Global Journal of Science Frontier Research: F
Mathematics and Decision Sciences
Volume 14 Issue 5 Version 1.0 Year 2014
Type : Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 2249-4626 & Print ISSN: 0975-5896

Estimating Effects and Variance Components in Models of Quantitative Genetics in an Era of Sequenced Genomes

By Charles J. Mode
Drexel University, United States

Abstract- As in many other areas of research in genetics, the availability of sequenced genomes in samples of individuals has revolutionized the study of quantitative traits, because researches have developed statistical evidence regarding the lo- cations of genomic regions, loci, that have been implicated with the expression of a quantitative trait or traits. Therefore, in cases in which it is possible to develop operational definitions of at least two alleles at each locus, genomic regions, it becomes possible to identify the genotype of each individual with re- spect to a set of loci that have been shown in other experiments to influence the expression of a quantitative trait. As will be shown in this paper, by knowing the genotype of each individual in a sample with respect to a set of identified loci, it is now possible to directly estimate effects that are measures of not only intra-allelic interactions at each locus under consideration but also various types of epistatic effects that are measures of interactions among alleles at different loci, governing the inheritance of a quantitative trait.

Keywords: genomic regions implicated with a quantitative trait, loci,locus and alleles, known genotypes, effects as measures of intra-allelic and epistatic interactions, direct estimates of effects, phenotypic, genetic and environmental variance components, partitioning the genetic variance into additive, intra-allelic interaction and epistatic components of variance

GJSFR-F Classification : MSC 2010: 97K80

Strictly as per the compliance and regulations of :

Ref

[21] Stranger, B. E. , Eli A. , Stahl, E. A. and Raj. T. (2010) Progress and promise of genome-wide association studies for human complex trait genetics. 10.1534/genetics.110.120907

Estimating Effects and Variance Components in Models of Quantitative Genetics in an Era of Sequenced Genomes

Charles J. Mode

Abstract- As in many other areas of research in genetics, the availability of sequenced genomes in samples of individuals has revolutionized the study of quantitative traits, because researches have developed statistical evidence regarding the lo- cations of genomic regions, loci, that have been implicated with the expression of a quantitative trait or traits. Therefore, in cases in which it is possible to develop operational definitions of at least two alleles at each locus, genomic regions, it becomes possible to identify the genotype of each individual with re- spect to a set of loci that have been shown in other experiments to influence the expression of a quantitative trait. As will be shown in this paper, by knowing the genotype of each individual in a sample with respect to a set of identified loci, it is now possible to directly estimate effects that are measures of not only intra-allelic interactions at each locus under consideration but also various types of epistatic effects that are measures of interactions among alleles at different loci, governing the inheritance of a quantitative trait. These straight forward methods of estimation differ from those used in classical quantitative genetics, because such effects and corresponding variance components could be estimated indirectly, using analysis of variance procedures or some version of general lin- ear models that have been and are widely in statistical genetics. The direct method of estimation described in this paper, show promise towards shifting the working paradigm that has been used in classical models of the genetics of quantitative traits involving the estimation of variance components to a more direct approach and simpler approach.

Keywords: genomic regions implicated with a quantitative trait, loci,locus and alleles, known genotypes, effects as measures of intra-allelic and epistatic interactions, direct estimates of effects, phenotypic, genetic and environmental variance components, partitioning the genetic variance into additive, intra-allelic interaction and epistatic components of variance.

I. Introduction

In an interesting review paper by Stranger et al. (2010) [21], the impact of genome wide associations studies on the genetics of complex traits is discussed in depth. Among these complex traits are Alzheimer's disease (AD) and immune-mediated diseases such as rheumatoid arthritis. For the case of AD, in a recent paper Raj et al. (2012) [19] have reported that 11 regions of the human genome are involved in susceptibility to this disease, and, moreover, there is evidence that four of these regions form a protein network that is under natural selection. Similarly, in paper by Rossin et al. (2011) [20], it has been found that proteins encoded in genomic regions associated with immune-mediated disease physically interact and this interaction may also suggest some basic biological mechanisms underlying such diseases.

Author: Professor Emeritus, Department of Mathematics, Drexel University, Philadelphia, USA. e-mail: cjmode@comcast.net

There is also another technological development, the sequencing of entire genomes of individuals, that may lead to a deeper understanding of the relationships of phenotypes to genotypes. Suppose, for example, that a sample of individuals with symptoms of a disease, such as AD, is available and that genome of each individual in the sample has been sequenced. Furthermore, suppose that some quantitative measurement is made on each individual in the sample. These measurements will vary among individuals and let W denote a random variable characterizing this variation. Given that the genome of each individual in the sample has been sequenced, the genotype of each individual in the sample can, in principle, be identified with respect to the 11 loci under consideration for the case of AD. It will also be supposed that at each locus at least two alleles can be identified.

In classical quantitative genetics, the loci and alleles at each locus were treated abstractly, because an investigator did not, in general, know the location of the hypothesized loci in the genome of a species or the number of alleles at each locus. However for the case of AD cited above, the genotype of each individual in the sample can be identified with respect to each of the 11 loci, and in some cases it may be known with respect to combinations of the 11 loci or even all 11 loci. Such technological developments provide opportunities to extend some of the ideas of classical quantitative genetics into the age of sequenced genomes. Moreover, as will be demonstrated in subsequent sections of this paper, when the genotype of each individual in a sample is known, the estimation of parameters of the model may be carried out in a relatively simple and straight forward manner based on elementary methods of statistical estimation.

As is recognized among many who have worked in the field of quantitative genetics, the subject known as components of variance analysis began with the publication of a paper on correlations among relatives on the supposition of Mendelian inheritance by R. A. Fisher (1918) [7]. In his paper, Fisher attempted to reconcile existing biometrical theories with Mendelian genetics that led him to describe genetic variation in terms of components of variance. During the 1950s, other investigators published papers that were motivated by the paper by Fisher. Among these investigators was Kempthorne (1954) [10], who introduced an approach to components of variance analysis based on effects defined in terms of expectations of genetics values with respect to the genotypic distribution under the assumption that the population was in a Hardy-Weinberg equilibrium. An alternative approach was introduced by Cockerham (1954) [5] is also of historical interest, because it contains an extensions of Fisher's ideas to accommodate epistatic effects in terms of ideas depending on the concept of orthogonality. If a reader is interested in further details and development of the ideas of Fisher and other workers, it is suggested that the book Kempthorne (1957) [11] be consulted, where many of the themes of statistical genetics as they existed during the 1950s were summarized and extended.

The techniques introduced in these papers have also been applied in the current genomic era. Examples of the ideas introduced by Cockerham have been applied in the paper Kao et al. (2002) [9], and those of Kempthorne have been applied and extended in the paper Mao et al. (2006) [15]. The ideas of Kempthorne were also used and extended in the paper of Mode and Robinson (1959) [16] as well as in unpublished lecture notes by the author written and presented during the period 1960 to 1966. Furthermore, the roots of the ideas presented in this paper are extensions of the some of the unpublished material in the lecture notes complied by the author during the period 1960 to 1966.

During the years following Fisher's seminal work, an extensive literature on quantitative genetics has evolved. It is beyond the scope of this paper to review this literature and in what follows a few books on the subject will be cited. A book that has been very popular with quantitative geneticists is that

Ref

[10] Kempthorne, O. (1954) The Correlations Between Relatives in a Random Mating Population. Proc. Royal Soc. London, B 143: 103-113.

[6] Falconer, D. and MacKay, T. F. C. (1996) Introduction to Quantitative Genetics. Longman, New York.

of Falconer and MacKay (1996) [6] as well as earlier editions. Another book of interest on quantitative genetics is that of Bulmer (1980) [4]. Both of these books contain extensive lists of references on quantitative genetics. A more recent book on genetics and analysis of quantitative traits is that of Lynch and Walsh (1998) [14]. This influential tome consists of over 900 pages and contains what seems to be the most extensive treatment of the subject of quantitative genetics published in the 20-th century. The principal focus of this book is a biological and evolutionary point of view along with an extensive use of applied statistical methods. There is also an extensive list of papers on quantitative genetics that a reader, who is interested in quantitative genetics, may wish to peruse. The book by Liu [13] on statistical genetics focuses on statistical genetics along with linkage, mapping and quantitative trait linkage (QTL) analysis. Two recent books on statistical genetics are those of Laird and Lange [7] and Wu, Ma and Casella (2010) [22].

Historically, procedures for estimating components of the genetic and environmental variances have been based on experimental designs or observational data involving various types of relatives. If a reader is interested in an account of such experimental designs, it is suggested that chapter 6 of Bulmer (1980) [4] be consulted. An in depth account of estimation procedures in various genetic settings may be found in section III of the book by Lynch and Walsh (1998) [14]. In this paper, however, it will be shown that when the genotype of each individual in a sample is known at the DNA level, then it is possible to estimate various types of genetic parameters directly, including variance components, using elementary statistical ideas. It should also be mentioned that the ideas presented in this paper are extensions of techniques from unpublished notes on quantitative genetics written by the author during the period 1960 to 1966. In these notes, it was assumed that for the one locus case the population was in a Hardy-Weinberg equilibrium, and for the case of multiple loci, it was assumed that the population was in linkage equilibrium. In this paper, however, these assumptions have been relaxed.

When two or more quantitative traits are under consideration several measurements are taken on each individual. In this case, it is assumed that the autosomal loci under consideration may influence the expression of alleles for two or more traits. In classical genetics, such joint expressions of alleles for quatitative or qualitative traits is referred to pleiotropism. In a recent paper, Mode (2014) [17] this case has been worked out in detail.

II. The Case of One Locus with Multiple Alleles

Let $\mathbb{A}$ denote a finite set of alleles at some autosomal locus in a diploid species such as man. Elements of $\mathbb{A}$ will be denoted by the symbols x and y, and the genotype of an individual with respect to the locus will be denoted by (x, y), where $x \in \mathbb{A}$ and $y \in \mathbb{A}$ denote, respectively, the alleles contributed the maternal and paternal parent of the individual under consideration. As the technology underlying the sequencing of DNA evolves, it seems likely that it will be possible to distinguish the DNA contributed by each parent to an offspring. More precisely, let $\mathbb{A} \times \mathbb{A}$ denote the Cartesian product of the set $\mathbb{A}$ with itself. Then $\mathbb{G} = \mathbb{A} \times \mathbb{A}$ is the set of all possible genotypes at the locus under consideration and $(x, y) \in \mathbb{G}$ for every genotype $(x.y)$.

One of the objectives in formulating models in quantitative genetics is to provide a framework such that phenotypic measurements on a population of individuals may formally be connected with the genotype of each individual. For many decades it has been observed that phenotypic measurements among individuals with the same genotype in a given environment may vary. But, it has also been observed that in populations consisting of several genotypes responses of the genotypes to a given environment may also vary. Let W denote a random variable that takes values in the set $\mathbb{R}_W$ of real numbers that constitute the

set of possible phenotypic measurements of individuals in the population. In general, by assumption, the numbers in the set $\mathbb{R}_W$ will depend on the genotype of a homogeneous set of individuals.

Given a genotype $(x,y) \in \mathbb{G}$, let $f(w \mid (x,y))$ denote the conditional probability density function of the random variable W. Then,

$$E[W \mid (x,y)] = \mu(x,y) = \int_{\mathbb{R}_W} w f(w \mid (x,y))\, dw \tag{2.1}$$

is the conditional expectation of the random variable W, given the genotype (x,y). It will be assumed that $\mu(x,y)$ is finite for all genotypes $(x,y) \in \mathbb{G}$. Let $p(x,y)$ denote the probability, frequency, that an individual chosen at random from the population is of genotype (x,y). Then, the unconditional expectation of the random variable W is, by definition,

$$\mu = E[W] = \sum_{(x,y)} p(x,y) E[W \mid (x,y)] = \sum_{(x,y)} p(x,y) \mu(x,y). \tag{2.2}$$

It is assumed that $p(x,y) \geq 0$ for all $(x,y) \in \mathbb{G}$ and

$$\sum_{(x,y)} p(x,y) = 1. \tag{2.3}$$

In what follows, it will also be helpful to observe that the joint distribution of a random genotype (x,y) and the phenotypic random variable W is $g((x,y),w) = p(x,y) f(w \mid (x,y))$ for all $(x,y) \in \mathbb{G}$ and $w \in \mathbb{R}_W$.

Next observe that the equation

$$W = \mu + (\mu(x,y) - \mu) + (W - \mu(x,y)) \tag{2.4}$$

is valid and provides a linear relationship connecting an observed phenotypic measurements W with the expectation μ, a measurement of a genetic effect expressed by the deviation $(\mu(x,y) - \mu)$ and the term $(Z - \mu(x,y))$, which may be interpreted as a measure of deviation of the phenotypic measure W from $\mu(x,y)$ due to environmental conditions. By definition, the total phenotypic variance in the population is

$$var_P[W] = E\left[(W-\mu)^2\right]. \tag{2.5}$$

A widely used technique in quantitative genetics is to partition the total phenotypic variance into a genotypic variance measuring the variation among genotypes in their responses to environmental conditions and an environmental variance measuring the variation of the phenotypic measure W around the genotypic values $\mu(x,y)$ for every $(x,y) \in \mathbb{G}$. From equation (2.4), it follows that

$$(W-\mu)^2 = (\mu(x,y) - \mu)^2 + (W - \mu(x,y))^2 + 2(\mu(x,y) - \mu)(W - \mu(x,y)). \tag{2.6}$$

Therefore,

$$\begin{aligned} E\left[(W-\mu)^2 \mid (x,y)\right] &= (\mu(x,y) - \mu)^2 + (W - \mu(x,y))^2 \\ &\quad + 2(\mu(x,y) - \mu) E[(W - \mu(x,y)) \mid (x,y)]. \end{aligned} \tag{2.7}$$

But,

$$E[(W - \mu(x,y)) \mid (x,y)] = 0. \tag{2.8}$$

Therefore,

$$E\left[(W-\mu)^2 \mid (x,y)\right] = (\mu(x,y) - \mu)^2 + (W - \mu(x,y))^2. \tag{2.9}$$

In deriving equation (2.9), some well known properties of conditional expectations have been used. Namely, for any function $f(x)$ with domain a subset of $\mathbb{R}$, the set of real numbers, and range a subset of $\mathbb{R}$, it follows from well

known properties of conditional expectations that $E[f(X) \mid X] = f(X)$. It is also well known that if two random variables X and Y with range $\mathbb{R}$ are under consideration, then $E[XY \mid X] = XE[Y \mid X]$.

An equivalent representation of the phenotypic variance in equation (2.5) is

$$var_P[W] = \sum_{(x,y)} p(x,y) E\left[(W-\mu)^2 \mid (x,y)\right]. \tag{2.10}$$

Given (2.9), it seems reasonable to define and genetic variance due to genetic effects in the population as

$$var_G[W] = \sum_{(x,y)} p(x,y) (\mu(x,y) - \mu)^2. \tag{2.11}$$

Similarly, the variance due to environmental effects is, by definition,

$$var_E[W] = \sum_{(x,y)} p(x,y) E[(W - \mu(x,y))^2 \mid (x,y)]. \tag{2.12}$$

From equation (2.9) it follows, therefore, that

$$var_P[W] = var_G[W] + var_E[W]. \tag{2.13}$$

At this point in the development of the contents of this paper, it should be mentioned that equation (2.13) is not new to quantitative genetics, but its derivation is a departure from derivations that appeared in some papers and books on the subject. For example, in some formulations the effects $(\mu(x,y) - \mu)$ and $(W - \mu(x,y))$ are treated as abstract uncorrelated random variables and sometimes it is assumed that genetic and environmental effects are independent. But, it follows from the use of conditional expectations that these types of assumptions are not necessary in the derivation of (2.13).

From equation (2.13), it can be seen that the total phenotypic variance may be partitioned into two component variances; namely the genetic and environmental variance. If both sides of equation (2.13) are divided by the phenotypic variance, then it is easy to see that

$$1 = F_G + F_E,$$

where

$$F_G = \frac{var_G[Z]}{var_P[Z]}$$

and

$$F_E = \frac{var_E[Z]}{var_P[Z]}. \tag{2.14}$$

Some authors refer to F_G as a measure of the heritability of quantitative trait. In what follows, F_G will be denoted by H_G and referred to as a measure of heritability.

This latter ratio has been given various names by authors of books on quantitative genetics. For example, in the book by Falconer and Mackay (1996) [6] on page 123 an expression similar to the ratio H_G is called the degree of genetic determination. Other authors such as Wu et al. (2010) [22] refer to this ratio as heritability in the broad sense and provide an example in which this parameter may be estimated by an analysis of variance procedure based on a designed breeding experiment, see page 178. If a reader is interested in pursuing the subject of heritability further, it is suggested that the book by Liu (1998) [13] be consulted, in particular see pages 34 and 35. A more in depth treatment of the concept of heritability may be found in the book by Lynch and Walsh (1998) [14], see pages 170 to 175 and elsewhere in that book.

Several recent papers have also been devoted to applications of the concept of heritability. Among these papers is that of Zaitlen et al. (2013) [24], who use

Ref

[14] Lynch, M. and Walsh, B. (1998) Genetics and the Analysis of Quantitative Traits. Sinauer Associates, Inc. Sunderland, MA, 01375.

extended genealogies to estimate components of heritability for 23 quantitative and dichotomous traits, using closely and distantly related relatives. In an interesting paper Price et al. (2011 [11] estimated variance components using single tissue data on cross-tissue heritability gene expression on individuals related by descent and also unrelated individuals. In a paper by Yang et al. (2010) [23] the heritability of human height is studied. These authors show that by considering all common $SNPs$ simultaneously, 45% of the phenotypic variance in human height can be attributed to genetic variation. It should be mentioned, however, that quantitative model or models used by these authors were not as comprehensive as the structure that will be developed in subsequent sections of this paper.

In a subsequent section of this paper, procedures for estimating the components of variance just defined will be presented. It is recognized, however, that an investigator may be interested in testing statistical hypotheses as to whether the expectations and variances among the genotypes do indeed differ, but a discussion of tests of hypotheses is beyond the scope of this paper, which will be limited to a presentation of straight forward procedures for estimating the components of variance defined above.

Before preceding to a discussion of estimation procedures, however, it is interesting to note that, even though the rhetoric in this section was confined to the case on one autosomal locus with multiple alleles, the formulas can be easily extended to the case of some finite number of autosomal loci $n \geq 2$ with a finite number of alleles at each locus. For let $(\boldsymbol{x}, \boldsymbol{y})$ denote the genotype of an individual with respect to n loci, where $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)$ and $\boldsymbol{y} = (y_1, y_2, \cdots, y_n)$ denote, respectively, the alleles inherited from the maternal and paternal parent, and suppose $p(\boldsymbol{x}, \boldsymbol{y})$ is the probability of selecting an individual of genotype $(\boldsymbol{x}, \boldsymbol{y})$ at random in the population. Then, it is easy to show that equation (2.13) also holds for some number $n \geq 2$ of autosomal loci, but the details of proving this statement will be left as an exercise of the reader.

Given that it can be shown that equation (2.13) holds for any number of loci $n \geq 2$, it is interesting to note that with respect to AD the total phenotypic variance may be partitioned into the genetic and environmental components for any combination of the 11 loci that have been implicated with this disease. In particular, it would be of interest to estimate the heritability for each of the 11 loci or in combinations of loci in order to gain some insights as to whether heritability would increase as the number of loci under increases.

III. A Partition of the Genetic Variance Into the Additive and Intra-Allelelic Components for the Case of One Autosomal Locus

Again let $p(x, y)$ denote the probability of finding an individual of genotype (x, y) a population. This probability is also known as the frequency of genotype $(x.y)$ in a population. In most past formulations of models in quantitative genetics for the one locus case, it has been assumed that a population was in a Hardy-Weinberg equilibrium and there was no mutation or selection. Mutation and selection will not be considered in this paper, but the condition that a population is in a Hardy-Weinberg equilibrium will be relaxed. Let $p(x)$ and $p(y)$ denote, respectively, the frequencies of alleles x and y in a population. Then, a population is in a Hardy-Weinberg equilibrium if $p(x, y) = p(x)\, p(y)$ for all genotypes $(x, y) \in \mathbb{G}$. In this section, the condition that a population is in a Hardy-Weinberg equilibrium will not be assumed, because in many populations this assumption may not hold. It should be mentioned, however, that an investigator may wish to test whether a sample from a population passes a statistical test or tests for a Hardy-Weinberg equilibrium.

Ref

[11] Kempthorne, O. (1957) An Introduction to Genetic Statistics. John Wiley and Sons, New York.

To relax the assumption that a population is in a Hardy-Weinberg equilibrium, it will be necessary to deal with conditional probabilities and expectations. Let

$$p(x) = \sum_y p(x,y) \tag{3.1}$$

denote that marginal distribution for all maternal alleles $x \in \mathbb{A}$ in a population, and similarly let

$$p(y) = \sum_x p(x,y) \tag{3.2}$$

denote the marginal distribution for all paternal alleles $y \in \mathbb{A}$. Then, if $p(y) \neq 0$

$$p(x \mid y) = \frac{p(x,y)}{p(y)} \tag{3.3}$$

is the conditional distribution of the alleles $x \in \mathbb{A}$, given allele $y \in \mathbb{A}$. Similarly, if $p(x) \neq 0$, then

$$p(y \mid x) = \frac{p(x,y)}{p(x)} \tag{3.4}$$

is the conditional distribution of alleles $y \in \mathbb{A}$, given allele $x \in \mathbb{A}$. The formulas just derived may be summarized in the equation

$$p(x,y) = p(x)\, p(y \mid x) = p(y)\, p(x \mid y) \tag{3.5}$$

for all genotypes $(x,y) \in \mathbb{G}$.

Therefore, the conditional expectation of $\mu(x,y)$, given x is, by definition,

$$\mu(x) = \sum_y p(y \mid x)\, \mu(x,y). \tag{3.6}$$

Therefore, the unconditional expectation of $\mu(x)$ is

$$E[\mu(x)] = \sum_x p(x)\, \mu(x) = \sum_x \sum_y p(x,y)\, \mu(x,y) = \mu. \tag{3.7}$$

For a justification of this equation, see equation (2.5). In particular, if the population is in Hardy-Weinberg equilibrium, then $p(x,y) = p(x)\, p(y)$ for all $(x,y) \in \mathbb{G}$ and equation (3.6) becomes

$$\mu(x) = \sum_y p(y \mid x)\, \mu(x,y) = \sum_y p(y)\, \mu(x,y), \tag{3.8}$$

because in this case $p(y \mid x) = p(y)$. Thus, in formulations in which the assumption that a population is in Hardy-Weinberg equilibrium is in force, (3.8) is the definition of the average value of maternal allele x in a population. Similarly, by using techniques similar to those used in the derivation of a formula for $\mu(x)$, it is straight forward to derive a formula for $\mu(y)$, the average value for paternal allele y in the population that is not in a Hardy-Weinberg equilibrium.

To cast the formulation in terms of an analysis of variance structure, it is useful to define the effects of alleles x and y as the deviations

$$\begin{aligned} \alpha(x) &= \mu(x) - \mu \\ &\text{and} \\ \alpha(y) &= \mu(y) - \mu. \end{aligned} \tag{3.9}$$

Observe that the unconditional expectations of these deviations is $E[\alpha(x)] = E[\alpha(y)] = 0$. The deviation

$$\alpha(x,y) = \mu(x,y) - \mu - \alpha(x) - \alpha(y) \tag{3.10}$$

is a measure of interactions among the maternal and paternal alleles. In this case, it is also easy to see that the unconditional expectation of this deviation is $E[\alpha(x,y)] = 0$. Alternatively, the deviations just described can be written in the form of an analysis of the variance equation

$$\mu(x,y) = \mu + \alpha(x) + \alpha(y) + \alpha(x,y), \tag{3.11}$$

which holds for all genotypes $(x,y) \in \mathbb{G}$. This equation suggests that it seems reasonable to call the terms $\alpha(x)$ and $\alpha(y)$ the additive effects of alleles. With the exception of μ, the terms on the right side of equation (3.11) are known statistically as effects. For if $\alpha(x,y) = 0$ for all genotypes, then

$$\mu(x,y) = \mu + \alpha(x) + \alpha(y) \tag{3.12}$$

for all $(x,y) \in \mathbb{G}$ so that the effects of alleles x and y have an additive effect on the expectation $\mu(x,y)$. But, if $\alpha(x,y) \neq 0$ for all (x,y), then there are interactions among the maternal and paternal alleles.

Having defined additive and intra-allelic interaction effects, the next step in the formulation is to define the additive and intra-allelic interaction variances. The additive genetic variance in the population is defined by

$$var_G(A) = \sum_x p(x)\alpha^2(x) + \sum_y p(y)\alpha^2(y), \tag{3.13}$$

and intra-allelic interaction, IAI, variance is defined by

$$var_G(IAI) = \sum_{(x,y)} p(x,y)\alpha^2(x,y). \tag{3.14}$$

To connect these variances with the total genetic variance in a population write equation (3.11) in the form

$$\mu(x,y) - \mu = \alpha(x) + \alpha(y) + \alpha(x,y)$$

and square both sides. The result is

$$(\mu(x,y) - \mu)^2 = \alpha^2(x) + \alpha^2(y) + \alpha^2(x,y) + R(x,y), \tag{3.15}$$

where

$$R(x,y) = 2\alpha(x)\alpha(y) + 2\alpha(x)\alpha(x,y) + 2\alpha(y)\alpha(x,y). \tag{3.16}$$

By multiplying equation (3.15) by $p(x,y)$ and summing over all genotypes (x,y), it follows that

$$var_G[W] = var_G(A) + var_G(IAI) + E[R(x,y)], \tag{3.17}$$

where

$$E[R(x,y)] = T_1 + T_2 + T_3. \tag{3.18}$$

The explicit forms of the symbols on the right, which involve covariances, are as follows:

$$\begin{aligned} T_1 &= 2\sum_{(x,y)} p(x,y)\alpha(x)\alpha(y) \\ T_2 &= 2\sum_{(x,y)} p(x,y)\alpha(x)\alpha(x,y) \\ &\quad \text{and} \\ T_3 &= 2\sum_{(x,y)} p(x,y)\alpha(y)\alpha(x,y). \end{aligned} \tag{3.19}$$

In general, $E[R(x,y)] \neq 0$, but there is a case when $E[R(x,y)] = 0$. Suppose the population is in a Hardy-Weinberg equilibrium so the $p(x,y) = p(x)p(y)$ for all genotypes $(x,y) \in \mathbb{G}$. Then, T_1 may be written in the form

$$T_1 = 2\left(\sum_x p(x)\alpha(x)\right)\left(\sum_y p(y)\alpha(y)\right). \tag{3.20}$$

But,

$$\sum_x p(x)\alpha(x) = 0,$$

and therefore $T_1 = 0$. Similarly, T_2 may be written in the form

$$T_2 = 2\left(\sum_x p(x)\,\alpha(x)\right)\left(\sum_y p(y)\,\alpha(x,y)\right). \tag{3.21}$$

For every fixed x, consider

$$\begin{aligned}\sum_y p(y)\,\alpha(x,y) &= \sum_y p(y)\left(\mu(x,y) - \mu - \alpha(x) - \alpha(y)\right)\\ &= \alpha(x) - \alpha(x) - \sum_y p(y)\,\alpha(y) = 0\end{aligned} \tag{3.22}$$

for every $x \in \mathbb{A}$. Therefore, $T_2 = 0$, and by a similar argument it can be shown that $T_3 = 0$ so that $E[R(x,y)] = 0$.

Thus, for the case a population is in a Hardy-Weinberg equilibrium at some autosomal locus, it follows that the total genetic variance may be partitioned into the additive and intra-allelic interaction variances. In symbols,

$$var_G[W] = var_G(A) + var_G(IAI). \tag{3.23}$$

It is interesting to observe that when the genotype of each individual may be identified, then each of the component variances on the right may be estimated separately. But, before the age of genomics, in quantitative genetic studies, the genotype of each individual in a population could not be identified. Under such circumstances, experiments could be designed in such a way that components of variance in equation (3.23) could be estimated from mean squares in an analysis of variance table. It should be noted, however, the when the effects $\alpha(x), \alpha(y)$ and $\alpha(x,y)$ can be estimated from the data, then all the covariances terms in $E[R(x,y)]$ could also be estimated. In such cases, one could also estimate the term $E[R(x,y)]$ in equation (3.17), which would be of interest in its own right for the cases in which the population was not in a Hardy-Weinberg equilibrium at the autosomal locus under consideration.

There is a notationally more succinct way to represent the variances and covariances encountered in the above discussion. For each genotype $(x,y) \in \mathbb{G}$ let the

$$\mathbf{\Phi}(x,y) = \begin{pmatrix} \alpha(x) \\ \alpha(y) \\ \alpha(x,y) \end{pmatrix} \tag{3.24}$$

denote a 3×1 matrix whose elements are defined above. The transpose of this matrix is

$$\mathbf{\Phi}^T(x,y) = \begin{pmatrix} \alpha(x) & \alpha(y) & \alpha(x,y) \end{pmatrix}. \tag{3.25}$$

Next observe that

$$\mathbf{\Psi}(x,y) = \mathbf{\Phi}(x,y)\,\mathbf{\Phi}^T(x,y) \tag{3.26}$$

is a 3×3 matrix and the element in position $(1,1)$ is $\alpha^2(x)$, the element in position $(1,2)$ is $\alpha(x)\,\alpha(y)$ and, by proceeding in this way, all nine of the element in the matrix $\mathbf{\Psi}(x,y)$ as squares or products of the elements in the vector $\mathbf{\Phi}(x,y)$. Let $\mathbf{\Psi}_G$ denote the 3×3 genetic variance-covarince matrix for the autosomal locus under consideration. Then,

$$\mathbf{\Psi}_G = \sum_{(x,y)} p(x,y)\,\mathbf{\Psi}(x,y). \tag{3.27}$$

From now on $\mathbf{\Psi}_G$ will be called the genetic covariance matrix for the autosomal locus under consideration. It should be observed that the variance components on the right of equation of equation (3.13) are in the principal diagonal positions $(1,1)$ and $(2,2)$ of the matrix $\mathbf{\Psi}_G$. Moreover, the sum of all elements off the principal diagonal of this matrix is the term $E[R(x,y)]$ in (3.17). Given the genetic matrix $\mathbf{\Psi}_G$, it may be useful to compute the eigenvalues of this matrix as well as its principal components in addition to estimating the components of the matrix $\mathbf{\Psi}_G$. As will be seen in subsequent sections, the matrix approach to computing the genetic covariance matrix described in this section will make it possible to describe the computation of the genetic

Notes

covariance matrix for cases in which more than one autosomal locus is under consideration.

It can be seen from a perusal of books on statistical genetics that the approach used in this section and in subsequent sections of this paper to partition the genetic variance into components differs from that used in some books cited in the introduction. For example, on page 54 of the book Laird and Lange (2011) [12] a phenotypic measurement Y of a quantitative trait is represented as a linear combination of unknown parameters with indicator functions coefficients plus a random error term. Included in these terms are parameters for the additive effects of allele as well as a codominant effect, which appears to be related to the intra-allelic interaction term defined this section. Such models appear to belong to the class of generalized linear models that are widely used in numerous areas of applied statistics. In particular, in the books on statistical genetics cited in the introduction, linear models similar to that cited in Laird and Lange (2011) have been used. As can be seen from the derivations presented in this section, however, the additive and interaction effects of alleles in the case of one autosomal locus are defined in terms of conditional expectations with respect to the genotypic distribution. Moreover, this scheme of using conditional expectations in defining effects when partitioning the total genetic variance into components will be used extensively in subsequent sections of this paper and provides a methodology for estimating effects and corresponding variance components directly from data. Furthermore, as will be shown subsequently, the squared effects making up a component of variance may also be estimated directly from a data set such that the genomes of all individuals in this sample have been sequenced.

IV. Estimating of Parameters and Effects From Data

In this section, a procedure for estimating the parameters defined in the forgoing sections from phenotypic data will be outlined. Suppose in a sample of individuals, $n((x,y)) \geq 2$ individuals of genotype $(x,y) \in \mathbb{G}$ are observed and let the random variables $W_\nu(x,y)$, for $\nu = 1, 2, \cdots, n(x,y)$, denote a sample of phenotypic measurements on the $n((x,y))$ individuals of genotype (x,y) with respect to some quantitative trait. Usually, the set of phenotypic measurements will belong to some set $\mathbb{R}$ of continuous real numbers will also be supposed that these random variable are independently and identically distributed according to common but unknown distribution with a finite expectation and variance. Let

$$n = \sum_{(x,y)} n(x,y) \tag{4.1}$$

denote the total number of individuals in the sample, where the sum runs over all genotypes $(x,y) \in \mathbb{G}$. Then, the random variable

$$\widehat{p}(n(x,y)) = \frac{n(x,y)}{n} \tag{4.2}$$

is an estimator of the frequency $p(x,y)$ of genotype (x,y) in a population or subpopulation from which a sample of individuals was drawn. It is interesting to note that if it is assumed the that the numbers $n(x,y)$ for $(x,y) \in \mathbb{G}$ are viewed as realizations from a multinomial distribution with probabilities $p(x,y)$ for $(x,y) \in \mathbb{G}$ and sample size n, then $E[\widehat{p}(n(x,y))] = np(x,y)/n = p(x,y)$ so that $\widehat{p}(n(x,y))$ is an unbiased estimator of $p(x,y)$ for all $(x,y) \in \mathbb{G}$.

Similarly, the random variable

$$\widehat{\mu}(x,y) = \frac{1}{n(x,y)} \sum_{\nu=1}^{n(x,y)} W_\nu(i,j) \tag{4.3}$$

is an estimator of the parameter $\mu(x,y)$. This estimator is conditionally unbiased, because $\widehat{E}[\widehat{\mu}(x,y) \mid (x,y)] = n(x,y)\mu(x,y)/n(x,y) = \mu(x,y)$ for all genotypes $(x,y) \in \mathbb{G}$. Therefore, the random variable

Ref
[12] Laird, N. M. and Lange, C. (2011) The Fundamentals of Modern Statistical Genetics. DOI 10.1007/978-1-4419-7338-2, Springer New York, Dordrecht, Heidelberg, London.

$$\widehat{\mu} = \sum_{(x,y)} \widehat{p}(n(x,y))\widehat{\mu}(x,y) \tag{4.4}$$

is an estimator of the parameter μ. From these definitions, it follows that the random variable

$$\widehat{var}_G[W] = \sum_{(x,y)} \widehat{p}(x,y)\left(\widehat{\mu}(x,y) - \widehat{\mu}\right)^2 \tag{4.5}$$

is an estimator of the genetic variance in (2.11).

To estimate the environmental variance defined in (2.12), let

$$\sigma^2(x,y) = E\left[\left(W(x,y) - \mu(x,y)\right)^2 \mid (x,y)\right] \tag{4.6}$$

for all genotypes $(x,y) \in \mathbb{G}$. Then

$$\widehat{\sigma}^2(x,y) = \frac{1}{n(x,y)-1} \sum_{\nu=1}^{n(x,y)} \left(W(x,y) - \widehat{\mu}(x,y)\right)^2 \tag{4.7}$$

is a conditionally unbiased estimator of $\sigma^2(x,y)$, given the genotype (x,y). Therefore,

$$\widehat{var}_E[W] = \sum_{(x,y)} \widehat{p}(x,y)\,\widehat{\sigma}^2(x,y) \tag{4.8}$$

is an estimator of the environmental variance defined in (2.12). From (2.13), it follows that an estimator of the phenotypic variance may be obtained by adding the estimators in (4.5) and (4.8), or this variance component could be estimated directly.

Given the estimators $\widehat{\mu}(x,y)$ for all genotypes in the sample, it would be straight forward to derive estimators of the three effects in the column vector $\mathbf{\Phi}(x,j)$ in (3.24) for all genotypes (x,y) in the sample. Let $\widehat{\mathbf{\Phi}}(i,j)$ denote the estimator of the vector $\mathbf{\Phi}(x,j)$ for all genotypes under consideration. Then, let

$$\widehat{\mathbf{\Psi}}(x,y) = \widehat{\mathbf{\Phi}}(x,y)\,\widehat{\mathbf{\Phi}}^T(x,y) \tag{4.9}$$

denote an estimator of the matrix $\mathbf{\Psi}(x,y)$ in (3.26) for all genotypes (x,y). Given these definitions of estimators, it follows that

$$\widehat{\mathbf{\Psi}}_G = \sum_{(x,y)} \widehat{p}(x.y)\,\widehat{\mathbf{\Psi}}(x,y) \tag{4.10}$$

is an estimator of the genetic covariance matrix defined in (3.27). It should also be noted that an investigator would be free to estimate each component of the matrix $\widehat{\mathbf{\Psi}}_G$ separately.

It is also possible to estimate H_G, the measure of heritability defined in section 2. From (2.13), it follows that

$$\widehat{var}_P[W] = \widehat{var}_G[W] + \widehat{var}_E[W] \tag{4.11}$$

is an estimator of the phenotypic variance. Therefore,

$$\widehat{H}_G = \frac{\widehat{var}_G[W]}{\widehat{var}_P[W]} \tag{4.12}$$

is an estimator of H_G, a measure of heritability.

At any step in the development of software to implement the ideas under discussion, one could proceed in a number of directions. Suppose, for example, an investigator was not inclined to estimate the matrix $\widehat{\mathbf{\Psi}}_G$ in (4.10). An alternative approach would be that of considering a remainder estimate $\widehat{R}_G$ which is defined by the equation

$$\widehat{var}_G[W] = \widehat{var}_G(A) + \widehat{var}_G(IAI) + \widehat{R}_G, \tag{4.13}$$

where $\widehat{var}_G(A)$ and $\widehat{var}_G(IAI)$ are estimates of the additive and intra-allelic-interactions variance components defined in (3.13) and (3.14). The remainder

term $\widehat{R}_G$ would be a direct measure of the departure of the population from a Hardy-Weinberg equilibrium when equation (3.23) is valid. Observe that $\widehat{R}_G$ is the sum of all elements in the matrix $\widehat{\mathbf{\Psi}}_G$ off the principal diagonal.

If an investigator were interested in investigating whether the off diagonal elements in this estimator of the covariance matrix would change significantly under the assumption that the population from which the sample was derived was in Hardy-Weinberg equilibrium, the following procedure could be executed. Let $\widehat{p}(x)$ be an estimator of the marginal frequency of maternal alleles $x \in \mathbb{A}$ in the sample, and let the marginal frequency $\widehat{p}(y)$ be defined similarly for paternal alleles $y \in \mathbb{A}$ in the sample. Then, the next step in a computer simulation experiment with a goal of recomputing the estimate of the matrix $\widehat{\mathbf{\Psi}}_G$, under the assumption that the population was in a Hardy-Weinberg equilibrium, would be that of computing the product

$$p^*_{HW}(x,y) = \widehat{p}(x)\,\widehat{p}(y) \tag{4.14}$$

for all genotypes $(x,y) \in \mathbb{G}$ in the sample. Given this trial set of genotypic frequencies, the calculation procedures outlined above could be used to compute an alternative estimate of the covariance matrix $\mathbf{\Psi}_G$, symbolized by $\mathbf{\Psi}_{GHW}$, under the assumption that the population was in a Hardy-Weinberg equilibrium so that one would expect that the remainder term $\widehat{R}_G$ would be zero.

The direct method of estimation described above has many advantages when compared with classical methods of estimating variance components , because the effects defined in section 3 may also be estimated directly from the data. By way of illustrative example, the direct estimator of the conditional expectation $\mu(y)$ is

$$\widehat{\mu}(x) = \textstyle\sum_y \widehat{p}(y \mid x)\,\widehat{\mu}(x,y), \tag{4.15}$$

where

$$\widehat{p}(y \mid x) = \frac{\widehat{p}(x,y)}{\widehat{p}(x)} \tag{4.16}$$

for $\widehat{p}(x) \neq 0$. Therefore, the direct estimator of the additive effect defined in (3.9) is

$$\widehat{\alpha}(x) = \widehat{\mu}(x) - \widehat{\mu} \tag{4.17}$$

for all alleles $x \in \mathbb{A}$. A formula for the direct estimator of the effect $\alpha(y)$ is analogous to that of $\widehat{\alpha}(x)$. Given the estimators $\widehat{\alpha}(x)$ and $\widehat{\alpha}(y)$, a direct estimator of the measure of interaction between alleles x and y defined in (3.10) is

$$\widehat{\alpha}(x,y) = \widehat{\mu}(x,y) - \widehat{\mu} - \widehat{\alpha}(x) - \widehat{\alpha}(y) \tag{4.18}$$

for all genotypes $(x,y) \in \mathbb{G}$.

As can be seen from (3.13) and (3.14), the squares of the estimators of the effects defined above would be terms in the estimators of the additive and intra-allelic interaction components of variance so that if attention was focused only the estimates of these variance components, an investigator may miss detecting the largest of the squared effects which would be of interest in their own right. It is recommended, therefore, that the squares in the sets

$$\left\{\widehat{\alpha}^2(x), \widehat{\alpha}^2(y) \mid x \in \mathbb{A}, y \in \mathbb{A}\right\} \tag{4.19}$$

be calculated and inspected to get an idea as to which allele produces the largest additive effect. Similarly, it is recommended that the set of squares of measures of interaction

$$\left\{\widehat{\alpha}^2(x,y) \mid (x,y) \in \mathbb{G}\right\} \tag{4.20}$$

also be calculated and inspected to get an idea of which genotype has the largest measure of interaction of alleles.

It should also be mentioned that it would be desirable to work out the statistical properties of the estimators defined in this section. Included in these properties of these estimators would be consistent as sample size becomes large and whether an estimator is unbiased. There is also a need for statistical tests to assess whether a particular estimate of a parameter was significantly different from zero. It is recognized that the working out of these statistical properties would be important, but a full response to such statistical issues is beyond the scope of this paper. In this connection, it is interesting to note that computer intensive methods are now being used extensively in judging the statistical significance as to whether some region of a genome is implicated in some quantitative trait. For example, an interested reader may wish consult the papers Raj et al. (2012) [19] and Rossin et al. (2011) [20] in which permutation tests have been used in assessing statistical significance of hypothesized protein and other networks. It should also be mentioned that such computer intensive methods as jack-knifing and boot-strapping could also be used to assess the statistical significance of an estimate of an effect or variance component.

V. The Case of Two Autosomal Loci

Let $\mathbb{A}_1$ and $\mathbb{A}_2$ denote the set of alleles at locus 1 and 2, respectively. It will be assumed that each of these sets contains at least two alleles. In a diploid species with two sexes, such as humans, at every locus there is an allele contributed by the female parent and another allele contributed by the male parent. For the case of two autosomal loci, a genotype will be represented by the symbol (x_1, y_1, x_2, y_2), where (x_1, x_2) denotes the maternal alleles at the two loci and (y_1, y_2) are the corresponding paternal alleles. The set $\mathbb{G}$ all genotypes with respect to the two loci under consideration is the product set

$$\mathbb{G} = \mathbb{A}_1 \times \mathbb{A}_1 \times \mathbb{A}_2 \times \mathbb{A}_2. \tag{5.1}$$

To lighten the notation in what follows, let the vector $\boldsymbol{z} = (x_1, y_1, x_2, y_2)$ denote a genotype $\boldsymbol{z} \in \mathbb{G}$, and let $p(\boldsymbol{z})$ denote the frequency of genotype $\boldsymbol{z} \in \mathbb{G}$ in the population. For some quantitative trait or character under consideration, let the W denote a random variable describing the phenotypic variation with respect to some quantitative measurement among the individuals in a population. Then, given some genotype $\boldsymbol{z} \in \mathbb{G}$, let the conditional expectation

$$\mu(\boldsymbol{z}) = E[W \mid \boldsymbol{z}] \tag{5.2}$$

denote the genetic value for this genotype. Just as in the case of one locus, this conditional expectation will play a basic role in defining measurements of the effects of each allele as well as the interactions among the at the two loci under consideration.

In general, one would not expect that the population under consideration would be in linkage equilibrium; consequently, it will be necessary to define a number of marginal and conditional distributions, that will be derived using the set

$$\mathfrak{D}_G = \{p(\boldsymbol{z}) \mid \boldsymbol{z} \in \mathbb{G}\} \tag{5.3}$$

of genotypic frequencies, which from now on will be called the genotypic distribution. For example, for allele x_1 suppose we wish to derive a formula for the conditional expectation of $\mu(x_1, y_1, x_2, y_2)$, given x_1 with respect to the genotypic distribution. A first step in this derivation, would be to calculate the marginal distribution

$$p(x_1) = \sum_{(y_1, x_2, y_2)} p(x_1, y_1, x_2, y_2) \tag{5.4}$$

for all $x_1 \in \mathbb{A}_1$. By definition, the conditional distribution of $\mu(x_1, y_1, x_2, y_2)$, given x_1 is

[20] Rossin E. J., Lage K., Raychaudhuri S., Xavier R. J., Tatar D, et al. (2011) Proteins Encoded in Genomic Regions Associated with Immune-Mediated Disease Physically Interact and Suggest Underlying Biology. PLoS Genet 7(1): e1001273. doi:10.1371/journal.pgen.1001273

$$p(y_1, x_2, y_2 \mid x_1) = \frac{p(x_1, y_1, x_2, y_2)}{p(x_1)} \tag{5.5}$$

for $p(x_1) \neq 0$. Let $\mu(x_1)$ denote the conditional expectation of $\mu(x_1, y_1, x_2, y_2)$, given x_1. Then, by definition

$$\mu(x_1) = \sum_{(y_1, x_2, y_2)} p(y_1, x_2, y_2 \mid x_1)\, \mu(x_1, y_1, x_2, y_2) \tag{5.6}$$

for all $x_1 \in \mathbb{A}_1$. The unconditional expectation of the $\mu(\boldsymbol{z})$ with respect to the genotypic distribution $\mathfrak{D}_G$, as expressed in a more succinct notation, is

$$\mu = \sum_{\boldsymbol{z} \in \mathbb{G}} p(\boldsymbol{z})\, \mu(\boldsymbol{z}). \tag{5.7}$$

Therefore, in analogy with the case of one allele, the additive effect of allele x_1 in the population will be defined as

$$\alpha(x_1) = \mu(x_1) - \mu \tag{5.8}$$

for all $x_1 \in \mathbb{A}_1$.

An analogous effect could be defined for each of the alleles y_1, x_2, and y_2, by applying the methods described for defining $\alpha(x_1)$. But, as will be demonstrated, for the case of two autosomal loci there are many more interactions terms that need to be defined. For example, for the case of a diploid species, there are four positions to be considered when classifying and defining effects and interactions among alleles. Consider, for example, the set of four alleles in each genotype $\boldsymbol{z} = (x_1, y_1, x_2, y_2) \in \mathbb{G}$, and let $\mathfrak{S} = \{1, 2, 3, 4\}$ denote the set of four positions that need to be considered with respect to two loci with two alleles at each locus that were contributed by the maternal and paternal parent respectively. To provide a framework for describing various types of interactions among the alleles at the two loci under consideration, it will be helpful to consider the class of all subsets of the four positions. Let $\mathfrak{T}$ denote the class of all subsets of $\mathfrak{S}$. Included in the class $\mathfrak{T}$ is the empty set φ as well as subsets containing 1, 2,3 and 4 elements of the set $\mathfrak{S}$. As is well known from combinatorial analysis, the total number of sets in $\mathfrak{T}$ is $2^4 = 16$, and, as is also well known from combinatorics, that the equation

$$\binom{4}{0} + \binom{4}{1} + \binom{4}{2} + \binom{4}{3} + \binom{4}{4} = 2^4 = 16 \tag{5.9}$$

is valid. For $\nu = 0, 1, 2, 3, 4$, let $\mathfrak{T}_\nu$ denote the subclass of sets in $\mathfrak{T}$ that contain ν elements. Then, as can be seen form equation (5.9), each of the subclasses $\mathfrak{T}_0$ and $\mathfrak{T}_4$ contain one set; namely φ and $\mathfrak{S}$, respectively. Similarly, each of the subclasses $\mathfrak{T}_1$ and $\mathfrak{T}_3$ contain 4 sets, and the subclass $\mathfrak{T}_2$ contains 6 sets. Recall that

$$\binom{4}{2} = 6. \tag{5.10}$$

To describe a framework in which to quantify the ideas of intra-allelic interactions and epistatic interactions among alleles at different loci, it will be helpful to enumerate the sets in the subclasses $\mathfrak{T}_1$, $\mathfrak{T}_2$ and $\mathfrak{T}_3$ in terms of elements of the set $\mathfrak{S}$. For example,

$$\mathfrak{T}_1 = (\{1\}, \{2\}, \{3\}, \{4\}) \tag{5.11}$$

is the class of singletons, which are subsets that contain only one element of $\mathfrak{S}$. It is this subclass of sets that was used to define the additive effects mentioned above. The subclass $\mathfrak{T}_2$ of sets has the explicit form

$$\mathfrak{T}_2 = (\{1, 2\}, \{1, 3\}, \{1, 4\}, \{2, 3\}, \{2, 4\}, \{3, 4\}). \tag{5.12}$$

At this point recall that positions 1 and 2 in the set $\mathfrak{S}$ are those for the two alleles at locus 1, and positions 3 and 4 in this set are those for the two alleles at locus 2. Therefore, the two sets of positions in subclass

$$\mathfrak{T}_{2IAI} = (\{1,2\},\{3,4\}) \tag{5.13}$$

will be used to define effects that measure intra-allelic interactions at the two loci under consideration. On the other hand, the pairs of positions in the subclass

$$\mathfrak{T}_{2EPI} = (\{1,3\},\{1,4\},\{2,3\},\{2,4\}) \tag{5.14}$$

represent positions from different loci. Consequently, sets in this class will form a basis for defining effects that measure epistatic interactions among the alleles at the two loci under consideration. The subsets in the subclass $\mathfrak{T}_3$ are as follows

$$\mathfrak{T}_3 = (\{1,2,3\},\{1,2,4\},\{1,3,4\},\{2,3,4\}). \tag{5.15}$$

The sets in this class form a basis for defining effects that measure the effect that an allele at one locus may affect or modify intra-allelic interactions at another locus. For example, the two sets in the subclass

$$\mathfrak{T}_{31EPI} = (\{1,2,3\},\{1,2,4\}) \tag{5.16}$$

would form a basis for defining an effect measuring intra-allelic interactions at locus 1 that may affect the expression of alleles in positions 3 and 4 at locus 2. Similarly, the sets subclass

$$\mathfrak{T}_{32EPI} = (\{1,3,4\},\{2,3,4\}) \tag{5.17}$$

would form a basis for defining an effect measuring intra-allelic interactions at locus 2 that may be affected by alleles at positions 1 and 2 at locus 1.

For cases in which many alleles can be recognized at each locus, it would be necessary to develop a nomenclature to describe many types of interactions among the alleles at the two autosomal loci under consideration as will be illustrated below. In this connection, an interested reader may wish to consult the pioneering work of Cockerham (1954) [5] that describes a nomenclature for various epistatic effects and components of the genetic variance. For example, effects and variance components corresponding to the sets in the class $\mathfrak{T}_{2IAI}$ would be labeled dominant for either the effects or variance components and would be denoted by the symbol D. Whereas those in the class $\mathfrak{T}_{2EPI}$ would be labeled additive by additive effects or variance components and denoted by the symbol AA. One could proceed in this way to develop a nomenclature of the 15 effects and variance components under consideration. But, this type of nomenclature will, however not be used in this paper and epistasis will be described in terms of sets and effects as well as variance components.

The first step in defining these effects is to derive a formula for the conditional expectation of a genetic value $\mu(\boldsymbol{z})$, given every set A of positions such that

$$A \in E = \bigcup_{\nu=1}^{3} \mathfrak{T}_{\nu}. \tag{5.18}$$

To define these effects, it will be helpful to introduce a succinct notation. For every set A of positions, let A^c denote the complement of this set with respect to the set $\mathfrak{S}$, and let $z(A)$ and $z(A^c)$ denote subsets of alleles in z corresponding to the positions in the sets A and A^c, respectively. In what follows, the symbol $z(A), z(A^c)$ will stand for the union of the positions in the two sets. Given this notation, the marginal distribution $p(z(A))$ is defined by

$$p(z(A)) = \textstyle\sum_{z(A^c)} p(z(A), z(A^c)) \tag{5.19}$$

for every $z(A) \in \mathbb{G}(A)$,where $\mathbb{G}(A)$ is a subset of $\mathbb{G}$ containing only those alleles corresponding to the positions in the set A. Thus, in this succinct notation,

$$p(z(A^c) \mid z(A)) = \frac{p(z(A), z(A^c))}{p(z(A))} \tag{5.20}$$

[5] Cockerham, C. C. (1954) An extension of the concept of partitioning the hereditary variance for analysis of covariance among relatives when epistasis is present. Genetics 39:859-882.

is the conditional distribution of $z(A^c)$, given $z(A)$ for $p(z(A)) \neq 0$. Let $\mu(z(A))$ denote the conditional expectation of $\mu(z)$, given $z(A)$. Then,

$$\mu(z(A)) = \sum_{z(A^c)} p(z(A^c) \mid z(A)) \mu(z(A), z(A^c)) \tag{5.21}$$

for every $A \in E$.

Given formula (5.21), one may proceed systematically through each of the sets in the union E in (5.18) to calculate $\mu(z(A))$ for every $A \in E$. For example, suppose $A = \{1\}$. Then, $\mu(z(A)) = \mu(x_1)$ for all $x_1 \in \mathbb{A}_1$. By continuing in this manner, all the conditional pairs of expectations, $(\mu(x_\nu), \mu(y_\upsilon))$, for $\nu = 1, 2$ could be computed, and formula (5.8) could be used to compute the four effects: $\alpha(x_\nu), \alpha(y_\nu)$ for $\nu = 1, 2$.

Similarly, for every set $A \in \mathfrak{T}_2$, $\mu(z(A))$ would need to be calculated. Suppose, for example, $A = \{1, 2\}$. Then, $\mu(x_1, y_1)$ would need to be calculated for every $(x_1, y_1) \in \mathbb{A}_1 \times \mathbb{A}_1$. Then, as in the case of one locus, the intra-allelic effect $\alpha(x_1, y_1)$ would be defined by

$$\alpha(x_1, y_1) = \mu(x_1, y_1) - \mu - \alpha(x_1) - \alpha(y_1). \tag{5.22}$$

By continuing in this way, an effect $\alpha(z(A))$ could be defined for every subset $A \in \mathfrak{T}_3$. To illustrate how each of these four effects could be defined, consider the case $A = \{1, 2, 3\}$. In this case, $\mu(z(A)) = \mu(x_1, y_1, x_2)$ for all $(x_1, y_1, x_2) \in \mathbb{A}_1 \times \mathbb{A}_1 \times \mathbb{A}_2$. Then, by definition, the effect $\alpha(x_1, y_1, x_2)$ is

$$\begin{aligned} \alpha(x_1, y_1, x_2) \quad = \quad & \mu(x_1, y_1, x_2) - \mu - \alpha(x_1) - \alpha(y_1) - \alpha(x_2) \\ & -\alpha(x_1, y_1) - \alpha(x_1, x_2) - \alpha(y_1, x_2). \end{aligned} \tag{5.23}$$

Altogether, for the subclass $\mathfrak{T}_3$, four effects would need to be computed, using the procedure illustrated in (5.23). Note that all the effects on the right in this equation, were defined for each subset of the set of symbols $\{x_1, y_1, x_2\}$. This procedure may also be used to set down formulas for each of the three remaining subsets in the subclass $\mathfrak{T}_3$. Furthermore, in formulations in which more than two loci were under consideration, the procedure (5.23) used to define the effects for the case of two loci could be extended to defining effects for some number of loci $n \geq 3$. The last step in defining effects for the two loci case is to define the effect $\alpha(z(\mathfrak{S})) = \alpha(x_1, y_1, x_2, y_2)$ for all genotypes $z \in \mathbb{G}$. In this connection let $\alpha(z(\mathfrak{S})) = \alpha(z)$ be such that the equation

$$\mu(z) = \mu + \sum_{A \in \mathfrak{T}_1} \alpha(z(A)) + \sum_{A \in \mathfrak{T}_2} \alpha(z(A)) + \sum_{A \in \mathfrak{T}_3} \alpha(z(A)) + \alpha(z) \tag{5.24}$$

holds for all genotypes $z \in \mathbb{G}$.

Having defined the set of 15 effects for the case of two autosomal loci, the next step is that of defining components of the genetic variance. For example, the additive genetic variance is defined by

$$var_A[W] = \sum_{A \in \mathfrak{T}_1} E_{\mathfrak{D}_G}[\alpha^2(z(A))], \tag{5.25}$$

where the expectation is taken with respect to the genotypic distribution $\mathfrak{D}_G$. The intra-allelic interaction component of the genetic variance is defined by

$$var_{IAI}[W] = \sum_{A \in \mathfrak{T}_{2IAI}} E_{\mathfrak{D}_G}[\alpha^2(z(A))], \tag{5.26}$$

and epistatic component of genetic variance with respect to two loci is defined by

$$var_{EPI}[W] = \sum_{A \in \mathfrak{T}_{2EPI}} E_{\mathfrak{D}_G}[\alpha^2(z(A))] \tag{5.27}$$

For the case of three alleles, the equation

$$var_{IAI1}[W] = \sum_{A \in \mathfrak{T}_{31EPI}} E_{\mathfrak{D}_G}[\alpha^2(z(A))],$$

is the component of variance for intra-allelic interaction at the first locus that may be modified by an alleles at the second locus. Similarly, the component of the genetic variance for intra-allelic interaction at the second locus that may be modified by an allele at the first locus is

$$var_{IAI2}[W] = \sum_{A\in\mathfrak{T}_{32EPI}} E_{\mathfrak{D}_G}[\alpha^2(z(A))]. \tag{5.28}$$

Finally, the component of the genetic epistatic variance as measured by effects $\alpha(z)$ is defined by

$$var_{EPI4}[W] = \sum_{A\in\mathfrak{T}_4} E_{\mathfrak{D}_G}[\alpha^2(z(A))]. \tag{5.29}$$

It should be noted that the set of components of the genetic variance was defined arbitrarily, but a user of the ideas presented in this section may wish to adapt another nomenclature for the set of 15 effects and components of the total genetic variance.

An experimenter could test whether a sample of individuals whose genotypes had been determined with respect to two autosomal loci was in linkage equilibrium, but in any case it would be of interest to compute the genetic covariance matrix for the case under consideration. Let A denote any set of positions in the union

$$A \in \mathfrak{A} = \bigcup_{v=1}^{4} \mathfrak{F}_v \tag{5.30}$$

and let

$$\mathbf{\Phi}(\boldsymbol{z}) = (\alpha(z(A)) \mid A \in \mathfrak{A}) \tag{5.31}$$

denote a 15×1 vector of classes of effects. Observe that within each class of effects corresponding to a set A there would be a collection of effects corresponding the to the number of alleles at each locus. A useful ordering of the effects in this vector would be to let the subset of singletons be the first four elements of the vector, the 6 sets of pairs of positions would be the next 6 element in the vector, the next four elements of the vector would be the four effects corresponding to the sets of triples of positions and lastly the effect for the singleton $\mathfrak{S}$ would be the last 15-th effect in the column vector. As was tacitly used in the definitions of the components of the genetic variance listed above, each effect has the unconditional expectation

$$E_{\mathfrak{D}_G}[\alpha(A(z))] = 0, \tag{5.32}$$

for all $A \in \mathfrak{A}$. Let,

$$\mathbf{\Psi}(z) = \mathbf{\Phi}(z)\mathbf{\Phi}^T(z)$$

denote a 15×15 matrix of products of effects for the genotypes $z \in \mathbb{G}$. Then, by definition, the covariance matrix of the vector $\mathbf{\Phi}(z)$ of effects is

$$\mathbf{\Psi}_G = \sum_{z\in\mathbb{G}} p(z)\mathbf{\Phi}(z)\mathbf{\Phi}^T(z) = E_{\mathfrak{D}_G}[\mathbf{\Psi}(z)]. \tag{5.33}$$

As part of an analysis of data, at this point in the calculations, a data annalist may wish to compute the eigen values and vectors of the symmetric matrix $\mathbf{\Psi}_G$. It would also be of interest to inspect the off-diagonal components of the matrix $\mathbf{\Psi}_G$ to provide an assessment of the impact of effects on the components of the genetic variance when the population is not in linkage equilibrium at the two loci under consideration.

On the other hand, an investigator may not wish to compute and analyze the matrix $\mathbf{\Psi}_G$ in (5.33) and would be content with an estimate of the fraction

$$\frac{var_{EPI}[W] + var_{IAI2}[W] + var_{EPI4}[W]}{var_G[W]}, \tag{5.34}$$

where $var_G[W]$ is the total genetic variance. An estimate of this ratio would be of interest, because it would provide an investigator with some idea of the

significance of the contribution of epistatic effects to the total genetic variance. At the same time, it should be recognized that the estimate in (5.34) under the assumption that the population was not in linkage equilibrium and could be biased by negative covariance terms.

An investigator may, therefore, also wish to carry out a computer simulation experiment under the assumption that the population or sample was in linkage equilibrium. The first step in setting up such a computer experiment would be that of computing the marginal allele probabilities. Let $p_1(x_1)$ and $p_1(y_1)$ denote the marginal probabilities, respectively, for the maternal and paternal alleles at locus 1, and define the marginal $p_2(x_2)$ and $p_2(y_2)$ for locus 2 similarly. Then, the simulated population would be in linkage equilibrium if the genotypic probabilities $p(z)$ satisfied the equation $p(z) = p_1(x_1)\, p_1(y_1)\, p_2(x_2)\, p_2(y_2)$ for all genotypes $z = (x_1, y_1, x_2, y_2) \in \mathbb{G}$. Given these assigned genotypic probabilities, an investigator could carry out a computer simulation experiment under the assumption that the sample or population was in linkage equilibrium.

Just as in the one locus case considered in section 4, it is recommended that an investigator inspect the squares of all effects defined above for the case of two autosomal loci. For example, the set of squares of additive effects is defined by

$$\mathfrak{E}_1 = \left\{\alpha^2(z(A)) \mid A \in \mathfrak{T}_1\right\}. \tag{5.35}$$

It will be tacitly be assumed that the elements in the set $\mathfrak{E}_1$ are estimates of effects so as to simplify the notation. For the case each locus has two alleles, the set $\mathfrak{E}_1$ would contain a small number of elements so that an investigator could easily find the largest one. Similarly, the set of squared effects that are measures of intra-allelic interactions is defined by

$$\mathfrak{E}_{2IAI} = \left\{\alpha^2(z(A)) \mid A \in \mathfrak{T}_{2IAI}\right\}. \tag{5.36}$$

Like set $\mathfrak{E}_1$ for the case of two alleles at each of the two loci under consideration, the set $\mathfrak{E}_{2IAI}$ would contain a small number of estimated squared effects so that an investigator could easily find the largest one. By continuing this way, set of estimated squared effects corresponding to each of the sub-classes of effects defined above for various types of epistatsis could also be defined but the enumeration of these sets will be left as an exercise for an interested reader.

In an interesting paper, Hemani et al. (2013) [8] an evolutionary perspective on epistasis and the missing heritability was the focus of attention. These authors assert that results of genome wide association studies may improved if epistatic effects may be searched for explicitly. It is suggested that the epistatic effects defined in this section may be also be useful in genome wide association studies.

VI. An Overview of the Case of Eleven Autosomal Loci

As mentioned in the introduction, there is an interesting and important case in human genetics pertaining to Alzheimer's disease (AD) in which there is a developing consensus that eleven autosomal regions, loci, of the human genome have been implicated in this disease. It is suggested that an interested reader may wish to consult the paper by Raj et al. (2012) [19] and the literature cited therein for more details regarding these genomic regions. In studies of patients with AD, quantitative measurements are often made on each patient so that AD may be viewed as a quantitative trait in humans. It is, therefore, of interest to provide an overview of an extension of the structure for the case of two autosomal loci developed in section 5 to the case of 11 autosomal loci.

For a diploid species such as humans, two alleles occupy each locus so that for the case of 11 loci, there are $11 \times 2 = 22$ positions to consider in the set

$$\mathfrak{S} = (s \mid s = 1, 2, \cdots, 22) \tag{6.1}$$

of positions. Therefore, in this case the class $\mathfrak{T}$ of all subsets of $\mathfrak{S}$ contains

$$2^{22} = 4,194,304 \tag{6.2}$$

sets. Included in $\mathfrak{T}$ is the empty set φ so, just as for the case of two loci, no effect will be associated with φ. It follows, therefore, that in theory, $2^{22} - 1 = 4,194,303$ effects could be defined for the case of 11 autosomal loci, but it is unlikely that any investigator would attempt to estimate such a large number of effects.

Notes

When dealing 11 or more autosomal loci, it is also important to remember that for the case of many loci, one should keep in mind the caveat that the number of possible genotypes under consideration may be quite large and exceed the sample size that is available to an investigator or investigators. For the case of 11 autosomal loci and two alleles per locus, each vector in the pair $(\boldsymbol{x}, \boldsymbol{y})$ denoting a genotype would contain 11 alleles contributed by the maternal and paternal, respectively. Thus, if it were possible to determine parental source of each allele, one could in principle identify four genotypes per locus. Therefore, if 11 autosomal loci were under consideration, the number of genotypes that could be identified would be

$$4^{11} = 4,194,304. \tag{6.3}$$

Observe that this is the same number as that in (6.2), and, moreover, it in all likelihood exceeds the number of individuals in any sample of individual whose genomes have been sequenced that are presently available to investigators.

Consider, for example, the case of 11 autosomal loci with two alleles at each locus and suppose that an investigator identifies three genotypes per locus; namely two homozygotes and one heterozygote at each locus. In such circumstances, an investigator may not be able to determine whether any alleles was contributed by the maternal or paternal parent. Under this assumption that only three genotypes can be identified per locus, it follows that the total number of "genotypes" that could be identified with respect to 11 autosomal loci would be

$$3^{11} = 177,147. \tag{6.4}$$

A number of this magnitude would in all likelihood exceed the sample size available to present day investigators, particularly if it is required that all individuals in the sample have had their genomes sequenced. If a sample size is considerably smaller than the number in (6.4), then it is recommended that an investigator confine attention to some sub-set $\mathfrak{S}_1$of loci and individuals in a sample such that for each identifiable genotype $(\boldsymbol{x},\boldsymbol{y})$, the number of individuals, $n(\boldsymbol{x},\boldsymbol{y}) \geq 1$, with this genotype is sufficiently large so that one may make reliable and statistically significant genetic inferences based on the available data.

For the case of 11 autosomal loci, a sample available to an investigator may not be sufficiently large to accommodate the set of possible genotypes, because the number of individuals of all genotypes may not be sufficiently large to draw reliable statistical inferences. However, when attention is focused on a sub-set of loci, the number of individuals for each genotype with respect to this sub-set of loci is sufficiently large to draw reliable statistical inferences. By way of an illustrative and hypothetical example, suppose that an investigator was able to find a sufficient sample size for each genotype with respect to six autosomal loci with three distinguishable genotypes at each locus. Let $\mathbb{G}_S$ denote the set of genotypes in the sample and let $n(\boldsymbol{x}, \boldsymbol{y})$ denote the number of individuals in the sample of genotype $(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{G}_S$. For the case of six autosomal loci, the total number of effects that may be defined is

$$2^{12} - 1 = 4,095. \tag{6.5}$$

It is doubtful that any investigation would have the persistence or interests to estimate $4,095$ effects, but it may be of interest to estimate only first, second

and third order effects. It is easy to see that the number of first order or additive effects would be

$$\binom{12}{1} = 12, \tag{6.6}$$

and as in (5.11) let $\mathfrak{T}_1$ denote the class of subsets of the set $\mathfrak{S} = (1, 2, \cdots, 12)$ of positions containing one position. It is straight forward to enumerate the sets in the class $\mathfrak{T}_1$. For the case of 12 positions, the number of sub-set of $\mathfrak{S}$ containing two positions is

$$\binom{12}{2} = 66. \tag{6.7}$$

Let $\mathfrak{T}_2$ denote the class of sub-sets of $\mathfrak{S}$ containing two positions. Similarly, the number of subsets of $\mathfrak{S}$ containing three positions is

$$\binom{12}{3} = 220. \tag{6.8}$$

Observe that if an investor chose to follow the procedure just outlined, the total number of effects that would need to be defined would be

$$12 + 66 + 220 = 298. \tag{6.9}$$

Let $\mathfrak{T}_3$ denote the class of sub-sets of $\mathfrak{S}$ containing three positions.

It is interesting to note that the enumeration of the sets in the classes $\mathfrak{T}_2$ and $\mathfrak{T}_3$ may be accomplished by using a type of recursive procedure. To describe this recursive procedure, it is helpful if the notation is extended to include the number of loci and positions under consideration. For example, let $\mathfrak{T}_2^{(\nu)}$ denote a class of sub-sets of two positions taken from the sets of positions $\mathfrak{S}_\nu$ for $\nu = 4, 6, \cdots, 12$ sets of positions corresponding to $l = 2, 3, \cdots, 6$ loci. Then, it follows that the containment relations

$$\mathfrak{T}_2^{(4)} \subset \mathfrak{T}_2^{(6)} \subset \mathfrak{T}_2^{(8)} \subset \mathfrak{T}_2^{(10)} \subset \mathfrak{T}_2^{(12)} \tag{6.10}$$

hold. Thus, if an investigator has enumerated the sub-sets in the class $\mathfrak{T}_2^{(4)}$ for the case of two loci, see section 5, then to extend this enumeration to case of 3 loci and 6 positions, one could add positions 5 and 6 to the set $\mathfrak{S}_4$ to obtain the set $\mathfrak{S}_6$ of position for the case of 3 loci. The next step in this recursive process would be that to adding to $\mathfrak{T}_2^{(4)}$ those sets with two positions that include position 5 and 6 to obtain all the sub-sets with two positions from the set $\mathfrak{S}_6$. By continuing in this recursive manner, the set of two positions in the class $\mathfrak{T}_2^{(12)}$ could be enumerated. It also of interest to note that the containments relations

$$\mathfrak{T}_3^{(4)} \subset \mathfrak{T}_3^{(6)} \subset \mathfrak{T}_3^{(8)} \subset \mathfrak{T}_3^{(10)} \subset \mathfrak{T}_3^{(12)} \tag{6.11}$$

for classes of sub-sets containing sets with three positions are also valid. Therefore, the class of sets $\mathfrak{T}_3^{(12)}$ could also be enumerated by using a recursive procedure. It is also highly plausible that a clever computer programmer could write code to accomplish the enumeration of the classes of sets $\mathfrak{T}_2^{(12)}$ and $\mathfrak{T}_3^{(12)}$.

Given the enumerated classes of sub-sets $\mathfrak{T}_1, \mathfrak{T}_2$ and $\mathfrak{T}_3$, the next step in providing an overview of the case of six autosomal loci is that of defining an effects for each set in the three classes of sub-sets. Briefly, the procedures used in defining and setting up algorithms to compute them are given implicitly in equations (5.19) , (5.20) and (5.21) . Let

$$\mathfrak{C}_1 = \{\alpha(z(A)) \mid A \in \mathfrak{T}_1\} \tag{6.12}$$

denote the set of first order effects. Similarly, let

$$\mathfrak{C}_2 = \{\alpha(z(A)) \mid A \in \mathfrak{T}_2\} \tag{6.13}$$

and

$$\mathfrak{C}_3 = \{\alpha(z(A)) \mid A \in \mathfrak{T}_2\} \tag{6.14}$$

denote, respectively, the class of second and third order effects. It should be noted that the formulas for computing the first, second and third order effects are outlined in formulas (5.22) and (5.23) for each combination of alleles.

Just as suggested for the case of two autosomal loci in section 5, it would be of interest to find the largest of the squares of each effects to get some idea as to which effect contributes the most to a variance component under consideration. An explicit form of the squares of first order effects is

$$\mathfrak{D}_1 = \left\{\alpha^2(\nu) \mid \nu \in \mathfrak{S}\right\}, \tag{6.15}$$

where $\mathfrak{S} = (s \mid s = 1, 2, \cdots, 12)$. For the sake of simplicity, suppose there are only two alleles at each the the six loci under consideration. Under this assumption, each of the 12 positions may be occupied by either of the two alleles at each locus. Therefore, the number of squared values in the set $\mathfrak{D}_2$ is 24. Let $\mathfrak{D}_2$ and $\mathfrak{D}_3$ denote, respectively, the set of squared effects from the sets $\mathfrak{C}_2$ and $\mathfrak{C}_3$. Suffice it to say that for the case of two alleles at each of the six loci, the number of squared effects in each of the sets $\mathfrak{D}_\nu$for $\nu = 1, 2, 3$ could be determined, but this exercise will be left to an interested reader.

If an investigator does not have a sufficiently large sample to work with for the case of 6 autosomal loci, then a reduced version of the ideas just outlined could be used to study a smaller number loci such that the number the sample size for each genotype would be sufficiently large to draw reliable statistical inferences. Given the ideas just outlined, a study of cases for 2, 3 or 4 loci may be feasible if there is insufficiently data to study the case of 5 or 6 autosomal loci. It should be noted that for the case that only three genotypes per locus may be identified, the number of effects that an investigator could estimate would significantly smaller than for the case that for in which four genotypes may be identified per locus. It is beyond the scope of this paper to consider the case of only three identifiable genotypes per locus, but the details for this case will be worked out in subsequent papers for one or more autosomal loci.

When one considers of the union of the sets $\mathfrak{D}_1, \mathfrak{D}_2$ and $\mathfrak{D}_3$, it is easy to see that many tests of statistical significance may need to be made if an investigator wishes to assess the statistical significance some chosen number of squared effects. It is beyond the scope of this paper to deal with the problem of making many statistical tests and computing measures of statistical significance, but it is suggested that a reader may wish to consult the literature on this subject. Included among the papers that would be interest to consult are Benjamini et al. (1995) [1], (2001) [2] and (2005) [3].

A version of equation (5.24) may also set down for the case of 6 autosomal loci under consideration and has the form

$$\mu(z) = \mu + \sum_{A\in\mathfrak{T}_1} \alpha(z(A)) + \sum_{A\in\mathfrak{T}_2} \alpha(z(A)) + \sum_{A\in\mathfrak{T}_3} \alpha(z(A)) + \alpha_R(z) \tag{6.16}$$

for all genotypes $z \in \mathbb{G}_S$, where $\alpha_R(z)$ is a remainder effect. In principle, if all the effects on the right hand side of equation have been estimated for all genotypes $z \in \mathbb{G}_S$, then the effect $\alpha_R(z)$ could be estimated for all genotypes $z \in \mathbb{G}_S$. Given these estimates, one could then proceed to estimate the variance component corresponding to the effect $\alpha_R(z)$, by using the formula

$$var_R[W] = \sum_{z\in\mathbb{G}} p(z)\,\alpha_R^2(z) \ . \tag{6.17}$$

Let $\widehat{var}_G[W]$ denote an estimate of the genetic variance defined in equation (2.11) and let $\widehat{var}_R[W]$ denote an estimate of the variance component in (6.17). Then, the ratio

$$\frac{\widehat{var}_R[W]}{\widehat{var}_G[W]} \tag{6.18}$$

[3] Benjamini, Y. and Yekutieli, D. (2005) False discovery rate adjusted multiple confidence intervals for selected parameters. J. Amer. Statist. Assoc. 100:71-93.

may be used as an estimate of the fraction of the total genetic variance that is attributable to the remainder effects $\alpha_R(z)$ for all genotypes $z \in \mathbb{G}_S$.

When interpreting this estimate, an investigator should also be aware of the possibility that the sample of individuals that constitute the data used to estimate all effects and variance components may not be in linkage equilibrium with respect to the 6 autosomal loci under consideration. In this case, it may be worthwhile to compute a version of the genetic covariance $\mathbf{\Psi}_G$ defined in (3.27) for the case of 6 loci. It can be shown that in terms of this matrix, the estimate $\widehat{var}_G[W]$ of the genetic variance may be represented in the form

$$\widehat{var}_G[W] = \mathbf{1}^T \widehat{\mathbf{\Psi}}_G \mathbf{1} \ , \tag{6.19}$$

where $\mathbf{1}$ is column of $1s$, T denotes the transpose of vector or matrix and $\widehat{\mathbf{\Psi}}_G$ is an estimate of $\mathbf{\Psi}_G$. Given this matrix, all variance components associated with equation (6.16) would be on the principal diagonal the the matrix $\mathbf{\Psi}_G$. Therefore, the trace of the matrix, the sum of the elements on the principal diagonal of $\widehat{\mathbf{\Psi}}_G$, is the sum of the variance components corresponding to the effects in equation (6.16). Let $\widehat{trace}\left[\widehat{\mathbf{\Psi}}_G\right]$ denote an estimate of the sum of these variance components. Then, the ratio

$$\frac{\widehat{trace}\left[\widehat{\mathbf{\Psi}}_G\right]}{\widehat{var}_G[W]}, \tag{6.20}$$

is an estimate of the fraction of the total genetic variance that is attributable to the variance components defined in connection with equation (6.16). It would also be of interest to inspect the elements in the matrix $\widehat{\mathbf{\Psi}}_G$ off the principal diagonal to make an assessment as to the affects that non-zero covariance terms contribute to the estimate of the total genetic variance in (6.19).

This ratio may be interpreted as a measure of the genetic variance that is attributable to the effects defined in connection with the construction of equation (6.16) , taking into account these effects may be correlated for the case the sample of individuals is not in linkage equilibrium. If this ratio is equal to one, then the variance components defined in connection with equation (6.16) are sufficient to account for all the genetic variance in the quantitative trait under consideration. But, if this ratio is less than one, then these components of the genetic variance would not be sufficient to account for the total genetic variance. It is also appropriate to mention that the ratio in (6.20) may be computed without computing the matrix $\widehat{\mathbf{\Psi}}_G$, by computing each variance component corresponding to the effects in equation (6.16), using formulas analogous to (6.17).

It is recognized that an investigator who wished to apply the ideas on the estimation of effects and variance components set forth in this paper may also want to test some statistical hypotheses, but it is beyond the scope of this paper to suggest various types of statistical tests of significance in addition to those mentioned briefly above.

VII. Acknowledgements

A word of thanks is due Dr. Towfique Raj, Division of Genetics, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, USA, who called the author's attention to recent papers devoted to the estimation of heritability of various quantitative traits in humans, which have been cited in the paper. It should also be mentioned that a cooperative research effort involving the author and Dr. Raj's group is also in progress, with a goal of writing software to implement the ideas set forth in this paper, along with results not included in this paper, and applying them in a quantitative genetic analysis of data from samples of patients whose genomes have been sequenced.

Notes

References Références Referencias

[1] Benjamini, Y. and Hochberg, Y. (1995) Controlling false discovery rate: A practical and powerful approach to multiple testing. J. Roy. Soc. Ser. B 57:289-300.

[2] Benjamini, Y. and Yekutieli, D. (2001) The control of false discovery rate under dependency. Ann. Statist. 29:1165-1188.

[3] Benjamini, Y. and Yekutieli, D. (2005) False discovery rate adjusted multiple confidence intervals for selected parameters. J. Amer. Statist. Assoc. 100:71-93.

[4] Bulmer, M. G. (1980) The Mathematical Theory of Quantitative Genetics. Clarendon Press, Oxford.

[5] Cockerham, C. C. (1954) An extension of the concept of partitioning the hereditary variance for analysis of covariance among relatives when epistasis is present. Genetics 39:859-882.

[6] Falconer, D. and MacKay, T. F. C. (1996) Introduction to Quantitative Genetics. Longman, New York.

[7] Fisher, R. A. (1918) The correlation among relatives on the assumption of Mendelian inheritance. Trans. Royal Soc., Edinburgh 52: 399 -433.

[8] Hemani, G., Knott, S. and Haley, C. (2013) An Evolutionary Perspective on Epistasis and the Missing Heritability. PLoS Genet 9(2): e1003295. doi:10.1371/journal.pgen.1003295

[9] Kao, C-H and Zeng, Z-B. (2002) Modeling Epistasis of Quantitative Trait Loci Using Cockerham Model. Genetics 160: 1243 -1261.

[10] Kempthorne, O. (1954) The Correlations Between Relatives in a Random Mating Population. Proc. Royal Soc. London, B 143: 103-113.

[11] Kempthorne, O. (1957) An Introduction to Genetic Statistics. John Wiley and Sons, New York.

[12] Laird, N. M. and Lange, C. (2011) The Fundamentals of Modern Statistical Genetics. DOI 10.1007/978-1-4419-7338-2, Springer New York, Dordrecht, Heidelberg, London.

[13] Liu, B. H. (1998) Statistical Genomics – Linkage, Mapping and QTL Analysis. CRC Press, Boca Raton, London, New York and Washington, D. C.

[14] Lynch, M. and Walsh, B. (1998) Genetics and the Analysis of Quantitative Traits. Sinauer Associates, Inc. Sunderland, MA, 01375.

[15] Mao,Y., Nicole R., Ma, l.,Dvorkin, D. and Da. Y. (2006) Detection of SNP epistasis effects of quantitative traits using extended Kempthorne model. Physiol Genomics 28: 46 -52.

[16] Mode, C. J. and Robinson, H. F. (1959) Pleiotropism and the genetic variance and covariance. Biometrics 15: 518-537.

[17] Mode, C. J. (2014) Estimating statistical measures of pleiotropic and epistatic effects in the genomic era. International Journal of Statistics and Probability. Vol. 3, No. 2, 81 - 100.

[18] Price, A. L., Helgason, A. et al. (2011) Single Tissue and Cross-Tissue Heritability of Gene Expression via Identity-by-Descent in Related and Unrelated Individuals. PloS Genet 7(2) e1001317, doc:10.1371/journal.pgen.1001317.

[19] Raj, T., Shulman, J. M., Keenan, B. T. Lori B. Chibnik, L. B., Evans, D. A.., Bennett, D. A., Stranger, B. E. and De Jager, P. L. (2012) Alzheimer Disease Susceptibility Loci: Evidence for a Protein Network under Natural Selection DOI 10.1016/j.ajhg.2012.02.022. _2012 The American Society of Human Genetics.

[20] Rossin E. J., Lage K., Raychaudhuri S., Xavier R. J., Tatar D, et al. (2011) Proteins Encoded in Genomic Regions Associated with Immune-Mediated

Disease Physically Interact and Suggest Underlying Biology. PLoS Genet 7(1): e1001273. doi:10.1371/journal.pgen.1001273

[21] Stranger, B. E. , Eli A. , Stahl, E. A. and Raj. T. (2010) Progress and promise of genome-wide association studies for human complex trait genetics. 10.1534/genetics.110.120907

[22] Wu, Rongling, Ma, Chang-Xing, and Casella, G. (2010) Statistical Genetics of Quantitative Traits - Linkage, Maps and QTL. ISBN978-1-4419-1912-0, Springer Science.

[23] Yang, J., Benyamin, B. et al. (2010) Common SNPs explain a large proportion of heritability for human height. Nat Genet 2010 July ; 42(7): 565-569, doi:10.1038/ng.608.

[24] Zaitlen, N, Kraft, P., et al. (2013) Using Extended Genealogy to Estimate Components of Heritability for 23 Quantitative and Dichotomous Traits. PLoS Genet 9(5): e1003520.doi:10.1371/journal.pgen.1003520.

Notes