Review

Human Heredity

Hum Hered 2015;80:165–170
DOI: 10.1159/000446931

Published online: September 1, 2016

Concepts and Misconceptions about the Polygenic Additive Model Applied to Disease

Peter M. Visscher Naomi R. Wray

Queensland Brain Institute, The University of Queensland, Brisbane, Qld., Australia

Key Words
Threshold model · Heritability of liability · Disease risk

Abstract
It is nearly one hundred years, since R.A. Fisher published his now famous paper that started the field of quantitative genetics. That paper reconciled Mendelian genetics (as exemplified by Mendel's peas) and the biometrical approach to quantitative traits (as exemplified by the correlation and regression approaches from Galton and Pearson), by showing that a simple model of many genes of small effects, each following Mendel's laws of segregation and inheritance, plus environmental variation could account for the observed resemblance between relatives. In this review, we discuss a number of concepts and misconceptions about the assumptions and limitations of polygenic models of common diseases in human populations.

Historical Context

Quantitative genetics, as a scientific field of research, was firmly established by the 1960s, with theory broadly consistent with empirical data from selection experiments in model organisms and genetic improvement programs in animal and plant breeding. In particular, a polygenic model underlying quantitative trait genetic variation was widely accepted. For binary (0–1) traits, a threshold (liability) model had been proposed [1], although its uptake was not without controversy (see for example the review by Fraser [2]). Although this model was not widely adopted by human geneticists as a model for disease, there were exceptions, see for example Carter [3] and Gottesman and Shields [4], and some landmark papers, e.g. [4, 5], for schizophrenia provided empirical evidence which showed that the risk to relatives was consistent with such a model.

For common diseases in humans, such as psychiatric disorders, heart disease and hypertension, the prevailing paradigm in the 1970s was Mendelian, i.e. that the cause of disease in an affected individual is due to a single factor, usually a single mutation or sometimes an environ-

KARGER 125 YEARS

© 2016 S. Karger AG, Basel
0001–5652/16/0804–0165$39.50/0

E-Mail karger@karger.com
www.karger.com/hhe

Peter M. Visscher
Queensland Brain Institute, The University of Queensland
Brisbane, QLD 4072 (Australia)
E-Mail Peter.Visscher@uq.edu.au

mental insult (e.g. a head injury leading to a brain disease). In fact, for some researchers working in human genetics, this is still the paradigm today, despite strong empirical evidence against this model. Moreover, some of the misconceptions of polygenic models discussed in Fraser's 1976 review [2] are still just as relevant today.

Genetic Factors and Risk to Relatives

Diseases that we term 'common' may still affect only a small proportion of the population, and the first line of evidence that these diseases may be underpinned by genetic factors, at least in part, comes from an increased risk of disease in relatives of those affected. James [6] provided a simple framework to equate the observable data with standard quantitative genetic models. He showed that on the observed probability (or risk of disease) scale, the risk to relative with relationship R (K_R) can be expressed in terms of the risk in the population (K) and the phenotypic covariance between probands and their relatives on the observed 0–1 scale (cov_R)

$$K_R = K + cov_R/K. \quad (1)$$

This equation is completely general and does not depend on assumptions about the sources of the phenotypic covariance, nor about any underlying continuous scale. It can also be expressed as

$$\lambda_R - 1 = cov_R/K^2, \quad (2)$$

with λ_R being the relative risk to relatives (K_R/K) [7]. Equations 1 and 2 are sometimes referred to as the 'James' Identity' [8].

For many diseases, adoption studies imply only a small contribution from environmental factors to the phenotypic covariance between relatives. So if we assume a genetic model such that the only covariance between relatives is due to genetic factors, then the phenotypic covariance on the 0–1 scale can be decomposed into genetic variance components [6],

$$\mathrm{cov}_{\mathrm{R}} = \sum\nolimits_{k=0}^{\infty} \sum\nolimits_{l=0}^{\infty} r^k u^l V_{A(k)D(l)} ,$$

where $V_{A(k)D(l)}$ denotes the genetic variance components with k A and l D terms, given an additive genetic relationship coefficient of r and a dominance coefficient of relationship of u. So for R = monozygotic (MZ) twins, $r = 1$ and $u = 1$ and,

$$cov_{MZ} = V_A + V_D + V_{AA} + V_{AD} + V_{DD} + V_{AAD} + V_{AAA} + \ldots = V_G. \quad (3)$$

Likewise for R = full-sibs (FS), where $r = {}^1/_2$ and $u = {}^1/_4$,

$$\mathrm{cov}_{\mathrm{FS}} = \frac{V_A}{2} + \frac{V_D}{4} + \frac{V_{AA}}{4} + \frac{V_{AD}}{8} + \frac{V_{DD}}{16} + \frac{V_{AAD}}{16} + \frac{V_{AAA}}{8} + \ldots \quad (4)$$

James's [6] contribution is important because this formulation provides a framework to test the observed frequencies (or relative risks) in relatives against single-locus versus multi-locus models. A single-locus model for human disease – the prevailing view at that time – does not give rise to epistatic variance and hence gives different predictions of risk across different classes of relatives compared to multi-locus models [6]. Therefore, it provides a strategy, in principle (if the data sets are sufficiently large and if there are multiple classes of relatives), to test the validity of a single-locus model compared to a multi-locus model from observable data.

Liability Threshold Model

A particular multi-locus model is the threshold model that Falconer [9] and Crittenden [10] described building on the work of others. An unobserved liability threshold model to explain observations on discrete characters was first proposed by Sewall Wright in the context of the number of toes in guinea pigs [1]. The quantitative genetic theory that showed the correspondence between genetic parameters on a scale of liability and an observed binary scale was developed later [11], in particular in the appendix developed by Alan Robertson. The theory in this paper used a linear approximation from a Taylor series expansion to transform from an additive scale of liability to a discrete 0–1 scale. Assuming that the liability threshold model is a reasonable model to explain observations on a binary scale, the paper by Dempster and Lerner [11] had important implications for the response to selection in populations undergoing selection. In particular, if genetic variation on the scale of liability is fully additive, then genetic variation on a binary scale can be highly epistatic. This implies that the estimation of additive genetic variance on the observed scale from the resemblance between relatives is biased upwards, the bias depending on whether close relatives (e.g. full-sibs) or more distant relatives (e.g. half-sibs) are used for estimation. In addition, the predicted response to selection based upon narrow sense heritability on the 0–1 scale can be biased downwards or upwards, the bias being a function of heritability of liability, population prevalence and selection intensity.

The linear transformation of heritability from the observed 0–1 scale to that of liability is

$$h_l^2 \approx \frac{h_o^2 \, K\left(1-K\right)}{z^2}, \tag{5}$$

with K being the population lifetime prevalence and z the height of the normal curve at the truncation point pertaining to K [11]. In human studies (of the pre-genomics era), accurate estimates of narrow-sense heritability on the observed scale were hard to achieve, as the sample sizes were limited and disease status could only be recorded on close family members, e.g. identical or non-identical twin pairs.

Falconer [9] (and Crittenden [10]) showed that the resemblance between relatives on discrete scales can be framed in the theory of response to truncation selection, and derived the estimation of heritability on the scale of liability directly from the lifetime prevalence (called 'incidence' by Falconer) in probands (ascertained individuals who have the trait/disease of interest) and their relatives. Falconer's method to estimate heritability is a linear regression of mean liability of relatives of probands on mean liability of probands, both as a deviation from the population mean. This is analogous to a ratio of response to selection and selection differential, and he showed that heritability of liability could be estimated from two measures, risk of disease in the population (K) and risk of disease in relatives of those affected (K_R),

$$h_l^2 = \frac{T - T_R}{a_R i}, \tag{6}$$

where a_R is the additive genetic relationship between the relatives, $T = \Phi^{-1}(1 - K)$ and $T_R = \Phi^{-1}(1 - K_R)$ are the thresholds of the normal distribution that truncate proportions K and K_R, respectively. i is the mean liability of the diseased group in the population, calculated as $i = z/K$, where $z = \varphi(T)$, as in Fisher [12].

The derivation of the expected disease concordance rate for MZ twin pairs under a liability threshold model was an extension from Charles Smith [13]. Smith made an important observation that the expected concordance rate can be low even when the heritability of liability is high. Conversely, a low MZ concordance rate for a disease with prevalence of, say, ≤1% does not imply that genetic factors are unimportant. To this day, there is much confusion in human genetics about the relationship between heritability and disease concordance in relatives, particularly MZ twins. Smith expanded on this study by showing how the proband concordance rate can be used to estimate heritability of liability from a design including both dizygotic and MZ twin pairs [14].

Assumptions of Polygenic Models

In Fisher's model, a large number of variants of small effects leads, by the central limit theory, to normal distributions of genetic effects (subsequently called 'breeding values' in the quantitative genetics literature in the absence of non-additive variation, but initially called 'essential genotypes' by Fisher [12]). The model is sometimes called the infinitesimal model, but one does not need an infinite number of variants to approach normality (actually only a small number, of the order of 10 or so, are needed). The gene effects do not have to be the same nor does gene action need to be strictly additive. In fact, Fisher partitioned his genetic variance into additive and dominant components of variation.

Clearly, the liability threshold model is only one of many possible models that link multifactorial contributions to disease risk to the observed binary outcome. A key feature of any multi-locus model for disease that affects only a small proportion of the population is a highly non-linear relationship between burden of disease alleles and risk of disease [15]. The liability threshold model is the simplest representation, as it depends on only two parameters: risk of disease in the population and the proportion of variance on the liability scale attributed to genetic factors; hence avoiding modelling individual loci by representing many genetic architectures in terms of number and frequency of risk loci that each could equate to explain the same total variance. It is incorrect to state that the liability threshold model to disease, and the concept of a heritability of liability, cannot be tested with empirical data. If we knew all risk variants for a particular disease, and their effect sizes were estimated without error, then the liability threshold model could be tested by comparing observed and predicted probabilities of disease given genotype. In terms of modelling the relationship between multiple risk factors and binary outcome, the liability threshold model is (almost) identical to a probit model. In addition, an underlying continuous logistic scale, as routinely used in epidemiology, results in almost identical observations and inference [16]. Hence whether there is a hypothetical threshold on an unobserved normal or logistic scale makes little difference to inference about genetic factors on disease.

Additive and Non-Additive Variation

There is a great deal of confusion about the meaning and implication of additive effects and additive genetic variation. This is the case for quantitative traits and even worse for binary traits. Fisher parameterised his genetic variance from a regression of phenotype on genotype (here, genotype can be thought of as 0, 1 or 2 alleles at a particular locus). The regression variance is the additive genetic variation and the residual variance is the dominance variance. This partitioning is very useful because it leads to a natural prediction of the response to natural or artificial selection and the resemblance between relatives. The confusion arises because the relationship between gene action (additive, when the mean value of a heterozygous genotype is exactly between the mean value of the homozygous genotypes, or dominant/recessive, when it is not) and additive and dominance variation depends on allele frequency. Additive genetic variation is the variance of 'average effects' of an allele, which is the regression coefficient from the regression of phenotype on genotype. Because of the dependency of the variance components on allele frequency, a strong deviation from additive gene action can result in mostly additive genetic variation. This apparent paradox becomes more extreme when modelling higher-order epistatic interactions in multi-loci models: the more higher-order interactions, the more additive genetic variation [17]. This observation does not mean that the partitioning of observed variance components in additive and non-additive genetic variation is not useful or wrong. In practice, it means that loci can be detected from modelling simple additive relationships between the trait and genotypes (SNP dosage).

Non-additive effects and their resulting variance (if any) depend on scale because they are statistically interaction effects. For binary disease traits, this begs the question of what scale is appropriate for variance partitioning and heritability. Empirical observations on the recurrence risk of relatives and the effect of individual genetic variants on disease imply that a hypothetical continuous scale of risk (liability) fits the data better than an additive model on the binary scale [18]. This is no different from the conclusions from epidemiological models of environmental risk factors for common complex diseases. Hence, the transformation to the liability scale for disease provides a useful framework to model disease on an additive scale, which fully implies that on the disease scale the mode of action is highly non-additive. Many researchers set out to detect non-additive effects between loci, but the burden of multiple testing means that such studies are restricted to scans of two- or three-locus interactions. For diseases with a multi-genic or polygenic architecture, such approaches are misguided as the non-additive action that leads to disease is likely to reflect an increase in risk of disease associated with a burden of risk alleles, i.e. high-order interactions (or genotypic context [19]). Such a disease architecture is consistent with robustness to lower burden of risk alleles.

We Do Not Need to Know Disease Etiology to Quantify and Estimate Genetic Variation for Disease

From a statistical or genetic (biological) point of view, if the aim is to estimate genetic variation and to dissect it into contributions from individual loci, there is no reason to treat human disease differently from quantitative traits in humans or from disease in other species. There is nothing wrong in using data on the observed 0–1 scale as if it is a quantitative trait. Animal breeders have been selecting against common diseases in their species of interest doing just that, and with success (as measured by changes in mean incidence over generations). In recent years, the introduction of methods that estimate genetic variation contributing to human disease [20, 21] has provided an important stepping stone to the ultimate identification of individual risk loci. This is because smaller sample sizes are needed to estimate the total contribution of genetic loci which have provided confidence to continue investing in increasing sample sizes to allow studies that are sufficiently powered to detect individual loci.

A Model Is Just That

All statistical models make assumptions, and clearly some models cannot be true (e.g. there are not an infinite number of genetic variants contributing to disease risk). However, models of disease can be useful and make predictions that can be tested empirically. When GWAS experiments started to discover robust associations between SNP allele and risk to disease, with small effect sizes, a polygenic liability threshold model of disease would predict that with increasing sample size very many such associations would be detected, and this is exactly what happened subsequently. For example, for schizophrenia there was a single locus detected in 2009 from about 3,000 cases and 3,000 controls [22], yet by 2014 that number had risen to over 100 [23], consistent with theory.

There are a number of diseases for which there are common variants of relatively large effect (e.g. auto-immune diseases and dementia) that in combination with polygenic variation contribute to overall genetic risk. Assuming an infinitesimal model of liability to disease for such diseases would be inefficient when applied to risk prediction [24, 25]. However, it would not necessarily lead to biased estimation of genetic variance from GWAS SNP data.

Conclusion

Empirical data on human diseases and disorders have been used to quantify recurrence risk of relatives and to dissect population genetic variation into contributions from individual loci. It is outside of the scope of this paper to review that literature in detail, but there are some broad conclusions that can be drawn. Decades of studies in large epidemiological cohorts and twin studies [26] imply that most recurrence risk to relatives is due to genetic factors and that the data are consistent with a polygenic model of liability. Gene mapping studies in pedigrees have rarely led to the identification of mutations of polymorphisms with a large effect on risk in those pedigrees. Population-based gene mapping studies, such as GWAS, have detected robust associations between genetic variants and a disease or disorder. The identification of tens or hundreds of variants, each with a small effect, has provided empirical evidence that there are many more such variants to be detected and that one-third to two-thirds of genetic variation inferred from pedigree data is captured by SNP arrays. Very recently, a number of functional studies have led to the resolution of SNP-disease associations to the level of individual nucleotides [27, 28]. In addition, there is little empirical evidence to support widespread variation in liability to disease caused by either gene-gene or gene-environmental interactions on that scale. Clearly, there are many exceptions to these broad conclusions. They include the detection of near-Mendelian mutations by linkage studies in families, the contribution of a burden of rare coding mutations to some disorders (e.g. [29]) and the existence of genes of large effects for some common diseases/disorders (particularly autoimmune disorders). Nevertheless, across common diseases and disorders, a model such as the liability threshold model is remarkably consistent with empirical data and useful for statistical analysis to quantify and partition genetic variation in the population, detect new loci and to generate genetic predictors.

Acknowledgements

The authors are supported by grants from the Australian National Health and Medical Research Council, the Australian Research Council and the US National Institutes of Health.

Disclosure Statement

The authors have no conflicts of interest to declare.

References

1 Wright S: An analysis of variability in number of digits in an inbred strain of guinea pigs. Genetics 1934;19:506–536.
2 Fraser FC: Multifactorial-threshold concept – uses and misuses. Teratology 1976;14:267–280.
3 Carter CO: Genetics of common disorders. Br Med Bull 1969;25:52–57.
4 Gottesman, II, Shields J: A polygenic theory of schizophrenia. Proc Natl Acad Sci USA 1967;58:199–205.
5 McGue M, Gottesman II, Rao DC: The transmission of schizophrenia under a multifactorial threshold-model. Am J Hum Genet 1983;35:1161–1178.
6 James JW: Frequency in relatives for an all-or-none trait. Ann Hum Genet 1971;35:47–49.
7 Risch N: Linkage strategies for genetically complex traits. 1. Multilocus models. Am J Hum Genet 1990;46:222–228.
8 Lynch M, Walsh B: Genetics and Analysis of Quantitative Traits, ed 1. Sunderland, Sinauer Associates, 1998.
9 Falconer DS: The inheritance of liability to certain diseases, estimates from the incidence among relatives. Ann Hum Genet 1965;29:51–76.
10 Crittenden LB: An interpretation of familial aggregation based on multiple genetic and environmental factors. Ann NY Acad Sci 1961;91:769–780.
11 Dempster ER, Lerner IM: Heritability of threshold characters. Genetics 1950;35:212–236.
12 Fisher RA: The correlation between relatives on the supposition of Mendelian inheritance. Trans R Soc Edinburgh 1918;53:399–433.
13 Smith C: Heritability of liability and concordance in monozygous twins. Ann Hum Genet 1970;34:85–91.
14 Smith C: Concordance in twins – methods and interpretation. Am J Hum Genet 1974;26:454–466.
15 Slatkin M: Exchangeable models of complex inherited diseases. Genetics 2008;179:2253–2261.
16 Wray NR, Goddard ME: Multi-locus models of genetic risk of disease. Genome Med 2010;2:10.
17 Maki-Tanila A, Hill WG: Influence of gene interaction on complex trait variation with multilocus models. Genetics 2014;198:355–367.
18 McGue M, Gottesman II, Rao DC: Resolving genetic models for the transmission of schizophrenia. Genet Epidemiol 1985;2:99–110.
19 Sackton TB, Hartl DL: Genotypic context and epistasis in individuals and populations. Cell 2016;166:279–287.

20 Lee SH, Wray NR, Goddard ME, Visscher PM: Estimating missing heritability for disease from genome-wide association studies. Am J Hum Genet 2011;88:294–305.

21 Bulik-Sullivan BK, Loh PR, Finucane HK, Ripke S, Yang J; Schizophrenia Working Group of the Psychiatric Genomics Consortium, Patterson N, Daly MJ, Price AL, Neale BM: LD score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat Genet 2015;47:291–295.

22 Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF, Sklar P: Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature 2009;460:748–752.

23 Schizophrenia Working Group of the Psychiatric Genomics Consortium: Biological insights from 108 schizophrenia-associated genetic loci. Nature 2014;511:421–427.

24 Moser G, Lee SH, Hayes BJ, Goddard ME, Wray NR, Visscher PM: Simultaneous discovery, estimation and prediction analysis of complex traits using a Bayesian mixture model. PLoS Genet 2015;11:e1004969.

25 Abraham G, Rohmer A, Tye-Din JA, Inouye M: Genomic prediction of celiac disease targeting HLA-positive individuals. Genome Med 2015;7:72.

26 Polderman TJ, Benyamin B, de Leeuw CA, Sullivan PF, van Bochoven A, Visscher PM, Posthuma D: Meta-analysis of the heritability of human traits based on fifty years of twin studies. Nat Genet 2015;47:702–709.

27 Sekar A, Bialas AR, de Rivera H, Davis A, Hammond TR, Kamitaki N, Tooley K, Presumey J, Baum M, Van Doren V, Genovese G, Rose SA, Handsaker RE; Schizophrenia Working Group of the Psychiatric Genomics Consortium, Daly MJ, Carroll MC, Stevens B, McCarroll SA: Schizophrenia risk from complex variation of complement component 4. Nature 2016;530:177–183.

28 Claussnitzer M, Dankel SN, Kim KH, Quon G, Meuleman W, Haugen C, Glunk V, Sousa IS, Beaudry JL, Puviindran V, Abdennur NA, Liu J, Svensson PA, Hsu YH, Drucker DJ, Mellgren G, Hui CC, Hauner H, Kellis M: FTO obesity variant circuitry and adipocyte browning in humans. N Engl J Med 2015;373:895–907.

29 Chang J, Gilman SR, Chiang AH, Sanders SJ, Vitkup D: Genotype to phenotype relationships in autism spectrum disorders. Nat Neurosci 2015;18:191–198.