Published in final edited form as:

Genome. 2010 November ; 53(11): 869–875. doi:10.1139/G10-062.

Statistical genetic issues for genome-wide association studies¹

Bruce S. Weir

Department of Biostatistics, University of Washington, Box 357232, Seattle, WA 98195-7232, USA

Bruce S. Weir: bsweir@u.washington.edu

Abstract

Genotyping technology now allows the rapid and affordable generation of million-SNP profiles for humans, leading to considerable activity in association mapping. Similar activity is anticipated for many plant species, including *Brassica*. These plant association mapping activities will require the same care in quality control and quality assurance as for humans. The subsequent analyses may draw upon the same body of theory that is described here in the language of quantitative genetics.

Keywords

linkage disequilibrium; case-control; quantitative trait; additive variance; dominance variance; trend test

Introduction

In 1969, forty years ago, I worked with plant geneticist R.W. Allard on some plant data generated by the new technique of electrophoresis (Allard et al. 1972). Allard's laboratory had generated data at four distinct esterase allozyme loci for several generations of an experimental population of barley. The numbers of individual plants genotyped were in the thousands, but our pride in having four molecular markers seems quaint in the current climate of whole-genome association studies (GWAS). My colleagues and I have recently written about our work in preparing data for posting on the database of Genotype and Phenotype (dbGaP) of sets of one million single nucleotide polymorphisms (SNPs) typed on tens of thousands of humans (Laurie et al. 2010). Gore et al. (2009) have reported over two million SNPs among 27 inbred lines of maize. The potential for genetic improvement of crop plants from such whole-genome data set is enormous, but it comes with the cost of new statistical genetic challenges. These challenges range from the mechanical issues of managing large data sets to the genetic issues of having to allow for dependencies among genetic markers. In this discussion I will show, however, that the data allow new insights into the genomic architecture as well as the genetic architecture of complex traits. Although whole-genome sequence data are not yet economically feasible for plant species (but see

¹This article is one of a selection of papers from the conference "Exploiting Genome-wide Association in Oilseed Brassicas: a model for genetic improvement of major OECD crops for sustainable farming".

Corresponding Editor: B. Golding.

Varshney et al. 2009), we already have enough information to address some very long-standing problems in population and quantitative genetics.

GWAS data

Generating the data

A comparison of genomic DNA sequences in different individuals reveals some positions at which two, or in some cases more than two, bases can occur. These SNPs are highly abundant and they have been described as the “ultimate genetic marker” by Duran et al. (2008). Among the many excellent reviews of the technology being used to generate whole-genome sets of SNP genotypes is that of Syvänen (2001). Information on over 12 million SNPs for *Homo sapiens* is contained in the public dbSNP (available at <http://ncbi.nlm.nih.gov/projects/SNP/>), along with information for nine other species. For rice, barley, and *Brassica* there is the autoSNPdb (available at <http://autosnpdb.qfab.org.au>) described by Duran et al. (2008) that draws data from GenBank (available at <http://ncbi.nlm.nih.gov/Genbank/>). The success of GWAS has been described by The Wellcome Trust Case Control Consortium (2007) and many other authors.

Managing the data

We have found it convenient to store SNP data in the network Common Data Form (netCDF). NetCDF is a set of software libraries and machine-independent data formats, designed specifically for large sets of array-oriented scientific data, and its use is widespread in those sciences that require very large data sets. It is maintained by the Unidata program at the University Corporation for Atmospheric Research (available at <http://unidata.ucar.edu/software/netcdf/>). This portable binary file format is convenient since it allows for efficient multidimensional arrayed data. The key feature is the speed with which data can be accessed from the file, as this is not a trivial consideration when there are a million SNP genotypes for sample of several thousand individuals. Function documentation can be found at <http://cran.r-project.org/web/packages/ncdf/ncdf.pdf> and a tutorial that details the steps to create netCDF files for raw SNP intensity files and called SNP genotype files is under preparation by my GENEVA colleagues (<http://genevastudy.org>). This tutorial describes a set of R programs that provide an alternative GWAS analysis platform to the very useful PLINK package (Purcell et al. 2007).

Cleaning the data

The production of GWAS data involves industrial-scale processes for which systems of quality assurance and quality control are essential (Laurie et al. 2010). Many association mapping methods involve comparisons of SNP allele frequencies among classes of individuals. Even small artificial differences in allele frequencies can generate false-positive results and there have been several accounts of appropriate quality control/quality assurance (QA/QC) methods (Broman 1999; Chanock et al. 2007; The Wellcome Trust Case Control Consortium 2007; Manolio et al. 2007; Miyagawa et al. 2008).

Sample cleaning—In any large study there is the possibility of miss-annotation of individuals. For human studies these include wrongly identified gender, race and (or)

ethnicity, and family membership. Similar problems can occur with plants, and Yu et al. (2006) referred to possible discrepancies between pedigree records and genetically determined relationships. For human and plant studies the genetic markers themselves allow population or family outliers to be identified.

A search for population outliers and possible miss-annotations can be accomplished with principal components analysis (PCA) where the million or so SNPs can be reduced down to a small number of components and individuals plotted in two-dimensional spaces for pairs of the first few components. Methodology in the GWAS context has been described by Patterson et al. (2006) and Price et al. (2006) and their EIGENSOFT software is available at <http://genepath.med.harvard.edu/~reich/Software.htm>. An interesting finding from such analyses is that PCA can detect features of the genome: Novembre et al. (2008) reported that a principal component in their study detected polymorphism for an inversion on human chromosome 8. Similar results may be anticipated in plant populations and they raise the issue of how to select SNPs to include in PCA. It may or may not be desirable to have PCA reflect genomic structure.

High-density SNP data are also leading to an appreciation of the prevalence of aneuploidy in humans. Peiffer et al. (2006) and Maciejewski and Mufti (2008) have shown how the SNP intensity signals, and inferred copy number and heterozygosity levels, can identify large portions of a chromosome that do not occur in the diploid state. Similar analyses will be of importance in *Brassica* studies when there are ploidy issues. It is prudent to remove aberrant chromosomes from association analyses.

Finally, samples with low quantity DNA and (or) high missing SNP genotype rates are generally excluded from further analyses.

SNP cleaning—Even if a sample passes QA/QC checks, it is unlikely that all SNPs will have been genotyped with high quality. A SNP with a high missing rate over all samples is unlikely to be used for association mapping, any more than would be a SNP with a low concordance rate among duplicate samples or with high “Mendelian error” rate (where genotypes are not consistent with known pedigrees). SNPs that have outlying departures from Hardy–Weinberg equilibrium may not have been genotyped correctly and are often found to have substantial excesses or deficiencies of heterozygosity.

In these various SNP tests there must be attention paid to multiple testing issues. A simple Bonferroni correction, such as rejecting Hardy–Weinberg equilibrium for a SNP in a one-million SNP panel if the P value is less than 5×10^{-8} is unlikely to be as informative as rejecting for those SNPs with a P value that deviates from expectation in a Q-Q probability plot (Weir et al. 2004).

A final QA/QC check on SNPs is conducted after association tests are conducted. It is useful to examine the “cluster plots” that display all individuals, colour-coded by assigned genotype, on a grid with the numbers of the two alleles as axes. There should be tight clusters at (2,0), (0,2), and (1,1) for the two homozygotes and the heterozygote.

Association mapping

A valuable review of association mapping in crop plants was given by Sorkheh et al. (2008). The following treatment allows for continuous or discrete traits and uses the classical language of quantitative genetics.

Allelic association

Whether locating the genes affecting economic traits is regarded as the first step in determining the genetic basis for the traits or as the basis for marker-assisted selection, it is helpful to cast the activity in terms of linkage disequilibrium between marker and trait loci. If SNP marker **M** has alleles M, m with frequencies p_M, p_m and a trait locus **T** has alleles T, t with frequencies p_T, p_t then for outbreeding random-mating populations the joint genotype frequencies can be set out as in Table 1. Each genotype probability is the product of gamete probabilities and these, in turn, can be expressed in terms of allelic frequencies and linkage disequilibrium D_{MT} as shown in Table 2. The most useful measure of association between marker and trait alleles is the correlation coefficient ρ_{MT} defined as

$$\rho_{MT} = \frac{D_{MT}}{\sqrt{p_M p_m p_T p_t}}$$

If variable x_M (or y_T) is defined for gametes to take the value 1 if the gamete carries allele M (or T) and 0 if the gamete carries allele m (or t), then D_{MT} is the covariance of x_M and y_T and ρ_{MT} is the correlation of these two variables.

Trait association

The contribution of locus **T** to the trait of interest can be indicated by variable G and, still assuming random mating, the genetic mean and variance are

$$\begin{aligned} \mu_G &= p_T^2 G_{TT} + 2p_T p_t G_{Tt} + p_t^2 G_{tt} \\ \text{Var}(G) &= p_T^2 G_{TT}^2 + 2p_T p_t G_{Tt}^2 + p_t^2 G_{tt}^2 - \mu_G^2 \end{aligned}$$

and the variance can be partitioned into additive and dominance components,

$\text{Var}(G) = \sigma_{A_T}^2 + \sigma_{D_T}^2$ where

$$\begin{aligned} \sigma_{A_T}^2 &= 2p_T p_t [p_T (G_{TT} - G_{Tt}) + p_t (G_{Tt} - G_{tt})]^2 \\ \sigma_{D_T}^2 &= p_T^2 p_t^2 (G_{TT} - 2G_{Tt} + G_{tt})^2 \end{aligned}$$

For a marker locus **M** with alleles M, m we can assign values X_{MM} , X_{Mm} , and X_{mm} , and have a completely analogous decomposition of the variance of X over individuals in a random mating population. Unless the marker and trait loci are the same, i.e., **M** is a casual SNP, the nature and location of **T** are not known and the values of G are not known. The marker and its variable, by contrast, are known.

Association mapping rests on the result (Weir 2008)

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

Continuous traits—A simple approach for continuous traits is to suppose that trait values Y may be written as the sums of genetic and nongenetic components, $Y = G + E$. If these components are independent and if E has a mean of zero then $\text{Cov}(X, Y) = \text{Cov}(X, G)$. If trait values are regressed on the marker genotype values X , the regression coefficient is

$$\beta_{YX} = \frac{\text{Cov}(X, Y)}{\text{Var}(X)} = \frac{\rho_{MT} \sigma_{A_T} \sigma_{A_M} + \rho_{MT}^2 \sigma_{D_T} \sigma_{D_M}}{\sigma_{A_M}^2 + \sigma_{D_M}^2}$$

This reduces to $\beta_{YX} = \rho_{MT} \sigma_{A_T} / (2p_M p_m)^{1/2}$ for additive marker scoring such as $X_{MM} = 2$, $X_{Mm} = 1$, and $X_{mm} = 0$. In that case, $\sigma_{D_T}^2 = 0$. If, instead, the marker is scored as $X_{MM} = p_m$, $X_{Mm} = 0$, and $X_{mm} = p_M$, making $\sigma_{A_T}^2 = 0$, then it is only the nonadditive trait variance that influences the regression: $\beta_{YX} = \rho_{MT}^2 \sigma_{D_T} / p_M p_m$. For any scoring of the marker genotypes, a significant regression coefficient for phenotype on marker genotype implies a significant linkage disequilibrium ρ_{MT} between marker and disease loci and there is an implication of physical proximity of these loci (Myles et al. 2009). The extent to which additive and nonadditive trait genetic effects contribute to the regression depends on the choice of marker scaling.

An alternative approach is to work with the correlation ρ_{XY} of X and Y . Note that the variance of trait values is $\text{Var}(Y) = \text{Var}(G) + V_E$. For an additive marker variable

$$\rho_{XY} = \rho_{MT} h_Y^T$$

where $(h_Y^T)^2 = \sigma_{A_T}^2 / (\sigma_{A_T}^2 + \sigma_{D_T}^2 + V_E)$ is the heritability of trait Y owing to locus T .

Nonadditive marker scoring provides that $\rho_{XY} \propto \rho_{MT}^2$. Sample values r_{XY} for the correlation ρ_{XY} can be transformed to normal variables with Fisher's transformation

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

and the hypothesis $\rho_{XY} = 0$, implying that $\rho_{MT} = 0$, against the hypothesis $\rho_{XY} > 0$ is rejected with significance level α if $z > z_{1-\alpha}$. Here $z_{1-\alpha}$ is the $(1 - \alpha)$ percentile of the standard normal distribution. Standard theory for correlation coefficients provides that the sample size n needed for this test to have $(1 - \beta) \%$ power is (approximately)

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

For 90% power and 1% significance level, $z_\beta = -1.28$ and $z_\alpha = -2.33$. For a SNP with $\rho_{MT}^2 = 0.8$ to the disease gene and a trait with heritability $(h_Y^T)^2 = 0.2$ so that $\rho_{XY} = 0.4$ this sample size is about 73. Multiple-testing considerations for one-million SNP sets suggest that a significance level of about 10^{-8} may be appropriate and then the sample size increases to over 250. These calculations assume a single trait locus, although plant quantitative traits are more likely to be affected by multiple loci and the heritability associated with any one locus will be reduced. If heritability per locus is proportional to the number of loci affecting the trait, then so too will be the sample size needed to detect association of a marker to each locus. Yang et al. (2010) make the useful distinction between testing for significant marker-trait associations and estimating the combined effects of all markers on a trait. For human height they showed that most of the heritability for human height is accounted for when 300 000 SNPs are used in an analysis even though individual SNP effects are generally too small to pass stringent significance tests.

Both the linear regression and correlation approaches can provide single df tests for marker-trait association. Another test follows from analysis of variance of trait values among the three marker genotypic classes. The three trait means are

$$\begin{aligned} \varepsilon(Y|MM) &= (P_{MT}^2 G_{TT} + 2P_{MT} P_{Mt} G_{Tt} + P_{Mt}^2 G_{tt}) / p_M^2 \\ &= \mu_G + (p_M \rho_{MT} A + \rho_{MT}^2 D) / p_M^2 \\ \varepsilon(Y|Mm) &= \mu_G + [(p_m - p_M) \rho_{MT} A - 2\rho_{MT}^2 D] / 2p_M p_m \\ \varepsilon(Y|mm) &= \mu_G - (p_m \rho_{MT} A - \rho_{MT}^2 D) / p_m^2 \end{aligned}$$

where $A = \sigma_{A_T} \sqrt{2p_M p_m}$, $D = \sigma_{D_T} p_M p_m$, and μ_G is the trait mean. An analysis of variance F test will therefore also test that $\rho_{MT} = 0$ and the test will be affected by both additive and dominance effects at the trait locus.

Binary traits—If, however, the trait is binary (presence or absence) the trait genetic value can be interpreted as the probability of an individual having the trait (i.e., being a “case”), and marker genotype frequencies can be conditioned on trait status. Then the random mating assumption leads to

$$\begin{aligned} \Pr(MM|\text{case}) &= p_M^2 + (p_M \rho_{MT} A + \rho_{MT}^2 D) / \mu_G \\ \Pr(Mm|\text{case}) &= 2p_M p_m + [(p_m - p_M) \rho_{MT} A - 2\rho_{MT}^2 D] / \mu_G \\ \Pr(mm|\text{case}) &= p_m^2 - (p_m \rho_{MT} A - \rho_{MT}^2 D) / \mu_G \end{aligned}$$

where now μ_G is the trait prevalence or the probability that a random individual has the trait. For individuals not having the trait (i.e., being a “control”)

$$\begin{aligned}\Pr(MM|\text{control}) &= p_M^2 - (p_M \rho_{MT} A + \rho_{MT}^2 D) / (1 - \mu_G) \\ \Pr(Mm|\text{control}) &= 2p_M \rho_m - [(p_m - p_M) \rho_{MT} A - 2\rho_{MT}^2 D] / (1 - \mu_G) \\ \Pr(mm|\text{control}) &= p_m^2 + (p_m \rho_{MT} A - \rho_{MT}^2 D) / (1 - \mu_G)\end{aligned}$$

These expressions lead to case-control test statistics.

For example, allelic data may be set out in a 2×2 table of case/control versus M/m as in Table 3. A contingency-table test statistic is

$$X^2 = \frac{(n_{Ma}n_{mo} - n_{ma}n_{Mo})^2}{n_M n_m n_a n_o}$$

Under the null hypothesis of no marker-trait association, X^2 is distributed as chi-square with one df. If there is association, the noncentrality parameter is

$$\lambda = \frac{n \rho_{MT}^2 \sigma_{A_T}^2}{\mu_G (1 - \mu_G)}$$

This allows sample size calculations: if the marker-causal gene linkage disequilibrium is $\rho_M^2 = 0.8$, the probability that a random individual has the trait is $\mu_G = 0.1$; the causal allele frequencies are $p_T = 0.1$ and $p_t = 0.9$; and the probabilities that individuals with genotypes TT, Tt , and tt have the trait are 0.82, 0.42, and 0.02, respectively, where $\sigma_{A_T}^2 = 0.0288$ and $n \approx 3\lambda$. For a 1% significance level and a 90% power, n is about 40 and this rises to 140 for a significance level of 10^{-8} . There are occasions when single markers are associated with genes having such substantial effects on a human disease: the two SNPs reported for macular degeneration (Klein et al. 2005) were found in a case-control study with 96 cases and 50 controls. The more usual situation, however, is that much smaller effects are associated with each SNP: the top SNPs reported for seven common diseases by The Wellcome Trust Case Control Consortium (2007) had odds ratios typically around 1.5 and $\sigma_{A_T}^2$ would then drop substantially. It is generally recognized that sample sizes in the thousands are necessary for risk ratios 1.5 (Risch and Merikangas 1996).

Effects of inbreeding

Plant populations often have some degree of inbreeding and this means that reducing marker genotypes to alleles can be misleading. To illustrate the consequences for association mapping, we extend the discussion of case-control tests to that of trend tests. The usual notation (Sasieni 1997) is shown in Table 4. With that notation, the allelic case-control test statistic is

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

and the genotypic linear trend test statistic is

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

The trend test statistic can be regarded as the correlation between marker genotype (coded linearly as 2, 1, and 0 for MM , Mm , and mm), and case status (coded as 1, 0 for case, control).

Now suppose the population is inbred to an extent f so that $P_{MM} = P_M^2 + f p_M p_m$, $P_{Mm} = 2(1 - f)p_M p_m$, and $P_{mm} = p_m^2 + p_M p_m f$. Under the hypothesis of no association, $\rho_{MT} = 0$, the test statistics should have an expected value of 1 and indeed $E(X_T^2) \approx 1$, but $E(X_A^2) \approx (1+f)$. This confirms that the trend test, unlike the allelic case-control test is robust to departures from Hardy–Weinberg equilibrium. Note that f or F_{IS} is the within-population inbreeding coefficient.

These last results, and those in the next section, are most easily derived by expressing the test statistics as functions of sample allele frequencies. The numerator of the allelic case-

control test statistic X_A^2 can be written as $8NR^2S^2(\tilde{p}_{M,case} - \tilde{p}_{M,control})^2$ where sample values are indicated by tildes. The denominator of X_A^2 can be written as

$4N^2SR\tilde{p}_M(1 - \tilde{p}_M)$ where $\tilde{p}_M$ is the sample frequency of allele M in the total sample of size N . Under the null hypothesis that the marker allele has the same frequency p_M in cases and controls, and under the assumption that all individuals in the population are independent, the numerator has expectation $4N^2SRp_M(1 - p_M)(1 + f)$ and the denominator has expectation $4N^2SRp_M(1 - p_M)[1 - (1 + f)/2N]$ or approximately $4N^2SRp_M(1 - p_M)$, and the expected value of X_A^2 is approximated by the ratio of the expected values of numerator and denominator: $E(X_A^2) \approx (1+f)$.

The numerator for the linear trend test statistic X_T^2 is half that for X_A^2 and therefore has the expectation $2N^2SRp_M(1 - p_M)(1 + f)$ under the null hypothesis. The denominator of X_T^2 can be written as $2N^2SR(\tilde{p}_M + \tilde{p}_{MM} - 2\tilde{p}_M^2)$ where $\tilde{p}_M$, $\tilde{p}_{MM}$ are the sample frequencies of allele M and MM homozygotes in the total sample of size N . Under the null hypothesis that the marker allele has the same frequency p_M in cases and controls, and under the assumption that all individuals in the population are independent, the denominator has expectation $2N^2SR[p_M + [p_M^2 + p_M(1 - p_M)f] - 2(p_M(1 - p_M)(1 + f)/(2) + p_M^2)]$ or approximately $2N^2SRp_M(1 - p_M)(1 + f)$. The ratio of numerator and denominator expected values provides an approximation to the expected value of X_T^2 .

Effects of population structure

Whether or not a population is inbred there may be substructure and this can inflate association test statistics (Devlin and Roeder 1999; Cardon and Palmer 2003; Wen et al. 2009). A discussion of the effects of population structure involves the total inbreeding

coefficient F or F_{IT} and the related quantity θ or F_{ST} . The change in perspective is most easily explained by referring to the variance of allele frequencies. In the previous section, the within-population inbreeding coefficient f allows the variance of a sample allele frequency $\tilde{p}_M$ in a sample of n individuals to be expressed as $p_M p_m (1+f)/2n$. This variance is over repeated samples from the same population. In a structured population, there is variation of allele frequencies among subpopulations as well as among samples within subpopulations. The total variance of a sample allele frequency accommodates both sources and is given by $p_M p_m [\theta + (1 + F - 2\theta)/(2n)]$.

If the i th subpopulation furnishes R_i cases and S_i controls, with $N_i = R_i + S_i$, then it is convenient to introduce $x_i = R_i/R$, $y_i = S_i/S$, $z_i = N_i/N$ as the proportions of cases, controls and all samples from the i th subpopulation. If it is not known to which subpopulation an individual belongs, the expected values of the association test statistics are

$$\begin{aligned}\varepsilon(Z_A^2) &\approx \frac{2RS\theta \sum_i (x_i - y_i)^2 + N(1+F-2\theta)}{N(1-\theta \sum_i z_i^2) - (1+F-2\theta)/2} \\ \varepsilon(Z_T^2) &\approx \frac{2RS\theta \sum_i (x_i - y_i)^2 + N(1+F-2\theta)}{N[(1+F)-2\theta \sum_i z_i^2]}\end{aligned}$$

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, further, there are many subpopulations, we can ignore the term $F \sum_i z_i^2$ in the denominator and then

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

The trend test does not offer protection against population structure, and the inflation of association tests increases the more a subpopulation contributes unequally to the case and control samples. With equal representation, $x_i = y_i$, the test statistics may actually be deflated.

Effects of relatedness

Individual plants or lines derived from crosses involving the same founders are related by virtue of sharing some of their genome — they have alleles identical by descent from these founders. Plants in a finite population necessarily have common ancestors even if there is random outcrossing, so any distinction between family and evolutionary relatedness is a little artificial. Nevertheless, it is usual to use “relatedness” for plants in the same family or pedigree. The availability of extensive marker data means that relatedness can be estimated in the absence of pedigree information. Even if pedigree information is available, however, there may be a case for using genetic estimates to guard against pedigree errors or to accommodate differences in actual relatedness along the genome whether these are due to chance or to the effects of natural or artificial selection.

A treatment of relatedness rests on the concept of identity by descent. Two noninbred diploid plants may have received identical alleles for neither, either, or both of their maternal and paternal genes. The probabilities of these three events are written as k_0 , k_1 , and k_2 and these sum to one. For full siblings, $k_0 = 1/4$, $k_1 = 1/2$, and $k_2 = 1/4$. Association testing rests on the kinship coefficient, deliberately written as θ to emphasize the continuity between evolutionary and family relatedness, with $\theta = k_2/2 + k_1/4$.

The allelic case-control test was introduced above as being a contingency table test on the 2×2 table of marker allelic state versus case-control state. Equivalently, it can be constructed by assuming that marker allele frequencies are normally distributed and working with the difference of case- and control-sample frequencies ($\tilde{p}_{M, \text{case}} - \tilde{p}_{M, \text{control}}$). Specifically, the test statistic is

$$X^2 = \frac{(\tilde{p}_{M, \text{case}} - \tilde{p}_{M, \text{control}})^2}{\text{Var}(\tilde{p}_{M, \text{case}}) - 2\text{Cov}(\tilde{p}_{M, \text{case}}, \tilde{p}_{M, \text{control}}) + \text{Var}(\tilde{p}_{M, \text{control}})}$$

The denominator depends on the total inbreeding coefficients F_i for all individuals i in the sample and the coancestry coefficients θ_{ij} for all pairs of individuals i, j . It consists of three sums, over the n_a cases, the n_o controls and the $n_a n_o$ case-control pairs:

$$\begin{aligned} \text{Var}(\tilde{p}_{M, \text{case}}) &= \frac{p_M(1-p_M)}{n_a^2} \left[\sum_{i=1}^{n_a} (1+F_i) + \sum_{i \neq i'} \theta_{ii'} \right] \\ \text{Var}(\tilde{p}_{M, \text{control}}) &= \frac{p_M(1-p_M)}{n_o^2} \left[\sum_{j=1}^{n_o} (1+F_j) + \sum_{j \neq j'} \theta_{jj'} \right] \\ \text{Cov}(\tilde{p}_{M, \text{case}}, \tilde{p}_{M, \text{control}}) &= \frac{p_M(1-p_M)}{n_a n_o} \sum_{i=1}^{n_a} \sum_{j=1}^{n_o} \theta_{ij} \end{aligned}$$

Bourgain et al. (2003) suggested using the F_i 's and θ_{ij} 's derived from known pedigrees, along with allele frequencies p_M estimated from the combined case and control sample. Yu et al. (2006) and Choi et al. (2009) suggested using the SNP genotypes to estimate the inbreeding and coancestry coefficients.

For noninbred individuals, $F_i = 0$, maximum-likelihood estimation of the three identity coefficients k_0, k_1 , and k_2 is straightforward, although it requires iteration. The procedure follows from being able to express the joint genotype probabilities for pairs of individuals in terms of the k 's and allele frequencies, with the coefficients being assumed the same over a set of independent SNPs. A sufficiently dense collection of SNPs would allow for differences in coefficients in different regions of the genome. If the individuals are inbred, however, a set of nine identity by descent coefficients is needed (Weir et al. 2006).

Variation in the identity coefficients and their estimates raise potential problems with incorporating relatedness coefficients into association mapping. Estimation relies on allele frequencies which, in turn, have to be estimated from the set of individuals under

consideration. This need to estimate allele frequencies introduces variation into relatedness estimates but a larger source of variation has to do with the inherent variation of the coefficients among loci. At any particular locus, of course, the individuals have only one identity state, so actual probabilities $\widehat{k}_0$, $\widehat{k}_1$, and $\widehat{k}_2$ are either one or zero and exactly one of the three equals one. For full siblings, the variances of the three binary quantities are 3/16, 1/4, and 3/16, respectively. When several loci are considered, the variances of the average actual identity coefficients depend on the recombination fractions between pairs of loci. If they are estimated from a very dense set of markers over a genome of with m chromosomes with maplengths l_i , $i = 1, 2, \dots, m$ Morgans and total maplength L , then the average actual coancestry coefficient for full siblings is (Visscher 2009)

$$\text{Var}(\hat{\theta}^{\text{FS}}) = \frac{1}{256L^2} (4L + \sum_i e^{-4l_i} - m)$$

This result assumes the Haldane mapping function. Visscher et al. (2006) have proposed using estimated coancestries for estimation of components of genetic variance and, hence, heritabilities of quantitative traits.

Discussion

Association mapping for all plants, including *Brassica*, is entering a phase where the number of genetic markers is not limiting factor, although a more pessimistic note is sounded by Ganai et al. (2009). We have come a long way from the isozyme studies of the 1970s and even candidate gene studies in the near future will have much more detailed information than those that use fewer than 10 markers (e.g., Krouchi et al. 2008). The lament of Rostoks et al. (2006) that “whole-genome association studies in crop plants, with the exception of rice, are currently limited by the number of markers available, their format, and cost” is no longer necessary. Recent authors (Myles et al. 2009) have stressed that, especially with the imminence of whole-genome sequence data, attention should be directed towards the design of association studies rather than the number or genomic spacing of markers. The observation of Stich et al. (2007) that association mapping is less expensive than linkage mapping because data routinely collected in plant breeding programs can be used remains pertinent.

As marker sets develop, genome-wide association studies in *Brassica* can be expected to proceed as they have in *Arabidopsis* (Aranzana et al. 2005), barley (Cockram et al. 2008), and maize (Beló et al. 2008). The special features of *Brassica* will have to be taken into account but the work already in place for identifying aneuploidy and accommodating population structure and relatedness in humans should provide a foundation.

Acknowledgments

This work was supported in part by NIH grant GM 075091 and by a travel grant from OECD.

References

- Allard RW, Kahler AL, Weir BS. The effect of selection on esterase allozymes in a barley population. *Genetics*. 1972; 72(3):489–503. [PubMed: 17248584]
- Aranzana MJ, Kim S, Zhao KY, Bakker E, Horton M, Jakob K, et al. Genome-wide association mapping in *Arabidopsis* identifies previously known flowering time and pathogen resistance genes. *PLoS Genet*. 2005; 1(5):e60.10.1371/journal.pgen.0010060 [PubMed: 16292355]
- Beló A, Zheng P, Luck S, Shen B, Meyer DJ, Li B, et al. Whole genome scan detects an allelic variant of *fad2* associated with increased oleic acid levels in maize. *Mol Genet Genomics*. 2008; 279(1):1–10.10.1007/s00438-007-0289-y [PubMed: 17934760]
- Bourgain C, Hoffjan S, Nicolae R, Newman D, Steiner L, Walker K, et al. Novel case-control test in a founder population identifies P-selectin as an atopy-susceptibility locus. *Am J Hum Genet*. 2003; 73(3):612–626.10.1086/378208 [PubMed: 12929084]
- Broman KW. Cleaning genotype data. *Genet Epidemiol*. 1999; 17(Suppl 1):S79–S83. [PubMed: 10597416]
- Cardon LR, Palmer LJ. Population stratification and spurious allelic association. *Lancet*. 2003; 361(9357):598–604.10.1016/S0140-6736(03)12520-2 [PubMed: 12598158]
- Chanock SJ, Manolio T, Boehnke M, Boerwinkle E, Hunter DJ, Thomas G, et al. Replicating genotype-phenotype associations. *Nature*. 2007; 447(7145):655–660.10.1038/447655a [PubMed: 17554299]
- Choi Y, Wijsman EM, Weir BS. Case-control association testing in the presence of unknown relationships. *Genet Epidemiol*. 2009; 33(8):668–678.10.1002/gepi.20418 [PubMed: 19333967]
- Cockram J, White J, Leigh FJ, Lea VJ, Chiapparino E, Laurie DA, et al. Association mapping of partitioning loci in barley. *BMC Genet*. 2008; 9(1):16.10.1186/1471-2156-9-16 [PubMed: 18282287]
- Devlin B, Roeder K. Genomic control for association studies. *Biometrics*. 1999; 55(4):997–1004.10.1111/j.0006-341X.1999.00997.x [PubMed: 11315092]
- Duran C, Appleby N, Clark T, Wood D, Imelfort M, Batley J, Edwards D. *AutoSNPdb*: an annotated single nucleotide polymorphism database for crop plants. *Nucleic Acids Res*. 2008; 37:D951–D953. Database issue. 10.1093/nar/gkn650 [PubMed: 18854357]
- Ganal MW, Altmann T, Röder MS. SNP identification in crop plants. *Curr Opin Plant Biol*. 2009; 12(2):211–217.10.1016/j.pbi.2008.12.009 [PubMed: 19186095]
- Gore MA, Chia JM, Elshire RJ, Sun Q, Ersoz ES, Hurwitz BL, et al. A first-generation haplotype map of maize. *Science*. 2009; 326(5956):1115–1117.10.1126/science.1177837 [PubMed: 19965431]
- Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, et al. Complement factor H polymorphism in age-related macular degeneration. *Science*. 2005; 308(5720):385–389.10.1126/science.1109557 [PubMed: 15761122]
- Krouchi F, Gustavsson S, Sjödin P, Kruskopf-Österberg M, Lagercrantz U, Lascoux M. Association between *COL1* and flowering time in *Brassica nigra*: replication validation, and genotypic disequilibrium. *Int J Plant Sci*. 2008; 169(9):1229–1237.10.1086/591989
- Laurie CC, Doheny KF, Mirel DB, Pugh EW, Bierut LJ, Bhangale T, et al. Quality control and quality assurance in genotypic data for genome-wide association studies. *Genet Epidemiol*. 2010; 34(6):591–602.10.1002/gepi.20516 [PubMed: 20718045]
- Maciejewski JP, Mufti GJ. Whole genome scanning as a cytogenetic tool in hematologic malignancies. *Blood*. 2008; 112(4):965–974.10.1182/blood-2008-02-130435 [PubMed: 18505780]
- Manolio TA, Rodriguez LL, Brooks L, Abecasis G, Ballinger D, Daly M, et al. New models of collaboration in genome-wide association studies: the Genetic Association Information Network. *Nat Genet*. 2007; 39(9):1045–1051.10.1038/ng2127 [PubMed: 17728769]
- Miyagawa T, Nishida N, Ohashi J, Kimura R, Fujimoto A, Kawashima M, et al. Appropriate data cleaning methods for genome-wide association study. *J Hum Genet*. 2008; 53(10):886–893.10.1007/s10038-008-0322-y [PubMed: 18695938]
- Myles S, Peiffer J, Brown PJ, Ersoz ES, Zhang Z, Costich DE, Buckler ES. Association mapping: critical considerations shift from genotyping to experimental design. *Plant Cell*. 2009; 21(8):2194–2202.10.1105/tpc.109.068437 [PubMed: 19654263]

- Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, et al. Genes mirror geography within Europe. *Nature*. 2008; 456(7218):98–101.10.1038/nature07331 [PubMed: 18758442]
- Patterson N, Price AL, Reich D. Population structure and eigenanalysis. *PLoS Genet*. 2006; 2(12): 2074–2093.10.1371/journal.pgen.0020190
- Peiffer DA, Le JM, Steemers FJ, Chang W, Jenniges T, Garcia F, et al. High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. *Genome Res*. 2006; 16(9):1136–1148.10.1101/gr.5402306 [PubMed: 16899659]
- Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*. 2006; 38(8): 904–909.10.1038/ng1847 [PubMed: 16862161]
- Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 2007; 81(3): 559–575.10.1086/519795 [PubMed: 17701901]
- Risch N, Merikangas K. The future of genetic studies of complex human diseases. *Science*. 1996; 273(5281):1516–1517.10.1126/science.273.5281.1516 [PubMed: 8801636]
- Rostoks N, Ramsay L, MacKenzie K, Cardle L, Bhat PR, Roose ML, et al. Recent history of artificial outcrossing facilitates whole-genome association mapping in elite inbred crop varieties. *Proc Natl Acad Sci USA*. 2006; 103(49):18656–18661.10.1073/pnas.0606133103 [PubMed: 17085595]
- Sasieni PD. From genotypes to genes: doubling the sample size. *Biometrics*. 1997; 53(4):1253–1261.10.2307/2533494 [PubMed: 9423247]
- Sorkheh K, Malysheva-Otto LV, Wirthensohn MG, Tarkesh-Esfahani S, Martínez-Gómez P. Linkage disequilibrium, genetic association mapping and gene localization in crop plants. *Genet Mol Biol*. 2008; 31(4):805–814.10.1590/S1415-47572008005000005
- Stich B, Melchinger AE, Piepho HP, Hamrit S, Schipprack W, Maurer HP, Reif JC. Potential causes of linkage disequilibrium in a European maize breeding program investigated with computer simulations. *Theor Appl Genet*. 2007; 115(4):529–536.10.1007/s00122-007-0586-1 [PubMed: 17598084]
- Syvänen A-C. Accessing genetic variation: genotyping single nucleotide polymorphisms. *Nat Rev Genet*. 2001; 2(12):930–942.10.1038/35103535 [PubMed: 11733746]
- The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*. 2007; 447(7145):661–678.10.1038/nature05911 [PubMed: 17554300]
- Varshney RK, Nayak SN, May GD, Jackson SA. Next-generation sequencing technologies and their implications for crop genetics and breeding. *Trends Biotechnol*. 2009; 27(9):522–530.10.1016/j.tibtech.2009.05.006 [PubMed: 19679362]
- Visscher PM. Whole genome approaches to quantitative genetics. *Genetica*. 2009; 136(2):351–358.10.1007/s10709-008-9301-7 [PubMed: 18668208]
- Visscher PM, Medland SE, Ferreira MA, Morley KI, Zhu G, Cornes BK, et al. Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *PLoS Genet*. 2006; 2(3):e41.10.1371/journal.pgen.0020041 [PubMed: 16565746]
- Weir BS. Linkage disequilibrium and association mapping. *Annu Rev Genomics Hum Genet*. 2008; 9(1):129–142.10.1146/annurev.genom.9.081307.164347 [PubMed: 18505378]
- Weir BS, Hill WG, Cardon LR. Allelic association patterns for a dense SNP map. *Genet Epidemiol*. 2004; 27(4):442–450.10.1002/gepi.20038 [PubMed: 15543640]
- Weir BS, Anderson AD, Hepler AB. Genetic relatedness analysis: modern data and new challenges. *Nat Rev Genet*. 2006; 7(10):771–780.10.1038/nrg1960 [PubMed: 16983373]
- Wen W, Mei H, Feng F, Yu S, Huang Z, Wu J, et al. Population structure and association mapping on chromosome 7 using a diverse panel of Chinese germplasm of rice (*Oryza sativa* L.). *Theor Appl Genet*. 2009; 119(3):459–470.10.1007/s00122-009-1052-z [PubMed: 19458933]
- Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al. Common SNPs explain a large proportion of the heritability for human height. *Nat Genet*. 2010; 42(7):565–569.10.1038/ng.608 [PubMed: 20562875]

Yu JM, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet.* 2006; 38(2):203–208.10.1038/ng1702 [PubMed: 16380716]

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 1

Marker–trait genotype probabilities.

	<i>TT</i>	<i>Tt</i>	<i>tt</i>	Total
<i>MM</i>	P_{MT}^2	$2P_{MT}P_{Mt}$	P_{Mt}^2	P_M^2
<i>Mm</i>	$2P_{MT}P_{mT}$	$2P_{MT}P_{mt}+2P_{Mt}P_{mT}$	$2P_{Mt}P_{mt}$	$2pMp_m$
<i>mm</i>	P_{mT}^2	$2P_{mT}P_{mt}$	P_{mt}^2	P_m^2
Total	P_T^2	$2pTp_t$	P_t^2	1

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 2

Marker–trait gamete probabilities.

	<i>T</i>	<i>t</i>	Total
<i>M</i>	$P_{MT}=p_Mp_T + D_{MT}$	$P_{Mt}=p_Mp_t - D_{MT}$	p_M
<i>m</i>	$P_{mT}=p_m p_T - D_{MT}$	$P_{mt}=p_m p_t + D_{MT}$	p_m
Total	p_T	p_t	1

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 3

Allelic case-control counts and probabilities.

		Case	Control	Total
<i>M</i>	Counts	n_{Ma}	n_{Mo}	n_M
	Probabilities	$p_M \mu_G + \frac{1}{2} \rho_{MT} A$	$p_M (1 - \mu_G) - \frac{1}{2} \rho_{MT} A$	p_M
<i>m</i>	Counts	n_{ma}	n_{mo}	n_m
	Probabilities	$p_m \mu_G - \frac{1}{2} \rho_{MT} A$	$p_m (1 - \mu_G) + \frac{1}{2} \rho_{MT} A$	p_m
Total	Counts	n_a	n_o	$2n$
	Probabilities	μ_G	$1 - \mu_G$	1

Table 4

Marker genotype counts in cases and controls.

Genotype	<i>MM</i>	<i>Mm</i>	<i>mm</i>	Total
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

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript