Genetic Markers in Clinical Trials

B.S. Weir and P.J. Heagerty

Abstract The current availability of dense sets of marker SNPs for the human genome is having a large impact on genetic studies and offers new possibilities for clinical trials. This chapter offers a unified basis for the analysis of marker and response data, emphasizing the central importance of the correlation, or linkage disequilibrium, between SNP markers and the genes that affect response. It is convenient to phrase the development of association mapping in the language of quantitative genetics, using additive and non-additive components of variance. A novel feature of dense SNP data is that good estimates can be made of actual inbreeding and relatedness. These estimates are more relevant than values predicted from family pedigree, and are all that are available in the absence of family data.

The dimensionality of SNP marker datasets has required the development of new methods that are appropriate for a large number of statistical comparisons, and the development of computational methods that allow high-dimensional regression. These methods are reviewed here, as is the use of biological annotation for both viewing the relevance of empirical associations, and to structure analysis in order to focus on those markers with the highest expectation for association with the outcomes under study.

1 Introduction

This chapter explores the statistical issues surrounding the use of SNPs in clinical trials and genome-wide association studies, and it contains the material presented in a short course by the authors. It is based, in part, on their experience with two NHGRI-funded consortia: GENEVA [3, 12], a collection of genome-wide

B.S. Weir (✉) • P.J. Heagerty
Department of Biostatistics, University of Washington, Box 357232, Seattle,
WA 98195-7232, USA
e-mail: bsweir@uw.edu

T.R. Fleming and B.S. Weir (eds.), *Proceedings of the Fourth Seattle Symposium in Biostatistics: Clinical Trials*, Lecture Notes in Statistics 1205,
DOI 10.1007/978-1-4614-5245-4_12, © Springer Science+Business Media New York 2013

association studies, and GARNET, a collection of randomized clinical trials. It also reflects our work as a data coordinating center for a number of randomized clinical trials including evaluation of vertebroplasty for osteoporotic fractures [11] and surgery for carpal tunnel syndrome [10], performed through the Center for Biomedical Statistics at the University of Washington.

At the time of the Fourth Seattle Symposium on Biostatistics there were 199 clinical trials listed at www.clinicaltrials.gov that were collecting genetic information on participants. The entry for trial NCT01106144, for example, states:

> The main component in the treatment of acute myeloid leukemia (AML) is consist of anthracycline (such as daunorubicin or idarubicin) and cytarabine. Inter-individual variability of transport/ metabolism of the chemotherapeutic agent and several genetic pathways involved in the drug action might be associated with different response following the treatment for AML usually consisted of chemotherapy and/or transplantation. One of potential pathways involved in the drug action is DNA repair pathway, accordingly single nucleotide polymorphisms (SNPs) in the DNA repair machinery pathway might be a predictive marker for therapy outcomes in AML.

This chapter focusses on the use of SNPs for clinical trials.

2 Single Nucleotide Polymorphisms

Information about the genetic constitution of an individual in a study is often provided by technologies that reveal SNP profiles. Each of us receives one genome, including 23 chromosomes, from each of our parents and the genome can be described by the base type (A,C,G, or T) at each of the three billion nucleotides in the genomic DNA sequence. There can be constraints on which bases can be present at each nucleotide position but there is now documentation of 25 million or so positions at which there is variation among people (http://www.1000genomes.org). The low rate of change with which one base may be replaced by another means that most SNPs have only two possible states in a population, such as A and C. If the frequency of type A in a population is $p_A = 0.8$, then C is termed the minor allele and the minor allele frequency (MAF) is 0.2.

Individuals may be typed at specific target regions of the genome but it is generally cost-effective to type many SNPs with platforms that give whole-genome data. The OMNI5 chip produced by the Illumina company allows five million SNPs to be typed (http://www.illumina.com/products), most of them with MAF values over 0.01 in publicly available data sets such as HapMap or 1000 Genomes (www.hapmap.org, or www.1000genomes.org). In a recent review, [5] listed sets of studies where associations of SNPs with drug response had been sought, often resulting in highly significant results.

3 Associations

The use of genetic markers for mapping disease genes or as biomarkers in clinical trials depends on associations between genetic variants and observed or measured traits. We first examine associations between genetic variants before taking up the association of markers with traits or outcomes.

It is convenient to describe associations between pairs of alleles in terms of correlations. At one locus, the correlation coefficient is an *inbreeding coefficient* and at two loci, the correlation depends upon *linkage disequilibrium*. For marker and trait locus pairs, the squared correlation coefficient is the key parameter.

3.1 Allelic Association at One Locus

For a set of n individuals in a sample from one population, it is convenient to replace every allele by an indicator variable for, say, allele A at locus $\mathbf{A}$. For allele k $(k = 1, 2)$ in individual j $(j = 1, 2, \ldots, n)$, these indicator variables x_{jk} are defined as

$$x_{jk} = \begin{cases} 1 \text{ allele is of type } A \\ 0 \text{ otherwise} \end{cases}$$

Taking averages over all samples from the population of these Bernoulli variables is straightforward:

$$\begin{aligned} \mathcal{E}(x_{jk}) &= p_A \\ \mathcal{E}(x_{jk}^2) &= p_A \\ \mathrm{Var}(x_{jk}) &= p_A(1 - p_A) \end{aligned}$$

where p_A is the allele frequency for A.

Now the product of the two x's for one individual is nonzero only if the individual is homozygous AA, and this leads to the covariance of indicator variables within individuals:

$$\begin{aligned} \mathcal{E}(x_{jk}x_{jk'}) &= P_{AA}, \quad k \neq k' \\ \mathrm{Cov}(x_{jk}, x_{jk'}) &= P_{AA} - p_A^2 \end{aligned}$$

where P_{AA} is the genotype frequency for AA.

The (within-population) inbreeding coefficient f_A at locus $\mathbf{A}$ is defined to allow the reparameterization of genotype frequencies in terms of allele frequencies:

$$P_{AA} = p_A^2 + p_A(1 - p_A)f_A$$

$$\begin{aligned}P_{Aa} &= 2p_A(1-p_A) - 2p_A(1-p_A)f_A \\ P_{aa} &= (1-p_A)^2 + p_A(1-p_A)f_A\end{aligned}$$

This imposes no constraints on genotypic proportions and it preserves the usual reduction of genotypic frequencies to allele frequencies:

$$p_A = P_{AA} + \frac{1}{2}P_{Aa},\ p_a = P_{aa} + \frac{1}{2}P_{As}$$

The inbreeding coefficient can be seen to be the correlation coefficient of the indicator variables for the two alleles carried by an individual at a locus. This follows because

$$\begin{aligned}\mathrm{Var}(x_{jk}) &= p_A(1-p_A) \\ \mathrm{Cov}(x_{jk}, x_{jk'}) &= p_A(1-p_A)f_A,\ \ k \neq k' \\ \mathrm{Corr}(x_{jk}, x_{jk'}) &= f_A\end{aligned}$$

Because genotypic frequencies are bounded by allele frequencies above and zero below,

$$0 \leq P_{AA} = p_A^2 + p_A p_a f_A \leq p_A$$

there are bounds on the inbreeding coefficient

$$\max\left(-\frac{p_A}{p_a}, -\frac{p_a}{p_A}\right) \leq f_A \leq 1$$

Sample Values If a sample of n individuals is found to have counts n_{AA}, n_{Aa}, n_{aa} for genotypes AA, Aa, aa, the sample allele frequencies, denoted by tildes, are

$$\tilde{p}_A = \frac{1}{2n}(2n_{AA} + n_{Aa}) = \tilde{P}_{AA} + \frac{1}{2}\tilde{P}_{Aa}$$

and these have means and variances over all samples from the same population of

$$\begin{aligned}\mathcal{E}(\tilde{p}_A) &= p_A \\ \mathrm{Var}(\tilde{p}_A) &= \frac{1}{2n}p_A(1-p_A)(1+f_A).\end{aligned}$$

A structured population is shown in Fig. 1, where the subpopulations have distinct allele frequencies p^*. Among samples from the ith subpopulation the allele counts are binomially distributed : $(2n\tilde{p}_i) \sim$ Binomial$(2n, p_i^*)$ providing there is Hardy–Weinberg equilibrium $(f = 0)$ in that subpopulation. We will return to variation among subpopulations later.

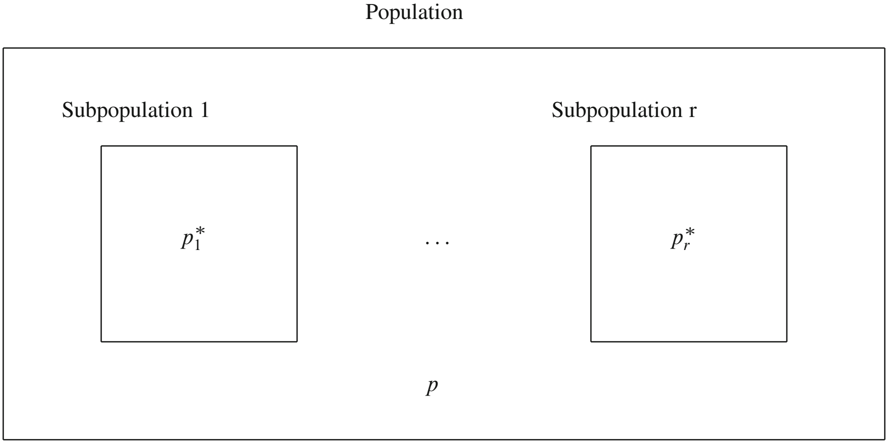

Fig. 1 Allele frequencies in a substructured population

3.2 Allelic Association at Two Loci

A *gamete* is the set of genetic material, in the egg or sperm, passed from parent to child and a *haplotype* is a relatively small region on one chromosome. For the present purposes the terms are essentially interchangeable, but "gamete" will be used as it has greater generality.

If data are available for a set of n gametes, indicator variables x_A and x_B can be defined for loci **A** and **B** in the same way as that above for one locus. For gamete j:

$$x_{j_A} = \begin{cases} 1 & \text{if gamete carries } A \\ 0 & \text{otherwise} \end{cases}$$

$$x_{j_B} = \begin{cases} 1 & \text{if gamete carries } B \\ 0 & \text{otherwise} \end{cases}$$

As the product $x_{j_A} x_{j_B}$ is nonzero only if the gamete is of type AB it has expectation $\mathcal{E}(x_{j_A} x_{j_B}) = P_{AB}$, where P_{AB} is frequency of that gamete type. Therefore

$$\begin{aligned} \mathcal{E}(x_{j_A}) &= p_A, \mathrm{Var}(x_{j_A}) = p_A(1 - p_A) \\ \mathcal{E}(x_{j_B}) &= p_B, \mathrm{Var}(x_{j_B}) = p_B(1 - p_B) \\ \mathcal{E}(x_{j_A} x_{j_B}) &= P_{AB}, \mathrm{Cov}(x_{j_A}, x_{j_B}) = D_{AB} = P_{AB} - p_A p_B \end{aligned}$$

The quantity D_{AB} is defined to be the (gametic) linkage disequilibrium between alleles A and B.

The correlation of indicator variables at two loci is

$$\rho_{AB} = \text{Corr}(x_{j_A}, x_{j_B}) = \frac{D_{AB}}{\sqrt{p_A(1-p_A)p_B(1-p_B)}}$$

and this is the two-locus analog of the inbreeding coefficient f. It is the parameter that determines the behavior of all genetic association tests. In practice, use is made of the squared sample value

$$r^2_{AB} = \frac{\tilde{D}^2_{AB}}{\tilde{p}_A(1-\tilde{p}_A)\tilde{p}_B(1-\tilde{p}_B)}$$

Because gamete frequencies are bounded above by allele frequencies and below by zero it can be seen that

$$0 \le D_{AB} + p_A p_B \le p_A \ , 0 \le -D_{AB} + p_A p_b \le p_A$$

$$0 \le \rho^2 \le \min\left(\tfrac{p_A p_b}{p_a p_B}, \tfrac{p_a p_B}{p_A p_b}\right) \le 1$$

The two loci need equal allele frequencies ($p_A = p_B$) for ρ^2_{AB} to attain the value 1.

Allelic Association at Two Unphased Loci It is generally the case that only genotypic data, rather than gametic data, are available and this suggests another measure of linkage disequilibrium. The indicator variables now need two subscripts: j ($j = 1, 2, \ldots, n$) for individual and k for gamete ($k = 1, 2$) within individual. Summing the indicator values at each locus:

$$X_{j_A} = x_{j1_A} + x_{j2_A}$$
$$X_{j_B} = x_{j1_B} + x_{j2_B}$$

provides indicators X_{j_A} having values $2, 1, 0$ with probabilities P_{AA}, P_{Aa}, P_{aa} and indicators X_{j_B} having values $2, 1, 0$ with probabilities P_{BB}, P_{Bb}, P_{bb}.

The means, variances, and covariances of the genotypic indicators are

$$\mathcal{E}(X_{j_A}) = 2p_A, \text{Var}(X_{j_A}) = 2p_A(1-p_A)(1+f_A)$$
$$\mathcal{E}(X_{j_B}) = 2p_B, \text{Var}(X_{J_B}) = 2p_B(1-p_B)(1+f_B)$$
$$\mathcal{E}(X_{j_A}X_{j_B}) = 4P_{AABB} + 2P_{AABb} + 2P_{AaBB} + P_{AaBb}$$
$$\text{Cov}(X_{j_A}, X_{j_B}) = 2D^c_{AB}$$

These equations introduce the "composite" linkage disequilibrium D_{AB}^c and

$$\mathrm{Corr}(X_{j_A}, X_{j_B}) = \frac{D_{AB}^c}{\sqrt{p_A p_a (1 + f_A) p_B p_b (1 + f_B)}}$$

With Hardy–Weinberg equilibrium, $f_A = f_B = 0$ and $D_{AB}^c = D_{AB}$. In this special case it is straightforward, although computationally challenging, to recover gamete frequencies from genotypic frequencies and it is possible to work only with gametic disequilibria.

3.3 Subgroup Analysis

Before considering association mapping methods that depend on linkage disequilibrium between marker and trait loci, we consider evaluating treatment effects in genotype subgroups. In a small modification of previous notation, the genotype indicator X_{lj} is the number of copies of the minor allele for the lth SNP in the jth individual, and we could consider a binary grouping: marker l is positive if $X_{lj} \geq 1$ and it is negative if $X_{lj} = 0$. Subjects can be placed into marker-positive and marker-negative subgroups.

Subgroup treatment effects Δ_{lg} are the differences in responses Y between the two treatment groups $Tx = 1$ and $Tx = 0$, for that subgroup:

$$\Delta_{lg} = \mathcal{E}(Y_j | X_{lj} = g, Tx = 1) - \mathcal{E}(Y_j | X_{lj} = g, Tx = 0) \tag{1}$$

The first question of interest is whether or not there are subgroups that have strong treatment effects, or evidence for harm: $H_0 : \Delta_{lg} = 0$. Within a subgroup, a test statistic for treatment effect is $z = \hat{\Delta}_{lg} / \sqrt{V_{lg}}$, where $\hat{\Delta}_{lg}$ is the observed effect for SNP l in that subgroup and V_{jg} is an estimate of the variance of the effect.

As an example of a study with which we have had experience, we refer to an evaluation of compound "X" by GlaxoSmithKline in which the main objective of the analysis is to identify genetic markers that influence the clinical efficacy of the compound for the treatment of disease "D."

A second question is whether or not there are subgroups that have treatment effects that are larger (or smaller) than the overall treatment effect. In other words, are there "enhanced" treatment effects? If $\bar{\Delta}$ is the treatment effect averaged over subgroups, or simply the overall marginal treatment effect:

$$\bar{\Delta} = \mathcal{E}(Y_j | Tx = 1) - \mathcal{E}(Y_j | Tx = 0)$$

this question is $H_0 : \Delta_{lg} = \bar{\Delta}$. This is equivalent to $H_0 : \Delta_{lg} = \bar{\Delta}_{lg^C}$ where g, g^C are the subgroups $X_{lj} = g, X_{lj} \neq g$.

4 Association Mapping

Association methods use random samples from a population and are alternatives to linkage methods based on pedigrees. The associations depend on linkage disequilibrium between marker and trait loci instead of depending on linkage between those loci as in pedigree methods.

Suppose that a quantitative trait locus **T** contributes to a trait of interest. The QTL genotype cannot be observed but maybe it can be inferred, and the location of the QTL estimated, from observations on the trait and the genotype at a genetic marker **M**. Individuals have observable marker genotypes and unobservable trait or response genotypes.

Each marker genotypic class M_uM_v is composed of a mixture of elements from each of the QTL classes, T_rT_s, where the proportion of QTL class T_rT_s contained within marker class M_uM_v is $Pr(T_rT_s|M_uM_v) = \Pr(T_rT_sM_uM_v)/\Pr(M_uM_v)$. With random mating, joint **TM** genotype frequencies are products of gamete frequencies as shown in Table 1, and gamete frequencies differ from products of allele frequencies because of linkage disequilibrium as shown in Table 2.

Trait Variables A treatment of association mapping also requires genetic variables Z and G for loci **M** and **T**. The values of Z are assigned for the marker, whereas the values G represent the genetic contributions to measured trait variables, disease status, or treatment response. The G's are not under control of the investigator. In either case, the Hardy–Weinberg assumption provides the following expressions for the means and variances:

$$\mathcal{E}(Z) = \mu_Z = p_M^2 Z_{MM} + 2p_M p_m Z_{Mm} + p_m^2 Z_{mm}$$
$$\mathcal{E}(G) = \mu_G = p_T^2 G_{TT} + 2p_T p_t G_{Tt} + p_t^2 G_{tt}$$

$$\text{Var}(Z) = \sigma^2_{A_M} + \sigma^2_{D_M}$$
$$\text{Var}(G) = \sigma^2_{A_T} + \sigma^2_{D_T}$$

Table 1 Two-allele genotypic frequencies

	TT	Tt	tt
MM	P_{MT}^2	$2P_{MT}P_{Mt}$	P_{Mt}^2
Mm	$2P_{MT}P_{mT}$	$2P_{MT}P_{mt} + 2P_{Mt}P_{mT}$	$2P_{Mt}P_{mt}$
mm	P_{mT}^2	$2P_{mT}P_{mt}$	P_{mt}^2

Table 2 Two-allele gametic frequencies

	T	t
M	$P_{MT} = p_M p_T + D_{MT}$	$P_{Mt} = p_M p_t - D_{MT}$
m	$P_{mT} = p_m p_T - D_{MT}$	$P_{mt} = p_m p_t + D_{MT}$

The “additive” and “dominance” components of variance are

$$
\begin{aligned}
\sigma^2_{A_M} &= 2p_M p_m[p_M(Z_{MM} - Z_{Mm}) + p_m(Z_{Mm} - Z_{mm})]^2 \\
\sigma^2_{A_T} &= 2p_T p_t[p_T(G_{TT} - G_{Tt}) + p_t(G_{Tt} - G_{tt})]^2
\end{aligned}
$$

$$
\begin{aligned}
\sigma^2_{D_M} &= p_M^2 p_m^2(Z_{MM} - 2Z_{Mm} + Z_{mm})^2 \\
\sigma^2_{D_T} &= p_T^2 p_t^2(G_{TT} - 2G_{Tt} + G_{tt})^2
\end{aligned}
$$

and the covariance of Z and G depends on the linkage disequilibrium ρ_{MT} between **M** and **T**:

$$\text{Cov}(G, Z) = \rho_{MT}\sigma_{A_T}\sigma_{A_M} + \rho^2_{MT}\sigma_{D_T}\sigma_{D_M}$$

If either Z or G are purely additive, then

$$\text{Cov}(G, Z) = \rho_{MT}\sigma_{A_T}\sigma_{A_M}$$

whereas if either is purely nonadditive

$$\text{Cov}(G, Z) = \rho^2_{MT}\sigma_{D_T}\sigma_{D_M}$$

The choice of marker coding, i.e. whether Z has only additive variance, only nonadditive variance, or a combination of the two, will determine the nature of trait genetic effects that can be detected by association mapping.

Suppose the measured trait or response variable has value Y where $Y = G + E$, the sum of the genetic effect G of locus **T** and all other effects E. These other effects may be supposed to have mean zero and to be independent of both G and the marker variable Z. Then

$$
\begin{aligned}
\mathcal{E}(Y) &= \mathcal{E}(G) \\
\text{Cov}(Y, Z) &= \text{Cov}(G, Z) \\
\text{Var}(Y) &= \sigma^2_{A_T} + \sigma^2_{D_T} + V_E
\end{aligned}
$$

4.1 Continuous Traits

Regression Trait values Y may be regressed on marker variables Z. The regression coefficient has parametric value

$$\beta_{YZ} = \frac{\text{Cov}(Y, Z)}{\text{Var}(Z)} = \frac{\rho_{MT}\sigma_{A_T}\sigma_{A_M} + \rho^2_{MT}\sigma_{D_T}\sigma_{D_M}}{\sigma^2_{A_M} + \sigma^2_{D_M}}$$

Marker variable Z is often chosen to be additive, e.g $Z_{MM} = 2, Z_{Mm} = 1, Z_{mm} = 0, \sigma^2_{D_M} = 0$, and then

$$\beta_{YZ} = \rho_{MT}\frac{\sigma_{A_T}}{\sigma_{A_M}}$$

Evidence for a nonzero regression slope is therefore evidence for linkage disequilibrium between trait and marker loci. This, in turn, is generally regarded as evidence for genomic proximity of these loci. A cluster of SNPs with nonzero regression coefficients is likely to delineate the region of a chromosome containing a trait locus.

The marker variable could also be made to have zero additive variance, e.g. $Z_{MM} = p_m, Z_{Mn} = 0, Z_{mm} = p_M$, and then

$$\beta_{YZ} = \rho^2_{MT}\frac{\sigma_{D_T}}{\sigma_{D_M}}$$

The size of the regression coefficient is lower than in the additive case, partly because $\rho^2_{AB} \leq \rho_{AB}$ and partly because it is generally the case that $\sigma^2_{D_T} \leq \sigma^2_{A_T}$.

For any scoring of the marker genotypes, a significant regression coefficient implies a significant linkage disequilibrium measure ρ_{MT} between marker and disease loci.

Correlation It may be more convenient to work with the correlation of Y and Z. For an additive marker variable

$$\text{Corr}(Y, Z) = \rho_{YZ} = \rho_{MT}h^{(T)}_Y$$

where $(h^{(T)}_Y)^2 = \sigma^2_{A_T}/(\sigma^2_{A_T} + \sigma^2_{D_T} + V_E)$ is the (narrow sense) *heritability* of trait Y due to locus **T**. Sample values r_{YZ} for the correlation ρ_{YZ} can be transformed to normal variables with Fisher's transformation

$$z = \frac{1}{2}\ln\left(\frac{1 + r_{YZ}}{1 - r_{YZ}}\right)$$

and then standard theory for correlation coefficients provides that, for $\alpha\%$ significance level and $(1 - \beta)\%$ power, the necessary sample size n is (approximately)

$$n = \left[\frac{2(z_{\alpha/2} + z_\beta)}{\ln\left(\frac{1+\rho_{YZ}}{1-\rho_{YZ}}\right)}\right]^2 + 3$$

For 90 % power, $z_\beta = 1.28$. For 90 % power and 1 % or 0.001 % significance level and for an SNP with $\rho^2_{MT} = 0.8$ to the disease gene and a trait with per-locus heritability $(h^{(T)}_Y)^2 = 0.2$ these sizes are about 85 or 185. Although heritabilities of

0.2 are not uncommon, these are values over all causal loci and the per-locus values are much smaller.

Analysis of Variance Instead of regressing trait values on marker scores, the trait means could be compared among marker classes. The expected trait means follow as

$$\begin{aligned}\mathcal{E}(Y|M_uM_v) &= \sum_{r,s} G_{rs}\,\mathrm{Pr}(T_rT_s|M_uM_v)\\ &= \sum_{r,s} Grs\,\mathrm{Pr}(T_rM_u, T_sM_v)/\,\mathrm{Pr}(M_uM_v)\end{aligned}$$

in general. For a trait locus with only two alleles, T, t, for marker homozygote MM and still assuming Hardy–Weinberg equilibrium

$$\mathcal{E}(Y|MM) = (G_{TT}P_{MT}^2 + 2G_{Tt}P_{MT}P_{Mt} + G_{tt}P_{Mt}^2)/p_M^2$$

The trait means among the three marker genotype classes are

$$\mathcal{E}(Y|MM) = \mu_G + 2\rho_{MT}A/p_M + \rho_{MT}^2 D/p_M^2$$

$$\mathcal{E}(Y|Mm) = \mu_G + \rho_{MT}A(1/p_M - 1/p_m) - \rho_{MT}^2 D/(p_M p_m)$$

$$\mathcal{E}(Y|mm) = \mu_G - 2\rho_{MT}A/p_m + \rho_{MT}^2 D/p_m^2$$

where $A = \sigma_{A_T}\sqrt{(p_M p_m)}$, $D = \sigma_{D_T}(p_M p_m)$, so that an analysis of variance will also test that $\rho_{MT} = 0$ and the test will be affected by both additive and dominance effects at the trait locus.

4.2 Dichotomous Traits

Case Only The case–control approach starts with independent samples of individuals who are either affected or not affected with a disease and compares marker frequencies between the two groups. The following development also applies to two arms of a clinical trial. The MM marker frequency among cases is

$$\mathrm{Pr}(MM|\mathrm{Case}) = p_M^2 + \frac{1}{\mu_G}\left[p_M\rho_{MT}A + \rho_{MT}^2\sigma_{D_T}D\right]$$

$$\mathrm{Pr}(Mm|\mathrm{Case}) = 2p_Mp_m + \frac{1}{\mu_G}\left[(p_m - p_M)\rho_{MT}A - 2\rho_{MT}^2 D\right]$$

$$\mathrm{Pr}(mm|\mathrm{Case}) = p_m^2 + \frac{1}{\mu_G}\left[-p_m\rho_{MT}A + \rho_{MT}^2 D\right]$$

Combining the genotypic frequencies to give allele frequencies:

$$\Pr(M|\text{Case}) = p_M + \frac{\rho_{MT}\sigma_{A_T}}{2\mu_G}\sqrt{2p_M p_m}$$
$$\Pr(m|\text{Case}) = p_m - \frac{\rho_{MT}\sigma_{A_T}}{2\mu_G}\sqrt{2p_M p_m}$$

Case-only HWE Testing The inbreeding coefficient at the marker locus in the case population, from the earlier definition of f, now written as $f = [\Pr(MM|\text{Case}) - \Pr(M|\text{Case})^2]/\{\Pr(M|\text{Case})[1 - \Pr(M|\text{Case})]\}$, is

$$f = \frac{\rho_{MT}^2(2\mu_G\sigma_{D_T} - \sigma_{A_T}^2)}{(\mu_G\sqrt{2p_M/p_m} + \rho_{MT}\sigma_{A_T})(\mu_G\sqrt{2p_m/p_M} - \rho_{MT}\sigma_{A_T})}$$

so that a test for Hardy–Weinberg equilibrium ($f = 0$) at the marker among cases is actually a test for linkage disequilibrium between marker and trait loci in the whole population. The power of this test depends on nf^2 which is proportional to ρ_{MT}^4 so the power will decrease quickly as ρ_{MT} decreases.

It is common for investigators to assume a multiplicative trait model (i.e., additive on a log scale), but that leads to Hardy–Weinberg equilibrium at marker loci among cases since then $2\mu_G\sigma_{D_T} = \sigma_{A_T}^2$.

Case–Control An argument similar to that above provides the marker genotype frequencies among controls:

$$\Pr(MM|\text{Control}) = p_M^2 - \frac{1}{1-\mu_G}\left[p_M\rho_{MT}A + \rho_{MT}^2 D\right]$$
$$\Pr(Mm|\text{Control}) = 2p_M p_m - \frac{1}{1-\mu_G}\left[(p_m - p_M)\rho_{MT}A - 2\rho_{MT}^2 D\right]$$
$$\Pr(mm|\text{Control}) = p_m^2 - \frac{1}{1-\mu_G}\left[-p_m\rho_{MT}A + \rho_{MT}^2 D\right]$$

Adding these to give allele frequencies:

$$\Pr(M|\text{Control}) = p_M - \frac{\rho_{MT}A}{2(1-\mu_G)}$$
$$\Pr(m|\text{Control}) = p_m + \frac{\rho_{MT}A}{2(1-\mu_G)}$$

The simplest case–control test compares marker allele frequencies between the two samples and it is clearly equivalent to testing that $\rho_{MT} = 0$ since

$$\Pr(M|\text{Case}) - \Pr(M|\text{Control}) \propto \rho_{MT}\sigma_{A_T}\sqrt{2p_M p_m}$$

The test is not affected by nonadditivity at the trait locus.

Table 3 Notation for trend test

	$i = 0$	$i = 1$	$i = 2$	
Marker genotype	MM	Mm	mm	Total
Marker variable	Z_0	Z_1	Z_2	
Case counts	r_0	r_1	r_2	R
Control counts	s_0	s_1	s_2	S
Total counts	n_0	n_1	n_2	N

If the allelic counts for M, m in cases and controls are laid out in a 2×2 table, the contingency-table chi-square test statistic has 1 df. An alternative is to work with the 3×2 table of marker genotype counts in cases and controls and calculate a 2 df chi-square test statistic. This test is affected by both additivity and nonadditivity at the trait locus but it is sensitive to errors in genotype calls for rare alleles. The main problem with the allelic case–control test is its sensitivity to departures from Hardy–Weinberg equilibrium, and for this reason there is preference for trend tests.

Trend Test The Armitage trend test is based on a score statistic U. Using the notation in Table 3:

$$U = \sum_{i=0}^{2} Z_i \left(\frac{S}{N} r_i - \frac{R}{N} s_i \right)$$

This is also the sample covariance between marker variable Z and disease status scored as 0 or 1 for case or control.

With random sampling, the case and control counts are multinomially distributed and the expected value of U is

$$\mathcal{E}(U) = \frac{SR}{N} \sum_i Z_i (R_i - S_i)$$

where R_i, S_i are the expected values of r_i, s_i. This expected value can be written as

$$\mathcal{E}(U) = \frac{1}{\mu_G(1-\mu_G)} \left[\rho_{MT} \sigma_{A_T} \sigma_{A_M} + \rho_{MT}^2 \sigma_{D_T} \sigma_{D_M} \right]$$

showing that it is zero when there is linkage equilibrium $\rho_{MT} = 0$.

The variance of U is

$$\text{Var}(U) = \frac{S^2 R}{N^2} \left(\sum_i Z_i^2 R_i - \left(\sum_i Z_i R_i \right)^2 \right) + \frac{SR^2}{N^2} \left(\sum_i Z_i^2 S_i - \left(\sum_i Z_i S_i \right)^2 \right)$$

Under the hypothesis of no association, $R_i = S_i, \mathcal{E}(U) = 0$ and

$$\text{Var}(U) = \frac{SR}{N} \left(\sum_i Z_i^2 R_i - \left(\sum_i Z_i R_i \right)^2 \right) = \frac{SR}{N} \left(\sigma_{A_M}^2 + \sigma_{D_M}^2 \right)$$

Assuming normality for U, the score test statistic is

$$X^2 = \frac{U^2}{\widehat{\text{Var}}(U)} = \frac{N(N\sum_i r_i Z_i - R\sum_i n_i Z_i)^2}{SR[N\sum_i n_i Z_i^2 - (\sum_i n_i Z_i)^2]}$$

and this is distributed as $\chi^2_{(1)}$ under the hypothesis $H_0 : \rho_{MT} = 0$.

It is usual to consider a linear trend test, say $Z_0 = 0, Z_1 = 1, Z_2 = 2$, so that $\sigma^2_{D_M} = 0$ and

$$X^2 = \frac{N[N(r_1 + 2r_2) - R(n_1 + 2n_2)]^2}{SR[N(n_1 + 4n_2) - (n_1 + 2n_2)^2]}$$

This will provide a test for additive effects at the disease locus. Setting $X_0 = p_m, X_1 = 0, X_2 = p_M$ gives $\sigma^2_{A_M} = 0$ and a test for nonadditive effects.

Effects of Inbreeding From the form of the allelic case–control test statistic

$$X^2_A = \frac{2N[N(r_1 + 2r_2) - R(n_1 + 2n_2)]^2}{SR[2N(n_1 + 2n_2) - (n_1 + 2n_2)^2]}$$

and the previous form of the genotypic linear trend test statistic it can be shown that

$$\mathcal{E}(X^2_A) \approx (1 + f)$$

$$\mathcal{E}(X^2_T) \approx 1$$

when there is inbreeding to extent f in the population. The trend test is therefore robust to departures from Hardy–Weinberg equilibrium. The other general concern about association tests are that they are sensitive to population structure.

Effect of Population Structure The effect of population structure on association tests can be phrased in terms of the variation in allele frequencies over subpopulations (Fig. 1). This variation reflects the dependence among individuals imposed by the history of the population. If x_{jk} indicates the allelic state of allele k in individual j, it is now necessary to consider that these states are dependent among individuals, as indicated by joint probability $P_{A,A}$ for two individuals each carrying A:

$$\mathcal{E}(x_{jk}x_{j'k'}) = P_{A,A}, \quad j \neq j', k \neq k'$$

This is no longer p^2_A as it was for the analyses within populations.

The expected value of squared sample allele frequency for a single subpopulation is changed to:

$$\mathcal{E}(\tilde{p}^2_A) = P_{A,A} + \frac{1}{2n}(p_A + P_{AA} - 2P_{A,A})$$

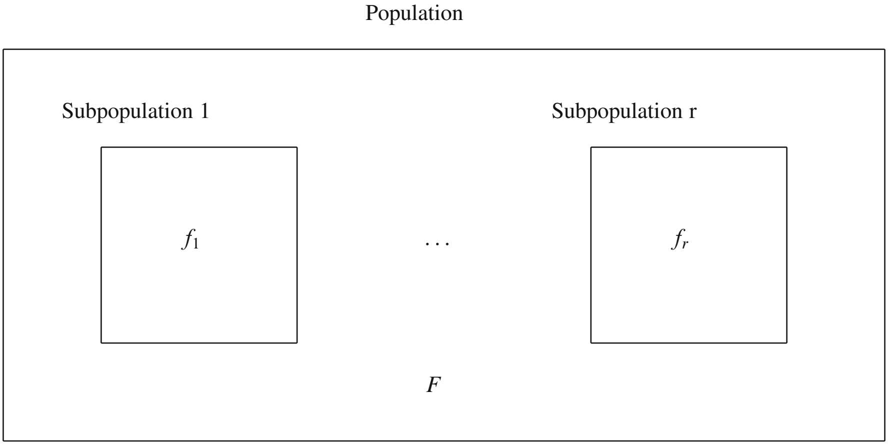

Fig. 2 Inbreeding coefficients in a structured population. Within a subpopulation: $P^*_{AA} = (p^*_A)^2 + fp^*_A(1 - p^*_A)$. Over all subpopulations: $P_{AA} = p_A^2 + Fp_A(1 - p_A)$

so that the (total) variance is

$$\text{Var}(\tilde{p}_A) = (P_{A,A} - p_A^2) + \frac{1}{n}(P_{AA} - P_{A,A}) + \frac{1}{2n}(p_A - P_{AA})$$

With this evolutionary perspective there is need for a new parameterization for joint allele frequencies to reflect the variation over subpopulations as well as over samples from one subpopulation:

$$P_{AA} = p_A^2 + p_A(1 - p_A)F$$

$$P_{A,A} = p_A^2 + p_A(1 - p_A)\theta$$

The total inbreeding coefficient F still refers to alleles within individuals but is for all individuals in the collection of subpopulations (see Fig. 2). The coancestry coefficient θ_i is for alleles in two individuals in the ith subpopulation. A common value θ is assumed here for all subpopulations and the subpopulations are assumed to be independent ($\theta_i = \theta, \theta_{ii'} = 0$ in Fig. 3). The frequency p_A is now the average over all subpopulations, and the total variance is

$$\text{Var}(\tilde{p}_A) = p_A(1 - p_A)\left[\theta + \frac{1}{n}(F - \theta) + \frac{1}{2n}(1 - F)\right]$$

There are three components of variance: among populations $p_A(1 - p_A)\theta$, among individuals within populations $p_A(1 - p_A)(F - \theta)$, and among alleles within individuals $p_A(1 - p_A)(1 - F)$, with a total variance of $p_A(1 - p_A)$.

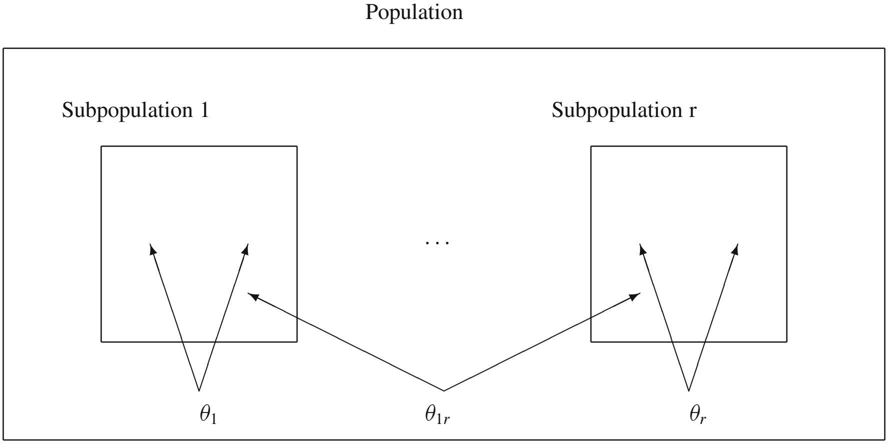

Fig. 3 Coancestry coefficients in a structured population. Within the ith subpopulation: $P_{A,A} = p_A^2 + \theta_i p_A(1 - p_A)$. Between the i, i'th subpopulations: $P_{A,A} = p_A^2 + \theta_{ii'} p_A(1 - p_A)$

Within a subpopulation it is still the case that

$$P^*_{AA} = (p^*_A)^2 + f p^*_A(1 - p^*_A)$$

Taking expectation values over populations, and using $f = (F - \theta)(1 - \theta)$:

$$\begin{aligned}
\mathcal{E}(p^*_A) &= p_A \\
\mathcal{E}[(p^*_A)^2] &= p_A^2 + \theta p_A(1 - p_A) \\
\mathcal{E}(P^*_{AA}) &= [p_A^2 + \theta p_A(1 - p_A)] + f[p_A - p_A^2 - \theta p_A(1 - p_A)] \\
&= p_A^2 + F p_A(1 - p_A)
\end{aligned}$$

If x_i, y_i, z_i are the proportions of cases, controls and all samples from the ith subpopulation, the case–control and trend test statistics have expectations

$$\mathcal{E}(X_A^2) \approx \frac{2RS\theta \sum_i (x_i - y_i)^2 + N(F - \theta) + N(1 - \theta)}{N(1 - \theta \sum_i z_i^2) - (F - \theta)/2 - (1 - \theta)/2}$$

$$\mathcal{E}(X_T^2) \approx \frac{2RS\theta \sum_i (x_i - y_i)^2 + N(F - \theta) + N(1 - \theta)}{N[(1 + F) - 2\theta \sum_i z_i^2]}$$

The behavior of these test statistics is therefore affected by both inbreeding (within subpopulations) and population structure (the existence of subpopulations).

If there is random mating within each subpopulation, $F = \theta$, and if there are equal numbers of cases and controls $R = S = N/2$. If there are many subpopulations, it is possible to ignore the term $F \sum_i z_i^2$ in the denominator and then

$$\mathcal{E}(X_A^2) = \mathcal{E}(X_T^2) \approx 1 + \frac{RF \sum_i (x_i - y_i)^2 - 2F}{1 + F}$$

as was given by Pritchard and Donnelly [14].

If there is only one subpopulation ($x_1 = y_1 = z_1 = 1$):

$$\mathcal{E}(X_A^2) \approx 1 + f$$

$$\mathcal{E}(X_T^2) \approx 1$$

as before.

5 Treatment Effects

In the previous section we considered individuals categorized by case/control status or by either of two treatment arms in a clinical trial. We now generalize this to allow for continuous treatments and we consider the joint effects of treatment and genotype on response. We preserve the notation of Y for response and X for genotype, and introduce S for treatment. For the jth subject and lth SNP we let μ_{lj} be the conditional mean response, recognizing that this depends on both genotype and treatment:

$$\mu_{lj} = \mathcal{E}(Y_j | X_{lj}, S_j)$$

with link function

$$g(\mu_{lj}) = \beta_0 + \beta_1 X_{lj} + \gamma_0 S_j + \gamma_1 X_{lj} S_j$$

For linear regression, $g(\mu) = \mu$ and for logistic regression $g(\mu) = \ln[\mu/(1 - \mu)]$.

In the previous section we accommodated a range of genetic models with the marker variable Z. We now express these models in terms of subgroup treatment effects as in Eq. (1). For an additive model, such as the genotype variable being the number of minor alleles, and for linear regression

$$\Delta_{l0} = \gamma_0$$

$$\Delta_{l1} = \gamma_0 + \gamma_1$$

$$\Delta_{l2} = \gamma_0 + 2\gamma_1$$

whereas for logistic regression the Δ's are treatment log-odds ratios specific for each genotype subgroup.

A dominant model regards the effects of one or two major alleles being the same and X_{lj} is replaced by 1 if there are one or no minor alleles and 0 if there are two minor alleles:

$$g(\mu_{lj}) = \beta_0 + \beta_1 \mathbf{1}(X_{lj} \geq 1) + \gamma_0 S_j + \gamma_1 \mathbf{1}(X_{lj} \geq 1) S_j$$

where the variable $\mathbf{1}(a)$ is 1 if a is true. This implies

$$\Delta_{l0} = \gamma_0$$

$$\Delta_{l1} = \gamma_0 + \gamma_1$$

$$\Delta_{l2} = \gamma_0 + \gamma_1$$

Conversely, a recessive model regards the effects of one or two minor alleles being the same and X_{lj} is replaced by 1 if there are no minor alleles and 0 if there are one or two minor alleles:

$$g(\mu_{lj}) = \beta_0 + \beta_1 \mathbf{1}(X_{lj} = 2) + \gamma_0 S_j + \gamma_1 \mathbf{1}(X_{lj} = 2) S_j$$

This implies

$$\Delta_{l0} = \gamma_0$$

$$\Delta_{l1} = \gamma_0$$

$$\Delta_{l2} = \gamma_0 + \gamma_1$$

A general, or nominal, model assigns different effect levels to each marker genotype:

$$\begin{aligned} g(\mu_{lj}) = \beta_0 &+ \beta_1 \mathbf{1}(X_{lj} = 1) + \beta_2 \mathbf{1}(X_{lj} = 2) \\ &+ \gamma_0 S_i + \gamma_1 \mathbf{1}(X_{lj} = 1) S_j + \gamma_2 \mathbf{1}(X_{lj} = 2) S_j \end{aligned}$$

This implies

$$\Delta_{l0} = \gamma_0$$

$$\Delta_{l1} = \gamma_0 + \gamma_1$$

$$\Delta_{l2} = \gamma_0 + \gamma_2$$

Table 4 Likelihoods for testing gene by treatment interaction

Model	Likelihood	Number of parameters	Maximum $\ln(L)$
Full	$X + S + X \times S$	$p + q$	$\ln(L_1)$
Null	$X + S$	p	$\ln(L_0)$

Table 5 Likelihood ratio tests for specific models

Model	Null parameters	LR test
Additive	γ_1	$\chi^2_{(1)}$
Dominant	γ_1	$\chi^2_{(1)}$
Recessive	γ_1	$\chi^2_{(1)}$
Nominal	γ_1, γ_2	$\chi^2_{(2)}$

Table 6 Power for gene by treatment interaction

	Test assumes		
Truth	Additive	Nominal	Dominant
Additive	0.882	0.800	0.746
Dominant	0.607	0.611	0.746

5.1 Testing for Gene by Treatment Interaction

A general testing procedure uses the likelihood ratio test based on alternative pairs of models for treatment and genetic effects. To test for no gene by treatment interaction, the likelihoods are displayed in Table 4. Under the null hypothesis of no interaction

$$2[\ln(L_1) - \ln(L_0)] \sim \chi^2_{(q)}$$

In Table 5 we display the interaction test parameters and distributions for each of the genetic models.

Designing a study to test for gene by treatment interaction is made complicated by the true genetic model being unknown. There is a body of literature to suggest that complex traits are additive, reflecting the independent and additive effects of alleles at the trait loci [8] but also a suggestion that at least some genes act in a nonadditive fashion [1,20]. In Table 6 we show powers of likelihood ratio tests under different model assumptions when the true model is either additive or dominant. For simulations we used data sets consisting of 780 cases and 780 controls. For an additive structure we assumed additivity on the probability scale with a prevalence of 0.4 when $X_l = 0$ and 0.3, 0.2 for $X_l = 1, 2$, respectively. For a dominant model we used a prevalence of 0.4 when $X_l = 0$ and 0.28 when X_l is either 1 or 2. Use of logistic regression with an additive genetic effect on the log odds scale is only an approximation to the true data-generating model (e.g., additive on probability scale).

5.2 Multiple Comparisons

With millions of SNPs being scored or imputed for each study participant, the issue of multiple testing needs to be considered. If α^* is the per-test (per-SNP) false-positive error rate, then a set of L tests under the null is expected to produce $\alpha^* L$ false positives. For $L \sim 10^6$ this number can be large. The family-wise error rate (FWER) α is the probability of at least one false positive in a set of tests where all the null hypotheses are true. The Bonferroni approach sets the per-test error to $\alpha^* = \alpha/L$, or 5×10^{-8} for an FWER of 0.05 and one million tests. The Sidak approach uses $\alpha^* = 1 - (1 - \alpha)^{1/L}$, with the very similar value of 5.13×10^{-8} in this case.

These simple multiple comparison corrections are conservative and they may assume independent tests even though linkage disequilibrium prevents whole-genome sets of SNPs being independent. Permutation procedures can yield correct FWER values for dependent tests, at the expense of being computationally intensive and dataset dependent. These procedures keep the genetic profiles intact and permute the outcomes Y among individuals: in essence destroying any genotype-trait association and producing data for which the null hypotheses at each SNP are true. Repeated permutations lead to reference distributions for the single-SNP test statistics. Approximations were proposed by Nyholt [13].

5.3 Bayes' Factors

Genetics research is embracing "evidence" criteria such as likelihood ratios. In linkage studies, where the transmission of trait values and genotypes are traced down pedigrees, LOD scores based on likelihood ratios have long been used. The appeal of an alternative that does not rely on p-values, the probabilities of data assuming the null hypothesis to be true, is that power (probabilities when the null is not true) and sample size can be considered when choosing criteria for evaluating tests.

The Bayes Factor is the ratio of the probability of the data under the null to the probability under the alternative hypothesis. Wakefield [17] showed that an Approximate Bayes Factor (ABF) is given by

$$\text{ABF} = \sqrt{\frac{V + W}{V}} \exp\left(-\frac{Z^2}{2} \frac{W}{V + W} \right)$$

where the maximum likelihood estimate $\hat{\theta}$ of a parameter of interest is normally distributed with mean θ and variance V. The test statistic Z for the hypothesis $H_0 : \theta = 0$ is $Z = \hat{\theta}/\sqrt{V}$. The Bayesian aspect of the analysis is to assign a prior distribution for parameter θ, such as normal with mean 0 and variance W, and to decide in favor of the alternative hypothesis if

$$\text{ABF} \times \text{PO} < R \tag{2}$$

The prior odds PO is the ratio of the probabilities of the hypotheses before the data are collected, PO $= \Pr(H_0)/\Pr(H_1)$ and R is the ratio of costs: the cost of a false non-discovery divided by the cost of a false discovery. For example, R is the cost of a Type II error divided by that of a Type I error.

In practice, values are assigned to (PO/R) and then the ABF threshold is determined. It may be that PO is 10,000 and $R = 1$, so that the threshold is 10^{-4}. The ABF threshold is translated into a threshold for the test statistic from Eq. (2).

5.4 Bioinformatics Tools

The biological significance of an association found between an SNP and an outcome can be phrased in terms of the biological function of an SNP or the biological pathway to which it belongs. There are a variety of tools available to help this activity:

- UCSC Genome Bioinformatics (http://ucsc.genome.edu).
- Fast SNP (http://fastsnp.ibm.sinica.edu.tw).
- Gene Ontology (GO) (http://www.geneontology.org).

These and other tools were reviewed by Bansal et al. [1].

6 Analyses with Multiple Markers

For a set of individuals j there is information Y_j on the outcome of interest, on the treatment or dose S_j, and on the genetic profile $\mathbf{X}_j = \{X_{lj}, l = 1, 2, \dots L\}$. A series of questions can be posed:

- How can the genetic markers be used to predict outcome?
- How can the genetic markers be used to score the individuals with respect to treatment benefit?
- How can the genetic markers be used to create a treatment decision function?

We address these questions in a generalized linear model framework:

$$\mathcal{E}(Y_j|\mathbf{X}_j, S_j) = \mu_j$$
$$g(\mu_j) = \beta(\mathbf{X}_j) + \gamma(\mathbf{X}_j) \times S_j$$

Hastie and Tibshirani [7] introduced a "varying coefficient model." As a simple example

$$g(\mu_j) = (\beta_0 + \beta_1 X_{1j} + \dots + \beta_L X_{Lj}) + (\gamma_0 + \gamma_1 X_{1j} + \dots + \gamma_L X_{Lj}) \times S_j$$

Major challenges to this work include

- How do we select SNPs to include in $\beta(\mathbf{X}_j)$ and $\gamma(\mathbf{X}_j)$?
- Should we also consider epistasis: gene by gene interactions, $X_{lj} \times X_{l'j}, l \neq l'$ or higher-order interaction?
- How can we fit a model when L, the number of SNPs, is much greater than n, the number of individuals?
- How do the model choice criteria reflect the ultimate clinical goal of the model (for example, prediction versus treatment selection)?

6.1 Regularization Methods

Tibshirani [16] introduced "lasso" for regression shrinkage and selection. He discussed maximizing an objective function, such as a likelihood, subject to constraints or a penalty. More generally, a set of parameters ` is estimated as

$$\hat{\boldsymbol{\theta}} = \underset{\theta}{\text{argmax}} \left(\sum_j \ln \Pr(Y_j|\mathbf{X}_j, S_j, `) - \lambda \sum_j |\theta_j|^p \right)$$

There are three special cases:

- $p = 1$: Lasso [16] with penalty $\lambda \sum_j |\theta_j|$.
- $p = 2$: Ridge Regression [9] with penalty $\lambda \sum_j |\theta_j|^2$.
- $p = 1, 2$: Elastic Net [19] with penalty $\lambda_1 \sum_j |\theta_j| + \lambda_2 \sum_j |\theta_j|^2$.

Some comments about regularization methods are:

- Lasso tends to "select" variables by keeping $\hat{\beta}_l = 0$.
- Ridge regression tends to include all variables, but with small coefficients. This is essentially no selection.
- Lasso will not estimate a model with more nonzero coefficients than there are individuals.
- Fast algorithms exist for calculating regularization paths.
- Lasso tends to select only one variable from a set of highly correlated predictors.

6.2 Example

Wu et al. [18] presented an analysis of SNP data in a case–control setting, and conducted simulations to demonstrate the feasibility of allowing for interactions among SNPs. In their example they analyzed $n = 2{,}200$ subjects with 778 having Coeliac Disease and 1,422 as controls. Using LASSO Wu et al. [18] constructed

Table 7 Treatment selection example

Genotype	$\bar{Y}_j(0)$	$\bar{Y}_j(1)$	A_0	A_1	A^*
0 (30 %)	10	20	0	1	1
1 (50 %)	10	10	0	1	0
2 (20 %)	15	5	0	1	0
Population mean			11	12	14

a multi-marker predictive model using $L = 310,637$ SNPs. In order to explore models of increasing dimension the authors chose the L_1 penalty parameter λ to obtain a fixed number (e.g., 5, 10, 20, 50) of predictors with nonzero coefficients, and used cross-validation to evaluate the accuracy of models with increasing dimension. Using a sequence of models has the attractive property that subsets of markers can be ordered in terms of their inclusion in the regression models. Rather than focusing on a specific model selection criterion such as the area under the ROC curve, or a statistical loss function, Wu et al. [18] evaluate a sequence of models of a specified dimensionality (such as using 50 SNPs). These authors clearly demonstrate the ability of modern penalized regression methods to consider development and evaluation of multi-marker models, and provide guidance for model development and the potential evaluation of interactions.

6.3 *Treatment Selection*

How can genetic marker analysis provide a scoring for treatment selection? To answer this question it is necessary to state the goals in statistical terms. Gunter et al. [6] formulated an action function and then defined the resulting population mean outcome that would result when using the specified action function:

$$\text{Action function: } A(\mathbf{X}_j) = a$$

$$\text{Population result: } \mathcal{E}_a \mathcal{E}_Y[Y_j(a) | A(\mathbf{X}_j) = a] = \mu_A$$

Here $Y_j(a)$ is the potential outcome for subject j if treated with choice a. For example, $a = 1$ may be "treat" and $a = 0$ may be "do not treat." Alternatively, a may be a dose level.

As a small example, consider the values shown in Table 7 for a single marker. There are three genetic marker values (0,1,2) and two treatment values (0,1). The function A_0 always assigns $A = 0$, while A_1 always assigns $A = 1$. However, the action function A^* is optimal in the sense it maximizes μ_A over all possible functions A.

With a vector $\mathbf{X}_j$ of genetic values the optimal action rule is

$$A^*(\mathbf{X}_j) : \text{argmax}_A \mu_A = \text{argmax}_A \mathcal{E}_a \mathcal{E}_Y[Y_j(a) | A(\mathbf{X}_j) = a]$$

The goal is to determine which components of $\mathbf{X}_j$ are prescriptive markers, i.e. those with qualitative interactions rather than simply having quantitative interactions with treatment.

The space of functions $\{A(x)\}$ has high dimension: with each genotype taking three values, and with L total genotypes there are 3^L possible genotypes that the function $A(x)$ can evaluate. In addition, there are two possible outcomes for each genotype evaluated (e.g., treat, or not treat) leading to $(3^L)^2$ binary actions a. Therefore, computation methods are needed to define model search strategies that can maximize the intended performance metric such as μ_A in order to identify an optimal allocation rule, A^*.

Gunter et al. [6] suggested that the following marginal characteristics of a marker are important for that marker to have prescriptive potential:

- Fraction with benefit:

$$p_{l1} = \mathcal{E}\left\{\mathbf{1}\left(\text{argmax}_a \mathcal{E}[Y_j(a)|X_{lj}, A(X_{lj}) = a] = 1\right)\right\}$$

- Interaction magnitude:

$$\begin{aligned} D_l &= \max_j \left\{\mathcal{E}[Y_j(a)|X_{lj}, a = 1] - \mathcal{E}[Y_j(a)|X_{lj}, a = 0]\right\} \\ &\quad - \min_j \left\{\mathcal{E}[Y_j(a)|X_{lj}, a = 1] - \mathcal{E}[Y_j(a)|X_{lj}, a = 0]\right\} \end{aligned}$$

Since p_{l1} doesn't capture the number of subjects who have their treatment changed because of their genetic profile, Gunter et al. [6] suggest using $P_l = p_{l1}(1-p_{l1})$. The motivation for considering marginal measures for X_{lj} is to provide an algorithm that can search among candidate functions $A(\mathbf{X}_j)$ using an attractive subset of $\mathbf{X}_j$.

Gunter et al. [6] then suggest two criteria for ranking markers:

$$\begin{aligned} U_l &= \text{Scale}(P_l) \times \text{Scale}(D_l) \\ \text{Scale}(x_l) &= (x_l) - \min_k x_k)/(\max_k x_k - \min_k x_k) \end{aligned}$$

The second criterion measures the impact on the mean outcome using X_{lj} optimally to direct treatment:

$$\begin{aligned} T_l &= \mathcal{E}_X\left\{\max_a \mathcal{E}[Y_j(a)|X_{lj}, A = a]\right\} - \max(\mu_{A_0}, \mu_{A_1}) \\ &= \mu_{A^*} - \max(\mu_{A_0}, \mu_{A_1}) \end{aligned}$$

With all these quantities defined, we can state the selection algorithm of Gunter et al. [6]:

1. Use lasso with K-fold cross-validation to obtain an additive model estimate of $\mathcal{E}[Y_j|\mathbf{X}_j, S_l]$.
2. Estimate U_l and/or T_l, then rank $\mathbf{X}_l$.
3. For $h = 1, 2, \ldots, H$:
    - Use lasso with top h markers, main effects from step 1, and interactions between top h markers and treatment.
    - Estimate λ_1 based on CV with a focus on $\mu^h_{A^*}$ obtained from h markers: $A^*(\mathbf{X}^h_l)$.
4. Choose h that maximizes $\mu^h_{A^*}$. Done.

Step 1 of this algorithm is suggested in order to stabilize estimates of the mean function $\mathcal{E}[Y_j|\mathbf{X}_j, S_j]$ used to estimate U_l and T_l. It may not be needed for genotype markers. This model selection targets out-of-sample estimation of the optimal population outcome. The algorithm is limited to a small number of candidate models. There may be other search procedures but Gunter et al. [6] used simulations to compare their algorithm to standard lasso and found it performs slightly better.

The work of Gunter et al. [6] offers a focus on model development with the goal of defining a treatment selection function. These authors target an optimal result at the population level.

7 Resemblance Between Relatives

Some individuals in a clinical trial may be related to each other, whether or not this is by design. Here "related" means members of the same family. There is also a low-level evolutionary relatedness that results in a low level of inbreeding within a study population. Because relatedness depends on previous generations it is necessary to work with the coefficients F and θ, and extensions of these, rather than the within-population coefficient f.

To extend the earlier treatment, it may now be supposed there is an arbitrary number of alleles at a trait/response locus, and the genetic value for genotype T_rT_s is written as

$$G_{rs} = \mu_0 + \alpha_r + \alpha_s + \delta_{rs}$$

where

$$\begin{aligned}
\mu_0 &= \sum_r \sum_s p_r p_s G_{rs} = G_{..} \\
\alpha_r &= \sum_s p_s G_{rs} - \mu = G_{r.} - G_{..} \\
\delta_{rs} &= G_{rs} - \mu - \alpha_r - \alpha_s = G_{rs} - G_{r.} - G_{s.} + G_{..}
\end{aligned}$$

These imply that $\sum_r p_r\alpha_r = 0, \sum_r \delta_{rs} = 0$. The variance components are

$$\sigma_A^2 = 2\sum_r p_r\alpha_r^2$$
$$\sigma_D^2 = \sum_r\sum_s p_r p_s \delta_{rs}^2$$

If several loci contribute to a trait, the effects of all alleles can be summed, and interactions (epistasis) introduced between loci. For an individual with alleles $T_{lr}, r = 1, 2$, at locus l:

$$Y = \mu + \sum_l \left[\sum_r \alpha_{lr} + \sum_r\sum_{r'} \delta_{lrr'}\right]$$
$$+ \sum_{l\neq l'}\left[\sum_r\sum_{r'}(\alpha\alpha)_{lr,l'r'} + \sum_r\sum_s\sum_{s'}(\alpha\delta)_{lr,l'ss'} + \sum_r\sum_{r'}\sum_s\sum_{s'}(\delta\delta)_{lrr',l'ss'}\right] + \ldots$$

The total genetic variance becomes

$$\text{Var}(G) = \sigma_A^2 + \sigma_D^2 + \sigma_{AA}^2 + \sigma_{AD}^2 + \sigma_{DD}^2 + \ldots$$

where the variance components are

$$\begin{aligned}
\sigma_A^2 &= 2\sum_l\sum_r p_{lr}\alpha_{lr}^2 \\
\sigma_D^2 &= \sum_l\sum_r\sum_{r'} p_{lr}p_{lr'}\delta_{lrr'}^2 \\
\sigma_{AA}^2 &= 2\sum_{l\neq l'}\sum_r\sum_{r'} p_{lr}p_{l'r'}(\alpha\alpha)_{lr,l'r'}^2 \\
&\ldots
\end{aligned}$$

7.1 Trait Mean in Inbred Populations

Inbreeding, whether due to individuals having related parents or simply a consequence of populations being finite, affects the trait or response mean in a population. If random members of a population are inbred to an extent F relative to a reference or founder population, the genotype frequencies are

$$P_{rr} = p_r^2 + Fp_r(1 - p_r)$$

$$P_{rs} = 2p_r p_s(1-F), \;\; r \neq s$$

The expected trait value is, therefore,

$$\begin{aligned}
\mu_F &= \sum_r \sum_s P_{rs}(\mu_0 + \alpha_r + \alpha_s + \delta_{rs}) \\
&= \sum_r \left[Fp_r + (1-F)p_r^2\right](\mu_0 + 2\alpha_r + \delta_{rr}) \\
&\quad + \sum_{r \neq s} [p_r p_s(1-F)]\,(\mu_0 + \alpha_r + \alpha_s + \delta_{rs}) \\
&= F\sum_r p_r(\mu_0 + 2\alpha_r + \delta_{rr}) + (1-F)\sum_r \sum_s p_r p_s(\mu_0 + \alpha_r + \alpha_s + \delta_{rs}) \\
&= F(\mu_0 + \sum_r p_r\delta_{rr}) + (1-F)(\mu_0) \\
&= \mu_0 + FH
\end{aligned}$$

This result uses the notation $H = \sum_r p_r\delta_{rr}$ where the δ's terms are as defined in the non-inbreeding case. The mean in an inbred population changes with the degree of inbreeding and the degree of dominance.

For a trait affected by multiple loci, the mean is also affected by the two-locus inbreeding coefficient and the degree of dominance by dominance epistasis.

7.2 Genetic Variance in Inbred Populations

For the trait variance, it is necessary to find the expected value of the square of the linear model for trait values. Using a similar approach as that for the mean:

$$\begin{aligned}
\mathcal{E}(G_{rs}^2) = F\sum_r p_r(\mu_0^2 + 4\alpha_r^2 + \delta_{rr}^2 + 4\mu_0\alpha_i + 2\mu_0\delta_{rr} + 4\alpha_r\delta_{rr}) \\
+ (1-F)\sum_{r,s} p_r p_s(\mu_0^2 + \alpha_r^2 + \alpha_s^2 + 2\alpha_r\alpha_s + 2\mu_0\alpha_r + 2\mu_0\alpha_s \\
+ 2\mu_0\delta_{rs} + 2\alpha_r\delta_{rs} + 2\alpha_r\delta_{rs} + \delta_{rs}^2)
\end{aligned}$$

$$\begin{aligned}
= F\left[\mu_0^2 + 4\sum_r p_r\alpha_r^2 + \sum_r p_r\delta_{rr}^2 + 2\mu_0\sum_r p_r\delta_{rr} + 4\sum_r p_r\alpha_r\delta_{rr}\right] \\
+ (1-F)\left(\mu_0^2 + 2\sum_r p_r\alpha_r^2 + \sum_{r,s} p_r p_s\delta_{rs}^2\right)
\end{aligned}$$

The genetic variance becomes

$$\begin{aligned}\sigma_G^2 &= 2(1+F)\sum_r p_r\alpha_r^2 + (1-F)\sum_{r,s} p_r p_s \delta_{rs}^2 + F\sum_r p_r\delta_{rr}^2 \\ &\quad + 4F\sum_r p_r\alpha_r\delta_{rr} - F^2\left(\sum_r p_r\delta_{rr}\right)^2 \\ &= (1+F)\sigma_A^2 + (1-F)\sigma_D^2 + 4FD_1 + FD_2 + F(1-F)H^2\end{aligned}$$

to introduce $D_1 = \sum_r p_r\alpha_r\delta_{rr}$ and $D_2 = \sum_r p_r\delta_{rr}^2$. This expression for variance allows for inbreeding but not for linkage disequilibrium among the trait loci. For two equally frequent alleles, $D_1 = 0, D_2 = 0$. For additive traits, $H = D_1 = D_2 = 0$.

For a trait affected by multiple loci, the variance in an inbred population involves the two-locus inbreeding coefficient F_{11} as well as the usual one locus coefficient:

$$\begin{aligned}\sigma_G^2 &= (1+F)\sigma_A^2 + (1-F)\sigma_D^2 + (1+2F+F_{11})\sigma_{AA}^2 \\ &\quad + (1-F_{11})\sigma_{AD}^2 + (1-2F+F_{11})\sigma_{DD}^2 + \ldots\end{aligned}$$

It is not always the case that $F_{11} = F^2$. This expression for variance allows for inbreeding but not for linkage disequilibrium among the trait loci.

7.3 Genetic Covariance for Two Individuals

If J, J', with genotypes T_rT_s and $T_{r'}T_{s'}$, are two members of a population, the covariance of trait values for the two individuals rests on the covariance of their genetic values which, in turn, rests on their joint genotypic frequencies:

$$\begin{aligned}\text{Cov}(G_J, G_{J'}) &= \mathcal{E}(G_J G_{J'}) - \mathcal{E}(G_J)\mathcal{E}(G_{J'}) \\ \mathcal{E}(G_J G_{J'}) &= \sum_{r,s,r',s'} P_{rs,r's'} G_{rs} G_{r's'}\end{aligned}$$

For one individual P_{rs} can be written in terms of allele frequencies and the inbreeding coefficient F. For two individuals $P_{rs,r's'}$ needs an expanded set of probabilities that alleles are identical by descent (ibd).

A complete description of the ibd status among four alleles a, b, c, d carried by two individuals J, J' with alleles ab and cd requires 15 measures, as opposed to the two, F and $1-F$, for one individual. If there is no need to distinguish between the identity status for maternal and paternal alleles, the 15 ibd states can be collapsed into the nine states shown in Fig. 4. Solid lines in that figure join alleles that are identical by descent. State S_3, shown as identity among alleles a, b, c also represents identity among alleles a, b, d. Note here that a, b, c, d are labels to distinguish one allele from another: they do not indicate allelic type.

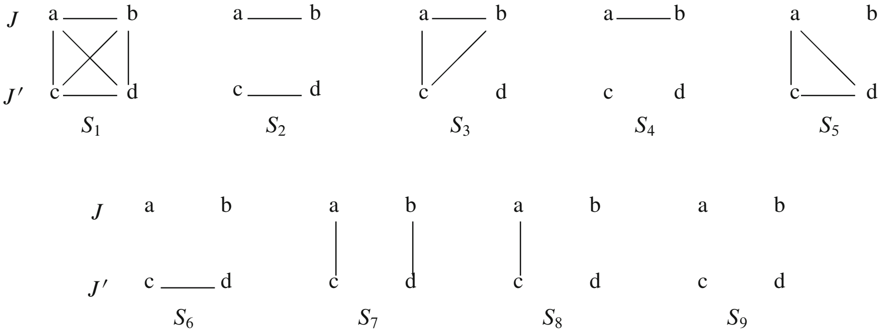

Fig. 4 Reduced identity states S for two individuals $J(a, b)$ and $J'(c, d)$

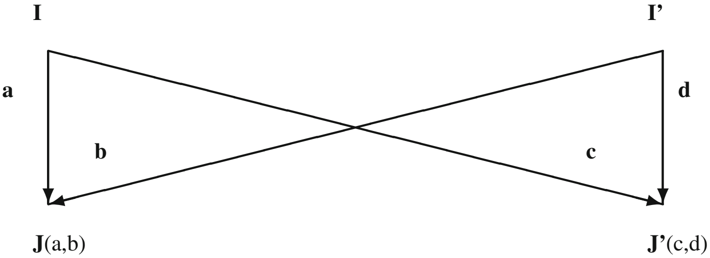

Fig. 5 IBD coefficients for Sibs J, J' from unrelated parents I, I'

The coancestry coefficient $\theta_{JJ'}$ referred to earlier is the probability that a random allele from $J(ab)$ is ibd ($\equiv$) to a random allele from $J'(cd)$:

$$\begin{aligned}\theta_{JJ'} &= \frac{1}{4}\left[\Pr(a \equiv c) + \Pr(a \equiv d) + \Pr(b \equiv c) + \Pr(b \equiv d)\right] \\ &= \Delta_1 + \frac{1}{2}(\Delta_3 + \Delta_5 + \Delta_7) + \frac{1}{4}\Delta_8\end{aligned}$$

For non-inbred relatives

$$\theta = \frac{1}{2}\Delta_7 + \frac{1}{4}\Delta_8$$

As an example, the pedigree for two non-inbred sibs J, J' with parents I, I' is shown in Fig. 5.

Non-inbred Relatives When neither individual is inbred, neither a, b nor c, d are ibd. There are only three states and the three probabilities are often written as $k_2 = \Delta_7, k_1 = \Delta_8$ or $k_0 = \Delta_9$ to indicate the number of pairs of ibs alleles carried by the two individuals. Values of these three probabilities for some common relationships are shown in Table 8.

Table 8 Identity coefficients for common non-inbred relatives

Relationship	k_2	k_1	k_0	$\theta = \frac{1}{2}k_2 + \frac{1}{4}k_1$
Identical twins	1	0	0	$\frac{1}{2}$
Full sibs	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$
Parent child	0	1	0	$\frac{1}{4}$
Double first cousins	$\frac{1}{16}$	$\frac{3}{8}$	$\frac{9}{16}$	$\frac{1}{8}$
Half sibs[a]	0	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{8}$
First cousins	0	$\frac{1}{4}$	$\frac{3}{4}$	$\frac{1}{16}$
Unrelated	0	0	1	0

[a] Also grandparent grandchild and avuncular (e.g., uncle niece)

Table 9 SNP genotype probabilities for pairs of relatives

Genotypes	General	Non-inbred
AA, AA	$\Delta_1 p_A + (\Delta_2 + \Delta_3 + \Delta_5 + \Delta_7)p_A^2$ $+(\Delta_4 + \Delta_6 + \Delta_8)p_A^3 + \Delta_9 p_A^4$	$k_0 p_A^4 + k_1 p_A^3 + k_2 p_A^2$
aa, aa	$\Delta_1 p_a + (\Delta_2 + \Delta_3 + \Delta_5 + \Delta_7)p_a^2$ $+(\Delta_4 + \Delta_6 + \Delta_8)p_a^3 + \Delta_9 p_a^4$	$k_0 p_a^4 + k_1 p_a^3 + k_2 p_a^2$
Aa, Aa	$2\Delta_7 p_A p_a + \Delta_8 p_A p_a + 4\Delta_9 p_A^2 p_a^2$	$4k_0 p_A^2 p_a^2 + k_1 p_A p_a + 2k_2 p_A p_a$
AA, Aa	$\Delta_3 p_A p_a + (2\Delta_4 + \Delta_8)p_A^2 p_a + 2\Delta_9 p_A^3 p_a$	$2k_0 p_A^3 p_a + k_1 p_A^2 p_a$
Aa, AA	$\Delta_5 p_A p_a + (2\Delta_6 + \Delta_8)p_A^2 p_a + 2\Delta_9 p_A^3 p_a$	$2k_0 p_A^3 p_a + k_1 p_A^2 p_a$
aa, Aa	$\Delta_3 p_A p_a + (2\Delta_4 + \Delta_8)p_A p_a^2 + 2\Delta_9 p_A p_a^3$	$2k_0 p_A p_a^3 + k_1 p_A p_a^2$
Aa, aa	$\Delta_5 p_A p_a + (2\Delta_6 + \Delta_8)p_A p_a^2 + 2\Delta_9 p_A p_a^3$	$2k_0 p_A p_a^3 + k_1 p_A p_a^2$
AA, aa	$\Delta_2 p_A p_a + \Delta_4 p_A p_a^2 + \Delta_6 p_A^2 p_a + \Delta_9 p_A^2 p_a^2$	$k_0 p_A^2 p_a^2$
aa, AA	$\Delta_2 p_A p_a + \Delta_6 p_A p_a^2 + \Delta_4 p_A^2 p_a + \Delta_9 p_A^2 p_a^2$	$k_0 p_A^2 p_a^2$

7.4 Joint Genotypic Probabilities for Relatives

The set of identity measures $\Delta_i = \Pr(S_i)$ for identity states S_i allow the joint genotypic probabilities to be written as in Table 9 for SNPs. These in turn allow for the covariance in trait values to be found for any pair of relatives.

7.5 Genetic Covariance for Non-inbred Relatives

The general expression for covariance can be shown to be

$$\begin{aligned}\mathcal{C}_{JJ'} &= 2\theta_{JJ'}\sigma_A^2 + \Delta_7\sigma_D^2 + (4\Delta_1 + \Delta_3 + \Delta_5)D_1 \\ &+ \Delta_1 D_2 + (\Delta_1 + \Delta_2 - F_J F_{J'})H^2\end{aligned}$$

Table 10 Genetic covariances of common non-inbred relatives

Relationship	Genetic covariance
Identical twins	$\sigma_A^2 + \sigma_D^2$
Full sibs	$\frac{1}{2}\sigma_A^2 + \frac{1}{4}\sigma_D^2$
Parent-child	$\frac{1}{2}\sigma_A^2$
Double first cousins	$\frac{1}{4}\sigma_A^2 + \frac{1}{16}\sigma_D^2$
Half sibs[a]	$\frac{1}{4}\sigma_A^2$
First cousins	$\frac{1}{8}\sigma_A^2$
Unrelated	0

[a] Also grandparent–grandchild and avuncular (e.g., uncle–niece)

When J and J' are the same individual, $\theta_{JJ'} = (1 + F)/2$, $\Delta_1 = F$ and $\Delta_7 = (1 - F)$. The other seven Δ's are zero, so

$$\mathcal{C}_{JJ} = \mathrm{V}_I = (1 + F)\sigma_A^2 + (1 - F)\sigma_D^2 + 4FD_1 + FD_2 + F(1 - F)H^2$$

as expected.

For non-inbred relatives

$$\mathcal{C}_{JJ'} = \left(k_2 + \frac{1}{2}k_1\right)\sigma_A^2 + k_2\sigma_D^2$$

and values for common relationships are shown in Table 10.

7.6 *Heritability*

Trait and response values have both genetic and environmental components. The simplest model of $Y = G + E$ leads to the variance of trait values among individuals J in a non-inbred population of unrelated individuals:

$$\mathrm{Var}_J = \sigma_A^2 + \sigma_D^2 + \sigma_E^2$$

This is also referred to as the phenotypic variance σ_P^2.

For an additive trait and for individuals that have no shared environment, the variance–covariance matrix for a sample of related pairs I, I' and inbred individuals I has elements

$$\mathrm{Var}_J = (1 + F_I)\sigma_A^2 + \sigma_E^2$$
$$\mathrm{Cov}_{JJ'} = 2\theta_{JJ'}\sigma_A^2$$

The narrow-sense heritability h^2 is defined as $h^2 = \sigma_A^2/\sigma_P^2$. The correlation of additive trait values for pairs of non-inbred individuals related to an extent $\theta_{JJ'}$ is, therefore,

$$\rho_{JJ'} = 2\theta_{JJ'}h^2$$

Traditional methods for estimating heritability have used trait values measured for sets of individuals whose relationship is known from their family membership. It is common, for example, to take measurements on monozygotic (MZ) and dizygotic (DZ) twins and make use of the relationships

$$\begin{aligned} \mathrm{Var}_J &= \sigma_A^2 + \sigma_D^2 + \sigma_E^2 \\ \mathrm{Cov}_{MZ} &= \sigma_A^2 + \sigma_D^2 \\ \mathrm{Cov}_{DZ} &= \frac{1}{2}\sigma_A^2 + \frac{1}{4}\sigma_D^2 \end{aligned}$$

where any environmental correlations have been ignored. It is a simple matter to estimate the three variance components, and hence heritability, by the method of moments from these three equations although maximum likelihood methods are used in practice.

The SNP profiles of individuals in a study can be used to estimate the *actual* inbreeding and coancestry coefficients. These, in turn, lead to estimates of the additive genetic variance and hence the heritability of a complex trait. Although heritability is a statistical construct, depending on allele frequencies in the study population, rather than a biological quantity, it is of interest. The heritability explained by markers found to be associated with a trait or response variable can be compared to prior values based on family pedigrees in order to check on the completeness of the genetic study. There has been discussion of "missing heritability" when the genetic estimates are less than prior values ([20] and references therein). This chapter will conclude with a discussion of estimating inbreeding and relatedness.

7.7 Estimation of Actual Inbreeding

Individual-specific inbreeding coefficients F can be estimated under the assumption that all loci have the same coefficient, interpreted as the probability of identity by descent (ibd). Many loci are needed. At locus l, write p_l for the frequency of A_l and code the genotypes A_lA_l, A_la_l, a_la_l as $X_l = 2, 1, 0$. These coding variables have the properties $\mathcal{E}(X_l) = 2p_l, \mathrm{Var}(X_l) = 2p_l(1 - p_l)(1 + F)$.

Then one moment estimator is formed by summing over loci $l, l = 1, 2, \ldots L$:

$$\hat{F}_1 = \frac{1}{L}\sum_{l=1}^{L} \frac{(X_l - 2p_l)^2}{2p_l(1 - p_l)} - 1$$

Another one is

$$\hat{F}_2 = \frac{1}{L}\sum_{l=1}^{L} \frac{X_l^2 - (1 + 2p_l)X_l + 2p_l^2}{2p_l(1 - p_l)}$$

If the p_l are known, both these are unbiased. The second one has a smaller variance.

The variances can be reduced by an alternative weighting over loci:

$$\hat{F}_1^a = \frac{\sum_{l=1}^{L}(X_l - 2p_l)^2}{\sum_{l=1}^{L} 2p_l(1-p_l)} - 1$$

$$\hat{F}_2^a = \frac{\sum_{l=1}^{L}[X_l^2 - (1+2p_l)X_l + 2p_l^2]}{\sum_{l=1}^{L} 2p_l(1-p_l)}$$

To avoid having to choose among different moment estimates, and to reduce variance, it may be preferable to use maximum likelihood estimation. An iterative method makes use of Bayes' theorem. If F represents the probability the individual in question has two ibd alleles at a locus, i.e. is inbred at that locus,

$$\begin{aligned} \Pr(A_lA_l|\text{inbred}) &= p_l, \Pr(A_lA_l|\text{Not inbred}) = p_l^2 \\ \Pr(A_la_l|\text{inbred}) &= 0, \Pr(A_la_l|\text{Not inbred}) = 2p_l(1-p_l) \\ \Pr(a_la_l|\text{inbred}) &= 1-p_l, \Pr(a_la_l|\text{Not inbred}) = (1-p_l)^2 \end{aligned}$$

From Bayes' theorem then

$$\begin{aligned} \Pr(\text{inbred}|A_lA_l) &= \frac{\Pr(A_lA_l|\text{inbred})\Pr(\text{inbred})}{\Pr(A_lA_l)} = \frac{F}{F + p_l(1-F)} \\ \Pr(\text{inbred}|A_la_l) &= 0 \\ \Pr(\text{inbred}|a_la_l) &= \frac{F}{F + (1-p_l)(1-F)} \end{aligned}$$

This suggests an iterative scheme: assign an initial value to F, and then average the updated values over loci. If G_l is the genotype at locus l, the updated value F' is

$$F' = \frac{1}{L}\sum_{l=1}^{L}\Pr(\text{inbred}|G_l)$$

This value is then substituted into the right-hand side and the process continues until convergence.

7.8 Estimation of Actual Relatedness

A moment estimate makes use of observed identity in state (ibs), as shown in Table 11. If N_0, N_1, N_2 are the number of loci in ibs state $i; i = 0, 1, 2$ then

$$\Pr(\text{ibs} = 0) = \Pr(\text{ibs} = 0|\text{ibd} = 0)\Pr(\text{ibd} = 0)$$

Table 11 Identity in state categories for two individuals

ibs state	Genotypes	Probability[a]
2	(AA, AA), (aa, aa), (Aa, Aa)	$(p^2+q^2)^2k_0 + k_1(p^3+pq+q^3) + k_2$
1	(AA, Aa), (Aa, AA), (aa, Aa), (Aa, aa)	$4pq(p^2+q^2)k_0 + 2pqk_1$
0	(AA, aa), (aa, AA)	$2p^2q^2k_0$

[a] $q = 1 - p$

summed over loci to provide

$$N_0 = \Pr(\text{ibd} = 0) \sum_l 2p_l^2(1-p_l)^2$$

leads to a moment estimate

$$\Pr(\text{ibd} = 0) = \frac{N_0}{\sum_l 2p_l^2(1-p_l)^2}$$

From

$$\Pr(\text{ibd} = 1) = \Pr(\text{ibs=1}|\text{ibd} = 0)\Pr(\text{ibd} = 0) + \Pr(\text{ibs=1}|\text{ibd} = 1)\Pr(\text{ibd} = 1)$$

summed over loci to provide

$$N_1 = \Pr(\text{ibd} = 0) \sum_l 4p_l(1-p_l)[p_l^2 + (1-p_l)^2] + \Pr(\text{ibd} = 1) \sum_l 2p_l(1-p_l)$$

a moment estimate of k_1 is obtained. Use is made of the previously estimated k_0:

$$\Pr(\text{ibd} = 1) = \frac{N_1 - \sum_l 4p_l(1-p_l)[p_l^2 + (1-p_l)^2]\Pr(\text{ibd} = 0)}{\sum_l 2p_l(1-p_l)}$$

The remaining coefficient k_2 is found from the result $k_0 + k_1 + k_2 = 1$. In practice, this method is not robust to small allele frequencies and it can return invalid estimates.

A moment estimator for the coancestry $\theta_{jj'}$ between individuals j and j', rather than the three k's is:

$$\hat{\theta}_{jj'} = \frac{1}{L} \sum_{l=1}^{L} \frac{(X_{lj} - 2p_l)(X_{lk} - 2p_l)}{2p_l(1-p_l)}$$

where $X_{lj}, X_{lj'}$ are $2, 1, 0$ if j, j' are (AA, Aa, aa), respectively, at locus l. An alternative way of combining over loci is

$$\hat{\theta}_{jj'a} = \frac{\sum_{l=1}^{L}[(X_{lj} - 2p_l)(X_{lj'} - 2p_l)]}{\sum_{l=1}^{L}[2p_l(1-p_l)]}$$

These are both unbiased but they have different variances.

An iterative procedure for maximum likelihood estimation of relatedness is analogous to that for the inbreeding coefficient, and it uses all six distinct pairs of genotypes shown in Table 9 (combining pairs of rows with the same probabilities) with probabilities depending on allele frequencies for that SNP and on a set of three k parameters that are assumed to be the same for all SNPs.

If S is the observed pair of genotypes, Table 9 provides the conditional probabilities $\Pr(S|D_i)$ where the D_i represent the identity states (the relationship). The probability of ibd state D_i is k_i. An iterative algorithm for estimating the k's from observed genotypes S_l at SNP l is based on Bayes' theorem for the probability of descent state $D_i, i = 0, 1, 2$:

$$\Pr(D_i|S_l) = \frac{\Pr(S_l|D_i)\Pr(D_i)}{\Pr(S_l)}$$

The procedure begins with initial estimates of the $k_i = \Pr(D_i)$. The denominator is calculated from the law of total probability by adding over the three descent states:

$$\Pr(S_l) = \sum_i \Pr(S_l|D_i)\Pr(D_i) = \sum_i \Pr(S_l|D_i)k_i$$

The updated estimates are obtained by averaging over L loci:

$$k_i' = \frac{1}{L}\sum_{l=1}^{L}\left(\frac{\Pr(S_l|D_i)k_i}{\sum_j \Pr(S_l|D_j)k_j}\right), \quad i = 0, 1, 2$$

These updated values are then substituted into the right-hand side and the process continued until the likelihood no longer changes (or changes by less than some specified small amount) where

$$\text{Likelihood} = \prod_{l=1}^{L}\left[\sum_i \Pr(S_l|D_i)k_i\right]$$

It will be better to monitor changes in the log-likelihood.

8 Discussion

The current availability of dense sets of marker SNPs for the human genome is having a large impact on genetic studies and offers new possibilities for clinical trials. This chapter offers a unified basis for the analysis of marker and response data, emphasizing the central importance of linkage disequilibrium between marker locus and the genes that affect response. It is convenient to phrase the development of association mapping in the language of quantitative genetics, using additive and nonadditive components of variance.

A novel feature of dense SNP data is that good estimates can be made of actual inbreeding and relatedness. These estimates are more relevant than values predicted from family pedigree, and all that are available in the absence of family data.

In biomedical research genetic markers can be used both to infer causes of disease and to identify treatments that are tailored to the individual. However, the dimensionality of genomic markers has challenged us to develop new methods that are appropriate for a large number of statistical comparisons and to develop computational methods that allow high-dimensional regression. In the broader context, the use of biological annotation is also essential for both viewing the relevance of empirical associations and to structure analysis in order to focus on those markers with the highest expectation for association with the outcomes under study.

The continued expansion of molecular technologies will challenge the biostatistical community to develop appropriate methodology so that reliable conclusions can be obtained from new measurements. Ultimately, meaningful collaboration between quantitative scientists and biomedical investigators will lead to the understanding of the mechanisms leading to disease onset, progression, and response to treatment.

Acknowledgements This work was supported in part by NIH grants GM 075091, HG 004464, HG 005157, HL 072966 and TR 000423.

References

1. Bansal V, Libiger O, Torkamani A, Schork NJ (2010) Statistical analysis strategies for association studies involving rare variants. Nature Reviews Genetics 11:773–785
2. Burton PR, Clayton DG, Cardon LR, Craddock N, Deloukas P et al (2007) Genome-wide association stidy of 14,000 cases of seven common diseases and 3,000 shared controls. Natire 447:661–676.
3. Cornelis MC, Agrawal A, Cole JW, Hansel NN, Barnes KC, Beaty TH, Bennett SN, Bierut LJ, Boerwinkle E, Doheny KF, Feenstra B, Feingold E, Fornage M, Haiman CA, Harris EL, Hayes MG, Heit JA, Hu FB, Kang JH, Laurie CC, Ling H, Manolio TA, Marazita ML, Mathias RA, Mirel DB, Paschall J, Pasquale LR, Pugh EW, Rice JP, Udren J, van Dam RM, Wang X, Wiggs JL, Williams K, Yu K (2010) The gene, environment association studies consortium (GENEVA): maximizing the knowledge obtained from GWAS by collaboration across studies of multiple conditions. Genet Epidemiol 34:364–372
4. Dahlman I, Eaves IA, Kosoy R et al (2002) Parameters for reliable results in genetic association studies in common disease. Nat Genet 30:149–150
5. Daly AK (2010) Genome-wide association studies in pharmacogenomics. Nat Rev Genet 11:241–246
6. Gunter L, Zhu J, Murphy S (2007) Variable selection for optimal decision making. Artif Intell Med 4594:149–154
7. Hastie T, Tibshirani R (1993) Varying coefficient models. J R Stat Soc Series B 55:757–796
8. Hill WG, Goddard ME, Visscher PM (2008) Data and theory point to mainly additive genetic variance for complex traits. PLoS Genet 4:e1000008
9. Hoerl AE (1962) Application of ridge analysis to regression problems. Chem Eng Prog 58: 54–59
10. Jarvik JG et al (2009). Surgery versus non-surgical therapy for carpal tunnel syndrome: a randomized parallel group trial. Lancet 374:1074–1081

11. Kallmes DF, Comstock B, Heagerty PJ et al (2009) A randomized clinical trial for vertoblasty for osteoporotic compression fractures. N Engl J Med 361:569–579
12. Laurie CC, Doheny KF, Mirel DB, Pugh EW, Bierut LJ, Bhangale T, Boehm F, Caporaso NE, Cornelis MC, Edenberg HJ, Gabriel SB, Harris EL, Hu FB, Jacobs K, Kraft P, Landi MT, Lumley T, Manolio TA, McHugh C, Painter I, Paschall J, Rice JP, Rice KM, Zheng X, Weir BS for the GENEVA Investigators (2010) Quality control and quality assurance in genotypic data for genome-wide association studies. Genet Epidemiol 34:591–602
13. Nyholt DR (2004) A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. Am J Hum Genet 74:765–769
14. Pritchard JK, Donnelly P (2001) Case? control studies of association in structured or admixed populations. Theor Popul Biol 60:227–237
15. Sitlani C, Heagerty PJ, Longitudinal structural mixed models as tools for characterizing the accuracy of markers used to select treatment (submitted)
16. Tibshirani R (1996) Regression shrinkage and selection via the lasso. J R Stat Soc Series B 58:267–288
17. Wakefield JC (2009) Bayes factors for genome-wide association studies: comparison with P-values. Genet Epidemiol 33:79–86
18. Wu TT, Chen YF, Hastie T, Sobel E, Lange K (2009) Genome-wide association analysis by lasso penalized logistic regression. Bioinformatics 25:714–721
19. Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. J Roy Stat Soc Series B 67:301–320.
20. Zuk O, Hechter E, Sunyaev SR, Lander ES (2012) The mystery of missing heritability: genetic interactions create phantom heritability. Proc Natl Acad Sci USA, vol 109, pp 1193–1198