanced polymorphisms that can be maintained in a natural population. *Genetics, 55*, 459–481.

Sved, J. A. (1968). The stability of linked systems of loci with a small population size. *Genetics, 59*, 543–563.

Svejgaard, A., Platz, P. J., Ryder, L. P., Staub Nielsen, L., & Thomson, M. (1975). HL-A and disease associations. A survey. *Transplantation Research, 22*, 3–43.

Tarkoff, M. P., Kueppers, P., & Miller, W. F. (1968). Pulmonary emphysema and alpha1-antitrypsin. deficiency. *American Journal of Human Genetics, 45*, 220–228.

Tomlinson, I., et al. (2007). A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. *Nature Genetics, 39*, 984–988.

Treloar, S. A., Wicks, J., Nyholt, D. R., Montgomery, G. W., Bahlo, M., Smith, V., Dawson, G., Mackay, I. J., Weeks, D. E., Bennett, S. T., Carey, A., Ewen-White, K. R., Duffy, D. L., O'Connor, D. T., Barlow, D. H., Martin, N. G., & Kennedy, S. H. (2005). Genomewide linkage study in 1,176 affected sister pair families identifies a significant susceptibility locus for endometriosis on chromosome 10q262005. *American Journal of Human Genetics, 77*, 365–376.

Tsankova, N., Renthal, W., Kumar, A., & Nestler, E. J. (2007). Epigenetic regulation in psychiatric disorders. *Nature Reviews Neuroscience, 8*, 355–367.

van Heel, D. A., et al. (2007). A genome-wide association study for celiac disease identifies risk variants in the region harbouring IL2 and IL21. *Nature Genetics, 39*, 827–829.

Vogel, F. (1970). ABO blood groups and disease. American Journal of Human Genetics, 22, 464–475.

Wang, D. L., Zhu, J., Li, Z. K., & Paterson, A. H. (1999). Mapping QTL with epistatic effects and QTL x environment interactions by mixed model approaches. *Theoretical and Applied Genetics, 99*, 1255–1264.

Weatherall, D. J. (2007). *George Fraser and the 'malaria hypothesis' in fifty years of human genetics a Festschrift and liber amicorum to celebrate the life and work of George Robert Fraser*. Adelaide:Wakefield Press.

Wellcome Trust Consortium. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature, 447*, 661–683.

Wood, E. T., Stover, D. A., Slatkin, M., Nachman, M. W., & Hammer, M. F. (2005). The b-globin recombinational hotspot reduces the effect of strong selection against HbC, a recently arisen mutation providing resistance to malaria. *American Journal of Human Genetics, 77*, 637–642.

Zanke, B. W., et al. (2007). Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. *Nature Genetics, 39*, 989–994.